Model Evaluation Report

Credit Card Fraud Detection

Name: Maviya Shaikh

Student number: 100766785

Date: December 18th,2020

Table of Contents

1. Introduction

In this digital era where physical cash flow is reduced and we are wrapped around the global pandemic of covid19 that pushes us to use only electronic methods for transaction, there is always a risk of misuse of this method that leads to higher losses to the user. In response to deal with this, the project aims to develop a system to detect fraudulent credit card transaction to prevent such cases by analysing their historical pattern. Specifically, this project is tailored to identify the credit card frauds as use of credit card has increased now-a-days because it provides instantaneous money without having actual money in the pocket.

2. Problem Statement

The credit card companies should be able to recognize fraudulent credit card transactions so that users are not charged for items that they did not purchase. For this, "**The analysis will be conducted to identify whether transaction occurred is fraudulent or not**".

3. Key Questions

This report will provide answer to the below major questions of the project:

1. **To identify fraudulent transaction**

   The solution develop should be answerable to detect all the fraud cases as misclassification of this cases is not tolerated at any cost because the customers money is at stake

2. **To identify normal transaction**

   It is also important to detect normal payment otherwise customer has to go through all the policies of financial institution that might block their card considering that it is fraud transaction

4. Data Sources

The dataset used for the analysis of this project is downloaded from Kaggle which is well known website in data science community to explore the open source datasets. Originally, the dataset has been collected and analysed during a research collaboration of Worldline and the Machine Learning Group of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

5. Data Requirements

The datasets consist of 284,807 transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. There are 31 variables in the dataset which are explained below:

- Due to privacy reason, variables **‘V1, V2, … V28’** are transformed with Principal Component Analysis (PCA)
- The variable '**Time'** contains the seconds elapsed between each transaction and the first transaction in the dataset.
- The variable '**Amount**' is the transaction amount
- The dependent variable '**Class**' is the response variable that has value 1 in case of fraud and 0 otherwise.

6. Feature Engineering

Feature Engineering is an important step to validate that all the features used are important in the prediction of the model and it allows us to add or remove the features according to the business problem requirements. Here, two variables Time and Amount are not scaled whereas all other variables are scaled and transformed using Principal Component Analysis. Thus, we will scale the variables using Standard Scaler which is the technique to rescales the value of the data points in the range of mean 0 and standard deviation 1.

Time → Scaled Time
Amount → Scaled Amount

Furthermore, the original time and amount is dropped from the dataset and scaled time and scaled amount is added in the dataset. Thus, our dataset is ready for model prediction.

7. Model Analysis

For the model analysis, training and test size of the model is decided as 80% to train the model and 20% to test the model. In the prediction of the model, Class 0 represent *‘No Fraud’* and Class 1 represent *‘Fraud’.*

Models

Three models are used to detect the fraud and no fraud cases which are as follows:

- **Logistic Regression**: It is used to predict the categorical dependent variable using a given set of independent variables
- **Decision Tree Classifier:** The general idea of Decision Tree is to create a training model which is used to predict class or value of target variables by learning decision rules inferred from training data
- **Random Forest Classifier:** It creates a set of decision trees from randomly selected subset of training set.

Analysis

The models are run in three development stages to find the accurate model. The confusion matrix and classification report of all the models are plotted in the notebook. Here for the simplicity, we will use confusion matrix to compare models.

The Confusion Matrix will be as shown for no fraud and fraud cases. True positive and True negative are correctly classified no fraud and fraud cases. Type I error is misclassification of no fraud cases whereas Type II error is misclassification of fraud cases. **So, our first priority is to reduce the Type II error i.e. reduce the misclassification of the fraud cases.**

		Predicted Values	
		No Fraud	**Fraud**
Actual Values	**No Fraud**	True Positive (correct)	Type I error (misclassify no fraud cases)
	Fraud	Type II error (misclassify fraud cases)	True Negative (correct)

Table 1: Confusion Matrix

Three Development stages are:

1) Models built on the original dataset

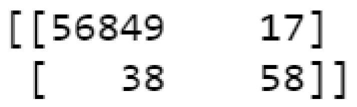
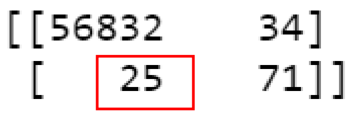
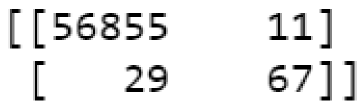
Logistic Regression	Decision Tree	Random Forest
 [[56849 17] [38 58]]	 [[56832 34] [25 71]]	 [[56855 11] [29 67]]

Table 2: Result of first development stage

Above are the results of the models, as our first priority is to reduce the misclassification of the fraud cases, decision tree had misclassified least fraud cases compare to other models.

2) Models built after performing Synthetic Minority Oversampling Technique (SMOTE)

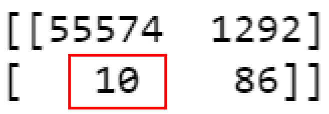
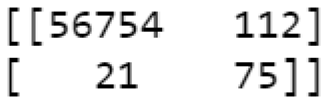
Logistic Regression	Decision Tree
 [[55574 1292] [10 86]]	 [[56754 112] [21 75]]

Table 3: Result of second development stage

SMOTE technique is used to balance the class in which synthetic samples are created for minority class while training the model so that model will not give biased result. As we have 2,84,807 data points in our dataset, it will be costly to run the random forest algorithm with SMOTE technique because random forest will run multiple decision trees that will increase the time complexity.

Logistic regression performs better, and it misclassifies only 10 fraud cases that is quite good performance than above models.

3) Models built after performing Random Under-sampling

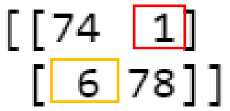
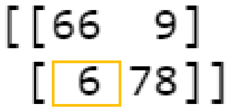
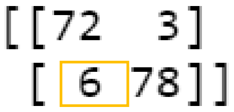
Logistic Regression	Decision Tree	Random Forest
 [[74 1] [6 78]]	 [[66 9] [6 78]]	 [[72 3] [6 78]]

Table 4: Result of third development stage

In random under-sampling, the majority class is down sampled to match the length of minority class. The one drawback of this method is that there will be information loss. However, it is performing better than previous models, all the models misclassify only 6 fraud cases. So, to select the best model type I error is checked and it was found that logistic regression misclassifies only 1 no fraud. Therefore, logistic regression is selected as best model in this stage.

8. Conclusion

The solution developed provides the answers to two key questions of the project i.e. detecting fraud and no fraud cases. Moreover, Logistic Regression is chosen as the best model for accurately detecting fraud and no fraud cases in both the cases with SMOTE technique and Random Under-sampling. I would recommend everyone to use the SMOTE technique because there will not be an information loss such as in random under-sampling. However, if there is time constraint then you can use random under-sampling as it is quick and accurate.

References

1. Machine Learning Group -ULB. (2018, March 23). Credit Card Fraud Detection. Retrieved October 30, 2020, from https://www.kaggle.com/mlg-ulb/creditcardfraud
2. Narkhede, S. (2019, May 26). Understanding AUC - ROC Curve. Retrieved October 30, 2020, from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5
3. Jaadi, Z. (n.d.). A Step by Step Explanation of Principal Component Analysis. Retrieved October 31, 2020 from https://builtin.com/data-science/step-step-explanation-principal-component-analysis