LUDWIG MAXIMILIAN UNIVERSITY OF MUNICH

DEPARTMENT OF STATISTICS

Algorithmic Profiling in the Austrian Labor Market:

Performance and fairness evaluation of various model classes and variable sets.

Master's Thesis

Author Viktoria Szabo

Supervisors Prof. Dr. Frauke Kreuter & Prof. Dr. Christoph Kern

Date Munich, September 30, 2022

Declaration of Originality

I confirm that the submitted thesis is original work and was written by me without further assistance. Appropriate credit has been given where reference has been made to the work of others. The thesis was not examined before, nor has it been published. The submitted electronic version of the thesis matches the printed version.

Viktoria Szabo

Abstract

This study compares the effects of using various model classes and different feature sets for the prediction of labor market chances of young unemployed. For this comparison the original Austrian labor market algorithm AMAS, which is based on a logistic regression und trained upon administrative data, is evaluated on a dataset of young unemployed in Vienna and then compared to additional model classes and various survey-based covariate sets as a benchmark in regard to different performance and fairness measures. Overall, choosing a diverse set of variables using not only administrative data but also survey data with multiple variable categories, like behavior, work attitudes and personality, did improve performance accuracy and fairness over all model classes. Nevertheless, a general trade-off between overall accuracy and fairness was also observed, consistent with numerous other publications. On average, the model class "extreme gradient boosting" performed best in regard to performance and fairness over all tasks.

Contents

1 Introduction

Algorithmic and statistical profiling is a widely used method in various fields such as policing (Julia Angwin and Kirchner, 2016), ad placement (Datta et al., 2014) or loan applications (Ramadorai et al., 2017) only to name a few examples. Recently, such algorithms are also getting more and more popular among Public Employment Services (PES) in various countries across the world to allocate resources efficiently and find suitable treatments (Desiere et al., 2019; Lechner and Smith, 2007). Multiple publications showed that algorithms are often better in accurately predicting job market chances than simple rule-based or human based decisions (Arbetsförmedlingen, 2014). But predictions solely based on administrative data usually ignore many dimensions of the jobseeker's profile, like personality traits and motivation, which most likely have a great impact on the job outlook of that person. Furthermore, administrative data focuses on variables which are observable from the outside and often not influenceable for the unemployed, like gender, age or ethnic background. Being treated differently based on such characteristics is commonly perceived as unfair and also often banned by legislation (Council of European Union).

Since 2018, Austria has been testing such an algorithm to evaluate the chances of job placement for registered unemployed by the Public Employment Service Austria (Arbeitsmarktservice – AMS) (Holl et al., 2018). The predictions are planned to be used for allocating money, spent on the unemployed, more efficiently by using it on those who can supposedly benefit most from the AMS programs. In the algorithm, these are groups who have "medium" chances on the labor market compared to those with "high" and "low" chances. According to their model, being a woman, old or with a migration background lowers your chances for finding a job and therefore could lead to a categorization in the "low" segment, which could in turn lead to limited access to support measures. Neglecting the low chance category could cause a great disadvantage for already marginalized groups in the labor market (Allhutter et al., 2020b). This is a major point why the model has been heavily criticized in public and was ultimately suspended by the Austrian data protection authority (Wimmer, 2018). The case now lies at the Higher Administrative Court (Fanta, 2021).

Since the classification into three categories can have consequences for the support offered to the unemployed individual, it is important to achieve high prediction accuracy. Furthermore, checking for fairness of the model should have a high priority to avoid

the above-mentioned marginalization of certain groups. Only a few publications about unemployment predictions evaluated fairness in addition to performance so far. For a Dutch statistical profiling system similar to the AMS algorithm, Desiere and Struyven (2021) showed that when the system is optimized for overall accuracy, subgroups like unemployed persons with a migration background are misclassified up to 2.6 times more often than with the long-established rule-based approach. And Kern et al. (2021) found while modeling long-term unemployment risks with German PES administrative data, that modeling choices and classification policies do have significant impact on fairness aspects. Allhutter et al. (2020a) gives a deep insight and evaluation about the AMS algorithm and fairness aspects of it. They are not able to work with the original model and data but conclude, based on the documentation, that multiple technical and sociocultural aspects of the algorithms could lead to discrimination of already marginalized groups.

This study is comparing the effects of using various model classes and different feature sets for the prediction of labor market chances of young unemployed. A special focus lies on fairness in regard to gender and migration background, which is still an often-neglected aspect while modeling unemployment risks. This is done by using the original AMS logistic regression, solely based on administrative data, as a baseline and comparing additional model classes with various survey-based covariate sets to it with regard to different performance and fairness measures.

Young people are especially vulnerable, because their careers are in an early stage and their entire future might be heavily impacted. It is very important to closely evaluate model outcomes for this group. Therefore, the analysis in this paper focuses on unemployed persons below 30 years. Overall, I found that a diverse set of variables using not only administrative data but also survey data with multiple variable categories, like behavior, work attitudes and personality, did improve performance accuracy and fairness over all model classes. Furthermore, variable sets composed of features that were often used in other PES algorithms and were already proven to have major impact on unemployment risks, did not only have good performance accuracy but also performed especially well in regard to fairness. Nevertheless, a general trade-off between overall accuracy and fairness was also observed, consistent with numerous other publications (Corbett-Davies et al., 2017; Desiere and Struyven, 2021)

In the following chapter the original AMS algorithm is analyzed and evaluated, while

section 3 gives a short overview of the data. Chapter 4 presents the methodology, where modeling and variable choices, as well as performance and fairness evaluation are discussed. Section 5 presents the results. And chapter 6 concludes in a summary of discussion points and results, while also pointing out limitations and further research.

2 AMS Algorithm

In 2018, the Public Employment Service Austria (Arbeitsmarktservice – AMS) started to test an algorithm, which was developed by the external company Synthesis GmbH. The goal was to evaluate the chances of job placement for registered unemployed and use this information for effective budget management (Holl et al., 2018, 2019; Gamper et al., 2020). The algorithm groups the jobseekers into three categories („high (H)“, „medium (M)“ and “low (L)”) based on the calculated chances for a job placement, which is called the “Integration Score (IC)”. Chances for job placement are looked at in a short-term perspective, finding a job for **more than 90 days in the next 7 months**, and long-term perspective, finding a job for **more than 180 days in the next 24 months**. The formation of the three groups is based on the consideration that not all clients have the same need for AMS support and services. Therefore, differing support measures will be offered to the jobseekers based on their categorization. The “high” chance group (H), with an **integration score of over 66% in the short-term perspective**, receives less support since it is assumed they will likely find employment without further help and training. The “low” chance group (L), with an **integration score of below 25% in the long-term perspective**, also receives less support from the AMS, instead they will be assigned to an external institution for special care. Main focus lies on the “medium” chance group (M), which comprises jobseekers neither falling into H or L, since they are seen as the group with the highest return on support measures. This categorization mechanism is depicted in figure 1 . The semi-automated classification is explicitly introduced to distribute scarce resources in the active labor market program in an efficient way. But the final decision for categorization lies upon the case worker (Gamper et al., 2020).

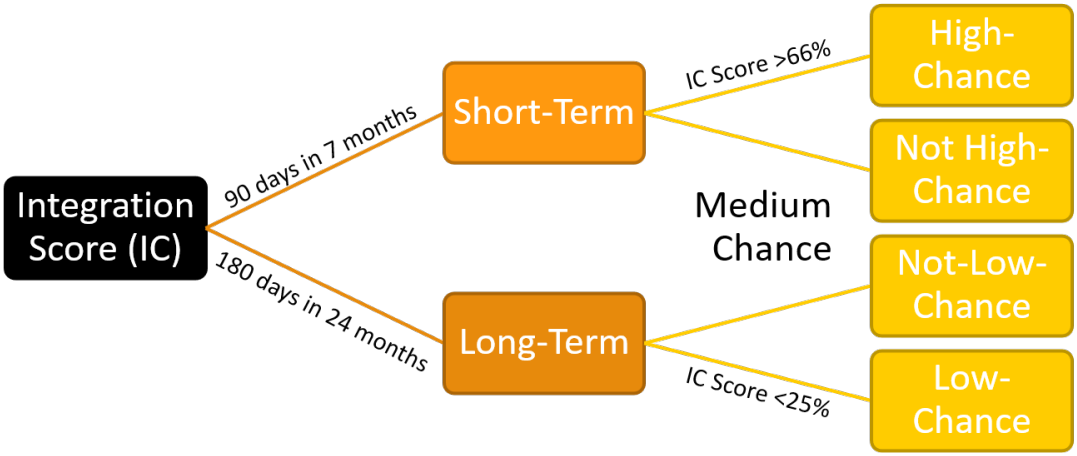

Figure 1: AMS categorization based on integration score

The calculation of the integration scores is done through a stratification procedure. It is performed by grouping the unemployed by different characteristics, and looking at the ratio of jobseekers leaving unemployment compared to those staying unemployed in the specific groups in different time-frames - short 7 months or long 24 months. The unemployed are grouped based on on the manifestations of 12 different characteristics as shown in table 1 (Gamper et al., 2020).

In addition to the stratification, the AMS is providing a logistic regression. Mainly as an easy descriptive way to indicate the influence of the various variables and to identify the main reasons for good or bad chances on the labor market, which can then be reported to the jobseeker in a consultation (Allhutter et al., 2020b). Equally to the stratification, 12 main characteristics are used. But in contrast to the stratification, only main effects are included in the regression and no interaction of variables is examined. Models are computed for the short- and long-term criterion, as well as for the so-called partially and fully estimable groups, which are explained below (Gamper et al., 2020). Since the stratification procedure is not documented in detail, only the published coefficients of the logistic regressions will be used as comparison in this study.

For the estimation, some of the variables need the employment history up to four years before the current business case. This information is not always available, because e.g., some only just migrated less than four years ago into Austria or just left the school and have not worked for 4 full years yet. Therefore, Synthesis defined four populations: fully valid estimable, partially valid estimable immigrants, partially valid estimable youth and partially valid estimable with fragmented working career. For each population, a separate

Variable	Characteristics
EMPLOYMENTDAYS	Days of employment (target variable)
GENDER	M/F
AGEGROUP	<30/30-49/50+
STATEGROUP	native/EU/other
EDUCATION	Compulsory schooling/Vocational training/Matura or higher
CHILDCARE	Yes/No (only women!)
IMPAIRMENT	Yes/No
RGS	Type 1-5
OCCUPATIONGROUP	Production/Service
EMPLOYMENTHIST	<75% />75% employment days in last 4 years
BUSINESSCASEDUR	0 Cases at AMS >6 months/ 1+ Cases >6 months
BUSINESSCASEFRE	no case in last 4 years/ 1 case in 4 years/ 2 cases in 4 years/ 3+ cases in 4 years
SUPPORTMEASURE	0 measures/ 1+ support measures/ 1+ qualifying measure/ 1+ employment promotion measure

Table 1: Original AMS Algorithm variables (Gamper et al., 2020)

model with partly reduced feature set is estimated (Gamper et al., 2020). Since using a dataset with 18- to 28-year-old only the full and youth model is looked at in the upcoming analysis.

3 Data

The data for this study is based on a panel dataset, that combines survey data based on a sample of young new entrants and re-entrants to registered unemployment in Vienna with register data from the Public Employment Service Austria (AMS). The survey was conducted by Steiber et al. between 2014 and 2015 and is called “Jung und auf der Suche nach Arbeit in Wien (JuSAW)”. The surveyed individuals are between 18 and 28 years old and were interviewed first at the time of entering the registered job search and then a second time after about one year (Steiber et al., 2015, 2017). Some alterations were made to the data for privacy reasons, like simplifying employment histories. For the first interview there is data for about 1133 individuals after cleaning. Furthermore, the data set has a total of 326 variables, of which 134 are only available for the second wave of observations, which contains significantly less data points and thus were not considered further. Table 2 shows some example features by categories. The online-appendix 1 gives an overview and description of all 127 variables used for further analysis. More detailed explanations and descriptive explorations of all variables can be found in Steiber et al. (2015, 2017).

Following the original AMS algorithm, the prediction goal is the probability estimation of young unemployed finding a job for at least 90 days, in the following 7 months after

Categories	**Example Features**
Socio-demographics	Age
	Gender
	Health
Previous employment/unemployment history	Group of occupation
	Employment history
	Job offer
Desired workplace attributes and motivation	Importance of job securtiy
	Value of leisure compared to work
	Working only for money
Children/Family	Having young children
	Planing further children
Migration Background	Origin in state groups (native, EU, other)
	High school in Austria
Education	Education
	Last grade in German
	Last grade in Maths
Ability	Symbol-numbers test
	Recall of wordlist
	Calculation task
Attitudes, characteristics and feelings	Locus of control: Life outcome depending on myself
	Depression risk score
	Big five: hardworking
Behavior	Alcohol consumption
	Computer games
	Sport
Social components	Meeting with friends
	Relationship status
	Living alone
Parents	Education father
	Education mother
	Parents interested in school performance

Table 2: Example variables in categories

the first interview. Though only the short-term timespan of 7 months is available, all group classifications (High, Medium, Low) will be estimated. Of the 1133 individuals only approximately 1/3 – 385 exactly – managed to sustain employment for at least 90 days in the 7 months time span after the first interview.

Figure 2 shows the proportions of different groups of jobseekers reaching the goal of 90 days, compared to the general population, with 95 % confidence intervals. The red line shows the proportion of unemployed having more than 90 days employment in the whole dataset. The confidence intervals are computed by testing the null hypothesis that the proportions in several groups are the same.

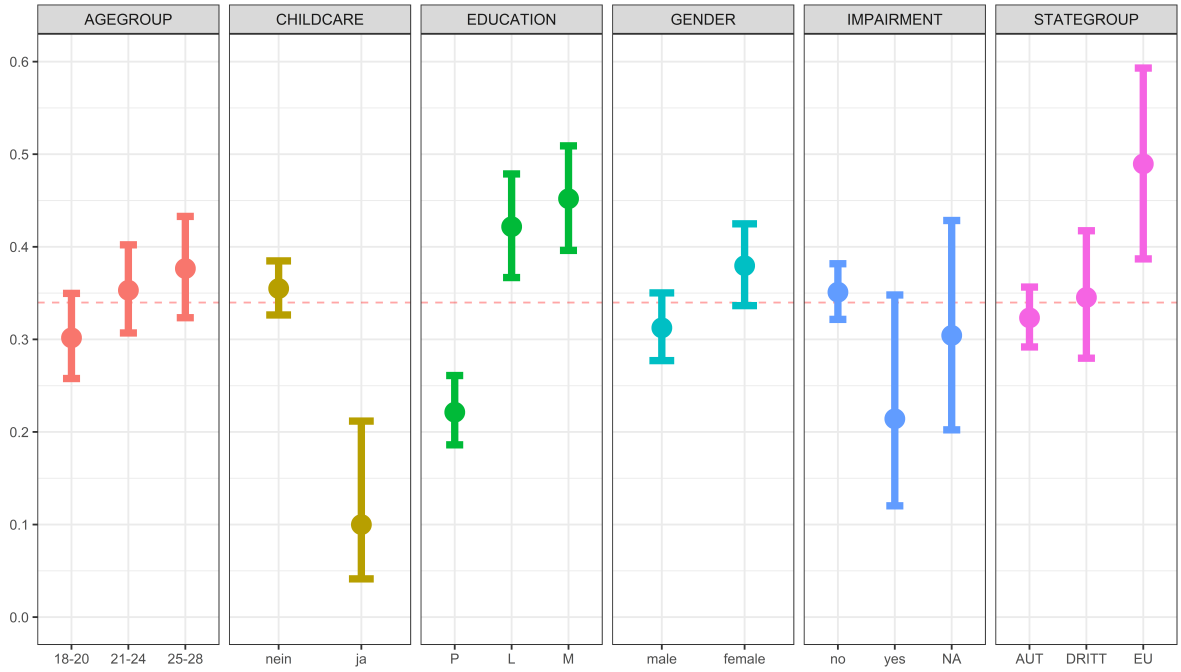

Figure 2: Proportions of different groups of jobseekers reaching the goal of 90 days

Gender and migration background (stategroup) were used as protected features in the fairness evaluation, following European anti-discrimination regulation. Contrary to the general AMS predictions for the full population, where being a woman is having a negative effect on the probability to find a job, in this dataset, where only younger unemployed are looked at, proportionally more women found and had a job for more than 90 days in 7 months after unemployment. A chi-square test reveals that this relation is significant on an alpha level of 0.05. Figure 2 shows the respective proportions for female and male unemployed obtaining employment for more than 90 days with confidence bands. Around 30% of the male jobseekers reached that goal, while almost 40% of the female jobseekers

reached it.
A similar relationship can be seen for native and non-native Austrians. Contrary to the general AMS predictions for the full population, non-Austrians, from other european countries (EU) or non-EU countries (DRITT), actually find jobs more often than native Austrians in the younger population, with being form another EU-country raises the proportions to almost 50%. Again, the chi-square test is significant on a level of 0.05.
For this younger population it is also apparent that older jobseekers, have higher proportions for successful job integration. The same pattern can be observed with the educational degree. Having health impairments aggravate chances on the job market. While the most considerable difference can be seen between mothers with young children and those without. It is important to note, that the AMS algorithm only uses the childcare variable for women.

4 Methodology

This section lists and explains the different modeling and variable selection strategies, named tasks, which were in the comparison to the original AMS model. Furthermore, strategies for evaluating performance and fairness and the corresponding metrics are described. All coding was done in R version 4.1.2 with following packages: For model building and comparison, the package mlr3verse was used (Lang and Schratz, 2022). mlr3fairness (Pfisterer et al., 2022) and yardstick (Kuhn and Vaughan, 2021) for evaluation. The code is available in the following github repository `https://github.com/vik-sz/Algorithmic-Profiling`

4.1 Modeling

This study compares the effects of using various model classes and different feature sets for the prediction of labor market chances of young unemployed. The goal is to predict if newly registered unemployed are able to find a job for more than 90 days in the next 7 months. Therefore, the performance of the original short-term AMS logistic regression, which coefficients were published in Gamper et al. (2020) as Odds-Ratios, is evaluated on this dataset and used as a benchmark. Only the model coefficients for the full and youth population will be used for this. The short-term perspective is evaluated, since the

survey data is not available for 24 months, which would be necessary for the long-term criterion. Furthermore, a featureless model, which always predicts the most frequent label – not finding a job for 90 days in 7 months in this case – will be used as a second benchmark for comparison with all trained models. Following the recommendations in Fernández-Delgado et al. (2014), who showed that in many cases, a small but diverse set of learners is sufficient to choose one ML algorithm that performs sufficiently well, the following six model classes were tested on different variable sets (see also publications for unemployment risk predictions fromKern et al. (2021) and de Troya et al. (2018)):

Logistic regression (LogReg): Common logistic regression, analogous to the original AMS documentation. Likewise, only main effects and no interaction terms were included. Works best on a limited number of covariates and results in interpretable coefficients. Does not require any hyperparameter tuning.

Penalized Logistic Regression with elastic net regularization (PenLogReg): Logistic regression fitted via penalized maximum likelihood (Friedman et al., 2010). The regularization path is computed for elastic net penalty, at a grid of values, for the regularization parameter lambda. In contrast to the unpenalized logistic regression larger feature sets can be used without loss of prediction performance, since irrelevant covariates will be sorted out automatically depending on the value of lambda. Tuning is required to find the best lambda value.

Random Forests (RF): Random Forests are ensembles of multiple decision trees grown on bootstrap samples using random feature subsets (Breiman, 2001). They utilize both bagging and feature randomness to create an uncorrelated forest of decision trees. For classification tasks, the output of the random forest is the class selected by most trees. Multiple hyperparameters can and usually need to be tuned. Random Forests are able to compute variable importance measures and can therefore be used for evaluation of the features.

Extreme Gradient Boosting (xgboost): Extreme Gradient Boosting is a more efficient implementation of the gradient boosting algorithm, which is an ensemble created from decision trees added sequentially to the model to correct the prediction errors made by

prior models. XGboost typically has a lot of hyperparameters that need to be tuned for the model to work efficiently. Similar to Random Forest, xgboost is able to compute variable importance measures for feature selection (Chen and Guestrin, 2016).

k-Nearest-Neighbor (KNN): The k-Nearest-Neighbor algorithm uses the k nearest observations to make a classification. Proximity is computed by a distance metric using the independent features. The label that is most frequently represented in the neighboring points is used for classification. Hyperparameters, like the number of neighbors k, need to be tuned(Hechenbichler and Schliep, 2004).

For hyperparameter tuning, recommendations from Bischl et al. (2021) were followed where possible. The package mlr3tuningspaces was used to obtain predefined search spaces and recommended tuning parameters for the selected learners (Becker, 2022). The hyperparameter search spaces used in this analysis can be found in online-appendix 2. The hyperparameter settings in mlr3tuningspaces were tested and explained in Bischl et al. (2021). Model accuracy was used as the tuning measure. Random search was adopted to find best hyperparameter settings, with the termination number for training being fixed at 1000 evaluations to comply with available computing capacity.
As for the sampling strategy, the data was first divided into a training and test set. Additionally, nested resampling is needed to ensure unbiased estimates of the generalization error during hyperparameter optimization. Although Bischl et al. (2021) recommend to use a high number of cross validation splits for small data sizes like this, holdout was chosen as outer sampling strategy to be able to evaluate all different tasks on the same test data set. This holdout test set was selected randomly among the observations which had no missing values for most of the features, and was then fixed for all tasks to allow better comparability. It comprises 160 observations, which leads to a test/train ratio of at least 0.14 to max 0.23. The split ratio is not equal across tasks, since only complete cases within the respective feature sets were used and therefore lead to slighty varying training set sizes. A more detailed explanation to this can be found below. Exact split ratios for every task are listed in table 3. Stratification ensured that the ratio for the target variable employment days is comparable in the training and test data set. Model training and tuning was done with a 3-fold cross-validation on the training data.

4.2 Variable Selection

A first exploratory analysis gave clues about which variables were worth considering for training. Features with too many missing values or too many categories with few observations, were not looked into further. A total of 127 features remained, which are listed by groups in online-appendix 1 . The final feature selection was done by handpicking and testing diverse variable groups as well as using two different supervised methods – embedded/intrinsic and filter. All selected variable sets are shown in table 3.

Embedded/Intrinsic methods: Through internal evaluation mechanisms some of the used algorithms do feature selection automatically during training. From the used model classes, Penalized Logistic Regression, Random Forest and Extreme Gradient Boosting have such intrinsic methods. In every training round these models automatically choose variables for each of the tasks.

Filter methods: Filtering in contrast, is done before training and is handled independently of the learner. Filter methods select a subset of features based on their relationship with the target variable. There are two methods for this, filtering measures and variable importance filters. Only filtering measures were used in this study for variable selection. Depending on the feature types, there are various statistical metrics available to calculate scores for filtering measures. Since the majority of variables in the dataset are categorical, following measures were used to rank them according to their association with the target variable - days of employment: "disr", "jmi", "mrmr", "cmim" and "relief".
The relief filter is implemented in the FSelectorRcpp package (Zawadzki and Kosinski, 2021), while all other filters are from the praznik package (Kursa, 2021). All filters were used through the mlr3verse package (Lang and Schratz, 2022). The praznik filter algorithms are based on mutual information, which quantifies the amount of information obtained about one variable, through the other variable. For two discrete variables it is calculated by the double sum (Cover and Thomas, 1991):

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} P_{(X,Y)}(x,y) log(\frac{P_{(X,Y)}(x,y)}{P_{(X)}(x) P_{(Y)}(y)})$$

where $P_{X,Y}$ is the joint probability function of X and Y, and P_X and P_Y are the marginal probability functions of X and Y respectively. The relief algorithm calculates weights

based on feature value differences between nearest neighbor pairs. Further explanations to the exact methods can be found in the corresponding package documentations. The filters produced ranked lists of features according to the measured association with the target variable. From those rankings, only the top 15 features were chosen for each filter measure as variable sets for training to keep the model parsimonious.

Handpicked variables: As for the handpicked variables, different strategies were followed to attain a useful variable set. First, the original AMS variables were used to construct three sets. "AMS full" uses all the variables for the adult fully estimable population. "AMS youth" is the reduced variable set for the jobseekers aged below 26, not using stategroup or employment history (Gamper et al., 2020). And "AMS extended" comprises the original AMS variables, but with enhanced characteristics. For example, all job classes instead only production and service groups.

Second, a literature review about unemployment prediction models from PES in other countries was done, to determine what features were mainly picked for such prediction tasks. Online-appendix 3 shows all the respective variable sets from Australia (Lipp, 2005; Caswell et al., 2010), Denmark (STAR, 2018; sta), Netherlands (Wijnhoven and Havinga, 2014; Brouwer et al., 2015), Ireland (McGuinness et al., 2014) and Sweden (Arbetsförmedlingen, 2014) and the corresponding closest available features in my dataset. A joint variable set was constructed using those matches ("other PES" in table 3). Third, the variables were put into categories, which were used to construct five different sets. The first set called "diverse" tries to cover all categories as best as possible, while still incorporating only the most important features, often index variables composed with other features in the category. The four other sets only comprise features of one or two categories and are called "Character", "Behavior", "Attitudes" and "Personality". Since the "Character" set has a larger number of features, it was broken down using the above-mentioned filter methods first before using it for modeling. Even though these sets are not expected to outperform a more diverse feature set, interesting conclusions can be made about how predictive attitudes, personality and behavior is for unemployment.

Depending on the variable sets, different numbers of observations are available for training, since not all survey questions were answered by everyone. While this compromises the comparability of the models, only using observations fully available for all tested features would reduce their number drastically and would therefore lead to massive problems in

Name	#Variables	#Observations	Test/Train ratio
AMS full	12	950	0,17
AMS youth	9	950	0,17
AMS ext	15	1009	0,16
Diverse	32	710	0,23
Filtering Char	25	1090	0,15
Behavior	12	1112	0,14
Attidues	23	1133	0,14
Personality	23	1082	0,15
other PES	27	736	0,22
Filtering disr	15	980	0,16
Filtering cmim	15	892	0,18
Filtering jmi	15	980	0,16
Filtering mrmr	15	904	0,18
Filtering relief	15	948	0,17

Table 3: Variable sets with respective number of variables, observations and split ratios

training and hyperparameter tuning of the model. Some solutions to this problem are proposed in the conclusion under further research, but were not implemented in this study. The different variable selection strategies described above resulted in the sets depicted in table 3. It shows a summary of all variable sets with their respective number of observations and variables, while online-appendix 1 lists the exact features contained in these sets.

4.3 Performance Evaluation

For the evaluation on the test set, various metrics are looked into to assess performance accuracy. Predictions for each task are paired and compared to the real outcomes in a confusion matrix as shown in table 4. While having achieved more than 90 days of employment in 7 months after the interview is seen as the positive outcome and having less than or equal to 90 days as the negative outcome. I will be estimating two threshold variants following the original AMS algorithm. First, High vs. Medium chance group with a threshold of 66%. Jobseekers are labeled as High-Chance, if the probability of finding a job predicted by the model is higher or equal to 66%. If it is less than 66% the jobseeker is labeled in the Medium-Chance group. Second, Medium vs. Low chance will be evaluated. Jobseekers with a predicted probability of less than 25% will be categorized into the Low chance group, while jobseeker with more will get into the Medium chance group. In both set-ups the Medium chance group is receiving more support measures, while High and

Low chance groups obtain less or no support.

Performance may then be evaluated based on measures of this confusion table (accuracy, TPR, TNR, FPR, FNR, PPV, NPV) or based on the original probabilities (ROC-AUC), which were used for class labeling. The confusion matrix measures can further be divided into two groups depending from which direction the relationship between prediction ($\hat{Y}$) and target (Y) is looked at. If the true target Y is given and the fit of prediction $\hat{Y}$ should be evaluated, recall (true positive rate) and the other rates like true negative, false positive, false negative can be examined. If on the contrary the prediction $\hat{Y}$ is given and we want to see how well it predicts the true class Y, then precision, that means positive and negative predictive values, is looked at.

	Labeled H Group	**Labeled M Group**
>90 Days Employment (N)	True Positive – TP	False Negative – FN
≤90 Days Employment (P)	False Positive – FP	True Negative – TN

	Labeled M Group	**Labeled L Group**
>90 Days Employment (P)	True Positive – TP	False Negative – FN
≤90 Days Employment (N)	False Positive – FP	True Negative – TN

Table 4: High vs. Medium chance groups with 66% threshold (first table) and Medium vs. Low chance groups with 25% threshold (second table)

Prevalence: Prevalence is the proportion of all positive values of the target variable (jobseekers with over 90 days of employment) divided by the total number of observations. This formula can also be used for the evaluation of the predictions, that means all positively labeled jobseekers divided by all predictions.

$$Prevalence = \frac{P}{P+N}$$

Accuracy: Accuracy (ACC) is one of the most popular measures used for performance evaluation. It is calculated as the number of all correct predictions divided by the total number of predictions. The range is 0 to 1, with higher values being better. An imbalanced target (as in this case) can lead to problems with accuracy, since a high accuracy score could be reached by simply classifying all observations as the majority class (featureless

model).

$$ACC = \frac{TP + TN}{P + N}$$

True positive rate (Sensitivity or Recall): The true positive rate (TPR) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called sensitivity or recall and the range is between 0 and 1, with higher numbers indicating better performance. In this study, this measure displays how many jobseekers, who found a job for more than 90 days, were correctly predicted positively (i.e., High or Medium chance) compared to all jobseekers with successful integration.

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} = 1 - FNR$$

False negative rate: The false negative rate (FNR) is calculated as the number of false negative predictions divided by the total number of positives. It ranges between 0 and 1, with lower numbers indicating better performance. In this study, this measure illustrates how many jobseekers, who actually found a job for more than 90 days, were falsely predicted negatively (i.e., Medium or Low chance) compared to all jobseekers with successful integration. This measure should be especially low for the Medium vs. Low chance case (25% threshold), since falsely labeling jobseekers into the Low chance group should be avoided as much as possible.

$$FNR = \frac{FN}{P} = \frac{FN}{TP + FN} = 1 - TPR$$

True negative rate (Specificity): True negative rate (TNR) is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called specificity (SP) and the range is between 0 and 1, with higher numbers indicating better performance. In this study, this measure shows how many jobseekers, who did not find a job for more than 90 days, were correctly predicted negatively (i.e., Medium or Low chance) compared to all jobseekers with unsuccessful integration.

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} = 1 - FPR$$

False positive rate: False positive rate (FPR) is calculated as the number of incorrect positive predictions divided by the total number of negatives. The range lies between 0 to 1, with lower numbers indicating better performance. In this study, this measure reveals how

many unemployed, who did not find a job for more than 90 days, were falsely predicted positively (i.e., High or Medium chance). For the High vs. Medium case this would mean that the falsely H labeled would not receive support measures even though they needed it. This measure should be as low as possible for H vs. M (66% threshold), since it is more important to support people who need help than to avoid giving help to those who would not need it.

$$FPR = \frac{FP}{N} = \frac{FP}{FP+TN} = 1 - TNR$$

Positive predictive value (Precision): The positive predictive value (PPV) is calculated as the number of correct positive predictions divided by the total number of positive predictions, with range 0 to 1. Not to be confused with TPR, where the positive predictions are divided by the positive cases. This measure displays how many of the jobseekers labeled positively (i.e., High or Medium chance) were indeed jobseekers with over 90 days of employment.

$$PPV = \frac{TP}{TP+FP}$$

Negative predictive value: The negative predictive value (NPV) is calculated as the number of correct negative predictions divided by the total number of negative predictions, with range 0 to 1. It displays how many of the negatively labeled jobseekers (i.e., Medium or Low chance) were indeed jobseekers with less than 90 days of employment.

$$NPV = \frac{TN}{TN+FN}$$

ROC-AUC: ROC-AUC is the area under the receiver operating characteristic (ROC) curve. It has a range of 0 to 1. The ROC curve depicts the sensitivity (TPR) against 1 – specificity (FPR) at various threshold settings and is therefore based on the calculated probabilities. The goal is to express how much the model is capable of distinguishing between classes. The higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1. If the AUC is 0, the model predicts positive classes as negative and negative classes as positive. If it is exactly 0.5, the predictions are equal to random class pickings.

4.4 Fairness Evaluation

In the EU and many other countries legislation prohibits the discrimination based on certain human features. Article 21 of the EU Charter of Fundamental Rights states following: "Any discrimination based on any ground such as sex, race, color, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited." (Council of European Union)

But how can antidiscrimination law like this be translated into modeling and algorithmic decisions? What makes an algorithm fair? Fairness in machine learning is a relatively new research area, but has grown exponentially in the last couple of years. While there are some publications giving good overviews of popular measures and fairness evaluation methods like Mitchell et al. (2021); Makhlouf et al. (2021); Barocas et al. (2017); Verma and Rubin (2018), new definitions and concepts get published regularly, some of them mutually incompatible (Chouldechova, 2017; Kleinberg et al., 2016; Barocas et al., 2017). In this chapter I will review some popular concepts to find the best fairness measures for the underlying task of employment prediction. Table 5 shows the discussed fairness metrics and their classification in the respective concepts.

Fairness and statistical independence: As a first step, mentioned characteristics in anti-discrimination law can be translated into protected features. Most existing fairness definitions define some kind of protected features (A) as well as privileged and unprivileged groups. In general, one can argue that the algorithm is fair if it behaves the same way for privileged and unprivileged groups of protected features. Translated into the classification setting, this "equal behavior" of the algorithm could be measured by equal performance metrics across groups. But which performance metrics should be inspected? As mentioned in section 4.3 performance measures based on the confusion matrix can be categorized depending from which direction the relationship between prediction ($\hat{Y}$) and target (Y) is looked at. When looking at recall and the "rates" (FPR, FNR, TPR, TNR), that means the true target Y is given and the fit of prediction $\hat{Y}$ is evaluated, equalization between groups of protected features would lead to *"equalized odds"* or *"error rate balance"* in the fairness setting (Verma and Rubin, 2018). In the language of statistical independence this fairness setup would be called **separation**: Given true target Y, prediction $\hat{Y}$ is independent

of group membership $\hat{Y} \perp A \mid Y$ (Barocas et al., 2017).
On the opposite side, when we are looking at precision (positive and negative predictive values), that means the prediction $\hat{Y}$ is given and we want to see how well it predicts the true class Y, equalization across groups would lead to *"predictive parity"* in the fairness setting (Verma and Rubin, 2018). Following Barocas et al. (2017) again this would be called **sufficiency** in the language of statistical independence: Given prediction $\hat{Y}$, target Y is independent of group membership $Y \perp A \mid \hat{Y}$.
In conclusion, Equality of opportunity (**separation**) means that people similar in real life (Y) will be classified similarly ($\hat{Y}$). Predictive parity (**sufficiency**) on contrast means that people who got the same classification ($\hat{Y}$) are, in fact, similar in real life (Y).

Worldviews: But there is also a more philosophical or societal side to the evaluation of fairness. The question of whether our data, that means the measured features, are reflecting the true happenings that one wants to be measured. Friedler et al. (2016) proposes a framework to incorporate the underlying beliefs about the true world (construct space) and our data (observed space). The construct space contains the actual value of the decision criterion, which would be in this case, the ability to find a job in a certain timespan. While the observed space only contains the measurement, which would be the time until a person found a job. Based on the beliefs about the connection between these spaces, they derive two definitions. The first one is "we are all equal" (WAE), where construct and observed space do not align and where differences between privileged and unprivileged groups of protected features only appear in the observed space, while there are no true differences in the underlying world of the construct space since in reality, we are all equal. And second, "what you see is what you get" (WYSIWYG), where the observations accurately reflect the construct and therefore all observable differences between groups are the true differences of what we want to measure. In the context of classification, the WYSIWYG view would urge us to look at *equalized odds* and *predictive parity* to maintain a fair algorithm. While WAE could for example be attained through another, not yet discussed, criterion called *demographic parity*, also known as statistical parity. *Demographic parity* requires the same classification rates for unprivileged and privileged groups. This criterion should be used if WAE is assumed to be true, since the underlying assumptions state that no true differences are apparent in the construct world, so positive classification rates should be equalized. Following Barocas et al. (2017) this criterion would mean **independence** between $\hat{Y}$ and

the protected attributes $\hat{Y} \perp A$.

Translating the world views into the context of PES this could mean: if the AMS believes that their observed space features (time until a person found a job) accurately represent the construct space features (ability to find a job), this scenario aligns with the WYSIWYG axiom. Even if we assume, that everyone is equally capable in doing their jobs, if the AMS is right about the structural discrimination in the hiring process, then it would not matter or even make it worse if one would try to equalize the predicted chances of finding a job based on fairness considerations, albeit nothing is changed in the hiring process of the companies. This could lead to jobseekers not being able to receive support measures even though they would need them. Therefore, to be able to counteract these discriminatory tendencies through support measures it is crucial to first correctly reflect them in modeling. In this case the focus should lie on *equalized odds* or *predictive parity*, so that everyone receives equally accurate predictions. No groups of jobseekers should have a disproportionate risk of an erroneous prediction.

If we believe that any systemic group differences in the observed space are inaccuracies, measurement errors or false discrimination, this scenario follows the WAE axiom. For example, if women in the past received more help in PES, because they were more eager to ask for support measures as men were, then it could have prolonged the unemployment phase for those who did not actually need it. The model based on past data would learn that women generally are longer unemployed, even though their actual ability to find a job is equal to those of men. It would therefore falsely suggest more support measures for women. In this case, the focus should be to achieve *demographic parity*, so men could receive an equal amount of support.

But these were only two examples of how one could argue for both worldviews. Without further research and expert reviews, it is not entirely clear if there is an underlying structural bias in the data or not. So how to decide which measures to employ?

Incompatibility: When deciding over the fairness metrics to use, three more publications revealed relevant information. First Chouldechova (2017); Kleinberg et al. (2016) showed that **separation** (*equalized odds*) and **sufficiency** (*predictive parity*) are impossible to maintain at the same time when prevalence of the target feature differs across groups. Which is true for both protected features evaluated here (gender and stategroup), as was apparent in figure 2 .

Furthermore, Yeom and Tschantz (2021) show that if a decision is to be made about which of the criterions to follow, then **separation** (i.e., *equalized odds*) and **independence** (i.e., *demographic parity*) should be preferred over **sufficiency** (i.e., *predictive parity*). Since **separation** is necessary and **sufficient** to avoid disparity amplification under the WYSIWYG worldview. **Independence** is necessary and sufficient to avoid disparity amplification under the WAE worldview. While **sufficiency** is neither necessary nor sufficient to avoid disparity amplification, irrespective of the worldview. That leaves us with **separation** and **independence**. But Barocas et al. (2017) show that given different prevalence rates, it is also impossible to simultaneously satisfy those two.

Model fairness and decision justness: At this point another recently published article of Kuppler et al. (2021) can help to decide which criterion to focus on. They propose a distinction between model fairness and decision justness. They build their argument on the notion that algorithmic decision-making is a two-step process, where first a prediction is made with observed features – the prediction task - and second a decision is made based on that prediction – the decision task (see also Mitchell et al. 2021Mitchell et al. (2021)). Based on this distinction they define justice as the property of the allocation principle (decision task) and fairness as the property of the prediction algorithm (prediction task). Justice is given with adherence of decision rules to well-justified distributive justice principles and fairness with equal risk of prediction error across individuals. They argue that: "Candidates may deserve different decisions due to differences in the decision criterion. But all candidates deserve that the value of the decision criterion is determined with an equal margin of error." (Kuppler et al., 2021). Based on this they appeal for the case of PES allocation of support measures: "Accordingly, we should prefer a prediction model that correctly represents true differences in Y in the predictions $\hat{Y}$. If female jobseekers really have a lower probability of re-employment and, therefore, need more support, the prediction model should reflect this gender difference. Only then can the distribution rule allocate more support to female jobseekers." (Kuppler et al., 2021).

Conclusion: Based on this argument, achieving accurate performance and minimizing error measures across protected groups will have priority in the following examination of fairness. This will be achieved by mainly focusing on *equalized odds* and *error rate balance*. While the discussion of just distribution of support measures is left to domain

experts and decision makers. Nevertheless, the results section will display all mentioned fairness metrics, while the core target for selecting a fair algorithm will lie on *equalized odds* and *error rate balance*.

Individual vs. Group: Another often-made differentiation in fairness that was not discussed yet is individual vs. group-wise perspective. Individual fairness is given, if individuals who are similar based on their characteristics are treated equally. Group-wise fairness on the contrast demands similar treatment in regard to whole demographic groups. All of the above-mentioned fairness criterions can be categorized in group-wise metrics. Individual fairness metrics would be for example consistency or counterfactual fairness. Nevertheless, this study will only analyze group measures, while individual measures will be left to further research.

Fairness Language	Statistical Independence Language	Metrics	Worldviews	Group/ Individual
Separation	Equalized Odds; Error rate balance	TPR, TNR, FPR, FNR	WYSIWYG	Group
Sufficiency	Predictive Parity	PPV, NPV	WYSIWYG	Group
Independence	Demographic Parity	Prevalence	WAE	Group

Table 5: Overview of fairness concepts and measures

5 Results

A total of 73 models, five model classes with 14 different variable settings plus two models from the AMS documentation and a featureless prediction, were evaluated in regard to performance and fairness on the hold-out test set of 160 jobseekers. All tasks were evaluated for the High vs. Medium chance groups with a 66% threshold and Medium vs. Low chance groups with a 25% threshold. In the following, results of three variable sets will be looked at in detail, while all other results can be found in online-appendix 4. For this comparison one model with AMS variables, one with handpicked additional features and one with variables inspired by other PES algorithms will be examined. The results for performance and group-wise performance are depicted in figures 3 - 8. The figures display all the performance measures mentioned in section 4.3 grouped by models and tasks. The first plot for each task shows the performance overall, while the second one depicts differences between male (privileged) and female (unprivileged) jobseekers and the third one differences between Austrian (privileged) and non-Austrian (unprivileged) unemployed.

The differences were computed by subtracting the supposedly unprivileged groups value from the privileged groups value. This means positive values appear if privileged groups have higher values and negative if unprivileged groups have higher values. For the "true" rates (TPR, TNR), PPV, accuracy and ROC-AUC positive (blue) values mean favoring privileged groups and negative (red) values mean favoring unprivileged groups. For the "false" rates (FNR, FPR) it is exactly the other way around. If the differences are around zero, the fields are colored yellow.

Generally, when comparing all models, it was apparent that models with better performance measures tended to perform a little worse in fairness relevant metrics. This trade-off was already observed in multiple publications (Corbett-Davies et al., 2017; Desiere and Struyven, 2021). Furthermore, specialized data sets, as the one for personality or attitude features, performed worse, while tasks with a diverse feature set did better. Only the models with behavioral features did perform decently compared to models with more diversified predictor sets. Therefore, it seems that while personality and attitudes are not as good predictors for unemployment when used by themselves, behavioral sets, with variables like alcohol consumption, are more predictive. Note that a more detailed investigation is required to evaluate the exact strength of each feature and especially if there are any causal relationships.

Overall, the diverse variable set performed best across model classes in performance measures. But the "other PES" task did almost as good and had less group-wise differences between privileged and unprivileged groups and could therefore be seen as fairer. Overall, using additional features from the survey, as compared to only AMS/register data, was beneficial for performance and fairness. Looking at the model classes, xgboost was the only class always under the best three models in regard to accuracy and fairness. Accuracy between all models and tasks for the 0.66 threshold, excluding the AMS documentation and featureless models, varied between 0.59 and 0.67. ROC-AUC had maximum scores up to 0.72. Performance for most tasks is therefore comparable to other reported PES and unemployment prediction algorithms, even though the data used did not have as many observations (Desiere et al., 2019; Kern et al., 2021). Almost all models and tasks managed to outperform the featureless model, which only classifies jobseekers into the majority group of the target variable.

The figures below show the results for the variable sets: AMS youth, Diverse and other

PES, for the categorization in High vs. Medium chance group and Medium vs. Low chance group. The jobseeker was classified in High, if the score surpassed 66% and into Low if below 25%. A positive outcome indicates the jobseeker getting a job for more than 90 days in 7 months. With a threshold of 66% almost no model predicted any positive outcomes. That means almost no one from the test set was categorized into the High chance group. On average the model classes performed best for the diverse feature set. The best accuracy was reached with 0.67 by the logistic regression with a diverse set and 66% threshold. Generally, the 25% threshold leads to worse accuracy levels overall. The best ROC-AUC (for both thresholds the same, since it is computed from the scores) of 0.72 was reached by a random forest model with the diverse variable set as well. The FPR is important to look at in the High vs. Medium case, since it displays how many unemployed were falsely predicted into the High chance group, and would therefore not receive any help and support measures even though they needed it. This rate should therefore be as low as possible. But since almost no positive classifications were made with this threshold, every task has FPR around 0. For the Medium vs. Low case, the FNR is more important to examine, since jobseekers in the Low chance group don't get access to AMS support measures and will be sent to an external institution. Therefore, the FNR should be as low as possible, so nobody will be falsely excluded from support measures.

AMS: Comparing the three AMS variable models - AMS youth, AMS full and AMS extended - , the best results in average over the model classes, which were trained on this data, were reached by the AMS full set. But the AMS documentation model for the adult population (full variable set) did perform worse than the model for the younger population, which is not surprising, since the dataset here contains only jobseekers aged 28 and younger. Interestingly the AMS extended task, which contains enhanced variables (e.g. more categories per variable) did perform worst of the AMS tasks. But overall the differences between these three sets are quite small, therefore only the results for AMS youth will be discussed further (s. figures 3, 4). The accuracy for the AMS youth models varies between 0.6-0.63 for the 66% threshold and 0.58-0.59 for the 25% threshold. The ROC-AUC lies between 0.61 and 0.68, with logistic regression performing best. Almost all of the models have a score of 1 for TNR and 0 for TPR in the H vs. M case, since no positive predictions were made. For the 25% threshold, the AMS youth task does not perform very well, with highest accuracy lying by 0.59. Furthermore, the FNR suggests

it favors female and native Austrian jobseekers, since it falsely predicts more males and non-native jobseekers into the Low chance group, which leads to less support from AMS. For the 66% threshold only the AMS documentation model seems to favor male and Austrian jobseekers. All the other models hover around zero differences in regard to gender or state-group.

Diverse: The figures 5 and 6 depict the results for the diverse variable set. This task seems to perform best regarding overall performance measures, with accuracy as high as 0.67 with a logistic regression and a ROC-AUC of 0.72 for the penalized logistic regression. In this task multiple models managed to predict some positive values with a 66% threshold. Of those labeled into the High chance group, 73% to 100% were indeed jobseekers with over 90 days of employment (precision/PPV). But still only 3% to 25% of all jobseekers with over 90 days were detected as positives (recall/TPR). And despite having more positively labeled values here, the FPR is still around 0, which is important so no one needing help would be denied of it. For the 25% threshold, the accuracy lies between 0.54 and 0.64, with logistic regression being the best again. The average FNR is higher than for the AMS models, which is a negative point. As for gender and state group fairness, a lot of the models do have higher groupwise differences, favoring women and native Austrians. But the average bias for all the models is still around zero. For the 25% case more male and non-Austrian get falsely labeled into the Low chance group.

Other PES: The figures 7 and 8 depict the results for the "other PES" variable set. The variable set using features recommended by other PES did perform above average with accuracy ranging between 0.6 and 0.62 for 66% and 0.51 to 0.61 for 25% threshold, while the ROC-AUC is ranging from 0.65 to 0.71. The gender and state group differences are even lower as for the diverse variable set for both thresholds, with most measures fluctuating around zero. Therefore, even though this task did not outperform the diverse-set in overall performance, the combined variable set of other PES algorithms can be considered as one of the best tasks overall, because of the higher fairness.

Demographic Parity: Figures 9 and 10 show the mean percentage of jobseekers being predicted into the higher group conditional on their group belonging, with the mean being formed over all self-trained models (logistic regression, penalized logistic regression,

random forest, xgboost and knn). When the predicted percentage across protected groups is equal, this is called demographic parity. There is a trend visible in both threshold cases, that tasks that did not use gender or stategroup as training features, like attitudes, behavior, characteristics and personality, lead to group categorizations more similar across the privileged and unprivileged groups. But these tasks also had the worst general prediction accuracy and performance. This trend is more visible in the 25% threshold case, since a lot of the models and tasks in the 66% case did not predict anyone into the High chance group, therefore it is harder to see clear differences there. Tasks that had better performance measures, like the diverse set, lead to higher differences in class categorization between the protected groups. But equally to the performance measure evaluation above, the “other PES” task achieved high performance and still performing better at demographic parity than the diverse set. Therefore, the general trade-off between accuracy and fairness is again evident, with the “other PES” set being some kind of exemption.

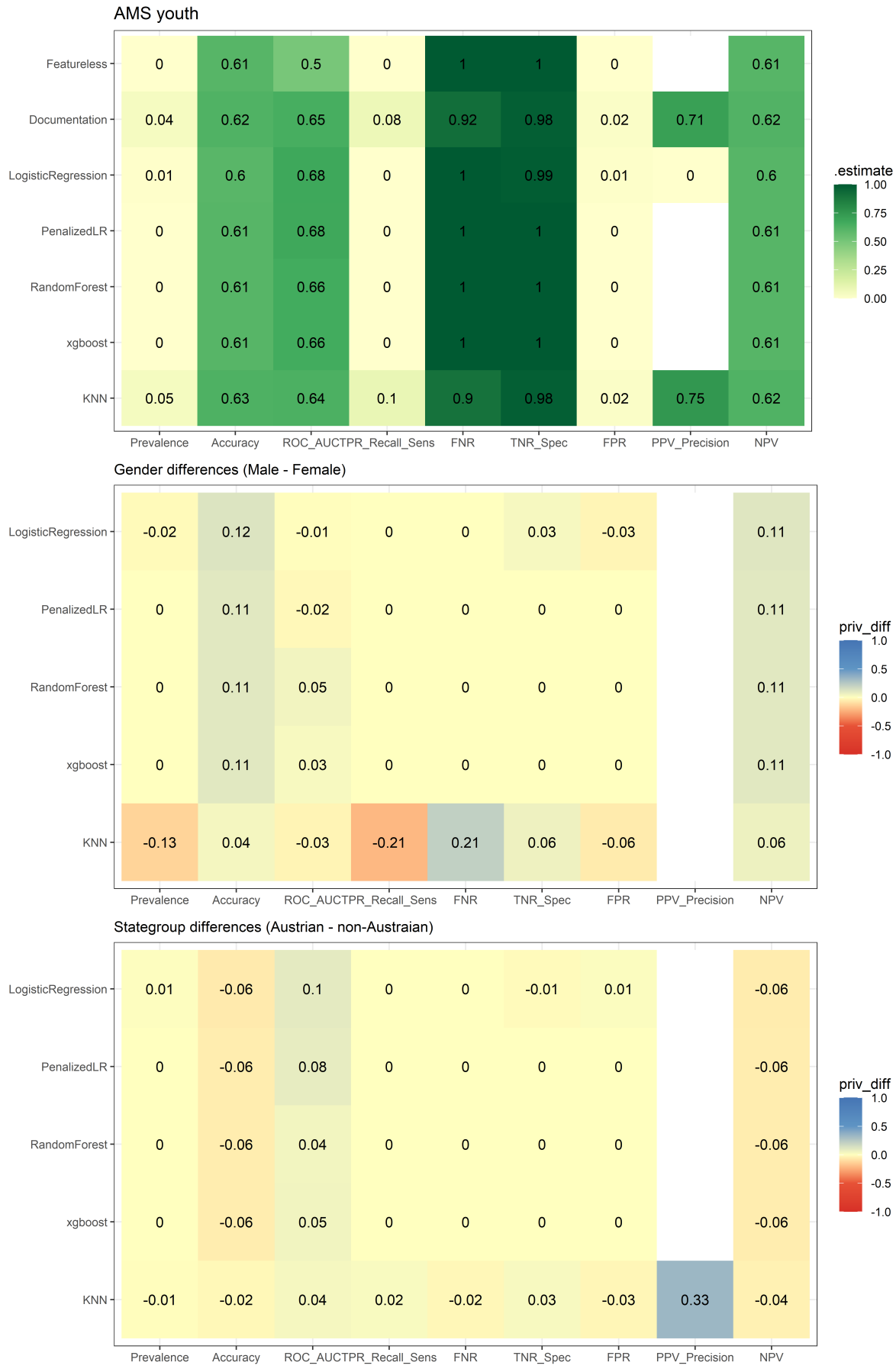

Figure 3: Heatmap of various performance and fairness measures for all models for the "AMS youth" task with 66% threshold

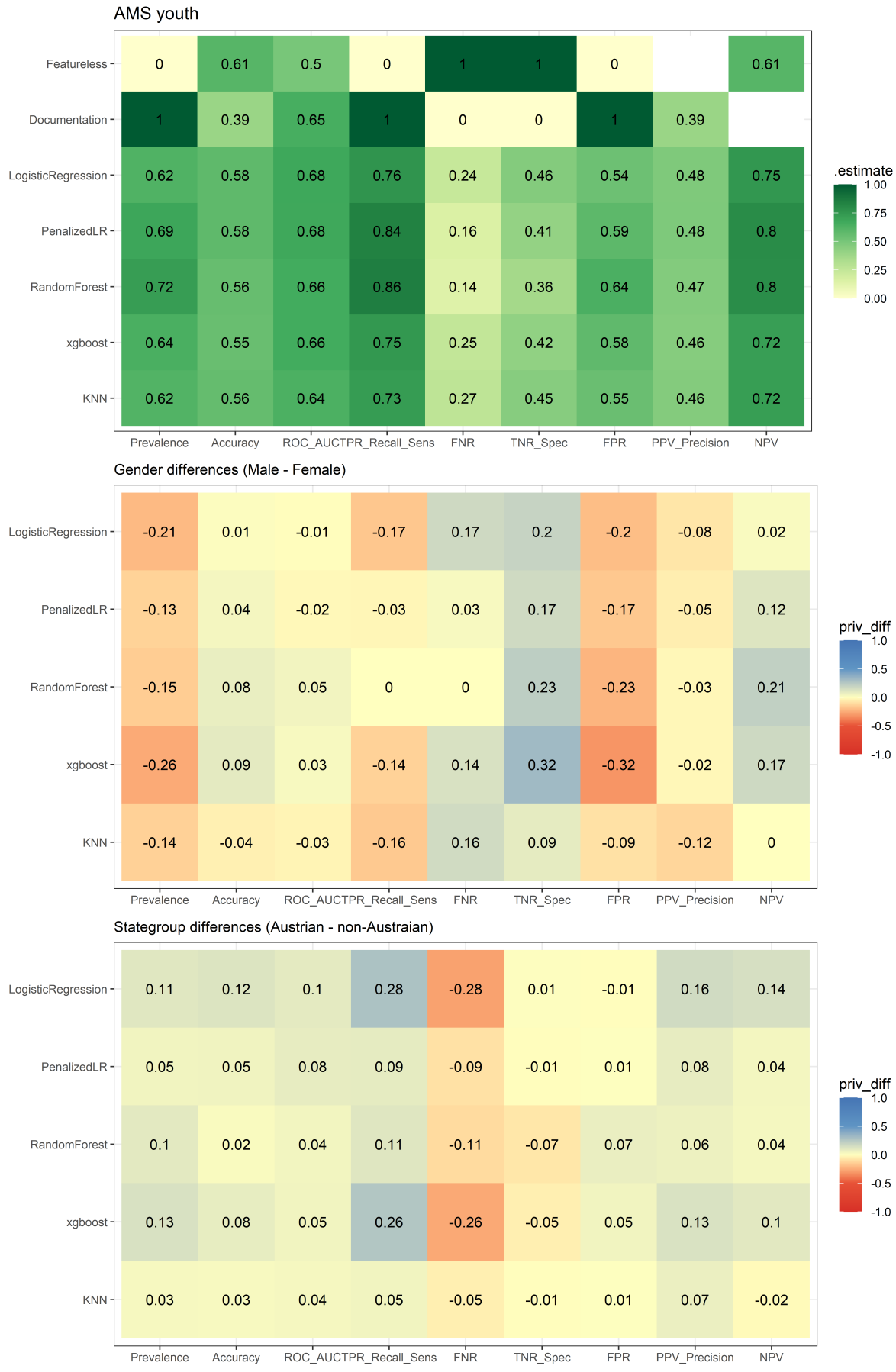

Figure 4: Heatmap of various performance and fairness measures for all models for the "AMS youth" task with 25% threshold

Figure 5: Heatmap of various performance and fairness measures for all models for the "diverse" task with 66% threshold

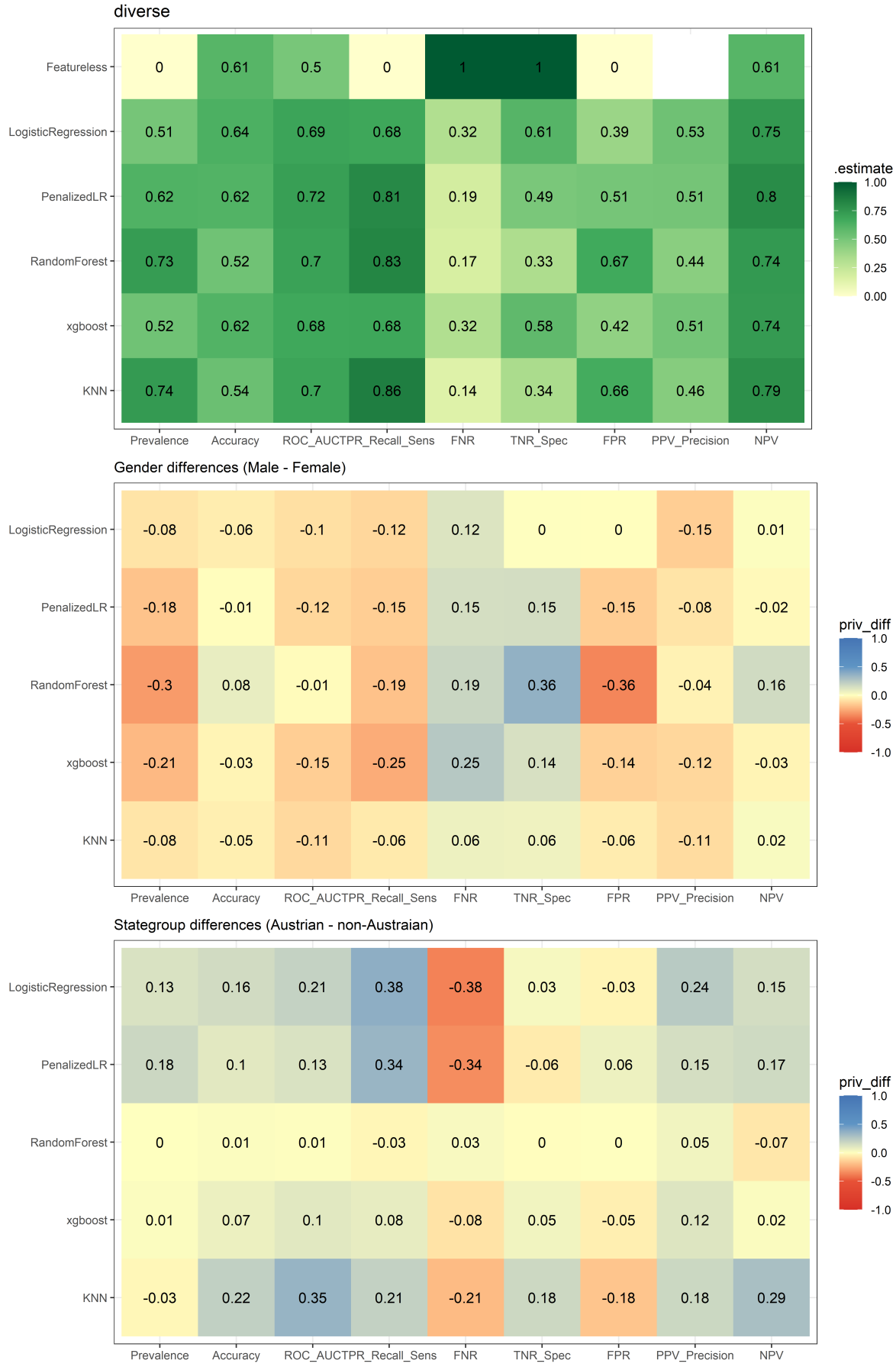

Figure 6: Heatmap of various performance and fairness measures for all models for the "diverse" task with 25% threshold

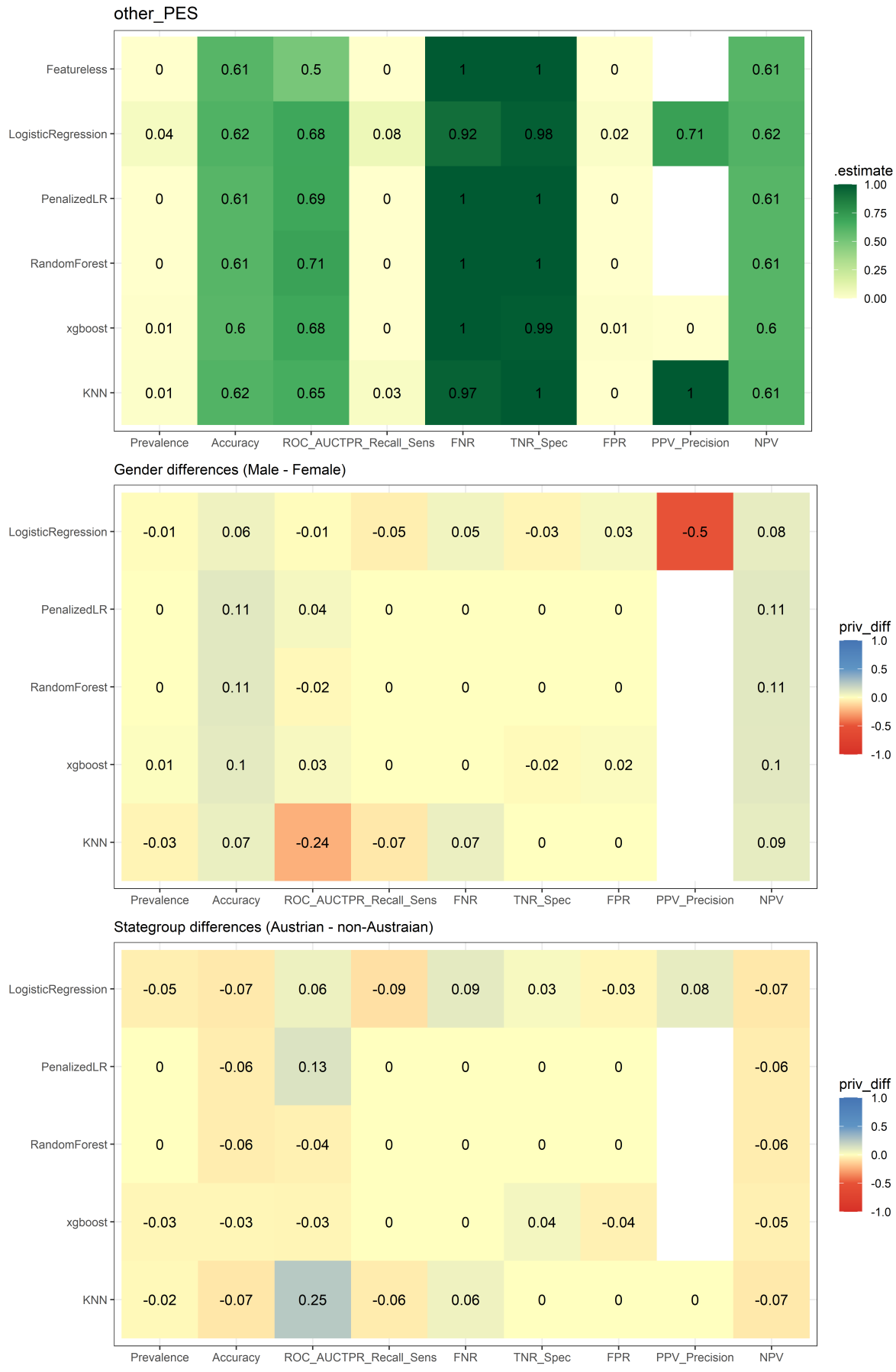

Figure 7: Heatmap of various performance and fairness measures for all models for the "other PES" task with 66% threshold

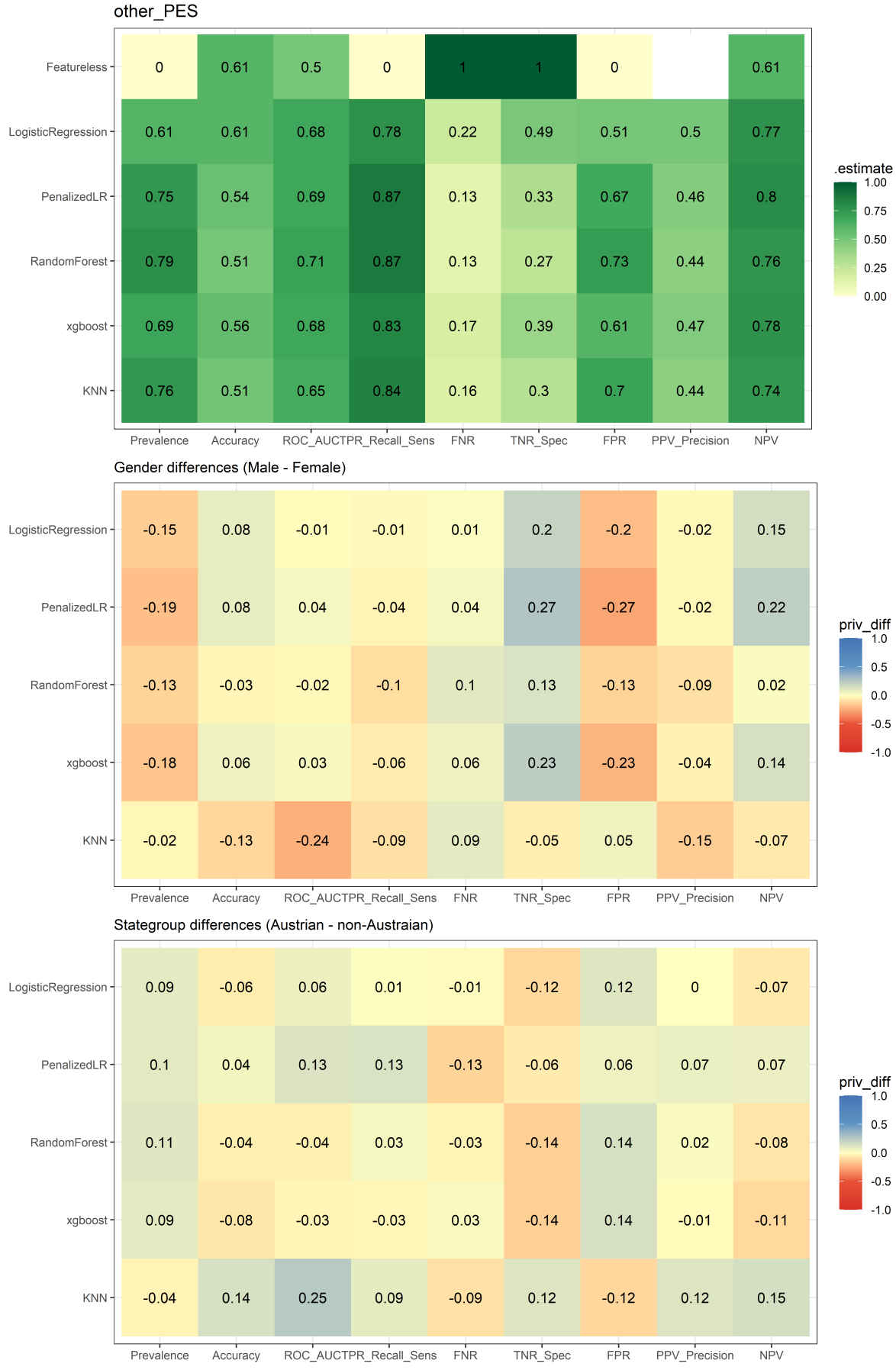

Figure 8: Heatmap of various performance and fairness measures for all models for the "other PES" task with 25% threshold

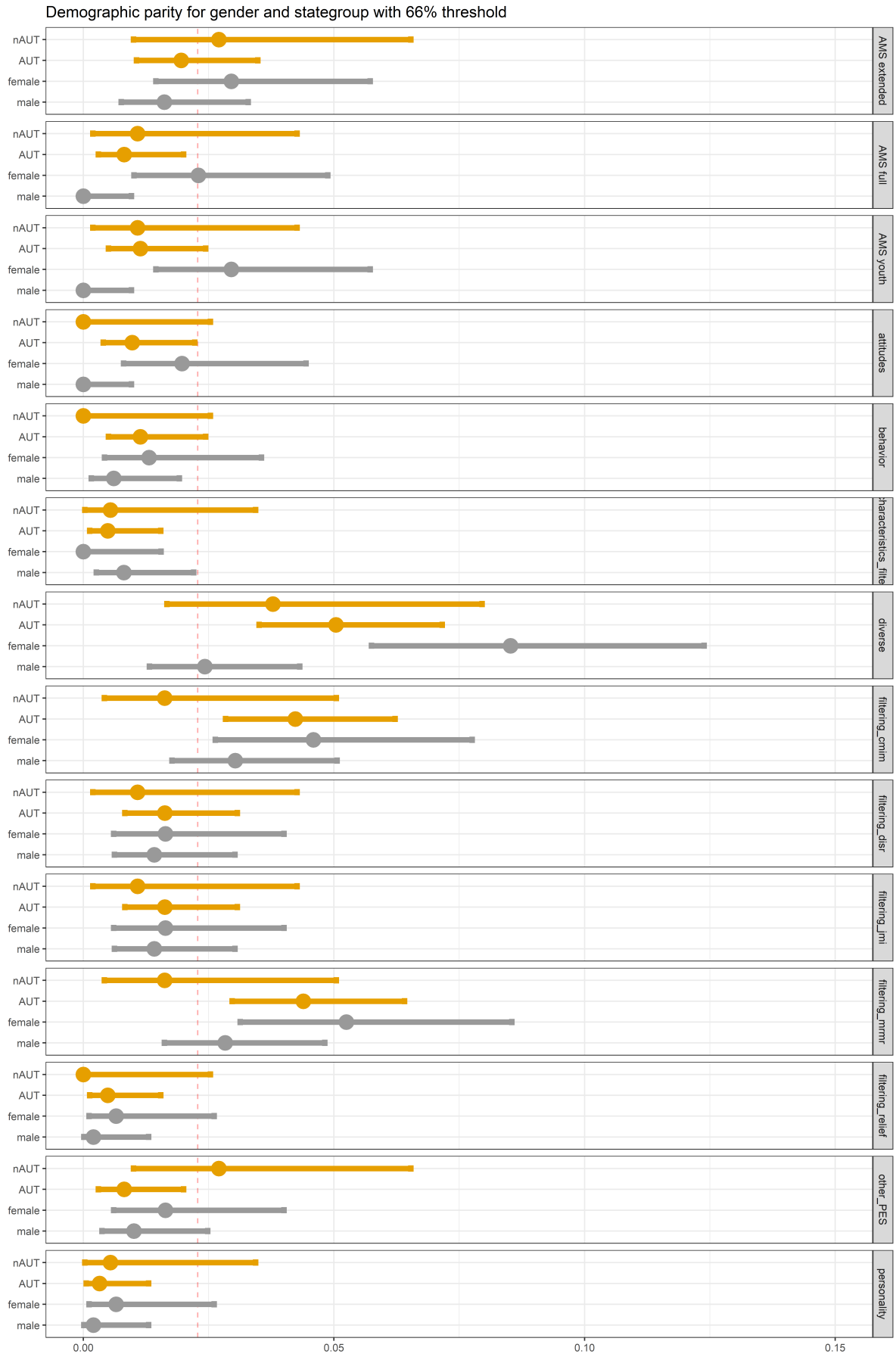

Figure 9: Proportions of protected groups predicted into the High chance group (66% threshold case)

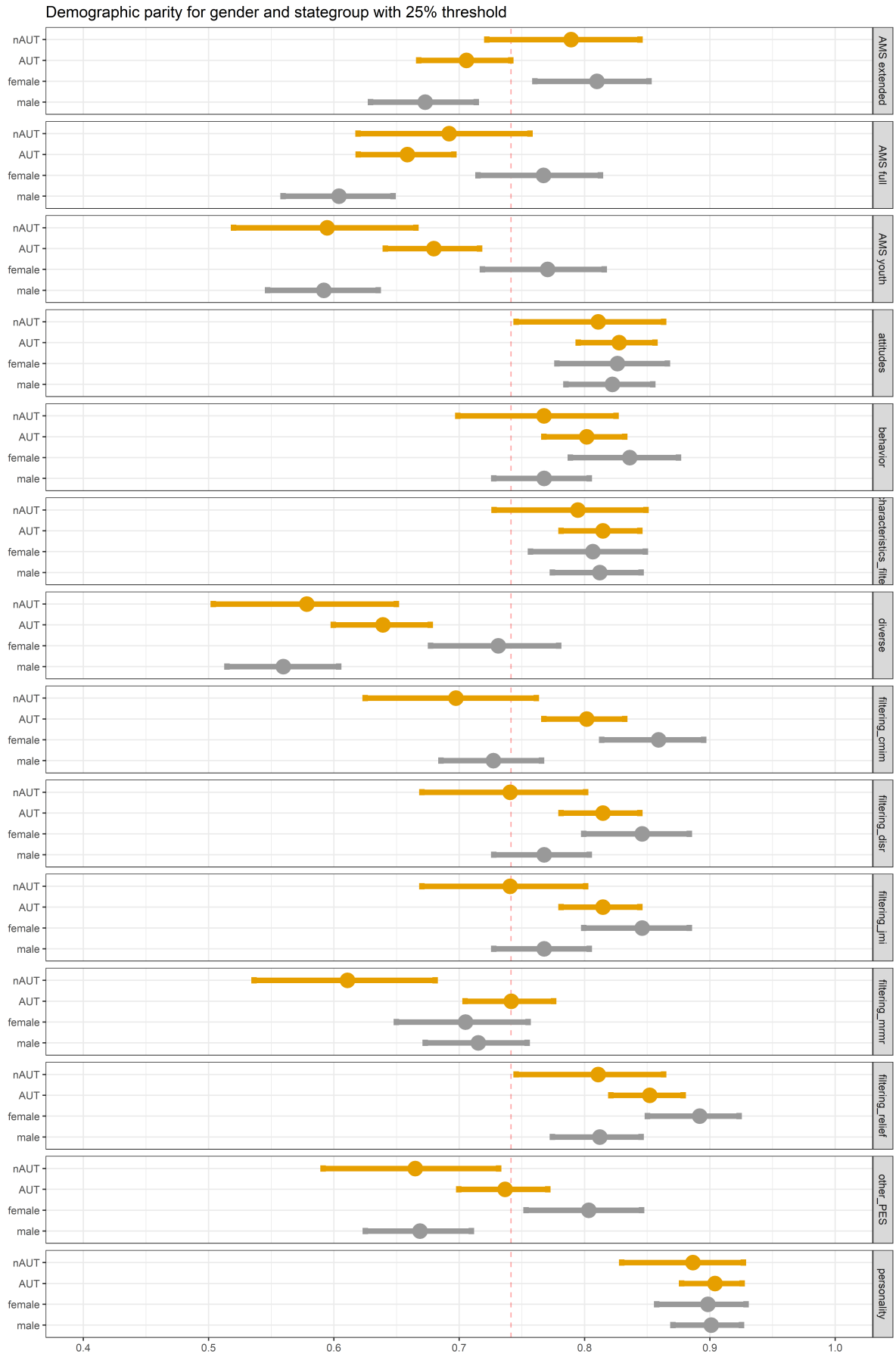

Figure 10: Proportions of protected groups predicted into the Medium chance group (25% threshold case)

6 Conclusion

Summary: Comparing all model estimations, it was apparent that additional survey features were beneficial to performance and fairness, compared to exclusively using AMS/register data. While the most diverse feature set, covering all available feature categories, performed best in overall performance accuracy, the set including features commonly used by PES of other countries did not only perform well with general performance measures, but also proved to be fair in the chosen fairness measures in regard to gender and migration background. Concluding the literature review, the focus of the fairness evaluation lied on *equalized odds* and *error rate balance* (**separation**) over *predictive* (**sufficiency**) and *demographic parity* (**independence**).

The specialized feature sets, like personality or work attitudes, did perform worse in general. Only the models with behavioral features appeared to be more predictive for unemployment risks. Note that a more detailed investigation would be necessary to evaluate the exact strength of the behavioral feature associations and especially for detecting causal relationships. Comparing model classes, xgboost performed best on average over all tasks in both performance and fairness. Accuracy for all models and tasks varied between 0.59 and 0.67 and ROC-AUC had maximum scores up to 0.72. Therefore, the performances are generally comparable to other reported PES and unemployment prediction algorithms, even though the data used less observations (Desiere et al., 2019; Kern et al., 2021). As in other publications a trade-off between model accuracy and fairness was found (Corbett-Davies et al., 2017; Desiere and Struyven, 2021).

Limitations: Because of the relatively small dataset and many missing values for some variable combinations, it was not possible to use the same number of observations for every task. Therefore, only one pre-set test set was determined to maintain a minimum level of comparability of the results. To avoid unstable results, it would have been preferable to make multiple cross-validation splits in the outer and inner resampling, as Bischl et al. (2021) is recommending for data sets of this size.

There are two other limiting factors of the dataset. First, for most PES long-term unemployment is more relevant than short-term unemployment, since it causes more costs for the state. But usually, long-term unemployment is defined as a period of more than 12 and often even 24 months of unemployment. Here, data was only available for 7 months

after unemployment registration. While the survey was repeated 12 months after the first interview and has a question of reemployment in it, only a few observations are available for this feature. Therefore, only the 7 months period was used in this analysis.
Second, since the dataset only comprises jobseekers 28 and younger living in Vienna, the results in this study are not easily transferable to the entire population of jobseekers. It was already visible when comparing the AMS documentation models, that using the coefficients for the full population lead to massive performance loss. Nevertheless, since multiple PES in other countries already demonstrated the usefulness of survey data, one can assume that the main conclusions would hold for the entire population.

Further research: A possible improvement to the problem of small dataset size, would be the imputation of missing values. With no missing values the number of observations would not change from task to task, therefore test and train-set splittings would be easier to perform. Another possible method would be to include the missing values as an additional category, since the majority of available variables are from categorical type. This is valid under the assumption that missing values have some information in them. For example, people might not answer health impairment questions because of possible repercussions. Both of these approaches could be tested in further research.
As already seen when comparing the 66% threshold with the 25% threshold, different threshold settings can have considerable influence on performance and fairness. This point was also proved by Kern et al. (2021) for unemployment risk prediction with German PES data. Therefore, an evaluation of more than two threshold settings should also be an important aspect in upcoming research.
Another relevant research point are individual variable effects. This study focused on distinct variable groups, but did not examine individual variable importance. For more refined statements about the importance of certain features and investigation of causal relationships for unemployment risk prediction, a more detailed analysis should be performed in the future.
Another relevant research area is the evaluation of various kinds of support measures. In this study, all support was assumed to be equally useful for everyone. But this is not necessarily the case. To generate the highest impact on job chances, the allocation of appropriate measures is an important objective and should therefore be included into any decision making about support distribution.

Conclusion: In conclusion, it should be a valid consideration for other PES to implement surveys to improve their unemployment risk assessment predictions. Additional survey data is already used by multiple PES in Europe (Desiere et al., 2019) and proved to be not perceived as an additional burden by jobseekers during unemployment registration (Wijnhoven and Havinga, 2014). An overall high performance should be an essential goal, since false categorizations could lead to limited access to support measures and therefore disadvantages on the job market. Groups already receiving discrimination on the labor market, like jobseekers with migration background or older women, should receive special attention in the modeling process. Therefore, it is a necessity to check prediction fairness for multiple demographic groups.

References

The danish star algorithm. https://ledighedsalgoritmen.dk/. Last accessed: 2022-09-25.

Doris Allhutter, Florian Cech, Fabian Fischer, Gabriel Grill, and Astrid Mager. Algorithmic profiling of job seekers in austria: how austerity politics are made effective. *frontiers in Big Data*, 3:5, 2020a.

Doris Allhutter, Astrid Mager, Florian Cech, Fabian Fischer, and Gabriel Grill. Der ams-algorithmus. eine soziotechnische analyse des arbeitsmarktchancen-assistenz-systems (amas). endbericht. 2020b.

Arbetsförmedlingen. Arbetsförmedlingens Återrapportering 2014: Insatser för att förhindra långvarig arbetslöshet (arbetsförmedlingen reports 2014: Efforts to prevent long-term unemployment). 2014.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2, 2017.

Marc Becker. *mlr3tuningspaces: Search Spaces for Hyperparameter Tuning*, 2022. URL https://CRAN.R-project.org/package=mlr3tuningspaces. R package version 0.1.1.

Bernd Bischl, Martin Binder, Michel Lang, Tobias Pielok, Jakob Richter, Stefan Coors, Janek Thomas, Theresa Ullmann, Marc Becker, Anne-Laure Boulesteix, et al. Hyperparameter optimization: Foundations, algorithms, best practices and open challenges. *arXiv preprint arXiv:2107.05847*, 2021.

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

Sandra Brouwer, RH Bakker, and JMH Schellekens. Predictors for re-employment success in newly unemployed: A prospective cohort study. *Journal of Vocational Behavior*, 89: 32–38, 2015.

Dorte Caswell, Greg Marston, and Jørgen Elm Larsen. Unemployed citizen or 'at risk'client? classification systems and employment services in denmark and australia. *Critical social policy*, 30(3):384–404, 2010.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

Council of European Union. Eu charter of fundamental rights - article 21: Non-discrimination. `https://fra.europa.eu/en/eu-charter/article/21-non-discrimination`.

Thomas M Cover and Joy A Thomas. Information theory and statistics. *Elements of information theory*, 1(1):279–335, 1991.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. *arXiv preprint arXiv:1408.6491*, 2014.

Íñigo Martínez de Rituerto de Troya, Ruqian Chen, Laura O Moraes, Pranjal Bajaj, Jordan Kupersmith, Rayid Ghani, Nuno B Brás, and Leid Zejnilovic. Predicting, explaining, and understanding risk of long-term unemployment. In *32nd Conference on Neural Information Processing Systems*, 2018.

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

Sam Desiere, Kristine Langenbucher, and Ludo Struyven. Statistical profiling in public employment services: An international comparison. 2019.

Alexander Fanta. Jobcenter-algorithmus landet vor höchstgericht. `https://netzpolitik.org/2021/oesterreich-jobcenter-algorithmus-landet-vor-hoechstgericht/`, 2021. Accessed: 2022-09-27.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness (2016). *arXiv preprint arXiv:1609.07236*, 2016.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

Jutta Gamper, Günter Kernbeiß, and Michael Wagner-Pinter. Das assistenzsystem amas. zweck, grundlagen, anwendung, 2020.

Klaus Hechenbichler and Klaus Schliep. Weighted k-nearest-neighbor techniques and ordinal classification. 2004.

J Holl, G Kernbeiß, and M Wagner-Pinter. Personenbezogene wahrscheinlichkeitsaussagen ("algorithmen"). Technical report, Technical report, Synthesis Forschung Gesellschaft mbH, 2019.

Jürgen Holl, Günter Kernbeiß, and Michael Wagner-Pinter. Das ams-arbeitsmarktchancen-modell. *Arbeitsmarktservice Österreich, Wien*, 2018.

Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. Machine bias. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`, 2016. Accessed: 2022-09-27.

Christoph Kern, Ruben L Bach, Hannah Mautner, and Frauke Kreuter. Fairness in algorithmic profiling: A german case study. *arXiv preprint arXiv:2108.04134*, 2021.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Max Kuhn and Davis Vaughan. *yardstick: Tidy Characterizations of Model Performance*, 2021. URL `https://CRAN.R-project.org/package=yardstick`. R package version 0.0.9.

Matthias Kuppler, Christoph Kern, Ruben Bach, and Frauke Kreuter. From fair predictions to just decisions? conceptualizing algorithmic fairness and distributive justice in the context of automated decision-making. *Frontiers in Sociology*, page 155, 2021.

Miron B. Kursa. Praznik: High performance information-based feature selection. *SoftwareX*, 16:100819, 2021. URL `https://doi.org/10.1016/j.softx.2021.100819`.

Michel Lang and Patrick Schratz. *mlr3verse: Easily Install and Load the 'mlr3' Package Family*, 2022. URL `https://CRAN.R-project.org/package=mlr3verse`. R package version 0.2.4.

Michael Lechner and Jeffrey Smith. What is the value added by caseworkers? *Labour economics*, 14(2):135–151, 2007.

Robert Lipp. Job seeker profiling: The australian experience. In *EU-Profiling Seminar*, 2005.

Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of machine learning fairness notions. *ACM SIGKDD Explorations Newsletter*, 23(1):14–23, 2021.

Seamus McGuinness, Elish Kelly, John R Walsh, et al. Predicting the probability of long-term unemployment in ireland using administrative data. *Economic and Social Research Institute (ESRI) Research Series*, 51:1–29, 2014.

Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

Florian Pfisterer, Wei Siyi, and Michel Lang. *mlr3fairness: Fairness Auditing and Debiasing for 'mlr3'*, 2022. https://mlr3fairness.mlr-org.com, https://github.com/mlr-org/mlr3fairness.

Tarun Ramadorai, Andreas Fuster, Paul Goldsmith-Pinkham, and Ansgar Walther. Predictably unequal? the effects of machine learning on credit markets. 2017.

STAR. Profilafklaringsvaerktøj til dagpengemodtagere (profiling tool for unemployment benefit recipients). 2018.

Nadia Steiber, Monika Mühlböck, Stefan Vogtenhuber, and Bernhard Kittel. Suche nach arbeit in wien. 2015.

Nadia Steiber, Monika Mühlböck, Stefan Vogtenhuber, and Bernhard Kittel. Jung und auf der suche nach arbeit in wien: Beschreibung des jusaw-paneldatensatzes und analysen von verläufen zwischen den beiden umfragezeitpunkten. endbericht modul 2. 2017.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

Martijn A Wijnhoven and Harriët Havinga. The work profiler: A digital instrument for selection and diagnosis of the unemployed. *Local Economy*, 29(6-7):740–749, 2014.

Barbara Wimmer. Der ams-algorithmus ist ein ̈paradebeispiel für diskriminierung. https://futurezone.at/netzpolitik/der-ams-algorithmus-ist-ein-paradebeispiel-fuer-diskriminierung/400147421, 2018. Accessed: 2022-09-27.

Samuel Yeom and Michael Carl Tschantz. Avoiding disparity amplification under different worldviews. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 273–283, 2021.

Zygmunt Zawadzki and Marcin Kosinski. *FSelectorRcpp: 'Rcpp' Implementation of 'FSelector' Entropy-Based Feature Selection Algorithms with a Sparse Matrix Support*, 2021. URL https://CRAN.R-project.org/package=FSelectorRcpp. R package version 0.3.8.