ILLINOIS INSTITUTE OF TECHNOLOGY

MATH 569 Final Project

Online Payment Fraud Detection

Muhammad Jamal Tariq
mtariq5@hawk.iit.edu

Shahrukh Sohail
ssohail3@hawk.iit.edu

Prof. Lulu Kang

May 1, 2022

IT IS DECLARED WITH MUTUAL AGREEMENT THAT PERCENT OF CONTRIBUTION FOR EACH MEMBER IS 50%

Table of Contents

1. Abstract

Online fraudulent transactions are a significant criminal violation. Every year, it costs people and financial institutions billions of dollars. It emphasizes the crucial importance of financial institutions in detecting and preventing fraudulent acts. Machine learning algorithms provide a proactive way mechanism to prevent online transaction frauds with high accuracy.

Online transaction fraud is a simple and easy target. E-commerce and other online sites have increased the number of online payment methods, raising the danger of online fraud. With the rise in fraud rates, machine learning approaches can be used to identify and evaluate fraud in online transactions. The primary goal of this project is to implement supervised machine learning models for fraud detection, with the goal of analyzing prior transaction information. Where transactions are classified into distinct groups based on the type of transaction. Following that, various classifiers are trained independently, and models are assessed for correctness. The classifier with the highest rating score can then be picked as one of the best approaches for predicting fraud. We worked with the Kaggle ***Synthetic Financial Datasets for Fraud Detection*** dataset collected by *Edgar Lopez-Rojas*.

In this project K Nearest Neighbor, Logistic Regression, Support Vector Machine (SVM), Decision Tree, and Random Forest Machine Learning models are implemented for detection of fraudulent transactions. A comparative analysis of these algorithms is performed to identify an optimal solution.

2. Introduction

Machine learning is frequently divided into two categories: unsupervised and supervised learning. An algorithm evaluates data and modifies its parameters based on what it has learnt from the data in unsupervised learning. In the supervised learning, the learning algorithm is given a set of training data, modifies parameters to match that data, and then applies generalizations learnt from the training set to a larger amount of data, sometimes referred to as the categorization set.

Machine learning has numerous objectives. Regression, classification, and clustering are the most prevalent. A machine learning system delivers continuous output in regression, such as fitting an equation to a point plot. The machine learning method divides items into groups based on their resemblance to one another in clustering. Classification is similar to clustering in that the machine learning algorithm seeks to categorize objects based on previously stated criteria supplied by training data. Clustering and classification are examples of supervised and unsupervised machine learning. Because the learning process has no prior knowledge of groups or classes, clustering is often unsupervised. However, classification is supervised.

Machine learning has numerous objectives. Regression, classification, and clustering are the most prevalent. A machine learning system delivers continuous output in regression, such as fitting an equation to a point plot. The machine learning method divides items into groups based on their resemblance to one another in clustering. Classification is similar to clustering in that the machine learning algorithm seeks to categorize objects based on previously stated criteria supplied by training data. Clustering and classification are examples of supervised and unsupervised machine learning. Because the learning process has no prior knowledge of groups or classes, clustering is often unsupervised. However, classification is supervised. The project uses Python programming language to for coding.

Table 2: Prediction Results

Model	Accuracy
K Nearest Neighbor	99.89%
Logistic Regression	99.81%
Support Vector Machine	99.92%*
Decision Tree	99.92%
Random Forest	99.91%*

As a result, the Decision Tree model had the greatest prediction accuracy of 99.92% and recall of 86.96%

3. Data Source

The first step of our project work was determining the right data set. Many online resources exist with access to plethora of financial fraud analysis datasets with transaction information without personal user information. We came across many data sets like *datahack* and *dataworld* data set. We selected the ***Synthetic Financial Datasets for Fraud Detection*** dataset collected by *Edgar Lopez-Rojas* for our task.

PaySim simulates mobile money transactions based on a sample of real transactions extracted from one month of financial logs from a mobile money service implemented in an African country. The original logs were provided by a multinational company, who is the provider of the mobile financial service which is currently running in more than 14 countries all around the world. This synthetic dataset was scaled down to a quarter of the original dataset and it is created just for Kaggle.

This data source is obtained from Kaggle for the detection of fraudulent online transactions. At present it consists of 6,362,620 recordings of 5 different types of transactions and 11 columns. Among the total transactions 6,354,407(99.87%) are legal transactions whereas 8,213(0.13%) are fraudulent transactions, which is understandable as only a very small percentage of the total transactions are fraud.

The 11 columns of the dataset and what each column represents:

1. step: represents a unit of time where 1 step equals 1 hour
2. type: type of online transaction
3. amount: the amount of the transaction
4. nameOrig: customer starting the transaction
5. oldbalanceOrg: balance before the transaction
6. newbalanceOrig: balance after the transaction
7. nameDest: recipient of the transaction
8. oldbalanceDest: initial balance of recipient before the transaction
9. newbalanceDest: the new balance of recipient after the transaction
10. isFraud: fraud transaction
11. isFlaggedFraud - transfer of more than 200,000 in a single transaction.

Table 1: Data types of columns

Column	Data Type	Data Type
step	Integer	int64
type	Object	object
amount	Float	float64
nameOrig	Object	object
oldbalanceOrg	Float	float64
newbalanceOrig	Float	float64
nameDest	Object	object
oldbalanceDest	Float	float64
newbalanceDest	Float	float64
isFraud	Integer	int64
isFlaggedFraud	Integer	int64

The goal is to predict the ***isFraud*** target variable whether a transaction is a legal or a fraud transaction. The main technical challenge it poses to predicting fraud is the highly imbalanced distribution between positive and negative classes in 6 million rows of data. The goal of this analysis is to solve both these issues by a detailed data exploration and cleaning followed by choosing a suitable machine-learning algorithm to deal with the skew.

4. Proposed Methodology

The goal is to predict whether a transaction is a legal transaction or a fraudulent transaction, this falls under the scope of a classification problem. We intend to deploy Supervised Machine Learning models in order to achieve the highest prediction accuracy. The data analysis process for the deployment of classification models is based on the following steps.

- **Data Acquisition**
- **Exploratory Data Analysis**
- **Feature Engineering**
- **Data Processing**
- **Supervised Machine Learning Model Deployment**
- **Results Analysis**

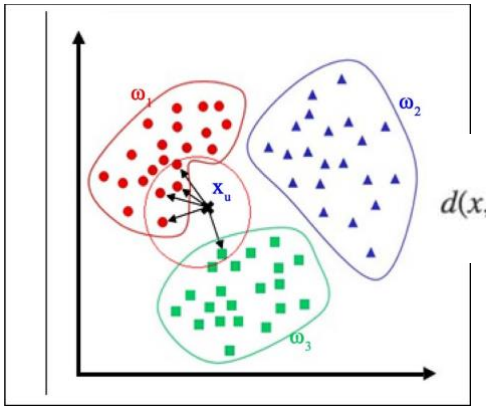
Note that the *visualization* action is performed in each step, in order to highlight new insights about the underlying patterns and relationships contained within the data.

4.1. Machine Learning Models

We will deploy 5 supervised machine learning classification algorithms to predict fraudulent transactions.

4.1.1. K Nearest Neighbor

KNN is a non-parametric classification approach for solving classification and regression issues. KNN does not do any generalization, resulting in a relatively quick training procedure. Because of the lack of generalization, the KNN training phase is either small or retains all of the training data. The value *k* (number of nearest neighbors) is user defined.

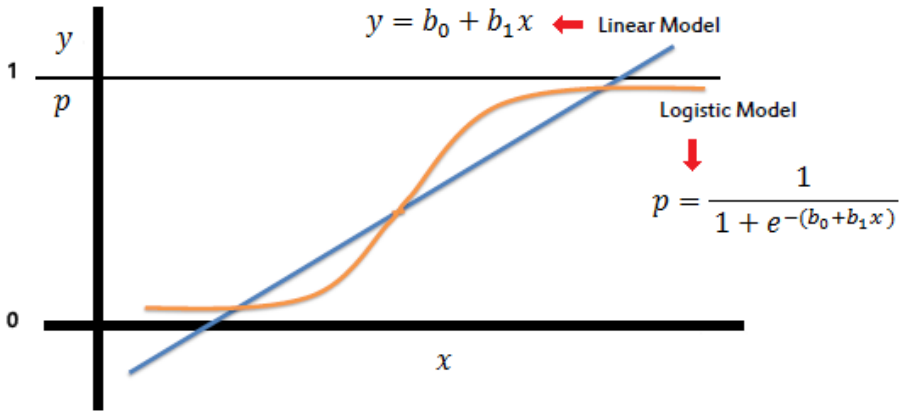


$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

K Nearest Neighbor algorithm is suitable for classification of fraud transactions, so by selecting the optimal nearest neighbor we can use K nearest neighbor to classify a transaction as legal or fraudulent.

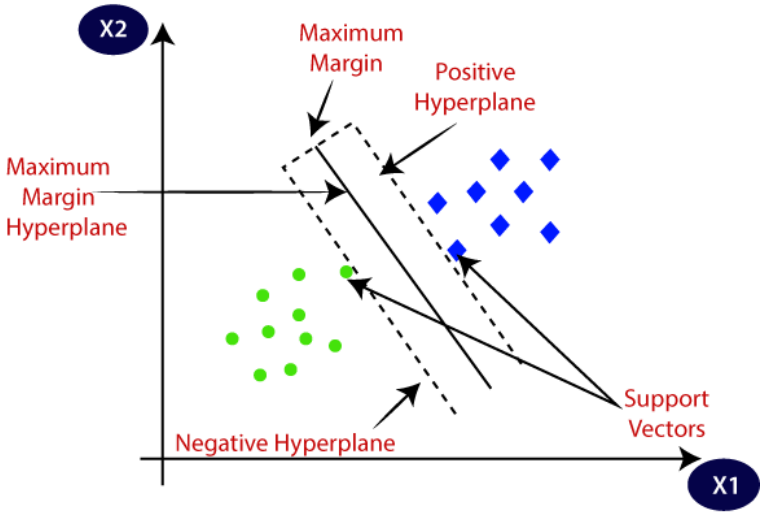
4.1.2. Logistic Regression

Logistic regression is a classification procedure that is used to forecast the likelihood of a target variable. The target or dependent variable has a dichotomous character, which means there are only two potential classes. The representation for logistic regression is an equation. To anticipate an output value, input data are linearly mixed with coefficient values. The output value is represented as a binary value, which distinguishes it from linear regression.



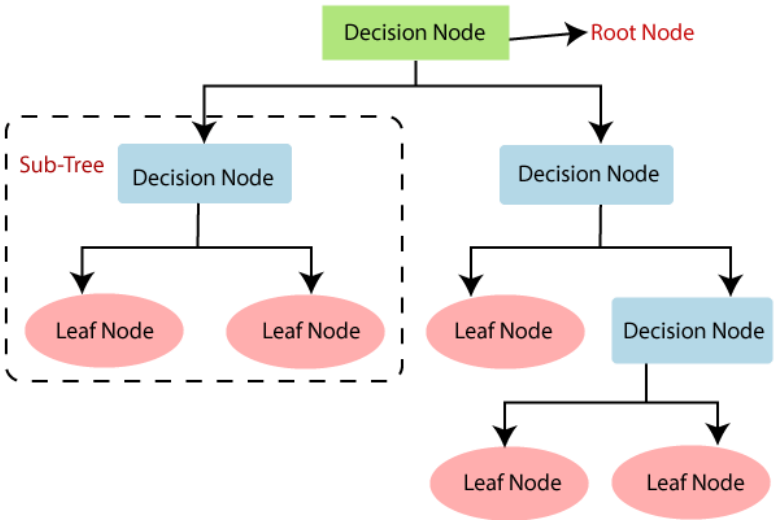
4.1.3. Support Vector Machine (SVM)

Support vector machine is a set of supervised learning methods used for classification, regression, and outlier detection. Different planes (hyperplanes) could be chosen, to separate the data points into two classes. Given a series of training examples, each labeled as belonging to one of two categories, an SVM training method constructs a model that assigns future instances to one of the two categories, resulting in a non-probabilistic binary linear classifier.



4.1.4. Decision Tree

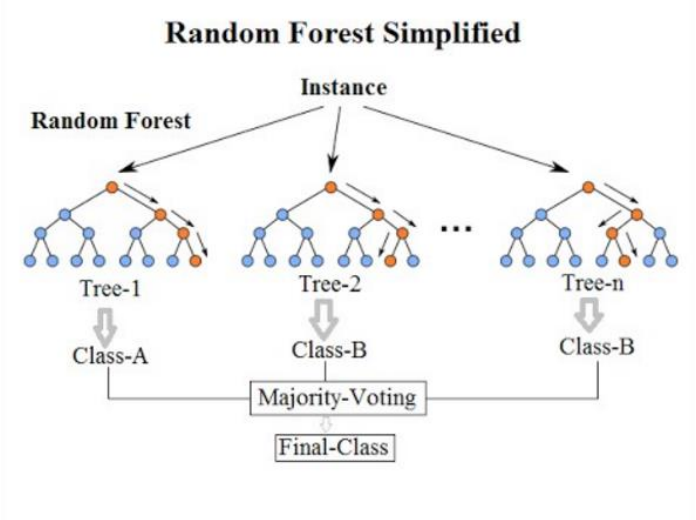
A decision tree is a decision-making tool that employs a tree-like model of decisions and their potential outcomes, such as chance event outcomes, resource costs, and utility. It is one method of displaying an algorithm that consists solely of conditional control statements. Decision trees are a prominent method in machine learning and are often used in operations research, notably in decision analysis, to assist determine the approach most likely to achieve a goal.



4.1.5. Random Forest

Random forest is an ensemble approach, because of its versatility and simplicity, it is one of the most often used algorithms. This model employs a large number of decision trees. Each of these decision trees separates a class of predictions, and the class with the most votes becomes our model's final output prediction. While growing trees in a

random forest, rather than looking for the most significant characteristics for splitting, it seeks for the best features from a random selection of features for splitting the nodes. This results in a wide range of variety, which will provide us with a more accurate model. Because there is no association between the many models developed, the models provide ensemble forecasts that are more accurate than any of the individual projections. This is due to the fact that although certain trees may be incorrect.



5. Analysis and Results

Visualizations are performed in each step, in order to highlight new insights about the underlying patterns and relationships contained within the data. The data analysis process for the deployment of classification models is based on the following steps.

- **Data Acquisition**
    1. Download data
    2. Upload data in Python environment
- **Data Exploration**

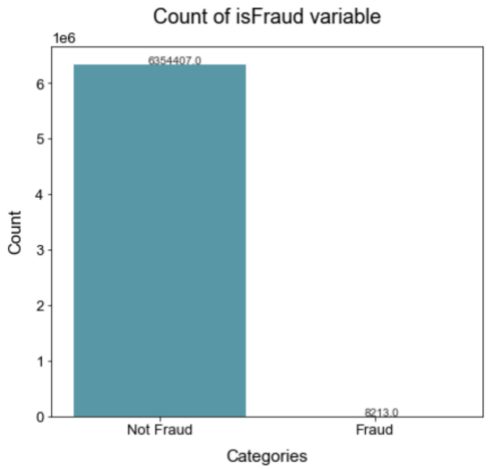
    Checking data head, info, summary statistics and null values

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0

	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
count	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06	6.362620e+06
mean	2.433972e+02	1.798619e+05	8.338831e+05	8.551137e+05	1.100702e+06	1.224996e+06	1.290820e-03	2.514687e-06
std	1.423320e+02	6.038582e+05	2.888243e+06	2.924049e+06	3.399180e+06	3.674129e+06	3.590480e-02	1.585775e-03
min	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	1.560000e+02	1.338957e+04	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
50%	2.390000e+02	7.487194e+04	1.420800e+04	0.000000e+00	1.327057e+05	2.146614e+05	0.000000e+00	0.000000e+00

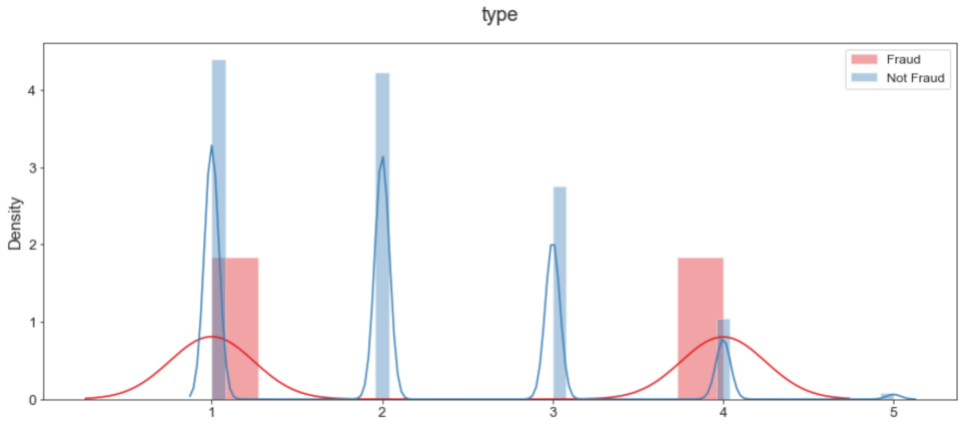
```
step              False
type              False
amount            False
nameOrig          False
oldbalanceOrg     False
newbalanceOrig    False
nameDest          False
oldbalanceDest    False
newbalanceDest    False
isFraud           False
isFlaggedFraud    False
dtype: bool
```

The dataset consists of several predictor variables and one target variable, isFraud. The dataset consists of 6362620 entries with non-null values.



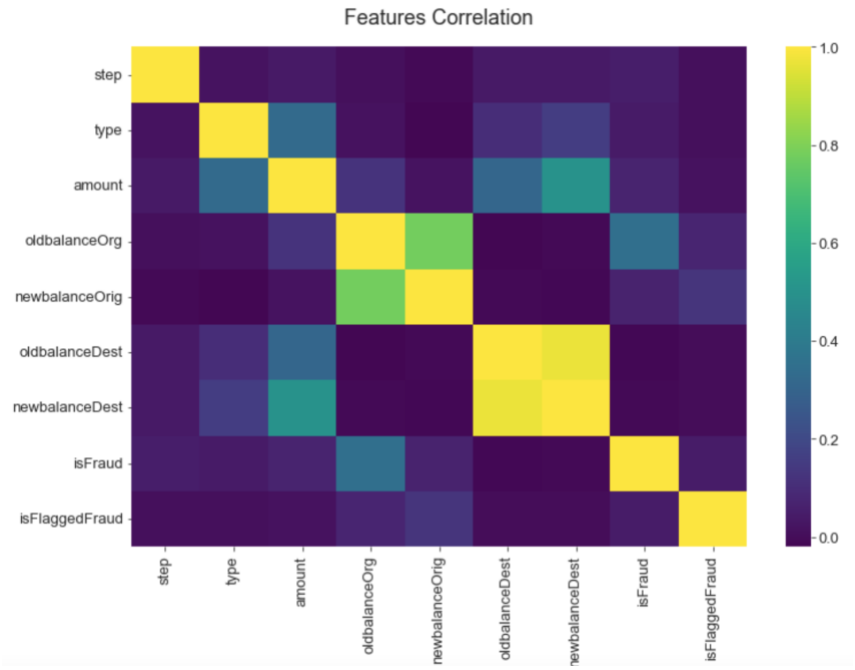
From the above graph we can say that only a handful of transactions are Fraud with a percentage of 0.1%. Some features contain 0 in the data. It doesn't make sense here because there is no change in the old balance or the new balance in some transactions. By further looking into the data, we found that it is because there are a number of factors involved in the data to correctly access the reason for the change in balance, one of them being the *type* feature.

- **Feature Engineering**



We can see from the above data that only two *type* of transactions are classified as fraud so we will drop the remaining types to generalize the data and we will only keep *Cash_out* and Transfer type.

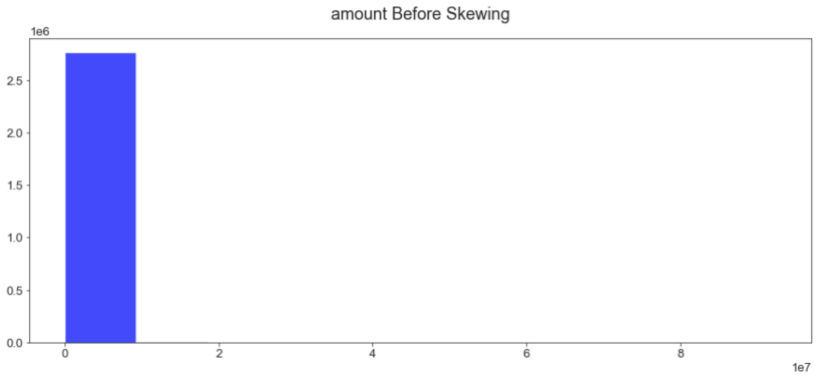
The Type feature in our data is categorical so we will map it to convert it to numerical data 6,3544,407 transactions were **Not Fraud** transactions with 2762196 Not Fraud transactions after considering only two types which are relevant with only 0.3% Fraud transactions. This shows us that we have a very imbalanced data.

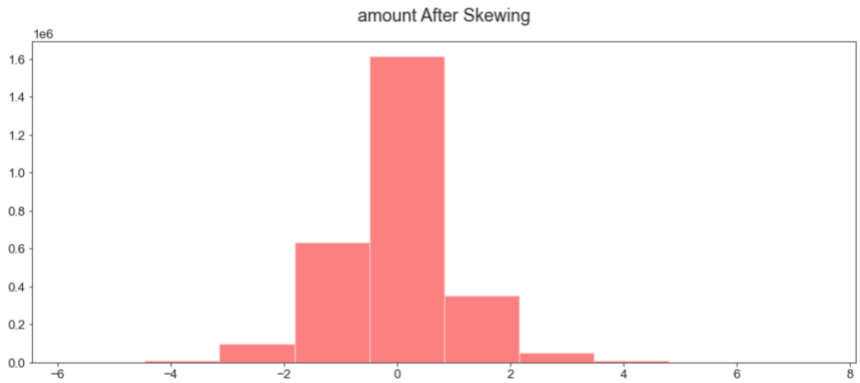


From the above correlation graph, we can see that highly correlated features are newbalanceOrig vs oldbalanceOrg, newbalanceDest vs oldbalanceDest and newbalanceDest vs amount.

- **Data Processing**

After feature engineering we scaled the data.

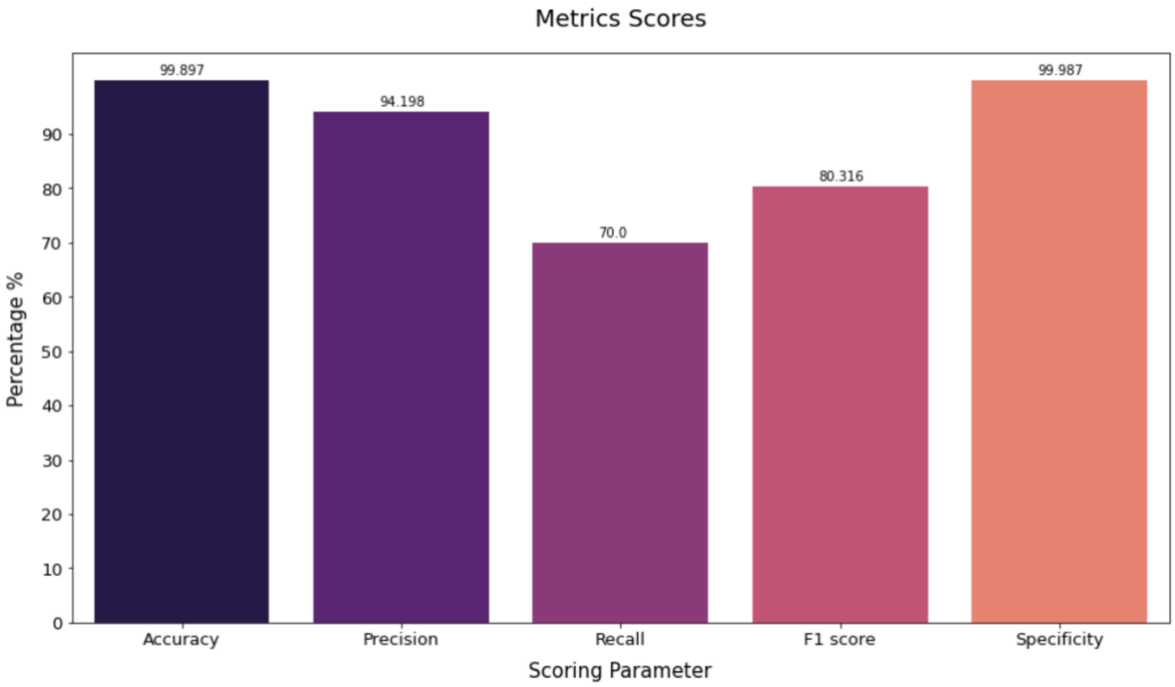




We dropped the uninterested and unscaled features from the dataset.

- **Supervised Machine Learning**
    1. Modeling of the classifier system
    2. Validation of the model
    3. Visualization of the model
    4. Visualization and interpretation of the results
- **Results analysis**
    1. Deployment of the models
    2. Comparing prediction accuracy of ML models
    3. Visualization of the results

Results were compiled and visualized, KNN model results were as follows:



6. Conclusions

The goal was to predict whether a transaction is a legal transaction or a fraudulent transaction, this falls under the scope of a classification problem. We intend to deploy Supervised Machine Learning models in order to achieve the highest prediction accuracy.

K Nearest Neighbor, Logistic Regression, Support Vector Machine, Decision Tree and Random Forest models were trained using k-fold technique, training contained total 5 folds and with each fold accuracy of the model kept increasing up to 5th fold. After the 5th fold, accuracy started decreasing because our dataset was not sufficient enough for more than 5 folds. So, the final model was trained on 5 folds with 88.55% average accuracy. This means that if someone would train Random Forest with a bigger data set using the k-fold technique then the average accuracy of the model would be even higher.

Table 3: Prediction Results

Model	Accuracy	Recall
K Nearest Neighbor	99.89%	70.0%
Logistic Regression	99.81%	41.55%
Support Vector Machine	99.92%*	---
Decision Tree	99.92%	86.96%
Random Forest	99.91%*	---

As a result, the Decision Tree model had the greatest prediction accuracy of 99.92% and recall of 86.96%

Due to huge amount of data models for Support Vector Machine and Random Forest were unable to compile, even on Google Collab. Further work can be done by under sampling of data by 50:50, that would reduce data size even more and as a result SVM and Random Forest results can be compiled accurately.

*Initial results, Final results could not be compiled due to lack of sufficient computing power.

7. Bibliography and Credits

7.1. Research Papers

[1] Design and development of financial fraud detection using machine learning. (2020). International Journal of Emerging Trends in Engineering Research, 8(9), 5838–5843.
https://doi.org/10.30534/ijeter/2020/152892020

[2] Rucco, M., Giannini, F., Lupinetti, K., & Monti, M. (2019). A methodology for part classification with supervised machine learning. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, *33*(1), 100–113. https://doi.org/10.1017/S0890060418000197

[3] Saarikoski, J., Joutsijoki, H., Järvelin, K., Laurikkala, J., & Juhola, M. (2015). On the influence of training data quality on text document classification using machine learning methods. International Journal of Knowledge Engineering and Data Mining, 3(2), 143.
https://doi.org/10.1504/IJKEDM.2015.071284

7.2. Dataset

[11] https://www.kaggle.com/code/netzone/eda-and-fraud-detection/data