IBM Supervised Learning: Classification Project – Credit Card Fraud Detection

This study looks into using machine learning classification techniques in order to detect credit card fraud. Techniques such as logistic regressions, support vector machines, decision trees and random forests are explored. Due to the heavy imbalance of the dataset, randomised undersampling is used.

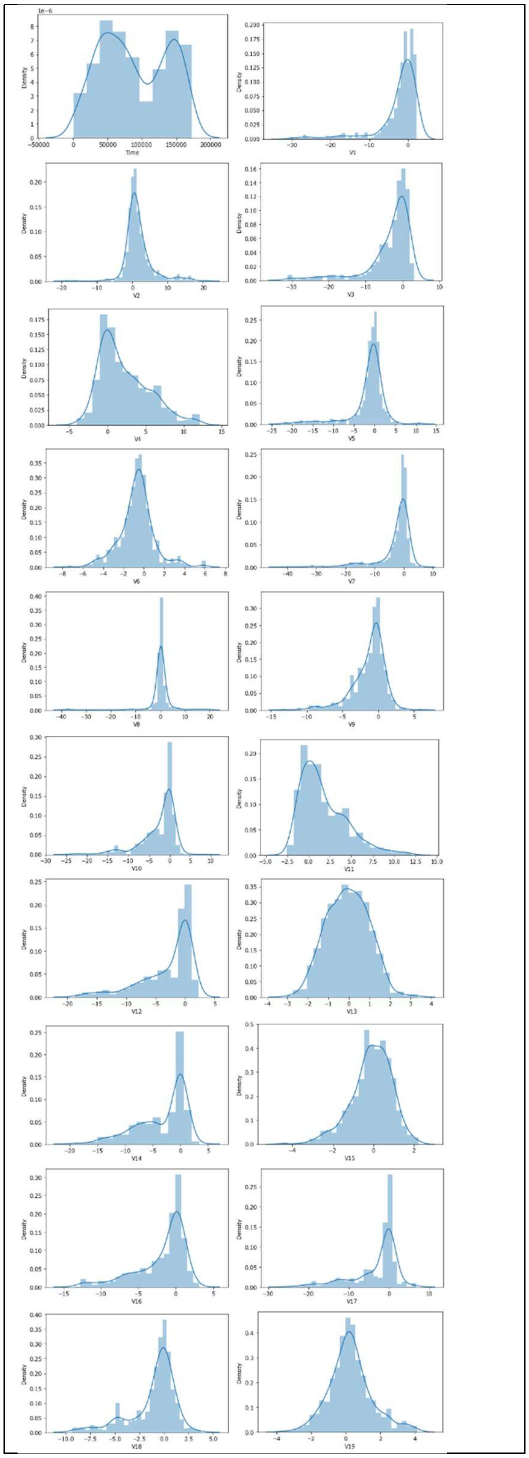
1 Introduction

Data was taken from the Kaggle Credit Card Fraud Detection Dataset available at https://www.kaggle.com/mlg-ulb/creditcardfraud. The aim is to be able to classify transactions as fraudulent on not fraudulent. Patterns in some of these transactions may not particularly obvious to the human eye and many may go missed – not to mention the difficulty and time one must spend to do it by hand. Alternatively, rules set by humans are sometimes used to identify fraudulent transactions but is there are way to be even more accurate? Machine learning can be used to improve up this process and here we demonstrate some of the common classification algorithms used.

2 Data Exploration and Cleaning

The dataset is a record of fraudulent and non-fraudulent transactions. It has 30 features and a target variable, 'Class', that specifies whether a transaction was fraudulent (1) or not (0). There are 28 features named V1, V2… V28 which are the principal components obtained from a PCA transformation in order to protect anonymity and maintain privacy. However, the 'Time' and 'Amount' features have not been transformed.

The dataset contained had no missing or N/A values. In order to better understand the data, the distribution of the variables was plotted as shown in Figure 1.



2.1 Randomized Undersampling

Upon analysis of the target variable, it became very clear that the dataset was heavily imbalanced as shown in Figure 2.

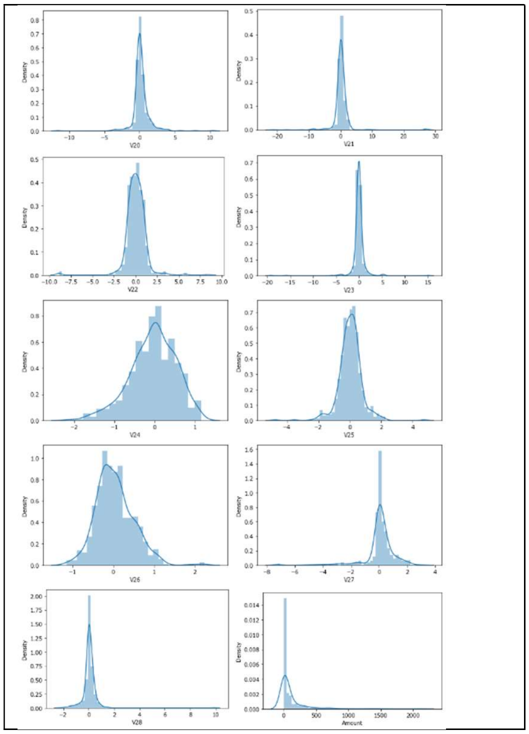


Figure 1: Feature distributions

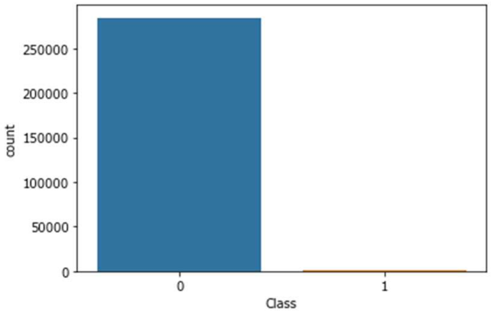


Figure 2: Imbalanced dataset

In order to balance this dataset, randomised undersampling was used. This was done by shuffling the dataset and then removing the required number of rows from the majority class until the number of fraudulent and non-fraudulent transactions was equal. The result is shown in Figure 3.

The splitting of the train and test set was done prior to this. It was also important to note this imbalance later when deciding on a metric to measure the effectiveness of our models by

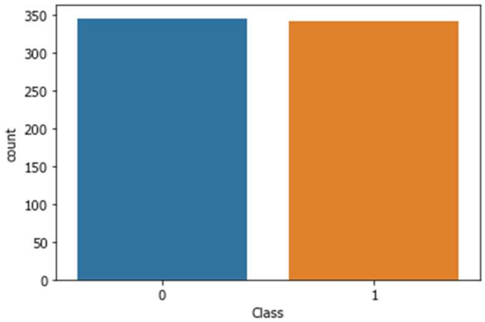


Figure 3: Balanced dataset

Additionally, to explore the dataset further, a correlation matrix was plotted as shown in Figure 4.

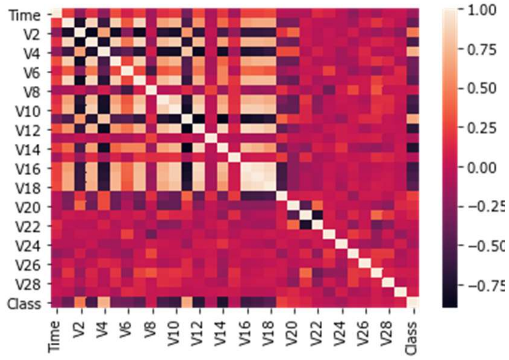


Figure 4: Correlation matrix of balanced dataset

Data was also scaled for the support vector machine model.

4 Classification Models

Four classification models were compared after being trained on our dataset: a logistic regression, a support vector machine, a decision tree, and a random forest.

4.1 Logistic Regression

Logistic regression utilises the sigmoid function as shown in Figure 5.

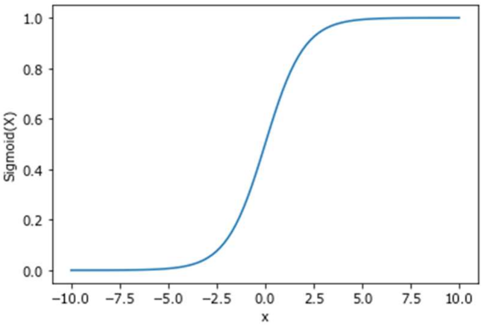


Figure 5: The sigmoid function

4.2 Support Vector Machines

Support Vector Machines classify data by creating a hyperplane that represents the largest margin between two classes. By utilising the kernel trick and moving to higher dimensions, SVMs can be used to classify non-linear data.

A pipeline was used with a standard scaler in order to scale the data

4.3 Decision Tree

A decision tree is made up of nodes. At each node, the data is split into two subsets which then move to the next node and so on till a leaf node is reached. The maximum number of nodes that data passes is known as the depth of the decision tree.

4.4 Random Forest

A random forest is an extension of bagging (bootstrap aggregating) where there are multiple decision trees that each use a random subset of features of the data in each sample.

5 Key Findings

The effectiveness of the models can be determined from the area under the ROC curve as shown in Figure 6 – cross-entropy loss is not suitable in this case due to the highly imbalanced dataset.

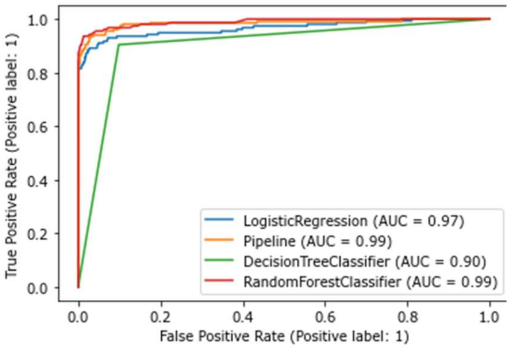


Figure 6: ROC curve for the models; note the pipeline is that of a standard scaler and a support vector machine

Logistic regression was used as a baseline to measure the performance of our models. It is clear that the decision tree classifier is the worst model. This is expected as they are often relatively inaccurate and unstable. In contrast the random forest classifier is much better as it averages over many decision trees. The support vector machine also appears to be a better model than logistic regression for this dataset.

6 Possible Flaws

Logistic regression constructs linear boundaries and assumes linearity between the target variable and the features. As mentioned before, decision trees are often inaccurate. Random Forests are not particularly interpretable and often act as a black box similarly to neural networks. Support vector machines are not suitable for large datasets as the kernel trick requires lots more computation. It also performs poorly when there is a lot of noise causing classes to “overlap”.

7 Next Steps

As an alternative to the randomized under sampling techniques we used, one could use oversampling techniques such as SMOTE in the hope that this would improve the accuracy of the models. It would also be beneficial to expand the dataset to include more instances where fraud has taken place. One could also consider using ensemble methods, combining these models. Additionally it might be worth considering the use of a neural network or clustering techniques to aid classification.

References

Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson and Gianluca Bontempi. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

Dal Pozzolo, Andrea; Caelen, Olivier; Le Borgne, Yann-Ael; Waterschoot, Serge; Bontempi, Gianluca. Learned lessons in credit card fraud detection from a practitioner perspective, Expert systems with applications,41,10,4915-4928,2014, Pergamon

Dal Pozzolo, Andrea; Boracchi, Giacomo; Caelen, Olivier; Alippi, Cesare; Bontempi, Gianluca. Credit card fraud detection: a realistic modeling and a novel learning strategy, IEEE transactions on neural networks and learning systems,29,8,3784-3797,2018,IEEE

Dal Pozzolo, Andrea Adaptive Machine learning for credit card fraud detection ULB MLG PhD thesis (supervised by G. Bontempi)

Carcillo, Fabrizio; Dal Pozzolo, Andrea; Le Borgne, Yann-Aël; Caelen, Olivier; Mazzer, Yannis; Bontempi, Gianluca. Scarff: a scalable framework for streaming credit card fraud detection with Spark, Information fusion,41, 182-194,2018,Elsevier

Carcillo, Fabrizio; Le Borgne, Yann-Aël; Caelen, Olivier; Bontempi, Gianluca. Streaming active learning strategies for real-life credit card fraud detection: assessment and visualization, International Journal of Data Science and Analytics, 5,4,285-300,2018,Springer International Publishing

Bertrand Lebichot, Yann-Aël Le Borgne, Liyun He, Frederic Oblé, Gianluca Bontempi Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection, INNSBDDL 2019: Recent Advances in Big Data and Deep Learning, pp 78-88, 2019

Fabrizio Carcillo, Yann-Aël Le Borgne, Olivier Caelen, Frederic Oblé, Gianluca Bontempi Combining Unsupervised and Supervised Learning in Credit Card Fraud Detection Information Sciences, 2019