Credit Card Fraud Detection

Practical Motivation

- Credit card fraud attempts increased by 46% year-on-year
- Substantial financial losses for both institutions and individuals

$43 billion by 2060

Practical Motivation

Reports of unauthorised online banking and card transactions in Singapore jump 460% in 2020

- **2,782** cases of credit card fraud were reported in 2020 alone, resulting in a **collective loss of over SGD 16 million**

Problem Statement

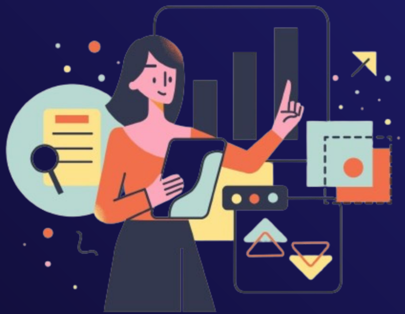
To **enhance** credit card fraud detection by developing **reliable** and **accurate** fraudulent transaction detection mechanisms using **Classification and Machine Learning algorithms** to minimise financial losses for financial institutions and individuals.

Defining Credit Card Frauds

- Unauthorised transactions made using someone else's credit card or credit card details
- Fraudsters use a variety of methods to obtain credit card information, which include:

→ Database hacking
→ Phishing scams
→ Skimming devices (duplicating of information located on the magnetic strip of the card)
→ Stealing of physical credit cards
→ Fraudulent telemarketing

Sample Collection



Dataset with a mix of categorical and numerical data types:
https://www.kaggle.com/datasets/kelvinkelue/credit-card-fraud-prediction

Variables Examined

Preparation of Data

1. Checking for Null Values

```
Unnamed: 0             0
trans_date_trans_time  0
cc_num                 0
merchant               0
category               0
amt                    0
first                  0
last                   0
gender                 0
street                 0
city                   0
state                  0
zip                    0
lat                    0
long                   0
city_pop               0
job                    0
dob                    0
trans_num              0
unix_time              0
merch_lat              0
merch_long             0
is_fraud               0
dtype: int64
```

2. Checking for Duplicates

Duplicate Rows :

Visualize

	Unnamed: 0 int64	trans_date_trans_t...	cc_num float64	merchant object	category object	amt float64	first object

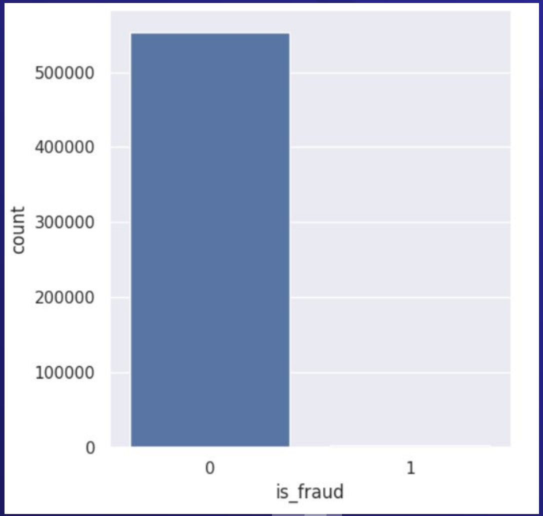
0 rows, showing 10 per page

Page 0 of 0

Determining Number of Fraudulent Transactions

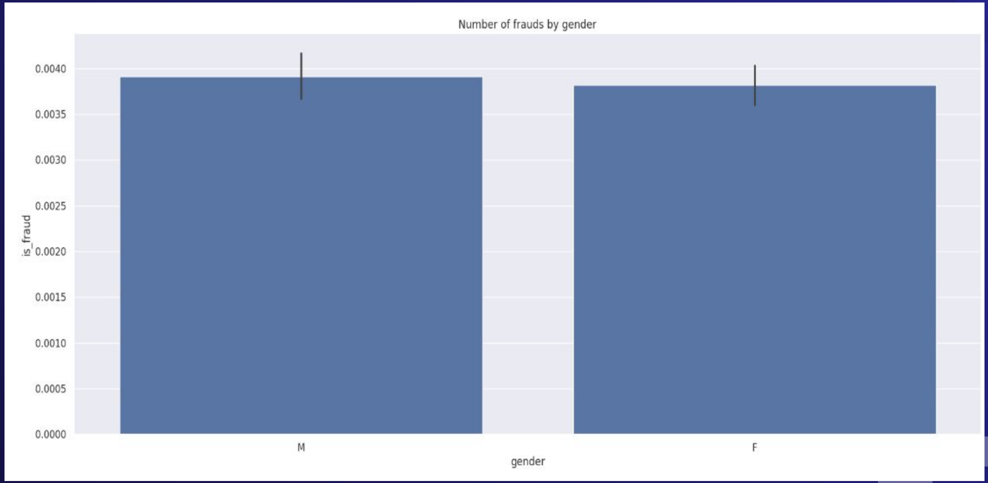
There are **2145** fraudulent transactions out of a total of 553,574 transactions

```
0    553574
1      2145
Name: is_fraud, dtype: int64
```



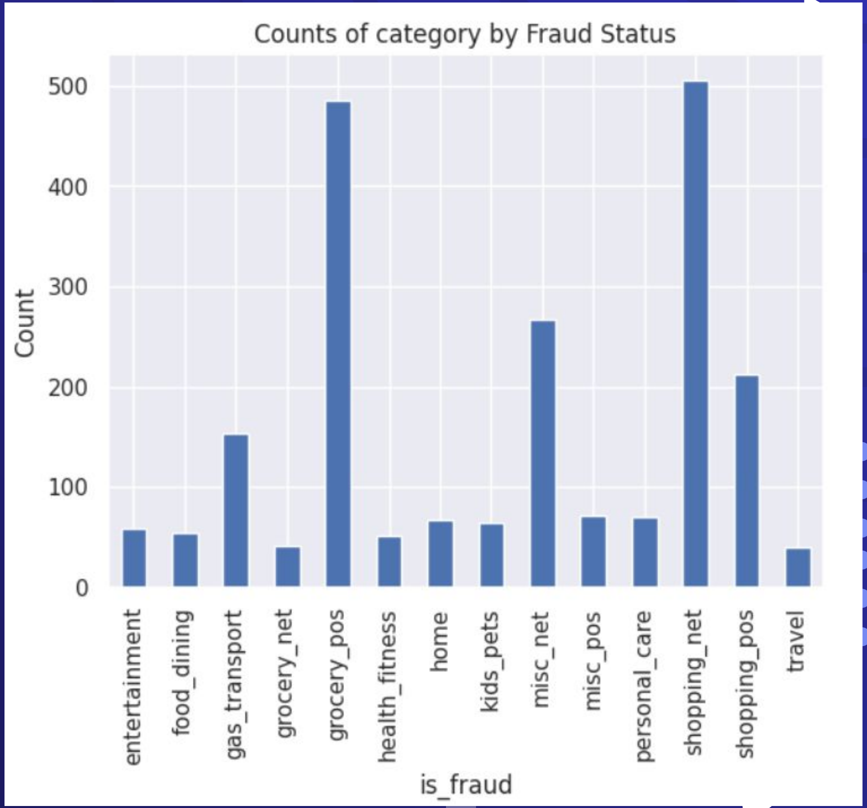
Involvement of Different Genders in Fraudulent Transactions

It can be seen that **Males** take up a higher proportion of individuals involved in fraudulent transactions



Number of Frauds By Category

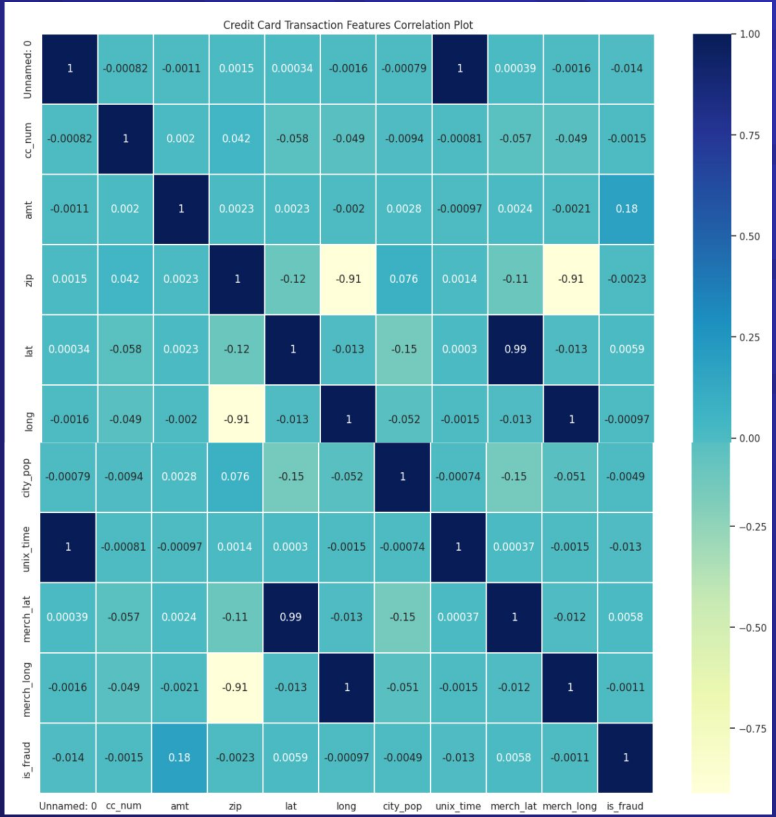
- Shopping_net makes up the **greatest proportion** of fraudulent transaction
- This highlights the need for individuals to be extra cautious when shopping online



Correlation Between the Different Variables and Fraudulent Transactions

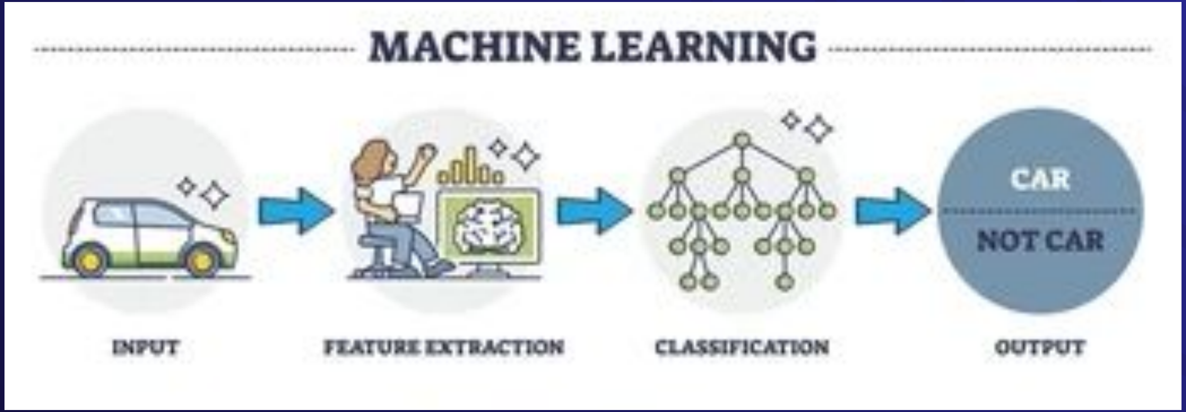
- The correlations between the variables from our raw data and fraudulent transactions are **relatively weak**

- There is a need to for **feature engineering** to develop better variables to increase the accuracy and reliability of our model



What is Feature Engineering?

- The process of **selecting, manipulating and transforming raw data** into features that can be used in **supervised learning**.



What We Did For Feature Engineering

01.

Created a **new variable**, Age, from data on date-of-birth and date of transaction

02.

Extracted features relevant to credit card fraud

03.

Did **undersampling** to resolve the imbalanced data

04.

Categorised the time of purchasing into time categories with one-hour interval

05.

Performed one-hot encoding on gender, category of purchases (category) and time (time_category)

06.

Cleaned the dataset to obtain a **reliable gender ratio** for fraudulent and non-fraudulent transactions

How We Did It

1. Converting date-of-birth to age

- Date-of-birth of individuals, dob, is used to generate the age of individuals (`age`)
- Allows us to see whether age has an influence on susceptibility to fraudulent transactions

```
data['age']= data['trans_date_trans_time'].dt.year- data['dob'].dt.year
```

How We Did It

2. Data Extraction

- Unnecessary variables are dropped out
- Only included relevant variables (ie. the category of objects, gender, age, amount of transaction [amt], transaction time and date [trans_date_trans_time], and whether the transaction is fraud [is_fraud])

```
data=data[["category","gender","is_fraud","age","amt","trans_date_trans_time"]]
```

How We Did It

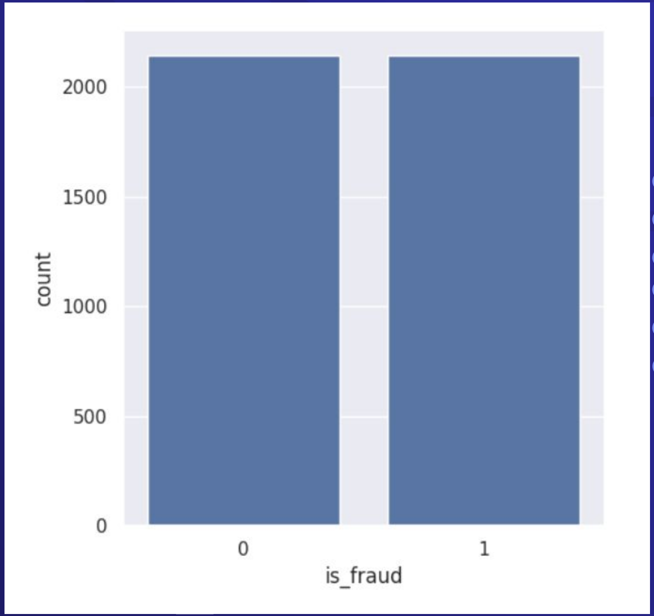
3. Data Balancing Using Undersampling

- Data is **highly imbalanced** ⟶ need to sample out non-fraudulent transactions so that the value is the same as the fraudulent transactions

- **Undersampling** is used by eliminating examples belonging to the majority class
    - Undersampling prevents data from being overfitted, unlike SMOTE and Oversampling

The data for is_fraud will now be **balanced**.

```
0     2145
1     2145
Name: is_fraud, dtype: int64
```



How We Did It

4. Categorising Transaction Timings into One-Hour Intervals

- Allows us to analyse during which period fraudulent transactions occur the most often

time_category
7am-8am
8pm-9pm
2pm-3pm
7am-8am
1am-2am

How We Did It

5. One-hot encoding

- Transforms categorical variables into a format that can be **understood and processed** by algorithms
- **Avoids** introducing **implicit ordering** in categorical variables
- Done for the variables **`gender`, `Category`** and **`Time_Category`**

```
data_encoded = pd.get_dummies(cleaned_data, columns=['category', 'gender','time_category'], dtype=int, drop_first=True
```

category_food_din...	category_gas_tran...	category_grocery_...	category_grocery_...	category_health_fi...	category_home int...	category_kids_pets i.	categ
0	0	0	0	0	0	0	
0	0	0	0	1	0	0	
0	0	0	0	0	0	0	
0	0	0	0	0	0	0	
0	0	0	0	0	0	0	

How We Did It

6. Gender Ratio for Cleaned Data

- After cleaning the dataset, we can see that there is a **greater proportion of females** involved in performing fraudulent transactions.

```
0     2308
1     1982
Name: gender_M, dtype: int64
```

Exploratory Data Analysis

- Categorical variables → perform chi-squared test and hypothesis testing to examine significance of variables + heatmap and barplots

- For numerical variables → directly plot the heatmap to get the correlation coefficient to examine the relationship between the variables and fraudulent transactions.

Correlation of Category of Products With Fraudulent Transactions

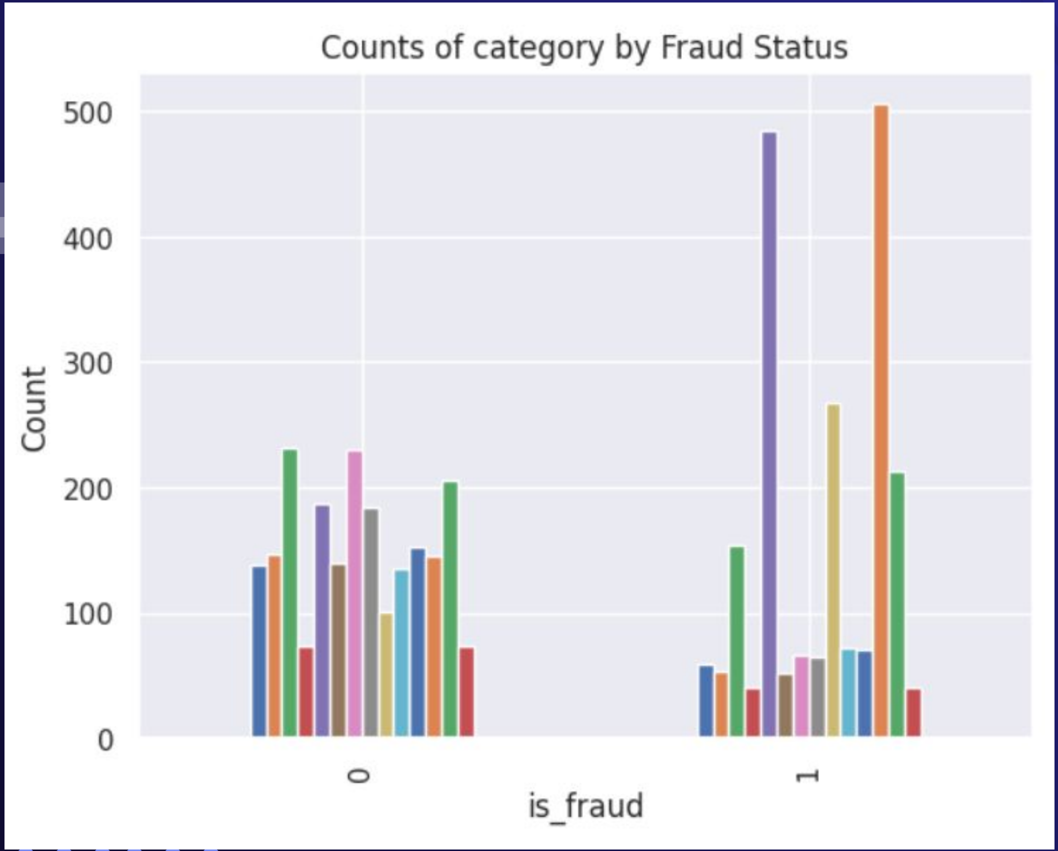
- **Chi-squared test and Hypothesis Testing**

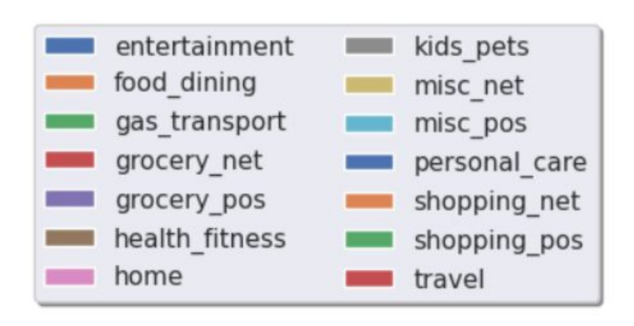
```
Chi-square statistic: 751.5528866806904

p-value: 3.24820512315656e-152
```

→ **significant association** between the two variables

Correlation of Category of Products With Fraudulent Transactions





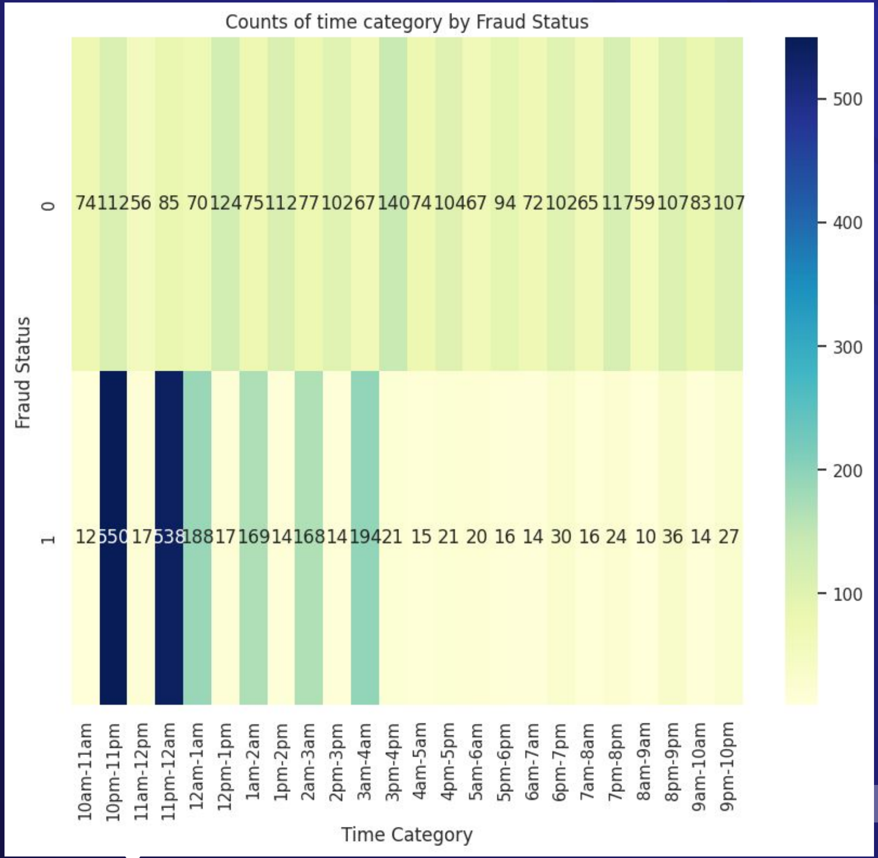
Correlation between Time of Transaction in a Day and Fraudulent Transactions

- **Chi-squared test and Hypothesis Testing**

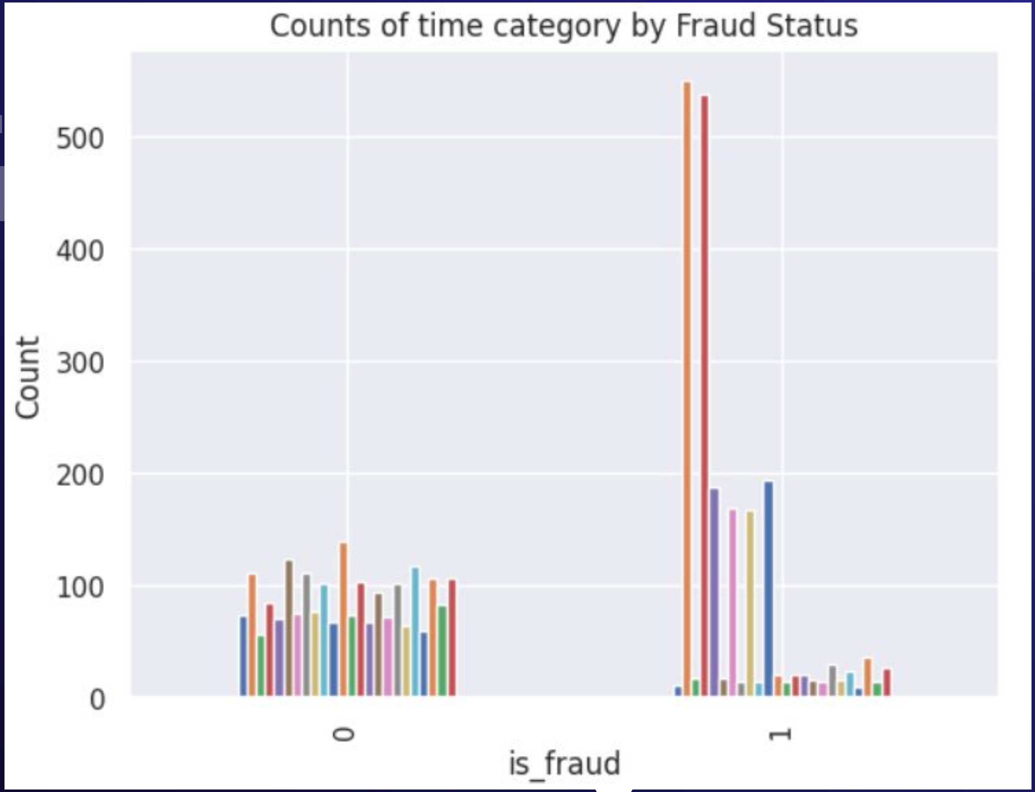
```
Chi-square statistic: 1693.8194640855882
p-value: 0.0
```

→ **significant association** between the two variables

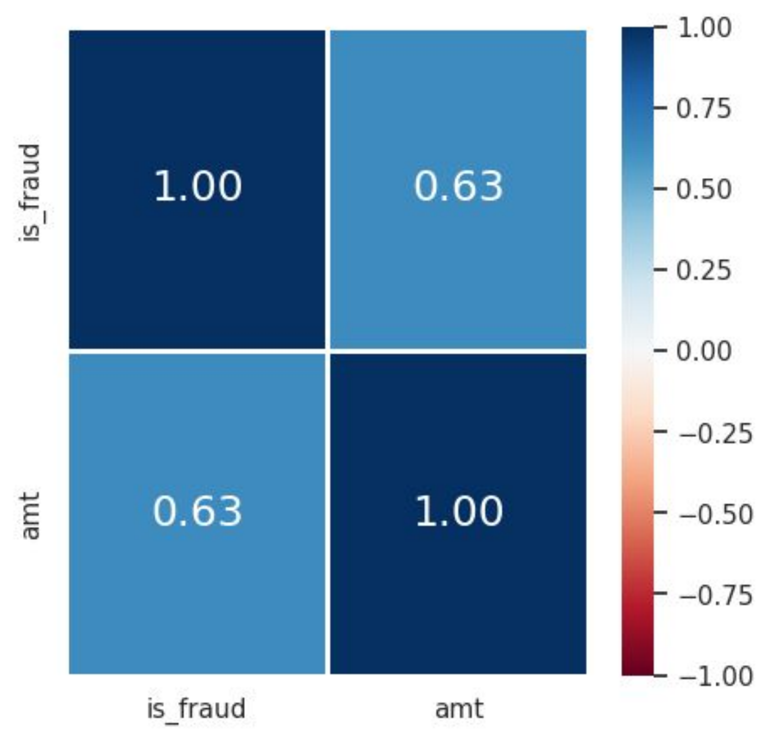
Correlation between Time of Transaction in a Day and Fraudulent Transactions

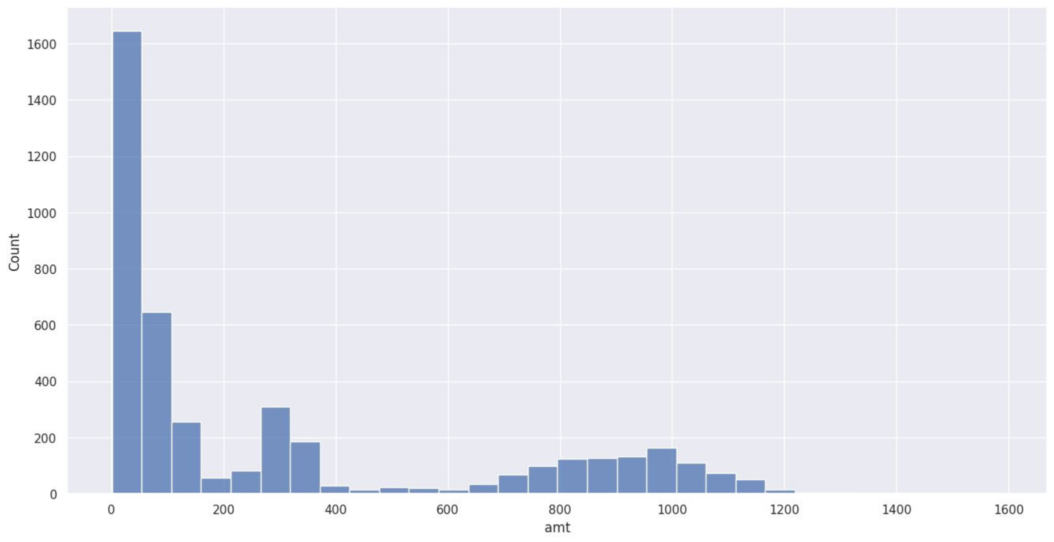


Correlation between Time of Transaction in a Day and Fraudulent Transactions

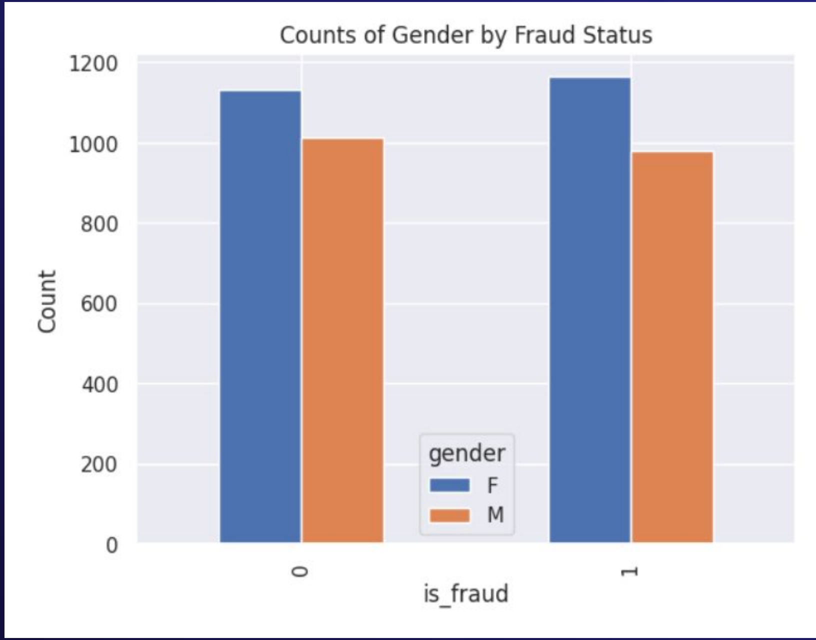


Correlation Between Transaction Amount and Fraudulent Transactions





Correlation Between Gender and Fraudulent Transactions

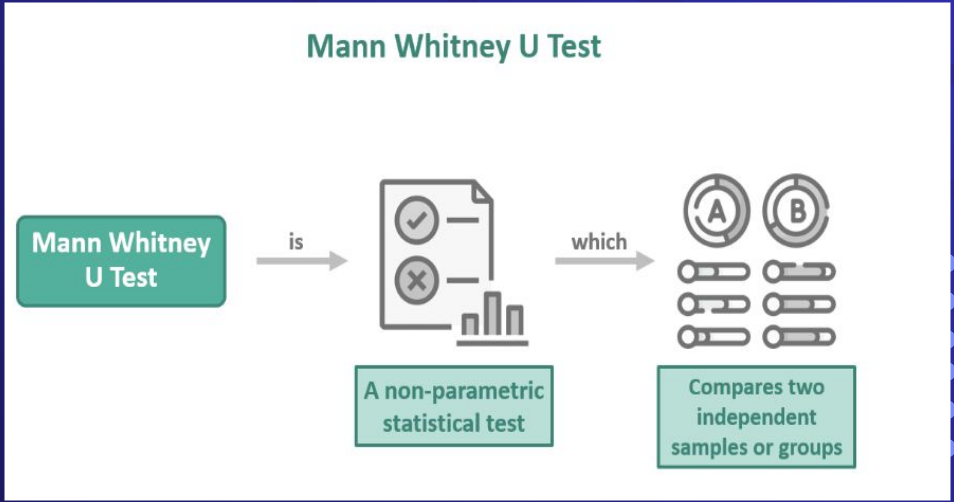



```
Counts of Gender by Fraud Status:
gender        F     M
is_fraud
0          1132  1013
1          1164   981
```

→ **Females** now take up a greater proportion of fraudulent transactions.

Correlation Between Age and Fraudulent Transactions

- Instead of performing the usual statistical tests such as Chi-squared tests and Hypothesis testing, we performed **Mann-Whitney U Test**
    - Mann-Whitney U Test does not rely on assumptions about the distribution of the data → can be used when the data are **not normally distributed** or when the **variances are unequal**
    - Suitable for **comparing** the distributions of a **continuous variable** (age) **between two independent groups** (fraud and non-fraud)



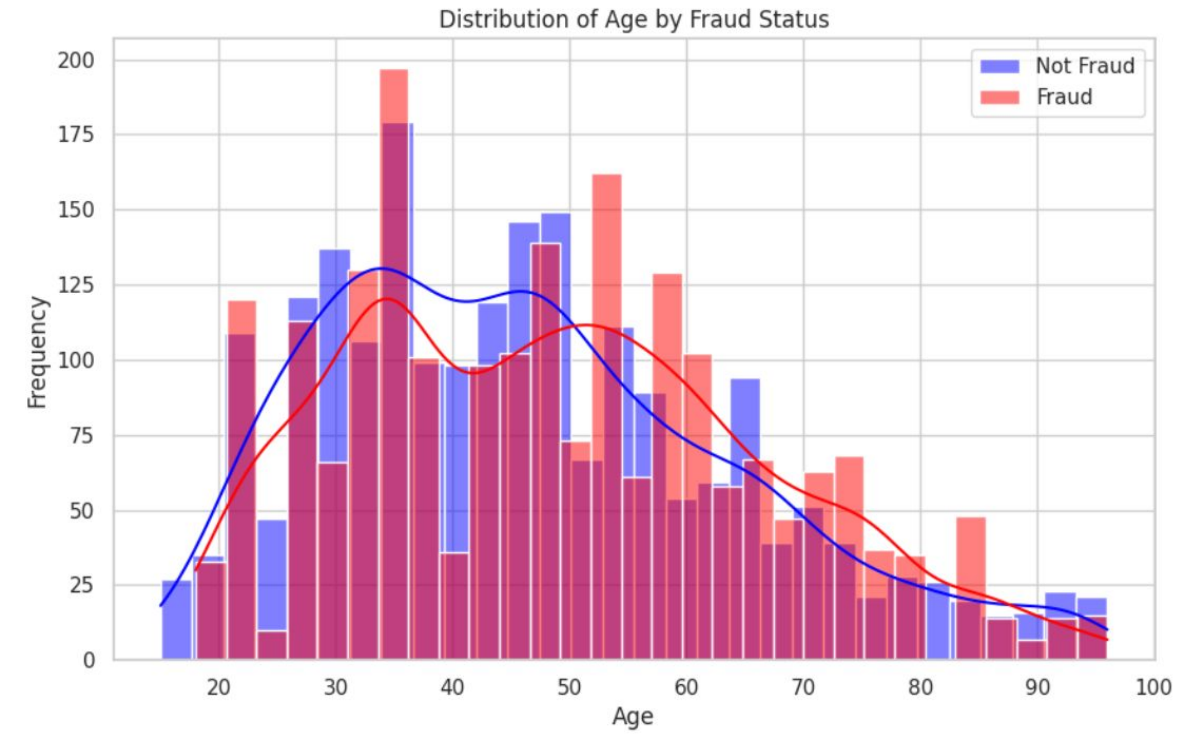
Correlation Between Age and Fraudulent Transactions

```
Mann-Whitney U Test U-statistic: 2493018.5
p-value: 2.0656195581016276e-06
```

- A higher U-statistic suggests a **significant distinction in age distributions**
- The p-value also indicates strong evidence against the null hypothesis and we can conclude that there is a **significant difference in age distributions** between fraudulent and non-fraudulent transactions

  → the ages of individuals involved in fraudulent transactions tend to **differ** significantly from those involved in non-fraudulent transactions

Correlation Between Age and Fraudulent Transactions



Feature Selection

Variables used for building our fraud detection model are

- Category of Products (`category`)
- Time of Transaction (`time_category`)
- Transaction Amount (`amt`)
- Gender (`gender`)
- Age (`age`)

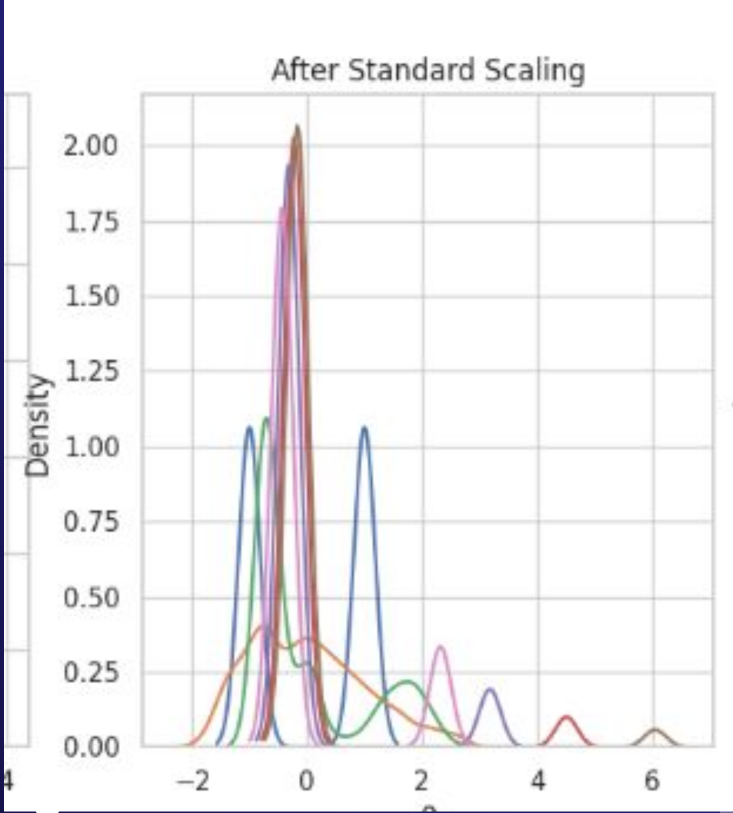




Feature Scaling

- The dataset contains some variables with large range of values as compared to other variables such as transaction amount
- In order to prevent such variables dominating other variables, we perform **Standard Scaling** on the dataset.
    - Other scaling methods, such as MinMax Scaling and Robust Scaling, have their disadvantages
    - Standard Scaling preserves the data distribution and is more compatible with machine learning algorithms

Feature Scaling



Machine Learning Models

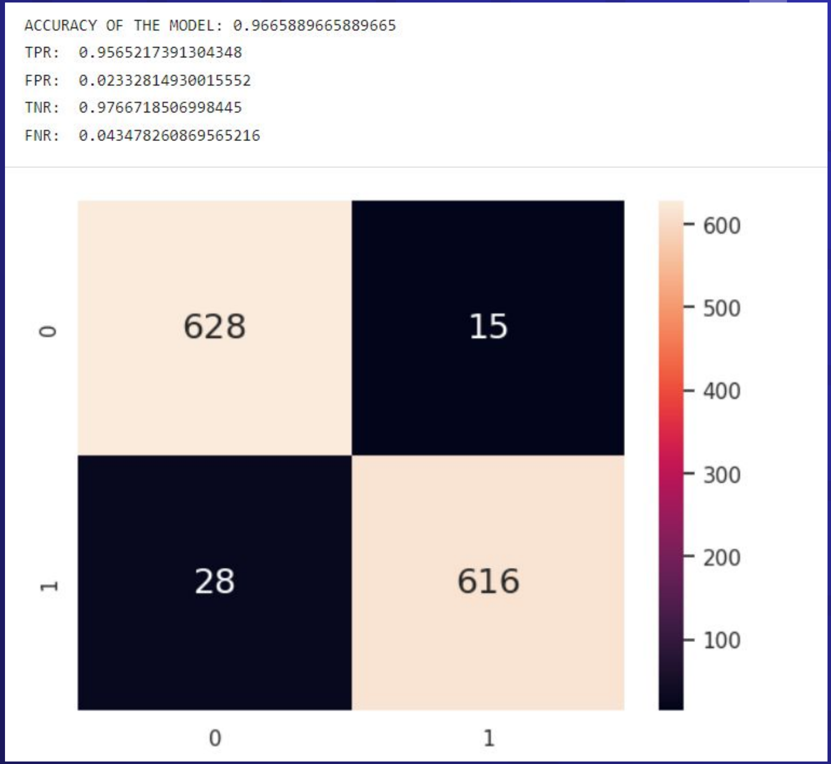
The models that we have chosen are

- Random Forest Classifier
- Logistic Regression
- Multi-layer Perceptron
- K-Nearest Neighbours.

Random Forest Classifier

- Combines predictions of multiple decision trees for better generalisation performance compared to individual trees

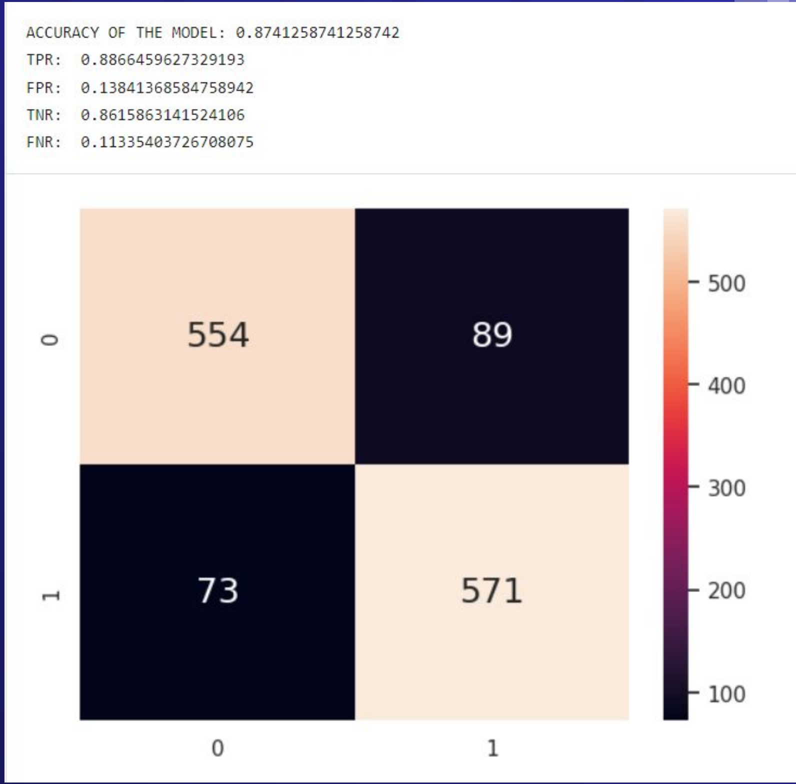
- Provides a measure of feature importance, indicating the **significance of each feature** to the model's predictive performance



Logistic Regression

- Provides insight into the impact of each feature on the likelihood of fraud crucial for prevention strategies

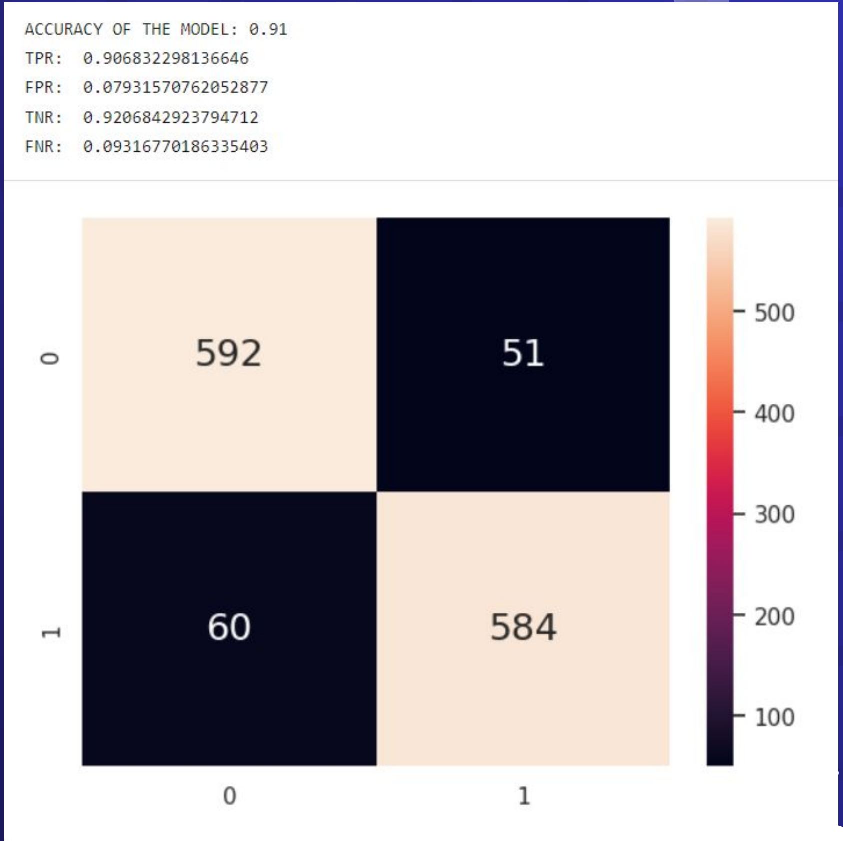
- Provides **probabilistic outputs**, allowing for the estimation of the likelihood that a transaction is fraudulent to set decision thresholds for automated fraud detection systems



Multi-Layer Perceptron

- Highly adaptable and can adjust their internal representations in response to changes in the data distribution, allowing it to **maintain its high performance**

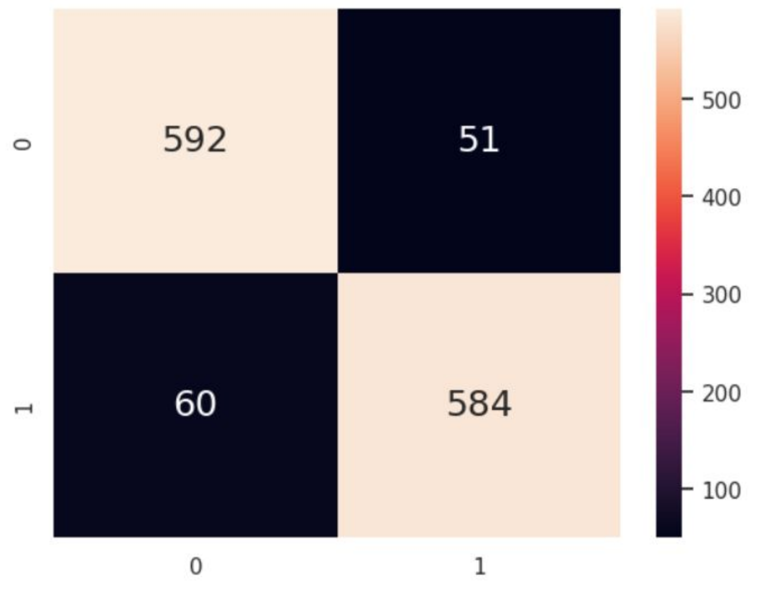
- By learning in an unsupervised or semi-supervised manner, MLP-based anomaly detection models can **identify previously unseen or novel fraud schemes**



K-Nearest Neighbours

- **Flags** potential fraudulent transactions for further investigation
- Capable of capturing **non-linear relationships** between input features and the target variable, making it suitable for modeling such complex data.
- It is a non-parametric algorithm that makes no assumptions about the underlying data distribution

```
ACCURACY OF THE MODEL: 0.91
TPR:  0.906832298136646
FPR:  0.07931570762052877
TNR:  0.9206842923794712
FNR:  0.09316770186335403
```



Evaluation of Models

1. Classification Accuracy

- Random Forest Classifier Model has the highest accuracy of 0.97 (Logistic Regression: 0.87, Multi-layer Perceptron:0.91, K-Nearest Neighbour: 0.85)

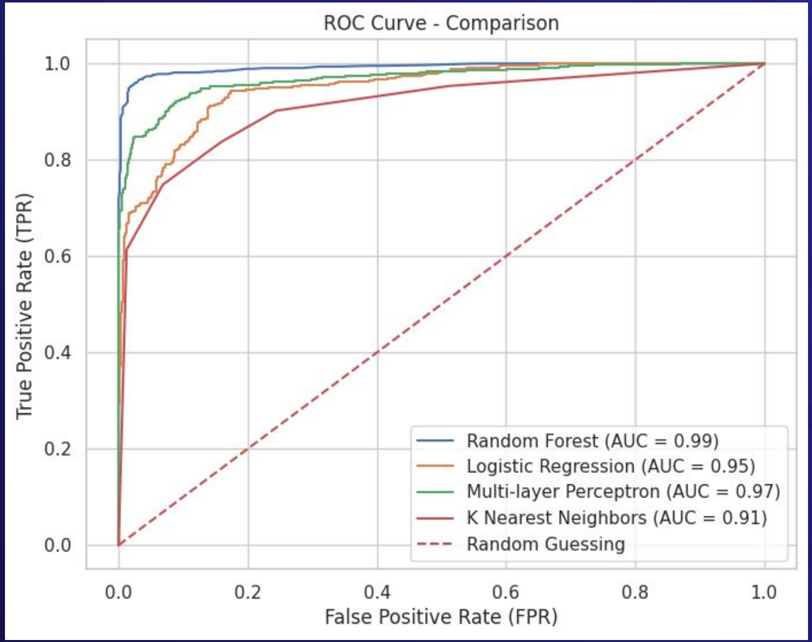
2. Confusion matrix

- Random Forest Classifier Model has the highest True Positive Rate of 0.95 (Logistic Regression: 0.89, Multi-Layer Perceptron: 0.91, K-Nearest Neighbour: 0.84)
- Random Forest Classifier Model has the lowest False Positive Rate of 0.02 (Logistic Regression: 0.14, Multi-Layer Perceptron: 0.08, K-Nearest Neighbour: 0.16)

Evaluation of Models

3. AUROC Evaluation

- A higher AUC score indicates better discrimination ability of the model in distinguishing between positive and negative cases
- AUC value of Random Forest Classifier is the highest at 0.99



Final Analysis

Random Forest Classifier model is the **best** and **most suitable model** in credit card fraud prediction as it fared the best out of the four models that we have analysed

Outcome of Project

1. **Improved Fraud Detection Accuracy**
   - By implementing robust and relevant machine learning models and techniques, the system can effectively identify suspicious activities and minimise false positives and false negatives

2. **Reduced Financial Losses**
   - Effective fraud detection prevents unauthorised transactions, reducing the risk of financial losses

Outcome of Project

3. Operational Efficiency

- Using machine learning algorithms can reduce manual effort needed for monitoring transactions
- More manpower can be redirected to improving other facets of financial transactions

4. Data Insights

- Insights gained from data can be incorporated into fraud prevention strategies

Insights

1. **Imbalanced Data Handling**
   - Learnt about how significant class imbalances can affect data analysis
   - The implementation of different strategies such as undersampling, oversampling and SMOTE

2. **Methods to Test for Statistical Significance**
   - Explored the different tests available, such as Chi-Squared Test and Mann-Whitney U Test, and the various conditions that need to be satisfied

3. **Methods for Model Evaluation**
   - New evaluation methods, such as AUROC Evaluation were also learnt

Recommendations

1. **Implementing greater surveillance and scrutiny during high-risk periods and on high-risk platforms**
   - Allows for early detection and prevention of unauthorised transactions

2. **Collaborate with Industry Partners**
   - Foster collaboration and information sharing with industry partners, payment networks, and law enforcement agencies to stay updated on emerging fraud trends and tactics.

3. **Raise Awareness**
   - Educate the public on how these frauds occur and how individuals are susceptible to such frauds by providing them with relevant statistics

Conclusion

Further **research and innovation** is needed to **stay ahead** of emerging and sophisticated credit card frauds and ensure the **integrity and security** of financial systems