Credit Card Fraud Detection

**Phase 3 Submission Document**

**Project: Credit Card Fraud Detection**

Introduction:

* **Protecting Consumers: Prevents unauthorized transactions andidentity theft.**
* **Protecting Businesses: Safeguards against revenue loss and reputational damage.**
* **Reducing Financial Losses: Minimizes financial impact onndividuals and organizations.**
* **Maintaining Trust in the Financial System: Ensures confidencein payment networks.**
* **Future Developments :**we need credit card fraud detection techniques**to protect the cardholders from false activity.** India is on its way to becoming a developed country. To achieve this, the Government of India (GoI) has launched several initiatives and one of these is Digital India Campaign.

Content of Project Phase 3 :

For analyzing data, we need some libraries. In this section, we are importing all the required libraries like pandas, NumPy, matplotlib, plotly, seaborn, and word cloud that are required for data analysis. Check the below code to import all the required libraries.

Data Source

A good data source for credit card fraud detection should be Accurate, Complete, Covering the geographic area of interest, Accessible.

# Data Collection and Preprocessing:

* Data collection : With Credit Card Fraud Detection, this project demonstrates the modelling of a data collection using **machine learning**. Modeling prior credit card transactions with data from those that turned out to be fraudulent is part of the Credit Card Fraud Detection Problem. The model is then used to determine whether or not a new transaction is fraudulent.

# Exploratory Data Analysis ( EDA ):

# This case study is focused to give you an idea of applying Exploratory Data Analysis (EDA) in a real business scenario. In this case study, apart from applying the various Exploratory Data Analysis (EDA) techniques, you will also develop a basic understanding of risk analytics and understand how data can be utilized in order to minimise the risk of losing money while lending to customers.

# Feature Engineering:

**Feature engineering** is a crucial step in credit card fraud detection. It involves selecting and transforming the most relevant features from the dataset to improve the performance of machine learning models. [In credit card fraud detection, feature engineering can help identify patterns and anomalies in transaction data that are indicative of fraudulent activity](https://medium.com/dataman-in-ai/how-to-create-good-features-in-fraud-detection-de6562f249ef).

Advanced Regression Techniques:

# There are several machine learning techniques that can be used for credit card fraud detection. One such technique is **logistic regression**. [In a study, researchers investigated the use of logistic regression to detect fraudulent credit card transactions in an imbalanced dataset where only a small fraction of transactions are fraudulent](https://ieeexplore.ieee.org/document/10112302/).

Partial Dependency Plots (PDP):

PDPs illustrate how a model's output changes with variations in a specific feature while holding other features constant. This allows users to see the relationship between individual features and the model's predictions.

SHAP (SHapley Additive exPlanations):

SHAP values provide a unified measure of feature importance. They can explain how each feature contributes to a specific prediction and can also give a global view of feature importance.

Decision Trees and Rule-Based Models:

Decision trees are inherently interpretable. By visualizing the tree structure, users can see how decisions are made at each node based on specific features. Rule-based models, like decision sets, provide a set of easy-to-understand rules for predictions.

Model-Agnostic Methods:

Techniques like SHAP, LIME, and anchor explanations work with any type of model. They provide a way to explain complex models, even if you don't know the internal workings of the model.

Visual Explanations:

Visualization tools can be used to present model interpretations in a more understandable and user-friendly manner. For instance, you can use heatmap visualizations to show how feature values influence predictions.

Global vs. Local Interpretability:

Models can provide global explanations that apply to the entire model or local explanations that focus on a specific prediction. Both types are important depending on the context and audience.

Regulatory Compliance:

In regulated industries like finance, ensuring that a model's behavior aligns with legal and compliance requirements is critical. Transparent and interpretable models are often favored in such scenarios.

Stakeholder Collaboration:

Collaboration between data scientists, domain experts, and business stakeholders can help bridge the gap between technical understanding and practical interpretation.

Documented Model Behavior:

Clearly documenting a model's behavior, including its input features and how it makes predictions, can aid in understanding and interpretation.

# Deployment and Prediction:

Deploy the chosen regression model to credit card fraud detection

Develop a user-friendly interface for users to input property features.

Program:

### Credit Card Fraud Detection

$ pip install sklearn==0.24.2 imbalanced-learn numpy pandas matplotlib seaborn

Let's import the necessary libraries:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from matplotlib import gridspec

Now we read the data and try to understand each feature's meaning. The Python module [pandas](https://pandas.pydata.org/) provide us with the functions to read data. In the next step, we will read the data from our directory where the data is saved, and then we look at the first and last five rows of the data using head(), and tail() methods:

dataset = pd.read\_csv("creditcard.csv")

dataset.head().append(dataset.tail())

╔════════╤══════════╤════════════╤═══════════╤═══════════╤═══════════╤═══════════╤═══════════╤═══════════╤═══════════╤═══════════╤═════╤═══════════╤═══════════╤═══════════╤═══════════╤═══════════╤═══════════╤═══════════╤═══════════╤════════╤═══════╗

║ │ Time │ V1 │ V2 │ V3 │ V4 │ V5 │ V6 │ V7 │ V8 │ V9 │ ... │ V21 │ V22 │ V23 │ V24 │ V25 │ V26 │ V27 │ V28 │ Amount │ Class ║

╠════════╪══════════╪════════════╪═══════════╪═══════════╪═══════════╪═══════════╪═══════════╪═══════════╪═══════════╪═══════════╪═════╪═══════════╪═══════════╪═══════════╪═══════════╪═══════════╪═══════════╪═══════════╪═══════════╪════════╪═══════╣

║ 0 │ 0.0 │ -1.359807 │ -0.072781 │ 2.536347 │ 1.378155 │ -0.338321 │ 0.462388 │ 0.239599 │ 0.098698 │ 0.363787 │ ... │ -0.018307 │ 0.277838 │ -0.110474 │ 0.066928 │ 0.128539 │ -0.189115 │ 0.133558 │ -0.021053 │ 149.62 │ 0 ║

╟────────┼──────────┼────────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼─────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼────────┼───────╢

║ 1 │ 0.0 │ 1.191857 │ 0.266151 │ 0.166480 │ 0.448154 │ 0.060018 │ -0.082361 │ -0.078803 │ 0.085102 │ -0.255425 │ ... │ -0.225775 │ -0.638672 │ 0.101288 │ -0.339846 │ 0.167170 │ 0.125895 │ -0.008983 │ 0.014724 │ 2.69 │ 0 ║

╟────────┼──────────┼────────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼─────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼────────┼───────╢

║ 2 │ 1.0 │ -1.358354 │ -1.340163 │ 1.773209 │ 0.379780 │ -0.503198 │ 1.800499 │ 0.791461 │ 0.247676 │ -1.514654 │ ... │ 0.247998 │ 0.771679 │ 0.909412 │ -0.689281 │ -0.327642 │ -0.139097 │ -0.055353 │ -0.059752 │ 378.66 │ 0 ║

╟────────┼──────────┼────────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼─────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼────────┼───────╢

║ 3 │ 1.0 │ -0.966272 │ -0.185226 │ 1.792993 │ -0.863291 │ -0.010309 │ 1.247203 │ 0.237609 │ 0.377436 │ -1.387024 │ ... │ -0.108300 │ 0.005274 │ -0.190321 │ -1.175575 │ 0.647376 │ -0.221929 │ 0.062723 │ 0.061458 │ 123.50 │ 0 ║

╟────────┼──────────┼────────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼─────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼───────────┼────────┼───────╢

║

The Time is measured in seconds since the first transaction in the data collection. As a result, we may infer that this dataset contains all transactions recorded during two days. The features were prepared using PCA, so the physical interpretation of individual features does not make sense. 'Time' and 'Amount' are the only features that are not transformed to PCA. 'Class' is the response variable, and it has a value of 1 if there is fraud and 0 otherwise.

## Data Exploration and Visualization

Now we try to find out the relative proportion of valid and fraudulent credit card transactions:

print("Fraudulent Cases: " + str(len(dataset[dataset["Class"] == 1])))

print("Valid Transactions: " + str(len(dataset[dataset["Class"] == 0])))

print("Proportion of Fraudulent Cases: " + str(len(dataset[dataset["Class"] == 1])/ dataset.shape[0]))

data\_p = dataset.copy()

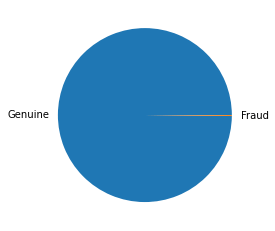
data\_p[" "] = np.where(data\_p["Class"] == 1 , "Fraud", "Genuine")

data\_p[" "].value\_counts().plot(kind="pie")

Fraudulent Cases: 492

Valid Transactions: 284315

Proportion of Fraudulent Cases: 0.001727485630620034



There is an imbalance in the data, with only 0.17% of the total cases being fraudulent.

Now we look at the distribution of the two named features in the dataset. For Time, it is clear that there was a particular duration in the day when most of the transactions took place:

f, axes = plt.subplots(1, 2, figsize=(18,4), sharex = True)

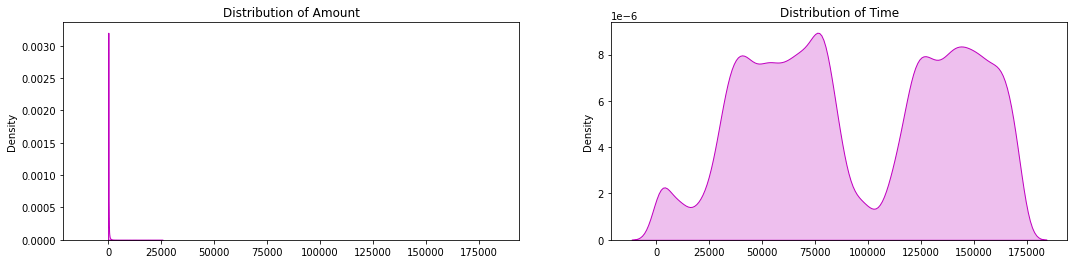
amount\_value = dataset['Amount'].values # values

time\_value = dataset['Time'].values # values

sns.distplot(amount\_value, hist=False, color="m", kde\_kws={"shade": True}, ax=axes[0]).set\_title('Distribution of Amount')

sns.distplot(time\_value, hist=False, color="m", kde\_kws={"shade": True}, ax=axes[1]).set\_title('Distribution of Time')

plt.show()



Let us check if there is any difference between valid transactions and fraudulent transactions:

print("Average Amount in a Fraudulent Transaction: " + str(dataset[dataset["Class"] == 1]["Amount"].mean()))

print("Average Amount in a Valid Transaction: " + str(dataset[dataset["Class"] == 0]["Amount"].mean()))

Average Amount in a Fraudulent Transaction: 122.21132113821133

Average Amount in a Valid Transaction: 88.29102242225574

As we can notice from this, the average money transaction for the fraudulent ones is more. It makes this problem crucial to deal with. Now let us try to understand the distribution of values in each feature. Let's start with the Amount:

print("Summary of the feature - Amount" + "\n-------------------------------")

print(dataset["Amount"].describe())

Summary of the feature - Amount

-------------------------------

count 284807.000000

mean 88.349619

std 250.120109

min 0.000000

25% 5.600000

50% 22.000000

75% 77.165000

max 25691.160000

Name: Amount, dtype: float64

The rest of the features don't have any physical interpretation and will be seen through histograms. Here the values are subgrouped according to class (valid or fraud):

data\_plot = dataset.copy()

amount = data\_plot['Amount']

data\_plot.drop(labels=['Amount'], axis=1, inplace = True)

data\_plot.insert(0, 'Amount', amount)

columns = data\_plot.iloc[:,0:30].columns

plt.figure(figsize=(12,30\*4))

grids = gridspec.GridSpec(30, 1)

for grid, index in enumerate(data\_plot[columns]):

ax = plt.subplot(grids[grid])

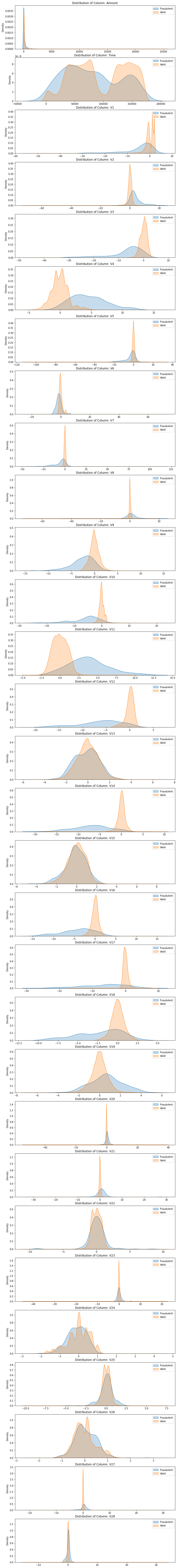
sns.distplot(data\_plot[index][data\_plot.Class == 1], hist=False, kde\_kws={"shade": True}, bins=50)

sns.distplot(data\_plot[index][data\_plot.Class == 0], hist=False, kde\_kws={"shade": True}, bins=50)

ax.set\_xlabel("")

ax.set\_title("Distribution of Column: " + str(index))

plt.show()



Importance of loading and processing dataset:

Loading and processing datasets is a critical step in many data-driven tasks, especially in the fields of machine learning, data analysis, and artificial intelligence. The importance of this step can be summarized as follows:

1. **Data Quality Assurance:** Loading and processing datasets allow you to examine and clean the data. This is crucial because real-world data is often messy, containing missing values, outliers, and errors. Data preprocessing helps ensure the quality and reliability of the data used for analysis or modeling.
2. **Feature Engineering:** Processing data involves selecting relevant features and transforming them into a suitable format. Proper feature selection and engineering can significantly impact the performance of machine learning models. This step is critical for enhancing the predictive power of models.
3. **Normalization and Standardization:** Data preprocessing often includes techniques like normalization and standardization, which can make the data more suitable for various machine learning algorithms. This step helps ensure that the scales of different features are comparable.
4. **Dimensionality Reduction:** High-dimensional data can be challenging to work with. Dimensionality reduction techniques, like Principal Component Analysis (PCA) or feature selection, help reduce the number of features while retaining the most critical information. This can lead to faster training times and better model performance.
5. **Data Integration:** In many cases, data comes from various sources and in different formats. Loading and processing data enable you to integrate diverse datasets and prepare them for analysis or modeling. This is essential for holistic insights and decision-making.
6. **Data Understanding:** As you load and process a dataset, you gain a better understanding of its characteristics, such as the distribution of values, statistical properties, and relationships between variables. This understanding is crucial for choosing appropriate algorithms and making informed decisions.
7. **Data Visualization:** Preprocessing often includes data visualization, which helps you identify patterns, trends, and correlations in the data. Visualizing the data can provide insights that guide subsequent analysis and modeling tasks.
8. **Computational Efficiency:** Efficient data processing can significantly reduce the time and resources required for training machine learning models. Properly processed data can lead to faster convergence and reduced computational costs.
9. **Model Performance:** The quality of data preprocessing can directly impact the performance of machine learning models. Well-preprocessed data can lead to more accurate, robust, and interpretable models.
10. **Ethical Considerations:** Data preprocessing also plays a role in addressing ethical and privacy concerns. It includes steps to anonymize or de-identify sensitive information and ensure that models are not biased or discriminatory.

In summary, loading and processing datasets are foundational steps in the data analysis and machine learning pipeline. The quality and effectiveness of these steps can have a profound impact on the success of data-driven projects, making them crucial for extracting valuable insights, making informed decisions, and building accurate predictive models.

## Data Preparation

Since the features are created using PCA, feature selection is unnecessary as many features are tiny. Let's see if there are any missing values in the dataset:

dataset.isnull().shape[0]

print("Non-missing values: " + str(dataset.isnull().shape[0]))

print("Missing values: " + str(dataset.shape[0] - dataset.isnull().shape[0]))

Non-missing values: 284807

Missing values: 0

As there are no missing data, we turn to standardization. We standardize only Time and Amount using [RobustScaler](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html" \t "_blank" \o "RobustScaler):

from sklearn.preprocessing import RobustScaler

scaler = RobustScaler().fit(dataset[["Time", "Amount"]])

dataset[["Time", "Amount"]] = scaler.transform(dataset[["Time", "Amount"]])

dataset.head().append(dataset.tail())

10 rows × 31 columns

Next, let's divide the data into features and targets. We also make the train-test split of the data:

y = dataset["Class"] # target

X = dataset.iloc[:,0:30]

from sklearn.model\_selection import train\_test\_split

X\_train, X\_test, y\_train, y\_test = train\_test\_split(

X, y, test\_size = 0.2, random\_state = 42)

X\_train.shape, X\_test.shape, y\_train.shape, y\_test.shape

Output:

((227845, 30), (56962, 30), (227845,), (56962,))

Let's import all the necessary libraries for the tutorial:

from sklearn.model\_selection import StratifiedKFold

from sklearn.model\_selection import GridSearchCV, cross\_val\_score, RandomizedSearchCV

kf = StratifiedKFold(n\_splits=5, random\_state = None, shuffle = False)

from imblearn.pipeline import make\_pipeline ## Create a Pipeline using the provided estimators .

from imblearn.under\_sampling import NearMiss ## perform Under-sampling based on NearMiss methods.

from imblearn.over\_sampling import SMOTE ## PerformOver-sampling class that uses SMOTE.

from sklearn.metrics import roc\_curve, roc\_auc\_score, accuracy\_score, recall\_score, precision\_score, f1\_score

from sklearn.linear\_model import LogisticRegression

from sklearn.svm import SVC

from sklearn.neighbors import KNeighborsClassifier

from sklearn.tree import DecisionTreeClassifier

from sklearn.ensemble import RandomForestClassifier

# Model Interpretability:

Credit card fraud is a growing problem in the financial industry, with the potential to cause significant

financial losses to both customers and financial institutions. As a result, there has been a significant

amount of research in recent years on developing effective fraud detection systems. These systems rely on

a combination of statistical techniques, machine learning algorithms, and deep learning models to identify

fraudulent transactions.One of the most commonly used approaches for credit card fraud detection is

rule-based systems. These systems use predefined rules to identify transactions that are deemed suspicious.

However, rule-based systems have limitations, as they are only as good as the rules that have been

predefined, and they may not be able to detect new types of fraud. To overcome these limitations, machine

learning algorithms and statistical techniques have been applied to credit card fraud detection. These

techniques are based on analysing transaction-related data, such as the transaction amount, location, and

time, as well as other relevant factors, such as the customer’s transaction history and account details. In

recent years, deep learning models, such as convolutional neural networks (CNNs) and recurrent neural

networks (RNNs), have also been applied to credit card fraud detection. These models have shown

promising results in identifying fraudulent transactions by learning patterns in the data and improving the

accuracy of fraud detection. Overall, credit card fraud detection is a critical area of research in the financial

industry, with significant potential for improving fraud detection rates and reducing financial losse

Credit card fraud is a growing problem in the financial industry, with the potential to cause significant

financial losses to both customers and financial institutions. As a result, there has been a significant

amount of research in recent years on developing effective fraud detection systems. These systems rely on

a combination of statistical techniques, machine learning algorithms, and deep learning models to identify

fraudulent transactions.One of the most commonly used approaches for credit card fraud detection is

rule-based systems. These systems use predefined rules to identify transactions that are deemed suspicious.

However, rule-based systems have limitations, as they are only as good as the rules that have been

predefined, and they may not be able to detect new types of fraud. To overcome these limitations, machine

learning algorithms and statistical techniques have been applied to credit card fraud detection. These

techniques are based on analysing transaction-related data, such as the transaction amount, location, and

time, as well as other relevant factors, such as the customer’s transaction history and account details. In

recent years, deep learning models, such as convolutional neural networks (CNNs) and recurrent neural

networks (RNNs), have also been applied to credit card fraud detection. These models have shown

promising results in identifying fraudulent transactions by learning patterns in the data and improving the

accuracy of fraud detection. Overall, credit card fraud detection is a critical area of research in the financial

industry, with significant potential for improving fraud detection rates and reducing financial losse

Credit Card Fraud can be defined as a case where a person uses someone else’s credit card for personal reasons while the owner and the card-issuing authorities are unaware of the fact that the card is being used. Due to the rise and acceleration of E-Commerce, there has been a tremendous use of credit cards for online shopping which led to High amount of frauds related to credit cards. In the era of digitalization, the need to identify credit card frauds is necessary. Fraud detection involves monitoring and analyzing the behaviour of various users to estimate detect or avoid undesirable behaviour. To identify credit card fraud detection effectively, we need to understand the various technologies, algorithms and types involved in detecting credit card frauds. The algorithm can differentiate transactions which are fraudulent or not. Find fraud, they need to passed dataset and knowledge of the fraudulent transaction. They analyze the dataset and classify all transactions. Fraud detection involves monitoring the activities of populations of users to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting. Machine learning algorithms are employed to analyses all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent. The investigators provide feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time.

# Project Conclusion:

In the Phase 2 conclusion: Recap Key Points: Stress the importance of fraud detection form financial security.Emphasize the Ongoing Need for Vigilance: Highlight that fraud prevention is an evolving process

Future Work: we need credit card fraud detection techniques**to protect the cardholders from false activity.** India is on its way to becoming a developed country. To achieve this, the Government of India (GoI) has launched several initiatives and one of these is Digital India Campaign. .