Loan Default Prediction

Yuanhao Jiang

Executive summary：

The project is aiming to find the relationship between features to better classify if the borrower will default the loan using the borrower’s financial history. We use visualization and model to achieve them. For the visualization, we could see most people have a loan term of 3 or 5 years. A 3-year loan is less than that for a 5-year loan. Half of the defaulters have one or even two houses. The vast majority of the data is 0. This indicates that their recent credit is good, and the majority of them did not default on their loans intentionally. Employees in the ordinary company often loan and have high default rates. Employees of Fortune 500 companies and higher education rarely loan and have low default rates. Cities with low-interest rates will have lower default rates. For the model part, we compare many models that we find Random Forest model which has better performance than other models. We put all the features into models to train that we get the feature importance. Stakeholders could mention those features that are significantly related to classify if the borrower will default the loan.

A. Design thinking:

To further advance the promotion of financial inclusion on the ground, financial institutions need to serve many new customer segments. Banks, as an industry with high requirements for risk control, often become an important obstacle to financial inclusion because of the lack of understanding of new customer segments and the handling of risk control for new segments. How to use banks' existing credit behaviour data to serve new scenarios and new customer segments has become a valuable research direction. If some people want to loan, stakeholders have financial markets. Such as insurance companies, it is important to be able to predict the risk of loans.

B. Get the data we need

To predict the default situations, we need some data containing the customer's personal status, economic situation and historical behaviour. During our search, we found a contest called CFF (China Computer Federation) Big Data & Computing Intelligence Contest. The organization release the dataset collected from Zhongyuan Bank. Because this data has fewer anonymous features and more explainable features. It is easy to understand the meaning of features. We choose this dataset. The dataset which provides consists of three parts: train internet, train public and test public. It consists of 770,000 observations and 35 features.

The “train internet” part is collected by an internet company that provides online loans to individuals. The other two parts are provided by a bank. These three datasets include many loan records and default records from 2007 to 2018. There are approximately 40 features we can use to build a model to make the prediction of future loans situation.

C. Problem solving with data

As inclusive financing develops, the types and numbers of target customers are increasing. Loan defaults have become a serious problem. Default is the failure to repay a debt, including interest or principal, on a loan. Therefore, a credit analysis should be performed. In this project, we want to use the bank's historical credit data to find the relations between customers’ features and loan default situations to assist decision making.

Banks and internet companies, as the funders are the stakeholders in this project. In addition, many investment companies and financial institutions will suffer losses from loan default. Governments could also be the stakeholders. We hope that we can use some statistical methods and machine learning models to predict loan defaults.

1. Stakeholder:

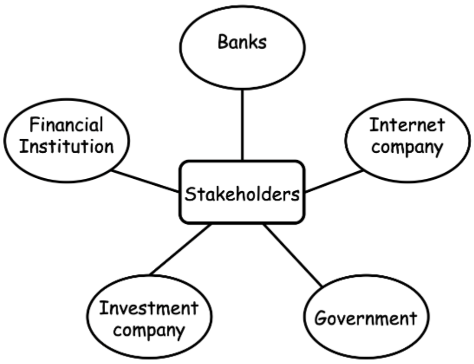

Figure1

2. Project aims:

Explore which features in the dataset have the greatest impact on loan defaults and find the relationship between these features and loan defaults.

3. Exploratory data analysis:

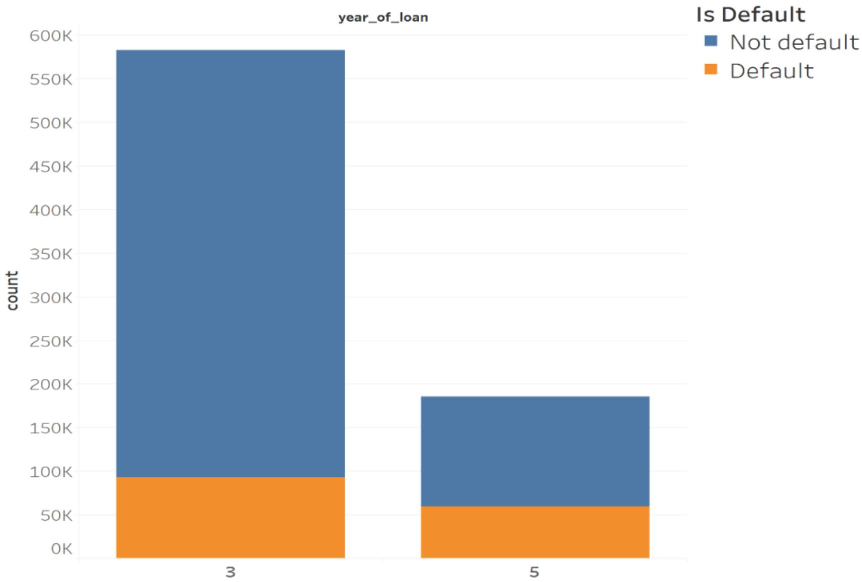

Figure2

At first, we could see if they are default and the distribution of the year of the loan from the figure7.

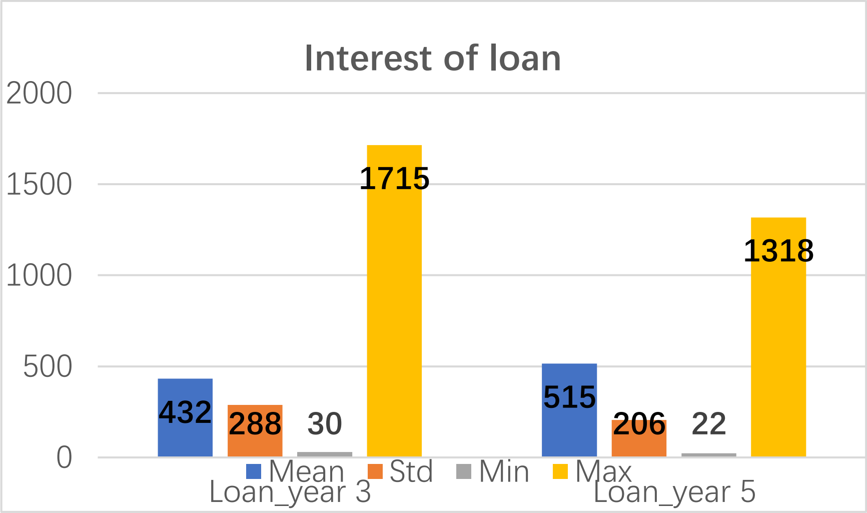

Figure3

From the figure3, we could see a 3-year loan is less than that for a 5-year loan.

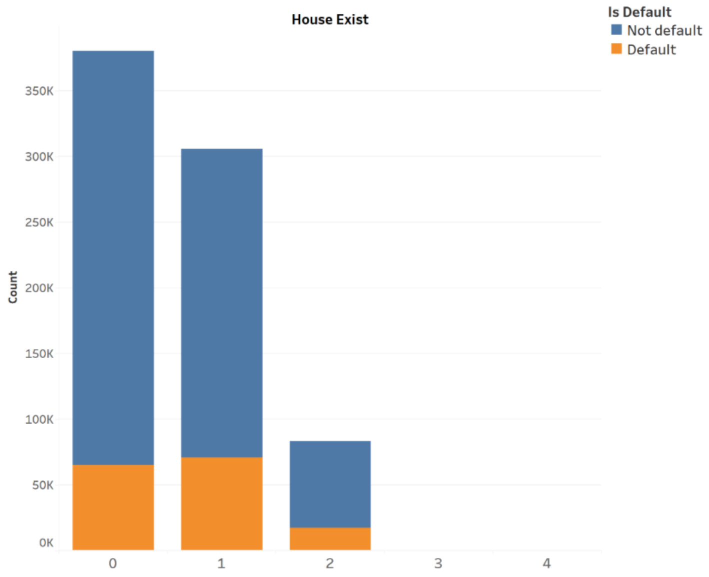

Figure4

From the figure4, we could see half of the defaulters have one or even two houses.

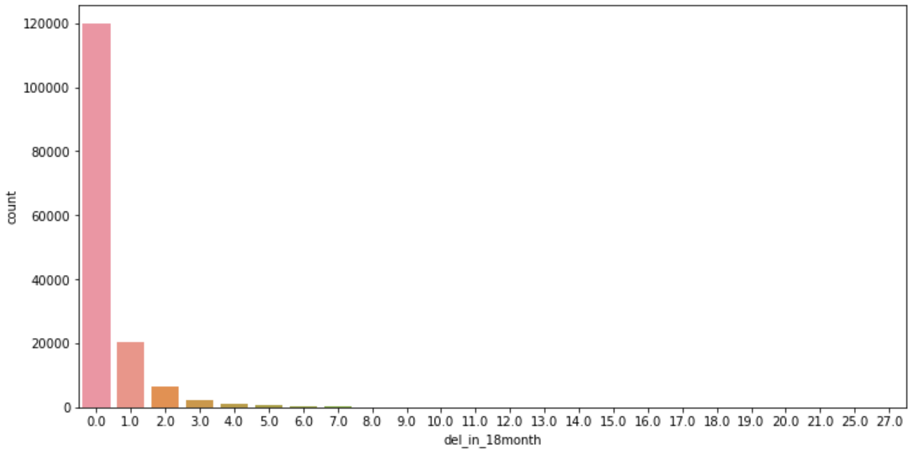

Figure5

From the figure5, we could see the vast majority of the data is 0. This indicates that their recent credit is good, and the majority of them did not default on their loans intentionally.

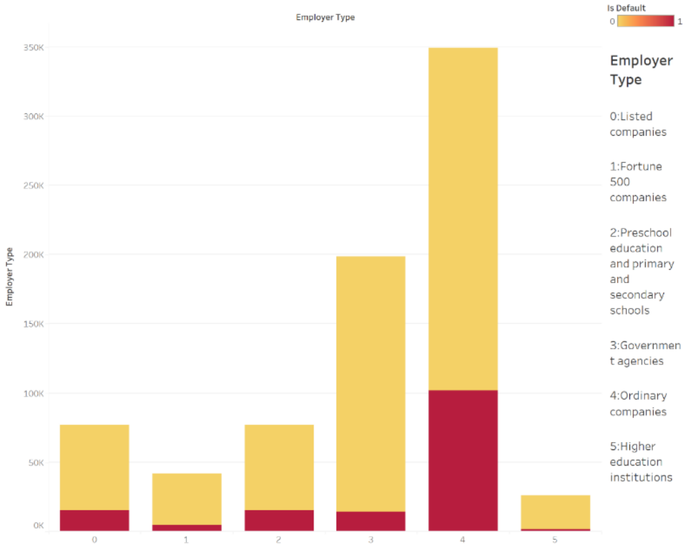

Figure6

From the figure6, we could see employees in ordinary company often loan and have high default rates. Employees of Fortune 500 companies and higher education rarely loan and have low default rates.

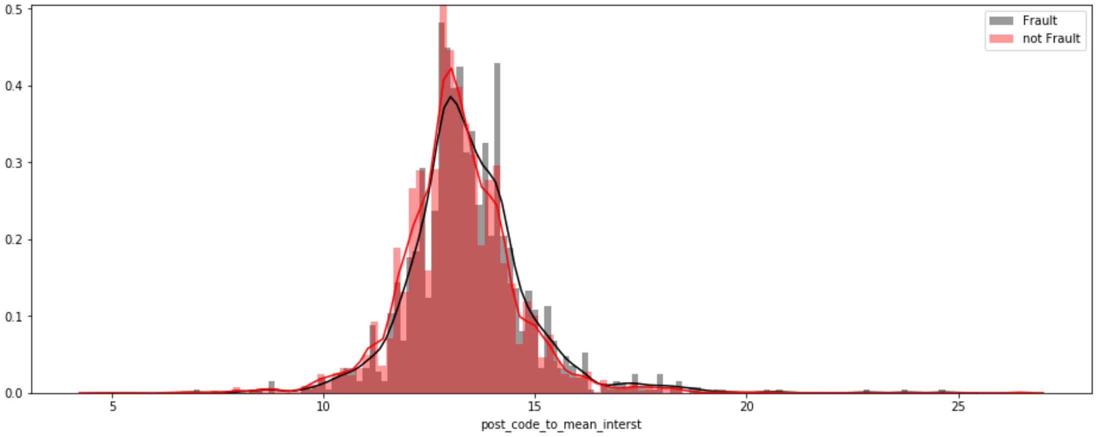

Figure7

From the figure7, we could see cities with low interest rates will have lower default rates.

D. Is my data fit for use

After getting the dataset, we can see that the label "1" means that the loan is the default. The value "0" means that the loan is repaid in time. After that, we want to find the unique features in every chart and remove them. Then we can see that there are still some redundant features in our prediction model. F0-F4 are anonymous features with too many missing values. Therefore, we delete them. The feature "region" and "post_code" both represent the location so that we can delete one of them. The feature "use" means the loan purpose categories but we don't have any explanation for its classes. As for "censor statues", "title" and "policy code", we don't know the

meaning of them. Therefore, we also delete all these.
After deleting redundant features, we can see that the remaining features don't have many missing values. Therefore, we directly delete these samples. This is the whole process of data cleaning before building the prediction model of the loan situation.

1. Data Engineering:

For this project, we have collected data from CCF Big Data & Computing Intelligence Contest, which is founded by the Chinese Computer Society in 2013. The dataset consists of 770,000 observations and 35 features. Out of the 35 features in our dataset, many of them were empty or invalid. We have removed all such features. Also, the features which didn't seem relevant to our goal were removed.

- String values have been formatted to integers.
- Categorical values have been transformed to numerical.
- Redundant variables have been dropped.
- Filled NAN values with mean values of corresponding columns.
- All the numerical values have been scaled to a range between -1 and 1.

E. Make data confess

1. Problem Statement:

Problem statement is to classify if the borrower will default the loan using borrower's finance history. That means, given a set of new predictor variables, we need to predict the target variable as 1 -> Defaulter or 0 -> Non-Defaulter.

2. Predictor Variables (Input):

On the above 35 features, we have implemented Univariate Feature Selection to get the best 6 features as input data.
The table1below mentioned are the features used for our model:

Predictor Variables	Description
year_of_loan	Duration of the loan
class	Grades of credit evaluation
early_return	Number of early repayments by the borrower
interest	Current Loan Rates
early_return_amount_3mon	Early repayment amount within the last 3 months
early_return_amount	Accumulated amount of early repayment by the lender

Table1

3. Target Variable (Output):

The target variable in our dataset is 'isDefault' which shows the status of the loan.

During the process, we predict the target variable as 1 ->Defaulter or 0 -> Non-Defaulter, which is the output data as well.

4. Models Applied:

Random Forests Classification:

The final output will be the mode of the outputs of all its decision trees which has better results than decision trees (which can possibly overfit). Hence, we choose to start our classification with random forests.

Multi-Layer Perceptron:

MLP utilizes backpropagation for training. Its multiple layers and non-linear activation function help us distinguish data that is not linearly separable.

Logistic Regression:

With logistic regression, outputs have a nice probabilistic interpretation, and the algorithm can be regularized to avoid overfitting. Hence, we choose to build logistic regression classifier.

KNN:

In k-NN classification, an object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

5. Evaluation Approach and Model comparison:

We plot the Roc curve (figure8) to get the AUC score of models. And we calculate the accuracy (figure9) of models. Then we find the Random Forest model which has better performance than other models. We put all the features into models to train that we get the feature importance (feature importance). Stakeholders could mention those features that are significantly related to classify if the borrower will default the loan.

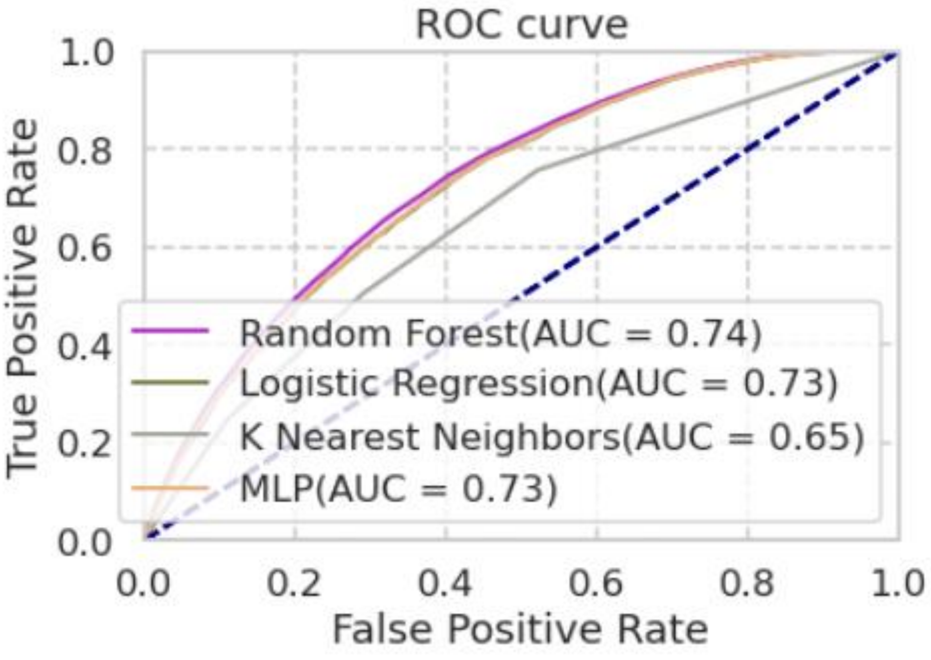

Figure8

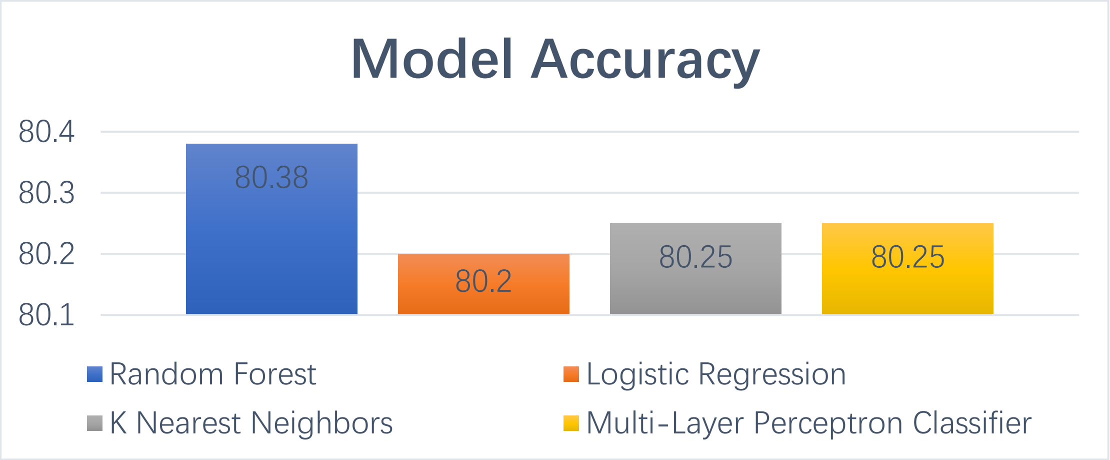

Figure9

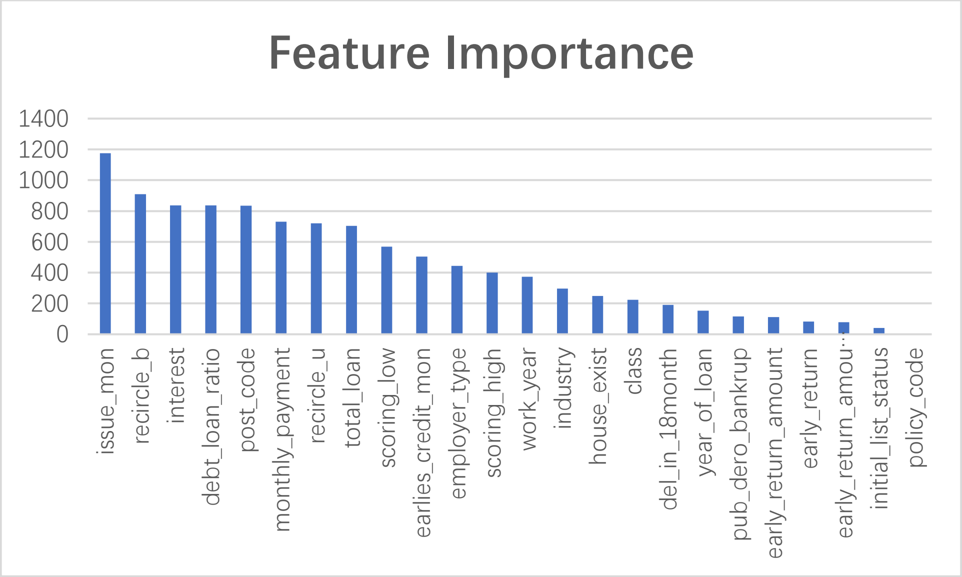

Figure10

6. Limitations:

In the process of feature selection, we wanted to try Grid search CV and Randomized search and RFE (Recursive Feature Elimination) methods. But it requires powerful computing power, which our computer cannot support. We gave up these methods. Cross-validation for the model here requires high computational time.

F. Summary:

1. Feedback:

1.who are stakeholders

2. For the get data I need, not clear description. When it collected? What was the year? What type of fields it has?

3.Is my data for use, explain why those features redundant.

4.For the EDA, it should not have two graphs in one slide

5.For the feature selection, explain why those feature important, which features are predictors of my output
6.storytelling, the graph should show the axes and numbers
7.make the task clear, could repeat the subject matter a few times.
Solution:
We fix those problem in final presentation. Clearly describing the task and explaining how to get the data we need. Explaining the reason why drop those redundant data. Changing the figure of EDA. Explaining why those features are important in detail.

2.Conclusion

The project is aiming to find the relationship between features to better classify if the borrower will default the loan using the borrower's financial history. We use visualization and model to achieve them. For the visualization, we could see most people have a loan term of 3 or 5 years. A 3-year loan is less than that for a 5-year loan. Half of the defaulters have one or even two houses. The vast majority of the data is 0. This indicates that their recent credit is good, and the majority of them did not default on their loans intentionally. Employees in the ordinary company often loan and have high default rates. Employees of Fortune 500 companies and higher education rarely loan and have low default rates. Cities with low-interest rates will have lower default rates. For the model part, we compare many models that we find Random Forest model which has better performance than other models. We put all the features into models to train that we get the feature importance. Stakeholders could mention those features that are significantly related to classify if the borrower will default the loan.

Reference

[1] The model of random forest, logistical regression, KNN, MLP from
scikit-learn: machine learning in Python — scikit-learn 1.0.1 documentation
[2] The clean the data and data imputation from https://pandas.pydata.org
[3] The dataset download from https://www.datafountain.cn/competitions/530