Name: Xueqi Zhou, Jiaxi Jiang, Meijiao Wang,
Proposal: Overdue Rate Prediction based on Personal and Behavior Information of Clients

Based on user's historical behavior data of the Chinese Credit Platform to predict the probability of users overdue in the next 6 months. In normal life, some people find it hard to obtain loans because of their no or less credit history. In order to increase the tolerance of loans for people with no bank account or credit history, credit institutions will use various alternative data: Telecommunications, transaction information, and other historical behavior data of customers can be a standard to predict customer repayment ability. Based on these data, various machine learning methods are used to make these predictions to calculate and reduce the expected loss level of credit.

We will use the database from Mirror Risk Control System, which is the first real risk control model based on big data in the industry, it includes 13.6 million online users and 12.86 billion pieces of data accumulated in 8 years. It is large enough to train machine learning.

There are several reasons why the proposal is worthwhile, cause Bank-like financial systems need to know their clients and try to avoid debts.

(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.

To put it simply, we want to establish fast screening algorithms to quickly determine whether to lend to this person or not, thereby improving the efficiency and profit of the financial institution. So, the important thing is to build a system, we need to extract the information from data, and find a good classifier.

We evaluate the user's current credit status from an average of 400 data dimensions, score each borrower's credit, find the probability that the user is overdue within 6 months, convert it into a 2 classification problem, and use AUC as an indicator for evaluation.

Methodology:

(1) Start with data cleaning, and deal with multi-dimensional missing values to eliminate outliers.
(2)Feature engineering, extract feature information, rank features by training
(3)Select features to deal with the imbalance of categories, use the method like over sampling.
(4)Design the model, use models like xgboost, logistic reg, svm， neural network and its transformation, we will try ANN, maybe RNN, we will use python, because it's relatively easy and open-source.
(5)Adjust parameters, combine the model, get the relatively good model.

In order to ensure the validity model, we will find similar projects on Kaggle as a reference.

After the necessary manipulation, the AUC should increase, which means in real life that the accuracy for the model to predict whether the person will overdue increases. The prediction can be a strong basis for bankers to decide whether to give a loan, and largely lower the risks for overdue debt.

Detailed Research Timetable

Data cleaning	17th-20th
Feature engineering	21th-23th
Feature selection	24th
Training model and parameter optimization	25th
Model diagnosis	26th
Model ensembling	27th-29th
Report and presentation	30th