Biases in Credit Risk Classification

Ian Rolls and **Helena von Leupoldt**
{iarolls,hevonleupoldt}@davidson.edu
Davidson College
Davidson, NC 28035
U.S.A.

Abstract

Accurately predicting loan repayment is essential for a bank to maintain profitability. With the increasing popularity of machine learning in loan risk prediction due to its accuracy, we want to know if these models are equitable. Our objective was to create a precise risk prediction model and analyze any inherent demographic biases. We used two different datasets from different periods to examine bias changes over time.

Our highest-performing models were an SVM model with an F1 score of 0.829 and a Random Forest model, which obtained a comparable score of 0.905. Our bias analysis found that demographic information could be accurately predicted with other features, meaning models without explicit demographic data can still use that information. This indicates the presence of demographic bias even in models that adhere to current anti-discrimination regulations.

1 Introduction

Accurately predicting a potential loaner's likeliness of repaying a loan is vital to the profitability of a loaning institution. Individual banks have different methods of approving loans, as maintaining a safer portfolio is crucial to a bank's profitability. The usage of machine learning techniques amongst financial industries has been growing more popular as machine learning developments have made models increasingly accurate. Because of this, it is likely that most banks use machine learning in some way during the loan approval process. Our aim is to create an accurate risk prediction model to analyze if and how the model biases risk based on certain demographic features, specifically age, and gender. Our hope is that, by understanding these biases, institutions can take steps to correct these models before using them to guide decisions in the real world.

There is a large literature base on loan risk prediction models. Rather than predicting a variable score (i.e. credit score, credit rating), some papers perform a binary classification (i.e. 0 for didn't default, 1 for defaulted). Moscato et al. developed models for lending in a peer-to-peer network (P2P) with a binary target. They used a dataset containing 900k observations and 151 features, including state, income, and employment. They did not include any demographic data in their model. They found that the random forest was the best-performing model, with a ROC AUC of 0.717 (F1 score of 0.654) (Moscato, Picariello, and Sperlì 2021). Alonso et al. analyzed model prediction on loan datasets with differing levels of information to find the best-performing models at the different levels. They also found that random forest outperformed all other models for virtually all levels of information (Alonso and Carbó 2020).

Hassani et al. specifically examined the potential for reenforcing demographic biases in credit risk machine learning models. If datasets capture biased data, this can pollute model evaluation with bias. Then, if these models are used to evaluate the same systems, they re-enforce the bias. Regardless of whether personal demographic data is explicitly contained in the dataset, this data can be correlated with other data points like income, employment, etc. This means regardless of the explicit inclusion/exclusion of demographic features, it is important to analyze models for bias. This study used all non-demographic features within their dataset to predict an individuals demographic features and found that gender and ethnicity could be predicted very accurately, with F1 scores of 0.843 and 0.986, respectively (Hassani 2021). This means that demographic data can be well-entrenched in non-demographic data. Therefore, regardless of privacy and fairness laws, demographic bias can still implicitly affect machine learning models.

2 Dataset

We analyzed two datasets containing the information of loan applicants who were **approved** for loans. This means we cannot comment on biases in the loan approval process outside of a machine-learning model. However, we can still examine what biases exist within the model itself.

German Credit Data

The first dataset, German Credit Data, was collected from the UCI Machine Learning Repository and contains loan data from Germany from 1994 provided by Professor Dr. Hans Hofmann from the University of Hamburg. It includes $10,000$ observations, 20 variables, and a binary label differentiating between good and bad loans. The seven numerical variables are: duration in months, credit amount, installment rate in percentage of disposable income, present residence since, age in years, number of existing credits at this bank, and the number of people being liable to provide maintenance for. The thirteen categorical variables are: status of existing checking account, credit history, purpose,

savings account/bonds, present employment since, personal status and sex, other debtors/guarantors, property, other installment plans, housing, job, telephone, and foreign worker.

Bondora Dataset

The second dataset is the Public Reports Loan Dataset from Bondora, which consists of 276, 382 unique entries and 112 features, including gender, employment status, employment area, years at current employment, marital status, country, income, previous loan data, age, number of dependents, and many more.

Unfortunately, a lot of this dataset had missing entries, especially for key variables like employment status, marital status, and employment duration. We removed these variables, leaving us with 36, 064 entries. Additionally, we removed some sparse and irrelevant features from the dataset, leaving us with 35 features. Specifically, we removed features that contained information that lenders would not have had at the point of application, like the number of late payments, amount of late payments, etc.

3 Exploratory Analysis

German Credit Data

To effectively use the dataset for classification purposes, we encoded the categorical variables by assigning each category to a corresponding numerical value. Then, we encoded the numerical values with higher max values than the average variable, namely duration in months, credit amount, and age in years, into four categories, according to the statistics seen in Table 1. The four categories relate to the distribution of the data. The categories are based on percentiles: values below the 25th percentile, between the 25th and 50th percentile, between the 50th and 75th percentile, and above the 75th percentile.

Variables	25 perc.	50 perc.	75 perc.
Duration in Months	12	18	24
Credit Amount	1365.5	2319.5	3972.25
Age in Years	27	33	42

Table 1: Percentiles for Numerical Variables in German Data

The Job variable distribution (Figure 1) shows that most people in the dataset are skilled employees or officials. There are almost the same amount of unskilled - residents, as management/self-employed/etc., while there are barely any entries of people that are unemployed/unskilled.

The Foreign worker distribution (Table 2) shows that most people in the dataset are foreign workers. Thus, if there is a bias towards foreign workers in the classifier, that would only account for a small subset of data. Consequently, we cannot draw any conclusions about bias towards foreign workers.

The Social status and sex variable (Figure 2) show a bias in either data collection or loan application in 1994 since no single females are in the dataset. The most prominent group is single males, with 409 observations. Furthermore,

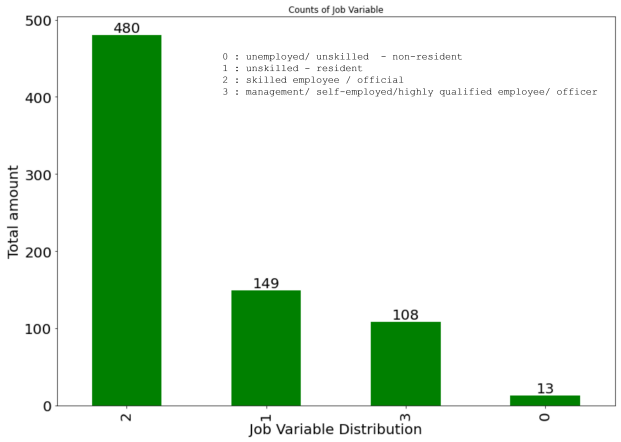

Figure 1: Counts Job Variable German Data

Foreign Worker	Counts
yes	721
no	29

Table 2: Counts Foreign Worker Variable German Data

men are distinguished between married/widowed and divorced/separated, while non/single females are put into one category. This further biases the dataset since the algorithm will distinguish females and males on different levels.

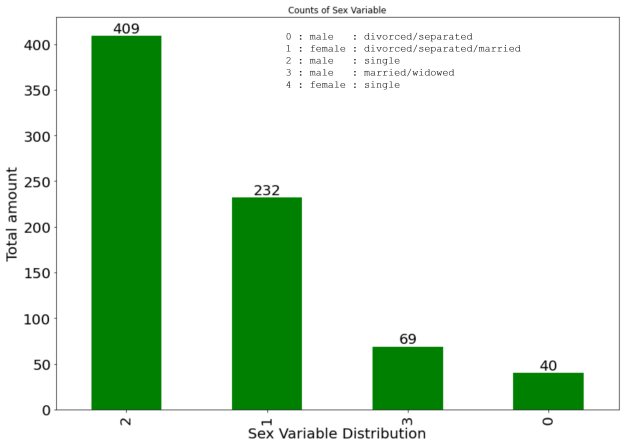

Figure 2: Counts Social Status and Sex Variable German Data

Lastly, the distribution of the labels (Table 3) shows that there are more than two times the number of good loans compared to the number of bad loans. This means that no matter how good the performance of the resulting classifier is, concerning future predictions, it will always be overfitted towards good loans and less tuned to recognize characteristics of bad loans.

Label	Counts
Good (1)	525
Bad (2)	225

Table 3: Counts Label German Data

Bondora Dataset

Table 4 shows the distribution of countries in the dataset. The data only contains observations from four European countries, and the number of observations varies from country to country. Identical to the first dataset, our target is a binary variable: we assign each individual a 1 if they successfully paid off their debt and a 2 if they defaulted on their loan. All of the data is extremely recent, as all data is from loans that ended in late 2022 or early 2023.

Country	Observations
Estonia	22,057
Spain	7,378
Finland	6,334
Slovakia	295

Table 4: Distribution of Countries

Unlike the German Credit data, Gender and Marital status variables are separated in the Bondora dataset. The distribution of entries between men and women is much closer, and there is an additional column for individuals who don't identify as either a man or a woman (Figure 3).

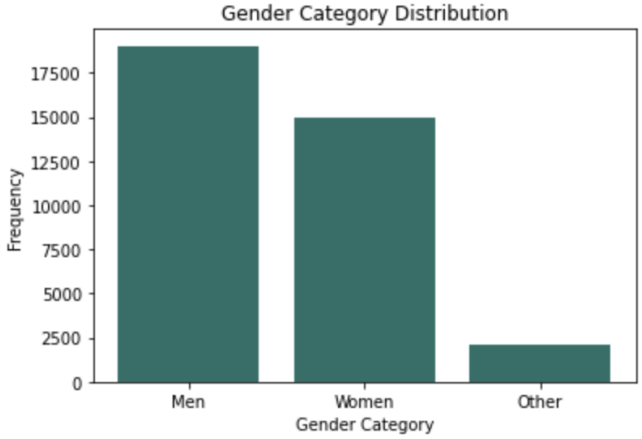

Figure 3: Gender Distribtion of the Bondora Dataset

Table 5 shows the distribution of marital status entries for each of the gender categories. Note that the majority of men are single, whereas the majority of women are married. Thus, we might expect a smaller percentage of women to default on their loans because a married couple is more likely to be financially stable as there is another person responsible for paying the loan.

Gender	MaritalStatus	Frequency
Male	Single	38.9%
	Married	28.8%
	Cohabitant	24.2%
	Divorced	7.3%
	Widow	.05%
Female	Married	31.3%
	Cohabitant	29.9%
	Single	24.1%
	Divorced	11.5%
	Widow	3.2%
Other	Single	57.0%
	Married	23.9%
	Divorced	10.5%
	Cohabitant	7.3%
	Widow	1.2%

Table 5: Frequency of Marital Status Entries for Each Gender Category

4 Experiments

We first performed a train test split of 3:1 on both datasets to evaluate our models in an intermediate stage without contaminating the test data.

German Credit Data

For the German data model development, we considered four classifiers, a support vector machine (SVM), a Logistic Regression model evaluated using stochastic gradient descent (SGD), a random forest classifier, and a k-nearest neighbors classifier (k-NN). We used an F1-score and 10-fold grid search for all models to determine the best model given a specific parameter grid. After training the models on the training data, we compared the F1 scores across the models to choose the best-fitting model.

The SGD classifier was trained with a log-loss function and Lasso (l1) or Ridge (l2) regularization, using learning rates (eta0) of 0.0001-0.015 and regularization sensitivity (alpha) of 0.00005-0.00009. The best estimator used an l2 penalty with alpha=0.00007 and eta0=0.005, achieving an F1-score of 0.662.

The k-NN classifier was trained with the options of a Manhattan distance (p1) or Euclidean distance (p2), with a number of neighbors (n) of 15, 20, and 25. The best estimator had a p1 penalty with n equals 20 and an F1-score of 0.752.

The SVM classifier was trained using a linear or an RBF kernel, with regularization (C) values of 10, 20, 30, 40, and 100 and kernel coefficients (gamma) of 0.05, 0.01, 0.005, and 0.001. The best estimator had an RBF kernel with C=20 and gamma=0.01, achieving an F1-score of 0.829.

The Random Forest classifier achieved an F1-score of 0.663.

Consequently, as Table 6 shows, the SVM is the best classifier for German Credit Data.

Model	F1 Score
SGD	0.662
KNN	0.752
SVM	0.829
Random Forest	0.663

Table 6: F1 Scores for the German Credit Data Models

Bondora Dataset

For the Bondora dataset, we tested 5 different classification models using a 10-fold cross-validation on the split test data. SVM, SGD, k-NN, AdaBoost, and a random forest. Similarly, we used the F1 score and grid search for these models. We also looked at precision and recall as incorrectly predicting someone to be a good debtor is much more harmful to a bank than incorrectly predicting someone as a defaulter.

We found that SVM scored highest using an rbf kernel, which makes sense because of the model's relatively low number of features. We found for both the AdaBoost and random forest models that n_estimators=200 was the best-performing parameter. For the k-NN model, we found k = 13 to be the highest-scoring value with Euclidean distance.

Model	F1 Score
SGD	0.708
KNN	0.8189
SVM	0.8469
Adaboost	0.9000
Random Forest	0.9219

Table 7: F1 Scores for the Bondora Dataset Models - Intermediate CV Testing

For the Bondora dataset, table 7 shows the various scores for each machine learning model. As you can see, the models with boosting seemed to outperform the others, as both Adaboost (.900) and Random Forest (.922) scored very high in the intermediate cross-validation testing. We chose to focus the rest of our analysis on our best-performing models.

5 Results

German Credit Data

The SVM classifier for the German Credit Data optimized with 229 support vectors for the first class, 191 support vectors for the second class, an F1 score of 0.829, and a mean accuracy score of 0.870 for the training data.

Figure 4 shows the confusion matrix for the training data. The biggest issue seems to be the classification of observations with an actual Class 2 label. This can stem from the fact that there are fewer Class 2 observations, which might lead to the model overfitting for Class 1 observations. This is particularly critical as accurately predicting bad loaners is much more important than predicting good loaners in real-life applications.

For the test data, the SVM classifier achieves a mean accuracy score of 0.744, a false discovery rate of 0.308, and a true positive rate of 0.672.

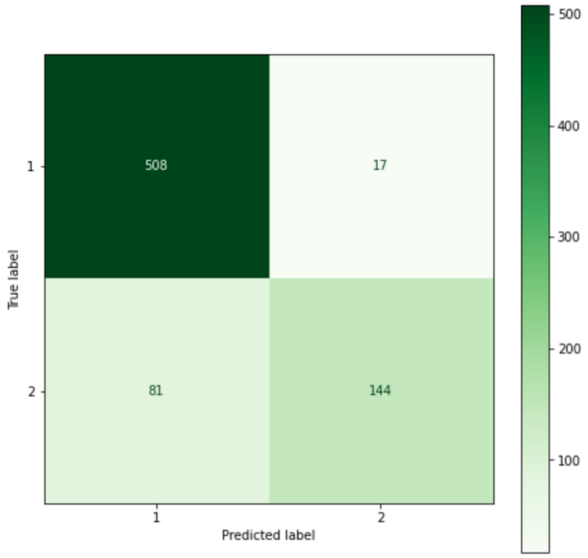

Figure 4: Confusion Matrix SVM on Train Data German Data

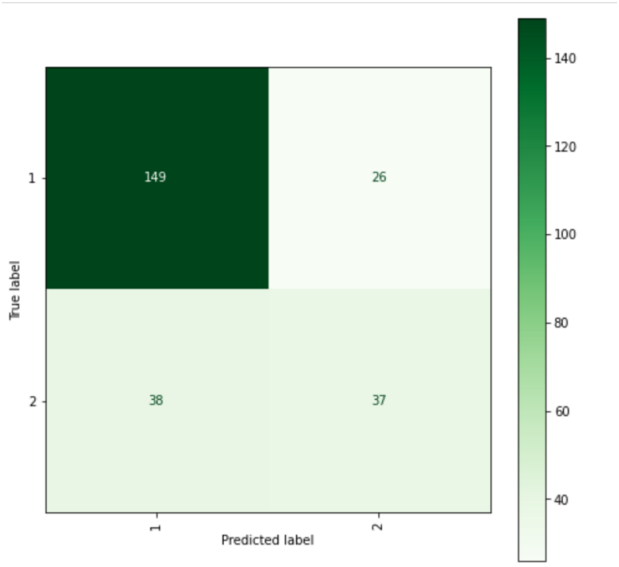

Figure 5: Confusion Matrix SVM on Test Data German Data

Figure 5 shows the confusion matrix for the test data and makes the initial problem of misclassifying objects of Class 2 even more prominent. Here the classifier successfully classifies as many Class 2 observations correctly as incorrectly.

To further understand the structural biases in the dataset, we drew out the best decision tree from the random forest classier. Since this classifier is built based on each variable's explanatory value, the idea is to understand better which variables might have some implicit biases.

This first subtree (Figure 6) shows that non-foreign workers' observations are more quickly classified as good

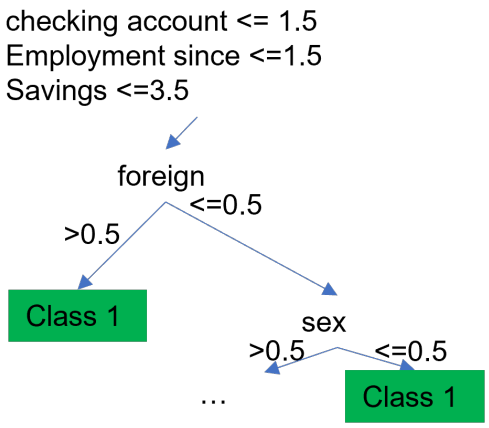

Figure 6: RandomForest Subtree Graph

loans. This is relevant as there are fewer observations of non-foreign workers than of foreign workers. Furthermore, looking one branch down, the graph shows that divorced/separated men are more likely to have a good loan given that they are foreign workers.

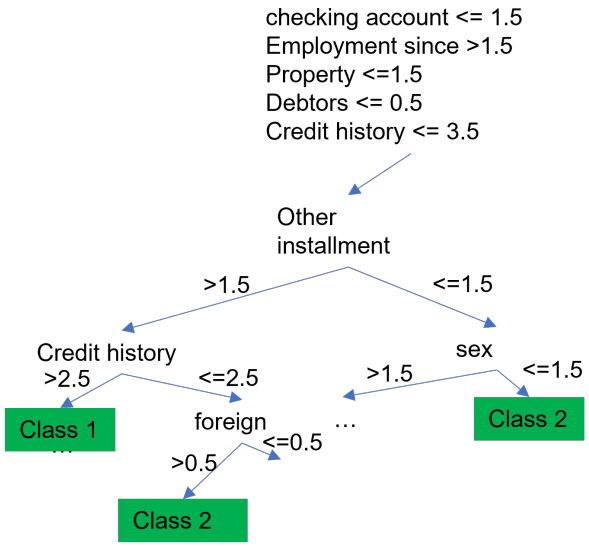

Figure 7: RandomForest Subtree Graph

This second subgraph (Figure 7) shows one possible origin of the misclassification problem. Here, an observation with a bad credit history (Credit history > 2.5, which means a delay in paying off previous credits or the account being labeled as critical) is labeled as having a good loan. Furthermore, a person with a good credit history (all credits paid back duly to the point of application) that is not a foreigner is labeled as having a bad loan. On the other side of the tree, the rightmost branch displays a bias concerning gender and social status, as females and separated/divorced men (sex $\leq$ 1.5) are more quickly classified for a bad loan.

The preceding subgraph (Figure 8) shows that young people are more quickly classified as having a good loan when they don't have other debtors (debtors > 0.5), are property owners (property $\leq$ 1.5), and have worked for more than four years (employment since > 1.5). On the contrary, older people (age > 2.5) with a bad credit history (credit history > 3.5) who ask for a higher amount (amount > 0.5) if they do not have a telephone (telephone $\leq$ 0.5) in their name are classified as bad loans.

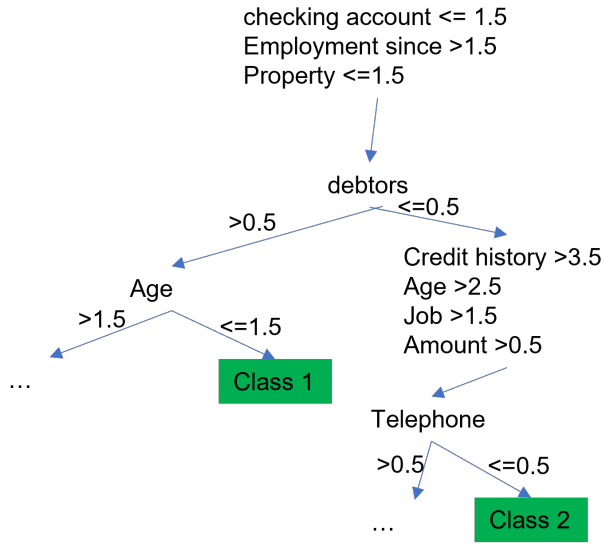

Figure 8: RandomForest Subtree Graph

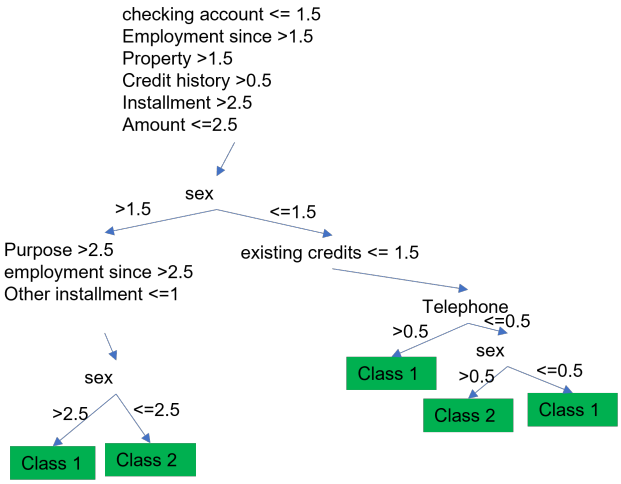

Figure 9: RandomForest Subtree Graph

As the subtree above (Figure 9) shows, given all other aspects being the same, a married man (sex > 2.5) will be classified as having a good loan, while a single man (1.5 < sex $\leq$ 2.5) will be classified as having a bad loan. The right branch also shows the importance of a telephone again (assumably a measure of technology). Telephone holders are more quickly classified with a good loan. If they do not own a phone, we see another example of gender bias: females

(0.5 < sex ≤ 1.5) get classified as having bad loans while males (sex ≤ 2.5) have good loans.

Lastly, we ran a classifier that predicted the Gender and Age variable separately. This was done in an effort to examine inherent gender and age bias in the dataset. An SVM classifier with an Rbf kernel, C = 100, and gamma = 0.01 achieved an F1-score of 0.847 for predicting the gender column. This value is above the accuracy score for the original classifier and thus suggests a high bias towards gender in the dataset, which would still persist to a similar degree if the gender column were to be removed from the dataset. A similar result can be found for predicting the Age variable. An SVM classifier with an Rbf kernel, C = 10, and gamma = 0.05 achieved an impressive F1-score of 0.973. Consequently, the dataset is even better suited to predict the age of the people in the dataset. Overall, this suggests that any classifier trained on the German Credit Dataset will have internalized biases toward Gender and Age. Thus, this paper suggests a first idea of what a classifier for credit data could accomplish, but in order for it to be a fair classifier, further alterations are necessary to reduce the inherent bias.

Bondora Dataset

We decided to select our three best-performing models for our intermediate CV scoring to test the final data on since we knew SGD and k-NN were not going to perform nearly as well as the others, given the complexity of the dataset. Table 8 shows the final testing results. What is most surprising is the drop in the random forest score. This suggests that the random forest model slightly overfits the training data. The AdaBoost and SVM models both maintained a similar performance on the testing data, but overall random forest was still the most accurate classification model.

Model	F1 Score
SVM	0.8556
Adaboost	0.8944
Random Forest	0.9051

Table 8: Final F1 Scores for the Bondora Dataset Models

Additionally, for the two boosting models, the random forest model had a much higher recall than AdaBoost. The random forest model misclassified 195 out of the 6697 total individuals predicted as good. On the other hand, AdaBoost misclassified twice the number of people, at 388. This gave the random forest model a recall score of 0.9708 versus Adaboost's score of .9420. This can be seen in figures 10 and 11, which display the true positive/negative and false positive/negative count for each model.

We first looked at the gender bias of our model. For the random forest model, the scikit-learn package includes a built-in attribute that returns the gini importance factor for each feature. This is a measure of how important a feature is to the model, which is calculated by how often a feature is used in a split and how well the feature is able to split the data. In the model containing gender and age features, both age and gender have a high importance ranking, with age the 10th/35 most predictive feature and gender the 14th/35

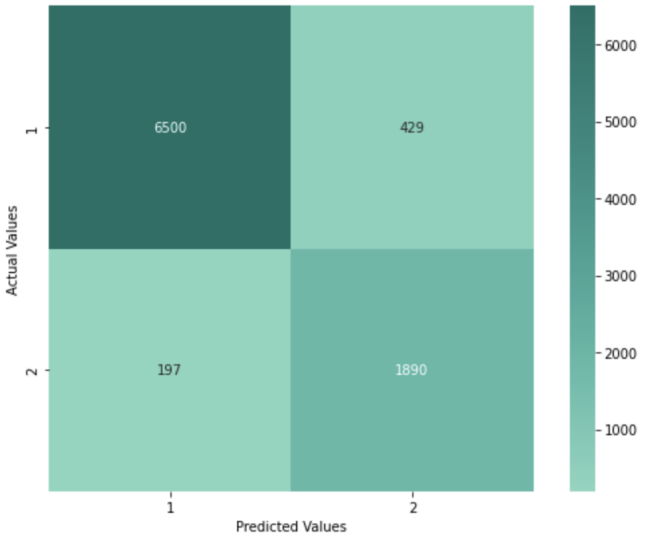

Figure 10: Confusion Matrix for Random Forest Model Predictions

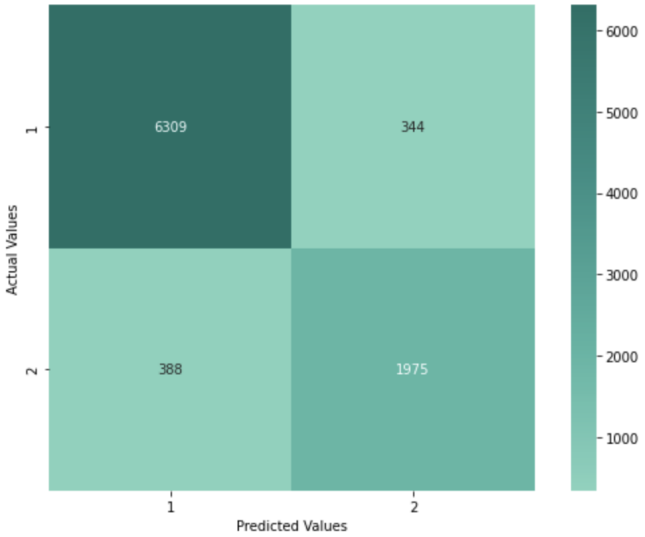

Figure 11: Confusion Matrix for Adaboost Model Predictions

most predictive feature. This shows that the random forest model, including gender and age, uses gender category differences to make biased splits. However, existing legislation prevents lenders from using demographic information to actively bias decisions, so it is necessary to examine how these models perform/bias without access to demographic information (FDIC 2021).

Since our highest-performing model was the random forest, we ran two additional datasets through our random forest model. The first does not contain Gender but contains Age, while the second contains neither Gender nor Age. Each model is fit on the entirety of the training data and evaluated on final testing data. Table 9 shows the F1 scores of

the random forest model tested on the original dataset compared to the model tested on the two other datasets.

Model	F1 Score
Gender and Age	0.9050
No Gender, Age	0.9068
No Gender, No Age	0.9070

Table 9: F1 Scores for Demographic Models

As one can see, all the models are, statistically speaking, identically performing. We have two hypotheses for why this occurred. The first is that demographic features do not significantly affect the model, which we know is not true because both gender and age are in the top 15 most important features. This leads us to believe the second hypothesis—the different models see identical performance because demographic data is implicitly contained within the other data, as the Hassani paper concludes. To investigate this hypothesis, we used two additional random forest models, which each have the same features minus Gender and Age. Each model tries to predict the gender and age of an individual, given all other features. This will enable us to see how much gender/age information implicitly exists within the other features.

Model	F1 Score
Gender Prediction	0.757
Age Prediction	0.672

Table 10: F1 Scores for Age and Gender Prediction Models

Table 10 shows the results of each prediction model. For the age prediction model, we had to split the age category into three buckets (0-31, 31-42, 42-77), as the random forest could not accurately predict specific age. The F1 score for the age prediction model was 0.672, which is twice as good as a baseline model, which would have an F1 score of 0.333 (as all entries are evenly distributed between the three categories). Figure 12 shows the confusion matrix for this model. Unsurprisingly the model is best at distinguishing people from the youngest and oldest categories, meaning most of the incorrect labeling is across adjacent age buckets. We assume that the error exists primarily for individuals who sit closer to the age border of these buckets and, therefore, still can predict relative age well.

The gender prediction model performed better than the age prediction model, with an F1 score of 0.757. However, the baseline model would have a higher score given the data is not evenly split among all three categories. Figure 13 shows the confusion matrix for this model. Interestingly, the model seems to have high accuracy when it comes to predicting males but much lower when predicting females, which might be due to the category's data distribution being skewed toward men. Nonetheless, both models have a high enough F1 score to argue the existence of implicit gender information within models.

Since we discovered that the model contains gender information, even when this information is not explicitly included, we can now examine the biases that the model exhibits towards different gender categories. Table 11 shows the random forest model's share of predicted good debtors by gender category. Interestingly enough, women have the highest percentage at 77.8%, being predicted to not default. However, the Other category percentage is much lower than the other two categories. We examined the mean/mode values for the most important features by gender category and found that some of these values were very different, which is a likely result of the small sample size of the Other category, meaning we cannot draw meaningful conclusions.

Regardless, the existence of gender information within the other features means that gender bias in the real world

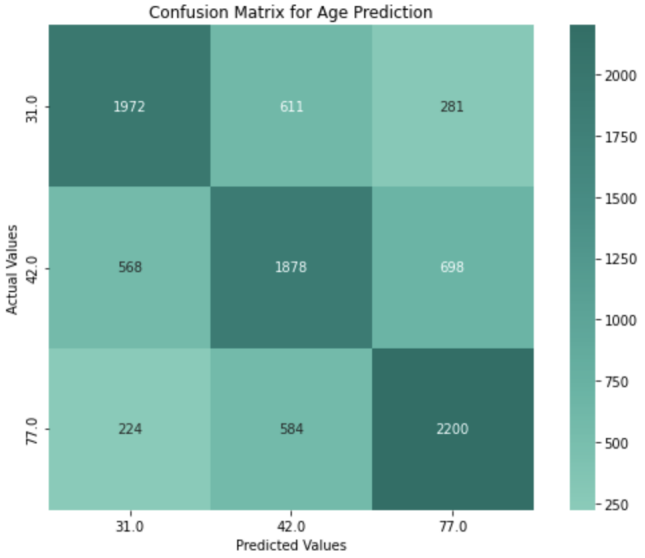

Figure 12: Confusion Matrix for Age Prediction Model. Note: Buckets are (0-31, 31-42, 42-77)

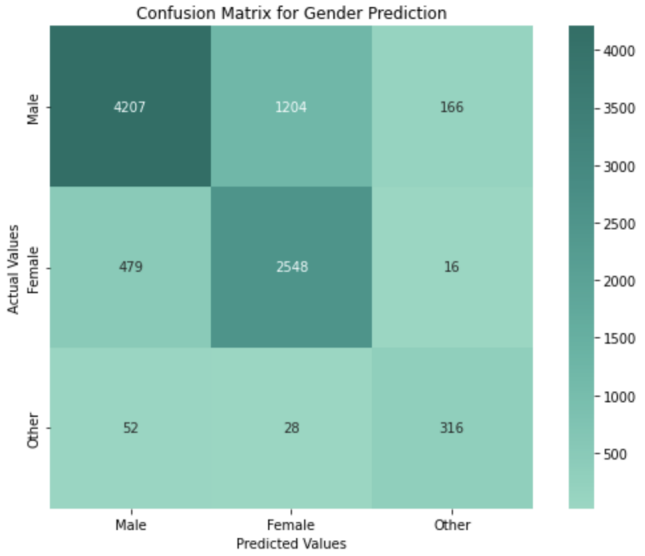

Figure 13: Confusion Matrix for Gender Prediction Model

Gender Category	% Good Debtors
Men	75.4%
Women	77.8%
Other	28.3 %

Table 11: Percentage of Individuals Within Each Gender Category Predicted as Good Debtors

will make its way into these models. Thus, a potential solution could be a third-party intervention into these models (which in itself could be biased) as well as a continual effort to reduce bias in the real world (which is both vague and extremely difficult).

The results of our best-performing models from each dataset point towards a decrease of gender/age bias over time, as this information is much more accurately predicted by our German data model than the Bondora data model (F1 scores of 0.973 and 0.847 vs. 0.672 and 0.757, respectively). This makes sense, given the large time gap between datasets and steady improvements in gender equality over the past 30 years. However, both models still display a high prediction accuracy, meaning we need to analyze these models regardless of the explicit presence of demographic features.

6 Broader Impacts

Credit Risk Data Classifiers have the potential to bring numerous benefits to the financial industry. One such advantage is that it can provide more precise risk assessments, quicker loan approvals, and reduced credit losses. This, in turn, can increase access to credit and financial services for underrepresented groups, including low-income borrowers or those without traditional credit histories. It can also increase credit access for businesses and stimulate economic growth and development. Credit Risk Data Classifiers can also enhance decision-making processes, leading to higher profits and better financial health for financial institutions.

However, if the data used to train the algorithm is biased or incomplete, it could result in unfair or discriminatory outcomes, particularly for underrepresented groups. The classifier may flag factors that disproportionately affect minority borrowers. For instance, in the German Dataset, there were no single women represented, which indicates the presence of gender bias. Moreover, a Credit Risk Classifier may not consider unique circumstances or personal qualities, potentially excluding borrowers who could have demonstrated their creditworthiness if allowed to explain their situation.

To address these concerns, future research should focus on strategies to mitigate bias and discrimination.

7 Conclusions

We successfully created accurate machine-learning models for both datasets, with our SVM model for the German Dataset achieving an F1 score of 0.829. For the Bondora dataset, our Random Forest model achieved a comparable score of 0.905.

Our analysis of both datasets found that demographic information was still contained within other features, regardless of the explicit presence of demographic features in our models. For the German model, we could predict age with an F1 score of 0.973 and gender with an F1 score of 0.847. For the Bondora model, we could predict age with an F1 score of 0.672 and gender with an F1 score of 0.757. We saw similar results to the Hassani paper, which was able to predict gender with an F1 score of 0.843 (Hassani 2021). It is essential to note that although we cannot accurately pinpoint the extent of the demographic bias, the implicit presence of gender and age within our models suggests that bias exists.

In conclusion, Credit Risk Data Classifiers have the potential to bring several benefits to the financial industry, including mitigating the risks of default and enhancing financial stability. However, addressing the risks associated with these classifiers, such as bias and discrimination, is essential to promote fairness and inclusivity.

References

Alonso, A., and Carbó, J. M. 2020. Machine learning in credit risk: measuring the dilemma between prediction and supervisory cost. Working Papers 2032, Banco de España.

FDIC. 2021. FDIC consumer compliance examination manual. Technical report, FDIC.

Hassani, B. K. 2021. Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics* 1(3):239–247.

Moscato, V.; Picariello, A.; and Sperlì, G. 2021. A benchmark of machine learning approaches for credit score prediction. *Expert Systems with Applications* 165:113986.