Loan User Repayment Forecast

1. Introduction

1.1 Background

Internet finance has been extremely hot in recent years, and a large amount of capital and talents have poured into this field to discover rich value. Regardless of whether it is investment or wealth management or loan lending, risk control is always the core foundation of the business. Among all Internet financial products, micro-borrowing is considered to be the most risky segment because of the particularity of its main service target.

1.2Problem

This report to analyze the credit status of borrowers who applied for “micro-credit” from user behavior data to determine whether they are overdue.

2. Data understanding and cleaning

2.1 Data sources

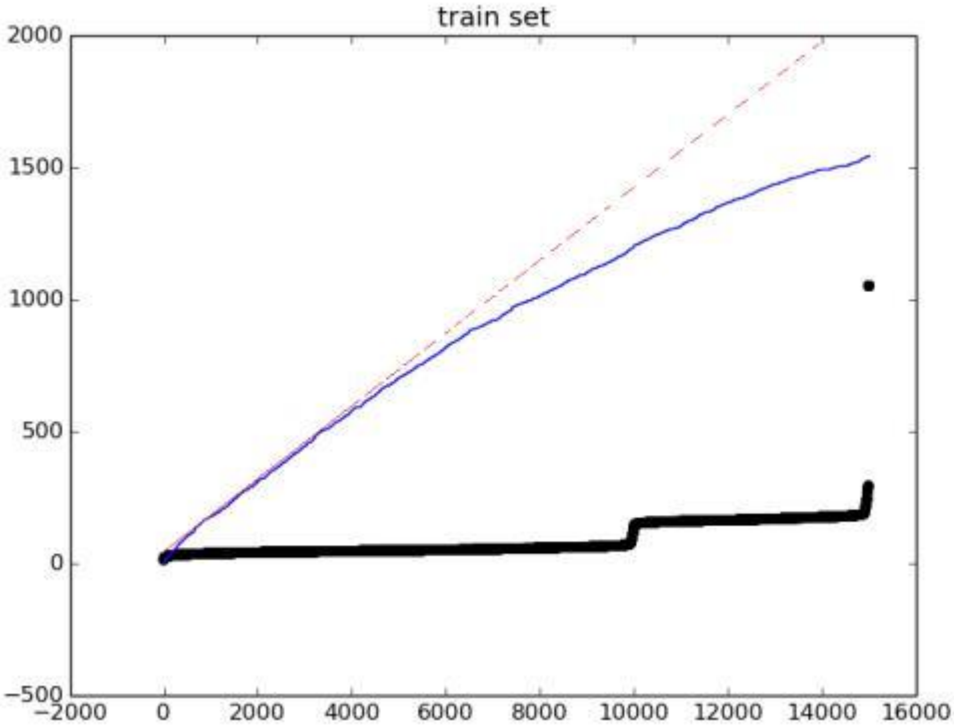
The data source comes from the actual combat data of the micro-borrowing industry, not only conventional labeled data but also unlabeled data. There are 1138 features in the data set, which are dominated by users' multi-dimensional behavior data. There are both numerical and categorical features.

2.2 Data cleaning

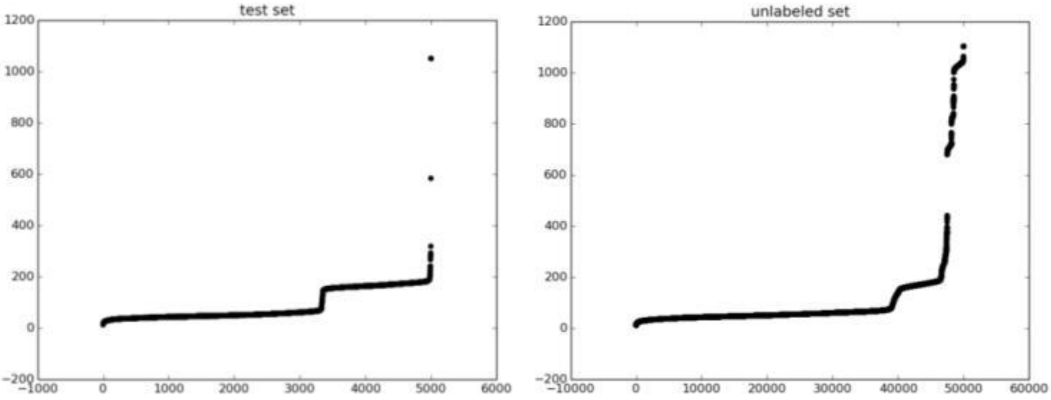
Mainly the processing of missing values. Most of the samples in the question data have missing values, and there are many missing values, and some samples even have thousands of missing values. The commonly used method of missing value processing is missing value filling (using the feature mean, median, etc. of the same category of data), but because the test set also has a large number of missing values, this method is not used. Since the training set and test

Sets all have this characteristic, so it is better to treat missing values as a feature.

I counted the number of missing values for each sample in the training set, and sorted them according to the number of missing values. The serial number is the abscissa and the number of missing values is the ordinate.



Further statistics of missing values on test and unlabeled data sets.



Further, I deleted the data with more than 194 missing values in the training data (that is, the samples with a discrete value of 5 missing values). This part of the sample contains too many missing values, which makes the learning of the model difficult, and even introduces noise, causing overfitting. So after removing this part of the data, the auc on the line increased by 0.002, which is not small Promote.

3 Feature Engineering

Feature engineering is very important, I invested a lot of time and energy in this part.

3.1 Sorting features

Sort the 1045-dimensional numeric features in the original features from small to large to get the 1045-dimensional sorted features. Sorting features are more robust to abnormal data, making the model more stable and reducing the risk of overfitting.

3.2 Discrete feature

There are two methods for feature discretization: one is equal division (equal division according to the value range), and the other is equal division (equal division according to the number of samples). The numeric type features are discretized in equal amounts: first, each dimension feature is sorted according to the numerical value, and then it is evenly divided into 10 intervals, that is, discretized into 1 ~ 10.

3.3 Counting feature

The features have been discretized earlier, taking the sample with uid 1 as an example, after discretization, its features are 5, 3, 1, 3, 3, 3, 2, 4, 3, 2, 5, 3, 2 , 3,2 ... 2,2,2,2,2,2,2, you can further count the number of occurrences of 1 ~ 10 in the discrete features ni (i = 1,2,…, 10), and get 10-dimensional counting features. The classifier was trained based on the 10-dimensional features, and the online score was about 0.58, indicating that the 10-dimensional features have good discriminability.

3.4 Category feature coding

The data contains 93-dimensional category features. Many algorithms (such as logistic regression, SVM) can only deal with numerical features. In this case, the category features need to be encoded. One-Hot encoding is used to obtain the 01 feature.

3.5 Cross feature

Such multi-dimensional features may cause dimensional disaster on the one hand, and may easily lead to voer fitting,so requires dimensionality reduction. Common dimensionality reduction methods include PCA and t-SNE (high computational complexity). I tried PCA, and the effect was not good. The reason is that most features contain missing values and there are too many missing values.

In addition to using dimensionality reduction algorithms, feature selection can also be used to reduce feature dimensions. There are many methods for feature selection: maximum information coefficient (MIC), Pearson correlation coefficient (measures linear correlation between variables), regularization method (L1, L2), and model-based feature ranking method. The more efficient is the last method, which is the feature ranking method based on the learning model. This method has a benefit: the process of model learning and the process of feature selection are performed simultaneously.

4 Predictive Modeling

4.1 Model design

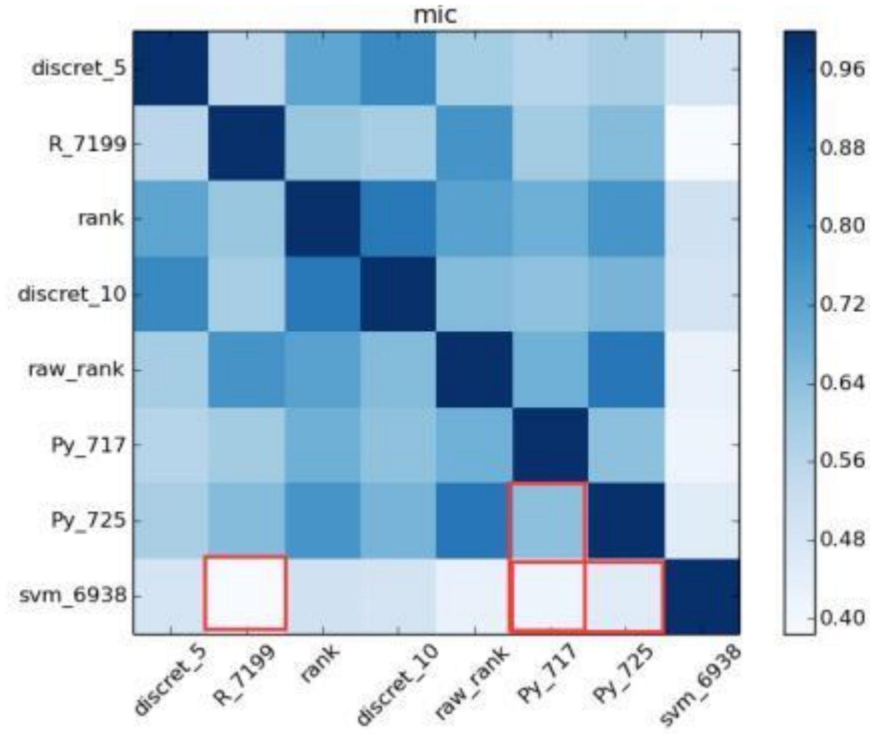
After obtaining the ranking features, discrete features, and counting features based on the original features, we perform feature selection on these features. The method of feature selection has been introduced earlier. Based on xgboost, the process of training xgboost is to perform feature importance Sorting process.

After getting the importance of the features, we can keep the top N1 original features, top N2 ranking features, and top N3 discrete features. (Counting features are only 10-dimensional, so feature selection is not done).

4.2 Model fusion

The method is simple weighted fusion.

For model fusion to achieve good results, a single model is required to be diverse (ie, different). In order to visually observe the differences between the single models, we calculated the maximum information coefficient (MIC) between them, and drew it in the form of a confusion matrix (the lighter the color, the smaller the correlation)


mic
discret_5
R_7199
rank
discret_10
raw_rank
Py_717
Py_725
svm_6938
discret_5
R_7199
rank
discret_10
raw_rank
Py_717
Py_725
svm_6938