Unce Shahid and Dara Wang

College of Arts and Sciences: New York University

Audit of a Default-Risk Predicting ADS

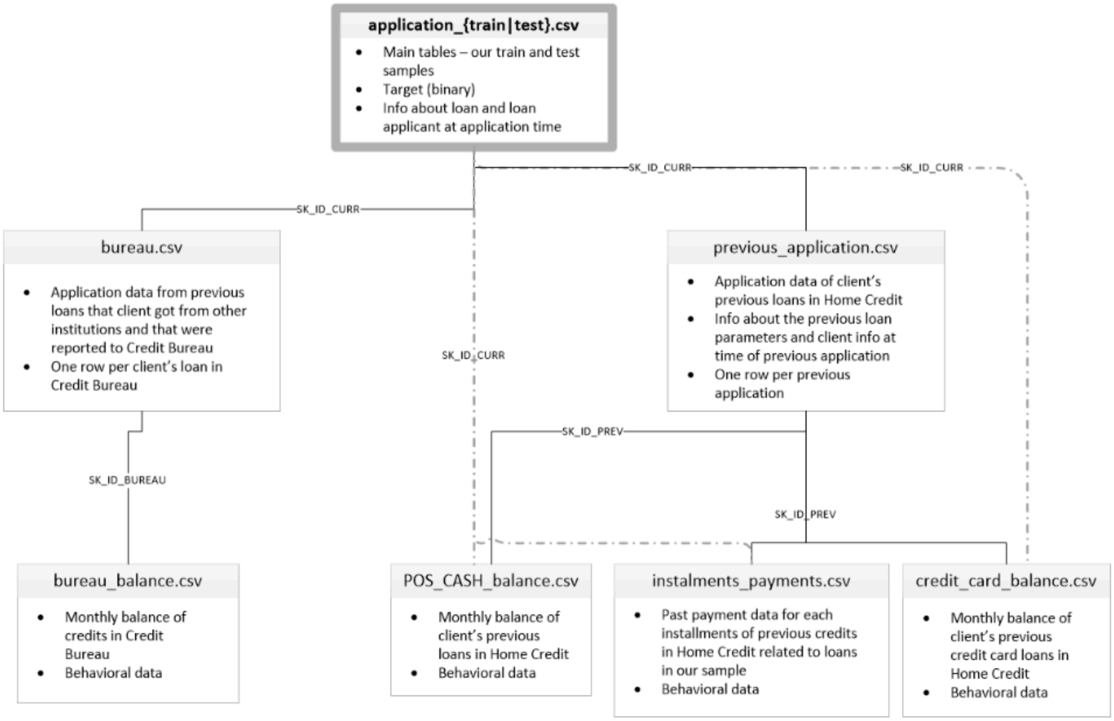
May 9, 2023

Background

The automated decision system we have selected is Omar Osman's solution for Home Credit Group's Kaggle Competition for predicting Home Credit Default Risk. Loans are a method of borrowing money, helping people pay for immediate expenses when they don't currently have the money to do so. The expenses can vary in many forms, be it grocery shopping with a credit card, taking a loan to pay for down payment on a new car, or to pay for wedding expenses. Loans have been heavily integrated into our society, especially through the use of credit cards. Credit cards follow a "buy now, pay later" structure, allowing people to buy immediate necessities, such as groceries, before they receive their paycheck. Loans, instead, are more likely to be used for emergency payments, like fixing a vehicle after a car crash or paying for an emergency medical treatment. However, loans are typically given to people with good credit scores, meaning that being well-off financially is often a prerequisite. To obtain a good credit score, you have to make use of borrowing money often and always repay the money on time. In that sense, a person would repay the money borrowed so fast that they wouldn't have needed to borrow money in the first place. However, the people who need loans the most are those who end up struggling the most financially. If you don't have any spare money to pay for emergencies now, it will still be very difficult to pay back the borrowed amount in the future. As such, trusted loan lenders typically require a high credit score in order to offer people loans so they can be certain that most of the loans will end up repaid. However, there are a multitude of possible reasons as to why a person may have a low credit score. This Kaggle Competition is focused on finding alternative measures to credit score that will help predict the risk of a loan borrower defaulting, thus allowing people with low credit scores to find a reputable business to obtain their loans from. The only alternative people with low credit scores have is to borrow money from untrustworthy lenders and thus they end up being taken advantage of. Home Credit seeks to find a viable alternative to help clients get approved for loans from reputable sources instead.

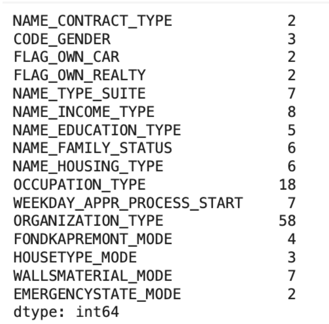
Input and Output

Input: The data comes from two different sources: loan applicants for Home Credit Group and loan applicants that applied to the Credit Bureau. Specifically each row in the data is an application for a loan, including various data such as type of loan, whether they own a car, if they have children, how much their income is, and much more with a total of 122 columns. Home Credit is an international financial institution that operates in 9 countries and focuses on delivering greater financial inclusion and serving people with low credit history. A credit bureau is an agency that collects, researches, and sells individual credit information. The provided credit bureau data is specifically of the loan applicants already in the train/test dataset, allowing more information to be gleaned if said client has a history of previous credits. Only 20% of the total data is provided publicly for each Kaggle submission to be trained on, followed by a private testing performed on the other 80% of the data and is scored for performance. A thorough exploratory data analysis was already performed by Omar Osman and will be used in this report. The following visualization was provided by Home Credit and illustrates the provided data and their connection to provided alternative data.

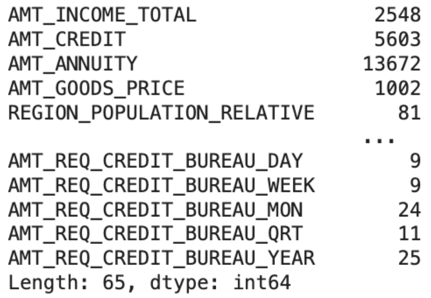

application_{train|test}.csv
Main tables – our train and test samples
Target (binary)
Info about loan and loan applicant at application time
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
SK_ID_CURR
bureau.csv
Application data from previous loans that client got from other institutions and that were reported to Credit Bureau
One row per client's loan in Credit Bureau
previous_application.csv
Application data of client's previous loans in Home Credit
Info about the previous loan parameters and client info at time of previous application
One row per previous application
SK_ID_BUREAU
SK_ID_PREV
SK_ID_PREV
bureau_balance.csv
Monthly balance of credits in Credit Bureau
Behavioral data
POS_CASH_balance.csv
Monthly balance of client's previous loans in Home Credit
Behavioral data
instalments_payments.csv
Past payment data for each installments of previous credits in Home Credit related to loans in our sample
Behavioral data
credit_card_balance.csv
Monthly balance of client's previous credit card loans in Home Credit
Behavioral data

The bureau dataset is used to find any pre-existing data of the clients available from the Credit Bureau. The previous application dataset is used to find any history the client has with Home Credit. Both of these history-checking datasets also contain balance information to see the monthly performance of each applicant in order to determine clients' historical loan-repayment strategies. Additionally, there is more information provided for clientele history in Home Credit, such as installments payments and credit card payments. Having a history of credit repayment helps to increase trust in the client repaying their loans. Overall, Home Credit provides loan applications and as much historical information they have available about each applicant.

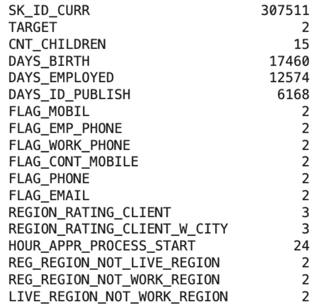
Input feature datatypes include float64, int64, and object, with specifics shown below (some flag_document variables are not shown due to space limitations but there are 19 of them, all consisting of the datatype int64).



```
NAME_CONTRACT_TYPE             2
CODE_GENDER                    3
FLAG_OWN_CAR                   2
FLAG_OWN_REALTY                2
NAME_TYPE_SUITE                7
NAME_INCOME_TYPE               8
NAME_EDUCATION_TYPE            5
NAME_FAMILY_STATUS             6
NAME_HOUSING_TYPE              6
OCCUPATION_TYPE               18
WEEKDAY_APPR_PROCESS_START     7
ORGANIZATION_TYPE             58
FONDKAPREMONT_MODE             4
HOUSETYPE_MODE                 3
WALLSMATERIAL_MODE             7
EMERGENCYSTATE_MODE            2
dtype: int64
```

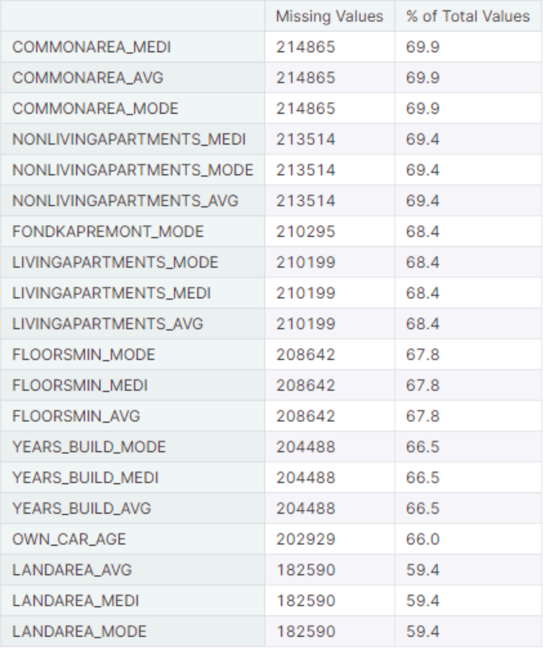


```
AMT_INCOME_TOTAL               2548
AMT_CREDIT                     5603
AMT_ANNUITY                   13672
AMT_GOODS_PRICE                1002
REGION_POPULATION_RELATIVE       81
                                ...
AMT_REQ_CREDIT_BUREAU_DAY         9
AMT_REQ_CREDIT_BUREAU_WEEK        9
AMT_REQ_CREDIT_BUREAU_MON        24
AMT_REQ_CREDIT_BUREAU_QRT        11
AMT_REQ_CREDIT_BUREAU_YEAR       25
Length: 65, dtype: int64
```



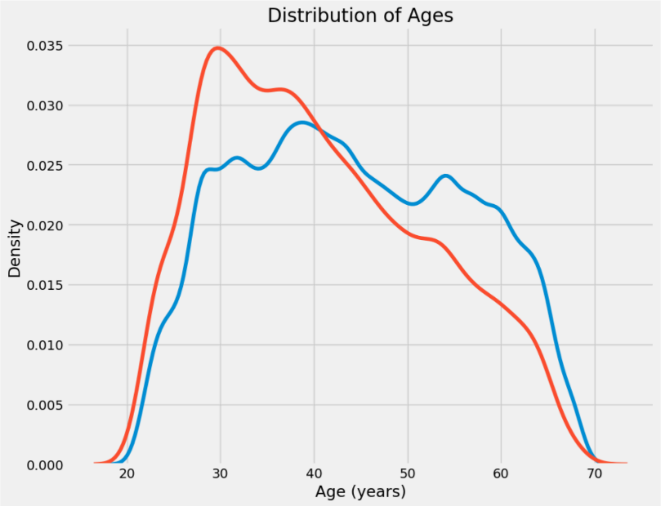
```
SK_ID_CURR                   307511
TARGET                            2
CNT_CHILDREN                     15
DAYS_BIRTH                    17460
DAYS_EMPLOYED                 12574
DAYS_ID_PUBLISH                6168
FLAG_MOBIL                        2
FLAG_EMP_PHONE                    2
FLAG_WORK_PHONE                   2
FLAG_CONT_MOBILE                  2
FLAG_PHONE                        2
FLAG_EMAIL                        2
REGION_RATING_CLIENT              3
REGION_RATING_CLIENT_W_CITY       3
HOUR_APPR_PROCESS_START          24
REG_REGION_NOT_LIVE_REGION        2
REG_REGION_NOT_WORK_REGION        2
LIVE_REGION_NOT_WORK_REGION       2
```

Missing Values: The input data the ADS uses are all of the 122 columns in the provided training dataset and all 6 additionally provided datasets. Overall, there are too many features to describe each individually. Instead, we will look at a portion of them. The below chart displays the top 20 columns with missing values. There are many, many missing values within the dataset that will end up handled. 67 of the 122 columns had missing values.

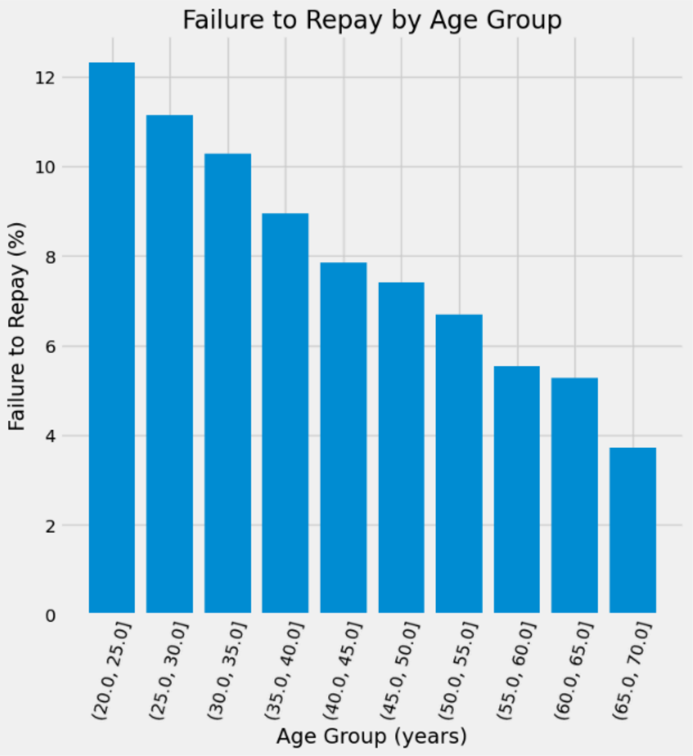


	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8
FLOORSMIN_AVG	208642	67.8
YEARS_BUILD_MODE	204488	66.5
YEARS_BUILD_MEDI	204488	66.5
YEARS_BUILD_AVG	204488	66.5
OWN_CAR_AGE	202929	66.0
LANDAREA_AVG	182590	59.4
LANDAREA_MEDI	182590	59.4
LANDAREA_MODE	182590	59.4

Age: One interesting exploratory analysis is to see whether age has an influence on whether a person will repay or default on their loan. Below is a graph of two kernel density estimate (KDE) plots showcasing the distribution of ages of clients who ended up repaying their loans in blue and the age distribution of clients who end up defaulting in red.

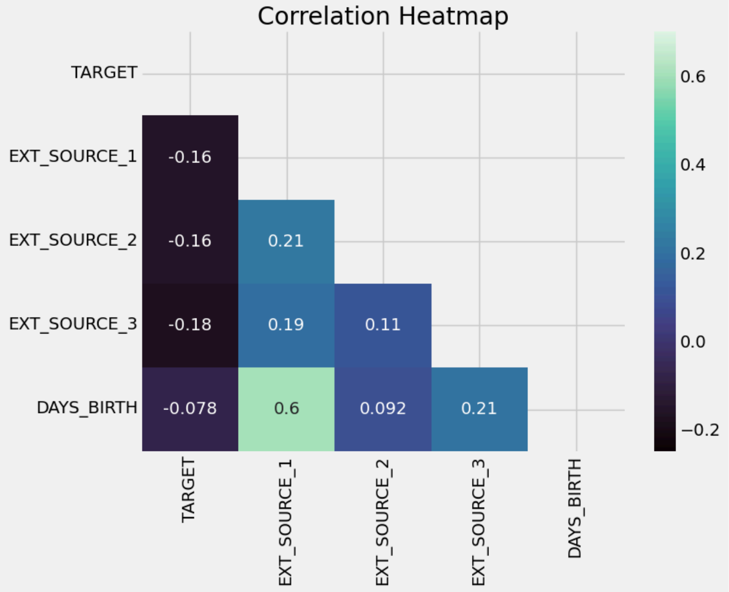


The rate of defaulting is skewed towards younger people in the age range from 25-35, while the rate of paying back loans is skewed towards older people from the age range of 45-65. Younger people are more likely to default, while older people are less likely to default. This is accompanied with the negative correlation between age and target of -.07, illustrating that age has a slight negative influence on whether a person will default. This trend is further supported by the following visual:



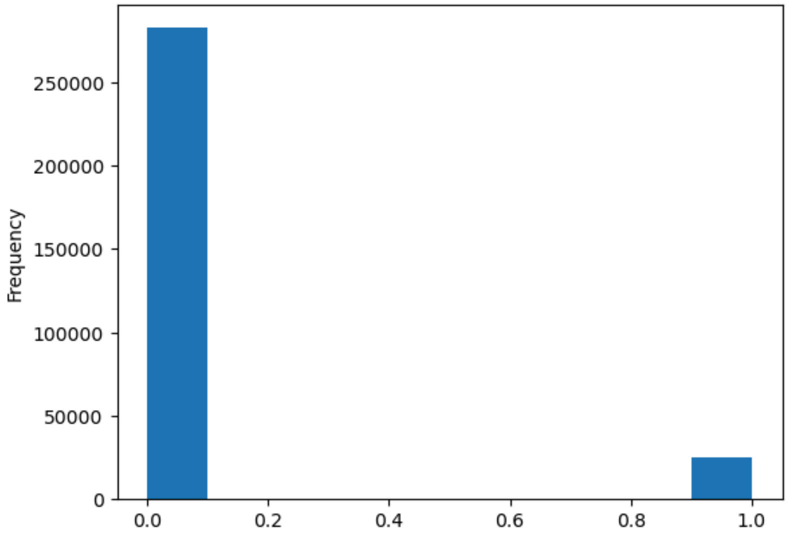
There is a steady monotonic trend of decreasing rate of defaulting as age increases. The youngest bin of ages 20-25 has triple the rate of defaulting as compared to the highest age bin of 65-70. It can be reasonably assumed that older people are more likely to pay back their loans.

Exterior Data: Exterior data is the data from the three datasets POS_CASH_balance, installments_payments, and credit_card_balance. The following plot is a heatmap correlation plot between the information contained in these three datasets, age, and the target variable of whether the applicant defaulted or paid back their loan.



All three exterior sources ended up with a negative correlation between -.16 and -.18 when correlated with the target variable. This can be translated into meaning that people with a better history of repaying loans in any of the three ways are also less likely to end up defaulting. Additionally exterior source 1 has a very large correlation with age, while exterior source 3 has a sizable correlation with age and exterior source 2 has a slight correlation with age. People of higher age are more likely to have a better credit history and thus are more likely to end up repaying their loans. Overall, the use of these three exterior sources will increase the impact age has on predicting whether a client will default or not.

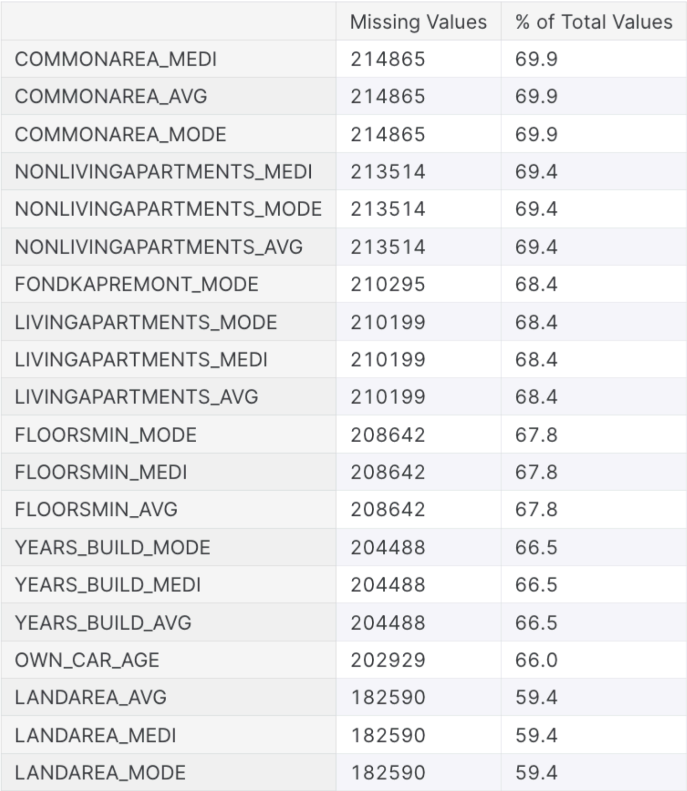
Output: The target feature to be predicted is the likelihood of an individual to default. The output is a range of values from 0 to 1 to indicate likelihood. However, in the provided train dataset, whether a client defaults or not is binary. Thus, the performance of the model will be tested by the AUROC. The distribution of the actual default rate is provided below.



The above plot distribution illustrates how many clients ended up paying back the loan or defaulting. Most loans were paid back, while some ended up defaulting. There is a notable imbalance between the two outcomes and that must be considered when assessing the ADS's ability to accurately predict whether a client will default or not.

Implementation and Validation: Cleaning & Preprocessing

Data Cleaning: The data cleaning is as follows. First, the datatypes of the columns in the datasets POS_CASH_balance, installments_payments, and credit_card_balance have been reformatted to use less memory. The function get_balance_data() was used to read the datasets into memory with specified data types, reducing memory consumption significantly. Then, to handle missing values, the function missing_values_table() was used to analyze the missing data in the training set app_train. As shown below, it identified columns with missing values and reported their percentage of missing data.



	Missing Values	% of Total Values
COMMONAREA_MEDI	214865	69.9
COMMONAREA_AVG	214865	69.9
COMMONAREA_MODE	214865	69.9
NONLIVINGAPARTMENTS_MEDI	213514	69.4
NONLIVINGAPARTMENTS_MODE	213514	69.4
NONLIVINGAPARTMENTS_AVG	213514	69.4
FONDKAPREMONT_MODE	210295	68.4
LIVINGAPARTMENTS_MODE	210199	68.4
LIVINGAPARTMENTS_MEDI	210199	68.4
LIVINGAPARTMENTS_AVG	210199	68.4
FLOORSMIN_MODE	208642	67.8
FLOORSMIN_MEDI	208642	67.8
FLOORSMIN_AVG	208642	67.8
YEARS_BUILD_MODE	204488	66.5
YEARS_BUILD_MEDI	204488	66.5
YEARS_BUILD_AVG	204488	66.5
OWN_CAR_AGE	202929	66.0
LANDAREA_AVG	182590	59.4
LANDAREA_MEDI	182590	59.4
LANDAREA_MODE	182590	59.4

The identified missing values were then handled appropriately, with columns having a high percentage of missing values being considered for removal or imputation. Imputation was performed using the mean-value strategy. Finally, anomalies such as mis-typed numbers, errors in measuring equipment, or extreme measurements in the dataset, particularly found within the column DAYS_EMPLOYED, were flagged and replaced with NaN.

Pre-Processing: For categorical variables, the code applied two types of encoding, label encoding and one-hot encoding. Label encoding was used for categorical variables with two or fewer unique categories like CODE_GENDER, while One-Hot Encoding was used for categorical variables with more than two unique categories. This process increased the dimensionality of the data.

Data Alignment: After one-hot encoding was performed, more columns were in the training data than existed in the test set since some categories were not represented in the testing data. As there needs to be the same features (columns) in both the training and testing data, the ADS code aligned the training and testing data by only retaining column labels that are present in both via an inner join.

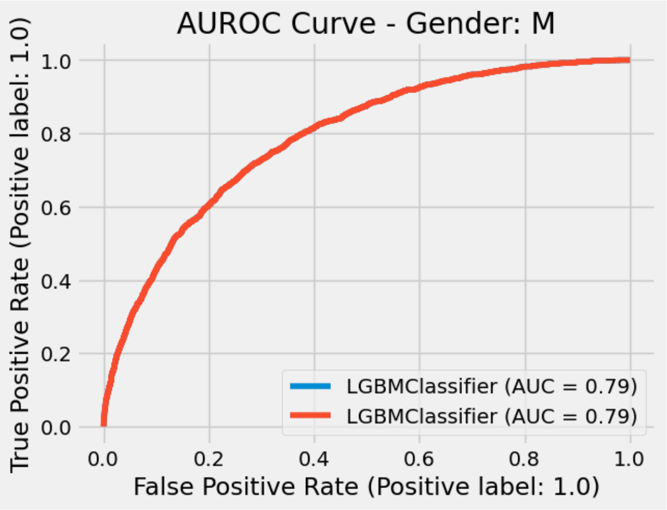
System Implementation: The ADS system is built using several machine learning models to predict the probability of loan defaults. The system used a variety of models, including Logistic Regression, Random Forest, and XGBoost. The target variable was TARGET, which indicated whether a loan applicant had repayment difficulties (1) or not (0). The models were evaluated based on their ability to predict this target variable. These models are trained on the training dataset (app_train) and evaluated on the testing dataset (app_test). Each model is assessed for feature importance to determine which factors have the greatest impact on predicting loan defaults. For instance, the Random Forest model computes feature importances, which are then visualized to identify the top contributing features. This insight helps in understanding the factors that influence loan repayment behavior, aiding in effective risk management strategies.

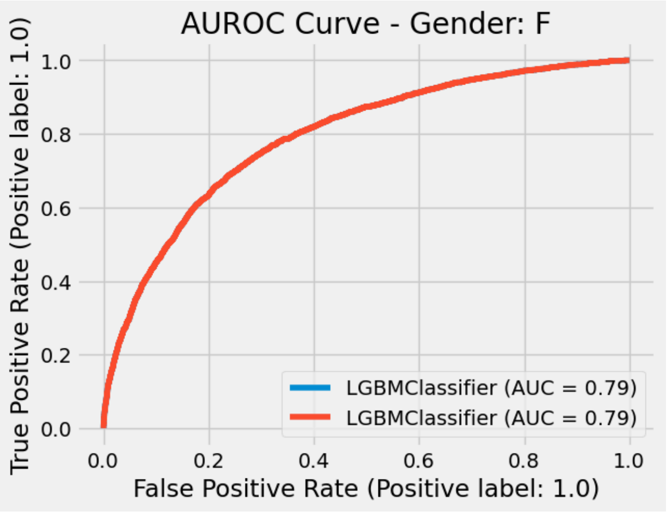
The implementation utilizes domain knowledge to engineer new features based on existing data. For example, CREDIT_TERM is created by dividing AMT_ANNUITY by AMT_CREDIT, representing the effective loan term, and DAYS_EMPLOYED_PERCENT is created by dividing DAYS_EMPLOYED by DAYS_BIRTH, showing the proportion of a client's life spent in employment. These new features offer valuable insights into a client's financial behavior, potentially enhancing the model's predictive capabilities.

ADS Validation: For the ADS, validation is achieved through a combination of train-test split and cross-validation. The dataset is initially divided into a training set (app_train) and a testing set (app_test). The model is trained on the training set and evaluated on the testing set, which it has not seen before, ensuring the model's ability to generalize to new, unseen data. Cross-validation further strengthens this process by dividing the dataset into multiple subsets and training and testing on different combinations, providing a robust assessment of model performance across various data segments.

Outcomes

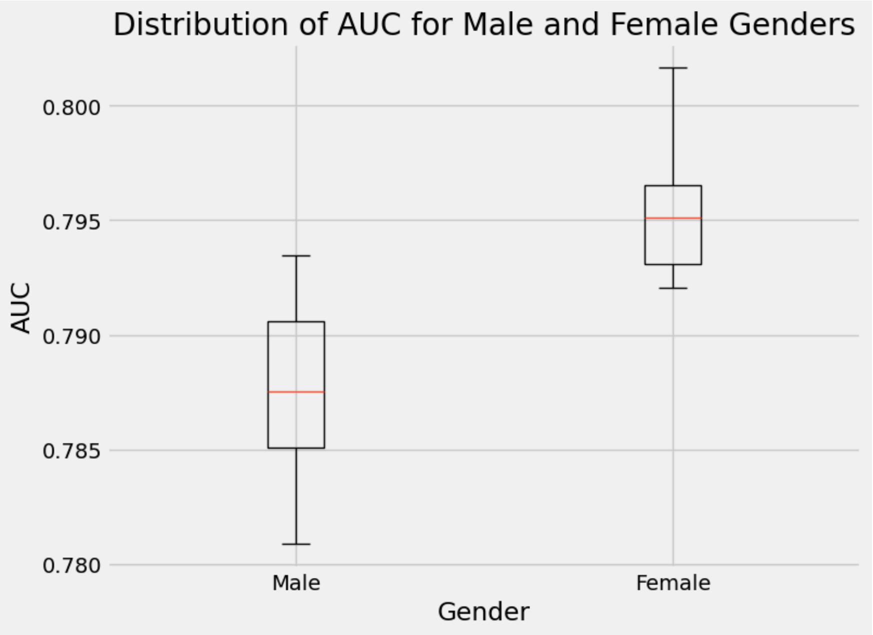
Accuracy: All of the ADS submitted to the Kaggle competition are evaluated and ranked based on AUROC, so that is the measure we will use to check for accuracy. Overall, the model has an AUROC value of 0.79. Below are two AUROC graphs, one for males and one for females. The AUROC specifically for the male gender is 0.789 and the AUROC for the female gender applicants is 0.793. Both of the genders have similar AUROC scores, so the ADS is similarly accurate for both of them.





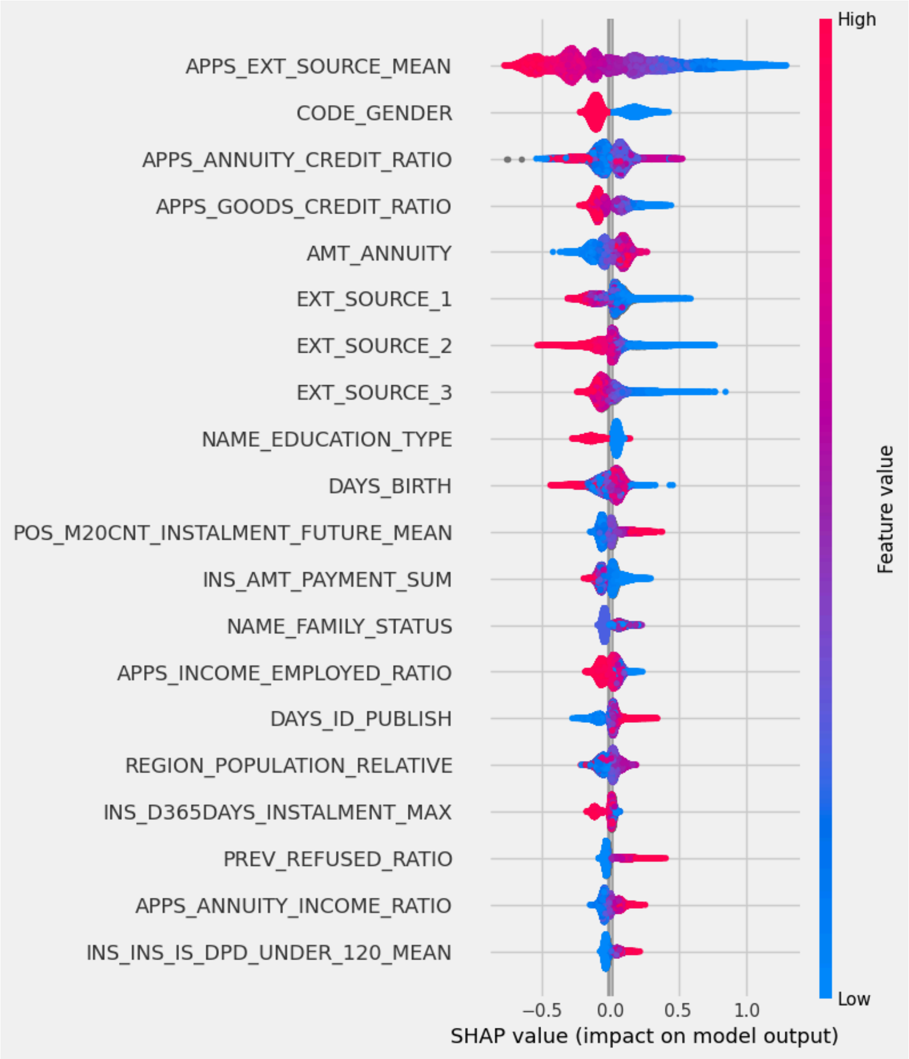
Because the target label is heavily imbalanced (much more loans were repaid than defaulted), we did not choose accuracy as a metric.

Stability: We determined the stability of the model by running 10 different train/test splits and seeing how much variation there was between the AUC in each gender.



As shown by the box plot above, the female gender has a very tight interquartile range, while the male gender has a larger, but still small interquartile range. There is not much variation between the AUC scores for both genders, so the model is consistent.

Fairness: We measured fairness via the equalized odds ratio and equalized odds difference as equalized odds is a good comparison as to how different genders get treated. Equalized odds ensures false positive rate parity and false negative rate parity. We obtained an equalized odds ratio of 0.570 and an equalized odds difference of 0.022 via fairlearn metrics. An ideal equalized odds ratio is a value of 1.0, meaning that the true positive rate and false positive rate are completely equal to each other for both the male and female gender. However, we have a very low ratio of 0.57, which is not good at all. There is a substantial difference between the model's ability to predict male loan default rates and female loan default rates. It may be confusing as to why both genders have similar AUROC scores, but end up with very different equalized odds. One potential explanation is that overall the model is similarly accurate for both genders in regards to both true positive rate and false positive rate, but one gender experiences a larger true positive rate and smaller false positive rate compared to the other gender at different levels, thus making the equalized odds ratio much smaller and making the AUC score relatively similar. The disparity in gender is further corroborated by the SHAP visual explainer plot specifically for the misclassified predictions of an applicant defaulting on their loan.



According to SHAP, the feature with the second highest importance on whether the model incorrectly predicts a person will default or not is their gender. The model likely saw that one gender defaulted more than the other gender, so gender started having an influence on the predictions. Gender influencing the prediction of a model inherently makes the model unfair, even more so when it plays a role in incorrectly classifying individuals. There is a third gender category, XAN, that consists of only 4 members. As there were so few people in this category, we did not perform any fairness checks on them. This limitation not only masks potential biases but also highlights the dataset's inadequacy in representing non-binary or gender-diverse individuals. Such underrepresentation can perpetuate existing societal biases and lacks transparency in the model's decision-making processes. We recommend expanding future data collection to include a more balanced representation of all gender identities, ensuring the automated decision system operates equitably across all demographic groups. Given these issues, deploying this ADS without addressing these fairness concerns could be problematic, especially in sectors requiring equitable treatment.

Summary

The data provided as input for this ADS is very appropriate. The data contains a lot of information about applicants and whether they ended up repaying loans or defaulting. The additional information from external datasets consisted of applicant history, which is also very appropriate. This is the same data that people in real life would use to determine whether to accept or reject a loan application. The model used for this ADS is relatively accurate, an AUC score of .0.79 is not bad at all. The usage of an AUC score benefits the banks and other loan lenders, whilst also benefiting applicants who will repay their loans. A higher AUC allows banks to trust that the ADS is accurate, while also accepting loan applicants who do intend to repay their loans. However, we do not believe the model to be fair. While the AUC scores for both genders are relatively similar, the equalized odds ratio is subpar. The model scored an EO ratio score of 0.57, which is very far from the ideal of 1.0. There are differing true positive and false positive rates between the genders, and the SHAP explanation plot illustrates that gender plays an important role into the decision making of the system, which is harmful to the people whom the model is used upon. No one wants to be treated differently solely because of their gender. As such, we would not be comfortable in deploying this ADS into the public sector. A model needs to be both accurate and fair to be worth deploying and using publicly. Improvements that we would suggest are for all three departments. There has been much useful data provided, albeit with a large number of missing values, and even external data sources have been used. However, there is very little representation of other gender types in the dataset, with the XAN category only having 4 members. There is an underrepresentation of this category of genders, which may lead to unfair treatment of this gender in particular. The model slightly lacks when it comes to accuracy and largely lacks when regarding fairness. There does not appear to have been any fairness analysis performed on this ADS, so we suggest changing analysis methodology to include fairness measures. Lastly, we would suggest incorporating less historical data into the ADS model. The goal of ADS is to help people with low to no credit history, yet the feature most important to the model when predicting who will repay or default on their loan is the data from external sources. The model relies heavily on a person's history when attempting to find an alternative to credit score, which increases as a client's history expands. The alternative this ADS presents is not much of an alternative to using credit score, but is rather based on the same features that affect a credit score.