Artificial Intelligence in Public Employment Services: Machine Learning, Job Match Quality, and Fairness in Labour Market Matching

Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaft

eingereicht an der

Fakultät für Wirtschaftswissenschaften der Universität Regensburg

vorgelegt von Sabrina Mühlbauer

Berichterstatter:

Prof. Dr. Enzo Weber, Universität Regensburg
Prof. Dr. Rolf Tschernig, Universität Regensburg

Tag der Disputation: 30. September 2026

Acknowledgements

First and foremost, I would like to thank Prof. Enzo Weber for giving me the opportunity to pursue this dissertation under his supervision. Over the past years, I have had the privilege of developing as a researcher in economics in an environment that was supportive and collegial. Through numerous discussions and joint research projects, I benefited greatly from his guidance and was able to grow into academic research.

Moreover, I would like to thank my second supervisor, Prof. Rolf Tschernig. He sparked my interest in econometrics and statistics and played an important role in my academic development. I am grateful for the methodological foundation he provided and for showing me the path toward an academic career.

Furthermore, I would like to thank my colleagues at the Institute for Employment Research (IAB), particularly those in the Department of Macroeconomics and Forecasting. I was fortunate to work in an environment characterised by collegiality, openness, and a genuine willingness to support one another. Both professionally and personally, I have benefited enormously from this atmosphere. I am especially grateful to Hermann Gartner, who served as my mentor during my time in the graduate program. Our regular and fruitful discussions and his constant support helped me to find my way into academic research practice. I am also particularly grateful to Anja Bauer, Yasemin Yilmaz, Anja Warning, and Paula Ziethmann, whose collaboration, encouragement, and assistance greatly contributed to my research.

I would also like to acknowledge the Graduate Program (GradAB), which supported me as a doctoral fellow for more than four years and provided the foundation upon which this dissertation was built. In particular, I am grateful to Sandra Huber for her guidance, support, and excellent coordination throughout this period.

Last but certainly not least, I would like to thank my family, especially Markus and my parents. Your unwavering support, patience, and encouragement have been invaluable throughout this journey. You have always been there whenever help was needed – whether with practical matters or moral support. This dissertation would not have been possible without you. And now, should anyone ask again what your daughter actually does as a research assistant, you have a concrete answer.

Contents

Acknowledgements **2**

List of Figures **6**

List of Tables **7**

Introduction **10**

1 Machine Learning for Labor Market Matching **16**

1.1 Introduction 16
1.2 Related Literature 19
1.3 Setting and Data Structure 20
1.3.1 Defining Matching 21
1.3.2 Data Description 21
1.3.3 Endogenous Variable 22
1.3.4 Choosing Samples for Estimation 23
1.3.5 Exogenous Variables 24
1.4 Estimation Strategy 26
1.4.1 Preparation 26
1.4.2 OLS and Logit 28
1.4.3 Random Forest 29
1.4.4 K-Nearest Neighbours 30
1.4.5 Estimation Methods - Findings 31
1.5 Empirical Results 33
1.5.1 Sample Splitting by Year 34
1.5.2 Test-Train Split by Person 35
1.5.3 Additional Methods 36
1.5.4 An Illustrative Scenario: Hypothetical Effects of Using RF Instead of Logit 37
1.6 Robustness Checks 39
1.6.1 Check list of Suggestions 39

1.6.2 Extending the Number of Occupational Groups 39
1.6.3 Adding the Skill Level to the Job Categories 40
1.7 Limitations and Potential Biases 41
1.7.1 Discrimination 41
1.7.2 Limitations 42
1.8 Conclusion 42

Appendix - Chapter 1 **46**
1.A Number of observations per job category 46

2 Quality-Adjusted Job Recommendations: A Machine Learning Approach Using Administrative Data **52**
2.1 Introduction 52
2.2 Setting 54
2.2.1 Data 54
2.2.2 Defining Job Match Quality 55
2.2.3 Explanatory Variables 57
2.3 Empirical Strategy and Results 57
2.3.1 Prediction of Job Stability 57
2.3.2 Prediction of Wages 61
2.3.3 Data Structure and Implications for Estimation Methods 63
2.3.4 Alternative Machine Learning Estimation Methods 64
2.4 Creating a Matching Index 65
2.5 Linking the Matching Index to Labour Market Dynamics 69
2.6 Conclusion 71

Appendix - Chapter 2 **74**
2.A Imputation of Wages above the Contribution Limit 74

3 Distributional Intersectional Fairness in AI-Supported Job Matching: Evidence on Amplification, Trade-Offs, and Residual Risk **76**
3.1 Introduction 76
3.2 Background and Related Literature 78
3.2.1 Existing Strategies on Algorithmic Fairness in Employment Research 78
3.2.2 Intersectionality as an Analytical Framework for Understanding Inequality 79
3.2.3 Intersectional Analysis of Data and AI 81
3.3 Empirical Study: Fairness Metrics, Estimation, and Model Selection . . . 82
3.3.1 Data and Intersectional Setup 83
3.3.2 Prediction Framework 84

3.3.3 Distributional Fairness and Baseline Model Comparison 84
3.3.4 Predictive Fairness, Fairness-Aware Estimation, and Formal Metrics 88
3.3.5 Model Selection under Retained Utility Constraints 91
3.4 Residual Distributional Risk, Explainability, and Governance 93
3.4.1 Residual Distributional Risk and Flagged Occupations 93
3.4.2 Explainability, Human Oversight, and Jobseeker Participation . . 96
3.5 Counterfactual Averaging and Distributional Fairness 97
3.5.1 Aggregate Fairness Effects 97
3.5.2 Predictive Performance under Counterfactual Averaging 98
3.5.3 Occupation-Level Comparison for the Strongest Regularisation Setting 99
3.6 Discussion: Integrating Fairness Dimensions in AI-Supported Job Matching 101
3.6.1 Distributional and Predictive Fairness Do Not Coincide 101
3.6.2 Technical Mitigation Compresses Rather Than Eliminates Fairness Risk 102
3.6.3 Implications for Deployment, Explainability, and Participatory Design in PES 103
3.7 Conclusion 104

Appendix - Chapter 3 **106**
3.A Additional Tables 106

Conclusion **110**

List of Figures

1.1 Out-of-sample error [proportion of misclassified observations] plotted against the number of neighbours (k) for datasets of different sizes. 31
1.2 Finding the optimal training period for minimizing the out-of-sample error. 35

2.1 Cumulative density function for the matching index (equal weights) and probability index 68
2.A.1 Density functions of daily wages in 2018 before and after imputation . . . 75

3.1 Occupation-level distributional shifts in the unconstrained baseline model. Values above the 45-degree line indicate amplification of observed intersectional disparities, while values below indicate attenuation. 86
3.2 Fairness–performance trade-off: log loss and equal opportunity gap as functions of λ. 90
3.3 Retained utility relative to the baseline model. The dashed line indicates the 95% admissibility threshold. 92

List of Tables

1.1 Levels of occupations — classification of occupations 2010 22
1.2 Descriptive statistics for the different samples 24
1.3 Explanatory variables 27
1.4 Error rates for OLS and RF 36
1.5 Prediction errors for OLS and RF 36
1.6 Prediction errors (TOP 1 and TOP 3) for the four-digit occupational groups for OLS and RF 40
1.A.1 Occupational categories — number of transitions (trans.) and total number of observations of jobs subject to social security (obs. comp.) 46

2.1 Proportion of observations per job category [in percent] 56
2.2 Descriptive statistics for the wage data 56
2.3 Explanatory variables 58
2.4 Predictive performance for job stability 60
2.5 Predictive performance for log daily wages 62
2.6 Predictive performance for daily wages 62
2.7 Exchange ratios for different weights of quality variables and matching probabilities 67
2.8 Mean values for probability index (prob) and matching index (qual) . . . 69

3.1 Occupations with the largest amplification and attenuation of intersectional disparities ($\lambda = 0$) 87
3.2 Performance–fairness trade-off across regularisation levels 90
3.3 Flagged occupations by regularisation level 94
3.4 Flagged occupations in the aggregate-fairest model ($\lambda = 0.010$) 95
3.5 Comparison of flagged occupations under the standard prediction rule and counterfactual averaging (CF). 98
3.6 Predictive performance comparison between the standard prediction rule and counterfactual averaging for $\lambda = 0$. 99
3.7 Flagged occupations under the standard prediction rule and counterfactual averaging for $\lambda = 0.010$. 100

3.A.1 Distribution of intersectional groups over persons 106
3.A.2 Jensen-Shannon divergence in the observed occupational data 106

Introduction

Helping people enter employment is one of the central objectives of labour market policy and constitutes the core mission of Public Employment Services (PES). Every day, caseworkers support unemployed individuals in identifying occupations that offer realistic employment opportunities and sustainable career prospects. Their decisions influence not only whether individuals return to work, but also the occupations in which they are employed and, consequently, their future labour market trajectories. Improving the matching process is therefore of considerable importance both for individual employment prospects and for the efficient allocation of labour within the economy.

From an economic perspective, however, successful labour market policy cannot be assessed solely by the number of individuals leaving unemployment. Employment relationships differ substantially with respect to their stability, productivity, earnings, and long-term career prospects. A recommendation that leads to rapid employment entry is not necessarily associated with favourable labour market outcomes if the resulting employment relationship is short-lived, poorly paid, or followed by repeated unemployment. Effective labour market policy therefore requires both facilitating transitions into employment and supporting occupational matches that are productive, durable, and beneficial for both workers and employers.

These considerations are closely related to search and matching theory, which explains persistent unemployment by the existence of search frictions, incomplete information, and heterogeneity among workers and firms (Mortensen and Pissarides, 1994). Because neither job seekers nor employers possess complete information, mutually beneficial employment relationships are not formed automatically. Thus, labour market institutions play an important role in reducing information frictions and improving matching efficiency.

PES constitute one of the central institutions through which these frictions are addressed. By providing counselling, occupational guidance, vacancy placement, and active labour market programmes, they support unemployed individuals in identifying suitable employment opportunities while simultaneously assisting employers in recruiting appropriate workers. Occupational recommendations are therefore one of the core instruments through which PES influence labour market outcomes. The quality of these recommendations affects not only individual employment prospects but also the efficiency with which workers are allocated across occupations.

At the same time, the institutional environment in which PES operate has changed considerably. Labour markets have become increasingly dynamic as technological change, demographic developments, digitalisation, and evolving skill requirements continuously reshape occupational structures. The administrative labour market data collected by the German Employment Agency captures these patterns by documenting employment biographies, educational qualifications, occupations, competencies and earnings for the whole workforce. Recent research increasingly combines ML with labour market matching and occupational recommendation. Existing studies demonstrate the potential of AI-supported recommendation systems to improve the matching process using large-scale labour market data (van Belle et al., 2018; Van Huynh et al., 2020; Le Barbanchon et al., 2023; Rosenberger et al., 2025).

Unlike traditional econometric approaches, ML algorithms are specifically designed to exploit complex nonlinear relationships, high-dimensional interactions, and heterogeneous prediction patterns without imposing restrictive functional-form assumptions. These properties are particularly relevant for administrative labour market data, which combine very large numbers of observations with complex employment histories and a fixed set of administrative variables. Consequently, ML provides promising opportunities to improve occupational recommendations in PES by exploiting predictive information that remains difficult to capture using conventional statistical methods.

Despite these developments, evaluating AI-supported occupational recommendation systems solely according to predictive performance remains insufficient. PES pursue multiple objectives that extend beyond maximising the probability of employment. Occupational recommendations should not only increase the likelihood of entering employment but should also contribute to stable and productive employment relationships. Moreover, because these recommendations influence access to employment opportunities, they must satisfy broader societal expectations regarding fairness and equal treatment. The evaluation of AI-supported occupational recommendation therefore requires considering predictive performance, job match quality, and fairness jointly rather than in isolation.

Against this background, this dissertation investigates how ML can support occupational recommendation in PES while accounting for the multiple objectives of labour market policy. Rather than viewing predictive performance, job match quality, and fairness as competing objectives, the dissertation views them as complementary dimensions of successful occupational recommendation. The central research question is how AI-supported recommendation systems can improve the placement process and how they should be evaluated before they can be responsibly deployed within PES. Evaluating occupational recommendation requires a broad empirical framework that progressively incorporates predictive effectiveness, job match quality, and fairness.

The dissertation addresses this objective through three cumulative empirical studies. Although each chapter represents an independent scientific contribution, the three studies

are closely connected and build upon one another. Each chapter extends the evaluation framework for AI-supported occupational recommendation in PES.

Chapter 1 is based on joint work with Enzo Weber. It contributes to the growing literature on machine learning for labour market matching by extending search-and-matching approaches with large-scale predictive modelling using administrative data (Mortensen and Pissarides, 1994; Petrongolo and Pissarides, 2001; Broecke, 2023). This chapter establishes the predictive foundation of the dissertation. Occupational recommendation is formulated as a multiclass prediction problem in which ML algorithms estimate the probability that an individual enters a particular occupation. Using comprehensive administrative employment biographies covering the German labour market, the chapter compares conventional statistical models with several ML approaches to evaluate their ability to predict occupational transitions.

The empirical analyses show that tree-based ML methods substantially outperform conventional econometric approaches. In particular, Random Forest achieves the highest predictive performance and provides a robust basis for generating individualised occupational recommendations. More broadly, the chapter demonstrates that large-scale administrative labour market data contain sufficient information to support data-driven occupational recommendation and illustrates the potential of ML for improving decision support in PES.

While chapter 1 identifies occupations with a high probability of successful labour market entry, prediction alone does not determine whether these occupations represent desirable recommendations. Individuals may face several occupations with comparable employment probabilities but markedly different employment prospects after labour market entry. Recommendation systems based exclusively on matching probabilities therefore provide only a partial assessment of occupational attractiveness.

Chapter 2 is based on joint work with Enzo Weber. It relates to the literature on job match quality by combining predicted matching probabilities with expected employment stability and wages into a unified recommendation measure (Allen and Van der Velden, 2001; van den Berg and Vikström, 2014). This chapter extends the evaluation from matching probability to job match quality. Job match quality is composed of two factors, employment stability and wages. Finally, a matching index is created that combines the three complementary dimensions of labour market success: the probability of entering employment, expected employment stability, and expected wages. The resulting index therefore balances immediate placement opportunities against longer-term labour market outcomes.

The empirical analyses demonstrate that incorporating job match quality changes occupational recommendations in economically meaningful ways. Compared with recommendation systems based exclusively on predicted matching probabilities, the matching index favours occupations associated with more stable employment relationships and higher expected earnings, even when these occupations exhibit slightly lower immediate employment probabilities. Consequently, the chapter broadens the evaluation of occupational

recommendation from successful labour market entry towards sustainable labour market integration.

However, recommendation systems that successfully predict occupational transitions and account for employment quality are not necessarily appropriate for deployment within PES. Because occupational recommendations influence access to employment opportunities, they must additionally satisfy the normative expectations associated with public institutions. Recommendation systems should therefore not only perform well but should also avoid reinforcing existing inequalities within the labour market.

Chapter 3 is based on joint work with Enzo Weber and Paula Ziethmann. It contributes to the emerging literature on algorithmic fairness in labour market matching and public sector AI by extending existing fairness evaluations toward an intersectional and deployment-oriented perspective (Bied et al., 2023a; Bauer et al., 2025; Broecke, 2023). This chapter extends the evaluation framework by considering the fairness implications of AI-supported occupational recommendation. Administrative employment biographies reflect historical labour market outcomes and therefore inevitably contain existing patterns of occupational segregation and unequal labour market opportunities. Recommendation systems trained on such data may reproduce or amplify these patterns, even when they achieve high predictive performance. Consequently, evaluating occupational recommendation requires considering not only how accurately recommendations predict labour market outcomes, but also how these recommendations are distributed across different groups of job seekers.

To address this challenge, chapter 3 develops a multidimensional framework for evaluating fairness in AI-supported occupational recommendation. Workers are not characterised by isolated demographic attributes but by combinations of characteristics that jointly shape employment opportunities and occupational sorting. The intersectional perspective therefore provides a more realistic representation of labour market heterogeneity than separate analyses of gender, nationality, or age alone. The empirical analyses investigates distributional fairness, predictive fairness, and residual operational fairness risk. Rather than relying on a single statistical fairness criterion, the chapter recognises that different dimensions of fairness capture different aspects of algorithmic behaviour and therefore provide complementary information for evaluating recommendation systems in PES.

The empirical analyses demonstrate that improving one dimension of fairness does not necessarily improve others. Fairness-aware estimation reduces certain predictive disparities but simultaneously generates measurable trade-offs with predictive performance and does not eliminate all distributional differences across intersectional groups. The analyses further show that some fairness concerns remain even after technical mitigation and therefore require institutional governance, transparency, and human oversight during practical deployment. Consequently, fairness should not be interpreted as a binary property of recommendation systems but rather as a multidimensional evaluation criterion that combines statistical performance with institutional considerations.

Taken together, the three empirical studies progressively broaden the evaluation of AI-supported occupational recommendation. Chapter 1 establishes whether occupational transitions can be predicted accurately using ML. Chapter 2 asks whether these predicted transitions are associated with desirable labour market outcomes. Chapter 3 finally examines whether recommendation systems satisfying these objectives are also consistent with the normative expectations placed upon PES. Rather than representing independent research questions, the three studies therefore address complementary dimensions of a common objective: developing an empirically grounded framework for evaluating AI-supported occupational recommendation in PES.

This dissertation makes three contributions to the literature. First, it contributes to the literature on labour market matching by demonstrating that ML substantially improves the prediction of occupational transitions using large-scale administrative employment biographies. The results show that flexible ML algorithms are particularly well suited to exploiting the complexity and scale of administrative labour market data for occupational recommendation.

Second, the dissertation extends the evaluation of occupational recommendation beyond successful labour market entry by introducing a quality-adjusted Matching Index that jointly considers predicted matching probabilities, expected employment stability, and expected wages. By integrating these complementary dimensions into a single recommendation measure, the dissertation connects predictive modelling with economically meaningful labour market outcomes.

Third, the dissertation contributes to the literature on algorithmic fairness in labour market policy by proposing a multidimensional evaluation framework that combines distributional fairness, predictive fairness, and residual operational fairness risk. Rather than evaluating fairness exclusively through statistical prediction metrics, the framework explicitly considers how recommendation systems influence the distribution of occupational opportunities and highlights the institutional responsibilities associated with AI deployment in PES.

Taken together, these contributions demonstrate that AI-supported occupational recommendation should not be evaluated according to predictive performance alone. Accurate prediction constitutes the empirical foundation of recommendation systems, but effective occupational recommendation in PES additionally requires considering the quality of employment outcomes and the fairness implications of algorithmic decision support. Evaluating these dimensions jointly provides a broader perspective on how ML can contribute to more effective occupational recommendation while remaining consistent with the objectives of labour market policy and the institutional responsibilities of PES.

The remainder of this dissertation is organised as follows. Chapter 1 examines the prediction of occupational transitions using ML methods and compares their predictive performance with conventional econometric approaches. Chapter 2 develops a quality-adjusted Matching Index that combines predicted matching probabilities with expected

employment stability and expected wages. Chapter 3 investigates the fairness implications of AI-supported occupational recommendation by proposing a multidimensional evaluation framework based on an intersectional perspective. The dissertation concludes by synthesising the findings of the three empirical studies, discussing their implications for research and labour market policy, and outlining directions for future research.

Chapter 1

Machine Learning for Labor Market Matching

1.1 Introduction

Matching is the key task of labour markets. Having suitable job alternatives represents a crucial condition for successful employment matches. Approaches based on matching theory (Mortensen and Pissarides, 1994; Petrongolo and Pissarides, 2001) consider the relationship between unemployment and vacancies as inputs into a production function of job findings. Beyond market size, hiring outcomes crucially depend on matching efficiency, i.e., the productivity parameter of the matching function. Improving matching efficiency therefore requires improving individual-level matching probabilities rather than aggregate market conditions alone.

Public employment services (PES) play a central role in this process by generating job suggestions based on the available information. Additional and better-targeted job recommendations can be an incentive for both job seekers and caseworkers to expand and refine search strategies. The practical challenge faced by PES is therefore not the lack of information, but the lack of methods capable of translating rich administrative information into accurate, individualized job recommendations.

In this paper, we examine how algorithm-based methods can support and improve the placement process in PES by estimating individual matching probabilities. We exploit large administrative data covering complete employment biographies of the entire German workforce. The scale and longitudinal depth of this dataset make it possible to observe systematic patterns in transitions into employment that are not identifiable in survey data or small administrative samples. The estimation of matching probabilities can be used to create ranked lists of job recommendations for each job seeker. This could assist caseworkers in their search for suitable occupations and lead to more efficient and flexible placement and thus with lower unemployment and more stable employment relationships.

Choosing a suitable algorithm is crucial for optimizing the results. Therefore, we systematically compare traditional statistical models with machine learning (ML) algorithms. The central motivation of this paper is to assess whether and why ML methods – especially random forest – are better suited than standard econometric approaches for extracting predictive information from administrative data. The estimated matching probabilities can be used to generate ranked lists of job recommendations for each job seeker, directly supporting caseworkers' decisions and potentially leading to more efficient placement and lower unemployment durations.

We contribute to a developing literature by optimizing individual matching probabilities in the setting of large administrative datasets and tackling the critical issues arising from the structure and properties of the data. Since job recommendations must be provided as specifically as possible in order to be useful for caseworkers, this means the predictions should cover the three- and four-digit level of the occupation classification, requiring the model to distinguish between up to 700 occupational categories with highly uneven frequencies. While standard applications often do not allow for specifying such a large number of categories, we show which algorithms are able to estimate such models. Typically, algorithm-based job recommendations are mostly designed for a limited group of people or depend solely on self-assessment, for example, due to data limitations or other constraints. We develop a model that can be used for the whole workforce including those who are looking for suitable apprenticeships, training or alternative jobs within a single unified framework. The institutional setting is relevant to many countries and allows the tapping of large administrative datasets.

Second, we are able to use a large-scale administrative dataset that covers the whole German work force. Due to the availability of the full universe of employment and unemployment spells, we open up the potential to use the full richness of variables for training in order to optimize flow matching. This approach incorporates substantially more information by considering flows into employment than models based solely on unemployment-to-employment transitions. As a result, the ML models achieve significantly better predictive performance while simultaneously reducing the risk that predictions are driven by past unemployment status alone. Thus, we develop ways how to estimate models combining a large dataset and a large amount of variables in an efficient way. We show how to efficiently estimate RF or k-nearest-neighbours (kNN) and how to specify hyperparameters for the kNN algorithm when the model becomes too large for the conventional approach.

Third, we provide a systematic empirical comparison of traditional econometric models (OLS and logit) and ML algorithms (RF and kNN) in a setting characterised by extreme dimensionality, strong heterogeneity, and severe class imbalance. We show that random forests clearly outperform both traditional statistical methods and alternative ML algorithms. In particular, the random forest out-of-sample classification error is approximately 30 percent lower than that of OLS, demonstrating economically meaningful performance

gains.

Fourth, we demonstrate that ML methods – and random forests in particular – much better translate the scale of administrative data into predictive gains. While conventional models either fail to converge or do not improve with additional data once functional form restrictions bind, random forests continue to benefit from increasing sample size. This finding directly links data volume and data structure to model performance and explains why ML dominates standard approaches in this context.

Administrative labour market data are characterised by a rigid and institutionally fixed set of covariates, limited flexibility in feature engineering, and strict data protection requirements. From a traditional econometric perspective, this rigidity could be considered a drawback. However, we argue that it constitutes a key advantage for ML applications: when feature spaces are fixed, model performance depends crucially on the ability to exploit nonlinearities, interactions, and heterogeneity within a very large number of observations.

ML methods offer concrete methodological advantages in this setting. They can capture complex nonlinear relationships and interactions between individual characteristics, employment histories, and occupational attributes without imposing restrictive functional form assumptions. This is particularly important for job-search processes, which are inherently nonlinear and heterogeneous. Moreover, ML algorithms scale efficiently to extremely large datasets and can therefore fully exploit the richness of administrative employment biographies.

Among the ML approaches considered, random forests are particularly well suited to this task. Random forests are robust to noise, multicollinearity, and highly imbalanced outcome distributions, all of which are prominent features of administrative labour market data. They are also able to handle an exceptionally large number of outcome categories. This is crucial for the underlying type of application.

An example that aligns closely with the findings in our study is given by a study that compares the ability of traditional rule-based profiling systems to statistical and ML models predicting long-term unemployment using comprehensive administrative data from Spain (Junquera and Kern, 2025). RF substantially outperform the rule-based approaches by capturing complex interactions and nonlinear patterns in employment histories.

Taken together, these methodological advantages explain why ML – and random forests in particular – outperform standard econometric and statistical approaches in the context of large-scale administrative data on employment biographies. ML does not merely offer more flexible functional forms; it enables a fundamentally different use of administrative data by converting scale, rigidity, and heterogeneity into predictive accuracy. The use of algorithm-based systems, especially when using personal data, also poses a major challenge from an ethical perspective. Since our model includes variables with discriminatory potential, it is imperative to examine where and to what extent discrimination occurs.

Thus, we elaborate on the role of potential discrimination due to data-based algorithms reproducing patterns. We demonstrate and propose an approach that does not neglect critical variables, but shows a way how to control for their influence. This is crucial for the introduction of algorithm-based models in the public sector.
This paper is structured as follows. In section 1.2, we give an overview on the related literature and in section 1.3 we describe the general setting and the dataset. The estimation strategy and the estimation methods are then described in section 1.4, which also discusses the advantages of ML applied to administrative data. In section 1.5, we present and interpret the empirical results. Section 1.6 outlines some robustness checks and section 1.7 shows limitations and potential biases. Section 1.8 concludes the paper.

1.2 Related Literature

Our study is related to several strands of literature. Mismatch unemployment occurs when worker qualifications differ from job requirements (Sahin et al., 2014; Bauer, 2013). One important factor is the impact of mismatch on the matching efficiency (Hutter and Weber, 2017). In this regard, we provide a method that can be used to generate recommendations for jobs that might serve as alternatives for job seekers. Studies that investigate job suggestions show that having more alternatives leads to a "broader" search strategy and that increasing the number of job interviews reduces mismatch (Belot et al., 2019; Blundell et al., 2004). A field experiment also shows that having more information is positively linked to labour market success (Altmann et al., 2018). Designing a recommender system based on ML techniques has a positive effect on the searching activity (Le Barbanchon et al., 2023). For example, a system called CareerBERT leverages the power of unstructured textual data sources to provide job recommendations (Rosenberger et al., 2025).
In this context, one option is skill proximity to suggest jobs that are close to the profiles of job seekers. While this has conceptual advantages, skill proximity relies on predefined measurements of similarity and does not directly capture the real-world choices people make. By contrast, realized outcomes – as in our approach – reflect observed decisions, including options that may not be relevant according to skill proximity but were nevertheless selected by workers. Our model covers actual transitions and thus, identifies real-world patterns that skill proximity might overlook.
Big data (Gerunov, 2014) and machine learning methods are already used in different fields of labour market research. For example, for policy evaluation (Huber, 2021), for analysing which skills are required for different occupations (Colombo et al., 2018), for the investigation of web job vacancies (Boselli et al., 2017) or for matching vacancies to candidates (van Belle et al., 2018; Fang, 2015; Braun, 2017). These studies try to select people who will fit a certain vacancy perfectly by using ML. The algorithm learns from the company's historical placement data to order a list of candidates for a given job offer.

Further, ML methods are applied for ranking purposes (Li, 2011; Cao et al., 2007) and online recruitment systems (Faliagka et al., 2012). For example, job recommendations are generated based on auxiliary behaviour data and user's explicit preference data simultaneously (Saito and Sugiyama, 2024). Further, ML techniques, in this case, neural networks, are used for determining suitable jobs (Van Huynh et al., 2020). Many of these models work with a predefined set of variables that are recorded specifically for the research project and are not necessarily transferable.

Some countries are using artificial intelligence (AI) already for the public employment service. AI is mostly used for matching job seekers with vacancies, designing job postings, profile job seekers or to support clients (Brioscú et al., 2024a). To be more precise, 20 percent of the OECD countries are already using matching tools that assist in recommending matches between job seekers and vacancies. For example, in France the public employment service uses a job recommendation algorithm (Bied et al., 2023b), involving a vast majority of low-qualified job seekers and an extremely sparse interaction matrix. The algorithm does not lead to unfairness in terms of the gender gap.

Our results can be placed in the existing literature as follows: The increasing use of ML methods, especially in labour market research, is due to the fact that they provide new possibilities for estimating models and significantly improve the results. We also observe that the underlying research question can be answered better using ML models. They produce lower error rates and the estimation and specification of the model can be done in an automated and more efficient way. A review of the literature shows that it is beneficial to present job seekers with several alternatives. On the one hand, it broadens their search behaviour, and on the other, it reduces mismatch. Our approach delivers such recommendations, which often include the jobs that similar individuals have taken in the past. The literature also shows that algorithm-based recommender systems have positive effects on the matching process. In Section 1.5.4, we use an illustrative scenario to show that our results could lead to a reduction in the unemployment rate. There would also be an effect using conventional methods, but a smaller one than the ML-based approach. AI applications introduced in other countries demonstrate that it is also important to examine the model from a fairness perspective. We show that variables with discriminatory potential have an important influence on our model and present approaches to address this problem.

1.3 Setting and Data Structure

In Germany, data on the whole workforce is collected by the Federal Employment Agency (BA). Each person who becomes unemployed receives an invitation to a consultation and is registered as a client. For both receiving financial benefits and active support, this is obligatory and takes place in the regional job centers or in employment agencies. There, a profile is created and caseworkers inform the clients about their duties, for example,

bringing a proof of active job search. Caseworkers support unemployed persons by searching for and suggesting jobs that have been reported to the BA. Additionally, the job search can be conducted via the online portal of the BA or other private job platforms. For employed persons, employers have to record the time and the occupational category if someone starts a new job or changes occur. Due to this procedure the BA is able to provide comprehensive information on the employment biographies which is included in a research dataset. The job recommendations are based on the caseworkers' impressions from the personal interview. When searching for suitable job ads, they can use a keyword search and a system that compares job and person requirements. Statistical methods could support this process in the future by providing additional or alternative recommendations to job seekers or people who are looking for a new job.
Indeed, the setting with a public labour market policy, in which administrative data is collected, can be found in many countries. Many administrative datasets are comparable with the German data. Our application therefore has potential well beyond the selected use case. Thus, both the practical environment and the data availability reflect typical constellations.

1.3.1 Defining Matching

The main idea is to match people with certain characteristics to a certain occupational group. The model is given by

$$P(M_i = j) = f(\mathbf{X}_i, \mathbf{Y}_i), \tag{1.1}$$

with $j = 1, \ldots, J$ where J is the number of different occupational groups, $i = 1, \ldots, N$, N is the number of observations (i.e. spells), $\mathbf{M}$ is a Nx1-vector containing the occupational group someone is employed in, so M_i denotes the occupation obtained in spell i. Further, $\mathbf{X}_i$ is a 1xS-vector, where S is the number of variables denoting the characteristics of a person in spell i. $\mathbf{Y}_i$ is a 1xR-vector, where R is the number of variables denoting the characteristics of the jobs a person in spell i. Thus, we estimate the probability of a person being employed in a certain occupation and assume that the process is a function of personal and job characteristics, both of which included in the dataset.
A match is considered successful when a person who was unemployed in the previous period (i.e. an unemployment spell is recorded in the previous period), starts a job. We define a match as the occupation in which a person enters employment. The main idea is to provide a list of recommendations of possible occupations. Thus, it is possible to provide job seekers with alternatives. In practice, we use two different samples for estimation (see section 1.3.4). The first represents transitions from unemployment into employment. The second covers all flows into employment.

1.3.2 Data Description

The underlying dataset is a random ten percent sample of the Integrated Employment Biographies (IEB) (Schneider, 2020). This is an administrative dataset that covers the German workforce covered by the social security system from 1980 onwards and gives information about the characteristics of a certain person as well as the characteristics of their jobs. Using this dataset allows us to capture the hard facts in the employment history, because the German Employment Agency collects a large number of administrative characteristics for a long time. The dataset combines different data sources. For the present research question, the Job seeker Histories (ASU/XASU) and the Employee History (BeH) are required. The dataset is also assumed to be large enough because it contains a large number of occupational groups with a different number of observations. For each group a sufficient number of observations is crucial. The distribution of people (and observations) across jobs varies widely. While some occupational categories contain many observations, there are also occupational categories that contain comparatively few observations. Too little information (i.e. observations) would not provide a sufficient basis for valid estimations for these groups, because a certain number of observations is needed that a statistical method has the possibility to recognize a specific pattern. We consider workers to be employed if they were in a job subject to social security and to be unemployed if they are registered as unemployed with the public employment service (excluding those temporarily unfit for work). In general, the basis for our estimations are employment spells. Thus, we obtain the characteristics of jobs and people at the beginning of each spell.

1.3.3 Endogenous Variable

The jobs are grouped according to the German classification of occupations 2010 (KldB2010)[1] (Paulus and Matthes, 2013). The five breakdown levels are shown in Table 1.1.

In this project, the estimation results should be able to support the placement process. The differentiation between occupational areas or occupational main groups is not precise enough, because this classification would provide no additional information for placement officers. Thus, it is necessary to define the occupational groups as precisely as possible, but at the same time, the number of observations per category has to be sufficiently large for estimation purposes. Bearing this in mind, we use the different occupational groups defined in the occupational groups (three-digit) on the one hand and the occupational sub-groups (four-digit) on the other as endogenous variables. More specifically, in the three-digit case a classification problem with 144 different classes has to be solved, while

[1] More detailed information can be found in: https://statistik.arbeitsagentur.de/DE/Navigation/Grundlagen/Klassifikationen/Klassifikation-der-Berufe/KldB2010-Fassung2020/Systematik-Verzeichnisse/Systematik-Verzeichnisse-Nav.html

Table 1.1: Levels of occupations — classification of occupations 2010

Group	**Number of different occupational fields**
Occupational areas (1-digit)	10
Occupational main-groups (2-digit)	37
Occupational groups (3-digit)	144
Occupational sub-groups (4-digit)	700
Occupational types (5-digit)	1288

Source: Statistics provided by the Federal Employment Agency.

in the four-digit model 700 different classes have to be dealt with.
The number of employed persons varies greatly across occupations, because certain occupations require more workers than others. Furthermore, the distribution of persons across occupations can change over time due to changes in occupational profiles and the demand for specific types of work. As a result, some occupations may become more or less relevant or the skill and task profile for a given occupation may also shift. For this reason, it is crucial to select an appropriate time period in order to avoid capturing outdated patterns. In addition, the modeling approach should handle the imbalanced distribution of persons across occupational groups.

1.3.4 Choosing Samples for Estimation

Our estimations are based on two different samples. Both contain information from the period between 1975 and 2018. The first sample, called the unemployment flows in the following, is adjusted to the present research question. Here we only consider those employment spells (i.e. people who start a job subject to social security) that are preceded by an unemployment spell. Table 1.2 shows some descriptives for the three-digit occupations for the whole dataset.
There might be a problem with potential discrimination due to the unemployment status. Providing job recommendations based on a sample of unemployed persons could be misleading, because one could argue that predictions are made based on persons with bad matches. If the matches would have been good, these people probably would not be unemployed. The second sample, called the total flows in the following mitigates this problem. It contains all spells in our sample reporting all flows into employment subject to social security. Therefore the reason someone starts a new job, especially if they were unemployed in the preceding period is not of interest. Thus, here we do not distinguish between the employment status in the preceding time period or the incentive for starting a new job. Consequently, unemployment flows represent a proper subset of the total flows.
Information about the number of children and the marital status of an individual is only

available in the Job seeker Histories. This means, that there is a missing for people who have no unemployment spell before an employment spell (in this case the information is carried forward by one period).

For seasonal workers, naturally there is a high number of transitions. Consequently, these classes may appear disproportionately often, but the feature vectors for these workers are highly similar. Intuitively, one might expect that this could bias the model toward over-predicting the seasonal classes. However, several properties of RF mitigate this risk. First, each tree is trained on a bootstrap sample, which introduces variability in class composition and prevents any single seasonal pattern from dominating the training process. Second, at each split, the algorithm considers only a randomly selected subset of predictors. This stochastic feature selection reduces the chance that trees base their structure solely on signals that recur in many observations. Third, the splitting criterion – Gini impurity – aims to choose splits that best separate observations belonging to different classes. This ensures that a class is not predicted more frequently because it appears more often in the data. The number of predictions depends on the features that allow the model to draw a clear distinction between the classes. Thus, bootstrap sampling, randomized feature selection, and impurity-based splitting ensure that the classifier does not automatically over-suggest occupations with seasonal workers.

Table 1.2: Descriptive statistics for the different samples

	complete[1]	**stock**[2]	**Unemployment flow**[3]
No. of observations total	141,327,440	54,781,854	6,019,817
No. of different persons	3,251,949	2,883,188	2,359,075
Median*	162,465.5	121,682.5	11,491.5
Mean*	498,905.2	380,429.5	41,804.28
Standard deviation*	1,020,538	750,014.1	84,801.27
Min*	77	27	6
Max*	7,761,145	6,053,898	532,939

Notes: * Number of observations per occupational group.
[1] complete: all observations for the entire time frame (1975–2018) are considered.
[2] stock: observations of persons having a job subject to social security during the entire time frame (1975–2018) are considered.
[3] unemployment flow: observations of persons having transitions from unemployment to employment during the entire time frame (1975–2018) are considered.
Source: Own calculations.

1.3.5 Exogenous Variables

In the following, the explanatory variables are described in more detail. Table 1.3 shows the exogenous variables. For consistency with the target group of the study, the distributions are calculated based on individuals experiencing unemployment flows. Focusing on

unemployment flows ensures that the reported descriptive statistics accurately reflect the characteristics of the participants relevant to the research question. First, a person's gender and the federal state they live in are included in the model. This information is available for each person. The number of observations across federal states reflects the difference in population size. Another explanatory variable is the nationality. For non-Germans, we distinguish between the nationalities of asylum seekers[2] and EU and non-EU nationalities. Whether a person really has a migrant background or not cannot be determined from the data.

The marital status and children variables are only captured in the Job seeker Histories because the information is relevant for the employment agency if a person becomes unemployed. The Employment History (BeH) does not include this information. Thus, for the stock data, information on marital status and children is carried forward under the assumption of temporal stability. While this approach may introduce inaccuracies, it remains preferable to having no information at all. Retaining these variables allows for a consistent characterization of individuals even when they are not in unemployment, and ensures that analyses can incorporate marital status without systematically excluding cases. The information is updated when a new unemployment spell occur, which mitigates cumulative errors over time. Overall, including these variables enhances the analytical richness of the dataset, whereas omitting it entirely would result in a loss of potentially relevant explanatory information. The two variables were not both captured in the same way from the beginning and had to be re-coded (Schneider, 2020). According to the data description, due to the reclassification process or incorrect reporting some information might be lost or transformed incorrectly by creating the new variables. For marital status less than 1 percent of the observations have missing values. In the case of the children variable a distinction cannot always be drawn between "nonresponse" and "no children". The reference category thus captures both cases. In cases where family composition does not affect the job search process, the variable may be left unrecorded. Concerning marital status, we only distinguish between living alone or in a relationship. In our dataset with profiles of job seekers the share of singles is larger than in the entire population. We assume that the unemployment status can correlate with marital status, and consequently, the observed proportions likely reflect the interplay between partnership patterns and job-seeking status rather than the broader population distribution.

To improve the education variable, which is fraught with missing values and inconsistencies (i.e. a person has a lower level of education in a spell that comes after a spell with a higher level of education), we rely on a well-established imputation process (Fitzenberger et al., 2006). After imputation the "education" variable contains only 2.24 percent missing

[2]Here we are referring to people whose nationality is equal to one of the eight countries of origin with the highest number of asylum seekers in July 2020 (source: https://de.statista.com/statistik/daten/studie/154287/umfrage/hauptherkunftslaender-von-asylbewerbern/). In particular, the countries are Syria, Iraq, Afghanistan, Turkey, Nigeria, Iran, Eritrea, Somalia and Georgia.

values. A further explanatory variable is the occupational group in which a person attends their vocational training. We assume that a person who attended vocational training for more than two years in the same occupation, has completed the vocational training in this category. We cannot observe, if someone definitely finished the vocational training. Even if this is the case, someone has experience in the corresponding occupation which could also be relevant for future job search. Further, for each person we record the age at the beginning of an employment spell and the number of days in unemployment before finding a new occupation. The age distribution of individuals exiting unemployment shows that the majority are aged 20–39, with 33 % aged 20–29 and 27 % aged 30–39. Participants aged 40–49 account for 23 %, those aged 50–59 for 14 %, while individuals younger than 20 and those 60 or older constitute only 2 % of flows. This distribution reflects the actual composition of the target population, with the largest shares in early and mid-career stages. In summary, the explanatory variables are predominantly binary or categorical and include sociodemographic characteristics and characteristics of previous jobs. The variables that are only observed during unemployment spells are carried forward to non-unemployment periods under the assumption of temporal stability. Education variables are imputed to address missing values and inconsistencies. Such data features require estimation strategies that are robust to unequal category sizes and sparsely filled variables. In addition, administrative real-world data reflect institutional processes rather than controlled data collection for research purposes. The data collection process can not be extended or modified, which imposes specific constraints on empirical modeling. Thus, the analysis must rely on the available variable definitions, observation windows, and reporting practices, even when these lead to missing values or inconsistencies. Consequently, the model design should focus on the structure of the existing data rather than on the possibility of refining or enriching the covariate set.

1.4 Estimation Strategy

We estimate the model using a range of approaches, including both traditional methods and ML techniques. While we present results from all methods for comparison, we conclude with a detailed discussion of why machine learning approaches – particularly RF – are the most appropriate choice for our administrative data, given their ability to flexibly capture complex, high-dimensional relationships.

1.4.1 Preparation

Before estimating the model, the dataset is split up into a train set and test set which is a common procedure in ML. The train set is used for estimating the model or fitting the parameters while the test set is independent of the train set but follows the same probability

Table 1.3: Explanatory variables

Variable	**Description of values/ Proportion (in parentheses)**
Gender	female (39.95%), male (60.05%)
Federal state	Nordrhein-Westfalen (17.51%), Bayern (13.43%), Schleswig-Holstein (3.31%), Sachsen-Anhalt (4.16%), Hessen (5.82%), Baden-Württemberg (9.02%), Brandenburg (4.08%), Mecklenburg-Vorpommern (3.16%), Thüringen (3.83%), Sachsen (6.60%), Niedersachsen (8.90%), Bremen (0.83%), Berlin (5.51%), Hamburg (2.22%), Saarland (0.98%), Rheinland-Pfalz (4.05%)
Nationality	German (88.66%), EU (3.65%), 8 main migration countries (0.90%), Europe without EU (4.76%), other countries (2.03%)
Marital status	single / lives alone (54.90%), in relationship / married (44.80%)
Children	at least one child under 15 years (26.44%), no children / unknown (73.56%)
Education	no school leaving certificate (0.02%), primary/lower school or equivalent (11.91%), primary/lower school with vocational training (65.33%), upper secondary without vocational training (1.29%), upper secondary with vocational training (9.15%), university of applied sciences (2.64%), university (7.42%)
Skill level	Skill level required for preceding employment(s); KLDB 2010 5-digit classification: expert / complex problem solving, advanced knowledge / application, basic operational tasks, manual or routine tasks.
Preceding occupation	Occupational group(s) employed in before starting a new job.
Vocational training	Occupational group of vocational training.
Age	Age at the start of employment.
Days in unemployment	Number of days unemployed before starting a new job.

Note: All distributions are calculated for individuals in unemployment flows only, as this subgroup constitutes the primary target population of the study.
Source: IEB and own calculations.

distribution. In classification tasks, the algorithm determines how the set of variables is combined into decision rules, either through linear weighting (OLS, logit) or through data-driven, nonlinear partitioning of the feature space (kNN, RF). As the main goal is to make predictions on unknown data, the performance of a specified classifier is evaluated on the test set, reflecting its ability to generalize within a fixed administrative feature space that cannot be extended or modified. Based on the results we make predictions for the training and the test data. Results from the train set are called in-sample, while those from the test set are called out-of-sample. We then compare the results in particular, and especially take a closer look at the difference between OLS and RF try to determine their practical benefits (i.e. reduction of the unemployment rate).
Sometimes there are multiple observations from the same person in the dataset. Thus, all observations from the same person are always either in the train set or in the test set. After having estimated the model, we predict the outcome for the train set and the test set. As a measure of goodness, we use the in-sample and the out-of-sample error, which enables us to calculate the percentage of wrong predictions. To be more precise, we define the occupation with the highest predicted probability as the first best job suggestion and check if this outcome coincides with the occupation someone is really employed in. If this is the case, the prediction is defined as right and otherwise it is defined as wrong.
Note that the underlying classification problem is imbalanced between the classes. This means that the sample size of several classes is significantly higher or lower in comparison to other classes. This may negatively affect the performance because relatively balanced class distributions are assumed for classification problems. Imbalanced class distribution might be no problem if the dataset is large enough because it contains also more information about the small class (Sun et al., 2009). Hence, pattern recognition is also possible for small classes because distinguishing rare samples from the majority is possible. Nevertheless, in decision trees, some branches that predict the small classes are probably removed because the new leaf node is labeled with a dominant class.

1.4.2 OLS and Logit

As multinomial models cannot be estimated with OLS or logit, we estimate a separate one-versus-rest equation for every occupational group, which reflects the high-dimensional and purely categorical outcome structure of the administrative data. Concretely, 144 equations for the 3-digit and over 700 equations for the 4-digit level, reflecting the respective numbers of occupational groups defined in the 2010 classification of occupations. First, we define a ixj-matrix $\mathbf{M}$, where

$$M_{ij} = \begin{cases} 1, & \text{if person } i \text{ is employed in occupation } j \\ 0, & \text{otherwise.} \end{cases} \tag{1.2}$$

The OLS model is

$$P(M_{ij} = 1) = \mathbf{X}_i\beta + \mathbf{Y}_i\gamma \tag{1.3}$$

and the logistic model

$$P(M_{ij} = 1) = \frac{exp(\mathbf{X}_i\beta + \mathbf{Y}_i\gamma)}{1 + exp(\mathbf{X}_i\beta + \mathbf{Y}_i\gamma)}, \tag{1.4}$$

where β is a Sx1-vector and γ is a Rx1-vector. Thus, in practice J single equations have to be estimated.
After estimating the models we make predictions for each occupational group in the train and the test set. Afterwards comparing the results with the RF results is possible, as a ranked list of occupations can be obtained from all estimation methods. The category with the highest predicted probability is chosen as the best suggestion for a person. It is also possible to compare the predicted category with the real one. Note that in the case of OLS the outcome values for predictions cannot be interpreted as probabilities as in the logit case. The resulting values do not lie between 0 and 1. However, the interpretation of coefficients is not of interest in this application, the focus lies on the predictive ranking performance.

1.4.3 Random Forest

For constructing a classification model a learning algorithm that obtains the relationship (i.e. the pattern) between the attribute set and the class label is crucial. The model that fits the training data best is generated. Getting good predictions on unseen data is important but is only possible if the model fits the data. Random forests (Hastie et al., 2017; Breiman, 2001) are constructed by using bagging[3] to build de-correlated trees that are averaged in the end. By definition the RF classifier is a collection of tree-structured classifiers given by

$$h(\mathbf{z}, \Theta_k), k = 1, \ldots, \tag{1.5}$$

where Θ_k are identically distributed random vectors and $\mathbf{z}$ is the input. In the underlying context, $\mathbf{z}$ contains the variables denoted by $\mathbf{X}_i$ and $\mathbf{Y}_i$. Hence, for classification each tree which is built from a random vector of parameters gives a vote for the most popular class. Finally, the RF classifies using the majority vote.
Our RF model was estimated using the *ranger* package in R (a high-performance implementation of RF designed for large and high-dimensional datasets). The final model was trained with 100 trees, an mtry value of 12 (the number of predictors randomly sampled at each split), a minimum node size of 10 (the minimum number of observations required for further splitting), permutation-based variable importance (a method that measures the

[3]For reducing the variance, bagging is used to average a large number of noisy but unbiased models.

accuracy loss after randomly shuffling a predictor), and the Gini split rule (an impurity measure used to determine the best splitting variable). For such a high-dimensional dataset, characterised by numerous categorical variables and long employment histories, the model relies strongly on the ability of individual trees to identify locally informative partitioning variables. The variable importance measure indicates that the most influential predictor is the previous occupation, followed by the previous skill level, with a clear gap between these two variables. Other important predictors are several variables capturing occupations further back in an individual's employment history. Gender appears at rank 19. Tuning is not always recommended for multiclass classification by comparing the results of estimations with between 11 and 2000 trees (Probst and Boulesteix, 2018). To check potential gains for our model, we tested alternative hyperparameter configurations for the underlying dataset which led to only minor performance changes, suggesting that the model is relatively stable with respect to its tuning choices. This finding is consistent with the way *ranger* handles tree depth: rather than imposing a fixed maximum depth, the effective depth emerges from the interaction of the minimum node size, the impurity-based split rule, and the underlying data structure. Given the large sample size and the heterogeneous features, the chosen minimum node size of 10 allows trees to grow sufficiently deep to capture non-linear and interaction effects without overfitting.
The algorithm produces a vector whose top entry corresponds to the class with the highest predicted probability. Thus, this vector can be directly compared with the vector of real occupations. To obtain Top-N accuracy (the share of cases where the true occupation appears among the top N predicted classes), the full probability matrix across all classes must be computed.

1.4.4 K-Nearest Neighbours

Another promising machine learning method is kNN. Finding the right value for k is one of the most important steps in the estimation procedure. In this application, the standard hyperparameter tuning is not feasible due to memory and runtime constraints in combination with the high dimensionality induced by categorical variables. Instead, we estimate the model with different k's and calculate the classification error. As the out-of-sample error is of interest, we choose the value that minimizes the out-of-sample prediction. Nevertheless, due to technical restrictions this procedure is only possible up to 100,000 observations. Thus, we estimate the optimal k for random samples of 1000, 10,000 and 100,000 observations and use this information to estimate the optimal k for with 1,000,000 observations. Fig. 1.1 shows the out-of-sample errors for different k for all numbers of observations. Here we achieve a reduction in the out-of-sample error with an increasing number of observations and the shapes of the graphs look very similar. It can also be seen that choosing the appropriate k is important for optimising the prediction

quality. We find that the best value for k depends also on the number of observations. Here a linear relationship might be suspected. Under the assumption of a linear relationship, the right k-value for 1,000,000 observations would be 15,924. So, as soon as we are able to estimate the model with an optimal k, this method has the potential to produce equal or maybe better results than RF.

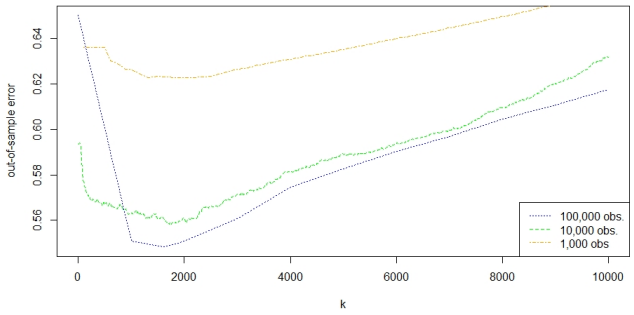


Figure 1.1: Out-of-sample error [proportion of misclassified observations] plotted against the number of neighbours (k) for datasets of different sizes.

Source: own calculations.

1.4.5 Estimation Methods - Findings

When predicting a large number of occupational fields using historical employment biographies, we find that tree-based methods consistently outperform all other methods. Among all evaluated models, RF has the strongest predictive accuracy. The richness of our data – many categorical variables, employment histories, and strong class imbalance across occupational groups – leads to complex, nonlinear, and often heterogenous relations between predictors and outcomes. The advantage of RF arises from the ability to capture nonlinear relationships, model complex interaction structures, and process high-dimensional feature spaces without requiring prior functional assumptions. The structure of RF to build an ensemble of decision trees using bootstrap samples and random feature selection at each split, thereby combining the strengths of flexible decision rules with robustness against overfitting enables the algorithm to exploit heterogeneous patterns in employment histories, such as nonlinear effects and interactions. In contrast, logit and OLS models impose strict linearity and additivity assumptions and require explicit specification of interactions and nonlinearities. This explains the weaker performance for the imbalanced structure of occupational data. Further, due to the large sample size and rich feature set of the IEB machine learning methods – especially RF – achieve much better out-of-sample performance and provide a more accurate basis for job recommendations. The OLS outputs show that the same variables are highly significant as in the RF case. This holds true for the previous

occupation just as for a lot of other variables that are also highly significant. Since the OLS estimation consists of 144 single OLS models, naturally, there are differences in the variables which are most relevant in a specific model.

Further, we find that RF outperforms other ML methods. Our estimations show that kNN struggles with the imbalancedness and the tuning of k for such large sample sizes. Also SVMs require careful tuning and scale poorly with very large sample sizes. Regarding hyperparameter tuning for RF, we see that the default implementations (we use the R package "ranger") produces performances that are close to each other for different numbers of observations.

Concluding, RF is a robust, efficient, and high-performing method for occupational classification on complete administrative labour market data, outperforming both classical statistical models and other ML alternatives. For efficiency reasons, RF require substantial computational resources, particularly CPU capacity. The training phase involves repeated sampling, tree construction, and node-level split optimisation across many parallel branches. This results in significantly higher computational complexity compared to other statistical methods. If the model is used in practice, regular model retraining and updating would be necessary. The model must be retrained, re-tuned, and re-evaluated. This ensures model stability in a long term.

Our results highlight a fundamental difference between machine learning models and classical statistical models in how they respond to increasing data availability. Traditional models rely on strong functional form assumptions and limited interaction structures; as a consequence, their predictive performance quickly saturates once the main effects are estimated with sufficient precision. In contrast, machine learning models exhibit a strong and persistent performance gradient with respect to sample size. As more data are introduced, these models are able to learn increasingly complex, high-order relationships between worker characteristics, job attributes, and labour market outcomes. This effect is particularly pronounced in administrative datasets, where sample sizes are large enough to support the estimation of complex decision rules without overfitting. Even in smaller subsamples, machine learning methods already deliver slightly superior results, but as the dataset grows, the performance gap widens substantially, illustrating the intrinsic scalability of machine learning.

We implement RF as a core ML approach and show that their algorithmic properties align especially well with the characteristics of administrative labour market data. RF efficiently handle high-dimensional feature spaces with limited scope for manual feature selection, are robust to noisy and correlated inputs, and naturally capture nonlinear effects and interactions between the exogenous variables without explicit a-priori specification. Their ensemble structure – i.e., the aggregation of predictions from many individual decision trees – stabilizes predictions and reduces variance, which is crucial when working with administratively recorded labour market data characterised by highly heterogeneous

employment biographies of the entire German workforce. This means that while individual trees may be sensitive to particular samples or noise, combining them in a forest produces more reliable and robust predictions across diverse subgroups. From an applied machine learning perspective, RF also offer important practical advantages: they scale well computationally, require comparatively little hyperparameter tuning, and can be deployed in automated pipelines with minimal manual intervention. These properties make them particularly suitable for operational settings where models must be retrained regularly and applied consistently across large populations, further strengthening the case for machine learning in administrative data applications.

Existing benchmark evidence highlights fundamental differences between ML methods and classical statistical models in how they exploit information contained in complex datasets. A large-scale comparison (Couronné et al., 2018) of RF and logistic regression across 243 real-world datasets, shows that RF consistently achieve superior predictive performance, particularly in settings characterised by nonlinear relationships, interaction effects, and heterogeneous data structures. These performance differences reflect a core methodological distinction: while classical statistical models rely on strong parametric assumptions and explicitly specified functional forms, ML methods such as RF learn flexible decision rules directly from the data. As a result, they are far better able to adapt to unknown or highly complex data-generating processes. This supports the findings in the underlying paper. Administrative data is typically characterised by a rigid, institutionally determined set of covariates, which constrains model design to algorithms that can extract complex patterns from the existing feature space rather than relying on tailored feature construction. At the same time, administrative datasets provide population-wide coverage and extremely large sample sizes, capturing substantial heterogeneity in worker and job characteristics. In this context, traditional models quickly exhaust the informational content of the data, as their limited functional flexibility prevents further performance gains once main effects are precisely estimated. In contrast, RF are able to exploit the scale of administrative data more effectively by using large sample sizes to reliably estimate complex split structures even for rare combinations of categorical characteristics.

Importantly, the benchmark results (Couronné et al., 2018) suggest that these advantages are not merely theoretical but empirically robust across a wide range of applications. When applied to administrative labour market data, ML methods therefore offer a decisive advantage: they transform large-scale but structurally rigid datasets – where variables are fixed and partly noisy – into meaningful predictive improvements by leveraging data-driven partitioning rather than parametric assumptions. This makes them particularly well suited for labour market matching tasks, where accurate predictions depend on capturing subtle interactions across a highly heterogeneous population.

1.5 Empirical Results

We use both the total flows and the unemployment flows for our estimations. The predicted probability for each occupation allows us to create a ranked list with job suggestions for each individual. First, we only consider the category with the highest probability as a prediction. The measure of goodness is the classification error that denotes the percentage of wrong predictions. We start with the three-digit occupations (i.e. 144 different classes) as the endogenous variable. Amongst the ML algorithms, RF achieves the lowest error rate and highest efficiency, which is consistent with the high-dimensional, predominantly categorical structure of the underlying data, while OLS is the best traditional method. Thus, in the following section, we compare RF and OLS.

1.5.1 Sample Splitting by Year

There was a change in the German classification system of occupations in 2011, which means there might be a structural break at this point in time. The goal of the new classification (i.e. KldB2010) was to better map the structure of occupations in Germany. Further, there is a better comparability with the international classification of occupations (ISCO-08). To check for this, we restrict the sample and consider only employment spells beginning after January 1, 2012. To be more precise, we use a subsample of the original data containing spells with specific job recordings after the change in the coding scheme. Additionally, in contrast to the common[4] test-train split, we adjust the splitting of the dataset to the actual situation. We use historical data to make predictions about future employments, reflecting the real-world use of administrative data in operational placement processes. Therefore, we take the years between 2012 and 2017 as the train set and 2018 as the test set. As a robustness check, we took several other time periods for a test-train split. We find that using data from 2012 on minimizes the error rate which is consistent with the point in time when the structural break occurs.[5] Fig. 1.2 shows the different classification errors according to the different time periods. Naturally, this procedure for finding the optimal time period for the train set could be repeated every year. This would guarantee that the train set is always chosen so that it minimizes the out-of-sample error for the preceding year.

We begin with calculations based on the total flows. The train set contains 12,054,684 observations while the test set contains 2,143,097 observations. This setup allows us to generate predictions for unemployed persons capturing all flows into employment. For the purpose of model comparison, the error rate for the total flows model is additionally

[4]Generally, as a rule of thumb, 66 percent of the data is used for training a model and 33 percent is used for testing it.

[5]We obtain the same results when taking the period up to 2016 as a train set and the years 2017 and 2018 as test set. Thus, in general, taking data from 2012 on is the best choice when the year of starting a new job is taken into account

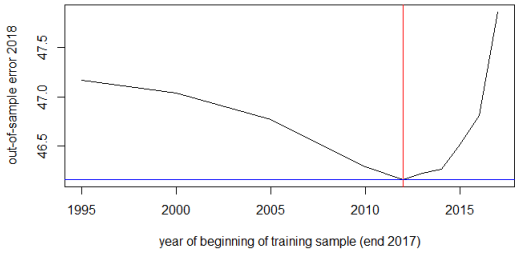


Figure 1.2: Finding the optimal training period for minimizing the out-of-sample error.

Source: own calculations.
Out-of-sample (OOS) error [in percent] for models trained on samples starting in different years and ending in 2017. The red vertical line marks the optimal start year (2012) that minimizes the OOS error for 2018; the blue horizontal line indicates the corresponding minimum OOS error value.

computed only on the unemployment flow portion of the test data. These results are particularly relevant for the PES, because caseworkers there focus on unemployment flows. According to table 1.4, the RF error rate (15.32 %) being 5.21 percentage points lower than for OLS (16.14 %). When restricting attention to the unemployment flows subset within the total flows, the reduction becomes even more pronounced, as RF error rate (42.20 %) outperforms OLS (60.70 %) by 35.96 percent.

Table 1.4 also reports information on the model estimated with unemployment flows. The train set contains 1,435,474 observations while the test sample contains 740,266 observations. Again, RF error rates (46.06 %) are than OLS errors (47.25 %), which stands for a reduction of 2.55 percent.

Finally, RF is included alongside conventional methods such as OLS and other ML techniques. Compared to OLS, RF can provide superior predictive performance in terms of out-of-sample error, particularly in high-dimensional settings with complex, nonlinear relationships. While RF requires computational resources for training, it is relatively robust to hyperparameter choices and does not require explicit specification of functional forms, making it well-suited for administrative datasets where variable interactions and nonlinearities are difficult to model explicitly. Thus, RF offers a practical balance of predictive accuracy and modeling flexibility. The total flows should be preferred for model training, as they expose the algorithms to a broader range of occupational transitions and categorical feature combinations, which improves learning in large administrative datasets. Furthermore, this avoids discrimination in the predictions due to the unemployment status.

Table 1.4: Error rates for OLS and RF

Condition (Train → Test)	**RF error**	**OLS error**
unemployment flow → unemployment flow	46.06%	47.25%
total flow → unemployment flow	**42.20%**	60.70%
total flow → total flow	15.32%	16.14%

Source: IEB and own calculations. Train: 2012–2017, Test: 2018. A prediction is correct if the occupation with the highest predicted probability matches the realized occupation. Unemployment flows are a subset from the total flows into employment; for the total flows two error rates are calculated – one for the total flows and one restricted to unemployment flows – to allow direct comparison.

1.5.2 Test-Train Split by Person

Next, we investigate model performance when the test-train split is performed at the individual level. Unlike the year-based split, this approach ensures that no person appears in both sets. We therefore randomly assign two-thirds of all individuals to the training set and one-third to the test set, using data from 2012 onwards.

The results for this setup are shown in table 1.5. For the total flows, RF error rate (16.29 %) clearly outperforms OLS (19.03 %), and thus being 15.52 percent lower. When we focus again on the unemployment flows subset contained within the total flows, RF (38.70 %) yields an error rate that is 35.63 percent lower than that of OLS (55.48 %).

We then repeat the analysis using the unemployment flows and see that RF (46.69 %) again performs better than OLS (50.03 %), reducing the prediction error by 6.91 percent. These results align closely with those obtained using the year-based split, indicating that the performance advantage of RF is robust to alternative sample constructions within the same administrative data structure, and underlining the conclusion that the total flows should be preferred for training the model.

Table 1.5: Prediction errors for OLS and RF

Condition (Train → Test)	**RF error**	**OLS error**
unemployment flow → unemployment flow	46.69%	55.48%
total flow → unemployment flow	**38.70%**	50.03%
total flow → total flow	16.29%	19.03%

Source: own calculations. Two-thirds of individuals are assigned to the train set, one-third to the test set. A prediction is correct if the occupation with the highest predicted probability matches the realized occupation. Unemployment flows are a subset from the total flows into employment; for the total flows two error rates are calculated – one for the total flows and one restricted to unemployment flows – to allow direct comparison.

1.5.3 Additional Methods

Logit provides another illustration of common methods. Here, we also estimate a single equation for each category. A ranked list can then be created by comparing the predicted probabilities. Here, the results (i.e. the list with suggestions) are almost identical to the OLS results. Although OLS yields very similar recommendations to logit despite not producing probabilities.
For the kNN algorithm. We obtain an out-of-sample error of 54.9 percent for $k = 2000$. Since we find that the optimal k is 16,000 and that there is a linear relationship for the underlying dataset the out-of-sample error could be reduced to 47.8 percent. RF still performs better and is computationally more efficient for large administrative datasets with high-dimensional categorical feature spaces than the KNN model. Comparing the results with corresponding OLS results shows, that the error rate is reduced by 2.3 percentage points. Thus, also for kNN, we find that machine learning performs better than microeconometric methods. We prefer RF, because the error rates are lower, it needs less computational power and is less time consuming.
Beyond the practical benefits of the paper, the main focus lies on the comparison of different estimation methods. Especially in ML, there are a lot of different methods someone can apply. Thus, we tried also Lasso, support vector machines (SVM) and neural networks (NN). We obtain different problems like the number of observations or independent variables and the computing time. Finally, we have to make some restrictions because estimating the whole train set and the specification of hyperparameters is not possible. The estimations work for instance for smaller train sets or less independent variables. Unfortunately the predictions errors are very high, so we decided to focus on the methods mentioned above.

1.5.4 An Illustrative Scenario: Hypothetical Effects of Using RF Instead of Logit

For illustrative purposes, we calculate the potential consequences in terms of unemployment rate of using the RF instead of common statistical methods. We take the best result for RF and the common methods (see Table 1.4[6]). Evidently, such calculations can only be carried out under assumptions in a hypothetical scenario. An extreme assumption would be that in a given month an unemployed person finds a job if they receive the top one recommendation (TOP 1) and otherwise they do not. However, clearly, not receiving the TOP 1 recommendation does not necessarily imply an unsuccessful job search. We only start with this assumption for illustrative purposes and provide a more plausible variant

[6]In this case the logit error rates are lightly higher than the OLS error rates. Since probabilities are required for the calculation, we use the logit results for the comparison with RF.

below.

In the three-digit case, the share of correct recommendations from RF is 18 percent higher than from logit, i.e. the job-finding rate would also be 18 percent higher. Equilibrium unemployment in the standard case of two labour market states (employed and unemployed) with job-finding rate f and separation rate s results in $u = \frac{s}{f+s}$. We use a standard calibration for the German labour market with monthly $f = 0.08$ and $s = 0.006$ (Hartmann et al., 2021). Note that since neither logit nor RF have been used in reality so far (and thus are not part of the data generating process), this can only be assumed to serve as a reference point. Increasing f by 18 percent reduces u by 1.0 percentage points.

Now, we replace the extreme assumption from above. As a measure of the proximity of the two recommendations from RF and logit, we can use their probability difference. This follows the idea that the sharper the distinction drawn by the RF model between two jobs, the higher the probability of the job search being more successful with a correct recommendation. In detail, both for the RF and the logit TOP 1, we sum the probabilities from the RF model (see equation 1.5) for the cases where RF TOP 1 is true and logit TOP 1 is not. The sum for logit TOP 1 is 35 percent of the sum for RF TOP 1. The 18 percent mentioned above is therefore scaled down, multiplying by $1 - 0.35$. Logically, the u-difference is also smaller, namely 0.3 percentage points. Naturally, the calculated difference is lower than above because incorrect recommendations are punished only in so far as the probability is lower than the RF TOP 1.

Note that the reduction refers to a pure comparison of two methods in order to illustrate the differences under assumptions. The real-world effect on job seekers' outcomes and caseworkers' decision-making can only be evaluated in a field experiment. Existing matching procedures in public employment services (PES) typically rely on rule-based comparisons between vacancy requirements and individual characteristics. Such systems could be limited because they depend on predefined classification schemes and often fail to capture the heterogeneity of the placement process on the labour market. In contrast, our ML approach allows the model to exploit complex, nonlinear patterns in administrative data, incorporate large numbers of variables simultaneously, and learn from observed flows into employment rather than relying solely on keyword matching. Thus, we would expect an ML-based system to contribute to more accurate and diverse recommendations. A field experiment is essential to find out to what extent the system affects job seekers and caseworkers in practice.

A closely related real-world example is a large-scale field experiment with the French PES (Behaghel et al., 2022). In this randomized trial, job seekers and firms were assigned to receive targeted job recommendations and prioritized visibility on the PES platform. The intervention generated substantial increases in applications, interview invitations, and subsequent hires, demonstrating that directing job seekers towards suitable vacancies can produce measurable improvements in matching outcomes. The study illustrates that

recommendation systems implemented within PES can generate significant positive effects by actively steering search behaviour rather than relying solely on unguided job search.

1.6 Robustness Checks

1.6.1 Check list of Suggestions

Because the empirical results can be used to advised job job seekers, not only the first-best suggestion (TOP 1)is relevant but the top N suggestions. The caseworker should be able to provide several alternative suggestions, allowing the job seeker to choose which job ist most suitable. This makes the decision process more flexible. The caseworker can consider the clients' preferences. For example, an individual may look for employment close to where they live, while for others the suggested jobs may provide new ideas. An unemployed person might also be encouraged to retrain. We also consider the top three categories (TOP 3). We find that the out-of-sample error is given by 12.71 percent for OLS and 11.20 percent for RF. Thus, the error rate for RF is 13.26 percent lower. This shows that the conclusions for the TOP 1 hold for the TOP 3, and job recommendations should be calculated with RF.

1.6.2 Extending the Number of Occupational Groups

For the caseworkers it is most informative if the occupations are divided into as specific subcategories as possible. Therefore, it would also be interesting for them to have job suggestions based on the four-digit category. Using the four-digit category results in an increase in the number of different classes from 144 to 696. Table 1.6 shows the results. Making predictions for the TOP 1 category leads to an almost identical in-sample and out-of-sample error of about 58 percent. Comparing this with the three-digit results shows that the error rate increases by about eight percentage points. The error rate for both samples for the TOP 3 is about 27 percent which means that the rate is about 14 percentage points higher than in the three-digit case. Due to capacity problems, it is not possible to estimate the four-digit model on the whole train set with RF. Thus, we take ten random samples of 30 percent.[7] After having completed the forecasts on the remaining sub-sample and the whole test set, we compute the error rate in the same way as in the three-digit case. Note that some of the categories with a small number of observations drop out when only a 30 percent sample is used. We also obtain similar results for all sub-sample estimations. This indicates that taking a 30 percent sample is sufficient to answer the research question for the majority of jobs. In order to map occupations that do not have many entries, it would nevertheless be advantageous to be able to estimate the full model. The in-sample

[7] 30 percent is the maximum amount of data for the machine to be able to compute

error for the first best category is 27 percent and the out-of-sample error is 57 percent. Looking at TOP 3 reduces the in-sample error to 2.7 percent and the out-of-sample error to 24.8 percent. Thus, compared to the three-digit estimation, the in-sample error increases by five percentage points, while the out-of-sample error increases by ten percentage points. For the TOP 3, the in-sample error is two percentage points lower, while the out-of-sample error is 13 percentage points lower than in the four-digit case.
In summary, OLS performs very consistently in any case. Neither in the three- or four-digit cases, nor comparing only the TOP 1 or top N categories results in a big difference between the error rates. Nevertheless, the error rates (especially the out-of-sample error rates) always lie above the RF results.
The out-of-sample error of the RF prediction of the first best category is 9.7 percent larger

Table 1.6: Prediction errors (TOP 1 and TOP 3) for the four-digit occupational groups for OLS and RF

Sample no.	**Out-of-sample error**
RF TOP 1	56.872%
RF TOP 3	24.795%
OLS TOP 1	58.47%
OLS TOP 3	26.85%

Train set: 2012–2017, Test: 2018. Source: own calculations.

in the four-digit case than in the three-digit case. Keeping in mind that the number of categories increases by 556, RF still handles a much larger class space efficiently.

1.6.3 Adding the Skill Level to the Job Categories

Another crucial aspect in the matching process is the skill level. Thus, we combine skill levels with the three digit job categories. For each job category there are up to four different skill levels. In the classification of occupations there is a differentiation between unskilled/ semi-skilled tasks, skilled tasks, complex tasks or highly complex tasks. This means that the algorithm has to distinguish between 430 different categories. Again, we compare RF and OLS and use the total flows of the years 2012 until 2017 to train our model and then make predictions for 2018. For RF the classification error is 17.47 percent while the OLS error is 87.61 percent. If we exclude transitions from unemployment to employment, the classification error for RF is 53.97 percent while the OLS error is 90.72 percent. This confirms our qualitative results also in case the skill level is considered. Indeed, this again shows that OLS performs worse as the model becomes more complex. Another positive side effect is that RF becomes relatively more efficient than OLS as the number of categories increases.

1.7 Limitations and Potential Biases

1.7.1 Discrimination

It is well known that algorithms can also involve discrimination against certain people or certain groups of persons. The most challenging aspects are the robustness of the algorithm, human rights and transparency and explainability of the model. Naturally, human decision making takes place on an individual basis. In contrast, the algorithm works based on the underlying data. Thus, stereotypes and prejudices can be picked up (Pandey and Caliskan, 2021). On the other hand, algorithms also have the potential to detect and prevent discrimination (Kleinberg et al., 2020).

Evidently, recommendations based on realised data may perpetuate certain labour market patterns that may be undesired. This so-called historical bias can be related to characteristics such as gender, nationality, region, marital status, children or age. Instead of excluding such variables entirely, we argue that they should be included in the estimation so that their influence can be quantified. Once estimated, their effect ca be neutralized in probability calculations if desired. In contrast, if the variables were already neglected during estimation, their explanatory power would be partly taken over by other correlated variables. It would then be impossible to control the effects and deciding about their usage. For instance, gender is strongly linked to qualification choice. However, for generating matching recommendations, one may prefer to estimate the pure qualification effect separately from gender.

For reasons of transparency, we illustrate the influence of gender, nationality and region characteristics. To this end, we simulate artificial variables by randomly drawing from the realised distribution, i.e., estimation is performed using the true data, but recommendations are generated based on the artificial data. First, we determine the out-of-sample errors based on artificial data and compare them with the errors from Table 1.4. We find that the results clearly deteriorate if the gender variable is perturbed. Thus, to achieve the aim of minimising the classification error, keeping the variables in the model is important. As argued above, in practice it is crucial to check if the algorithm is discriminatory. One way to deal with the issue of potential discrimination in practice might be to combine different job recommendations. In addition to the job recommendations produced with full information, the system would provide job recommendations based on neutralized information of each person. So, all variables that might cause discrimination are not included when the job recommendations are generated. Thus, caseworkers would receive recommendations maximising the matching probability, but would also be informed what changes if, e.g., the gender effect is neutralised. This could be taken into account as a suggestion for alternative paths in the labour market.

Besides the design of the algorithm and the results, minimising the behavioural bias is one of the most important points. In general, this sort of bias occurs because each person

reacts in a different way to suggestions made by algorithms. Both groups, caseworkers and job seekers have to be well-informed how the job recommendations are generated and how to use them. The two extreme scenarios, which means that the list is accepted without questioning or categorically rejected, must be prevented in any case. A potential concern may be that recommendations based on realised data only replicate the past behaviour of placement officers. However, the universe of labour market data we use to train the algorithms contains a variety of job findings that were not mediated by the employment service. Indeed, studies based on the German Job Vacancy Survey show that the public employment agency search is the decisive search channel in just 7.4 percent of all hires (Kubis, 2021).

1.7.2 Limitations

While the data has clear advantages for the application at hand, it is also subject to some minor limitations. In general, civil servants not subject to social security contributions are excluded, but the public sector is covered. Due to data protection regulations, only data that is needed for social security and labour market policy is collected, for example, or that must be collected for other reasons. Since the complete employment history, including the CV, as well as the knowledge and skills of the person are recorded, we can still access a very rich set of variables.

The dataset uses different data sources (ASU/XASU and BeH) for which the data is collected in different ways. Variables like age, place of residence or nationality are taken directly from the internal database. Other variables are filled in by job seekers, caseworkers or employers via self assessment. Naturally, this could lead to inconsistencies. The education variable is often reported differently by different employers. This is well known and therefore an established imputation procedure (Fitzenberger et al., 2006) is applied. While vocational training can be observed, there is no information on the field someone has a university degree. Nevertheless, we can observe if someone has a university degree or not. There are also variables that are only recorded for job seekers, like marital status and the number of children under 15 years. Assuming that marital status does not change that often, this information is extrapolated for the periods in employment. If a person becomes unemployed again, this information is updated again. The occupational group (5-digit) is recorded by the employer, which is not necessarily handled identically by all establishments.

In general, the variables are very well filled, since the recording is obligatory.

1.8 Conclusion

In this paper, we have shown that applying empirical methods, especially machine learning (ML) algorithms, can play an important role in improving the matching on the labour market. Using data that covers past matches and the corresponding information about the characteristics of people and jobs for estimating different models results in recommendations for job seekers. Thus, job suggestions from two sources, on the one hand, the placement officer who can use their experience and is able to acquire an overview of the job seekers' motivation and skills and on the other, the data-driven algorithm that makes suggestions based on a sample that covers the whole population could have a positive effect on the unemployment rate. Using administrative data on employment biographies covering the full universe of employment spells allows us to build statistical models that effectively map the current situation on the labour market. For estimating these models, we use OLS, logit, k-nearest-neighbours (kNN) and random forest (RF). We use two different samples for our estimations. The total flows that covers all people with jobs subject to social security and the unemployment flows that contains only transitions from unemployment to employment. We find that using the total flows minimizes the error rates more than using the unemployment flows. Furthermore, RF perform best in every modification. The in-sample error, in particular, differs markedly. Based on the total flows the error rates for transitions are about 30 percent lower for RF than for OLS. This clearly shows, that RF should be favoured above other methods.

Another advantage is the adaptability of the RF algorithm. Estimating the model with different modifications shows that RF always performs best. The ability of the algorithm to find interactions itself, for example, eliminates the need for a time-consuming manual modification of the model. Thus, the model can easily be adapted to different datasets without losing predictive power. Taken together, this study provides clear evidence of the advantages of machine learning for applied prediction tasks in large administrative datasets. We show that machine learning models not only outperform classical methods in terms of predictive accuracy, but that their advantage is fundamentally linked to their ability to convert large volumes of administratively collected, fixed-feature data into additional predictive information. In settings with fixed feature libraries and population-wide coverage, ML methods continue to improve as more data become available, while traditional models exhibit diminishing returns. Beyond accuracy gains, machine learning offers substantial benefits in terms of scalability, robustness, adaptability, and automation potential, which are central concerns in applied ML research. These findings underscore the relevance of ML for real-world applications using administrative data and highlight its value for both methodological development and practical deployment in labour market analytics.

To clarify the relevance of the results, we take a closer look at the difference between RF and OLS. Our aim is try to establish the importance of our results for the unemployment

rate. Naturally, our calculations have to rely on assumptions in order to determine an impact on a theoretical basis. Thus, our calculations are for illustrative purposes only. In terms of the unemployment rate, hypothetically, the advantage of RF over common methods would represent a difference of 0.3-1.0 percentage points.

The main goal of our predictions is to support the placement process in practice. This means that the traditional procedure that a job seeker gets some suggestions from a caseworker stays the same. Additionally, personal information on the job seeker can be fed into the system or taken directly from the employment biographies. Afterwards, the system provides a list of job recommendations individually for each job seeker based on the predictions described in this paper. Thus, personal preferences, the impression of the caseworker, and the algorithmically generated list of job recommendations jointly contribute to identifying suitable job offers which could further increase the job finding rate and lead to more stable employment relationships.

Concluding, we made contributions in several areas. We demonstrate how ML methods can be effectively deployed in large-scale administrative settings characterised by extreme heterogeneity and class imbalance. We develop and evaluate a job recommendation system that produces recommendations at the three- and four-digit occupation level. A key point of our approach is the availability of the full universe of employment and unemployment spells to construct training that optimizes the flows into employment. This strategy allows us to incorporate substantially more information by considering flows into employment. We provide practical guidance on implementing RF and kNN algorithms for estimating such data. Our empirical comparison of traditional econometric models and ML algorithms reveals large and economically meaningful performance differences and shows that RF consistently outperforms the other models.

Taken together, our findings highlight the strong potential of ML – especially RF – for policy-relevant labour market applications based on administrative data. Beyond job recommendations, the proposed framework can be adapted to other matching problems in labour economics and public policy where high-dimensional data and institutional constraints play a central role. From a policy perspective, our results suggest that the availability of administrative data and the application of ML algorithms can yield substantial efficiency gains in public employment services, enabling more targeted job recommendations and support for job seekers across the entire workforce.

Matching probabilities mainly cover the quantitative aspect; this means we cannot determine whether mismatch occurs. To control for this, we plan to extend the model by job match quality. The matching probabilities calculated in the present paper should be combined with quality measures like job stability and job duration. Indeed, one possibility for improving the model could be to add information about skills. The Federal Employment Agency also collected data on skills based on self-assessment . In the past, it has been shown that information about competencies can enhance the estimation model

(Chakravarty et al., 2020; Rentzsch and Steneva, 2020). Therefore, we intend to include the analysis of competencies in future research.

In a future paper, we aim to complement our analysis of matching probabilities by explicitly addressing the quality dimension of matches. In particular, we consider job stability and wages as key indicators of job match quality. Evidence for the importance of this dimension is also provided by recent work showing that ML methods reveal substantial heterogeneity in earnings losses following job displacement, and that job match quality plays a crucial role in this context (Gulyas and Pytka, 2025).

In future research, the underlying models could be further developed. While ML algorithms have already achieved clear progress using a data set of standard properties, incorporating further data, for example on specific competencies, may represent a promising approach to increase practical usefulness. Further, we plan to extend the current matching of job seekers and occupational categories to a matching of concrete jobs and job seekers. The model should also be considered thoroughly from a fairness perspective. Finally, it should be examined which prerequisites must be met for algorithm-based applications to be successfully introduced in practice.

Appendix - Chapter 1

1.A Number of observations per job category

Table 1.A.1 shows all 3-digit job categories and the corresponding number of observations for the full sample and the unemployment-to-employment transitions.

Table 1.A.1: Occupational categories — number of transitions (trans.) and total number of observations of jobs subject to social security (obs. comp.)

3-digit no.	Occupation	Trans.	Obs. comp.
13	Junior non-commissioned officers	6	77
14	Armed forces personnel in other ranks	10	121
12	Senior non-commissioned officers and higher	18	260
411	Occupations in mathematics and statistics	218	3,442
116	Occupations in vini-and viticulture	355	3,710
11	Commissioned officers	458	2,156
931	Occupations in product- and industrial design	488	4,629
421	Occupations in geology, geography and meteorology	521	5,209
912	Occupations in the humanities	536	4,310
114	Occupations in fishing	543	5,260
947	Technical and management occupations in museums and exhibitions	548	6,523
936	Occupations in musical instrument making	554	8,875
625	Sales occupations (retail) selling books, art, antiques, musical instruments, recordings or sheet music	575	8,672
824	Occupations in funeral services	655	10,819
822	Occupations providing nutritional advice or health counselling and occupations in wellness	731	9,311
423	Occupations in environmental protection management and environmental protection consulting	741	6,758

3-digit no.	Occupation	Trans.	Obs. comp.
933	Occupations in artisan craftwork and fine arts	750	9,700
712	Legislators and senior officials of special interest organisations	845	22,848
934	Artisans designing ceramics and glassware	905	14,869
532	Occupations in police and criminal investigation, jurisdiction and the penal institution	1,015	11,334
943	Presenters and entertainers	1,194	13,449
922	Occupations in public relations	1,394	14,240
815	Occupations in veterinary medicine and non-medical animal health practitioners	1,404	13,999
816	Occupations in psychology and non-medical psychotherapy	1,703	18,268
113	Occupations in horsekeeping	1,765	17,621
523	Aircraft pilots	1,796	22,194
211	Occupations in underground and surface mining and blasting engineering	2,041	75,535
833	Occupations in theology and church community work	2,126	46,217
941	Musicians, singers and conductors Occupations in environmental protection engineering	2,152	34,336
522	Drivers of vehicles in railway traffic	2,309	35,707
514	Service occupations in passenger traffic	2,388	26,313
511	Technical occupations in railway, aircraft and ship operation	2,489	52,430
524	Ship's officers and masters	2,568	36,065
512	Occupations in the inspection and maintenance of traffic infrastructure	2,589	25,331
935	Artisans working with metal	2,742	48,610
214	Occupations in industrial ceramic-making and -processing	2,885	59,890
946	Occupations in stage, costume and prop design	2,999	10,986
634	Occupations in event organisation and management	3,085	34,743
613	Occupations in real estate and facility management	3,179	36,361
533	Occupations in occupational health and safety administration, public health authority, and disinfection	3,204	34,543
291	Occupations in beverage production	3,311	40,516
414	Occupations in physics	3,523	71,764
112	Occupations in animal husbandry	3,594	45,479
312	Occupations in surveying and cartography	3,771	50,632

3-digit no.	Occupation	Trans.	Obs. comp.
261	Occupations in mechatronics, automation and control technology	3,781	41,964
923	Occupations in publishing and media management	3,862	71,509
911	Occupations in philology	4,030	74,267
944	Occupations in theatre, film and television productions	4,123	23,346
433	Occupations in IT-network engineering, IT-coordination, IT-administration and IT-organisation	4,199	61,955
283	Occupations in leather- and fur-making and -processing	4,544	120,743
913	Occupations in economics	4,762	46,193
233	Occupations in photography and photographic technology	4,871	90,776
243	Occupations in treatment of metal surfaces	4,955	88,927
711	Managing directors and executive board members	5,056	68,874
432	Occupations in IT-system-analysis, IT-application-consulting and IT-sales	5,189	68,948
733	Occupations in media, documentation and information services	5,813	63,418
812	Laboratory occupations in medicine	5,900	107,984
213	Occupations in industrial glass-making and -processing	5,911	87,687
115	Occupations in animal care	6,024	69,927
434	Occupations in software development and programming	6,175	76,895
914	Occupations in economics	6,417	91,538
515	Occupations in traffic surveillance and control	6,477	107,314
281	Occupations in textile making	7,004	173,895
412	Occupations in biology	7,678	108,594
624	Sales occupations (retail) selling drugstore products, pharmaceuticals, medical supplies and healthcare goods	8,656	163,979
845	Driving, flying and sports instructors at educational institutions other than schools	9,113	122,790
842	Teachers for occupation-specific subjects at vocational schools and in-company instructors in vocational training	9,175	97,971
924	Occupations in editorial work and journalism	9,941	213,791
231	Technical occupations in paper-making and -processing and packaging	10,380	176,488
122	Occupations in floristry	10,489	139,359

3-digit no.	Occupation	Trans.	Obs. comp.
932	Occupations in interior design, visual marketing, and interior decoration	11,200	147,306
241	Occupations in metal-making	11,783	213,270
818	Occupations in pharmacy	12,491	180,538
245	Occupations in precision mechanics and tool making	12,604	299,508
825	Technical occupations in medicine, orthopaedic and rehabilitation	12,904	241,159
715	Occupations in human resources management and personnel service	12,949	124,624
731	Occupations in legal services, jurisdiction, and other officers of the court	13,083	164,694
631	Occupations in tourism and the sports (and fitness) industry	13,755	191,814
222	Occupations in colour coating and varnishing	14,008	179,903
723	Occupations in tax consultancy	14,193	245,535
844	Teachers at educational institutions other than schools (except driving, flying and sports instructors)	14,769	132,286
516	Management assistants in transport and logistics	15,130	239,075
843	Teachers and researcher at universities and colleges	15,548	303,146
234	Occupations in printing technology, print finishing, and book binding	15,811	318,150
232	Occupations in technical media design	17,426	271,166
282	Occupations in the production of clothing and other textile products	18,106	400,700
117	Occupations in forestry, hunting and landscape preservation	18,685	128,607
262	Technical occupations in energy technologies	18,933	200,060
272	Draftspersons, technical designers, and model makers	20,103	381,590
814	Occupations in human medicine and dentistry	20,299	232,594
945	Occupations in event technology, cinematography, and sound engineering	20,612	168,029
331	Floor layers	20,675	160,952
343	Occupations in building services and waste disposal	22,819	254,916
311	Occupations in construction scheduling and supervision, and architecture	26,379	342,979
341	Occupations in building services engineering	26,391	353,622
942	Actors, dancers, athletes and related occupations	26,841	253,592

3-digit no.	Occupation	Trans.	Obs. comp.
841	Teachers in schools of general education	26,854	189,484
212	Conditioning and processing of natural stone and minerals, production of building materials	27,900	201,234
732	Occupations in public administration	28,235	377,345
413	Occupations in chemistry	28,235	521,959
721	Occupations in insurance and financial services	29,539	884,677
623	Sales occupations (retail) selling foodstuffs	31,231	302,481
722	Occupations in accounting, controlling and auditing	31,730	556,684
817	Occupations in non-medical therapy and alternative medicine	31,886	354,357
622	Sales occupations (retail trade) selling clothing, electronic devices, furniture, motor vehicles and other durables	32,546	355,842
632	Occupations in hotels	37,625	393,951
273	Technical occupations in production planning and scheduling	38,236	633,787
821	Occupations in geriatric care	39,400	409,794
823	Occupations in body care	40,979	657,375
431	Occupations in computer science	42,568	575,724
271	Occupations in technical research and development	43,260	593,950
612	Trading occupations	43,663	935,017
221	Occupations in plastic- and rubber-making and -processing	46,516	564,279
111	Occupations in farming	48,258	316,228
525	Drivers and operators of construction and transportation vehicles and equipment	49,616	396,185
322	Occupations in civil engineering	51,935	433,222
611	Occupations in purchasing and sales	57,665	738,450
263	Occupations in electrical engineering	57,949	1,069,966
921	Occupations in advertising and marketing	59,381	450,172
832	Occupations in housekeeping and consumer counselling	61,436	636,524
333	Occupations in the interior construction and dry walling, insulation, carpentry, glazing, roller shutter and jalousie installation	63,813	522,652
342	Occupations in plumping, sanitation, heating, ventilating, and air conditioning	67,320	833,165
811	Doctors' receptionists and assistants	68,470	1,213,193

3-digit no.	Occupation	Trans.	Obs. comp.
292	Occupations in the production of foodstuffs, confectionery and tobacco products	75,211	967,994
233	Occupations in photography and photographic technology	76,780	90,776
713	Occupations in business organisation and strategy	87,772	1,230,525
813	Occupations in nursing, emergency medical services and obstetrics	98,193	1,503,772
531	Occupations in physical security, personal protection, fire protection and workplace safety	112,333	1,059,851
252	Technical occupations in the automotive, aeronautic, aerospace and ship building industries	115,268	1,727,693
244	Occupations in metal constructing and welding	124,952	1,177,706
242	Occupations in metalworking	129,373	1,400,857
332	Painters and varnishers, plasterers, occupations in the water-proofing of buildings, preservation of structures and wooden building components	130,860	869,263
251	Occupations in machine-building and -operating	143,800	1,932,378
121	Occupations in gardening	145,079	765,348
293	Cooking occupations	188,256	1,801,060
633	Gastronomy occupations	188,412	2,296,617
831	Occupations in education and social work, and pedagogic specialists in social care work	213,960	2,141,165
541	Occupations in cleaning services	271,847	4,006,181
521	Driver of vehicles in road traffic	287,743	2,760,959
321	Occupations in building constructions	359,606	2,538,077
621	Occupations in retail trade (without product specialisation)	378,624	5,299,499
513	Occupations in warehousing and logistics, in postal and other delivery services, and in cargo handling	487,495	4,805,261
714	Office clerks and secretaries	532,939	7,761,145

Source: own calculations.

Chapter 2

Quality-Adjusted Job Recommendations: A Machine Learning Approach Using Administrative Data

2.1 Introduction

Job seekers face a broad set of potential employment opportunities across occupations. In this setting, not only the probability of entering employment but also the quality of the resulting match is key for labour market outcomes and individual decision making. This point is emphasized in a strand of literature showing that search channels affect subsequent match quality. Several studies use survey data (see Simon and Warner (1992), Olsen et al. (2010)). Many studies find that search via newspapers, personal contacts, and especially online channels is associated with better job matches (see Holzer (1988), Freeman (2002)). Based on individual-level survey data, Mang (2012) shows that online job search yields better matches than search via newspapers, job agencies, or personal recommendations. Similarly, Cappellari and Tatsiramos (2015) analyse how friendship ties affect both job-finding probabilities and job match quality.

The main goal of this paper is to construct a ranked list of job recommendations for individual job seekers. These recommendations are designed to combine two dimensions that are often considered separately in practice: the probability of entering a given occupation and the expected quality of the resulting job match. At present, caseworkers typically rely on experience when making job suggestions. We aim to complement this process by providing model-based recommendations derived from historical employment biographies and by identifying additional occupational alternatives for current job seekers.

This paper makes three main contributions. More broadly, the paper contributes to the literature on labour market matching by moving beyond recommendation rules that focus exclusively on job entry. We show that occupational recommendations can be formulated

as a quality-adjusted allocation problem that jointly considers entry probabilities, expected stability, and expected earnings. First, we show that the choice of estimation method is closely tied to the structure of administrative labour market data. We demonstrate that estimator performance depends critically on the interaction between model architecture and data structure. Using rich administrative employment biographies, we show that flexible tree-based methods are particularly well suited to prediction tasks involving heterogenous employment histories, high-dimensional categorical information, and interaction-rich labour market environments.

Second, we develop a novel individual-level matching index that combines occupation-specific job-finding probabilities with predicted job stability and expected wages. By integrating both entry probabilities and post-match outcomes the index provides a unified measure of quality-adjusted employment prospects and directly links micro-level predictions to core labour market dynamics.

Third, we translate these predictions into a practical decision tool for employment services. The resulting ranking of occupations differs substantially from rankings based solely on matching probabilities, implying that focusing exclusively on job entry may lead to systematically different –and potentially less stable– placement outcomes.

A key feature of our setting is the nature of the data. We use rich administrative data on employment biographies in Germany, which combines a very large sample size with a high-dimensional set of worker and job characteristics. Such data are likely to contain nonlinearities, threshold effects, and interactions across employment histories, occupations, skill levels, and regional labour market conditions. This makes the choice of estimation method substantive rather than merely technical: model performance depends on how well a given method matches the structural properties of the data-generating process.

Building on these predictions, we construct a combined indicator of matching probability and job match quality. Specifically, we combine predicted wages and predicted job stability with the individual matching probability, following Mühlbauer and Weber (2026). All components are derived from individual characteristics observed in the administrative data. The resulting index yields a ranked list of occupational recommendations for each job seeker that incorporates both the likelihood of placement and the expected quality of the match. This information may help caseworkers broaden the set of occupations they consider and may help job seekers identify alternatives that offer more sustainable employment prospects.

Taken together, this paper develops and empirically validates an algorithmic framework for generating quality-adjusted occupational recommendations in a public employment service context. Rather than focusing solely on prediction accuracy, we show how predictive models can be translated into operational decision support tools that balance job-entry probabilities with expected job match quality.

Our study is related to several strands of literature. A broad literature analyses how institu-

tional settings and policy reforms affect job match quality, especially through the design of unemployment benefits (e.g., Caliendo et al. (2013)). In this literature, job match quality is often measured by wages or employment stability (e.g., Van Ours and Vodopivec (2006), Giannelli et al. (2013), van den Berg and Vikström (2014), Nekoei and Weber (2017), Gartner et al. (2022)).

We also connect to the growing literature applying ML methods to labour market research. In practice, ML is used to process web texts, address high-dimensional prediction problems, and support policy evaluations. For example, Amato et al. (2015) and Boselli et al. (2018) analyse vacancies via text classification, while other studies highlight the role of ML in complex labour market environments (see, e.g., Kianpisheh et al. (2012), Gulyas et al. (2019)).

The remainder of the paper is structured as follows. Section 2.2 describes the data and defines the variables used to measure job match quality. Section 2.3 presents the empirical strategy and the prediction results. Section 2.4 introduces the matching index and discusses its labour market interpretation. Section 2.5 links the matching index to labour market dynamics and section 2.6 concludes.

2.2 Setting

Our approach builds on work that focuses on predicting the probability of entering specific occupations (Mühlbauer and Weber, 2026). While that framework captures the likelihood of occupational transitions, it does not evaluate the quality of the resulting matches. The present paper extends this approach along two dimensions. First, we incorporate predictions of employment stability and expected wages as measures of job match quality. Second, we combine these components into an integrated recommendation framework that explicitly captures the trade-off between rapid job entry and longer-term employment outcomes.

2.2.1 Data

We use a ten percent sample of a large administrative dataset, namely, the Integrated Employment Biographies (IEB)[1] (see Antoni et al. (2016)). The IEB contains individuals in the German labour market who were covered by the social security system from 1980 onward and provides information on personal characteristics as well as job characteristics. Using this dataset allows us to observe hard facts regarding employment histories. The dataset combines different data sources. For the present research question, jobseeker histories (ASU/XASU) and employee histories (BeH) are required.

Reliable information on occupations is crucial for our analysis. In 2010, a new occupa-

[1] Version 14.00.00-190927

tional classification system (classification of occupations 2010) was introduced. From November 2011 onward, this new classification was applied to the data. Due to recoding, errors occurred in the occupation variable during the transition phase. Therefore, for our estimations, we use data from 2012 to 2018, analogously to Mühlbauer and Weber (2026). The dataset allows us to include information on gender, nationality, education, age, federal state of residence, the number of days spent in unemployment before finding a new job, the occupational category of vocational training, children (children under 15 years old), previous skill levels, previous occupations, and tenure.

2.2.2 Defining Job Match Quality

In this paper, we use two indicators to define job match quality, namely, job stability and wages (see Nekoei and Weber (2017), Gartner et al. (2022)). Importantly, both indicators can be measured accurately using rich administrative process data.

Job Stability

We use the duration an individual is employed in an occupation as a measure of job stability. We distinguish between two different concepts. The first is the duration of employment in the same occupation (i.e., job category), which we call "occupation duration". Here, a change in the job category or a transition into unemployment is treated as an interruption. The second is the duration of employment without interruption by unemployment, which we call "employment duration". We define the binary outcome variable as follows:

$$\text{duration} = \begin{cases} \text{less than 6 months}, & \text{short-term employment} \\ \text{more than 6 months}, & \text{long-term employment.} \end{cases} \tag{2.1}$$

The six-month threshold is motivated by institutional features of the German labour market, where probation periods typically last up to six months. After this period, the Protection Against Dismissal Act (Kündigungsschutzgesetz) becomes applicable. Thus, six months is a practically relevant threshold that is also used by Heisz (1999), for example. Moreover, using a relatively short threshold implies that only observations within six months of the end of the observation window have to be excluded. More precisely, we use observations from 2012 to 2016 to estimate the model and observations from 2017 to evaluate it. As a robustness check, we additionally consider a one-year threshold, yielding qualitatively similar results.

For some individuals, the sample contains multiple observations. When counting employment days and days in the same occupation, both ongoing and terminated employment spells are included. The reported means therefore do not represent the true mean duration of completed employment spells, because the final duration of ongoing employment cannot

be observed.

Table 2.1 shows that the distribution of short-term and long-term employment differs

Table 2.1: Proportion of observations per job category [in percent]

	employment duration		occupation duration	
sample*	short-term	long-term	short-term	long-term
Train set	43.54	56.46	58.05	41.95
Test set	39.11	60.89	51.64	48.36

*Short-term and long-term employment are defined in Equation 2.1.
The table reports the shares of observations in each outcome class.
Source: own calculations

somewhat across the two stability concepts. For employment duration, long-term employment dominates in both the training and test sets, accounting for 56.46 percent and 60.89 percent of observations, respectively. In contrast, occupation duration exhibits a larger share of short-term outcomes, particularly in the training sample, where 58.05 percent of observations fall below the six-month threshold. The distributions are broadly comparable across the training and test sets, although the test sample shows slightly higher shares of long-term outcomes in both cases.

Wages

The second indicator of job match quality is the daily wage. Since the working-time information only allows us to distinguish between full-time and part-time employment, we restrict the analysis to full-time employment. Wages are reported as daily wages in the data, that is, annual earnings divided by the number of days in the corresponding year.
For data protection reasons, wages above the contribution limit for social security are right-censored. We therefore apply the imputation procedure described in Gartner (2005) (see 2.A). The wage distributions shown in table 2.2 are similar across the training and test samples. Mean daily wages increase slightly from 113.65 Euro in 2017 to 115.82 Euro in 2018, while median wages rise from 98.84 Euro to 101.02 Euro. The difference between mean and median wages, together with positive skewness values above two, indicates a right-skewed wage distribution driven by a relatively small number of high-wage observations. The number of observations and distinct persons is of similar magnitude, reflecting repeated observations for some individuals in the administrative data.

2.2.3 Explanatory Variables

The IEB data contain information on both personal and job characteristics, including age, gender, marital status, children, and residence at the federal-state level.

Table 2.2: Descriptive statistics for the wage data

	Train 2017	Test 2018
N observations	1,050,210	1,092,315
Mean wage	113.65	115.82
Standard deviation	48.72	49.81
Minimum	64.50	64.50
P1	65.11	65.21
P25	80.45	81.70
Median	98.84	101.02
P75	129.82	132.54
P99	288.50	294.76
Maximum	1,007.49	881.00
Skewness	2.16	2.10
N persons	856,636	896,990

Wages are measured as daily wages in Euro.
The sample is restricted to full-time employment.
Source: own calculations.

We also observe nationality. Naturally, the largest group consists of German nationals. For non-Germans, we distinguish among nationals from the main countries of origin of asylum seekers,[2] EU nationals, and non-EU nationals. Table 2.3 provides an overview of the explanatory variables. The data also contain information on education, vocational training, the skill level required in previous jobs, preceding occupations, and tenure.

2.3 Empirical Strategy and Results

2.3.1 Prediction of Job Stability

Model

To predict job stability, we solve a binary classification problem. The model is given by

$$\text{P(duration } > 6 \text{ months)} = f(\mathbf{X}, \mathbf{Y}), \tag{2.2}$$

where $\mathbf{X}$ is an $n \times i$ matrix containing i characteristics of every person and $\mathbf{Y}$ is an $n \times j$ matrix containing j characteristics of the jobs of every person. Furthermore, n is the number of observations (i.e., spells). The predicted probability refers to long-term employment.

[2]Here, we refer to people whose nationality is equal to one of the eight countries of origin with the highest number of asylum seekers in July 2020. These eight countries are Syria, Iraq, Afghanistan, Turkey, Nigeria, Iran, Eritrea, Somalia, and Georgia.

Table 2.3: Explanatory variables

Variable	**Description**
Gender	Female, male
Federal state	Nordrhein-Westfalen, Bayern, Schleswig-Holstein, Sachsen-Anhalt, Hessen, Baden-Württemberg, Brandenburg, Mecklenburg-Vorpommern, Thüringen, Sachsen, Niedersachsen, Bremen, Berlin, Hamburg, Saarland, Rheinland-Pfalz
Nationality	German, EU, 8 main migration countries, Europe without EU, remaining countries
Marital status	Single/lives alone, in relationship/married
Children	Minimum of one child under 15 years, no children/unknown
Education	No school leaving certificate, primary/lower school, intermediate school without vocational training, primary/lower school, intermediate school with vocational training, upper secondary school leaving certificate without vocational training, upper secondary school leaving certificate with vocational training, university of applied sciences, university
Skill level	Skill level required for preceding employment(s)
Preceding occupation	Occupational group(s) employed in before starting a new job
Vocational training	Occupational group of vocational training
Age	Age at the start of employment
Days in unemployment	Number of days unemployed before starting a new job

Source: IEB; own calculations ©IAB

Estimation

Because the aim of our analysis is to predict future employment outcomes using historical employment biographies, we apply a test-train split by year. The training set contains observations from 2012 to 2016, and the test set contains observations from 2017. This design mirrors the intended practical application, in which caseworkers would rely on predictions based on the most recent historical data.

We estimate the model using both common statistical methods, namely OLS and logit, and ML approaches. The classical benchmark is a logit model, which is commonly used for binary classification problems. We generate predictions for the test set and determine the optimal cutoff value for classifying the outcome. For both samples, the cutoff value is close to 50 percent.

At the same time, the structure of the administrative data suggests that the determinants of job stability are likely to be highly nonlinear and interactive. Whether a match dissolves quickly may depend on combinations of age, unemployment duration, previous occupation, skill level, family characteristics, and regional labour market conditions in ways that are difficult to represent parsimoniously in a conventional parametric model. We therefore additionally estimate tree-based methods, namely random forest and extreme gradient boosting (XGB). In our application, XGB performs best.

This result is consistent with the properties of the underlying data. Administrative employment biographies contain many categorical predictors, heterogeneous employment histories, and potentially strong interaction effects across worker and job characteristics. XGB is particularly well suited to this setting because it can flexibly capture nonlinearities, threshold effects, and interactions without requiring the researcher to impose them ex ante. We choose the optimal XGB specification by minimizing the classification error on the test set.[3] Given the high dimensionality of the data and the large number of potential hyperparameter combinations, we adopt a sequential tuning strategy. In each step, one hyperparameter is varied while the others are held fixed at their current values. The value that minimizes the classification error is retained and used as the new default. We repeat this procedure iteratively. Once improvements are confined to the third decimal place, an exhaustive grid search becomes computationally disproportionate relative to the expected gain in predictive performance. This approach represents a computationally feasible and empirically robust compromise, allowing us to explore the relevant parameter space without incurring prohibitive computational costs.

[3] Different error measures are available. For predicting job stability, we use the classification error as the relevant criterion for model selection.

Results

We evaluate predictive performance using the classification error (CE), we report further performance measures that are standard in classification analysis, namely the area under the ROC curve (AUC), precision, recall, the F1 score, and the Brier score. These measures provide complementary information on discrimination ability, classification quality, and probability calibration. To be more precise, CE captures overall prediction accuracy, AUC reflects ranking quality, precision and recall describe classification trade-offs, and the Brier score evaluates probability calibration. Because the proposed recommendation framework relies on both ranking and predicted probabilities, evaluating multiple metrics is particularly important in our setting.
Table 2.4 reports the predictive performance of logit and XGB. For both samples, XGB

Table 2.4: Predictive performance for job stability

Sample	Metric	Logit	XGB
Occupation duration	CE	27.79	24.49
	AUC	0.74	0.81
	Precision	0.73	0.78
	Recall	0.78	0.82
	F1	0.76	0.80
	Brier score	0.18	0.16
Employment duration	CE	21.99	19.97
	AUC	0.78	0.85
	Precision	0.80	0.84
	Recall	0.84	0.87
	F1	0.82	0.86
	Brier score	0.15	0.14

The models are evaluated on the test set. Besides classification error (CE), we report discrimination measures (AUC), classification quality measures (precision, recall, F1), and calibration performance (Brier score). Lower values indicate better performance for CE and Brier score, whereas higher values indicate better performance for AUC, precision, recall, and F1.
Source: own calculations © IAB

clearly outperforms logit across all performance dimensions. The reduction in classification error is accompanied by sizeable improvements in discrimination ability, as reflected in higher AUC values, as well as better classification quality measured by precision, recall, and F1 scores. The performance gains are not only statistically visible but also economically meaningful from a recommendation perspective. For occupation duration, XGB lowers the CE by 3.3 percentage points relative to logit, corresponding to roughly a 12 percent reduction in prediction error. For employment duration, the reduction amounts to about two percentage points. Given the large number of observations in the administrative data, these improvements translate into a substantial number of observations in the

administrative data, these improvements translate into a substantial number of additional correctly classified employment outcomes.

The increase in AUC is equally relevant. An improvement from 0.74 to 0.81 for occupation duration implies markedly better ranking performance, meaning that XGB is more effective in distinguishing between employment relationships with higher and lower risks of early separation. This property is particularly important in a recommendation setting, where occupations are ultimately ranked rather than simply classified.

Lower Brier scores further indicate superior probability calibration. This matters because the matching index relies directly on predicted probabilities rather than only binary classifications. Well-calibrated probabilities improve the comparability of ranking components and therefore contribute to a more reliable construction of the quality-adjusted recommendation index.

Taken together, the results suggest that the advantage of XGB is not confined to a single performance criterion. Rather, the flexible tree-based architecture consistently improves predictive accuracy, ranking performance, and probability calibration relative to the parametric benchmark.

To obtain deeper insights into the prediction results, we examine variable importance. Gain is defined as the relative contribution of a variable to the model, aggregated over all trees. In both XGB specifications, marital status, age, and the number of days in unemployment before reemployment emerge among the most important predictors of job stability.

2.3.2 Prediction of Wages

Model and Estimation

Predicting wages is a regression problem. The model is given by

$$E[\ln(\text{wage})|\text{X, Y}] = f(\mathbf{X}, \mathbf{Y}). \tag{2.3}$$

We solve this regression problem using both OLS and XGB. OLS provides the standard benchmark because wage equations are often modeled as approximately linear in log wages. At the same time, the richness of the administrative data suggests that nonlinearities and interactions may matter, so that a more flexible regression method may yield better predictive performance.

The data structure is again highly relevant for the choice of the estimation method. Daily wages are shaped by a large number of worker and job characteristics, and the effects of these characteristics are unlikely to be purely linear or additively separable. Returns to education may differ across occupations, the effect of previous unemployment may vary across skill groups, and regional labour market conditions may interact with worker characteristics in complex ways. In such a setting, a flexible nonparametric regression

method can be advantageous because it can approximate these patterns without requiring explicit manual specification.

We therefore compare OLS with XGB as a regression algorithm. The tuning procedure follows the same logic as in the classification problem described above. To mirror the intended practical application, we again implement a temporal train–test split. Observations from 2017 are used for model training, while observations from 2018 are reserved for out-of-sample evaluation.

Results

To evaluate predictive performance, we use four standard regression measures, namely, the mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and the coefficient of determination (R^2). As a benchmark, we first estimate the linear regression model

$$\ln(\text{wage}) = \mathbf{X}\beta + \mathbf{Y}\gamma + \epsilon, \tag{2.4}$$

where β and γ contain the coefficients of $\mathbf{X}$ and $\mathbf{Y}$, respectively, and ϵ denotes the error term.

In the next step, we estimate XGB as a regression model. The regression metrics capture complementary dimensions of prediction quality. RMSE penalizes larger prediction errors more strongly, MAE reflects average absolute deviations, and R-squared summarizes explained variation. Using several measures is important because the recommendation system ultimately depends on predicted wage levels rather than a single goodness-of-fit criterion. Table 2.5 reports predictive performance for log daily wages.

To improve interpretability, Table 2.6 reports the corresponding prediction errors trans-

Table 2.5: Predictive performance for log daily wages

Metric	OLS	XGB
MSE	0.0624	0.0586
RMSE	0.2499	0.2420
MAE	0.1927	0.1876
R^2	0.5061	0.5367

Lower values indicate better predictive performance for MSE, RMSE, and MAE. Higher values indicate better explanatory power for R^2.
Source: own calculations © IAB

formed into Euro terms. The results show that XGB also outperforms OLS in the wage prediction task. Both the MSE and the MAE are lower for XGB than for OLS, both in logs and in Euro terms. RMSE and R^2 lead to the same conclusion. Thus, the superiority of

Table 2.6: Predictive performance for daily wages

Metric	OLS	XGB
MSE	1212.02	1152.09
RMSE	34.81	33.94
MAE	23.11	22.54
R^2	0.4893	0.5145

This table shows the transformation of the results for log(wage) into Euro terms to improve economic interpretability.
Source: own calculations © IAB

XGB in our application is not confined to classification, but also extends to regression. The magnitude of the prediction improvements is moderate but economically relevant. In Euro terms, XGB reduces RMSE from 34.81 to 33.94 Euro and lowers MAE from 23.11 to 22.54 Euro. While the absolute differences may appear modest at the individual level, they become meaningful in a recommendation framework that generates predictions across a large number of occupations for many job seekers.
The increase in the coefficient of determiniation from 0.49 to 0.51 also suggests that flexible methods capture additional structure in the wage-generating process beyond conventional linear specifications. Although a substantial share of wage variation remains unexplained, this is unsurprising given the complexity of individual wage formation in administrative labour market data.
Put differently, the superiority of XGB is closely linked to the structure of the underlying administrative data: the larger and richer the predictor space, the more likely it is that flexible methods can uncover relevant patterns that remain only imperfectly captured by linear regression.

2.3.3 Data Structure and Implications for Estimation Methods

The structure of the underlying administrative data has direct implications for the choice of the estimation method. The Integrated Employment Biographies combine several features that are empirically valuable but methodologically demanding: a very large number of observations, repeated employment spells for some individuals, a rich set of worker and job characteristics, and a high-dimensional predictor space with many categorical variables describing occupations, education, skill levels, regions, and employment histories. In addition, the economic relationships of interest are unlikely to be purely linear or additively separable. Job stability and wages are plausibly shaped by thresholds, nonlinearities, and interactions. For example, the effect of unemployment duration may differ across age groups, the return to education may vary across occupations, and the influence of previous employment histories may depend on regional labour market conditions.

These features matter because different estimators impose different restrictions on how information in the data can be translated into predictions. Classical econometric models such as logit and OLS are useful benchmark methods because they are transparent, computationally convenient, and familiar in applied labour economics. At the same time, they rely on comparatively restrictive functional-form assumptions. In the present context, this means that nonlinearities and interactions either have to be specified ex ante by the researcher or remain unmodeled. If the true relationship between worker characteristics, job characteristics, and labour market outcomes is more complex than the imposed functional form, these models may suffer from specification bias.

By contrast, tree-based methods are particularly well suited for the structure of the underlying data. XGB can flexibly approximate nonlinear relationships, detect interaction effects automatically, and capture threshold-type mechanisms without requiring the researcher to specify them in advance. This is especially relevant in administrative labour market data, where many predictors are discrete, high-dimensional, and economically likely to interact. The method is also well suited to very large samples because the abundance of observations allows the algorithm to learn complex patterns while limiting the variance costs that more flexible methods might otherwise entail. In our setting, the empirical advantage of XGB is therefore not simply predictive flexibility in the abstract, but the fact that its statistical architecture closely matches the structural properties of the data.

The suitability of the estimator differs somewhat across tasks, but the same basic logic applies in both cases. For job stability, the outcome is binary and the underlying process is likely governed by interaction-rich and threshold-driven mechanisms. Whether a job ends quickly may depend on combinations of prior unemployment, age, occupation, tenure, family background, and regional conditions in ways that are difficult to capture parsimoniously in a standard logit specification. This is precisely the type of problem for which boosted trees are well suited.

For wages, the prediction target is continuous, but the same complexity of the predictor space remains relevant. Wage formation depends on heterogeneous returns to human capital, occupation-specific structures, and interaction effects across worker and job characteristics. XGB can account for this complexity in regression mode without imposing a globally linear conditional mean. Our results confirm that this flexibility matters empirically: XGB also dominates OLS in the regression setting.

More generally, our findings imply that method choice should be guided by the interaction between estimator properties and data structure. In datasets such as the IEB, where the predictor space is large, heterogeneous, and interaction-rich, flexible methods are particularly attractive for both classification and regression tasks. Conventional parametric models remain useful benchmarks, but in our setting they are dominated by an estimator whose architecture is more closely aligned with the structure of the underlying labour market data.

2.3.4 Alternative Machine Learning Estimation Methods

In addition to tree-based methods, we also considered alternative ML approaches including artificial neural networks (NNs), k-nearest neighbours (KNN), and support vector machines (SVMs). While these methods are widely used and well suited for many prediction problems, their application in our setting is more challenging due to the specific structure of the administrative labour market data.

First, NNs offer a high degree of flexibility and are capable of approximating complex nonlinear relationships. However, in our application, their implementation proved to be computationally demanding. The combination of a very large number of observations and a high-dimensional feature space requires extensive hyperparameter tuning and substantial computational resources. In practice, even relatively simple network architectures led to long training times, while predictive performance remained not better than XGB. This suggests that, in our setting, the additional flexibility of neural networks does not translate into clear empirical gains relative to their computational costs.

Second, KNN relies on distance-based similarity measures to generate predictions. In high-dimensional settings with a very large number of observations, defining an appropriate distance metric becomes difficult. The administrative data used in this study combine heterogeneous features such as occupation, education, employment history, and personal characteristics, which are not easily mapped into a single meaningful notion of distance. As a result, while KNN is conceptually simple, its practical implementation is less straightforward in this context and tends to be less effective than tree-based approaches in capturing the relevant structure of the data.

Third, SVMs are a powerful class of methods that can perform well in a range of classification tasks. However, their performance depends critically on the choice of kernel functions and tuning parameters. In large-scale datasets with complex nonlinearities and many categorical predictors, this tuning process becomes computationally intensive. In our application, SVMs yielded reasonable results, but did not match the predictive accuracy and computational efficiency achieved by XGB.

The relative advantage of tree-based methods can be understood in relation to the characteristics of the data. The administrative employment biographies are high-dimensional, contain many categorical variables, and exhibit nonlinear and interaction-rich relationships between predictors and outcomes. Tree-based algorithms are particularly well suited to such settings because they naturally handle mixed data types, capture interaction effects and threshold structures, and remain computationally efficient in large samples.

2.4 Creating a Matching Index

Jobs with high wages, high stability, and high matching probabilities are naturally attractive to job seekers. Thus, recommendations based on a combination of these dimensions may be highly relevant for placement decisions. For example, Allen and Van der Velden (2001), Green and Zhu (2010), Mavromaras et al. (2013), and Pecoraro (2014) show that mismatch can lead to negative labour market outcomes such as wage penalties, absenteeism, or high turnover. Incorporating variables that describe job match quality is therefore crucial for generating meaningful job recommendations.

In Mühlbauer and Weber (2026), individual matching probabilities were predicted using the same type of employment-biography data as in the present paper. For each person, the authors estimate the probability of becoming employed in a certain occupation (i.e., a 3-digit occupation in the classification of occupations 2010). They find that random forest performs best for predicting matching probabilities in this multiclass classification problem with 144 occupational categories. In practice, however, it is useful for caseworkers and job seekers to complement this information with quality aspects. Combining both dimensions therefore yields a more comprehensive recommendation rule than considering matching probabilities alone.

We construct a matching index to provide such combined information. Job match quality consists of stability and wages, and both are combined with the matching probability. For the time being, we assign equal weights to matching probability and job match quality, and analyse this choice in the following. Within the quality block we assign equal weights to stability and wages. The matching index Q_{rs} is therefore defined as

$$\begin{aligned} Q_{rs} = \omega_P \cdot \tilde{P}(M_r = s \mid X_r, Y_r) + \omega_S \cdot \tilde{P}(\text{duration}_r > 6 \text{ months} \mid s, X_r, Y_r) \\ + \omega_W \cdot \tilde{E}[\text{wage}_r \mid s, X_r, Y_r], \end{aligned} \tag{2.5}$$

where $\tilde{P}$ and $\tilde{E}$ denote normalized components on a common scale, and $\omega_P + \omega_S + \omega_W = 1$. In our baseline specification, we set $\omega_P = 0.5$ and $\omega_S = 0.25 = \omega_W = 0.25$. Here, $r = 1, \ldots, N$ indexes observations, where N is the number of observations, and $s = 1, \ldots, S$ indexes occupations, where S is the number of occupations. The index combines three occupation-specific components for person r: the probability of entering occupation s, the probability of remaining employed in occupation s for more than six months, and the expected wage in occupation s. Thus, for each person, we obtain S index values. More concretely, the index links matching chances and job match quality in a single summary measure. Since matching probabilities, stability probabilities, and wages are measured in different units, the predicted components are expressed on a common scale before aggregation. This step serves only to ensure comparability across components within the index allowing us to rank occupations separately for each person.

Naturally, there is no universally optimal index construction. Weights cannot be generalized

Table 2.7: Exchange ratios for different weights of quality variables and matching probabilities

weight	0.75 PROB	0.5 PROB	0.333 PROB	0.25 PROB	0.167 PROB
exchange ratio TOP1	-0.11	-0.57	-1.16	-1.30	-1.37
exchange ratio TOP2	-0.15	-0.33	-0.38	-0.72	-0.90
exchange ratio TOP3	-0.12	-0.24	-0.38	-0.42	-0.54

Source: own calculations © IAB

across individuals because preferences differ. Nevertheless, the weighting scheme can be explored systematically. To evaluate the practical implications of alternative weighting schemes, we calculate an exchange ratio. The exchange ratio summarizes the trade-off between improvements in expected job match quality and the corresponding reduction in matching probability. It therefore provides an intuitive measure of how much placement probability must be sacrificed to achieve better labour market outcomes. We compute the exchange ratio for the TOP 3 recommendations, defined as

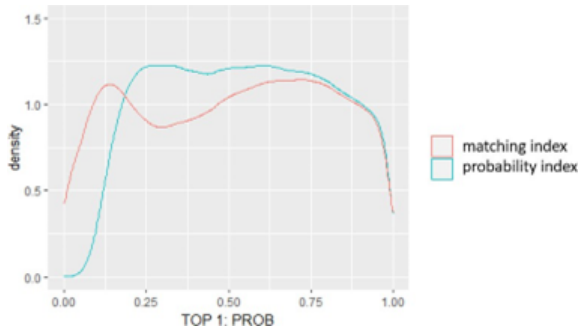
$$\frac{PROB_{\text{matching index}} - PROB_{\text{probability index, } w}}{QUAL_{\text{matching index}} - QUAL_{\text{probability index, } w}}, \tag{2.6}$$

where w denotes the specific weight combination. To implement this comparison, we calculate the mean wage, duration probability, and matching probability separately for the TOP 3 recommendations. To ensure consistency, with the construction of the matching index, the same scaled components are used in the calculation of the exchange ratios. The exchange ratio is then calculated according to Equation 2.6. The resulting pattern is similar across all TOP 3 recommendations. Table 2.7 shows that the exchange ratios deteriorate considerably once the weight on job match quality exceeds the 50 percent benchmark used in Equation 2.5.

The objective of the matching index is not to maximize agreement with observed occupational transitions. Such an objective would primarily reproduce realized matching outcomes and would naturally favor recommendations based solely on matching probabilities. Instead, the purpose of the index is to identify occupations that jointly offer favorable employment prospects and expected job quality. We therefore evaluate alternative weighting schemes in terms of the resulting trade-off between matching probability and job match quality. As shown in Table 2.7, assigning equal weights to matching probability and job match quality provides a balanced compromise: larger weights on the quality dimensions lead to increasingly large reductions in matching probability relative to the associated quality gains, while larger weights on matching probability move the index closer to a pure placement-probability ranking. We therefore use the equal-weight specification as our baseline.

To calculate the matching index according to Equation 2.5, we require predictions for

Figure 2.1: Cumulative density function for the matching index (equal weights) and probability index



each observation in 2017. For this purpose, wages are predicted based on a test set from 2016, which yields 144 predicted wages for each observation. In addition, the matching probabilities from Mühlbauer and Weber (2026) have to be re-estimated for 2017.
Mühlbauer and Weber (2026) define matching as a job seeker entering employment in a specific occupation. Using the same employment-biography data as in the current project, they predict matching probabilities with statistical methods. The model is given by

$$P(M_r = s \mid X, Y) = f(\mathbf{X}_r, \mathbf{Y}_r) \tag{2.7}$$

with $s = 1, \ldots, S$, where S is the number of occupations, $r = 1, \ldots, N$, and N is the number of observations. $\mathbf{M}$ is an $N \times 1$ vector containing the occupational group into which someone enters employment, so that M_r denotes the observed occupation for individual r. Furthermore, $\mathbf{X}_r$ is a $1 \times U$ vector containing the characteristics of person r, and $\mathbf{Y}_r$ is a $1 \times V$ vector containing job characteristics. Equation 2.7 therefore estimates the probability that a person enters a certain occupation as a function of observed personal and job characteristics. The authors find that random forest performs best in this multiclass classification setting with 144 occupational groups.
Since each estimation is based on slightly different samples of persons or spells, the matching index can only be calculated for observations that are available in all underlying samples. More precisely, only full-time jobs subject to social security contributions can be considered. This restriction reflects data availability and comparability requirements for wage measurement. In the end, the matching index can be calculated for 129,840 observations. While this limits direct generalizability to some labour market segments, the framework itself is not conceptually restricted to full-time employment and could be extended in future applications.
In analogy to the matching index, we also construct a matching probability index. This means that the matching probabilities for each person are scaled analogously to the matching index, so that the two indices can be compared directly. Figure 2.1 shows the density of the first-best recommendation for each index. As expected, once quality dimensions

Table 2.8: Mean values for probability index (prob) and matching index (qual)

	probability [in %]		stability [in %]		wage [in Euro]	
	prob	qual	prob	qual	prob	qual
TOP 1	54.68	50.83	53.39	60.48	99.74	100.38
TOP 2	10.73	8.76	11.69	20.63	98.55	107.90
TOP 3	5.64	3.24	9.80	21.78	98.10	108.59

Source: own calculations © IAB

are taken into account, some probability mass shifts to the left. This difference reflects a trade-off between immediate matching probability on the one hand and higher expected job stability and wages on the other hand.

Table 2.8 illustrates this trade-off by reporting the mean values of all three dimensions – probability, stability and, wage– for the first to third recommendation (TOP 3) under both the probability index and the matching index. On average, the matching probabilities are higher under the probability index, whereas wage and stability values are higher under the matching index. These differences arise precisely because the matching index incorporates job match quality and therefore changes the ranking of occupations in a meaningful way. We also take a closer look at the overlap in the TOP 3 recommendations. Ignoring the order, we find that in 56 percent of the cases, two recommendations differ across the two indices, while in 31 percent of the cases one recommendation differs. This indicates that matching probability remains important, but that the quality dimensions add economically meaningful information beyond pure placement chances.

2.5 Linking the Matching Index to Labour Market Dynamics

The matching index can be interpreted not only as a recommendation device but also as a reduced-form measure of labour market dynamics at the individual-occupation level. In standard search-and-matching models, labour market outcomes are governed by two central transition margins: the job-finding rate, which determines flows from unemployment into employment, and the separation rate, which determines flows from employment back into unemployment. The contribution of our index is that it combines information on both margins in a single metric and complements them with a measure of expected match surplus.

The matching-probability component, $P(M_r = s)$, is directly related to the individual job-finding rate into occupation s. It captures the likelihood that job seeker r enters employment in that occupational category. A recommendation rule based only on this component

would therefore prioritize occupations with a high immediate placement probability. In labour market terms, such a rule is primarily designed to maximize short-run exits from unemployment.

The stability component adds a second margin that is essential from an economic perspective. The probability of remaining employed beyond six months is closely related to the inverse of the short-term separation rate. A high predicted stability implies a low probability that a newly formed match dissolves quickly and results in a return to unemployment. Put differently, while the matching probability captures the inflow into employment, the stability measure captures the expected persistence of that inflow. This distinction is crucial because not all successful placements contribute equally to durable labour market integration.

The wage component provides a further dimension of match quality. In search-theoretic models, wages reflect, at least in part, the productivity and surplus associated with a worker-job match. Matches with higher expected wages are therefore not only preferable in terms of earnings but may also be more robust because they generate a larger joint surplus and are less likely to dissolve. In this sense, the wage prediction complements the stability prediction by capturing an economically meaningful dimension of match quality that is not exhausted by survival alone.

Taken together, the index combines three conceptually distinct but closely related objects: the probability of entering a given occupation, the probability of remaining employed in that occupation, and the expected wage conditional on that match. The resulting measure can therefore be interpreted as a quality-adjusted job-finding metric. Relative to a pure probability ranking, it reorders occupations by trading off immediate placement chances against expected post-entry outcomes.

This interpretation also clarifies the difference between the probability index and the matching index. The probability index focuses on the extensive margin of job entry. By construction, it tends to favor occupations with relatively high short-run job-finding probabilities, even if these occupations are associated with lower expected wages or a higher risk of early separation. The matching index, by contrast, incorporates the intensive margin of match quality. It may assign a lower rank to occupations with the highest immediate placement probability if these occupations imply weaker expected stability or lower wages, and it may favor occupations with slightly lower entry probabilities if they offer more durable and better-paid employment relationships.

From the perspective of labour market policy, this distinction is central. Employment agencies are often evaluated on rapid placements, which naturally creates a focus on the job-finding margin. However, an exclusive emphasis on immediate exits from unemployment may lead to recommendations that perform poorly once subsequent separations are taken into account. If a placement is followed by a quick return to unemployment, the observed job-finding success is only temporary. By incorporating stability and wage

information, the matching index shifts the decision criterion from short-run placement toward sustainable labour market integration.
This interpretation is consistent with the broader labour economics literature. Jovanovic (1979) emphasizes that match quality is revealed over time and his framework implies that match quality is a key determinant of employment stability and unemployment dynamics. Mortensen and Pissarides (1994) show that search frictions and matching efficiency jointly shape unemployment dynamics and turnover. Empirical work on unemployment insurance and search duration likewise finds that better job matches are associated with higher wages and more stable employment relationships (e.g., Nekoei and Weber (2017), Caliendo et al. (2013)). Against this background, our index can be understood as an operational tool that brings these labour market mechanisms into the recommendation process.
In practical terms, recommendations based on the matching index may in some cases imply a modest reduction in immediate placement probabilities relative to recommendations based purely on matching probabilities. However, this short-run trade-off is the counterpart of an expected gain in match stability and earnings. Economically, this means that the index is designed to lower expected short-term separations and reduce repeated unemployment spells, even if it does not always maximize the probability of the very first transition into employment. In the medium run, this should improve cumulative employment duration and expected earnings, which is precisely the notion of labour market success that is most relevant from both an individual and a policy perspective.

2.6 Conclusion

In this paper, we analyse whether ML methods can improve the prediction of job match quality in a large administrative labour market dataset. We measure job match quality by two outcomes, job stability and wages. The underlying data consist of rich employment biographies with facts on both workers and jobs. This setting provides an informative basis for prediction, but it also creates a demanding methodological environment because the predictor space is large, heterogeneous, and likely to contain nonlinearities, thresholds, and interaction effects.
Our estimations are based on historical data and reflect the structure of the labour market during the observation period. As labour market conditions evolve, periodic re-estimation is required to maintain predictive accuracy. The framework is therefore inherently dynamic and designed to adapt to changing economic conditions. Our results show that the choice of estimation method is closely related to the structure of the data. For job stability, we compare logit and XGB. We distinguish between employment duration and occupation duration. In both cases, XGB substantially reduces the classification error relative to logit. The advantage of XGB is particularly pronounced in the employment-duration sample. This suggests that flexible tree-based methods are well suited to binary prediction problems

in administrative labour market data when the underlying relationships are nonlinear and interaction-rich.

For wage prediction, we compare OLS and XGB in a regression setting. Here, too, XGB yields lower prediction errors than the classical benchmark, both in terms of MSE and MAE. Thus, the superiority of XGB in our application is not confined to classification, but also extends to regression. This finding supports the view that the wage-generating process in large administrative datasets is complex enough for flexible methods to outperform linear specifications.

Taken together, our findings also show that common statistical methods and alternative ML methods are generally applicable, tree-based approaches are particularly well suited for the combination of data characteristics and prediction tasks considered in this study.

A central contribution of the paper is to link these predictions to a practical recommendation framework. We combine predicted wages and predicted job stability with the occupation-specific matching probability from Mühlbauer and Weber (2026) to construct a matching index. This index produces a ranked list of occupational recommendations for each individual and therefore translates the prediction results into an operational placement tool. Our results show that recommendations based on the matching index differ substantially from those based solely on matching probabilities. This demonstrates that job match quality contains economically meaningful information beyond immediate placement chances.

The labour market interpretation of the index is equally important. The matching-probability component captures the job-finding margin, while the stability component is closely related to the inverse of the short-term separation rate. The wage component captures an additional dimension of expected match surplus. The matching index can therefore be interpreted as a quality-adjusted job-finding metric that balances immediate employment entry against subsequent employment stability and earnings prospects. More generally, the framework illustrated how predictive methods can be integrated into economically interpretable recommendation systems that explicitly connect labour market prediction tasks to search-and-matching mechanisms. In this sense, it shifts the focus from rapid placement toward sustainable labour market integration.

In practical terms, this information may be valuable for caseworkers in employment agencies. The index can complement their own assessments by identifying occupational alternatives that may offer slightly lower short-run placement probabilities but better medium-run outcomes in terms of stability and wages. More broadly, the results suggest that ML methods can be particularly useful in labour market applications where large administrative datasets meet complex outcome-generating processes.

The use of predictive models in public employment services raises important fairness concerns, including the potential reproduction of existing labour market inequalities. Recent research on algorithmic fairness shows that AI systems can reproduce and amplify patterns

of discrimination (Buyl et al. (2022); Poe and El Mestari (2024)). For example, Bied et al. (2023a) conduct a fairness analysis for a comparable AI-supported matching project in the French public employment service, focusing primarily on gender-based disparities. Mühlbauer and Weber (2026) and Bauer et al. (2025) discuss one approach to this issue. Because potentially sensitive variables may be correlated with other observed characteristics, discrimination cannot be addressed simply by omitting a small set of variables. A more promising approach is to measure the influence of critical variables and then allow users, such as caseworkers or job seekers, to decide which information should be used or displayed.

In practical implementation, the system is designed as a decision-support tool rather than an automated decision mechanism. Caseworkers are presented with a ranked list of occupations together with information on the underlying components allowing them to assess trade-offs transparently. The final decision remains with the caseworker and the job seeker, ensuring a human-in-the-loop approach.

Overall, the matching index makes the trade-off between rapid job entry and match quality explicit at the individual level. It shows that occupational recommendations should not be evaluated solely by immediate placement probabilities, but also by expected stability and earnings.

The analysis does not estimate the causal effect of using the recommendation system on labour market outcomes. Instead, it evaluates the predictive and operational validity of the proposed framework. Assessing causal impacts would require a field experiment or pilot implementation, which is beyond the scope of this study. The framework developed here should therefore be understood as a precondition for policy implementation. This distinction is particularly important in public sector applications, where predictive accuracy alone does not imply welfare improvements.

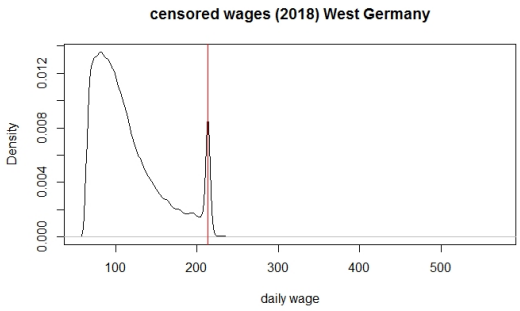
Appendix - Chapter 2

2.A Imputation of Wages above the Contribution Limit

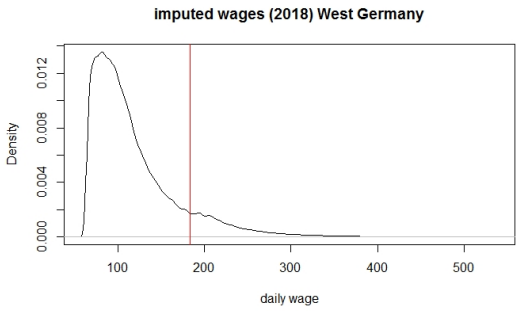
Wages are right censored due to the contribution limit in social security. For reasons of data protection, it is only necessary to capture wages until this threshold is reached. Consequently, all values that lie above the contribution limit are automatically set to this value. Additionally, there are some outliers in the data with extremely high daily wages. This can occur, for example, if the wage for a long-term project of artists is reported only for one day. In contrast, there are also some implausibly low wages due to errors in the calculation of daily wages. To eliminate these values, we define a lower limit for daily wages. From 2012 to 2014, the limit is two times the so-called marginal earnings threshold (Geringfügigkeitsgrenze). Starting in 2015, the statutory minimum wage can serve as a threshold.

In the next step, we have to estimate the wages that lie above the contribution limit. Since we have right-censored wages, we use a tobit model. To cover small inaccuracies in the calculation of daily wages, we use 95 percent of the contribution limit as a threshold for the tobit estimation. Furthermore, we estimate separately for year, gender and educational group. Since the contribution limit differs for East and West Germany, the estimations are conducted separately for each region. Therefore, the daily wage is predicted for 112 subsamples. To calculate the expected value, we apply the imputation of wages described in Gartner (2005). After this procedure, we are able to replace daily wages that lie above the contribution limit with imputed values. Figure 2.A.1 shows the density function of daily wages for 2018 before and after the imputation for West and East Germany. The graphs for the remaining years look similar.

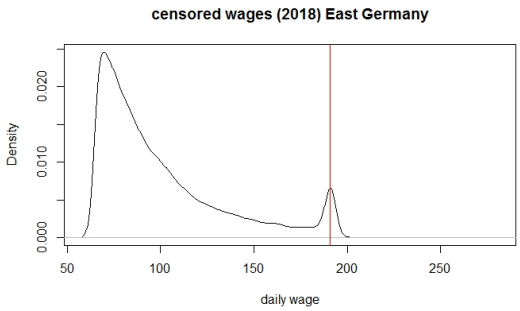
(a) Non-imputed daily wages, West Germany



(b) Imputed daily wages, West Germany



(c) Non-imputed daily wages, East Germany



(d) Imputed daily wages, East Germany

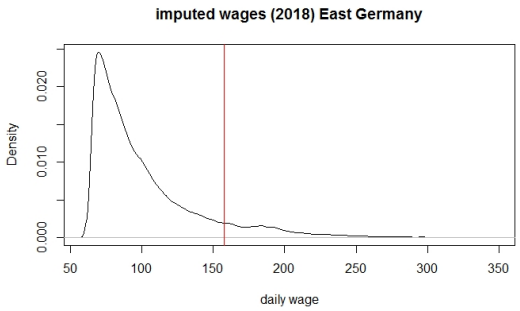


Figure 2.A.1: Density functions of daily wages in 2018 before and after imputation

Source: own calculations © IAB

Chapter 3

Distributional Intersectional Fairness in AI-Supported Job Matching: Evidence on Amplification, Trade-Offs, and Residual Risk

3.1 Introduction

Research on algorithmic fairness has accelerated as it has become unmistakably clear that AI systems can reproduce and amplify patterns of discrimination (Buyl et al., 2022; Poe and El Mestari, 2024). In many technical contexts, such "discrimination" is understood in a narrow, operational sense: machine learning (ML) models are expected to distinguish, rank, or classify, and differential treatment of cases is therefore both inherent and, at times, intended. Yet when claims that "AI discriminates" are interpreted in a social sciences register, the meaning shifts. Here, discrimination refers to the unequal distribution of opportunities or burdens along socially salient lines – such as gender, ethnicity, migration background, or age – shaped by historical inequalities and embedded power relations. The tension between these technical and social-scientific understandings of discrimination and fairness has become a defining challenge in the field (Zollo et al., 2025).

This challenge is particularly pronounced in employment-related AI systems. Recommendation, matching, and profiling models used in labour market contexts influence access to occupations, training opportunities, and employment trajectories. At the same time, labour markets themselves are already structured by complex patterns of social sorting, occupational concentration, and unequal opportunity. Fairness analysis in such settings therefore faces a dual problem: it must account both for inequalities embedded in the observed data and for distortions introduced through the predictive transformation performed by the model itself.

A growing literature addresses algorithmic fairness in employment and public-sector AI through technical mitigation strategies, governance frameworks, transparency requirements, and human oversight mechanisms (Sharma et al., 2025; Davidson et al., 2026; Broecke, 2023). Yet existing approaches overwhelmingly rely on isolated protected attributes, prediction-level metrics, or aggregate model evaluations. As a consequence, they provide only limited insight into how AI systems reshape outcome distributions across intersecting social groups. This limitation is particularly consequential in labour market matching, where occupational allocation patterns emerge through the interaction of multiple demographic and socio-economic characteristics.

The importance of intersectionality for understanding discrimination has long been emphasized in the humanities and social sciences. As formulated by Crenshaw (1991), disadvantage does not arise along single and separable dimensions, but through the interaction of multiple social positions. Nevertheless, as recent reviews of fairness research demonstrate, the systematic integration of intersectionality into technical fairness analysis remains limited (Chen et al., 2024; Mehrotra et al., 2022). Most empirical studies continue to rely on single-axis subgroup comparisons, despite extensive evidence that disadvantage frequently emerges through intersecting structures of inequality.

Our work is situated within a project that develops an AI-based recommendation system for public employment services (PES). Previous work has already outlined the ethical, sociotechnical, and discrimination-sensitive challenges associated with such systems (Bauer et al., 2025). The authors argued there that responsible system development requires not only normative reflection but also empirically grounded approaches capable of identifying and evaluating fairness risks in real-world deployment settings. Related work exists in adjacent contexts. For example, Bied et al. (2023a) conduct a fairness analysis for an AI-supported matching project in the French PES, focusing primarily on gender-based disparities. While such studies represent important advances toward fairness-aware system design, they also illustrate a broader limitation of current practice: fairness is usually assessed along single dimensions of discrimination, even though labour market outcomes are shaped by multiple and interacting social characteristics.

To address this gap, we develop a framework for evaluating intersectional fairness in AI-supported labour market matching. Our contribution is threefold. Conceptually, we shift the focus from individual prediction errors and single-attribute subgroup comparisons to the distributional consequences of algorithmic recommendations across intersectional social groups. Methodologically, we integrate distributional and predictive fairness measures with occupation-level diagnostics and counterfactual robustness analysis, thereby extending fairness evaluation beyond aggregate metrics. Empirically, we apply this framework to administrative labour market data covering more than three million observations and 141 occupations, providing evidence on the amplification of intersectional disparities, the fairness-performance trade-off associated with fairness-aware regularisation, and the

persistence of residual risks despite technical mitigation.

Empirically, we find that unconstrained models systematically amplify intersectional disparities in occupational allocations. Fairness-aware regularisation reduces predictive disparities across groups but introduces a pronounced fairness-performance trade-off and leaves residual distributional distortions concentrated in a small number of occupations. Counterfactual averaging (CF) further reduces demographic sensitivity while largely preserving predictive performance, suggesting a more favourable fairness-performance trade-off than direct fairness regularisation. Together, these findings highlight both the potential and the limits of technical mitigation: while counterfactual averaging substantially reduces demographic sensitivity at comparatively low utility costs, residual risks persist and require complementary monitoring, explainability, and human oversight.

These findings carry broader implications for fairness research and public-sector AI deployment. First, they suggest that fairness evaluation in employment recommendation systems cannot be reduced to a single metric or a single level of analysis. Second, they indicate that technical mitigation primarily redistributes or compresses fairness risks rather than removing them entirely. Third, they underline the importance of complementing fairness-aware modelling with explainability functions, monitoring mechanisms, human oversight, and governance structures that involve both caseworkers and job seekers as active participants in the recommendation process.

The paper proceeds as follows. Section 3.2 reviews the literature on algorithmic fairness and intersectionality. Section 3.3 introduces the data, modelling framework, and fairness measures. Sections 3.4 and 3.5 present the empirical results, including fairness-aware estimation, residual risk analysis, and counterfactual robustness. Section 3.6 discusses implications for fairness evaluation and AI governance in Public Employment Services. The final section concludes and outlines implications for explainability, participatory design, and human-centered fairness governance in PES.

3.2 Background and Related Literature

3.2.1 Existing Strategies on Algorithmic Fairness in Employment Research

The literature on algorithmic fairness in the employment context approaches fairness through technical, organizational, legal, and governance-oriented perspectives. Across these strands, there is broad agreement that AI-supported recruitment and placement systems may reproduce or intensify existing labour market inequalities. A substantial body of work focuses on the technical origins of algorithmic bias and corresponding mitigation strategies (Sharma et al., 2025). Within the fair machine learning paradigm, discriminatory outcomes are often attributed to biased training data, label constructions,

or feature choices, and addressed through techniques such as reweighting, fairness-aware loss functions, or constrained optimisation. While these approaches provide important tools for bias mitigation, fairness is typically conceptualized along isolated dimensions, such as single protected attributes or specific bias sources (Agbasiere and Nze-Igwe, 2025). Beyond technical interventions, organizational and governance-oriented research emphasizes that fair outcomes depend not only on model design but also on institutional arrangements, oversight mechanisms, and decision-making practices (Davidson et al., 2026). Studies in the context of public employment services highlight that profiling and matching systems can achieve high predictive performance while simultaneously producing substantial distributive effects across population groups (Desiere and Struyven, 2021; Bach et al., 2023). Rather than relying solely on technical fixes, this literature stresses the importance of transparency, accountability, and deliberation around fairness trade-offs. User-centered research further demonstrates that perceived fairness depends on how algorithmic outputs are interpreted and integrated into decision-making processes (Malin et al., 2025). In public employment services, caseworkers act as intermediaries who contextualize and potentially override algorithmic recommendations, underscoring that fairness emerges through the interaction of system design, organizational routines, and human judgment (Flügge, 2021; Körtner and Bonoli, 2022). At a more macro level, policy-oriented research and OECD reports situate algorithmic fairness within broader labour market and social policy debates, framing it as a governance challenge that requires institutional steering, transparency, and stakeholder participation (Broecke, 2023; Brioscú et al., 2024b). Taken together, existing strategies on algorithmic fairness in employment research provide valuable technical and institutional insights, but they rarely operationalize fairness as a multidimensional and relational phenomenon. Intersectional perspectives remain the exception rather than the rule. As a result, current approaches offer limited tools for capturing how intersecting social positions jointly shape algorithmic outcomes in employment systems. The present study addresses this gap by applying an explicitly intersectional analytical framework to AI-supported labour market matching.

3.2.2 Intersectionality as an Analytical Framework for Understanding Inequality

Intersectionality, as formulated by Crenshaw (1991), rests on the insight that discriminatory disadvantage does not occur along a single, neatly isolable axis. Rather, it crystallises through the convergence of multiple dimensions of social differentiation. Crenshaw's intervention addressed a structural blind spot in both antidiscrimination law and feminist theory: by treating categories such as "gender" or "race" as separate and mutually exclusive, institutions and scholarship obscured the ways in which individuals situated at their intersections experience forms of disadvantage that cannot be reduced to either category

alone. Her analysis of Black women's lived experiences demonstrated that discriminatory dynamics operate in complex ways – interlocking, overlapping, and mutually reinforcing – and thus reveal inequality as a fundamentally multidimensional phenomenon.

In contemporary intersectionality research more broadly, the central object of inquiry is the emergence and reproduction of social inequalities. Social inequality is understood as the uneven distribution of resources, power, and opportunities for social participation that constrains individuals' life chances over time. Individuals become marginalised when discriminatory practices—whether through formal exclusion, subtle sanctioning, or routine disadvantage – limit their ability to participate fully in social, economic, or political life. Intersectional analysis examines how such marginalisation emerges in particular contexts and how multiple axes of difference jointly shape these processes (Hill Collins and Bildge, 2020).

Which categories of differentiation matter under which conditions is not predetermined (Page and Yip, 2020). Intersectionality theory posits that experiences of discrimination and privilege are relationally shaped by social markers such as race, class, gender, dis/ability, sexuality, or age, while acknowledging that these markers do not operate uniformly across contexts. The analytical focus therefore lies in understanding how these differentiations interact – how they are embedded in, and sustained by, power relations and institutional arrangements that position groups unevenly within social hierarchies. The aim is not to establish a fixed list of relevant categories but to investigate how multiple dimensions of difference become salient within a given setting. Recognising that AI systems reproduce social inequalities and power structures underscores the importance of transferring this conceptual openness into the examination of algorithmic systems themselves. Because inequality manifests differently across domains—labour markets, education, housing, healthcare – the categories through which it should be analysed inevitably vary. Intersectional analysis thus resists rigid systematisation, insisting instead on attentiveness to context and to the historically contingent nature of social differentiation (Vethman et al., 2025). Making intersecting structures of inequality visible is not merely a descriptive task. By tracing how discriminatory mechanisms arise and interweave, intersectionality points directly to potential sites of intervention within relations of domination. Visibility becomes a precondition for change: only when the layered constitution of inequality is rendered analytically graspable can strategies for its disruption be developed (Bauer et al., 2025). This methodological orientation also explains why intersectionality has never been tied to a single research procedure. As Davis (2008) emphasises, there are no universal rules governing how intersectional research must proceed. The field is characterised by methodological pluralism –from qualitative ethnography to quantitative modelling – by an openness to new analytical tools capable of rendering complex constellations of power intelligible. Intersectionality is therefore best understood not as a fixed methodology, but as an epistemic commitment: to analysing inequality in its relational, situated, and

multidimensional form (Hancock, 2007). Our own analysis focuses on the dataset and the model we developed for use as a recommendation system in public employment services. Section 3.3.1 outlines in detail the composition of our data, the rationale for applying intersectional research to our analytical context, and the way in which we operationalise this perspective in our study.

3.2.3 Intersectional Analysis of Data and AI

While intersectionality is firmly established in the social sciences and humanities, its systematic integration into data science and AI research remains limited. Only recently have scholars begun to translate intersectional insights into methodological guidance for quantitative and computational contexts. A prominent contribution in this direction is provided by Nielsen et al. (2025), who conceptualize intersectionality not merely as a normative concern but as a methodological orientation that informs variable selection, modelling decisions, and interpretation across the entire analytical pipeline.

A central challenge identified in this literature concerns the context-dependent identification and operationalization of relevant social categories. Social attributes such as gender, age, or nationality are socially constructed and may interact in non-linear ways, complicating their representation in quantitative models. These difficulties are compounded by data limitations, proxy variables, and reduced statistical power in fine-grained subgroup analyses. Rather than treating these constraints as disqualifying, recent work emphasizes the importance of transparent handling of limitations and empirically grounded subgroup constructions. Our research design follows this logic by combining theory-informed variable selection with large-scale administrative data and by explicitly acknowledging representational gaps in the dataset. In parallel, a growing body of AI fairness research has operationalized intersectionality through subgroup-based performance comparisons, interaction-sensitive statistical testing, or extensions of established fairness concepts (An et al., 2025; Mohammadi et al., 2025; Wang et al., 2024). Seminal work such as Buolamwini and Gebru (2018) Gender Shades study demonstrates how accuracy disparities emerge at the intersection of gender and skin tone, while subsequent contributions propose statistical frameworks to detect interaction effects between multiple sensitive attributes (Suk and Han, 2023). Although these approaches have yielded important insights, they are often tailored to low-dimensional classification tasks and balanced subgroup definitions.

In contrast, the present study addresses a large-scale, highly imbalanced, multiclass matching setting. Rather than focusing on individual-level prediction errors, we adopt a distributional perspective that captures how intersectional groups are structurally positioned across occupational outcomes. This approach allows us to operationalize intersectionality in a way that is compatible with real-world labour market data and scalable decision-support systems.

While the preceding discussion has focused on methodological and technical challenges, it is important to emphasize that intersectionality is not merely a question of combining variables, but a fundamentally different way of conceptualizing inequality in data-driven systems. Intersectional analysis does not assume that social categories such as gender, age, or nationality exert independent and additive effects. Instead, it highlights that disadvantage emerges through their interaction and that these interactions are shaped by broader structural and institutional dynamics. From this perspective, intersectional group constructions are not neutral statistical artifacts, but representations of socially embedded positions. Their empirical manifestation in data reflects both structural constraints and individual decision-making processes, including self-selection into occupations. As a result, observed disparities cannot be interpreted purely as outcomes of discrimination, but must be understood as the joint result of preferences, opportunities, and institutional sorting mechanisms.

This has direct implications for fairness analysis in machine learning. First, it suggests that fairness cannot be adequately assessed by comparing marginal distributions or isolated group outcomes. Instead, it requires attention to how multiple dimensions of difference jointly structure outcome distributions. Second, it implies that attempts to eliminate all observed disparities through technical interventions may introduce new distortions if they ignore the underlying socio-economic processes that generate these patterns.

In large-scale applied settings such as labour market matching, this perspective calls for analytical approaches that remain sensitive to structural inequality while being compatible with high-dimensional data and complex prediction tasks. The distributional approach adopted in this paper follows this logic by focusing on how intersectional groups are positioned across outcomes, rather than on isolated prediction errors. This allows us to capture not only whether disparities exist, but how they are reshaped by the predictive model.

3.3 Empirical Study: Fairness Metrics, Estimation, and Model Selection

This section introduces the empirical framework used to analyse how occupational outcomes emerge from observed transition patterns and how these patterns interact with fairness considerations in AI-supported job matching. The empirical strategy distinguishes two complementary fairness dimensions. First, we evaluate *distributional fairness* by examining whether the model amplifies or attenuates intersectional concentration patterns across occupational outcomes. This dimension is measured using Jensen-Shannon divergence (JSD). Second, we evaluate *predictive fairness* by examining whether different intersectional groups receive equally accurate occupational recommendations conditional

on the true occupational outcome. This dimension is measured using the equal opportunity gap (EOG). The empirical strategy proceeds in four steps. First, we identify status-quo transition patterns into occupations that arise through self-selection and existing labour market sorting mechanisms and examine their intersectional structure in the observed data. Second, we estimate an unconstrained baseline model and compare its implied occupational distributions to these empirical patterns. Third, we introduce fairness-aware model variants and evaluate their effect on predictive fairness. Fourth, we select a preferred model under a retained-utility criterion.

3.3.1 Data and Intersectional Setup

The empirical analysis is based on administrative data from the Integrated Employment Biographies (IEB), which provide longitudinal labour market histories for the German workforce. The dataset is reconstructed to approximate the informational environment of the Placement and Counseling Information System (VerBIS), thereby aligning the estimation sample with the institutional context of AI-supported counseling. The analytical sample consists of a 10% random sample with complete information, comprising 3,174,204 individuals. The outcome variable is the occupational group after unemployment, measured at the three-digit level of the German Classification of Occupations 2010 (KldB 2010). While the classification distinguishes 144 occupations, we aggregate military occupations into a single category due to small sample sizes, resulting in $K = 141$ occupational classes. At the core of this analysis is an intersectional perspective. Rather than examining social characteristics in isolation, intersectionality emphasizes that multiple dimensions jointly shape labour market outcomes. Analyzing gender, age, or nationality separately risks masking important patterns that only emerge through their interaction. Intersectional analysis therefore captures compounded or non-additive effects and provides a more accurate representation of inequality structures. In this study, we focus on three variables with discriminatory potential: gender, nationality, and age. Gender is recorded as a binary variable (male/female), reflecting institutional data constraints. Age is observed with high precision and grouped into meaningful categories capturing life-cycle differences. Nationality serves as a proxy for broader stratification patterns in the absence of direct measures such as race. We distinguish between five groups: German, ten main migration countries, EU, Europe without EU and remaining countries. Individuals belonging to one the ten main migration countries are assigned exclusively to this category and are therefore not represented in one of the other groups. Each individual can belong to exactly one nationality group.

We define intersectional group membership as the joint realization of these three variables.

For individual i, let

$$G_i = \Big(\text{gender}_i, \text{nationality}_i, \text{age}_i\Big), \tag{3.1}$$

where gender_i, nationality_i, and age_i denote the categorical realizations of the respective attributes. The resulting set of possible combinations is denoted by $\mathcal{G}$, with cardinality $|\mathcal{G}| = 60$.

This definition implies that intersectional group membership enters the empirical analysis exclusively as a joint categorical construct. In particular, the empirical design does not include the underlying variables separately once the intersectional grouping is formed. This avoids implicitly imposing additive structures and ensures that all effects are captured at the level of the combined categories. Importantly, these attributes are not used as direct inputs in the main prediction model but are reserved for the fairness analysis. This separation ensures that predictive performance and distributional evaluation remain conceptually distinct.

To characterize baseline inequality, we analyse how individuals from different intersectional groups are distributed across occupations. Let $Y_i \in \{1, \dots, K\}$ denote the observed occupational outcome for individual i. The population share of intersectional group $g \in \mathcal{G}$ is denoted by $\pi_g = P(G_i = g)$, and the corresponding share within occupation k by $\pi_{g|k} = P(G_i = g \mid Y_i = k)$. Both quantities are estimated empirically from the data using relative frequencies and provide the benchmark against which model-induced changes are evaluated.

3.3.2 Prediction Framework

The predictive estimation framework builds directly on Mühlbauer and Weber (2024, 2026). In these studies, job recommendation is formulated as a large-scale multiclass prediction problem and estimated using gradient-boosted decision trees tailored to highly imbalanced occupational outcome structures. We adopt the established estimation framework and extend it through a multi-level intersectional fairness evaluation of the resulting occupational recommendations.

We model occupational outcomes as a multiclass prediction problem. Let X_i denote the feature vector of individual i. For each regularisation level $\lambda \geq 0$, the model produces predicted class probabilities

$$\hat{p}_{ik}(\lambda) = P(\hat{Y}_i = k \mid X_i; \lambda), \qquad k = 1, \dots, K. \tag{3.2}$$

The case $\lambda = 0$ corresponds to the unconstrained baseline model. Predicted occupational assignments are obtained as

$$\hat{Y}_i(\lambda) = \arg \max_{k \in \{1,\dots,K\}} \hat{p}_{ik}(\lambda). \tag{3.3}$$

For each occupation k, we define the predicted distribution of intersectional groups as

$$\hat{\pi}_{g|k}(\lambda) = P(G_i = g \mid \hat{Y}_i(\lambda) = k), \tag{3.4}$$

estimated empirically via relative frequencies.

3.3.3 Distributional Fairness and Baseline Model Comparison

To evaluate whether the prediction model reshapes observed intersectional allocation structures, we compare occupational group compositions in the observed data with those implied by model predictions.

We measure deviations between occupational and population compositions using the Jensen-Shannon divergence (JSD). For each occupation k, define

$$\mathrm{JSD}_k^{\mathrm{true}} = H\left(\frac{\pi_{\cdot|k} + \pi_{\cdot}}{2}\right) - \frac{1}{2}H(\pi_{\cdot|k}) - \frac{1}{2}H(\pi_{\cdot}), \tag{3.5}$$

where $\pi_{\cdot} = (\pi_g)_{g \in \mathcal{G}}$ and $\pi_{\cdot|k} = (\pi_{g|k})_{g \in \mathcal{G}}$ denote the full distributions over intersectional groups, and $H(p) = -\sum_{g \in \mathcal{G}} p_g \log p_g$ denotes Shannon entropy for a discrete distribution p on $\mathcal{G}$. JSD is used because the empirical object of interest is distributional rather than individual-level: we ask whether occupational group compositions become more or less intersectionally concentrated under model prediction. Unlike asymmetric divergence statistics such as the Kullback–Leibler divergence, JSD is symmetric and bounded, which is important in a highly imbalanced multiclass setting with sparse subgroup support in some occupations. This makes it suitable for comparing observed and predicted occupational compositions across many classes without giving disproportionate weight to very small probability cells. Higher values of $\mathrm{JSD}_k^{\mathrm{true}}$ indicate stronger deviations of an occupation's intersectional composition from the population benchmark. These deviations reflect the status-quo outcome of self-selection and structural sorting processes and therefore constitute the empirical reference point for evaluating model-induced distortions.

Appendix Table 3.A.2 documents the extent of these disparities across occupations. The results reveal substantial heterogeneity. Highly specialized occupations exhibit particularly strong deviations, indicating concentrated and selective transition patterns, whereas larger and more standardized occupational groups tend to display lower divergence. Overall, these findings highlight that occupational outcomes are shaped by structured transition dynamics across intersecting social dimensions. This underscores why an intersectional

framework is essential for a meaningful fairness evaluation.
Using the Jensen-Shannon divergence defined above, the predicted divergence is

$$\mathrm{JSD}_k^{\mathrm{pred}}(\lambda) = \mathrm{JSD}\Big(\hat{\pi}_{\cdot|k}(\lambda), \pi_{\cdot}\Big), \tag{3.6}$$

where $\hat{\pi}_{\cdot|k}(\lambda) = (\hat{\pi}_{g|k}(\lambda))_{g\in\mathcal{G}}$. The model-induced shift relative to the empirical benchmark is

$$\Delta\mathrm{JSD}_k(\lambda) = \mathrm{JSD}_k^{\mathrm{pred}}(\lambda) - \mathrm{JSD}_k^{\mathrm{true}}. \tag{3.7}$$

Positive values indicate that the model amplifies existing distributional disparities, whereas negative values indicate attenuation. Figure 3.1 shows, for each occupational group, the JSD between the intersectional group distribution within the occupation and the overall population distribution, comparing observed outcomes to model predictions in the unconstrained baseline model ($\lambda = 0$). The 45-degree line indicates perfect reproduction of the empirical distributional structure. Points above this line indicate occupations in which the model amplifies existing intersectional concentration patterns, whereas points below the line indicate attenuation. The figure shows that the unconstrained model does not simply reproduce the observed intersectional distributions. Instead, it induces systematic shifts across occupations. A non-trivial share of occupations lies above the 45-degree line, indicating prediction-induced amplification of existing disparities, while only a small number of occupations show attenuation. This pattern implies that the baseline model is not distribution-preserving but acts as a selective amplifier of existing labour market sorting mechanisms. To summarize the most substantively relevant effects, Table 3.1 reports the ten occupations with the largest positive and negative values of $\Delta\mathrm{JSD}_k(0)$. The results reveal a strongly asymmetric pattern. The magnitude of amplification substantially exceeds that of attenuation, with the largest positive shifts being considerably larger than the negative ones. This indicates that the unconstrained model systematically increases intersectional concentration in selected occupations. Amplification occurs both in occupations with already high structural concentration and in occupations with only moderate baseline divergence. This suggests that the model captures predictive patterns that reinforce and, in some cases, intensify existing sorting mechanisms. By contrast, attenuation effects are limited and generally small in magnitude. Even in occupations with high baseline divergence, the model rarely produces substantial reductions. The pronounced attenuation observed for economics occupations appears to be an isolated case rather than part of a broader pattern.
Overall, the baseline model acts as a selective amplifier of intersectional disparities. Rather than passively reflecting observed labour market structures, it actively reshapes them. This distinction is critical: it shows that fairness concerns arise not only from the data-generating process, but also from the predictive transformation applied by the model. The baseline

Table 3.1: Occupations with the largest amplification and attenuation of intersectional disparities ($\lambda = 0$)

Occupation	$\mathrm{JSD}_k^{\mathrm{true}}$	$\Delta\mathrm{JSD}_k(0)$
Largest amplifications		
Military	0.303	0.636
Legislators and senior officials of special interest organisations	0.523	0.362
Trading occupations	0.251	0.355
Occupations in publishing and media management	0.300	0.298
Technical occupations in paper-making and -processing and packaging	0.265	0.284
Occupations in artisan craftwork and fine arts	0.341	0.275
Occupations in philology	0.485	0.235
Occupations in mathematics and statistics	0.391	0.231
Sales occupations (retail) selling foodstuffs	0.306	0.230
Occupations in event organisation and management	0.236	0.228
Largest attenuations		
Occupations in economics	0.404	-0.404
Occupations in musical instrument making	0.548	-0.086
Aircraft pilots	0.618	-0.029
Occupations in the production of foodstuffs, confectionery and tobacco products	0.238	-0.009
Occupations in veterinary medicine and non-medical animal health practitioners	0.508	-0.006
Occupations in hotels	0.277	-0.004
Occupations in body care	0.410	-0.000
Occupations in horsekeeping	0.397	0.003
Painters and varnishers, plasterers, occupations in the waterproofing of buildings, preservation of structures and wooden building components	0.414	0.007
Technical occupations in medicine, orthopaedic and rehabilitation	0.202	0.010

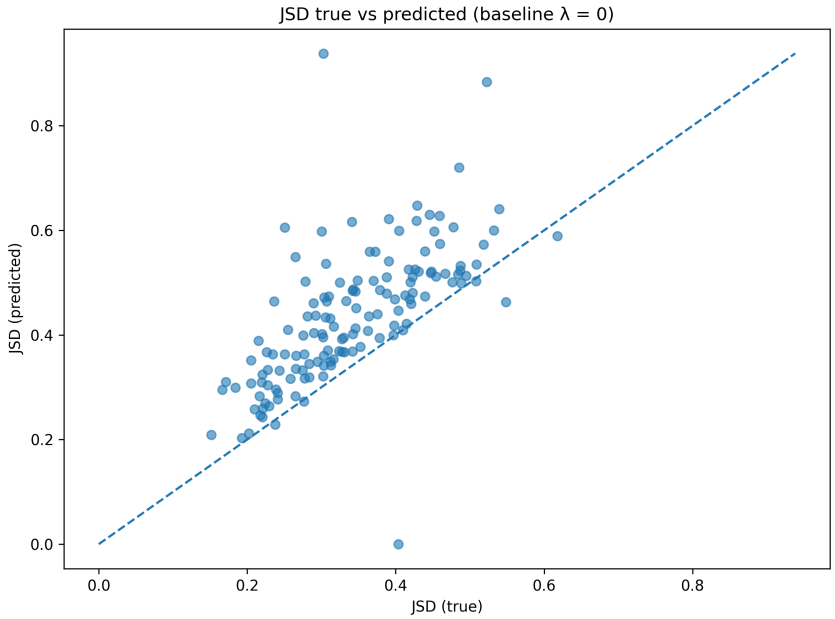


Figure 3.1: Occupation-level distributional shifts in the unconstrained baseline model. Values above the 45-degree line indicate amplification of observed intersectional disparities, while values below indicate attenuation.

comparison therefore provides the essential reference point for evaluating fairness-aware model adjustments in the subsequent analysis.

3.3.4 Predictive Fairness, Fairness-Aware Estimation, and Formal Metrics

To evaluate whether the prediction model reshapes observed intersectional allocation structures, we compare occupational group compositions in the observed data with those implied by predictions. The baseline analysis has shown that the unconstrained model does not merely reproduce observed labour market structures but actively reshapes them, often amplifying intersectional disparities. This raises the central question of whether such distortions can be mitigated through fairness-aware model design, and at what cost in terms of predictive performance.

We estimate a sequence of models indexed by $\lambda \in \Lambda \subset \mathbb{R}_{\geq 0}$. Model training is based on

$$\mathcal{L}(\lambda) = \mathcal{L}_{\text{pred}} + \lambda \mathcal{P}_{\text{fair}}, \tag{3.8}$$

where $\mathcal{L}_{\text{pred}}$ denotes the class-weighted multiclass log loss and $\mathcal{P}_{\text{fair}}$ penalizes disparities across intersectional groups. The parameter λ governs the trade-off between predictive accuracy and fairness.

The equal opportunity gap $\text{EOG}(\lambda)$ captures predictive fairness. It measures the extent to which individuals from different intersectional groups have unequal chances of being

correctly assigned to their true occupation by the model, conditional on that occupation being the correct outcome. For occupation k and group g, let

$$\mathrm{TPR}_{k,g}(\lambda) = P\Big(\hat{Y}_i(\lambda) = k \mid Y_i = k, G_i = g\Big) \tag{3.9}$$

denote the group-specific true positive rate. The occupation-specific equal opportunity gap is then

$$\mathrm{EOG}_k(\lambda) = \max_{g \in \mathcal{G}_k} \mathrm{TPR}_{k,g}(\lambda) - \min_{g \in \mathcal{G}_k} \mathrm{TPR}_{k,g}(\lambda), \tag{3.10}$$

where $\mathcal{G}_k \subseteq \mathcal{G}$ denotes the set of intersectional groups with sufficient support in occupation k. We aggregate this measure as

$$\mathrm{EOG}(\lambda) = \frac{1}{|K^*|} \sum_{k \in K^*} \mathrm{EOG}_k(\lambda), \tag{3.11}$$

where K^* denotes the set of sufficiently supported occupations.

Equal opportunity is selected because the central policy concern in Public Employment Services is not equal occupational assignment rates across demographic groups, but equal predictive access to correct recommendations conditional on the true occupational outcome. Demographic parity would be difficult to justify in this setting because occupational transitions are strongly qualification-dependent and empirically heterogeneous. Equalized odds is substantially harder to interpret in a high-dimensional multiclass setting with many occupation-by-group combinations and sparse subgroup support. Calibration, in turn, addresses the reliability of predicted probabilities rather than equality in successful recommendation performance. Equal opportunity therefore provides a transparent measure of predictive fairness that directly captures disparities in correct occupational recommendations across intersectional groups.

In the context of AI-supported job matching, this quantity is directly linked to substantive fairness. A large gap implies that some groups are systematically less likely to receive correct recommendations for occupations that are in fact suitable for them. This can be interpreted as unequal access to accurate job matching, which may translate into differential labour market opportunities. For example, a value of $\mathrm{EOG} = 0.74$ implies that, on average, the difference between the most and least advantaged groups within an occupation is 74 percentage points. Conversely, a reduction of EOG to values around 0.19 implies a much more balanced distribution of predictive accuracy across groups, although this improvement may come at significant cost in predictive performance.

Importantly, $\Delta\mathrm{JSD}_k(\lambda)$ and $\mathrm{EOG}(\lambda)$ capture conceptually distinct fairness dimensions and need not move in parallel. JSD evaluates whether predicted occupational assignment structures become more or less intersectionally concentrated relative to observed labour market distributions. EOG evaluates whether predictive performance differs systematically

across intersectional groups. A model may therefore improve predictive fairness while still increasing distributional concentration, or it may reduce distributional sensitivity while leaving performance disparities across groups largely unchanged. Evaluating both dimensions jointly is necessary because AI-supported occupational recommendation systems can affect both who is predicted accurately and how groups are distributed across predicted occupational outcomes. Predictive performance is evaluated using complementary metrics reflecting the imbalanced multiclass nature of the problem. Log loss captures probabilistic accuracy. Standard accuracy is reported but is not central due to class imbalance. Balanced accuracy averages recall across occupations and is therefore more appropriate for evaluating class-sensitive performance. Top-k accuracy, reported for $k = 3, 5, 10$, is particularly relevant because the system provides ranked occupation recommendations rather than deterministic assignments. Macro one-vs-rest PR-AUC captures class discrimination under imbalance, while the Matthews correlation coefficient provides a balanced overall summary.

Table 3.2 summarizes the joint performance and fairness outcomes across λ. Figure 3.2

Table 3.2: Performance–fairness trade-off across regularisation levels

	Performance							*Fairness*
λ	LogLoss	Acc	BalAcc	Top-3	Top-5	PR-AUC	MCC	EOG
0.000	2.43	0.524	0.312	0.641	0.722	0.284	0.271	0.747
0.0005	2.45	0.522	0.310	0.639	0.720	0.282	0.269	0.744
0.001	2.47	0.519	0.307	0.636	0.718	0.279	0.265	0.739
0.005	2.81	0.472	0.270	0.590	0.675	0.245	0.231	0.598
0.010	3.45	0.381	0.198	0.480	0.560	0.180	0.162	0.184

visualizes the core trade-off between log loss and $\mathrm{EOG}(\lambda)$. The results reveal a clear

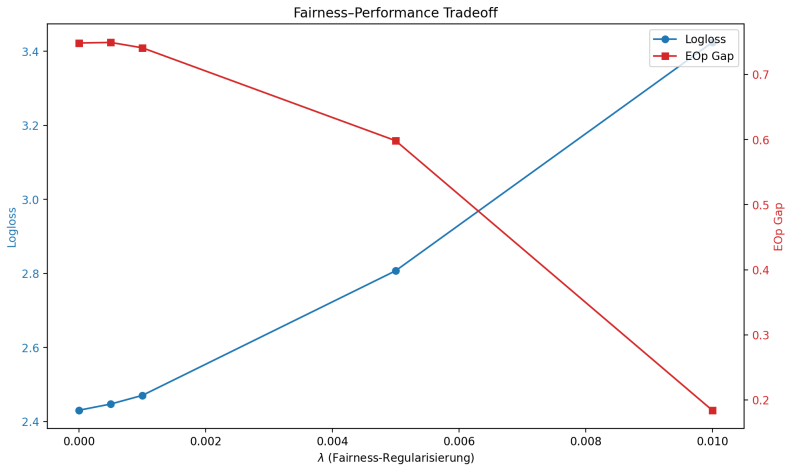


Figure 3.2: Fairness–performance trade-off: log loss and equal opportunity gap as functions of λ.

and systematic trade-off between predictive performance and predictive fairness. For

very small values of λ ($\lambda \leq 0.001$), predictive performance remains largely stable across all metrics, while fairness improvements are negligible. For intermediate regularisation ($\lambda = 0.005$), the equal opportunity gap decreases from 0.74 to 0.60, but balanced accuracy, top-k accuracy, and PR-AUC decline noticeably. For strong regularisation ($\lambda = 0.010$), predictive fairness improves substantially, with the equal opportunity gap reduced to 0.19, but predictive performance deteriorates to a degree that is unlikely to be operationally acceptable in a recommendation setting.

Two central insights emerge. First, predictive fairness improvements are costly in this setting: reducing disparities in correct recommendations across intersectional groups requires the model to deviate from predictive optimality. Second, the trade-off is nonlinear. Small increases in λ preserve performance but have little fairness effect, whereas substantial fairness gains require disproportionately large sacrifices in predictive quality. Top-k accuracy makes this trade-off especially tangible for the application context: at high levels of regularisation, even the recommendation set no longer provides reliable guidance.

3.3.5 Model Selection under Retained Utility Constraints

The previous section established a clear trade-off between predictive fairness and predictive performance. Model selection in this context cannot be reduced to a single-objective optimisation problem. Instead, it requires an explicit criterion that balances fairness improvements against the preservation of predictive usefulness.

We formalize this trade-off using retained utility relative to the unconstrained baseline model:

$$\mathrm{RU}(\lambda) = \min \left\{ \frac{\mathrm{BalAcc}(\lambda)}{\mathrm{BalAcc}(0)}, \frac{\mathrm{Top3}(\lambda)}{\mathrm{Top3}(0)} \right\}. \tag{3.12}$$

Here, $\mathrm{BalAcc}(\lambda)$ denotes balanced accuracy and $\mathrm{Top3}(\lambda)$ denotes top-3 accuracy for the model with regularisation parameter λ. The baseline values $\mathrm{BalAcc}(0)$ and $\mathrm{Top3}(0)$ serve as normalization benchmarks.

The use of top-3 accuracy is particularly important in the present application. $\mathrm{Top3}(\lambda)$ measures whether the true occupation is contained within the three highest-ranked model recommendations. In an AI-supported job matching system, recommendations are not used as deterministic assignments but as ranked suggestions for caseworkers and job seekers. Preserving top-3 performance therefore ensures that the system continues to provide relevant and actionable guidance.

A model is considered technically admissible if

$$\mathrm{RU}(\lambda) \geq 0.95. \tag{3.13}$$

The 95% threshold is motivated by the operational context. In a high-stakes application such as public employment services, predictive performance is directly linked to labour market outcomes. The primary objective of the system is to support effective job matching and thereby contribute to reducing unemployment duration. A model that substantially degrades predictive quality, even if it achieves strong fairness improvements, would undermine this objective.

Figure 3.3 shows the retained utility profiles for balanced accuracy and top-3 accuracy across λ. The figure reveals that both performance measures remain close to baseline levels

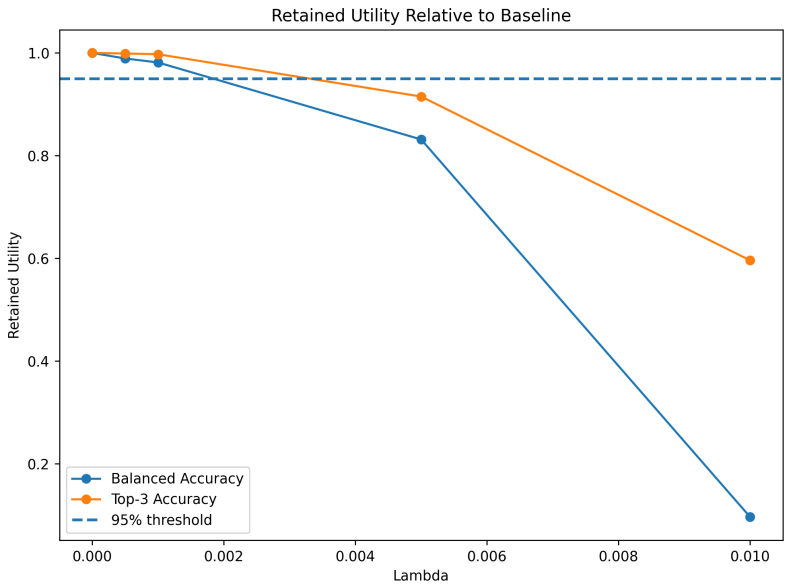


Figure 3.3: Retained utility relative to the baseline model. The dashed line indicates the 95% admissibility threshold.

for small values of λ, but diverge sharply as regularisation increases. Balanced accuracy deteriorates more rapidly than top-3 accuracy, indicating that class-level discrimination is more sensitive to fairness constraints than recommendation quality. For $\lambda = 0.001$, both measures remain above the 95% threshold. At $\lambda = 0.005$, top-3 accuracy is close to the threshold, while balanced accuracy already falls below it. For $\lambda = 0.010$, both measures collapse substantially.

Combining these results with the fairness analysis yields a clear selection outcome. While predictive fairness improves monotonically as λ increases, only small values of λ satisfy the retained utility constraint. Among these, $\lambda = 0.001$ provides the strongest fairness improvement without violating the performance requirement. We therefore select $\lambda = 0.001$ as the technically preferred model. By contrast, $\lambda = 0.010$ can be interpreted as a stress-test scenario: it demonstrates the upper bound of fairness improvements achievable through regularisation, but at a level of predictive degradation that renders the model unsuitable for deployment.

This result highlights a central tension in fairness-aware AI systems. The objective is not

to achieve maximal fairness in isolation, but to improve fairness within the constraints of practical usefulness. In the present application, the ultimate goal is to identify suitable occupations for job seekers and thereby improve labour market outcomes. A perfectly fair model that fails to provide accurate recommendations would be counterproductive.
Moreover, observed disparities partly reflect endogenous labour market sorting, including self-selection by individuals. Job seekers do not enter occupations randomly; their choices are shaped by preferences, qualifications, and constraints. Eliminating all observable disparities without accounting for these mechanisms may therefore introduce a different form of distortion. Not all observed differences are necessarily unjust, and not all reductions in disparity improve real-world outcomes.

3.4 Residual Distributional Risk, Explainability, and Governance

The preceding section identified $\lambda = 0.001$ as the technically preferred model under the retained-utility criterion and showed that stronger regularisation can improve aggregate predictive fairness only at substantial performance cost. This section returns to the distributional fairness perspective and asks where residual occupation-level distortions remain concentrated after regularisation. The objective is not only to quantify remaining risks, but also to translate them into audit, explainability, and governance mechanisms for Public Employment Services.

3.4.1 Residual Distributional Risk and Flagged Occupations

While the equal opportunity gap summarizes predictive fairness at the model level, the fairness flag identifies occupations in which the model induces unusually large distributional distortions relative to the empirical benchmark.
Formally, for each λ, an occupation k is flagged if

$$\Delta\mathrm{JSD}_k(\lambda) > \tau(\lambda), \qquad \tau(\lambda) = \mu_\Delta(\lambda) + \sigma_\Delta(\lambda), \tag{3.14}$$

where $\mu_\Delta(\lambda)$ and $\sigma_\Delta(\lambda)$ denote the mean and standard deviation of $\Delta\mathrm{JSD}_k(\lambda)$ across sufficiently supported occupations.
The fairness flag is not a legal or normative classification of discrimination. Rather, it constitutes an operational audit signal that identifies occupations in which the model induces unusually large distributional distortions relative to the empirical benchmark. As such, it provides a systematic link between aggregate fairness evaluation and localized diagnostic analysis, enabling targeted explainability, monitoring, and governance interventions.
Importantly, the flag is defined relative to the distribution of occupation-level distortions

within each model specification. The threshold depends on both the mean distortion $\mu_\Delta(\lambda)$ and the dispersion of distortions $\sigma_\Delta(\lambda)$. A larger standard deviation increases the threshold and therefore makes it more difficult for an occupation to be flagged, whereas a smaller standard deviation lowers the threshold and makes comparatively smaller residual deviations more visible. The flag should therefore not be interpreted as a fixed absolute measure of unfairness across different values of λ. Instead, it identifies occupations whose distributional distortions are unusually large relative to the overall distortion profile of the respective model.

This distinction is particularly important for the strongly regularized model with $\lambda = 0.010$. In this specification, the threshold is substantially higher than in the weakly regularized models. Occupations that remain flagged under this condition therefore exceed a much stricter benchmark. This does not imply that these occupations are necessarily treated less fairly than under the unconstrained baseline in every respect. Rather, it indicates that, even after strong regularisation has improved aggregate predictive fairness, these occupations remain exceptional residual cases in distributional terms. In substantive terms, fairness-aware regularisation appears to reduce broad-based predictive disparities while leaving a smaller set of occupations with unusually persistent intersectional concentration patterns. These occupations should therefore be interpreted as priority cases for further audit, explainability, and human oversight rather than as direct evidence of legal discrimination. Table 3.3 reports the number of flagged occupations for each value of λ.

Table 3.3 shows that the number of flagged occupations must be interpreted together

Table 3.3: Flagged occupations by regularisation level

λ	Number Flagged occupations	Threshold	Mean flagged ΔJSD	Max ΔJSD
0.0000	20	0.119	0.186	0.789
0.0005	20	0.124	0.197	0.767
0.0010	21	0.126	0.196	0.548
0.0050	21	0.177	0.253	0.808
0.0100	5	0.580	0.637	0.905

with the threshold and the severity of the flagged distortions. For weak and intermediate regularisation levels, the threshold remains relatively low, between 0.119 and 0.177, and around twenty occupations are flagged. For $\lambda = 0.010$, by contrast, the threshold increases sharply to 0.580. This substantially higher threshold means that occupations must exhibit much larger residual distributional distortions to be flagged at this level of regularisation. The decline from around twenty flagged occupations to five therefore does not mean that all other distributional risks disappear mechanically; rather, it indicates that only a small set of occupations remains sufficiently extreme to exceed a much stricter benchmark.

At the same time, the mean flagged ΔJSD rises to 0.637 and the maximum ΔJSD reaches 0.905. This pattern indicates that strong regularisation compresses broad-based distortions

into fewer but more severe residual cases. In other words, stronger fairness regularisation improves aggregate predictive fairness and reduces the number of occupations that stand out relative to the model-specific threshold, but the occupations that still exceed this threshold exhibit particularly pronounced distributional shifts. These flagged occupations should therefore be interpreted as structurally persistent residual risks rather than as evidence that the overall model has become less fair.

The results show that stronger regularisation does not monotonically eliminate flagged occupations. For small and intermediate values of λ, the number of flagged occupations remains broadly stable. Only for the strongest specification, $\lambda = 0.010$, does the number of flagged occupations fall substantially. However, the remaining flagged occupations exhibit large distortions. This indicates that regularisation compresses fairness risks into fewer, but more extreme, residual cases.

The model with $\lambda = 0.010$ achieves the lowest aggregate equal opportunity gap and can therefore be interpreted as the fairest model in predictive terms. Nevertheless, five occupations remain flagged from a distributional perspective. Table 3.4 lists these occupations. These results are substantively important. Even under the strongest fairness regularisation,

Table 3.4: Flagged occupations in the aggregate-fairest model ($\lambda = 0.010$)

Code	Occupation	N_k^{true}	N_k^{pred}	ΔJSD_k	Main overrepresented group
633	Gastronomy occupations	15,976	343	0.736	Female, Remaining, 25–34
293	Cooking occupations	14,129	403	0.670	Male, Remaining, 55–64
814	Human medicine and dentistry	4,259	205	0.541	Female, Remaining, 25–34
621	Retail trade without product specialisation	29,441	34,806	0.643	Female, Germany, 15–24
921	Advertising and marketing	10,680	54	0.593	Female, Remaining, 25–34

residual distributional distortions remain concentrated in specific occupations. The pattern is not random. The flagged occupations include hospitality, cooking, retail, medicine, and advertising. These are occupational fields in which labour market sorting is often strongly structured by gender, nationality, age, sectoral experience, and employment histories. Such variables may enter the model as legitimate predictors, but they can also operate as proxies for intersectional group membership.

The relationship between true and predicted class counts is also informative. For some occupations, such as advertising and marketing, human medicine and dentistry, gastronomy, and cooking, the model predicts far fewer cases than are observed in the data. This suggests that strong regularisation narrows the model's assignment behaviour and concentrates predictions on a small subset of profiles. For retail, by contrast, predicted counts exceed

observed counts, indicating that the model reallocates individuals into a broad service occupation. Both patterns can generate large changes in intersectional composition.
The key implication is that fairness-aware regularisation does not remove structural inequality; it changes its location and form. At $\lambda = 0.010$, predictive fairness improves substantially, but some occupations remain strongly distorted in distributional terms. This matters because the policy objective is not to produce a formally fair model in isolation. The purpose of the system is to support suitable job matching and thereby contribute to reducing unemployment. A model that is highly regularized but no longer provides reliable recommendations would fail this objective.

3.4.2 Explainability, Human Oversight, and Jobseeker Participation

The fairness flag is intended to support sociotechnical governance rather than replace professional or individual judgment. In Public Employment Services, algorithmic recommendations are embedded in a decision process that involves not only caseworkers, but also jobseekers themselves. Caseworkers interpret recommendations in light of counselling goals, institutional rules, local labour market conditions, and available support instruments. Jobseekers, in turn, bring preferences, constraints, qualifications, care responsibilities, mobility limitations, health conditions, and occupational aspirations into the matching process. A fairness-sensitive recommendation system must therefore support both professional interpretation and meaningful participation by jobseekers.
Flagged occupations can serve as targeted triggers for explainability and oversight mechanisms. First, they can be prioritized for local explanation methods that identify which features drive a recommendation and whether employment history, qualifications, sectoral experience, regional labour market information, or other variables act as proxies for intersectional group membership. Second, flagged recommendations can be accompanied by explanation cards that distinguish qualification-related drivers from potentially sensitive or proxy-related patterns. Third, uncertainty indicators can inform caseworkers and jobseekers when the model produces recommendations in occupational fields with known residual distributional distortions. Fourth, contrastive explanations can show which factors would make alternative occupations more likely, thereby supporting counselling conversations rather than presenting recommendations as fixed assignments.
These mechanisms are important because fairness risks do not arise only from the internal model structure, but also from how model outputs are interpreted and used. Without explanation, a flagged recommendation may be accepted mechanically, even if it reflects residual occupational concentration rather than an individually appropriate match. With appropriate explanation, the same recommendation can become a starting point for critical reflection: caseworkers can assess whether the recommendation is driven by relevant qualifications, jobseekers can contest or contextualize the suggestion, and both parties can consider

alternative occupational pathways. For jobseekers, explainability is particularly important because occupational recommendations may shape perceived opportunities and future search behaviour. If a system repeatedly recommends occupations that reflect historical sorting patterns, it may narrow rather than expand labour market opportunities. Explanation interfaces should therefore not only justify why an occupation was recommended, but also support agency by making visible which alternative occupations are plausible, which skills or qualifications would increase access to them, and whether recommendations are influenced by structurally patterned employment histories.

From a governance perspective, the fairness flag therefore provides an operational bridge between statistical fairness diagnostics and practical decision support. Rather than auditing all occupations equally, governance resources can be focused on occupations where the model produces the strongest residual distortions. These occupations can be monitored over time, reviewed with domain experts, and incorporated into interface-level safeguards. In this sense, the fairness flag complements the aggregate equal opportunity gap: the equal opportunity gap shows whether predictive fairness improves on average, while the fairness flag shows where residual distributional risks remain. Together, they demonstrate that technical mitigation must be combined with explainability, monitoring, jobseeker participation, and human judgment.

3.5 Counterfactual Averaging and Distributional Fairness

To further evaluate the robustness of the proposed fairness-aware estimation framework, we additionally implement a counterfactual averaging prediction strategy. In this approach, predictions are repeatedly generated under alternative intersectional identity assignments and subsequently averaged before the final occupation recommendation is produced. Importantly, the analysis remains fully intersectional. The counterfactual procedure operates on the complete intersectional representation, which jointly captures combinations of gender, nationality, and age group. The method therefore evaluates whether occupational predictions remain stable across alternative intersectional group assignments.

Conceptually, counterfactual averaging differs fundamentally from fairness regularisation. Whereas the regularized model constrains the estimation process directly through the penalty parameter λ, counterfactual averaging modifies the otherwise unconstrained prediction mechanism ex post by averaging predicted probabilities across alternative intersectional identities. Counterfactual averaging therefore primarily targets residual distributional sensitivity rather than predictive parity. Its effects are evaluated mainly through the occupation-level divergence framework, while changes in predictive performance metrics are interpreted as secondary robustness outcomes.

Residual demographic sensitivity captures the remaining dependence of occupational recommendations on intersectional identity assignments after holding all qualification- and

employment-related characteristics fixed.

3.5.1 Aggregate Fairness Effects

Table 3.5 compares the standard prediction rule with counterfactual averaging across regularisation levels. The results reveal a differentiated relationship between fairness

Table 3.5: Comparison of flagged occupations under the standard prediction rule and counterfactual averaging (CF).

λ	Flagged (Base)	Flagged (CF)	Mean ΔJSD (Base)	Mean ΔJSD (CF)
0.0000	20	22	0.186	0.134
0.0005	20	24	0.197	0.133
0.0010	21	22	0.196	0.138
0.0050	22	18	0.253	0.186
0.0100	5	2	0.637	0.282

regularisation and counterfactual averaging. For small values of λ, counterfactual averaging does not systematically reduce the number of flagged occupations and occasionally increases it slightly. However, even in these settings, the average severity of the flagged distortions declines noticeably. This indicates that counterfactual averaging primarily weakens the intensity of occupation-level concentration patterns rather than uniformly eliminating them.

The strongest effect emerges for $\lambda = 0.010$. Under the standard prediction rule, five occupations remain flagged, with an average residual divergence of 0.637. Counterfactual averaging reduces the number of flagged occupations to two and lowers the mean residual divergence to 0.282. Substantively, this implies that the remaining prediction distortions become less than half as large after counterfactual smoothing.

This pattern suggests that fairness regularisation and counterfactual averaging address different components of the fairness problem. The regularisation term primarily suppresses the formation of strongly group-dependent prediction structures during estimation, whereas counterfactual averaging reduces the sensitivity of final predictions to specific intersectional identity assignments. The two mechanisms therefore appear to reinforce each other particularly in highly regularized models.

Counterfactual averaging successfully reduces measured distributional disparities by lowering the dependence of occupational on intersectional identity assignments. However, these reductions should not automatically be interpreted as the complete removal of underlying structural inequalities. Because the approach modifies the prediction mechanism itself through probability averaging across identity perturbations, part of the observed fairness gain reflects reduced demographic sensitivity in prediction space rather than a full disappearance of labour market inequalities embedded in the underlying data-generating process.

3.5.2 Predictive Performance under Counterfactual Averaging

Table 3.6 compares selected predictive performance metrics between the standard prediction rule and counterfactual averaging for the baseline setting ($\lambda = 0$). The performance

Table 3.6: Predictive performance comparison between the standard prediction rule and counterfactual averaging for $\lambda = 0$.

Metric	Baseline	Counterfactual Averaging
Balanced Accuracy	0.274	0.262
Top-3 Accuracy	0.603	0.581
MCC	0.360	0.347

results indicate that counterfactual averaging slightly reduces predictive accuracy relative to the standard prediction rule. This pattern is theoretically plausible because the averaging procedure intentionally smooths prediction probabilities across alternative intersectional identity assignments. As a consequence, predictions become less specialized toward specific demographic profiles, which can moderately reduce classification precision and ranking performance.

At the same time, the observed utility reductions remain comparatively limited relative to the substantial fairness improvements documented in the occupation-level divergence analysis. In particular, counterfactual averaging substantially lowers residual demographic sensitivity, meaning that occupational recommendations become less dependent on specific combinations of gender, nationality, and age group while preserving most of the overall predictive structure of the model.

The two mitigation strategies operate through fundamentally different mechanisms. Fairness-aware regularisation directly constrains model outcomes and therefore reduces disparities irrespective of their source. Counterfactual averaging, by contrast, targets only the contribution of protected characteristics while leaving labour market-relevant predictors unchanged. As a result, some distributional disparities remain even after counterfactual adjustment. However, these residual differences are more likely to reflect variation in qualifications, employment histories, or other job-relevant characteristics rather than the direct influence of protected attributes. This distinction helps explain why counterfactual averaging achieves substantial reductions in distributional disparity while maintaining predictive performance. For PES, this trade-off may be particularly attractive because recommendation systems operate in institutionally sensitive environments where procedural consistency and fairness are central operational requirements. Counterfactual averaging therefore appears to provide an additional robustness layer that can complement fairness-aware estimation while maintaining comparatively strong predictive utility.

3.5.3 Occupation-Level Comparison for the Strongest Regularisation Setting

Table 3.7 reports the occupations that remain flagged under the strongest regularisation setting ($\lambda = 0.010$) and compares the standard prediction rule with counterfactual averaging. The occupation-level results clarify the magnitude of these effects. For trading

Table 3.7: Flagged occupations under the standard prediction rule and counterfactual averaging for $\lambda = 0.010$.

Occupation	JSD_{true}	JSD_{base}	JSD_{cf}	ΔCF vs. Base
Trading occupations	0.251	0.888	0.514	-0.374
Occupations in mechatronics, automation and control engineering	0.404	0.932	0.650	-0.282

occupations, the predicted intersectional distribution under the standard prediction rule deviates very strongly from the observed distribution, with a JSD of 0.888 relative to an observed benchmark of 0.251. Counterfactual averaging reduces this divergence by 0.374, corresponding to a reduction of approximately 42%. A similar pattern emerges for occupations in mechatronics, automation and control engineering, where the divergence declines from 0.932 to 0.650, implying a reduction of roughly 30%. In substantive terms, the results suggest that part of the residual unfairness in highly regularized models stems from prediction sensitivity to specific intersectional identities rather than exclusively from broader structural imbalances in the training data. At the same time, the remaining divergence levels after counterfactual averaging remain substantially above the observed benchmark distributions. This is important because it indicates that even aggressive post-processing interventions cannot fully eliminate occupation-level concentration patterns. The persistence of residual divergence therefore supports the interpretation that some disparities reflect underlying labour market structures that are already embedded in the observed data-generating process.

Importantly, the persistence of some distributional disparities under counterfactual averaging should not automatically be interpreted as a fairness failure. Because the intervention neutralizes protected attributes while preserving labour market-relevant information, remaining disparities may partly reflect differences in qualifications, employment histories, or occupational specialization rather than demographic characteristics themselves.

The counterfactual results additionally highlight several properties that are particularly relevant for operational deployment in PES. Counterfactual averaging reduces the sensitivity of occupational recommendations to demographic characteristics while preserving the broader predictive structure of the model. In practical terms, this means that occupational recommendations become more robust against demographic perturbations and less dependent on specific combinations of gender, nationality, and age.

This property is especially valuable in PES because recommendation systems are typically used in highly heterogeneous populations characterised by diverse educational, social, and demographic backgrounds. Counterfactual averaging therefore offers an additional safeguard against the emergence of strongly group-dependent recommendation patterns and may improve the perceived procedural fairness and consistency of algorithmic support systems. The empirical results show the effects when counterfactual averaging is combined with fairness-aware estimation. Whereas regularisation reduces the formation of structurally unfair prediction patterns during model training, counterfactual averaging attenuates residual demographic sensitivity at the prediction stage. The two mechanisms therefore appear to complement each other by operating at different stages of the modeling pipeline. At the same time, the moderate reduction in predictive performance observed under counterfactual averaging is theoretically plausible. By construction, the procedure smooths prediction probabilities across alternative demographic constellations and thereby partially suppresses information that may still carry predictive signal within the observed data-generating process. As a result, prediction outputs become less specialized toward particular intersectional profiles, which can slightly reduce classification precision and ranking performance. Importantly, however, the results indicate that these utility reductions remain comparatively limited relative to the observed fairness gains.

Comparing both approaches reveals an important difference. Fairness-aware regularisation reduces disparities primarily through stronger constraints on the prediction process, which eventually leads to substantial declines in predictive performance. Counterfactual averaging instead removes demographic sensitivity while preserving the informational contribution of non-protected predictors. Consequently, it achieves a more favourable fairness-performance trade-off, reducing occupation-level disparities without requiring the extensive predictive sacrifices observed under strong regularisation.

For practical deployment in PES, counterfactual averaging would nevertheless require additional supporting mechanisms. In particular, XAI components remain important to ensure transparency and accountability of recommendations. Caseworkers and jobseekers must be able to understand how recommendations are generated, which qualification-related characteristics remain most influential, and how fairness interventions affect prediction probabilities. Overall, the findings suggest that counterfactual averaging represents a promising complementary fairness mechanism for PES. The approach appears particularly valuable as an additional robustness layer that reduces residual demographic sensitivity while maintaining comparatively strong predictive utility.

3.6 Discussion: Integrating Fairness Dimensions in AI-Supported Job Matching

The empirical results highlight that fairness in AI-supported job matching is inherently multidimensional and cannot be adequately captured by a single metric or analytical perspective. Our findings show that distributional fairness, predictive fairness, and operational fairness risk are related but distinct properties of the modeling pipeline. Considering these dimensions jointly substantially changes how fairness interventions must be interpreted and evaluated in labour market recommendation systems.

3.6.1 Distributional and Predictive Fairness Do Not Coincide

A central finding of the analysis is that distributional fairness and predictive fairness capture different aspects of model behaviour and should therefore not be conflated. The Jensen–Shannon divergence framework evaluates how occupational outcome distributions across intersectional groups are reshaped by the predictive model relative to the empirical benchmark. By contrast, the equal opportunity gap evaluates whether predictive performance is distributed evenly across groups conditional on the true occupational outcome. The empirical results demonstrate that these fairness dimensions do not necessarily move together. The baseline model exhibits substantial amplification of occupational concentration patterns in distributional terms while simultaneously maintaining comparatively strong predictive performance. Conversely, strong fairness regularisation substantially reduces predictive disparities as measured by the equal opportunity gap but does not fully eliminate occupation-level distributional distortions. This distinction is substantively important for employment-related AI systems. A model may provide relatively balanced predictive accuracy across groups while still reallocating occupational opportunities in ways that increase intersectional concentration. Similarly, reductions in aggregate predictive disparities do not automatically imply preservation of observed occupational compositions. The results therefore suggest that fairness evaluation in labour market recommendation systems requires a structured multi-level framework rather than reliance on isolated fairness metrics.

3.6.2 Technical Mitigation Compresses Rather Than Eliminates Fairness Risk

The results further indicate that technical fairness interventions primarily reshape the location and form of fairness risk rather than removing it entirely. Fairness-aware regularisation reduces predictive disparities across intersectional groups but introduces a pronounced and nonlinear trade-off with predictive performance. Small regularisation levels preserve

most predictive utility while producing limited fairness improvements, whereas substantial reductions in predictive disparities require increasingly large sacrifices in recommendation quality. The occupation-level fairness flag reveals an additional layer of complexity. Aggregate fairness improvements do not imply uniform reductions in residual risk across occupations. Instead, stronger regularisation tends to compress fairness risk into a smaller number of occupations characterised by unusually large residual distortions. Importantly, the threshold defining flagged occupations depends on the distribution of occupational shifts within a given specification. Because the threshold is defined relative to the mean and standard deviation of $\Delta\mathrm{JSD}$, cross-model comparisons must be interpreted carefully. A smaller number of flagged occupations under strong regularisation therefore does not necessarily imply that all occupations become uniformly fairer. Rather, it indicates that fairness risk becomes concentrated in fewer but more pronounced residual cases relative to the distributional benchmark of that specification. The counterfactual robustness analysis reinforces this interpretation. Counterfactual averaging reduces residual demographic sensitivity, particularly in highly regularized models, indicating that some residual distortions remain sensitive to alternative intersectional identity assignments. At the same time, substantial residual divergence persists even after counterfactual smoothing. This suggests that fairness-aware estimation and counterfactual averaging address different components of the fairness problem. Regularisation constrains the formation of group-dependent prediction structures during estimation, whereas counterfactual averaging attenuates demographic sensitivity at the prediction stage. Neither mechanism fully removes the influence of underlying labour market sorting structures embedded in the observed data.

This observation also helps explain why the two mitigation strategies exhibit markedly different fairness-performance trade-offs. An important distinction is that the two mitigation strategies do not target the same source of disparity. Fairness-aware regularisation reduces disparities by directly constraining model outcomes, regardless of whether these differences originate from protected attributes or from labour market-relevant characteristics. Counterfactual averaging operates differently. By averaging predictions across alternative intersectional identity assignments, it neutralizes the direct influence of protected characteristics while preserving the predictive contribution of qualifications, employment histories, and other job-relevant information. Consequently, some distributional disparities remain even after counterfactual adjustment. However, these residual differences are more likely to reflect labour market-relevant heterogeneity than demographic sensitivity itself. This distinction helps explain why counterfactual averaging achieves meaningful reductions in occupation-level disparities while maintaining comparatively strong predictive performance.

Taken together, these findings suggest that technical mitigation should not be understood as a mechanism for eliminating fairness concerns. Rather, fairness interventions modify how predictive influence, demographic sensitivity, and occupational concentration patterns

are distributed throughout the recommendation system.

3.6.3 Implications for Deployment, Explainability, and Participatory Design in PES

The findings have direct implications for the deployment of AI-supported recommendation systems in PES. In such environments, algorithmic outputs are embedded within broader sociotechnical decision processes involving institutional constraints, professional judgement, and heterogeneous user populations. Recommendations are not implemented automatically but interpreted, contextualized, and potentially contested within counselling interactions. This deployment context has important implications for fairness governance. First, aggregate fairness metrics alone are insufficient for operational monitoring. Occupation-level diagnostics, fairness flags, and robustness analyses provide complementary tools for identifying localized fairness risks that may otherwise remain hidden behind aggregate improvements. Second, the results strengthen the case for integrating explainability functions into fairness-sensitive recommendation systems. Explainability is relevant not merely as a transparency add-on but as a mechanism for understanding how recommendation outputs are generated, how occupational predictions change under fairness interventions, and which qualification-related variables remain influential after mitigation. In particular, explanation interfaces could support the interpretation of flagged occupations, uncertainty communication, and the identification of variables acting as potential proxies for intersectional group membership. Third, the results underline the importance of user-centered and participatory system design in PES. Algorithmic fairness in counselling systems affects not only caseworkers but also jobseekers as active participants in the recommendation process. Future development should therefore investigate how fairness diagnostics and explanation functions can be embedded into interfaces co-designed with end users. Such approaches may improve transparency, interpretability, and perceived procedural legitimacy while ensuring that fairness interventions remain aligned with practical counselling needs. More broadly, the results suggest that responsible AI deployment in labour market institutions requires moving beyond purely technical fairness optimisation toward iterative, human-centered, and governance-oriented fairness management. In this sense, intersectional fairness analysis should not be interpreted as a one-time evaluation exercise but as an ongoing component of deployment, monitoring, and participatory system development.

3.7 Conclusion

This paper developed a multi-level framework for evaluating intersectional fairness in AI-supported job matching. Using large-scale administrative labour market data and a mul-

ticlass recommendation setting with 141 occupational groups, we distinguished between distributional fairness, predictive fairness, and residual operational fairness risk.

Our findings demonstrate that fairness in employment-related AI systems cannot be reduced to a single metric or analytical level. The unconstrained baseline model does not merely reproduce observed labour market structures, but systematically reshapes occupational outcome distributions across intersectional groups. Fairness-aware regularisation improves predictive fairness as measured by the equal opportunity gap, but introduces pronounced trade-offs with predictive performance and does not eliminate distributional distortions. Rather than removing inequality, technical mitigation primarily compresses and redistributes fairness risks across occupations. The occupation-level fairness flag and the counterfactual robustness analysis further show that residual fairness risks persist even under strong regularisation. Some occupational fields remain unusually sensitive to intersectional composition shifts or identity perturbations, indicating that aggregate fairness improvements may coexist with localized distributional distortions. These findings underscore the importance of complementing aggregate fairness metrics with diagnostic mechanisms capable of identifying where residual risks remain concentrated.

Limitations of this study should be considered in interpreting these results. The analysis is conditioned by the structure of administrative labour market data, which – while enabling large-scale intersectional analysis through variables such as age, gender, and nationality – does not systematically capture other socially salient dimensions, including disability status or racialised experiences beyond nationality. Gender is additionally recorded in a binary format, which reflects institutional classification practices and limits the representation of gender diversity within the data. Moreover, the empirical focus on a specific class of fairness-aware regularisation strategies and evaluation metrics provides a structured basis for comparison, but does not exhaust the broader space of possible fairness definitions or intervention approaches.

The implications extend beyond model optimisation. In public employment services, algorithmic recommendations are embedded within sociotechnical decision environments involving caseworkers, institutional constraints, and jobseekers as active participants in the counselling process. Fairness evaluation should therefore not be understood as a purely technical exercise, but as part of a broader governance challenge that includes explainability, monitoring, and human oversight.

Future research should investigate how intersectional fairness diagnostics can be operationalized within deployment-oriented AI systems. This includes the development of explanation interfaces that support both caseworkers and jobseekers, uncertainty communication mechanisms for fairness-sensitive recommendations, and participatory design approaches that involve end users throughout system development, evaluation, and governance. More broadly, our results suggest that responsible AI deployment in labour market institutions requires moving beyond static fairness metrics toward iterative, human-

centered, and context-sensitive fairness governance.

Appendix - Chapter 3

3.A Additional Tables

Table 3.A.1: Distribution of intersectional groups over persons

Gender, nationality, and age group	Share
Male, Germany, 25–34	12.79
Female, Germany, 25–34	10.50
Male, Germany, 35–44	8.25
Female, Germany, 35–44	7.74
Male, Germany, 15–24	7.74
Male, Germany, 45–54	6.41
Female, Germany, 45–54	6.34
Female, Germany, 15–24	5.90
Male, Germany, 55–64	4.71
Female, Germany, 55–64	4.26
Male, 10 main migration countries, 25–34	3.24
Male, 10 main migration countries, 35–44	2.26
Male, Remaining, 25–34	1.61
Female, 10 main migration countries, 25–34	1.48
Male, 10 main migration countries, 45–54	1.46
Male, 10 main migration countries, 15–24	1.40
Female, 10 main migration countries, 35–44	1.40
Male, Remaining, 35–44	1.12
Female, 10 main migration countries, 45–54	1.00
Female, Remaining, 35–44	0.77
Female, Remaining, 25–34	0.71
Female, 10 main migration countries, 15–24	0.62
Male, EU, 25–34	0.59
Male, Remaining, 45–54	0.59
Male, Remaining, 15–24	0.54
Male, 10 main migration countries, 55–64	0.52

Table 3.A.2: Jensen-Shannon divergence in the observed occupational data

JSD	Class count	Occupational group
0.6181	112	Aircraft pilots

Continued on next page

JSD	Class count	Occupational group
0.5483	29	Occupations in musical instrument making
0.5395	46	Occupations in fishing
0.5325	10 210	Doctors' receptionists and assistants
0.5226	57	Legislators and senior officials of special interest organisations
0.5183	497	Occupations in floristry
0.5087	1044	Floor layers
0.5084	247	Occupations in veterinary medicine and non-medical animal health practitioners
0.4952	2997	Occupations in plumping, sanitation, heating, ventilating, and air conditioning
0.4884	131	Ship's officers and masters
0.4874	9303	Occupations in building construction
0.4870	897	Occupations in psychology and non-medical psychotherapy
0.4852	25	Occupations in philology
0.4840	43	Occupations in vini- and viticulture
0.4777	5871	Occupations in housekeeping and consumer counselling
0.4764	2403	Occupations in civil engineering
0.4668	3228	Occupations in the interior construction and dry walling, insulation, carpentry, glazing, roller shutter and jalousie installation
0.4598	4150	Drivers and operators of construction and transportation vehicles and equipment
0.4590	6538	Occupations in building services engineering
0.4541	5159	Occupations in metal constructing and welding
0.4520	176	Occupations in underground and surface mining and blasting engineering
0.4483	278	Occupations in environmental protection engineering
0.4473	157	Occupations in stage, costume and prop design,
0.4459	170	Occupations in police and criminal investigation, jurisdiction and the penal institution
0.4397	4752	Technical occupations in energy technologies
0.4394	1522	Sales occupations (retail) selling drugstore products, pharmaceuticals, medical supplies and healthcare goods
0.4310	794	Drivers of vehicles in railway traffic
0.4292	35	Artisans designing ceramics and glassware
0.4281	160	Occupations in beverage production
0.4261	2830	Occupations in building services and waste disposal
0.4230	644	Occupations in the inspection and maintenance of traffic infrastructure
0.4226	134	Sales occupations (retail) selling books, art, antiques, musical instruments, recordings or sheet music
0.4208	3341	Occupations in legal services, jurisdiction, and other officers of the court
0.4198	7200	Technical occupations in the automotive, aeronautic, aerospace and ship building industries
0.4188	1000	Conditioning and processing of natural stone and minerals, production of building materials
0.4174	328	Technical occupations in railway, aircraft and ship operation
0.4144	4352	Painters and varnishers, plasterers, occupations in the waterproofing of buildings, preservation of structures and wooden building components
0.4128	25 634	Driver of vehicles in road traffic
0.4100	3745	Occupations in body care
0.4044	1663	Occupations in mechatronics, automation and control technology
0.4038	66	Occupations in economics
0.4036	972	Occupations in colour coating and varnishing
0.3988	550	Occupations in forestry, hunting and landscape preservation
0.3977	171	Musicians, singers and conductors
0.3969	227	Occupations in horsekeeping
0.3906	102	Occupations in mathematics and statistics
0.3906	936	Occupations in metal-making
0.3880	953	Occupations in precision mechanics and tool making

Continued on next page

JSD	Class count	Occupational group
0.3878	666	Occupations in public relations
0.3787	187	Occupations in geology, geography and meteorology
0.3784	4259	Occupations in human medicine and dentistry
0.3756	3794	Occupations in wood-working and -processing
0.3727	481	Occupations in treatment of metal surfaces
0.3701	114	Occupations in the humanities
0.3651	396	Occupations providing nutritional advice or health counselling, and occupations in wellness
0.3638	3969	Occupations in non-medical therapy and alternative medicine
0.3627	250	Occupations in funeral services
0.3524	1278	Occupations in editorial work and journalism
0.3487	323	Occupations in environmental protection management and environmental protection consulting
0.3466	23 235	Occupations in cleaning services
0.3459	2368	Occupations in technical research and development
0.3459	3575	Occupations in software development and programming
0.3427	1672	Managing directors and executive board members
0.3424	16 986	Occupations in machine-building and -operating
0.3419	1981	Occupations in pharmacy
0.3419	797	Occupations in traffic surveillance and control
0.3410	82	Occupations in artisan craftwork and fine arts
0.3331	1139	Teachers for occupation-specific subjects at vocational schools and in-company instructors in vocational training
0.3308	29 492	Occupations in education and social work, and pedagogic specialists in social care work
0.3300	816	Occupations in event technology, cinematography, and sound engineering
0.3278	219	Occupations in product and industrial design
0.3277	2676	Occupations in IT-network engineering, IT-coordination, IT-administration and IT-organisation
0.3248	209	Occupations in industrial ceramic-making and -processing
0.3238	12 155	Occupations in public administration
0.3167	772	Occupations in biology
0.3160	3705	Teachers in schools of general education
0.3126	1122	Laboratory occupations in medicine
0.3121	12 110	Occupations in metalworking
0.3120	431	Occupations in physics
0.3098	625	Occupations in media, documentation and information services
0.3086	4526	Occupations in human resources management and personnel service
0.3070	400	Occupations in animal care
0.3062	7589	Sales occupations (retail) selling foodstuffs
0.3057	380	Occupations in interior design, visual marketing, and interior decoration
0.3036	856	Occupations in the production of clothing and other textile products
0.3033	12 145	Occupations in geriatric care
0.3030	570	Occupations in industrial glass-making and -processing
0.3028	81	Military
0.3022	34 068	Office clerks and secretaries
0.3021	2989	Teachers and researcher at universities and colleges
0.3009	156	Presenters and entertainers
0.3002	415	Occupations in publishing and media management
0.2946	421	Occupations in theatre, film and television productions
0.2922	3725	Occupations in computer science
0.2896	1419	Teachers at educational institutions other than schools (except driving, flying and sports instructors)
0.2892	169	Technical and management occupations in museums and exhibitions

Continued on next page

JSD	Class count	Occupational group
0.2835	12 956	Occupations in nursing, emergency medical services and obstetrics
0.2833	3124	Occupations in IT-system-analysis, IT-application-consulting and IT-sales
0.2812	206	Occupations in theology and church community work
0.2783	242	Occupations in occupational health and safety administration, public health authority, and disinfection
0.2771	2117	Occupations in tax consultancy
0.2768	287	Occupations in surveying and cartography
0.2766	4956	Occupations in hotels
0.2755	200	Occupations in photography and photographic technology
0.2746	6552	Occupations in accounting, controlling and auditing
0.2656	6865	Occupations in electrical engineering
0.2653	4455	Occupations in plastic- and rubber-making and -processing
0.2649	5717	Occupations in gardening
0.2648	827	Technical occupations in paper-making and -processing and packaging
0.2582	47 541	Occupations in warehousing and logistics, in postal and other delivery services, and in cargo handling
0.2547	2154	Occupations in the social sciences
0.2507	1119	Occupations in tourism and the sports (and fitness) industry
0.2506	1635	Trading occupations
0.2432	172	Artisans working with metal
0.2410	14 129	Cooking occupations
0.2410	2637	Draftspersons, technical designers, and model makers
0.2384	2277	Occupations in technical media design
0.2376	5372	Occupations in the production of foodstuffs, confectionery and tobacco products
0.2363	844	Occupations in event organisation and management
0.2343	5221	Technical occupations in production planning and scheduling
0.2295	2948	Occupations in construction scheduling and supervision, and architecture
0.2278	320	Occupations in animal husbandry
0.2277	10 587	Occupations in physical security, personal protection, fire protection and workplace safety
0.2260	1904	Occupations in farming
0.2244	10 680	Occupations in advertising and marketing
0.2207	4122	Occupations in insurance and financial services
0.2206	631	Actors, dancers, athletes and related occupations
0.2204	29 441	Sales occupations in retail trade (without product specialisation)
0.2193	1572	Occupations in real estate and facility management
0.2173	282	Occupations in leather- and fur-making and -processing
0.2168	1723	Driving, flying and sports instructors at educational institutions other than schools
0.2151	557	Occupations in textile making
0.2098	12 535	Occupations in purchasing and sales
0.2051	790	Occupations in printing technology, print finishing, and book binding
0.2050	19 909	Occupations in business organisation and strategy
0.2022	1738	Technical occupations in medicine, orthopaedic and rehabilitation
0.1926	15 976	Gastronomy occupations
0.1841	7160	Sales occupations (retail trade) selling clothing, electronic devices, furniture, motor vehicles and other durables
0.1713	3285	Management assistants in transport and logistics
0.1663	768	Service occupations in passenger traffic
0.1517	3573	Occupations in chemistry

Note: The table reports the Jensen-Shannon divergence between the intersectional group distribution within each occupation and the overall population distribution. Higher values indicate stronger structural concentration in the observed data. Occupations are sorted in descending order of JSD.

Conclusion

The overarching contribution of this dissertation is the development of a broader empirical framework for evaluating machine learning-supported occupational recommendation in public employment services (PES). Rather than assessing recommendation systems exclusively according to their predictive performance, the dissertation demonstrates that successful occupational matching is inherently multidimensional. Effective occupational recommendations should identify occupations with high employment probabilities but should also account for the expected quality of employment relationships and the fairness implications of algorithmic decision support. By progressively integrating these dimensions, the dissertation extends the empirical evaluation of occupational recommendation beyond labour market entry towards a more comprehensive assessment of successful labour market matching.

The dissertation makes three complementary contributions. First, it contributes to the literature by extending the empirical analysis of occupational matching. While search-and-matching theory emphasises the importance of efficiently allocating workers to suitable jobs, empirical evidence on how administrative labour market data can be used to support occupational recommendation remains limited. Using comprehensive employment biographies, the dissertation demonstrates how modern prediction methods can improve the empirical analysis of occupational transitions and provide new insights into labour market matching at the occupational level.

Second, the dissertation demonstrates how recent advances in machine learning can be integrated into labour market research using large-scale administrative data. Rather than replacing established econometric approaches, machine learning complements them by providing flexible prediction methods that are particularly well suited to modelling complex occupational transitions in high-dimensional administrative datasets. The analyses show that these methods substantially improve predictive performance and therefore strengthen the empirical foundation of occupational recommendation.

Third, the dissertation broadens the evaluation of occupational recommendation by integrating job match quality and fairness into a common empirical framework. The matching index extends occupational recommendation beyond the probability of labour market entry by incorporating expected employment stability and wages, while the multidimensional fairness framework recognises that recommendation systems implemented in PES must

additionally account for the distributional consequences of algorithmic decision support. Together, these extensions demonstrate that evaluating occupational recommendation requires considering multiple, complementary dimensions of successful labour market matching.

Taken together, the three empirical studies progressively broaden the concept of successful occupational matching. The first study establishes whether occupational transitions can be predicted accurately. The second asks whether these predicted transitions are associated with desirable labour market outcomes. The third examines whether recommendation systems satisfying these objectives also meet the institutional requirements associated with their use in PES. In doing so, the dissertation connects advances in predictive modelling with established concepts from labour economics and the practical challenges of labour market policy. It thereby provides an integrated empirical framework for analysing occupational recommendation that reflects both the economic objectives of labour market matching and the institutional responsibilities of PES.

The findings of this dissertation suggest several promising directions for future research. While the present dissertation focuses on developing and evaluating prediction models, an important next step concerns their integration into practical decision-support systems. Future research should investigate how algorithmic recommendations can be incorporated into counselling processes in a way that complements rather than replaces caseworkers' expertise. Understanding how caseworkers interact with algorithmic recommendations and how such recommendations influence counselling decisions represents an important avenue for future research.

Although the analyses demonstrate the predictive potential of AI-supported occupational recommendation, the actual labour market effects of such systems remain an empirical question. Future research should therefore evaluate whether algorithmically supported occupational recommendations improve employment outcomes when implemented in practice through a field experiment within PES.

The increasing use of algorithmic decision support also raises important questions regarding explainability. While the present dissertation focuses on predictive performance, job match quality, and fairness, future research should investigate how explainable AI can improve transparency, strengthen trust in algorithmic recommendations, and facilitate communication between caseworkers and job seekers.

More broadly, the findings suggest that advances in predictive modelling create new opportunities to strengthen the empirical analysis of labour market matching. Their long-term contribution, however, will depend not only on further methodological progress but also on how these methods are integrated into the institutional environment of PES and aligned with the broader objectives of labour market policy. Continued research at the intersection of labour economics, data-driven decision support, and public policy therefore offers considerable potential for improving occupational matching while maintaining transparency,

fairness, and public accountability.

Bibliography

Agbasiere, C. and Nze-Igwe, G. (2025). Algorithmic fairness in recruitment: Designing ai-powered hiring tools to identify and reduce biases in candidate selection. *Path of Science*, 11(4):5001–5021.

Allen, J. and Van der Velden, R. (2001). Educational mismatches versus skill mismatches: effects on wages, job satisfaction, and on-the-job search. *Oxford economic papers*, 53(3):434–452.

Altmann, S., Falk, A., Jäger, S., and Zimmermann, F. (2018). Learning about job search: A field experiment with job seekers in germany. *Journal of Public Economics*, 164:33–49.

Amato, F., Boselli, R., Cesarini, M., Mercorio, F., Mezzanzanica, M., Moscato, V., Persia, F., and Picariello, A. (2015). Challenge: Processing web texts for classifying job offers. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, pages 460–463.

An, J., Huang, D., Lin, C., and Tai, M. (2025). Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. *PNAS Nexus*, 4(3):pgaf089.

Antoni, M., Ganzer, A., and vom Berge, P. (2016). Sample of integrated labour market biographies (siab) 1975-2014. *Documentation on Labour Market Data 201604*.

Bach, R. L., Kern, C., Mautner, H., and Kreuter, F. (2023). The impact of modeling decisions in statistical profiling. *Data & Policy*, 5:e32.

Bauer, A. (2013). Mismathch unemployment: Evidence from germany, 2000-2010. *IAB-Discussion Paper*.

Bauer, B., Mühlbauer, S., Schlögl-Flierl, K., Weber, E., and Ziethmann, P. (2025). Ethical integration in public sector ai: The case of algorithmic systems in the public employment service in germany. IAB-Discussion Paper 12/2025, Nürnberg.

Behaghel, L., Dromundo, S., Gurgand, M., Hazard, Y., and Zuber, T. (2022). Encouraging and directing job search: Direct and spillover effects in a large scale experiment. Technical report, Banque de France.

Belot, M., Kircher, P., and Muller, P. (2019). Providing advice to jobseekers at low cost: An experimental study on online advice. *The Review of Economic Studies*, 86(4):1411–1447.

Bied, G., Gaillac, C., Hoffmann, M., Caillou, P., Crépon, B., Nathan, S., and Sebag, M. (2023a). Fairness in job recommendations: estimating, explaining, and reducing gender gaps. In *Proceedings of the 1st Workshop on Fairness and Bias in AI co-located with 26th European Conference on Artificial Intelligence (ECAI 2023)*, volume 3523, Krakow, Poland. CEUR-WS.org.

Bied, G., Nathan, S., Perennes, E., Hoffmann, M., Caillou, P., Crépon, B., Gaillac, C., and Sebag, M. (2023b). Toward job recommendation for all. In *IJCAI 2023-The 32nd International Joint Conference on Artificial Intelligence*, pages 5906–5914. International Joint Conferences on Artificial Intelligence Organization.

Blundell, R., Costa Dias, M., Meghir, C., and Van Reenen, J. (2004). Evaluating the employment impact of a mandatory job search program. *Journal of the European Economic Association*, 2(4):569–606.

Boselli, R., Cesarini, M., Mercorio, F., and Mezzanzanica, M. (2017). Using machine learning for labour market intelligence. pages 330–342.

Boselli, R., Cesarini, M., Mercorio, F., and Mezzanzanica, M. (2018). Classifying online job advertisements through machine learning. *Future Generation Computer Systems*, 86:319–328.

Braun, H. (2017). Applying learning-to-rank to human resourcing's job-candidate matching problem: A case study. Master's thesis, Radboud Universiteit.

Breiman, L. (2001). Random forests. *Machine Learing*, 45(1):5–32.

Brioscú, A., Lauringson, A., Saint-Martin, A., and Xenogiani, T. (2024a). A new dawn for public employment services: Service delivery in the age of artificial intelligence. Technical report, OECD Publishing.

Brioscú, A., Lauringson, A., Saint-Martin, A., and Xenogiani, T. (2024b). A new dawn for public employment services: Service delivery in the age of artificial intelligence. *OECD Artificial Intelligence Papers*, 19.

Broecke, S. (2023). Artificial intelligence and labour market matching. *OECD Publishing*, 284.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Buyl, M., Cociancig, C., Frattone, C., and Roekens, N. (2022). Tackling algorithmic disability discrimination in the hiring process: An ethical, legal and technical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, pages 1071–1082, New York, NY, USA. Association for Computing Machinery.

Caliendo, M., Tatsiramos, K., and Uhlendorff, A. (2013). Benefit duration, unemployment duration and job match quality: a regression-discontinuity approach. *Journal of applied econometrics*, 28(4):604–627.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136.

Cappellari, L. and Tatsiramos, K. (2015). With a little help from my friends? quality of social networks, job finding and job match quality. *European Economic Review*, 78:55–75.

Chakravarty, S., R., K. T., and Lindsay, L. (2020). Increasing employment through the partial release of information.

Chen, Z., Zhang, J. M., Hort, M., Harman, M., and Sarro, F. (2024). Fairness testing: A comprehensive survey and analysis of trends. *ACM Trans. Softw. Eng. Methodol.*, 33(5).

Colombo, E., Mercorio, F., and Mezzanzianica, M. (2018). Applying machine learning tools on web vacancies for labour market and sill analysis. *Terminator or the Jetsons*.

Couronné, R., Probst, P., and Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, 19(1):270.

Crenshaw, K. W. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43(6):1241–1299.

Davidson, J., Schnepp, J., Buttrick, H., and Gupta, A. (2026). A framework for ethical ai-hrm development and use. *AI Ethics*, 6(28).

Davis, K. (2008). Intersectionality as buzzword: A sociology of science perspective on what makes a feminist theory successful. *Feminist Theory*, 9(1):67–85.

Desiere, S. and Struyven, L. (2021). Using artificial intelligence to classify jobseekers: The accuracy-equity trade-off. *Journal of Social Policy*, 50(2):367–385.

Faliagka, E., Ramantas, K., Tsakalidis, A., and Tzimas, G. (2012). Application of machine learning algorithms to an online recruitment system. In *Proc. International Conference on Internet and Web Applications and Services*, pages 215–220. Citeseer.

Fang, M. (2015). Learning to rank candidates for job offers using field relevance models. Master's thesis, University of Groningen and Saarland University.

Fitzenberger, B., Osikominu, A., and Völter, R. (2006). Imputation rules to improve the education variable in the iab employment subsample. *Journal of Contextual Economics–Schmollers Jahrbuch*, (3):405–436.

Flügge, A. A. (2021). Perspectives from practice: Algorithmic decision-making in public employment services. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '21 Companion, page 253–255, New York, NY, USA. Association for Computing Machinery.

Freeman, R. (2002). The labour market in the new information economy. *Oxford Review of Economic Policy*, 18:288–305.

Gartner, H. (2005). The imputation of wages above the contribution limit with the german iab employment sample. *FDZ-Methodenreport*, (02).

Gartner, H., Rothe, T., and Weber, E. (2022). The quality-weighted matching function: Did the german labour market reforms trade off efficiency against job quality? *Journal of Money, Credit and Banking*.

Gerunov, A. (2014). Big data approaches to modeling the labour market.

Giannelli, G. C., Jaenichen, U., and Rothe, T. (2013). Doing well in reforming the labour market? recent trends in job stability and wages in germany. Technical report, IZA Discussion Papers.

Green, F. and Zhu, Y. (2010). Overqualification, job dissatisfaction, and increasing dispersion in the returns to graduate education. *Oxford economic papers*, 62(4):740–763.

Gulyas, A. and Pytka, K. (2025). Understanding the heterogeneity of earnings losses after job displacement: A machine-learning approach. *Journal of Labor Economics*.

Gulyas, A., Pytka, K., et al. (2019). Understanding the sources of earnings losses after job displacement: A machine-learning approach. Technical report, University of Bonn and University of Mannheim, Germany.

Hancock, A.-M. (2007). Intersectionality as a normative and empirical paradigm. *Politics & Gender*, 3:248–254.

Hartmann, M., Klaus, A., Beckmann, R., and Stephani, J. (2021). Monatsbericht zum arbeits- und ausbildungsmarkt. *Berichte: Blickpunkt Arbeitsmarkt*.

Hastie, T., Tibshirani, R., and Friedman, J. (2017). *The Elements of Statistical Learning – Data Mining, Inference, and Prediction*.

Heisz, A. (1999). Changes in job duration in canada. *Relations industrielles*, 54(2):365–387.

Hill Collins, P. and Bildge, S. (2020). *Intersectionality*. Polity, Cambridge.

Holzer, H. J. (1988). Search method use by unemployed youth. *Journal of Labour Economics*, 6:1–20.

Huber, M. (2021). An introduction to flexible methods for policy evaluation. In *Handbook of Research Methods and Applications in Empirical Microeconomics*, pages 82–111. Edward Elgar Publishing.

Hutter, C. and Weber, E. (2017). Mismatch and the forecasting performance of matching functions. *Oxford Bulletin of Economics and Statistics*, 79(1):101–123.

Jovanovic, B. (1979). Job matching and the theory of turnover. *Journal of political economy*, 87(5, Part 1):972–990.

Junquera, Á. F. and Kern, C. (2025). From rules to forests: rule-based versus statistical models for jobseeker profiling. *Journal for Labour Market Research*, 59(1):1–27.

Kianpisheh, S., Jalili, S., and Charkari, N. M. (2012). Predicting job wait time in grid environment by applying machine learning methods on historical information. *International Journal of Grid and Distributed Computing*, 5(3):11–22.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. R. (2020). Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences*, 117(48):30096–30100.

Körtner, J. and Bonoli, G. (2022). Predictive algorithms in the delivery of public employment services. *SocArXiv*, page 19.

Kubis, A. (2021). Iab-stellenerhebung 1/2021: Betriebe sind bei neueinstellungen weiter zurückhaltend. (1).

Le Barbanchon, T., Hensvik, L., and Rathelot, R. (2023). How can ai improve search and matching? evidence from 59 million personalized job recommendations.

Li, H. (2011). A short introduction to learning to rank. *IEICE TRANSACTIONS on Information and Systems*, 94(10):1854–1862.

Malin, C., Fleiß, J., Ortlieb, R., and Thalmann, S. (2025). Rejected by an ai? comparing job applicants' fairness perceptions of artificial intelligence and humans in personnel selection. *Frontiers in Artificial Intelligence*, Volume 8 - 2025.

Mang, C. (2012). Online job search and matching quality. *ifo Working Paper*, (147).

Mavromaras, K., McGuinness, S., O'Leary, N., Sloane, P., and Wei, Z. (2013). Job mismatches and labour market outcomes: Panel evidence on university graduates. *Economic Record*, 89(286):382–395.

Mehrotra, A., Pradelski, B. S. R., and Vishnoi, N. K. (2022). Selection in the presence of implicit bias: The advantage of intersectional constraints. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 599–609, New York, NY, USA. Association for Computing Machinery.

Mohammadi, E., Cai, Y., Novin, A., Vera, V., and Soltanmohammadi, E. (2025). Who is a scientist? gender and racial biases in google vision ai. *AI Ethics*, pages 4993–5010.

Mortensen, D. T. and Pissarides, C. A. (1994). Job creation and job destruction in the theory of unemployment. *The review of economic studies*, 61(3):397–415.

Mühlbauer, S. and Weber, E. (2024). Predicting job match quality: A machine learning approach. Technical report, IAB-Discussion Paper.

Mühlbauer, S. and Weber, E. (2026). Machine learning for labor market matching. *Machine Learning with Applications*, page 100861.

Nekoei, A. and Weber, A. (2017). Does extending unemployment benefits improve job quality? *American Economic Review*, 107(2):527–61.

Nielsen, M. W., Gissi, E., Heidari, S., Horton, R., Nadeau, K. C., Ngila, D., Noble, S. U., Paik, H. Y., Tadesse, G. A., Zeng, E. Y., et al. (2025). Intersectional analysis for science and technology. *Nature*, pages 329–337.

Olsen, K. M., Kalleberg, A. L., and Nesheim, T. (2010). Perceived job quality in the united states, great britain, norway and west germany, 1989-2005. *European Journal of Industrial Relations*, 16(3):221–240.

Page, S.-J. and Yip, A. K. (2020). *Intersecting Religion and Sexuality: Sociological Perspectives*. Brill, Leiden, Niederlande.

Pandey, A. and Caliskan, A. (2021). Disparate impact of artificial intelligence bias in ridehailing economy's price discrimination algorithms. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 822–833.

Paulus, W. and Matthes, B. (2013). Klassifikation der berufe - struktur, codierung und umsteigeschlüssel. Technical Report 201308 (de), Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg [Institute for Employment Research, Nuremberg, Germany].

Pecoraro, M. (2014). Is there still a wage penalty for being overeducated but well-matched in skills? a panel data analysis of a s wiss graduate cohort. *Labour*, 28(3):309–337.

Petrongolo, B. and Pissarides, C. A. (2001). Looking into the black box: A survey of the matching function. *Journal of Economic literature*, 39(2):390–431.

Poe, R. L. and El Mestari, S. Z. (2024). The conflict between algorithmic fairness and non-discrimination: An analysis of fair automated hiring. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, pages 1907–1916, New York, NY, USA. Association for Computing Machinery.

Probst, P. and Boulesteix, A.-L. (2018). To tune or not to tune the number of trees in random forest. *Journal of Machine Learning Research*, (18):1–18.

Rentzsch, R. and Steneva, M. (2020). Skills-matching and skills intelligence through curated and data-driven ontologies. In *Proceedings of the DELFI Workshops*, pages 46–58. Gesellschaft für Informatik eVz Bonn.

Rosenberger, J., Wolfrum, L., Weinzierl, S., Kraus, M., and Zschech, P. (2025). Careerbert: Matching resumes to esco jobs in a shared embedding space for generic job recommendations. *Expert Systems with Applications*, 275:127043.

Sahin, A., Song, J., Topa, G., and Violante, G. L. (2014). Mismatch unemployment. *American Economic Review*, 104(11):3529–3564.

Saito, Y. and Sugiyama, K. (2024). Leveraging multiple behaviors and explicit preferences for job recommendation. *Expert Systems with Applications*, 258:125149.

Schneider, A. (2020). Ieb integrierte erwerbsbiografien. *DIM Datenreport*.

Sharma, R., Rajput, A., and Srivastav, V. K. (2025). Job description and resume matching system using nlp. In *2025 International Conference on Next Generation Information System Engineering (NGISE)*, volume 1, pages 1–5.

Simon, C. J. and Warner, J. T. (1992). Matchmaker, matchmaker: The effect of old boy networks on job match quality, earnings, and tenure. *Journal of Labor Economics*, 10:306–330.

Suk, Y. and Han, K. C. T. (2023). Evaluating intersectional fairness in algorithmic decision making using intersectional differential algorithmic functioning. *PsyArXiv*.

Sun, Y., Wong, A. K. C., and Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(4):687–719.

van Belle, A., Dehling, E., and Foster, D. (2018). Improving candidate to job matching with machine learning.

van den Berg, G. J. and Vikström, J. (2014). Long-run effects of active labor market policy programs. In *the Fifth Workshop of the Centre for Research in Active Labor Market Policy Effects (CAFE)*. Citeseer.

Van Huynh, T., Van Nguyen, K., Nguyen, N. L.-T., and Nguyen, A. G.-T. (2020). Job prediction: From deep neural network models to applications. pages 1–6.

Van Ours, J. C. and Vodopivec, M. (2006). How shortening the potential duration of unemployment benefits affects the duration of unemployment: Evidence from a natural experiment. *Journal of Labor economics*, 24(2):351–378.

Vethman, S., Smit, Q. T. S., van Liebergen, N. M., and Veenman, C. J. (2025). Fairness beyond the algorithmic frame: Actionable recommendations for an intersectional approach. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 3276–3290, New York, NY, USA. Association for Computing Machinery.

Wang, Z., Wu, Z., Guan, X., Thaler, M., Koshiyama, A., Lu, S., Beepath, S., Ertekin, E., and Perez-Ortiz, M. (2024). Jobfair: A framework for benchmarking gender hiring bias in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3227–3246. Association for Computational Linguistics.

Zollo, T., Rajaneesh, N., Zemel, R., Gillis, T., and Black, E. (2025). Towards effective discrimination testing for generative ai. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pages 1028–1047, New York, NY, USA. Association for Computing Machinery.