ChaoFu_DLL_assignment

CHAO FU

17 June, 2022

1 Objective

Predict the probability of default for each customer in a given portfolio.

2 Methods

2.1 Model selection

The dependent variable is discrete (default and not default). The objective is to predict the probability of default. The value of probability is continuous. Hence, the logistic regression model can be applied to solve this problem. The logistic regression equation is as follows:

$$logit(\hat{Y}) = ln(p/1-p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_p X_p = \beta^T X$$

The interpretation of this equation is given below:

1) $\hat{Y}$ is the estimated continuous value matrix$(1 \times N)$.

2) β_0 is the intercept which is regarded as a constant value. $X_1, X_2, \dots X_p$ are p number of explanatory variables. $\beta_1, \beta_2, \dots \beta_p$ are their regression coefficients. β is a regression coefficient matrix$((p+1) \times 1)$ and X is a sample matrix$((p+1) \times N)$ in which N is the number of observations.

3) $p = P(Y_i)$ is the estimated probability of each response variable value occurring.

2.2 Dummy variables

The independent variables should be easy to understand for a bank's customers. The customers can understand the dummy variables ('income:3k-6k' and 'income:6k-9k') better than the independent variable ('income'). Hence, all the independent variables should be transformed into their own dummy variables.

2.3 Fine classing and weight of evidence(WoE)

The number of dummy variables transformed from the independent variables can be significantly large. The large number of variables in a model can cause many serious problems, such as overfitting and unstable performance. Some dummy variables of an independent variable are similar with each other or have few samples. Then, these dummy variables should be combined. The weight of evidence and the number of samples can be used to determine which dummy variables should be combined. The formula of weight of evidence is as follows:

$$WoE_i = \ln\left(\frac{\%NotDefault_i}{\%Default_i}\right)$$

i means the ith dummy variable of an independent variable.

But the model can have a bad performance for the dummy variables trap. Hence, the reference dummy variables should be removed from the model.

2.4 Feature selection

The weight of evidence can not determine whether the independent variable is useful for the model. Hence, the p-values of each dummy variable's coefficient can be used to choose the useful independent variables. If the p-values of most dummy variables from an independent variable are larger than 0.05, this independent variable should be removed from the model. In this assignment, the statsmodels module is used to calculate the p-values.

2.5 Model estimation and probability prediction

In this assignment, the sklearn module is used to train, predict and estimate the model. The F1-score is used to estimate the accuracy of the model on the test data. The false-positive rate means that the model predicts the default customer to non-default. It can cause a high risk for the bank. Hence, a good model should have a low false-positive rate. The receiver operating characteristic curve(ROC) and area under roc curve(AUC) are used to assess the performance of a classification model.

3 Data Description

There are ten independent variables and one dependent variable in this assignment. The independent variables are continuous. The dependent variable is discrete. The train and test data have 120000 and 30000 samples respectively.

4 Data Processing

There are seven steps in the data processing.

1) Solve the doubts about data. In the logistic regression, 1 in the dependent variable means non-default while 0 means default. But the train and test data have the contrary results.
2) Check the balance of the data. The train data can be considered to be balanced.
3) Solve the missing data. Fill the missing data with mode.
4) Solve the outlier. There is an outlier in the test data. Then, it can be removed.
5) Calculate the weight of evidence and fine classing based on the train data. Then the reference dummy variables are removed from the model.
6) The test data has the same dummy variables as the train data.
7) Feature selection. The p-values of dummy variables from 'num_dependents' are all larger than 0.05. Hence, this independent variable should be removed from the model. It means that both train and test data exclude the 'num_dependents' variable.

5 Result

1) The accuracy of the model on the test data is 94%.
2) The false-positive rate on the test data is 5.3%.
3) The AUC value is 0.86.

Conclusion

A well-calibrated and robust model should have a high accuracy while a low false-positive rate on the test data. Moreover, the AUC value should be larger than 0.8. My results indicate that the probability of default (PD) model which is developed in this assignment is well calibrated and robust.