**Prediction via Machine Learning**

With the credit card fraud dataset, I use machine learning to develop a logistic regression model that attempts to predict whether a transaction is a fraud.**Credit card fraud** is when someone uses your credit card or credit account to make a purchase you didn't authorize.

The dataset consists of **3492** observations and  **30 variables**.I use the seed No. of 321 to generate a sequence of random numbers and ensure that I obtain the same result given the same seed No. I use 70% of original data rows randomly as the trainData, and the remaining of random rows as the testData. And then, I make the naïve guess about Fraud probability using the sum of Fraud divided by the total rows of trainData.With this Fraud probaility, I make a random prediction from trainData with sample function and then use mean fuction to get the accuracy of random guess is 0.76

After, I make the logistic regression model to train the data and see which variables of trainData are most relevant for the probability of fraud. Then, I realize V2,V3, V4, V5, V6,V7, V8, V9, V10, V11, V12, V14, V15,V16, V17, V18, V19, V21,V22,V27,V28 are significant variables for prediction with stars, I make a new logistic regression with these important variables.

I use test data to predict the probability values of Credit fraud based on the trained model.

In order to map a logistic regression value to a binary category, I define a **classification threshold** (also called the **decision threshold**). A value above that threshold indicates "Non-Fraud"; a value below indicates "Fraud" If we set the threshold =0.5, the accuracy of our model with test data is 98.09%, while we set the threshold higher or lower than 0.5, the accuracy of model would be lower.

The predictive outcome of two category classification with Threshold of 0.5:

|  |  |  |  |
| --- | --- | --- | --- |
| Threshold=0.5 |  | Actual | |
| Logit.pred |  | Fraud | Non Fraud |
| Prediction | Fraud | 138 (True Positive) | 4(False Positive) |
| Non Fraud | 16(False Negative) | 890(True Negative) |

**Accuracy** =(True P + True N)/total cases =98.09%, and the **recall/sensitivity**=True P/(total pos. transactions) = True P / (True P + False N) =89.61%. **Precision** = True P / (True P + False P) =97.18%

For this case, Sensitivity is cucial for us, we want to decrease the number of False Negative as much as possible. So, if we enhance the threshold to 0.9, the False Negative would be only 12, although the acurracy is 95.61% lower than the model with threshold of 0.5, the sensitivity is raised to 92.20%. But the Precision is decreased to 80.68% from 97.18% dramatically.

|  |  |  |  |
| --- | --- | --- | --- |
| Threshold=0.9 |  | Actual | |
| Logit.pred |  | Fraud | Non Fraud |
| Prediction | Fraud | 142 (True Positive) | 34(False Positive) |
| Non Fraud | 12(False Negative) | 860(True Negative) |

I prefer the prediction model with threshold of 0.5.Although I want to identify the people with Fraud as many as possible, I don’t want to see nearly 20% of people with non-Fraud behavior, are treated as Fraud.

For **the confusion matrix** with threshold of 0.5, it directly visualize the performance of an algorithm:

Reference

Prediction Fraud NonFraud

Fraud 138 4

NonFraud 16 890

* Accuracy is 0.9809, showing it is a good model (Explained above)
* sensitivity /recall is 0.8961. Just like what I explained above, it is a critical matrix in this case, the higher the better (Explained above)
* P-value of 2e-16 (<0.05, Significant effect) P Value is a probability score that is used in statistical tests to establish the statistical significance of an observed effect
* Specificity of 0.9955, which is calculated as the number of correct negative predictions divided by the total number of negatives. It is also called true negative rate (TNR). The result indicates a good model.

It is helpful to know that the**F1 Score**is a measure of how accurate a model is by using Precision and Recall following the formula of: F1\_Score = 2 \* ((Precision \* Recall) / (Precision + Recall)),and Precision =True P/(total pos. tests) =TP/ (TP + FP). The result is 0.9324, which shows the model is very good.

In addition, I Plot **ROC curve** using ROCR package. We want to see true positive rate is close to 1 as much as possible and false positive rate is close to 0. The result shows the line is very close to 1. So, ROC curve indicates it is a good model.

Last but not least, I compute **AUC (Area under curve**), we want to the area under ROC curve is as much as possible. The result is 0.9837, which shows the model is pretty good for prediction of Fraud.

In conclusion, All of the matrix show that this logistic regression model I build is a good model.