Implementing machine learning techniques to identify poor quality borrowers

Alexander Gryga - 500400597

Table of Contents

[Introduction 3](#_Toc46742678)

[Literature Review 4](#_Toc46742679)

[Data Overview 6](#_Toc46742680)

[Descriptive Statistics 8](#_Toc46742681)

[Exploratory Data Analysis 12](#_Toc46742682)

[Approach 25](#_Toc46742683)

[**Step 1: Feature elimination using domain knowledge** 25](#_Toc46742684)

[**Step 2: Data pre-processing & cleaning** 25](#_Toc46742685)

[**Step 3: Joining tables** 25](#_Toc46742686)

[**Step 4: Data Normalization & Encoding** 26](#_Toc46742687)

[**Step 5: Upsampling** 26](#_Toc46742688)

[**Step 6: Feature Selection** 26](#_Toc46742689)

[**Step 7: Model development** 26](#_Toc46742690)

[**Step 8: Model scoring** 26](#_Toc46742691)

[Results 27](#_Toc46742692)

[Conclusion 38](#_Toc46742693)

[References 39](#_Toc46742694)

[Appendix 41](#_Toc46742695)

[Current Applicants Table 41](#_Toc46742696)

[Previous Applications Table 48](#_Toc46742697)

# Introduction

With growing amounts of consumers taking loans to fund purchases, financial institutions seek technological innovations for efficiencies in minimizing lending risks associated with borrowers lagging on payments or defaulting. Depending on client profiles, credit history and personal circumstances, loans could be denied, reduced or lending terms adjusted with higher interest rates, shorter repayment schedules, or other pre-cautions for higher risk clients. Correctly classifying problematic clients allows financial institutions to minimize lending risk and grow revenue.

Utilizing supervised learning methods to develop classification models, clients and future applicants will be classified on their likelihood of successfully repaying loans. Employing logistic regression, random forests and gradient boosted decision trees, this analysis aims to identify which model would identify problematic applicants most precisely and accurately. Additionally, using machine learning to gain a deeper understanding of key drivers correlated with problematic applicants provides insights and opportunities to guide business steering.

The full GitHub repository can be found at <https://github.com/AGryga/Loan_Delinquency_ML>

The dataset is available for download at https://www.kaggle.com/mishra5001/credit-card#application\_data.csv

# Literature Review

This literature review looks to research academic journals to better understand how similar researchers approached credit and lending related machine learning problems. This provides learning opportunities to better understand machine learning frameworks and techniques applied, as well as addressing challenges with unbalanced datasets and general approaches to data pre-processing and cleaning.

Liang, Wang & Jin (2019) sought solutions for when applicants had sparse credit history, using machine learning to identify which applicants can successfully repay the loan. The dataset contained the loan’s status, employment & credit histories, and other personal features. experimenting with using logistic regression, random forests, gradient boosted decision trees, Naïve Bayes, and Neural Networks to identify the ideal model framework. Incorporating PCA dimensionality reduction techniques, 800 features were reduced to key drivers of fraud. The authors compared accuracy, precision, recall and F1 scores of the various models. The authors ultimately look to focus on accuracy, precision and F1 scores of the models to understand the performance and results trade-offs of the various models while finding no distinctively ideal model.

Ma, Sha, Yang, Yu, Yang & Niu (2018) studied implementing machine learning to identify defaulters on a P2P lending platform, Lending Club to ultimately reduce default rates. Using a multi-dimensional dataset containing multiple classifications, the dataset’s features contains loan details, personal information, financial situation details and credit datapoints. Using gradient boosted decision trees to successfully identify that borrowing details and financial situations were key drivers to identify defaulters. Using accuracy as the primary scoring method of the models, they identified that LightGBM models had outperformed XGboost models, correctly classifying applicants more frequently while maintaining a lower level of standard deviation.

Researching an article covering credit card fraud, Stolfo, Fan, Lee & Prodromidi (1997) use various decision trees and neural networks to identify fraudulent credit card transactions, developing their model with a credit card transaction dataset obtained from an American financial institution. Implementing BAYES, CART, RIPPER and ID3 techniques while simultaneously balancing the training data with an equal proportion of fraud vs. non-fraudulent transactions. The authors used the True Positive and False Positive rates to decide on the most effective model, looking to maximize correct classifications, finding that Bayes models performed best. The authors noted if the proportional split of the training data begins to skew to an unbalanced training dataset, the accuracy begins to fall.

Bhagat (2018) implemented logistic regression to classify loan applicants on their likelihood to default on a loan. Following deployment of the logistic regression model, the author seeks to utilize precision, accuracy, True Positive and False Positive rates as the indicators for the model’s effectiveness. With the dataset being unbalanced and skewing towards successful repayments, the author utilized k-folds cross validation to minimize overfitting and underfitting, experimenting with manual weighting. Logistic regression was successful after incorporating manual weighting. Bhagat identifies random forests and neural networks could also be implemented as alternative techniques to identify at risk applicants.

Addo, Guegan & Hassani (2018) studied the use of machine learning algorithms to predict the likelihood of a client to default on their loan. Identifying the 10 key features selected by each model, the authors implemented logistic regression, random forests, gradient boosted trees, and neural networks to identify likely to default individuals. For quantifying the performance of these models, the authors utilized AOC and AUC, F1 scores as well as precision and accuracy scores. When comparing results, tree-based algorithms (random forest & GBTs) performed strongest with neural networks performing weakest.

Coser, Maer-Matei & Albu (2019) studied the use of machine learning algorithms for identifying loan delinquency in loaning scenarios, comparing various features around the loan and credit history. Employing logistic regression, gradient boosted trees, and random forests the team looked to identify the strongest performing model to increase profitability by identifying and rejecting delinquent clients. Comparing accuracy, precision, recall and the AUC score, the team identified that random forests performed strongest.

Nui, Li & Ren (2019) similarly looked to develop credit scores for consumers lacking credit histories. Utilizing social media and other personal information, they implemented random forests and gradient boosted trees to identify the likelihood of default when using P2P lending platforms. When comparing the performance of the model, correlation was used for logistic regression and decision tree-oriented models were compared based on AUC, F1 score and accuracy. LightGBM algorithms performed the best although logistic regression was also able to identify strongly correlated features linked to loan default.

Vieira, Barboza, Sobreiro & Kimura (2019) look to predict mortgage defaults amongst low income families who had their loan approved. Engaging income, mortgage and loan details, and demographic oriented features to train and test the models they deployed bagging, boosted trees and random forests models. The authors compared the performance of the various algorithms prioritizing AUC, KS, Brier, and precision & accuracy scores. Random forests incrementally outperformed bagging and boosting algorithms with the bagging and boosted models performing equally.

# Data Overview

Using a publicly available Kaggle dataset (2019), the aim is to develop a machine learning model to identify clients at risk of being unable to repay loans in a timely manner. This dataset consists of two tables: one containing current applicant data and the second containing data of client’s prior loan applications.

Within the current applicant table, there are 122 features, including the target variable. The remaining 121 features span continuous quantitative variables such as the amount of credit issued and the annuity, the client’s income, the price of the goods being purchased (if applicable), age of the borrower and a discrete count of credit bureau inquiries. Alongside loan specific details are qualitative demographic features such relationship status and family information supplemented with qualitative socio-economic features providing an overview of the client’s education, employment and living situation. Binary values are present to flag if the client has car or real estate assets. Additional features around the application itself are present, using binary flag to identify whether contact information provided during application are valid, as well as binary flags for whether the client submitted specific documents during the application process.

The previous loan table contains 37 features around previous loan applications made by clients. Continuous quantitative variables are present for credit requested by the applicant, the amount approved by the lending institution and if applicable, the price of the goods being purchased, and the down payment made by the client. Qualitative attributes exist as well, identifying the approval status of the loan, details around the application channels and other marketing details.

A full list of all features and their description for both tables is available within the appendix.

Eliminating features from the initial dataset using domain knowledge, there are a variety of different features which can be immediately removed. Within the current applicants table, the following features can be removed.

1. External scores

2. Housing features

3. Occurrences of days past due on loans within applicant’s social circle

With lack of documentation around the external score feature, we would remove this due to interpretability issues and remove these 3 features which potential use the output of a previous machine learning model which scored applicants to reduce inference from other models.

The housing feature dataset contains mean, median and mode values of the number of floors, elevators, square footage, age of the building and more of the building the client lives in. Due to this data encompassing the building a client lives, this is irrelevant to the client and the loan applied for.

Finally, the features which count the number of occurrences where a friend or family member of the client had a late payment on a loan were also removed. Including these in the model could potentially raise ethical and legal difficulties due to discrimination. As these set of features focus on the social circle of the client and not the client themselves, these features are directly irrelevant and can create issues around potential discrimination lawsuits in the Canadian market.

Additionally, within the previous application table, the following features can be removed.

1. Day and time of application

2. Interest rates on prior loans

3. Previous reason the loan was rejected

4. Intended use for the loan

5. Marketing related information

The above features are not directly correlated with the current application which could introduce additional noise. Additionally, with applicants ranging from 0 to multiple previous applications, we look to minimize potential noise introduced by irrelevant features.

# Descriptive Statistics

Removing the unnecessary features identified within the feature elimination stage, the current applicant’s dataset has been reduced to 48 remaining features. Specifically focusing on non-binary features, overviews are provided below for the quantitative and qualitative features.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Feature | Count | Unique | Top | Freq |
| NAME\_CONTRACT\_TYPE | **307,511** | **2** | **Cash loans** | **278,232** |
| CODE\_GENDER | **307,511** | **3** | **F** | **202,448** |
| NAME\_TYPE\_SUITE | **306,219** | **7** | **Unaccompanied** | **248,526** |
| NAME\_INCOME\_TYPE | **307,511** | **8** | **Working** | **158,774** |
| NAME\_EDUCATION\_TYPE | **307,511** | **5** | **Secondary / secondary special** | **218,391** |
| NAME\_FAMILY\_STATUS | **307,511** | **6** | **Married** | **196,432** |
| NAME\_HOUSING\_TYPE | **307,511** | **6** | **House / apartment** | **272,868** |
| OCCUPATION\_TYPE | **211,120** | **18** | **Laborers** | **55,186** |
| WEEKDAY\_APPR\_PROCESS\_START | **307,511** | **7** | **TUESDAY** | **53,901** |
| ORGANIZATION\_TYPE | **307,511** | **58** | **Business Entity Type 3** | **67,992** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Feature | Count | Mean | Std | Min | 25% | 50% | 75% | Max |
| TARGET | **307,511** | **0.08** | **0.27** | **-** | **-** | **-** | **-** | **1.00** |
| CNT\_CHILDREN | **307,511** | **0.42** | **0.72** | **-** | **-** | **-** | **1.00** | **19.00** |
| AMT\_INCOME\_TOTAL | **307,511** | **168,798** | **237,123** | **25,650** | **112,500** | **147,150** | **202,500** | **117,000,000** |
| AMT\_CREDIT | **307,511** | **599,026** | **402,491** | **45,000** | **270,000** | **513,531** | **808,650** | **4,050,000** |
| AMT\_ANNUITY | **307,499** | **27,109** | **14,494** | **1,616** | **16,524** | **24,903** | **34,596** | **258,026** |
| AMT\_GOODS\_PRICE | **307,233** | **538,396** | **369,446** | **40,500** | **238,500** | **450,000** | **679,500** | **4,050,000** |
| REGION\_POPULATION\_RELATIVE | **307,511** | **0.02** | **0.01** | **0.00** | **0.01** | **0.02** | **0.03** | **0.07** |
| DAYS\_BIRTH | **307,511** | **16,037** | **4,364** | **25,229** | **19,682** | **15,750** | **12,413** | **- 7,489** |
| DAYS\_EMPLOYED | **307,511** | **63,815** | **141,276** | **17,912** | **2,760** | **1,213** | **289** | **365,243** |
| DAYS\_REGISTRATION | **307,511** | **4,986** | **3,523** | **24,672** | **7,480** | **4,504** | **2,010** | **-** |
| DAYS\_ID\_PUBLISH | **307,511** | **2,994.20** | **1,509.45** | **7,197.00** | **4,299.00** | **3,254.00** | **1,720.00** | **-** |
| OWN\_CAR\_AGE | **104,582** | **12.06** | **11.94** | **-** | **5.00** | **9.00** | **15.00** | **91.00** |
| CNT\_FAM\_MEMBERS | **307,509** | **2.15** | **0.91** | **1.00** | **2.00** | **2.00** | **3.00** | **1.00** |
| REGION\_RATING\_CLIENT | **307,511** | **2.05** | **0.51** | **1.00** | **2.00** | **2.00** | **2.00** | **1.00** |
| REGION\_RATING\_CLIENT\_W\_CITY | **307,511** | **2.03** | **0.50** | **1.00** | **2.00** | **2.00** | **2.00** | **1.00** |
| DAYS\_LAST\_PHONE\_CHANGE | **307,510** | **962.86** | **826.81** | **4,292** | **1,570.00** | **757.00** | **274.00** | **1.00** |

When looking at the descriptive statistics of the quantitative features. We can see that the target variable is highly skewed towards clients who will be non-delinquent, with 92% have no issues with repayment. The average applicant will apply for a 600,000 rupee loan while having an average yearly income of 168,789 rupees.

Based on birthday data, applicants are roughly 44 years old at the time of application.

Focusing on the categorical variables within the dataset, value frequency for each feature is represented on the left-hand side with analysis taking place on the right-hand side.

|  |  |
| --- | --- |
| Target | Count |
| 0 | 282,686 |
| 1 | 24,825 |

Within the target feature column, we see that the dataset is highly imbalanced, with 92% of having no issues with repaying the loan and 8% having some sort of challenge.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Gender | Count | | F | 202,448 | | M | 105,059 |  |  |  | | --- | --- | | Loan Type | Count | | Cash loans | 278,232 | | Revolving loans | 29,279 |  |  |  | | --- | --- | | Real Estate Ownership | Count | | N | 213,312 | | Y | 94,199 | | The gender ratios of the applicants skew towards female, where there is a 2:1 ratio of females to males amongst the dataset’s clients.  With the vast majority of loans being cash loans, these are typically used to purchase goods where as revolving loans are essentially treated as open credit for a client.  Roughly 30% of applicants own real estate, providing additional value to their net worth. |
| |  |  | | --- | --- | | Car Ownership Count | | | N | 202,924 | | Y | 104,587 | | Roughly 33% of applicants own a vehicle. |
| |  |  | | --- | --- | | Housing | Count | | House / Apartment | 272,868 | | With Parents | 14,840 | | Municipal Apartment | 11,183 | | Rented Apartment | 4,881 | | Office Apartment | 2,617 | | Co-Op Apartment | 1,112 | | A sizable portion of applications live in a house or apartment. With no distinction being made for either a house or apartment in the ‘House / Apartment’ category, this reduces some of the interpretability of this feature and reduces the opportunity of doing additional analysis for applicants living in houses vs. apartments.  Ideally, if the data capture process could provide greater distinctions, additional analyses could be conducted. |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Family Status | Count | | Married | 196,432 | | Unmarried | 45,444 | | Civil Marriage | 29,775 | | Separated | 19,770 | | Widow | 16,088 | | With 73% of applicants having a partner, summarized regardless of the legal status of the relationship. Of those 73% of applicants, additional income could be captured by their partner or spouse, supplementing the individual’s income with their partners. Similarly, this could add additional risk as other potential loans and debt obligations their partner may have would be undocumented and uncaptured within this dataset. |
| |  |  | | --- | --- | | Education | Count | | Higher Education | 248,526 | | Incomplete Education | 40,149 | | Lower Secondary | 11,370 | | Academic Degree | 3,267 | | 87% of applicants have completed some level of education, with 82% completing higher levels of education. Ideally, applicants with higher levels of education should see greater income and stability, |

|  |  |  |  |
| --- | --- | --- | --- |
| Target | Loan Type | Cash loans | Revolving loans |
| **0** | 255,011 | 27,675 |
| **1** | 23,221 | 1,604 |

When comparing the ratio of the type of loans applied for cash loans vs. revolving loans, the dataset regardless of the target values skew towards a cash loan. Both types of loans are heavily skewed towards the loan being successfully repaid.

|  | **Target** | **Car Ownership** | **N** | | **Y** | |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Realty Ownership** | **N** | **Y** | **N** | **Y** |
|  | **0** | 56,400 | 129,275 | 29,957 | 67,054 |
|  | **1** | 5,572 | 11,677 | 2,270 | 5,306 |

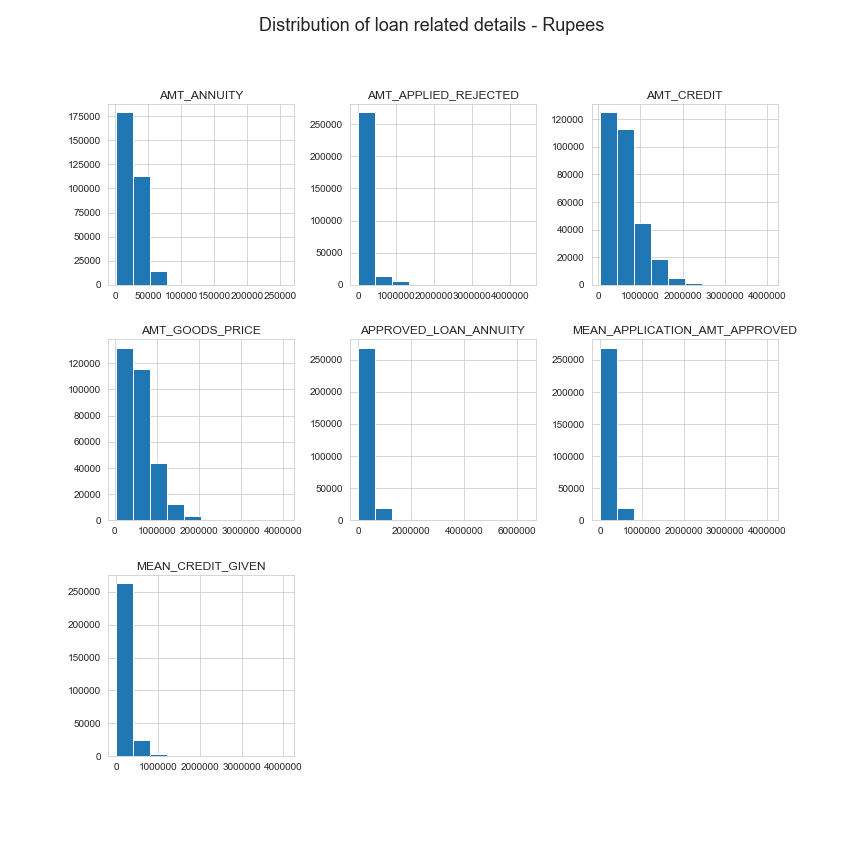
When comparing the ratio of car and real estate ownership with loan delinquency, the proportion of assets compared against delinquency remains consistent throughout the various features though there is a larger propensity for individuals in this dataset to own real estate.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Housing | Family Status | Civil marriage | Married | Separated | Single / not married | Widow |
| **Co-op apartment** | 75 | 678 | 66 | 276 | 27 |
| **House / apartment** | 26,057 | 178,290 | 17,178 | 36,409 | 14,930 |
| **Municipal apartment** | 1,178 | 6,385 | 957 | 1,875 | 786 |
| **Office apartment** | 227 | 1728 | 173 | 382 | 107 |
| **Rented apartment** | 616 | 2,593 | 340 | 1,238 | 94 |
| **With parents** | 1,621 | 6,755 | 1,056 | 5,264 | 144 |

From the cross-tabs table above, the majority of applicants will be living in a house or apartment. The majority of individuals are married or in a civil marriage, potentially indicating that multiple incomes could also support repayment of loans.

# Exploratory Data Analysis

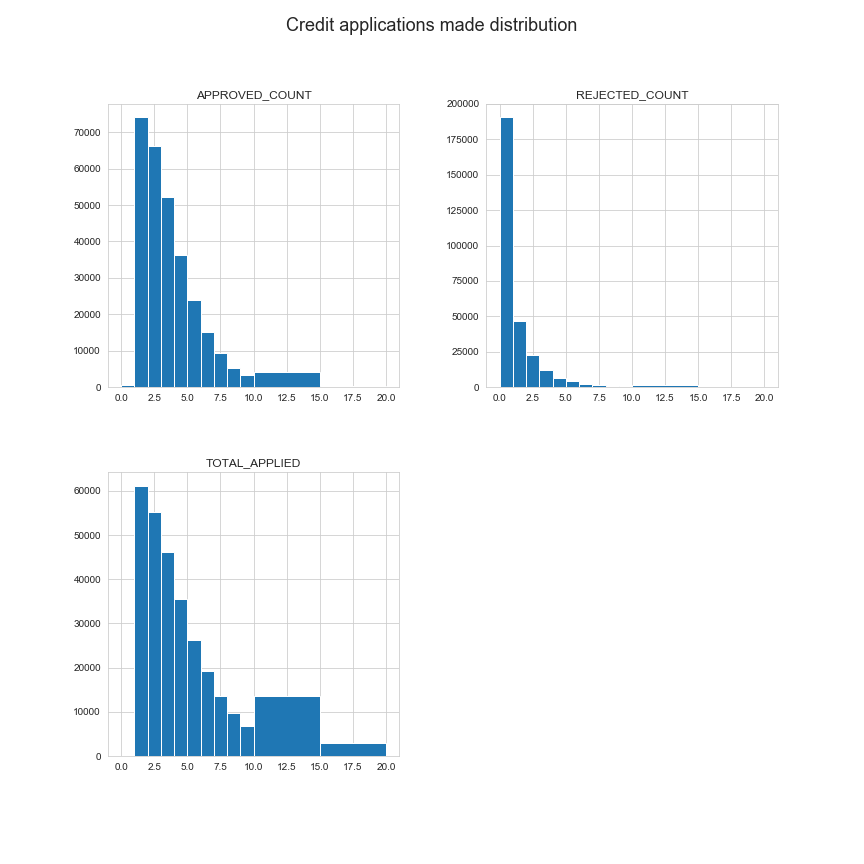
Following data preprocessing and data preparations where unnecessary features or features containing significant amounts of null values are removed, redundant values are reassigned for clarity and the current loan applications and previous loan history datasets are combined. Previous loan history is aggregated, summarizing the dataset into the total count of previous applications, the count of how many were approved or rejected and the corresponding average loan they had applied for and credit they had received, if applicable. These aggregations are then appended onto the current applicant dataset, if a history is present.



When looking at the current loan details (AMT\_CREDIT, AMT\_GOODS\_PRICE & AMT\_ANNUITY), we can identify that current loans are primarily in larger denominations than previous applications, both loans that were previously rejected and well as previously approved.

One key discovery is that based on the current credit being offered to the customer vs. the loan annuity of the current loan provided, we can see that the count of payment cycles is not included in the dataset. Because loan annuity values are smaller than the credit offered, we are missing a key variable as banks would not offer free money to applicants.

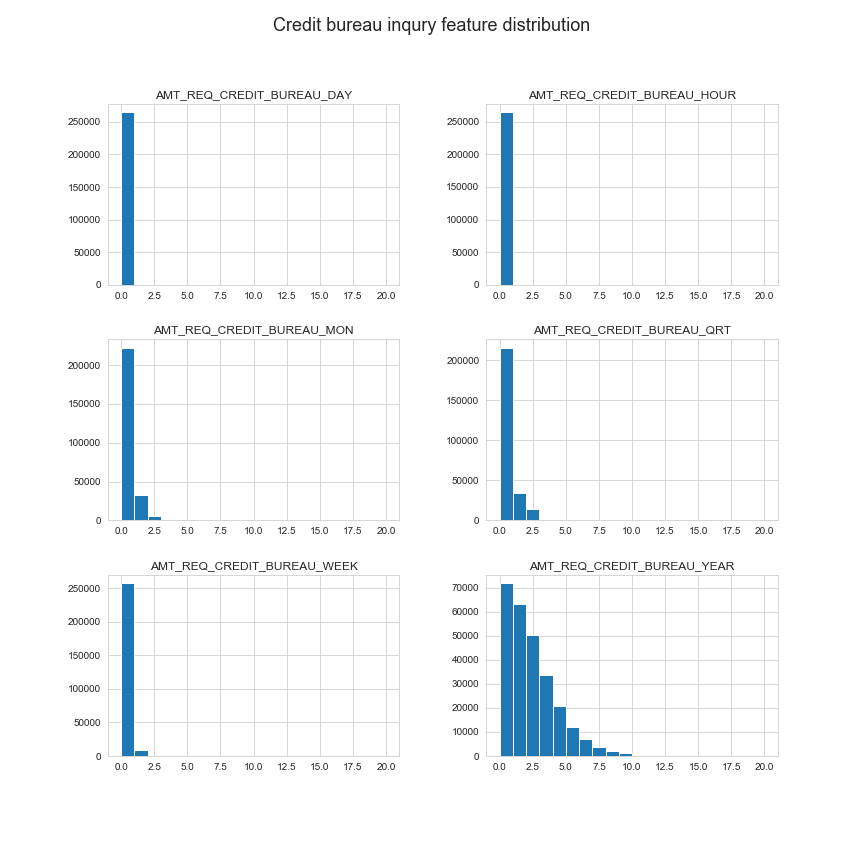
Additionally, there seems to be no distinction between the size of the loan requested and whether or not that would be critical to approve or reject an applicant based on previous approvals and previous rejections.



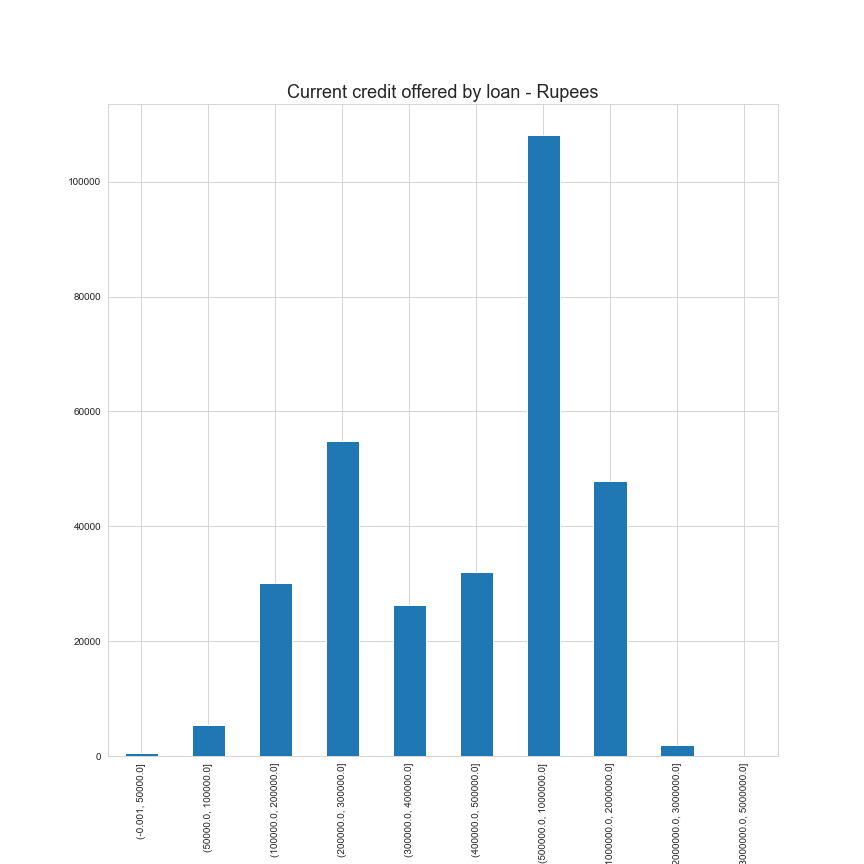
When looking at available loan history data for reoccurring applicants we can see that the distribution is skewed to the left, with the majority of applicants not applying to more than 5 loans. For applicants there does seem to be a potential likelihood that not all applications had been approved.

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | TOTAL\_APPLIED | APPROVED\_COUNT | REJECTED\_COUNT |
| count | 290,830 | 290,830 | 290,830 |
| mean | 3.968848 | 3.12509 | 0.843758 |
| std | 3.209301 | 2.181529 | 1.814422 |
| min | 1 | 0 | 0 |
| 25% | 2 | 1 | 0 |
| 50% | 3 | 3 | 0 |
| 75% | 5 | 4 | 1 |
| max | 72 | 28 | 68 |

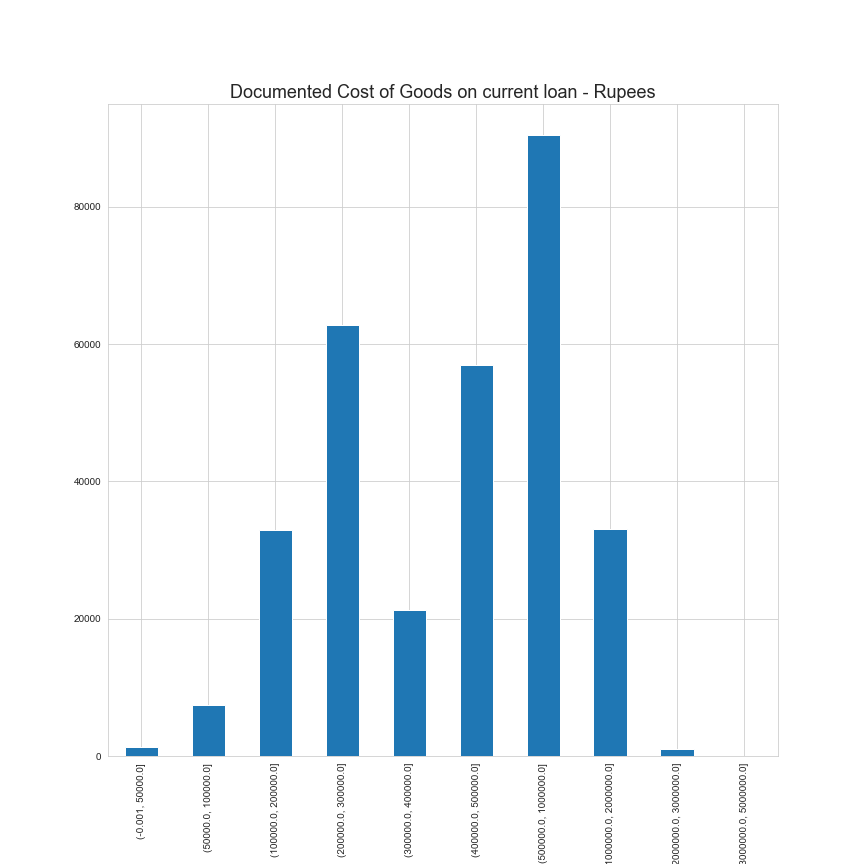
Of the 307, 511 current applicants, 290,830 of them have previous loan history on file. For the 94% of applicants with an established loan history, an average of 4 loans have been applied for and typically 3 are approved. The high rejection percentage may be skewed due to few problematic applicants with multiple rejected loan applications. Until the 75th percentile, we see that applicants typically have no rejected loans, indicating a sizable portion of previous applicants had not been rejected.



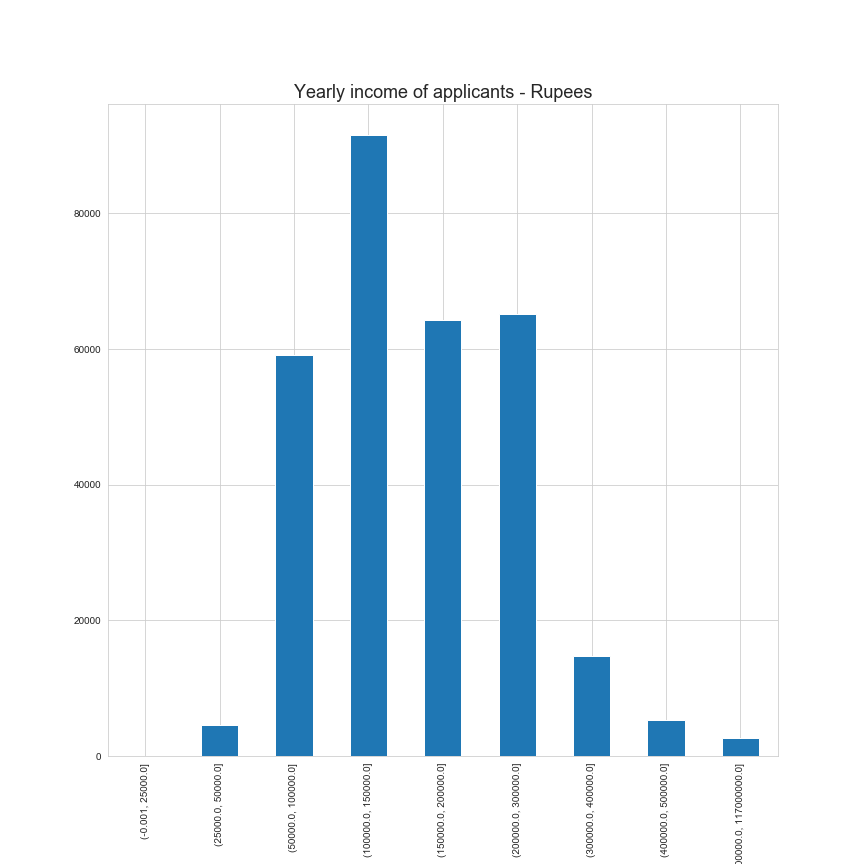
Based on available credit bureau inquiries within the last hour, day, week, month, quarter and year, we can see there is no particular trend or immediate concern with applicants who have had multiple credit bureau inquiries in a short time frame that typically indicates fraud. As the time horizon grows, we see a gradual increase in credit inquiries.



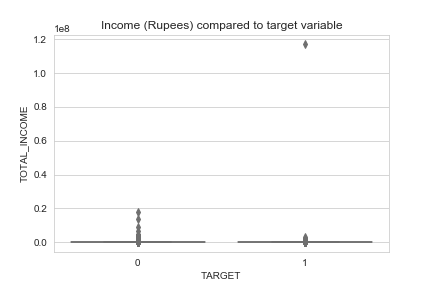
Based on the credit being offered in the current loan, the dataset is skewed to the right, with the majority of loans being 500,000 rupees to 1,000,000 rupees. There is still plenty of applicants applying for smaller loans, however the majority of loans are higher value.



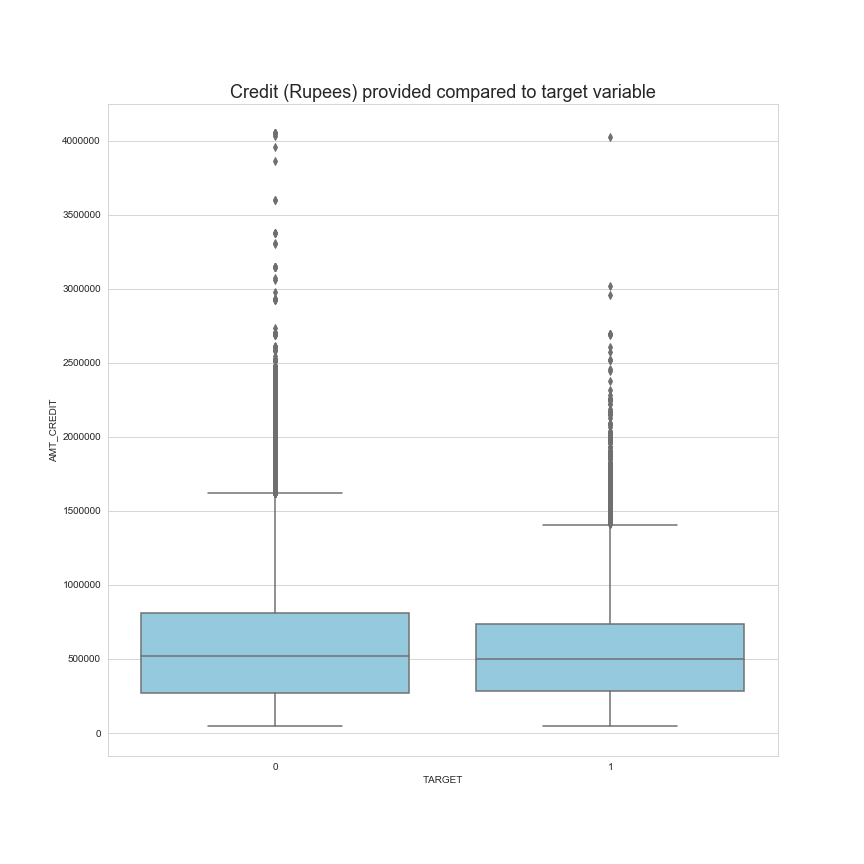
Similarly, when looking at the cost of goods that the loan will be utilized for, it follows a similar distribution to the credit provided by loans, indicating that the loan is near equal to the item being purchased.



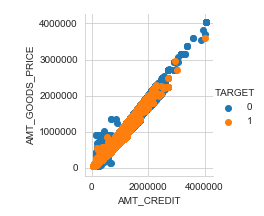
Comparing the current loan details against the income of the clients, we see that generally, average income is around 150,000 rupees, indicating that applicants are leveraging for purchasing some of these goods, indicating that they are high value purchases.



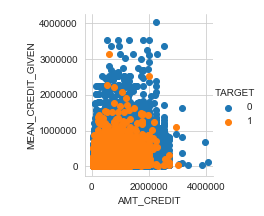
When analyzing the income of an applicant relative to the target variable, denoting those who will have loan repayment difficulties or default, we see that the majority of those who will have challenges will be lower income applicants with some outliers of high-income applicants. This would support the insights above where applicants are becoming leveraged, taking on a high value loan relative to their income.



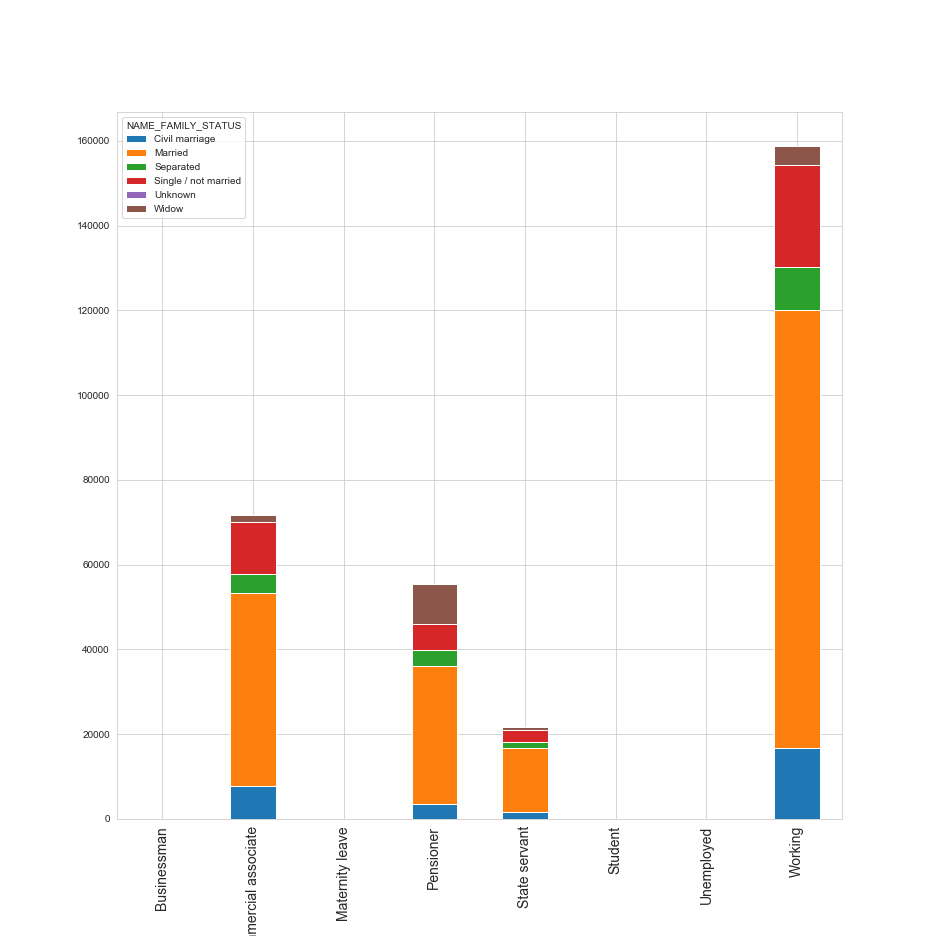
Examining the credit offered to applicants, outside of outliers, there does not seem to be a stark difference in the amount of credit being extended to those who have repayment difficulties versus those who successfully repay the loan in a timely fashion.



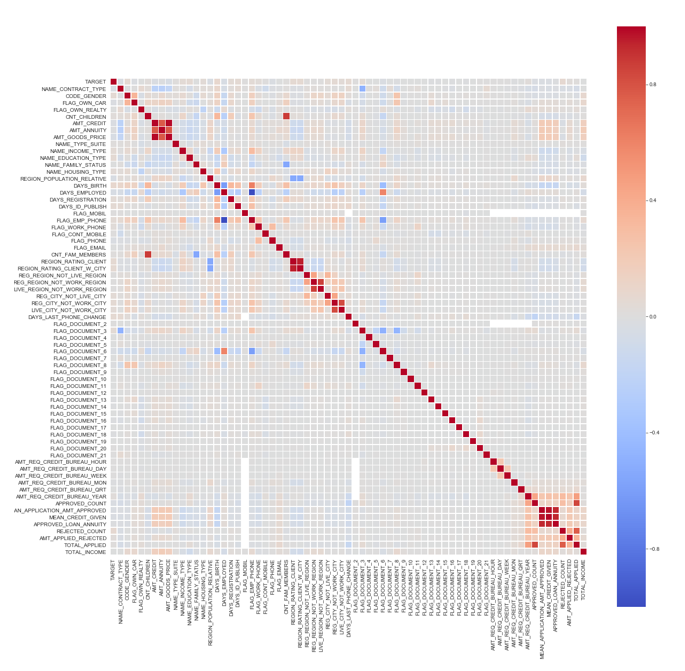
When looking at the visualizations of the price of goods versus credit extended, we can see that the two marry up closely, indicating that the credit is tightly aligned with the cost of the good being purchased with little distinction between both target values.



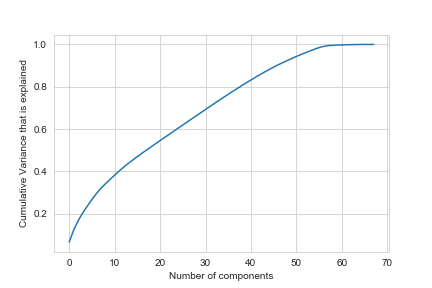
For those with a loan history, examining the visualization of current credit extended versus the mean credit they received in previous applications, we see that generally the current credit is greater than what was historical provided in previous loans. Those who struggle to repay the loan seem to be offered more credit than previously provided which supports the theory of becoming overextended with this new loan.



The vast majority of applicants are in a relationship, providing an additional level of financial support. The one point of concern may be within the Pensioner category, as there is roughly 20% of Pensioner applicants who are single, there are no additional members of the family who could potentially bring in income. These individuals will have to rely on social security payments and existing assets or investments to bring in income.



When looking at the correlation between all of the available features and the target feature, we see that there is minor correlation with gender, days since birth (age), income type, education, the region ratings of the client and count of previously rejected loan applications.



When utilizing PCA, we can see that roughly 10 features compose 40% of the variance within the dataset, adding an additional 10 features with incorporate an increment 15%, having 20 features account for 55% of the variance within the dataset.

# Approach

**Step 1: Feature elimination using domain knowledge**

Using domain knowledge of the lending industry, we look to slim the dataset down to relevant features. Within the current applicant & previous applications tables, there are features which are irrelevant to predicting payment challenges. Within the current applicant dataset, there are features which could cause potential ethical and legal issues. Additionally, within the previous applications table there are features which are irrelevant to future loans and applications.

Any personally identifying information or primary keys will be removed after the two tables are joined.

**Step 2: Data pre-processing & cleaning**

Following the removal of irrelevant features, we look to clean the dataset. For any categorical or qualitative data, these values will be reassigned using encoding to reduce errors when training the models. For any remaining numerical values and non-binary values, these remaining values will be normalized.

For any columns where there are entirely homogeneous values, an abundance of NULL or missing values, these columns will be removed as they are not adequate for machine learning purposes (principle of garbage in – garbage out).

**Step 3: Joining tables**

Following the initial feature selection for the previous applications table, the remaining features will be aggregated, summarized and appended onto the current applications table as a series of new columns.

Following the joining, any identification and other unneeded variables will be removed.

**Step 4: Data Normalization & Encoding**

Once the full dataset is consolidated, containing the features from the current application and previous application datasets, categorical features will have labels encoded and the full dataset will be normalized.

Categorical labels will be encoded using LabelEncoder function, any boolean features using non-binary values will be converted to binary values to allow for normalization of the full dataframe.

**Step 5: Upsampling**

One final step will take place during the pre-processing phase. Due to the unbalanced nature of the dataset where the target feature contains 282,686 0’s and 24,825 1’s, the dataset will need to be downsampled or upsampled to increase model performance.

Opting to upsample the data, the full dataset is split into two dataframes, one containing records for all 0s in the target column and another dataset for all the 1s in the target column. The dataframe containing the 1 values are then resampled, raising the number of records to 282,686, creating an equal proportion of the target values.

**Step 6: Feature Selection**

Post pre-processing, key features will be identified for each model. Learnings from the literature review indicate that for logistic regression and random forest models, dimensionality reduction increases model performance. Each model will be allowed to select its own key features, with any unselected features being removed.

Each model will select its own key features, preparing a model utilizing only those features selected. Based on the Principal Component Analysis conducted during the EDA phase, we can identify that there should be a maximum of 20 key features selected for each model.

**Step 7: Model development**

Sklearn will be used to develop the random forest and logistic regression models, utilizing the prepared dataset from the steps above. The xgboost library will be utilized to develop gradient boosted trees, providing 3 different models to approach this problem.

The training data will 50/50 split of classifiers to reduce potential overfitting of the model and provides a balanced dataset for the models to train on.

**Step 8: Model scoring**

Following the development of these 3 models, we can use precision, accuracy, F1 scores and the area under curve to identify if a specific model clearly outperforms the others, placing greater importance on the area under curve and F1 scores.

# Results

In preparation of developing the various machine learning models, the full dataset is split into a training and testing dataset, proportioning 80% for the training data and 20% for the testing dataset. Both datasets contain an equal proportion of both classes in the target feature.

The logistic regression and random forest models will select the most important features within the dataset. Two types of models will be developed for each technique, providing performance results with and without feature selection incorporated. Based on the best parameters found using GridSearch, both models will use the same parameters.

GridSearch will be employed to explore a variety of different parameters for the various machine learning techniques, prioritizing the highest area under the curve scores to select the best model parameters.

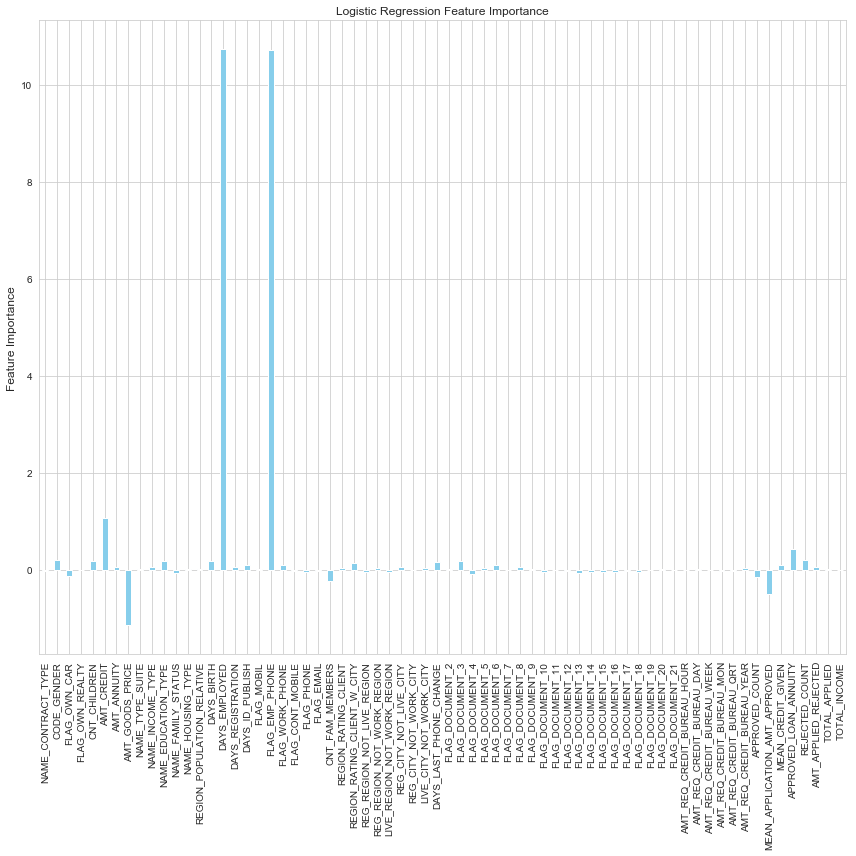
Finally, all models will employ K Folds cross-validation in each model, experimenting with incorporating between 5 and 10 folds to increase model accuracy.

**Logistic Regression**

For the logistic regression model, utilizing GridSearch, two parameters were experimented with, the learning rate, ranging from values of 0.001 to 0.1 as well as l1 and l2 regularization. A learning rate of 0.1 and l1 regularization was chosen as the best parameters.

Allowing the logistic regression model to identify the most significant features, the following 6 features are selected as the most important.

|  |  |
| --- | --- |
| Feature | Feature Description |
| AMT\_CREDIT | For the current loan, how much credit is extended |
| AMT\_GOODS\_PRICE | For the current loan, what is the price of the goods being purchased |
| DAYS\_EMPLOYED | Number of days the applicant has been employed |
| FLAG\_EMP\_PHONE | Boolean value for whether applicant provided a work phone alongside the application |
| MEAN\_APPLICATION\_AMT\_APPROVED | For prior loans, what is the mean amount the applicant has previously applied for? |
| APPROVED\_LOAN\_ANNUITY | For prior loans, what was the mean amount to be repaid |

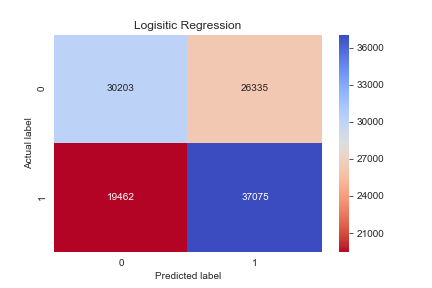


Of the various features, days employed and the flag for whether the applicant provided their work phone are most critical to the logistic regression model. Aside from the 6 features identified above, the other features are less impactful.

Examining the impact of dimensionality reduction, the model which did not utilize feature selection was a superior performer. Interestingly, the removal of insignificant features does not boost the performance of the model.

Below we compare the results of the two logistic regression models, with the model incorporating feature selection on the left-hand side. Within the confusion matrices, charts A1 and A2, the classification of 0’s are the most significant differentiator in terms of both model’s performance. The model incorporating no feature selection is more effective at correctly classifying individuals who are able to successfully repay the loan, while being incrementally less effective at correctly classifying those who will have repayment challenges.

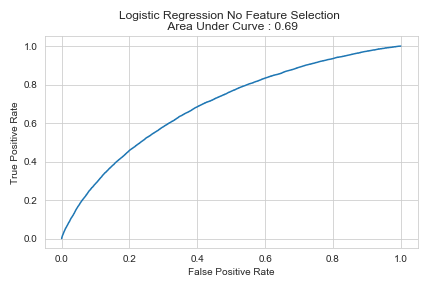
Within a business context, it would be wiser to correctly classify an incremental 6,000 applicants as having no issues compared to incrementally identifying 600 of those who would have issues. In this scenario, significant loss in the model’s precision would detrimentally affect revenue in contrast with marginally decreasing loan delinquency rates.



A2

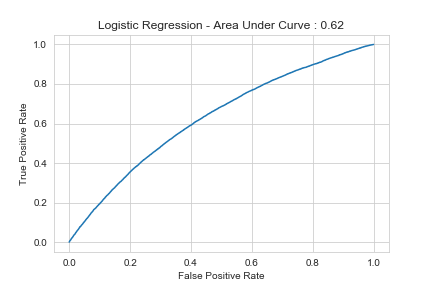
A1





A4

A3



When comparing the area under curve performance in charts A3 & A4, the no feature selected model performs superiorly, accurately identifying the class of each record 69% of the time, an incremental 7% higher than the feature selection model.

Expanding the comparison of results to incorporate precision, accuracy and F1 scores, we can see the following results for the two models.

|  |  |  |
| --- | --- | --- |
| Metric | Logistic Regression  Feature Selection | Logistic Regression  No Feature Selection |
| Accuracy | 59% | 64% |
| Precision | 61% | 64% |
| F1 Score | 57% | 64% |
| Area Under Curve | 62% | 69% |

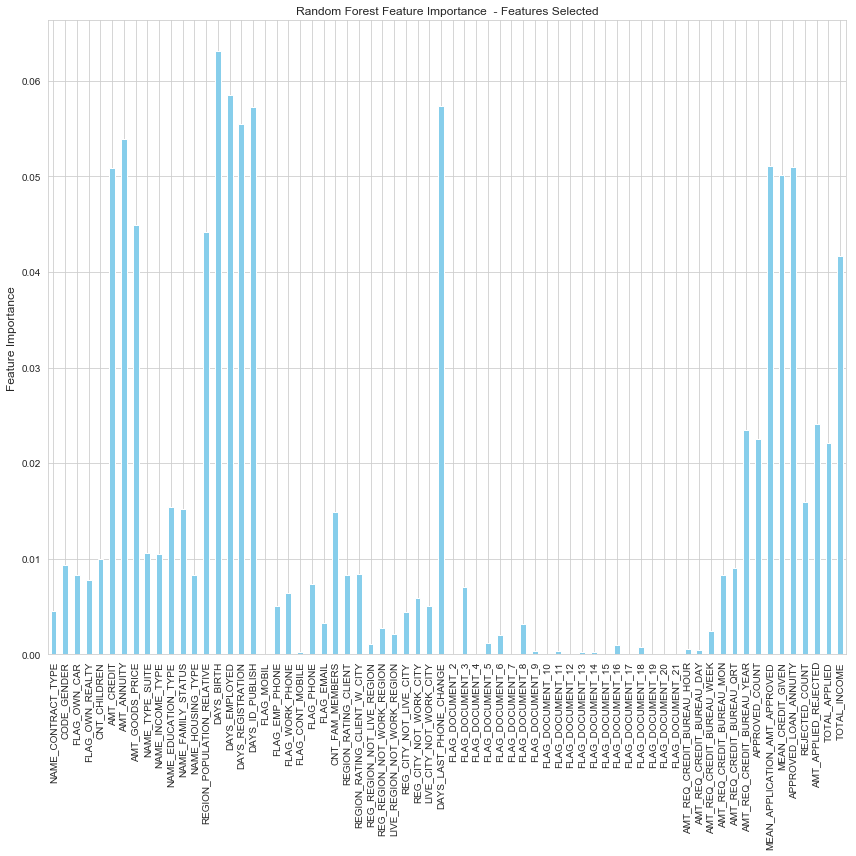
Based on the performance results in table above, stronger performance is achieved by not incorporating feature selection. This results in the model accurately classifying individuals 64% of the time and sees the same performance across precision and recall scores.

Holistically, the logistic regression model with no features selected outperforms the model that incorporates feature selection. Incorporating k fold cross-validation into the models yields no performance increase when comparing 5 to 10 K Folds being incorporated.**Random Forest**

When rigorously testing for the strongest performing parameters, GridSearch was utilized, experimenting with the number of estimators, max depth of the trees, the max number of features, and the criterion function for the quality of each split. Ultimately, the strongest performing model utilized 400 estimators, a max depth of 20 features, automatically chosen max features and the Gini function for the criterion.

Similarly, two random forest classifier models are developed, one incorporating feature selection and one without. The random forest classifier has identified 21 features as being significant to the model, with features revolving around the current loan, socio-demographic data, documentation details submitted during the application process and finally details around their previous loans.

|  |  |
| --- | --- |
| Feature | Feature Description |
| AMT\_CREDIT | For the current loan, how much credit is extended |
| AMT\_ANNUITY | For the current loan, how much must be repaid |
| AMT\_GOODS\_PRICE | For the current loan, what is the price of the goods being purchased |
| NAME\_EDUCATION\_TYPE | Highest level of education completed by applicant |
| NAME\_FAMILY\_STATUS | Relationship status of applicant (i.e. married, single, etc.) |
| REGION\_POPULATION\_RELATIVE | Rating of population density where the applicant lives |
| DAYS\_BIRTH | Client’s age in days relative to application date |
| DAYS\_EMPLOYED | Number of days the applicant has been employed |
| DAYS\_REGISTRATION | Number of days since applicant changed registration details relative to application date |
| DAYS\_ID\_PUBLISH | Number of days since applicant changed ID documents relative to the application date |
| CNT\_FAM\_MEMBERS | Number of family members the applicant has |
| DAYS\_LAST\_PHONE\_CHANGE | How many days since the applicant changed their phone number |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | Number of credit bureau enquiries for the applicant within a year of the application date |
| APPROVED\_COUNT | Total count of previously approved loans |
| MEAN\_APPLICATION\_AMT\_APPROVED | Average amount of credit that applicant had applied for based on previous loan history where they were approved |
| MEAN\_CREDIT\_GIVEN | Average amount of credit the applicant was given based on previous loan history where they were approved |
| APPROVED\_LOAN\_ANNUITY | Average amount needing to be repaid based on previous loan history where they were approved |
| REJECTED\_COUNT | Total count of previously rejected loans |
| AMT\_APPLIED\_REJECTED | Average amount of credit that applicant had applied for based on previous loan history where they were rejected |
| TOTAL\_APPLIED | Total count of all previous loan applications submitted |
| TOTAL\_INCOME | Total yearly income of the client |



The most important features identified by the model are age, days employed, days since the last phone change, change to registration details and change to government ID and other documentation. Details around the current loan as well as prior loan history are also identified are highly significant to the model.

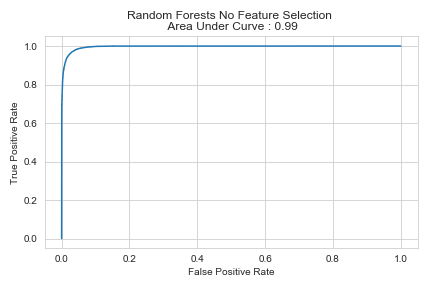
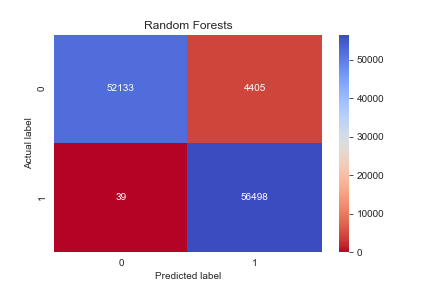
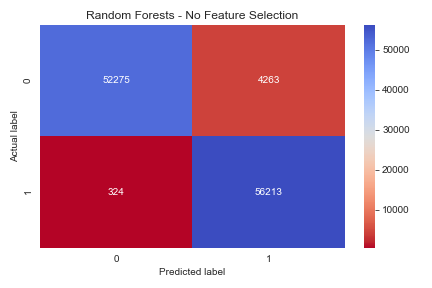
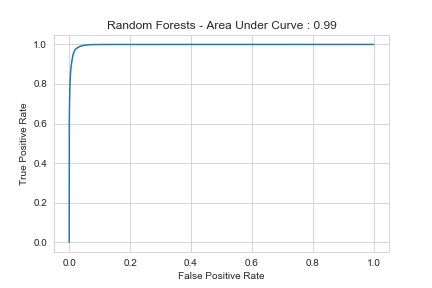
Comparing the performances of both models in charts B1 and B2, both have performed strongly. The model incorporating feature selection offers marginal performance gains, correctly classifying those with more accurately classifying those with repayment issues. Holistically, both performed near identically with the overall performance of the model which incorporated feature selection performing marginally better. However, both models correctly identified the class of an applicant and successfully classified them with 99% accuracy.

B4

B3

B2

B1

Analyzing the various performance metrics of the random forest models, the results for the performance metrics remain similar with only precision results differing.

|  |  |  |
| --- | --- | --- |
| Metric | Random Forest  Feature Selection | Random Forest  No Feature Selection |
| Accuracy | 96% | 96% |
| Precision | 100% | 99% |
| F1 Score | 96% | 96% |
| Area Under Curve | 99% | 99% |

Both models are able to correctly predict True Positive and False Positive values 96% of the time. Both models near perfectly classify individuals who will successfully repay the loan. Both models face greater challenges with correctly classifying applicants who will have issues with loan repayment.

Both models also have strong F1 scores, indicating there are exceptional results within both recall and precision. As the precision of the model is 99%, relative to the F1 score of 96%, we can identify that recall is the metric where performance is weaker between the two. The weaker performance in recall indicates that the model has challenges with False Negatives.

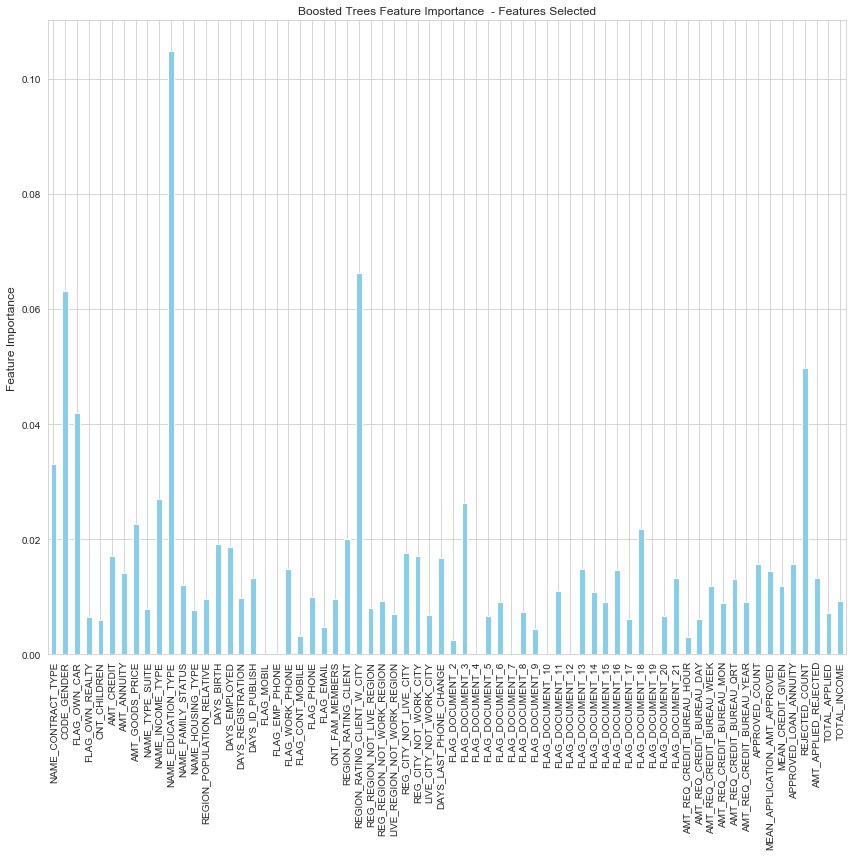
With regards to overall identification and classification of both types of applicants, the random forest models are able to correctly classify the corresponding class 99% of the time.

When incorporating K fold cross-validation, 5 folds provided the same results as incorporating 6 to 10 folds. With 5 folds providing consistent results as models with a greater number of folds, 5 folds was chosen to minimize run times.

**Gradient Boosted Trees**

When searching for the strongest performing parameters; the learning rate, the max depth and number of estimators were experimented with to identify the greatest area under curve score. Searching through the combinations of parameters, 400 estimators with a max depth of 20 and a learning rate of 0.1 was chosen as the parameters for the strongest performing model.

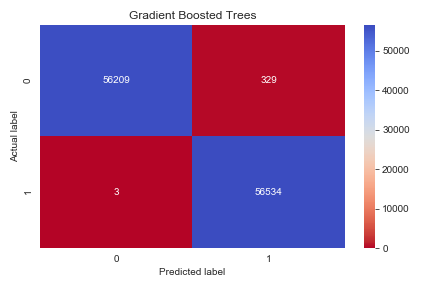
For the gradient boosted trees, no feature selection is incorporated as it’s renown for strong performance amongst high dimensionality datasets. Strong performance while retaining high dimensionality provides a functional model allowing for greater interpretability and more robustness when deployed, accommodating a large breadth of different datapoints.



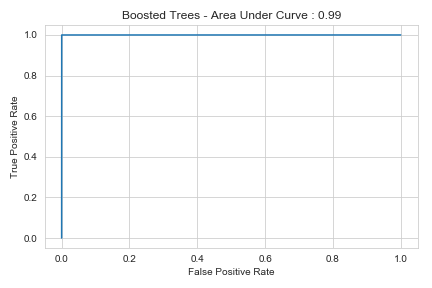
|  |  |
| --- | --- |
| Feature | Feature Description |
| NAME\_CONTRACT\_TYPE | Whether it is an open loan or a loan for specific goods |
| CODE\_GENDER | Gender of applicant |
| FLAG\_OWN\_CAR | Boolean value for applicant owning a vehicle |
| NAME\_EDUCATION\_TYPE | Highest level of education completed by applicant |
| REGION\_RATING\_CLIENT\_W\_CITY | Rating of the region and city where the applicant lives |
| REJECTED COUNT | Total count of previously rejected loans |

Reviewing the importance of features with the gradient boosted trees model, 6 features are the primary drivers within the model. The type of contract; a loan specific to goods to be purchased or an open loan, the applicant’s gender, whether they own a car, their highest level of education completed, the credit rating for their region & city and the amount of previously rejected loans they have are identified as most important to the model.

When analyzing the confusion matrix results in chart B1, the model performs amicably, only mis-classifying 332 applicants overall. Nearly all misclassifications occurring where good applicants are misclassified as having repayment issues.



B1



B2

Based on the charts above, the model performs near perfectly across all metrics. Correctly classifying applicants near 100% of the time, indicating that there may be problems with overfitting.

|  |  |
| --- | --- |
| Metric | Gradient Boosted Trees |
| Accuracy | 99% |
| Precision | 100% |
| F1 Score | 100% |
| Area Under Curve | 99% |

Utilizing differing amounts of K folds ranging between 5 to 10 folds, 5 folds was the optimal number of folds. Utilizing 5 folds returned the same results as a greater number of folds but benefitted from quicker run times.

# Conclusion

In conclusion, after developing the various machine learning models and analyzing their results, random forests and gradient boosted trees had performed strongest based on accuracy, precision, F1 and area under curve scores.

Within this specific project, while feature selection did improve model performance, it had made fairly minimal impact on the random forest classifiers. Due to the nature of decision tree-based algorithms, particularly random forests, which generates a multitude of random combinations of features to create decision trees, the nature of feature selection may see less impact in results due to the final results being chosen democratically. Additionally, incorporating the max depth parameter creates a natural feature selection like parameter, which creates a maximum ceiling for the number of features that could be used in the trees.

Comparing results between the random forest classifier and the gradient boosted tree model, the results are similar with the xgboost model performing marginally stronger. My inclination would be performance is superior due to how xgboost boosted tree models are constructed, by democratically selecting each layer of the decision tree. In contrast to random forests which the algorithm independently develops each decision tree and then concludes the best model based on the majority of similarly oriented trees.

The inclusion of K Fold cross-validation aided the development in models, increasing overall accuracy. However, experimenting with the various machine learning techniques, utilizing 5 folds provided the same results as incorporating 6 to 10 folds. Utilizing 5 folds in the final models allowed for increased run times and returned the same results.

With the performance of random forest and boosted tree models performing nearly perfectly, overfitting may be an issue present within the models. Experimenting with using SMOTE and under sampling techniques may provide more realistic results. Additionally, experimenting with lowering the number of estimators and the learning rate may also support with lowering potential overfitting.

# References

Vieira, J. R., Barboza, F., Sobreiro, V. A., & Kimura, H. (2019). Machine learning models for credit analysis improvements: Predicting low-income families’ default. *Applied Soft Computing,* *83*, 105640. doi:10.1016/j.asoc.2019.105640

Liang, Y., Jin, X., & Wang, Z. (2019). Loanliness: Predicting Loan Repayment Ability by Using Machine Learning Methods. *Loanliness: Predicting Loan Repayment Ability by Using Machine Learning Methods*. Retrieved June 8, 2020, from https://pdfs.semanticscholar.org/b3e2/836b546e80f537c50c16cc57ab426b1e2adf.pdf

Ma, X., Sha, J., Wang, D., Yu, Y., Yang, Q., & Niu, X. (2018). Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications,* *31*, 24-39. doi:10.1016/j.elerap.2018.08.002

Stolfo, S. J., Fan, D. W., Lee, W., & Prodromidis, A. L. (n.d.). Credit Card Fraud Detection Using Meta-Learning: Issues ... Retrieved June 8, 2020, from https://www.aaai.org/Papers/Workshops/1997/WS-97-07/WS97-07-015.pdf

Abhishek, A. (2018). *Predicting Loan Defaults using Machine Learning Techniques* (Unpublished master's thesis). CALIFORNIA STATE UNIVERSITY, NORTHRIDGE. Retrieved June 8, 2020, from http://scholarworks.csun.edu/bitstream/handle/10211.3/203343/Bhagat-Abhishek-thesis-2018.pdf?sequence=1

Addo, P., Guegan, D., & Hassani, B. (2018). Credit Risk Analysis Using Machine and Deep Learning Models. *Risks,* *6*(2), 38. doi:10.3390/risks6020038

Coser, A., Maer-Matei, M. M., & Albu, C. (2019). Predictive Models for Loan Default Risk Assessment. *Economic Computation And Economic Cybernetics Studies And Research,* *53*(2/2019), 149-165. doi:10.24818/18423264/53.2.19.09

Niu, B., Ren, J., & Li, X. (2019). Credit Scoring Using Machine Learning by Combing Social Network Information: Evidence from Peer-to-Peer Lending. *Information,* *10*(12), 397. doi:10.3390/info10120397

Mishra5001. (2019, July 15). Credit Card Fraud Detection. Retrieved June 08, 2020, from <https://www.kaggle.com/mishra5001/credit-card>

# Appendix

**Current Applicants Table**

The following table contains all of the features contained within the primary dataset, current\_applicants.csv. This dataset contains 122 different features consisting data around the loan, details about the client, covering socio-economic, housing and application details.

|  |  |  |  |
| --- | --- | --- | --- |
| SK\_ID\_CURR | | ID of loan in our sample | |
| TARGET | | Target variable - Does the client have late or missed payments? (1 = TRUE, 0 = FALSE) | |
| NAME\_CONTRACT\_TYPE | | Cash loan or revolving credit | |
| CODE\_GENDER | | Gender of the client | |
| FLAG\_OWN\_CAR | | Flag if the client owns a car | |
| FLAG\_OWN\_REALTY | | Flag if client owns a real estate asset | |
| CNT\_CHILDREN | | Number of children the client has | |
| AMT\_INCOME\_TOTAL | | Income of the client | |
| AMT\_CREDIT | | Credit amount of the loan | |
| AMT\_ANNUITY | | Loan annuity | |
| AMT\_GOODS\_PRICE | | Cost of goods that the loan was used for (if applicable) | |
| NAME\_TYPE\_SUITE | | Who was accompanying client when he was applying for the loan | |
| NAME\_INCOME\_TYPE | | Employment classification | |
| NAME\_EDUCATION\_TYPE | | Level of highest education the client achieved | |
| NAME\_FAMILY\_STATUS | | Family status of the client | |
| NAME\_HOUSING\_TYPE | | Living arrangement | |
| REGION\_POPULATION\_RELATIVE | | Normalized population of region where client lives (higher number means the client lives in more populated region) | |
| DAYS\_BIRTH | | Client's age in days at the time of application | |
| DAYS\_EMPLOYED | | How many days before the application the person started current employment | |
| DAYS\_REGISTRATION | | How many days before the application did client change his registration | |
| DAYS\_ID\_PUBLISH | | How many days before the application did client change the identity document with which he applied for the loan | |
| OWN\_CAR\_AGE | | Age of client's car | |
| FLAG\_MOBIL | | Did client provide their mobile phone (1 = TRUE, 0 = FALSE) | |
| FLAG\_EMP\_PHONE | | Did client provide their work phone (1 = TRUE, 0 = FALSE) | |
| FLAG\_WORK\_PHONE | | Did client provide their home phone (1 = TRUE, 0 = FALSE) | |
| FLAG\_CONT\_MOBILE | | Was mobile phone reachable (1 = TRUE, 0 = FALSE) | |
| FLAG\_PHONE | | Did client provide their home phone (1 = TRUE, 0 = FALSE) | |
| FLAG\_EMAIL | | Did client provide an email (1 = TRUE, 0 = FALSE) | |
| OCCUPATION\_TYPE | | Client's occupation | |
| CNT\_FAM\_MEMBERS | | Count of client's family members | |
| REGION\_RATING\_CLIENT | | Internal rating of the region where client lives | |
| REGION\_RATING\_CLIENT\_W\_CITY | | Internal rating of the city & region where client lives | |
| WEEKDAY\_APPR\_PROCESS\_START | | On which day of the week did the client apply for the loan | |
| HOUR\_APPR\_PROCESS\_START | | Approximately at what hour did the client apply for the loan | |
| REG\_REGION\_NOT\_LIVE\_REGION | | Flag if client's permanent address does not match contact address (1=different, 0=same, at region level) | |
| REG\_REGION\_NOT\_WORK\_REGION | | Flag if client's permanent address does not match work address (1=different, 0=same, at region level) | |
| LIVE\_REGION\_NOT\_WORK\_REGION | | Flag if client's contact address does not match work address (1=different, 0=same, at region level) | |
| REG\_CITY\_NOT\_LIVE\_CITY | | Flag if client's permanent address does not match contact address (1=different, 0=same, at city level) | |
| REG\_CITY\_NOT\_WORK\_CITY | | Flag if client's permanent address does not match work address (1=different, 0=same, at city level) | |
| LIVE\_CITY\_NOT\_WORK\_CITY | | Flag if client's contact address does not match work address (1=different, 0=same, at city level) | |
| ORGANIZATION\_TYPE | | Type of organization where client works | |
| EXT\_SOURCE\_1 | | Normalized score from external data source | |
| EXT\_SOURCE\_2 | | Normalized score from external data source | |
| EXT\_SOURCE\_3 | | Normalized score from external data source | |
| APARTMENTS\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| BASEMENTAREA\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| YEARS\_BEGINEXPLUATATION\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| YEARS\_BUILD\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| COMMONAREA\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| ELEVATORS\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| ENTRANCES\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| FLOORSMAX\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| FLOORSMIN\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| LANDAREA\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| LIVINGAPARTMENTS\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| LIVINGAREA\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| NONLIVINGAPARTMENTS\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| NONLIVINGAREA\_AVG | | Normalized information about building where the client lives. What is average apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| APARTMENTS\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| BASEMENTAREA\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| YEARS\_BEGINEXPLUATATION\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| YEARS\_BUILD\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| COMMONAREA\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| ELEVATORS\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| ENTRANCES\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| FLOORSMAX\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| FLOORSMIN\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| LANDAREA\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| LIVINGAPARTMENTS\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| LIVINGAREA\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| NONLIVINGAPARTMENTS\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| NONLIVINGAREA\_MODE | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| APARTMENTS\_MEDI | | Normalized information about building where the client lives. What is modal apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| BASEMENTAREA\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| YEARS\_BEGINEXPLUATATION\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| YEARS\_BUILD\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| COMMONAREA\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| ELEVATORS\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| ENTRANCES\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| FLOORSMAX\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| FLOORSMIN\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| LANDAREA\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| LIVINGAPARTMENTS\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| LIVINGAREA\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| NONLIVINGAPARTMENTS\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| NONLIVINGAREA\_MEDI | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| FONDKAPREMONT\_MODE | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| HOUSETYPE\_MODE | | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| TOTALAREA\_MODE | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| WALLSMATERIAL\_MODE | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| EMERGENCYSTATE\_MODE | Normalized information about building where the client lives. What is median apartment size, common area, living area, age of building, number of elevators, number of entrances, state of the building, number of floors | |
| OBS\_30\_CNT\_SOCIAL\_CIRCLE | Count of payments 30 days past due within client's social circle | |
| DEF\_30\_CNT\_SOCIAL\_CIRCLE | Count of 30 days past due defaults within client's social circle | |
| OBS\_60\_CNT\_SOCIAL\_CIRCLE | Count of payments 60 days past due within client's social circle | |
| DEF\_60\_CNT\_SOCIAL\_CIRCLE | Count of 60 days past due defaults within client's social circle | |
| DAYS\_LAST\_PHONE\_CHANGE | How many days before application did client change phone | |
| FLAG\_DOCUMENT\_2 | Did client provide document # 2 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_3 | | Did client provide document # 3 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_4 | | Did client provide document # 4 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_5 | | Did client provide document # 5 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_6 | | Did client provide document # 6 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_7 | | Did client provide document # 7 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_8 | | Did client provide document # 8 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_9 | | Did client provide document # 9 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_10 | | Did client provide document # 10 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_11 | | Did client provide document # 11 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_12 | | Did client provide document # 12 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_13 | | Did client provide document # 13 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_14 | | Did client provide document # 14 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_15 | | Did client provide document # 15 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_16 | | Did client provide document # 16 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_17 | | Did client provide document # 17 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_18 | | Did client provide document # 18 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_19 | | Did client provide document # 19 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_20 | | Did client provide document # 20 (1 = TRUE, 0 = FALSE) | |
| FLAG\_DOCUMENT\_21 | | Did client provide document # 21 (1 = TRUE, 0 = FALSE) | |
| AMT\_REQ\_CREDIT\_BUREAU\_HOUR | | Count of credit bureau enquiries regarding the client one hour before application | |
| AMT\_REQ\_CREDIT\_BUREAU\_DAY | | Count of credit bureau enquiries regarding the client one day before application (excluding one hour before application) | |
| AMT\_REQ\_CREDIT\_BUREAU\_WEEK | | Count of credit bureau enquiries regarding the client one week before application (excluding one day before application) | |
| AMT\_REQ\_CREDIT\_BUREAU\_MON | | Count of credit bureau enquiries regarding the client one month before application (excluding one week before application) | |
| AMT\_REQ\_CREDIT\_BUREAU\_QRT | | Count of credit bureau enquiries regarding the client 3 month before application (excluding one month before application) | |
| AMT\_REQ\_CREDIT\_BUREAU\_YEAR | | Count of credit bureau enquiries regarding the client one day year (excluding last 3 months before application) | |

**Previous Applications Table**

The following table contains all of the features contained within the supplementary dataset, previous\_applications.csv. This dataset contains 38 different features consisting of details from client’s previous applications. This dataset contains datapoints around the approval status of a previous application, the details of the loan, supportive application information and dates around the first & last drawing and first payment.

|  |  |
| --- | --- |
| Feature | Description |
| SK\_ID\_PREV | ID of previous loan within database (both successful and unsuccessful applications) |
| SK\_ID\_CURR | ID of loan in our sample |
| NAME\_CONTRACT\_TYPE | Type of previous loan |
| AMT\_ANNUITY | Annuity of previous application |
| AMT\_APPLICATION | Amount of credit that the client has previously requested |
| AMT\_CREDIT | Amount of credit that the client had been previously approved for |
| AMT\_DOWN\_PAYMENT | Down payment on the previous application |
| AMT\_GOODS\_PRICE | Goods price of good that client asked for (if applicable) on the previous application |
| WEEKDAY\_APPR\_PROCESS\_START | On which day of the week did the client apply for previous application |
| HOUR\_APPR\_PROCESS\_START | Approximately at what day hour did the client apply for the previous application |
| FLAG\_LAST\_APPL\_PER\_CONTRACT | Flag if it was last application for the previous contract. Sometimes by mistake of client or our clerk there could be more applications for one single contract |
| NFLAG\_LAST\_APPL\_IN\_DAY | Flag if the application was the last application per day of the client. Sometimes clients apply for more applications a day. Rarely it could also be error in our system that one application is in the database twice |
| NFLAG\_MICRO\_CASH | Flag Micro finance loan |
| RATE\_DOWN\_PAYMENT | Down payment rate normalized on previous credit |
| RATE\_INTEREST\_PRIMARY | Interest rate normalized on previous credit |
| RATE\_INTEREST\_PRIVILEGED | Interest rate normalized on previous credit |
| NAME\_CASH\_LOAN\_PURPOSE | Purpose of the cash loan |
| NAME\_CONTRACT\_STATUS | Previous loan status (approved, cancelled, etc.) |
| DAYS\_DECISION | Days between current application and approval decision of previous application |
| NAME\_PAYMENT\_TYPE | Payment method that client chose to pay for the previous application |
| CODE\_REJECT\_REASON | Reason previous application was rejected |
| NAME\_TYPE\_SUITE | Who accompanied the client when applying for the previous application |
| NAME\_CLIENT\_TYPE | Was the client old or new client when applying for the previous application |
| NAME\_GOODS\_CATEGORY | What kind of goods did the client apply for in the previous application |
| NAME\_PORTFOLIO | Intended use for previous loan |
| NAME\_PRODUCT\_TYPE | Was the previous application a cross-sell or walk-in application? |
| CHANNEL\_TYPE | Channel the client was acquired through on the previous application |
| SELLERPLACE\_AREA | Location of the seller of goods which the client intended to purchase from |
| NAME\_SELLER\_INDUSTRY | The industry of the seller |
| CNT\_PAYMENT | Term of previous credit at application of the previous application |
| NAME\_YIELD\_GROUP | Grouped interest rate into small medium and high of the previous application |
| PRODUCT\_COMBINATION | Detailed product combination of the previous application |
| DAYS\_FIRST\_DRAWING | Days since the first disbursement of the previous application |
| DAYS\_FIRST\_DUE | Days since the first due payment of the previous application |
| DAYS\_LAST\_DUE\_1ST\_VERSION | Days since the first payment of the previous application |
| DAYS\_LAST\_DUE | Days since the last due date of the previous application |
| DAYS\_TERMINATION | Days since the expected termination of the previous application |
| NFLAG\_INSURED\_ON\_APPROVAL | Did the client requested insurance during the previous application |