A PROJECT REPORT

On

“Credit Card Fraud Detection”

BY

ANKIT ANSHUMAN MOHAPATRA 2006006

SCHOOL OF COMPUTER ENGINEERING

KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY

BHUBANESWAR, ODISHA - 751024

May 2020

Section 1 ~ Introduction

Credit card fraud is a growing concern for both consumers and financial institutions. As the use of credit cards for online transactions continues to increase, so does the risk of fraudulent activity. In response, the field of data science has developed sophisticated algorithms and models to detect and prevent credit card fraud. In this project report, we present our efforts to build a credit card fraud detection system using machine learning models. By leveraging a dataset of credit card transactions, we aim to develop a model that can accurately identify fraudulent transactions in real-time, enabling swift action to be taken to prevent further losses. Through this report, we provide a comprehensive overview of the methods used, the results obtained, and the potential impact of our credit card fraud detection system. Along with that, we will also aim to contribute to the ongoing efforts to combat credit card fraud and provide a valuable tool for financial institutions to protect their customers and prevent losses.

Keywords:
Credit Card Fraud, Machine Learning, R-Programming, Transactions, Deep Learning, Data Analysis, Gradient Booster.

Section 2 ~
Data-Set Description (creditcard.csv)

The creditcard.csv dataset is a robust dataset used for credit card fraud detection and contains anonymous credit card transactions made by European cardholders. The dataset has 284,807 observations (rows) and 31 variables (columns), including the following:

1. **Time:** Number of seconds elapsed between this transaction and the first transaction in the dataset.
2. **V1-V28:** anonymized variables representing different features of the credit card transactions, which were transformed by a principal component analysis (PCA) transformation for confidentiality purposes.
3. **Amount:** transaction amount in euros.
4. **Class:** binary variable indicating whether the transaction was fraudulent (1) or not (0).

The Class variable is the response variable in this dataset, and it is important to note that the vast majority of transactions in this dataset are non-fraudulent (Class=0), with only a small percentage of transactions being fraudulent (Class=1). This dataset is often used to train and test machine learning models for credit card fraud detection, where the goal is to accurately identify fraudulent transactions based on the available features.

Section 3 ~ Proposed Algorithms

Logistic Regression: Logistic Regression is a statistical method used to analyze the relationship between a dependent binary variable (such as fraudulent or non-fraudulent transactions) and one or more independent variables (such as transaction amount or time). It is a popular method for binary classification problems and is used to estimate the probability of a transaction being fraudulent. In the credit card fraud detection project, logistic regression is used to build a predictive model that detects fraudulent transactions based on the available features.

Decision Tree: Decision Tree is a type of supervised machine learning algorithm that is used for both classification and regression tasks. It creates a tree-like model of decisions and their possible consequences, where each internal node represents a decision based on a particular feature, and each leaf node represents a class label. In the credit card fraud detection project, decision tree is used to build a model that can classify transactions as fraudulent or non-fraudulent based on the available features.

Artificial Neural Network: Artificial Neural Network (ANN) is a type of machine learning algorithm inspired by the structure and function of the human brain. It consists of multiple layers of interconnected nodes (neurons) that process and transmit information. In the credit card fraud detection project, ANN is used to build a model that can detect fraudulent transactions based on the available features. ANN has shown promising results in detecting credit card fraud, especially for complex and non-linear patterns.

Gradient Boosting: Gradient Boosting is a type of ensemble learning algorithm that combines multiple weak models to create a strong model. It works by building a sequence of decision trees, where each new tree is trained to correct the errors of the previous tree. In the credit card fraud detection project, gradient boosting is used to build a model that can accurately identify fraudulent transactions based on the available features. Gradient boosting has been shown to be effective in detecting credit card fraud and is commonly used in the industry.

Section 4 ~
Justification - Algorithm Analysis

Logistic Regression: Logistic Regression is used to build a predictive model that detects credit card fraud based on the features available in the creditcard.csv dataset, such as transaction amount, time, and the various V1 to V28 features that are transformed via PCA. Logistic regression provides interpretable results and is useful in identifying which features are most important for detecting credit card fraud.

Decision Tree: Decision Tree algorithm is used to build a model that can classify credit card transactions as fraudulent or non-fraudulent based on the available features. Decision tree creates a tree-like structure where each node represents a decision based on a particular feature, and each leaf node represents a class label (fraudulent or non-fraudulent). Decision tree also provides interpretable results and is useful in identifying which features are most important for detecting credit card fraud.

Artificial Neural Network: Artificial Neural Network (ANN) is used to build a model that can detect credit card fraud based on the available features in the creditcard.csv dataset. ANN is a powerful algorithm that can capture complex and non-linear relationships between the features and the target variable. ANN is also useful in feature selection, as it can automatically identify which features are most important for detecting credit card fraud.

Gradient Boosting: Gradient Boosting is used to build a model that can accurately classify credit card transactions as fraudulent or non-fraudulent based on the available features. Gradient Boosting works by building a sequence of decision trees, where each new tree is trained to correct the errors of the previous tree. Gradient Boosting captures complex relationships between the features and the target variable and can provide accurate predictions. However, Gradient Boosting may not provide interpretable results, and feature selection may require additional analysis.

Section 5 ~ Visualization Analysis

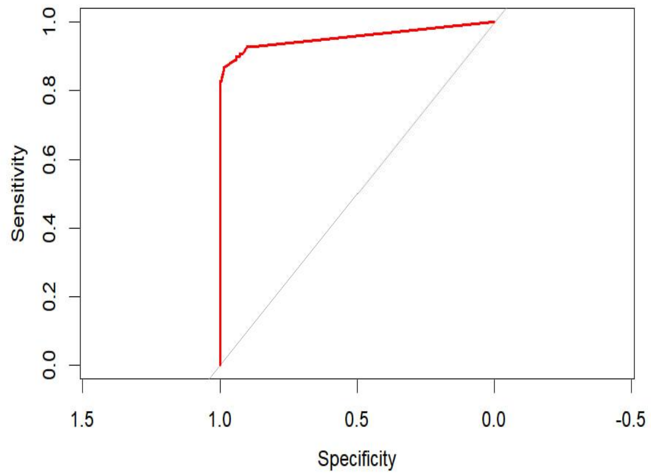
Confusion Matrices: Confusion matrices can be used to visualize the performance of the models. Confusion matrices show the number of true positives, false positives, true negatives, and false negatives. This can be useful in evaluating the sensitivity, specificity, precision, and recall of the models.

ROC Curves: ROC curves can be used to visualize the performance of the models. ROC curves plot the true positive rate (sensitivity) against the false positive rate (1-specificity) for different classification thresholds. This can be useful in selecting the optimal threshold that balances the trade-off between false positives and false negatives.

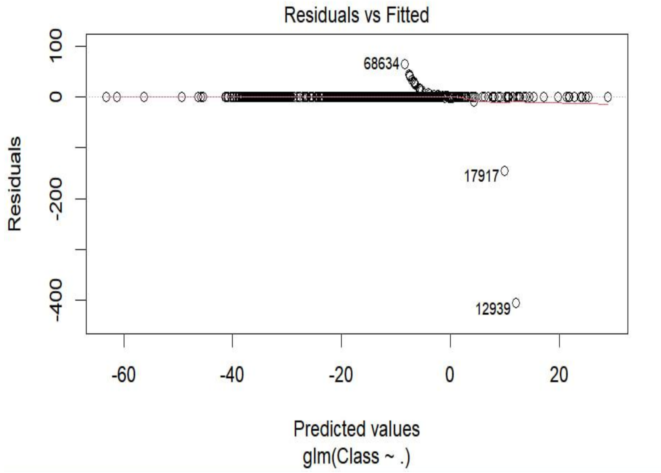
Precision-Recall Curves: Precision-Recall curves can be used to visualize the performance of the models in situations where the class distribution is highly imbalanced (i.e., there are many more non-fraudulent transactions than fraudulent transactions). Precision-Recall curves plot the precision (positive predictive value) against the recall (sensitivity) for different classification thresholds.

Feature Importances: Feature importances can be used to visualize the importance of each feature in the models. Feature importances can be calculated using different methods such as permutation importance or feature importance based on impurity reduction in decision trees. This can be useful in identifying the most important features for credit card fraud detection.

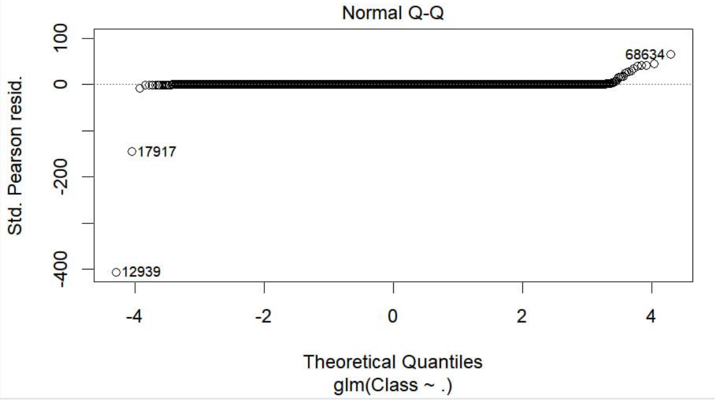
Decision Trees: Decision trees can be visualized to show the decision-making process of the models. Decision trees can show the importance of each feature, the thresholds for each feature, and the class predictions for each leaf node. This can be useful in interpreting the models and identifying potential patterns that can be exploited for credit card fraud detection.



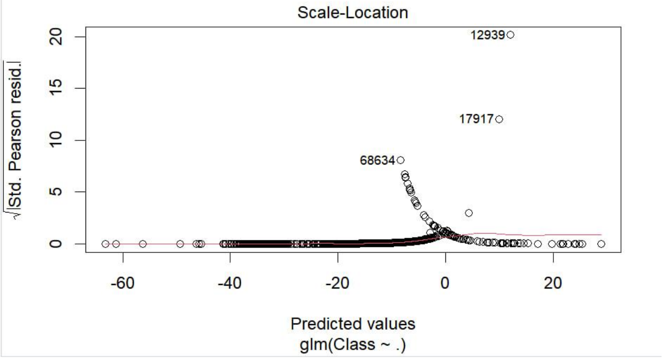
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.5
1.0
0.5
0.0
-0.5
Specificity



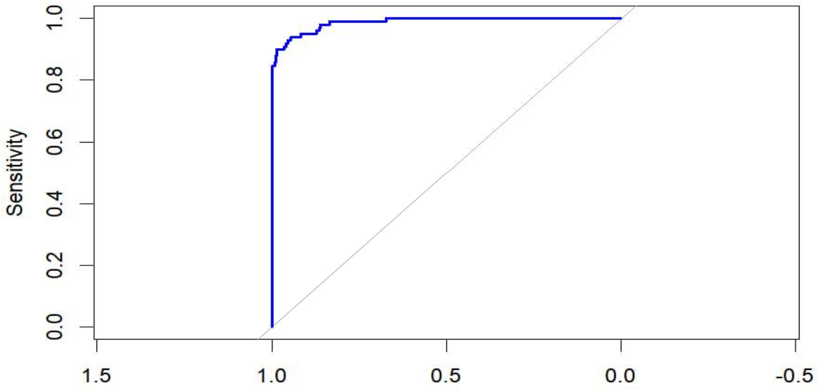
Residuals vs Fitted
Residuals
100
0
-200
-400
68634
17917
12939
-60
-40
-20
0
20
Predicted values
glm(Class ~ .)



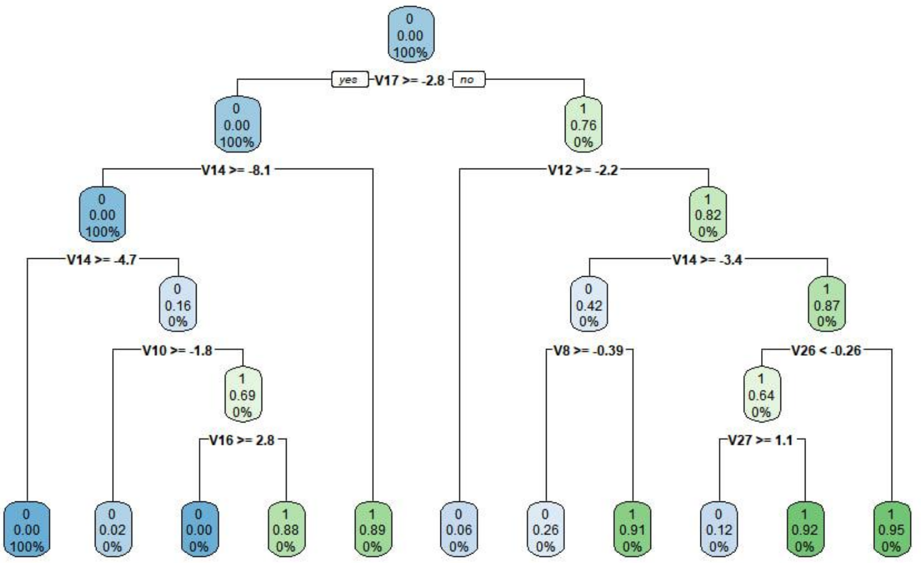
Normal Q-Q
Std. Pearson resid.
100
0
-200
-400
68634
17917
12939
-4
-2
0
2
4
Theoretical Quantiles
glm(Class ~ .)



Scale-Location
√|Std. Pearson resid.|
20
15
10
5
0
12939
17917
68634
-60
-40
-20
0
20
Predicted values
glm(Class ~ .)



Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1.5
1.0
0.5
0.0
-0.5



0 0.00 100%
yes
V17 >= -2.8
no
0 0.00 100%
1 0.76 0%
V14 >= -8.1
V12 >= -2.2
0 0.00 100%
1 0.82 0%
V14 >= -4.7
V14 >= -3.4
0 0.16 0%
0 0.42 0%
1 0.87 0%
V10 >= -1.8
V8 >= -0.39
V26 < -0.26
1 0.69 0%
1 0.64 0%
V16 >= 2.8
V27 >= 1.1
0 0.00 100%
0 0.02 0%
0 0.00 0%
1 0.88 0%
1 0.89 0%
0 0.06 0%
0 0.26 0%
1 0.91 0%
0 0.12 0%
1 0.92 0%
1 0.95 0%

Section 6 ~ References

[1] International Journal of Engineering Research & Technology (IJERT) IJERTV8IS090031 ISSN: 2278-0181 http://www.ijert.org

[2] .A machine learning based credit card fraud detection using the GA algorithm for feature selection ~ Emmanuel Ileberi, Yanxia Sun & Zenghui Wang

[3] - https://scholarworks.rit.edu/cgi/viewcontent.cgi?article=12455&context=theses

[4] The IEEE website. [Online]. Available: http://www.ieee.org/

[5] Panesar A. Machine Learning and AI on Credit_Card_Fraud_Detection_System (1–73) Apress, Coventry, UK (2019)