Loan Data Solution

In Supporting Loan Application

Business Understanding and Analitycal approach

In this critical phase, data scientists formulate an analytical approach, considering statistical techniques and machine learning, after clarifying the target customers and understanding business needs. The success of a project hinges on the quality of questions posed, emphasizing the importance of asking the right questions to narrow down data acquisition. Statistical problem clarification aids in determining the necessary trends for an efficient problem resolution, with various approaches like predictive modeling, descriptive analysis, and statistical analysis considered based on needs. Analyzing loan data from 2007-2014, the objective is to predict individuals' ability to repay loans based on their background, addressing potential challenges with respective interest rates. The ultimate goal is prediction, leading to the utilization of a machine learning approach, utilizing loan data from the United States during the specified period.

Data Requirements and Collections

LAs stated before, this loan application data is from United States with period of 2007-2014 and is accessible in Kaggle at https://www.kaggle.com/datasets/devanshi23/loan-data-2007-2014

Data Understanding

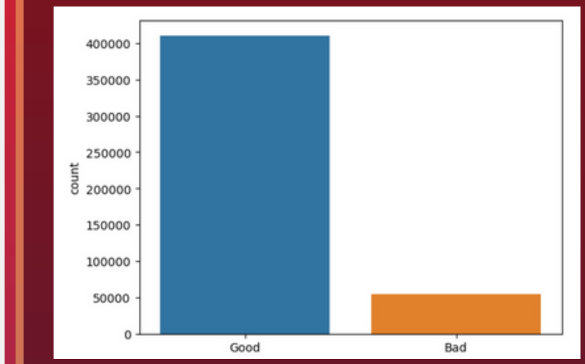
The data consists of 466,285 rows with 73 features and 1 target variable. Upon inspection, it was found that many data entries do not conform to their expected types, for instance, 'emp_length' is still in text format, among other issues. Additionally, there are numerous missing values, exceeding 40%. Despite the absence of duplicate data, the process of data cleaning is prolonged due to these challenges.

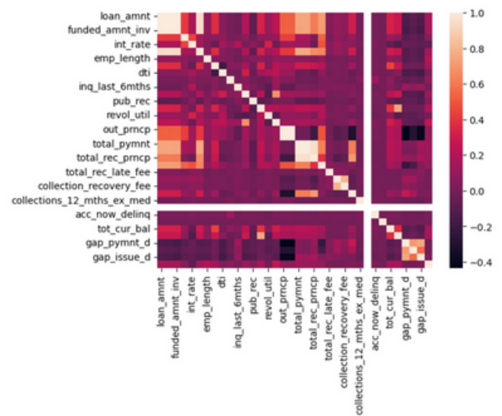
Data Preparation

In this step, data preparation is performed, which includes removing certain columns based on criteria such as: missing values > 40%, single unique values, those not deemed necessary based on common sense, not used in modeling, having too many unique values, and being purely text-based. After carrying out those steps, the process continues with the creation of preprocessing for text and date data, which should be categorized as well as numerical. Next is the deletion of variables and imputation for missing values. If there are no more missing values, the final step is to categorize loans into good and bad based on information in the metadata. ['Current', 'Fully Paid', 'Does not meet the credit policy. Status:Fully Paid'] are categorized as good loans because they have been fully paid. Meanwhile, ['Charged Off', 'Late (31-120 days)', 'In Grace Period', 'Late (16-30 days)', 'Default', 'Does not meet the credit policy. Status:Charged Off'] represent problematic loans.

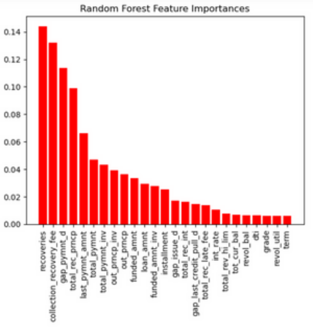
EDA

Two crucial aspects in this Exploratory Data Analysis (EDA) process are understanding the distribution of the target variable and assessing its correlation. Below is a bar chart indicating an imbalance between good and bad loans, which is unfavorable in machine learning. Secondly, in the heatmap and correlation analysis, the generated correlations are quite poor and excessive, necessitating an alternative method for feature selection, namely, the use of wrapping methods. Ps: The graph can be see on the next page





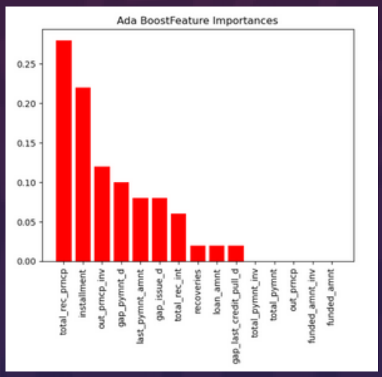
Feature Engineering and Selection

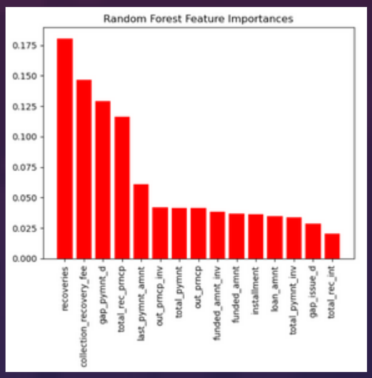


It's time; in this process, some target variables that are still in text form are converted into numerical values with certain criteria: label encoder for ordinal data and dummy encoder for nominal variables. The data is then separated into target and features, and scaling is performed to avoid bias in modeling, using the standard scaler. Perform undersample sampling to address imbalanced data and conduct a train-test split with a ratio of 0.2. Based on the results from the random forest, several features have significant importance, as shown in the above bar plot. Next, the top 15 features with the highest feature importances will be selected using both Random Forest and Ada Boost.

Modelling

It is found that the random forest model performed the best with an accuracy of 98.6%, slightly higher than Ada Boost (98,15%). Interestingly, there is a slight difference in the feature importance between the two models. The Random Forest results indicate that all 15 features contribute to their importance. In contrast, in Ada Boost, the last 5 features do not have any contribution to importance, with distinct feature importance values.





Findings

Best model is obtained using Random Forest, slightly higher than Ada Boost. The features ['recoveries', 'collection_recovery_fee', 'gap_pymnt_d', 'total_rec_prncp', 'last_pymnt_amnt', 'out_prncp_inv', 'total_pymnt', 'out_prncp', 'funded_amnt_inv', 'funded_amnt', 'installment', 'loan_amnt', 'total_pymnt_inv', 'gap_issue_d', 'total_rec_int'] hold the highest importance according to Random Forest. Meanwhile, according to Ada Boost, ['total_rec_prncp', 'installment', 'out_prncp_inv', 'gap_pymnt_d', 'last_pymnt_amnt', 'gap_issue_d', 'total_rec_int', 'recoveries', 'loan_amnt', 'gap_last_credit_pull_d'] are the most important features."

Suggestions

For stakeholders, this can be implemented in your banking system to facilitate bankers in selecting loan applications. Additionally, you can communicate these findings to bankers, encouraging them to focus more on the identified important features during the loan application evaluation process. As for further development, computational limitations may have hindered optimal Feature Selection and Hyperparameter Tuning. In the future, this aspect can be addressed. The use of Auto-SkLearn is also highly recommended to explore models with the most effective methods.