Vol: 1 Issue: 1

International Journal of Data Science and Advanced Analytics

Predicting Credit Card Fraud on a Imbalanced Data

Soh Wei Wen[1], Rika Mohd Yusuf [2]
[1, 2] School of Computing,
[1, 2] Asia Pacific University of Technology & Innovation,
Kuala Lumpur, Malaysia
[1]swwen5148@gmail.com, [2]ika.rikayusuf@gmail.com

Received: 05-Dec-18; Accepted: 25-Feb-19; Published: 21-Apr-19
Corresponding author email: swwen5148@gmail.com

***Abstract*— Credit card fraud is increasing considerably with the development of modern technology and the global superhighways of communication. Credit card fraudsters continuously try to come out with a new tactic challenged the present technology and system. It cost both, providers and consumers a lot of money. Thus, quick and accurate model become essential for companies and credit card providers, to decrease their financial and customer trust losses. However, there is a lack of published literature on credit card fraud detection techniques, due to the unlabeled credit card transactions dataset for researchers. High dimensional data refer to data that have multiple variables. The dataset consist of the credit card details, amount transaction, location, time, and personal details of the cardholders that are anonymized. Thus, in this study, a real-world dataset (European Credit card), with PCA transformation applied is being used. The common problem happened in this kind of research is the data tend to be imbalanced. Imbalanced data will often introduce bias which the accuracy of the prediction is not accurate. In this study, the dataset has been train with an oversampling pre-processing technique called SAS Sample and various data mining technique such as Random Forest, KNN, Decision Tree, and Logistic Regression. After several trials, we found out that the regression technique has the best performance among the others.**

***Keywords*— Credit card fraud; PCA; Data Mining; Imbalanced data; Classification**

1. Introduction

Credit card is considered as the most common type of fraud since it takes very less time for an attacker to steal the information and most of the time, the fraud is discovered few days after it happened. Credit card fraud is a significant issue and has taken a considerable cost for banks and card issuer companies. This issue has been taken seriously which a highly sophisticated security system assigned to monitor the transaction fraud quickly when it happened. There are various methods used for fraud detection, each of them tries to increase the detection rate while keeping the false alarm rate at a minimum. Different method and approach have been applied to different data such as Bayesian Algorithm, K-Nearest Neighbor. Statistical fraud detection methods have been divided into two broad categories, supervised and unsupervised. The approach is vary depending on the problem and dataset. A relevant data set has been experimented and tested using Random forest, KNN, Regression, and Decision Tree in this paper. In the 3.0 section, the previous work with common solution have been discussed. In order to proceed with the high dimensional data with several attributes such as card holder's personal details, transaction amount, time, date, location, and credit card details are needed. However, due to confidential regulation, further details can't be provided. Thus, real-world data of European credit card that has been transformed using PCA transformation has been used. Unfortunately, the most common problem faced by the researcher is the dataset tend to be unbalanced. Imbalanced data refers to a dataset that is not balanced which mean the ratio between the classes is not the same. Imbalanced data will often introduce bias which the accuracy of the prediction is not accurate. Thus, an oversampling method is applied before the experiments. The aim of this study is to develop a supervised learning algorithm to detect credit card fraud on oversampling.

This paper is organized as follows. In section 1.0, we have presented the introduction of the problem along with the most common solution. In section 2.0, we provide the related work previously done for credit card fraud detection by data mining technique with oversampling. In section 3.0, we have stated the research methodology or details of the experimental setup and details explanation of algorithm purposed. Section 4.0 deals with details of analysis done from the background of the dataset, tools, and comparative finding of all the techniques on the basis of the result obtained from the experiment done. In section 5.0, the conclusion of the study described.
.

2. Related Works

Due to the advance of technology, credit card fraud has become a threat to the financial company. Credit card fraud refers to the misuse of information or physical credit card by another person without the owner's acknowledgment [14]. There are few types of frauds that have been discussed by [8][13][3] which are credit card fraud, telecommunication fraud, computer intrusion, bankruptcy fraud, theft fraud, application fraud, and behavioral fraud. In this study, we only focus on behavioral fraud because the data set is about the transaction which indicates the behavior of using a credit card. The main challenge that comes with credit card data set is highly unbalanced data which has been mentioned by [11][25][16][3]. However, there are few sampling techniques that can be used to process the unbalanced data. These techniques are undersampling, oversampling, SOMTE, roughly balancing (RB) and random balance [11]. Undersampling works with decreasing the size of the majority class whereas oversampling works with increasing the size of the minority class [11][25][28]. Besides that, SMOTE is a technique use to create new synthetic examples of the minority class [11][25][9]. Roughly balancing refers to the class is balancing in determining the sampling probability of each class [11][21]. RNB randomly selects a class scale and uses SMOTE to oversample one of the classes and undersample the other [11]. Undersampling has been used in most of the paper [3] because undersampling has been proved to perform better than oversampling. In this study, we will use oversampling just to measure how well the model can perform.

In [11][24], they trying to solve imbalanced data by using bagging ensemble based on threshold-moving, while [36] trying to solve it by introducing DeepBalance approach where deep belief networks tanned with balanced bootstraps and random features selection. Imbalanced data refers to a dataset that is not balanced which mean the ratio between the classes is not the same. Imbalanced data will often introduce bias which the accuracy of the prediction is not accurate. Besides that, how undersampling can induce a bias in the posterior probabilities generated by machine learning methods has been discussed by [25]. Other than that, [38] had discussed how to detect credit card fraud using data mining technique. [12] had mentioned the problem of estimating the probability for a skier to suffer injuries while skiing. Other than that, [27][7] stated that the common issue they are trying to approach is looking for the best technic to detect the fraud efficiently due to the increase of fraud activities reported. This issue supported by [2] discussed that most of the time the fraud can only be detected after it happened. The late detection of fraud cost both side, companies and users a lot of money. The late detection of fraud cost both side, companies and users a lot of money. Even though there is a lot of systems introduced, but the fraudsters constantly trying to create a new tactic to do the illegal action. The users tend to lose trust towards the providers and it is not good for the business. Lastly, the data slicing problem had been discussed [10].

3. Methodology

KDD refers to Knowledge Discovery in Database, it is a type of methodology that allows transforming raw data to some useful knowledge/ pattern. Phases of KDD are data selection, data pre-processing, data transformation, data mining and interpretation evaluation. The stages of KDD is shown in Figure 1.

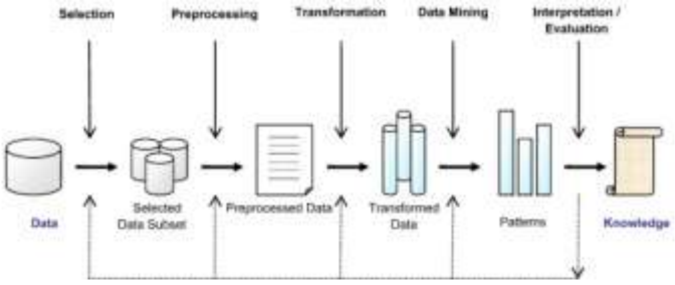

Figure 1: Stages in KDD

3.1 Data Selection

Dataset was obtained from [25]. This dataset consists of 100,000 instances with 30 attributes which describe the characteristics of fraud detection. The dataset consists of numerical and binary variables, most of the variables are anonymized.

3.2 Data Pre-processing and Data Transformation

Data pre-processing also known as data cleaning. It is a process of cleaning dirty data, outlier, missing value and skewness. In this dataset, there's no missing value, the standard deviation of most of the variables is larger than mean, and some of the variables are quite skewed. Thus, there is no data cleaning process being done. The visualization takes place using SAS stat explore to see the skewness and to make it easier to understand the data. The data is very unbalanced. Transformation of data will not perform in this dataset because PCA was implemented for this dataset. PCA stands for principal component analysis. This is an unsupervised linear feature extraction technique, the purpose of PCA is to reduce the size of data by taking only the features that have most information [4]. PCA use features, in the dataset to create new features, the new features also known as principal components. The principal components are then using to build the model. Besides, [4] stated that variables that are more or equal to 1.221 standard deviations will retain.

Data visualization is done before the data mining process and at the comparison stage. Data visualization or visual analytics is performed to gain insights into the data structure. The data is visualized using charts and graphs that automatically generated by SAS Enterprise Miner to show the distribution of the data. This process will help the researchers to have more detail understanding of the data and the next suitable prediction model which will be used is selected. According to [4], an interactive dashboard is very helpful to business users to understand their data. The dashboard will comprise of several graphs relating to the data set on the screen. As such, trends and patterns in the data set can be studied while showing the relationships between different attributes. In short, a summary of the data can be seen in one view.

3.3 Data Mining

Data mining is a process of looking for a pattern or useful knowledge from the data. Data mining has been applied to detect fraud because of its effectiveness. The prediction model is trained by the historical data and will be used to detect every unknown data, to classify whether it is a fraud. With this prediction model, company, bank, or credit card provider can react faster and provide the best support to their customers. Data mining apply in this dataset is to know what are the characteristics that has the most worth to affect the result. Other than that, data mining can provide trends or characteristics when there's a fraud. The algorithms that will be using are K-Nearest Neighbor (KNN), decision tree, random forest, and logistic regression. These are belonging to the classification category.

3.3.1 Random Forest

Random forest is one of the well-known ensemble classifiers. Ensemble classifier refers to an algorithm consists of multiple classifiers instead of one classifier [1]. The random forest consists of a set of decision trees and it classifies the unknown data by majority vote [22]. [6] introduced a random forest which adds randomness to bagging. Bagging is a method that uses to create a different version of classifiers and come out with an aggregate predictor [5]. In random forest, each node is split by choosing the best in a subset of predictors chosen randomly instead of splitting the nodes by choosing the best split among all the variables [1][22]. [35] stated that "The prediction performance of RF compares well to other classification algorithms such as support vector machines (SVMs), artificial neural networks, Bayesian classifiers, logistic regression, k-nearest-neighbours, discriminant analysis such as Fisher's linear discriminant analysis and regularized discriminant analysis, partial least squares (PLS) and decision trees such as classification and regression trees (CARTs)." The algorithm of random forest is stated below: - [37]

Step 1: In a training set, a new sample set is extracted randomly by repeating N times. [6] introduced the margin function is as below: -

$$mg(\mathbf{X}, Y) = av_k I(h_k(\mathbf{X}) = Y) - \max_{j \neq Y} av_k I(h_k(\mathbf{X}) = j).$$

where *I (·)* is the indicator function. *hk is* the ensemble of classifier (*h*1(**x**), *h*2(**x**), . . . , *hK* (**x**), X,Y is the vector where the training set drawn at random from. The margin measures the extent to which the average number of votes at **X**, *Y* for the right class exceeds the average vote for any other class. The larger the margin, the more confidence in the classification.

Step 2: Decision trees are build based on a sample set from step 1.

Step 3: Let every tree in the forest to vote for x.
Step 4: The majority vote is the classification for x.

3.3.2 K Nearest Neighbors (KNN)

K nearest neighbors (KNN) is a sample-based learning algorithm [17]. It measures the distance between training data and the unknown data and classifies according to the similarity score. The model of KNN will not be built during the training phase, only training tuples with class label are stored and the model will be built during the classification phase [34]. The measurement that use to measure the distance between data is Euclidean distance. The formula of Euclidean distance is stated below: -

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Only Euclidean distance calculation will be used to calculate the distance because SAS Enterprise Miner only has one calculation method.
The algorithm of KNN is stated below: -
Step 1: Given an unknown data x, calculate the distance of training data and x.
Step 2: X is classified based on the majority vote of the nearest neighbors.

3.3.3 Decision Tree

Decision Tree is one of the most common and popular algorithms for classification and prediction, due to the easy to understand hierarchy. However, irrelevant attributes may affect badly the construction of a decision tree. According to [15] decision tree methodology is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable. This method classifies a population into branch-like segments that construct an inverted tree with a root node, internal nodes, and leaf nodes. The algorithm is non-parametric and can efficiently deal with large, complicated datasets without imposing a complicated parametric structure. When the sample size is large enough, study data can be divided into training and validation datasets. Using the training dataset to build a decision tree model and a validation dataset to decide on the appropriate tree size needed to achieve the optimal final model. There are few frequently used algorithms, used to develop decision trees including CART, C4.5, CHAID, and QUEST. SPSS and SAS programs that can be used to visualize tree structure.

Illusion how the attribute selection measures:
Step 1: Data divided into training and validation data
Step 2: Entropy is calculated

$$Entropy(S)= \sum_{i=1}^{C} p_i \log_2 p_i$$

Step 3: Then information gain is calculated to select the most useful attribute for classification

$$Gain(D,A)=Entropy(D)-\sum_{j=1}^{v}\frac{|Dj|}{|D|}Entropy(D_j)$$

Where,
D: A given data partition
A: Attribute
V: Suppose we partition the tuples in D on some attribute A
having v distinct values
D is split into v partition or subsets, {D1, D2,Dj} were Dj
contains those tuples in D that have outcome aj of A.
The attribute that has the highest information gain is chosen.

3.3.4 Regression

The regression model has many applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, biomedical and drug responses modeling, and environmental modeling [31]. The process of training a regression model involves finding the best parameter values for the function that minimize a measure of the error, for example, the sum of squared errors. Here is the equation involved.

$$y = F(x, \theta) + e$$

It shows that regression is the process of estimating the value of a continuous target (y) as a function (F) of one or more predictors (x1, x2, ..., xn), a set of parameters (θ1, θ2 , ..., θn), and a measure of error (e).

Multivariate Linear Regression

$$y = \theta 1+\theta 2\ x1+ \theta 3\ x2 +.... + \theta n\ xn\text{-}1+ e$$

3.4 Interpretation Evaluation

Performance matrix is used to measure the performance of a model. Different kind of algorithm has different performance matrix. There is two types of the prediction model, numerical and categorical. For the numerical model, the performance is measured by the mean square error whereas for categorical model the performance is measured by the misclassification rate which is also known as error rate. The outcome to achieve is to find out a suitable model to detect fraud with less than 20% error rate.

4. Analysis and Discussion

SAS Enterprise Miner was used throughout the study progression. SAS Enterprise Miner is an analytical tool provided by SAS Institute. SAS data mining process follow SEMMA steps: Sample, Explore, Modify, Model and Assess. SAS enterprise miner provides a graphical user interface (GUI) which make the process of the building model and testing model more efficient and easier. This is because SAS Enterprise Miner provides nodes which user can drag and drop to create a flow for the process. User can modify the properties for each of the nodes with the properties window available on the left [30].

4.1 Dataset Description

Dataset was obtained from [25], consists of 100,000 instances with 30 attributes which describe the characteristics of fraud detection. The datasets contain transactions made by credit cards in September 2013 by European cardholders. The dataset presents transactions that occurred in two days. The dataset consists of numerical and binary variables, most of the variables are anonymized. The dataset is highly unbalanced as the negative class hold 99777 records whereas positive class hold 223 records. Due to confidentiality issues, the original features of the data can't be provided. The only features that are not been transformed with PCA are 'Time' and 'Amount'. According to [23], Attribute 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset, while the attribute 'Amount' represents the transaction amount. Attribute 'Class' is the response variable and it takes value 1 in case of fraud while 0 otherwise.

First, we will import the data set into enterprise miner, and set the class as a target. Then we use statexplore to go through the statistics of the data.

The data set is highly imbalanced. Thus, the data set has been balanced by oversampling. There are no missing values in the data set. Although most of the standard deviation is higher than the mean, no transformation will be done because of PCA is already implemented in this dataset. According to [4], variables with a standard deviation of 1.221 or more will be kept. Finally, our dataset with 984 instances is ready for the data mining process.

4.2 Data Mining

4.2.1 Random Forest

Several random forest models with different properties has been compared. The model properties vary with the number of trees which are 10,20,30,40 and 50. The misclassification rate of the models are 0.086294, 0.076142, 0.071066, 0.086294 and 0.086294 respectively. As the result shows that all random forests are having validation error rate higher than training error rate, this indicates that all the random forests have been overfitted. Thus, no model will be chosen.

4.2.2 KNN

Five KNN model with a different number of k (6,7,8,9,10) has been created and compared. The properties of the model vary with the number of neighbors. The accuracy for number of k: 6,7,8,9,10 are 0.40404, 0.41077, 0.39394, 0.40741 and 0.40067 respectively. The result shows that all the models are over fitted. Thus, no model will be chosen.

4.2.3 Logistic Regression

Several Logistic Regression models with different properties has been compared. The model properties vary with the data partition (70:30, 60:40, 80:20) and selection model (stepwise, forward, backward). The misclassification rate shows that Logistic regression performance is better among others with 0.6% error rate.

4.2.4 Decision Tree

Five Decision Tree with different splitting rules approach has been created and compared. The property of data partition and splitting rules different from one and another. The lowest misclassification rate is chosen in which model decision tree (variance, Entropy, Entropy).

The comparison between the chosen Regression model and the Decision Tree model. The result shows that Logistic Regression with stepwise splitting rules has outperformed the Decision Tree with only 0.6% error rate.

5. Conclusion

Credit card fraud is increasing considerably with the development of modern technology and the global superhighways of communication. In this study, we proposed 4 models to detect credit card fraud. First, we go through the KDD process which is data selection, data preprocessing, we never transform the data because PCA has been implemented. After the dataset is balanced by using the oversampling method. We move on to the data mining and evaluation stages. We build a model by using a different algorithm which are KNN, decision tree, random forest, and logistic regression. Different properties of certain model have been compared to get the best performance. The result shows that the random forest and KNN are overfitting. Thus, only the decision tree and logistic regression have been compared. The best performing model is a logistic regression. In the future, we would like to use different sampling technique such as undersampling, SMOTE or roughly balancing to compare the result.

References

[1] Akar, Ö. & Güngör, O. (2012). Classification of multispectral images using Random Forest algorithm. *Journal of Geodesy and Geoinformation*.

[2] Aravindh, Venkatesan & Kumaravel, Technology, I. & Nadu, T. (2012). Online Credit Card Fraudulent Detection Using Data Mining. p.pp. 1–7.

[3] Bhattacharyya, S., Jha, S., Tharakunnel, K. & Westland, J.C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*. [Online]. 50 (3). p.pp. 602–613. Available from: http://dx.doi.org/10.1016/j.dss.2010.08.008.

[4] Boodhun, N. & Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems*. 4 (2). p.pp. 145–154. Available from: http://link.springer.com/10.1007/s40747-018-0072-1.

[5] Breiman, L. (1996). Bagging predictors. *Machine Learning*. 24 (2). p.pp. 123–140.

[6] Breiman, L. (2001). Random forests. *Machine Learning*. 45 (1). p.pp. 5–32.

[7] Chandrahas, Dharmendra, & Raghuraj, 2017. Credit Card Fraud identification Using Artificial Neural Networks. 4(7). P.pp. 151-159.

[8] Chaudhary, K., Yadav, J. & Mallick, B. (2012). A review of Fraud Detection Techniques: Credit Card. *International Journal of Computer Applications*. 45 (1). p.pp. 975–8887.

[9] Chawla, N. V., Bowyer, K.W., Hall, L.O. & Kegelmeyer, W.P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 16 (January). p.pp. 321–357.

[10] Chung, Y., Kraska, T., Whang, S.E. & Polyzotis, N. (n.d.). Slice Finder : Automated Data Slicing for Model Interpretability. *SysML 2018*. [Online]. p.pp. 1–3. Available from: https://arxiv.org/pdf/1807.06068. [Accessed: 15 July 2018]

[11] Collell, G., Prelec, D. & Patil, K.R. (2018). A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*. 275. p.pp. 330–340.

[12] Dallagiacoma, M. (2017). *Predicting the risk of accidents for downhill skiers.* [Online]. Available from: https://pdfs.semanticscholar.org/ea65/8aa3c754c3ed6170ce0eccc43ef200e8427d.pdf. [Accessed: 19 July 2018]

[13] Delamaire, A. (2009). Credit card fraud and detection techniques : a review Title Credit card fraud and detection techniques : a review Credit card fraud and detection techniques: a review. *Banks and Bank Systems*. 4 (2).

[14] Gulati, A., Dubey, P., Mdfuzail, C., Norman, J. & Mangayarkarasi, R. (2017). Credit card fraud detection using neural network and geolocation. *IOP Conference Series: Materials Science and Engineering*. 263 (4).

[15] Gupta, B., Rawat, A., Jain, A., Arora, A. & Dhami, N. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computers Applications. 163 (8). p.pp. 15–19.

[16] Jain, N. & Khan, V. (2007). Credit Card Fraud Detection using Recurrent Attributes. *International Advanced Research Journal in Science, Engineering and Technology ISO*. 3297 (2). p.pp. 43–47.

[17] Jiang, S., Pang, G., Wu, M. & Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. *Expert Systems with Applications*.

[18] Ketan, Pranali, Prajal, & Neha, 2017. A Novel Idea for Credit Card Fraud Detection using Decision Tree. 161(13). P.pp. 6-9.

[19] Konyushkova, K., Raphael, S. & Fua, P. (2017). Learning Active Learning from Data. (Nips).

[20] Available from: https://papers.nips.cc/paper/7010-learning-active-learning-from-data.pdf

[21] Lango, M. & Stefanowski, J. (2018). Multi-class and feature selection extensions of Roughly Balanced Bagging for imbalanced data. *Journal of Intelligent Information Systems*. [Online]. 50 (1). p.pp. 97–127. Available from: https://link.springer.com/article/10.1007/s10844-017-0446-7. [Accessed: 15 July 2018]

[22] Liaw, a & Wiener, M. (2002). Classification and Regression by randomForest. *R news*. [Online]. 2 (December). p.pp. 18–22. Available from: https://www.researchgate.net/publication/228451484_Classification_and_Regression_by_RandomForest.

[23] Machine Learning Group (2013). *Credit Card Fraud Detection*. [Online]. September 2013. Available from: https://www.kaggle.com/mlg-ulb/creditcardfraud/home . [Accessed: 10 July 2018]

[24] Mishra, C., Gupta, D.L. & Singh, R. (2017). Credit Card Fraud Identification Using Artificial Neural Networks. 04 (07). p.pp. 151–159. Available from: file:///C:/Users/User/Desktop/dmpm/1.%20Application%20of%20Credit%20Card%20Fraud%20Detection_%20Based%20on%20Bagging%20Ensemble%20Classifier.pdf

[25] Pozzolo, A.D., Caelen, O., Johnson, R.A. & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced

classification. *Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*. p.pp. 159–166.

[26] Pumsirirat, A. & Yan, L. (2018). Credit Card Fraud Detection using Deep Learning based on Auto-Encoder and Restricted Boltzmann Machine. 9 (1). p.pp. 18–25.

[27] Roberto & Salvatore, 2017. Evaluating Credit Card Transactions in the Frequency Domain for a Proactive Fraud Detection Approach.

[28] Sahin, Y. & Duman, E. (2011). *Detecting Credit Card Fraud by Decision Trees and Support Vector Machines*. [Online]. I. p.pp. 315–319.

[29] Save, Prajal; Tiwarekar, Pranali; Jain, Ketan N; Mahyavanshi, N. (2017). A Novel Idea for Credit Card Fraud Detection using Decision Tree. International Journal of Computer Applications. [Online]. 161 (13). p.pp. 975–8887..

[30] SAS, n.d. *SAS® Enterprise Miner.* [Online]
Available at:
https://www.sas.com/content/dam/SAS/en_us/doc/factsheet/sas-enterprise-miner-101369.pdf

[31] Singh, N., Raw, R.A.M.S. & Chauhan, R.K. (2012). Data Mining With Regression Technique. 3 (1). p.pp. 199–202.

[32] Sznitman, R., Sznitman, R., Unibe, A. & Fua, P. (2015). Learning Active Learning from Real and Synthetic Data. (2).

[33] Tan, P.-N., Steinbach, M. & Kumar, V. (2006). Classification : Basic Concepts , Decision Trees , and. Introduction to Data Mining. [Online]. 67 (17). p.pp. 145–205. Available from: http://www-users.cs.umn.edu/~kumar/dmbook/index.php.

[34] Taneja, S., Gupta, C., Goyal, K. & Gureja, D. (2014). An Enhanced K-Nearest Neighbor Algorithm Using Information Gain and Clustering. *2014 Fourth International Conference on Advanced Computing & Communication Technologies*. (February 2016). p.pp. 325–329.

[35] Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M. & Sacha van Hijum, A.F.T. (2013). Data mining in the life science swith random forest: A walk in the park or lost in the jungle? *Briefings in Bioinformatics*. 14 (3). p.pp. 315–326.

[36] Xenopoulos, P. (n.d.). Introducing DeepBalance : Random Deep Belief Network Ensembles to Address Class Imbalance. Available from: https://arxiv.org/abs/1709.10056

[37] Yao, D., Yang, J. & Zhan, X. (2013). An Improved Random Forest Algorithm for Class-Imbalanced Data Classification and its Application in PAD Risk Factors Analysis. *The Open Electrical & Electronic Engineering Journal*. p.pp. 62–70.

[38] Yee, O.S., Sagadevan, S., Hashimah, N. & Hassain, A. (n.d.). *Credit Card Fraud Detection Using Machine Learning As Data Mining Technique*. 10 (1). p.pp. 23–27.