Credit Card Fraud Detection

Jay Shukla
College Of Engineering
Northeastern University
Toronto, ON
Shukla.j@northeastern.edu

A: Introduction

Credit Card Frauds in banking (2013) explores the credit card fraud and methods of it, and gives information about what to do in case of encountering credit card fraud by chargeback topic. In this paper it is studied on the types of credit card fraud such as application fraud, lost stolen cards, account takeover, fake and counterfeit cards. Also it includes part of gaining information by taking reports and data from different and safe official sources. Besides that, paper investigated about how often occurrence of these methods.

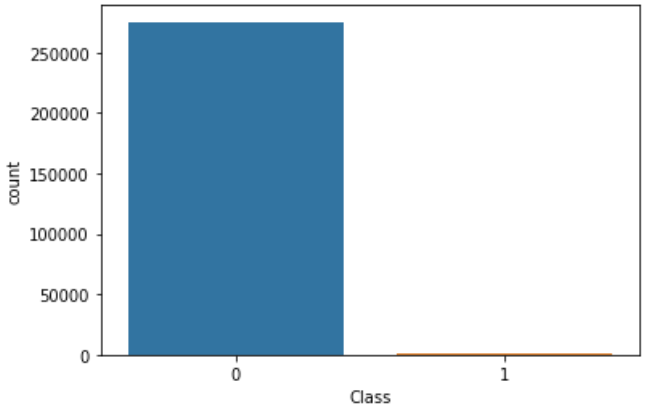
B: Datasets

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

The datasets I have used for training in this project are from Kaggle. The dataset from Kaggle has 31 features, 28 of which have been anonymised are labelled V1 through V28. The remaining three features are the time and amount of transaction as well as a label whether that transaction was fraudulent or not. The variables have been anonymised (as these are actual European card holder transactions) in the form of a PCA. The dataset contains 284,807 transactions. The mean value of transactions is 88.35 USD and the largest transaction value is 25,691.16 USD. Most of the transactions are quite small as you would expect in everyday transactions. The time is recorded in number of seconds elapsed since the first transaction in the dataset. The transactions are over a period of 2 days. 99.83% of transactions in this dataset were not fraudulent while only 0.17% were fraudulent. There is also minimal correlation between variables – This may be as a result of PCA transformed variables. Hence I don't need to account for any multi collinearity in my model.

C: Implementation

As the data is so large first of all I have decided to display top 5 rows of the dataset and last 5 rows of the dataset. So that whole data can be verified. After that I will find the shape of our dataset (number of Rows and columns). Now I need to get information about our dataset like total number of rows, total number of columns, Datatypes of each column and memory requirement. We need to check the null values in the dataset which includes feature scaling, removing duplicated values. Which includes not handling imbalanced values.



D: Methods and Software Requirements

The methods includes splitting datasets into the training set. As the unbalanced data was shown above we have to handle the unbalanced dataset with undersampling, Logistic Regression, Decision Tree Classifier and Random Forest Classifier etc. I used Windows 10 as my operating system and used jupyter Notebook as my IDE. Python as my programming language. Anaconda as my framework and numpy, matplotlib, pandas, scikit-learn as my libraries.

E: References

1.S. Maes, K. Tuyls, B. Vanschoenwinkel, and B. Manderick, "Credit Card Fraud Detection Using Bayesian and Neural Networks," in Proceedings of the 1st International Naiso Congress on Neuro Fuzzy Technologies, pp. 261–270, 2002.

2. K. Fu, D. Cheng, Y. Tu, and L. Zhang, "Credit Card Fraud Detection Using Convolutional Neural Networks," in Neural Information Processing, vol. 9949 of Lecture Notes in Computer Science, pp. 483–490, Springer International Publishing, 2016.

3. S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: a comparative study," Decision Support Systems, vol. 50, no. 3, pp. 602– 613, 2011.

4. N. Carneiro, G. Figueira, and M. Costa, "A data mining based system for credit-card fraud detection in e-tail," Decision Support Systems, vol. 95, pp. 91–101, 2017.

5. W. Yin, K. Kann, M. Yu, and H. Schtze, "Comparative Study of Cnn and Rnn for Natural Language Processing," 2017, https://arxiv.org/abs/1702.01923.