DATA7001 Introduction to Data Science

GROUP PROJECT (40%)

Semester 2, 2021

Loan Default Prediction

Group5 member:

Yuanhao Jiang

Jiahui Zhang

Zixiong Du

Xinxin Zhang

Motivation:

To further promote 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 behavior data to serve new scenarios and new customer segments has become a valuable research direction.

Get the data I need

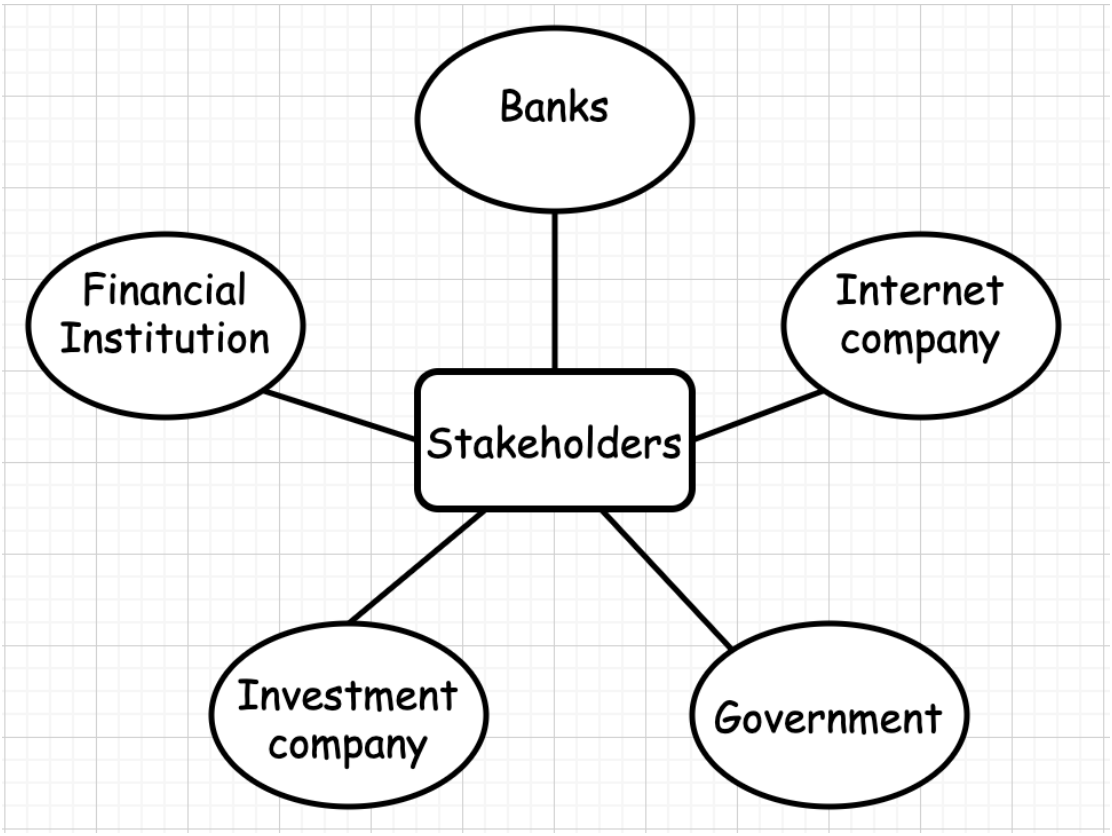
To predict the default situations, We need some data containing the customer's personal status, economic situation and historical behavior. During our search, we found a contest called CFF Big Data & Computing Intelligence Contest. The dataset it provides consists of three parts: train internet, train public and test public. It consists of 770,000 observations and 35 features.
The “train internet” part is collected by an internet company that provide online loan 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.

Problem solving with data

As inclusive financing develops, the types and numbers of target customers are increasing. Loan default has become an innegligible 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. Therefore we hope that we can use some statistical methods and machine learning models to predict loan defaults.

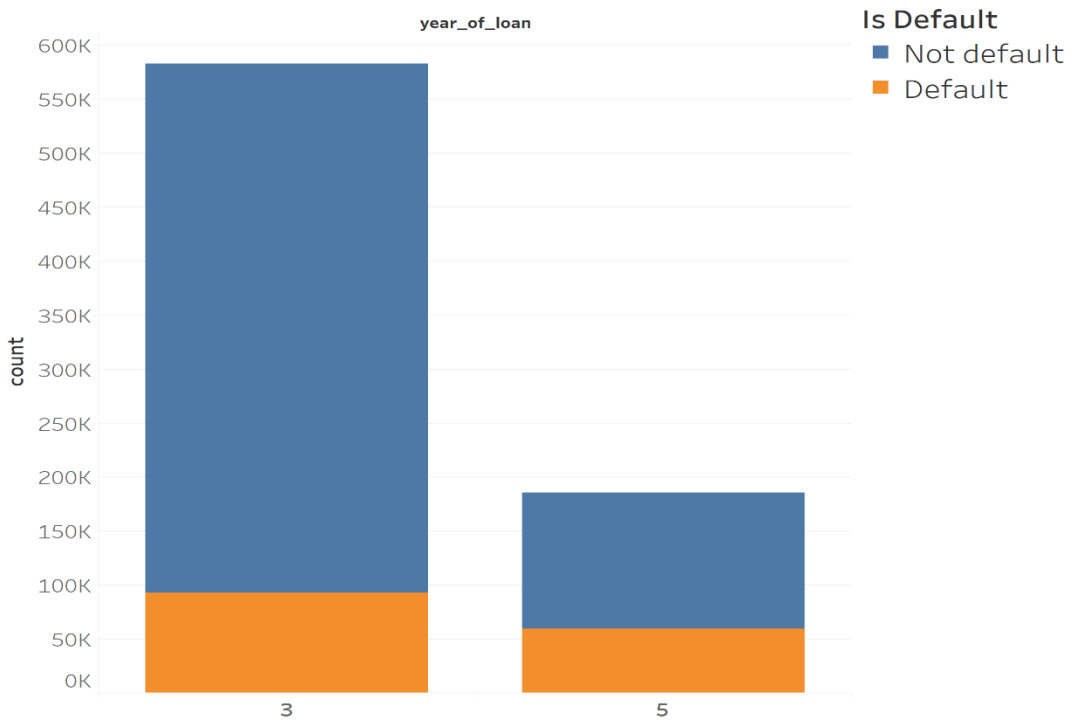
Stakeholder:



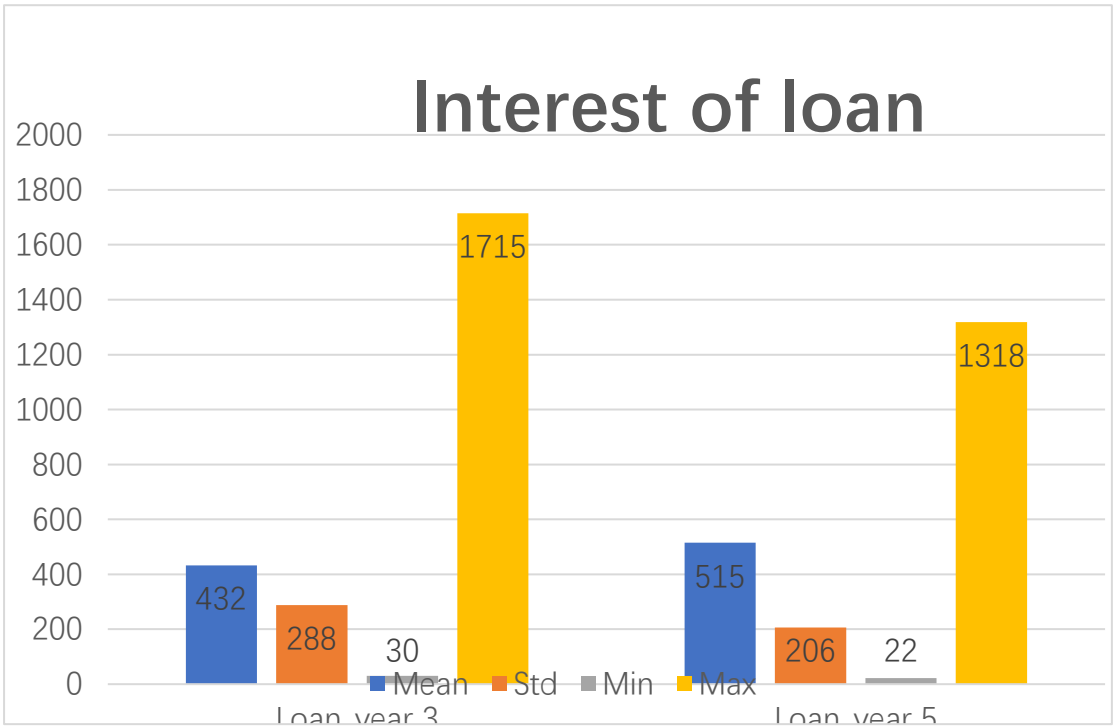
Task:

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

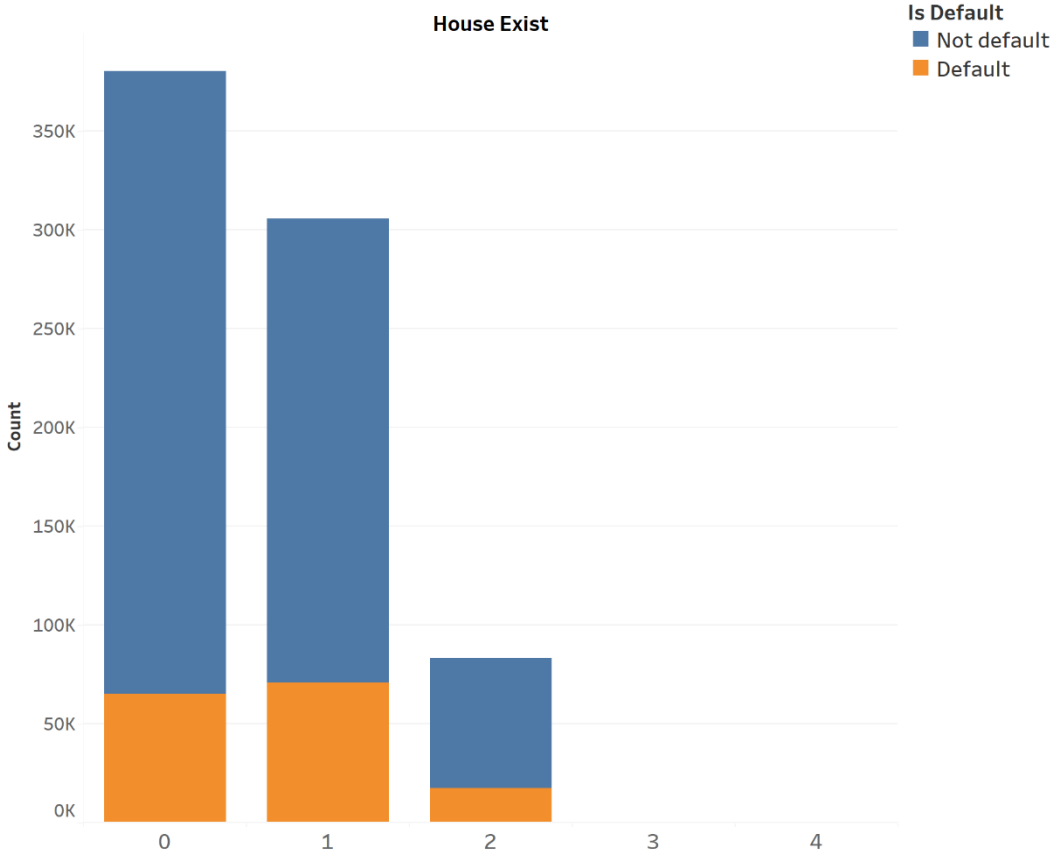
Exploratory data analysis:



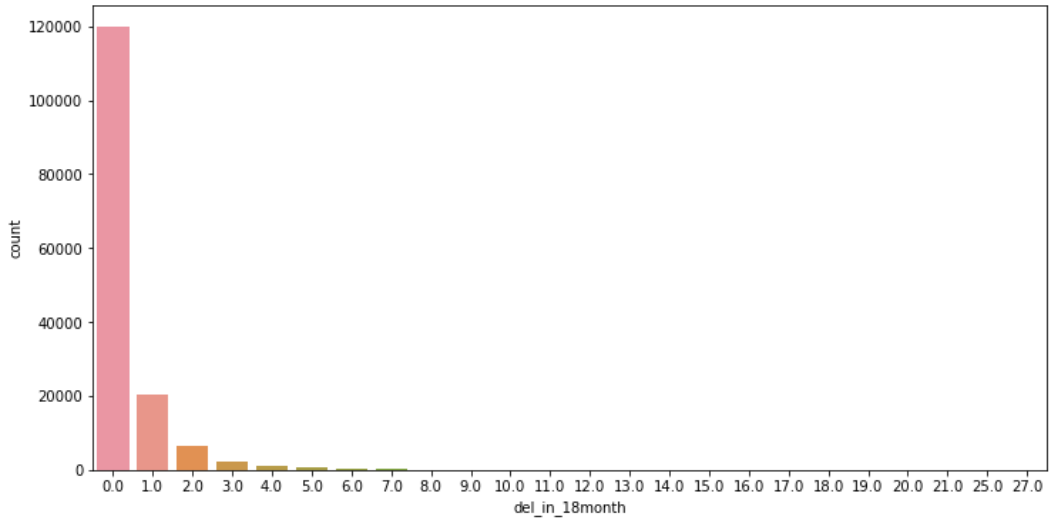
At first, we could see if they are default and the distribution of year of loan.



A 3-year loan is less than that for a 5-year loan.

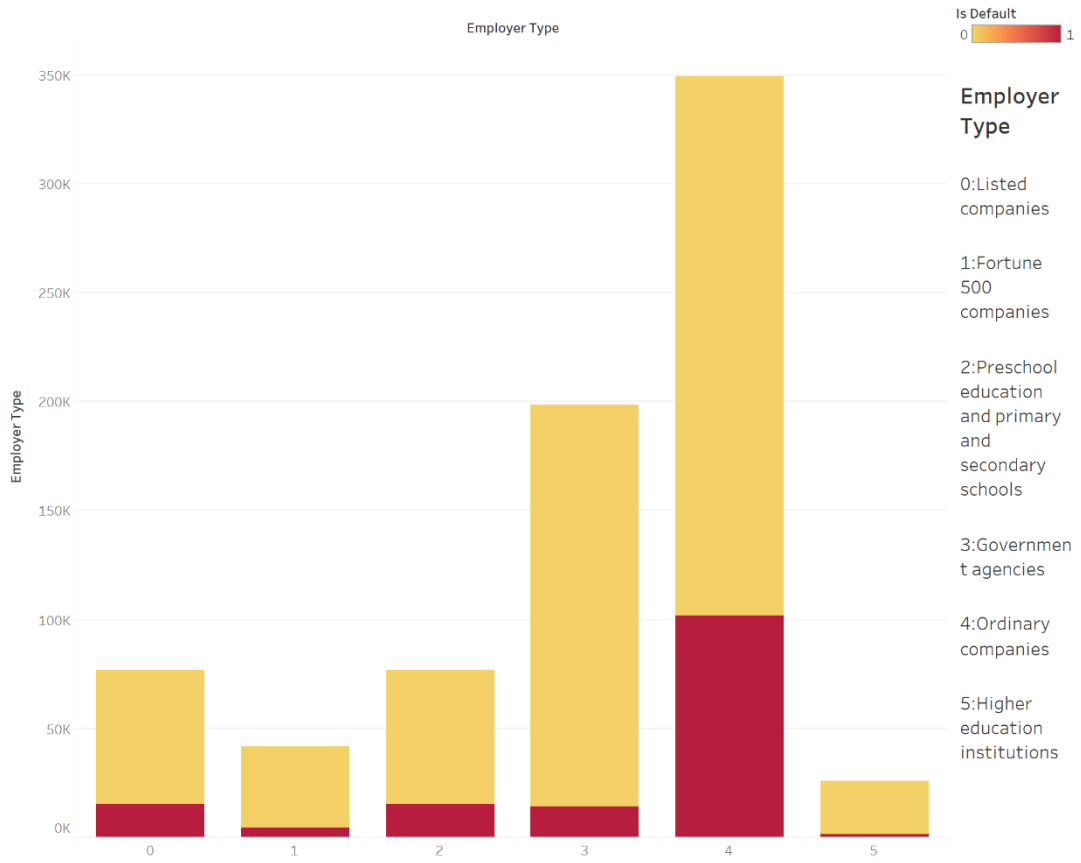


Half of the defaulters have one or even two houses.

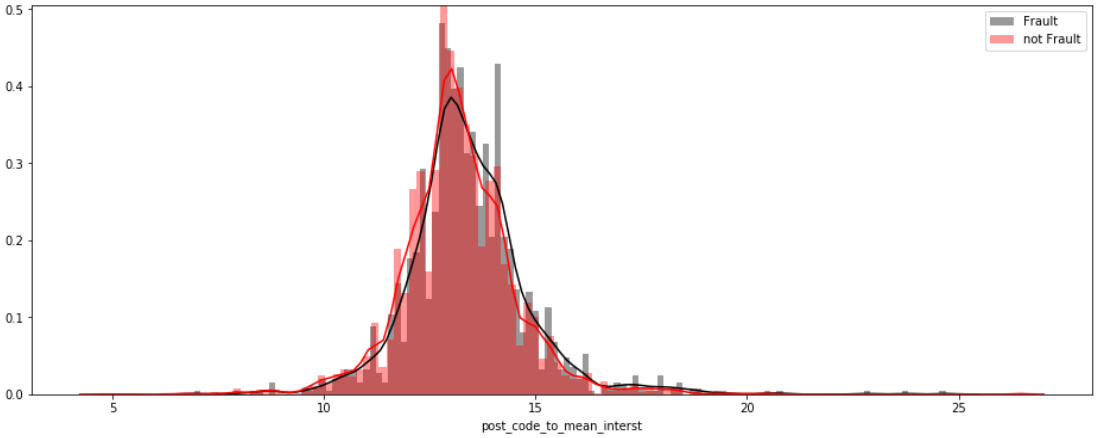



```
count   151133.000
mean         0.354
std          0.944
min          0.000
25%          0.000
50%          0.000
75%          0.000
max         27.000
Name: del_in_18month, dtype: float64
```

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 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.

Is my data fit for use

After get the dataset, we can see that the label “1” means that the loan is 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 ones.
After deleting redundant features, we can see that the remaining features don’t have much missing values. Therefore, we directly delete these samples. This is the whole process of data cleaning before building the prediction model of loan situation.

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.

Make data confess

Problem Statement:

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.

Predictor Variables(Input):

On the above 35 features, we have implemented Univariate Feature Selection to get the best 6 features as input data.
Below 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

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.

Models Applied and Motivation:

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.

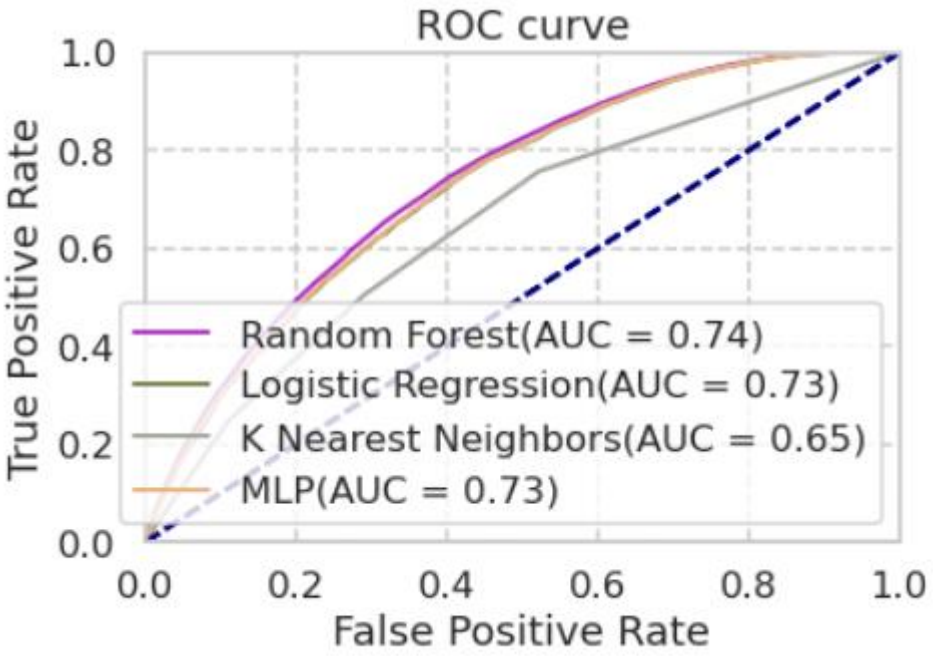
Logistic Regression:

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.

Evaluation Approach:



Model Accuracies:

Random Forest----80.38

Logistic Regression----80.20
K Nearest Neighbors----80.25
Multi-Layer Perceptron Classifier ----80.25

Limitations:

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