Research on Investment Model of Internet Financial Loan Platform

Hua-ling LIU\textsuperscript{1,*}, Hai-zhen CHEN\textsuperscript{1}, Yu-jie DING\textsuperscript{2} and Xue ZHANG\textsuperscript{2}

\textsuperscript{1}Shanghai University of International Business and Economics, Shanghai, China
\textsuperscript{2}Shanghai University of International Business and Economics
*Corresponding author

Keywords: Data mining, P2P lending, Portfolio.

Abstract. In recent years, Peer-to-peer (P2P) lending platforms are in a stage of rapid development in China, which gradually become one of the important investment options for investors. A key challenge for personal investors in P2P lending market is how to get the expected benefit under certain risks. In this paper, we present the efficient frontier for investors with the combination theory of random forest, nuclear regression and Markowitz portfolio theory. The final experimental result reveals that the frontier can effectively provide investors more investment options.

Introduction

Due to the rapid development of new internet technology recently, P2P mode of network lending and small-scale lending have increasingly came into people's lives. Peer-to-peer (P2P) lending was the practice of lending money to individuals or businesses through online services that match lenders with borrowers. P2P lending is playing an important role in template, modified in the loan market now, which can effectively solve the financing difficulties for small business.

In the process of development, there will be problems of adverse selection and moral hazard because of asymmetric information. Therefore, it is very important to evaluate the income and risk of loan funds effectively so as to obtain expected return under the certain risks. Based on the situation, this paper will study credit default rate through the real data of Paipai Lending, using the nuclear regression model to give different weights to different probability of successful lending. At the same time, the Markowitz portfolio theory is adopted to obtain the benefit and risk of loan, providing efficient frontier for investors, giving the optimal combination of the recommended bids for the lenders under different risk preferences.

Data and Model

Data Source

An open API from Paipai Lending is used to call the customer ID information, and get total 300000 data which has 39 index dimensions, and spanning from December 2015 to February 2017. The following is a description of some of the variables.

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Type of data</th>
<th>Variable name</th>
<th>Type of data</th>
</tr>
</thead>
<tbody>
<tr>
<td>LastBidTime</td>
<td>DateTime</td>
<td>BorrowName</td>
<td>String</td>
</tr>
<tr>
<td>AuditingTime</td>
<td>DateTime</td>
<td>Gender</td>
<td>Int</td>
</tr>
<tr>
<td>RemainFunding</td>
<td>Decimal</td>
<td>EducationDegree</td>
<td>String</td>
</tr>
<tr>
<td>DeadLineTime</td>
<td>String</td>
<td>GraduateSchool</td>
<td>String</td>
</tr>
<tr>
<td>CreditCode</td>
<td>String</td>
<td>StudyStyle</td>
<td>String</td>
</tr>
<tr>
<td>ListingId</td>
<td>Int</td>
<td>SuccessCount</td>
<td>Int</td>
</tr>
<tr>
<td>Amount</td>
<td>Decimal</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Model Construction

1) Decision Tree and Random Forest

Decision tree is a common machine learning method. For example, if you want to classify a new sample through a model which is learned from a given data set, the classification task can be regarded as whether the current sample is positive or not. In general, a decision tree contains a root node, a number of internal nodes, and a number of leaf nodes. The leaf node corresponds to a decision process. And we hope that the branches of the decision tree contain as much of the same kind as possible, that is, the high purity of the nodes.

Random forest is a classifier that contains multiple decision trees, whose output category is determined by the mode of classes that individual trees outputs. Random forest algorithm is a combination of Bagging algorithm and Random Subspace algorithm, the basic constituent unit is a decision tree, which can be a two-fork tree, or a multi-fork tree, through the combination of multiple decision trees, and the \( h_1(x), h_2(x), \ldots, h_{nTree}(x) \) (the classification is obtained by the Random forest classifier (figure.1)), which is classified by the final voting for the unknown class. If the regression problem is solved, the average of all decision tree results is output.

![Figure 1. Random Forest Model Construction.](image)

2) Nuclear Regression

The kernel regression method is a non-parametric regression method and does not make any assumptions about the original data. And each point is given a different weight based on the distance from \( x \). Then get an intuitive loss function:

\[
\text{Loss} = \sum_i w_i(x)(y_i - c)^2
\]

According to \( \frac{\partial \text{Loss}}{\partial c} = 0 \), we can get a new sample point,

\[
c^* = \frac{\sum_i w_i y_i}{\sum_j w_j} = \frac{\sum_i \sum_j w_j(x)y_i}{\sum_j w_j(x)}
\]

Here \( K(\bullet) \) is the nuclear function, and \( w_i(x) = K(x, x_i) \), the estimate point is used as the core, and to control the proportion of each sampling point, the smoothing the parameter \( h \) is used to control the size of the kernel, the function \( K(\bullet) \) takes the y-axis as the axis of symmetry and takes maximum value when independent variable is 0. In this paper, we choose Radial Basis Function as a kernel function. The form is as follows:

\[
K(x, z) = e^{-\frac{(x-z)^2}{h}}
\]
**Empirical Analysis**

**Data Processing**

1) Data Deletion

Delete the data which are still in the repayment period.

2) Data Identification

According to the repayment condition, if the overdue period is over 30, the sample mark is 1, otherwise the mark is 0.

3) One-hot Coding

Transform categorical variables by one-hot encoding. For example, \{sex: male, female, unknown\} will be converted into the form of \{\{1,0,0\},\{0,1,0\},\{0,0,1\}\}.

**Feature Combination**

This article uses the Apriori algorithm to combine the categorical variables of the original data. By looking for frequent items that are fraudulent and non-fraud categories, respectively, and comparing frequent items of different fraud categories. According to the support and confidence parameters, this article found a total of 21 frequent items. As shown in the table below.

<table>
<thead>
<tr>
<th>Combinations name</th>
<th>Variables of combination</th>
</tr>
</thead>
<tbody>
<tr>
<td>cond0</td>
<td>creditvalidate_0,graduateschool_0,ncicidentitycheck_0,phonevalidate_1,educatevalidate_0</td>
</tr>
<tr>
<td>Cond1</td>
<td>edu_degree_0,creditvalidate_0,ncicidentitycheck_0,phonevalidate_1,educatevalidate_0</td>
</tr>
<tr>
<td>Cond2</td>
<td>videovalidate_0,creditvalidate_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0</td>
</tr>
<tr>
<td>Cond3</td>
<td>edu_degree_0,videovalidate_0,creditvalidate_0,phonevalidate_1,educatevalidate_0</td>
</tr>
<tr>
<td>Cond4</td>
<td>videovalidate_0,creditvalidate_0,graduateschool_0,phonevalidate_1,educatevalidate_0</td>
</tr>
<tr>
<td>Cond5</td>
<td>creditvalidate_0,ncicidentitycheck_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0</td>
</tr>
<tr>
<td>Cond6</td>
<td>edu_degree_0,videovalidate_0,creditvalidate_0,ncicidentitycheck_0,phonevalidate_1,educatevalidate_0</td>
</tr>
<tr>
<td>Cond7</td>
<td>videovalidate_0,creditvalidate_0,graduateschool_0,ncicidentitycheck_0,phonevalidate_1,educatevalidate_0</td>
</tr>
<tr>
<td>Cond8</td>
<td>edu_degree_0,videovalidate_0,creditvalidate_0,certificatevalidate_0,phonevalidate_1,educatevalidate_0</td>
</tr>
</tbody>
</table>

We will use frequent items as new variables, then the original data set will have 21 new variables.

**Feature Selection**

In the random forest algorithm mechanism, because each node selection is random, then the redundant variables will reduce the chances of important variables being selected. Therefore, the redundant variables will result in the results of the algorithm. So, we need select the variables which are more important. In this paper, we select important variables through Apriori algorithm.
Evaluation of Network Borrowing Overdue Rate

We choose the overdue rate of P2P network loan (the overdue value is set to 1, not overdue to 0) as dependent variable, and then study the problem of overdue rate of network borrowing through Random forest algorithm, finally we get the probability of every loan. And we give the confusion matrix of the test set, it can be seen that in the actual fraud, the algorithm of fraud recognition finds about 90% of the fraud target. So, the algorithm is effective.

Investment Recommendation Based on Nuclear Regression

For each new loan $L_i$, assuming its yield is $\mu_i$, and based on the kernel weight model, the expression formula of its yield is as follows:

$$\mu_i = \sum \omega_y(d_j) \times R_j$$  \hspace{1cm} (4)

Define the standard deviation as follows:

$$\sigma_i = \sqrt{\sum \omega_y(R_j - \mu_i)^2}$$  \hspace{1cm} (5)

Then we give an effective frontier of investment based on the $\mu_i, \sigma_i$ and portfolio theory. As shown below:

Table 3. Confusion matrix.

<table>
<thead>
<tr>
<th></th>
<th>Prediction of No-Fraud</th>
<th>Prediction of Fraud</th>
</tr>
</thead>
<tbody>
<tr>
<td>No actual fraud</td>
<td>62148</td>
<td>5045</td>
</tr>
<tr>
<td>Actual fraud</td>
<td>280</td>
<td>2459</td>
</tr>
</tbody>
</table>
Conclusions and Perspectives

It is clear that P2P network lending has developed rapidly, which is really an effective complement to traditional finance. With the further development of Internet, it gradually become an important financial choices for investors. So it is important for the investors to construct the optimum portfolio to maximize the benefits under the certain risk.

In this paper, we give the efficient frontier which can give investors the personalized portfolio recommendation based on machine learning and large data analysis technology, it can help investors avoid irrational investment behavior.

Acknowledgement

This work was supported in part by the project of National Social Science Foundation-Research on Internet Information Behavior and Governance for National Public Security (No.16ZDA055), the “Big data-driven management and decision research” project of National Natural Science Foundation—Study on the Risk Identification Decision of the Three Typical Business Market Driven by Big Data (No.91546121),and projects of National Natural Science Foundation Research on the Management of Web2.0 User Authoring Content Oriented to Uncertainty(No.71231002).

References