Final Project-Checkpoint

What We Having Done

In view of the current widely used mobile payment scenarios, the work our team has to do is to predict the risks in the payment process. This work is of great significance.

Collecting data and Preprocessing data:

AliPay is a very popular mobile payment platform. The first work we do is data collection and data processing. Collected about 1.2G of desensitized mobile payment data sets from the public data of the platform, but after data analysis, it was found that it contains a lot of useless dirty data, we need to perform simple preliminary preprocessing and screening of the data Out of the dataset we can use. In the end, we generated the part_train_data.csv file, which contains preliminary training data sets that can be used. (Author: Zijun He)

Data Analysis and Feature extraction:

After the data set exists, we have to further analyze the data set, combine the actual application scenarios, select features, find useful features, and prepare for the risk prediction model behind. Finally, a 285-dimensional feature vector is generated. (Author: Zijun He)

Modeling:

We abstract risk prediction into a classification problem and predict whether there is a risk. Therefore, we use a variety of classification models to train, such as: a simple single model such as DecisionTree, and some popular integrated models, AdaBoost and GBDT, There are attempts to compare the experimental results and analyze the results. (Author: Erdun E)

Evaluating model:

After the model is established, we must evaluate the performance of the model and establish an evaluation standard to determine the direction of model optimization. This is a two-category problem, but it is not a simple classification problem. Risk prediction is a small probability event, but it is unreasonable to use accuracy. Therefore, we will combine several classification evaluation indicators as an evaluation of our problem Standard, 0.4 * TPR1 + 0.3 * TPR2 + 0.3 * TPR3. In addition, we also visualize the model results to more intuitively evaluate model performance, using ROC curves. (Author: Erdun E)

At present, the score of the decision tree model is 0.29, the score of the GBDT model is 0.41, and the score of the AdaBoost model is 0.36.

Further planning

Index	What to do	Author
1	The error analysis of the experimental results of several models is carried out to find the bottleneck existing in the performance of the current model and the direction of optimization.	Zijun He
2	Use the currently popular Bagging integration method Random Forest model to train and improve model performance	Erdun E
3	Use neural network methods to train models, such as: MLP and DNN models to Improve the performance of risk prediction models	Erdun E
4	The stacking method combining GBDT and DNN is used to train the model. The GBDT method can be used as a feature selection, and the output is used as the input of the DNN model, thereby further improving model performance.	Zijun He

The final output is the experimental result of higher performance, and the relevant code supplement file is in Model.ipynb.