DS-UA 202, Responsible Data Science: Course Project
A Nutritional Label for an ADS Predicting Loan Default Risk

Erin Choi, Vicky Lin

May 10, 2021

1 Background: General Information, Purpose and Stated Goals

The ADS we are analyzing is one that predicts the loan repayment abilities of customers of Home Credit, an international consumer finance provider whose goal is to broaden financial inclusion for underbanked and low-income populations by focusing on lending to people with little or no credit history. Home Credit uses a variety of data and machine learning methods to predict clients' repayment abilities in order to ensure that the unbanked population has a positive borrowing experience. Home Credit and other financial institutions could use this ADS and its results to increase access to credit for underbanked and low-income populations, so NYC or other municipalities could refer to our nutritional label to explain how this ADS operates to such institutions.

The ADS was found on the Kaggle competition page "Home Credit Default Risk," and the implementation of this ADS that we are using is the solution "Start Here: A Gentle Introduction" by Kaggle user Will Koehrsen (willkoehrsen).

2 Input and Output

2.1 Description of Data

The data we are using is available on the Kaggle competition page. It was provided by Home Credit. There is no information on the page about how the data was collected, so we are assuming it is a sample of actual individuals who applied for loans from Home Credit with some identifying information removed. While there are eight input datasets available, the solution that we are using primarily makes use of two of them: application_train.csv, which is the main training data, and application_test.csv, the testing data. These two files contain static data for all applications, with each row representing one loan case in the data sample.

2.2 Input Features

The 122 features in the training set include clients' income, clients' highest level of education, credit amount of the loan, and information about clients' previous applications, among many others. The testing set includes all the features in the training set excluding the target variable, which the ADS predicts for each applicant. Sensitive attributes in the data include clients' gender and age in days at the time of the application. Based on data exploration code provided by the ADS author, there are 67 features out of the 122 in the training set that have missing values. The number and percentage of missing values in a particular column are displayed in the data frame missing_values.

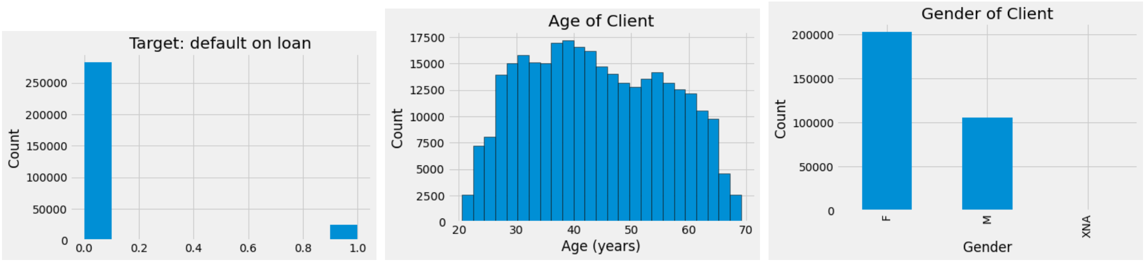


Figure 1: Training set: distributions of target variable, age, and gender

The target variable in the training set represents the clients' repayment abilities; a value of "1" indicates that a client had payment difficulties, and "0" represents all other cases. In the training set, there are many more applicants who successfully repaid their loans than those who failed to.

The sensitive attribute of age, represented in the column DAYS_BIRTH, is the negative of the number of days since an applicant's birth at the time of the application. The absolute value of this column is divided by 365 to get the age in years of each applicant. The data contains DAYS_BIRTH information for all clients. There appears to be a reasonable distribution of age in the training set, with applicant age ranging from about 20 to 70. The sensitive attribute of gender is represented in the column CODE_GENDER. The values in this column are either 'M' for male or 'F' for female, and gender information is available for every client in the training data. In the training set, there appears to be about twice as many female applicants as males.

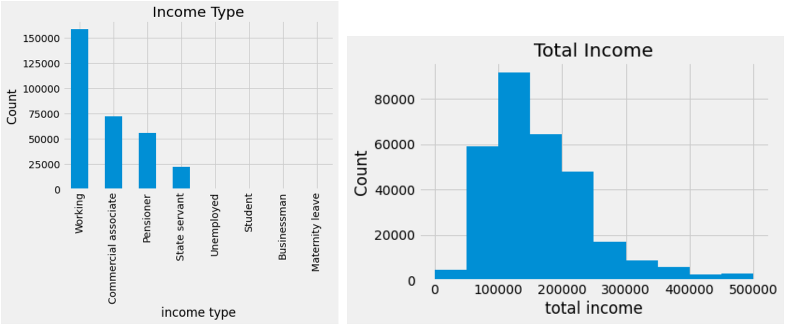


Figure 2: Training set: distributions of income type and income amount

Some other features of interest are income type and sensitive attribute of income amount. There are no missing values for either feature. In the distribution for income type (NAME_INCOME_TYPE), the "Working" income type is the most frequent, followed by"Commercial associate," "Pensioner," and "State servant." Other income types have much lower frequencies. The distribution of total income amount (AMT_INCOME_TOTAL) is skewed to the right, with the most frequent incomes being at about $150,000 to $175,000. This skew in the data is consistent with the purpose of Home Credit, which is to support low-income populations. Not all income amounts are displayed in this graph because there are some extreme values that would make most of the data impossible to observe in detail if they were included; the maximum income in the training set is $117,000,000.

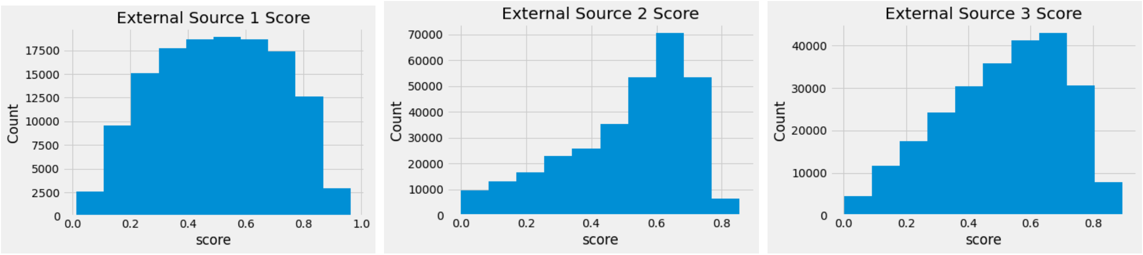


Figure 3: Training set: distributions of external source variables 1-3

When all the features are correlated with the target variable, three of the variables that are most strongly, negatively correlated with the target variable represent normalized scores from external sources (EXT_SOURCE_1, EXT_SOURCE_2, and EXT_SOURCE_3). There is not much information given about these features except that they are “a cumulative [...] credit rating made using numerous sources of data.” Each of these features are of data type float with values between 0 and 1. About 56.4% of EXT_SOURCE_1 values, 0.2% of EXT_SOURCE_2 values, and 19.8% of EXT_SOURCE_3 values are missing from the training set. The distribution of External Source 1 appears to be close to a bell curve; however, this distribution should not be taken at face value since External Source 1 values are missing for over half of the clients in the data. The distributions of External Sources 2 and 3 are both skewed to the left.

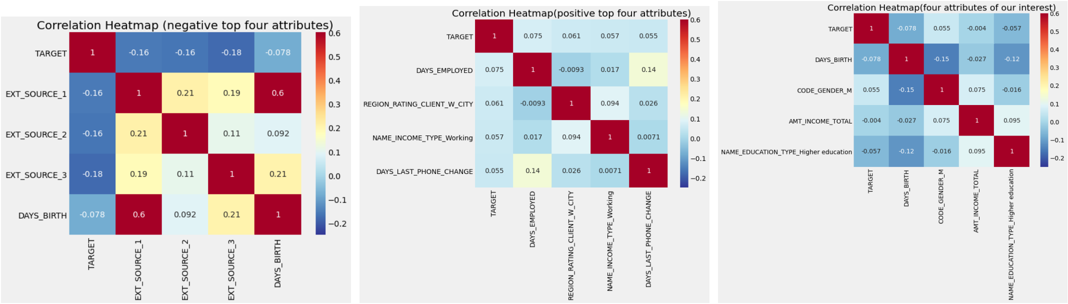


Figure 4: Heatmaps of some features correlated with the target variable

Along with External Source features 1-3, the absolute value of DAYS_BIRTH is also one of the most strongly, negatively correlated features with the target variable. These four features were correlated with each other in a heatmap by the author of the ADS. The absolute value of DAYS_BIRTH is fairly strongly and positively correlated with External Source 1. Higher values for these features correspond with higher probability of clients being classified with 0 for the target variable, meaning they are more likely to repay their loans on time.

We created a heatmap showing the pairwise correlations between the four features that were most strongly positively correlated with the target variable: DAYS_EMPLOYED (number of days applicant has been employed at their current job), REGION_RATING_CLIENT_W_CITY (Home Credit’s rating of the region where the client lives, considering city), NAME_INCOME_TYPE_Working (One Hot Encoded variable - whether an applicant’s income type is ”Working”), and DAYS_LAST_PHONE_CHANGE (number of days before the application since the client changed phones). All pairs of these features, excluding REGION_RATING_CLIENT_W_CITY with DAYS_EMPLOYED, are very weakly positively correlated with each other. Another heatmap we created shows the pairwise correlations of our chosen sensitive attributes and another variable of choice that could be of interest. Out of these features, the most strongly positively correlated variables are AMT_INCOME_TOTAL and NAME_EDUCATION_TYPE_Higher_education. The most strongly negatively correlated variables are CODE_GENDER_M (whether applicant is male or not) and the absolute value of DAYS_BIRTH.

In his data exploration, the author observed the potential effect of age on repayment ability. (The graph can be found in the "ADS: Effect of Age on Repayment" section in our Colab notebook.) He found that younger applicants are more likely to default on loans, and this information helps us analyze the ADS by thinking of younger applicants as an "unprivileged" group for the sensitive feature of age.

2.3 Output

The target variable in the training set is the clients' repayment abilities; a value of "1" indicates that a client had payment difficulties, and "0" represents all other cases. In line with this, the output of the system is a table containing the IDs of clients in the test set along with predictions (values between 0 and 1) of this target variable on the clients in the test set. The predictions represent the probability that the loan will not be repaid (probability of the client being classified with the value 1). If we wanted to use these probabilities to classify the applicants, we would set a threshold such as 0.5 for determining that a loan is risky. Any applicant with a probability below or above this threshold would be classified as 0 or 1, respectively.

3 Implementation and Validation of ADS Code

3.1 Data Cleaning and Pre-processing

Firstly, the number and percentage of missing values in the training set were examined. Of the 122 columns in the training data, 67 have missing values. The data type for each column was also examined. The data consists of 65 float features, 41 integer features, and 16 object/categorical features.

The ADS used two encoding methods to convert the categorical data into numerical data. For any categorical variable with two unique categories, the system used label encoding, which assigns each unique category in a categorical variable to an integer, not creating any new columns. For any categorical variable with more than two categories, the ADS used one-hot encoding, which creates a new column for each unique category; the presence of particular categories is marked with a 1 in their respective columns, and all other categories are marked with a 0. Since some new columns were created in the training set by one-hot encoding, the system used the align function to match the size of the test set with the training set. After the aligning process, there are 240 features in the training set and 239 in the testing set, which is compatible for model training and prediction.

The author also examined the data for anomalies, which are values which appear to be abnormal due to mistyped numbers, errors in measurement, and outliers (valid but extreme values). For example, after converting DAYS_BIRTH into years, it was found that the maximum age is about 1000 years old, which does not fall into the reasonable human age range. For the DAYS_EMPLOYED feature, clients with non-anomalous values for this column have a 8.66% rates of default, but the anomalous clients have an unexpectedly low rate of default of 5.4%. This observation indicates that clients with anomalous values for certain columns still have some importance in affecting the output, so the ADS fills the anomalous values with np.nan values. Anomalous values were imputed in the same way as missing values are later on.

The author used two feature construction methods to engineer some new features that may improve model performance. He created 35 polynomial features, which are features that are powers of and/or interaction terms (products) between existing features. These features were added to copies of the original training and testing data. However, these features were not used in the final random forest model because they did not contribute to a higher AUC score. He also created "domain knowledge" features, which are columns that combine existing features to produce features that are more specific to the subject of credit. Because these features are more specific to this particular field, they may be important in predicting whether clients will default on a loan or not. The four new features - CREDIT_INCOME_PERCENT, ANNUITY_INCOME_PERCENT, CREDIT_TERM, and DAYS_EMPLOYED_PERCENT - were added to copies of the original training and test sets called app_train_domain and app_test_domain.

Before fitting the data to the random forest model, missing values were imputed by filling them in with the median of their respective columns. The data was then scaled using MinMaxScaler to normalize the ranges.

3.2 Description of Implementation

A random forest classifier was created with 100 trees. The random_state parameter was set to 50 to make the same results of the model reproducible. The verbose parameter was set to 1 to enable some logging, and the n_jobs parameter was set to -1 to use all processors. The model was then trained on the pre-processed training data. The predict_proba function was used to predict the probability of a 0 or 1 value for each client in the pre-processed test set, and only the probabilities of the client getting a value of 1 were saved since that is the target value. An output dataframe containing the ID of each client ('SK_ID_CURR') and their predicted target value was created. This model using only the original features scored about 0.678.

The process of imputing, min-max scaling, training, and generating predictions was repeated with the training sets containing polynomial features (app_train_poly) and domain knowledge features (app_train_domain). The model using the new polynomial features scored the same as the original model (0.678), so the author concluded that polynomial feature construction did not help improve the system. The model using domain knowledge features along with the original features scored very slightly higher than the original model (about 0.679), so this is the final model we evaluated.

3.3 ADS Validation

The submissions are evaluated on area under the ROC curve (AUC) between the predicted probability and the observed target. The ROC (Receiver Operating Characteristic) graphs the true positive rate versus the false positive rate. Moving along the line indicates changing the threshold used for classifying a positive class, and if multiple models are being compared, the curve that is to the left of and above other curves represents the better model. AUC (Area Under the Curve) is between 0 and 1; an AUC closer to 1 is ideal. Overall, the ADS uses AUC score as its validation metric - the higher the AUC, the more accurate the model. As discussed above, the random forest model with original features has an AUC score of 0.678. The most accurate model is the random forest model using data that includes the domain knowledge features, with an AUC score of 0.679, indicating a slightly higher accuracy than just using the original data.

4 Outcomes

4.1 Accuracy

We identified the overall accuracy as well as the accuracy in different sub-populations based on three sensitive attributes: age (DAYS_BIRTH), gender (CODE_GENDER), and total income (AMOUNT_TOTAL_INCOME). Age and gender are commonly used protected attributes, and income amount was chosen since this feature may be important in determining whether someone may get a loan or not. Home Credit is also interested in supporting low-income populations, so we want to verify that the ADS does not discriminate against this group.

After making three copies of the original training data with domain features, one for each sensitive attribute, we converted the sensitive attribute for each copy into a binary variable based on the median value if applicable. For age, applicants at least 43 years old (slightly below the median since it was a decimal number) were marked with a 1, and applicants younger than 43 years old were marked with a 0. Gender had already been label encoded in pre-processing, so we used the CODE_GENDER_M column as the sensitive attribute column, in which male and female clients were marked with 1 and 0, respectively. For income, all clients with at least the median income amount ($147,150) were marked with a 1 while clients with income lower than the median were marked with a 0.

Overall and Group Accuracy Based on Sensitive Attributes

	Age	Gender	Income
Overall Accuracy	0.919305	0.919228	0.919321
Privileged Group Accuracy	0.93736	0.897438	0.922877
Privileged Group Accuracy	0.901104	0.930683	0.915758

Because the provided test set did not contain the true labels, we re-split the copies of the training sets into new training and test sets for each sensitive attribute, with 80% of the data in the training sets and 20% in the test sets. We then used the new training sets to train the ADS and analyzed the accuracy and AUC score of the model overall as well as the accuracy for each privileged and unprivileged group. Since the ADS's target variable was originally a probability (generated using the .predict_proba function), we also came up with predictions as binary outcomes (using .predict, which sets the threshold for classification at 0.5 by default). The overall accuracy when the models were re-trained using the three new training groups was 0.9193, 0.9192, and 0.9193 with age, gender, and total income amount as the binary sensitive attribute in the training set, respectively. The AUC score was 0.719 (age), 0.715 (gender), and 0.695 (income). The overall accuracy and AUC score are approximately the same, though AUC was lowest when income was turned into a binary variable. The accuracy is greater than 0.9 in all cases, which means the ADS has a good overall accuracy even when our three sensitive attributes are converted into binary variables.

The performance of the ADS for different sub-populations was analyzed using the aif360 package's ClassificationMetric functions. All groups previously marked with a 1 (age >= 43, male, income >= $147,150) were called the privileged groups, while their binary counterparts (age < 43, female, income < $147,150) were considered the unprivileged groups.

For the sensitive attribute of age, the accuracy for the privileged group was 0.937, which is higher than the accuracy for the unprivileged group (0.901). The higher accuracy for the privileged group indicates the performance of the ADS is better for older applicants. For the sensitive attribute of gender, the accuracy for the privileged group was 0.897, which is unexpectedly lower than that for the unprivileged group (0.931), indicating that the ADS performed better for female applicants than male applicants in this test set. This may be attributed to the fact that there were almost twice as many females as males in the original test set. For the sensitive attribute of income amount, the accuracy of the privileged group was 0.923, which is greater than the unprivileged group's accuracy at 0.916. The higher accuracy for the privileged group indicates the performance of the ADS is better for clients with a greater income amount. In conclusion, the ADS performed better when the applicants were older than 43 years old, female, and had an income amount greater than $147,150. Since we chose not to drop other sensitive attributes when analyzing each sensitive attribute, these conclusions may not be independent of each other, and the test set may not be representative of the population of clients. Though we should be careful in drawing these conclusions, we can clearly see that there are groups for which the ADS is more likely to predict the target variable correctly.

4.2 Fairness and Diversity

We continued to use the same privileged and unprivileged sub-populations identified above to analyze the ADS's fairness and diversity. We chose disparate impact as one of our fairness measures, defined as the probability of the unprivileged group being classified as 1 over the probability of the privileged group being classified as 1. Disparate impact was analyzed since we want to ensure that membership in a particular group doesn't lead to discrimination and that all people have a fair chance of being given a loan. As seen in the data exploration by the ADS's author, younger applicants are more likely to default on loans; this trend in the training data could lead to many more younger applicants being classed with a 1 in the test set. Another measure that we chose is false positive rate difference, defined as the false positive rate for the unprivileged group minus that for the privileged group. A false positive in the context of our ADS would mean the system predicted that an individual will have a hard time repaying a loan (1), when in actuality they will repay the loan with no problems (0). A high false positive rate for a group would then mean a qualified individual who is part of this group may have less of a chance of taking out a loan. We want to make sure qualified applicants are not being denied loans due to certain group membership, so we want to evaluate the difference in this metrics between subgroups based on protected attributes.

Disparate Impact and FP/FN Rate Difference Based on Sensitive Attributes

	Age	Gender	Income
Disparate Impact	8.064516	0.347657	2.337966
FP Rate Difference	0.000036	-0.000026	0.000036
FN Rate Difference	-0.00179	0.001239	-0.001052

Between the privileged and unprivileged age groups, the disparate impact is about 8.065. This number shows that the ADS is biased between different age groups when making predictions; since the outcome 1 represents difficulties repaying a loan and 0 represents successfully repaying a loan, this value indicates that the unprivileged group (younger than 43 years) is much more likely to be classified with a 1 (or assigned a value that is closer to 1 if the target value is a probability), which is the unfavorable outcome in this context. The false positive rate difference is very small but still positive; this indicates that the false positive rate is slightly greater for the unprivileged group, meaning applicants younger than 43 years old are slightly more likely to incorrectly be classified with a 1.

Between the privileged and unprivileged genders, the disparate impact is about 0.348. This is much smaller than the value for the age groups. However, it is a positive number, so it indicates that females are more likely to be classified with a 1 than males. The false positive rate difference is negative, meaning males have a very slightly greater false positive rate than females and thus are slightly more likely to be incorrectly classified with a 1. This agrees with the values for accuracy that were found earlier, since the accuracy for females was higher than that for males. Again, there were many more females than males in the original training data, so these values could differ if the data were more balanced.

Between the privileged and unprivileged groups for income amount, the disparate impact is about 2.338. Again, this value is positive, so it indicates that the unprivileged group, or applicants with less than $147,150 total income, have a high probability of being classified with a 1 compared to the privileged group. The false positive rate is same as between the two groups for age (0.000036), applicants who earn less are slightly more likely to be incorrectly classified with a 1.

4.3 Other ADS Performance Properties

As an additional measure of fairness and of accuracy, false negative rate difference was analyzed. The magnitude of the false negative rate differences were consistently higher than that of the false positive rate differences, meaning the system incorrectly classes certain groups as likely to repay a loan more often than it incorrectly classes the other group as likely to default on a loan. The false negative rate differences for age and income are both between -0.001 and -0.002, meaning it is slightly more likely for the privileged groups (older than 43 years or more than median income) to be incorrectly classified with a 0. For gender, the false negative rate difference is a small positive number, indicating that it is slightly more likely for females to be incorrectly classified with a 0. It appears that in this test set, females are "privileged" in comparison to males, but again, since there are many features included in the data, other features could be affecting these results.

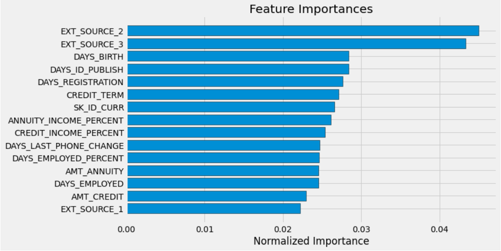
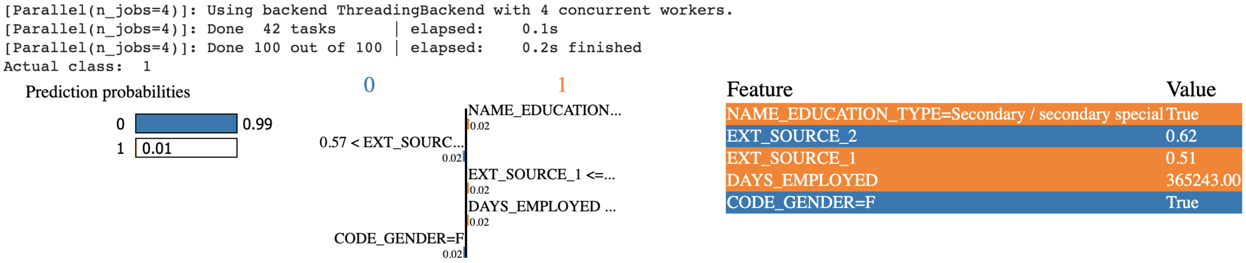


Figure 5: Feature importances for ADS

In the ADS implementation, the author includes a chart of feature importances for the random forest model; this provides insight about which features are the most relevant and decisive for the outcome. The features are sorted in descending order based on their normalized importance value. All four of the features which are most negatively correlated with the target (EXT_SOURCE_1, EXT_SOURCE_2, EXT_SOURCE_3, and DAYS_BIRTH) are present in this chart and are thus among the 15 most important features for our model.

We then used the LIME (Local Interpretable Model-agnostic Explanations) package to analyze explanations for the system's predictions for some of the most important instances in the data. To make the data appropriate for processing in LIME, we encoded the categorical data in the original training data (including domain features) using LabelEncoder, split the original training set into new training and test set with 80% of data in the training set, then retrained the model. After making predictions on the test set, we used the LIME Tabular Explainer and Submodular Picker to choose useful, representative examples that give a global explanation for the model. The five chosen examples were at indices 0, 8, 1, 4, and 7. Of these instances, clients 8, 1, 4, 7 were correctly classified, while client 0 was misclassified.

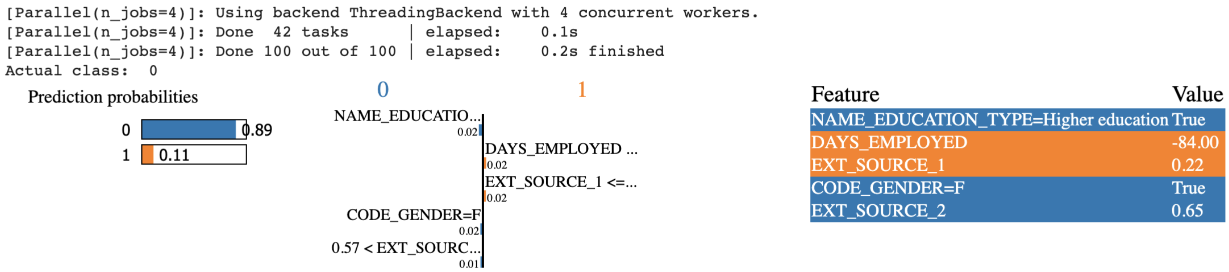


```
[Parallel(n_jobs=4)]: Using backend ThreadingBackend with 4 concurrent workers.
[Parallel(n_jobs=4)]: Done  42 tasks      | elapsed:    0.1s
[Parallel(n_jobs=4)]: Done 100 out of 100 | elapsed:    0.2s finished
Actual class:  1
```

Feature	Value
NAME_EDUCATION_TYPE=Secondary / secondary special	True
EXT_SOURCE_2	0.62
EXT_SOURCE_1	0.51
DAYS_EMPLOYED	365243.00
CODE_GENDER=F	True

Figure 6: LIME Explanation for Client 0

The classification for client 0 is a misclassification. The actual class of the client is 1 (had difficulties repaying), but the classification shows a 99% probability towards the favorable outcome 0. Of the top five features, the EXT_SOURCE_2 AND CODE_GENDER=F features contribute to the favorable outcome (0), which contribute the misclassification. The NAME_EDUCATION_TYPES, EXT_SOURCE_1, and DAYS_EMPLOYED features contribute to the correct unfavorable outcome (1). We also discovered the EXT_SOURCE_3 and FLAG_DOCUMENT_9, 11, and 17 features also strongly shift the classification towards the incorrect outcome 0 when we extend the feature-value table to show the top 10 features.



```
[Parallel(n_jobs=4)]: Using backend ThreadingBackend with 4 concurrent workers.
[Parallel(n_jobs=4)]: Done  42 tasks      | elapsed:    0.1s
[Parallel(n_jobs=4)]: Done 100 out of 100 | elapsed:    0.2s finished
Actual class:  0
```

Feature	Value
NAME_EDUCATION_TYPE=Higher education	True
DAYS_EMPLOYED	-84.00
EXT_SOURCE_1	0.22
CODE_GENDER=F	True
EXT_SOURCE_2	0.65

Figure 7: LIME Explanation for Client 7

The classification for client 7 is a correct classification, since the actual class is 0, and the prediction is consistent with a 89% probability towards this outcome. Of the top five features, the NAME_EDUCATION_TYPE =Higher education, CODE_GENDER=F, and EXT_SOURCE_2 features contribute to the correct label (0), while the DAYS_EMPLOYED and EXT_SOURCE_1 features contribute to the unfavorable outcome (1).

5 Summary of Findings

5.1 Appropriateness of Data

We do believe the data was appropriate for this ADS. The data was provided by Home Credit, which is the company that actually wants to make use of an ADS like this in their practice. There were a very large number of features and many applicants included in the data. While it is not clear how this data was obtained from the Kaggle competition page - whether it was collected from actual clients the company considered and/or worked with or the ”clients” are fictional but realistic - it was data deemed appropriate by the company themselves to distribute in order for competitors to build an system that could potentially be deployed. Since the data was for a competition, the target values were not provided for the test set, creating limitations that we had to work around while auditing the ADS. However, in an environment where the company or a partner of the company is trying to improve the ADS, this data would be available.

5.2 Discussion of Robustness, Accuracy, and Fairness

We believe the ADS implementation is fairly accurate. When we retrained the model on our new training sets for each sensitive attribute, the overall accuracy score was about 0.92 for each new testing set, and the AUC was around or greater than 0.7 for each as well. Additionally, for each privileged and unprivileged group, the accuracy was close to or greater than 0.9. Though of course these values could be even higher, all the accuracy scores are high values, and the AUC scores are satisfactorily high as well. These measures would be found appropriate both by those using the ADS, i.e. Home Credit or other financial institutions, and those affected by it, including everyone who applies for loans.

Some of the robustness of this ADS comes from pre-processing, where anomalous values (potentially including outliers) are handled by being replaced with medians. However, as seen in the feature importance graph, the external source variables, particularly 2 and 3, appear to hold a lot of weight in the model’s predictions. Due to this, the system may not be as accurate for clients that have ”abnormal” values for these features that aren’t handled by pre-processing. For example, the LIME explanation for client at index 0 showed that EXT_SOURCE_2 pushed the prediction in the opposite direction as the client’s true class. Though we don’t have much information on what the external source features are, we can see that they may play a role in decreasing the robustness of the system somewhat.

In contrast to its high accuracy, we do not believe this implementation is currently fair enough. The accuracy scores for privileged and unprivileged groups based on our chosen sensitive attributes are different; they may be similar enough to overlook especially since they are close to the overall accuracy and the new test sets were not as large as the original test set. The magnitudes of the false positive rates were quite small and thus acceptable, though it would be ideal to further minimize these values if possible. However, disparate impact values, particularly for age and income, were large, meaning membership in the unfavorable groups based on these features could lead to an unfair chance of being assigned the unfavorable label (1). These fairness measures are important for Home Credit and potentially other institutions utilizing the ADS to consider, especially because Home Credit's goal is to increase service for low-income and underbanked populations. There should be an improvement in at least the disparate impact based on income before the ADS is deployed if this goal is to be attained. These metrics are also very important to those looking to apply for loans because everyone wants and deserves a fair chance; it would be disappointing if a low-income individual approached Home Credit because he/she identified with their goals but got turned away because of the result of this system.

5.3 Potential of Deployment of ADS

We do not think this ADS should be deployed, at least in its current state. This decision is mostly based on the contradiction between Home Credit's mission and the results of analyzing the system's fairness. The system may be accurate, but it is not fair to unprivileged groups based on age and income, which is contradictory to what Home Credit aims to do for the international community. As discussed above, the ADS's results as of now could lead to unintentional discrimination based on group membership, but Home Credit wants to work towards mitigating such bias instead of perpetuating it, especially for low-income and underbanked populations. If this ADS were to be deployed, the institutions using it would not be able to use its predictions reliably to make decisions on whether to provide loans for clients in marginalized groups. In any case, rather than discriminating against anyone, Home Credit would likely want to provide additional guidance for those who may be less likely to repay loans on time, such as younger clients, and potentially take precautionary measures in case they do default on loans.

5.4 Potential Improvements

There should be increased transparency for the data collection. While there is good metadata in terms of detailed descriptions of most of the features in the datasets, there is almost no information on the Kaggle competition page about who collected the data and when it was collected. This improvement would ensure the dataset is more reusable and credible. There should also be more description of the External Source features, as there is currently no information about what they are even though they are important to the system's predictions. Even if we as auditors may not be able to fully understand what they are, since they are more industry-specific, the data and system would be more interpretable if these were described in detail.

In pre-processing, the proportion of clients in the privileged and unprivileged groups based on sensitive features (such as gender) could be balanced before training/prediction is done. If there is still unfairness after this step, it would show that the bias is not caused by imbalanced amounts of information the model was trained on. In the implementation stage of the system, it would be ideal to do some hyperparameter tuning based on both some accuracy measure and disparate impact in order to improve fairness while maintaining accuracy.