Loan Prediction on LendingClub Issued Loans Dataset

Alex Marin Felices
U162444
17/12/21

Introduction and motivation

Loans are a huge important part of adult life. To buy a car, a home, to get money for education, health treatments, … they are used by millions of people and are very important in today's life. It is not only important for the companies providing the loans, it is also important for the people that issue a request for a loan. It is important for the people investing to know if they are likely to get their money back or not, but it is also important for the people who ask for it to know if they will be able to give it all back or if they will get in debt for life.

Lots and lots of people end up with huge amounts of debt because they do not really know if they will be able to pay it back. In this case, I will be trying to help people who are trying to invest their money in giving loans to people who may need it, and so they can help people, while also earning some money without taking huge risks.

Description, objectives and expected benefits of the project

At the start of the project I defined some of the goals and objectives I proposed to accomplish during the project. Below it can be seen a general overview of the definition of the project as well as some expected benefits.

- Develop a classification model to predict if a future customer will pay back the loan
- Create a training dataset formed by internal data without outliers
- Create a cluster method to identify main characteristics of customers that paid back the loan or not
- Identify the most important causes that are related with customers that did not pay back the loan
- Benefits: Assess whether or not a new customer is likely to pay back the loan.

Required data sources

For the realization of the project, it was important to gather useful data, such as the one that can be seen below.

- Customer data: Home ownership, annual income
- Loan data: Loan amount, term, interest rate, grade, loan status, ...
- Location

Expected delivery/output

Finally, I proposed to deliver certain things at the end of the project which I believe will be useful to try and improve the loan recommender systems that make such important decisions on whether to accept or reject a loan issue.

- Propensity model
- List of most important variables
- List of characteristics to refuse/accept a loan issue

Visualizations

This is a Visual Analytics course, and therefore the main goal is to provide a good visual analysis of the data so that it is easy and fast to understand the underlying problems or characteristics of the data we are dealing with. Therefore I have used different visualization methods to show off the data analysis:

- Plots inside the Jupyter Notebook file, with matplotlib or seaborn
- Profiling tool by pandas
- Dashboard in Tableau with all the important knowledge gained from the code analysis, to create a significant and relevant visualization based on the most important features.

Dataset

LendingClub is a US peer-to-peer lending company, headquartered in San Francisco, California. It was the first peer-to-peer lender to register its offerings as securities with the Securities and Exchange Commission (SEC), and to offer loan trading on a secondary market.

Lending Club enables borrowers to create unsecured personal loans between $1,000 and $40,000. The standard loan period is three years. Investors can search and browse the loan listings on Lending Club website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. Lending Club makes money by charging borrowers an origination fee and investors a service fee.

For this project I collected all the data between 2007 and 2017.

Python Code in Jupyter Notebook

In order to do this project I used python with the Jupyter Notebook tool, in order to have a deep understanding of the data and the most important variables so that I can make a correct data analysis. Apart from this I trained different models with the goal of trying to classify and predict whether an issued loan will be good or bad. A good loan is one that is Fully Paid within the time limit or within the Grace Period, which accounts for at most 15 days. All other loans are considered to be bad.

Dataset Exploratory Data Analysis (EDA)

For this dataset, with around 1.7 million rows, we eliminate the current loans, since we are not able to know if these will turn up into good or bad loans. This results in a dataset with around 370 thousand entries, from which 276,817 are good loans and 93,885 are bad loans.

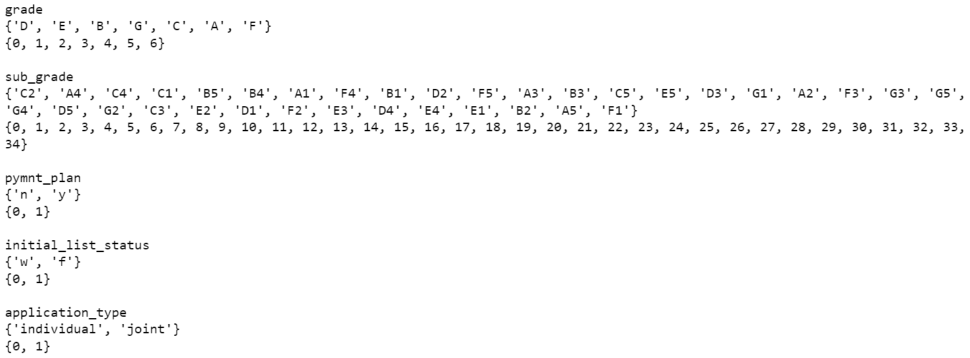
Also this dataset contains 74 features. But quite a good number of them get eliminated because of poor usability. Some of them are variables whose values will be unknown at the time of the loan issue, such as the number of recoveries, the amount paid, the amount still outstanding, …. Also, I formatted all the columns that had improper formats, I eliminated all the columns that had more than a 20% of its values as missing (nan, null, …); as well as taking care of the columns that were in date format.

After that I defined what a good or a bad loan is and created a reference column that lets us identify the good and the bad loans from the dataset. As mentioned before, I did not make use of the current loans.

In the end I finished with a dataset with 370,702 entries and 50 columns. With this resulting dataset I will create the Tableau dashboard and make the visualization analysis.

Data Wrangling

After having the dataset for the visual analysis, it is important to know that this data can not be fed into a classification model because it is not able to deal with the categorical data. For that reason pre-processed the data. I used LabelEncoder() and OneHotEncoding() to generate binary classes for all the relevant categorical variables. These are some of the changes I made.



```
grade
{'D', 'E', 'B', 'G', 'C', 'A', 'F'}
{0, 1, 2, 3, 4, 5, 6}

sub_grade
{'C2', 'A4', 'C4', 'C1', 'B5', 'B4', 'A1', 'F4', 'B1', 'D2', 'F5', 'A3', 'B3', 'C5', 'E5', 'D3', 'G1', 'A2', 'F3', 'G3', 'G5',
'G4', 'D5', 'G2', 'C3', 'E2', 'D1', 'F2', 'E3', 'D4', 'E4', 'E1', 'B2', 'A5', 'F1'}
{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33,
34}

pymnt_plan
{'n', 'y'}
{0, 1}

initial_list_status
{'w', 'f'}
{0, 1}

application_type
{'individual', 'joint'}
{0, 1}
```

Since I was dealing with an imbalanced dataset, we will perform a big down-sampling of the most common class, and a bit of over-sampling for the least common one. I made use of a pipeline to do this process. Finally I obtained 215,301 entries for the good loan class, and 193,771 entries for the bad loan class, which is quite balanced.

With all this process, I do a final analysis of the correlation between the variables. I decided to eliminate quite a big number of variables that were really correlated between them, so they did not provide any new useful data, and therefore I only kept one of them.

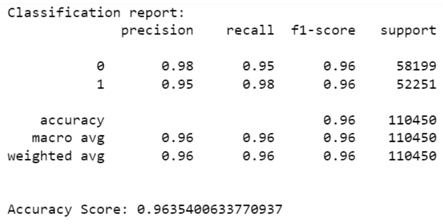
After this, I used the Profiling tool from pandas to get an even better understanding of the data, and since my data is ready to train classification models, that is what I will do next.

Classification Models

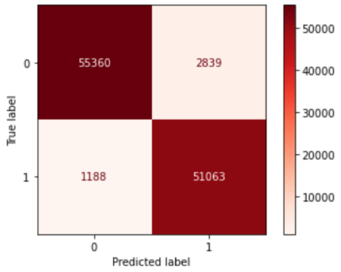
In this case I decided to run 4 different models, to try and classify the loans as good or bad ones.

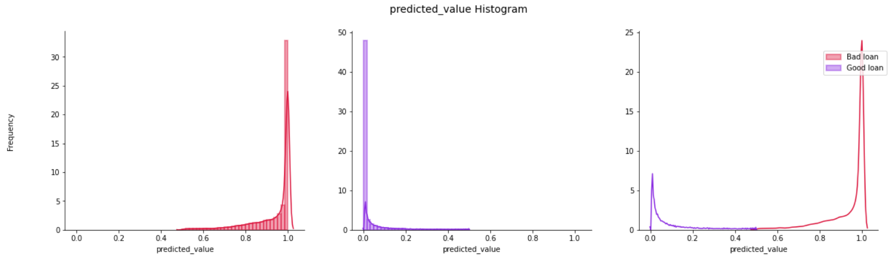
1. Logistic Regression

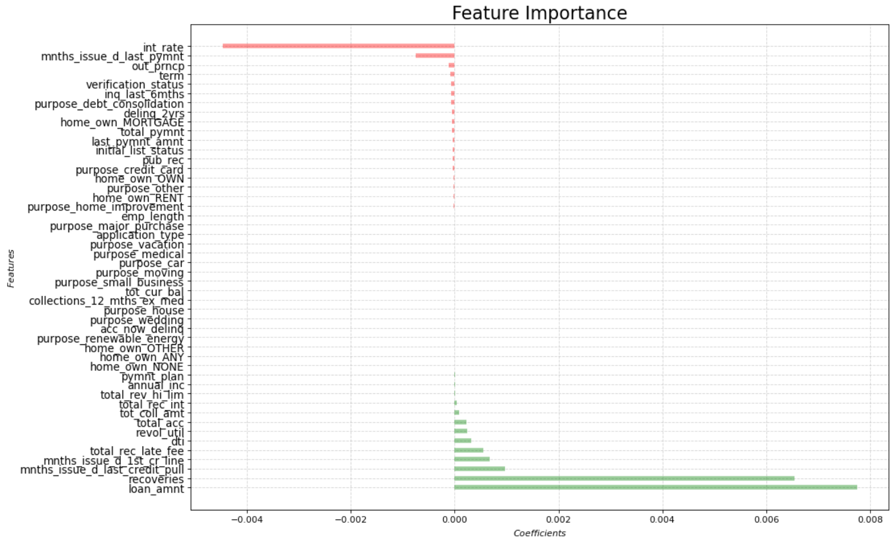
I ran two different models with these one, since after the first one I realized that there were still some variables that are not relevant, or that are only defined at the end of the loan and therefore we should not use them. As it can be seen, the model is too good because it is taking into account some variables that are defined for finished loans, and therefore it is a perfect example of data leakage. This is where you use data that is impossible to obtain at the time of deciding whether to give or not the loan, usually because it is a direct consequence of an event that you are trying to predict.



```
Classification report:
              precision    recall  f1-score   support

           0       0.98      0.95      0.96     58199
           1       0.95      0.98      0.96     52251

    accuracy                           0.96    110450
   macro avg       0.96      0.96      0.96    110450
weighted avg       0.96      0.96      0.96    110450

Accuracy Score: 0.9635400633770937
```

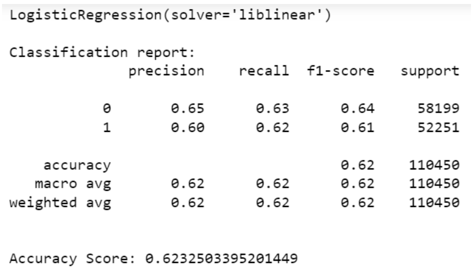




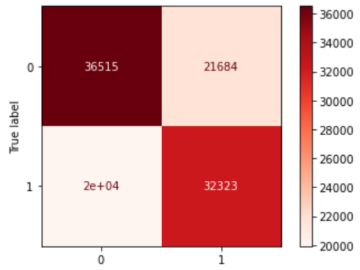


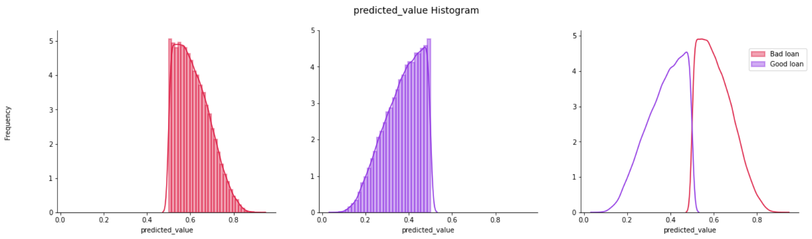
Here it can be seen that some of the most important variables are determined after the end of the loan, or in the middle of it, such as, recoveries, total_rec_late_fee, mnths_issue_d_last_credit_pull, …

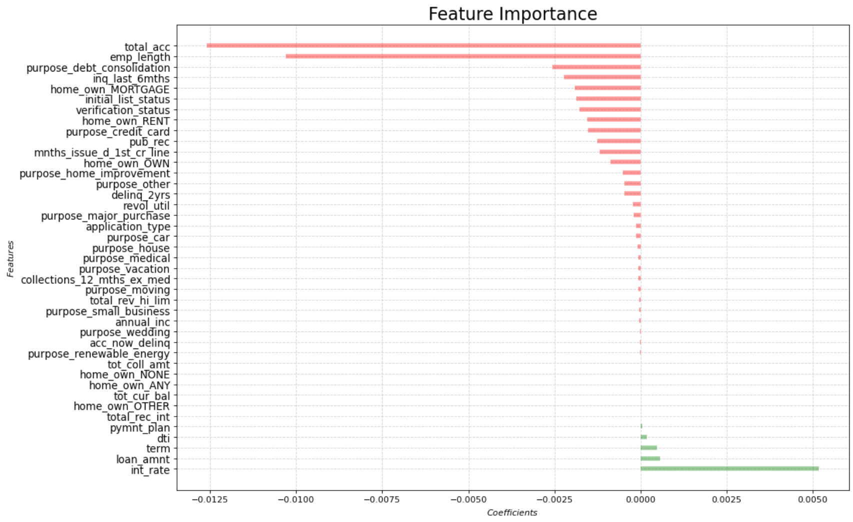
After dealing with these variables, the second one came like this:



```
LogisticRegression(solver='liblinear')

Classification report:
              precision    recall  f1-score   support

           0       0.65      0.63      0.64     58199
           1       0.60      0.62      0.61     52251

    accuracy                           0.62    110450
   macro avg       0.62      0.62      0.62    110450
weighted avg       0.62      0.62      0.62    110450


Accuracy Score: 0.6232503395201449
```

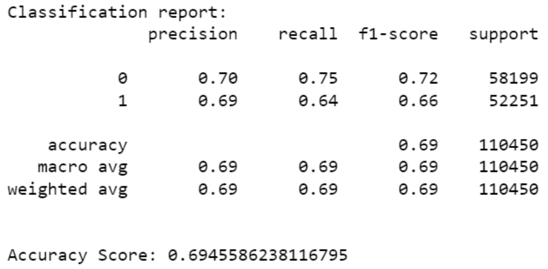




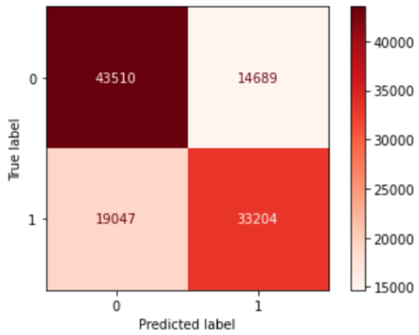


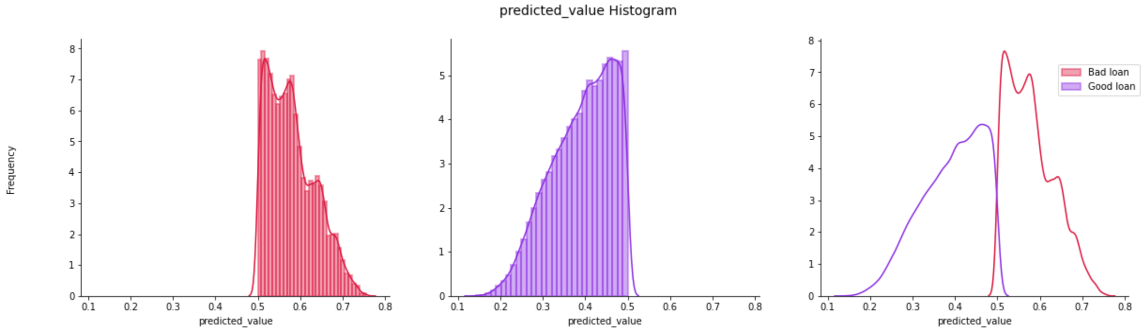
Here we can now see some of the important variables of the dataset: int_rate, term, loan_amount, payment_plan and the dti. This has of course worse performance, but it is more representative.

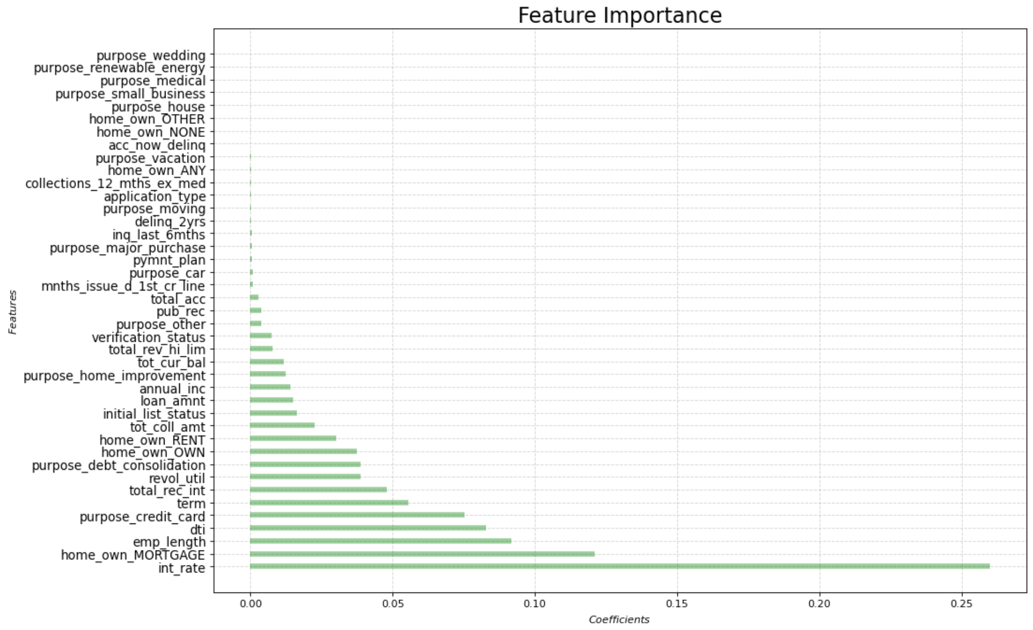
2. Random Forest Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.70      0.75      0.72     58199
           1       0.69      0.64      0.66     52251

    accuracy                           0.69    110450
   macro avg       0.69      0.69      0.69    110450
weighted avg       0.69      0.69      0.69    110450


Accuracy Score: 0.6945586238116795
```

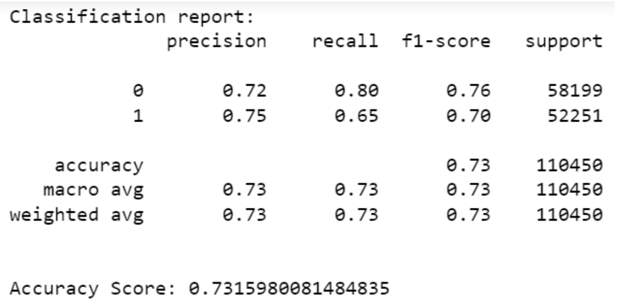




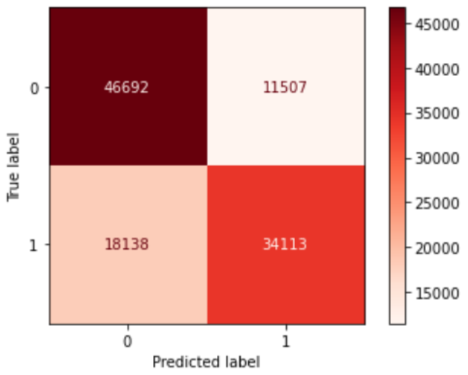


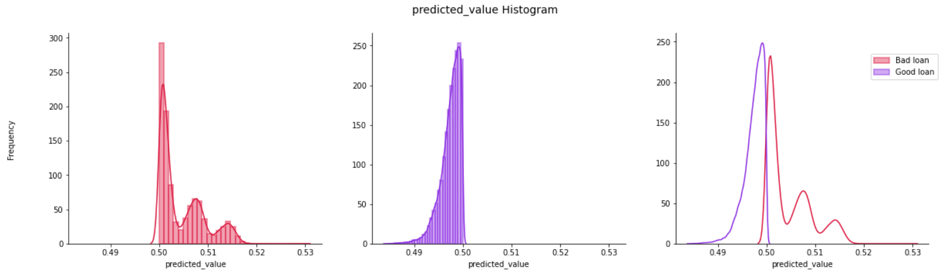
Here we have a very similar performance of the model before but the relevant features vay a little bit. However we can also see interest rate, dti and term among the top 5 most important features.

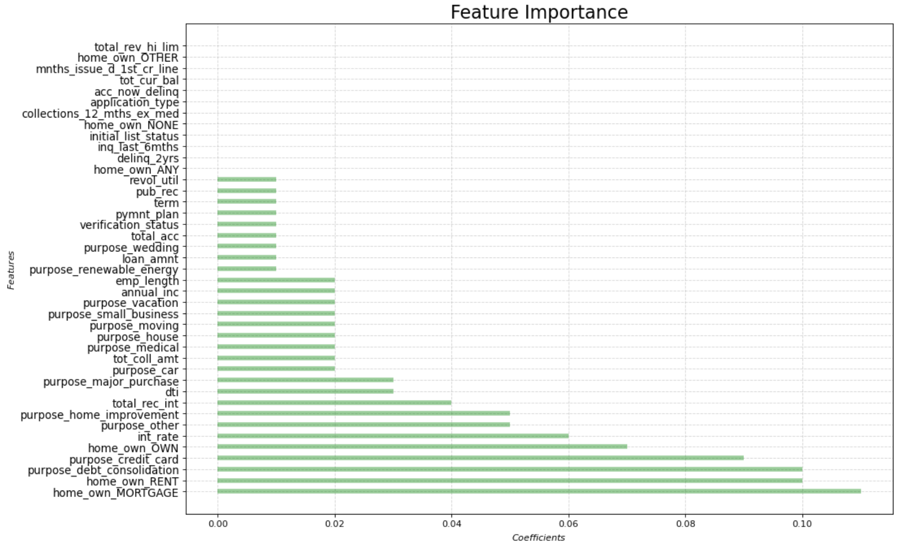
3. AdaBoost Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.72      0.80      0.76     58199
           1       0.75      0.65      0.70     52251

    accuracy                           0.73    110450
   macro avg       0.73      0.73      0.73    110450
weighted avg       0.73      0.73      0.73    110450


Accuracy Score: 0.7315980081484835
```

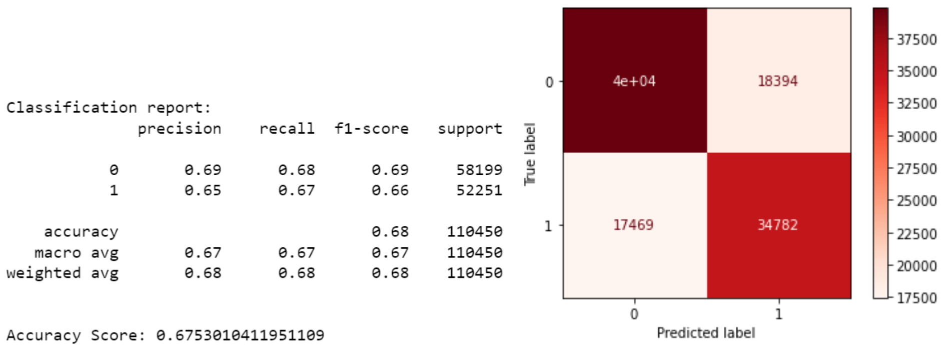




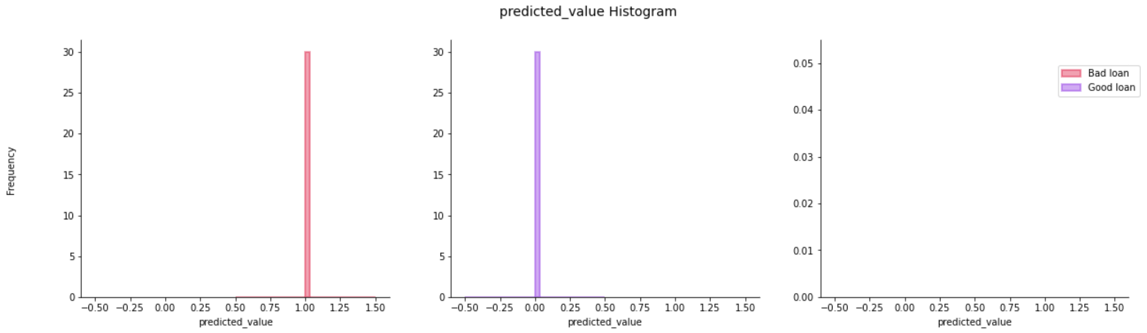


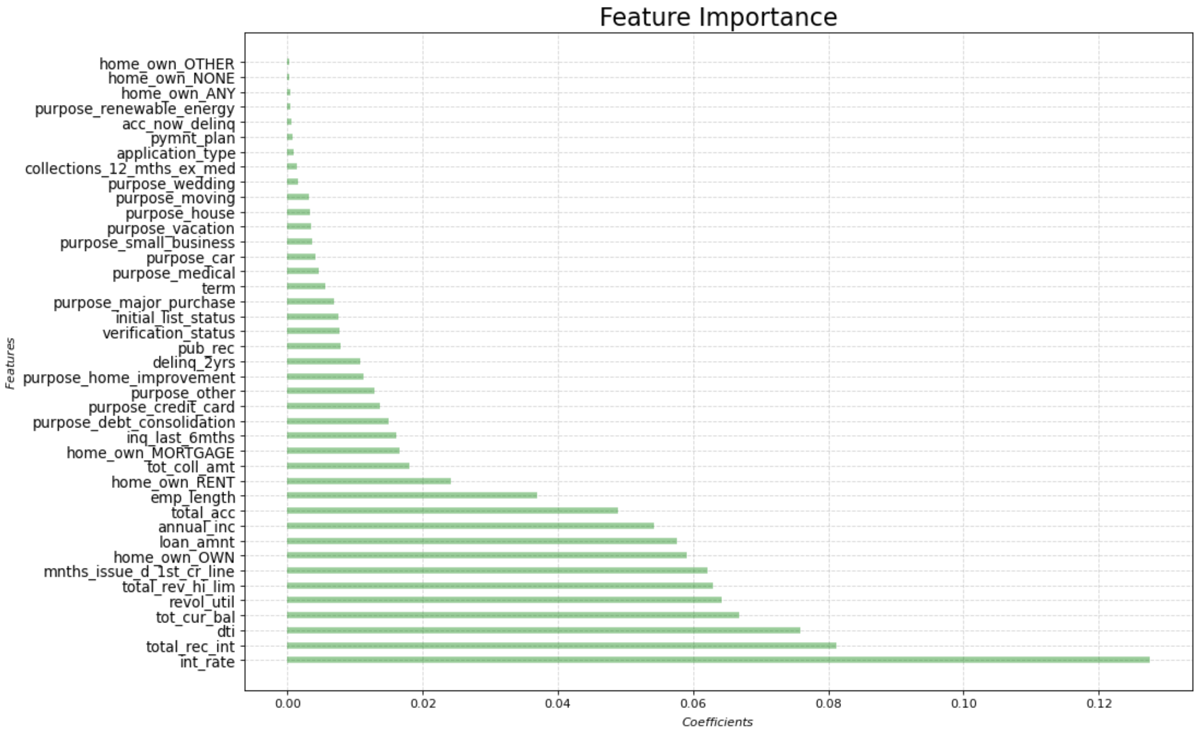
This model already has a better performance, which is quite decent considering the amount of entries in our dataset. As well as before, interest rate, dti are one of the most important columns. But now we can see home_ownership Mortgage is actually a relevant feature, which was also in the top in the Random Forest Classification.

4. Decision Tree Classifier



```
Classification report:
              precision    recall  f1-score   support

           0       0.69      0.68      0.69     58199
           1       0.65      0.67      0.66     52251

    accuracy                           0.68    110450
   macro avg       0.67      0.67      0.67    110450
weighted avg       0.68      0.68      0.68    110450


Accuracy Score: 0.6753010411951109
```





This is a way more simple algorithm than the Random Forest or AdaBoost as they are ensembles of models. That is why this model is not so good. It is the worst until now. Nevertheless, this model also has interest rate and dti as the key aspects of the dataset.

Dashboard

Here I provide the link to the Tableau dashboard:

There with the use of filters I created an interactive dashboard where people can get with little time the most important aspects of the dataset without having to give much explanation:

- The most important features for loan classification
- How the dataset is distributed
- How bad/good loans are characterized by
- An overall view of the most relevant data

This way someone can by just taking a glance at the dashboard have a deep understanding of the problem, the dataset, its main features and characteristics, and what defines a good loan and what defines a bad loan. This way any person could have a better understanding of whether a loan can be fully paid back in time or not.

Results and Evaluation

Overall, and as we have seen in the dashboard and in the analysis of the data, the main characteristics that define a good loan or a bad loan are encoded in the interest rate, the term in which this loan has to be paid back, whether that person owns already a home o has a mortgage on it. Also it is important the amount of the loan but as seen, it only appeared as an important feature in 1 of the 5 different models that I have used.

Therefore a person can look into these key features and analyse whether or not the loan he is issuing or investing into has a high chance of being paid back or if it will end up in debt.

Bibliography

- https://www.kaggle.com/husainsb/lendingclub-issued-loans
- https://www.kaggle.com/ethon0426/lending-club-20072020q1