Responsible Data Science Final Project

Larry Li
NETID: zl2902

Kayan Shih
NETID: ks5250

1 Background

In the banking business, credit score cards are a frequent risk control approach. It predicts the likelihood of future defaults and credit card borrowings based on personal information and data provided by credit card applicants. The bank has the authority to determine whether or not to offer the applicant a credit card. Credit scores can be used to objectively measure the severity of a risk. In this case, we want to build an ADS of Credit Card Approval Prediction, which predicts if an applicant is a 'good' or 'bad' client and decides whether to issue a credit card to the applicant, and the definition of 'good' or 'bad' is not given, which should be constructed based on the developers.

However, unwanted societal biases towards unprivileged groups reflected in datasets of applicants would influence a model's behavior, it can cause severe consequences by reproducing or amplifying the bias. In this case, unprivileged groups, such as females, poor communities, or communities of color, may have limited access to credit due to the unjust system. Hence, we want to improve the accuracy of the ADS results while mitigating its bias to avoid discriminatory outcomes. However, there could be a trade-off between accuracy and fairness.

2 Input and Output

2.1 Description of Data

The data used in the ADS is from the Credit Card Dataset for Machine Learning in Kaggle's Credit Card Approval Prediction. There are two tables in the dataset, application record and credit record, which contain information about the applicants and one's credit history respectively. There are 438557 rows and 18 columns in the application record table, and 1048575 rows and 3 columns in the credit record table. These two tables could be merged by the feature ID.

Feature Name	Datatype
ID	ID
CODE_GENDER	Categorical (Binary)
FLAG_OWN_CAR	Categorical (Binary)
FLAG_OWN_REALTY	Categorical (Binary)
CNT_CHILDREN	Numerical
AMT_INCOME_TOTAL	Numerical
NAME_INCOME_TYPE	Categorical
NAME_EDUCATION_TYPE	Categorical
NAME_FAMILY_STATUS	Categorical
NAME_HOUSING_TYPE	Categorical
DAYS_BIRTH	Numerical
DAYS_EMPLOYED	Numerical
FLAG_MOBIL	Categorical (Binary)
FLAG_WORK_PHONE	Categorical (Binary)
FLAG_PHONE	Categorical (Binary)
FLAG_EMAIL	Categorical (Binary)
OCCUPATION_TYPE	Categorical
CNT_FAM_MEMBERS	Numerical

Table 1: Input Features for Applicant Dataset

Feature Name	Datatype
ID	ID
MONTHS_BALANCE	Numerical
STATUS	Categorical

Table 2: Input Features for Credit Dataset

2.2 Description of Input Features

2.2.1 Datatype

Table 1 and Table 2 show detailed information of the input features in applicant and credit datasets.

2.2.2 Missing Values

From Table 3, we can see that the Occupation Type column contains $134,203$ missing values, which is approximately 30.60% of all the values. Upon inspecting all occupation types, we observed that the occupation type does not include "Unemployed." Therefore, it was unable to know whether the missing values indicate the applicants are unemployed or were caused by the applicants' decisions not to report their occupations.

From Table 4, we can see that the credit dataset contains no missing values. Since the applicant and credit datasets are in a one-to-many relationship, we also checked if all applicants are in the credit dataset. We found that of all the 438,510 unique applicant IDs, the credit dataset only contains $36,457$ applicants' credit records. In other words, approximately 91.69% of the applicants' credit records are missing. Since it is not specified in the data dictionary, we cannot know whether the missing records indicate that the applicants have not made any loans or is it because the agency does not have access to the applicants' credit records.

Feature Name	Missing Values
ID	0
CODE_GENDER	0
FLAG_OWN_CAR	0
FLAG_OWN_REALTY	0
CNT_CHILDREN	0
AMT_INCOME_TOTAL	0
NAME_INCOME_TYPE	0
NAME_EDUCATION_TYPE	0
NAME_FAMILY_STATUS	0
NAME_HOUSING_TYPE	0
DAYS_BIRTH	0
DAYS_EMPLOYED	0
FLAG_MOBIL	0
FLAG_WORK_PHONE	0
FLAG_PHONE	0
FLAG_EMAIL	0
OCCUPATION_TYPE	134203
CNT_FAM_MEMBERS	0

Table 3: Missing Value Count for Applicant Dataset

Feature Name	Missing Values
ID	0
MONTHS_BALANCE	0
STATUS	0

Table 4: Missing Value Count for Credit Dataset

2.2.3 Value Distributions

Figure 1 shows the value distribution for all the features in the applicant dataset (excluding applicant ID). For numerical features, it is worth noting that we observed some extreme values for the number of children (`CNT_CHILDREN`), annual income (`AMT_INCOME_TOTAL`), and family size (`CNT_FAM_MEMBERS`). For the categorical feature "is there a mobile phone" (`FLAG_MOBIL`), we observed that the feature has only a single value of 1, indicating that all applicants have mobile phones. Therefore, this feature will not contribute to the ADS's prediction.

Figure 2 shows the value distribution of all the features in the credit dataset (excluding applicant ID). We observed that the top-3 most common status are loans have been paid off that month (C), 1-29 days past due (0), and no loans for the month (X).

2.2.4 Pairwise Correlations Between Features

Figure 3 shows the heatmap for pairwise attributes mutual information in the applicant dataset. We observed that most of the feature pairs are either uncorrelated or very weakly correlated. Some observed correlations are between family size and the number of children, the number of days since birth and number of days employed, marital status and family size, the number of days since birth and annual income, and the number of days employed and annual income.

2.3 Output of the System

The status of credibility, which is a binary variable with two possible values of 0 and 1, is the result of the ADS. In this case, status 0 indicates that the consumer has good credit, and status 1 indicates that the customer has bad credibility. It is worth noting here that the definition of "good" or "bad" is not given in the original dataset. In later sections, we will discuss and analyze the author's technique of defining the "good" or "bad" credibility.

3 Implementation and Validation

3.1 Data Cleaning and Pre-processing

In data pre-processing, the author defines users who are overdue for more than 60 days as target risk users and the other users as good credit users because she does a vintage analysis and concludes that generally, users at risk should be within 3%. In the status column of the credit dataset, values 2, 3, 4, and 5 suggest that the user is overdue more than 60 days, so the author assigns these values to the label ”Yes” and the other values in this column to the label ”No.” Then, the author copies the clients' ID to assign them with the label ”Yes” or ”No” into a new column and merges this column into the credit dataset by replacing ”Yes” with 1 and ”No” with 0. In this case, the author counts that there are 98.5%

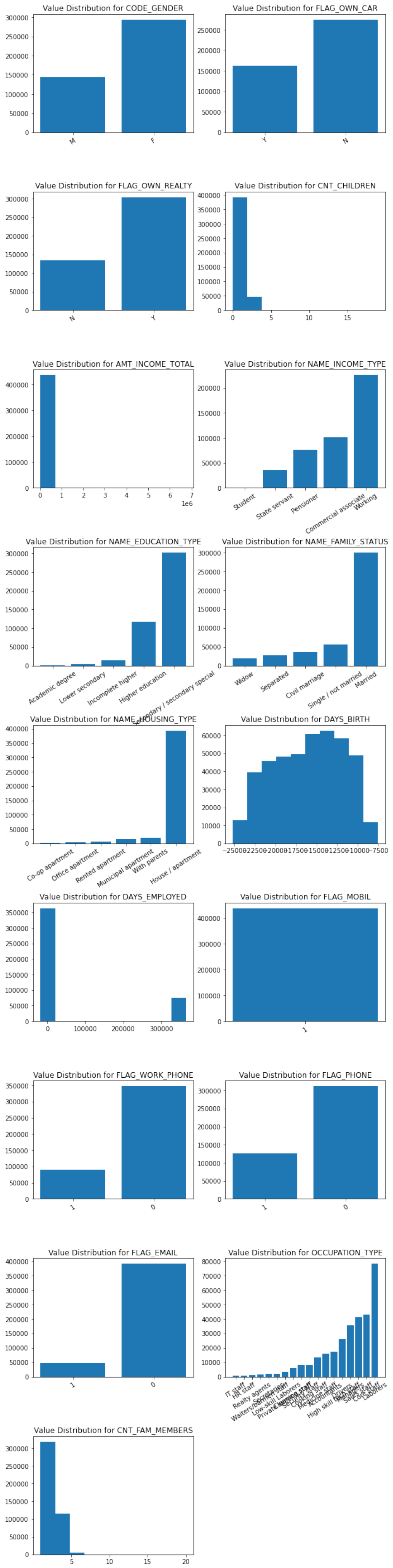

Figure 1: Value Distributions for Applicant Dataset

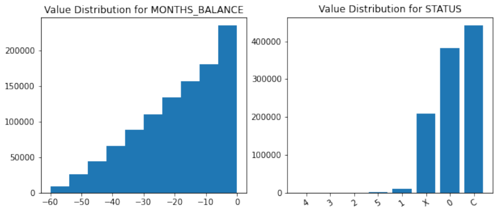

Figure 2: Value Distributions for Credit Dataset

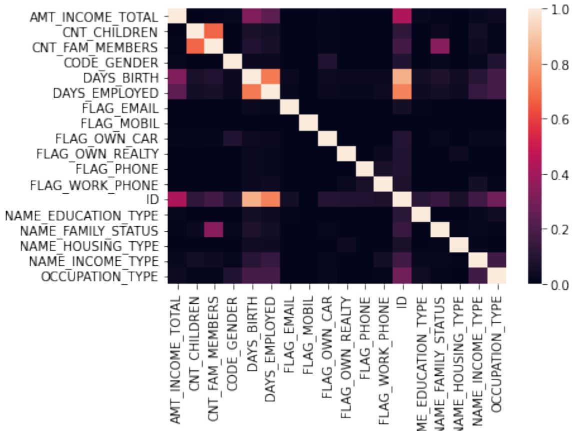

Figure 3: Heatmap Pairwise Attributes Mutual Information in Applicant Dataset

good clients and 1.45% bad clients, which shows an imbalanced sampling. Thus, to avoid over representation of the majority group in the classifier, the author uses Synthetic Minority Over-Sampling Technique (SMOTE) to overcome the sample imbalance problem. After applying SMOTE, the author claimed that the number between 1 and 0 is balanced. However, we find that, before applying SMOTE, the author converts several features, such as gender and whether they have a work phone, from integers into string values, which will generate some results into decimal numbers instead of 0 or 1. In this case, it is unsure about the author's claim that the sampling is balanced. Also, the accuracy of the ADS could be negatively influenced.

For the pre-processing of null values, the author does not use missing values imputation or other approaches to pre-process null values but simply drops all null values, resulting in a loss of information that could be important in predicting.

Besides, the author calculates the Information Value of all the features in the applicant dataset to rank variables based on their importance so that the variables can be selected based on their predictive power. Before calculating the Information Value, the author uses the Weight of Evidence (WOE) to transform a continuous independent variable into a set of bins based on the similarity of the dependent variable distribution.

3.2 Implementation of the System

After the data pre-processing, the author breaks the features into distinct columns, but we find that some features are dropped without explanation. The author only includes three binary features and drops some continuous and categorical feature groups, such as ”agemedium,” ”laborwk” et cetera. In this case, some applicants will only have 0 value in some features, which may raise concerns about fairness. For example, applicants with the occupation type of labor work will have 0 values both in ”worktype_high_tech” and ”work_type_office.”

Then, the author fits these selected features to the XGBoost model, a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library, and assigns a series of hyperparameters to the XGBClassifier. However, the author did not explain the choices of hyperparameters, so it is unsure whether these hyperparameters are optimal for predicting.

3.3 Validation of the System

The author used accuracy as the validation metric, and the XGBoost model shows an accuracy score of 87.8%. In this situation, the XGBoost model has a high accuracy score so that banks can better examine the credit status of their applicants while applying risk controls. We could say that it meets its stated goal.

4 Outcomes

Features gender and age are selected as sensitive attributes to the performance of the ADS.

4.1 Effectiveness of the ADS

4.1.1 Overall Performance

As shown in Figure 4, the test set contained a total number of 14828 applicant records. The ADS correctly identified 6181 good clients and 6770 bad clients out of those records. The ADS misclassified 644 bad clients as good clients (i.e., false negative) and 1233 good clients as bad clients (i.e., false positive).

The ADS achieved an overall accuracy of 87.34%, positive predictive value (i.e., PPV) of 84.59%, false positive rate (i.e., FPR) of 16.63%, and false negative rate (i.e., FNR) of 8.69%.

We observed that the ADS performed well overall from the metrics above, measuring accuracy. The ADS was twice more likely to misclassify good clients as bad clients than to misclassify bad clients as good clients when considering error rates. The bank, as a stakeholder, may be willing to incur a higher FPR to help lower the FNR because a higher FNR means higher risks for the bank of receiving bad debts. However, a higher FPR harms applicants because some of the qualified applicants' credit card applications are denied.

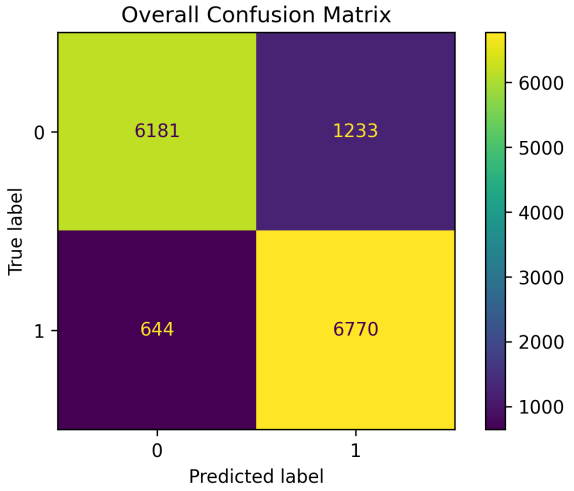

Figure 4: Overall Confusion Matrix

4.1.2 Performance Based on Gender

As shown in Figure 5, there are two gender values in the dataset: 8975 female applicants and 5853 male applicants.

The ADS's accuracy was 86.97% for female applicants and 87.90% for male applicants. The PPV was 84.60% for female applicants and 84.59% for male applicants. The ADS's performance showed minimal differences (less than 1%) in accuracy and PPV between the two subpopulations.

The ADS was more likely to make FP errors than FN for both gender groups, similar to the overall performance. For female applicants, the FPR was 15.53%, while the FNR was 10.40%. For male applicants, the FPR was 18.44%, while the FNR was 6.22%.

4.1.3 Performance Based on Age

In the pre-processing stage, the author divided age into five groups of equal length: lowest (19.95 to 29.4), low (29.4 to 38.8), medium (38.8 to 48.2), high (48.2 to 57.6), and highest (57.6 to 67.0).

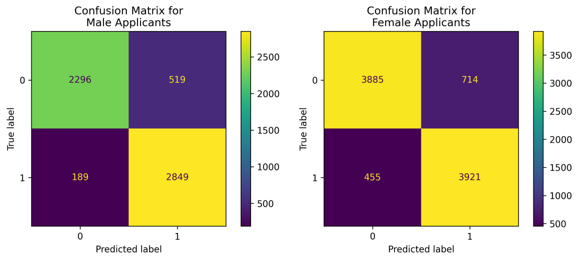

Figure 5: Confusion Matrix for Gender Subpopulations

For the simplicity of the analysis, referring to the previous German Credit Risk dataset analysis, we further grouped the applicants into two age groups: ”young” and ”middle-aged/senior.” We defined young applicants as individuals with age lower than or equal to 29.4 and middle-aged/senior applicants as individuals older than 29.4. The logic behind such a definition was that young applicants were more likely to be presumed to have poor credits than older applicants. The confusion matrix is shown in Figure 6.

We observed that both the system’s accuracy and PPV for young applicants were slightly higher than that for middle-aged/senior applicants. The ADS’s accuracy was 88.16% for young applicants and 86.71% for middle-aged/senior applicants. The PPV was 85.69% for young applicants and 83.27% for middle-aged/senior applicants.

The ADS was more likely to make FP errors than FN for both age groups, similar to the overall performance. For young applicants, the FPR was 15.77%, while the FNR was 8.01%. For middle-aged/senior applicants, the FPR was 16.80%, while the FNR was 9.50%.

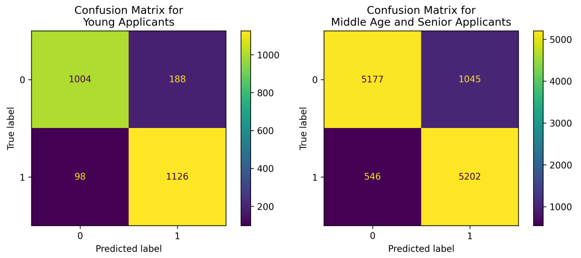

Figure 6: Confusion Matrix for Age Subpopulations

4.2 Fairness of the ADS

Judging from our understandings of society’s preexisting bias and other related analyses, we chose male applicants as the privileged class for gender and middle-aged/senior applicants as the privileged class for age.

4.2.1 Differences in Mean Outcomes

The mean difference is computed as the difference in the rate of being predicted as bad clients received by the unprivileged group to the privileged group.

The difference in mean outcomes between unprivileged (i.e., female) and privileged (i.e., male) gender groups is -0.0590, indicating that the female applicants had worse credits on average than male applicants.

The difference in mean outcomes between unprivileged (i.e., young) and privileged (i.e., middle-aged/senior) age groups is -0.0220, indicating that the young applicants had worse credits on average than middle-aged/senior applicants.

4.2.2 Disparate Impact

The disparate impact is computed as the ratio of the rate of being predicted as bad clients for the unprivileged group to that of the privileged group.

The disparate impact based on gender is 0.8975, indicating a female applicant would only get a good credit prediction 0.8975 times as often as a male applicant.

The disparate impact based on age is 0.9540, indicating a young applicant would only get a good credit prediction 0.9540 times as often as a middle-aged/senior applicant.

By calculating differences in mean outcomes and disparate impact, we were able to quantify the level of group disparities in the ADS’s prediction. The results aligned with our assumptions about the privileged and unprivileged groups for gender and age.

4.2.3 FPR/FNR Imbalance

From the FPR/FNR reported in Section 4.1, we could calculate how FPR/FNR differed between subpopulations.

The FPR imbalance between female and male applicants was -2.91%, indicating that the ADS was more likely to misclassify a qualified male applicant as a bad client. The FNR imbalance between female and male applicants was 4.18%, indicating that the ADS was more likely to misclassify a bad credit female applicant as a good client.

The FPR imbalance between young and middle-aged/senior applicants was -1.02%, indicating that the ADS was more likely to misclassify a qualified middle-aged/senior applicant as a bad client. The FNR imbalance between young and middle-aged/senior applicants was -1.49%, indicating that the ADS was also more likely to misclassify a bad credit middle-aged/senior applicant as a good client.

4.3 Additional Methods for Analyzing ADS Performance: Explaining Black-box Model

4.3.1 Feature Importance

Figure 7 shows the feature importance measured by the mean SHAP value.

The binary indicator of whether an applicant has their property had the highest feature importance because an applicant with property is presumably more financially successful and resilient than those that do not own any property.

Other features among the top-5 were if the applicant had three or more family members, if the applicant had short employment days, if the applicant did work type occupation, and if the applicant had a work phone.

Our primary features of interest - age and gender - ranked eighth and ninth among all features, indicating that they do not have extreme predictive power.

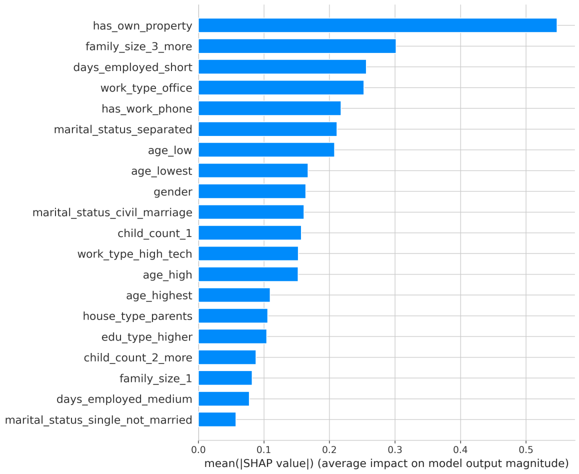

Figure 7: Feature Importance

4.3.2 Explaining Predicted Label for Two Applicants

Figure 8 shows features, each contributing to pushing the model output from the base value, for the applicant of index 0. The true label indicated that the applicant was a good client, and the model correctly predicted that. The two features that contributed the most to pushing the output label toward good were the applicant had an office type work, and the applicant had their own property. The top two features that pushed the output label toward bad were that the applicant's total number of family members was less than three, and the applicant did not have a work phone. The base value was -0.158, and the model's output value was -0.286. Therefore, the model's output label was good.

Figure 9 explains the model's output for the applicant of index 14809. The true label indicated that the applicant was a good client. However, the model misclassified the applicant as a bad client. The top two features that pushed the output label toward bad were that the applicant was single, and the applicant's total number of family members was less than three. The top two features that pushed the output label toward good were that the applicant had their own property and the applicant was young. The base value was -0.158, and the model's output value was 1.704. Therefore, the model's output label was bad.

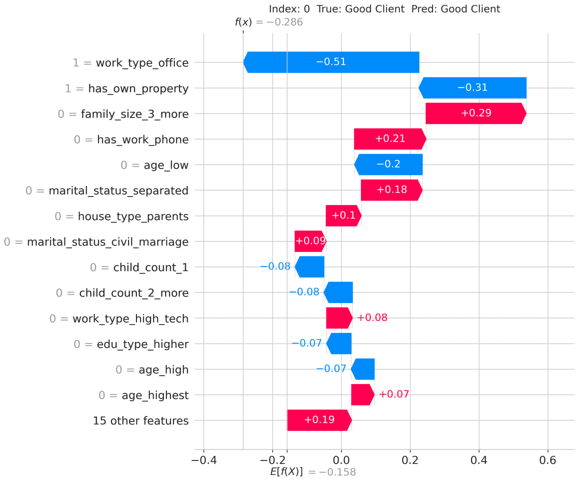

Figure 8: SHAP Explanation for Applicant 0

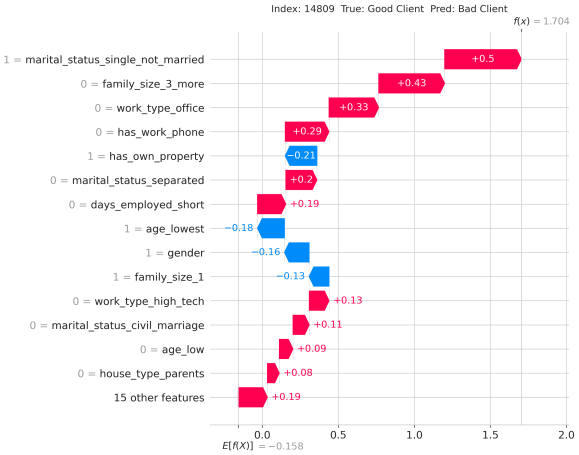

Figure 9: SHAP Explanation for Applicant 14809

4.3.3 Features Contributing to the Misclassification of Applicants

In order to improve the accuracy of the system and lower error rates, it would be helpful to know what features were contributing to the misclassification of applicants. To investigate this, we defined features that helped misclassify bad clients as good clients as those with negative SHAP values for the misclassified applicants. Similarly, we defined features that helped misclassify good clients as bad clients as those with positive SHAP values for the misclassified applicants.

Figure 10 shows the absolute SHAP values of all the features that contributed to the misclassification of bad clients as good clients. The top three features are whether the applicant had their own property, whether the applicant had an office type work, and whether the applicant had three or more family members.

Figure 11 shows the SHAP values of all the features that contributed to the misclassification of good clients as bad clients. The top three features are whether the applicant had three or more family members, whether the applicant had an office type work, and whether the applicant's marital status was separated.

From the two figures, we observed an overlap of two features among the top three that were contributing to the misclassification of applicants in opposite directions, indicating that the two features, whether the applicant had an office type work and whether the applicant's marital status was separated, could be introducing noise to the model's prediction.

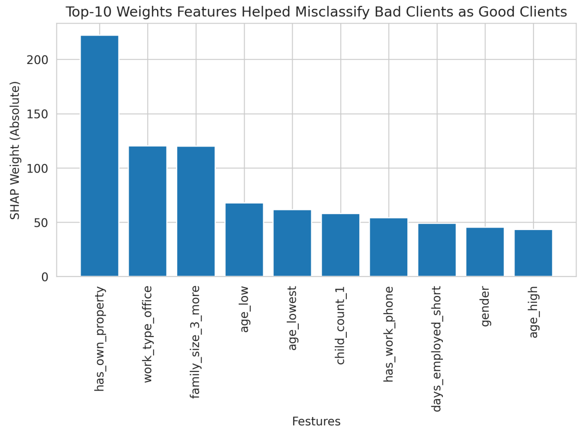

Figure 10: Features Contributed to Misclassify Bad Clients as Good Clients

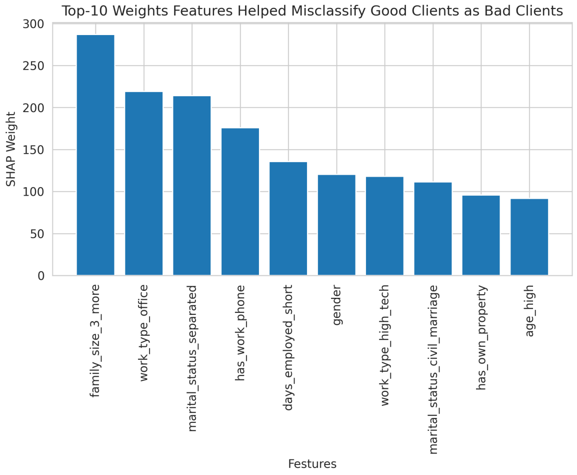

Figure 11: Features Contributed to Misclassify Good Clients as Bad Clients

5 Summary

5.1 Data Appropriateness

The dataset contained many features that comprehensively described the credit card applicants in multiple dimensions, allowing the system to predict the applicants' credit more accurately. It also included sensitive features like gender and age, which are not common in similar datasets. Including sensitive features allowed us to audit the system to see if there were fairness issues.

However, the dataset is highly imbalanced, which requires the author to apply resampling tools. As mentioned in Section 3.1, in this case, the resampling process introduced errors and noises into the generated dataset. It was unknown how well the original distributions and correlations in the original dataset were preserved.

Overall speaking, the dataset was appropriate for this ADS.

5.2 Evaluation of Implementation

The ADS performed well, measuring accuracy. It achieved an overall accuracy of 87.34%. High accuracy could benefit the bank as they can trust the system's output to be accurate. It could also benefit the applicants as an accurate system could reduce human bias when evaluating the applications.

The ADS also had relatively high PPV (84.59%), which could benefit the bank as high PPV reduced the risk of receiving bad debts.

The ADS had an overall FPR of 16.63% and FNR of 8.69%. The bank may be willing to incur a higher FPR to help lower the FNR because a higher FNR means higher risks for the bank of receiving bad debts. However, a higher FPR harms applicants because some of the qualified applicants' credit card applications are denied.

The ADS also had reasonably well performance, measuring fairness. We observed minimal differences in mean outcomes for both gender and age subpopulations. When examining FPR/FNR imbalance, we observed that the maximum difference in FPR/FNR between subpopulations was around 4%, with most differences around 1%. This level of difference was justifiable due to the random sampling process. The fairness issue that we found worth noting was that we observed disparate impact based on gender: a female applicant would only get a good credit prediction 0.8975 times as often as a male applicant; and disparate impact based on age: a young applicant would only get a good credit prediction 0.9540 times as often as a middle-aged/senior applicant.

5.3 Evaluation of Deployment Appropriateness

Based on the analysis of the system, we would argue that it would be appropriate to deploy the ADS in the industry under the condition that sensitive and protected features (e.g., gender and sex) should be removed from the model's training process. The logic behind such a decision was that the system had high accuracy/PPV while age and gender did not have high predictive power on the outcome, and we found no significant fairness issues upon preliminary inspection. However, since the system is trained and evaluated using resampled data, we would argue that it would be necessary to conduct regular audits on the system's performance both in accuracy and fairness to see if there exist issues when used in the industry and to make adjustments to the system accordingly.

5.4 Improving Current Design

For the data collection stage, the data collector did not reveal the source of the dataset or its license. They also did not reveal whether the applicants were informed and agreed that application information was released and used for analysis. The data collector should include information on privacy-preserving, specifically, if any methods were taken to protect the applicants' privacy, and provide proof that the degree of protection was adequate (e.g., beyond simple de-identification).

For the data preprocessing stage, the author should provide justifications when creating new categorical features (i.e., age groups and employment duration groups). The author should also be more cautious when dealing with missing values rather than dropping them directly. Similar suggestions also apply to the feature selection process: the author should provide more reasonings and analysis rather than drop features solely based on information value.

For the analysis stage, the author should provide information on how the hyperparameters were chosen (e.g., is it the result of some hyperparameters tuning tools). Providing such information would make it more convincing that the model was performing at its peak.