Credit card fraud detection using random forest

Aruna Vedula

Harrisburg University of Science and Technology

Author Note : Correspondence concerning this article should be addressed to Aruna Vedula.

Department of Data Analytics, Harrisburg University of Science and Technology. E-mail:

avedula@my.harrisburgu.edu

Abstract

Though many machine learning algorithms have been extensively studied, the use of random forests for fraud detection has not been researched enough. This paper examines the advantages of using random forests in the detection of credit card fraud particularly on large, unbalanced datasets and proposes that additional research is conducted to better understand the use of random forests in credit card fraud detection. The unbalanced feature of the data is left untreated in this research, in order to examine how random forests, handle this feature. Further, the metric used to evaluate the performance of the random forests classifier is the area under the ROC Curve (AUC) metric which is insensitive to imbalanced data and reflects the overall ranking performance of a classifier.

Keywords: Random forests, machine learning, Logistic Regression, Neural Networks (NN), Bayesian Network (BN), Fuzzy Logic, Decision Trees (DT), Support Vector Machines (SVM), ROC (receiver operating characteristic) Curve and AUC (area under the curve)

Credit card fraud detection using random forest

Numerous studies have been conducted to understand the use of various machine learning algorithms in detecting a fraudulent transaction of any nature (Albashrawi, M 2016; Bhattacharyya et al., 2011; Glancy and Yadav 2011; Masoumeh, Z 2013). However, the use of random forests for fraud detection has not been researched enough (Ngai et al.,2011). This paper analyzes the existing research on credit card fraud detection and extends it by proposing the use of random forests algorithm in the detection of credit card fraud where categorical variables are involved. For categorical variables, mostly logistic regression is used as it is the most popular and tried and tested algorithm (Peng, Lee & Ingersoll 2002). But there are issues like handling large datasets and noise in the data and the presence of nonlinear parameters that are handled by random forests better than logistic regression models (West et al., 2014). As Altendorf et. al (2009) mentioned that "The Random Forest algorithm has been found to be resistant to overfitting and provides a good estimate of the generalization error (without having to do cross-validation) through the "out-of-bag" error rate that it returns" (p. 3)

Literature review

What is Financial fraud?

Before we discuss what is financial fraud, let us first understand fraud. Vasiu L, Warren M and Mackay D (2003) encapsulate fraud as the act of intentionally representing a false information with an ulterior motive of gain or to cause loss to the other. They further state that the term "fraud" can be used to describe various acts like bribery, deception, forgery, corruption, extortion, theft, misappropriation, false representation, conspiracy, false representation and concealment of material facts.

Zhou W and Kapoor G (2011) describe fraud as an act committed with an intention of financial gain and involves financial transactions of any kind. Reurink, A. (2016) describes financial fraud as "acts and statements through which financial market participants misinform or mislead other participants in the market by deliberately or recklessly providing them with false, incomplete, or manipulative information related to financial goods, services, or investment opportunities in a way that violates any kind of legal rule or law, be it a regulatory rule, statutory law, civil law, or criminal law." (p. 11)

How does it impact companies?

Albashrawi, M. (2016) states that financial fraud has been a huge concern for many companies in all industries alike and in all countries since it can hugely devastate a business. The author further explains that financial industry is more prone to this kind of fraud given the nature of its business and there have been many instances of billions of dollars being lost yearly as a result of financial fraud. Dikmen and Küçükkocaoğlu (2010) site an example in their paper, of Bank of America, that agrees to pay $16.5 billion in order to resolve financial fraud case.CNBC.com states that according to a new report from Javelin Strategy and Research, around 15.4 million consumers had become victims of fraud or identity theft in the year 2017. This is 16% up from the previous year. According to Glancy et al. (2011) "Financial fraud is an issue that has wide-reaching consequences in both the finance industry and daily life. Fraud can reduce confidence in industry, destabilize economies, and affect people's cost of living." (p. 1)

Types of Financial Fraud.

Wells, B. J. (2001, December 1) categorizes financial fraud broadly into four types: Embezzlement, internal fraud, payoffs and kickbacks and skimming. The author explains that embezzlement is the fraud that happens when the person controlling the funds uses them illegally.

Further, he describes internal theft as the type of fraud where the company's employee steels the company of its assets and payoffs and kickbacks as the kinds of bribery where an employee makes gains by trading the company's internal information. Lastly skimming as the fraud where the employee fails to record the money from receipts in the company's financial books and uses it for personal use.

Depending on the context and nature of fraud, Glancy and Yadav (2011) in their paper identified fraud as financial statement fraud, transaction fraud, insurance fraud or credit card fraud. Here we will focus on credit card fraud. According to Masoumeh, Z. (2013) "Credit card fraud can be defined as the illegal use of any system or, criminal activity through the use of physical card or card information without the knowledge of the cardholder." (p.1)

Bhattacharyya et.al (2011); Kancherla R, Venkata R and Verma A. (2008) further divided credit card fraud into application fraud and behavioral fraud. Application fraud is the fraud that occurs when an individual obtains a credit card using false information. Behavioral fraud is when, the details of a legitimate card fall in the hands of a wrong person, mostly fraudulently and the card is used by the wrong person for his personal gains during the absence of the rightful card owner. Kancherla et al. (2008) assert that most credit card fraud is noted to be behavioral fraud.

Is there enough literature on credit card fraud detection methods?

Ngai et al. (2011) mention that even though credit card fraud has been a matter of big concern for many companies across various countries, surprisingly, there has not been a considerable amount of research in this field. This can probably be attributed to the lack of data for such research, due to consumer data privacy and confidentiality reasons. Albashrawi, M. (2016) asserts that "Although detecting financial fraud is considered a high priority for many organizations, the current literature lacks for an up-to-date, comprehensive and in-depth review

that can help firms with their decisions of selecting the appropriate data mining technique" (p.2). Glancy and Yadav. (2011) state "While academic fraud research has examined many business areas, very little effort has been made to use quantitative approaches to examine textual data for automated financial reporting fraud detection. Phua et al. concluded that the use of unstructured data in fraud detection is essentially unexplored." (p.1). Ngai et al. (2011) state "Although financial fraud detection (FFD) is an emerging topic of great importance, a comprehensive literature review of the subject has yet to be carried out." (p.1). Boltan and Hand (2002) also noted a lack of concise published works on credit card fraud detection. Masoumeh (2013) also expresses a dearth of a concise system that can surely predict a transaction to be fraudulent instead of just predicting the likelihood of any transaction to be a fraudulent one.

Ways of detecting fraud.

Zareapoor M, Seeja.K. R, and Alam M.A (2012) mention the following properties for any system to be good at fraud detection. It should have a high enough prediction accuracy, the time taken to detect a fraudulent transaction should be less, i.e. the system should be able to detect the fraud quickly and the number of false negative transactions should be very low, i.e. the system should not be classifying a genuine transaction to be a fraudulent one, in order to maintain customer loyalty and trust. Albashrawi, M. (2016); Bhattacharyya et.al (2011) have established that some of the popular choices of ways in detecting fraudulent transactions are, Logistic Regression, Neural Networks, Bayesian Network, Fuzzy Logic, Decision Trees, Support Vector Machines(SVM), Hidden Markov Model, Artificial Immune among which the popular ones are discussed below.

Logistic Regression. Albashrawi, M. (2016) conclude logistic regression to be the most preferred method for detecting any kind of fraud, possibly because it was one of the earliest

understood methods and is relatively easy to use. This method is used when the dependent variable is categorical in nature. Albashrawi, M. (2016) further states that “Logistic model can help in detecting financial fraud in automobile insurance, corporate insurance, financial statement, and credit card but it can be considered the best-performing method in the context of corporate insurance fraud”. (p.2)

Neural Networks (NN). Neural networks are the next most popular detection method (Albashrawi, M.,2016). The way neural networks work is, they are able to make predictions of events or values depending on the patterns it was learned. Zareapoor et al. (2012) claim that neural networks improve results with time as the model learns from the past and implements it the next time but they are not good with all kinds of data. Zareapoor et al (2012) claim “NNs can produce the best result for only large transaction dataset. And they need a long training dataset.” (p.1)

Bayesian Network (BN). According to Zareapoor et al (2012) “Bayesian networks are very effective for modeling situations where some information is already known and incoming data is unsure or partially unavailable” (p.2). The model works by constructing two Bayesian networks with a hypothesis for describing a user’s behavior. By using expert knowledge to set the fraud net, the first network is constructed to model the user’s behavior assuming the transaction is fraudulent and the second network models the behavior assuming the transaction is legitimate. Finally, the probability of fraud that was obtained from the training is used to set the alarm level. The drawback with the Bayesian network is they tend to be slow on testing datasets and take longer time. Zareapoor et al. (2012) express “BN is more accurate and much faster than neural network [57], but BBNs are slower when applied to new instances.” (p. 2)

Fuzzy Logic. Zareapoor et al (2012) further state that fuzzy logic is very good in dealing with very large and uncertain data but can be very expensive. In this process the data is

preprocessed from SQL database, extracted and distributed into three sections, training, prediction and detection.

Decision Trees. Bhattacharyya et.al (2011) explain that this method constructs a tree-like model based on certain classifiers. The tree keeps splitting depending on binary classifiers until it cannot be split any further and then it is pruned to increase accuracy and handle overfitting. Bhattacharyya et.al (2011) mention it has good interpretability, ease of use, is very fast and can handle various data attributes, but as the number of branches keeps growing, the model is susceptible to overfitting hence the accuracy decreases.

Support Vector Machines(SVM). Support vector machines are another popular method (Albashrawi, M.,2016). It is very similar to neural networks and works in a similar way except that it constructs a hyperplane for a decision plane which maximizes the distance between the positive and the negative nodes. Zareapoor et. Al (2012) mention that SVM shows very good prediction performance but is not as good at dealing with large data. According to Zareapoor et. Al (2012) "Performance evaluation of SVM with BPN in credit card fraud detection shows that when the data number is small, SVM can have better prediction performance than BPN in predicting the future data. But in large data, BPN has a good performance" (p.3)

Types of methods used to statistically detect fraud.

Bolton R.J, Hand D.J (2001) broadly divide statistical fraud detection methods into two categories, supervised and unsupervised. In supervised fraud detection methods, models are constructed using training data set with the prior knowledge of the existing data and this model is tested on the new data and the transactions are classified as either legitimate or fraudulent.

Unsupervised learning methods do not train the data set with any prior known information. This is an area which still needs to be researched and understood much as sometimes data in real life situations might not have historical information to fall back on. Unsupervised learning method works by modeling the underlying distribution or structure in the data. Bolton R.J, Hand D.J (2001) further categorize supervised learning methods as "classification" type-when the dependent variable is categorical and "regression" type- when the dependent variable is numeric types and unsupervised methods as "Clustering" types– when the objective is to understand the inherent grouping in the data and "association" type- when the objective is to discover underlying rules or associations. As Bolton and Hand (2001) state "Statistical methods for fraud detection are often classification (supervised) methods that discriminate between known fraudulent and non-fraudulent transactions; however, these methods rely on accurate identification of fraudulent transactions in historical databases – information that is often in short supply or non-existent." (p. 1)

What is data mining?

Albashrawi, M. (2016) defines data mining as "a process that uses statistical, mathematical, artificial intelligence, and machine learning techniques to extract and identify useful information and subsequently gain knowledge from a large database." (p. 2). It is an automated analytical method that combines computer science and analytics with an objective to discover hidden insights and patterns in large datasets and involves machine learning which automates the process and expedites the process of data exploration and analysis.

How have credit card fraud detection techniques evolved over time?

West et.al (2014) have well documented the evolution of various data mining techniques, starting with early forms of neural networks and game theory in the late 1990's and early 2000's then progressing to logistic regression and decision trees around 2007 and subsequently to Bayesian network and more hybrid models like Hidden Markov models and artificial immune models. Fig 1 on page no. 22 shows the chronological evolution of fraud detection techniques as depicted by West et al. (2014, p.4). Bhattacharyya et al. (2011) explain that though data mining techniques have greatly evolved over the past two decades, with newer models like Hidden Markov models discovered recently, the techniques used for detection of credit card fraud, however, have not evolved much. Bhattacharyya et al. (2011) put it as "Considering the profusion of data mining techniques and applications in recent years, however, there have been relatively few reported studies of data mining for credit card fraud detection" (p.1). One possible explanation for this could be attributed to the lack of data for such research, due to consumer data privacy and confidentiality reasons.

Which is the most implemented method?

Albashrawi, M. (2016) shows that logistic regression has been by far the most implemented method for any kind of fraud detection. This is possibly attributed to the fact that it is one of the early discovered, widely understood and researched method. Albashrawi, M. (2016, p.10) has detailed out the most implemented or widely used data mining techniques and states that logistic regression followed by neural networks, decision trees, support vector machines, naïve bayes and Bayesian networks are the most widely used techniques in the same order as mentioned above. Fig.2 in pg.23 is an extract from the table illustrated by Albashrawi (2016, p.10)

One point to note here is random forest method has not been implemented much (Albashrawi, M. 2016; West et.al.,2014). Random forest is an ensemble model that builds on decision tree model, and as a result of this feature, it is able to take care of all the issues that decision tree is not very good at handling (Altendorf et al.,2009).

Do some methods perform better than other methods?

Depending on the objectives and the nature of the data, the methods for prediction should be selected. Below are the summarized pros and cons of each method and depending on the specific nature of the problem one is working on, the type and nature of the data and the amount of information and resources available, one should choose the method to implement. Albashrawi (2016) noted logistic regression to be the most widely applied method. But there are some drawbacks to logistic regression. Logistic regression does not take into account variable interactions or any nonlinear features in the data and cannot handle unbalanced data as well as other models (Zareapoor et al.,2012).

Zareapoor et al (2012) compare different types of fraud detecting algorithms and explain that neural networks can produce the best result only with very large datasets and they need a long training dataset and that there are other models like the bayesian network that show higher accuracy. They further explain that the drawback of the bayesian network is, they tend to be slow on testing datasets and take longer time even though they show better accuracy than neural networks.

Further comparing the other models, Zareapoor et al (2012) state that fuzzy logic is very good at dealing with very large and uncertain data but can be very expensive and that the decision tree has good interpretability, ease of use, is very fast and can handle various data attributes, but as the number of branches keeps growing, the model is susceptible to overfitting hence the

accuracy decreases. They further state that SVM shows very good prediction performance but is not very good at dealing with large data.

When is using random forests better than any other method?

According to Ruiz and Villa (2008) when a greater rate of False alarm is allowed, then random forests are better compared to other models. Ruiz and Villa (2008) put it as "Focusing on the comparison between random forest and logistic regression, things are not as easy as the comparison of probability densities (Figure 9) could leave to believe. Actually, the optimal model strongly depends on the objectives: for a low FAR (between 15- 20%), the optimal model (that has the highest TS) is logistic regression. But if higher FAR is allowed (20- 30%), then random forest should be used for prediction". (p.8)

Random forest edge.

Random forest is a relatively recently developed algorithm which is used for classification and regression (West et.al.,2014). (Albashrawi M, 2016; Altendorf et al.,2009) describe random forest as an ensemble model, which has been built upon decision tree model where the output is based on a majority vote among the tree classifiers and each tree is trained by randomly sampling each subset. With this, a decision tree is built, but unlike in decision trees, in the random forest, the trees are not pruned. Random forest model can be trained very fast it can take into account many features as each tree is randomly and independently grown and is resistant to overfitting, unlike decision trees. It is more robust to any noise in the data and unbalanced data can be easily handled too. Altendorf et.al (2009) state, "Training is fast, even for large datasets with many features and data instances, because each tree is trained independently of the others. The Random Forest algorithm has been found to be resistant to overfitting and provides a good estimate of the

generalization error (without having to do cross-validation) through the “out-of-bag” error rate that it returns. Our data sets are quite imbalanced, which in general, can lead to problems during the learning process” (p. 3)

Breiman and Leo (2001) “Random forests are an effective tool in prediction. Because of the Law of Large Numbers, they do not overfit. Injecting the right kind of randomness makes them accurate classifiers and regressors.” (p. 25)

West et.al (2014) state that random forests produce results with high accuracy compared to the more popular and widely implemented models like Logistic Regression and Support Vector Machines (SVM). West et.al (2014) described a table showing the accuracy results of various fraud detection practices. Below is an extract of the table marked as Fig 3 on pg.24. From the table we can see, random forests projected a higher accuracy rate compared to the other two models.

Random forests individually have been researched thoroughly. However, further study of the use of random forests in credit card fraud detection will lead to a better understanding in choosing the right machine learning algorithm, one that is inexpensive too. Specifically, in understanding, if random forests can better detect credit card fraud, without overfitting when using large unbalanced datasets. Therefore, for my research, I hypothesize that random forest algorithm project a greater accuracy and therefore can detect credit card fraud better in large, unbalanced datasets without overfitting.

Methods

Participants and procedure

For our research, we shall use a publicly available dataset. The dataset we shall use was collected and analyzed by a machine learning group during a research collaboration of Worldline

(Bontempi G, Caelen O, Pozzolo D L and Reid A J.,2015). The authors have made the data available to the public. Bontempi G et al. (2015) "The dataset is available at http://www.ulb.ac.be/di/map/adalpozz/data/ credit card.Rdata" (p.7).This dataset contains records of transactions made by credit cards holders in September 2013. The transactions were made in an undisclosed location, somewhere in Europe. The records of the data present transactions that occurred in two days and a total of 492 frauds out of 284,807 transactions were identified. In order to preserve customer privacy and confidentiality, the names and other details of the customers are seen to be anonymized.

Measures

1. **Highly unbalanced data.**

A data is said to be highly unbalanced when the class of the dependent variable is not evenly distributed,i.e, one of the classes from the binary classification of the dependent variable outnumbers the other class (Bontempi G et al.,2015). (Bontempi G et al.,2015) further, explain that this can lead to the wrong precision and accuracy results as the machine learning algorithm tends to classify all the observations as a majority class. Bontempi G et al. (2015) state that "This translates into poor accuracy on the minority class (low recall), which is typically the class of interest". (p.2)

Our data also is highly unbalanced, as the number of legitimate transactions is way more than the number of fraudulent transactions. The positive class (frauds) account for only 0.172% of all transactions. In such cases, the sampling bias is handled by using one of the three techniques, which are, under-sampling the majority class, over-sampling the minority class, and the SMOTE technique (Bontempi G et al.,2015). (Albashrawi, 2016) mention that random

forest algorithm can handle large unbalanced data unlike models like logistic regressions. Hence, here, the unbalanced feature of our dataset will be left untreated.

2. Principal component transformation(PCA)

It can be seen from the data set, that due to consumer privacy and confidentiality reasons, the details of the variables are not available and the data contains only the variables with numeric values as input. This is possibly the result of PCA transformation. PCA transformation is done on a highly dimensional data to reduce its complexity but preserve the essential crux of the data (Jolliffe, I. T., & Cadima, J.,2016). In our data, only features "Time" and "Amount" have been left untransformed.

3. Tools

For the research, we shall use "R" which is a freely available statistical analysis tool and use the random forests package available in R.

4. Variable being measured.

We can see that the feature "class" takes values of 0 and 1. 0 is the case when any transaction has been classified as a legitimate one and 1 is the case when any transaction has been classified as a fraudulent one. This is our response variable or dependent variable. We shall use the variables "V1", "V2", ... "V28" as the independent variables as they are the principal components obtained with PCA.

Analysis

In data mining, many techniques, using various metrics can be used to evaluate the performance of the binary classifier and the metric used greatly influences the performance of the

classifier (Hossin, M, and Sulaiman, M.N.,2015). They put it as, “Evaluation metric plays a critical role in achieving the optimal classifier during the classification training. Thus, a selection of suitable evaluation metric is an important key for discriminating and obtaining the optimal classifier” (p.1). They further mention that most of the evaluation metrics employ, accuracy or the error rate to evaluate the performance of the classifier. They put it as, “Typically, most of the PS classifiers employ the accuracy or the error rate (1-accuracy) to discriminate and to select the best (optimal) solution. However, using the accuracy as a benchmark measurement has a number of limitations”. (p.2). They explain clearly the situations where the above-mentioned metric should not be used to evaluate the performance of the classifier, one being, when the data is highly imbalanced, i.e. when the class of the dependent variable is not evenly distributed, as in the case of our data. They state that “Similar to accuracy, the main limitation of MSE is this metric does not provide the trade-off information between class data. This may lead the discrimination process to select the suboptimal solution. Moreover, this metric is really dependent on the weight initialization process. In the extremely imbalanced class problem, if the initial weights are not proper selected (i.e. no initial weight to represent the minority class data), this may lead the discrimination process ends up with the sub-optimal solution due to lack information of minority class data although the MSE value is minimized (under-fitting or over-fitting)”. (p.5).

The area under the ROC Curve (AUC) metric, is insensitive to imbalanced data (Fawcett, T.,2005). (Hossin, M & Sulaiman, M.N.,2015) describe AUC as “AUC is one of the popular ranking type metrics. In [13, 17, 31] the AUC was used to construct an optimized learning model and also for comparing learning algorithms [28,29]. Unlike the threshold and probability metrics, the AUC value reflects the overall ranking performance of a classifier”. (p.5)

AUC ROC is robust to the highly imbalanced feature of the classes in datasets with

binary classification problems. Also, AUC is only sensitive to rank ordering and ROC curves can be extended to problems with three or more classes with what is called one versus all approach. (Fawcett, T.,2005)

Fawcett, T.,(2005) explain that AUC can be thought of as representing the probability that a classifier will rank a randomly chosen positive observation higher than a randomly chosen negative observation, and thus it is a useful metric for datasets even when the classes are highly unbalanced.

In real life situations, we see that most binary classification problems do not have balanced classes and the probability distribution of any classifier will not follow any particular distribution. Since our data is highly imbalanced, we decide to use the area under the ROC Curve (AUC) as the metric to evaluate the performance of the binary classifier, which in our case is random forests.

Results

In our thesis, since we are hypothesizing that Random forests is a better binary classification algorithm and is able to handle large unbalanced data better, there has to be some other classifier that the random forests classifier could be compared with, so as to be able to logically analyze our hypothesis. Since Logistic regression is the most popular and frequently used classifier (Peng, Lee & Ingersoll 2002), we build a model using logistic regression to serve as a base model so that we can, later on, compare the performance of the random forests model to the logistic regression model and see if the model using random forests algorithm performs better or not.

To start with, we first perform EDA on our dataset, to know the underlying structure and trends, that can’t is noted otherwise. Along with checking for missing values, normalizing the

variable “Amount”, a correlation among the variables was checked. Figure 4, shows the summary of the dataset. The process of EDA revealed some very interesting observations, out of which two are worth mentioning here. The first observation was that the response variable is highly imbalanced in nature. The number of legitimate transactions hugely outnumber the no of fraudulent transactions, which makes good sense. Secondly, we were also able to see some interesting correlations between the various variables. Figure 5 and Figure 6, below show the imbalanced nature of the dataset and the correlations among the variables. The variables “Amount” and “Class” were shown to be highly correlated. The dataset was divided into training and testing dataset with 75% of the observations being allocated to the training set, and the remaining to the test set.

The next step was to build a logistic regression model, that would serve as the base model for comparison purposes. Below is the regression model. The dataset was divided into training and testing dataset with 75% of the observations being allocated to the training set, and the remaining to the test set. The model was then fit to the testing dataset to see how the model was behaving with a dataset that had not been used to train, and a confusion matrix and a fourfold plot was plotted to better understand the predictions made. Figure 7, below shows the summary of the logistic regression model. Figure 8, shows the summary of the confusion matrix of the logistic regression model. Figure 9 shows the fourfold plot of the predictions of the logistic regression model. We can see from the fourfold plot, that the total False Positive and False negative cases are very low, 46 and 9 respectively, but they don’t give us a very good estimation of the accuracy. Because we fit a logistic regression with maximum likelihood estimation, the deviance residuals as seen in the summary of the logistic regression model do not make much sense and hence to check the accuracy of the model, we use the model to produce the ROC and then use the area under the curve of the ROC. Figure 10, below shows the Area Under the ROC Curve of the Logistic regression model.

As we can see below in the AUC of ROC plot, the model seemed to be very high in its prediction power, with as AUC measure of 0.982, which means that the model has a prediction power of a little over 98%, which is a very high measure.

The next step was to proceed with the same process of model building, but this time using the random forests algorithm, to compare and see how it performs against the logistic regression model. Figure 11, shows the summary of the Random Forests model. After building the model, the model was checked for accuracy using a confusion matrix and a fourfold plot to visually understand the predictions better. Figure 12, shows the confusion matrix using random forests model. Figure 13, shows the fourfold plot using the random forests algorithm and we can see, the random forest fourfold plot is identical to the logistic regression plot. Figure 14, shows the ROC curve drawn on the random forest model's predictions. From the AUC ROC score, we can see that random forests model performed very slightly better than the logistic regression model.

Lastly, on plotting a variable importance plot of the random forest model, we see that the three top most variables used in the model building, were V17, V12, V14 V10. Below is the Variable importance plot. This information could be useful in feature selection in the future which could enhance the model performance. Figure 15, shows variable importance plot of the random forest model

Discussions

A closer look at the data shows that all the anonymized variables were normalized with mean 0. The same transformation was applied to the variable "amount" so that it can fecilitate in building machine learning models. The get a better understanding of the data, the mean and median values of the fraudulent and nonfraudulent transactions were compared. This gave us an

understanding that the median of the legitimate transactions was higher, which was also confirmed with the boxplot of the legitimate and fraudulent transactions.

Logistic regression Vs Random forest model.

Logistic regression is a classification model that estimates the probability using the maximum likelihood estimation method, which is a step-by-step method, where non relevant variables are dropped from the model while building it and the more relevant ones are added leading to a final model that is built on minimal variables that are most relevant to the model. (Ruiz, A., & Villa, N. 2008).From the model summary of the logistic model shown in pg, we can see that the variables with three stars are the ones that are most significant hence more relevant too. This could be a very important information, in feature selection, which can be performed to enhance the predictability and accuracy of the model. Random forest is a classification or a regression tool that can be used with nonlinear data too and works like decision tree but random forests are an improvement to decision trees as they use the bagging technique to aggregate many under-efficient classification trees and a majority vote law is used on all the classification trees to reach the final classification decision .(Ruiz, A., & Villa, N. 2008). One point to note here is both the models have not been cross validated, the accuracy measure of both the models is likely to change after cross validation, and a model that has been cross validated is considered to be higher than a model that has not been cross validated. But since, the Random Forest method, uses bagging technique to aggregate many under-efficient classification trees and a majority vote law is used on all the classification trees to reach the final classification decision, it is very close to cross validation, and the accuracy of the random forests' model after cross validation is not likely to change as much as the logistic regressions model. Hence, the accuracy of the non-cross validated

random forest model can be seen as more reliable compared to the non-cross validated Logistic Regression model.

Comparing the two models, logistic regression is a much faster model with less computational time, but the accuracy of the two models is almost the same. Both the models could be further enhanced by performing feature selection and thereby using the most significant variables.

Limitations

There were a few limitations in the study and the biggest of which was not having enough time in hand to carry further enhancement of the model with methods like cross validation. Since the data is an anonymized data, there is very little knowledge of the description of the variables. Also, all the anonymized features seem to have been normalized with mean 0, which has been left as is which might not be the best way to normalize the variables. Further, no feature selection was done, which could have enhanced the performance of the models by using the most significant variables.

Conclusion

Both the models performed almost identically with random forest's AUC being only very slightly higher, though the confusion matrix and the fourfold plots for both the models look identical. Further research should follow to know how the two algorithms perform with cross validation and handle much larger dataset and nonlinear parameters if present. The calculated accuracy is not very relevant in the conditions where there is a very large unbalance between the number of fraud and non-fraud events in the dataset. In such cases, we can see a very large

accuracy. More relevant is the value of ROC-AUC (Area Under Curve for the Receiver Operator Characteristic). The value obtained (0.983) is relatively good, considering that we left the unbalanced nature of the dataset as is. To get a better performing model, the dataset's unbalanced nature should be handled better with some technique like the SMOTE.

Future work

This research can be carried forward by using cross validation methods and using appropriate feature selection techniques to see how it can affect the model's predictability and thereby enabling us to understand which machine learning algorithms perform better under given specific circumstances.

References

Albashrawi, M. (2016). Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015. Journal of Data Science, 14(3), 553-569.

Altendorf, E., Brende, P., Daniel, J., & Lessard, L. (2005). Fraud detection for online retail using random forests. Technical report. Stanford University.

Bai B, Yen J, Yang X. False Financial Statements: Characteristics of China's Listed Companies and CART Detecting Approach. International Journal of Information Technology & Decision Making 2008; 7: 339-359.

Bermúdez L, Pérez J, Ayuso M, Gómez E, Vázquez F. A Bayesian Dichotomous Model with Asymmetric Link for Fraud in Insurance. Insurance: Mathematics and Economics 2008; 42: 779-786

Bhattacharyya S, Jha S, Tharakunnel K, Westland, JC. Data Mining for Credit Card Fraud: A Comparative Study. Decision Support Systems 2011; 50: 602-613.

Bidder OR, Campbell HA, Gómez-Laich A, Urgé P, Walker J, Cai Y, Wilson RP. Love Thy Neighbour: Automatic Animal Behavioural Classification of Acceleration Data Using the K-Nearest Neighbour Algorithm. PLoS ONE 2014; 9: 1-7

Bolton R.J, Hand D.J (2001), Unsupervised profiling methods for fraud detection In: Proceedings of conference credit scoring and credit control VII, pp 5–7

Bontempi G, Caelen O, Pozzolo D.L, and Reid A.J. Calibrating Probability with Undersampling for Unbalanced Classification. In Symposium on Computational Intelligence and Data Mining (CIDM), IEEE, 2015

Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1), 5-32.

Caudill S, Ayuso M, Guill'en M. Fraud Detection Using A Multinomial Logit Model with Missing Information. The Journal of Risk and Insurance 2005; 72: 539-550

Caudill S, Ayuso M, Guill'en M. Fraud Detection Using A Multinomial Logit Model with Missing Information. The Journal of Risk and Insurance 2005; 72: 539-550.

Chan P.K., Fan W, Prodromidis A.L, Stolfo S.J, Distributed Data Mining in Credit Card Fraud Detection, Data Mining, (November/December), 1999, pp. 67-74.

Chen R.C, Chen T.S, Lin C.C, A new binary support vector system for increasing detection rate of credit card fraud, International Journal of Pattern Recognition 20 (2) (2006) 227-239.

Dechow P, Ge W, Larson C, Sloan R. Predicting Material Accounting Misstatements. Contemporary Accounting Research 2011; 28: 1-16. [13] Deng Q, Mei G. Combining Self-Organizing Map and K-Means Clustering for Detecting Fraudulent Financial Statements. In IEEE International Conference on Granular Computing 2009; 126-131.

Dharwa JN, Patel AR. A Data Mining with Hybrid Approach Based Transaction Risk Score Generation Model (TRSGM) for Fraud Detection of Online Financial Transaction. International Journal of Computer Applications 2011; 16: 18-25.

Dikmen B, Küçükkocaoğlu G. The Detection of Earnings Manipulation: The Three-Phase Cutting Plane Algorithm Using Mathematical Programming. Journal of Forecasting 2010; 29: 442-466

Fawcett, T. (2005). An introduction to ROC analysis. Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, CA 94306, USA Available online 19 December 2005

Glancy FH, Yadav SB. A Computational Model for Financial Reporting Fraud Detection. Decision Support Systems 2011; 50: 595-601

Glancy FH, Yadav SB. A Computational Model for Financial Reporting Fraud Detection. Decision Support Systems 2011; 50: 595-601.

Grant, K. B (2017) Identity theft, fraud cost consumers more than $16 billion. Retrieved from https://www.cnbc.com/2017/02/01/consumers-lost-more-than-16b-to-fraud-and-identity-theft-last-year.html

Hossin, M & Sulaiman, M.N. (2015). A review on evaluation metrics for data classification evaluations. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.2, March 2015.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences*, *374*(2065), 20150202. http://doi.org/10.1098/rsta.2015.0202

Kancherla R, Venkata R, Verma A. (February 2008). Behavioral Fraud Mitigation through Trend Offsets, Genpact India, (2008)

Ngai E, Hu Y, Wong Y, Chen Y, Sun X. The Application of Data Mining Techniques in Financial Fraud Detection: A Classification Framework and an Academic Review of Literature. Decision Support Systems 2011; 50: 559-569

Peng, C. J., Lee, K. L., & Ingersoll, G. M. (2002). An Introduction to Logistic Regression Analysis and Reporting. Journal of Educational Research, 96(1), 3.

Ravisankar, P., Ravi, V., Raghava RAO, G., & Bose, I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. Decision Support Systems, (2), 491.

Reurink, A. (2016). Financial fraud: A literature review. Discussion Paper 16/5. May, Max Planck Institute for the Study of Societies, Cologne.

Ruiz, A., & Villa, N. (2008). Storms prediction: Logistic regression vs random forest for unbalanced data. Case Studies in Business, Industry and Government Statistics, 2007, 1 (2), pp.91-101.

Vasiu, L., & Warren, M. & Mackay, D. (2003). Defining Fraud: Issues for Organizations from an Information Systems Perspective. 7th Pacific Asia Conference on Information Systems, 10-13 July 2003, Adelaide, South Australia

Wells, B. J. (2001, December 1). Enemies Within Asset misappropriation comes in many forms. Retrieved from Journal of Accountancy:

https://www.journalofaccountancy.com/issues/2001/dec/enemieswithin.html

West J, Bhattacharya M, Islam R "Intelligent Financial Fraud Detection Practices: An Investigation", Computers & Security, 57(3), pp. 47-66. 2014

Williams K, The Evolution of Credit Card Fraud: Staying Ahead of the Curve, eFunds Corporation, 2007

Zareapoor M, Seeja.K. R, and Alam M.A, “Analysis of Credit Card Fraud Detection Techniques: based on Certain Design Criteria”, International Journal of Computer Applications (0975 – 8887) Volume 52– No.3, August 2012

Zhou W and Kapoor G (2011) Detecting evolutionary financial statement fraud. Decision Support Systems 50, 570-5

Figures

Chronological document of the evolution of fraud detection techniques.

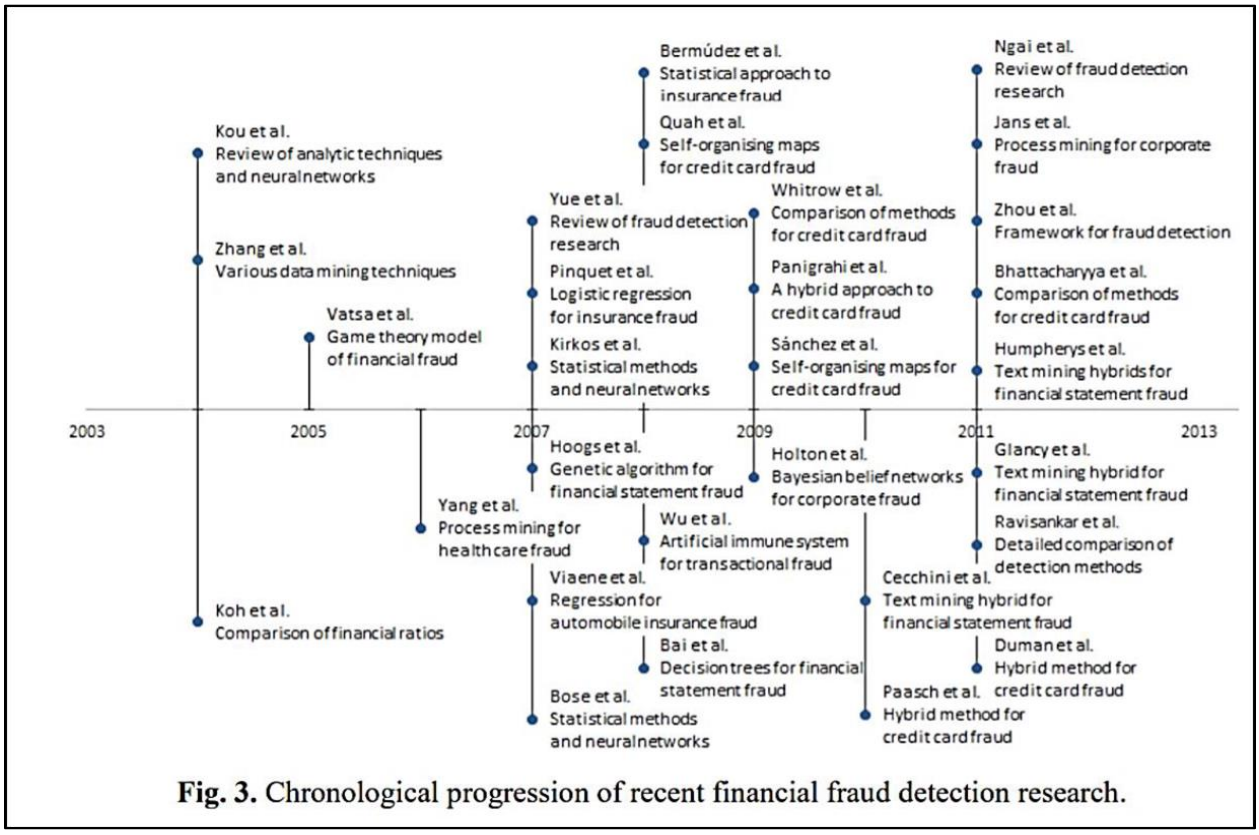


Fig. 3. Chronological progression of recent financial fraud detection research.

Figure 1. is extracted from West J, Bhattacharya M, Islam R "Intelligent Financial Fraud Detection Practices: An Investigation" 2014(p.4)

Six most popular data mining techniques widely implemented in detecting any nature of prediction.

No.	Method	Frequency	Description	Business Application
1	Logistic regression	17	It is a typical classification method used to generate dichotomous possible values [59].	Prediction of failure probability in selling a specific product
2	Neural network	15	ANN shows better results when testing large sets of data. It consists of neurons or nodes [43].	Credit Rating
3	Decision trees	15	Decision tree or classification tree is a method for assigning and classifying data points into predefined clusters via splitting rules [20].	Stock market prediction
4	Support vector machine	12	SVM is a statistical method that used for linear classification [4].	Bankruptcy prediction
5	Naïve Bayes	8	This tool has the capability of predicting group membership [26].	Sentiment analysis
6	Bayesian	7	"Directed acyclic graph, used to predict the	Tracking performance over

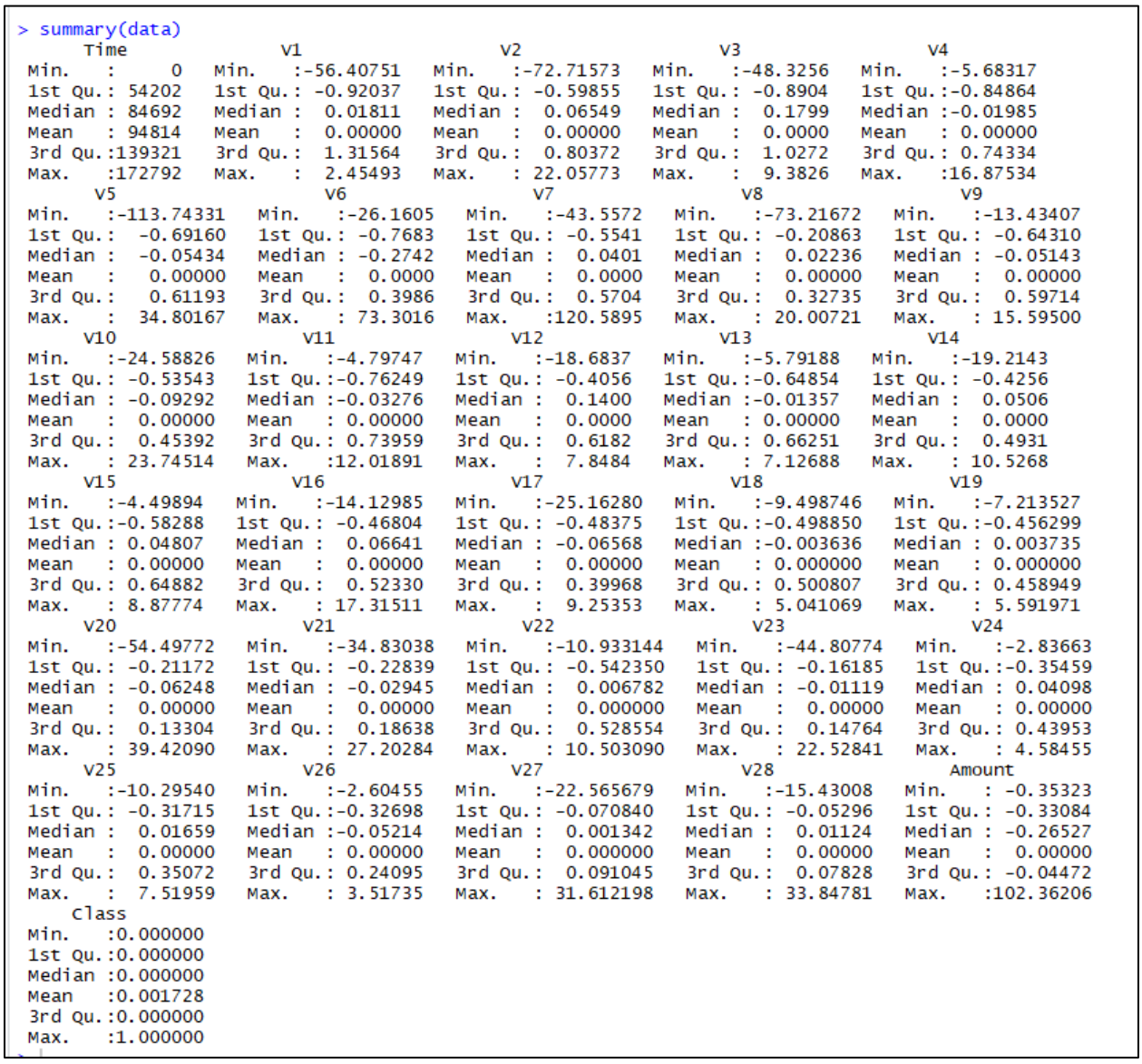
Figure 2. is extracted from Albashrawi, M. (2016). Detecting Financial Fraud Using Data Mining Techniques: A Decade Review from 2004 to 2015. (p.10)

Accuracy results of various fraud detection practices.

Research	Fraud Investigated	Method Investigated	Accuracy
[3]	Credit card transaction fraud from a real world example	Logistic model (regression)	96.6-99.4%
		Support vector machines	95.5-99.6%
		Random forests	97.8-99.6%

Figure 3. is extracted from West J, Bhattacharya M, Islam R "Intelligent Financial Fraud Detection Practices: An Investigation" 2014 (p. 5)

Figure 4



```
> summary(data)
      Time              V1                   V2                   V3                  V4
 Min.   :     0   Min.   :-56.40751   Min.   :-72.71573   Min.   :-48.3256   Min.   :-5.68317
 1st Qu.: 54202   1st Qu.: -0.92037   1st Qu.: -0.59855   1st Qu.: -0.8904   1st Qu.:-0.84864
 Median : 84692   Median :  0.01811   Median :  0.06549   Median :  0.1799   Median :-0.01985
 Mean   : 94814   Mean   :  0.00000   Mean   :  0.00000   Mean   :  0.0000   Mean   : 0.00000
 3rd Qu.:139321   3rd Qu.:  1.31564   3rd Qu.:  0.80372   3rd Qu.:  1.0272   3rd Qu.: 0.74334
 Max.   :172792   Max.   :  2.45493   Max.   : 22.05773   Max.   :  9.3826   Max.   :16.87534
       V5                  V6                   V7                   V8                  V9
 Min.   :-113.74331   Min.   :-26.1605   Min.   :-43.5572   Min.   :-73.21672   Min.   :-13.43407
 1st Qu.:  -0.69160   1st Qu.: -0.7683   1st Qu.: -0.5541   1st Qu.: -0.20863   1st Qu.: -0.64310
 Median :  -0.05434   Median : -0.2742   Median :  0.0401   Median :  0.02236   Median : -0.05143
 Mean   :   0.00000   Mean   :  0.0000   Mean   :  0.0000   Mean   :  0.00000   Mean   :  0.00000
 3rd Qu.:   0.61193   3rd Qu.:  0.3986   3rd Qu.:  0.5704   3rd Qu.:  0.32735   3rd Qu.:  0.59714
 Max.   :  34.80167   Max.   : 73.3016   Max.   :120.5895   Max.   : 20.00721   Max.   : 15.59500
      V10                 V11                 V12                 V13                 V14
 Min.   :-24.58826   Min.   :-4.79747   Min.   :-18.6837   Min.   :-5.79188   Min.   :-19.2143
 1st Qu.: -0.53543   1st Qu.:-0.76249   1st Qu.: -0.4056   1st Qu.:-0.64854   1st Qu.: -0.4256
 Median : -0.09292   Median :-0.03276   Median :  0.1400   Median :-0.01357   Median :  0.0506
 Mean   :  0.00000   Mean   : 0.00000   Mean   :  0.0000   Mean   : 0.00000   Mean   :  0.0000
 3rd Qu.:  0.45392   3rd Qu.: 0.73959   3rd Qu.:  0.6182   3rd Qu.: 0.66251   3rd Qu.:  0.4931
 Max.   : 23.74514   Max.   :12.01891   Max.   :  7.8484   Max.   : 7.12688   Max.   : 10.5268
      V15                 V16                 V17                 V18                 V19
 Min.   :-4.49894   Min.   :-14.12985   Min.   :-25.16280   Min.   :-9.498746   Min.   :-7.213527
 1st Qu.:-0.58288   1st Qu.: -0.46804   1st Qu.: -0.48375   1st Qu.:-0.498850   1st Qu.:-0.456299
 Median : 0.04807   Median :  0.06641   Median : -0.06568   Median :-0.003636   Median : 0.003735
 Mean   : 0.00000   Mean   :  0.00000   Mean   :  0.00000   Mean   : 0.000000   Mean   : 0.000000
 3rd Qu.: 0.64882   3rd Qu.:  0.52330   3rd Qu.:  0.39968   3rd Qu.: 0.500807   3rd Qu.: 0.458949
 Max.   : 8.87774   Max.   : 17.31511   Max.   :  9.25353   Max.   : 5.041069   Max.   : 5.591971
      V20                 V21                  V22                 V23                 V24
 Min.   :-54.49772   Min.   :-34.83038   Min.   :-10.933144   Min.   :-44.80774   Min.   :-2.83663
 1st Qu.: -0.21172   1st Qu.: -0.22839   1st Qu.: -0.542350   1st Qu.: -0.16185   1st Qu.:-0.35459
 Median : -0.06248   Median : -0.02945   Median :  0.006782   Median : -0.01119   Median : 0.04098
 Mean   :  0.00000   Mean   :  0.00000   Mean   :  0.000000   Mean   :  0.00000   Mean   : 0.00000
 3rd Qu.:  0.13304   3rd Qu.:  0.18638   3rd Qu.:  0.528554   3rd Qu.:  0.14764   3rd Qu.: 0.43953
 Max.   : 39.42090   Max.   : 27.20284   Max.   : 10.503090   Max.   : 22.52841   Max.   : 4.58455
      V25                 V26                 V27                 V28                Amount
 Min.   :-10.29540   Min.   :-2.60455   Min.   :-22.565679   Min.   :-15.43008   Min.   : -0.35323
 1st Qu.: -0.31715   1st Qu.:-0.32698   1st Qu.: -0.070840   1st Qu.: -0.05296   1st Qu.: -0.33084
 Median :  0.01659   Median :-0.05214   Median :  0.001342   Median :  0.01124   Median : -0.26527
 Mean   :  0.00000   Mean   : 0.00000   Mean   :  0.000000   Mean   :  0.00000   Mean   :  0.00000
 3rd Qu.:  0.35072   3rd Qu.: 0.24095   3rd Qu.:  0.091045   3rd Qu.:  0.07828   3rd Qu.: -0.04472
 Max.   :  7.51959   Max.   : 3.51735   Max.   : 31.612198   Max.   : 33.84781   Max.   :102.36206
     Class
 Min.   :0.000000
 1st Qu.:0.000000
 Median :0.000000
 Mean   :0.001728
 3rd Qu.:0.000000
 Max.   :1.000000
```

Figure 4, shows the summary of the dataset.

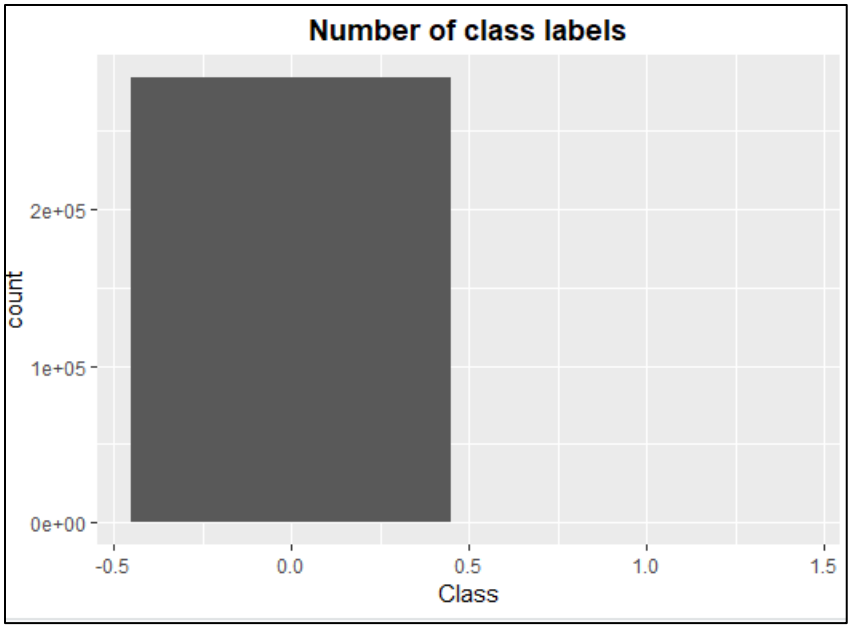


Figure 5, below shows the imbalanced nature of the dataset

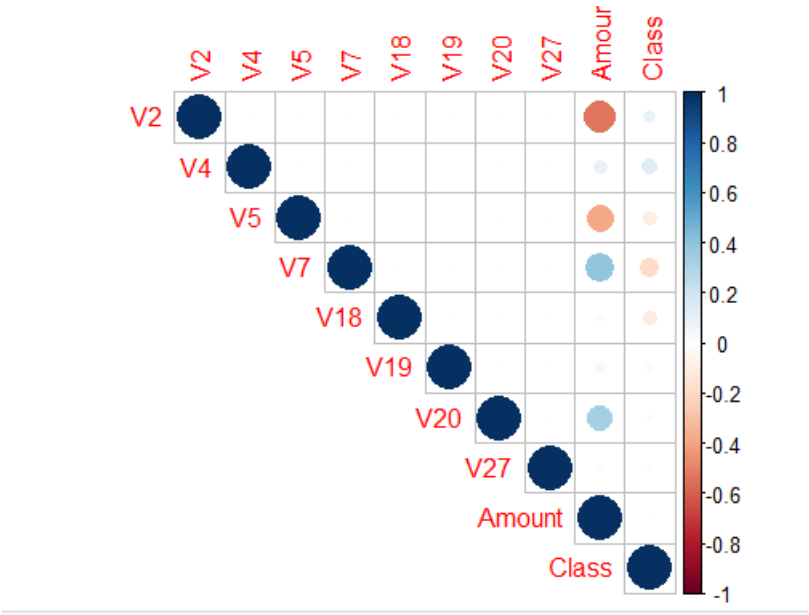
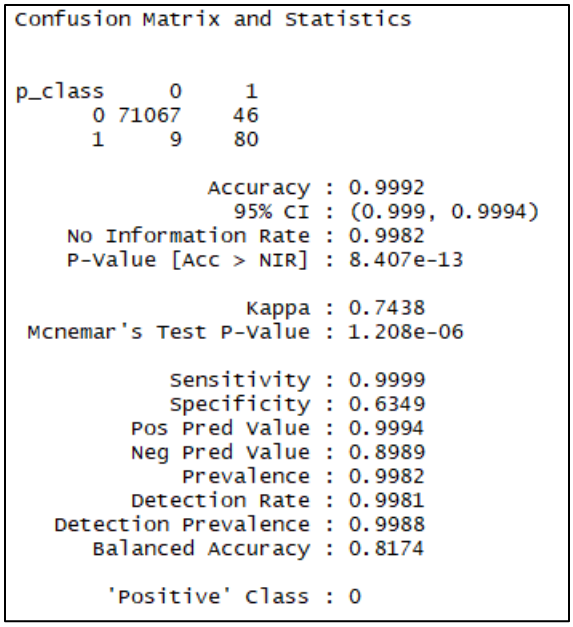


Figure 6, below shows the correlations among the variables.

```
Call:
glm(formula = Class ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9241  -0.0293  -0.0193  -0.0125   4.5231

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.409e+00  2.893e-01 -29.069  < 2e-16 ***
Time        -2.563e-06  2.600e-06  -0.986 0.324152
V1           6.453e-02  5.045e-02   1.279 0.200791
V2           3.270e-02  7.468e-02   0.438 0.661529
V3          -9.279e-04  6.179e-02  -0.015 0.988019
V4           7.053e-01  8.967e-02   7.866 3.67e-15 ***
V5           1.446e-01  7.940e-02   1.821 0.068594 .
V6          -8.096e-02  8.194e-02  -0.988 0.323132
V7          -6.179e-02  8.349e-02  -0.740 0.459268
V8          -1.808e-01  3.649e-02  -4.955 7.25e-07 ***
V9          -1.746e-01  1.335e-01  -1.308 0.190938
V10         -8.256e-01  1.158e-01  -7.129 1.01e-12 ***
V11         -5.250e-02  9.435e-02  -0.556 0.577906
V12          5.350e-02  1.005e-01   0.532 0.594554
V13         -2.645e-01  9.250e-02  -2.860 0.004243 **
V14         -5.440e-01  7.162e-02  -7.596 3.06e-14 ***
V15         -6.686e-02  9.970e-02  -0.671 0.502512
V16         -2.478e-01  1.482e-01  -1.672 0.094460 .
V17          3.396e-03  8.232e-02   0.041 0.967095
V18         -3.039e-02  1.515e-01  -0.201 0.841037
V19          4.837e-02  1.127e-01   0.429 0.667663
V20         -4.245e-01  1.050e-01  -4.043 5.28e-05 ***
V21          3.579e-01  6.989e-02   5.121 3.04e-07 ***
V22          5.641e-01  1.557e-01   3.623 0.000291 ***
V23         -4.041e-02  6.937e-02  -0.583 0.560225
V24          1.124e-01  1.663e-01   0.676 0.499203
V25         -9.667e-03  1.525e-01  -0.063 0.949441
V26         -4.377e-02  2.251e-01  -0.194 0.845814
V27         -7.072e-01  1.573e-01  -4.495 6.96e-06 ***
V28         -2.345e-01  1.006e-01  -2.330 0.019792 *
Amount       2.097e-01  1.295e-01   1.620 0.105199
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5495.4  on 213604  degrees of freedom
Residual deviance: 1669.2  on 213574  degrees of freedom
AIC: 1731.2

Number of Fisher Scoring iterations: 12
```

Figure 7, below shows the summary of the logistic regression model.



```
Confusion Matrix and Statistics


p_class     0     1
      0 71067    46
      1     9    80

               Accuracy : 0.9992
                 95% CI : (0.999, 0.9994)
    No Information Rate : 0.9982
    P-Value [Acc > NIR] : 8.407e-13

                  Kappa : 0.7438
 Mcnemar's Test P-Value : 1.208e-06

            Sensitivity : 0.9999
            Specificity : 0.6349
         Pos Pred Value : 0.9994
         Neg Pred Value : 0.8989
             Prevalence : 0.9982
         Detection Rate : 0.9981
   Detection Prevalence : 0.9988
      Balanced Accuracy : 0.8174

       'Positive' Class : 0
```

Figure 8, shows the summary of the confusion matrix of the logistic regression model.

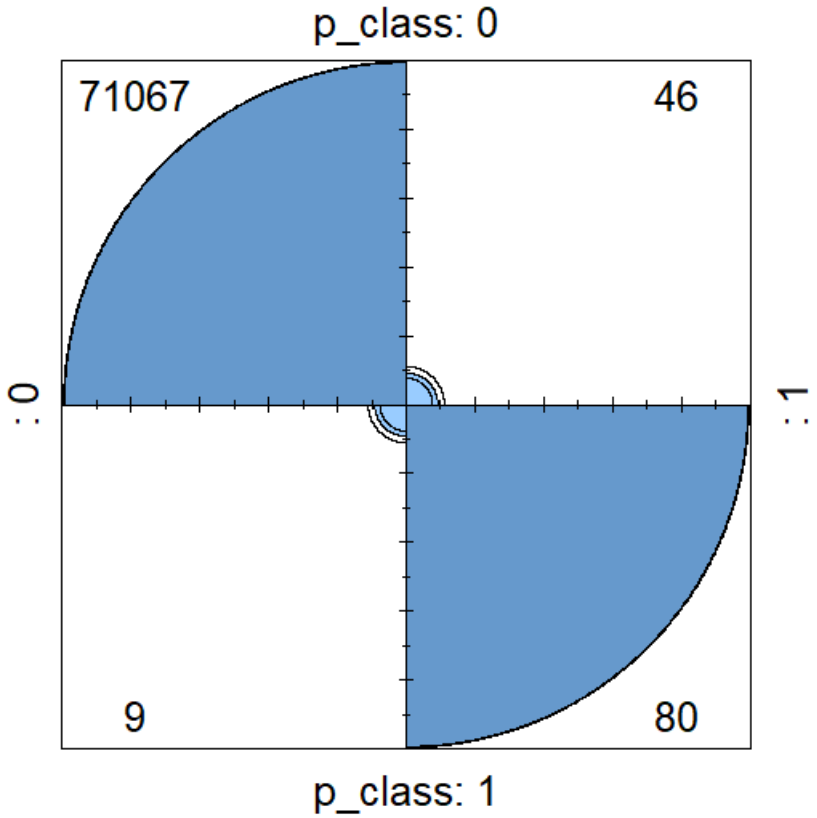


Figure 9 shows the fourfold plot of the predictions of the logistic regression model

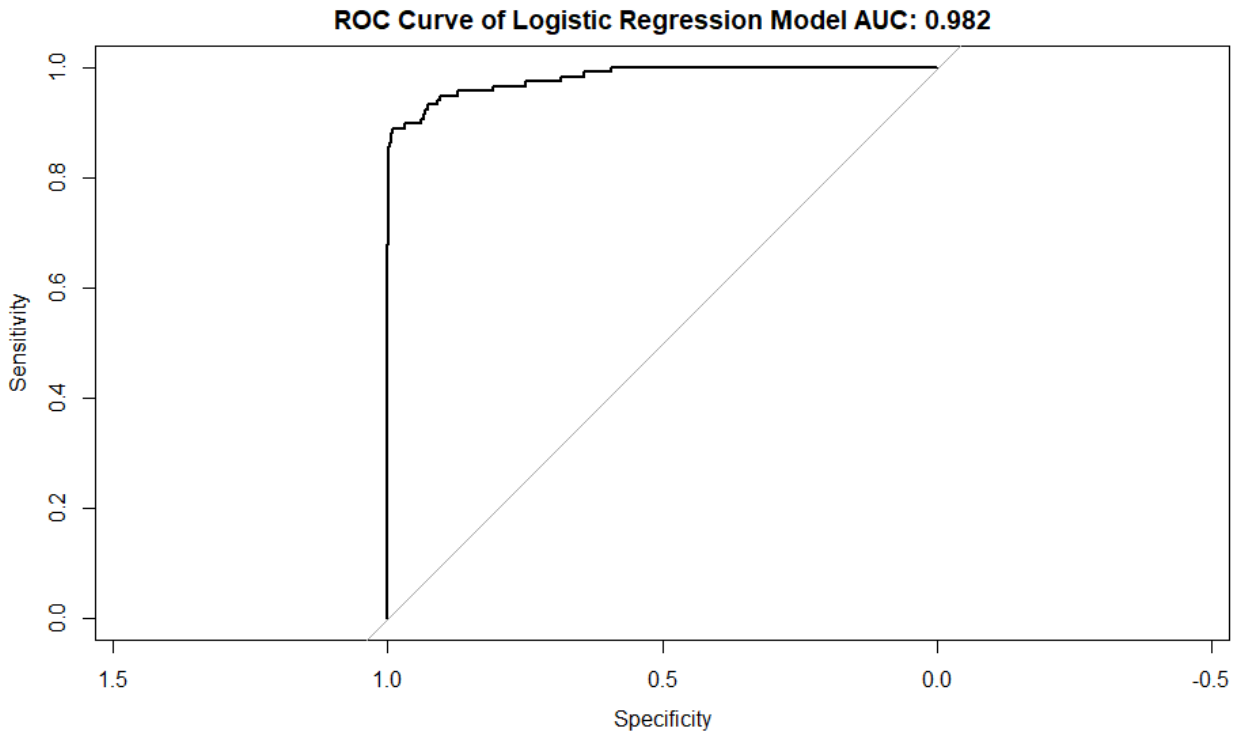
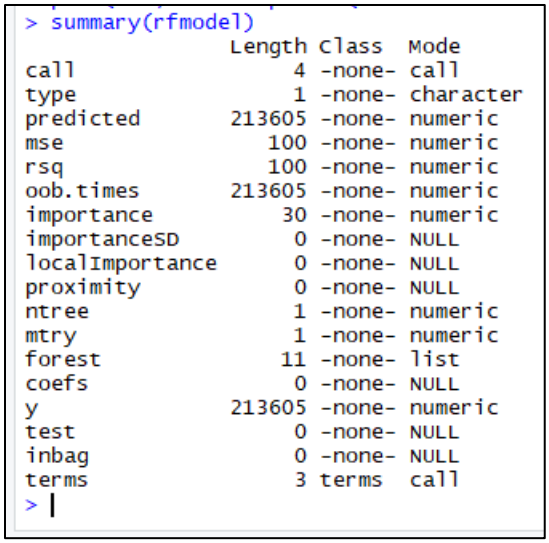
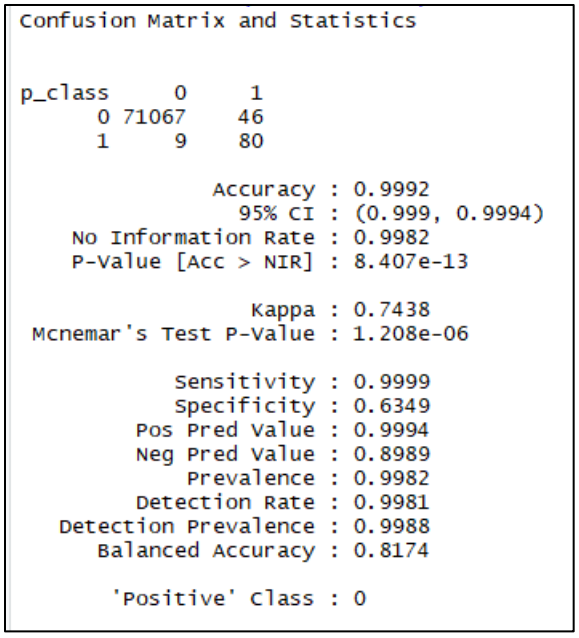


Figure 10, below shows the Area Under the ROC Curve of the Logistic regression model.



```
> summary(rfmodel)
                Length Class  Mode
call                 4 -none- call
type                 1 -none- character
predicted       213605 -none- numeric
mse                100 -none- numeric
rsq                100 -none- numeric
oob.times       213605 -none- numeric
importance          30 -none- numeric
importanceSD         0 -none- NULL
localImportance      0 -none- NULL
proximity            0 -none- NULL
ntree                1 -none- numeric
mtry                 1 -none- numeric
forest              11 -none- list
coefs                0 -none- NULL
y               213605 -none- numeric
test                 0 -none- NULL
inbag                0 -none- NULL
terms                3 terms  call
>
```

Figure 11, shows the summary of the Random Forests model.



```
Confusion Matrix and Statistics


p_class     0     1
      0 71067    46
      1     9    80

               Accuracy : 0.9992
                 95% CI : (0.999, 0.9994)
    No Information Rate : 0.9982
    P-Value [Acc > NIR] : 8.407e-13

                  Kappa : 0.7438
 Mcnemar's Test P-Value : 1.208e-06

            Sensitivity : 0.9999
            Specificity : 0.6349
         Pos Pred Value : 0.9994
         Neg Pred Value : 0.8989
             Prevalence : 0.9982
         Detection Rate : 0.9981
   Detection Prevalence : 0.9988
      Balanced Accuracy : 0.8174

       'Positive' Class : 0
```

Figure 12, shows the confusion matrix using random forests model.

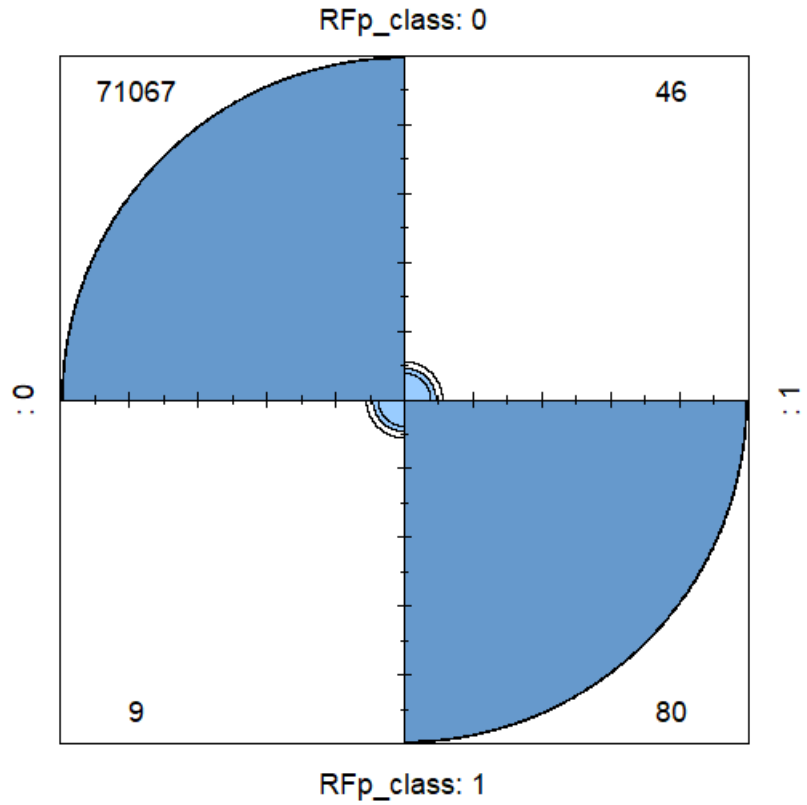


Figure 13, shows the fourfold plot using the random forests algorithm

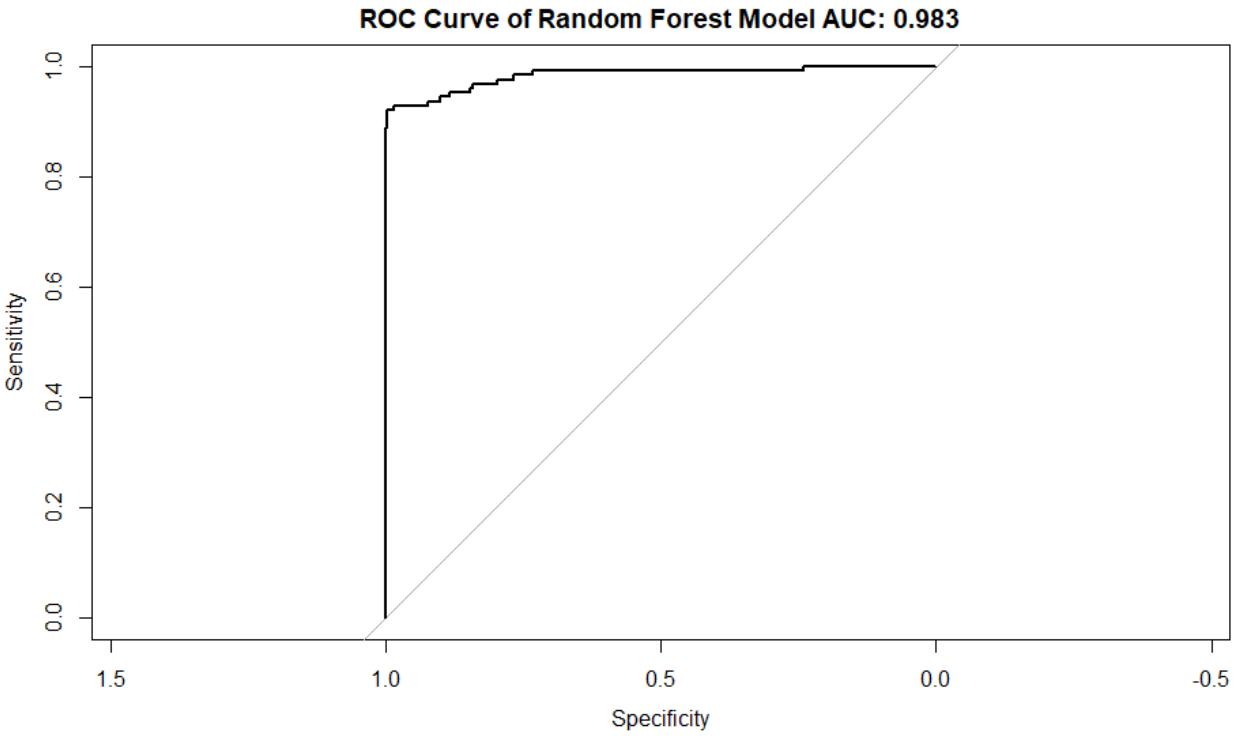


Figure 14, shows the ROC curve drawn on the random forest model's predictions.

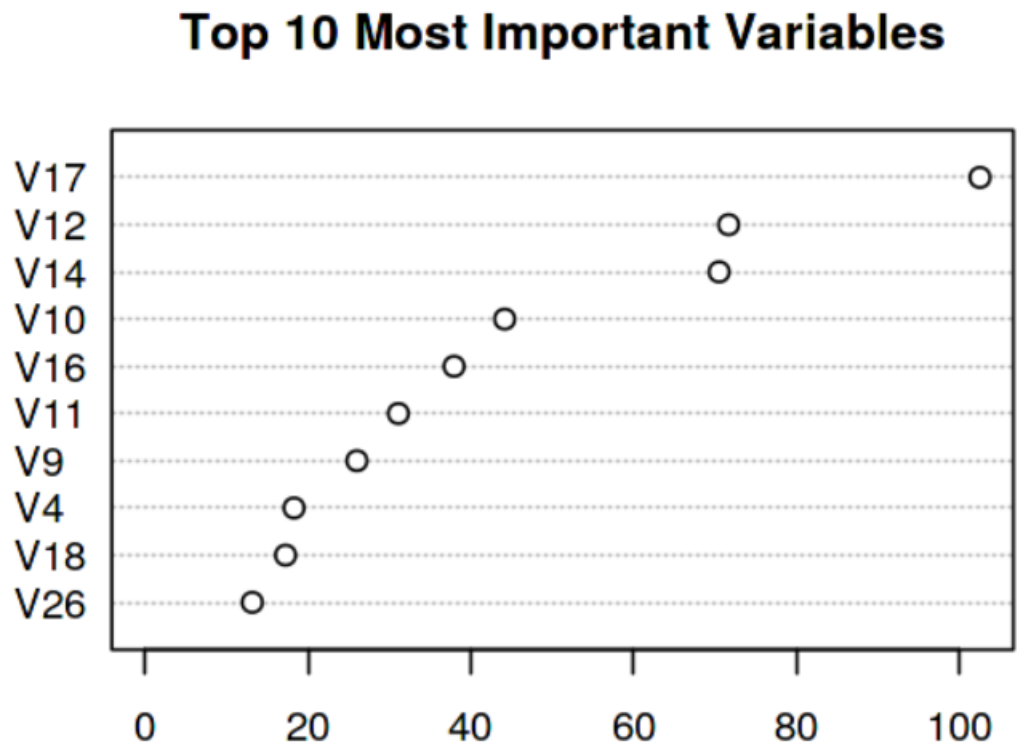


Figure 15, shows variable importance plot of the random forest model