CKME 136 Data Analytics: Capstone Course - Final Report

Palak Mehta (# 501003503)

Supervisor: Tamer Abdou, PhD

Ryerson University

CKME 136 Data Analytics: Capstone Course
Final Report
Chang School of Continuing Education, Ryerson University

Title: How the credit card companies can save their customers money by applying machine learning algorithms?
Data Set: Credit Card Fraud Detection

Palak Mehta
501003503
Supervisor: Tamer Abdou, PhD Email: tamer.abdou@ryerson.ca

Table of Contents

Introduction

These days the demand of credit cards have substantially increased on both online and offline purchases because of the rapid development of the e-commerce and the banking system. With the recent advancements in FinTech apps and Touchless payment number of credit transactions have exploded.

While the credit card offers obvious benefits like easier access to credit, earning rewards, better spending tracking and online shopping, it comes with its shortcomings too. A credit card is susceptible to cybercriminals which is causing credit card fraud.

When a person uses other person's credit card for his personal use and the authorized person of the credit card is unaware about it that is known as credit card fraud. This unauthorized access to credit card results in financial loss to the company and the customer. Fraud transactions are on the rise and reduce the faith everyone has on the process. To keep the system safe, credit card providers need strong fraud detection system but at the same time not to add too many hoops for genuine customers to jump through.

With this requirement in mind, a machine learning model will be prepared to address the research question- How the credit card companies can save their customers money by applying machine learning algorithms? In this project, an attempt will be made to classify Credit card transactions as legitimate or fraudulent.

Supervised Machine Learning classification algorithms will be applied on resampled dataset for classifying fraud and non-fraud transactions. Performance of various algorithms will be compared, and the optimal model which will be statistically better and give better performance in terms of accuracy will be determined. The optimal model for classifying legitimate and fraudulent transactions will be compared against benchmark model also.

Literature Review

As credit card fraud has become a major problem in the financial market, many credit cards companies have spent an enormous amount of money and formed the teams of the human experts to create fraud detection systems. Credit card fraud detection has drawn a lot of research interest and researchers has used few different tools to detect credit card fraud. Numerous literatures pertaining to this have been published that needs to be reviewed to ensure that this project is properly grounded in best practices. All the research papers referenced are set out below. See references at the end for full citations.

M. Zareapoor, P. Shamsolmoali et al., "Application of credit card fraud detection: Based on bagging ensemble classifier," [1]

The authors have presented an application pertaining to the bagging ensemble for detecting credit card fraud. The ensemble approach is formed on the basis of decision tree algorithm that had been used for the experimental step. Besides this, this paper highlights the detailed description of the methods used such as (NB) Naïve Bayes, (SVM) Support Vector Machine and (KNN) K Nearest Neigbor. 10-fold cross validation was used to evaluate the experiment. The results show that the bagging classifier based on the decision tree showed the best results.

Andrea Dal Pozzolo, Olivier Caelen, Yann-A¨el Le Borgne, Serge Waterschoot, and Gianluca Bontempi. Learned lessons in credit card fraud detection from a practitioner perspective [PCB+14] [2]

The authors have used two different approaches for detecting fraud: static approach- model used for detection is used in a seasonal manner and online approach highlights that the model is updated immediately once the new transaction data arrives. The study recommends that online learning approach is a preferable approach as the fraud behavior fluctuates time to time. They have suggested that Average Precision (AP), Area Under Curve (AUC), and Precision Rank are the best measures for detecting fraud.

[AAO17] John O. Awoyemi, Adebayo Olusola Adetunmbi, and Samuel Adebayo Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis[3]

This paper highlights the resampling technique (SMOTE) Synthetic Minority Over-sampling Technique which was applied on credit card transaction data after performing machine learning algorithms such as K- Nearest Neighbor, Logistic regression and Naïve Byes. Out of the 3 algorithms K-Nearest Neighbor showed the best results. Recall, precision, balanced classification rate, specificity, and Mathews correlation coefficient were used to measure the performance.

A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, and P. Beling, "Deep learning detecting fraud in credit card transactions,"[4]

In this study the authors used six classifiers on the dataset before and after the pre-processing phase. After applying the under-sampling technique results were significantly improved. Precision and recall were used to measure the performance of the classifiers applied on both the datasets.

The results show that precision for all the classifiers were significantly increased after using the under sampled dataset.

K. Seeja and M. Zareapoor, "Fraudminer: A novel credit card fraud detection model based on frequent itemset mining,"[5]

This study highlights the use of matching algorithm to compare the new and existing patterns of a new transaction for each customer. Legal and fraud patterns were formulated for each customer. After applying the Apriori algorithm authors used the largest frequent itemset to determine the legal and fraud patterns for each customer. So, whenever a customer makes a new transaction it is compared with both the patterns to find if it is a legal or fraudulent transaction.

"Credit card Fraud Detection with a neural network " by Ghosh and Reilly. IEEE"[6]

The authors have introduced neural network for detecting credit card fraud. A detection system which was trained on a huge sample size of labeled credit card account transactions was developed. Fraud occurred due to lost cards, stolen cards, application fraud, counterfeit fraud, mail-order fraud and non-received issue fraud were the part of those transactions.

Dataset

The dataset used for this analysis is sourced from the Credit Card Fraud Detection Dataset from Kaggle (https://www.kaggle.com/mlg-ulb/creditcardfraud). This dataset was created during a research collaboration between Worldline and the Machine Learning Group (http://mlg.ulb.ac.be) of ULB (Université Libre de Bruxelles) on big data mining and fraud detection.

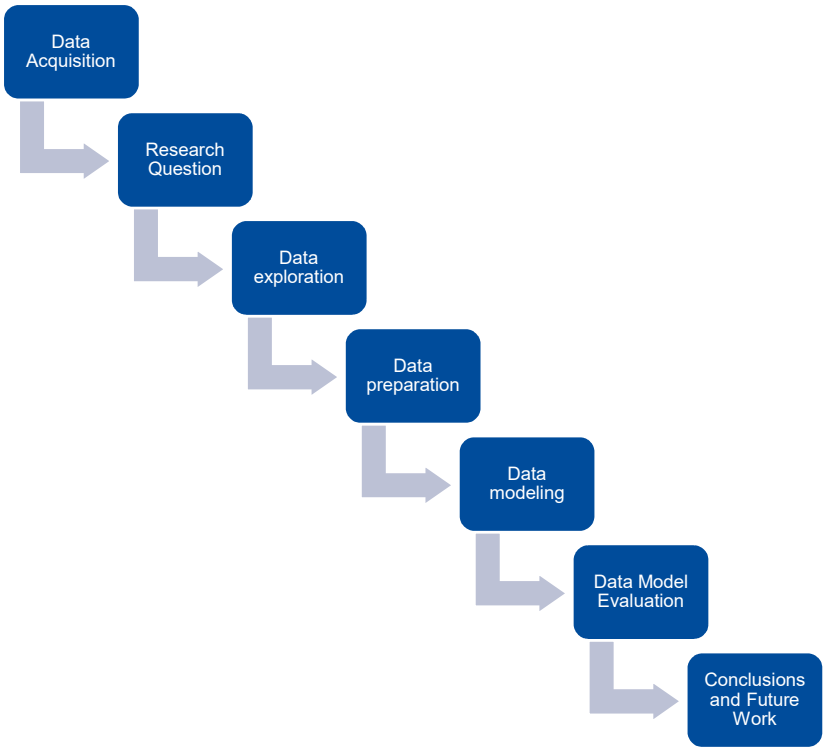
This dataset consists of 284,807 transactions made by European cardholders for two days in the month of September 2013. Out of the 284,807 transactions,492 (0.172%) are the fraudulent transactions making the dataset highly unbalanced. There are no missing values in the dataset so no need to handle those.

The dataset includes 30 numerical predictor variables and 1 numerical target variable. Due to privacy reasons, other than time, amount and Class, the information about the other attributes (V1-V28) is missing. The description of the data includes that all the V1-V28 features in the dataset are reconstructed after applying Principal Component Analysis (Dimensionality Reduction Technique) but does not include time and amount. In order to apply a PCA transformation all the features need to be scaled and in this dataset all the attributes except time and amount have been previously scaled.

S. No.	Feature	Description
1.	Time	Time in seconds to specify the elapses between current transaction and first transaction
2.	Amount	Transaction Amount
3.	Class (response variable)	0-Not Fraud 1-Fraud

Table 1: Non-Anonymized Attributes of Credit Card Fraud Detection Dataset

Approach



Step1: Data Acquisition

It is necessary that the credit card providers can classify credit card transactions as legitimate and fraudulent so that the customers do not have to pay the money for the items not purchased by them. So to solve this problem the publicly available dataset Credit Card

Fraud Detection Dataset has been downloaded from Kaggle (https://www.kaggle.com/mlg-ulb/creditcardfraud) to be used for the analysis.

Step 2: Research Question

The main aim behind this credit card fraud detection research is to classify and validate whether a specific transaction is legitimate or fraudulent which can help the banks in saving customers money. So, the guiding research question is the following: How the credit card companies can save their customers money by applying machine learning algorithms?

Step 3: Data Exploration

- During exploratory data analysis it was discovered that there were no missing values in the dataset.
- Legitimate class heavily dominated the fraudulent class which reflected that the data was unbalanced.
- Dataset consisted of 284,807 transactions. The average value of the transactions is $88.35 whereas the largest amount of the transaction is $25691.16. So, the distribution of the amount feature was right skewed, and distribution of time feature came from the bimodal distribution.
- Before data was uploaded to Kaggle, the anonymized variables (V1-V28) were modified in the form of a PCA (Principal Component Analysis), so all the anonymized features except time and amount have been scaled already. Therefore, time and amount features were focused in the study.
- Heatmap Correlation matrix was prepared to understand the variables and see the correlation between the predictor variables with respect to the class (target) variable before fixing the imbalance problem.

Step 4: Data Preparation

- Time and amount features are scaled like other features in order to bring all the features on the same magnitude and scaled features have been included in the data frame.
- Outliers in this data frame were the fraudulent transactions only that deviate from the behavior of normal transactions. So, no outliers were removed.
- To balance the class distribution which means to have equal cases of legitimate and fraud transactions various resampling techniques have been applied-Random Under sampling, Random Oversampling and SMOTE.
- Resampled data were split randomly into training and testing sets. 70% of the credit card data set was allocated for training and remaining 30% of the credit card data was allocated for testing.

- Feature selection was done on the resampled under sampled, over-sampled and SMOTE datasets by using Random Forest Classifier to improve the performance of the model, those features were identified which contributed much to the target variable, increased the speed of modelling process, improved the performance of the model and data subset was created with only the most important features.

Step 5: Data Modelling

Random Forest Classifier, Logistic Regression and KNN classifiers were applied on the resampled feature selected datasets in order to identify which classifier gave better accuracy in terms of classifying legitimate and fraudulent transactions.

Step 6: Data Model Evaluation and Results

- Evaluation metrics like Accuracy, Precision, Recall, F1 score, ROC AUC score and confusion matrix were used to explain the performance of the model.
- Model which gave the better performance was compared against the benchmark model. Random Forest Classifier was considered as the benchmark model and the performance of other models were compared against the benchmark model in order to identify which classifier gives a better accuracy in terms of classifying legitimate and fraudulent transactions.
- Cross validation technique was used to test the accuracy, stability and effectiveness of machine learning models.

Step 7: Conclusions and Future Work

Based on the results derived from the research, conclusion and future work was recommended.

Data Exploration

The dataset used in this analysis consisted of 284,807 rows and 31 columns in the credit card data frame. Out of 31 variables there are 30 numerical predictor variables and 1 numerical target variable. Due to privacy reasons, other than time and amount, the information about the other attributes (V1-V28) is missing. The description of the data includes that all the V1-V28 features in the dataset are reconstructed after applying Principal Component Analysis (Dimensionality Reduction Technique) but does not include time and amount. In order to apply a PCA transformation all the features need to be scaled and in this dataset all the attributes except time and amount have been previously scaled.

Exploratory analysis details

Target Variable	Class
Predictor Variables	V1-V28, Amount, Time
Rows and Columns	284807 rows and 31 columns
Missing Values	Null
Unique Values	0 1
Count of Unique Values	0-legitimate Transactions-284315 1-Fraudulent Transactions-492
Total No of Legitimate Transactions Total No of Fraudulent Transactions	0-legitimate Transactions-284315 1-Fraudulent Transactions-492

Table 2: Credit Card Data Frame Details

Distribution of Class Attribute

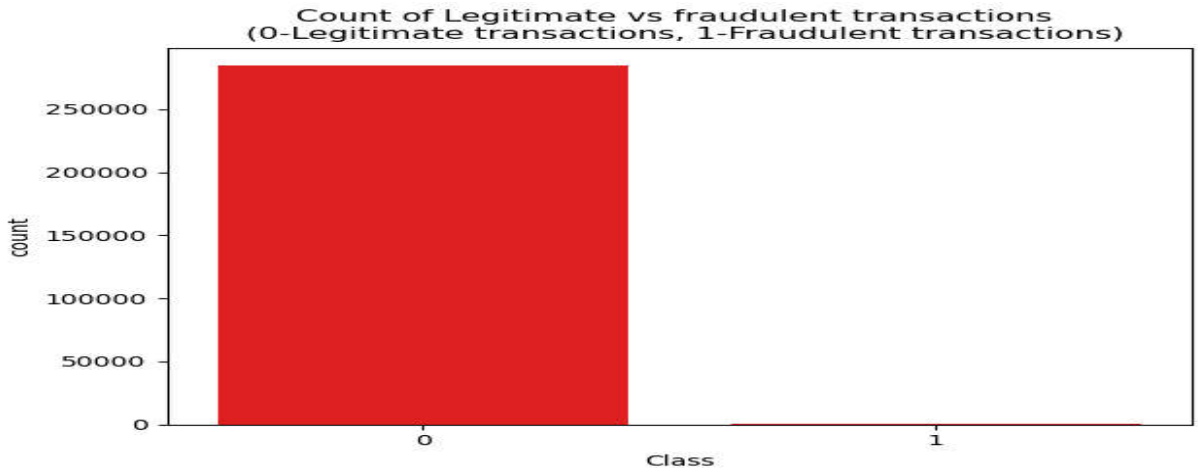


Figure 1 Distribution of Class Attribute (target variable)

The above figure shows that there is severe skewness in the class distribution with about 99.83 percent of the transactions marked as legitimate and 0.17 percent of the transactions marked as fraudulent. As total number of legitimate transactions are 284315 and total number of fraudulent

transactions are 492 out of total 284807 transactions which means that the data is highly imbalanced as most of the transactions are legitimate.

Correlation Matrix

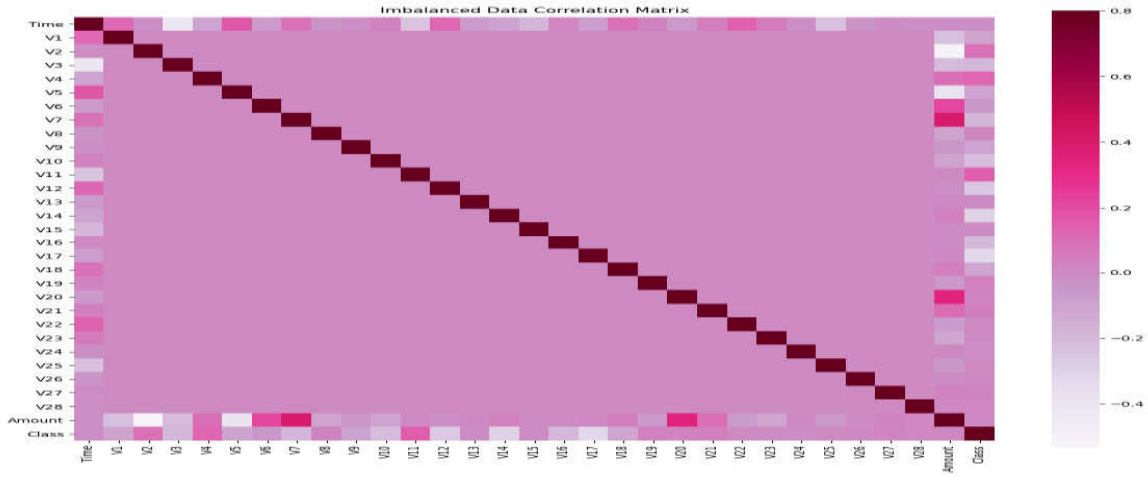


Figure 2: Heatmap Correlation Matrix on the Credit Card data frame (Imbalanced

Heatmap correlation matrix was applied on the original credit card data frame before fixing the imbalance problem to see the correlation between our predictor variables with regards to our target variable 'Class'. From the above figure it was discovered that there was no notable correlation between the features with regards to target variable 'Class'. This can be probably due to huge class imbalance.

Visualization of the dataset features

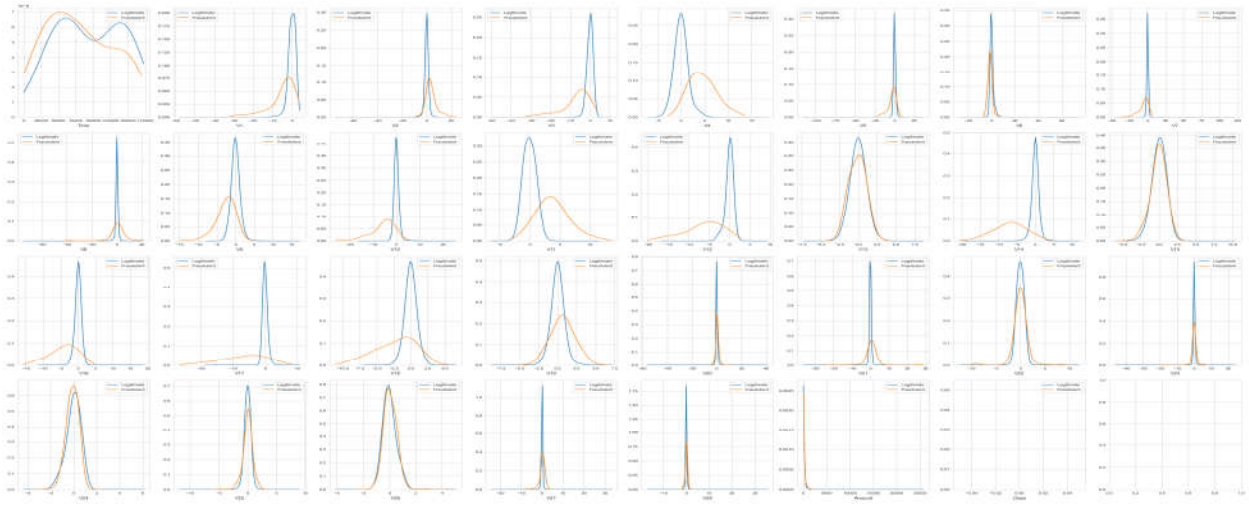


Figure 3: Kernel Density Plot for all the features of the Credit Card Data Frame

Density Plots were prepared to visualize the distribution of all the features of the credit card data frame over a continuous interval or time period.

Data Preparation

Outliers

Outliers in this credit card data frame are the fraudulent transactions only that deviate from the behavior of normal transaction. IQR-Interquartile Range was calculated for every feature of the data frame. Outliers in this case were the observations that were below (Q1-1.5*IQR) or above (Q3 + 1.5*IQR). After removing the outliers there were no fraudulent transactions so it was found that outliers in this credit card data frame were frauds only. Therefore, no outliers were removed.

Amount and Time Features

Due to privacy reasons, other than time and amount, the information about the other attributes (V1-V28) is missing. The description of the data includes that all the V1-V28 features in the dataset are reconstructed after applying Principal Component Analysis (Dimensionality Reduction Technique) but does not include time and amount. In order to apply a PCA transformation all the features need to be scaled and in this dataset all the attributes except time and amount have been previously scaled. So, amount and time features have been focused.

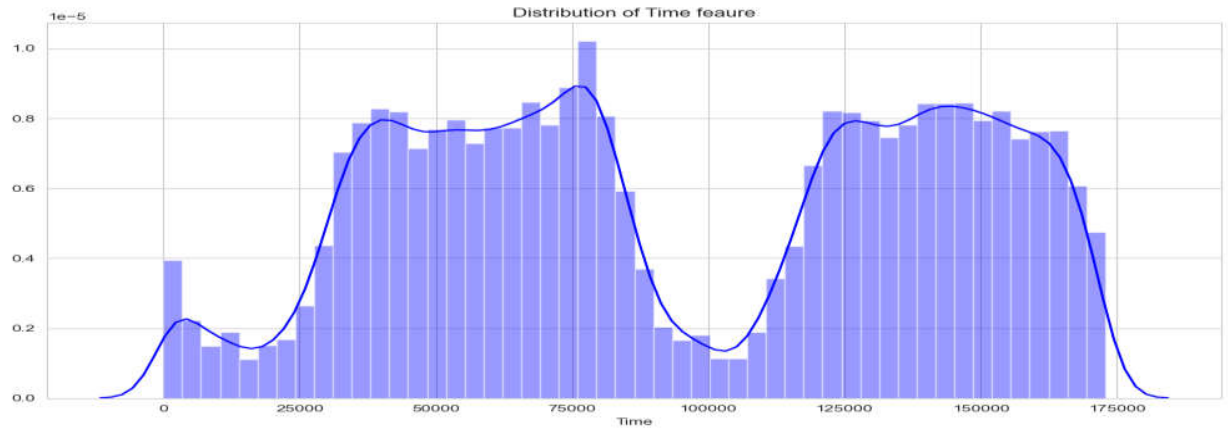


Figure 4: Visualizing the Time Feature

The above figure shows that distribution of time feature is bimodal. Time is recorded in seconds since the first transaction in the dataset. Therefore, it was concluded that this data frame included all the transactions recorded over the course of two days.

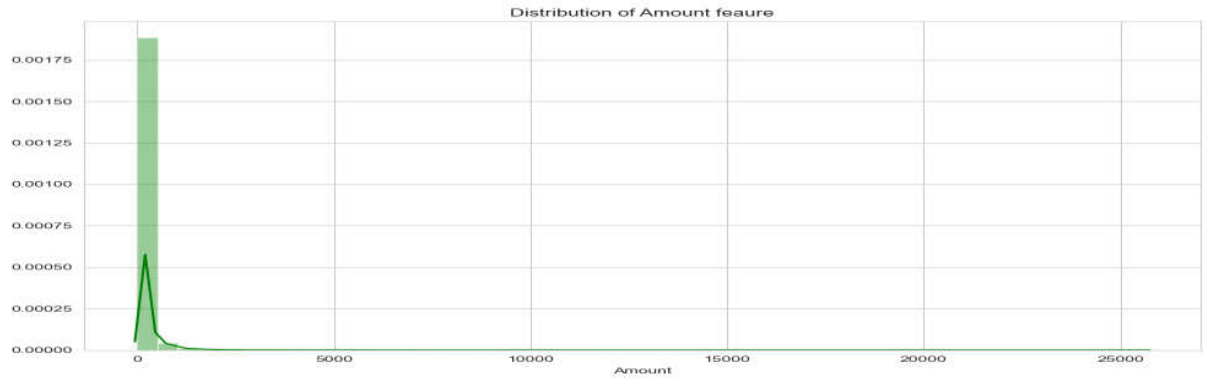


Figure 5: Visualizing the Amount Feature

The above figure shows that distribution of the amount of all the transactions is heavily right skewed. The amount of most of the transactions are relatively small and a very small fraction of the transactions are close to the maximum.

Scaling the Amount and Time Features

All the attributesV1-V28 except time and amount have been previously scaled in order to bring all the features on the same level of magnitude. So, amount and time features have been scaled using Robust Scaling Technique in order to normalize both the features in a particular range. Robust Scaling removes the median and scale the data according to the quantile range. This technique is robust to outliers.

Scaled amount and scaled time features were included in the credit card data frame and original amount and time features were dropped from the data frame.

Resampling techniques and Feature Selection

As total number of legitimate transactions are 284315 and total number of fraudulent transactions are 492 out of total 284807 transactions which means that the data is highly imbalanced as most of the transactions are legitimate. So, resampling techniques were applied to balance in the imbalanced data. Resampling techniques turn the dataset into a more balanced one by adding instances to the minority class or reducing instances from the majority class.

Feature Selection using Random Forest Classifier comes under the category of embedded methods which means that they are less prone to overfitting, generalize better, easily interpretable and gives more accurate results. The tree-based strategies used by Random Forest naturally ranks by how well they improve the purity of the node. This mean decrease in impurity over all trees is called

Gini Impurity. Subset of the most important features were created by pruning trees below a specific node.

The following resampling techniques were applied to balance the imbalanced credit card data:

I. Random Under sampling technique:

This technique randomly selects samples from the majority class to delete from the training set. Here samples were removed randomly from the overrepresented class (legitimate transactions).

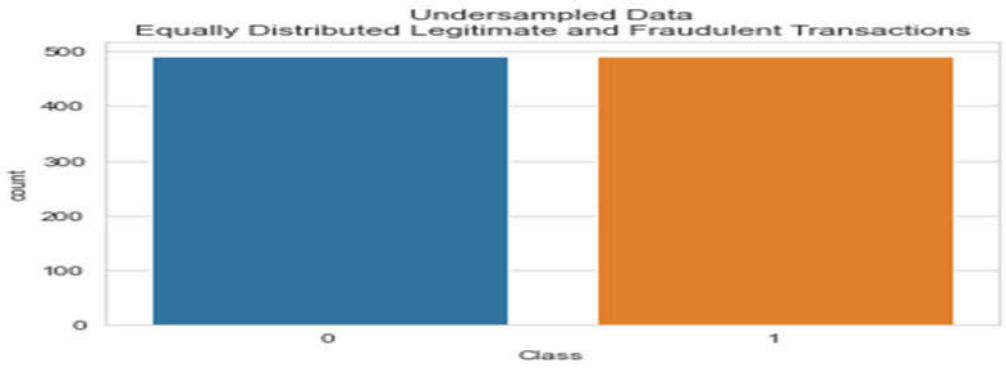


Figure 6: Legitimate and Fraudulent Transactions After applying Random Under sampling

The above figure shows that after applying under sampling the number of legitimate and fraudulent transactions are 492 each. So, the total number of transactions in under sampled credit card dataset are 984.

Split the Under sampled Credit Card data frame-

Under sampled credit card dataset was split randomly into training and test sets. 70% of the under sampled credit card data set was allocated for training and remaining 30% of the under sampled credit card data was allocated for testing. Model fitting was done on the training data and test data was used for model prediction.

Feature Selection in Under Sampled Credit Card Data using Random Forest Classifier-

Important features were selected according to the scores assigned as per their relative importance for making the prediction.

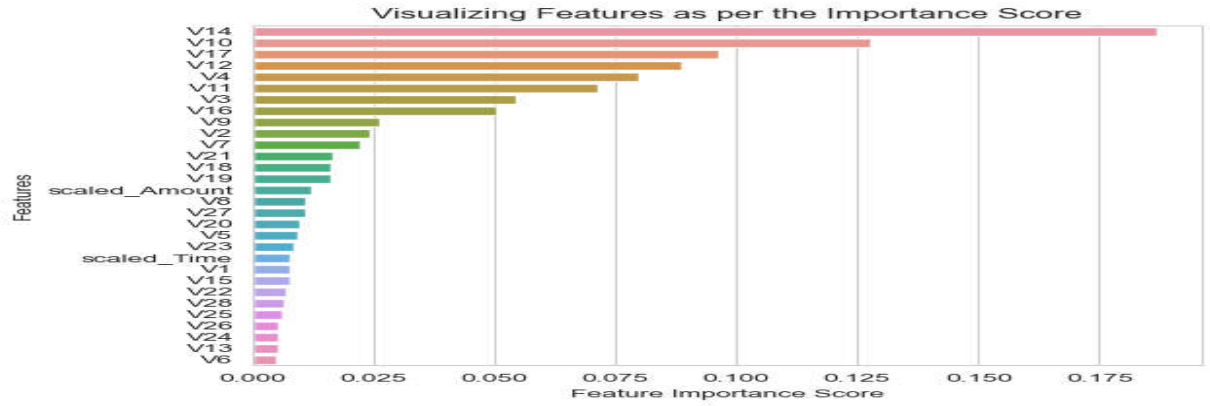


Figure 7: Features of the Random Under sampled Data as per the Importance Scores

1.	**V3**
2.	V4
3.	V10
4.	V11
5.	V12
6.	V14
7.	V16
8.	V17

Table 3: Selected Features in Random Under sampled Data

II. Random Over sampling technique:

This technique randomly duplicate records from the minority class. Here new instances of the minority class were created randomly by replicating current samples in order to increase the minority count (fraudulent transactions).

Split the Dataset before applying Oversampling – Credit card data frame was split randomly into training and test sets before applying oversampling technique to prevent the overfitting and poor generalization to the test data. 70% of the credit card data set was allocated for training and remaining 30% of the credit card data was allocated for testing.

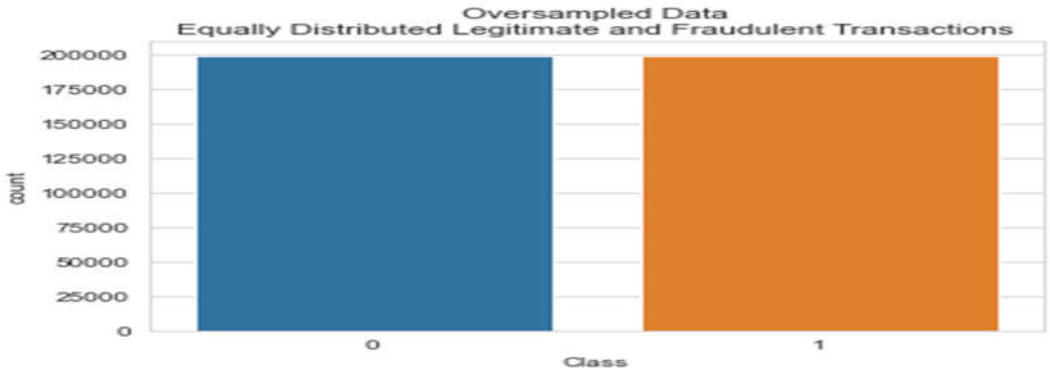


Figure 8: Legitimate and Fraudulent Transactions After applying Random Over sampling

The above figure shows that after applying random over sampling the number of legitimate and fraudulent transactions are 199027 each. So, the total number of transactions in random over sampled credit card dataset are 398054.

Feature Selection in Random Over Sampled Credit Card Data using Random Forest Classifier

Important features were selected according to the scores assigned as per their relative importance for making the prediction.

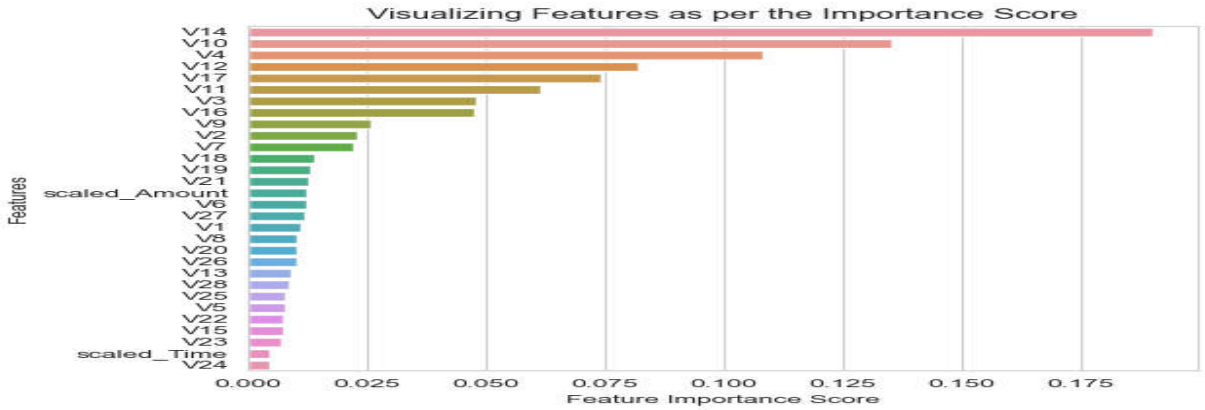


Figure 9: Features of the Random Over sampled Data as per the Importance Scores

1.	**V3**
2.	V4

1.	V3
3.	V10
4.	V11
5.	V12
6.	V14
7.	V16
8.	V17

Table 4: Selected Features in Random Over sampled Data

III. SMOTE (Synthetic Minority Oversampling Technique)

SMOTE technique generates synthetic elements from the minority class (fraudulent transactions) instead of creating the copies based on those that exist already**.** New samples were generated only in the training set to ensure our model generalized to unseen data.

Split the Dataset before applying SMOTE technique – Credit card data frame was split randomly into training and test sets before applying SMOTE technique to prevent the overfitting and poor generalization to the test data. 70% of the credit card data set was allocated for training and remaining 30% of the credit card data was allocated for testing.

After applying SMOTE technique the number of legitimate and fraudulent transactions are 199027 each.

Feature Selection in SMOTE Credit Card Data using Random Forest Classifier-

Important features were selected according to the scores assigned as per their relative importance for making the prediction.

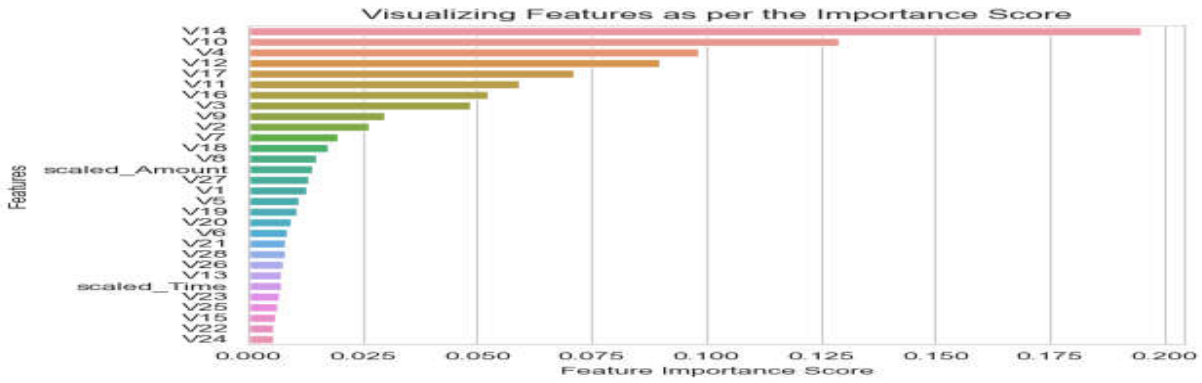


Figure 10: Features of the SMOTE Data as per the Importance Scores

1.	**V3**
2.	V4
3.	V10
4.	V11
5.	V12
6.	V14
7.	V16
8.	V17

Table 5:Selected Features in SMOTE Data

Data Modelling

Application of the Classifiers on the Random Under Sampled, Random Over-sampled and SMOTE Feature Selected Data-

Random Forest Classifier was applied on the under sampled, over sampled and SMOTE feature selected datasets. This algorithm is based on supervised learning algorithm and used decision trees for classification of the dataset. This algorithm is an ensemble of decision trees that predict collectively if a transaction was fraudulent or not. This algorithm has been found to provide a good estimate of the generalization error, resistant to overfitting and easy to use.

Logistic Regression Classifier was applied on the under sampled, over-sampled and SMOTE feature selected datasets. Logistic regression is a statistical model which helps in minimizing the cost of how wrong the prediction is. This algorithm is well suitable for binomial outcomes. In credit card data frame target variable Class is a binomial outcome 0(legitimate transactions) and 1(fraudulent transactions).

K Nearest Neighbors (KNN) Classifier has been considered as one of the best classifier algorithms that have been used in credit card fraud detection by computing its nearest point. It requires a distance or similar the measure defined between two data instances. So, it was applied on the under sampled, over-sampled and SMOTE feature selected datasets.

Classifiers	Accuracy	Precision	Recall	F1 Score	ROC AUC Score	Cross Validation Accuracy 5 Folds	Cross Validation Accuracy 10 Folds
Random Forest	93.58%	98%	88%	93%	97.26%	93.75%	93.61%
Logistic Regression	93.58%	96%	90%	93%	96.95%	92.88%	93.17%
KNN	93.24%	96%	89%	92%	96.7%	92.15%	92.44%

Table 6: Random Under Sampled Feature Selected Data Set- Results

Classifiers	Accuracy	Precision	Recall	F1 Score	ROC AUC Score	Cross Validation Accuracy 5 Folds	Cross Validation Accuracy 10 Folds
Random Forest	99.96%	94%	81%	87%	95.05%	100.0%	100.0%
Logistic Regression	97.44%	6%	92%	11%	98.2%	93.87%	93.88%
KNN	99.89%	64%	87%	74%	93.52%	99.95%	99.96%

Table 7:Random Over Sampled Feature Selected Data Set-Results

Classifiers	Accuracy	Precision	Recall	F1 Score	ROC AUC Score	Cross Validation Accuracy 5 Folds	Cross Validation Accuracy 10 Folds
Random Forest	99.91%	69%	86%	77%	97.58%	99.96%	99.97%
Logistic Regression	97.27%	6%	92%	11%	98.07%	93.53%	93.53%

Classifiers	Accuracy	Precision	Recall	F1 Score	ROC AUC Score	Cross Validation Accuracy 5 Folds	Cross Validation Accuracy 10 Folds
KNN	99.52%	26%	89%	40%	95.36%	99.75%	99.77%

Table 8: SMOTE Feature Selected Data Set-Results

Data Model Evaluation and Results

I. Random Under sampled Feature Selected Dataset-

From the above results it was clearly visible that the Random Forest Classifier (benchmark model) and Logistic Regression outperformed KNN classifier for this problem. Both the classifiers have the same accuracy of 93.58% and F1 score of 93%. The greater the accuracy and F1 score the better the model is which means that Random Forest (benchmark model) and Logistic regression scores were very promising for the under sampled dataset. KNN results in terms of accuracy, precision, recall and F1 score were not promising when compared to the Random Forest (benchmark model).

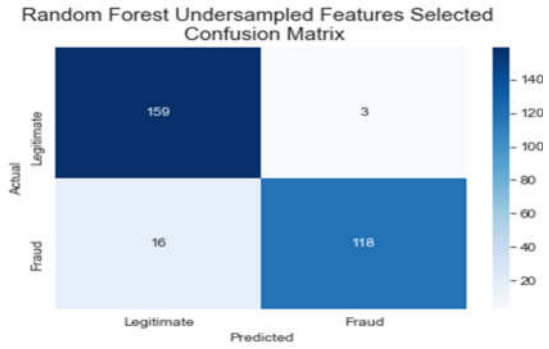


Figure 11: Random Forest Confusion Matrix

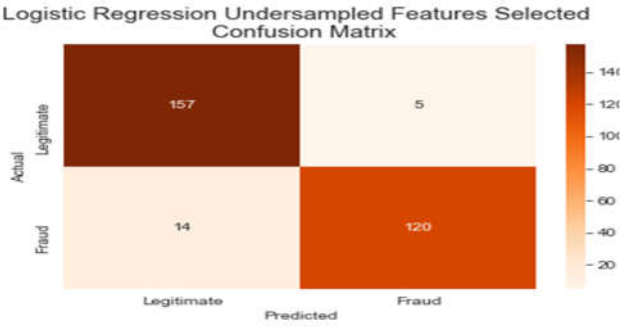


Figure 12: Logistic Regression Confusion Matrix

From the Random Forest confusion matrix, it was found that 118 of the 134 fraudulent transactions were captured whereas Logistic Regression has captured 120 of the 134 fraudulent transactions. Both the models had captured optimal amount of fraudulent transactions and have high true positive rate and low false positive rate which was really good.

Therefore, it was concluded that Random Forest (the benchmark model) and Logistic Regression turned out to be the best algorithms in under sampled dataset for classifying and validating whether a specific transaction is legitimate or fraudulent and help the banks in saving customers money.

II. Random Over sampled Feature Selected Dataset

From the above results it was clearly visible that the Random Forest Classifier clearly outperformed Logistic Regression Classifier and KNN classifier for this problem. Random Forest Classifier had the accuracy of 99.96%.as compared to Logistic regression accuracy of 97.44% and KNN accuracy of 99.89%. There was not much difference between the accuracy of Random Forest and KNN classifier but the F1 score of Random Forest model was 87% which was much higher than the KNN F1 score of 74%. The greater the accuracy and F1 score the better the model is which means that Random Forest (benchmark model) scores were very promising for the over sampled dataset as compared to Logistic regression and KNN classifier.

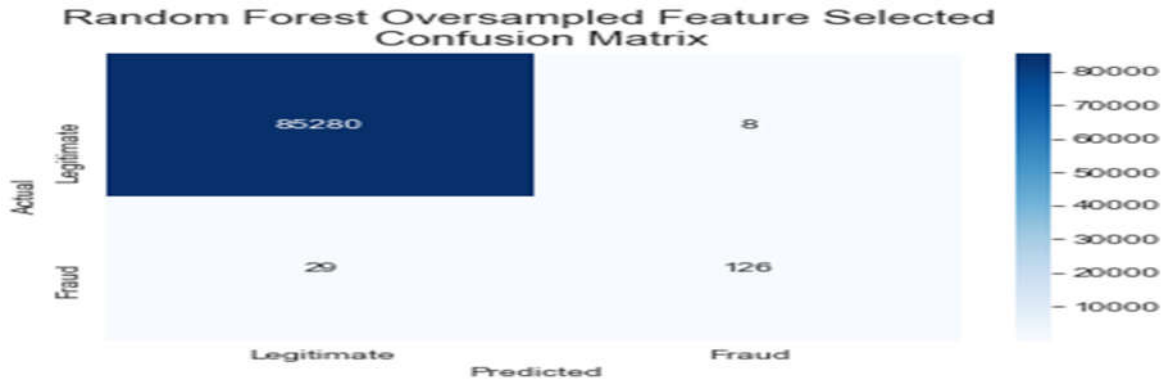


Figure 13: Random Forest Oversampled Feature Selected Confusion Matrix

From the Random Forest confusion matrix, it was found that 126 of the 155 fraudulent transactions were captured correctly. Random Forest had captured optimal amount of fraudulent transactions and have high true positive rate and low false positive rate which was really good.

Hence it was concluded that Random Forest Classifier (benchmark model) turned out to be the best algorithm in oversampled dataset for classifying and validating whether a specific transaction is legitimate or fraudulent and help the credit card companies in saving their customers money.

III. SMOTE Feature Selected Dataset

From the above results it was clearly visible that the Random Forest Classifier clearly outperformed Logistic Regression Classifier and KNN classifier for this problem. Random

Forest Classifier had the accuracy of 99.91%.as compared to Logistic regression accuracy of 97.27% and KNN accuracy of 99.52%. Precision and F1 scores of Random Forest were much better than the Logistic regression and KNN scores. Random Forest F1 score of 77% was much higher than the Logistic F1 score of 11% and KNN F1 score of 40%.

It means that Random Forest (benchmark model) scores were very promising for the SMOTE dataset as compared to Logistic regression and KNN classifier.

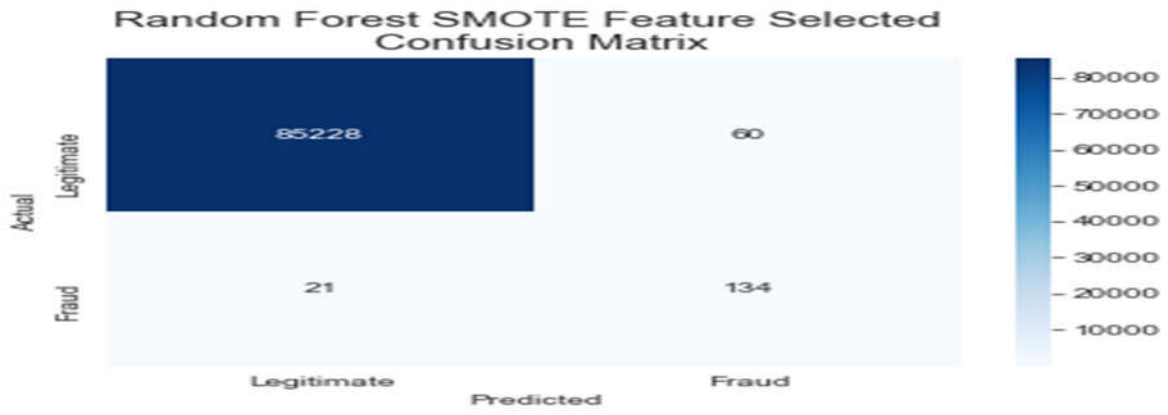


Figure 14: Random Forest Oversampled Feature Selected Confusion Matrix

From the Random Forest confusion matrix, it was found that 134 of the 155 fraudulent transactions were captured correctly. Random Forest had captured optimal amount of fraudulent transactions and have high true positive rate and low false positive rate which was really good.

Hence it was concluded that Random Forest Classifier (benchmark model) turned out to be the best algorithm in SMOTE dataset for classifying and validating whether a specific transaction is legitimate or fraudulent and help the banks in saving customers money.

Cross validation

It was decided to validate the models used in this study in order to have an assurance of the accuracy of the predictions that the models used were putting out.

Cross validation is a method of validating the model which splits the data in creative ways in order to obtain the better estimates of real-world model performance and reduce validation error. In this study it was used to validate the stability of the machine learning models, test the effectiveness of the models, make sure that they are not picking up too much on the noise, not overfitting and

estimate how the models will perform when predictions are to be made on the unseen data during training of the model.

This method is easy to understand, and it results in less biased estimate of the model skill than the other methods such as simple train/test split.

K Fold cross validation technique was used in this study to validate our models. Reducing the training data increases the risk of losing important patterns in the dataset. K Fold provides ample data for training and ample data for validation. As a result, it reduces biasness as most of the data is used for fitting and also reduces variance as most of the data is also used in validation set.

5 Folds and 10 Folds were used to validate the models used as they are widely used. As K gets larger, the difference in size between the training set and the resampling subsets gets smaller. By using K=5 and K=10 values were shown empirically to yield test error rate estimates that neither effect from excessive high bias nor from high variance.

By using 5 Folds and 10 Folds cross validation on all the 3 classifiers in under sampled feature selected, over-sampled feature selected and SMOTE feature selected datasets it was found that K Folds cross validation helped in providing more information about the algorithm's performance used in this study and more accurate estimate of performance of the models used. From the results shown in table- it was found that the models used were not overfitting and were not generalized as they generated pretty good accuracy such as RFC mean accuracy using 5 Folds was and 10 Folds in under-sampled dataset was 93.75% and 93.61%, in oversampled dataset was 100% each and in SMOTE dataset was 99.96%-5 Folds and 99.97%-10 Folds whereas other models such as Logistic Regression and KNN also generated pretty good accuracy.

Conclusion and Future Work

With the increased popularity of credit card system fraudulent transactions have increased at an alarming rate for over the past few years. To keep the system safe, credit card providers need strong fraud detection system but at the same time not to add too many hoops for genuine customers to jump through.

The overall objective behind this credit card fraud detection research was to ascertain the optimal algorithm for classifying and validating whether a specific transaction is legitimate or fraudulent which can help the banks in saving their customers money.

Imbalanced data was one of the challenges of the credit card fraud detection. In this research dataset was cleaned and integrated. When providing highly unbalanced class distribution data to the predictive model, the model gets biased towards the majority samples. Consequently, it misrepresents a fraudulent transaction as a genuine transaction. So, to deal with this class imbalance problem random under-sampling, random over-sampling and SMOTE resampling

techniques were applied to generate the balanced data. Dataset were randomly split into training and testing sets. Feature selection was performed using the Random Forest Classifier on the resampled datasets to improve the performance of the model, avoid overfitting and train the algorithms faster. In this research 3 machine learning algorithms-Random Forest Classifier, Logistic Regression and KNN classifiers were applied on the resampled data subsets consisting of the important features only in order to predict whether a credit card transaction is fraudulent or not. Set of evaluation metrics were used to evaluate the performance of the models.

The results from the machine learning classifiers show that Random Forest (the benchmark model) and Logistic Regression turned out to be the best algorithms in random under sampled feature dataset with the same accuracy of 93.58%, that Random Forest Classifier (benchmark model) turned out to be the best algorithm in random oversampled feature selected dataset with the accuracy of 99.96% and Random Forest Classifier (benchmark model) again turned out to be the best algorithm in SMOTE feature selected dataset with the accuracy of 99.91% for classifying and validating whether a specific transaction is legitimate or fraudulent. Therefore, based on this research credit card companies can implement these chosen machine learning models in order to detect whether a credit card transaction is fraudulent and save their customers money.

Future Work-

This research on detecting credit card fraud has great potential for future implications.

In this research 3 machine learning algorithms were applied on resampled feature selected datasets. For the future work other machine learning models like deep machine learning models can be implemented on with or without resampled data sets. In this study KNN classifier gave lesser promising results as compared to the benchmark model-Random Forest and Logistic regression. So the future researcher can do a detailed study to find out which factors contributed in giving good performance for Random Forest and Logistic regression model and poor performance for KNN model. Future researcher can also do the hyper parameter tuning to compare the performance of different machine learning algorithms by building a model for each possible combination of all of the hyperparameter values provided, evaluating each model, improvise the model performance and selecting the model architecture which produces the best accuracy.

References

[1] M. Zareapoor, P. Shamsolmoali et al., "Application of credit card fraud detection: Based on bagging ensemble classifier," Procedia computer science, vol. 48, no. 2015, pp. 679–685, 2015.

[2] [PCB+14] Andrea Dal Pozzolo, Olivier Caelen, Yann-A¨el Le Borgne, Serge Waterschoot, and Gianluca Bontempi. Learned lessons in credit card fraud detection from a practitioner perspective. Expert Syst. Appl., 41:4915–4928, 2014.

[3] [AAO17] John O. Awoyemi, Adebayo Olusola Adetunmbi, and Samuel Adebayo Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. 2017 International Conference on Computing Networking and Informatics (ICCNI), pages 1–9, 2017.

[4] A. Roy, J. Sun, R. Mahoney, L. Alonzi, S. Adams, P. Beling, "Deep Learning Detecting Fraud in Credit Card Transactions", In: Proc. of International Conference Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, pp.129- 134, 2018.

[5] Seeja KR, Zareapoor M, FraudMiner: a novel credit card fraud detection model based on frequent itemset mining. Sci World J, vol 2014, 2014.

[6] Ghosh, S., Reilly, D.L., (1994). Credit card fraud detection with a neural-network, in: Proceedings of the Twenty-Seventh Hawaii International Conference on System Sciences, IEEE, Wailea, HI, USA. pp. 621–630.

Appendix

Table of Figures:

Table of Tables