Machine Learning Approach for Analysing Anonymous Credit Card Fraud Patterns

Bushran Mohammed[1]

[1] Harrisburg University of Science and Technology

Author Note

MS in Data Analytics.

Correspondence concerning this article should be addressed to Bushran Mohammed, 326 Market St. Harrisburg, PA. E-mail: bmohammed@my.harrisburg.edu

Abstract

Credit card encourages the cardholder to acquire products and services depending on the pledge provided by the cardholder to pay for these products and services. Making purchases, and other transactions is simple, convenient, and futuristic. E-commerce and several other web platforms enhanced purchases online which increase the risk of fraud. Researchers began utilizing various machine learning approaches to identify and evaluate frauds in electronic purchases as a consequence of the exponential rise in fradulents. This analysis focuses almost exclusively on "Credit Card Fraud Analysis," and analyzes four common approaches utilized to evaluate consumer data from previous purchases to identify behavioural trends. Where cardholders are clustered into separate categories, depending on the volume of their purchase. Then using various methods behavioural patterns of the groups are extracted. Later different models are tested separably over categories. And then it is possible to choose the model with the best accuracy score as one of the best ways to predict fraud.

Keywords: logistic regression, decision tree, random forest, cluster, fraud, credit card, K-Means

Machine Learning Approach for Analysing Anonymous Credit Card Fraud Patterns

Paying with credit cards is a widespread practice. When credit card purchases expanded, fraudulent purchases often grow. This is not only a financial thing, but Identity Theft has also been a major problem for a few days now. However, the incidence of fraud is rocketing with the ever-increasing rise in online purchases, where the card simply stays unfeed or physically missing. As a reminder, online payment systems in 2015 have made out more than $31 trillion globally, credit card losses in the same line reported $21 billion in the same year. (Jiang, Song, Liu, Zheng, & Luan, 2018) It is projected that this trend will rise by 51% by 2025.

These days, the credit card is liable for billion-dollar size purchases. Global card business, had a financial value of about USD 28.84 trillion in 2014 alone (Manlangit, Azam, Shanmugam, Kannoorpatti, Jonkman, and Balasubramaniam, 2018). This implied rising the value of credit cards. It became part of the financial structures. Some of the main reasons being the comfort it has provided to customers.

Essentially, fraudulent activity involves using the credit card of others without their consent or permission. In certain instances, the suspect has no connection with the victim, nor will he even try to express the information about himself or the mechanism involved in the embezzlement; the money will never be recovered. (Prakash and Chandrasekar, 2013) Merchants are more at risk than consumers. The retailer is one of the main sufferers in the event of theft because his / her product is corrupted. Also, they have to compensate for the chargeback costs and run the possibility of closing their accounts. (Ganji and Mannem, 2012) Such can result in significant harm to the merchant's image, and may also face litigation of varying sort.

New fraud trends also have arisen, with the rise in the flexibility of payment methods. This has rendered the latest mechanisms for identifying fraud struggle. Another explanation why fraud detection mechanisms struggle is because individuals who commit fraud continuously alter habits while committing fraud, and this is exactly why the engineered barrier against fraudsters will use machine learning techniques, not just to combat but also to counter the ‘Concept Drift’ trend (Elitzur, Sai, 2010). Stolen credit card details may also be used by fraudulent operators to execute black market purchases, where cryptocurrencies such as bitcoin are now in wide usage.

Training the algorithm of fraud detection has a variety of important factors. Additionally, due to the privacy problem, public data is not always accessible. (Pozzolo, Caelen, Johnson, and Bontempi, 2015) The designed system also has to address factors such as non-stationary distribution of the data, decidedly imbalanced class distributions (skewed towards observations that are authentic) and unceasing streams of transactions.

It has been established from several studies that the need for purely reliable and high-performance fraud detection systems focused on automatic machine learning concepts that can hold or even outstrip the “SMOTE” is on the rise. SMOTE applies to create an area for the single fraud instances. The current rule-based standardized structures are too lagging behind the pace to deal with the criminal cartels ’ relentless process of creative theft tactics and leaves plenty to be anticipated. To address this gap, this study is based on analyzing anonymous credit card fraud patterns when the different number of fraudulent transactions is done the percentage value of SMOTE needs to be identified more clearly. Developing the current dataset to PCA format can make the performance of the data better and the possibility to be gained or lost. This work examines the proposed process on the binary dataset resulting in different performances.

Literature Review

Creating coherent trends per consumer reflect not only usual conduct, but also fraud trends that have been historically identified and verified as fraud transactions that encourage activity by studding fraudsters. An update to the existing Fraud Miner algorithm has been suggested. This improvement includes the implementation of a LINGO clustering data mining algorithm by removing the Apriori algorithm used in Fraud Miner to build regular trends and enabling the review of previous conduct of customers in either their legitimate or fraud transactions (Hegazy, Madian, & Ragaie, 2016).

Data mismatch is one of the main anomalies found in environments where anomaly detection is concerned. The identification of credit card fraud is one such field, the data of which is not free from this feature. Concentrating on variance and addressing it is one of the secrets to increasing every Classification algorithm's degree of precision (Nadarajan, and Ramanujam, 2016).

(Smith,2002) PCA is a valuable computational technique that has been used in fields including facial recognition and image compression and is a popular technique for detecting patterns in high-dimensional images.

Financial fraud is the fast-growing problem under the IoT ecosystem with the advent of mobile and online transfer services. Under the IoT setting, a highly efficient method of financial fraud identification is required, because financial fraud causes financial loss. As we analyzed financial fraud approaches, primarily from 2016 to 2018, utilizing machine learning and deep learning techniques, and suggested a framework for effective fraud identification focused on the advantages and disadvantages of each study (Choi, & Lee, 2018). The proposed process for detecting financial fraud and processing large amounts of financial data includes feature selection, sampling, and application of supervised and unsupervised algorithms.

Customers today choose the most approved payment method via credit card for the best way to buy online, paying bills in the simplest way. Around the same time the possibility of theft activity via credit card is a big challenge to prevent. There are tons of data processing methods needed to successfully mitigate these threats. Throughout the proposed research (Prakash, and Chandrasekar, 2012) semi Hidden Markov model (SHMM) anomaly detection algorithm is introduced to have improved precision and to prevent computational difficulty throughout fraud detection, which measures the difference between the processes tracked by the credit card detection method and the total normal processes.

As one of the core features of several security programs, data mining has evolved. Often used as a tool to identify theft, even to determine risk. The usage of data processing techniques to find new, true correlations as well as associations in massive data sets includes data mining. Across sectors such as finance, insurance, pharmacy and retailing, data mining has been commonly used to cut prices, improve analysis and boost revenue. Of this purpose, Fraud Detection requires tracking user/customer activity to predict, identify or prevent inappropriate conduct in the future (Chaudhary, and Mallick, 2012).

A popular question across disciplines motivates the increasing interest in data mining: how can you store, navigate, model, and eventually explain and understand very broad data sets? Historically, growing facets of data mining have been separately discussed by multiple disciplines. The concepts offer a tutorial description of and implementation of the principles underlying data mining algorithms. The data mining algorithms illustrate how algorithms are built rationally to solve problems. It also explains how much of the above research comes together as applied to data mining issues in the modern world (Hand, Mannila and, Smyth, 2001).

The increasing dependency of civilization on computers and digital technology has been accompanied by a spike in the scale and complexity of cyber-attacks perpetrated by Darknet-operating offenders. Security analysts have also developed an interest in scrutinizing the Darknet and other hidden online networks to build a deeper understanding of cybercrime and emerging threats. The DICE-E system offers a central point of reference and comprehensive guidance for researchers seeking to become involved in the study stream

Darknet (Benjamin, Valacich, & Hsinchun, 2019).

Organizations also have the expertise of policies to avoid or minimize threats to the protection of information technology. Nonetheless, such organizations cannot take appropriate steps to enforce such protocols and therefore remain prone to violations of protection. Based on interviews with IT managers at a multinational car distribution and marketing organization, possible explanations for this discrepancy in awareness of information protection and execution are given. Four methods are suggested to minimize this distance, along with a novel method for performing a laboratory trial to determine the efficacy of such methods, implemented individually and in combinations (Elitzur, Sai, 2010).

Since high-speed advancement in the online payment sector, the usage of credit cards has increased considerably. As the credit card continues to be the most common mode of payment for both electronic and daily transactions, instances of theft connected with it are often growing. The problem of not having an ideal state sequence for the underlying Markov cycle in regular Hidden Markov System even this observed series cannot be interpreted as training a system to match the observed data to the optimum. The key purpose of this work (Prakash, and Chandrasekar,2013) is to model the series of events in the handling of credit card purchases utilizing an Advanced Hidden Markov Method (AHMM), and to demonstrate how it may be used to identify fraud.

(Gayathri, & Malathi, 2020) These days the more reliable data sharing happens almost through the internet. Including the business corporations, the public often begin consuming the outlets of the network. Around the same time the possibility of successful data transmission always decreases. Some of the main problems among them are credit card fraud identification systems which currently have a large number of purchases that are classified as fraudulent. This can hinder the identification of fraudulent transactions. This requires into account Neural Network, Decision Tree, Naïve Bayes, K-nn, and Support Vector Machine to overcome the fraudulent.

The growing usage of credit cards in E-Banking communication networks for electronic and daily transactions is susceptible to credit card fraud. Data imbalance is indeed a major problem in preventing fraud. Proposing a smart two-level credit card fraud detection model from extremely imbalanced data sets, based on the semantic combination of k-means and artificial bee colony algorithm (ABC) to improve classification precision and speed up detection convergence. Besides, the classifier k-means that be surrounded by the optimum local, because it is prone to the initial state. Experimental findings (Darwish, 2020) show that the new model will improve the precision of classification against the probability of irregular transactions and have better consistency relative to conventional models.

For several years, banks have used early detection devices for fraud. Improved prevention of theft has since been critical to ensuring the integrity of the payment network. Outlier mining in data mining is an essential feature of the current algorithms that can be categorized into methods based on mathematical approaches, approaches based on size, methods based on density and methods based on variance (Ganji, and Mannem,2012).

(Dal Pozzolo, Boracchi, Caelen, Alippi, & Bontempi, 2017) Detecting payment card fee abuse is probably one of the toughest test grounds for algorithms in artificial intelligence. Nevertheless, the large majority of learning algorithms suggested for fraud detection are focused on-premises that barely fit in a real-world fraud detection system (FDS). There are two key reasons for this lack of realism: 1) the nature and pace of the distribution of supervised knowledge and 2) the methods used to determine the efficiency of fraud detection.

Analyzing specific evidence from several broad-scale randomized marketing trials of a major financial institution's pre-approved credit card solicitations, we notice that customers subscribing to the lower applications provided by the issuer have weaker attributes of credit rating. This result reinforces the claim that lenders of a riskier sort are limited by collateral or credit, and therefore have higher interest levels for reservation loans. Upon accounting for the observed risk factors, demographic features, and adverse economic fluctuations of a cardholder, we note that cardholders who reacted to the lower credit card offers are considerably more likely to default ex-post (Agarwal, Chomsisengphet, and Liu., 2010).

There are more and more credit card frauds, in terms of both quantity and number, according to foreign credit card organizations such as VISA. An anti-fraud project is built to solve the issue by merging two unsupervised algorithms: Principal Component Analysis and SIMPLEKMEANS Algorithm. For the proposed model (Lepoivre, Avanzini, Bignon, Legendre, and Piwele, 2016) on the generated test database, successful results are obtained by achieving the expected outcomes and the classification of potential frauds.

If the classification groups are not defined evenly fairly, a dataset is imbalanced. Real-world data sets are mostly comprised primarily of “normal” cases with just a limited number of “abnormal” or “interesting” ones. Under-sampling of the majority class has been suggested as a reasonable way of growing a classifier’s exposure to the minority class. This process of minority class over-sampling includes generating representations of fictitious minority groups (Chawla, Bowyer, Hall, & Kegelmeyer, 2002).

We found six methods of classification that can be split down into two specific categories: statistical and entropy-based. For model construction, four supervised learning algorithms are implemented, namely IB1, Naive Bayes, C4.5 decision tree, and the RBF network. Ranking approaches for numerous supervised learning algorithms in our experiments provide very different outcomes for a healthy accuracy. Our cases (Novaković, 2016) suggest that the usage of several specific indexes is advised to guarantee that a selection of features offering the best precision has been chosen.

Today, the increased usage of Internet credit cards in e-banking networks is vulnerable to credit card theft. Data imbalance often faces a major obstacle in identifying fraud. The reliability of current fraud detection programs is still at risk as it identifies illegal activity upon completion of the suspicious transaction. To overcome these difficulties, this article

(Darwish, 2020) provides an improved two-level credit card fraud monitoring model focused on the textual combination of k-means and the artificial bee colony (ABC) algorithm to improve recognition precision and speed detection convergence.

Since credit card fraud costs billions of dollars annually to the financial system, that the losses from credit card fraud is a significant catalyst for the industry and end-users. We concentrate on examining the actions of cardholder transactions and suggest a novel cardholder behavior model to identify credit card fraud. This program is named the Cardholder Behaviour Model (CBM). Regarding CBMs, two focus points are suggested and measured. It is to develop the model of conduct utilizing single-card transactions and multi-card transactions and incorporate holiday seasons as separate consumption times from the rest of the year (Kültür, & Çağlayan, 2018).

Lack of time for contact with the group is indulging several people to engage in social networking sites. Using this piece of technology consumers will communicate with each other. Ecommerce websites are now becoming common because consumers don't need to enter physical shops. As customers rising, frauds do the same. The primary objective of this literature (Ahuja, & Singh, 2017) is fraud detection. To achieve this function of KNN and Euclidean distance is hybridized.

(Wang, & Han, 2019) Currently credit card fraud is growing slowly with the popularization of credit cards. Built on this, this paper utilizes computational science to develop a forecast of credit card fraud focused on cluster analysis and optimized support vector machine. We checked the algorithm, and the outcome shows that the proposed algorithm effectively decreased the risk of unintended damage, which provides the card issuer with a tremendous opportunity to effectively reduce the economic damages incurred by credit card fraud, which has provided a strong theoretical base and realistic base for implementation.

An online credit card fraud identification framework focused on a neural classifier. Because it is built in a transactional center for service delivery, and not on a card issuing entity, it operates strictly on the details of the activity to be classified and its immediate prior background, and not on historical records of past cardholder operations. Some of the key characteristics of credit card activity is the broad imbalance between legitimate and illegal activities, and a strong degree of mixing of both (Dorronsoro, Ginel, and Sanchez, 1997).

Bank adds considerable significance to credit card fraud prevention as an inevitable outcome of multibillion-dollar cumulative damages suffered as a result of credit card fraud. The proposal (Kültür, & Çağlayan, 2017) also suggested the use of six established models, including Decision Tree, Random Forest, Bayesian Network, Naïve Bayes, Support Vector Machine, and K* models, to construct a credit card fraud detection ensemble. They based on the voting processes utilized by the group and suggested methods for optimistic, pessimistic and weighted votes.

Detection of credit card fraud is an integral aspect of monitoring suspicious purchases before card issuers approve them. While credit card fraud occurs extremely infrequently, it results in huge losses as the majority of fraudulent transactions have high values. A sufficient identification of crime helps prosecutors to take prompt steps that can deter future fraud or financial damage. The primary purpose of the fraud detection model is, thus, to produce reliable warnings and less false alarms and missing frauds. Throughout the analysis, it is performed an in-depth contrast between the hybrid ensemble and the approach of deep learning to decide whether or not to implement the latter in the framework of our collaborator currently working with the hybrid ensemble model (Kim, Lee, Shin, Yang, Cho, Nam, Song, Yoon, & Kim, 2019).

Methodology

Study's first component is the study of the dataset. The goal is to examine what method of anonymization has been used and how it has transformed the data. Before that the initial' Time' and' Amount ' functions are split down into different groupings. This will provide an overview of how many fraudulent transactions are involved in each class. Such groupings can be included in the scale of the planned process structure for the training. A fraud identification mechanism would be introduced. They would then check the new method. The first correlation would be between the dataset used in the original study and the scale of the data set for the fraud detecting method proposed. The next comparison would involve the usage of test data erroneously labeled in the initial analysis. The approach followed would involve the PCA-transformation of the function set to a subspace with limited loss of data. The chosen training set would then be implemented using the Synthetic Minority Oversampling Technique (SMOTE). After completion of the grouping, three models Logistic regression, Random Forest model and Desicion tree will be implemented for classification purposes, following which the tests will be analyzed.

Participants

The data set concerned had registered a total of 284,807 transactions over a 2-day duration. Of all transactions reported, 492 were listed as fraud (0.172% of total transactions). The dataset is strongly disequilibrated. (Pozzolo, Caelen, Johnson, and Bontempi, 2015) As per Han et al. (Han, Wang, and Mao, 2005), two types of imbalances in a dataset are generally observed. The first is within a class imbalance, which is an imbalance where one class has more samples than the other (as stated by Chawla), (Chawla, 2015) and the second is an imbalance within the class, where certain subsets within the class have less samples than the others in the class. (Weiss, 2004) The majority class is the one with plenty of samples and the minority class is the one with few samples.

Procedure

The dataset used in this research is from Kaggle.com. It is anonymous credit card details from holders of European credit cards. It has been anonymized to preserve credit card users' data. The data has been skewed in such a way that every user would not be recognizable. The dataset includes numerical input variables converted to PCA. (Pozzolo, Caelen, Johnson, and Bontempi, 2015) Principal Component Analysis (PCA) can be a basis for multivariate data analysis. One of PCA's aims is to find a connection between every data. (Wold, Esbensen, and Geladi, 1987)PCA translated the initial values into numbers, essentially shielding the credit card users' data. This also proposed sampling methods to fix imbalanced datasets.

182 International Journal of Electronic Commerce Studies The dataset comprises 30 columns, features V1 to V28 are values which have been converted into PCA to protect anonymity except for features such as “Time” and “Amount”. “Time” is the seconds that have passed in the data collection since the first transactions and “Amount” is the value of the transaction. The “Class” feature is the transaction description, showing that the transaction is a fraud. This thesis may use the features “Time” and “Amount” to recognize any trends that may contribute to future fraudulent transactions being recognized or detected. The Table 1 shows the features that are not transformed.

Table 1

Description of Features

Feature	Description
Time	The seconds elapsed between each transaction and the first transaction in the dataset.
Amount	The transaction Amount
Class	The response variable and it takes value 1 in case of fraud and 0 otherwise.

Measures

PCA allows for a detailed view of the connection between credit card purchases, including estimates. (Lepoivre, Avanzini, Bignon, Legendre, and Piwele, 2015) This can be extended to very broad databases where this method is capable of managing both the size and content. For credit card transactions, original data features are converted into a smaller subspace, without missing any data. (Lepoivre, Avanzini, Bignon, Legendre, and Piwele, 2015) The transformation of the original data into a smaller subspace can also be represented as dimensional reduction, which can be done easily using PCA. (Powell, 2015) In order to simplify the original data to one dimension, it was proposed that a principal component with the most variability be developed. High-dimensional data structures are always challenging to depict graphically but with PCA's capacity to minimize dimensions, analysis can be rendered much simpler and more intuitive. (Smith, 2002)

Time and Amount, two of the functions that were not converted using PCA, should be used to segment the dataset into several classes and consider the transaction distribution. Similar to fraudulent transactions, the summary would involve the fraudulent activity and the amount of fraudulent transactions.

The dataset covers purchases via credit card over two days. The unit of measure used for the 'time' function is in seconds. The period of each transaction makes use of the first transaction as the reporting reference point. Every transaction that occurs on the second day after 86,400 seconds would be listed as transactions and the last reported transaction in the dataset has a time of 172,792 seconds. This work sought to turn the seconds measurement unit into daytime hours, suggesting the first reported activity occurred at midnight on 12.

The same method will now be added to the' Amount' function. The knowledge gathered in this segment can expose the fraudulent transaction characteristics. Each segment would also use Sturges ' law to evaluate the amount of sections that should be required to separate the transactions. The category “Amount” will have 19 groupings. The category “Amount” will have 19 groupings. The findings revealed 1352.16 when calculating the class width. This is going to be rounded up into 1400. The maximum value corresponds to the highest transaction value in the dataset, while the lowest value relates to the lowest value in the dataset.

SMOTE will be included in the proposed fraud detection system. Fraud data has been described as an exception in the system, these are typically alone, and non-fraudulent data accompany them. Because the algorithm uses k-NN to identify the test data, the closest neighbors will decide the test data classification. While k can be a value of 1, it is easily affected by noise or irrelevant data. This is where the SMOTE samples will be positioned and neighbors will be generated around the fraud data, forming a clustered area around it. When the test data are in this area, it can be viewed as the test data having the same characteristics as the fraudulent data.

In this study three cost-insensitive classiffication algotrithms Logistic Regression, Random Forest and Desicion tree used as models for the pupose of implimentation. These cost-insensitive algortihms are using their own percpective libraries. Each model is well trained and tested with various dimentionality reduction.

Analysis

When a new transaction comes through the analysis begins. Each transaction should be translated to PCA in such a manner that it meets the PCA translation cycle of the initial dataset. A subset from the initial dataset is used as the training sample for the current transaction classification. Using the ‘Time’ function, the subset must follow the class groupings.

The next cycle is the implementation of the Synthetic Minority Oversampling Technique (SMOTE) to the chosen subset’s fraud class. SMOTE is an oversampling technique that produces synthetic samples instead of over-sampling of substitutions. Such synthetic samples are seen along the line segments of bonafide minority samples utilizing the nearby 190 adjacent International Journal of Electronic Commerce Studies k minority class. Neighbours from the nearest k neighbours are chosen randomly based on the defined number of synthetic samples that need to be made. In this research work SMOTE aims to build neighbours for the fraudulent data.

The next cycle is to establish a grouping of elements, the clustering distances of which are identical. Using k-NN, classification can then be added with the value of k defined in the previous section of this analysis. Then, it decides the algorithm's classification decision. Every grouping shows the list of closest neighbours dependent on k, with each grouping's decision on distance and classification.

Clustering is a common strategy for grouping related data points or items into clusters or classes. Clustering is an essential tool for outlier research. Cluster-based strategy works as a decrease of data here. Primarily, clustering methodology is used to combine data with identical properties. And then, for every group measure the centroids. The K-Means clustering produces a specific number of flat, disjoint clusters. It is well adapted for the development of globular clusters. The K-Means approach is linear, unsupervised, iterative and not deterministic. Clustering of K-means and outlier identification is merged to create a hybrid strategy.

A well-established mathematical approach for forecasting binomial or multinomial outcomes is logistic regression. Multinomial Logistic Regression algorithm can produce models when a set field with two or more possible values is the target field. Binomial Logistic Regression algorithm is limited to models where a flag, or binary field, is the target field. Since the logistic regression forecasts probabilities, we should suit it using probability rather than only groups. We have a set of functions and an observable class for each training data. Regardless on how the risk is measured, though, the equations do not take into account the various effects on misclassification. The cost-sensitive logistic regression is one model historically established to incorporate the different financial costs during the training process. This approach incorporates explanation-dependent costs into a logistic regression by turning the model's objective function into a cost-sensitive one.

Random Forest is a type of supervised algorithm for machine learning focused on ensemble learning. Ensemble learning is a method of learning where you several times combine different types of algorithms or the same algorithm to create a more efficient model of prediction. The algorithm for random forest incorporates several algorithms of the same form i.e. several decision trees, resulting in a trees forest, hence the term "Random Forest." Saia and Carta (Saia, Carta, 2019) suggested a novel approach using Fourier transform and Wavelettransform to evaluate and experiment with k-foldcross-validation data using the Random Forests process, compared to ten Data mining algorithms and concluded Random Forests as the best performing model. Dal Pozzolo et al.(Dal Pozzolo, Caelen, Le Borgne, Waterschoot, Bontempi, 2014) contrasted RandomForests with Neural Network and Support Vector Machine where Random Forests worked as anticipated and proposed that the accuracy could be enhanced by increasing the size of the training results. Carneiro et al.[21] applied 10-fold cross-validation to Random Forests, SVM and Logistic regression and then tested with balanced and unbalanced data to confirm that Random Forests achieved the best performance possible.

The random forest tree is supervised machine learning methodology used to solve regression and classification problem. A decision tree is a tree structure that tries to divide the given records into subgroups which are mutually exclusive. To do this, each node is split into child nodes starting from the root node in a binary or multi split fashion related to the method used based on the value of the attribute (input variable) which best separates the given records. Records in a node are divided recursively into child nodes until there is no separation that causes statistical difference in the node's distribution of the records, or the amount of records in a node is too low. Every decision tree approach uses its own algorithms of splitting and metric splitting. When the tree is formed, the resulting tree will overfit the training data and may include potential mistakes or noise, or any of the resulting tree branches may include abnormalities. The resulting tree should therefore be checked if the removal of some nodes, starting with the leaf ones, has a significant effect on the performance of the tree classification. This is called pruning operation.

Result

Dataset Distribution

There are totally 284807 records. There are 31 features for each record including V1, V2, V3 .... V28 which conceals the true sense of privacy as well as the "Time," "Amount" and "Class" features. For this project the "Time" function is meaningless. The "Amount" means how much money is expended on every purchase. The "Class" shows that the transaction is fraud "1" or not "0". In Fig. 2 it could be shown that there are 284315 records being the Class "0", and 492 records being the Class "1". So most records (99.83%) aren't fraudulent, the data is very imbalanced. There are essentially two methods to cope with the imbalanced data through additional data pre-processing: undersampling and oversampling. Until we do the undersampling, we will first delete the "Time" function and standardize the "Number" function like the other 28 features are standardized. The plot displays in Fig. 3 the bulk of the transaction is not fraud. Throughout the multivariate plot Fig. 4, it interprets that all the vector values follow the class value "0," since this indicates that the bulk of transactions are not fraudulent.

K-Means Algorithm

In the K-means algorithm scheme the input data is grouped into the number of groups defined. This is the unsupervised learning technique that is used where there is no previous information in a sample for a certain series of observations. This scheme classifies n data points into predetermined clusters of k. The data will be grouped together into k-clusters according to cluster similarities. It is important to identify k-clusters in the first step. For each cluster is selected randomly centroid in the second step. Centroid of a given cluster represents the cluster's mean worth. The data distance from each cluster's centroid is determined in the third stage. Information should be clustered into a different cluster dependent on the minimum data distance from the cluster centroid. Centroid should be recalculated for each cluster when various values arrive in cluster again each time. Such measures must be replicated until the performance will not shift.

Cluster Analysis

Clustering is a common approach used for grouping related data Points or points in clusters or in groups. Here cluster-based method serves as data reduction. First, the technique of clustering is used to group data having similar characteristics. So then, for each group measure the optimal number of clusters. The silhouette function in the cluster package is to compute the average silhouette width. There are three popular methods for determining the optimal number of clusters

1. Elbow Mthod 2. Silhouette method 3. Gap Statistics

In this study, silhouette method is utilized as its approach measure the quality of the clustering. This method requires the dataset to be scaled and normalized. By the use of Euclidean method the function calculates and returns the distance matrix by using specified distance measure to compute the distances between the rows of the matrix. As a result it shows 2 cluster maximize the average silhouette values in Fig. 5. The visualization of k-means confirms that the class "0" has the maximum number of frauds. There are two clusters Fraud and Non fraud as shown in Fig. 6.

Hierarchical Analysis

The data should have no missing values when doing this method of clustering study, should also be scalable and standardized. Hierarchical clustering has separate functions available in R for hierarchical clustering computation. The widely used hclust and diana functions would each have identical outputs. Agglomerative Hierarchical Clustering with hclust feature is used in this research. There is complete, average, single and ward for agglomerative methods. The complete linkage process is used here and a cluster dendogram plot has been produced. The dendogram with borders around the 8 clusters can be drawn using rect.hclust function as shown in Figure 7.

Logistic Regression Model

For each dataset, logistic regression was developed using the respective training data set. I have used the remaining transactions for each model in the relevant test sets to evaluate those models. The fraudulent transactions are identical in the sample test sets. Levels of accuracy were used to characterize the models' utility. Perhaps the most commonly used metric for measuring the performance of targeting models in classification applications is accuracy. The amount of fraudulent transactions detected, however, still constitutes an significant performance measure. The model itself actually predicts input outcome likelihood and does not conduct statistical classification (it is not a classifier), but it appears to be used to construct a classifier.

With the support of the Cutoff feature, optimum cut-off is calculated to boost the efficiency of the model. The cut-off is used by the latter formula for the most important variable which provides the strongest output for all variables in contrast to the maximum cut-off for logistic model. The most important variables displayed by the cutoff function are "V14" "V17" "V10" "V12" "V4" "V11" "V3" "V16" "V7" "V2" "V9" "V21" "V27" "V18" "V1".

Random Forest Model

By using this algorithm we extracted the exact fraud detection percentage from the given dataset by studying its behaviour. A confusion matrix is essentially a description of the prediction results or a table used to explain the output of the classifier on a collection of test data where true values are identified. It visualizes the output of an algorithm and enables fast class identification. As a result, most performance measures are computed by providing insight not only into the errors that are made by the classification model, but also into the type of errors that are made. This algorithm is used to make random samples. Using varImpPlot, the 15 most important variables are shown in Figure 8. Then, by re-creating the model, adding the most important variable each time, and checking its accuracy. With support for ggplot in Figure 9, all F1 scores are measurements to see which variable set generates the better pattern.

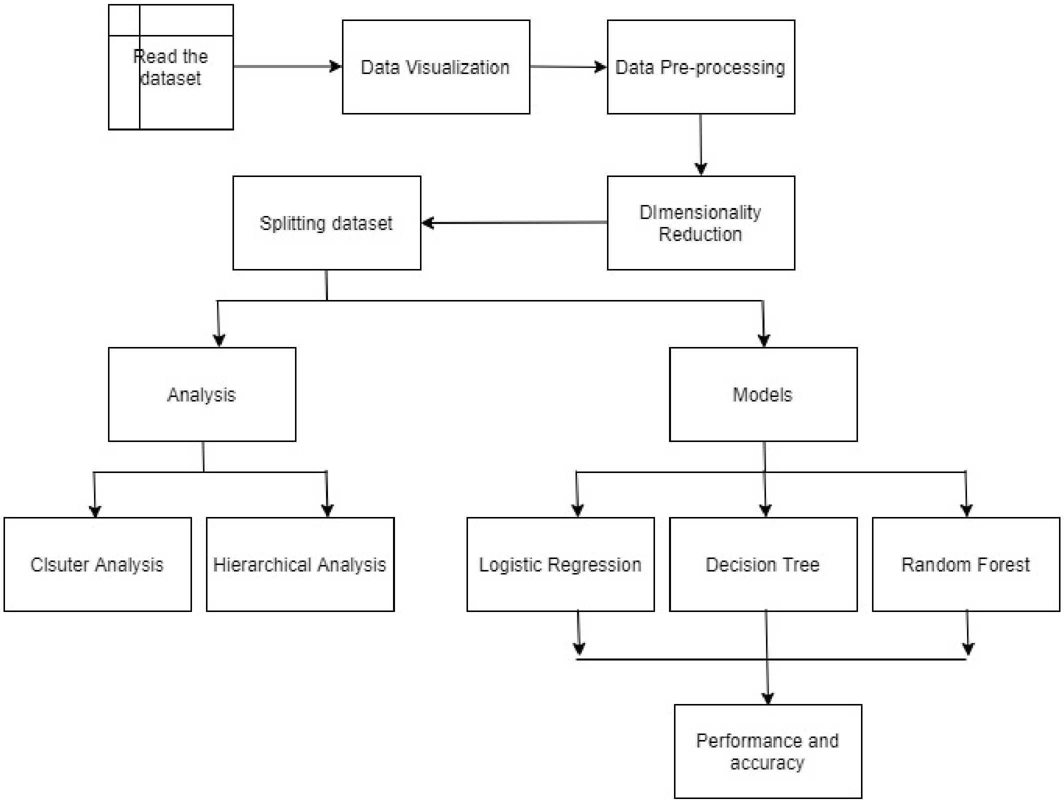
Desicion Tree Model

In this analysis, seven alternative models focused on decision tree algorithms and decision tree were built using the appropriate training data set for each sample. In order to validate these models, we used the remaining transactions in the related test sets for each model. Fraudulent transactions in the experimental study set are much the same. Accuracy rates have been used to explain the utility of models. Accuracy is perhaps the most widely employed parameter for calculating the efficiency of targeting models in classification applications. Decision trees use various algorithms to agree to divide the node into two or more sub-nodes. The formation of sub-nodes enhances the homogeneity of the subsequent sub-nodes. In other terms, we may conclude that the purity of the node is that with respect to the goal value. The decision tree splits the nodes on all relevant variables. We used a decision tree programming algorithm that can determine whether the incoming transaction is fraud or not. The decision tree is created by running a package rpart that takes Fraud as a answer variable and other variables are predictors that distinguish incoming transactions as not fraud and fraud. The function is model_DT <- rpart(Class ., train_balanced, control = rpart.control(maxdepth = 3)). In order to improve the performance of the decision tree, the most significant variables are taken from the trained model and the model is aligned with the most significant variables and the decision tree is designed to predict transaction class as fraud or not as shown in Figure 10.

Proposed Work

The techniques proposed for detection used in this paper are for Fraud in the credit card system. The distinction is rendered with various machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, to decide the algorithm better fits and can be modified by credit card traders to recognise fraud transactions. The Figure 1 represents the architectural design of the overall analysis.

Figure 1. Architecture Design



Related Work

There are several credit card fraud identification algorithms (Chen, Rong-Chang, 2004). Fraud identification requires tracking the actions of consumers in order to predict, track or prevent inappropriate behaviour. Financial fraud is increasing significantly in the development of modern technology and therefore fraud detection is a very important area. In order to combat credit card fraud effectively , it is necessary to understand the technologies involved in detecting credit card fraud and to identify different types of credit card fraud (Chen, Rong-Chang, 2004). Sanchez 'et al. (Sánchez, Daniel, 2009) defined the approach for detecting fraud in transactional repositories using the fuzzy association rule mining in information extraction. This approach is very successful and optimizes execution time and decreases the unnecessary generation of laws. Those are artificial neural-network models focused on artificial intelligence and deep learning (John, Kennedy O, Anele, Olajide, Chinyere Grace Kennedy, 2016), hierarchical data mining networks, sequence matching algorithm centred on the cardholder 's expenditure profile, adaptive decision-making algorithms based on artificial intelligence, Meta computing Era.

Original confidence in each incoming transaction may be determined through numerous facts originating from the law-based portion using Dempster 's theorem. S. Maes et al. (Maes, Sam, 2002) identified the fraud identification method utilizing BBN and ANN. They find that BBN is offering a stronger outcome than ANN. The preparation time for BNN is limited relative to ANN. Most of the credit card fraud identification systems listed above are focused on artificial intelligence, meta-learning and pattern matching.

(John, Kennedy O,Anele, Olajide, Chinyere Grace Kennedy, 2016) Many innovations used in the identification of credit card fraud are the Cloud Services-Based Collaboration Scheme for Credit Card Fraud Detection, in which participating banks may exchange awareness of fraud trends in a heterogeneous and global ecosystem to improve their capacity to identify fraud and minimize financial losses, the Credit Card Fraud Detection with the

Artificial Immune System. Chen et al. (Lu Q, Ju C, 2011) proposed a modern approach for identifying fraud in which QRT data were obtained via an electronic questionnaire. The support vector machine is used to train data and create QRT models that are used to determine whether new transactions are fraud or not.

Future Work

Efficient credit card fraud identification program is a crucial prerequisite for every card issuing bank. Credit card fraud identification has drawn a great deal of attention from the academic community, and a range of strategies have been developed to combat credit card faults. The Fuzzy Darwinian Fraud Detection System improves system accuracy. Although the fraud detection broad of Fuzzy Darwinian Fraud detection systems in terms of true positive is 100 percent and provides strong results in detective fraud transactions. CARD WATCH's neutral network demonstrates that the quality of fraud identification and processing speed is still high, but is restricted to one network per user. Fraud identification of the secret Markov model is very weak relative to other approaches. The hybridized algorithm called BLAH-FDS recognizes and prevents fraudulent transactions using the sequencing algorithm tool. The processing speed of BLAST-SAHA is high enough to enable on-line identification of credit card fraud. BLAH-FDS may be used successfully to combat theft in many fields, such as telecommunications and financial fraud identification. ANN and BNN are used for the detection of cell phone fraud, network intrusion. All the credit card fraud prevention methods mentioned in this study paper have their own strengths and disadvantages. Such a survey bill allows them to create a holistic solution to the detection of suspicious credit card purchases.

Conclusion

The broad implication of the present research is that it is necessary to use features that examine the customer behaviour of individual cardholders while designing a credit card fraud detection model. Also proposed as a flexible platform for storing robust quantities of data, the system features can be generalized to collect real-time data from multiple desperate sources. The derived data is then used to establish a powerful analytical model. By implementing the SMOTE, we attempted to stabilize the dataset, where we found that the classifiers were doing stronger than before. Pre-processing data in order to provide recent user activity, efficiency improves relative to the usage of direct purchase information only. As the final clustering outcome of the k-means clustering methods is strongly dependent on the collection of the initial centroids, it should be a comprehensive approach for deciding the initial centroids, which lets the k-means algorithm converge in a global optimal and special clustering outcome. Finally, Machine learning methods such as logistic regression, random forest and decision tree have been used to detect fraud in the credit card system. The accuracy for logistic regression, random forest and decision tree models is 94.9, 99.3 and 93.1 respectively. After evaluating the three approaches, it was decided that the random forest model was higher than the logistic regression and decision tree.

References

Hegazy, M., Madian, A., & Ragaie, M. (2016). Enhanced Fraud Miner: Credit Card Fraud Detection using Clustering Data Mining Techniques. Egyptian Computer Science Journal, 40(3), 72–81.

Nadarajan, S., and Ramanujam, B.(2016), Encountering imbalance in credit card fraud detection with metaheuristics. Advances in Natural and Applied Sciences, 10(8), p33-42.

Smith,L. I. (2002), A tutorial on Principal Components Analysis. Information Fusion, 51, 52.

Choi, D., & Lee, K. (2018). An Artificial Intelligence Approach to Financial Fraud Detection under IoT Environment: A Survey and Implementation. Security & Communication Networks, 1–15. https://doi-org.proxy-harrisburg.klnpa.org/10.1155/2018/5483472

Prakash, A. and Chandrasekar, C. (2012), A Novel Hidden Markov Model for Credit Card Fraud Detection. International Journal of Computer Applications, 59(3), p35-41.

Chaudhary, K. and Mallick, B.(2012), Exploration of Data mining techniques in Fraud. Detection: Credit Card. International Journal of Electronics and Computer Science Engineering, 1(3), p1765-1771.

Hand, D. J., Mannila, H. and Smyth, P. (2001), Principles of data mining, MIT Press.

Benjamin, V., Valacich, J. S., & Hsinchun Chen. (2019). Dice-E: A Framework for Conducting Darknet Identification, Collection, Evaluation with Ethics. MIS Quarterly, 43(1), 1–22. https://doi-org.proxy-harrisburg.klnpa.org/10.25300/MISQ/2019/13808

Elitzur, R., Sai, Y. (2010), A Laboratory Study Designed for Reducing the Gap between Information Security Knowledge and Implementation. International Journal of Electronic Commerce Studies, 1(1), p37-50.

Prakash, A., and Chandrasekar, C. (2013), A parameter optimized approach for improving credit card fraud detection. International Journal of Computer Science Issues, 10(1), p360-366.

Gayathri, R., & Malathi, A. (2020), (n.d.). Investigation of Data Mining Techniques in Fraud Detection: Credit Card.

Darwish, S. M. (2020). An intelligent credit card fraud detection approach based on semantic fusion of two classifiers. Soft Computing - A Fusion of Foundations, Methodologies & Applications, 24(2), 1243.

Ganji, V. R., and Mannem, S. N. P., (2012), Credit card fraud detection using anti-k nearest neighbor algorithm. International Journal on Computer Science and Engineering, 4(6), p1035-1039.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (n.d.). Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. IEEE Transactions on Neural Networks and Learning Systems, Neural Networks and Learning Systems, IEEE Transactions on, IEEE Trans. Neural Netw. Learning Syst, PP(99), 1–14. https://doi-org.proxy-harrisburg.klnpa.org/10.1109/TNNLS.2017.2736643

Agarwal, S., S. Chomsisengphet, and C. Liu. 2010. “The Importance of Adverse Selection in the Credit Card Market: Evidence from Randomized Trials of Credit Card Solicitations.” Journal of Money, Credit and Banking 42 (4): 743–754. doi: 10.1111/j.1538-4616.2010.00305.x

Lepoivre, M. R., Avanzini, C. O., Bignon, G., Legendre, L., and Piwele, A. K. (2016), Credit card fraud detection with unsupervised algorithms (Report). Journal of Advances in Information Technology, 7(1), 34. https://doi.org/10.12720/jait.7.1.34-38

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-Sampling Technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)

Novaković, J. (2016),Toward optimal feature selection using ranking methods and classification algorithms. Yugoslav Journal of Operations Research, 21(1), p119- 135. https://doi.org/10.2298/YJOR1101119N

Darwish, S. M. (2020). A bio-inspired credit card fraud detection model based on user behavior analysis suitable for business management in electronic banking. Journal of Ambient Intelligence and Humanized Computing

Kültür, Y., & Çağlayan, M. U. (2018). A Novel Cardholder Behavior Model for Detecting Credit Card Fraud. Intelligent Automation & Soft Computing, 24(4), 813.

Ahuja, M. S., & Singh, L. (2017). A Novel Approach of Detecting Frauds in

Ecommerce Sites by Hybridizing Knn and Euclidean Distance Mechanism. International Journal of Advanced Research in Computer Science, 8(5), 2169–2172.

Wang, C., & Han, D. (2019). Credit card fraud forecasting model based on clustering analysis and integrated support vector machine. Cluster Computing, 22, 13861.

Dorronsoro, J. R., Ginel, F., Sanchez et al., C. (1997), "Neural fraud detection in credit card operations," IEEE Trans. on Neural Networks, vol. 8, no. 4.

Kültür, Y., & Çağlayan, M. U. (2017). Hybrid approaches for detecting credit card fraud. Expert Systems, 34(2), n/a-N.PAG. https://doi-org.proxy-harrisburg.klnpa.org/10.1111/exsy.12191

Kim, E, Lee, J, Shin, H, Yang, H, Cho, S, Nam, S, Song, Y, Yoon, J & Kim, J 2019, 'Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and deep learning', Expert Systems with Applications, vol. 128, pp. 214–224, viewed 29 February 2020, http://search.ebscohost.com.proxy-harrisburg.klnpa.org/login.aspx?direct=true&db=aci&AN=136419749&site=eds-live&scope=site

Jiang, C., Song, J., Liu, G., Zheng, L., & Luan, W. (2018). Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism. IEEE Internet of Things Journal, Internet of Things Journal, IEEE, IEEE Internet Things J, 5(5), 3637–3647. https://doi.org/10.1109/JIOT.2018.2816007

Manlangit, S., Azam, S., Shanmugam, B., Kannoorpatti, K., Jonkman, M., and Balasubramaniam, A. (2018). An efficient method for detecting fraudulent transactions using classification algorithms on an anonymized credit card dataset. Intelligent Systems Design and Applications, Springer, 736, p418-429. https://doi.org/10.1007/978-3-319-76348-4_41

Elitzur, R., Sai, Y. (2010), A Laboratory Study Designed for Reducing the Gap between Information Security Knowledge and Implementation. International Journal of

Electronic Commerce Studies, 1(1), p37-50.

Pozzolo, A. D., Caelen, O., Johnson, R. A. and Bontempi, G. (2015), Calibrating Probability with Undersampling for Unbalanced Classification. In IEEE Symposium Series on 200 International Journal of Electronic Commerce Studies Computational Intelligence, pp: 159-166. https://doi.org/10.1109/SSCI.2015.33

Han, H., Wang, W., and Mao, B.(2005), Borderline-SMOTE: a new over-sampling method in imbalanced datasets learning. Advances in intelligent computing, p878-887. https://doi.org/10.1007/11538059_91

Chawla, N. V. (2015), Data mining for imbalanced datasets: An overview. Data mining and knowledge discovery handbook, Springer, Chap:40. https://doi.org/10.1007/978-0-387-09823-4_45

Weiss, G. (2004), Mining with rarity: A unifying framework. SIGKDD Explorations, 6(1), p7-19. https://doi.org/10.1145/1007730.1007734

Wold, S., Esbensen, K., and Geladi, P. (1987), Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), p37-52. https://doi.org/10.1016/0169-7439(87)80084-9

Powell, V. (2015), Principal Component Analysis. Retrieved from: http://setosa.io/ev/principal-component-analysis/.

Smith, L. I. (2002), A tutorial on Principal Components Analysis. Information Fusion, 51, 52.

Saia, R., Carta, S. (2019): Evaluating the benefits of using proactive transformed-domain-basedtechniques in fraud detection tasks. Future Gener. Comput. Syst. 93,18–32

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.A., Waterschoot, S., Bontempi, G. (2014): Learnedlessons in credit card fraud detection from a practitioner perspective. Expert Syst. Appl. 41(10), 4915–4928

http://www.iaeng.org/publication/IMECS2011/IMECS2011_pp442-447.pdf

Sánchez, Daniel, et al. (2009),"Association rules applied to credit card fraud detection." Expert systems with applications 36(2), 3630-3640.

Chen, Rong-Chang, et al. (2004) : "Detecting credit card fraud by using questionnaire-responded transaction model based on support vector machines." Intelligent Data Engineering and Automated Learning–IDEAL 800-806.

S.N. John, Kennedy O, C. Anele, F. Olajide, Chinyere Grace Kennedy, (2016) “Fraud Detection in the Banking Sector Using Data Mining Techniques Algorithm”, International Conference on Computational Science and Computational Intelligence

Lu Q, Ju C (2011) “Research on credit card fraud detection model based on class weighted support vector machine”, Journal Convergence Information Technology l 6, 62–68.

Chen, Rong-Chang, et al. (2004) : "Detecting credit card fraud by using questionnaire-responded transaction model based on support vector machines." Intelligent Data Engineering and Automated Learning–IDEAL 800-806.

Sánchez, Daniel, et al. (2009),"Association rules applied to credit card fraud detection." Expert systems with applications 36(2), 3630-3640.

A. Kundu, S. Sural, and A. Majumdar,(2006) "Two-stage credit card fraud detection using sequence alignment," Information Systems Security, Springer Berlin , Heidelberg, 260-275.

Appendix

Figures

Figure 2. Classification of Class Feature.

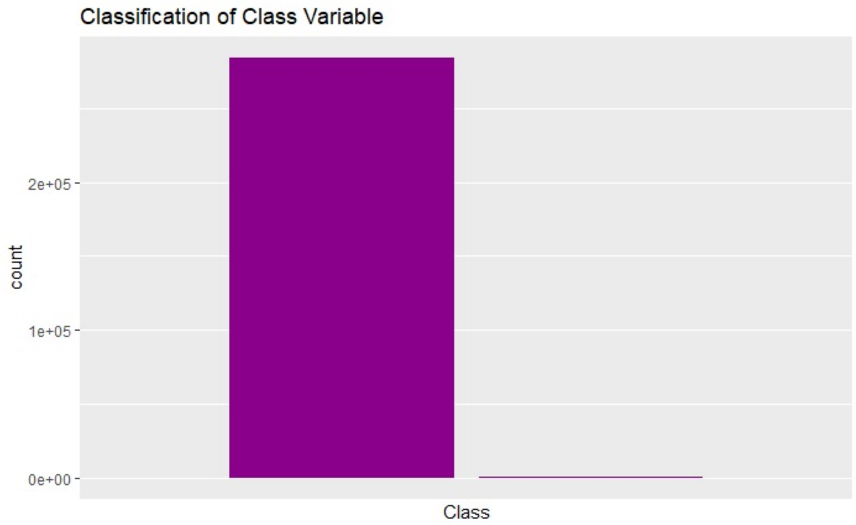


Figure 3. Classification of Amount by Class Feature.

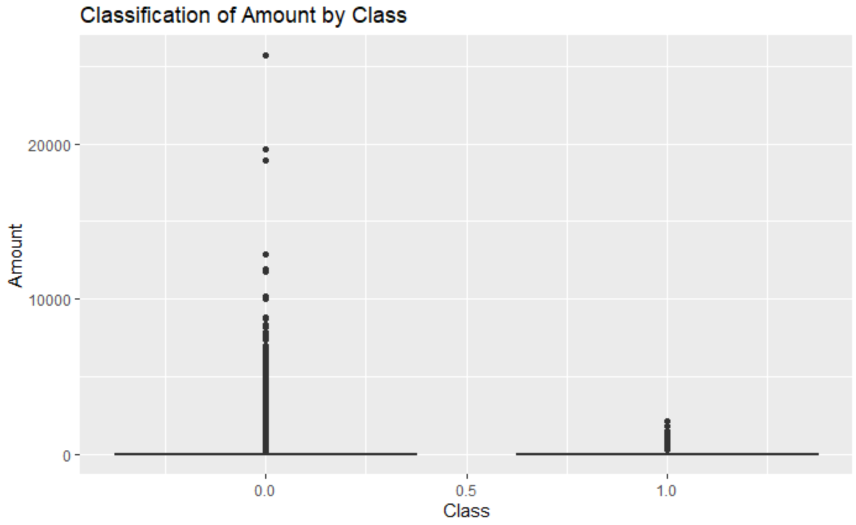


Figure 4. Classification of Non-Fraud and Fraud Class.

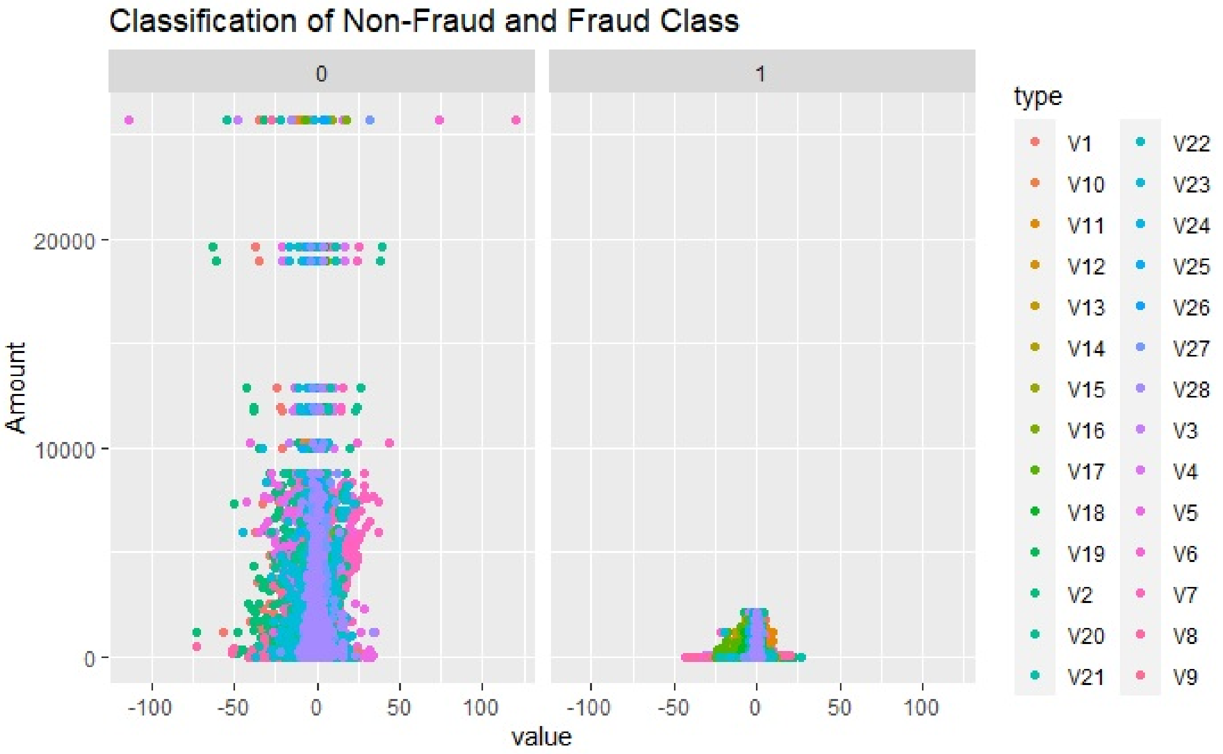


Figure 5. Number of optimal Clusters K.

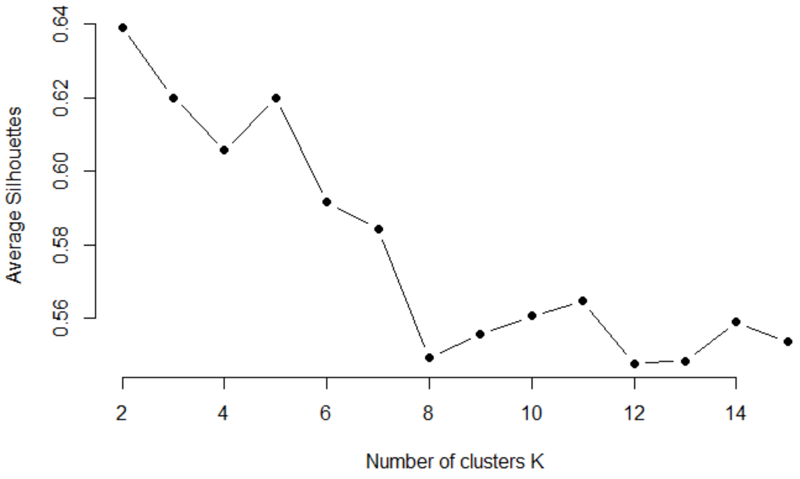


Figure 6. Clustering of the Class Feature.

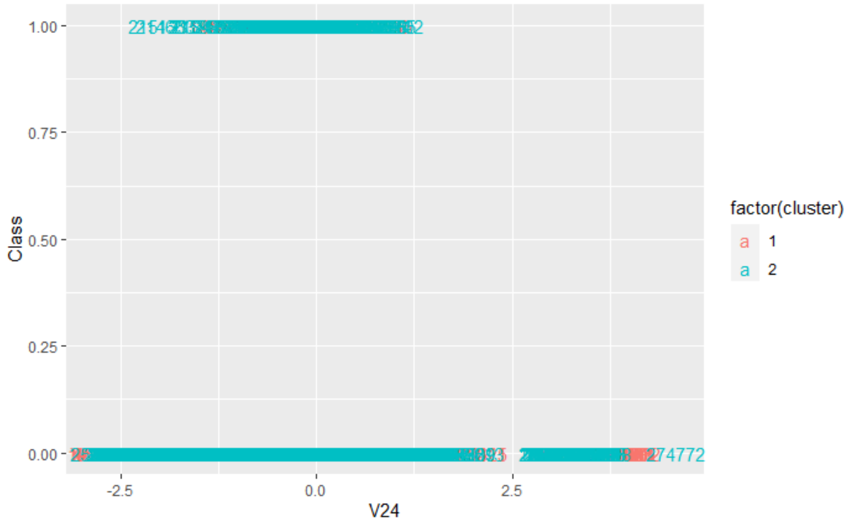


Figure 7. Hierarchical Clustering.

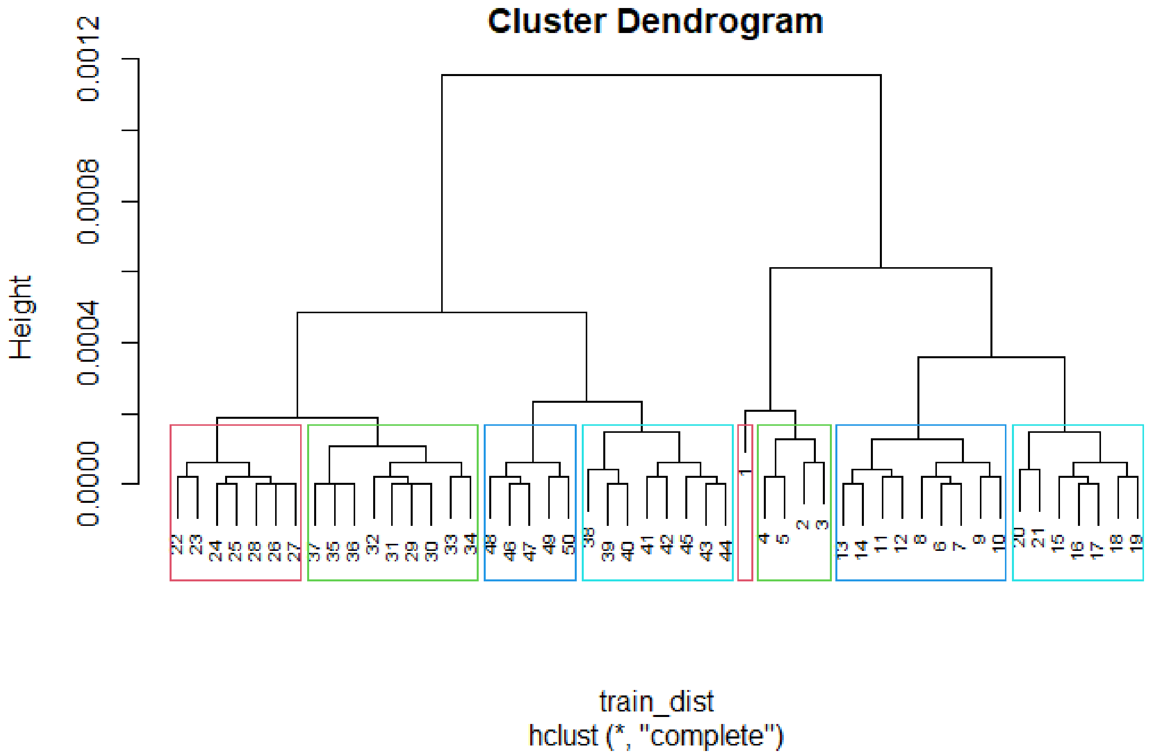


Figure 8. Important variables.

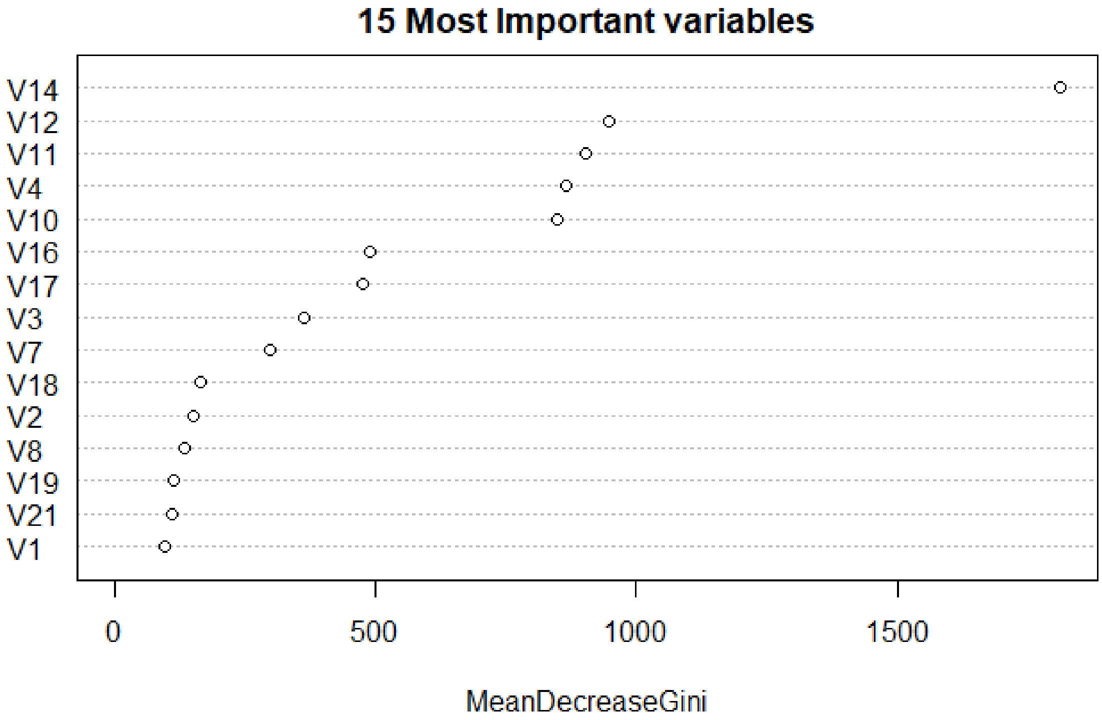


Figure 9. F1 Score.

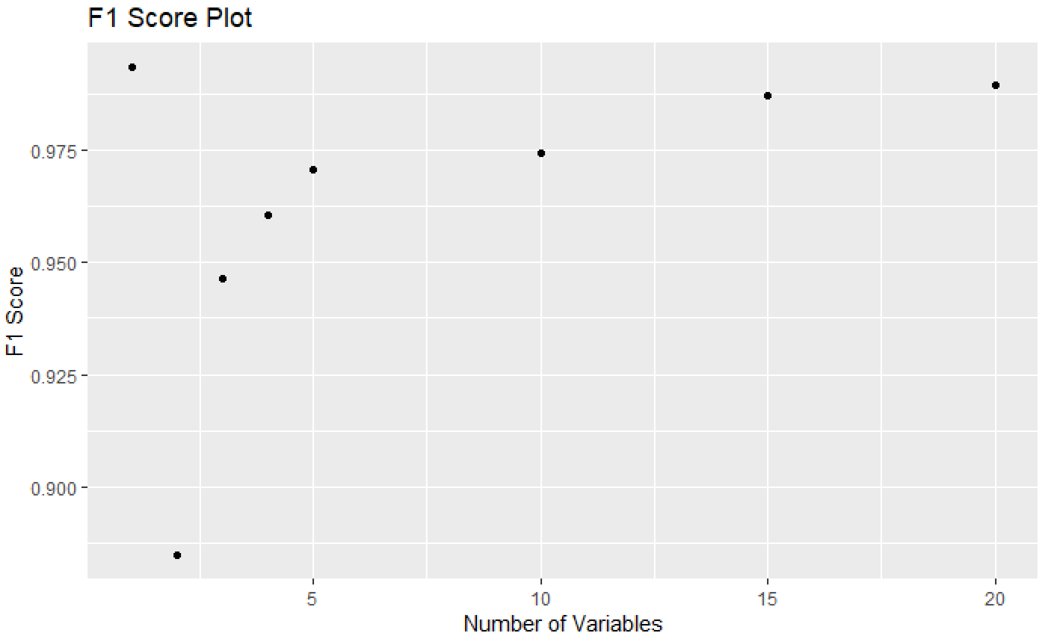
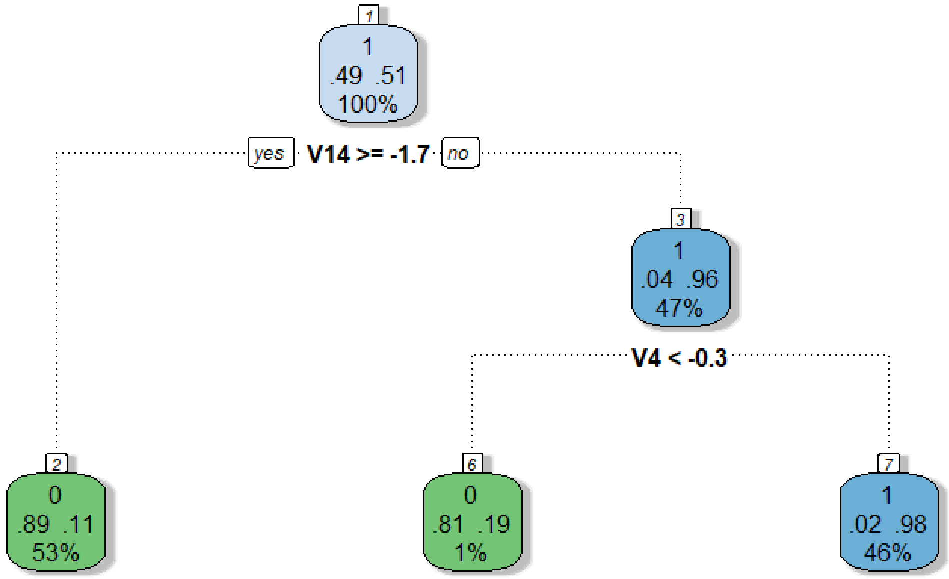


Figure 10. Decision Tree Analysis.



Rattle 2020-Jul-05 15:32:05 bushr

Code

Here is the link to my github: https://github.com/MBushran/ANLY699.git