# Overdue Rate Prediction based on Personal and Behavior Information of Clients

6202 Project

Professor: Amir Hossein Jafari

Jiaxi Jiang, Xueqi Zhou, Meijiao Wang

# Background

Historical behavior data of the Chinese Credit Platform

Aim:

Predict the probability of users overdue in the next 6 months

# Background

Mirror Risk Control System:

- the first real risk control model based on big data in the industry

- 13.6 million online users
- 12.86 billion pieces of data

# Reason of Research

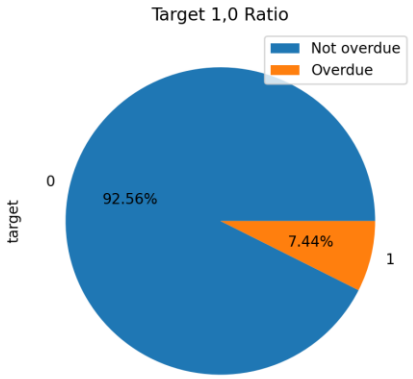
1. The labor cost is relatively high
2. The accuracy is difficult to guarantee by human
3. It is easy to cause corruption if there are no good standards.

# Table of Contents

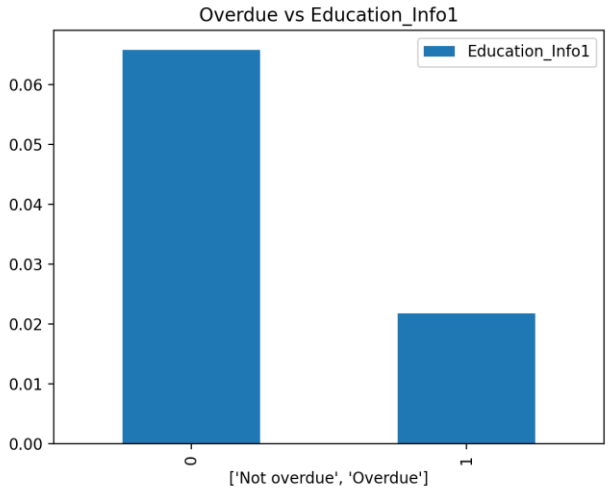

- Introduction
- Description of dataset
- Experimental setup and Results
    - (1) Data visualization
    - (2) Data cleaning
    - (3) Feature engineering
    - (4) Feature selection
    - (5) Model Building
    - (6) Accuracy
- Conclusion

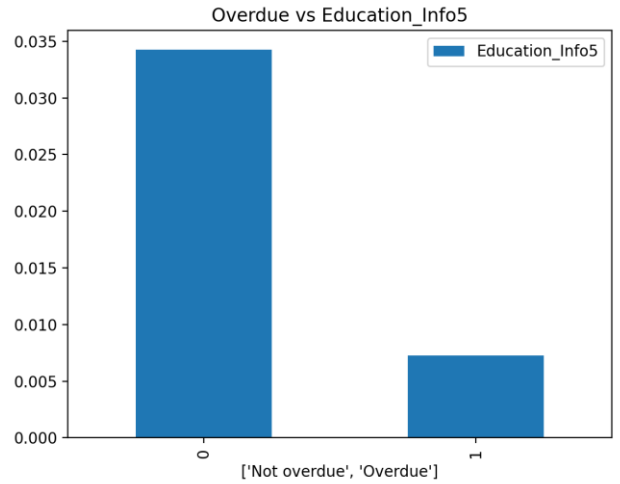
# Experimental setup and Results

- data visualization

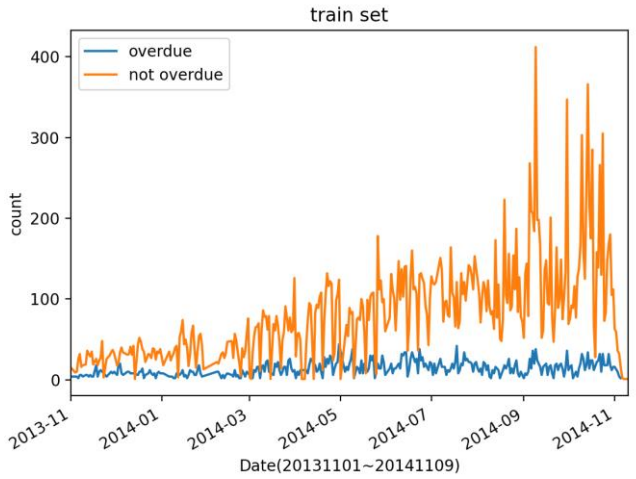



Graph1



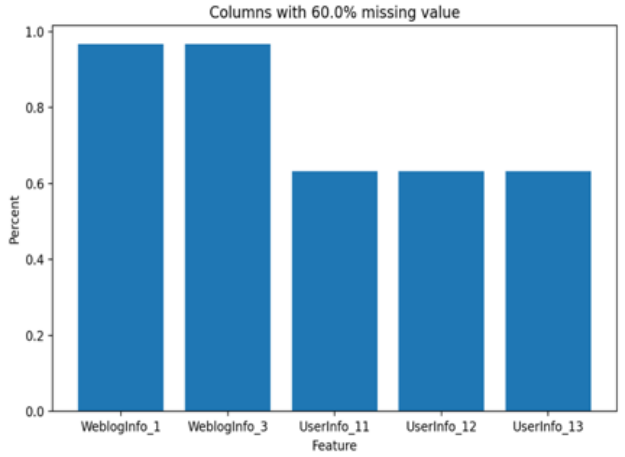

Graph2



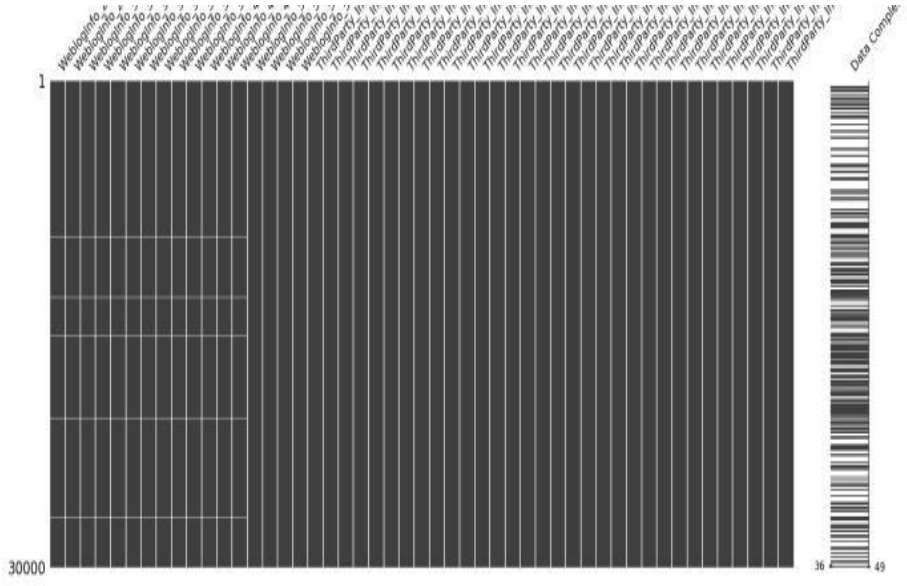

Graph 3

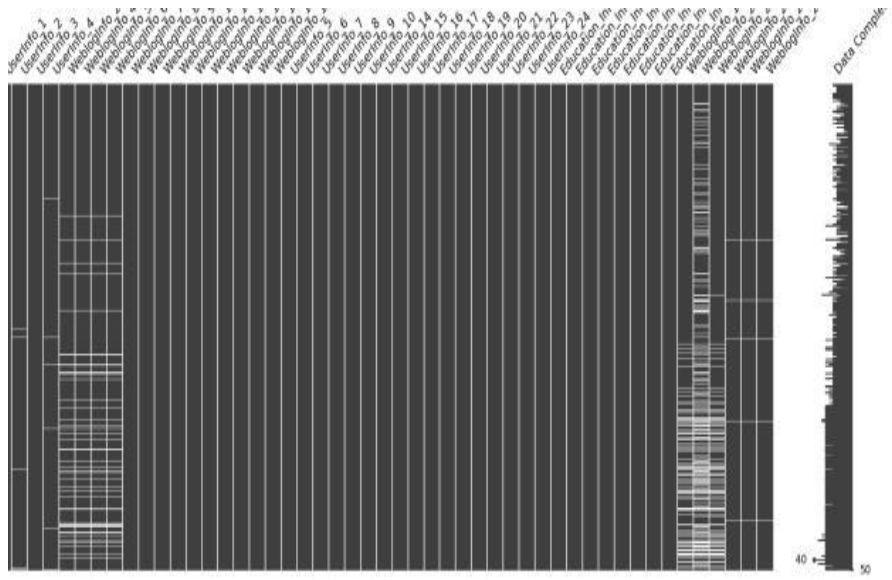
# Experimental setup and Results

- data cleaning
  - Delete features have missing values higher than 60
  - Delete features have standard deviation near 0
  - Change the format of city's name ("重庆" equals to "重庆市")



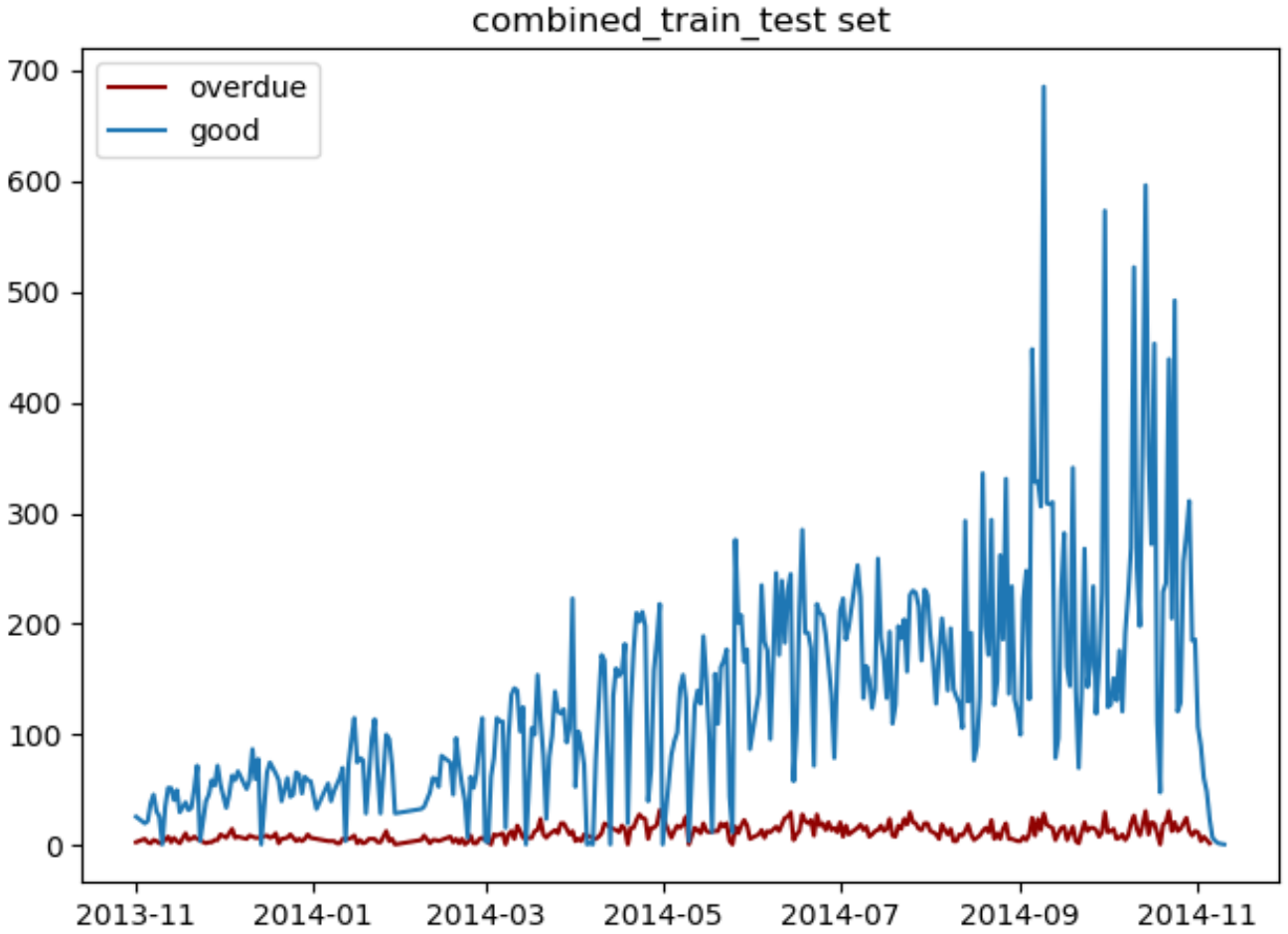
# Experimental setup and Results





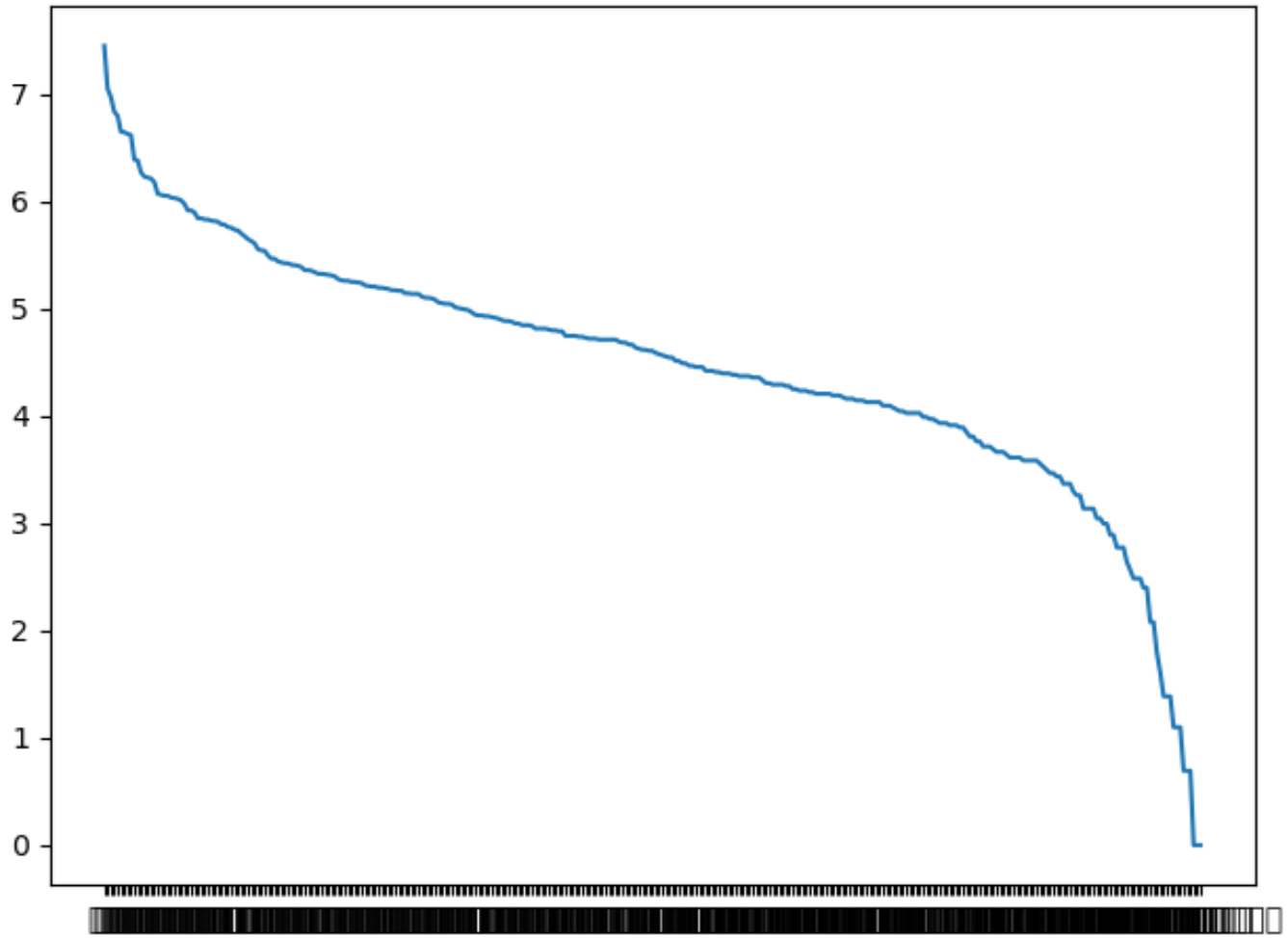
We found the missing value, the white represent the missing value, black means filled with values, and I just change the object value with mode, the int or float value with mean

# Experimental setup and Results




Plot the overdue with date and group by the overdue, 0 or 1. see whether there is some time series relationship between target and date. As we can see, there indeed have some time series information, so I change the date data from 1~10 to 1, 11~20 to 2,…..29991~30000 to 3000

# Experimental setup and Results




Plot the log(the numbers if cities) We can see there are three group of cities, it is clearly to see this in the plot, group 1 equal to 1, group 2 equal to 2, group 3 equal to 3.

# Experimental setup and Results

- feature engineering
  - (i) create new features, according to provinces having overdue ratio higher than 85%
  - (ii) whether the names are the same among every two columns

| Idx | UserInfo_2 | UserInfo_4 | UserInfo_7 | UserInfo_8 | UserInfo_19 | UserInfo_20 |
|---|---|---|---|---|---|---|
| 10005 | 广州 | 韶关 | 广东 | 广州 | 辽宁省 | 锦州市 |
| 10013 | 郴州 | 广州 | 广东 | 广州 | 湖南省 | 郴州市 |
| 10020 | 惠州 | 惠州 | 广东 | 惠州 | 四川省 | 广安市 |
| 10033 | 枣庄 | 枣庄 | 山东 | 枣庄 | 山东省 | 枣庄市 |
| 10035 | 深圳 | 南平 | 福建 | 南平 | 福建省 | 不详 |
| 10038 | 济宁 | 济宁 | 山东 | 济宁 | 山东省 | 济宁市 |
| 1004 | 连云港 | 连云港 | 江苏 | 连云港 | 江苏省 | 连云港市 |
| 10042 | 德州 | 德州 | 山东 | 滨州 | 山东省 | 德州市 |
| 10043 | 青岛 | 聊城 | 不详 | 不详 | 山东省 | 聊城市 |
| 10046 | 深圳 | 汕尾 | 广东 | 汕尾 | 广东省 | 汕尾市 |
| 1005 | 新乡 | 新乡 | 河南 | 新乡 | 河南省 | 新乡市 |

  - (iii) rank the cities
  - (iv) For the rest, we made them one hot encoder

# Experimental setup and Results

Dummy variables:

After we do all the data transformation and cleaning process, we change the left object variables into dummy variable. This step is fault, but we still do that, cause it is easy.

# Experimental setup and Results

- feature selection
  - 200 the most important features
  - Combine 5 models to rank the feature importance
    - Pearson correlation selector, chi-squared selector, random forest selector, logistic selector, and RFE

# Choose the model and use the stacking method

Modeling:

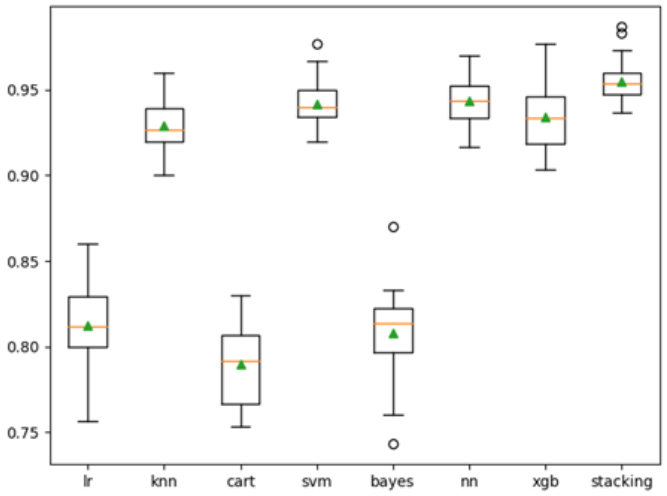
We choose the logisticregression, knn, decisiontree, svm, gaussianbayes, nn, xgb model and combine them to make a stacking model

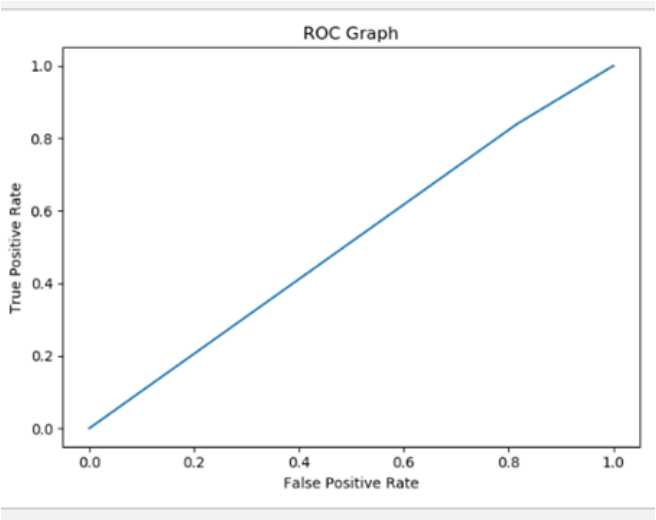
K-fold:
then we use the 10-fold and repeat = 3 to train the model.

Finally we can get the result, like the box plot or the mean(score)

# Experimental setup and Results

- Accuracy




```
[0 1 1 ... 1 0 1]
[[ 3481 15407]
 [  179   933]]
0.057099143206854344
0.8390287769784173
0.1069218427687371
0.5116628425341049
```

Prediction
Confusion matrix
precision score, recall-score, f1-score and AUC

# Conclusion

- Still have a lot to improve
  - Dummy variable (factorization)
  - Feature engineering
  - Building model