A PROJECT REPORT

ON

CREDIT CARD FRAUD DETECTION USING

MACHINE LEARNING ALGORITHMS

Submitted for the requirement of the award of

TRAINING

IN

Data Analytics,Machine Learning and AI using Python

Diginique TechLabs
Ideal for Geeks

Submitted By

Esakki Mathy U(SSN College Of Engineering,Chennai)

, Under the guidance of

Mr. Bipul Shahi

Abstract

The objective of this project is to detect all the fraudulent transactions while minimising incorrect fraud classification using Logistic Regression,Support vector machine and Random Forest.The evaluation and comparison of algorithms is performed using K-fold cross validation. F1 score and ROC-AUC are used as performance metrics.The F-score of Random Forest is better than the other two with decreasing ratio of fraud cases.

Problem Statement

Fraud detection in credit card is truly the process of identifying those transactions that are fraudulent into two classes of legit class and fraud class transaction.Comparative analysis of Logistic Regression,SVM and Random Forest is carried out.Credit card transaction datasets are highly imbalanced and skewed.Optimal feature selection for models,suitable metric is most important part to evaluate the performance of techniques on skewed credit card fraud data.

Technology and Concepts

Machine learning(ML) is an application of artificial **intelligence** (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning** focuses on the development of computer programs that can access data and use it to learn for themselves.

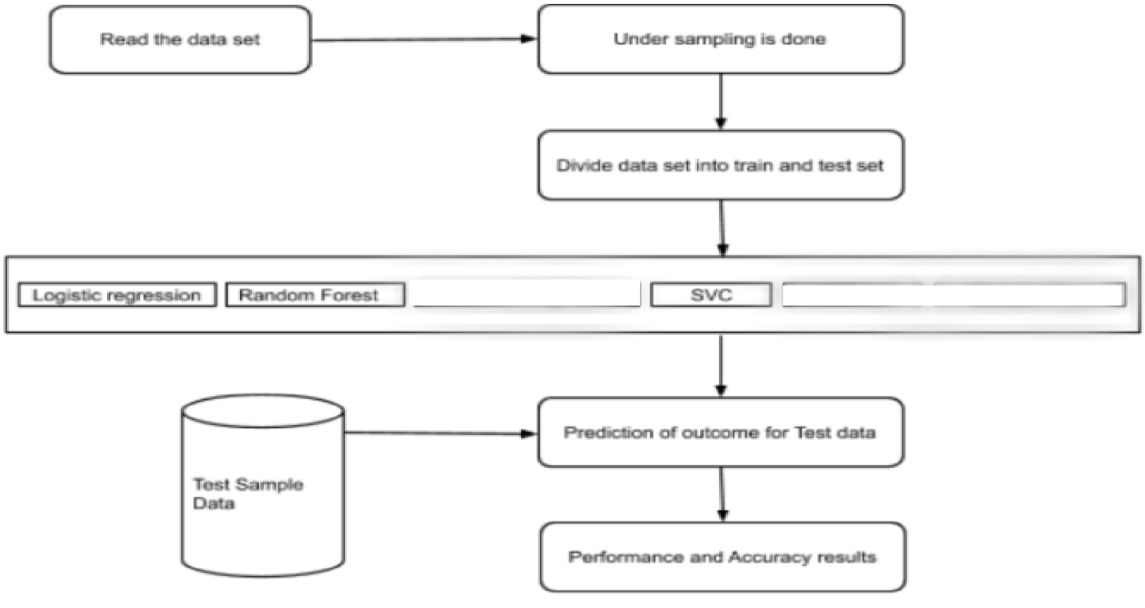

Read the data set
Under sampling is done
Divide data set into train and test set
Logistic regression
Random Forest
SVC
Test Sample Data
Prediction of outcome for Test data
Performance and Accuracy results

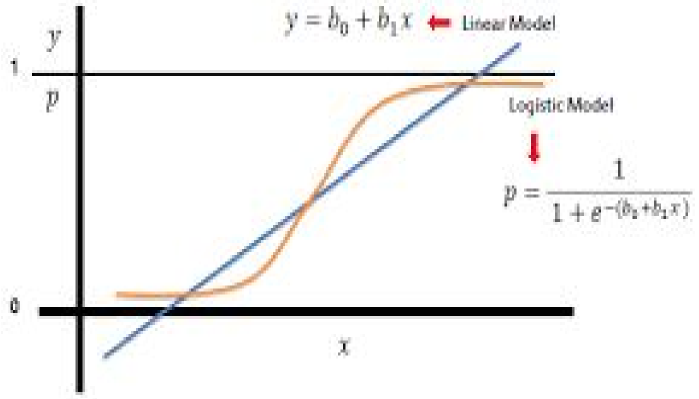
Figure 1: System Architecture

Classification Algorithms

Three algorithms - Logistic Regression, Support Vector Machine and Random Forest are used to solve the problem. A comparative study is then followed by evaluating each algorithm with respect to selected performance metrics. Simple logistic regression is used. For the parameters of Support Vector Machine and Random Forest, Grid Search technique is employed to tune the parameters.

A.Logistic Regression

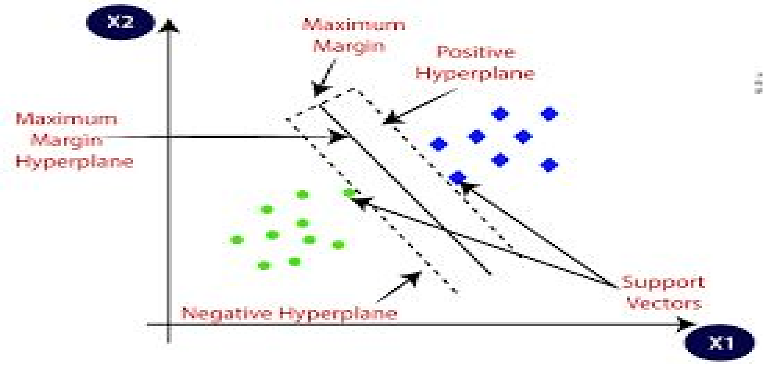
Logistic Regression is a supervised classification method that returns the probability of binary dependent variable that is predicted from the independent variable of dataset i.e. logistic regression predicts the probability of an outcome which has two values, either zero or one, no or yes and false or true. Logistic regression has similarities to linear regression, but, in linear regression a straight line is obtained, logistic regression shows a curve. The use of one or several predictors or independent variable is on what prediction is based, logistic regression produces logistic curves which plots the values between zero and one. Logistic Regression is a regression model where the dependent variable is categorical and analyzes the relationship between multiple independent variables. There are many types of logistic regression model such as binary logistic model, multiple logistic model, binomial logistic models. Binary Logistic Regression model is used to estimate the probability of a binary response based on one or more predictors.



This graph shows the difference between linear regression and logistic regression where logistic regression shows a curve but linear regression represents a straight line.

B. SVM Model (Support Vector Machine)

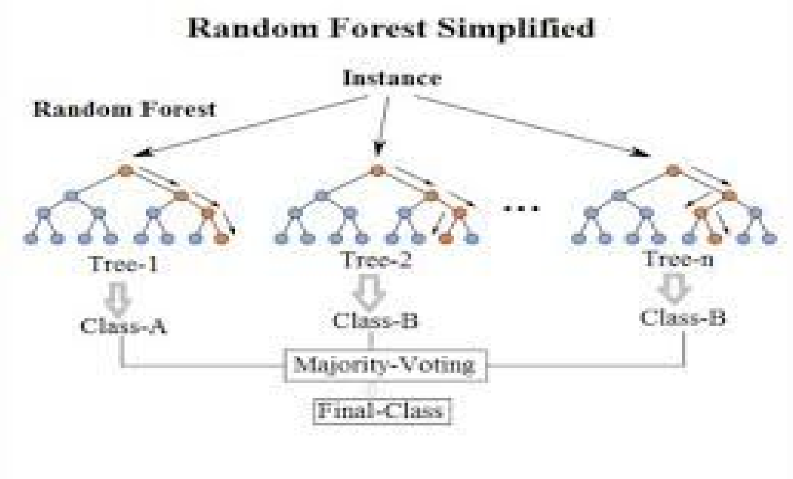
SVM is a one of the popular machine learning algorithm for regression, classification. It is a supervised learning algorithm that analyses data used for classification and regression. SVM modeling involves two steps, firstly to train a data set and to obtain a model & then, to use this model to predict information of a testing data set. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane where SVM model represents the training data points as points in space and then mapping is done so that the points which are of different classes are divided by a gap that is as wide as possible. Mapping is done into the same space for new data points and then predicted on which side of the gap they fall.



In SVM algorithm, plotting is done as each data item is taken as a point in n-dimensional space where n is number of features, with the value of each feature being the value of a particular coordinate. Then, classification is performed by locating the hyperplane that separates the two classes very well.

C. Random Forest

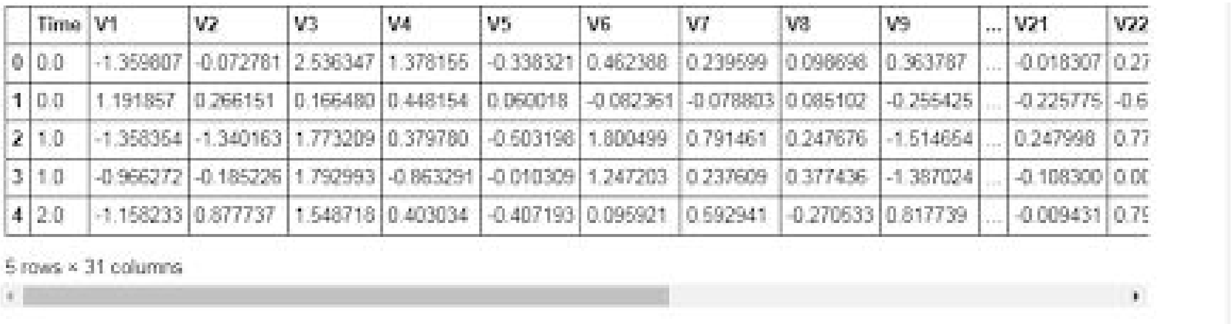
Random Forest is an algorithm for classification and regression. Summarily, it is a collection of decision tree classifiers. Random forest has advantage over decision tree as it corrects the habit of overfitting to their training set. A subset of the training set is sampled randomly so that to train each individual tree and then a decision tree is built, each node then splits on a feature selected from a random subset of the full feature set. Even for large data sets with many features and data instances training is extremely fast in random forest and because each tree is trained independently of the others. The Random Forest algorithm has been found to provides a good estimate of the generalization error and to be resistant to overfitting. Random forest ranks the importance of variables in a regression or classification problem in a natural way can be done by Random Forest.



Working of the project

Dataset

This dataset presents transactions which 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. Due to confidentiality issues, the original features and more background information about the data are not provided. 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. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

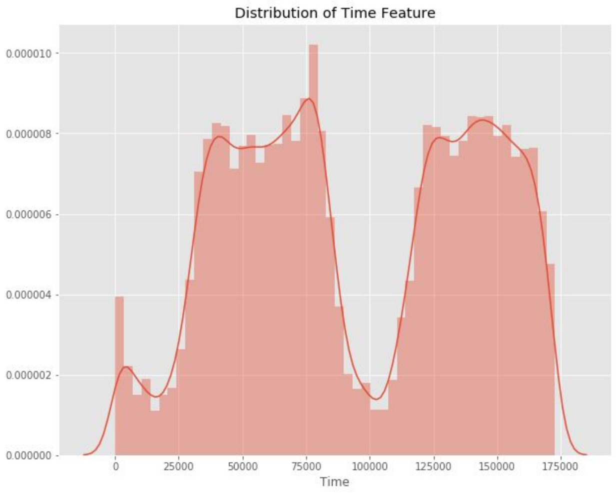


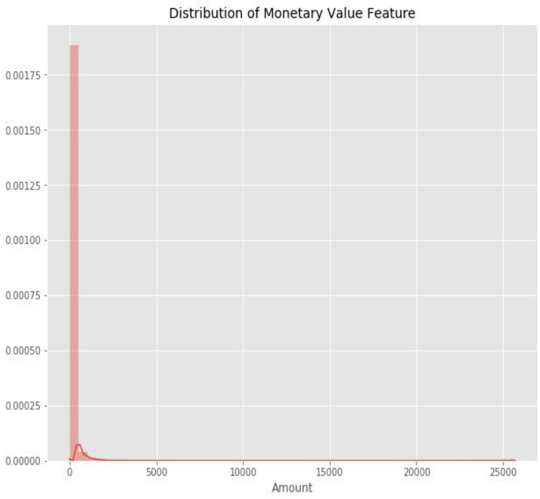
	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.27
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.6
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.77
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.00
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.79

5 rows × 31 columns

Scaling Amount and Time

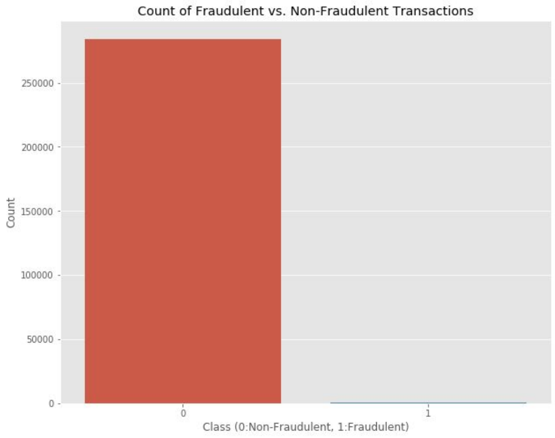
Anonymised features appear to have been scaled and centred around zero but Amount and Time have not been scaled. For Logistic Regression to perform well, the features have to be scaled.



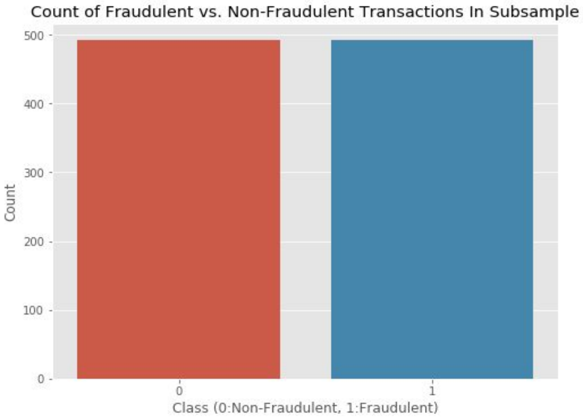


Random Under Sampling

The present dataset is highly imbalanced with the fraud cases forming an extremely small percentage of the whole. This poses a problem in training the algorithms. To deal with this problem, the technique of random undersampling - removal of samples from majority class, is used.



Before Sampling

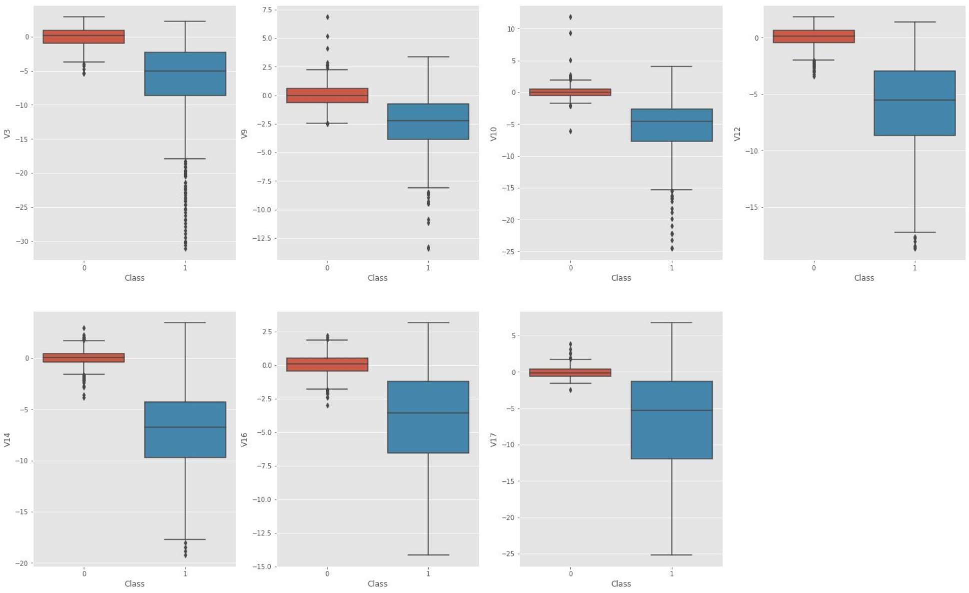


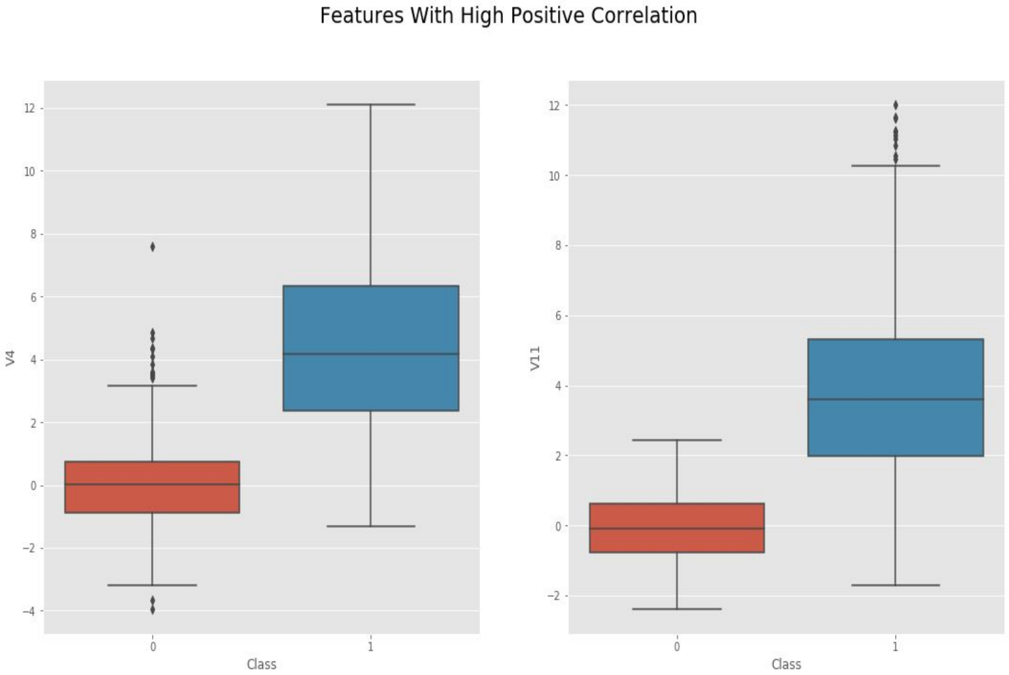
After Sampling

Outlier Detection and Removal

Here, features with high negative correlation and high positive correlation are identified. We then remove the outliers that lies outside 2.5 times the IQR in an attempt to eliminate the effect of outliers.

Features With High Negative Correlation





Features With High Positive Correlation

Parameter tuning & Evaluating Algorithms

For the parameters of Support Vector Machine and Random Forest, Grid Search technique is employed to tune the parameters.The evaluation and comparison of algorithms is performed using K-fold cross validation. F1 score and ROC-AUC are used as performance metrics. The mean of each parameter across the k sets is displayed.

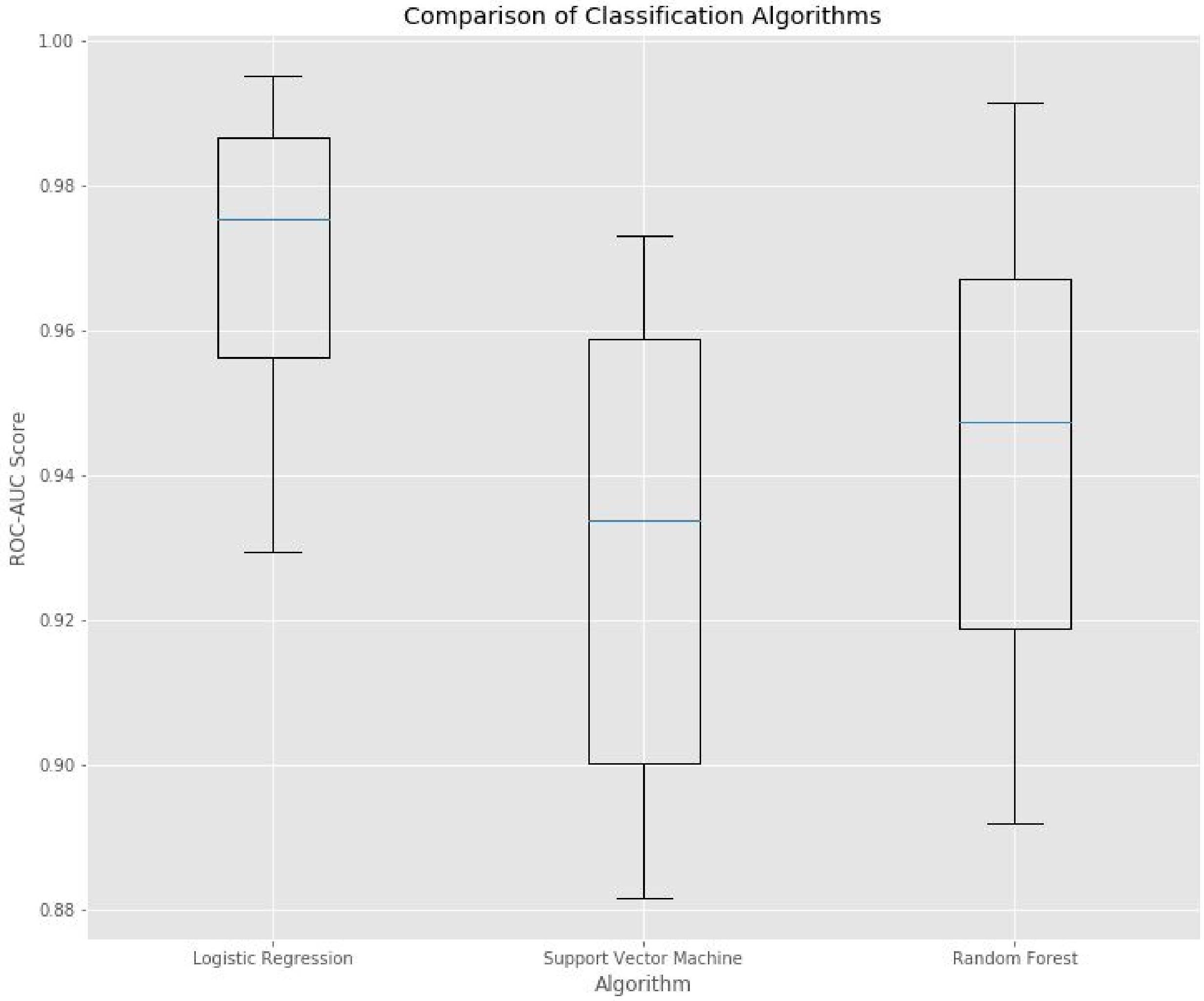
Result Analysis

Logistic Regression:
F1 Score: 0.9242971956405512
ROC-AUC: 0.9706274235691745

Support Vector Machine:
F1 Score: 0.9188589468803274
ROC-AUC: 0.929719914072749

Random Forest:
F1 Score: 0.9307026313613227
ROC-AUC: 0.9450204332273134

Comparing different algorithms



Conclusion

From the observation, the ratio of fraud cases decreases the ROC-AUC and F1 score of Logistic Regression and Support Vector Machine show a downward trend whereas those of Random Forest show better performance. The F-score of Random Forest is better than the other two with decreasing ratio of fraud cases. This is an important performance indicator as it shows that the system is correctly classifying frauds as well as minimising errors in incorrect classification, both of which are extremely relevant to the real world scenario.