Credit Card Fraud Detection

**AIM:** This report aims to develop an efficient and accurate model for detecting fraudulent credit card transactions using Machine Learning algorithms.

**Introduction:**

The need to investigate more sophisticated

techniques has arisen as sophisticated fraud has made clear the limitations of conventional

rule-based systems

The need to investigate more sophisticated

techniques has arisen as sophisticated fraud has made clear the limitations of conventional

rule-based systems

Banks used to provide only in-person services to customers until 1996 when the first internet banking application was introduced in the United States of America by Citibank and Wells Forgo Bank. After the introduction of internet banking, the use of credit cards over the internet was adopted.

A credit card generally refers to a card that is assigned to the customer (cardholder), usually allowing them to purchase goods and services within credit limit or withdraw cash in advance. Credit card provides the cardholder an advantage of the time, i.e., it provides time for their customers to repay later in a prescribed time, by carrying it to the next billing cycle.

Credit card Frauds are easy targets so as the frequency of transactions increases, the number of fraudulent activities also increases. Fraudulent activities in financial transactions pose significant risks and economic losses to businesses and individuals alike. Detecting such activities in a timely and accurate manner is crucial for minimizing these losses and maintaining trust in financial systems. With the advent of big data and machine learning, it is now possible to analyze vast amounts of transaction data to identify patterns indicative of fraud.

**Objectives:**

The objectives of this paper follow:

Investigate the use of machine learning algorithms for fraud detection in financial

transactions.

Investigate the use of machine learning algorithms for fraud detection in financial

transactions.

Investigate the use of machine learning algorithms for fraud detection in financial

transactions.

Investigate the use of machine learning algorithms for fraud detection in financial

transactions.

Investigate the use of machine learning algorithms for fraud detection in financial

transactions.

1. Investigate the use of machine learning algorithms for fraud detection in financial transactions such as credit cards.
2. **Identify Key Features for Fraud Detection**
3. Compare the Performance of Different Models

**Research Gap:**

Handle Imbalance in the dataset

**Literature Review:**

In recent years, there has been a lot of study on applying machine learning algorithms to

detect fraud in financial transactions. Various strategies and algorithms have been examined

in several research to increase the precision and effectiveness of fraud detection systems.

In recent years, there has been a lot of study on applying machine learning algorithms to detect fraud in financial transactions. Various strategies and algorithms have been examined in several kinds of research to increase the effectiveness and precision of fraud detection systems.

**Supervised Learning Approaches:**

Logistic regression is a technique that is used to predict an outcome variable that is binary. This technique does not demand that explanatory variables follow normal distribution, or correlated. The outcome variable in Logistic Regression models is qualitative. This method is the most common method for fraud detection tasks.

Decision tree is a non-linear classification technique that divides a sample into increasingly smaller subgroups using a collection of explanatory variables. At each branch of the tree, the process iteratively chooses the explanatory variable that, in accordance with a predetermined criterion, has the strongest correlation with the outcome variable.

Random Forest is a supervised machine learning algorithm that uses a group of decision tree models for classification and making predictions. Each decision tree is a weak learner because they have a low predictive power. It is based on ensemble learning, which uses many decision tree classifiers to classify a problem and improve the accuracy of the model. RF has the power of handling large datasets with higher dimensionality. It can handle thousands of input variables and identify most significant variables so it is considered as one of the dimensionality reduction methods. Further, the model outputs Importance of variable, which can be a very handy feature.

Support Vector Machine has ability to handle high-dimensional data and nonlinear relationships has led to their use in fraud detection as well. SVMs look for an ideal hyperplane that can distinguish between fraudulent and legal transactions with the greatest margin. at dealing with unbalanced datasets, SVMs have shown to perform well at classifying fraudulent transactions.

**Unsupervised Learning Approaches:**

For spotting fraud in numerous domains, unsupervised learning techniques like clustering and anomaly detection have been investigated. The goal of these strategies, which do not require labelled data, is to find patterns and anomalies in the data that may point to fraudulent activity. clusters are groups of similar features and anomaly detection (also called outlier detection) is any instance that has a low affinity to all the clusters is likely to be an anomaly.

These are some ways to implement fraud detection model, this study focuses on Random Forest and Logistic Regression.

**Methodology:**

**Data Description:**

The Dataset used for this study is a synthetic dataset generated to train different ML algorithms, it contains 6362620 rows of observations which has 11

columns of variables. The variables contain information on the type of the

transaction, amount of transaction, names of receiver and sender, account balance information before and after the transaction, and fraud status.

**Data Wrangling:**

In this stage, data is observed then cleaning begins which improves the data

quality. steps are as follows:

* Handling missing values: - Impute them with mean or median, simply drop
* Correcting data types: Format accordingly
* Removing duplicates: Drop
* Addressing outliers and anomalies: Format accordingly

The dataset used in this study is free of missing values, and duplicates. All variables have the correct data type.

To find outliers Interquartile Range (IQR) method was used, result of the test

shows potential outliers using IQR method are 2043214. Removing outliers and

Capping them are common method to treat outliers but both of these methods affect the target variable (‘isFraud’) resulting in loss of critical information needed to identify frauds hence researcher decided to keep outlier data points

**Exploratory Data Analysis:**

Data visualization can be a valuable step to gain insights into the dataset and

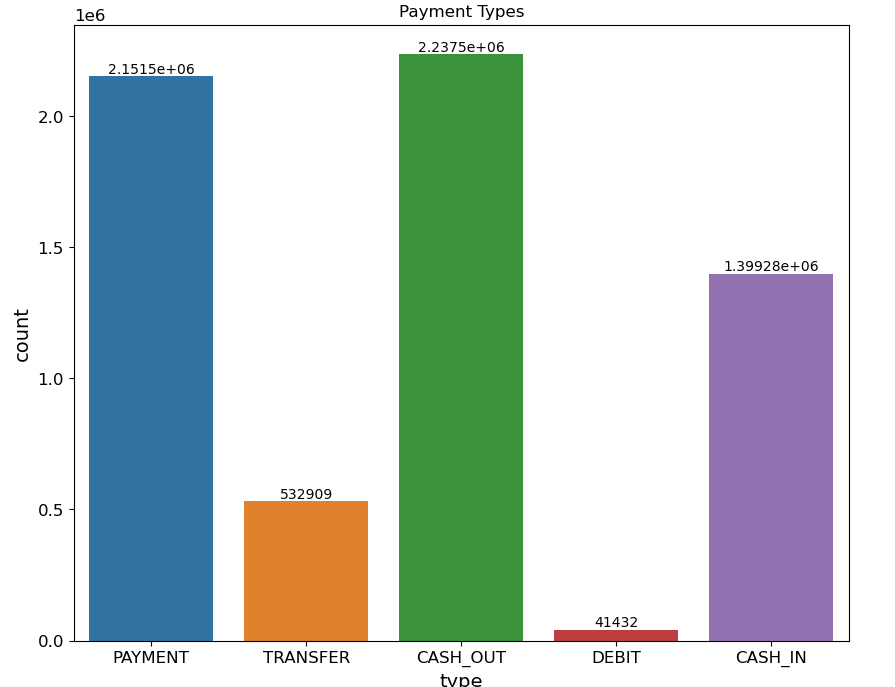
understand its characteristics. Seaborn and Matplotlib python libraries are

used for creating visualizations.

The box plot will show the distribution of transaction amounts for both fraudulent (isFraud=1) and non-fraudulent (isFraud=0) transactions



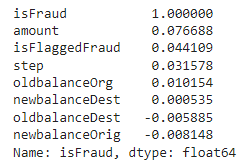
The count plot will show the frequency distribution of different transaction types (e.g., 'PAYMENT', 'TRANSFER', etc.) in the dataset.



Cash out and Payment are the most frequent transaction types.

The numerical variables in the dataset are highly skewed please refer google colab link at the end of this report for visualization of numerical data.

Now to check any correlation between variables ‘.corr()’ a pandas method will calculate the correlation coefficients between numerical columns in a DataFrame.



If the Correlation coefficient is closer to 1 it indicates a positive linear relationship, closer to – 1 indicates a negative linear relationship, and closer to 0 indicates no relationship.

**Data Preprocessing and Model Implementation:**

Before implementing ML model data must be ready for implementation and training.

Drop the unnecessary columns which helps in building a more efficient, effective, and interpretable model.

converts categorical variables into binary vectors, which is done by encoders or categorical encoding. Scikit-learn the Python module provides various types of encoding options like OneHotEncoder, TargetEncoding, and LabelEncoding. For this study ***‘OneHotEncoder’*** is used which converts categorical variables into a binary matrix representation. Each category is transformed into a column with binary values (0 or 1), indicating the presence of that category. This prevents the model from assuming any ordinal relationship between categories.

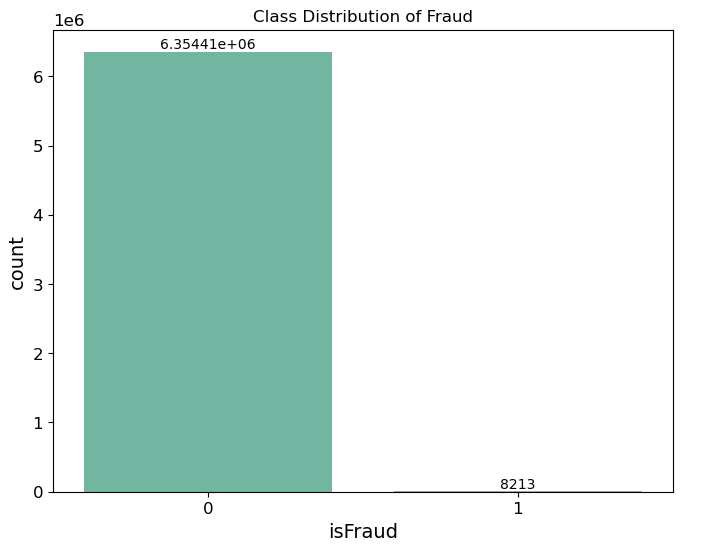
Normalization scales numerical features to a specific range, typically 0 to 1, to ensure that each feature contributes equally to the model. This process improves the performance of algorithms. ***‘StandardScaler()’*** standardizes numerical features to have a mean of 0 and a standard deviation of 1, which helps in improving the performance of many algorithms by normalizing the data.

Both these preprocessing operations are applied using ***‘Pipeline’***. Pipeline is a sequence of data processing steps arranged in a specific order, where the output of one step serves as the input to the next. This ensures a streamlined and reproducible workflow for data preprocessing, and to apply these preprocessing steps together in a pipeline ***‘ColumnTransformer’*** a scikit-learn tool is used.

The first model is ***‘RandomForestClassifier’.*** Random forest is an ensemble learning method which uses several decision trees. Each decision tree in the random forest performs the classification. The final output of the classification is determined by voting. Tree based classifiers generally work well with imbalanced data.

The Second model is ***‘LogisticRegression’*** a binary classification algorithm simple but struggles with class imbalance because it tends to be biased towards the majority class, leading to poor performance in predicting the minority class.

The count plot will display imbalance in dataset



Since data is highly imbalanced the researcher has used resampling methods like RandomUnderSampling, SMOTE, SMOTEENN.

* **Random Under-Sampling**: This technique reduces the number of majority class samples to balance the dataset by randomly removing some of the majority class instances. While simple and effective, it can result in loss of valuable information.
* **SMOTE (Synthetic Minority Over-sampling Technique)**: SMOTE generates synthetic samples for the minority class by interpolating between existing minority instances. This helps balance the dataset without losing information, improving model performance on the minority class.
* **SMOTEENN**: A combination of SMOTE and Edited Nearest Neighbors (ENN), this method first applies SMOTE to create synthetic minority class samples and then uses ENN to clean the dataset by removing noisy and misclassified samples from both classes, enhancing data quality and balance.

**Evaluation metrics:**

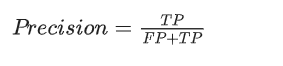
Accuracy = number of correct predictions made divided by total number of observations. Accuracy can be misleading in classification tasks



Recall = the number of positive predictions divided by the number of positive class values in the test data. A low recall indicates many false negatives.



**Precision** = It measures the ratio of correctly classified fraud transactions to the total transactions predicted to be fraud transactions.



Confusion matrix = It displays the counts of true positives (correctly predicted positives), true negatives (correctly predicted negatives), false positives (incorrectly predicted positives), and false negatives (incorrectly predicted negatives).

ROC AUC Score (Receiver Operating Characteristic - Area Under the Curve) = It measures the model's ability to distinguish between the positive and negative classes.

**Conclusion:**.

In order to categorize online credit card transactions as either fraud or not, this study built different classification models Logistics Regression, and Random Forest using supervised machine learning. To prevent class imbalance and overfitting and to ensure the model does not favor solely the majority class we balanced data using resampling techniques. This study also implemented Random Forest model which has the highest precision score of 0.978 but The model that performed best is the **Logistic Regression with SMOTEENN**. It has the highest ROC AUC score (0.9672), indicating strong overall performance in distinguishing between classes. Additionally, it has high recall (0.9807), ensuring it captures most of the fraudulent transactions, which is crucial for fraud detection.

**Google colab link:**

[**https://colab.research.google.com/drive/1QiMjzCVUaVddZg4LNAKvx-EuiPJIu21d?usp=sharing**](https://colab.research.google.com/drive/1QiMjzCVUaVddZg4LNAKvx-EuiPJIu21d?usp=sharing)

**Dataset Link:**

[**https://drive.google.com/file/d/1DyL5Pw0x-sqlZkgRvzP5Wfj\_mXXdVygD/view?usp=sharing**](https://drive.google.com/file/d/1DyL5Pw0x-sqlZkgRvzP5Wfj_mXXdVygD/view?usp=sharing)

**References:**

Joyce Annie George (2020). *Credit Card Fraud Detection: A Case Study for Handling Class Imbalance*. [online] Medium. Available at: https://medium.com/analytics-vidhya/credit-card-fraud-detection-a-case-study-for-handling-class-imbalance-f81abf997421 [Accessed 21 Jul. 2024].

https://www.facebook.com/jason.brownlee.39 (2016). *8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/>.

Afriyie, J.K., Tawiah, K., Pels, W.A., Addai-Henne, S., Dwamena, H.A., Owiredu, E.O., Ayeh, S.A. and Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. *Decision Analytics Journal*, 6(100163), p.100163. doi:https://doi.org/10.1016/j.dajour.2023.100163.

@phdthesis{phdthesis,

author = {Kazeem, Oladimeji},

year = {2023},

month = {09},

pages = {},

title = {FRAUD DETECTION USING MACHINE LEARNING},

doi = {10.13140/RG.2.2.12616.29441}

}

Normalization scales numerical features to a specific range, typically 0 to 1, to ensure that each feature contributes equally to the model. This process improves the performance of algorithms Normalization scales numerical features to a specific range, typically 0 to 1, to ensure that each feature contributes equally to the model. This process improves the performance of algorithmsThe ability of Support Vector Machines (SVMs) to handle high-dimensional data and

nonlinear relationships has led to their use in fraud detection as well. SVMs look for an ideal

hyperplane that can distinguish between fraudulent and legal transactions with the greatest

margin. at dealing with unbalanced datasets, SVMs have shown to perform well at

classifying fraudulent transactions.

The ability of Support Vector Machines (SVMs) to handle high-dimensional data and

nonlinear relationships has led to their use in fraud detection as well. SVMs look for an ideal

hyperplane that can distinguish between fraudulent and legal transactions with the greatest

margin. at dealing with unbalanced datasets, SVMs have shown to perform well at

classifying fraudulent transactions.