CREDIT CARD FRAUD DETECTION

AP18110010468
Bollu.Rajyalakshmi
CSE-H

Abstarct:

Fake digital transactions is generic in modern day society. However it's far obvious that the quantity of fraudulent digital transaction instances are continuously increasing in spite of the chip playing cards global integration and existing protection systems. That is why the hassle of fraud detection could be very essential now. In spite of fraud fees growing and cardholder self assurance reducing, monetary establishments need to do so to ensure that their organization and cardholders are safeguarded. These prices incorporate the tough losses that hurt the company's bottom line along with the price incurred to update playing cards, in addition to case investigations, consumer phone guide, and harm to the employer's popularity. In the occasion of a facts breach, the clients could avoid in addition transactions/enterprise with the agency. This challenge intends to illustrate a possible strategy to keep away from the fraudulent transactions using system studying strategies and r programming language.

Introduction

In cutting-edge day the fraud is one of the major causes of first rate monetary losses, no longer best for merchants, character customers also are affected. Many researches have used system gaining knowledge of algorithms to come across fraudulent transactions. As in step with the survey from 2014 to 2018 it became observed that losses because of counterfeit amounted to 3 billion u.S. Bucks in 2014 and they are projected to decrease to one.8 billion u.S. Dollars by way of the give up of 2018. Therefore, detecting technique in most effective manner is a time-consuming procedure and ordinarily is accomplished offline in static operation. The goal of this paper underlying the content is to triumph over the above-referred to trouble in an efficient manner by way of the usage of r programming language. As a consequence, the use of this project, we are able to examine and visualize huge amount of facts.

Challenges involved in credit card fraud detection are:

1. Large amount of records is processed each day and the version advanced have to be rapid sufficient to respond to the scam in time.
2. Imbalanced records i.E. Maximum of the transactions (ninety nine.8%) aren't fraudulent which makes it laborious for detecting the fraudulent ones
3. Information availability, because the facts is often private.
4. Misclassified information may be some other influential problem, as not all fraudulent transaction is detected and suggested.
5. Adaptive strategies used against the version by means of the scammers.

How do we address these challenges?

1. The model used need to be a easy and greater agile technique to hit upon the exception and classify it as a fraudulent transaction immediately
2. Asymmetry may be handled the usage of pre-processing methods on the way to be addressed inside the approaching segment

3. For protective the privateness of the person, the dimensionality of the information may be decreased
4. 4.A greater dependable source need to be taken which double-exams the facts, at the least for education the model

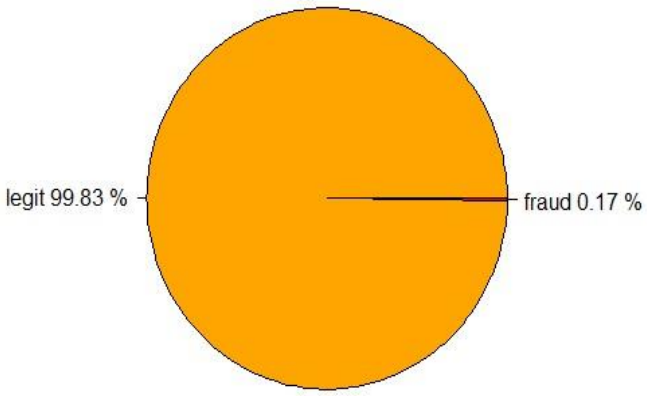
Literature survey:

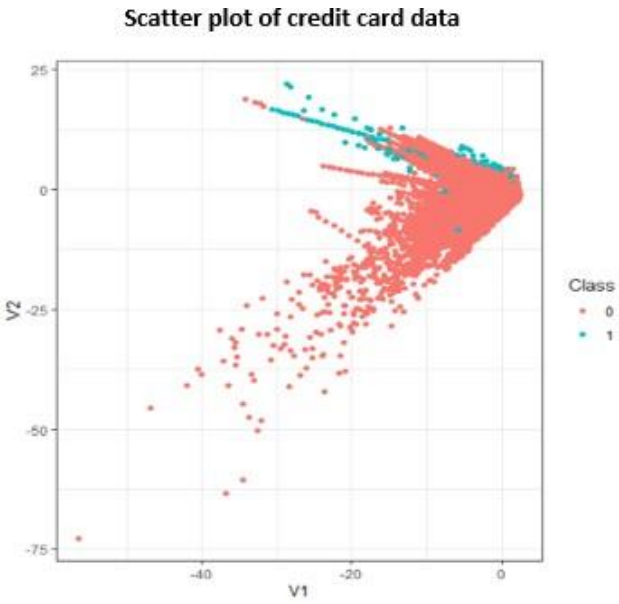
K. R. Seeja et al [2] the author propose a version to encounter fraudulent credit score rating rating card pastime in a excessive degree of disproportionate and nameless sample datasets. The writer has used the not unusual itemset mining technique to discover prison as opposed to fraudulent transactions. This technique is obtainable as similar to apriori algorithms, which returns the set of habitual interest units for each mate interest. The author moreover compares the wonderful fraud detection strategies based totally on their benefits and downsides to proves knn has maxim customer. An equal set of rules is used to pick out the first rate matching sample for the Incoming transaction. The paper has used the matthews correlation coefficient to make sure binary type. In line with the writer, this version has excessive accuracy for fraud detection and less faux brilliant or fake terrible type rate in imbalanced datasets.

Yiğit accurate sufficientültür and a further creator [6] have analysed the cardholder spending behaviour and proposed a card holder behaviour version to come across credit score rating card fraud. The paper started out off speaking approximately the present rule-based totally models for credit score rating card fraud detection and the manner most of the ones fashions are ignoring a essential issue of the problem i.E., the cardholder behaviour. Then a behaviour-primarily based version grow to be proposed, which inferred the transaction policies decided through the usage of the human fraud professionals. They known as this model cardholder behaviour version(cbm).

Data Exploration and Visualization

Data visualization and exploration is perhaps the quickest and most useful way to summarize and study more approximately our data. The dataset that we use here is received from european cardholders provides transactions that happened in two days, wherein we've got 492 frauds out of the 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for the 0.172% from the transactions. Inner this dataset, there are 31 columns out of which 28 are named as v1-v28 to defend sensitive data. The other columns constitute time, amount and sophistication. Time suggests the time gap between the primary transaction and the following one. The amount is the amount of cash transacted. Elegance 0 represents a legitimate transaction and 1 represents a fraudulent one.

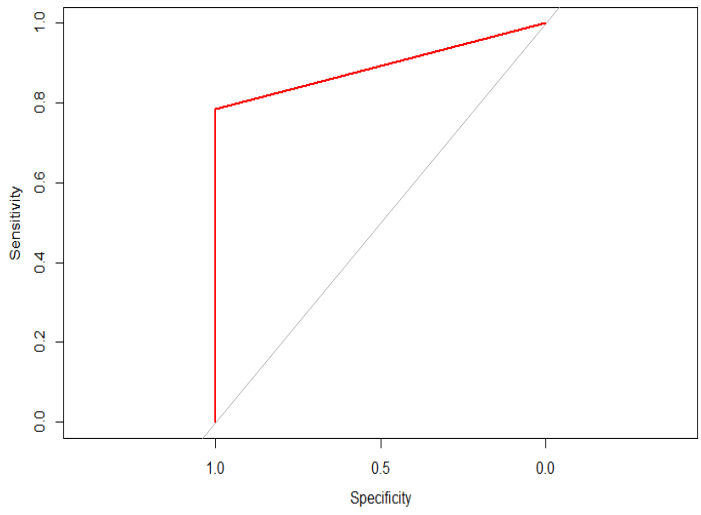




Different Techniques and Implementation:

Random Forest:

Random forest classifier is supervised class/regression ensemble algorithm. Ensembled algorithms are the ones which combines more than one algorithm of identical or exceptional kind for classifying objects. Random forest classifier creates a hard and fast of choice trees from randomly selected subset of training set. It then aggregates the votes from extraordinary selection timber to determine the final elegance of the test item. In a nutshell if we've got more timber in the forest robust the woodland looks as if. Within the equal way in the random wooded area algorithm the better the wide variety of bushes within the wooded area gives the high accuracy results. There are four benefits to illustrate why we use random wooded area algorithm. One is that it may be used for each classification and regression responsibilities. Over becoming is one crucial hassle that could make the effects worse, but for random wooded area algorithm, if there are sufficient trees within the woodland, the classifier gained't over match the version. The 0.33 benefit is the classifier of random woodland can cope with missing values, and the closing advantage is that the random wooded area classifier can be modeled for express values.



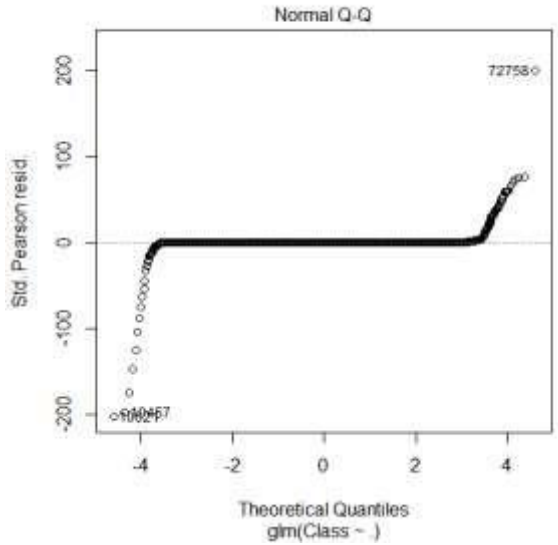
The Area under the curve: 0.8923

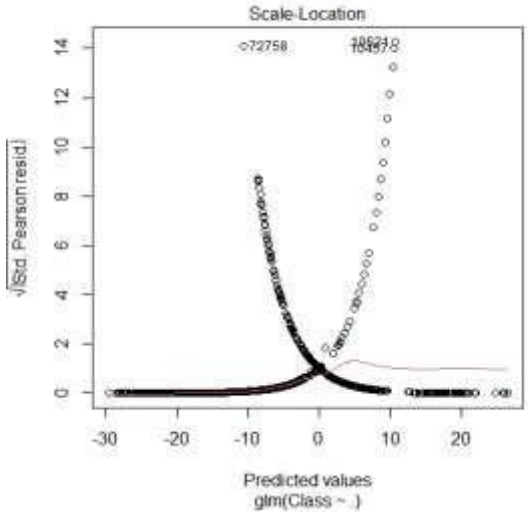
Logistic Regression:

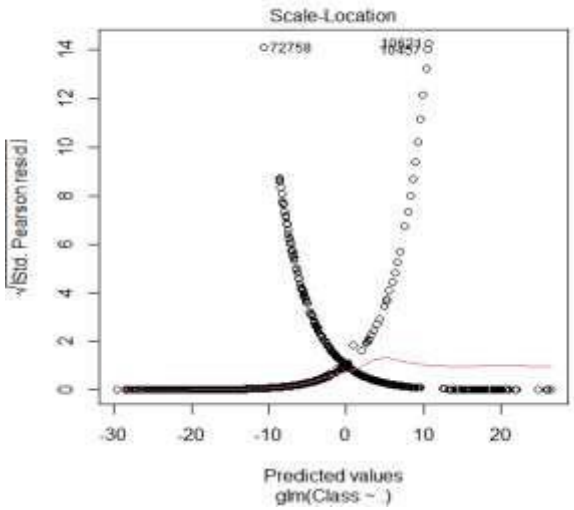
Logistic regression is a well-mounted statistical method for predicting binomial or multinomial results. Like any regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe information and to give an explanation for the relationship among one based binary variable and one or extra nominal, ordinal, c language or ratio-degree independent variables. Logistic feature:
P (y = 1) = 1/ (1 + e −β0 + β1x1 + β2x2 + ... + βkxk)
in which, x1, x2..., x*k are independent variables. P(y) = opportunity prediction. Even though logistic regression is a class set of rules, we expect chances which have to be transformed into a binary fee (zero or 1) that allows you to actually make a opportunity prediction.

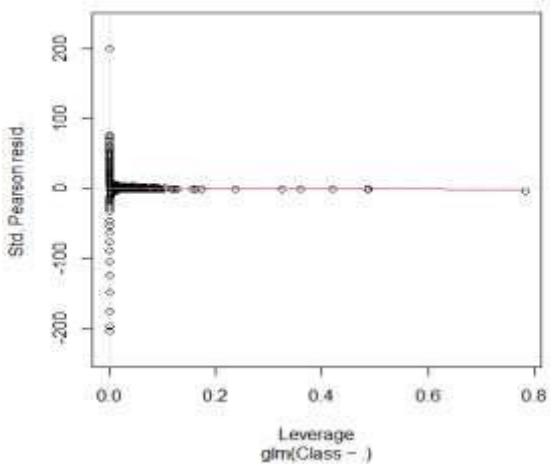
```
summary(LogisticModel)

Call: glm (formula = Class ~., family = binomial (), data =
trainingdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6108  -0.0292  -0.0194  -0.0125   4.6021

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.651305   0.160212 -53.999  < 2e-16 ***
V1           0.072540   0.044144   1.643 0.100332
V2           0.014818   0.059777   0.248 0.804220
V3           0.026109   0.049776   0.525 0.599906
V4           0.681286   0.078071   8.726  < 2e-16 *** V5
0.087938   0.071553   1.229 0.219079
V6          -0.148083   0.085192 -1.738 0.082170.
V7          -0.117344   0.068940 -1.702 0.088731 .
V8          -0.146045   0.035667  -4.095 4.23e-05 ***
V9          -0.339828   0.117595  -2.890 0.003855 **
V10         -0.785462   0.098486  -7.975 1.52e-15 ***
V11          0.001492   0.085147   0.018 0.986018
V12          0.087106   0.094869   0.918 0.358532
V13         -0.343792   0.092381  -3.721 0.000198 ***
V14         -0.526828   0.067084  -7.853 4.05e-15 ***
V15         -0.095471   0.094037  -1.015 0.309991
V16         -0.130225   0.138629  -0.939 0.347537
V17          0.032463   0.074471   0.436 0.662900
V18         -0.100964   0.140985  -0.716 0.473909
V19          0.083711   0.105134   0.796 0.425897
V20         -0.463946   0.081871  -5.667 1.46e-08 ***
V21          0.381206   0.065880   5.786 7.19e-09 ***
V22          0.610874   0.142086   4.299 1.71e-05 ***
```

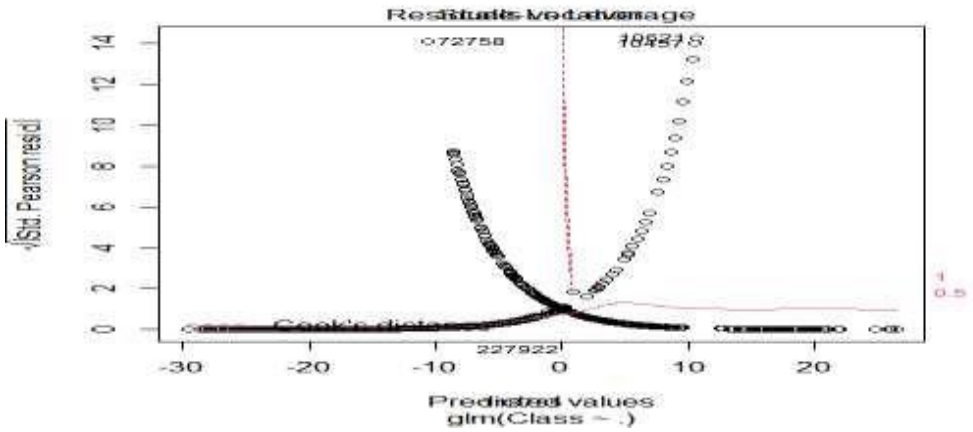
```
V23         -0.071406   0.058799  -1.214 0.224589
V24          0.255791   0.170568   1.500 0.133706
V25         -0.073955   0.142634  -0.519 0.604109
V26          0.120841   0.202553   0.597 0.550783
V27         -0.852018   0.118391  -7.197 6.17e-13 ***
V28         -0.323854   0.090075  -3.595 0.000324 ***
Amount       0.292477   0.092075   3.177 0.001491 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5799.1  on 227845  degrees of freedom
Residual deviance: 1790.9  on 227816  degrees of freedom
AIC: 1850.9
Number of Fisher Scoring iterations: 12
```



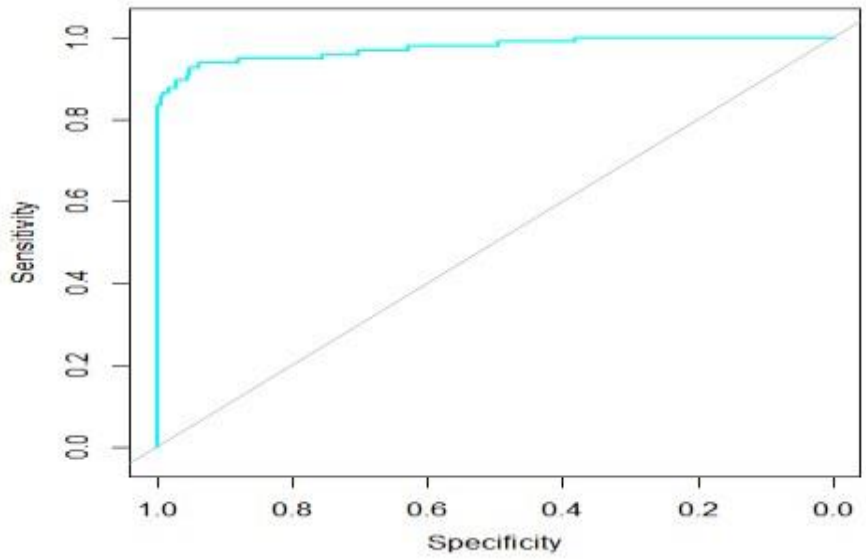








Performance of the model using ROC curve



>print(auc.gbm)

Call: roc.default(response = testdata$Class, predictor = LogiReg.Prediction, plot = T, col = "cyan")

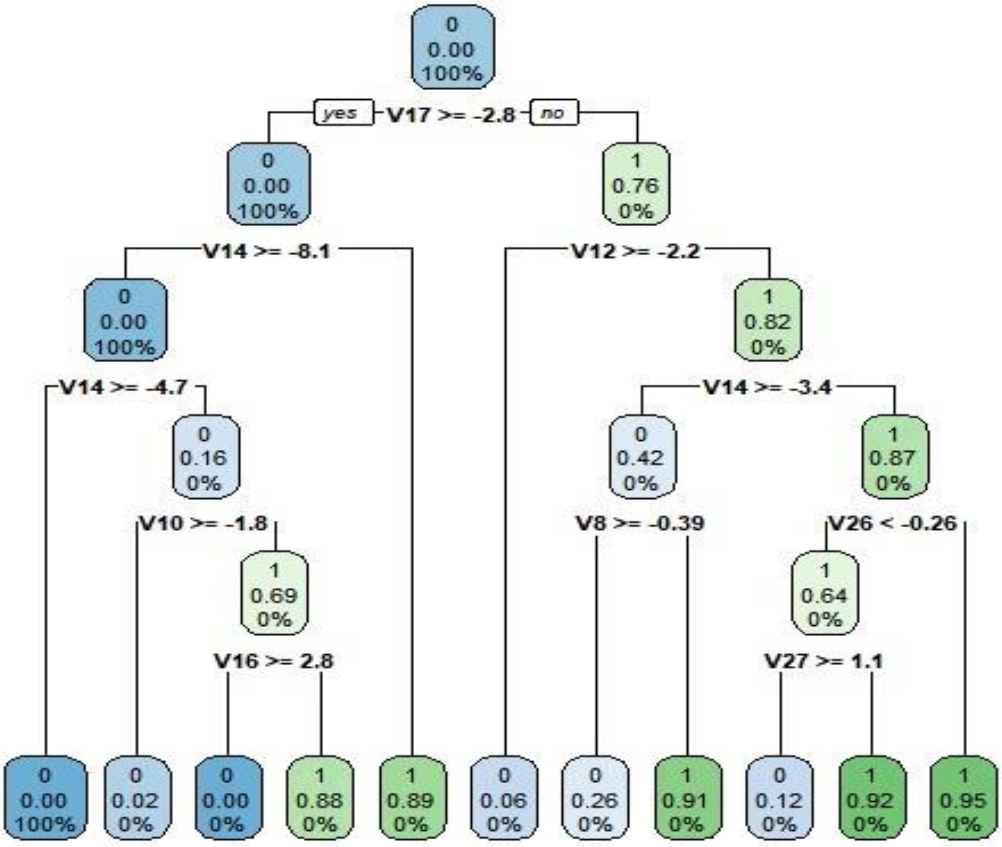
Data: LogiReg.Prediction in 56863 controls (testdata$Class 0) < 98 cases (testdata$Class 1).

Area under the curve: 0.9748

Decision Tree:

Decision trees is the most intuitive one among all the other machine learning algorithms. It is a supervised learning algorithm that can be used for solving both regression and classification problems. It solves the problem by representing the given data and the attributes as a tree. The internal nodes of the tree are the attributes, the branches are the conditional statements and the leaf nodes are the target

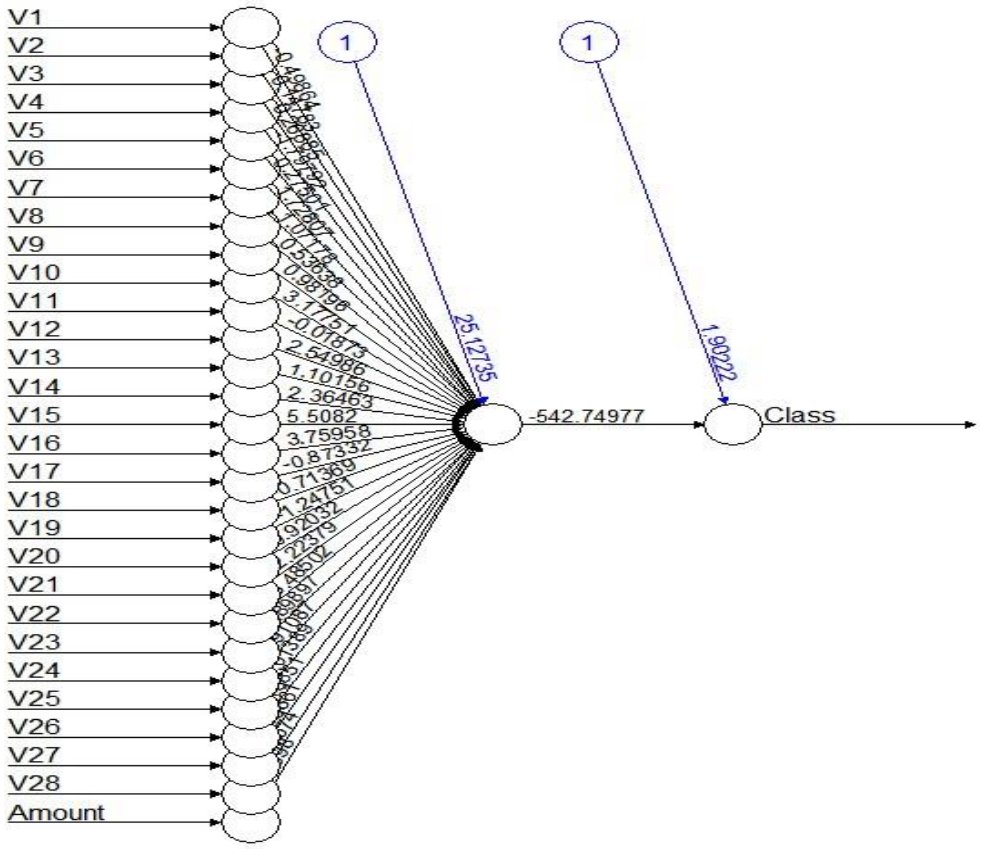
classes. The order of attributes in the tree is decided by calculating the contribution of that attribute using methods like Information Gain .




```
mean(predictedValue == NewData$Class) [1] 0.9995471
```

Artificial Neural Networks:

We are importing the neural net kit which helps us to implement our ANNs. Using historical data, the ANN models can learn the patterns, and can identify the input data. We then plotted it using the plot () function. In the case of ANN, we have a range of numbers that is between 1 and 0. We set a threshold as 'X' and values above 'X' will correspond to 1 and the rest will be 0.

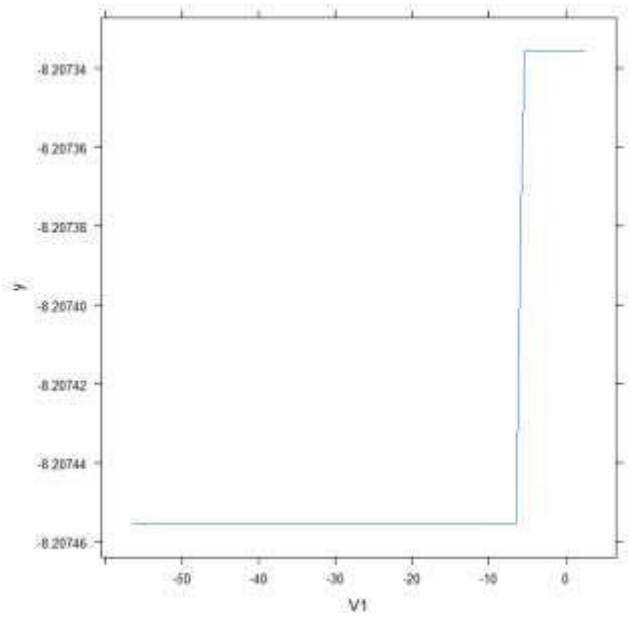


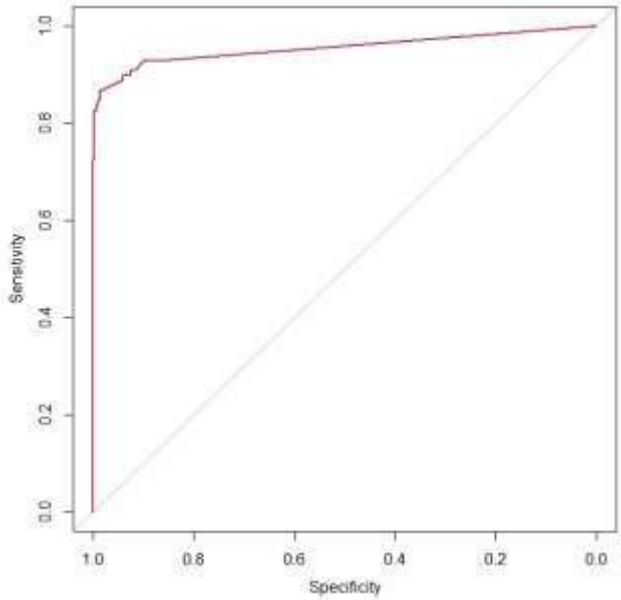
>mean(NNResult==testdata$Class) [1] 0.9993504 > table(NNResult, testdata$Class)

NNResult 0 1 0 56847 21 1 16 77

Gradient Boosting:

Gradient Boosting is a common algorithm for machine learning used to perform classification and regression tasks. This model consists of different underlying models such as weak decision trees. These decision trees combine to form a powerful gradient boosting model. In this method, each new tree is a fit on a modified version of the original data set.





>print(gbm_auc)

Call: roc.default(response = testdata$Class, predictor = gbm_test, plot = T, col = "maroon")

Data: gbm_test in 56863 controls (testdata$Class 0) < 98 cases (testdata$Class 1).

Area under the curve: 0.9555

CONCLUSION:

- Logistic regression has the first-class overall performance. Random forest isn't always favored as it overfits the education statistics and gbm is desired because it improves upon both models.

- ANN isn't always desired as opportunity algorithms inclusive of decision tree, regression which might be to be had and are simple, fast, smooth to teach. They also offer better overall performance.

Primarily based on the plots, outside research and the model output, transaction amount, total variety of declines consistent with day, overseas transaction to high-hazard international locations are the vital factors this is relatively correlated with fraud transactions. So, when there may be a worldwide transaction of an quantity better than the average transaction quantity per day by means of a merchant to a excessive risk country, then it has a excessive possibility of being a fraud transactions.

Our studies helped us understand the numerous fashions to correctly expect the fraudulent credit score card transactions. We built various statistical models that enables prevent suspicious activities and determined variable signs that increases red flag while fraud transactions are tried. However, it should be cited that during this contemporary day and age, fraudsters hold trying more modern methods to hold on these sports. For this reason, it is very critical to constantly include additional attributes into models that might help predict those sports. Also, the use of superior classifier strategies and statistical models and gear ensures reduction and prevention of capability fraud sports in a timely manner.

References

[2] K. R. Seeja and Masoumeh Aerator “Fraud Miner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining”, The Scientific World Journal, Volume 2014

[6] Yiğit Kültür, Mehmet Ufuk Çağlayan, "A novel cardholder behavior model for detecting credit card fraud", Application of Information and Communication Technologies (AICT) 2015 9th International Conference on, pp. 148-152, 2015.XI. APPENDIX

Thank you