CREDIT CARD FRAUD DEDUCTION USING DATA SCIENCE

PHASE-1

DATA SCIENCE

Data science is an interdisciplinary field that encompasses a set of techniques, processes, and methods for extracting valuable insights, knowledge, and patterns from data. It combines elements of statistics, computer science, domain expertise, and data engineering to collect, analyze, and interpret large and complex datasets. The ultimate goal of data science is to use data to inform decision-making, solve problems, and drive improvements in various domains, including business, healthcare, finance, and more. Data scientists use a combination of data analysis, machine learning, data visualization, and domain-specific expertise to uncover meaningful information from data and provide valuable insights for organizations and individuals.

Tools used

1. **Python**
   Python is often used as a support language for software developers, for build control and management, testing, and in many other ways.
2. **Python idle**
   IDLE can be used to execute a single statement and create, modify, and execute Python scripts. IDLE provides a fully-featured text editor to create Python scripts that include features like syntax highlighting, autocompletion, and smart indent.
3. **PyCharm**
   PyCharm is a dedicated Python Integrated Development Environment (IDE) providing a wide range of essential tools for Python developers, tightly integrated to create a convenient environment for productive Python, web, and data science development
4. **A Kaggle dataset**
   It also known as a Kaggle Kernel is a set of data provided by companies, students, and alum. These data sets allow competitors to work through problems, or use them as a practice simulation.
5. **MATLAB**
   In Data Science, MATLAB is used for simulating **neural networks** and fuzzy logic. Using the MATLAB graphics library, you can create powerful visualizations. MATLAB is also used in image and signal processing.
6. **Excel**
   Excel is a powerful **analytical tool for Data Science**. While it has been the traditional tool for data analysis, Excel still packs a punch. Excel comes with various formulae, tables, filters, slicers, etc. You can also create your own custom functions and

formulae using Excel. While Excel is not for calculating the huge amount of Data, it is still an ideal choice for creating powerful data visualizations and spreadsheets.

7. **Jupyter**
   Project **Jupyter** is an open-source tool based on IPython for helping developers in making open-source software and experiences interactive computing. Jupyter supports multiple languages like Julia, **Python**, and R.
8. **Matplotlib**
   Matplotlib is a plotting and visualization library developed for Python. It is the most popular tool for generating graphs with the analyzed data. It is mainly used for plotting complex graphs using simple lines of code. Using this, one can generate bar plots, histograms, scatterplots etc.
9. **Sciket-learn**
   Scikit-learn is a library-based in Python that is used for implementing Machine Learning Algorithms. It is simple and easy to implement a tool that is widely used for analysis and data science. Scikit-learn makes it easy to use complex machine learning algorithms. It is therefore in situations that require rapid prototyping and is also an ideal platform to perform research requiring basic Machine Learning. It makes use of several underlying **libraries of Python** such as SciPy, Numpy, Matplotlib, etc.

CREDIT CARD FRAUD DETECTION

Credit card fraud is a term that has been coined for unauthorized access of payment cards like credit cards or debit cards to pay for using services or goods. Hackers or fraudsters may obtain the confidential details of the card from unsecured websites. When a fraudster compromises an individual's credit/debit card, everyone involved in the process suffers, right from the individual whose confidential data has been leaked to the businesses (generally banks) who issue the credit card and the merchant who is finalizing the transaction with purchase. This makes it extremely essential to identify the fraudulent transactions at the onset. Financial institutions and businesses like e-commerce are taking firm steps to flag the fraudsters entering the system. Various advanced machine learning technologies are at play, assessing every transaction and stemming the fraud users in its nip using behavioral data and transaction patterns. The process of automatically differentiating between fraudulent and genuine users is known as "credit card fraud detection".

Lost/Stolen cards: People steal credit cards from the mail and use them illegally on behalf of the owner. The process of blocking credit cards that have been stolen and re-issuing them is a hassle for both customers and credit card companies. Some financial institutions keep the credit cards blocked until it is verified that the rightful owner has received the card.

Card Abuse: The customer buys goods and items on the credit card but has no intention to pay back the amount charged by the bank for the same. These customers stop answering the calls as the deadline to settle the dues approaches. Sometimes they even declare bankruptcy—this type of fraud results in losses of millions every year.

Identity Theft: The customers apply illegitimate information, and they might even steal the details of a genuine customer to apply for a credit card and then misuse it. In such cases, even card blocking can not stop the credit card from falling into the wrong hands.

Merchant Abuse: Some merchants show illegal transactions (that never occurred) for money laundering. For performing these illicit transactions, legal information of genuine credit card users is stolen to generate replicas of the cards and use it for illegal work.

CREDIT CARD FRAUD DETECTION USING DATA SCIENCE

Data science is used for credit card fraud detection because it offers powerful tools and techniques to identify suspicious patterns and anomalies in large volumes of transaction data. Here's why data science is particularly well-suited for this task:

1. **Handling Big Data:**
   Credit card transactions generate massive amounts of data. Data science techniques, including big data processing frameworks like Hadoop and Spark, enable the efficient handling and analysis of this data.
2. **Pattern Recognition:**
   Data science algorithms excel at recognizing patterns within data. They can Identify unusual behaviors or transactions that deviate from established patterns, which is crucial for fraud detection.
3. **Real-Time Analysis:**
   Many data science models can analyze transactions in real-time, allowing for **immediate** detection and response to suspicious activities as they occur.
4. **Adaptability:**
   Fraudsters constantly evolve their tactics, so fraud detection systems need to adapt. Data science models can be regularly retrained to stay up-to-date with emerging fraud patterns.
5. **Anomaly Detection:**
   Anomaly detection techniques in data science can flag transactions that are statistical outliers, helping to pinpoint potentially fraudulent activity.

6. **False Positive Reduction:**
   Data science can help reduce false positives (legitimate transactions mistakenly I identified as fraud), which is crucial to avoid inconveniencing genuine cardholders.
7. **Ensemble Techniques:**
   Combining multiple machine learning models using ensemble methods can improve fraud detection accuracy and reduce false alarms.
8. **Predictive Analytics :**
   Data science can be used not only to detect ongoing fraud but also to predict future fraudulent activities based on historical data and trends.

ALGORITHMS USED IN CREDIT CARD FRAUD DEDUCTION

1.Decision Tree

2. Predictive Analytics and Algorithms

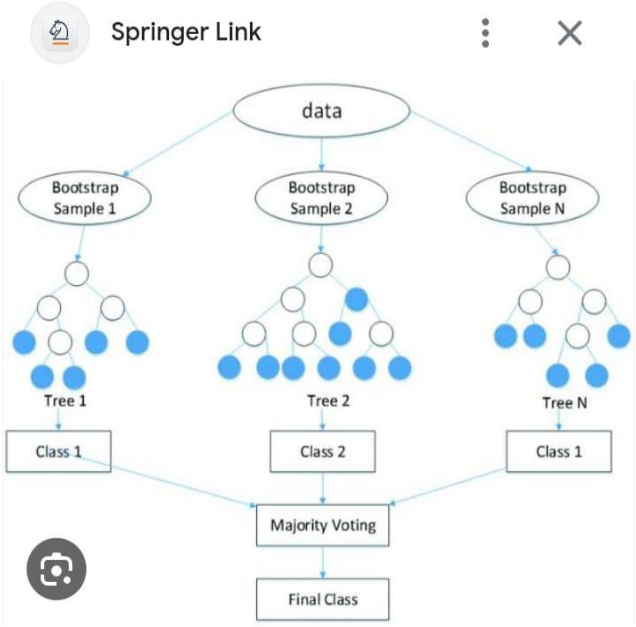
3. Clustering Techniques

4. K-Nearest Neighbour Algorithm

5.Neural Networks

6.Naive Bayes Classifiers

7.Support Vector Machines (SVMs)



Problem Definition:

The problem is to develop a machine learning-based system for real-time credit card fraud detection. The goal is to create a solution that can accurately identify fraudulent transactions while minimizing false positives. This project involves data preprocessing, feature engineering, model selection, training, and evaluation to create a robust fraud detection system.

Steps involved in design thinking:

1.DATA SOURCE

In credit card fraud detection using data science, the primary data source is typically a dataset containing transaction data. Here are the key data sources used in this type of project:

- ➢ **Transaction Data** : This is the most critical data source. It includes details of each credit card transaction, such as: - Transaction amount - Timestamp (date and time of the transaction) - Merchant information (merchant ID, location, category) - Card details (credit card number, expiration date)
- ➢ **Historical Data** : - Historical transaction data is used to train machine learning models. It provides a basis for the model to learn patterns of legitimate and fraudulent transactions. - The data sets are used to find the detection using kaggle website datasets

  Link: https://www.kaggle.com/datasets/mig-ulb/creditcardfraud
- ➢ **Label Data** : This dataset includes labels or tags that specify whether each transaction is fraudulent or not. Labeling can be done manually by fraud analysts or through historical fraud data.
- ➢ **Customer Information** : While not always included, customer data can be useful in fraud detection. It may include details such as customer demographics, transaction history, and account status.
- ➢ **External Data :** Some projects incorporate external data sources to enhance fraud detection. This could include data from third-party fraud detection services, economic indicators, or geolocation data.
- ➢ **data model** : We use the logistics Regression Model to find the fraud detection using data set .

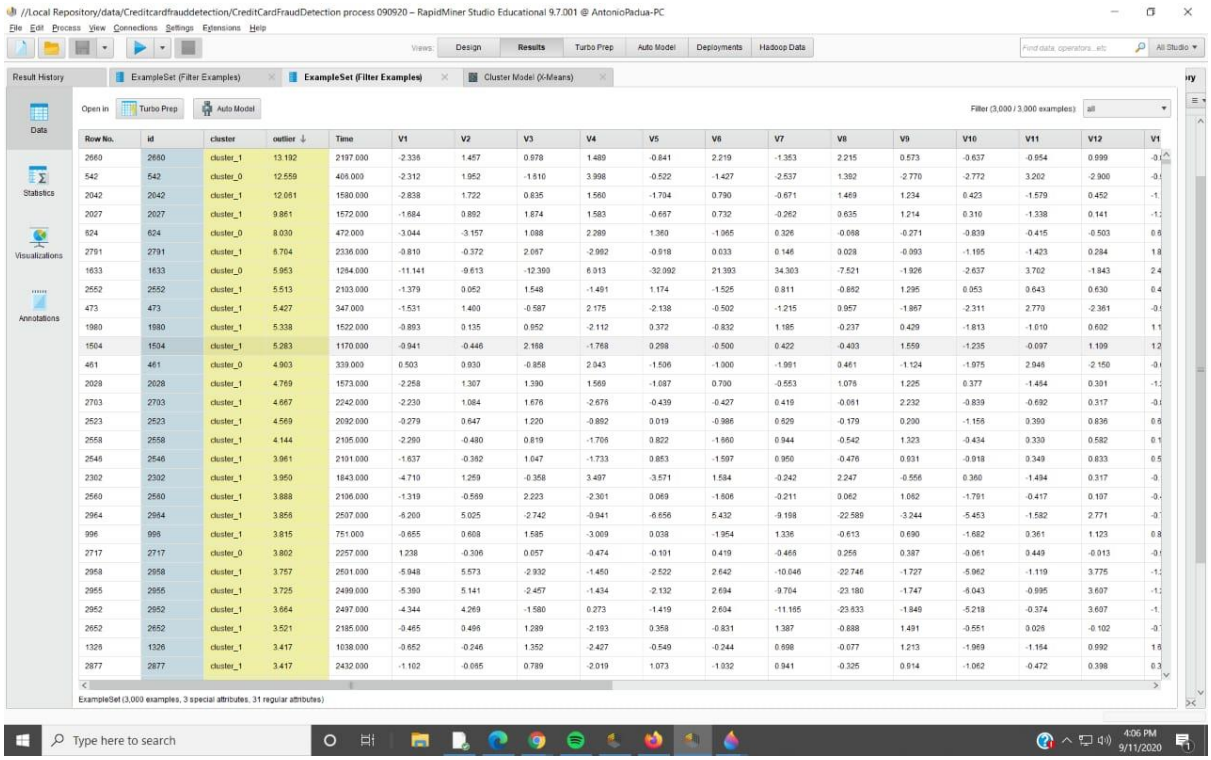
2.DATA PREPROCESSING :

Data preprocessing is a crucial step in credit card fraud detection using data science. It helps ensure that the data you feed into your machine learning model is clean, standardized, and ready for analysis. Here are the common data preprocessing steps for this types :

Handling Missing Values : - Identify and assess missing values in the dataset. Missing values can occur in various fields like merchant information or card details. - Decide on an appropriate strategy to handle missing values. Common approaches include imputation

(filling in missing values with statistical measures like mean or median) or removing rows or columns with too many missing values. - In this use the method to find the Missing Values are NaN values is

data.isnull().values.any() -> method used to handle missing values



Data Cleaning : - Check for and address data anomalies or errors, such as outliers that might indicate fraudulent transactions or incorrect entries. - Remove duplicates if they exist in the dataset. - Removing Missing values and NaN values .

Data Transformation : - Convert categorical variables into numerical format using techniques like one-hot encoding or label encoding, especially for features like merchant information or card details. - Normalize or scale numerical features to ensure they have similar scales. Common methods include Min-Max scaling or z-score standardization.

Handling Imbalanced Data : - Credit card fraud datasets often suffer from class imbalance, with a significantly higher number of non-fraudulent transactions compared to fraudulent ones. Address this issue using techniques like oversampling (creating more instances of the minority class), undersampling (reducing instances of the majority class), or using advanced methods like Synthetic Minority Over-sampling Technique (SMOTE).

Data Splitting : - Split the dataset into training and testing sets. The training set is used to train your machine learning model, while the testing set is used to evaluate its performance.

Handling Time-based Data : - If your dataset includes timestamps, consider converting them into useful features, **such** as time since the last transaction or time of day features

3.Feature Engineering : -

Create relevant features that may aid in fraud detection. For example, calculate transaction frequency, average transaction amount, or time-based features like the hour of the day or the day of the week.

➢ **Data Splitting** : - Split the dataset into training and testing sets. The training set is used to train your machine learning model, while the testing set is used to evaluate its performance.

➢ **Handling Time-based Data** : - If your dataset includes timestamps, consider converting them into useful features, such as time since the last transaction or time of day features.

➢ **Data Scaling :** - Normalize numerical features to ensure that they have similar scales. This can help improve the performance of some machine learning algorithms.

➢ **Data Privacy and Security** : - Implement data privacy measures to protect sensitive information, such as credit card numbers. This may involve tokenization, encryption, or anonymization of data

4.MODEL SELECTION:

Choose suitable machine learning algorithms

Example: logestic regression, random forest, gradient boosting

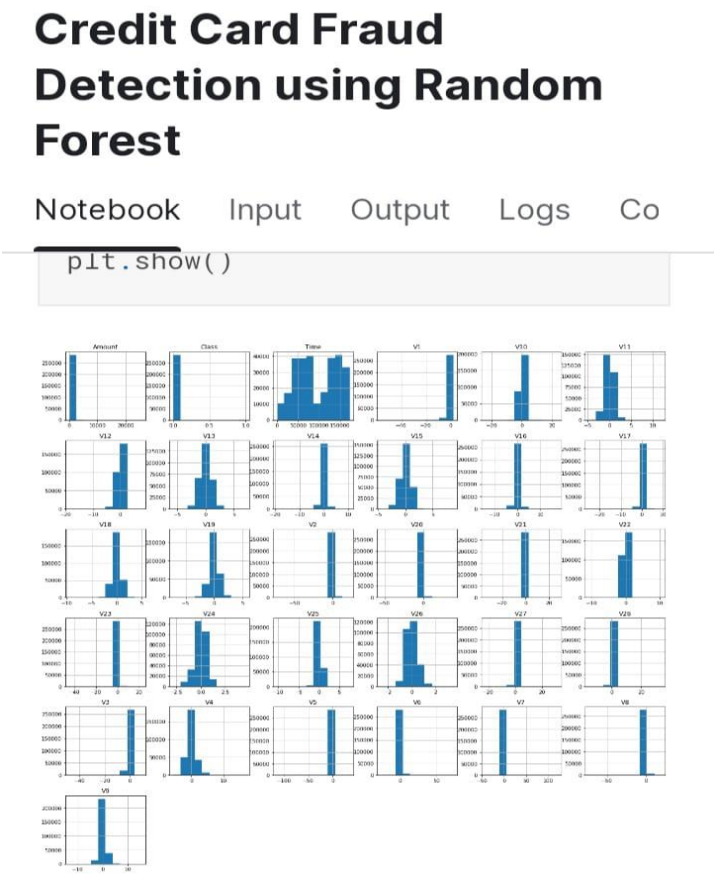
LOGESTIC REGRESSION:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. To ensure high accuracy detection, two main method are used to clean the data. The mean-based method Deals with missing values, and the clustering-based Method deals with outliers. Extensive experiments are conducted to train and test The proposed classifier using a standard database. Import the relevant libraries

```
Import numpy as np
Import matplotlib.pyplot as plt
Import pandas as pd
Import seaborn as sns
From sklearn.linear_model
import LogisticRegression
From sklearn.model_selection
 import _test_ssplit
```

RANDOM FOREST

Random forest is called a Random Forest because we use Random subsets of data and features and we end up building a Forest of decision trees (many trees). Random Forest is also a classic example of a bagging approach as we use different subsets of data in each model to make predictions.



Implementation of Random Forest Algorithm

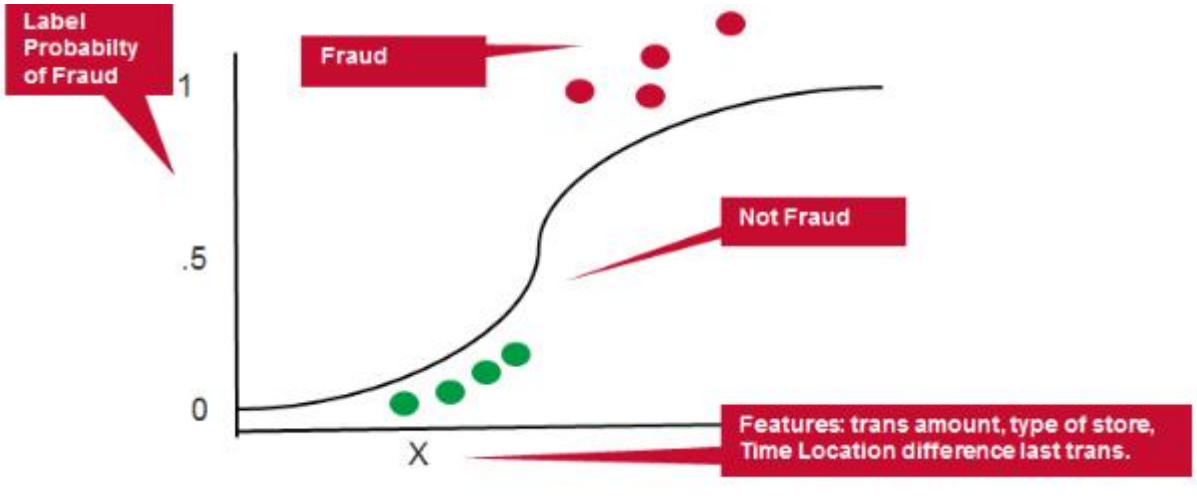
The procedure of the random forest algorithm execution is done there are several steps involved; first, there is a requirement togather the information and to store the information. The gathered information are in the form of data set in an excel sheet. In data exploration, the entire data set checked and removed the unnecessary data that is present. However, the data which is further treated using a random forest algorithm in two ways by using train data set and then using the test data set.

5.MODEL TRAINING

Data pre processing refers to the steps applied to make data more suitable for data mining. The steps used for Data Preprocessing usually fall into two categories: Selecting data objects and attributes for the analysis.Creating/changing the attributes.In this post I am going to walk through the implementation of Data Preprocessing methods using Python. It covers,

- Importing the libraries
- Importing the Dataset
- Handling of Missing Data
- Handling of Categorical Data
- Splitting the dataset into training and testing datasets
- Feature Scaling

For this Data Preprocessing script, I am going to use Anaconda Navigator and specifically Spyder to write the following code. If Spyder is not already installed when you open up Anaconda Navigator for the first time, then you can easily install it using the user interface.



6. EVALUATION

Confusion matrix, also known as an error matrix, is a performance measurement for assessing classification models. Below is an example of a two-class confusion matrix.

Prediction Within the confusion matrix, there are some terms that you need to know, which can then be used to calculate various metrics:

Negative: Outcome where the model correctly predicts the negative class.

True Positive: Outcome where the model correctly predicts the positive class.

True Positive (Type 1 Error): Outcome where the model incorrectly predicts the positive class.

False Negative (Type 2 Error): Outcome where the model incorrectly predicts the negative class.

Accuracy: equal to the fraction of predictions that a model got right.

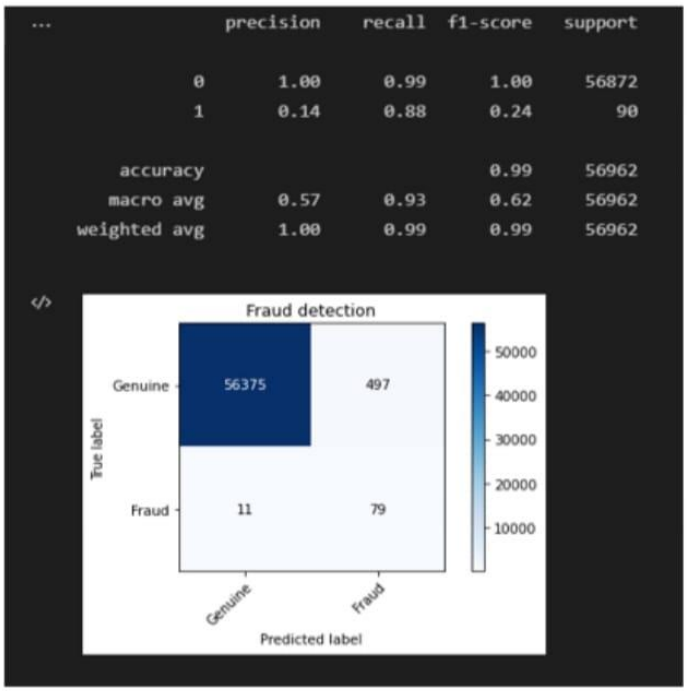


Figure: Confusion matrix for Naïve bayes

Conclusion:

Credit card fraud is the unauthorized use of a credit or debit card to make purchases. Credit card companies have an obligation to protect their customers' finances and they employ fraud detection models to identify unusual financial activity and freeze a user's credit card if transaction activity is out of the ordinary for a given individual. The penalty for mislabeling a fraud transaction as legitimate is having a user's money stolen, which the credit card company typically reimburses. On the other hand, the penalty for mislabeling a legitimate transaction as fraud is having the user frozen out of their finances and unable to make payments. There is a very fine tradeoff between these two consequences and we will discuss how to handle this when training a model where data science has been used widely so far.