Credit Card Fraud Detection Using Machine Learning

Abstract

The purpose of this project is to detect the fraudulent transactions made by credit cards by the use of machine learning techniques, to stop fraudsters from the unauthorized usage of customers' accounts. The increase of credit card fraud is growing rapidly worldwide, which is the reason actions should be taken to stop fraudsters. Putting a limit for those actions would have a positive impact on the customers as their money would be recovered and retrieved back into their accounts and they won't be charged for items or services that were not purchased by them which is the main goal of the project. Detection of the fraudulent transactions will be made by using three machine learning techniques Decision tree classifier, SVM and Logistic Regression, those models will be used on a credit card transaction dataset.

Keywords: Credit Card Fraud Detection, Fraud Detection, Fraudulent Transactions,
Support Vector Machine, Logistic Regression, Decision Tree Classifier

Table of Contents

List of Figures

List of Tables

Chapter 1

1.1 Introduction

With the increase of people using credit cards in their daily lives, credit card companies should take special care in the security and safety of the customers. According to (Credit card statistics 2021) the number of people using credit cards around the world was 2.8 billion in 2019, in addition 70% of those users own a single card at least.

Reports of Credit card fraud in the US rose by 44.7% from 271,927 in 2019 to 393,207 reports in 2020. There are two kinds of credit card fraud, the first one is by having a credit card account opened under your name by an identity thief,

reports of this fraudulent behavior increased 48% from 2019 to 2020. The second type is when an identity thief uses an existing account that you created, and it's usually done by stealing the information of the credit card. Reports on this type of fraud increased 9% from 2019 to 2020 (Daly, 2021). Those statistics caught my attention as the numbers are increasing drastically and rapidly throughout the years, which gave me the motive to try to resolve the issue analytically by using different machine learning methods to detect the credit card fraudulent transactions within numerous transactions.

1.2 Project goals

The main aim of this project is the detection of credit card fraudulent transactions, as it's important to figure out the fraudulent transactions so that customers don't get charged for the purchase of products that they didn't buy. The detection of the credit card fraudulent transactions will be performed with multiple ML techniques then a comparison will be made between the outcomes and results of each technique to find the best and most suited model in the detection of the credit card transaction that are fraudulent, graphs and numbers will be provided as well. In addition, exploring previous literature and different techniques used to distinguish the fraud within a dataset.

Research question: What is the most suited machine learning model in the detection of fraudulent credit card transactions?

Phase 1: Business Understanding

As stated before credit card fraud is increasing drastically every year, many people are facing the problem of having their credits breached by those fraudulent people, which is impacting their daily lives, as payments using a credit card is similar to taking a loan. If the problem is not solved many people will have large amounts of loans that they cannot pay back which will make them face a hard life, and they won't be able to afford necessary products, in the long run not being able to pay back the amount might lead to them going to jail. Basically, the problem proposed is the detection of the credit card fraudulent transactions made by fraudsters to stop those breaches and to ensure customers security.

Business Objective: Identification of fraudulent transactions to prohibit deduction from affected customers' accounts.

Phase 2: Data Understanding

In the Data understanding phase, it was critical to obtain a high-quality dataset as the model is based on it, the dataset was explored by taking a closer look into it which gave the knowledge needed to confirm the quality of the dataset, additionally to reading the description of the whole dataset and each attribute. It's also important to have a dataset that contains several mixed transaction types "Fraudulent and real" and a class to clarify the type of transaction, finally, identifiers to clarify the reason behind the classification of the transaction type. I made sure to follow all of those points during the search for the most suited dataset.

Phase 3: Data Preparation

After choosing the most suited dataset the preparation phase begins, the preparation of the dataset includes selecting the wanted attributes or variables, cleaning it by excluding Null rows, deleting duplicated variables, treating outlier if necessary, in addition to transforming data types to the wanted type, data merging can be performed as well where two or more attributes get merged. All those alterations lead to the wanted result which is to make the data ready to be modeled.

The dataset chosen for this project didn't need to go through all of the alterations mentioned earlier, as there were no missing or duplicated variables, there was no merging needed as well. But there was some change in the types of the data to be able to create graphs, in addition to using the application Sublime Text to be able to insert the data into Weka and perform analysis, as it needed to be altered.

Phase 4: Modeling

Three learning models were created in the modeling phase ,Decision Tree , SVM, Logistic Regression. A comparison of the results will be presented later in the paper to know which technique is most suited in the credit card fraudulent transactions detection. The dataset is sectioned into a ratio of 80:200, the training set will be the 70% and the remaining set will be the testing set which is the 20%.

Phase 5: Evaluation and Deployment

The final phase will show evaluations of the models by presenting their efficiency, the accuracies of the models will be presented in addition to any comment observed, to find the best and most suited model for detecting the fraud transactions made by credit card.

Chapter 2: Project Description

2.1 Introduction

In order to accomplish the objective and goal of the project which is to find the most suited model to detect credit card fraud several steps need to be taken. Finding the most suited data and preparing/preprocessing are the first and second steps, after making sure that the data is ready the modeling phase starts, where 3 models are created: Decision Tree , SVM and the last one is Logistic Regression. All Models were created in jupyter notebook.

2.2 Data Source

The dataset was retrieved from an open-source website, Kaggle.com. It contains data of transactions that were made in 2013 by credit card users in Europe, in two days only. The dataset consists of 31 attributes, 284,808 rows. 28 attributes are numeric variables that due to confidentiality and privacy of the customers have been transformed using PCA transformation, the three remaining attributes are "Time" which contains the elapsed seconds between the first and other transactions of each attribute, "Amount" is the amount of each transaction, and the final attribute "Class" which contains binary variables where "1" is a case of fraudulent transaction, and "0" is not as case of fraudulent transaction.

Dataset Link: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Chapter 3: Data Analysis

3.1 Data Preparation

The first figure below shows the structure of the dataset where all attributes are shown, with their type, in addition to glimpse of the variables within each attribute, as shown at the end of the figure the Class type is integer which I needed to change to factor and ide

ntify the 0 as Not Fraud and the 1 as Fraud to ease the process of creating the model and obtain visualizations.

```
'data.frame':   284807 obs. of  31 variables:
 $ Time  : num  0 0 1 1 2 2 4 7 7 9 ...
 $ V1    : num  -1.36 1.192 -1.358 -0.966 -1.158 ...
 $ V2    : num  -0.0728 0.2662 -1.3402 -0.1852 0.8777 ...
 $ V3    : num  2.536 0.166 1.773 1.793 1.549 ...
 $ V4    : num  1.378 0.448 0.38 -0.863 0.403 ...
 $ V5    : num  -0.3383 0.06 -0.5032 -0.0103 -0.4072 ...
 $ V6    : num  0.4624 -0.0824 1.8005 1.2472 0.0959 ...
 $ V7    : num  0.2396 -0.0788 0.7915 0.2376 0.5929 ...
 $ V8    : num  0.0987 0.0851 0.2477 0.3774 -0.2705 ...
 $ V9    : num  0.364 -0.255 -1.515 -1.387 0.818 ...
 $ V10   : num  0.0908 -0.167 0.2076 -0.055 0.7531 ...
 $ V11   : num  -0.552 1.613 0.625 -0.226 -0.823 ...
 $ V12   : num  -0.6178 1.0652 0.0661 0.1782 0.5382 ...
 $ V13   : num  -0.991 0.489 0.717 0.508 1.346 ...
 $ V14   : num  -0.311 -0.144 -0.166 -0.288 -1.12 ...
 $ V15   : num  1.468 0.636 2.346 -0.631 0.175 ...
 $ V16   : num  -0.47 0.464 -2.89 -1.06 -0.451 ...
 $ V17   : num  0.208 -0.115 1.11 -0.684 -0.237 ...
 $ V18   : num  0.0258 -0.1834 -0.1214 1.9658 -0.0382 ...
 $ V19   : num  0.404 -0.146 -2.262 -1.233 0.803 ...
 $ V20   : num  0.2514 -0.0691 0.525 -0.208 0.4085 ...
 $ V21   : num  -0.01831 -0.22578 0.248 -0.1083 -0.00943 ...
 $ V22   : num  0.27784 -0.63867 0.77168 0.00527 0.79828 ...
 $ V23   : num  -0.11 0.101 0.909 -0.19 -0.137 ...
 $ V24   : num  0.0669 -0.3398 -0.6893 -1.1756 0.1413 ...
 $ V25   : num  0.129 0.167 -0.328 0.647 -0.206 ...
 $ V26   : num  -0.189 0.126 -0.139 -0.222 0.502 ...
 $ V27   : num  0.13356 -0.00898 -0.05535 0.06272 0.21942 ...
 $ V28   : num  -0.0211 0.0147 -0.0598 0.0615 0.2152 ...
 $ Amount: num  149.62 2.69 378.66 123.5 69.99 ...
 $ Class : int  0 0 0 0 0 0 0 0 0 0 ...
```

Figure 1 - Dataset Structure

The second figure shows the distribution of the class, the red bar which contains 284,315 variables represents the non-fraudulent transactions, and the blue bar with 492 variables represents the fraudulent transactions.

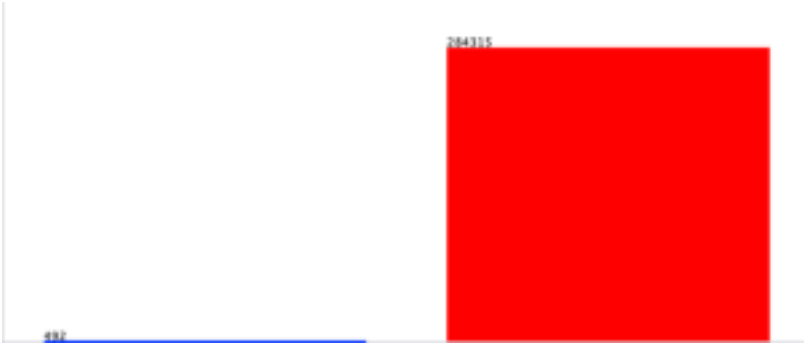

Figure 2 - Class Distribution

The correlations between all the of the attributes within the dataset are presented in the figure below.

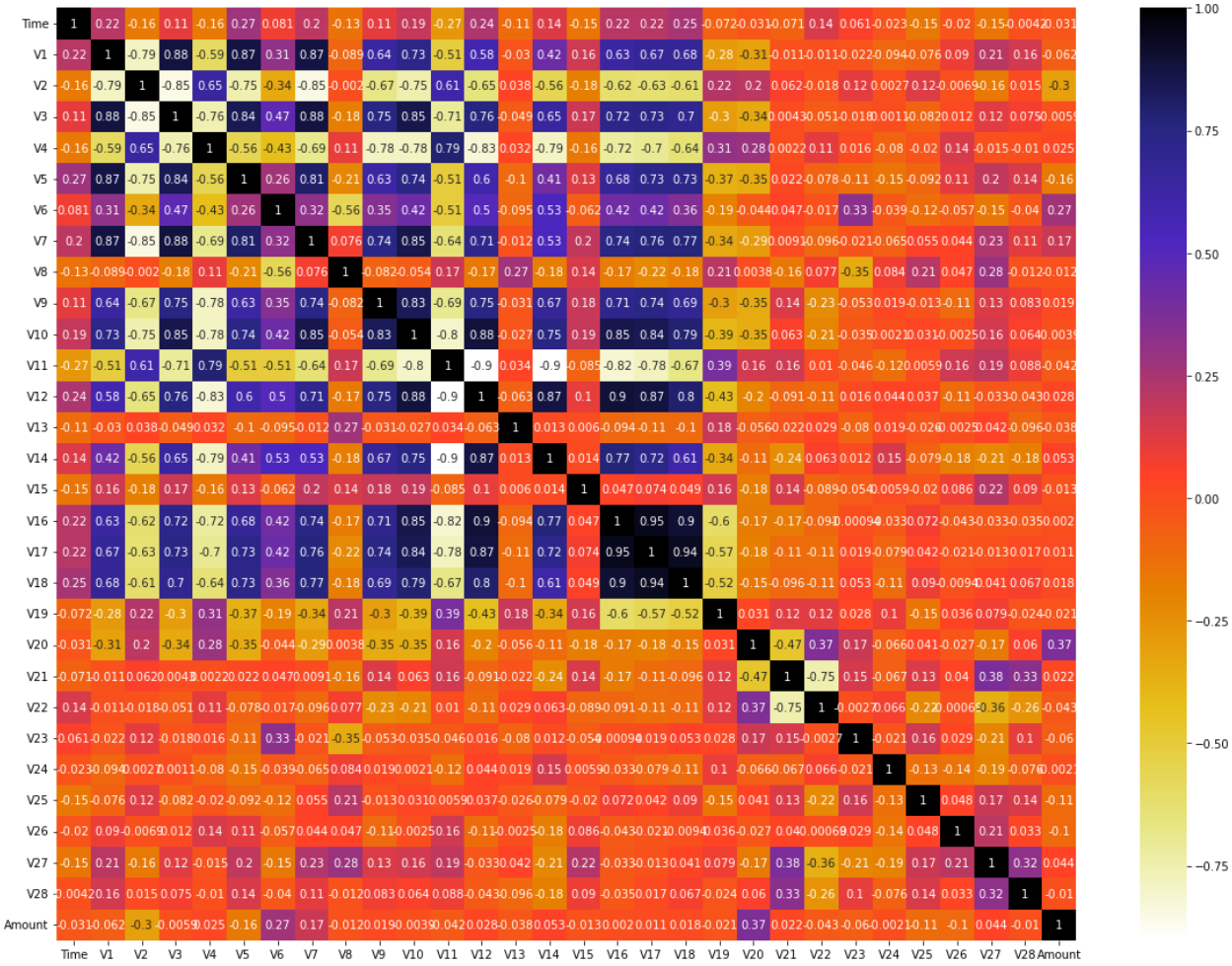

Figure 3 - Correlations

3.2 Data Preprocessing

As there are no NAs nor duplicated variables, the preparation of the dataset was having 28 variables which are pca to keep customer confidentiality and then cleaned.

3.3 Data Modeling

3.3.1 Logistic regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. The model has attained 0.9238578680203046 Accuracy score .

```
Accuracy 0.9238578680203046
Precision 0.9468085106382979
Recall 0.898989898989899
F1_score 0.9222797927461138
```

Figure 4- Metrics of Logistic regression

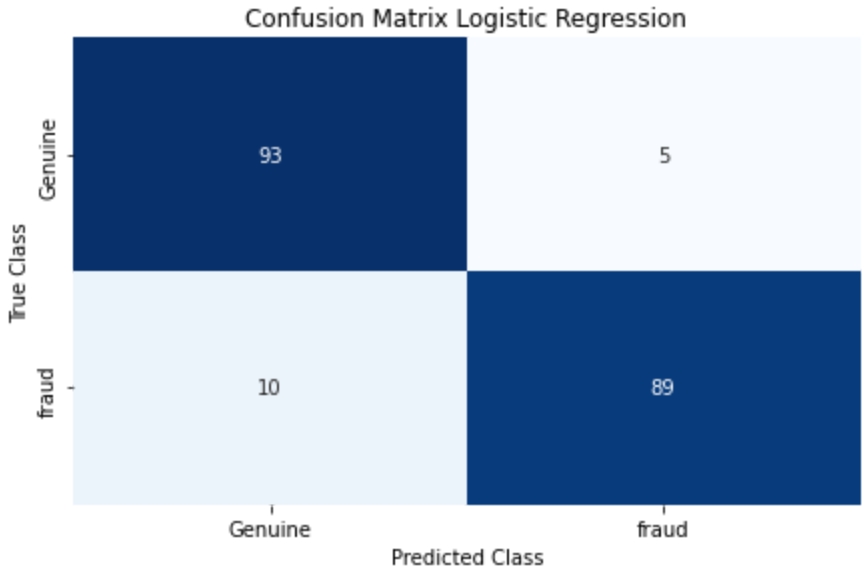

Figure 5-Confusion Matrix of logistic regression

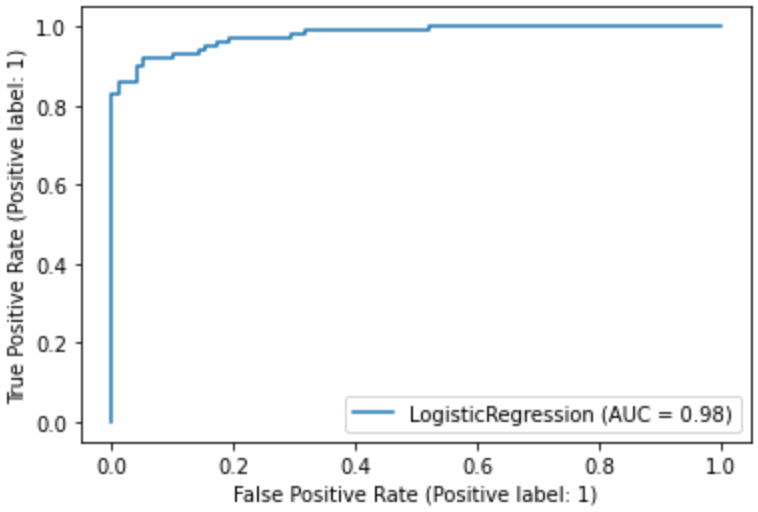

Figure -6:AUROC Curve of logistic regression

3.3.2 Decision Tree

Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

For instance, in the example below, decision trees learn from data to approximate a sine curve with a set of if-then-else decision rules. The deeper the tree, the more complex the decision rules and the fitter the model.

The model has achieved the accuracy score of 0.8883248730964467.

```
Accuracy 0.8883248730964467
Precision 0.8737864077669902
Recall 0.9090909090909091
F1_score 0.891089108910891
```

Figure 7 - Metrics of Decision tree Classifier

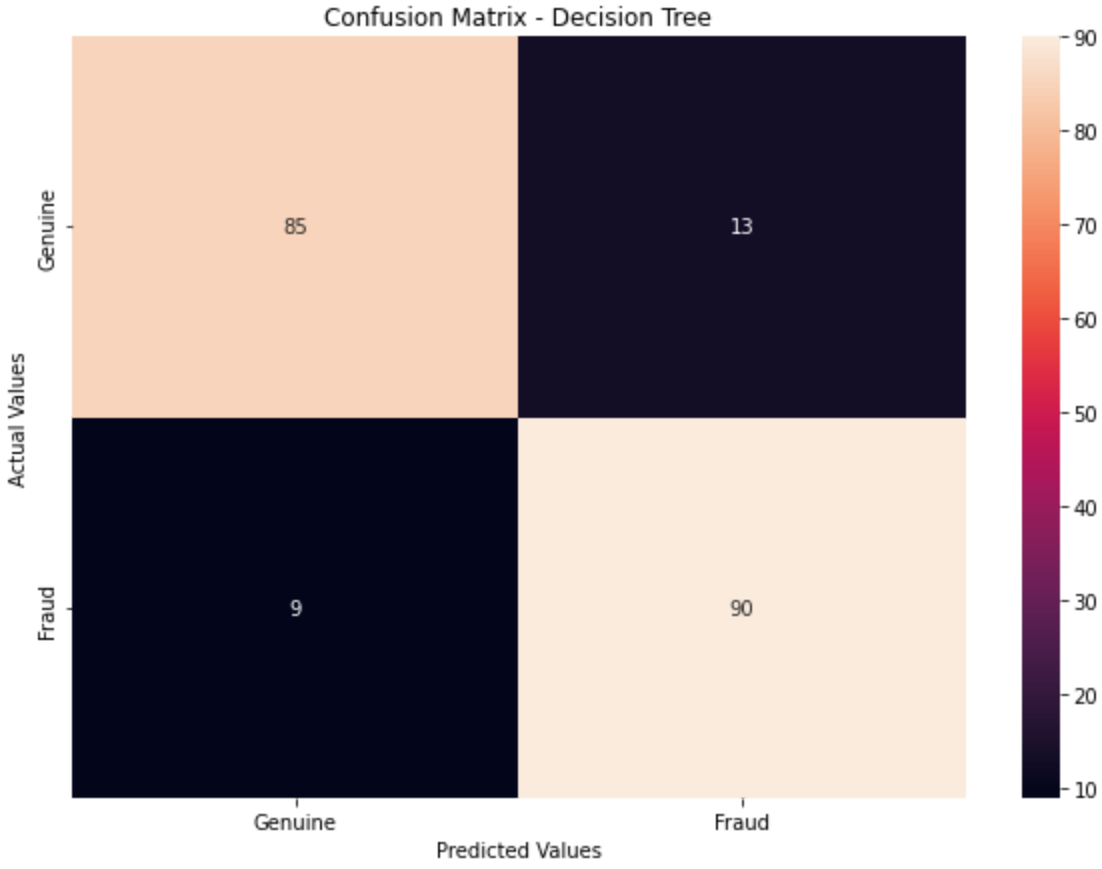

Figure 8 - Confusion matrix of Decision Tree

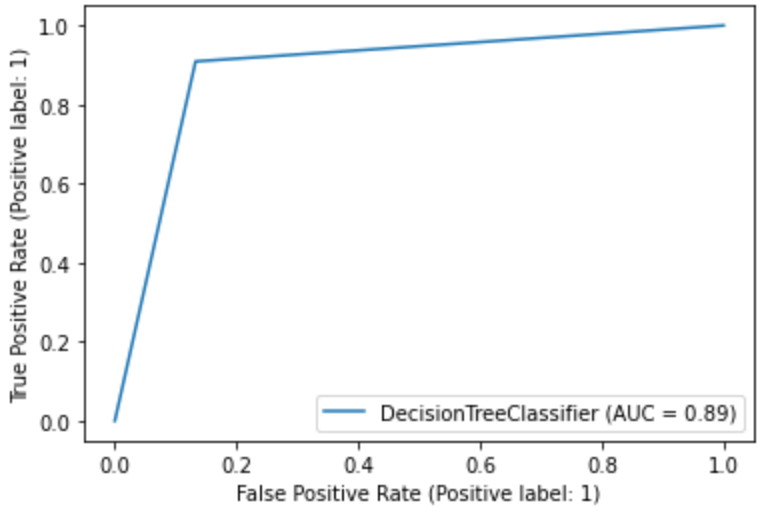

Figure 9- AUROC curve of Decision Tree classifier

3.3.1 Support Vector Machine

Support Vector machine is a supervised ML technique with connected learning algorithms which inspect data used for both classification and regression analyses, it also performs linear classification, additionally to non-linear classification by creating margins between the classes, which are created in such a fashion that the space between the margin and the classes is maximum which minimizes the error of the classification. The model has Accuracy of 0.5482233502538071.

```
Accuracy SVM: 0.5482233502538071
Precision SVM: 0.5454545454545454
Recall SVM: 0.6060606060606061
F1 Score SVM: 0.5741626794258374
```

Figure 10 -Metrics of SVM

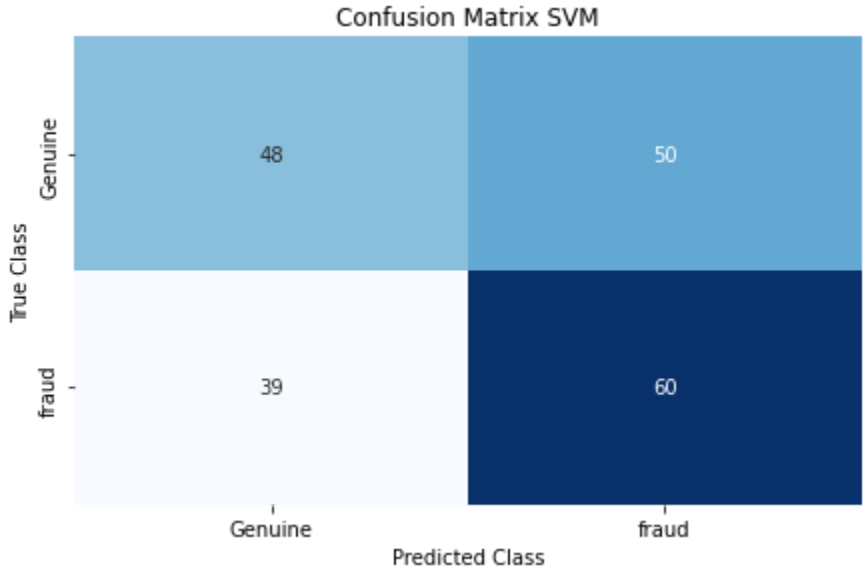

Figure 11-Confusion Matrix of SVM

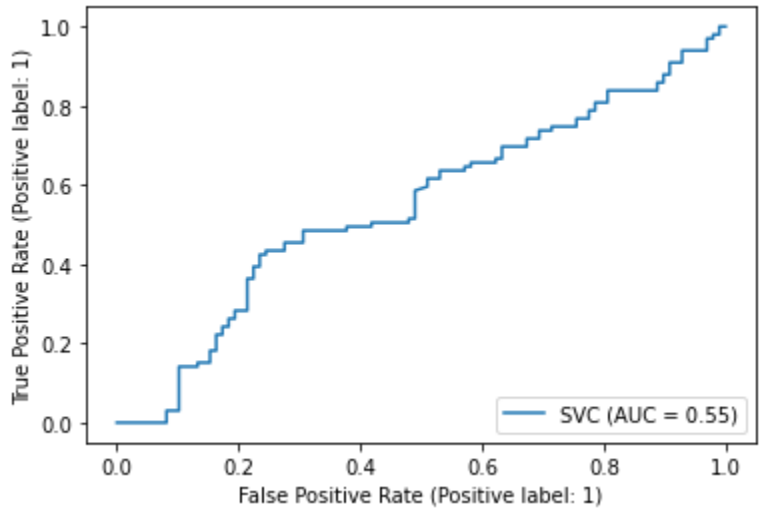

Figure 12 - AUROC Curve of SVC

3.4 Evaluation and Deployment

The last stage of the model is the evaluation and deployment stage, as presented in table 2 below all models are being compared to each other to figure the best model in identifying fraudulent credit card transactions.

Accuracy is the overall number of instances that are predicted correctly, accuracy is represented by a confusion matrix where it shows the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). True Positive represents the transactions that are fraudulent and was correctly classified by the model as fraudulent. True Negative represents the not fraudulent transactions that were correctly predicted by the model as Not fraudulent. The third rating is False positive which represents the transaction that are fraudulent but was misclassified as not fraudulent. And finally False Negative which are the not fraudulent transactions that were identified as fraudulent, table 1 below shows the confusion matrix.

Actual/Predicted	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 1 - Confusion Matrix

The table above shows all the components to calculate an accuracy of a model which is displayed in the below equation.

Model		Accuracy
Logistic Regression	Logistic Regression	92%
	Logistic Regression	
Support Vector Machine	SVM	55%
Decision tree classifier	DT	88%

Table 2-Accuracy values of model

4.Conclusion

4.1 Conclusion

In conclusion, the main objective of this project was to find the most suited model in credit card fraud detection in terms of the machine learning techniques chosen for the project, and it was met by building the three models and finding the accuracies of them all, the best model in terms of accuracies is Logistic Regression which scored 0.9238578680203046.