CREDIT RISK ASSESSMENT USING MACHINE LEARNING BASED ON PYTHON

Jiang Daqing [1], Губин Е.И. [2]
1 НИ ТПУ, ИШИТР, гр. 8ПМ2И, e-mail: dacin1@tpu.ru
2 НИ ТПУ, ИШИТР, ОИТ, доцент, e-mail: gubine@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. Minimizing the risk of default is a major concern for financial institutions. For this reason, commercial and investment banks, venture capital funds, asset management companies and insurance firms, to name a few, are increasingly relying on technology to predict which clients are more prone to stop honoring their debts. Machine Learning models have been helping these companies to improve the accuracy of their credit risk analysis, providing a scientific method to identify potential debtors in advance. In this paper, we introduce steps of create a credit risk model by using Machine Learning based on Python, contains the method of data preparation and the method of create model [1].

Data preparation

The dataset is derived from information on bank loan customers and contains 24 features with 3,000 records. First of all data cleaning, the quality of data cleaning has a direct impact on the accuracy of our results. In general, we need to deal with the following issues: missing values, outliers, duplicate values and multicollinearity. I recommend using python with Pandas, Numpy in this process.

Before setting up the machine learning algorithms, we still need to perform some preprocessing. Considering that most Machine Learning algorithms work better with numerical inputs, we'll preprocess our data using Scikit Learn's LabelEncoder for the binary variables and Pandas' get_dummies for the other categorical variables.

Before setting up the machine learning algorithms, we need to perform some preprocessing. Considering that most Machine Learning algorithms work better with numerical inputs, we'll preprocess our data using Scikit Learn's LabelEncoder for the binary variables and pandas' get_dummies for the other categorical variables. Final, we need to split the data into training (70 %) and test (30 %) sets [2].

Machine Learning Model

Logistic regression is a variation of ordinary regression which is used when the dependent variable is a binary variable and the independent variables are continuous, categorical, or both. Unlike ordinary linear regression, logistic regression does not assume that the relationship between the independent variables and the dependent variable is a linear one. Nor does it assume that the dependent variable or the error terms are distributed normally.

The form of the model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \quad (1)$$

where p is the probability that Y=1 and X_1, X_2, …, X_k are the independent variables (predictors). $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are known as the regression coefficients, which have to be estimated from the data. Logistic regression estimates the probability of a certain event occurring.

Logistic regression, thus, forms a predictor variable (log (p/(1-p)) which is a linear combination of the explanatory variables. The values of this predictor variable are then transformed into probabilities by a logistic function. This has been widely used in credit scoring applications due to its simplicity and explainability. Our assessment of risk is formally a binary classification problem, so we use logistic regression for modelling, and here we use Scikit-learn in Python, currently the most popular machine learning library.

In this model, each feature is given a different weight and we chose eight main features as input variables and trained the model [3].

	coef	std err	z	P>|z|	[0.025	0.975]
Age	-0.0434	0.004	-10.180	0.000	-0.052	-0.035
EC_card_holders	-0.9689	0.094	-10.269	0.000	-1.154	-0.784
Telephone	-0.8082	0.107	-7.573	0.000	-1.017	-0.599
Num_of_running_loans	0.2289	0.038	5.988	0.000	0.154	0.304
Num_Mybank_Loans	-0.2273	0.047	-4.848	0.000	-0.319	-0.135
Time_at_Job	-0.0012	0.000	-2.932	0.003	-0.002	-0.000
Num_in_Household	-0.6509	0.092	-7.041	0.000	-0.832	-0.470
Num_of_Children	0.6752	0.115	5.878	0.000	0.450	0.900
const	4.1578	0.267	15.568	0.000	3.634	4.681

Fig. 1. Generalized Linear Model Regression Results

After training model and tuning parameters, we could get a classification report of our model. As you can see, we have achieved a maximum accuracy of 72 % on test data in Figure 2.

Test - classification report:

	precision	recall	f1-score	support
0	0.61	0.71	0.66	407
1	0.72	0.63	0.67	493
accuracy			0.67	900
macro avg	0.67	0.67	0.67	900
weighted avg	0.67	0.67	0.67	900

Fig. 2. Test - classification report

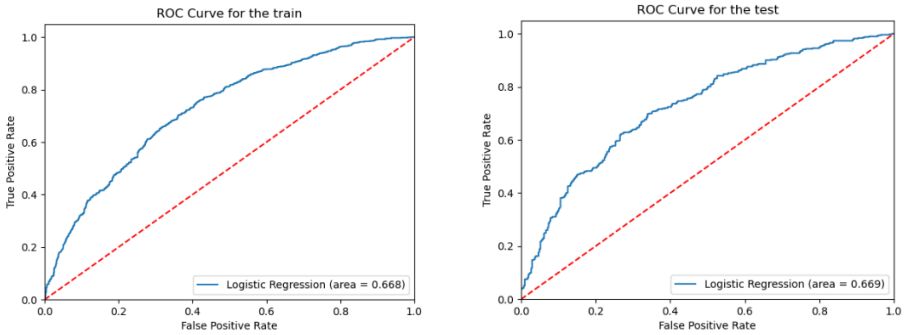

Fig. 3. Receiver operating characteristic curve

We can see that the closer the receiver operating characteristic (ROC) curve is to the top left corner in Figure 3, the more accurate the test is overall. Therefore, the receiver operating characteristic curve for the training data shows that it has 66.8 % accuracy, and for the test, data has 66.9 % accuracy. This result is quite well!

Conclusion

The main objective in this paper was to build credit risk model that would be able to identify potential defaulters and therefore reduce company loss.

References

1. Li Ke, Gubin E.I. Assessment of credit risk by using big data tools // Молодежь и современные информационные технологии: сборник трудов XIX Международной научно-практической конференции студентов, аспирантов и молодых ученых (21-25 марта 2022 г), г. Томск– Томск: Изд-во ТПУ, 2022. – С. 224-225.
2. E.I. Gubin, T. Phana, The most important variables for credit risk model // Молодежь и современные информационные технологии: сборник трудов XVIII Международной научно-практической конференции студентов, аспирантов и молодых ученых (Томск, 22–26 марта 2021 г.) г. Томск – Томск: Изд-во ТПУ, 2021. – С. 349-351.
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.