**Loan Default Analysis of Lending Club**

**Motivation**

P2P is peer-to-peer, person-to-person or person-to-person, and its spirit is to remove intermediaries. In terms of finance, after P2P implies financial intermediation, it is indeed very obvious that the emotional characteristics of P2P decline, and it can almost make everyone in the province. Have enough fun, so for some poor people, financial novices are a huge help.

It has been developed in Europe and the United States for nearly 10 years. The data set used here is a P2P company lending club that has been listed. However, this method that does not rely on intermediaries is not 100% safe. Through online matching, the United States may lend. The risk of the party is very high. If there is no management, the risk will be great. In 2014, P2P companies in mainland China that absconded investors’ money will have 200 days. Knowing that a good P2P is important and important, it is very important to break the model. Therefore, I want to rely on the technology of connected information engineering, coupled with my interest in finance, to come into contact with the following models:

• Run through the API of the bank and the bureau to provide relevant customer postage and joint collection information.

• Through internal cloud technology, through machine learning to model data to predict the default rate and convey the safest model model.

• Implement blockchain technology, combined with zero-knowledge proof, upload customer data after de-identification, so that the data is free from information security issues.

• The establishment of a P2P platform eliminates the need to wait for financing time. Funders who provide urgently needed funds have reliable financial support and credit at the same time, combined with debt transfer services, to create more investment opportunities and cash flows.

**Dataset**

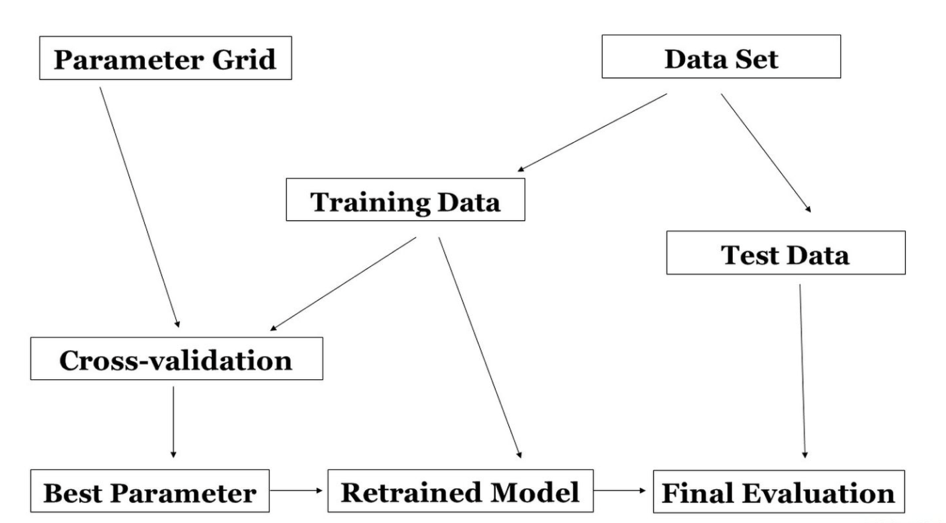
The dataset contains complete loan data for all loans issued through four seasons in 2018, including the current loan status (Current, Charged-off, Fully Paid) and latest payment information. Additional features include credit scores, number of finance inquiries, and collections among others. The file is a matrix of about 495,242 observations and 102 variables.

**Analyzing Tool**

Linear regression, random forest, XGBoost.

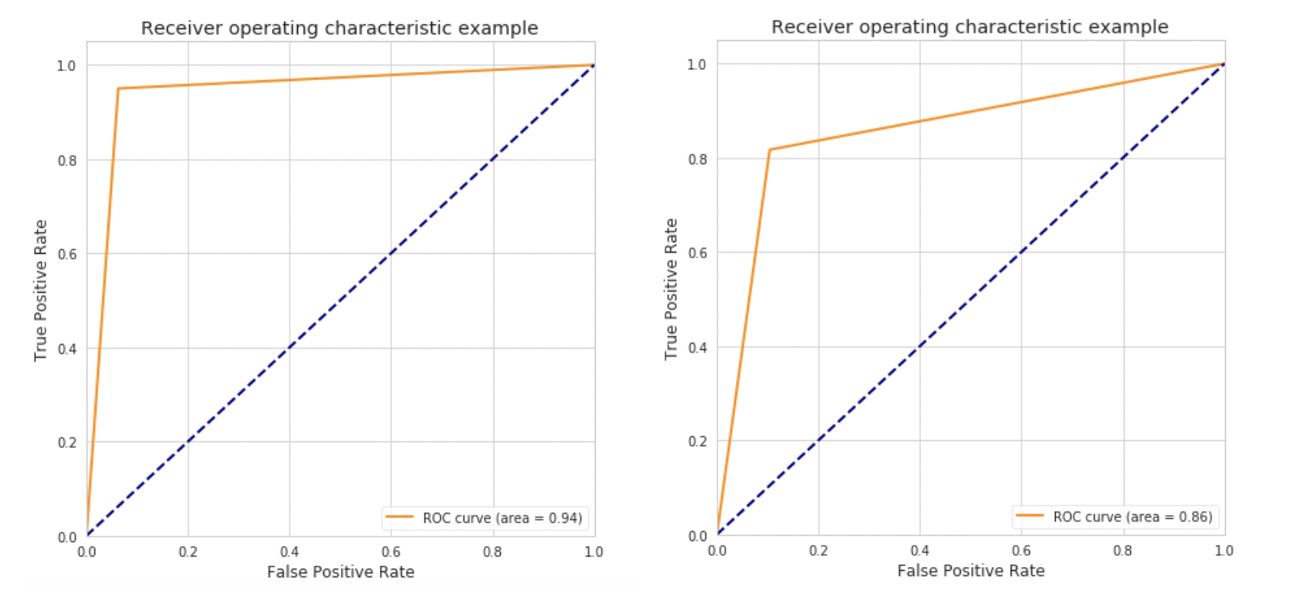
**Implementation**

In order to build a complete P2P lending default model platform, this project designs a three-tier system architecture. As shown in the figure above, the user's activity records will be collected and safely stored in the server and blockchain. We are hosted on Google Cloud Platform as the main server, serving as a data storage center and a relay for open data access. Google Cloud Platform provides secure network protection, data integrity and reliability of system execution under the Google domain. The main tasks are divided into four parts: data inventory acquisition; unsupervised learning model for estimating default rate; Ethereum blockchain service (Blockchain as a Service, BaaS) and other open data access engines; cash flow and Investment transaction systems, these systems provide complete services and security of the P2P lending ecological chain. Take security as an example. In the system, we split the user’s account data into two parts: data (Metadata) and transaction (Transaction), which are stored in the main server and the Ethereum blockchain service separately to avoid centralized Data management balance, etc. were attacked or tampered with by malicious persons.

In terms of training model, I designed a two-layer system architecture. The first layer is to use logistic regression as a classifier to predict default or non-default after the case comes in, avoiding the problem of data imbalance caused by the low default rate of P2P lending. . The second layer uses the random forest algorithm and XGBoost to predict the rate of return, and if it is greater than 0.6, it goes through the case. In the random forest prediction model, in order to reduce the impact of the decision tree with low training accuracy on the entire model, I designed a weighted random forest model. The core is to divide the training sample into two parts: one part is used as the training of the traditional random forest model The sample is used to train all random numbers; the other part is the pre-test sample. After the training is completed, each decision tree is tested separately and its classification accuracy is calculated. This accuracy is used as the weight of the corresponding decision tree. It is the weighted random forest model. In addition, the selection of various parameters such as the number of decision trees L, the number of pre-test samples X, etc., I introduced the particle swarm algorithm into the model to iteratively optimize the parameters in the weighted random forest algorithm.

**Evaluation Method**

In terms of models, random forest performs well, with an accuracy rate of about 95%, while XGBoost is only 85%. Compared with the following ROC picture, it can be seen that the predicted True Positive of random forest is much larger than XGBoost.





The left picture is the interface diagram of the website. We provide borrowers, lenders and corporate parties to work together and look forward to creating a platform for inclusive finance so that everyone can enjoy the rights and interests of financial services.

**Analysis Results**

* **Revenue**

The revenue of Lending Club is more dependent on the scale of loan transactions. The operating income of Lending Club is mainly loan transaction fees, investment service fees and account management fees. According to the 2018 financial report, loan transaction fees are US$449 million, accounting for 78.08%, and investment service fees are US$87 million, accounting for 15.16% , Lending Club’s revenue is dominated by loan transaction fees, and is more dependent on the scale of loan transactions.

* **Internal Risk Control Mechanism**

The mechanism of internal risk control of Lending Club is relatively mature. In the personal loan approval process, the Lending Club risk control mechanism is based on FICO credit score data. After the loan applicant submits an application, the platform will review it according to the standards. The applicant's access conditions are as follows:

FICO credit score is not less than 660 points;

The debt-to-income ratio is less than 40%;

In the credit report, at least two accounts are in normal use;

No more than 5 times of credit inquiry in the last 6 months;

Provide a credit history of at least 36 months.

* **Credit Rating Model**

LendingClub implements a differentiated risk pricing strategy, and its credit rating model is relatively mature. For standard personal loan business, Lending Club comprehensively evaluates the loan applicant’s FICO scores, credit records and credit risk indicator information in the loan application, verifies the applicant’s assets and income, and uses a proprietary credit rating model to risk the applicant Evaluation, combined with the amount and duration of the loan applied for, determines the loan credit rating and loan interest rate. Loan credit rating is divided into seven grades, A-G, and five sub-grades are set up under each grade. LendingClub comprehensively considers market benchmarks and credit risks when setting interest rates, and implements a differentiated risk pricing strategy. The lower the loan credit rating, the higher the loan interest rate. The lower the credit rating of a loan, the higher the loan default rate, and the Lending Club credit rating model is more mature.

**Opinion**

* **Risk Identification**

From the perspective of risk identification, the risk control model online of Lending Club relies more on quantitative indicators such as credit scores or annual income. For risk assessment, it should also consider the influence of American consumers' concepts of early consumption and overconsumption and seasonal factors. The higher the income, the greater the personal consumption expenditure, and may be more likely to fall into the dilemma of insolvency. Individuals applying for loans are mainly used for debt consolidation and credit card repayment, and loan customers have greater financial risks. In the fourth quarter, consumers' financial pressure on holiday spending has increased the possibility of overdue loans and bad debts.

* **Credit Rating Standard**

From the perspective of credit rating standards, according to the 2018 financial report, Lending Club's loan transaction scale and operating income grew slowly from 2015 to 2018. The platform adjusted credit rating standards and pricing strategies to maintain transaction scale development. Since 2015, the loan overdue rate has exceeded the average level of US bank consumer loans. While maintaining business growth, Lending Club should also implement risk control mechanisms to ensure loan quality, control default risks, and pay attention to operating efficiency.

* **High-Risk Market**

From the perspective of high-risk customers, under normal circumstances, P2P loan customers are mainly high-risk groups who cannot meet the traditional bank credit standards. The return of credit funds depends on the customer's repayment ability and willingness to repay. Lending Club should strengthen the screening and monitoring of customer segments with high default risk based on the customer characteristics of different loan credit ratings, promptly warn of the risk of overdue loans, and actively carry out collection work to reduce the risk of bad debts.