Project Report

VIVEKPANDIAN VEERAPANDIAN

VXV180020

Abstract

Now-a-days, Internet has become an important part of human's life, a person can shop, invest, and perform all the banking task online. Due to the rise and rapid growth of E-Commerce, use of credit cards for online purchases has dramatically increased and it caused an explosion in the credit card fraud. In an era of digitalization, credit card fraud detection is of great importance to financial institutions. In this presentation, we will analyze credit card fraud detection using several Machine Learning classification Algorithms. After analyzing, our aim is to compare all the different methods to handle imbalanced data. The obtained results from databases of credit card transactions show the power of these techniques in the fight against banking fraud comparing them to others in the same field. It was determined that the Support Vector Machine algorithm had the highest performance rate for detecting credit card fraud under realistic conditions.

Introduction:

Credit card fraud[1] is a form of identity theft that involves an unauthorized taking of another's credit card information for the purpose of charging purchases to the account or removing funds from it.

The financial services industry and the industries that involve financial transactions are suffering from fraud-related losses and damages[2]. 2016 was a banner year for financial scammers. In the US alone, the number of customers who experienced fraud hit a record 15.4 million people, which is 16 percent higher than 2015. Fraudsters stole about $6 billion from banks last year. A shift to the digital space opens new channels for financial services distribution. It also created a rich environment for fraudsters.

Although less than 0.1% of all credit card transactions are fraudulent, analysts predict that Credit card fraud losses incurred by banks and credit-card companies can surpass $12 billion in the United States in 2020. Evidently, there is a dire need for robust detection of card-present and card-not-present fraudulent transactions to minimize monetary losses.

Technically speaking, two approaches are used to combat these crimes, the fraud prevention systems (FPS) and fraud detection systems (FDS)[3]. The FPS are the pro-active hardware and software mechanisms to avoid the fraud occurrence, on the other hand, the FDS are used when the fraudsters have overtaken the FPS. The main objective of the FDS is identify the fraud as soon as possible in order to take the necessary actions to revert it.

The fraud detection problem will come under supervised classification task aiming to determine whether a new transaction is legitimate or fraudulent one. Many techniques have been applied to credit card fraud detection, artificial neural network, support vector machine , decision tree, random forest, naïve bayes, logistic regression, and K nearest neighbors. A comparative analysis of logistic regression and naïve bayes is carried out in [10].

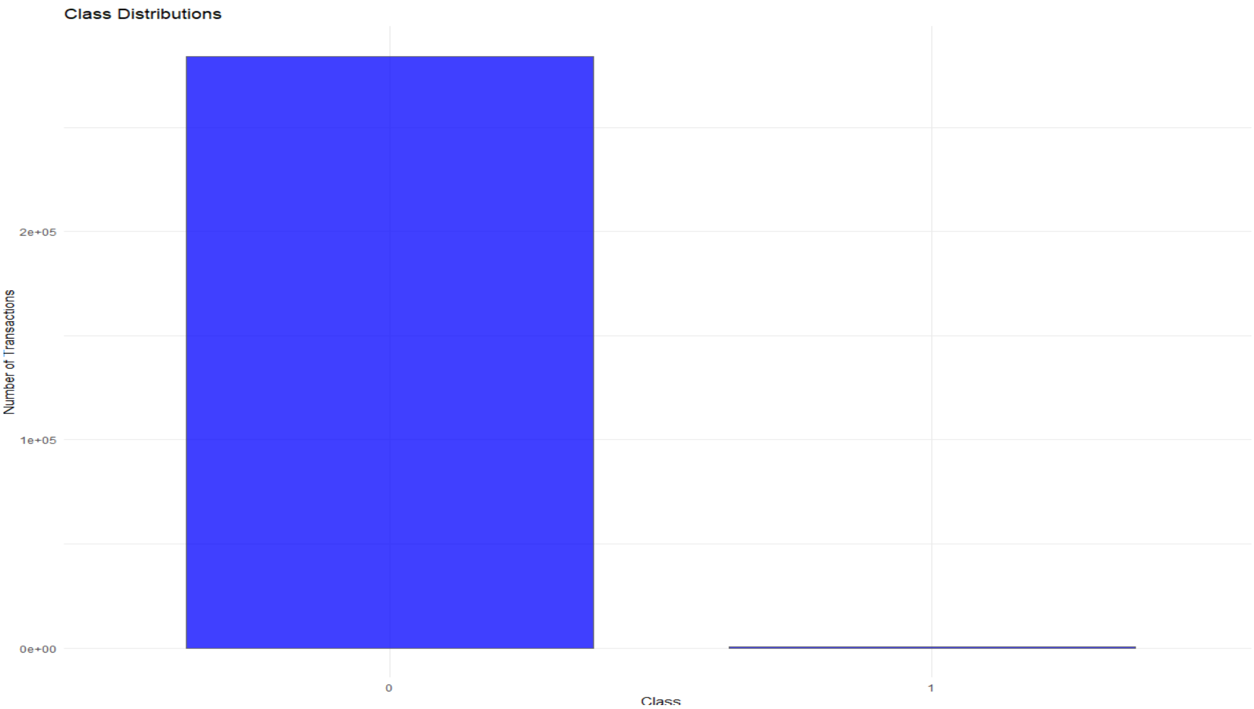
This project evaluates Seven advanced data mining approaches, support vector machines and random forests, KNN, together with logistic regression, as part of an attempt to better detect credit card fraud while neural network, naïve bayes and logistic regression is applied on credit card fraud detection problem.

Dataset:

The dataset[4] contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. There are no “NA” values in the dataset.

Data (Class):



Count of Fraudulent to Non-Fraudulent Datapoints

Number of Legitimate/Non-fraud Transactions: 284315

Percentage of Legitimate /Non-fraud Transactions: 98.8272

Number of Fraud transactions: 492

Percentage of Fraud transaction: 0.1727

This shows that the data is highly imbalanced.

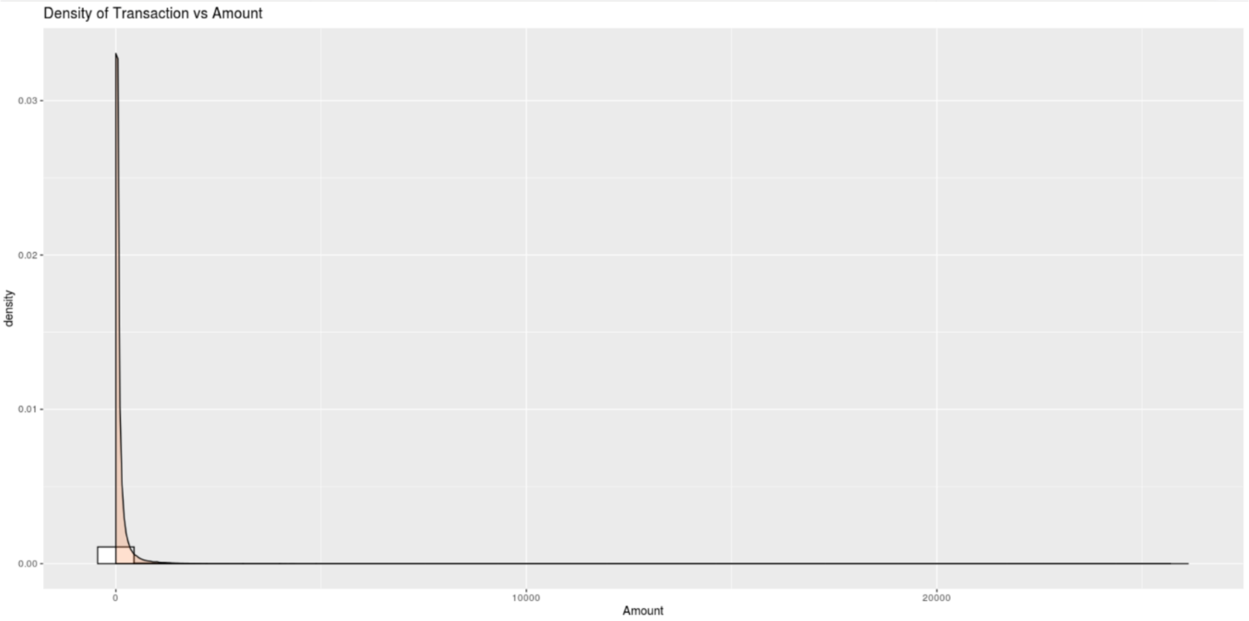
The imbalanced nature of the data can be rectified by one of the following methods:

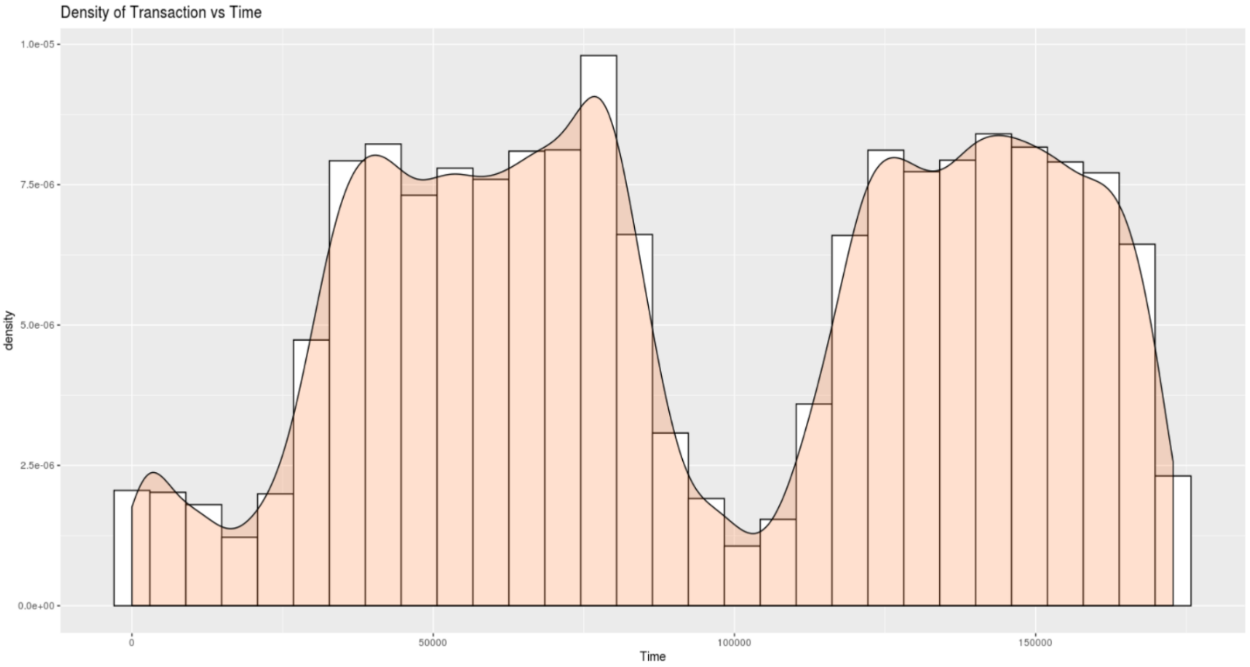
i) Random under sampler.

ii) Over sampling.

iii) Combined sampling (both over and under sampler)

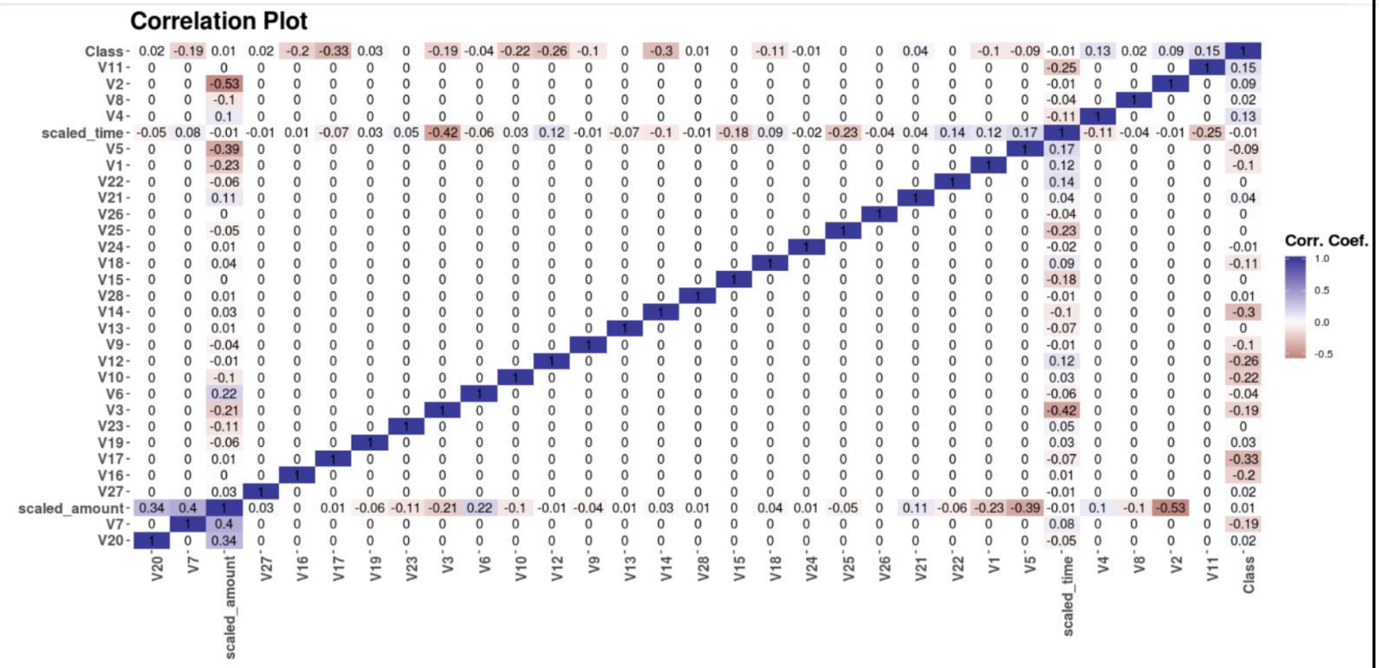
__Distribution:__


Density of Transaction vs Amount

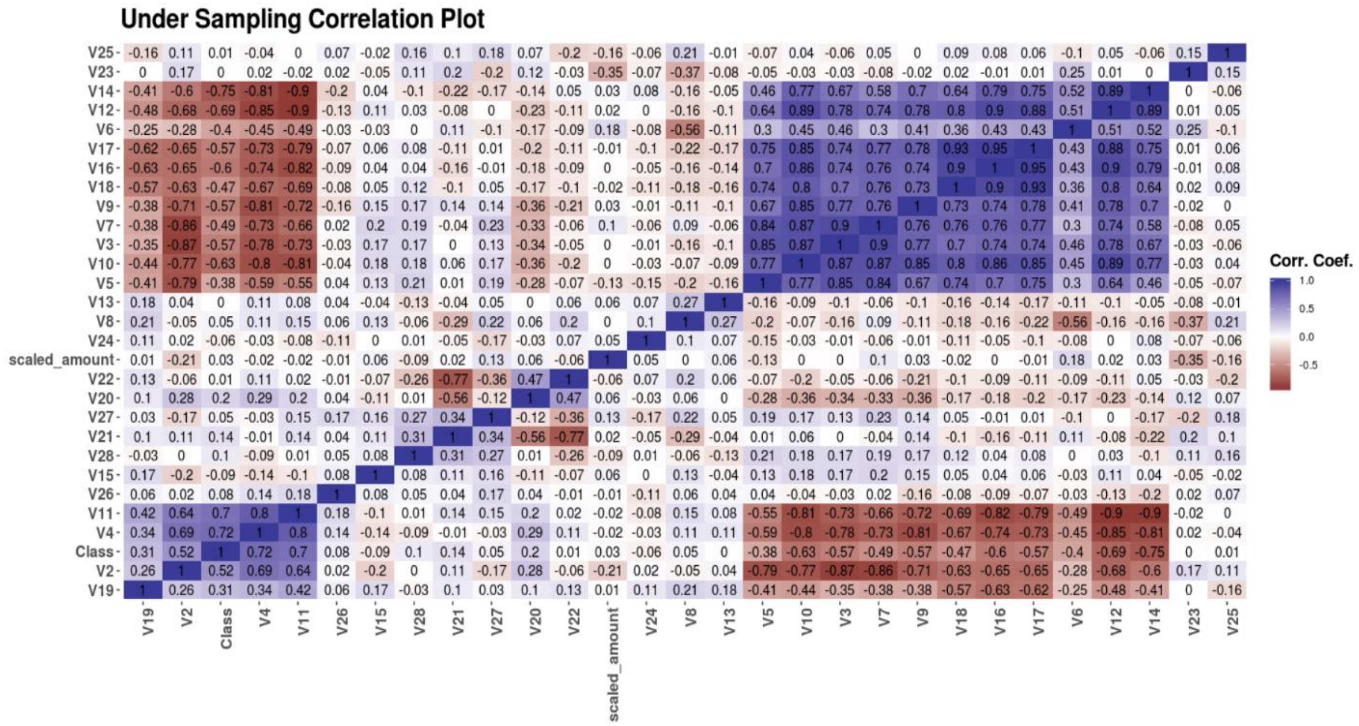

Density of Transaction vs Time

By seeing the distributions, we can have an idea how skewed are these features, we can also see further distributions of the other features. There are techniques that can help the distributions be less skewed which will be implemented in this notebook in the future.

Correlation:

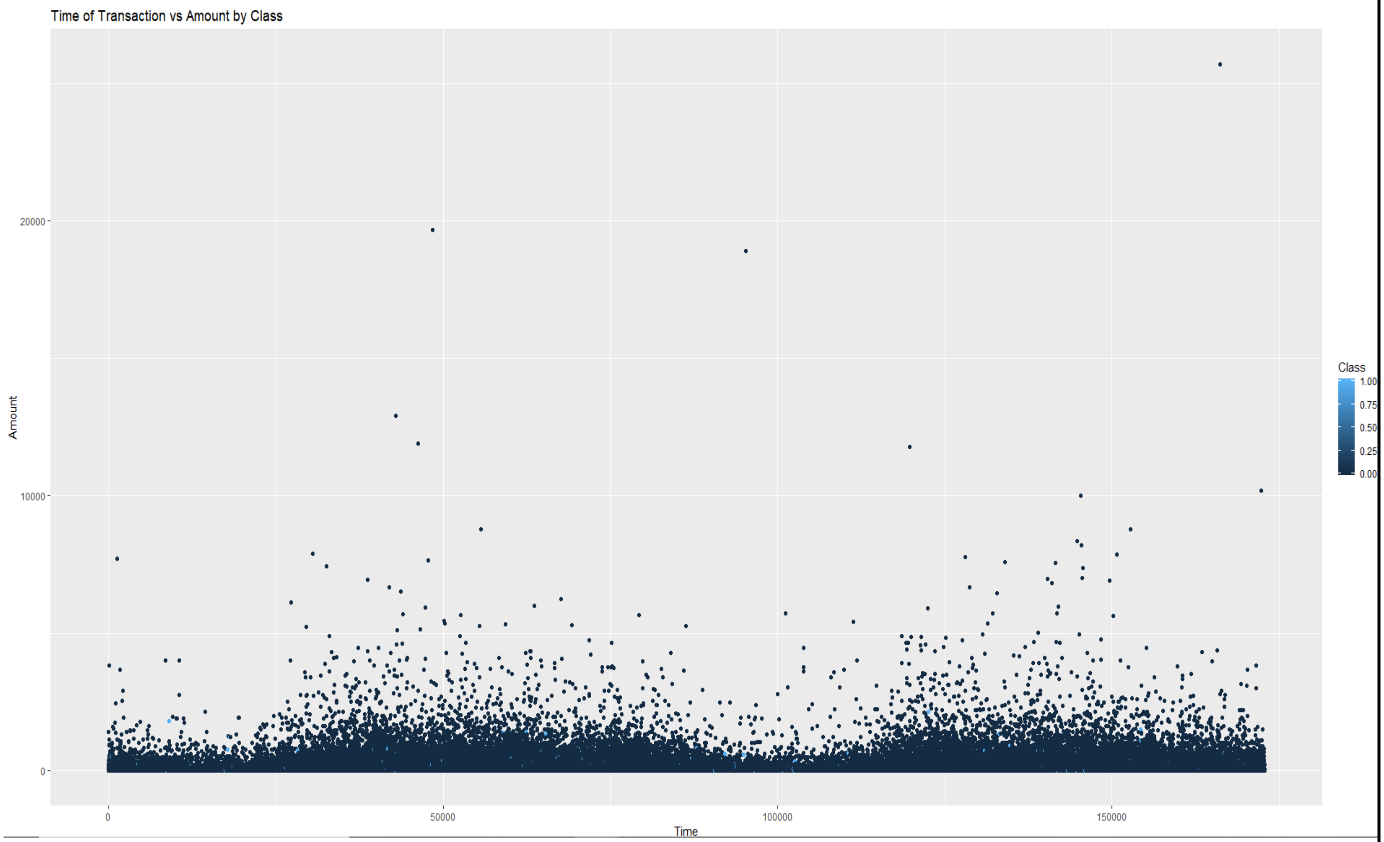


Under-Sampling Correlation:

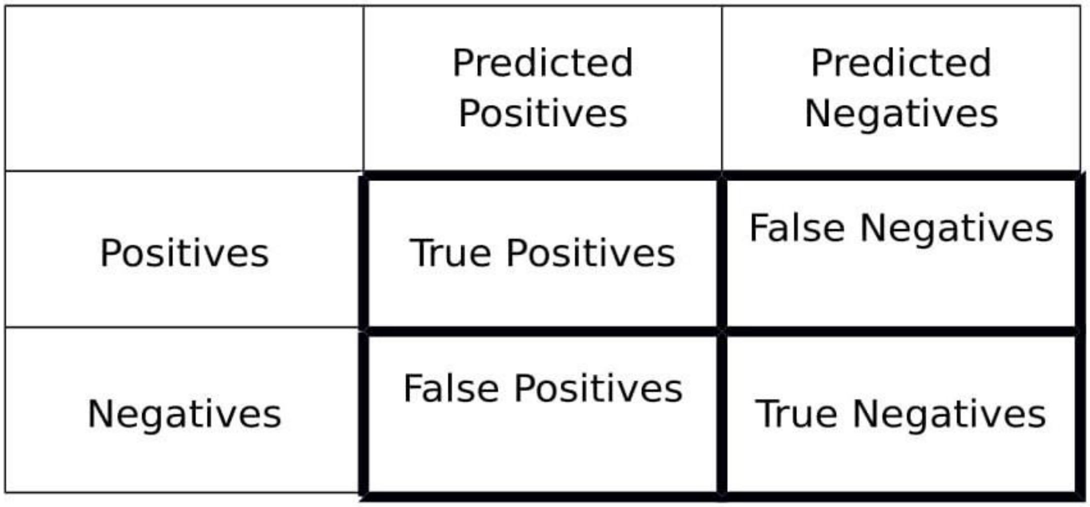


Time of Transactions Vs Amount by class:

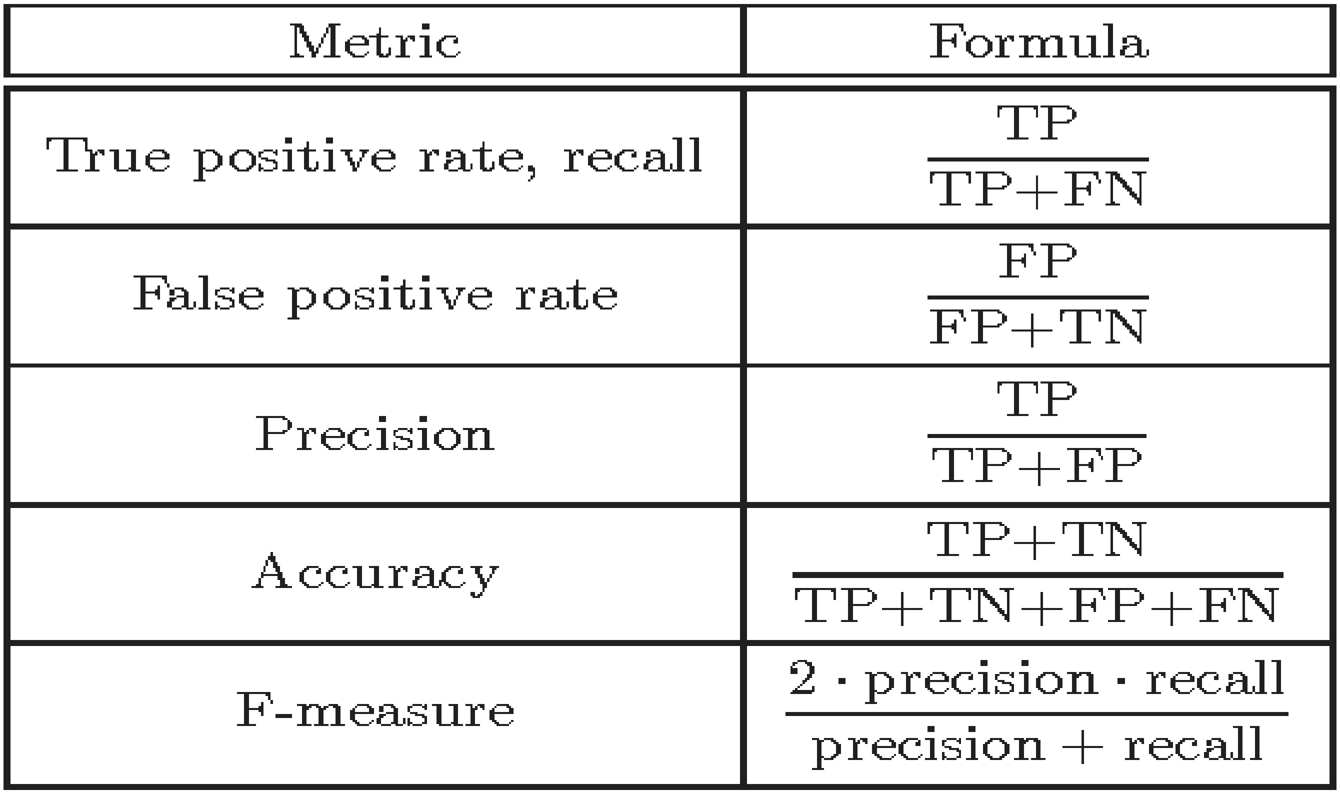
(light blue – fraud transaction, dark blue- valid transaction)



Evaluation Metrics:



	Predicted Positives	Predicted Negatives
Positives	True Positives	False Negatives
Negatives	False Positives	True Negatives



Metric	Formula
True positive rate, recall	$\frac{\text{TP}}{\text{TP+FN}}$
False positive rate	$\frac{\text{FP}}{\text{FP+TN}}$
Precision	$\frac{\text{TP}}{\text{TP+FP}}$
Accuracy	$\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$
F-measure	$\frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Machine learning Algorithms:

Supervised:

1) Logistic Regression.

2) K-Nearest Neighbors

3) Naïve Bayes

4) Decision Tree

5) Random Forest

6) Support Vector Machines

7) Artificial Neural Network

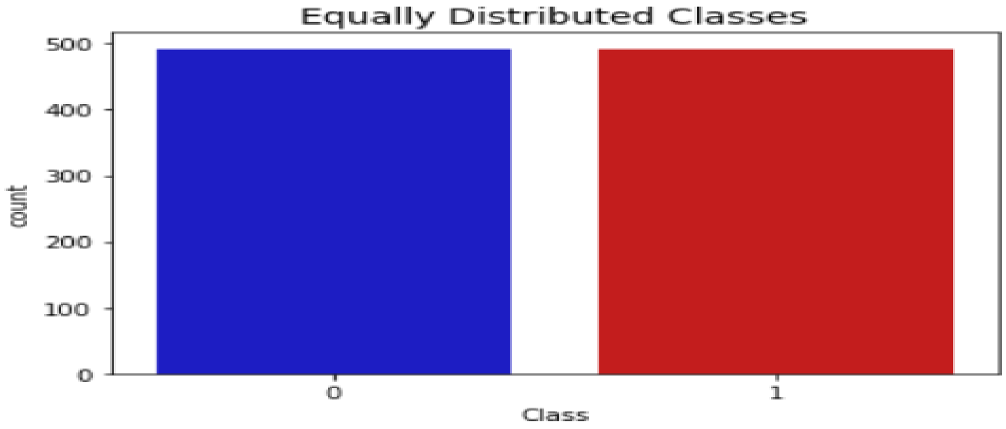
Un-Supervised:

1)K-means Clustering

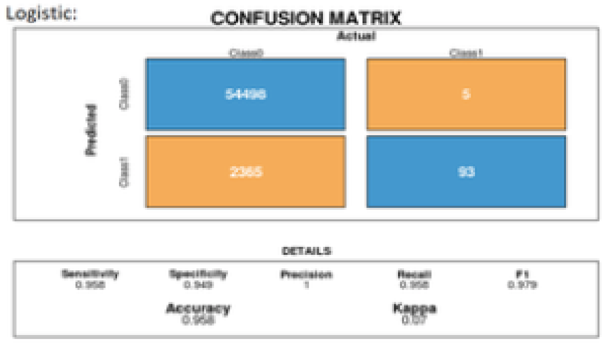
Random Under-Sampling:

"Random Under Sampling" which basically consists of removing data in order to have a more balanced dataset and thus avoiding our models to overfitting. Randomly remove samples from the majority class, with or without replacement. This is one of the earliest techniques used to alleviate imbalance in the dataset, however, it may increase the variance of the classifier and may potentially discard useful or important samples. [5]

Once we determine how many instances are considered fraud transactions (Fraud = "1") , we should bring the non-fraud transactions to the same amount as fraud transactions (assuming we want a 50/50 ratio), this will be equivalent to 492 cases of fraud and 492 cases of non-fraud transactions.



Logistic:



CONFUSION MATRIX

Actual

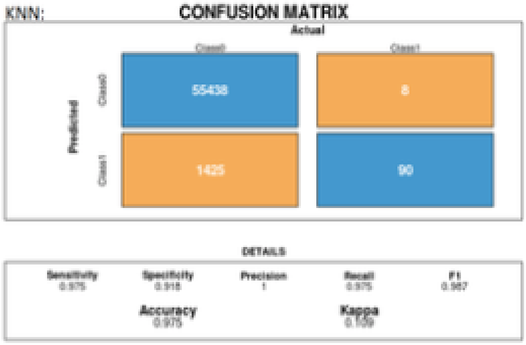
		Class0	Class1
Predicted	Class0	54496	5
	Class1	2365	93

DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.958	0.949	1	0.958	0.979

Accuracy	Kappa
0.958	0.07

KNN:



CONFUSION MATRIX

Actual

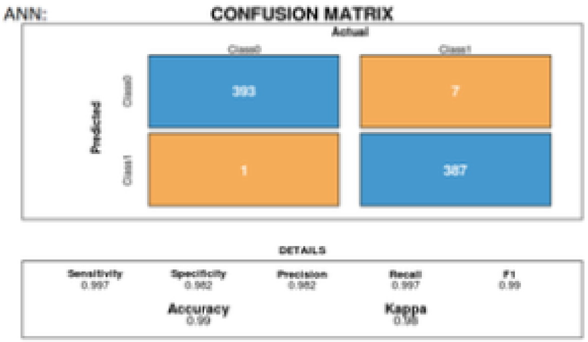
		Class0	Class1
Predicted	Class0	55438	8
	Class1	1425	90

DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.975	0.918	1	0.975	0.987

Accuracy	Kappa
0.975	0.109

ANN:



CONFUSION MATRIX

Actual

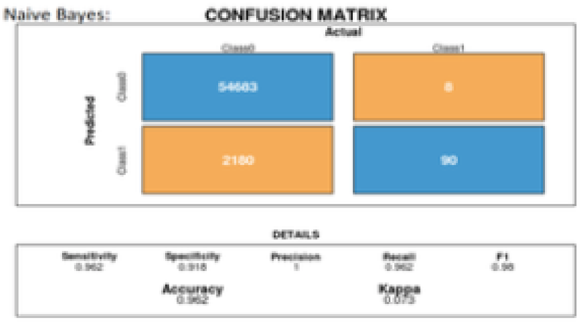
		Class0	Class1
Predicted	Class0	393	7
	Class1	1	387

DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.997	0.982	0.982	0.997	0.99

Accuracy	Kappa
0.99	0.98

Naive Bayes:



CONFUSION MATRIX

Actual

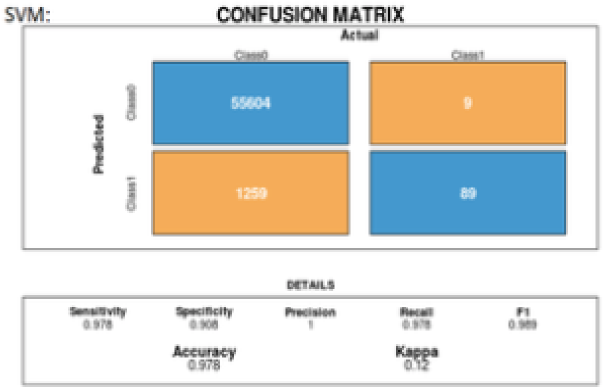
		Class0	Class1
Predicted	Class0	54683	8
	Class1	2180	90

DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.962	0.918	1	0.962	0.98

Accuracy	Kappa
0.962	0.073

SVM:



CONFUSION MATRIX

Actual

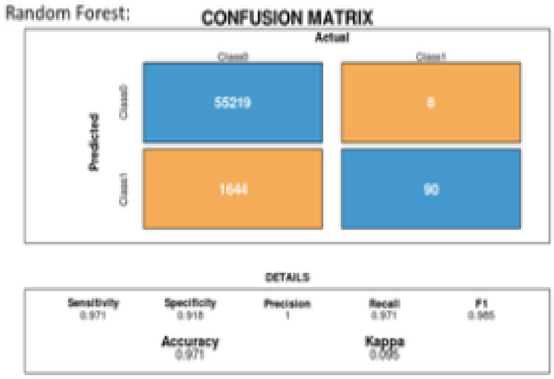
		Class0	Class1
Predicted	Class0	55604	9
	Class1	1259	89

DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.978	0.908	1	0.978	0.989

Accuracy	Kappa
0.978	0.12

Random Forest:



CONFUSION MATRIX

Actual

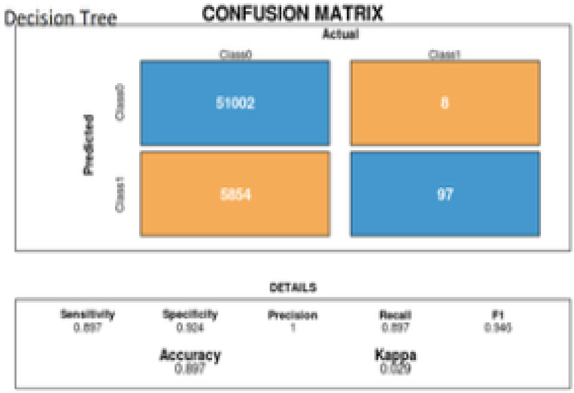
		Class0	Class1
Predicted	Class0	55219	8
	Class1	1644	90

DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.971	0.918	1	0.971	0.985

Accuracy	Kappa
0.971	0.095

Decision Tree



CONFUSION MATRIX

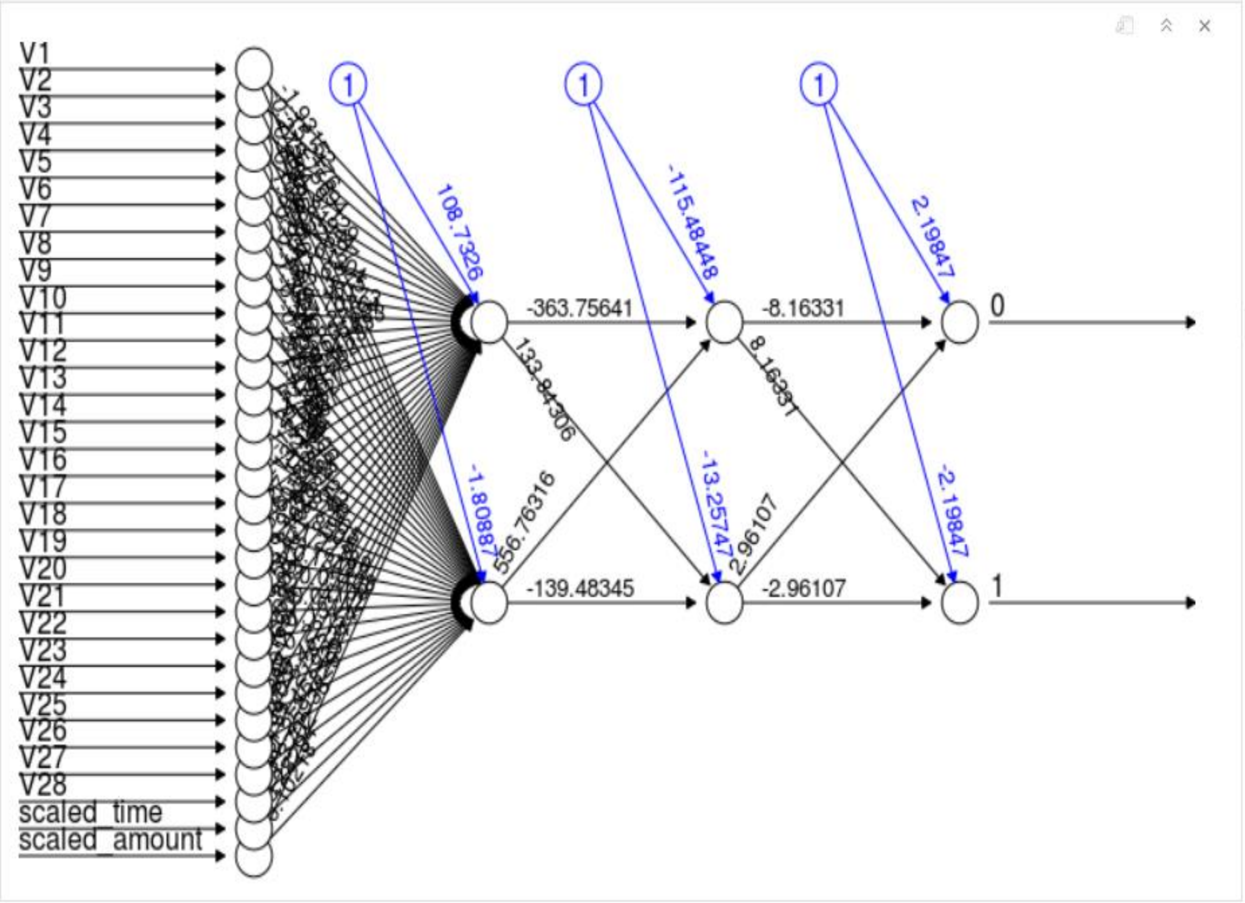
Actual

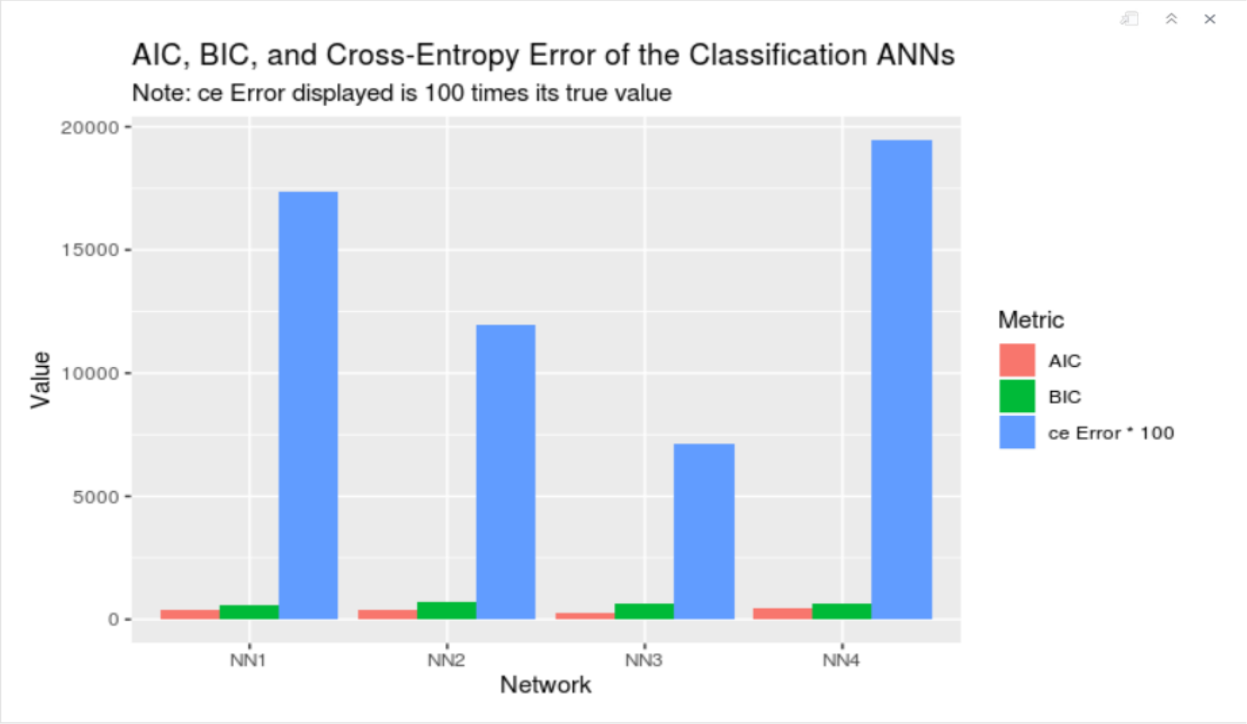
		Class0	Class1
Predicted	Class0	51002	8
	Class1	5854	97

DETAILS

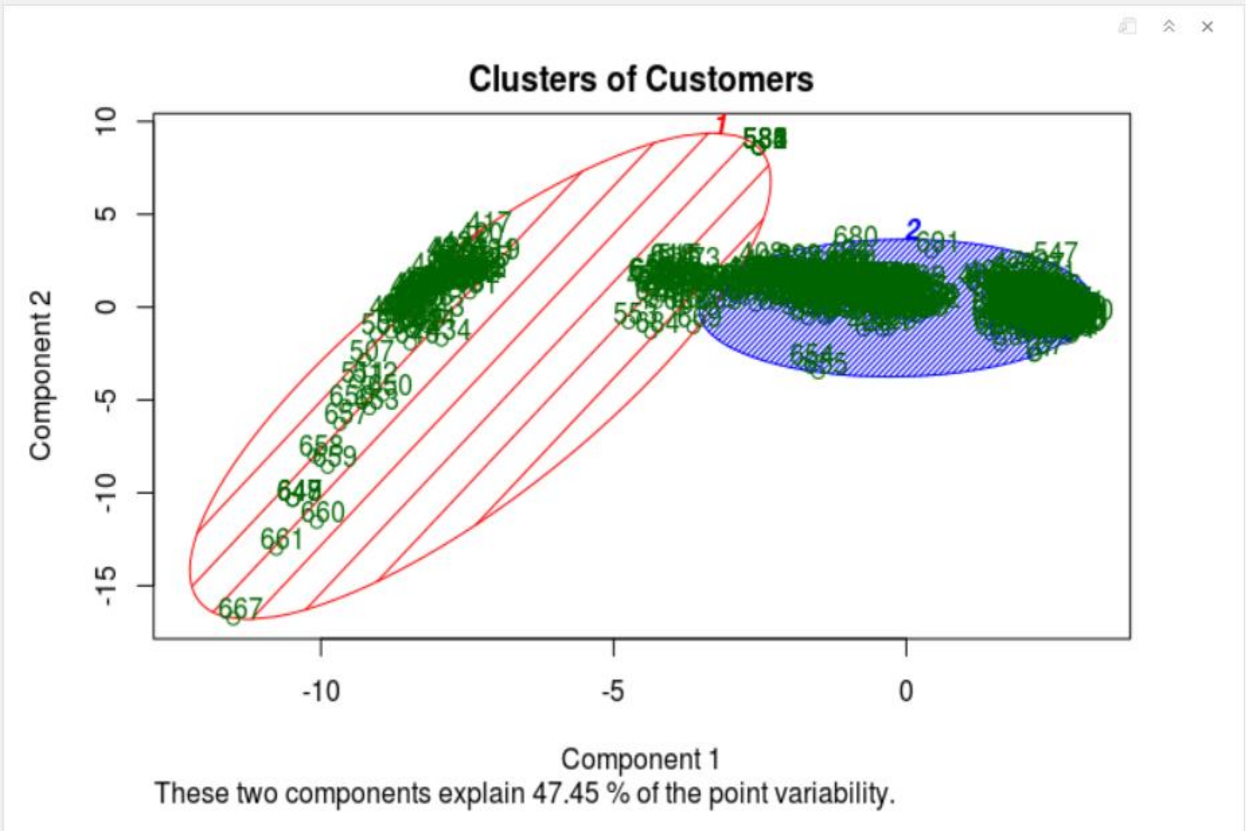
Sensitivity	Specificity	Precision	Recall	F1
0.897	0.924	1	0.897	0.946

Accuracy	Kappa
0.897	0.029

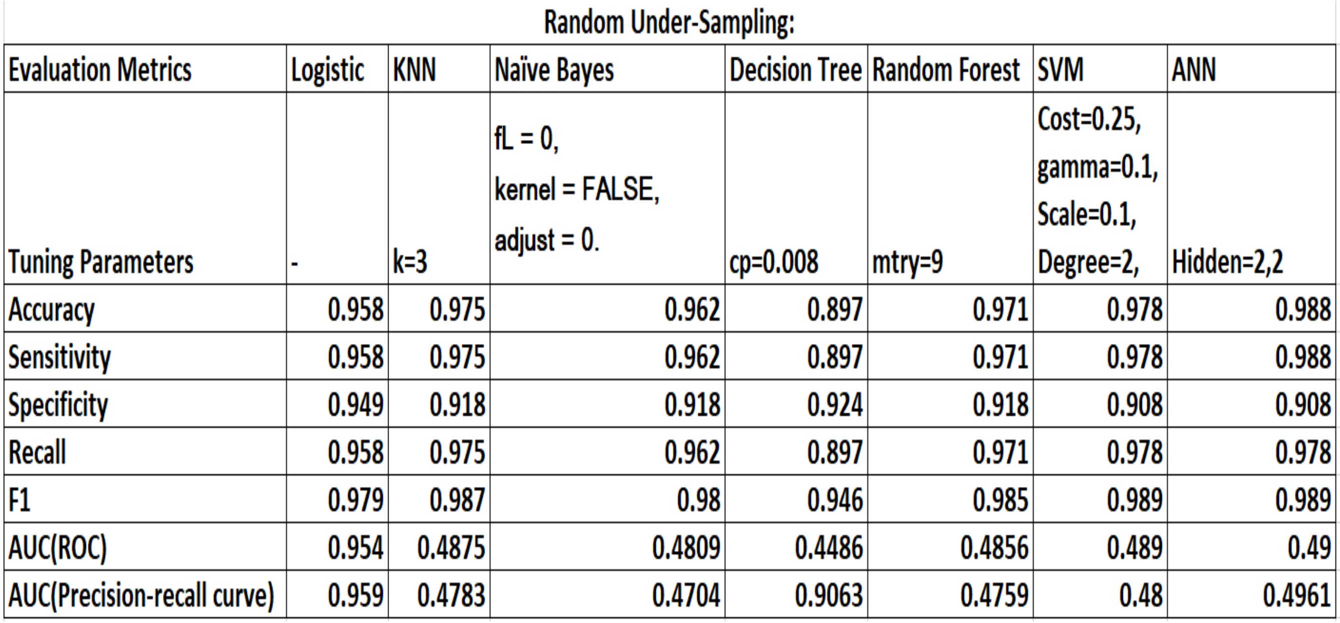

V1
V2
V3
V4
V5
V6
V7
V8
V9
V10
V11
V12
V13
V14
V15
V16
V17
V18
V19
V20
V21
V22
V23
V24
V25
V26
V27
V28
scaled_time
scaled_amount
1
1
1
108.7326
-1.80887
-115.48448
-13.25747
2.19847
-2.19847
-363.75641
133.84306
556.76316
-139.48345
-8.16331
8.16331
2.96107
-2.96107
0
1


AIC, BIC, and Cross-Entropy Error of the Classification ANNs
Note: ce Error displayed is 100 times its true value
Value
20000
15000
10000
5000
0
NN1
NN2
NN3
NN4
Network
Metric
AIC
BIC
ce Error * 100

Clustering:



These two components explain 47.45 % of the point variability.

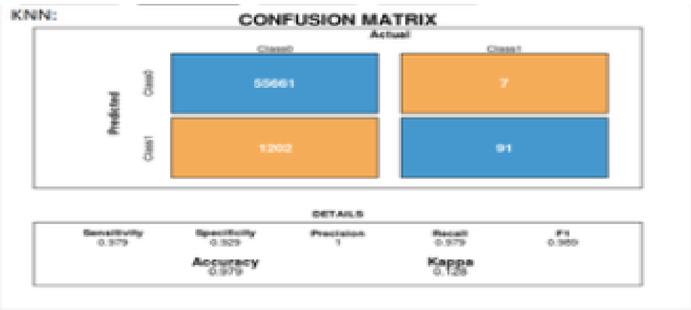


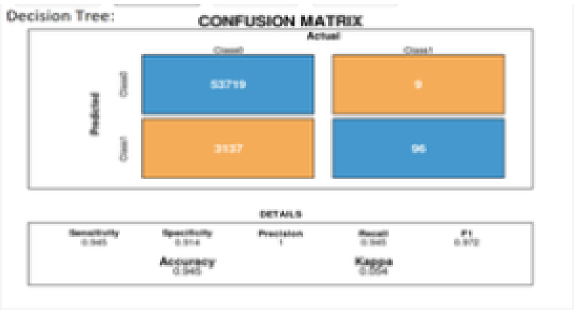
Random Under-Sampling:

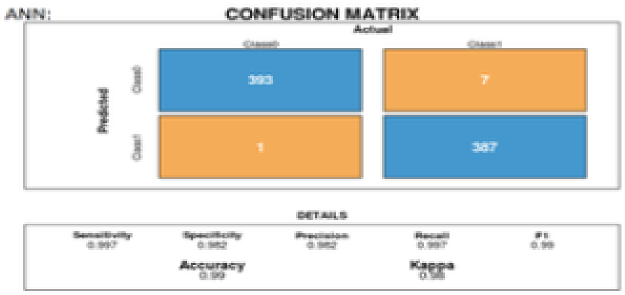
Evaluation Metrics	Logistic	KNN	Naïve Bayes	Decision Tree	Random Forest	SVM	ANN
Tuning Parameters	-	k=3	fL = 0, kernel = FALSE, adjust = 0.	cp=0.008	mtry=9	Cost=0.25, gamma=0.1, Scale=0.1, Degree=2,	Hidden=2,2
Accuracy	0.958	0.975	0.962	0.897	0.971	0.978	0.988
Sensitivity	0.958	0.975	0.962	0.897	0.971	0.978	0.988
Specificity	0.949	0.918	0.918	0.924	0.918	0.908	0.908
Recall	0.958	0.975	0.962	0.897	0.971	0.978	0.978
F1	0.979	0.987	0.98	0.946	0.985	0.989	0.989
AUC(ROC)	0.954	0.4875	0.4809	0.4486	0.4856	0.489	0.49
AUC(Precision-recall curve)	0.959	0.4783	0.4704	0.9063	0.4759	0.48	0.4961

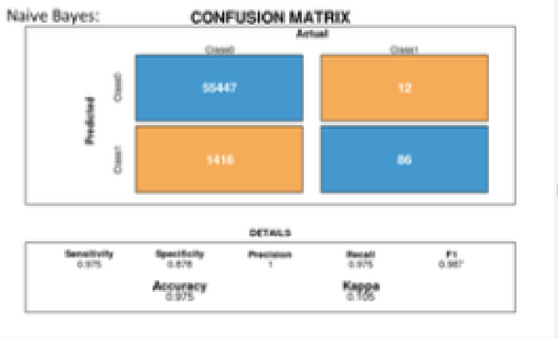
Over-Sampling:

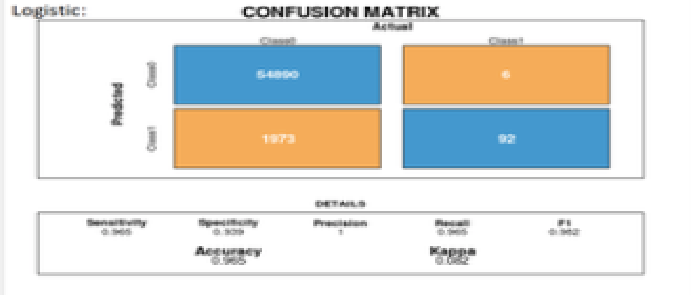
To oversample means to artificially create observations in our data set belonging to the class that is under represented in our data.One common technique is SMOTE — Synthetic Minority Over-sampling Technique. At a high level, SMOTE creates synthetic observations of the minority class (in this case, fraudulent transactions).

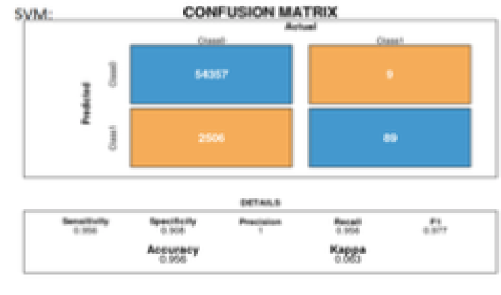


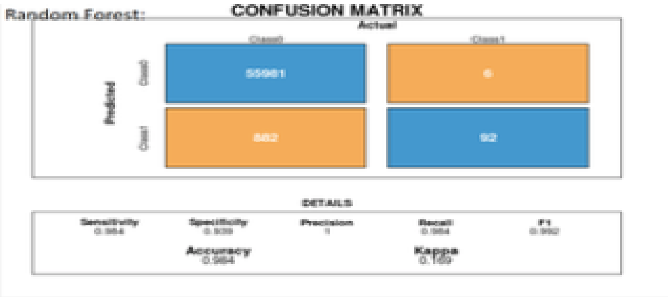


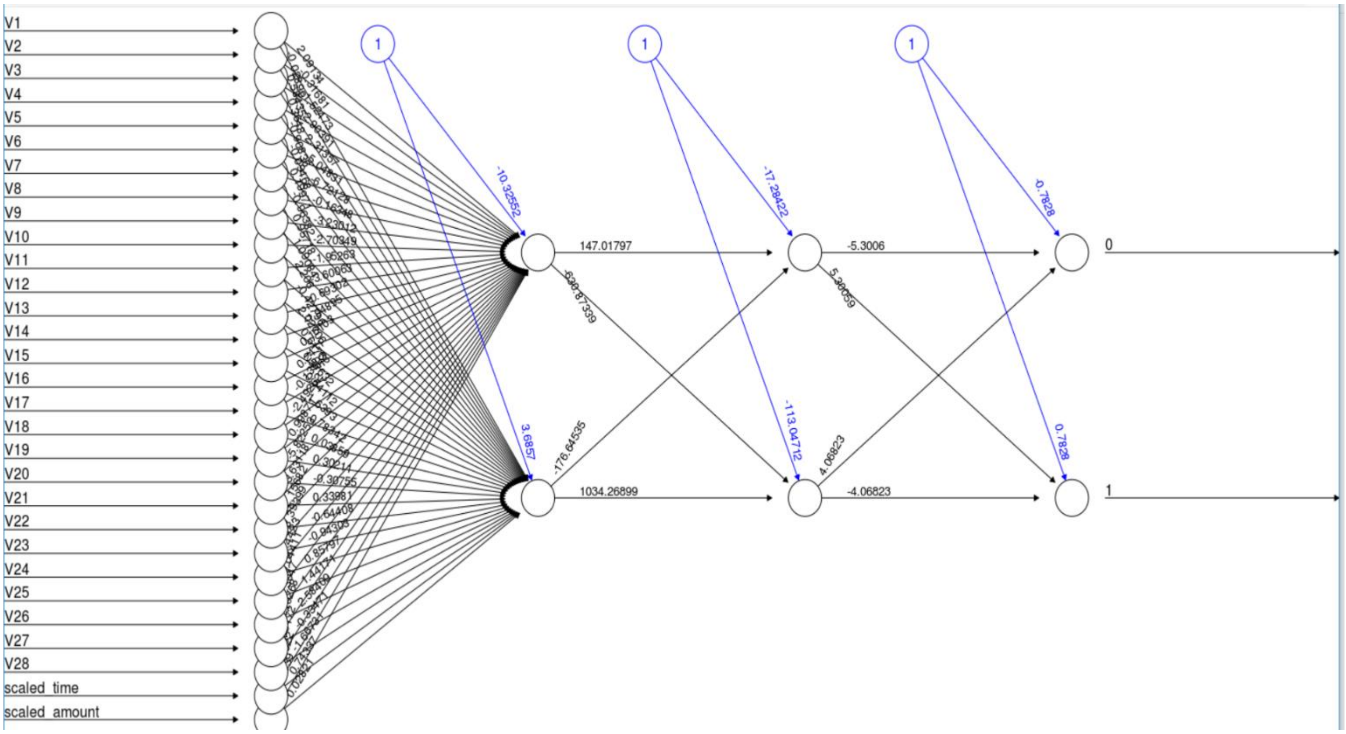


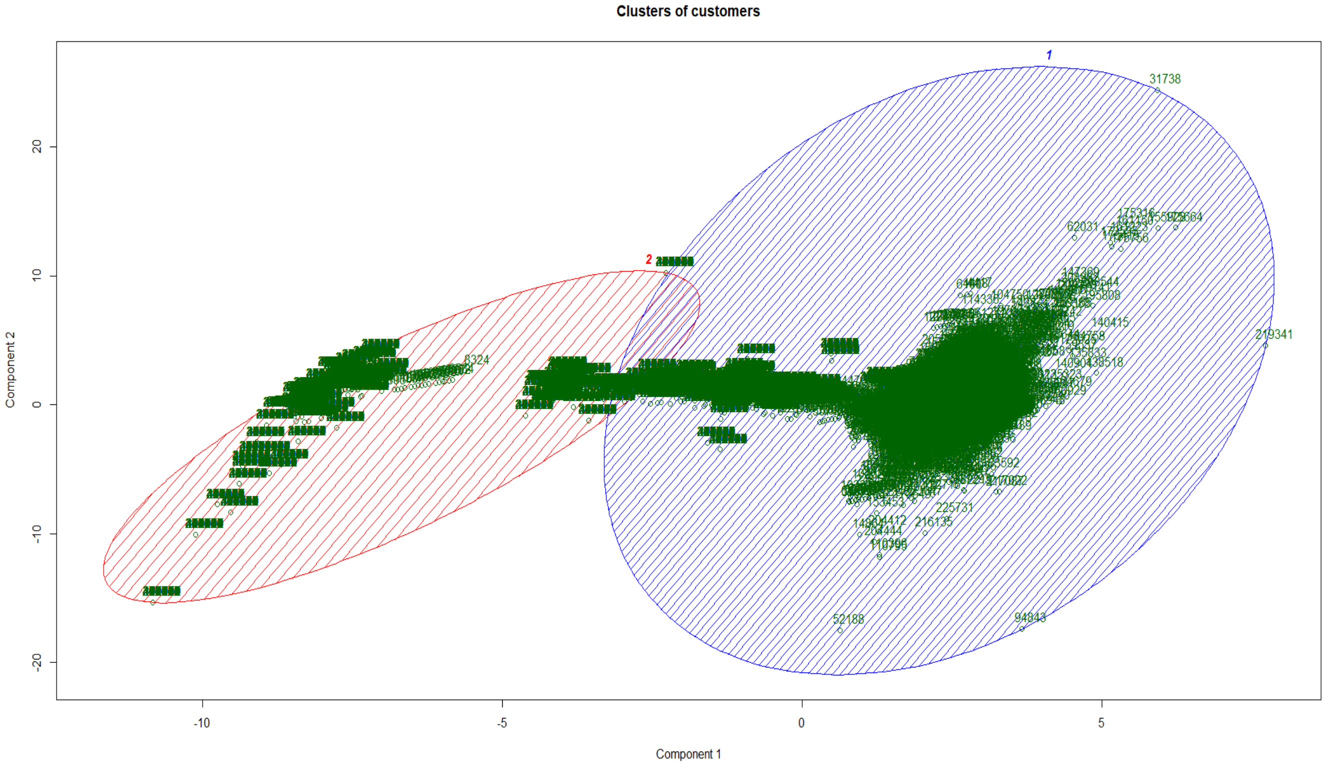












These two components explain 47.09 % of the point variability.

Conclusion:

Thus, by careful observation and anlaysis I conclude that the credit card dataset can be classified as legitimate or not by sampling the data using the undersampling, oversampling and combined methods.

These sampled data have been further analyzed using 7 machine learning algorithms to obtain and compare the best model using various metrics.

Future Work:

Though the dataset has been well analyzed, I would want to perform analysis using much advanced machine learning algorithms and considering a much larger dataset.

Reference:

1.https://www.law.cornell.edu/wex/credit_card_fraud

2.https://www.altexsoft.com/whitepapers/fraud-detection-how-machine-learning-systems-help-reveal-scams-in-fintech-healthcare-and-ecommerce/

3.Y. Sahin and E. Duman, “Detecting Credit Card Fraud by Decision

Trees and Support Vector Machines,” International Association of

Engineers, vol. I, 2011.

4.https://www.kaggle.com/mlg-ulb/creditcardfraud

5.Chawla, Nitesh V.; Herrera, Francisco; Garcia, Salvador; Fernandez, Alberto (2018-04-20). "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary". Journal of Artificial Intelligence Research. 61: 863–905