Analysis on Automated Decision System(ADS) Created by "Yingdan Li"

on Predicting Home Credit Default Risk

Jack Jia (zj829) and Jiewu Rao (jr5512)

May 10th, 2023

**Please note that results may differ on different runs, all graphs and numbers included in this report is based on one particular execution conducted on May 10th, 2023, not based on numbers currently showing in the .ipynb file*

I. Purpose and background of the ADS

The original challenge that we have chosen to work on involves predicting the capability of clients to repay loans, to ensure that clients capable of repaying the loans are not rejected by loan offerings, vice versa. In other words, the system aims to improve efficiency and accuracy of assessing loan applicants. The ADS uses clients' transaction history and fundamental information to predict the ability of repayment.

II. Input and output

1. Data selection

Data comes from several different csv files. The application csv is broken into two parts, which are the train part and the test part. For this project, we will only use the train data and split train data into new train data and test data, since the test data file lacks a "target" variable to be compared to. The training and test data include basic personal information regarding each applicant, including gender, age, work information, location, normalized external data source, car ownerships, housing information, whether he/she has submitted certain documents, etc. The data are connected to different sub-datasets through SK_ID_CURR.

The sub-datasets originate from multiple CSV files, each containing different types of data. Descriptions regarding all the sub-datasets related to the original challenge are portrayed in the following few paragraphs. Information regarding datasets utilized by the ADS will be described in the next subsection of this report.

The bureau.csv file contains information on the previous credits reported to the Credit Bureau, including credits from other financial institutions for clients who have a loan in the sample. For every loan in the sample, there are multiple rows in the bureau.csv file, corresponding to the number of credits the client had in the Credit Bureau before the application date.

The bureau_balance.csv file provides monthly balances of the previous credits reported to the Credit Bureau. This table has one row for each month of history for every previous credit, resulting in a total of (#loans in sample * # of relative previous credits * # of months with observable history) rows.

The POS_CASH_balance.csv file contains monthly balance snapshots of previous point of sales (POS) and cash loans that the applicant had with Home Credit. It includes one row for each month of history for every previous credit related to loans in the sample, including consumer credit and cash loans. The total number of rows in this table is (#loans in sample * # of relative previous credits * # of months with observable history).

Similarly, the credit_card_balance.csv file provides monthly balance snapshots of previous credit cards that the applicant has with Home Credit. It includes one row for each month of history for every previous credit card related to loans in the sample, resulting in a total of (#loans in sample * # of relative previous credit cards * # of months with observable history) rows.

The previous_application.csv file contains information on all previous applications for Home Credit loans of clients who have loans in the sample. There is one row for each previous application related to loans in the data sample.

The installments_payments.csv file provides repayment history for previously disbursed credits in Home Credit related to the loans in the sample. It includes one row for every payment that was made, and one row for each missed payment. Each row represents either one payment of one installment or one installment corresponding to one payment of one previous Home Credit credit related to loans in the sample.

Finally, the HomeCredit_columns_description.csv file contains descriptions for the columns in the various data files, providing additional information about the data included in the dataset.

These data are given and collected by Home Credit, including telco and transactional information, to predict their clients' repayment abilities. These data are given in the kaggle competition already so there is no need to search for more information outside of the competition.

2. Basic information on data used by the ADS

In the **application_train.csv**, there are 122 columns and 307511 rows of different data. These data contain the general information/demographics of clients. Therefore, there are essentially three different types of data: integers, float, and string. The document contains 57 features with more than 2000 missing values.
*See Appendix - 1 for statistical information on application_train
*See Appendix - 2 for data types on application_train
*See Appendix - 3 for missing values > 2000 on application_train

The **credit_card_balance.csv** has 23 columns of different data. This file mainly contains the activities about clients' credit cards, such as the amount drawn during the month of the previous credit, the amount the client paid during the month on the previous credit, the total amount of receivable on the previous credit, etc. For this dataset, there are 9 features with missing values.

```
SK_ID_PREV                int64
SK_ID_CURR                int64
NUM_INSTALMENT_VERSION  float64
NUM_INSTALMENT_NUMBER     int64
DAYS_INSTALMENT         float64
DAYS_ENTRY_PAYMENT      float64
AMT_INSTALMENT          float64
AMT_PAYMENT             float64
DAYS_INSTALMENT_DIFF    float64
AMT_PATMENT_PCT         float64
dtype: object
```

*See Appendix - 4 for statistical information on credit_card_balance
*See Appendix - 5 for data types and missing values on credit_card_balance

The **installments_payments.csv** has 8 different columns of data. This file mainly contains information about clients' installment payments, such as the actual amount paid on previous credit on installments, which installment had payments, and the time that the installments were paid out. There are 2 features with missing values of 2905. Considering the total number of data entries, the number of missing values in installments_payments is rather acceptable. The on image on the right shows data types and missing values of the features.

```
DAYS_ENTRY_PAYMENT     2905
AMT_PAYMENT            2905
dtype: int64
```

*See Appendix - 6 for statistical information on installments_payments.

We suspect that the creator of the ADS utilizes these 2 subsets of data to reduce complexity, and increase efficiency and accuracy when making predictions, provided that these data subsets already contain sufficient attributes for predictions. Looking at the information of the datasets employed by the ADS, we see that the historical information regarding applicants of loans are rather comprehensive with detailed data on personal information, payments, installments, and credit cards. However, there are significantly more missing values in the training data file compared to credit card and installments/payments information. This occurrence is reasonable as the training data sets contain extremely detailed personal information that certain applicants would not be willing to share or are not required to share.

After checking on the pairwise correlation between attributes of the three files, there are generally low correlations among attributes, except for certain attributes such as AMT_PAYMENT and AMT_INSTALLMENTS, AMT_DRAWINGS_ATM_CURRENT and AMT_DRAWINGS_CURRENT, etc. These occurrences are highly related in the context of real life.

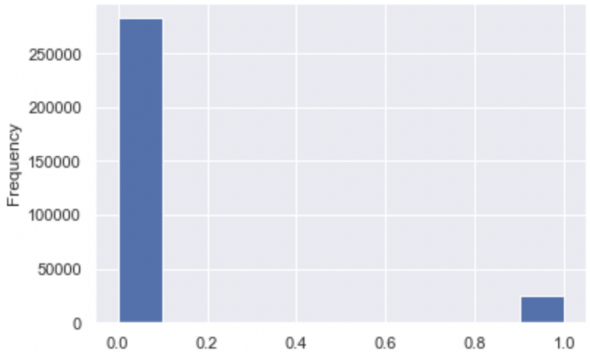
When visualizing the training data, see class 0 has a lot more records than class 1, creating an imbalance. When checking on the number of children, we can see that there are a number of outliers that shall be removed before applying the model.

3. Output of the system

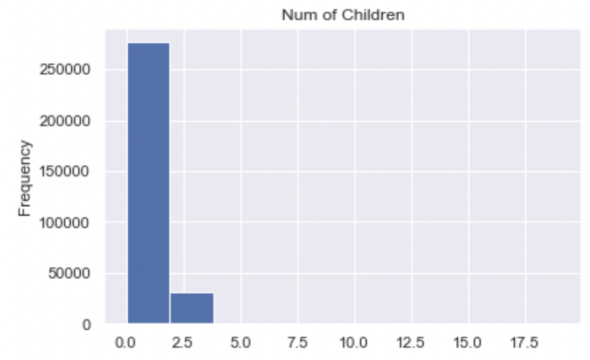
The system outputs the predicted results and probability through RandomTreeClassifier. Since the creator uses RandomTreeClassifier, the probability only outputs 1s and 0s as well. The creator checks the predicted values against the test values to find that the test group shows 80883 occurrences of 0s and 7199 occurrences of 1s, while the predicted results show 79954 counts of 0s and 8128 counts of 1s.

III. Implementation and Validation

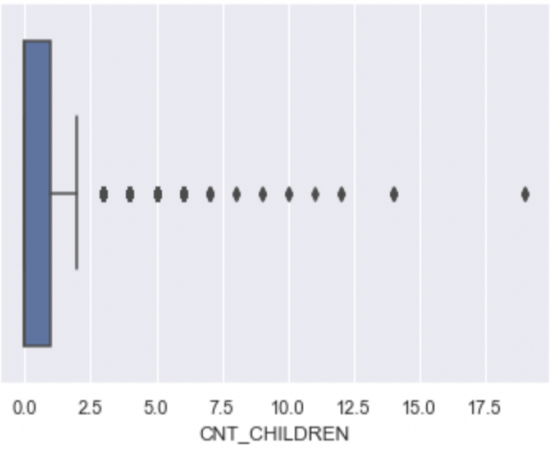
1. Data cleaning and preprocessing



The “target" feature in the application csv signifies the positive and negative classes for this paper, with “1”(positive) representing clients with payment difficulties, and “0”(negative) representing clients without payment difficulties.



The code first identifies and filters out columns or elements in the train dataset that have more than 2000 missing values. Then in order to know the data distribution better, the code plots a density distribution plot. From the plot, we could tell that there is more data in the 0 column, less in the 1 column. This means that the default rate on loan is about 8% in the given training dataset.When checking on the number of children, we can see that there are a number of outliers that shall be removed before applying the model. The mean is 0.417, and the 75% percentile is 1.



In the section "Preparing data for modeling”, the creator constructs new columns in the “Credit Card Balance” dataset that calculate the percentage of the credit limit used for different types of withdrawals and the percentage of the principal amount of the credit used for payments. Then, the code groups the data by 'SK_ID_CURR' and calculates the mean of various attributes such as average balance, average percentages of withdrawals, and the average number of different types of withdrawals.

The creator then builds two new columns in the 'Installments Payments' dataset that calculate the difference between the due date and the payment date and the percentage of the installment amount paid. Then, he/she groups the data by 'SK_ID_CURR' and calculates the mean of the two new attributes.

The resulting datasets, 'cc_use' and 'pmts_use', contain aggregated information about the credit card and installment payments of each individual, which are merged with the train dataset for the model.
*See Appendix - 7 for visualizations on other features

The author removes columns with missing values of more than 100000, outliers of individuals with more than 5 children, outliers of clients with more than 350000 of annual income. To further prepare the data, the creator fills in missing values with an extreme value of '-999', and converts data through one-hot encoding for categorical variables.

2. High-level information about the implementation of the system

The algorithm uses the joint dataset of the information processed above regarding the personal information and past credit history in binary form to predict whether an individual will have the ability to repay a loan or not; hence, Home Credit will be able to better decide if the firm should release the loan to the applicant or not. If the algorithm predicts an individual does not have the ability to repay a loan and the corresponding test data proves the same decision, the ADS has made a successful classification, vice versa.

3. How was the ADS validated

We would know how effective the model is by looking at its performance score(AUC, F1, Recall). This directly tells us how accurate the model is. Also, one can look at the ROC curve that the code runs to see how effective the model is. The stated goal is to predict if clients could repay the loan. If the model can classify positive and negative classes with a rather good accuracy (higher than one's expectation), the ADS meets its stated goal. The model also provides a rank of feature importance utilized by the system for a better illustration of the factors influencing the final results.

IV. Outcomes

1. Accuracy and Subpopulations

Please beware that in this analysis, "1"(positive) represents clients with payment difficulties, and "0"(negative) represents clients without payment difficulties. The ADS uses a decision tree model with 3 input files, with one including the test and train dataset. Firstly, we test only the accuracy of the original ADS which uses a decision tree classifier. We test on the subpopulations of gender, age, and whether the subject has children. We use AUC performance, F1 score, recall score to measure the accuracy. The AUC metric takes into account both TPR and FPR, which allows lenders to evaluate the trade-offs between correctly finding risky individuals and incorrectly rejecting worthy applicants. The F1 score is a metric that illustrates the balance

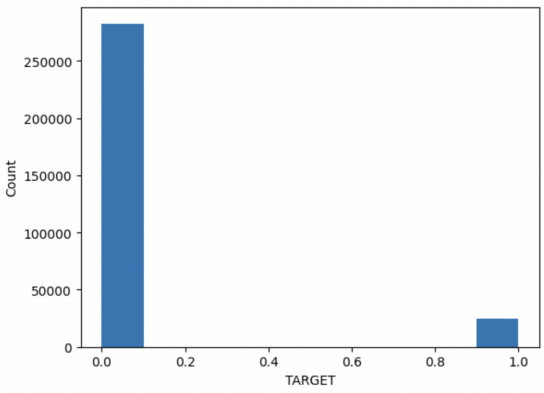
between precision and recall, allowing lenders to evaluate the trade-offs between minimizing false negatives (for instance, high risk applicants who are classified as creditworthy) and minimizing false positives. To ensure that high risk applicants are successfully identified, recall would help the lender to evaluate the performance.

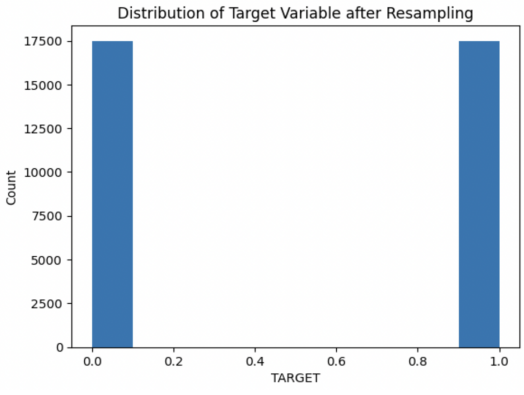
Original ADS:	Male Subpopulation v. Female:	
AUC Performance: 0.538	AUC Performance (Male): 0.537	(Female): 0.536
F1 Score: 0.150	F1 Score (Male): 0.170	(Female): 0.136
Recall Score: 0.853	Recall Score (Male): 0.824	(Female): 0.868

Age >=43(median) v. age < 43:	
AUC Performance (Old): 0.537	(Young): 0.539
F1 Score (Old): 0.170	(Young): 0.170
Recall Score (Old): 0.824	(Young): 0.827

With v. without children:	
AUC Performance (With Children): 0.540	(No Children): 0.536
F1 Score (With Children): 0.163	(No Children): 0.143
Recall Score (With Children): 0.839	(No Children): 0.858

Regarding subpopulations, we do not see big differences in accuracy scores in aspects of gender, age, and whether an applicant has children. However, overall, the ADS produces less than optimal AUC performance and F1 score with AUC performance scores around 0.5 and F1 scores between 0.13 to 0.17. To look for reasons behind the issue, we use different models and techniques. For instance, we try to improve the performance of the original ADS by finding the optimal parameters using RandomizedSearchCV. The outcome is AUC Performance: 0.684, F1 Score: 0.0, Recall Score: 0.921 Next, the new method trains a pruned Decision Tree model using a fixed ccp_alpha value = 0.00005. We find AUC Performance with pruning: 0.705, F1 Score with pruning: 0.0, Recall Score with pruning: 0.921. The two updated models perform better in terms of AUC Performance and Recall Score but have a worse F1 Score of 0. Upon checking, we found that these two models ignore any class 1. We also used oversampling and threshold adjustments, neither of which provide significant improvements.

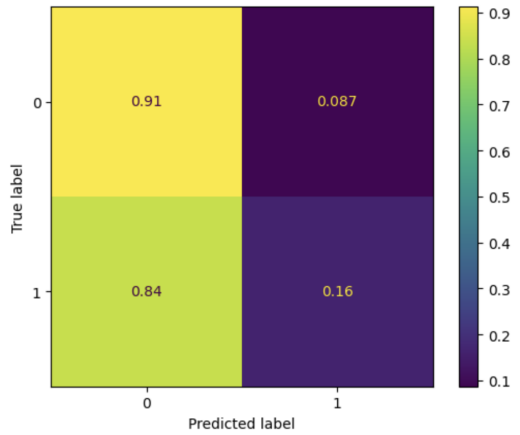




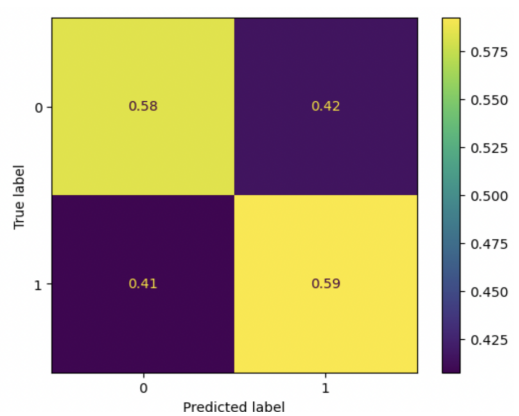
To look for reasons, we look for class imbalance and find significantly more class 0s than class 1s. We tried implementing undersampling to address the issue, and received results of AUC Performance: 0.588, F1 Score: 0.185, Recall Score: 0.584. Upon first glance, AUC performance increased by around 5%, F1 score increased by around 3%. However, the Recall Score decreased by around 24%.

To determine if undersampling resulted in a worthwhile tradeoff of improvement, we compared the confusion matrices.

The Original ADS:

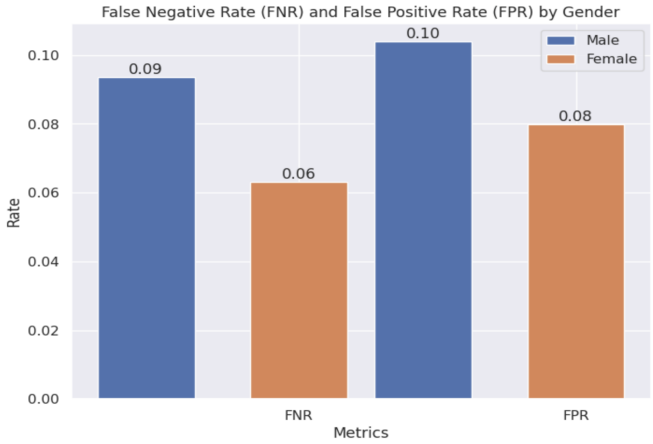


Undersampling:



Undersampling is conducted under the same feature engineering and pre-processing conditions. We see that when no class 1 or 0 is under-represented, the model produces high rates of false negative and false positive results, while the ADS produces significantly less false positives and true positives when no undersampling is conducted. The results indicate certain technical bias in the ADS; the model is likely influenced by the overwhelmingly larger proportion of class 0s that the ADS generates a strong inclination to predict the output as class 0, thus explaining the low AUC and F1 scores. In conclusion, the model is incapable of predicting the clients' capability to repay loans. The ADS fails to discern applicants who are able to repay loans as well as applicants who are unable to repay the loans.

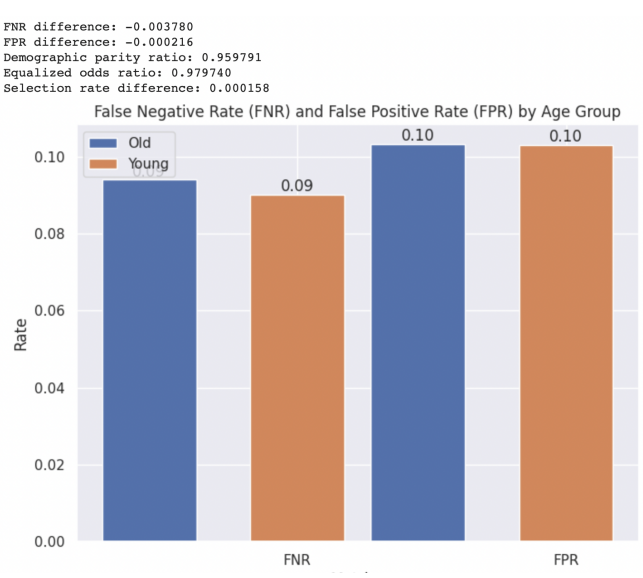
2. Fairness

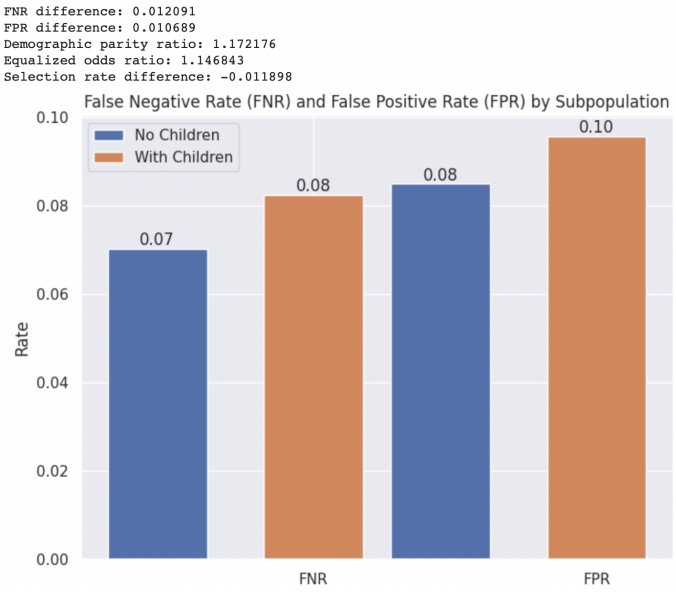


We analyzed fairness surrounding the sensitive features related to personal information(gender, age, whether one has children or not). We used metrics of FNR difference, FPR difference, demographic parity ratio, equalized odds ratio, and selection rate difference metrics for fairness analysis.

```
FNR difference: -0.030330
FPR difference: -0.024003
Demographic parity ratio: 0.675662
Equalized odds ratio: 0.724770
Selection rate difference: 0.026624
```

The model produces slightly higher FNR and FPR for males than females. Demographic parity ratio compares the proportion of positive outcomes for different demographic groups. In this case, the demographic parity ratio is rather high at 0.676, meaning that there's no significant bias. The model favors males slightly as more males receive positive outcomes than females. Equalized odds ratio measures whether the model has similar true positive rates and false positive rates across different groups. In this case, the equalized odds ratio is around 0.725, suggesting that the model is more likely to incorrectly predict a negative outcome for the female group. The selection rate difference is minimal. Overall, the model is slightly biased towards males based on the demographic parity ratio, equalized odds ratio, and selection rate difference; the model performs better with regards to correctly labeling female applicants.





With regards to the difference in age groups, the model produces minimal disparity with a low FNR difference, a low FPR difference, a demographic parity ratio close to 1, an equalized odds ratio close to 1, and a selection rate difference close to 0.

With regards to whether an applicant has children or not, the FNR and FPR differences are positive, indicating that the FNR and FPR for individuals with children is higher. The demographic parity ratio and equalized odds ratios are greater than 1, indicating that there is a bias in favor of individuals without children in the model. Overall, the model is slightly biased towards and slightly better at classifying applicants without children, considering selection rate difference is minimal.

3. Additional methods

Cross Validation:

To test the robustness of the ADS, we applied 5-fold cross validation through multiple splits of the data, to better estimate the model's ability to generalize to new, unseen data.

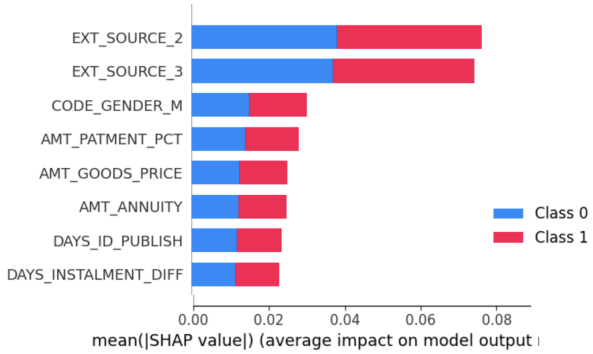
```
AUC Cross-validation scores:  [0.53771692 0.54000251 0.53839167 0.54126882 0.53827375]
Average AUC:  0.5391307338733664
```

The AUC scores are similar to the original model at around 0.53 to 0.54, indicating that the model is rather stable, but not necessarily robust, since AUC scores are around 0.5, meaning close to random.

Noise Injection:
We test the robustness of the original model by adding noise of different levels on it, and it shows that the ADS is pretty consistent in the performance metrics.

```
AUC Performance:  0.5359547344940291
F1 Score:  0.14673505128846637
Recall Score:  0.852104812393621
Noise level: 0.01
AUC Performance: 0.5346355946180555
F1 Score: 0.1456801470588235
Recall Score: 0.8387701781203586

Noise level: 0.05
AUC Performance: 0.5292554315481627
F1 Score: 0.14141320042852484
Recall Score: 0.8001647857243525

Noise level: 0.1
AUC Performance: 0.530695872871178
F1 Score: 0.1444458757353257
Recall Score: 0.7839897659392244

Noise level: 0.2
AUC Performance: 0.5274473095538511
F1 Score: 0.1427538713236693
Recall Score: 0.7557376871456294

Noise level: 0.5
AUC Performance: 0.5195669145438326
F1 Score: 0.14403618390375827
Recall Score: 0.6019774286922301
```

SHAP:



	Name	Importance
0	EXT_SOURCE_2	0.073561
1	EXT_SOURCE_3	0.055938
2	DAYS_BIRTH	0.046834
3	DAYS_REGISTRATION	0.046595
4	DAYS_INSTALMENT_DIFF	0.045476
5	DAYS_ID_PUBLISH	0.045212
6	AMT_ANNUITY	0.039208

We generated a summary plot that shows the impact of each feature through a random sample of 500 for efficiency. Comparing the SHAP plot against feature importance ranking, we conclude that EXT_SOURCE_2 AND EXT_SOURCE_3 are the main drivers of the decision making of the algorithm with significantly higher impact than other attributes. The two graphs are similar with slight differences as a result of the sampling process to create the SHAP plot. Each feature creates a similar impact on the two classes of the model's output, proving the conclusions we made before regarding the model's ineffectiveness. As a result, attributes such as gender which makes on average the third largest impact in this sample, or age(DAYS_BIRTH) which is ranked as third in importance, do not present significant disparities, even though they may seem important. However, as ethical data analysis practice, analysts should aim to remove sensitive features, like the two mentioned above, to avoid the algorithm from picking them up as key sources for classification.

LIME:

The graph above is generated through LIME regarding the instance with the highest predicted probability of belonging to the class 1. In this instance, the applicant is considered to have a high inability to repay. The factors contributed most to this decision is that the client isn't working for a cultural organization, the client is not an HR staff, and the client didn't submit document_7. The factors contributing most against the result is that the applicant is not working for an industry of type 6, and that he/she did not submit document_19.

The graph above is another instance. In this case, the applicant is considered as a target customer since he does not work for a housing organization, isn't an accountant, and has a normalized score of 0.72 from external data source 2. The attributes weighing against the decision are his numbers of draws at ATM during this month on the previous credit as well as his annuity of previous application. In this instance, the factors weighing against the classification are reasonable practically.

Other Models:

We then use the random forest model based on the current data imported. We see an improvement in the accuracy scores, especially in the AUC performance. The accuracy of the original ADS is below:

AUC Performance: 0.718 F1 Score: 0.000819 Recall Score: 0.921

In summary, the Random Forest model seems to perform better overall in terms of AUC, Recall. However, the Decision Tree model has a higher F1 score. This suggests that the Random Forest model may be a better choice for this particular task, but the low F1 score for both models indicates that there might be room for improvement in terms of identifying applicants correctly identified as risky(precision) and considering the high recall score.

Now using the random forest model, test the subpopulation accuracy metrics:

For age >= 43: AUC Performance: 0.693 F1 Score: 0.0 Recall Score: 0.937

For age < 43: AUC Performance: 0.706 F1 Score: 0.00218 Recall Score: 0.901

For gender:
AUC Performance(male): 0.707 F1 Score(male): 0.00185
Recall Score(male): 0.897
AUC Performance(female): 0.698 F1 Score(female): 0.000468
Recall Score(female): 0.930

Overall, the random forest model appears to perform better than the decision tree model in terms of AUC performance and recall score for both age groups. One reason why the decision tree model might have performed better in this case is that it can handle nonlinear relationships between the input features and the target variable better than the random forest model. This is because a decision tree can split the feature space into more complex regions than a random forest model.

Beyond the decision tree model that was implemented by the creator originally, we used the logistic model, LightGBM, and XGBoost model to fit the data.

Logistic Regression:	LightGBM:	XGBoost:
AUC Performance: 0.632	AUC Performance: 0.759	AUC Performance: 0.753
F1 Score: 0.0	F1 Score: 0.0363	F1 Score: 0.0668
Recall Score: 0.921	Recall Score: 0.921	Recall Score: 0.920

Compared to the base model, all three models (Logistic Regression, LightGBM, and XGBoost) have significantly better AUC performance and recall score, with varying degrees of reduction in F1 scores. Among the three models, LightGBM has the highest AUC performance, followed closely by XGBoost, while Logistic Regression has the lowest AUC performance. However, Logistic Regression has the highest recall score, indicating that it has the lowest number of false negatives compared to the other models. Logistic regression is a linear model that assumes a linear relationship between the features and the log-odds of the target variable, while LightGBM and XGBoost are tree-based models that can capture more complex nonlinear relationships. This flexibility in modeling nonlinear relationships can lead to improved performance in this context of classification problems.

V. Summary and Reflection

1. Data collection

At the first glance, the data was appropriate, as they are collected in an unbiased way. The data was treated in the pre-processing stage comprehensively through removing outliers, combining attributes, etc. The size of the data was also considerable, even in the case of using the training dataset only.

However, after closely examining data, we found that the data is actually class imbalanced. This means that there are more cases of 0s than 1s. Based on the interpretation, 1 - client with

payment difficulties: he/she had late payment more than X days on at least one of the first Y installments of the loan in our sample, 0 - all other cases. This explanation makes us less worried about the imbalance. The data itself is not that biased, but the model fails to discern applicants who are able to repay loans as well as applicants who are unable to repay the loans. I would say this arises from both the imbalanced data and the design of the ADS.

Besides these, the data amount is massive, as there are hundreds of thousands of rows and columns in these files. All in all, the data is appropriate, though not perfect for this ADS.

2. Implementation

The AUC scores are similar to the original model at around 0.53 to 0.54. In this perspective, the implementation is rather stable, but not necessarily robust. Since the AUC score is around 0.5, similar to random chance, along with a low F1 score, the model proves to be rather ineffective when applied to the dataset when identifying applicants correctly identified as risky. The model proves a trade-off between precision and recall; the algorithm sacrifices accuracy of positive predictions for completeness of positive predictions. Moreover, results of noise injection show that the ADS is consistent in its performance, though the values of the performance metrics are still not ideal. After trying different models, we find that most of the models can do better than the original ADS in terms of AUC score. However, most models exacerbate the trade-off between precision and recall, producing high recall values and low F1 scores. This is likely the result of an imbalanced data set, since F1 score increased after undersampling, as well as AUC.

In terms of fairness, when judging by different subpopulations divided by sensitive features, despite the differences in fairness being small, the overall fairness is achieved in each case. All the fairness metrics illustrated relatively low disparity among the subpopulations. Selection rate differences are minimal in all of the three analyses despite differences in the number of data entries.

3. Deployment

The metrics used to analyze fairness and accuracy are suitable for judging systems to be deployed in the sector of classifying individual's ability for loan repayment. For instance, A high AUC score indicates that the model is better at distinguishing between the two classes, which is important for identifying individuals who are likely to default on the loan. The F1 score is a useful metric for evaluating the balance between precision and recall. In the context of loan qualification, precision would measure the individuals who are actually unable to repay the loan among those who are predicted to be unable to repay, while recall would measure the proportion of individuals who are predicted to be unable to repay the loan among all those who are actually unable to repay. However, despite the metrics being suitable, the specific ADS is not accurate

enough to be deployed in the market, especially regarding qualifications for loans, which may lead to irrevocable consequences for Home Credit as a result of inaccurate classifications.

When it comes to the public sector, the data in this project, which is the sensitive information of real clients, needs to be clearly protected. In this case, the original data could be used directly in the public sector or in the industry. However, if one does so, it would leak the information of these clients. The data contains large amounts of information on different aspects for applicants. Therefore, even when intentionally blocking certain attributes or groups, differential privacy is unlikely to be achieved. One way to prevent this is through synthetic data, which simply generates fake data and uses it on the model in order to prevent information leaking.

In this case, this ADS did not use all of the input files. The files it used are credit card balance file, and installments_payments files. There are four more files that contain personal information that the ADS did not use. The feature engineering part calculates various percentages related to drawing amounts, principal receivable, and payment amounts. It also calculates the differences in days between installment and entry payment. The results are aggregated by the 'SK_ID_CURR' column, creating two new dataframes, cc_use and pmts_use, with mean values of these calculated metrics. However, this process did not directly help with data privacy and data protection. It focused on calculating and aggregating specific features and metrics for further analysis. To protect data privacy and ensure data protection, other techniques such as anonymization, pseudonymization, encryption, and differential privacy should be applied when handling sensitive information. This was not seen in the original ADS. Therefore, although this ADS has transformed and engineered the inputs in various ways, I still would not be comfortable with deploying this ADS in public.

4. Improvements

If we have test data in the project, we could have a better view on if the model would underfit or overfit. In the data preprocessing methods, checking for and removing certain highly correlated attributes could improve the model since the data unavoidably would have certain features with extremely high correlations due to practicality. Moreover, the program only builds the decision tree model but nothing else. Plus, the model did not find its optimal parameters.

Furthermore, the data has lots of NaN and missing values. The cleaning process may potentially harm the original data. The feature engineering also makes many adjustments before the data is used. Though it would be hard to achieve, especially considering the magnitude of the data, it could be better if data is more organized and contains less missing values through more detailed preprocessing and feature engineering to manage each column.

In terms of accuracy, we would not use a decision tree model if we were going to really take part in this competition. Instead, we would compare with other algorithms such as a random forest model, lightbgm model, xgboost, or others.

Finally, the original ADS only uses 4 input files, but in fact there are 8 to use. Therefore, to increase the accuracy of ADS, including more input files might yield better results since the ADS might have missed certain attributes of determining importance in other files.

References

"Home Credit Default Risk." *Kaggle*, www.kaggle.com/competitions/home-credit-default-risk. Accessed 10 May 2023.

yingdanli49. "Predict Home Credit Default Risk." *Kaggle*, 13 Dec. 2021, www.kaggle.com/code/yingdanli49/predict-home-credit-default-risk/notebook.

Appendix - 1

	SK_ID_CURR	TARGET	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE
count	307511.000000	307511.000000	307511.000000	3.075110e+05	3.075110e+05	307499.000000	3.072330e+05
mean	278180.518577	0.080729	0.417052	1.687979e+05	5.990260e+05	27108.573909	5.383962e+05
std	102790.175348	0.272419	0.722121	2.371231e+05	4.024908e+05	14493.737315	3.694465e+05
min	100002.000000	0.000000	0.000000	2.565000e+04	4.500000e+04	1615.500000	4.050000e+04
25%	189145.500000	0.000000	0.000000	1.125000e+05	2.700000e+05	16524.000000	2.385000e+05
50%	278202.000000	0.000000	0.000000	1.471500e+05	5.135310e+05	24903.000000	4.500000e+05
75%	367142.500000	0.000000	1.000000	2.025000e+05	8.086500e+05	34596.000000	6.795000e+05
max	456255.000000	1.000000	19.000000	1.170000e+08	4.050000e+06	258025.500000	4.050000e+06

REGION_POPULATION_RELATIVE	DAYS_BIRTH	DAYS_EMPLOYED	DAYS_REGISTRATION	DAYS_ID_PUBLISH	OWN_CAR_AGE	FLAG_MOBIL
307511.000000	307511.000000	307511.000000	307511.000000	307511.000000	104582.000000	307511.000000
0.020868	-16036.995067	63815.045904	-4986.120328	-2994.202373	12.061091	0.999997
0.013831	4363.988632	141275.766519	3522.886321	1509.450419	11.944812	0.001803
0.000290	-25229.000000	-17912.000000	-24672.000000	-7197.000000	0.000000	0.000000
0.010006	-19682.000000	-2760.000000	-7479.500000	-4299.000000	5.000000	1.000000
0.018850	-15750.000000	-1213.000000	-4504.000000	-3254.000000	9.000000	1.000000
0.028663	-12413.000000	-289.000000	-2010.000000	-1720.000000	15.000000	1.000000
0.072508	-7489.000000	365243.000000	0.000000	0.000000	91.000000	1.000000

FLAG_EMP_PHONE	FLAG_WORK_PHONE	FLAG_CONT_MOBILE	FLAG_PHONE	FLAG_EMAIL	CNT_FAM_MEMBERS	REGION_RATING_CLIENT	I
307511.000000	307511.000000	307511.000000	307511.000000	307511.000000	307509.000000	307511.000000	
0.819889	0.199368	0.998133	0.281066	0.056720	2.152665	2.052463	
0.384280	0.399526	0.043164	0.449521	0.231307	0.910682	0.509034	
0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	1.000000	
1.000000	0.000000	1.000000	0.000000	0.000000	2.000000	2.000000	
1.000000	0.000000	1.000000	0.000000	0.000000	2.000000	2.000000	
1.000000	0.000000	1.000000	1.000000	0.000000	3.000000	2.000000	
1.000000	1.000000	1.000000	1.000000	1.000000	20.000000	3.000000	

1

REGION_RATING_CLIENT_W_CITY	HOUR_APPR_PROCESS_START	REG_REGION_NOT_LIVE_REGION	REG_REGION_NOT_WORK_REGION	I
307511.000000	307511.000000	307511.000000	307511.000000	
2.031521	12.063419	0.015144	0.050769	
0.502737	3.265832	0.122126	0.219526	
1.000000	0.000000	0.000000	0.000000	
2.000000	10.000000	0.000000	0.000000	
2.000000	12.000000	0.000000	0.000000	
2.000000	14.000000	0.000000	0.000000	
3.000000	23.000000	1.000000	1.000000	

LIVE_REGION_NOT_WORK_REGION	REG_CITY_NOT_LIVE_CITY	REG_CITY_NOT_WORK_CITY	LIVE_CITY_NOT_WORK_CITY	EXT_SOURCE_1	EXT_SOURCE_2
307511.000000	307511.000000	307511.000000	307511.000000	134133.000000	3.068510e+05
0.040659	0.078173	0.230454	0.179555	0.502130	5.143927e-01
0.197499	0.268444	0.421124	0.383817	0.211062	1.910602e-01
0.000000	0.000000	0.000000	0.000000	0.014568	8.173617e-08
0.000000	0.000000	0.000000	0.000000	0.334007	3.924574e-01
0.000000	0.000000	0.000000	0.000000	0.505998	5.659614e-01
0.000000	0.000000	0.000000	0.000000	0.675053	6.636171e-01
1.000000	1.000000	1.000000	1.000000	0.962693	8.549997e-01

EXT_SOURCE_3	APARTMENTS_AVG	BASEMENTAREA_AVG	YEARS_BEGINEXPLUATATION_AVG	YEARS_BUILD_AVG	COMMONAREA_AVG	ELEVATORS_AVG
246546.000000	151450.00000	127568.000000	157504.000000	103023.000000	92646.000000	143620.000000
0.510853	0.11744	0.088442	0.977735	0.752471	0.044621	0.078942
0.194844	0.10824	0.082438	0.059223	0.113280	0.076036	0.134576
0.000527	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000
0.370650	0.05770	0.044200	0.976700	0.687200	0.007800	0.000000
0.535276	0.08760	0.076300	0.981600	0.755200	0.021100	0.000000
0.669057	0.14850	0.112200	0.986600	0.823200	0.051500	0.120000
0.896010	1.00000	1.000000	1.000000	1.000000	1.000000	1.000000

ENTRANCES_AVG	FLOORSMAX_AVG	FLOORSMIN_AVG	LANDAREA_AVG	LIVINGAPARTMENTS_AVG	LIVINGAREA_AVG	NONLIVINGAPARTMENTS_AVG
152683.000000	154491.000000	98869.000000	124921.000000	97312.000000	153161.000000	93997.000000
0.149725	0.226282	0.231894	0.066333	0.100775	0.107399	0.008809
0.100049	0.144641	0.161380	0.081184	0.092576	0.110565	0.047732
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.069000	0.166700	0.083300	0.018700	0.050400	0.045300	0.000000
0.137900	0.166700	0.208300	0.048100	0.075600	0.074500	0.000000
0.206900	0.333300	0.375000	0.085600	0.121000	0.129900	0.003900
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

NONLIVINGAREA_AVG	APARTMENTS_MODE	BASEMENTAREA_MODE	YEARS_BEGINEXPLUATATION_MODE	YEARS_BUILD_MODE	COMMONAREA_MODE
137829.000000	151450.000000	127568.000000	157504.000000	103023.000000	92646.000000
0.028358	0.114231	0.087543	0.977065	0.759637	0.042553
0.069523	0.107936	0.084307	0.064575	0.110111	0.074445
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.052500	0.040700	0.976700	0.699400	0.007200
0.003600	0.084000	0.074600	0.981600	0.764800	0.019000
0.027700	0.143900	0.112400	0.986600	0.823600	0.049000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

ELEVATORS_MODE	ENTRANCES_MODE	FLOORSMAX_MODE	FLOORSMIN_MODE	LANDAREA_MODE	LIVINGAPARTMENTS_MODE	LIVINGAREA_MODE
143620.000000	152683.000000	154491.000000	98869.000000	124921.000000	97312.000000	153161.000000
0.074490	0.145193	0.222315	0.228058	0.064958	0.105645	0.105975
0.132256	0.100977	0.143709	0.161160	0.081750	0.097880	0.111845
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.069000	0.166700	0.083300	0.016600	0.054200	0.042700
0.000000	0.137900	0.166700	0.208300	0.045800	0.077100	0.073100
0.120800	0.206900	0.333300	0.375000	0.084100	0.131300	0.125200
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

NONLIVINGAPARTMENTS_MODE	NONLIVINGAREA_MODE	APARTMENTS_MEDI	BASEMENTAREA_MEDI	YEARS_BEGINEXPLUATATION_MEDI
93997.000000	137829.000000	151450.000000	127568.000000	157504.000000
0.008076	0.027022	0.117850	0.087955	0.977752
0.046276	0.070254	0.109076	0.082179	0.059897
0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.058300	0.043700	0.976700
0.000000	0.001100	0.086400	0.075800	0.981600
0.003900	0.023100	0.148900	0.111600	0.986600
1.000000	1.000000	1.000000	1.000000	1.000000

YEARS_BUILD_MEDI	COMMONAREA_MEDI	ELEVATORS_MEDI	ENTRANCES_MEDI	FLOORSMAX_MEDI	FLOORSMIN_MEDI	LANDAREA_MEDI
103023.000000	92646.000000	143620.000000	152683.000000	154491.000000	98869.000000	124921.000000
0.755746	0.044595	0.078078	0.149213	0.225897	0.231625	0.067169
0.112066	0.076144	0.134467	0.100368	0.145067	0.161934	0.082167
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.691400	0.007900	0.000000	0.069000	0.166700	0.083300	0.018700
0.758500	0.020800	0.000000	0.137900	0.166700	0.208300	0.048700
0.825600	0.051300	0.120000	0.206900	0.333300	0.375000	0.086800
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

LIVINGAPARTMENTS_MEDI	LIVINGAREA_MEDI	NONLIVINGAPARTMENTS_MEDI	NONLIVINGAREA_MEDI	TOTALAREA_MODE	OBS_30_CNT_SOCIAL_CIRCLE
97312.000000	153161.000000	93997.000000	137829.000000	159080.000000	306490.000000
0.101954	0.108607	0.008651	0.028236	0.102547	1.422245
0.093642	0.112260	0.047415	0.070166	0.107462	2.400989
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.051300	0.045700	0.000000	0.000000	0.041200	0.000000
0.076100	0.074900	0.000000	0.003100	0.068800	0.000000
0.123100	0.130300	0.003900	0.026600	0.127600	2.000000
1.000000	1.000000	1.000000	1.000000	1.000000	348.000000

DEF_30_CNT_SOCIAL_CIRCLE	OBS_60_CNT_SOCIAL_CIRCLE	DEF_60_CNT_SOCIAL_CIRCLE	DAYS_LAST_PHONE_CHANGE	FLAG_DOCUMENT_2
306490.000000	306490.000000	306490.000000	307510.000000	307511.000000
0.143421	1.405292	0.100049	-962.858788	0.000042
0.446698	2.379803	0.362291	826.808487	0.006502
0.000000	0.000000	0.000000	-4292.000000	0.000000
0.000000	0.000000	0.000000	-1570.000000	0.000000
0.000000	0.000000	0.000000	-757.000000	0.000000
0.000000	2.000000	0.000000	-274.000000	0.000000
34.000000	344.000000	24.000000	0.000000	1.000000

FLAG_DOCUMENT_3	FLAG_DOCUMENT_4	FLAG_DOCUMENT_5	FLAG_DOCUMENT_6	FLAG_DOCUMENT_7	FLAG_DOCUMENT_8	FLAG_DOCUMENT_9
307511.000000	307511.000000	307511.000000	307511.000000	307511.000000	307511.000000	307511.000000
0.710023	0.000081	0.015115	0.088055	0.000192	0.081376	0.003896
0.453752	0.009016	0.122010	0.283376	0.013850	0.273412	0.062295
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

FLAG_DOCUMENT_10	FLAG_DOCUMENT_11	FLAG_DOCUMENT_12	FLAG_DOCUMENT_13	FLAG_DOCUMENT_14	FLAG_DOCUMENT_15	FLAG_DOCUMENT_16
307511.000000	307511.000000	307511.000000	307511.000000	307511.000000	307511.00000	307511.000000
0.000023	0.003912	0.000007	0.003525	0.002936	0.00121	0.009928
0.004771	0.062424	0.002550	0.059268	0.054110	0.03476	0.099144
0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000
1.000000	1.000000	1.000000	1.000000	1.000000	1.00000	1.000000

FLAG_DOCUMENT_17	FLAG_DOCUMENT_18	FLAG_DOCUMENT_19	FLAG_DOCUMENT_20	FLAG_DOCUMENT_21	AMT_REQ_CREDIT_BUREAU_HOUR
307511.000000	307511.000000	307511.000000	307511.000000	307511.000000	265992.000000
0.000267	0.008130	0.000595	0.000507	0.000335	0.006402
0.016327	0.089798	0.024387	0.022518	0.018299	0.083849
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
1.000000	1.000000	1.000000	1.000000	1.000000	4.000000

AMT_REQ_CREDIT_BUREAU_DAY	AMT_REQ_CREDIT_BUREAU_WEEK	AMT_REQ_CREDIT_BUREAU_MON	AMT_REQ_CREDIT_BUREAU_QRT
265992.000000	265992.000000	265992.000000	265992.000000
0.007000	0.034362	0.267395	0.265474
0.110757	0.204685	0.916002	0.794056
0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000
9.000000	8.000000	27.000000	261.000000

AMT_REQ_CREDIT_BUREAU_YEAR
265992.000000
1.899974
1.869295
0.000000
0.000000
1.000000
3.000000
25.000000

Appendix - 2

SK_ID_CURR	int64
TARGET	int64
NAME_CONTRACT_TYPE	object
CODE_GENDER	object
FLAG_OWN_CAR	object
FLAG_OWN_REALTY	object
CNT_CHILDREN	int64
AMT_INCOME_TOTAL	float64
AMT_CREDIT	float64
AMT_ANNUITY	float64
AMT_GOODS_PRICE	float64
NAME_TYPE_SUITE	object
NAME_INCOME_TYPE	object
NAME_EDUCATION_TYPE	object
NAME_FAMILY_STATUS	object
NAME_HOUSING_TYPE	object
REGION_POPULATION_RELATIVE	float64
DAYS_BIRTH	int64
DAYS_EMPLOYED	int64

DAYS_REGISTRATION	float64
DAYS_ID_PUBLISH	int64
OWN_CAR_AGE	float64
FLAG_MOBIL	int64
FLAG_EMP_PHONE	int64
FLAG_WORK_PHONE	int64
FLAG_CONT_MOBILE	int64
FLAG_PHONE	int64
FLAG_EMAIL	int64
OCCUPATION_TYPE	object
CNT_FAM_MEMBERS	float64
REGION_RATING_CLIENT	int64
REGION_RATING_CLIENT_W_CITY	int64
WEEKDAY_APPR_PROCESS_START	object
HOUR_APPR_PROCESS_START	int64
REG_REGION_NOT_LIVE_REGION	int64
REG_REGION_NOT_WORK_REGION	int64
LIVE_REGION_NOT_WORK_REGION	int64
REG_CITY_NOT_LIVE_CITY	int64

REG_CITY_NOT_WORK_CITY	int64
LIVE_CITY_NOT_WORK_CITY	int64
ORGANIZATION_TYPE	object
EXT_SOURCE_1	float64
EXT_SOURCE_2	float64
EXT_SOURCE_3	float64
APARTMENTS_AVG	float64
BASEMENTAREA_AVG	float64
YEARS_BEGINEXPLUATATION_AVG	float64
YEARS_BUILD_AVG	float64
COMMONAREA_AVG	float64
ELEVATORS_AVG	float64
ENTRANCES_AVG	float64
FLOORSMAX_AVG	float64
FLOORSMIN_AVG	float64
LANDAREA_AVG	float64
LIVINGAPARTMENTS_AVG	float64
LIVINGAREA_AVG	float64
NONLIVINGAPARTMENTS_AVG	float64

NONLIVINGAREA_AVG	float64
APARTMENTS_MODE	float64
BASEMENTAREA_MODE	float64
YEARS_BEGINEXPLUATATION_MODE	float64
YEARS_BUILD_MODE	float64
COMMONAREA_MODE	float64
ELEVATORS_MODE	float64
ENTRANCES_MODE	float64
FLOORSMAX_MODE	float64
FLOORSMIN_MODE	float64
LANDAREA_MODE	float64
LIVINGAPARTMENTS_MODE	float64
LIVINGAREA_MODE	float64
NONLIVINGAPARTMENTS_MODE	float64
NONLIVINGAREA_MODE	float64
APARTMENTS_MEDI	float64
BASEMENTAREA_MEDI	float64
YEARS_BEGINEXPLUATATION_MEDI	float64
YEARS_BUILD_MEDI	float64

COMMONAREA_MEDI	float64
ELEVATORS_MEDI	float64
ENTRANCES_MEDI	float64
FLOORSMAX_MEDI	float64
FLOORSMIN_MEDI	float64
LANDAREA_MEDI	float64
LIVINGAPARTMENTS_MEDI	float64
LIVINGAREA_MEDI	float64
NONLIVINGAPARTMENTS_MEDI	float64
NONLIVINGAREA_MEDI	float64
FONDKAPREMONT_MODE	object
HOUSETYPE_MODE	object
TOTALAREA_MODE	float64
WALLSMATERIAL_MODE	object
EMERGENCYSTATE_MODE	object
OBS_30_CNT_SOCIAL_CIRCLE	float64
DEF_30_CNT_SOCIAL_CIRCLE	float64
OBS_60_CNT_SOCIAL_CIRCLE	float64
DEF_60_CNT_SOCIAL_CIRCLE	float64

DAYS_LAST_PHONE_CHANGE	float64
FLAG_DOCUMENT_2	int64
FLAG_DOCUMENT_3	int64
FLAG_DOCUMENT_4	int64
FLAG_DOCUMENT_5	int64
FLAG_DOCUMENT_6	int64
FLAG_DOCUMENT_7	int64
FLAG_DOCUMENT_8	int64
FLAG_DOCUMENT_9	int64
FLAG_DOCUMENT_10	int64
FLAG_DOCUMENT_11	int64
FLAG_DOCUMENT_12	int64
FLAG_DOCUMENT_13	int64
FLAG_DOCUMENT_14	int64
FLAG_DOCUMENT_15	int64
FLAG_DOCUMENT_16	int64
FLAG_DOCUMENT_17	int64
FLAG_DOCUMENT_18	int64
FLAG_DOCUMENT_19	int64

```
FLAG_DOCUMENT_20                int64
FLAG_DOCUMENT_21                int64
AMT_REQ_CREDIT_BUREAU_HOUR    float64
AMT_REQ_CREDIT_BUREAU_DAY     float64
AMT_REQ_CREDIT_BUREAU_WEEK    float64
AMT_REQ_CREDIT_BUREAU_MON     float64
AMT_REQ_CREDIT_BUREAU_QRT     float64
AMT_REQ_CREDIT_BUREAU_YEAR    float64
dtype: object
```

Appendix - 3

OWN_CAR_AGE	202929
OCCUPATION_TYPE	96391
EXT_SOURCE_1	173378
EXT_SOURCE_3	60965
APARTMENTS_AVG	156061
BASEMENTAREA_AVG	179943
YEARS_BEGINEXPLUATATION_AVG	150007
YEARS_BUILD_AVG	204488
COMMONAREA_AVG	214865
ELEVATORS_AVG	163891
ENTRANCES_AVG	154828
FLOORSMAX_AVG	153020
FLOORSMIN_AVG	208642
LANDAREA_AVG	182590
LIVINGAPARTMENTS_AVG	210199
LIVINGAREA_AVG	154350
NONLIVINGAPARTMENTS_AVG	213514
NONLIVINGAREA_AVG	169682
APARTMENTS_MODE	156061
BASEMENTAREA_MODE	179943
YEARS_BEGINEXPLUATATION_MODE	150007
YEARS_BUILD_MODE	204488
COMMONAREA_MODE	214865
ELEVATORS_MODE	163891
ENTRANCES_MODE	154828
FLOORSMAX_MODE	153020
FLOORSMIN_MODE	208642
LANDAREA_MODE	182590
LIVINGAPARTMENTS_MODE	210199
LIVINGAREA_MODE	154350
NONLIVINGAPARTMENTS_MODE	213514
NONLIVINGAREA_MODE	169682
APARTMENTS_MEDI	156061
BASEMENTAREA_MEDI	179943
YEARS_BEGINEXPLUATATION_MEDI	150007
YEARS_BUILD_MEDI	204488
COMMONAREA_MEDI	214865
ELEVATORS_MEDI	163891
ENTRANCES_MEDI	154828
FLOORSMAX_MEDI	153020
FLOORSMIN_MEDI	208642
LANDAREA_MEDI	182590
LIVINGAPARTMENTS_MEDI	210199
LIVINGAREA_MEDI	154350
NONLIVINGAPARTMENTS_MEDI	213514
NONLIVINGAREA_MEDI	169682
FONDKAPREMONT_MODE	210295
HOUSETYPE_MODE	154297
TOTALAREA_MODE	148431
WALLSMATERIAL_MODE	156341
EMERGENCYSTATE_MODE	145755
AMT_REQ_CREDIT_BUREAU_HOUR	41519
AMT_REQ_CREDIT_BUREAU_DAY	41519
AMT_REQ_CREDIT_BUREAU_WEEK	41519
AMT_REQ_CREDIT_BUREAU_MON	41519
AMT_REQ_CREDIT_BUREAU_QRT	41519
AMT_REQ_CREDIT_BUREAU_YEAR	41519

Appendix - 4

	SK_ID_PREV	SK_ID_CURR	MONTHS_BALANCE	AMT_BALANCE	AMT_CREDIT_LIMIT_ACTUAL	AMT_DRAWINGS_ATM_CURRENT
count	3.840312e+06	3.840312e+06	3.840312e+06	3.840312e+06	3.840312e+06	3.090496e+06
mean	1.904504e+06	2.783242e+05	-3.452192e+01	5.830016e+04	1.538080e+05	5.961325e+03
std	5.364695e+05	1.027045e+05	2.666775e+01	1.063070e+05	1.651457e+05	2.822569e+04
min	1.000018e+06	1.000060e+05	-9.600000e+01	-4.202502e+05	0.000000e+00	-6.827310e+03
25%	1.434385e+06	1.895170e+05	-5.500000e+01	0.000000e+00	4.500000e+04	0.000000e+00
50%	1.897122e+06	2.783960e+05	-2.800000e+01	0.000000e+00	1.125000e+05	0.000000e+00
75%	2.369328e+06	3.675800e+05	-1.100000e+01	8.904669e+04	1.800000e+05	0.000000e+00
max	2.843496e+06	4.562500e+05	-1.000000e+00	1.505902e+06	1.350000e+06	2.115000e+06

AMT_DRAWINGS_CURRENT	AMT_DRAWINGS_OTHER_CURRENT	AMT_DRAWINGS_POS_CURRENT	AMT_INST_MIN_REGULARITY	AMT_PAYMENT_CURRENT
3.840312e+06	3.090496e+06	3.090496e+06	3.535076e+06	3.072324e+06
7.433388e+03	2.881696e+02	2.968805e+03	3.540204e+03	1.028054e+04
3.384608e+04	8.201989e+03	2.079689e+04	5.600154e+03	3.607808e+04
-6.211620e+03	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.523700e+02
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.702700e+03
0.000000e+00	0.000000e+00	0.000000e+00	6.633911e+03	9.000000e+03
2.287098e+06	1.529847e+06	2.239274e+06	2.028820e+05	4.289207e+06

AMT_PAYMENT_TOTAL_CURRENT	AMT_RECEIVABLE_PRINCIPAL	AMT_RECIVABLE	AMT_TOTAL_RECEIVABLE	CNT_DRAWINGS_ATM_CURRENT
3.840312e+06	3.840312e+06	3.840312e+06	3.840312e+06	3.090496e+06
7.588857e+03	5.596588e+04	5.808881e+04	5.809829e+04	3.094490e-01
3.200599e+04	1.025336e+05	1.059654e+05	1.059718e+05	1.100401e+00
0.000000e+00	-4.233058e+05	-4.202502e+05	-4.202502e+05	0.000000e+00
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
6.750000e+03	8.535924e+04	8.889949e+04	8.891451e+04	0.000000e+00
4.278316e+06	1.472317e+06	1.493338e+06	1.493338e+06	5.100000e+01

CNT_DRAWINGS_CURRENT	CNT_DRAWINGS_OTHER_CURRENT	CNT_DRAWINGS_POS_CURRENT	CNT_INSTALMENT_MATURE_CUM	SK_DPD	SK_DPD_DEF
3.840312e+06	3.090496e+06	3.090496e+06	3.535076e+06	3.840312e+06	3.840312e+06
7.031439e-01	4.812496e-03	5.594791e-01	2.082508e+01	9.283667e+00	3.316220e-01
3.190347e+00	8.263861e-02	3.240649e+00	2.005149e+01	9.751570e+01	2.147923e+01
0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
0.000000e+00	0.000000e+00	0.000000e+00	4.000000e+00	0.000000e+00	0.000000e+00
0.000000e+00	0.000000e+00	0.000000e+00	1.500000e+01	0.000000e+00	0.000000e+00
0.000000e+00	0.000000e+00	0.000000e+00	3.200000e+01	0.000000e+00	0.000000e+00
1.650000e+02	1.200000e+01	1.650000e+02	1.200000e+02	3.260000e+03	3.260000e+03

Appendix - 5

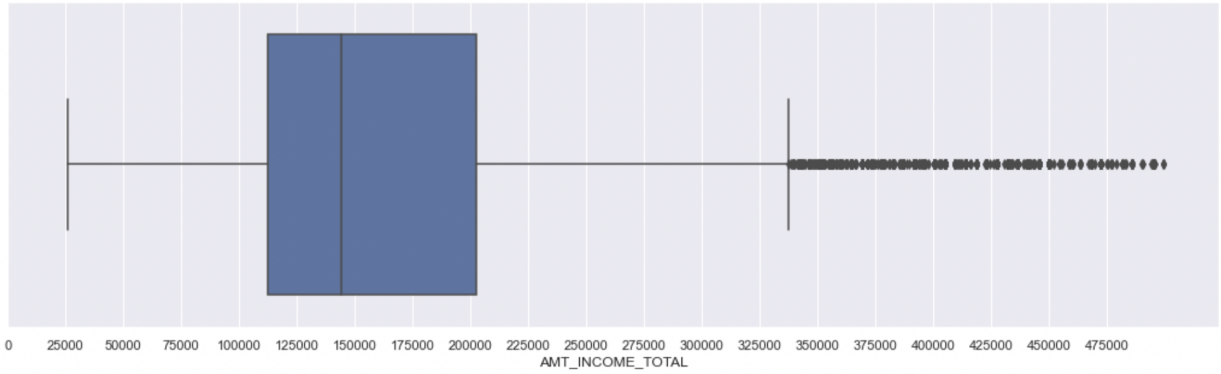
```
SK_ID_PREV                    int64
SK_ID_CURR                    int64
MONTHS_BALANCE                int64
AMT_BALANCE                 float64
AMT_CREDIT_LIMIT_ACTUAL       int64
AMT_DRAWINGS_ATM_CURRENT    float64
AMT_DRAWINGS_CURRENT        float64
AMT_DRAWINGS_OTHER_CURRENT  float64
AMT_DRAWINGS_POS_CURRENT    float64
AMT_INST_MIN_REGULARITY     float64
AMT_PAYMENT_CURRENT         float64
AMT_PAYMENT_TOTAL_CURRENT   float64
AMT_RECEIVABLE_PRINCIPAL    float64
AMT_RECIVABLE               float64
AMT_TOTAL_RECEIVABLE        float64
CNT_DRAWINGS_ATM_CURRENT    float64
CNT_DRAWINGS_CURRENT          int64
CNT_DRAWINGS_OTHER_CURRENT  float64
CNT_DRAWINGS_POS_CURRENT    float64
CNT_INSTALMENT_MATURE_CUM   float64
NAME_CONTRACT_STATUS         object
SK_DPD                        int64
SK_DPD_DEF                    int64
AMT_DRAWINGS_PCT            float64
AMT_DRAWINGS_ATM_PCT        float64
AMT_DRAWINGS_OTHER_PCT      float64
AMT_DRAWINGS_POS_PCT        float64
AMT_PRINCIPAL_RECEIVABLE_PCT float64
dtype: object
```

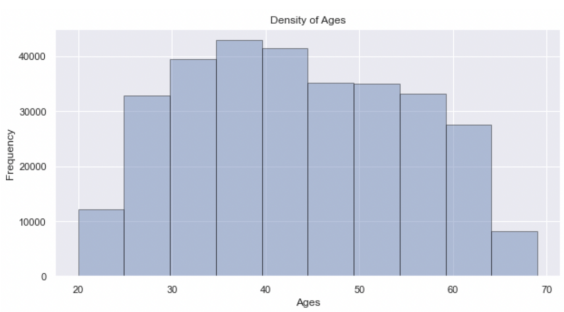
```
AMT_DRAWINGS_ATM_CURRENT      749816
AMT_DRAWINGS_OTHER_CURRENT    749816
AMT_DRAWINGS_POS_CURRENT      749816
AMT_INST_MIN_REGULARITY       305236
AMT_PAYMENT_CURRENT           767988
CNT_DRAWINGS_ATM_CURRENT      749816
CNT_DRAWINGS_OTHER_CURRENT    749816
CNT_DRAWINGS_POS_CURRENT      749816
CNT_INSTALMENT_MATURE_CUM     305236
dtype: int64
```

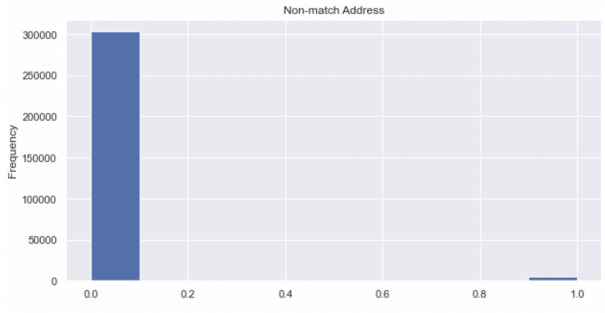
Appendix - 6

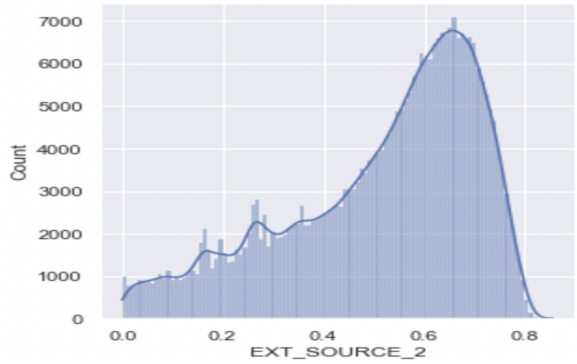
	SK_ID_PREV	SK_ID_CURR	NUM_INSTALMENT_VERSION	NUM_INSTALMENT_NUMBER	DAYS_INSTALMENT	DAYS_ENTRY_PAYMENT	AMT_INSTALMENT	AMT_PAYMENT
count	1.360540e+07	1.360540e+07	1.360540e+07	1.360540e+07	1.360540e+07	1.360250e+07	1.360540e+07	1.360250e+07
mean	1.903365e+06	2.784449e+05	8.566373e-01	1.887090e+01	-1.042270e+03	-1.051114e+03	1.705091e+04	1.723822e+04
std	5.362029e+05	1.027183e+05	1.035216e+00	2.666407e+01	8.009463e+02	8.005859e+02	5.057025e+04	5.473578e+04
min	1.000001e+06	1.000010e+05	0.000000e+00	1.000000e+00	-2.922000e+03	-4.921000e+03	0.000000e+00	0.000000e+00
25%	1.434191e+06	1.896390e+05	0.000000e+00	4.000000e+00	-1.654000e+03	-1.662000e+03	4.226085e+03	3.398265e+03
50%	1.896520e+06	2.786850e+05	1.000000e+00	8.000000e+00	-8.180000e+02	-8.270000e+02	8.884080e+03	8.125515e+03
75%	2.369094e+06	3.675300e+05	1.000000e+00	1.900000e+01	-3.610000e+02	-3.700000e+02	1.671021e+04	1.610842e+04
max	2.843499e+06	4.562550e+05	1.780000e+02	2.770000e+02	-1.000000e+00	-1.000000e+00	3.771488e+06	3.771488e+06

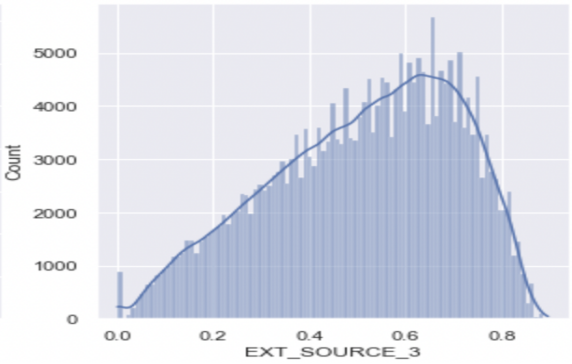
Appendix - 7

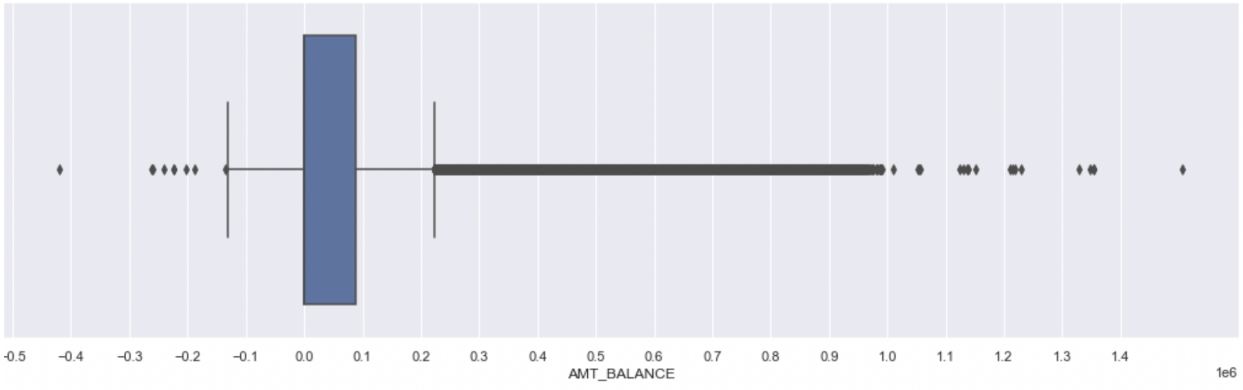

0 25000 50000 75000 100000 125000 150000 175000 200000 225000 250000 275000 300000 325000 350000 375000 400000 425000 450000 475000
AMT_INCOME_TOTAL

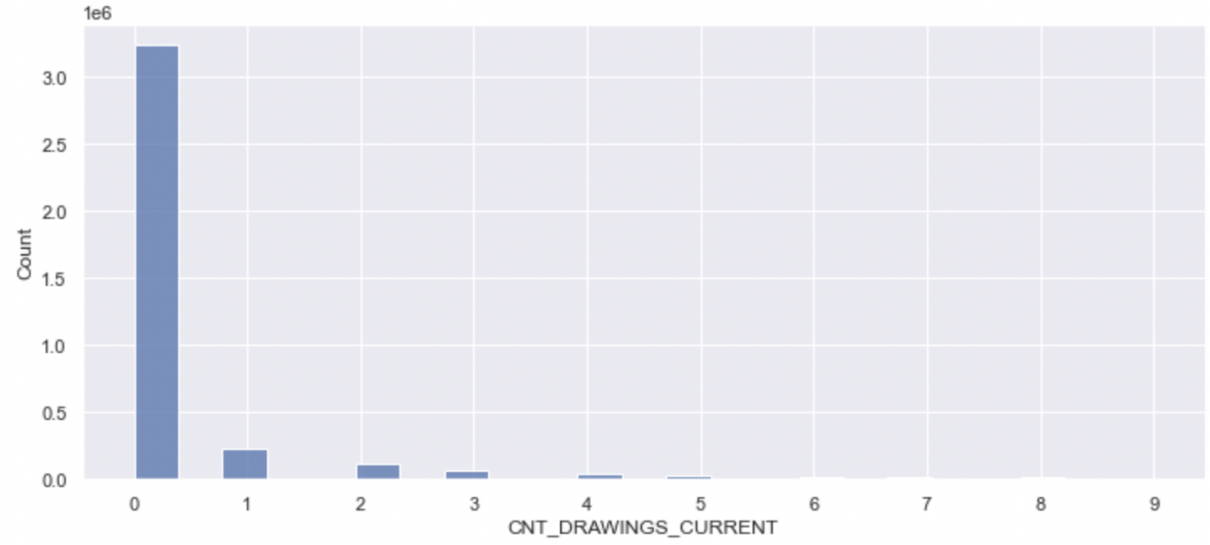

Density of Ages
Frequency
Ages

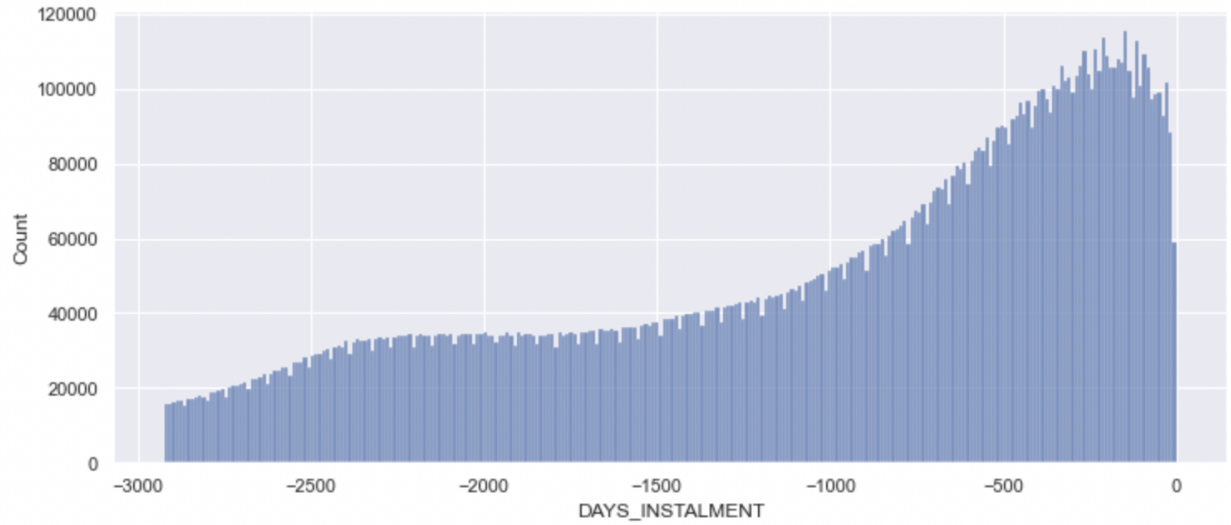

Non-match Address
Frequency

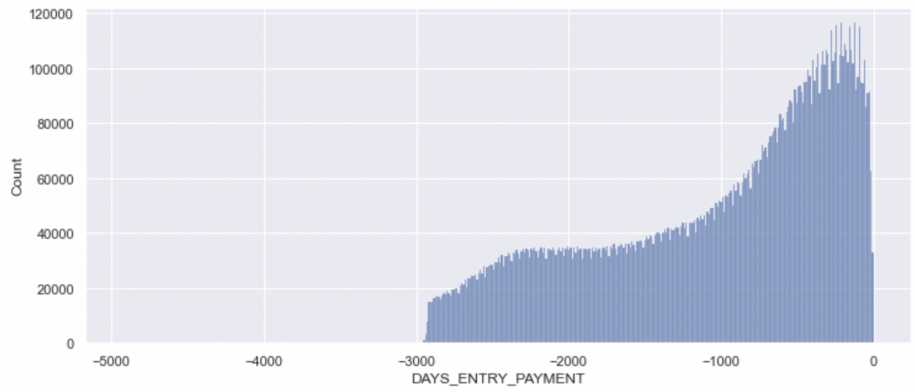

Count
EXT_SOURCE_2

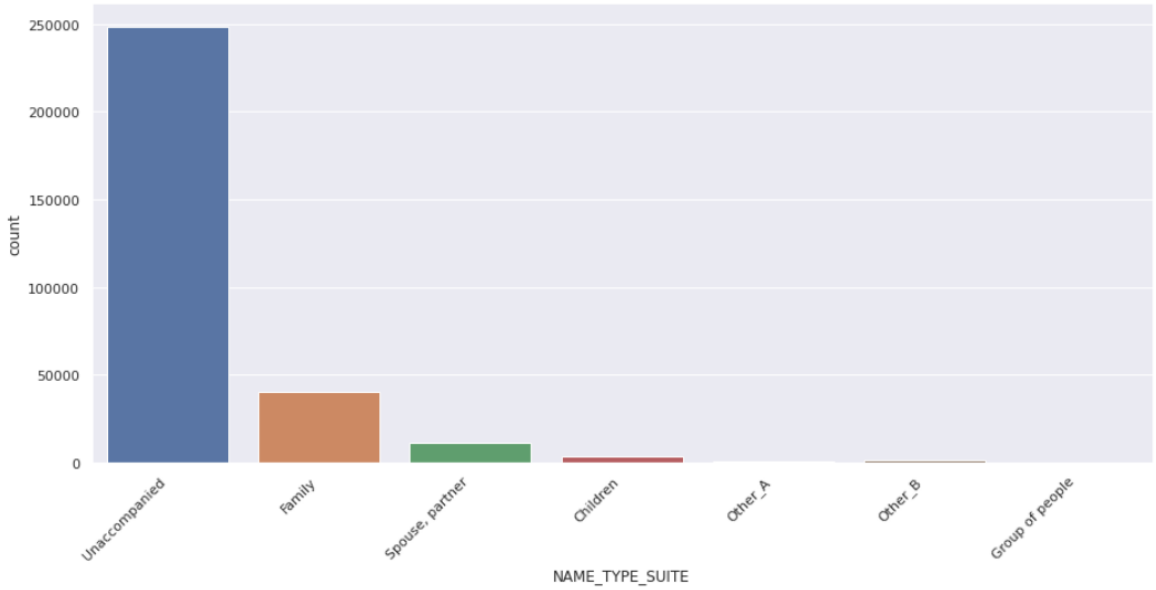

Count
EXT_SOURCE_3

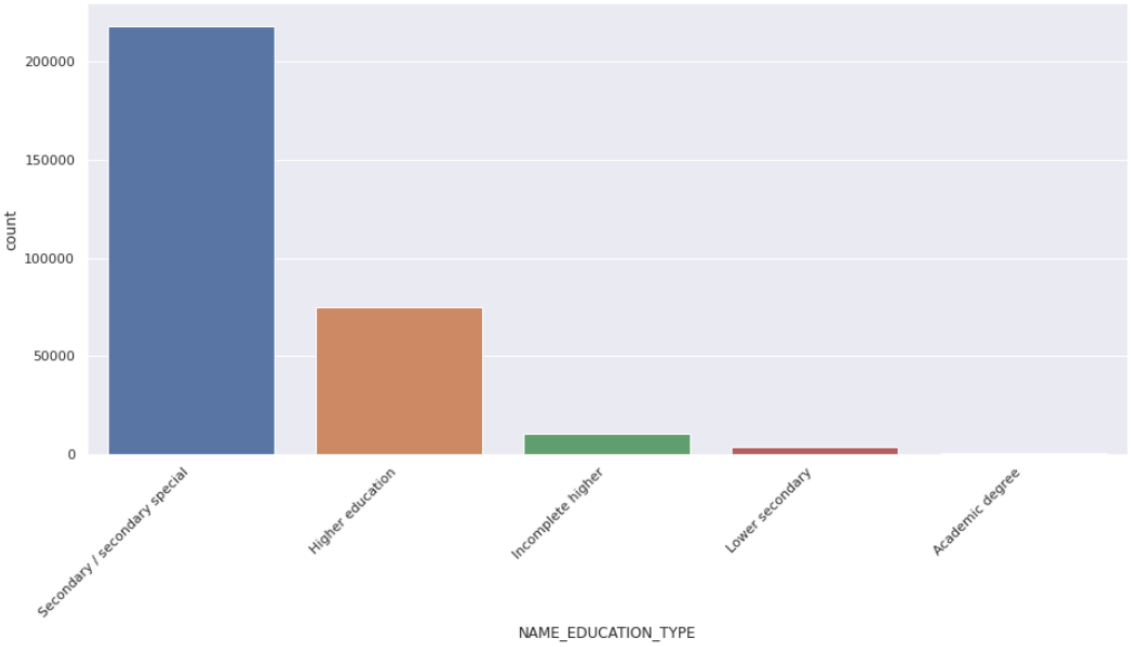

AMT_BALANCE
1e6

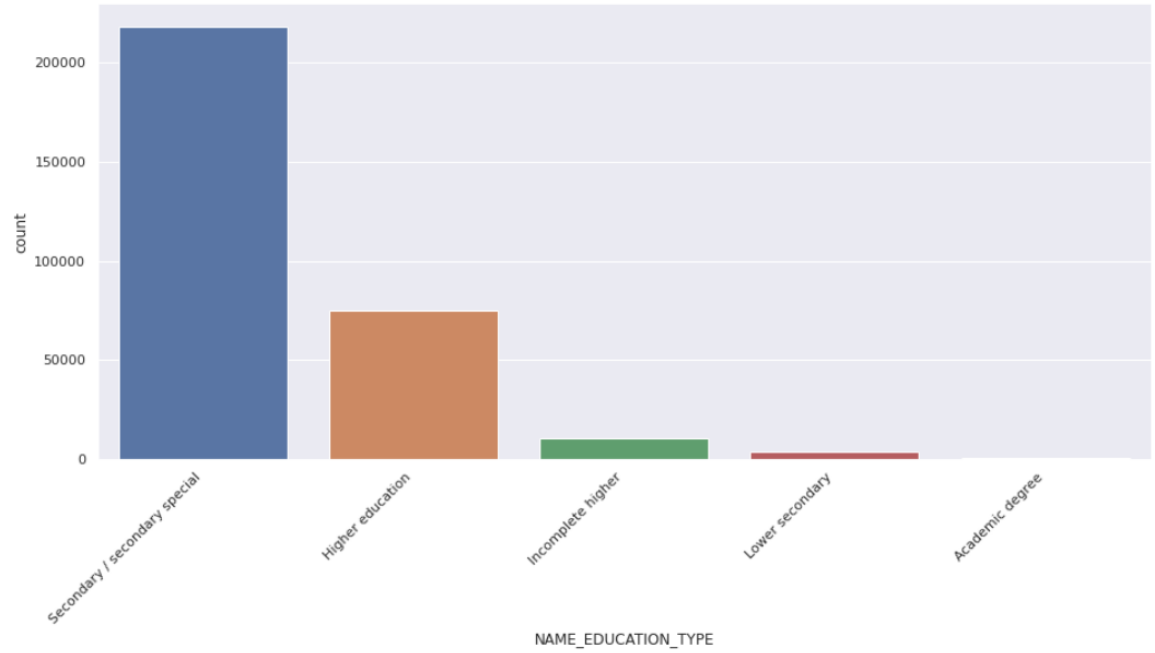

1e6
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Count
0
1
2
3
4
5
6
7
8
9
CNT_DRAWINGS_CURRENT

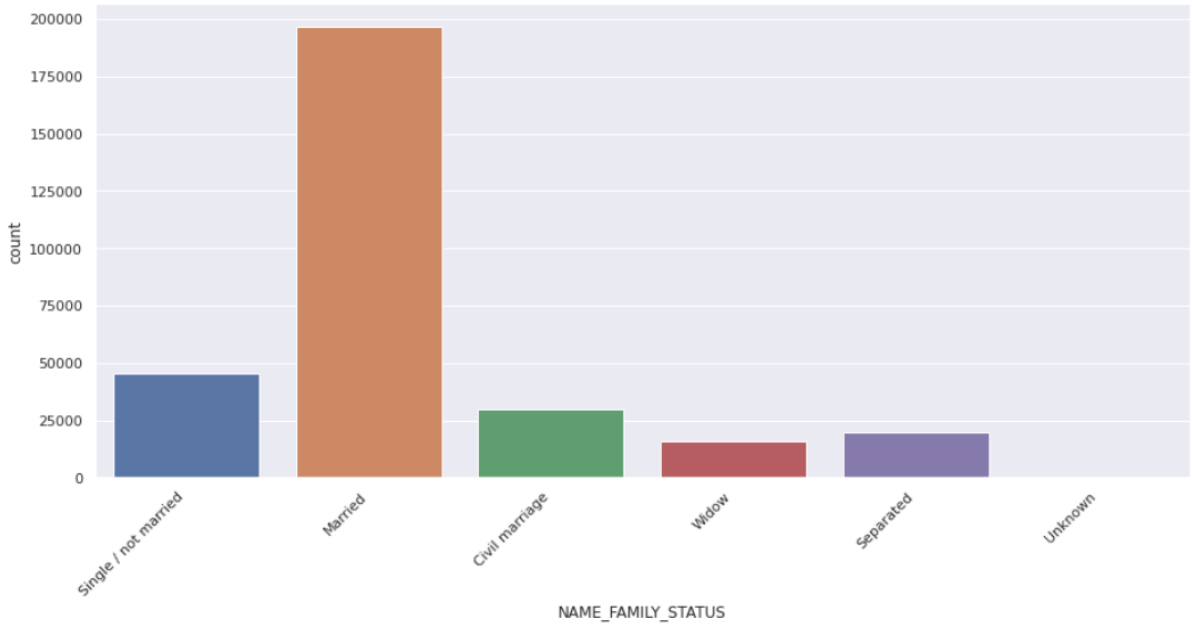

120000
100000
80000
60000
40000
20000
0
Count
−3000
−2500
−2000
−1500
−1000
−500
0
DAYS_INSTALMENT

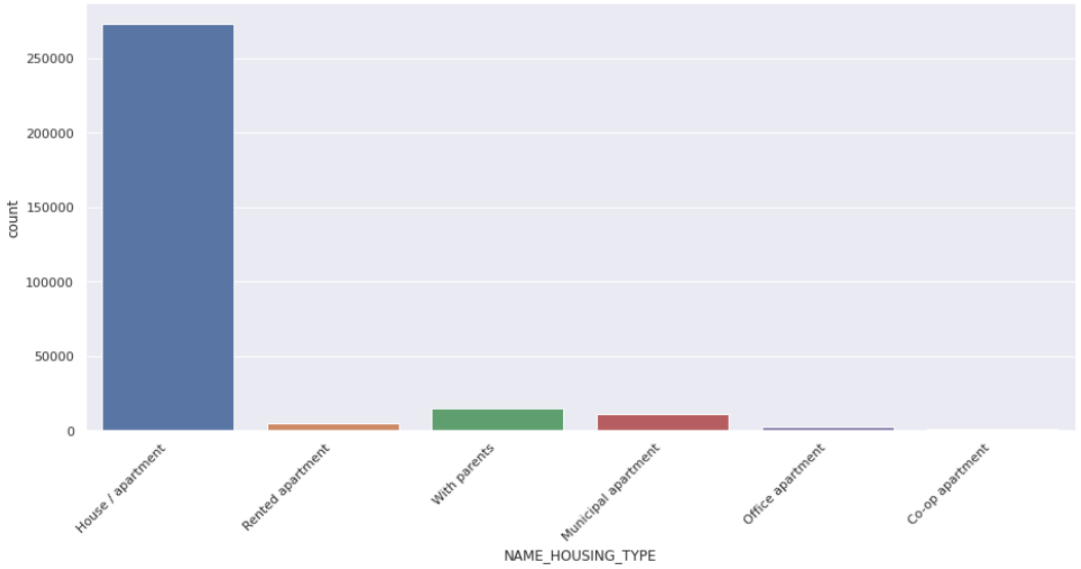

120000
100000
80000
60000
40000
20000
0
Count
−5000
−4000
−3000
−2000
−1000
0
DAYS_ENTRY_PAYMENT

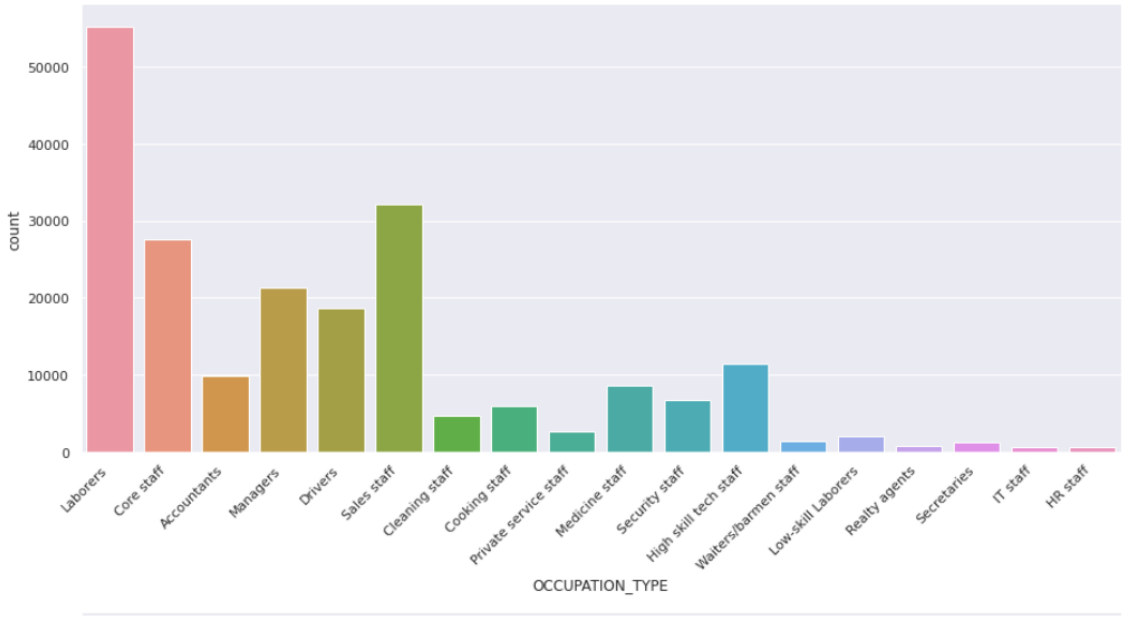

250000
200000
150000
100000
50000
0
count
Unaccompanied
Family
Spouse, partner
Children
Other_A
Other_B
Group of people
NAME_TYPE_SUITE


200000
150000
100000
50000
0
count
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

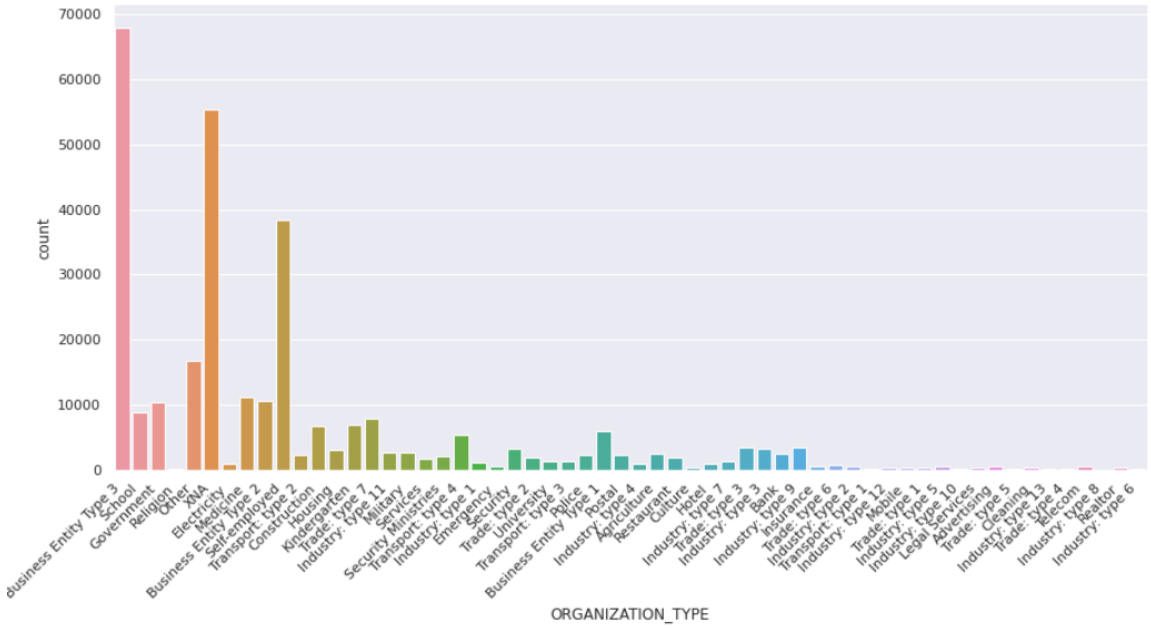

200000
150000
100000
50000
0
count
Secondary / secondary special
Higher education
Incomplete higher
Lower secondary
Academic degree
NAME_EDUCATION_TYPE

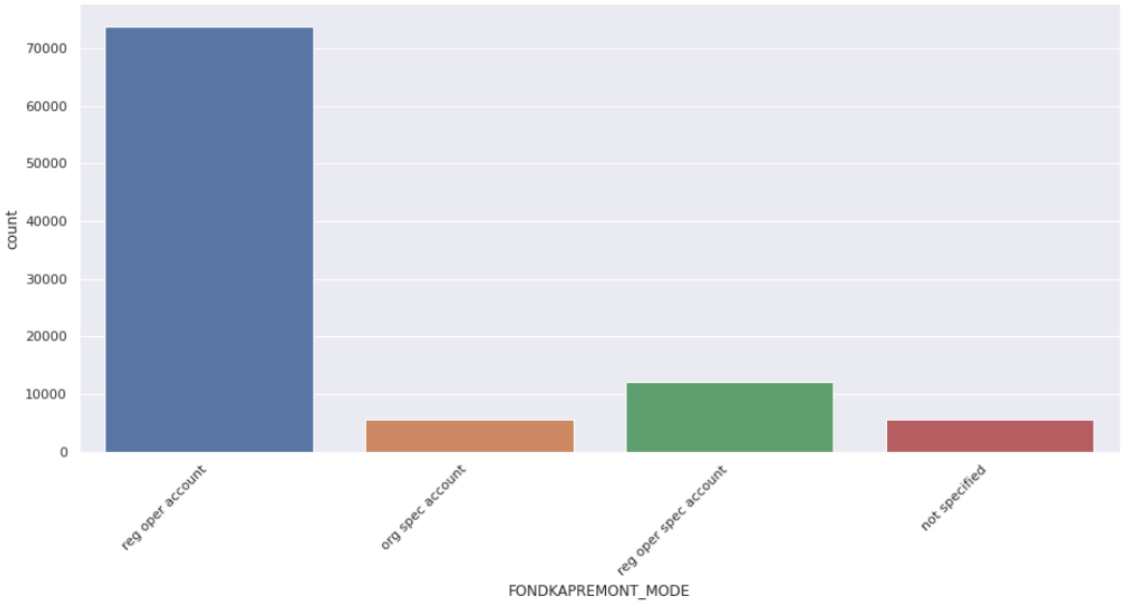

200000
175000
150000
125000
100000
75000
50000
25000
0
count
Single / not married
Married
Civil marriage
Widow
Separated
Unknown
NAME_FAMILY_STATUS

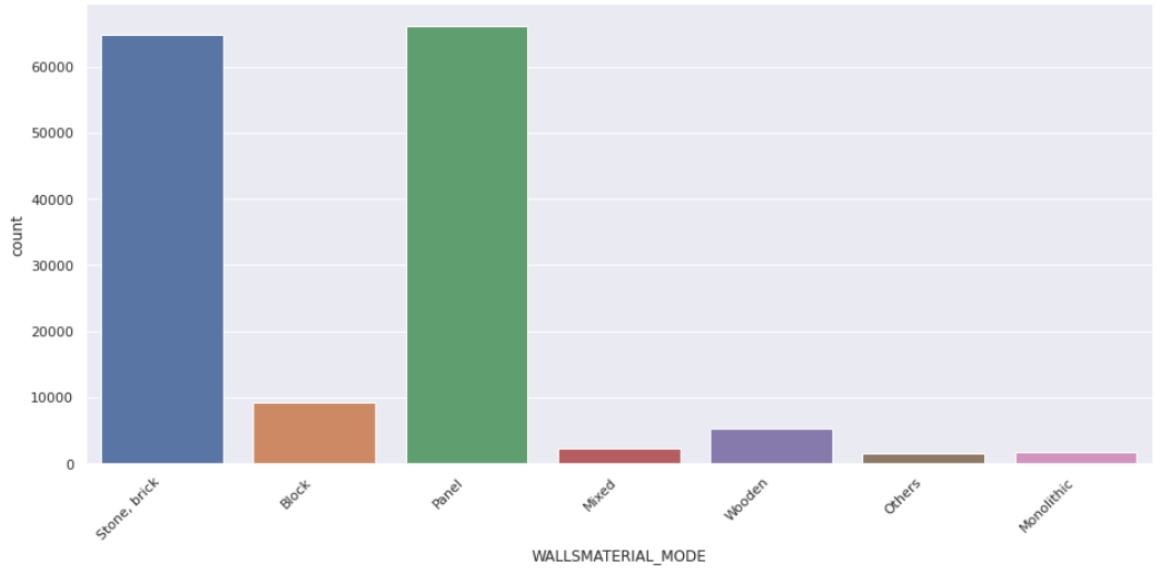

250000
200000
150000
100000
50000
0
count
House / apartment
Rented apartment
With parents
Municipal apartment
Office apartment
Co-op apartment
NAME_HOUSING_TYPE


50000
40000
30000
20000
10000
0
count
Laborers
Core staff
Accountants
Managers
Drivers
Sales staff
Cleaning staff
Cooking staff
Private service staff
Medicine staff
Security staff
High skill tech staff
Waiters/barmen staff
Low-skill Laborers
Realty agents
Secretaries
IT staff
HR staff
OCCUPATION_TYPE


count
50000
40000
30000
20000
10000
0
WEDNESDAY
MONDAY
THURSDAY
SUNDAY
SATURDAY
FRIDAY
TUESDAY
WEEKDAY_APPR_PROCESS_START


count
70000
60000
50000
40000
30000
20000
10000
0
Business Entity Type 3
School
Government
Religion
Other
XNA
Electricity
Medicine
Business Entity Type 2
Self-employed
Transport: type 2
Construction
Housing
Kindergarten
Trade: type 7
Industry: type 11
Military
Services
Security Ministries
Transport: type 4
Industry: type 1
Emergency
Security
Trade: type 2
University
Transport: type 3
Business Entity Type 1
Police
Postal
Industry: type 4
Agriculture
Restaurant
Culture
Hotel
Industry: type 7
Trade: type 3
Industry: type 3
Bank
Industry: type 9
Insurance
Trade: type 6
Industry: type 2
Transport: type 1
Industry: type 12
Mobile
Trade: type 1
Industry: type 5
Industry: type 10
Legal Services
Advertising
Trade: type 5
Industry: type 13
Cleaning
Trade: type 4
Telecom
Industry: type 8
Realtor
Industry: type 6
ORGANIZATION_TYPE


70000
60000
50000
40000
30000
20000
10000
0
count
reg oper account
org spec account
reg oper spec account
not specified
FONDKAPREMONT_MODE


60000
50000
40000
30000
20000
10000
0
count
Stone, brick
Block
Panel
Mixed
Wooden
Others
Monolithic
WALLSMATERIAL_MODE