Analysis of a Home Credit Default Risk ADS

Sangwon Baek and Simone Rittenhouse

Center for Data Science, New York University

DS-UA 202: Responsible Data Science

Prof. George Wood

May 9, 2022

1. Background

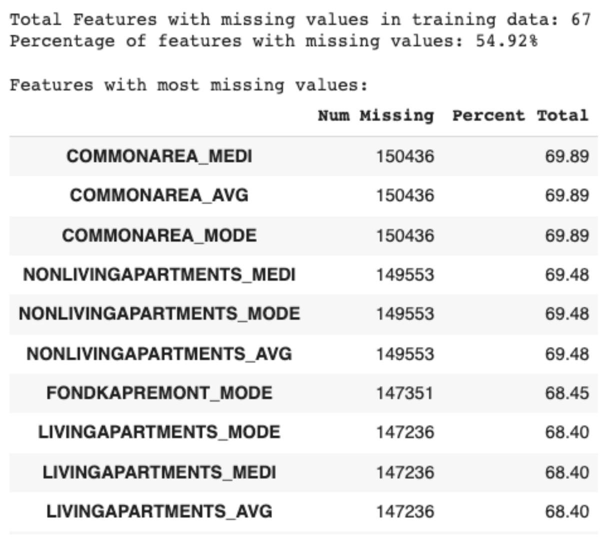
In this project, we explored an automated decision system (ADS) created by Will Koehrsen in response to Home Credit's loan default risk Kaggle competition. The competition was proposed by Home Credit to help make the loan application and repayment process more inclusive. Specifically, this ADS seeks to predict whether or not individuals will be able to repay a loan (i.e. the likelihood of timely repayment versus defaulting on the loan) using historical loan application data and a variety of other features like total income, age, and gender. These predictions are made in the hope that individuals with no credit history will be able to receive loans - providing equal opportunity to individuals new to banking. In addition to predicting whether or not an individual with little to no prior history will be able to repay their loan, the ADS was created as a training tool for those seeking to engage in data science projects. It therefore has a second goal of being as simple and easy to understand as possible.

There is an inherent trade-off between the system's dual goals of simplicity and accuracy in predicting whether or not an individual will repay a loan. The competition solution used two models, one being logistic regression and the other being a random forest model. Koehrsen mentions in his solution that the logistic regression model is a simpler model acting as a baseline against the random forest model. However, depending on the data used, a simpler linear model may not optimize accuracy. Additionally, in an effort to keep the solution as simple as possible, several of the available competition datasets were not used by the ADS. Not using all available datasets may have also lowered the accuracy of the model. Despite these points, our project will primarily focus on interpreting the random forest model, which achieved slightly higher performance than the logistic regression model.

2.1 Input

The competition provided seven different sources of data. Out of these seven, the ADS used data from a single source - the primary training/test loan application data. However, because the original test dataset did not include outcome labels, our analysis required splitting the training dataset into labeled training and test sets. The model was then trained on the new training data - which was 70% of the original training set. An additional file was included to describe each feature across all competition datasets. This information includes which dataset each feature is found in, a description of what the feature is measuring, and other 'special' information such as whether or not the feature has been normalized and the units in which it was recorded. Unfortunately, no other metadata was included, making it difficult to determine how or when this data was collected. The absence of metadata is potentially problematic because it is difficult to verify the validity of the dataset. Thus, such an exclusion prevents us from having a deeper understanding of the data.

The main features of the dataset include demographics such as gender, age, education, income, and family status. In addition, detailed credit history information is used - for example, the credit amount of the loan, loan annuity, housing situation, car owning status, type of work, etc. In total, the training dataset had 122 features: 106 were numeric and 16 were categorical. As shown in Figure 1a, 54.92% of the 122 features contained missing values. The input with the most missing values was COMMONAREA—referring to the common area of an individual's residential building—with 150,436 missing values.



```
Total Features with missing values in training data: 67
Percentage of features with missing values: 54.92%

Features with most missing values:
```

	Num Missing	Percent Total
COMMONAREA_MEDI	150436	69.89
COMMONAREA_AVG	150436	69.89
COMMONAREA_MODE	150436	69.89
NONLIVINGAPARTMENTS_MEDI	149553	69.48
NONLIVINGAPARTMENTS_MODE	149553	69.48
NONLIVINGAPARTMENTS_AVG	149553	69.48
FONDKAPREMONT_MODE	147351	68.45
LIVINGAPARTMENTS_MODE	147236	68.40
LIVINGAPARTMENTS_MEDI	147236	68.40
LIVINGAPARTMENTS_AVG	147236	68.40

Figure 1a. The top ten features with missing values

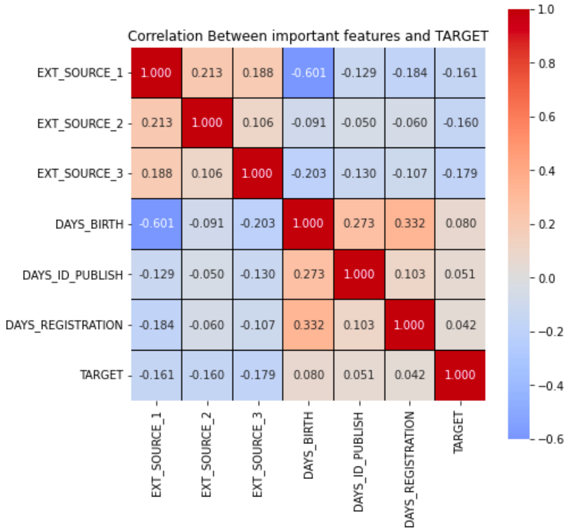


Figure 1b. Correlations between target and features

Due to the number of input features, it would be infeasible to show the distribution of each individual independent variable. Koehrsen's solution included feature importance for the ADS, showing some of the inputs that contributed heavily to the system's prediction. This was used to select a subset of features for exploratory analysis. The following heatmap shows the pairwise correlations in the training data between some of the most important features and the target (Figure 1b). Here, the three EXT_SOURCE features are negatively correlated with the outcome variable and have the strongest overall correlation. However, the description for the exit source variables is 'Normalized score from external data source' ("HomeCredit_columns_description.csv," 2018), making it unclear what these input features represent and why they would be strong predictors of loan repayment.

To explore the exit source features further, their frequency and probability density distributions were plotted for both those who repaid their loan and those who did not. Code from the ADS solution was used to plot the density of the variables (Koehrsen, 2018).

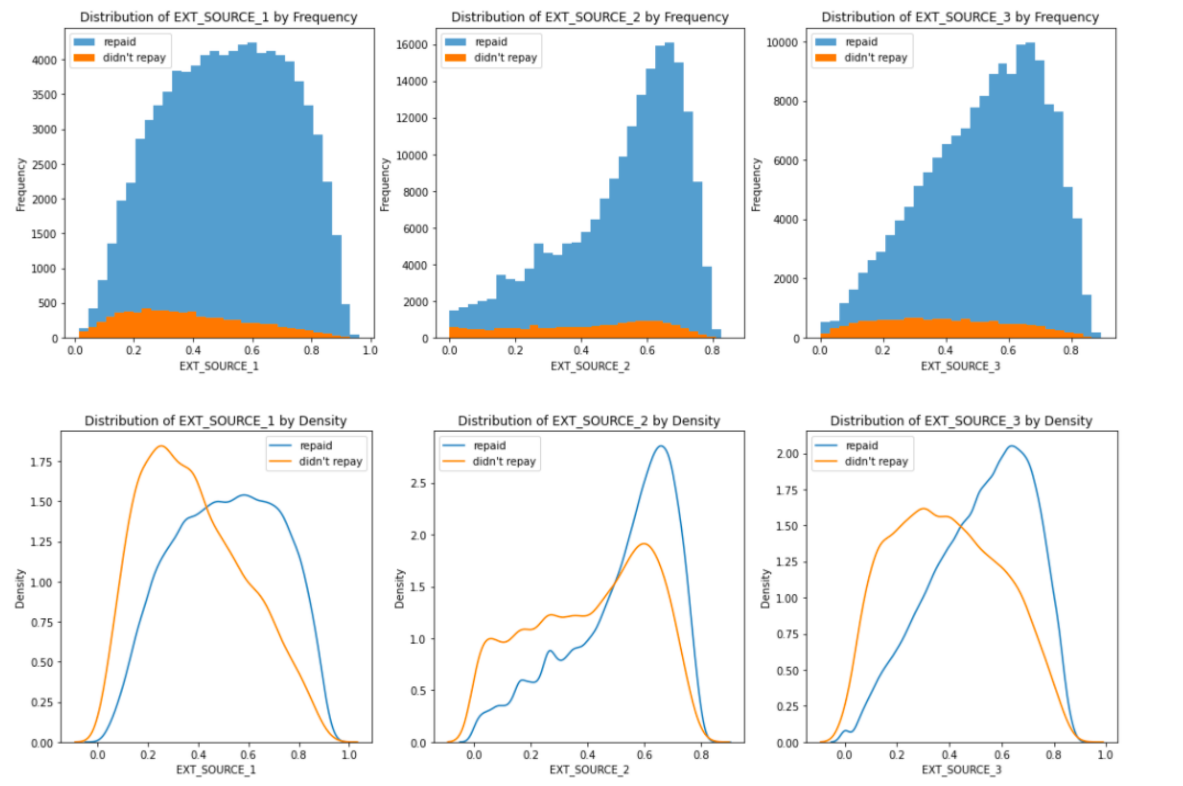


Figure 2. The distribution of three exit source features plotted by frequency (top) and density (bottom).

Observing the plots of the three exit source features, all cases reveal that more individuals did not repay their loans on time at lower values EXT_SOURCE (Figure 2). At higher values, more individuals were able to repay their loan on time.

The same exploratory analysis was repeated for the variable DAYS_BIRTH to examine how age might affect the ADS' predictions (Figure 3). The input feature DAYS_BIRTH measures an individual's age in days - counting backwards from the day the loan was taken. Therefore, the plots below show that at higher values of DAYS_BIRTH individuals were more likely to not repay their loans on time. This means that younger people were more likely to miss payment than older people.

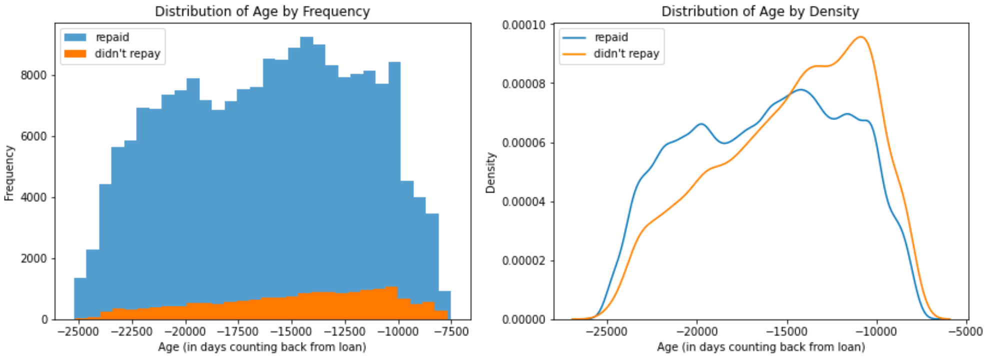


Figure 3. The distribution of the input feature DAYS_BIRTH plotted by frequency (left) and density (right).

Lastly, the value distribution for gender was plotted—a sensitive demographic attribute that could lead to biased predictions for or against a certain group. In both the training and test data, counts for women were almost twice as large as for men (Figure 4). Since men are underrepresented in the data, the ADS may make less accurate predictions for them due to a smaller sample size. This possibility will be further explored in our Outcomes section. Additionally, 7.03% of women and 10.23% of men in the training data failed to repay their loan on time. Therefore, women were slightly more likely to repay their loan - although the difference is minimal.

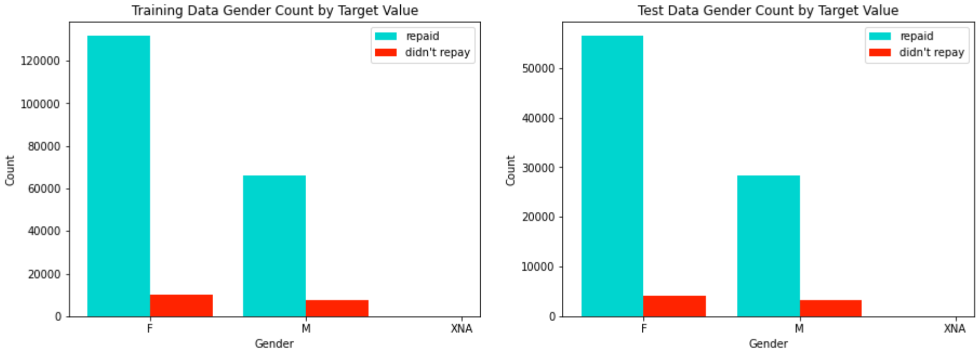


Figure 4. Gender count in the training data (left) and test data (right).

2.2 Output

The output of the ADS is a data frame of predictions for the test dataset with each test unit's ID and the prediction of whether or not they will be able to repay their loans. The predictions themselves are values between 0 and 1 representing the probability of not being able to repay a loan on time, with 0 being able to repay and 1 being unable to repay with certainty.

3. Implementation and Validation

Before building the logistic regression and random forest models, the training and test data were pre-processed. The first step in this process was encoding the sixteen categorical variables. Features with two categories were re-coded using label encoding, and features with more than two categories were re-coded using one-hot encoding. Because certain variables in the training set had categories not represented in the test set, the training and test datasets were then realigned to have the same number of features by dropping columns from the training set not found in the test data. Additional data cleaning was performed to deal with extreme values. Specifically, the maximum value of 'DAYS_EMPLOYED' was found to be 365,423 days (Koehrsen, 2018), which is impossible. These observations were replaced with nulls. Because 'DAYS_BIRTH' was measured counting backwards from the day a loan was taken, it was also reassigned as its absolute value to measure the positive age of an individual. Lastly, missing values were imputed using the feature's median, and the dataset was normalized by rescaling each feature in the range 0 and 1.

The random forest model itself was built using the RandomForestClassifier class from sklearn.ensemble. The hyperparameter n_estimators was set to 100, meaning that one-hundred individual decision trees were used in the model. Random forest models fit these individual trees to samples from the training data and then average the predictions of these trees to make a classification. This helps create a more accurate prediction than any individual decision tree alone, as well as helping to control for overfitting. This model was fit to the processed training data and the final probabilities of not repaying a loan were found using the test data and the model's predict_proba method.

To validate the performance of the ADS, a data frame containing applicant IDs and their predicted probability of loan repayment was submitted to the Kaggle competition. For this competition, Kaggle evaluated submissions using the area under the ROC curve ("Home credit default risk", 2018). Koehrsen stated that the original model, which used the entire original training dataset, should achieve an AUC score of 0.678. However, the model trained on our subsetted training and test set achieved an AUC score of 0.71 and an accuracy of 92.05%. This is slightly higher than Koehersen's predicted AUC as well as the AUC for his logistic regression model - which was 0.68 for the new training and test data. An AUC score of 0.71 shows relatively good performance, although it could be improved.

4.1 Outcomes: Performance Across Subpopulations

Because there was significant class imbalance between those who repaid and those who did not, both models predicted more individuals as being able to repay than not. The logistic regression model predicted that all individuals would repay the loan, and the random forest model predicted that only thirteen individuals would be unable to repay. Of these thirteen individuals, seven were male and six were female. This shows that a higher rate of men are predicted to not repay compared to women. Additionally, the average age in those predicted not to repay is much lower than the average age of the test data (0.194 compared to 0.481) - suggesting that younger individuals are more likely to be classified as unable to repay. The same analysis was considered for total income, but the average income for those predicted not to repay was essentially equal to the average of the entire test set (0.00118 compared to 0.00122).

To assess the performance of the random forest model, accuracy and AUC was computed across values of gender, income, and age. These attributes were chosen due to their particularly sensitive nature, as a system predicting loan repayment must not discriminate on the basis of age or gender. As shown in Figure 5, accuracy and AUC did not drastically differ for these subgroups. Those with higher income had a slightly higher accuracy but a lower AUC (accuracy = 92.5% compared to 91.6%; AUC = 0.705 compared to 0.714). The insignificant differences between accuracy and AUC for high and low income suggest the model performs equally well regardless of income level.

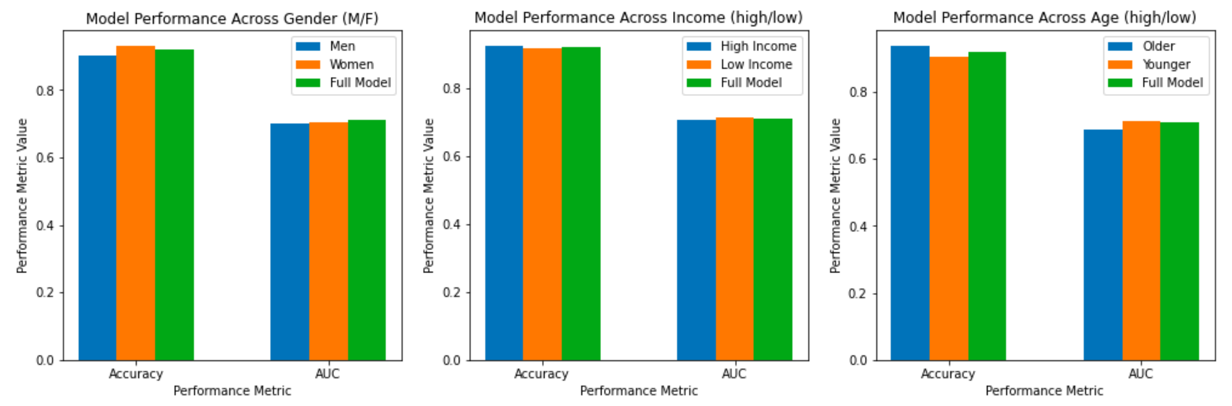


Figure 5. Accuracy and AUC across gender, income, and age groups.

Similarly, the model performed with higher accuracy but lower AUC for those with an age above the median (accuracy = 93.7% compared to 90.4%; AUC = 0.686 compared to 0.712). The difference between these two age groups is greater than the subgroups of income, meaning an individual's age impacted the model's performance more than their income. AUC may be higher for those who are younger due to the model predicting that more of them are unable to repay loans - which would decrease the false negative rate of this group relative to older individuals.

Lastly, men had both a lower AUC and accuracy than women - although the difference in AUC between the two groups is minimal. The model was 93.1% accurate for women and 90.1% accurate for men (AUC = 0.705 for women, 0.701 for men). The lower performance for men may be explained by the scarcity of men in the training and test data; however, this is problematic as men would be unfairly adversely impacted.

4.2 Outcomes: Fairness and Diversity Measures

Due to the class imbalance in the outcome variable, it is insufficient to assess the model on accuracy alone. The logistic regression model, for example, was able to achieve a high accuracy of 92% by predicting that all individuals would repay their loans. It is therefore vital to consider the error rates of this model when determining its fairness in order to see if a certain group was more likely to be incorrectly offered a loan despite not repaying or denied a loan despite being able to repay. Our chosen metrics include the false negative and positive rates for privileged and unprivileged groups, as well as their differences and the difference in overall error rate. In addition to assessing the model's errors across groups, it is also important that the probabilities of the positive outcome are fair. This is necessary because, for this system, a positive classification would predict that an individual could not repay their loan - potentially denying that individual of a financial opportunity. If the model systematically predicts this more often for one group over another, the model would be unfair. Mean difference and disparate impact were used to assess the prediction probabilities across groups. Mean difference is the difference between unprivileged and privileged groups in the probability of being classified as unable to repay, while disparate impact is the ratio of this quantity.

It was again important to see if the model was fair with respect to various protected attributes. These were gender and age. For gender, men were set as the unprivileged class since the model predicted they would be unable to repay at higher rates. Men had a slightly lower false positive rate than women (3.52e-5 vs. 3.54e-5) with a difference of -1.58e-7. Men also had a slightly lower false negative rate than women (0.998 vs. 0.999) with a difference of -9.65e-4. Importantly, women's higher false negative rate implies that the model more often incorrectly classifies women as able to repay their loan - which shows bias towards women. The overall error rate difference was 0.0299, showing that the error rates between men and women were very similar with a slight bias against men. The disparate impact was 2.25 and the mean difference was 1.23e-4. The large disparate impact shows that men were over twice as likely to be predicted as being unable to repay a loan, although the small mean difference implies that neither men nor women were very likely to be predicted as being unable to repay.

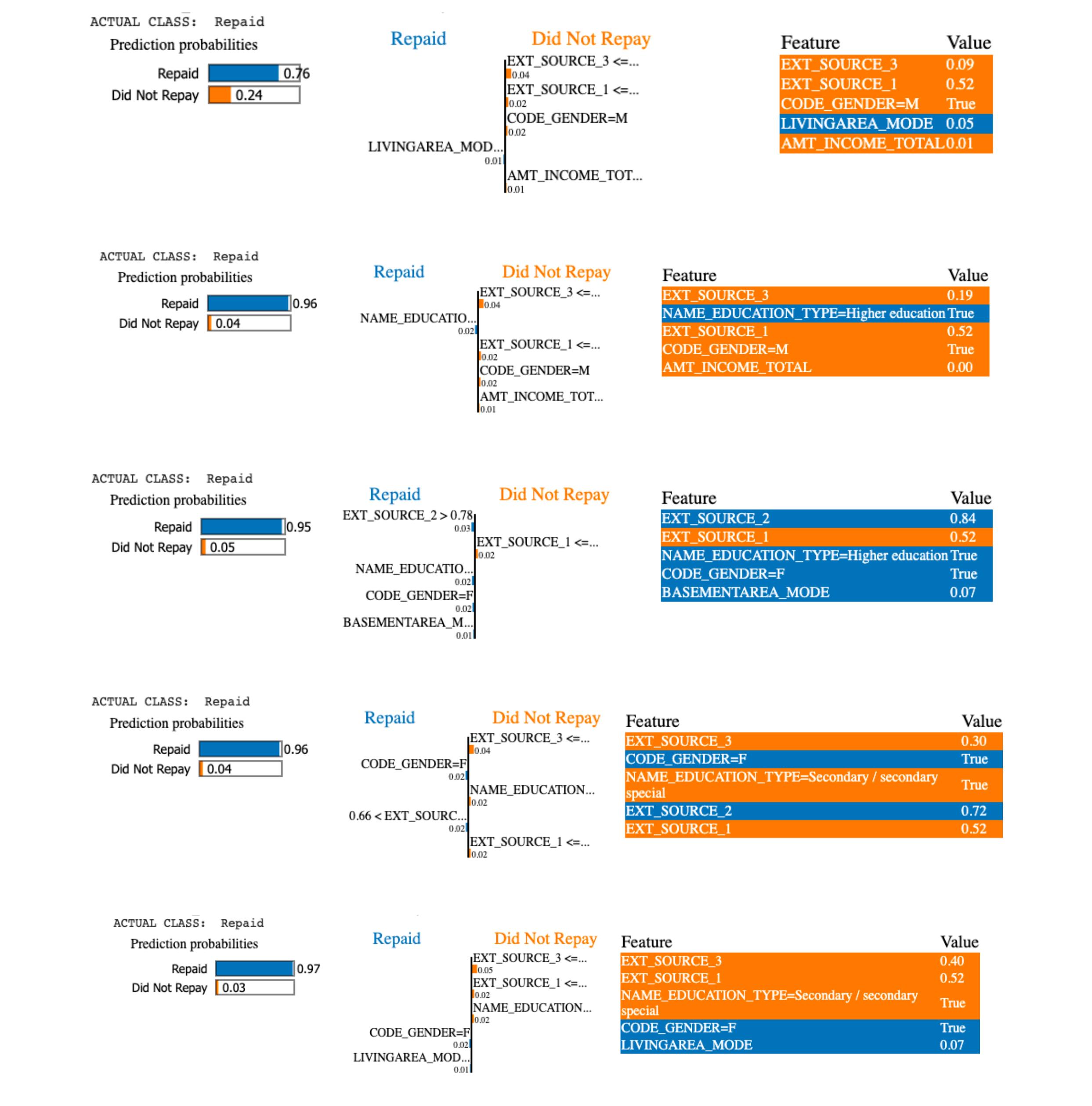
For age, those under the median were set as the unprivileged group since the average age of those predicted to default on a loan was much lower than the overall average. The false positive rate for the privileged, older group was 0, meaning that the model had no cases in which older individuals were incorrectly predicted to default on a loan. However, the false positive rate for the younger group was 7.20e-5, showing some unfair misclassification of younger individuals. Additionally, the younger group had a lower false negative rate (0.99797 vs. 0.99965), which means that they were less likely to benefit from being incorrectly classified as able to repay a loan. There was also a slightly higher overall error rate difference for age than gender, with a value of 0.033. These errors show a bias against the younger group. Similarly, the disparate impact was 12.0 and the mean difference was 2.38e-4. This shows that the younger group was much more likely to be classified as being unable to repay a loan, though once again the small mean difference suggests that neither group had a high likelihood of being predicted as defaulting.

4.3 Outcomes: Classification Explanations

To further assess the random forest model, SP-LIME was used to generate explanations for several relevant examples. This was done due to the inherent lack of interpretability in a random forest model. Because this model uses one-hundred individual trees, it is very difficult to examine how exactly the model has made a classification (i.e. what features were used and to what extent). When considering the appropriateness of a model, it is vital to take into account the transparency of the system and how easily its output will be understood by users. SP-LIME offers some clarity for black box models such as this by finding features that are important in explaining as many predictions of the model as possible and then choosing individual examples where these features are present.

In order to use LIME submodular pick, the scaled, imputed, and one hot encoded training and test data was converted back to un-encoded form. Categorical variables then received label encodings, which were stored and passed into LIME's tabular explainer as categorical_names. An explainer was created using the training data, after which a SubmodularPick object was instantiated using the test data. This object explained ten randomly chosen predictions from the test data and then selected five with globally important features. The indices of these five relevant explanations were then explained for the test dataset (Figure 6).

As seen previously in the feature importances from the random forest model, the three exit sources appeared as relevant features across all five explanations, with EXT_SOURCE 1 and 3 pushing the model towards a prediction of 'default' and EXT_SOURCE 2 pushing the model towards 'repay.' One unexpected yet important result is that gender was also used as an important feature in all five predictions. A gender of female pushed the model towards a prediction of repayment, while a gender of male pushed the model towards a prediction of defaulting. The model therefore has disparate impact, as it heavily considers an individual's gender when predicting default risk - which is very problematic and discriminatory.



Feature	Value
EXT_SOURCE_3	0.09
EXT_SOURCE_1	0.52
CODE_GENDER=M	True
LIVINGAREA_MODE	0.05
AMT_INCOME_TOTAL	0.01

Feature	Value
EXT_SOURCE_3	0.19
NAME_EDUCATION_TYPE=Higher education	True
EXT_SOURCE_1	0.52
CODE_GENDER=M	True
AMT_INCOME_TOTAL	0.00

Feature	Value
EXT_SOURCE_2	0.84
EXT_SOURCE_1	0.52
NAME_EDUCATION_TYPE=Higher education	True
CODE_GENDER=F	True
BASEMENTAREA_MODE	0.07

Feature	Value
EXT_SOURCE_3	0.30
CODE_GENDER=F	True
NAME_EDUCATION_TYPE=Secondary / secondary special	True
EXT_SOURCE_2	0.72
EXT_SOURCE_1	0.52

Feature	Value
EXT_SOURCE_3	0.40
EXT_SOURCE_1	0.52
NAME_EDUCATION_TYPE=Secondary / secondary special	True
CODE_GENDER=F	True
LIVINGAREA_MODE	0.07

Figure 6. Five relevant explanations generated by SP-LIME, with their five most explanatory features.

5. Summary

After reviewing the implementation and performance of this model, it is clear that the ADS has several shortcomings. Firstly, the lack of metadata makes it difficult to determine the appropriateness of the dataset for such a model. For example, the three ‘exit source’ variables are only described as coming from an external source, but it remains unclear as to what specific information these features represent. This is especially problematic as our analysis revealed that these variables are globally important and contribute largely to the model's predictions. Additionally, it is not appropriate for such a model to contain demographic features such as age or gender. These are protected attributes, and including these directly into the input leaves the model at risk of making loan repayment decisions in a discriminatory manner. It also may pose privacy concerns if it allows individuals in the dataset to be identified by adversaries - through linkage attacks, for example. Therefore, the dataset itself has several issues of fairness and transparency that make it inappropriate to use in such a system.

The model itself also had some issues of fairness and performance. Though the accuracy of both the logistic regression and random forest models was high (> 0.9), they achieved an AUC score of about 0.7 - which is not as high as it could have been. This is mainly because both models predicted the positive class at extremely low rates - potentially because neither model weighed the classes to correct for imbalance. The accuracy of the model was also lower for men and younger people. The model's fairness metrics also suggest that the model is biased against these two groups. Across all groups, the false negative rate was extremely high. The stakeholders that would find this measure appropriate would be those applying for loans - as it means they would more often be predicted as able to repay despite not being able to. This would mean they would be more likely to receive a loan. The false negative rate was slightly higher for women and older people, benefiting them more. Additionally, financial institutions would likely be more concerned with the false positive rate, as they would be more concerned with minimizing the false negative rate - even if it means denying capable individuals a loan. Younger people and women had a higher false positive rate, which disadvantages them more than other individuals. The disparate impact was also very high for both age and gender - suggesting that men and younger individuals were more likely to be classified as defaulting. The stakeholders that would be concerned with this metric would be the users in disadvantaged groups, as well as financial institutions and those creating the model since they would not want to refuse loans to certain demographics in a discriminatory manner.

Because of the issues of both the model and its input data, we would not be comfortable deploying this ADS in the public or financial sector. The model would have a disparate impact by systematically denying loans more often to men and younger individuals. This would not be legal were it to be used in industry, as it discriminates on the basis of age and sex. The model's performance is also lacking, as it almost always predicts that users will repay their loans. This would not be satisfactory for the financial institutions deploying the system since they would be at risk of offering loans to those who will default. Additionally, this solution has not specified how the data was collected or if any randomized privacy measures were taken - meaning it may also be inadequate in protecting users' potentially sensitive financial information. For these reasons, this ADS should not be used to make decisions on who should receive a loan.

Improvements could be made at several points in the implementation of this ADS. Firstly, potentially discriminatory features like age and gender should be removed from the training data - as the model should not make decisions on the basis of these attributes. Additionally, it is important that more thorough metadata is provided to explain exactly what information the model is taking into account and how it is collected. Randomized privacy measures should be introduced to protect users' information. The model itself could likely improve performance by correcting for class imbalance and reweighing the two classes. This would be more helpful to financial institutions as it would lower the false negative rate. The data was also processed by imputing the median for all null values. The model's fairness may be improved by handling null values on a feature-by-feature basis and considering if null values are uniformly distributed or differ systematically between subgroups and how this may affect predictions. Additionally, to mitigate the biases of this model, a processing technique - equalized odds, for example - could be used to ensure that subgroups of protected attributes have an equal likelihood of being denied a loan and minimize disparate impact. Lastly, the model was currently analyzed by Kaggle using AUC alone. Because this model has the potential to deprive individuals of important financial opportunities, it is vital that other metrics are taken into account to assess performance, such as the error rates across subgroups.

In conclusion, this ADS exhibits bias against certain groups in the protected attributes age and gender, making it problematic. An analysis with SP-LIME showed that gender was a globally important feature, while the model's built-in feature importances method showed that age was also an important factor. The heavy use of these independent variables leave the model prone to making discriminatory predictions. Though the solution achieved its stated goal of simplicity, it failed at its primary goal of creating a robust and fair predictive model. We do not recommend its use.

References

Baek, S., & Rittenhouse, S. (2022, March 20). *RDS: Final Project Code.ipynb*. Google Colab. Retrieved April 10, 2022, from https://colab.research.google.com/drive/RDS_sharing

HomeCredit_columns_description.csv. Kaggle. (2018, May 17). Retrieved April 10, 2022, from https://www.kaggle.com/competitions/home-credit-default-risk/data

Home credit default risk. Kaggle. (2018, May 17). Retrieved April 10, 2022, from https://www.kaggle.com/c/home-credit-default-risk/overview

Koehrsen, W. (2018, August 25). *Start Here: A Gentle Introduction*. Kaggle. Retrieved April 10, 2022, from https://www.kaggle.com/code/willkoehrsen/start-here-a-gentle-introduction#Introduction:-Home-Credit-Default-Risk-Competition