Why Financial Fraud Detection?

Financial fraud is a growing concern in the banking industry, as fraudsters continue to find new ways to exploit systems. Financial fraud detection is vital for the banking industry to protect customers and financial assets from increasingly sophisticated fraud schemes. It helps minimize monetary losses, maintain trust, and comply with regulatory requirements.

By leveraging advanced technologies like machine learning, banks can proactively identify and prevent fraud, ensuring a secure and stable financial ecosystem while reducing operational costs. Effective fraud detection safeguards both customers and institutions, reinforcing trust in the banking system.

Project Goal

The goal of this project is to uncover key insights into the characteristics of fraudulent transactions and the common behaviors of fraudsters. Additionally, the project aims to apply various machine learning algorithms to identify the most effective model for helping banks evaluate and detect fraudulent transactions, ensuring customer protection and safeguarding financial assets.

Dataset

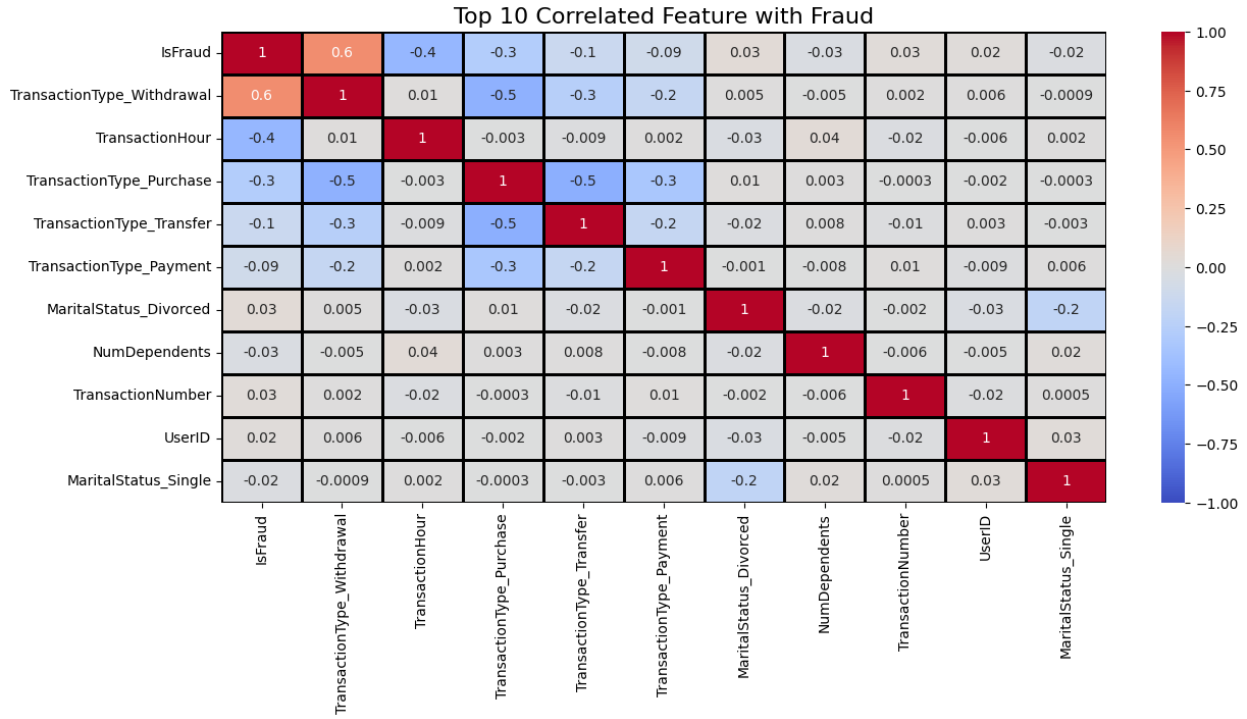
The dataset is provided by a major Melbourne-based bank in 2023, includes 10966 anonymized transaction records, customer demographics, and socio-economic indicators.

TABLE OF CONTENTS

1. Exploratory Data Analysis (EDA)

Dataset includes 10,966 transactions during 2023, with 24 unique features.

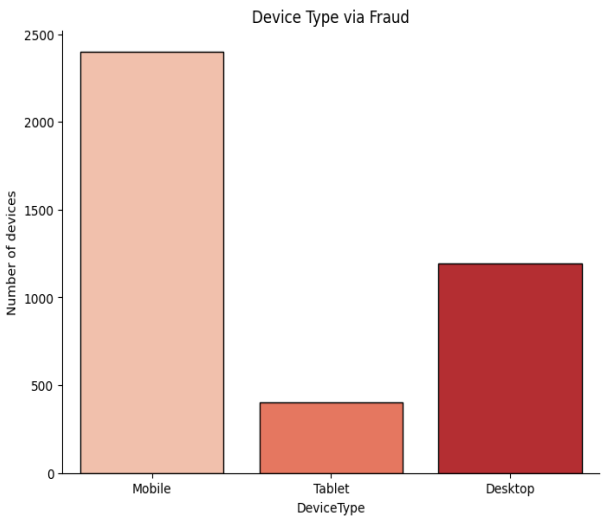
After conducting several data cleaning, I check for the top 10 correlated variables with fraud transactions.

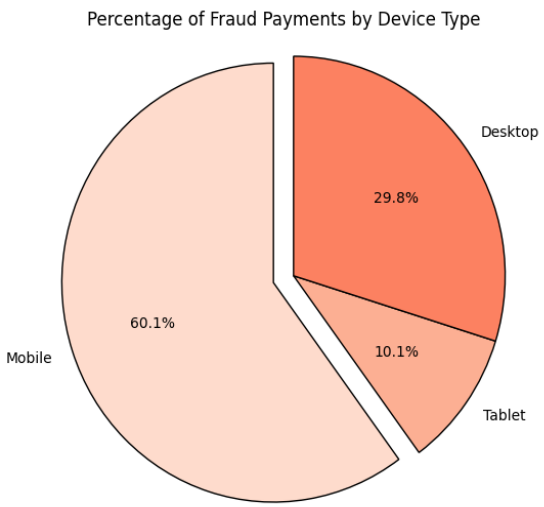


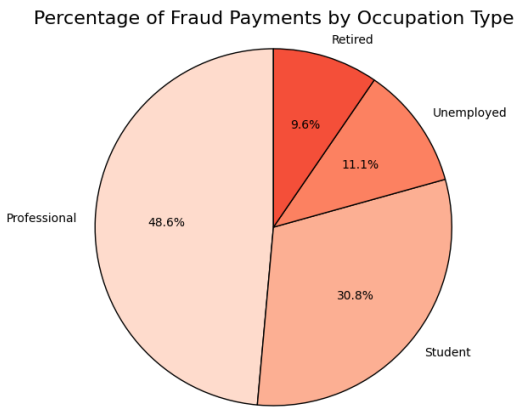
The correlation analysis reveals that **TransactionType_Withdrawal** exhibits the strongest positive correlation with fraudulent payments (**0.6**), followed by **TransactionHour** (**0.4**) and **TransactionType_Purchase** (**0.3**). This indicates that fraudulent transactions are most frequently associated with withdrawals and purchases. Furthermore, the time of the transaction plays a significant role in identifying potential fraud, suggesting that certain hours are more prone to fraudulent activities. These findings underscore the importance of focusing on transaction type and timing when designing fraud detection systems.

The analysis shows that **Marital Status** has a low overall correlation with fraudulent transactions, suggesting that marital status has minimal influence on fraud detection. However, among the marital status categories, individuals who are divorced (0.03) exhibit a slightly higher correlation with fraudulent transactions compared to single person (0.02). This indicates that divorced individuals might have a marginally higher tendency to be involved in fraudulent activities than single individuals. While this finding is interesting, it should be interpreted cautiously and in conjunction with other variables to avoid overgeneralization.

Visualizations for fraudsters insights and behaviors

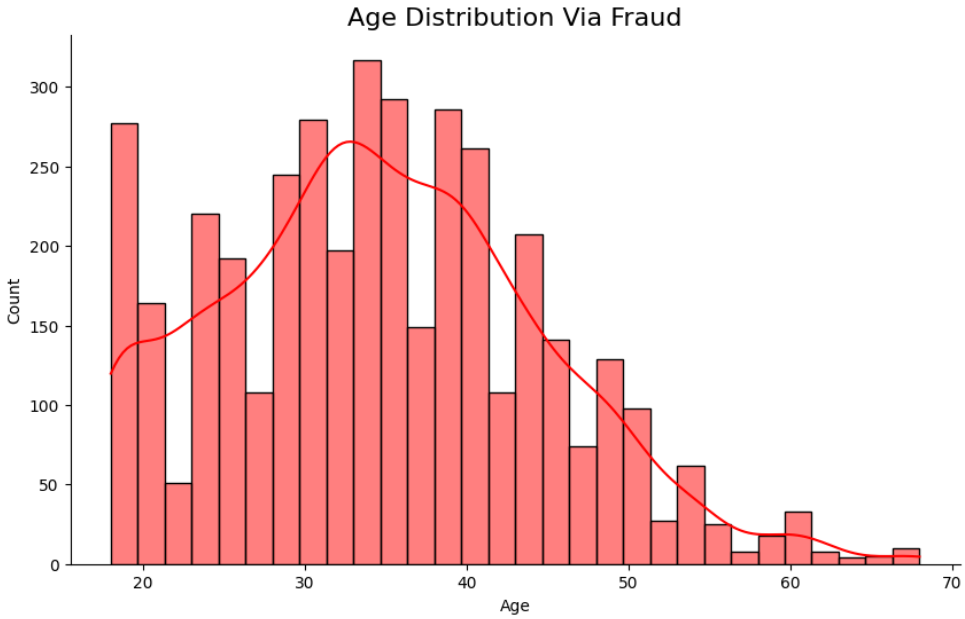






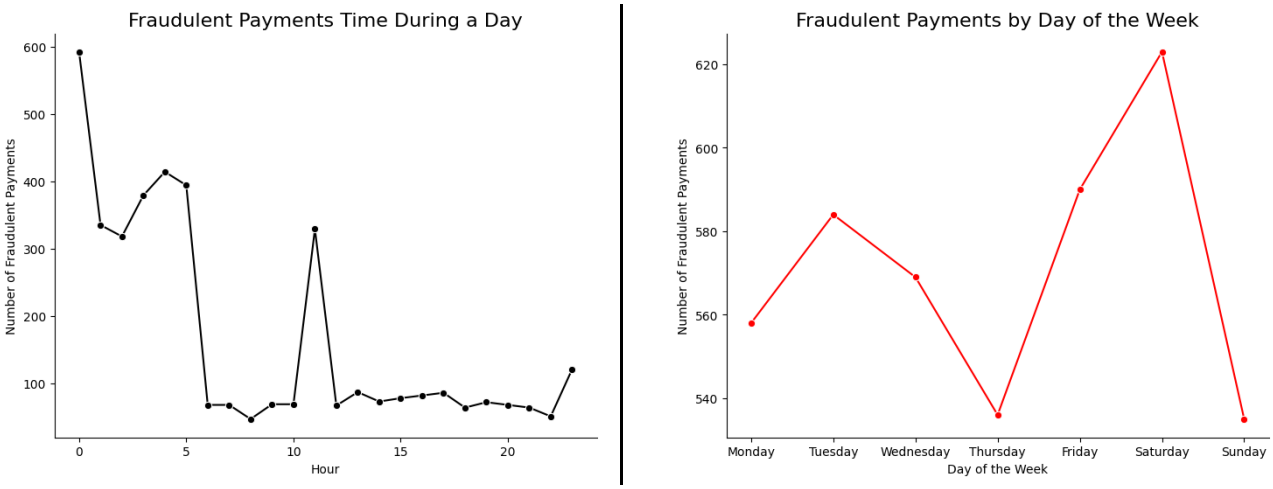
The data indicates that the majority of fraudulent transactions were conducted via **Mobile devices (60.1%)**, followed by **Desktop (29.8%)** and **Tablet (10.1%)**, highlighting the prevalence of mobile platforms in fraudulent activities.

Interestingly, a significant portion of fraudsters are categorized as **professionals (48.6%)**, suggesting that individuals with office jobs are disproportionately represented among those engaging in fraudulent behavior. **Students account for 30.8%**, while **unemployed and retired individuals** make up a relatively small proportion. These insights suggest that fraud prevention measures should prioritize mobile platforms and consider demographic factors like employment status to enhance detection and deterrence.



The age distribution indicates that most fraudsters are **primarily in the 30-40 age group**.

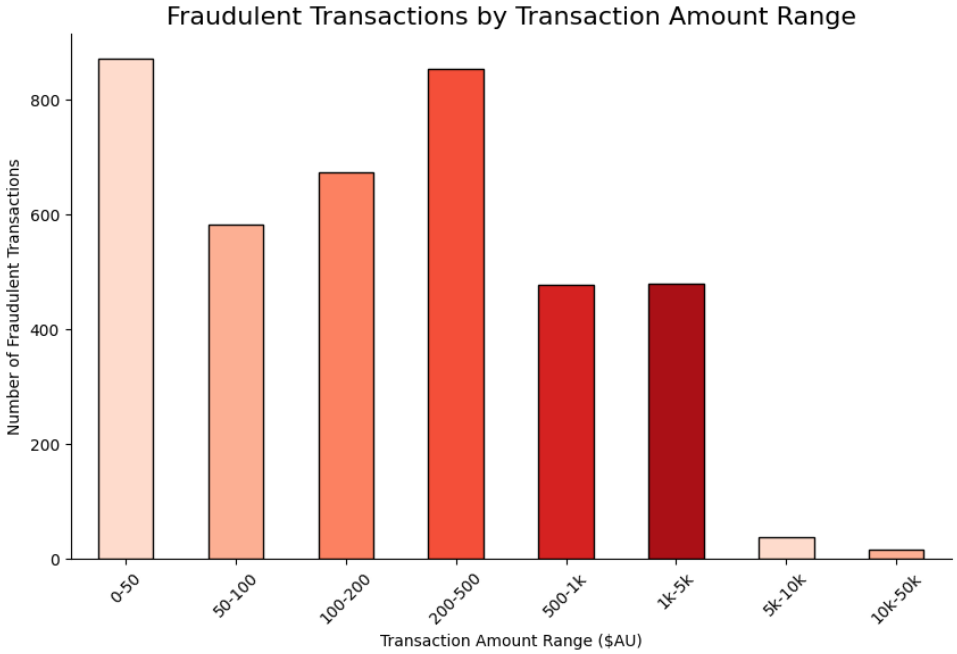
However, a significant number of younger individuals, around 18-20 years old, are also involved in fraudulent transactions. This trend could be attributed to students seeking ways to earn money but lacking the experience or qualifications for full-time jobs, leading them to resort to easier, unethical, means of income generation.



Most fraudulent payments occur late at night, particularly between 12:00 AM and 3:00 AM, and are significantly more frequent from Friday to the end of Saturday.

This pattern suggests that fraudsters strategically choose times when individuals are likely asleep and less vigilant about their transactions. Additionally, weekends provide an opportunity for

fraudsters as people tend to pay less attention to their bank accounts, and many companies are closed, making it harder to detect and respond to fraudulent activities promptly.



Most fraudulent transaction amounts fall within the range of $0 to $500, with only a small proportion involving larger amounts exceeding $5,000.

This indicates that fraudsters often target smaller, less noticeable amounts to reduce the likelihood of detection.

Most fraudsters have an annual income ranging from $50k to $100k, followed by those earning $5k to $50k and $100k to $150k.

This suggests that the majority of fraudsters fall within the average income bracket, indicating that fraudulent activities are not limited to any specific income level but are more common among those with moderate earnings.

In general, fraudsters and fraudulent transactions tend to share the following characteristics:

Most transactions are conducted through mobile

Fraudulent activities are concentrated late at night (12:00 AM to 3:00 AM) and during weekends (Friday to Saturday).

Most of fraudulent transactions are of the withdrawal type.

Fraud transaction amounts are typically flexible, ranging from $0 to $500, with only a small proportion exceeding $5,000.

Fraudsters are primarily aged between 30 and 40 years old.

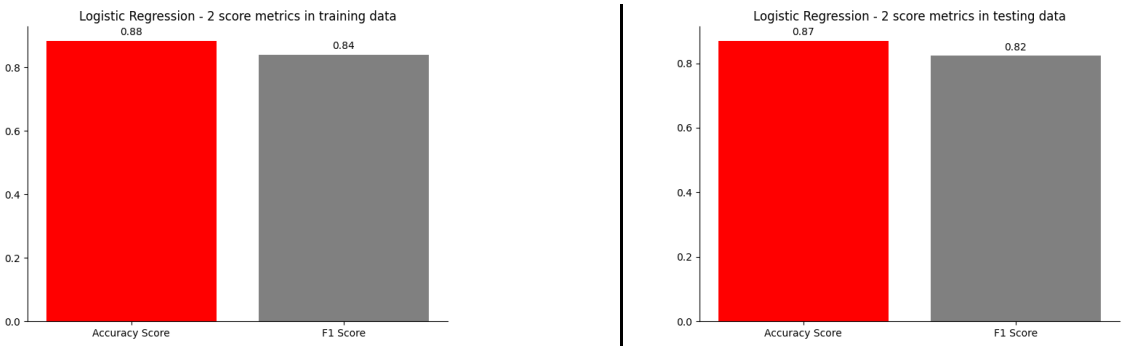
2. Apply Machine Learning Technique

To assist the bank in predicting whether a transaction is fraudulent, we applied several machine learning techniques. These models were evaluated and compared using accuracy and F1 scores to determine the most effective approach for fraud detection. This ensures a balanced assessment of the models' precision and recall, helping the bank choose the best model for reliable fraud prevention.

Accuracy score: is a metric used to evaluate a classification model's performance. It measures the proportion of correctly predicted instances (both true positives and true negatives) out of the total number of predictions.

F1 score: is the harmonic mean of precision and recall. It balances the trade-off between these two metrics and is especially useful for imbalanced datasets where detecting the minority class (e.g., fraud) is critical.

2.1 Logistic Regression



- The model performs well with both accuracy and F1 scores being consistently high on training and testing data.
- The minimal drop in scores from training to testing indicates that the model generalizes well and is not overfitting.
- The balance between accuracy and F1 suggests that this model can effectively predict fraudulent transactions while maintaining a good balance between precision and recall.

2.2 K-Nearest Neighbor (KNN)

To determine the optimal value of K for the best prediction model, I employed two tuning methods: **Recursive Feature Elimination (RFE)** and **GridSearch Cross-Validation**. These methods were used to identify the most meaningful features and the optimal number of K for the model.

For a detailed explanation of the tuning process and results, please refer to the accompanying Jupyter Notebook. It includes step-by-step implementation, evaluation metrics, and insights into feature selection and hyperparameter optimization

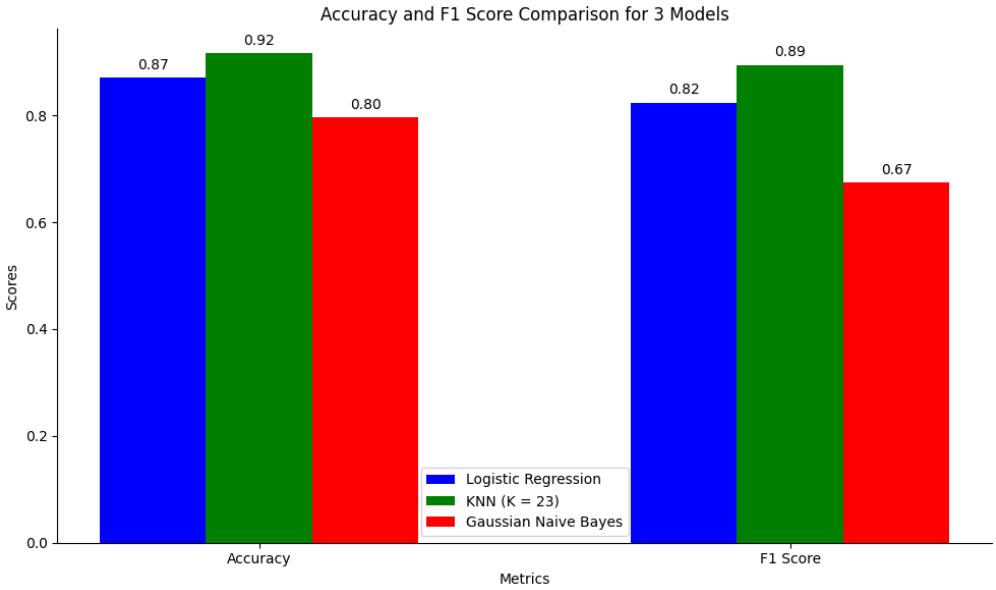
- The results demonstrate that K=23 is an excellent choice for the model, as it achieves both high accuracy and F1 scores.

- The consistency across training and testing datasets indicates that the model is reliable and robust, capable of handling unseen data effectively.

2.3 Gaussian Naïve Bays

- The **Gaussian Naive Bayes** model performs fairly well but lags behind other models like Logistic Regression and KNN in terms of F1 Score. This indicates that it may not handle the balance between precision and recall as effectively.
- The consistency between training and testing accuracy shows that the model is not overfitting, but the relatively low F1 Score suggests it may not be ideal for imbalanced datasets like fraud detection.

Comparison between 3 models



Best Performing Model: KNN (K = 23) achieves the highest scores in both accuracy and F1 Score, making it the most effective model for this dataset.

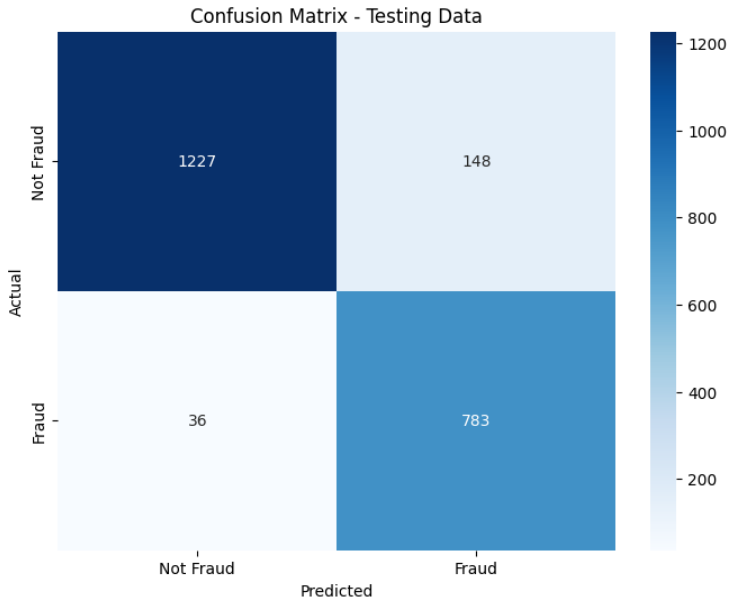
Runner-Up: Logistic Regression performs reasonably well and could be a simpler alternative with slightly lower performance.

Least Effective Model: Gaussian Naive Bayes has the lowest performance, likely due to its inability to handle the complexities or imbalances in the dataset.

3. Recommendation and Focus Area

Chosen model:
After comparing the performance of three models - Logistic Regression, KNN (K = 23), and Gaussian Naive Bayes - based on accuracy and F1 score, the **KNN (K = 23)** model emerged as the best-performing model.

It achieved the highest accuracy (92%) and F1 score (90%), indicating its strong ability to correctly classify fraudulent transactions while maintaining a good balance between precision and recall.



True Negatives (TN): 1227
- Transactions that are truly "Not Fraud" and were correctly predicted as "Not Fraud."

False Positives (FP): 148
- Transactions that are actually "Not Fraud" but were incorrectly predicted as "Fraud."
- These are cases of false alarms.

False Negatives (FN): 36
- Transactions that are actually "Fraud" but were incorrectly predicted as "Not Fraud."
- These are missed fraud cases.

True Positives (TP): 783
- Transactions that are truly "Fraud" and were correctly predicted as "Fraud."

Key factor and considerations for improving fraud detection:

1) **Focus on Late Night Transactions**:

- Fraudulent activities are more likely to occur during late-night hours (e.g., 12:00 AM to 3:00 AM).
- Implement stricter monitoring during these hours to detect unusual or suspicious transactions more effectively.

2) **Limit High Transaction Amounts**:

- Set transaction amount thresholds and require additional verification (e.g., three-sign authorization or multi-factor authentication) for transactions exceeding a significant limit.
- This reduces the risk of large-scale fraud and adds an extra layer of protection for high-value transactions.

3) **Enhanced Monitoring for Withdrawal Payments**:

- Withdrawals are a common transaction type for fraud.
- Introduce stricter scrutiny and advanced anomaly detection algorithms to flag unusual withdrawal patterns.