ASSESSMENT OF CREDIT RISK BY USING BIG DATA TOOLS

Li Ke, Postgraduate 8PM0I
E.I. Gubin Ph.D., Associate Professor
Tomsk Polytechnic University
E-mail: ke1@tpu.ru

Introduction

Credit risk assessment is estimating the probability of loss resulting from a borrower's failure to repay a loan or debt, it's very important and full of challenging. Today the study of financial credit risk assessment attracts increasing attentions in the face of one of the most severe financial crisis ever observed in the world. The accurate assessment of financial credit risk and prediction of business failure play an essential role both on economics and society [1]. In this paper, we introduce steps of create a credit risk model by using big data tools, contains the method of data preparation and the method of create model. After training model, we also analyze key factors of customers whether repays the loan on time. In this paper we used different big data tools like SAS, Python, it's good for big data beginner to study and learn.

Data Description

This dataset is about personal information of customers, it contains 24 variables, and 3000 observations. The target variable is GB, it means the customer is a good or bad borrowers. Other variables contain age, amounts of children, income, region, etc.

Data Preparation

The first step is data cleaning. Data cleaning is a very important step of data preparation, if we ignore this step, we will get wrong result in final [2]. In general, there are 6 problems we need to care about, they are missing data, mistake of data, outliers of data, duplicate cases, multicollinearity of data and digitalization of data [3]. In this part, we use SAS to clean data [4]. SAS is a statistical software for data analysis, and it is easy, safe and stable. In this dataset we find 500 missing values and some outliers, we fill up missing values by checking the frequency of the column, and drop these outliers.

After data cleaning, we split data into training data (75%) and test data (25%).

Credit Risk Model

The next step is creating credit risk model, and we use random forest model to classify good and bad customers. We use RandomForestClassifier to create this model, and also include cross validation. This method is come from the most popular library in machine learning area-scikit learn. Scikit-learn is a famous library in Python, also in this work we used Pandas, we could analyze dataset in it.

Random forest is a good algorithm to solve classification and regression problems. In our work, we need to classify good and bad customers, so we use random forest algorithm. Also, we could use logistic regression algorithm, because this is a binary classification problem.

After training our model and adjust parameters, we could get a classification report of our model. On Figure 1 we can see we have 70% accuracy on test data. That means we have 70% accuracy to predict whether a customer repays the loan on time.

	precision	recall	f1-score	support
0	0.69	0.70	0.69	351
1	0.71	0.70	0.70	366
accuracy			0.70	717
macro avg	0.70	0.70	0.70	717
weighted avg	0.70	0.70	0.70	717

Fig. 1. Classification report of random forest model

Key factors

In this model, each feature has different weights, we could analyze these weights to get key factors of a customer whether repays the loan on time.

On Figure 2 we can see top 10 features of our model, key factors of a customer are age, working time, cash, income, region, etc. That means these features we will consider at first when we want to loan to customers.

```
[(0.1392, 'AGE'),
 (0.1147, 'TMJOB1'),
 (0.0967, 'CASH'),
 (0.0941, 'TMADD'),
 (0.0697, 'INCOME'),
 (0.06, 'REGN'),
 (0.0474, 'LOANS'),
 (0.0468, 'PERS_H'),
 (0.0415, 'prof_'),
 (0.0397, 'product_'),
```

Fig. 2. Top 10 features and their weights

Conclusion

In this paper, we analyze our data and create a credit risk model, and finally we get 70% accuracy. We could use this model to predict whether a customer repays the loan on time. Also, we analyze key factors. We could know that when we want to loan to customers, we need to consider about these features like age, working time, cash, income, and region of customers.

References

1. Chen N, Ribeiro B, Chen A. Financial credit risk assessment: a recent review[J]. Artificial Intelligence Review, 2016, 45(1): 1-23.
2. Huang Shan. Data cleaning for data analysis / Huang Shan, E. I. Gubin // Молодежь и современные информационные технологии : сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых учёных, 3-7 декабря 2018 г., г. Томск. — Томск : Изд-во ТПУ, 2019. — [С. 387-388].
3. Губин Е. И. Методология подготовки больших данных для прогнозного анализа / Е. И. Губин // Современные технологии, экономика и образование : сборник трудов Всероссийской научно-методической конференции, г. Томск, 27-29 декабря 2019 г. — Томск : Изд-во ТПУ, 2019. — [С. 27-29].
4. Губин Е. И. Использование программных инструментов SAS для подготовки больших данных / Е. И. Губин // Современные технологии, экономика и образование : сборник материалов II Всероссийской научно-методической конференции, г. Томск, 2-4 сентября 2020 г. — Томск : Изд-во ТПУ, 2020. — [С. 52-54].