

CREDIT CARD FRAUD DETECTION USING DATA MINING TECHNIQUES

Maryann Inimfon Atakpa¹

¹Big Data Analytics

1947 Words

Student ID:20158627

Keywords: Credit card, Fraud, Data mining

10th January, 2021

Abstract

Many businesses, particularly the financial industry, have been able to expand their services to customers all over the world, thanks to the evolution of the internet. Users now find convenience in making payment for services online, using their credit cards. Due to this technological advancement, fraudsters are continuously improvising ways to steal money from consumers online, thereby making credit card fraud exponentially worse. The domain for this study is focused on Financial Data Security/Credit Card Security. Which is aimed at helping financial institutions to detect and predict suspicious activities. The key challenge is how to improve the accuracy of fraud detection with an increasing number of user-per-second transactions. The rise in the number of users and online transactions has caused these systems to have heavy workloads. In this study we are going to identify credit card fraudulent transactions using various types of supervised machine learning algorithms. To be able to determine which algorithm gives the best prediction, i'll be comparing three data mining techniques which are Logistic regression, Xgboost and Random forest.

1 Introduction

Many businesses, particularly the financial industry, have been able to expand their services to customers all over the world, thanks to the evolution of the internet. Users now find convenience in making payment for services online, using their credit cards.

Due to this technological advancement, fraudsters are continuously improvising ways to steal money from consumers online, thereby making credit card fraud exponentially worse.

Financial institutions have now been saddled with the responsibility of ensuring that customers/consumers of their services can perform safe transactions using their credit card for electronic payments of products and services delivered on the internet.

In detecting and combating these credit card frauds, data mining techniques have been introduced due to it's powerful techniques and algorithms that can be applied through information discovery to detect or predict fraud from unusual patterns extracted from data collected.

However, there are several other approaches in detecting fraudulent transactions which includes Decision Trees, Classification With Logistics Regression and so on.

In this study, we shall discuss the use of R programming language to analyse and compare three data mining techniques which are: Logistics Regression, Random Forest and Xgboost, in detecting fraudulent transactions.

2 Domain Description

The domain for this study is focused on Financial Data Security/Credit Card Security. Which is aimed at helping financial institutions to detect and predict suspicious activities.

Usually when transactions happen, there is a payer and a receiver, and a gateway (financial institutions, issuing/acquiring banks, credit card processors, or the switch) through which money is paid. These actors are potential targets for attackers or fraudsters.

Before a fraud can happen, fraudsters usually perform some information gathering (called reconnaissance), and then attempt to use this information to make fraudulent transactions on-behalf of the credit card owner.

There is an evolving need to protect or safeguard the money and credit information of users from attackers using datasets and patterns of suspicious activities.

3 Problem Definition

Detection of fraud in financial transactions is one of the most critical concerns facing financial firms.

A credit card is a plastic card that allows you access to credit that can be used on making transactions, reducing debt, and receiving benefits.

Credit card fraud happens when someone uses your credit card or credit account to make a payment that you did not approve.

The key challenge is how to improve the accuracy of fraud detection with an increasing number of user-per-second transactions. The rise in the number of users and online transactions has caused these systems to have heavy workloads.

In this study we are going to identify credit card fraudulent transactions using various types of supervised machine learning algorithms. To be able to determine which algorithm gives the best prediction, i'll be comparing three data mining techniques which are Logistic regression, Xgboost and Random forest.

3.1 Data Mining Techniques

Predicting fraud detection in this paper would use the 'Supervised Machine Learning Algorithm'.

3.1.1 Supervised Machine Learning Algorithm

'A supervised strategy works by exploiting the system's previous fraudulent and non-fraudulent transactions, using them to define a model capable of classifying new transactions into a specific class (i.e., legitimate or fraudulent)'

Supervised learning problems can be further classified into problems of regression and classification.

Popular examples of supervised machine learning algorithms are

- Logistics regression for regression problems
- Random forest for classification and regression problems
- Xgboost

3.1.2 Logistics Regression Algorithm

Logistic Regression is a machine learning algorithm, a predictive analysis algorithm focused on the idea of probability, which is used for classification problems.

An example of classification problems are Online transactions fraud or not fraud. The logistic regression is also called the linear regression it predicts the probability of occurrence by fitting a data to a logistics function;

$$O = e^{(I_0 + I_1 * x)} / (1 + e^{(I_0 + I_1 * x)}) \text{ Where,}$$

O is the predicted output

I₀ is the bias or intercept term

I₁ is the coefficient for the single input value (x).

Each column in the input data has an associated I coefficient (a constant real value) that must be learned from the training data. Logistic function is used in the logistic regression in which the cost function quantifies the error, as its model's response is compared with the true value.

3.1.3 Xgboost Algorithm

Xgboost stands for eXtreme Gradient Boosting. Xgboost is a supervised learning approach that is based on the approximation of functions by optimising unique loss functions, It is an implementation of the fast and efficient gradient boosted decision trees.

The implementation of the algorithm was designed for the utilisation of compute time and memory resources. A design goal was to make the best use of available resources to train the model.

Typically, the XGBoost is fast. When compared to other gradient boosting implementations, it is very fast.

'The boosting function consists of three simple steps:

- An initial model F₀ is defined to predict the target variable y. This model will be associated with a residual (y – F₀)
- A new model h₁ is fit to the residuals from the previous step
- Now, F₀ and h₁ are combined to give F₁, the boosted version of F₀. The mean squared error from F₁ will be lower than that from F₀:

To improve the performance of F₁, we could model after the residuals of F₁ and create a new model F₂:

$$F_1(x) <- F_0(x) + h_1(x)$$

$$F_2(x) <- F_1(x) + h_2(x)$$

This can be done for 'm' iterations, until residuals have been minimized as much as possible':

$$F_m(x) <- F_{m-1}(x) + h_m(x)$$

3.1.4 Random Forest Classifier

'Random forest is a supervised learning algorithm which constructs and fuses several decision-making trees to make predictions more accurate and steady.' Random forest can be used for both classification and regression problems. Below you can see how a random forest would look like with two trees:

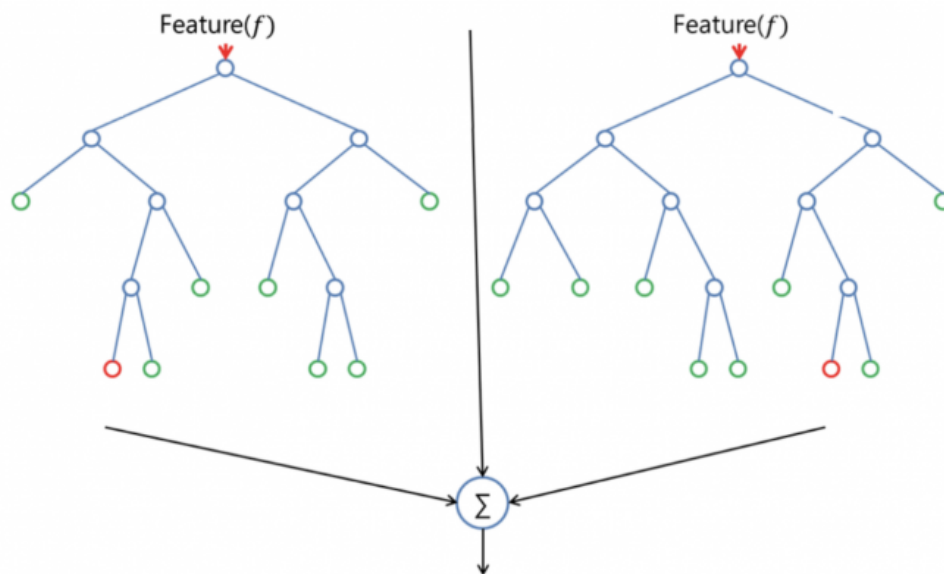


Fig. 1. Random Forest

3.2 AUC AND ROC CURVE

AUC stands for Area Under the Curve Plot it is a calculation of the relationship between false positive and true positive. The greater the AUC, the better the model usually is. The "steepness" of the curve, nevertheless, is also important to inspect, as this reflects the maximisation of the true positive rate while decreasing the false positive rate.

ROC stands for Receiver Operating Characteristic which is a measure of the predictive quality of a classifier that compares and visualises the trade-off between the sensitivity and specificity of the model. A ROC curve reveals the real positive rate on the Y-axis and the false positive rate on the X-axis, both on a global average and per-class basis, when plotted. Therefore, the optimal point is the top-left corner of the diagram where false positives are zero and true positives are one.

In this paper AUC and ROC's Curves are plotted for each Machine learning algorithm.

4 Data Set Description

'The dataset includes credit card purchases made by European cardholders in September 2013.

This dataset presents two-day transactions, in which we have 492 frauds out of 284,807 transactions. The dataset is very unbalanced, and 0.172 percent of all transactions are of the positive class (fraud).

It only includes numerical input variables resulting from the processing of a PCA. The key PCA components are features V1.V2. ...V28. 'Time' and 'Amount' are the only features that are not converted with PCA. The 'Time' feature includes the seconds between any transaction and the first dataset transaction.

The 'Amount' feature is the transaction amount, which can be used for example-depending on cost-sensitive learning. The "class" function is the response variable, which, in the event of fraud, takes value 1, or value 0.'

This data set was gotten from kaggle

5 Data Set Pre-processing

Our aim is to use 31 variables from the dataset and the description to predict a class variable.

5.1 Pre-Processing Steps

Below are the steps taken to process the data

- Step 1: Download the dataset
- Step 2: Read the dataset
- Step 3: Run an exploratory and statistical analysis of the dataset
- Step 4: Run the number of rows, fraudulent and non fraudulent count of the dataset
- Step 5: Split the dataset into 2 (training and verification dataset each)

listings

5.2 Install Packages and load libraries

```
1 library(dplyr) # for data manipulation
2 library(caTools) # for train/test split
3 library(ggplot2) # for data visualization
4 library(Rborist)# for random forest model
5 library(xgboost) # for xgboost model

1 # Input data files are available in the "../input/" directory.
```

Reading the dataset

```
1 dataset<- read.csv("../input/creditcard.csv")
```

This is the exploratory analysis of the data set

```
1 Str (df)
2 'data.frame': 284807 obs. of 31 variables:
3 $ Time : num 0 0 1 1 2 2 4 7 7 9 ...
4 $ V1 : num -1.36 1.192 -1.358 -0.966 -1.158 ...
5 $ V2 : num -0.0728 0.2662 -1.3402 -0.1852 0.8777 ...
6 $ V3 : num 2.536 0.166 1.773 1.793 1.549 ...
7 $ V4 : num 1.378 0.448 0.38 -0.863 0.403 ...
8 $ V5 : num -0.3383 0.06 -0.5032 -0.0103 -0.4072 ...
9 $ V6 : num 0.4624 -0.0824 1.8005 1.2472 0.0959 ...
10 $ V7 : num 0.2396 -0.0788 0.7915 0.2376 0.5929 ...
11 $ V8 : num 0.0987 0.0851 0.2477 0.3774 -0.2705 ...
12 $ V9 : num 0.364 -0.255 -1.515 -1.387 0.818 ...
13 $ V10 : num 0.0908 -0.167 0.2076 -0.055 0.7531 ...
```

```

14 $ V11 : num -0.552 1.613 0.625 -0.226 -0.823 ...
15 $ V12 : num -0.6178 1.0652 0.0661 0.1782 0.5382 ...
16 $ V13 : num -0.991 0.489 0.717 0.508 1.346 ...
17 $ V14 : num -0.311 -0.144 -0.166 -0.288 -1.12 ...
18 $ V15 : num 1.468 0.636 2.346 -0.631 0.175 ...
19 $ V16 : num -0.47 0.464 -2.89 -1.06 -0.451 ...
20 $ V17 : num 0.208 -0.115 1.11 -0.684 -0.237 ...
21 $ V18 : num 0.0258 -0.1834 -0.1214 1.9658 -0.0382 ...
22 $ V19 : num 0.404 -0.146 -2.262 -1.233 0.803 ...
23 $ V20 : num 0.2514 -0.0691 0.525 -0.208 0.4085 ...
24 $ V21 : num -0.01831 -0.22578 0.248 -0.1083 -0.00943 ...
25 $ V22 : num 0.27784 -0.63867 0.77168 0.00527 0.79828 ...
26 $ V23 : num -0.11 0.101 0.909 -0.19 -0.137 ...
27 $ V24 : num 0.0669 -0.3398 -0.6893 -1.1756 0.1413 ...
28 $ V25 : num 0.129 0.167 -0.328 0.647 -0.206 ...
29 $ V26 : num -0.189 0.126 -0.139 -0.222 0.502 ...
30 $ V27 : num 0.13356 -0.00898 -0.05535 0.06272 0.21942 ...
31 $ V28 : num -0.0211 0.0147 -0.0598 0.0615 0.2152 ...
32 $ Amount: num 149.62 2.69 378.66 123.5 69.99 ...
33 $ Class : int 0 0 0 0 0 0 0 0 0 0 0 ...

```

The dataframe has 284807 observations with 31 variables. The variable 'Class' indicates whether a transaction is fraudulent(1) or not (0).

Summary (dataframe) a statistical exploration of the data set

```

1   Time          V1          V2          V3
2   Min.      :    0   Min.   :-56.40751   Min.   :-72.71573   Min.
3   : -48.3256   Min.   :-5.68317
4   1st Qu.: 54202   1st Qu.: -0.92037   1st Qu.: -0.59855   1st Qu.:
5   -0.8904   1st Qu.: -0.84864
6   Median : 84692   Median :  0.01811   Median :  0.06549   Median :
7   0.1799   Median : -0.01985
8   Mean    : 94814   Mean    :  0.00000   Mean    :  0.00000   Mean    :
9   0.0000   Mean    :  0.00000
10  3rd Qu.:139320   3rd Qu.:  1.31564   3rd Qu.:  0.80372   3rd Qu.:
11  1.0272   3rd Qu.:  0.74334
12  Max.    :172792   Max.    :  2.45493   Max.    : 22.05773   Max.    :
13  9.3826   Max.    :16.87534
14          V5          V6          V7          V8
15          V9
16  Min.    :-113.74331   Min.    :-26.1605   Min.    :-43.5572   Min.
17  : -73.21672   Min.    :-13.43407
18  1st Qu.: -0.69160   1st Qu.: -0.7683   1st Qu.: -0.5541   1st Qu.:
19  -0.20863   1st Qu.: -0.64310
20  Median : -0.05434   Median : -0.2742   Median :  0.0401   Median :
21  0.02236   Median : -0.05143

```

12	Mean	:	0.00000	Mean	:	0.0000	Mean	:	0.0000	Mean	:
			0.00000	Mean	:	0.00000					
13	3rd Qu.:		0.61193	3rd Qu.:		0.3986	3rd Qu.:		0.5704	3rd Qu.:	
			0.32735	3rd Qu.:		0.59714					
14	Max.	:	34.80167	Max.	:	73.3016	Max.	:	120.5895	Max.	:
			20.00721	Max.	:	15.59500					
15	V10			V11			V12			V13	
			V14								
16	Min.	:	-24.58826	Min.	:	-4.79747	Min.	:	-18.6837	Min.	:
			-5.79188	Min.	:	-19.2143					
17	1st Qu.:		-0.53543	1st Qu.:		-0.76249	1st Qu.:		-0.4056	1st Qu.	:
			-0.64854	1st Qu.:		-0.4256					
18	Median	:	-0.09292	Median	:	-0.03276	Median	:	0.1400	Median	:
			-0.01357	Median	:	0.0506					
19	Mean	:	0.00000	Mean	:	0.00000	Mean	:	0.0000	Mean	:
			0.00000	Mean	:	0.0000					
20	3rd Qu.:		0.45392	3rd Qu.:		0.73959	3rd Qu.:		0.6182	3rd Qu.:	
			0.66251	3rd Qu.:		0.4931					
21	Max.	:	23.74514	Max.	:	12.01891	Max.	:	7.8484	Max.	:
			7.12688	Max.	:	10.5268					
22	V15			V16			V17			V18	
			V19								
23	Min.	:	-4.49894	Min.	:	-14.12985	Min.	:	-25.16280	Min.	:
			-9.498746	Min.	:	-7.213527					
24	1st Qu.:		-0.58288	1st Qu.:		-0.46804	1st Qu.:		-0.48375	1st Qu.	:
			-0.498850	1st Qu.:		-0.456299					
25	Median	:	0.04807	Median	:	0.06641	Median	:	-0.06568	Median	:
			-0.003636	Median	:	0.003735					
26	Mean	:	0.00000	Mean	:	0.00000	Mean	:	0.00000	Mean	:
			0.000000	Mean	:	0.000000					
27	3rd Qu.:		0.64882	3rd Qu.:		0.52330	3rd Qu.:		0.39968	3rd Qu.:	
			0.500807	3rd Qu.:		0.458949					
28	Max.	:	8.87774	Max.	:	17.31511	Max.	:	9.25353	Max.	:
			5.041069	Max.	:	5.591971					
29	V20			V21			V22			V23	
			V24								
30	Min.	:	-54.49772	Min.	:	-34.83038	Min.	:	-10.933144	Min.	:
			-44.80774	Min.	:	-2.83663					
31	1st Qu.:		-0.21172	1st Qu.:		-0.22839	1st Qu.:		-0.542350	1st Qu.:	
			-0.16185	1st Qu.:		-0.35459					
32	Median	:	-0.06248	Median	:	-0.02945	Median	:	0.006782	Median	:
			-0.01119	Median	:	0.04098					
33	Mean	:	0.00000	Mean	:	0.00000	Mean	:	0.000000	Mean	:
			0.00000	Mean	:	0.00000					
34	3rd Qu.:		0.13304	3rd Qu.:		0.18638	3rd Qu.:		0.528554	3rd Qu.:	
			0.14764	3rd Qu.:		0.43953					


```

35 Max.      : 39.42090    Max.      : 27.20284    Max.      : 10.503090    Max.      :
    22.52841    Max.      : 4.58455
36      V25              V26              V27              V28
    Amount
37 Min.      : -10.29540   Min.      : -2.60455   Min.      : -22.565679   Min.
    : -15.43008   Min.      : 0.00
38 1st Qu.: -0.31715   1st Qu.: -0.32698   1st Qu.: -0.070840   1st Qu.:
    -0.05296   1st Qu.: 5.60
39 Median : 0.01659   Median : -0.05214   Median : 0.001342   Median :
    0.01124   Median : 22.00
40 Mean      : 0.00000   Mean      : 0.00000   Mean      : 0.000000   Mean      :
    0.00000   Mean      : 88.35
41 3rd Qu.: 0.35072   3rd Qu.: 0.24095   3rd Qu.: 0.091045   3rd Qu.:
    0.07828   3rd Qu.: 77.17
42 Max.      : 7.51959   Max.      : 3.51735   Max.      : 31.612198   Max.      :
    33.84781   Max.      : 25691.16
43      Class
44 Min.      : 0.000000
45 1st Qu.: 0.000000
46 Median : 0.000000
47 Mean      : 0.001728
48 3rd Qu.: 0.000000
49 Max.      : 1.000000
50 >

```

Number of fraud/Non-fraud data of whole data

```

1  #nrow(transactions);
2  [1] 284807
3
4  #Non fraud count
5  > length(which(transactions$Class == 0))
6  [1] 284315
7
8  #Fraud count
9  > length(which(transactions$Class == 1))
10 [1] 492

```

In comparison to 284315, there are only 492 fraud cases in total of 284807, which amounts to 99,82 percent of the fraud cases without a blueprint.

Splitting the dataset

```

1  # Split data to training and verification sets 80:20
2  Where 80 is training set and 20 is verification set
3  library(caTools)
4  split <- sample.split(transactions, SplitRatio = 0.8)
5  training <- subset(transactions, split == TRUE)
6  verification <- subset(transactions, split == FALSE)

```

5.3 Experiments and Analysis Using Data Mining Techniques

5.4 Classification using logistics regression

```
1  # Classification using logistics regression
2 classifier <- glm(formula = training$Class ~.,
3                   family = binomial, data = training[, -30])
4 pred <- predict(classifier, type = 'response', newdata = verification[,
5                   -30])
6
7
8  # Run the ROCR functions for AUC calculation
9 roc_perf <- performance(prediction(pred, verification[, 30]), 'tpr', '
10   fpr')
11
12 roc_sens <- performance(prediction(pred, verification[, 30]), 'sens', '
13   spec')
14
15
16 roc_auc <- performance(prediction(pred, verification[, 30]), 'auc')
17
18 roc_err <- performance(prediction(pred, labels=verification[, 30]), '
19   err')
20
21
22 # AUC value
23 auc <- roc_auc@y.values[[1]]
24
25 #AUC of Precision and recall
26
27 prc_frame <- data.frame(pred, training$Class)
28 prc <- pr.curve(prc_frame[prc_frame$training.Class == 'YES',]$pred,
29   prc_frame[prc_frame$training.Class == 'NO',]$pred,
30   curve = TRUE)
31
32 plot(roc_perf)
33
34 plot(prc)
```

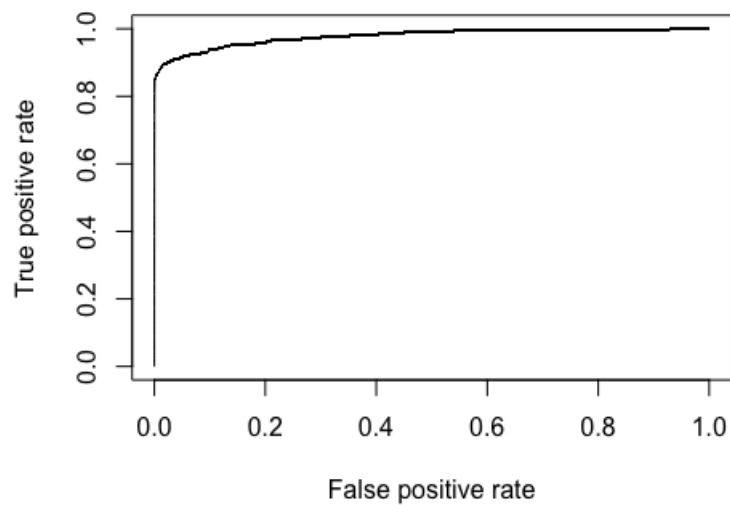


Fig. 2. ROC curve for Logistics Regression

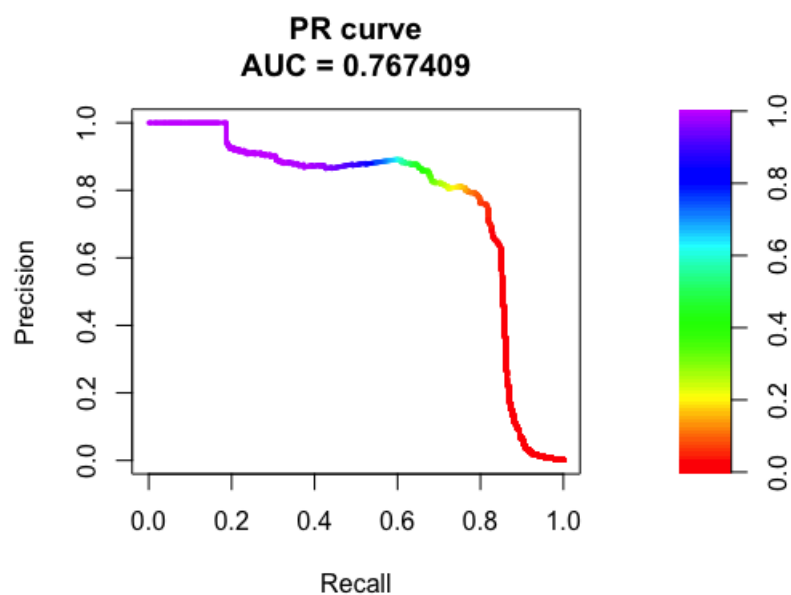


Fig. 3. PRAUC Curve for Logistics Regression

5.5 Classification using Xgboost

```

1 classifier_xg = xgboost(data = as.matrix(training[,-30]),label =
  training$Class,nrounds = 10)
2 pred_xg = predict(classifier_xg,type = 'response', newdata = as.matrix(
  verification[,-30]))
3
4 [1] train-rmse:0.352859
5 [2] train-rmse:0.247056

```

```
6 [3] train-rmse:0.173442
7 [4] train-rmse:0.122216
8 [5] train-rmse:0.086481
9 [6] train-rmse:0.061590
10 [7] train-rmse:0.044805
11 [8] train-rmse:0.033556
12 [9] train-rmse:0.026252
13 [10] train-rmse:0.021914
14
15 # AUC of ROC function
16
17 # Run the ROCR functions for AUC calculation
18
19 roc_perf <- performance(prediction(pred_xg, verification[, 30]), 'tpr',
20                          'fpr')
21
22 roc_sens <- performance(prediction(pred_xg, verification[, 30]), 'sens',
23                          'spec')
24
25 roc_auc <- performance(prediction(pred_xg, verification[, 30]), 'auc')
26
27 roc_err <- performance(prediction(pred_xg, labels=verification[, 30]),
28                          'err')
29
30 # AUC value
31
32 auc <- roc_auc@y.values[[1]]
33
34 #AUC of Precision and recall
35
36 prc_frame <- data.frame(pred_xg, training$Class)
37
38 prc <- pr.curve(prc_frame[prc_frame$training.Class == 'YES',]$pred_xg,
39                 prc_frame[prc_frame$training.Class == 'NO',]$pred_xg,
40                 curve = TRUE)
41
42 plot(roc_perf)
43
44 plot(prc)
```

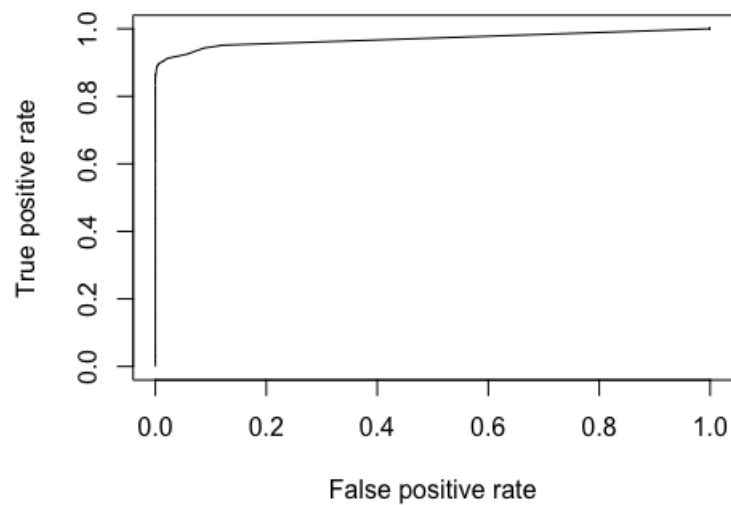


Fig. 4. ROC Curve For Xgboost

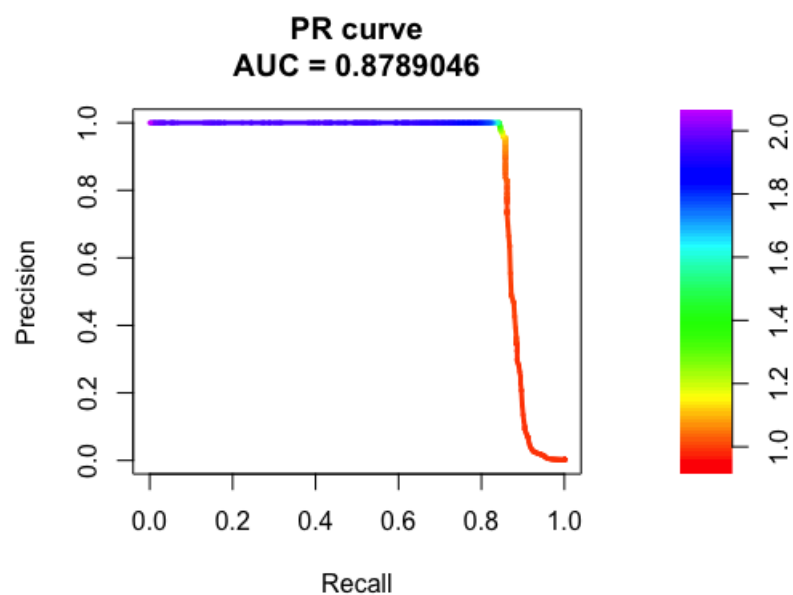


Fig. 5. PRAUC Curve for Xgboost

5.6 Classification using Random Forest

```

1  x = training[, -30]
2  y = training[,30]
3  rf_fit = Rborist(x, y, ntree = 1000, minNode = 20, maxLeaf = 13)
4
5  rf_pred <- predict(rf_fit, verification[, -30], ctgCensus = "prob")
6  prob <- rf_pred$prob
7

```

```

8 # roc.curve(training$Class, prob[,2], plotit = TRUE)
9
10 roc_perf <- performance(prediction(prob[,2], verification[, 30]), 'tpr'
    , 'fpr')
11 roc_sens <- performance(prediction(prob[,2], verification[, 30]), 'sens
    ', 'spec')
12
13 roc_auc <- performance(prediction(prob[,2], verification[, 30]), 'auc')
14 roc_err <- performance(prediction(prob[,2], labels=verification[, 30]),
    'err')
15
16 auc <- roc_auc@y.values[[1]]
17
18 prc_frame <- data.frame(prob[,2], training$Class)
19 prc <- pr.curve(prc_frame[prc_frame$training.Class == 'YES',]$prob
    ...2.,
20                 prc_frame[prc_frame$training.Class == 'NO',]$prob...2.,
21                 curve = TRUE)
22 plot(roc_perf)
23 plot(prc)

```

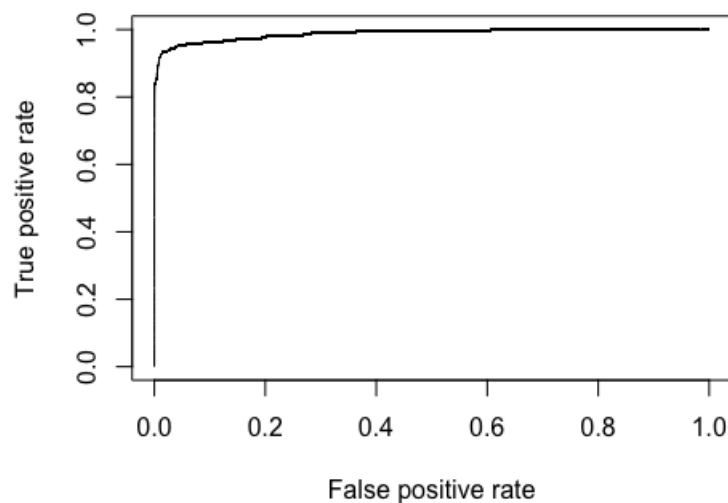


Fig. 6. ROC Curve for Random Forest

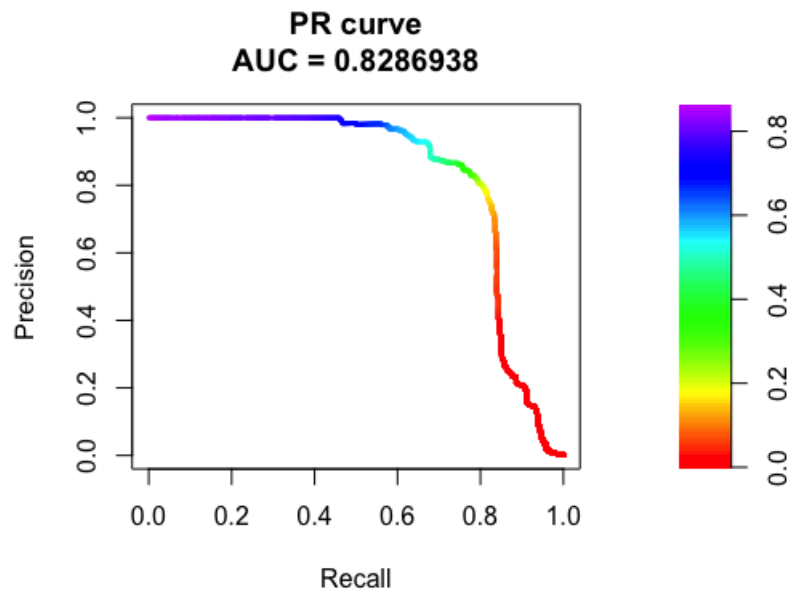


Fig. 7. PRAUC Curve for Random Forest

6 Results

The generated result shows that in a case of a high imbalanced ratio, AUC will not produce an accurate output, while PRAUC provides an accurate ratio of false positive to true positive

The XGBOOST model performed the best with an auc score of 0.878, while both the random forest and logistic regression models showed fair results.

7 Conclusion

In the era of digital transformation where electronic payment has been adopted in making delivery of services easy and much quicker, credit card fraud detection systems can play an essential role in minimizing the numbers of fraudulent transactions.

Furthermore, this study compares data mining techniques used to demonstrate unbalanced dataset of credit card fraud where fraudulent cases are few compared to regular transaction cases.

This paper used supervised machine learning algorithms.

The accuracy is 0.767, 0.878 and 0.828 for logistic regression, Xgboost and Random forest classifier, respectively. By contrasting all three techniques, the Xgboost model was found to perform better than the Random Forest and Logistic Regression models.

8 References

Card, C., 2019. What Is A Credit Card. [online] moneysupermarket.com. Available at: <https://www.moneysupermarket.com/credit-cards/what-is-a-credit-card> [Accessed 15 March 2019].

Kaggle.com. 2018. Credit Card Fraud Detection. [online] Available at: <https://www.kaggle.com/mlg-ulb/creditcardfraud> [Accessed 8 January 2021].

Ounacer, S., Ait El Bour, H., Oubrahim, Y., Ghoumari, M. and Azzouazi, M., 2018. Using Isolation Forest In Anomaly Detection: The Case Of Credit Card Transactions.

Patil, S., Nemade, V. and Soni, P., 2018. Elsevier. Predictive Modelling For Credit Card Fraud Detection Using Data Analytics, 132, pp.385-395.

Brownlee, J., 2021. Supervised And Unsupervised Machine Learning Algorithms. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/> [Accessed 8 January 2021].

Brownlee, J., 2021. A Gentle Introduction To Xgboost For Applied Machine Learning. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/gentle-introduction-xgboost-applied-machine-learning/> [Accessed 8 January 2021].

Sundaram, R., 2021. Xgboost Algorithm — Xgboost In Machine Learning. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/> [Accessed 6 September 2018].

Donges, N., 2020. A Complete Guide To The Random Forest Algorithm. [online] Built In. Available at: <https://builtin.com/data-science/random-forest-algorithm> [Accessed 3 September 2020].

Ghosh, S., 2019. Rpubs - Credit Card Fraud Detection. [online] Rpubs.com. Available at: https://rpubs.com/saugatoghosh/imabalance_data_card [Accessed 8 January 2019].

Kaggle.com. 2018. Calculating AUC And AUPRC For Credit Card Fraud. [online] Available at: <https://www.kaggle.com/shaliniyaramada/calculating-auc-and-auprc-for-credit-card-fraud> [Accessed 8 January 2021].

Saia, R. and Carta, S., 2017. Evaluating Credit Card Transactions in the Frequency Domain for a Proactive Fraud Detection Approach, [Accessed 10 July 2017].