CAPSTONE-PROJECT CREDIT CARD FRAUD DETECTION

PRESENTATION

NARAYANA ISANAKA

PROBLEM STATEMENT

BUSINESS PROBLEM OVERVIEW

For many banks, retaining high profitable customers is the number one business goal. Banking fraud, however, poses a significant threat to this goal for different banks. In terms of substantial financial losses, trust and credibility, this is a concerning issue to both banks and customers alike.

It has been estimated by Nilson Report that by 2020, banking frauds would account for **$30 billion** worldwide. With the rise in digital payment channels, the number of fraudulent transactions is also increasing in new and different ways.

In the banking industry, credit card fraud detection using machine learning is not only a trend but a necessity for them to put proactive monitoring and fraud prevention mechanisms in place. Machine learning is helping these institutions to reduce time-consuming manual reviews, costly chargebacks and fees as well as denials of legitimate transactions.

CREDIT CARD FRAUD DETECTION

- BANKING FRAUDS WOULD ACCOUNT FOR **$30 BILLION** WORLDWIDE.
- CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING HELPS:
  - Proactive monitoring
  - Fraud prevention mechanisms
  - Machine learning reduce time-consuming manual reviews

RISKS:

- **Detect Fraud transaction accurately.**
- **Detect low value high frequent transaction well**
- **Detect high value low frequent transactions well**
- **Do not classify a non-fraud transaction as fraudulent transaction**

OBJECTIVE

GOAL IS TO DEVELOP MULTIPLE MODELS TO CLASSIFY TRANSACTIONS AS FRAUD/NON-FRAUD. EVALUATE MODELS AND DETERMINE THE BEST MODEL.

EXPLORATORY DATA ANALYSIS (EDA)

CREDIT CARD FRAUD DETECTION

Exploratory Data Analysis

1. Data Cleaning
2. Data Analysis
3. Recommendations
4. Understanding the Data set

Understand the Data set

1. Nature of data, related data sets, domain, timeframe and size of the data set.
2. Metadata

Data Cleaning

1. Fix Rows and columns
2. Fix missing values
3. Standardise values
4. Fix invalid values
5. Filter data

Data Analysis

1. Perform Univariate Analysis
2. Segmented Univariate Analysis
3. Bivariate analysis
4. Derived metrics

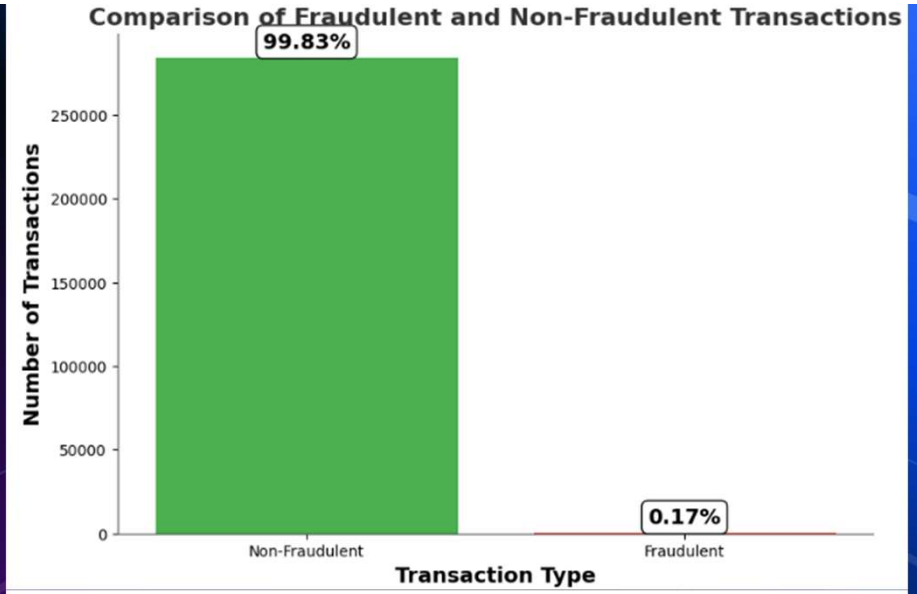
Insights & Recommendations

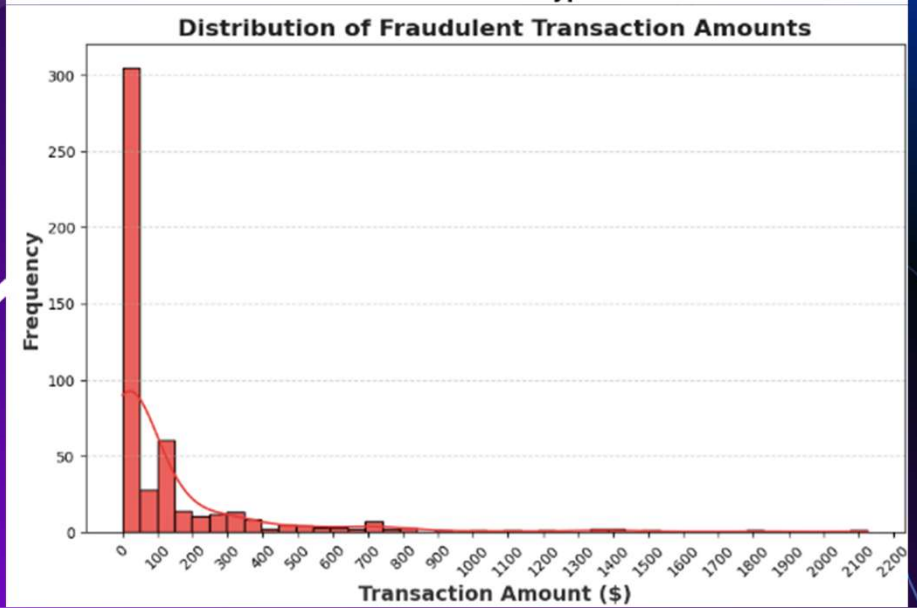
1. Extract insights and provide recommendations

UNDERSTANDING THE DATA SET.

KEY FINDINGS

1. The dataset has PCA performed data and has tow columns Time and Amount.
2. The dataset has total of 284807 transactions
3. There are 31 fields in total of them 29 PCA derived fields , additional Time and Amount fields.
4. The data set has Total of `284807 transactions:`
   `- with Non-Fraudulent Transactions: 284315 (99.83%)`
   `- Fraudulent Transactions: 492 (0.17%)`
5. `Most Fraud transactions occur at low $ values.`





DATA CLEANING.

KEY FINDINGS

1. There are no null values.
2. Time and Amount are the only normal values
3. All other fields are values obtained after performing PCA on original dataset.
4. All the columns are of type float64 except from column Class which is int64
5. Class value of 1 - represents Fraud transaction
6. Class value of 0 - represents Non-Fraud transaction

```
Class
0    284315
1       492
Name: count, dtype: int64
99.82725143693798 0.1727485630620034
```

Observation

- The data set has Total Number of Transactions : 284807
- with Non-Fraudulent Transactions: 284315 (99.83%)
- Fraudulent Transactions: 492 (0.17%)
- The dataset has very high class imbalance.
- Only 492 records are there among 284807 records which are labeld as fradudulent transaction.

DATA ANALYSIS.

KEY FINDINGS

1. There is significant overlap between the two distributions, meaning time alone is not a perfect differentiator for fraud detection.

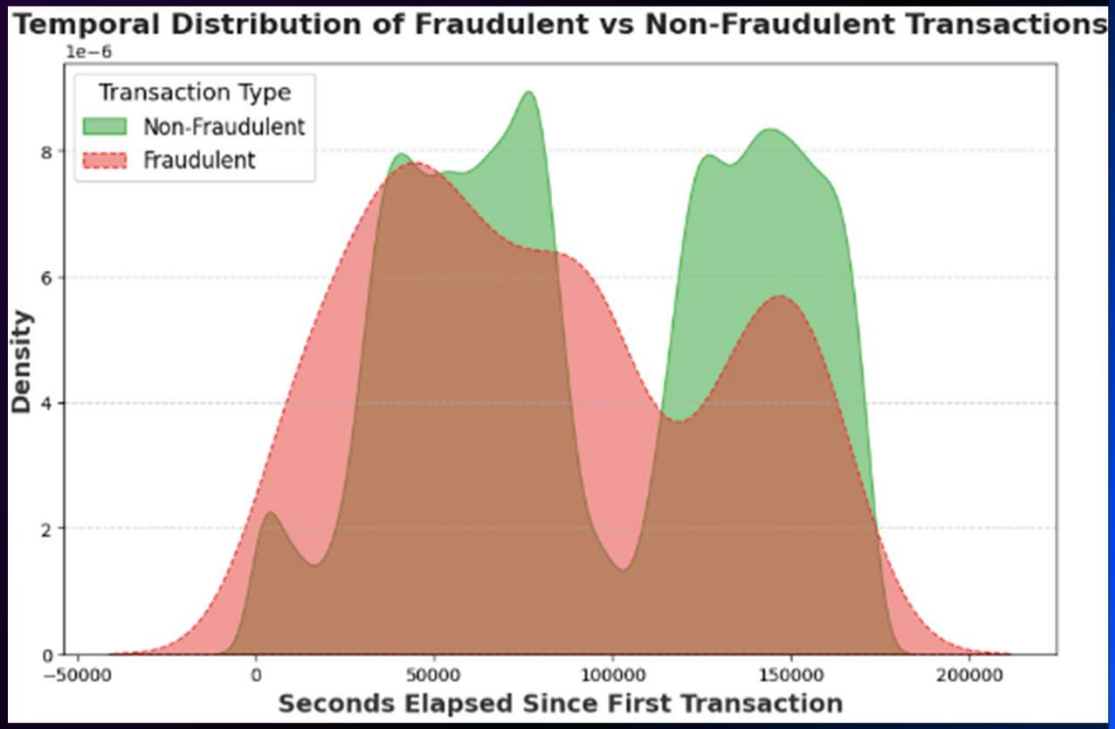
2. **Feature Engineering Ideas:**

    1. **Time since last transaction** – Fraudulent transactions might cluster within short intervals.

    2. **Transaction burst frequency** – Fraudsters may execute multiple transactions in quick succession.

    3. **Time-based anomaly detection** – Outliers in transaction timing could indicate fraudulent behavior.

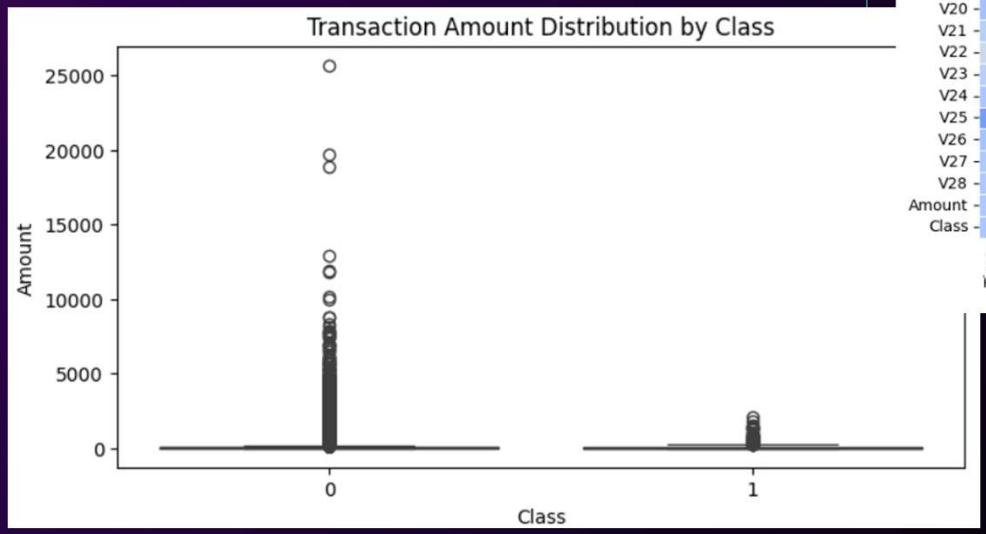
    4. **Time of day effects** – Fraudulent transactions may peak at unusual hours.

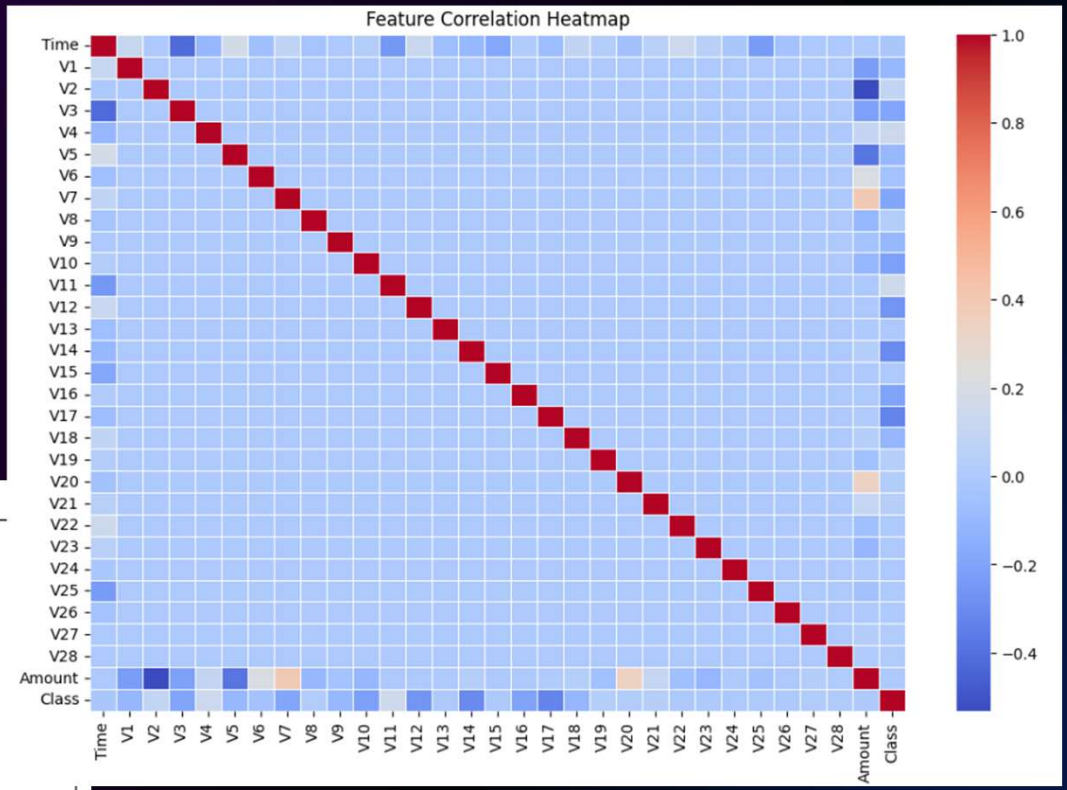


This is a kernel density estimation (KDE) plot showing the distribution of elapsed time (in seconds) between a given transaction and the first transaction, for both fraudulent and non-fraudulent transactions.

DATA ANALYSIS.

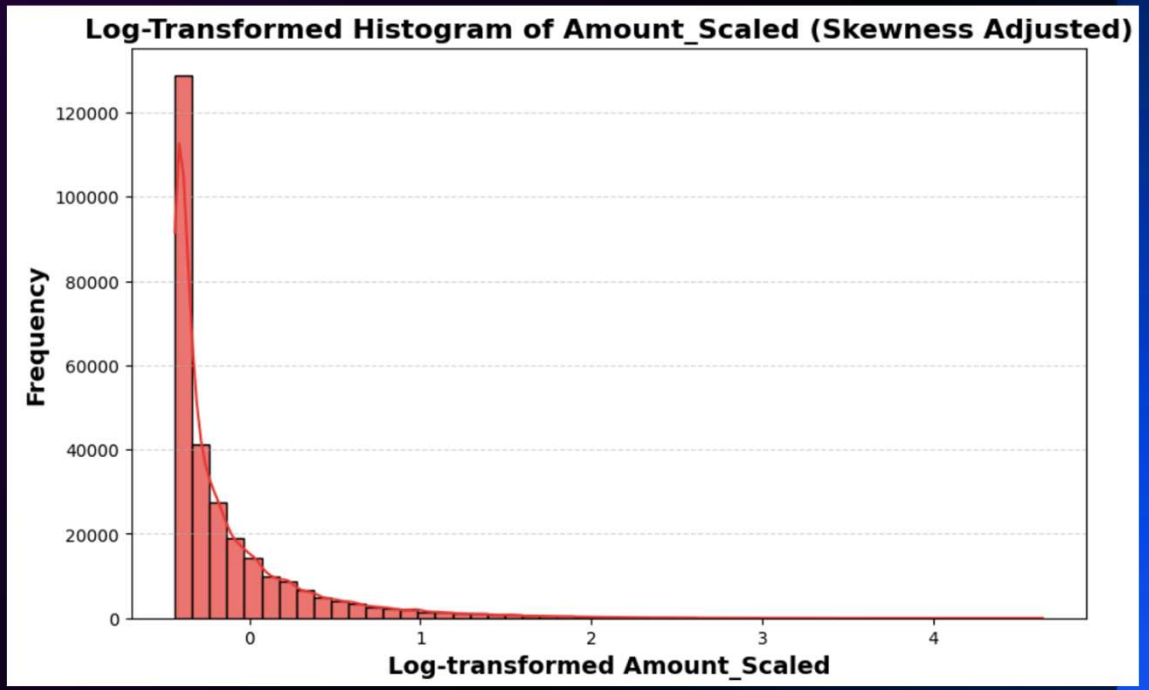
KEY FINDINGS





DATA ANALYSIS.

KEY FINDINGS



MODEL BUILDING

MODEL: LOGISTIC REGRESSION

Key findings

```
Logistic Regression Model Evaluation:

Model Performance for Logistic Regression:
Accuracy: 0.9623
Precision: 0.9700
Recall: 0.9155
F1 Score: 0.9419
ROC-AUC Score: 0.9506

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.99      0.97     56777
           1       0.97      0.92      0.94     28518

    accuracy                           0.96     85295
   macro avg       0.96      0.95      0.96     85295
weighted avg       0.96      0.96      0.96     85295
```

MODEL: RANDOM FOREST

Key findings

```
Random Forest Model Evaluation:

Model Performance for Random Forest:
Accuracy: 0.9998
Precision: 0.9996
Recall: 0.9999
F1 Score: 0.9998
ROC-AUC Score: 0.9999

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56777
           1       1.00      1.00      1.00     28518

    accuracy                           1.00     85295
   macro avg       1.00      1.00      1.00     85295
weighted avg       1.00      1.00      1.00     85295
```

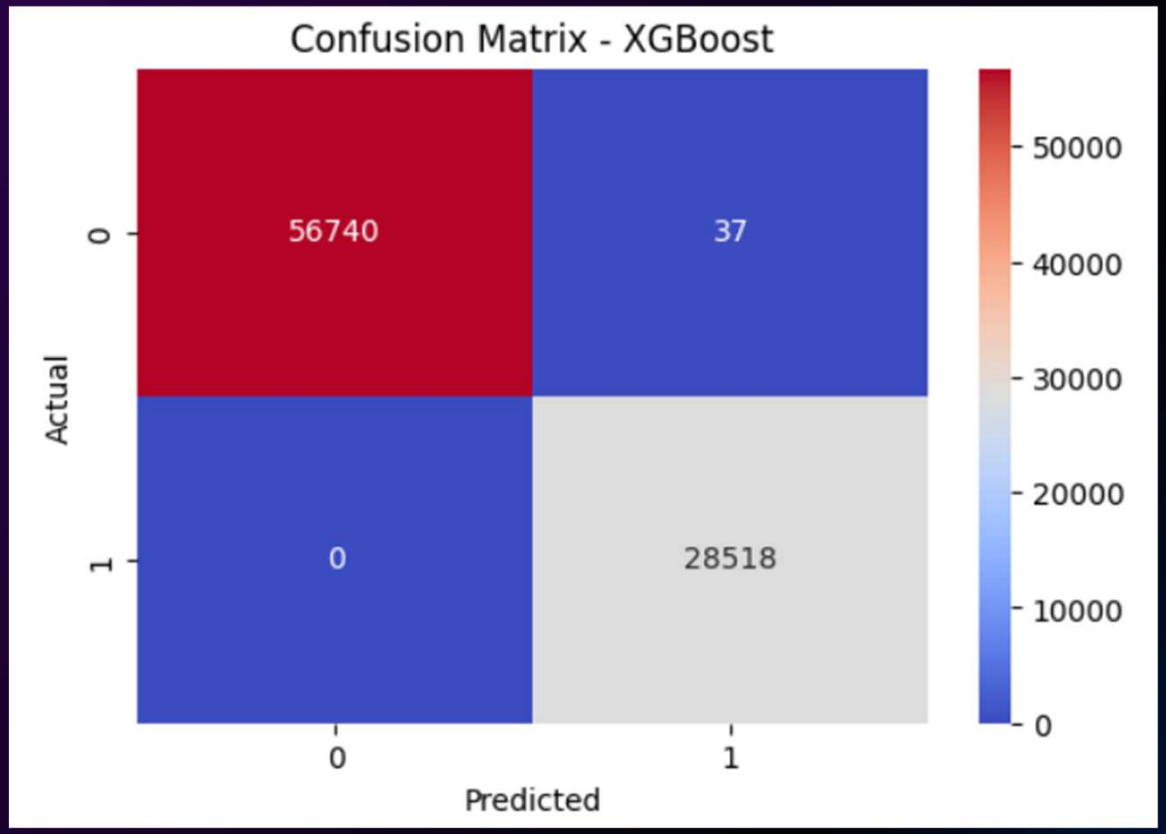
MODEL: XG BOOST

Key findings

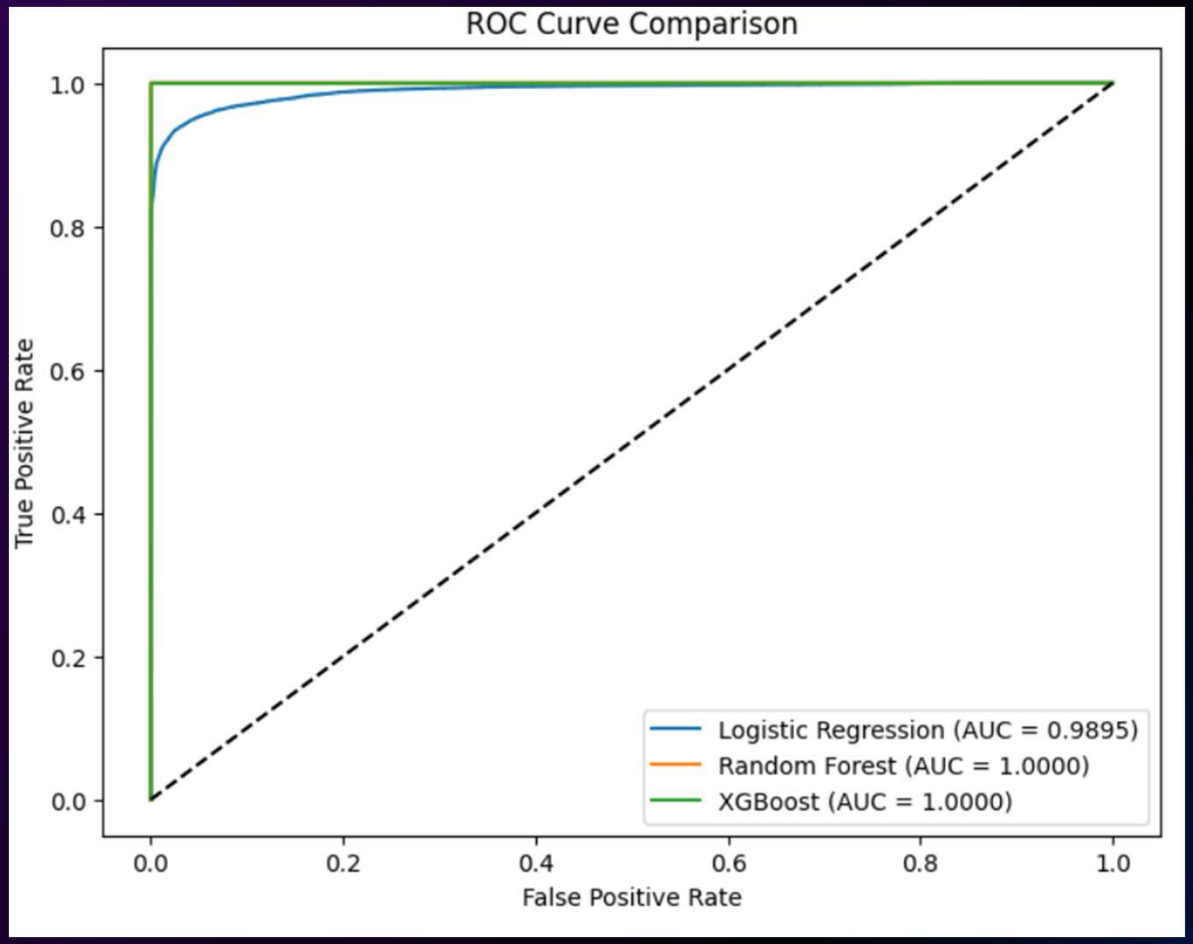
```
XGBoost Model Evaluation:

Model Performance for XGBoost:
Accuracy: 0.9996
Precision: 0.9987
Recall: 1.0000
F1 Score: 0.9994
ROC-AUC Score: 0.9997

Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     56777
           1       1.00      1.00      1.00     28518

    accuracy                           1.00     85295
   macro avg       1.00      1.00      1.00     85295
weighted avg       1.00      1.00      1.00     85295
```

MODEL: RANDOM FOREST

Key findings

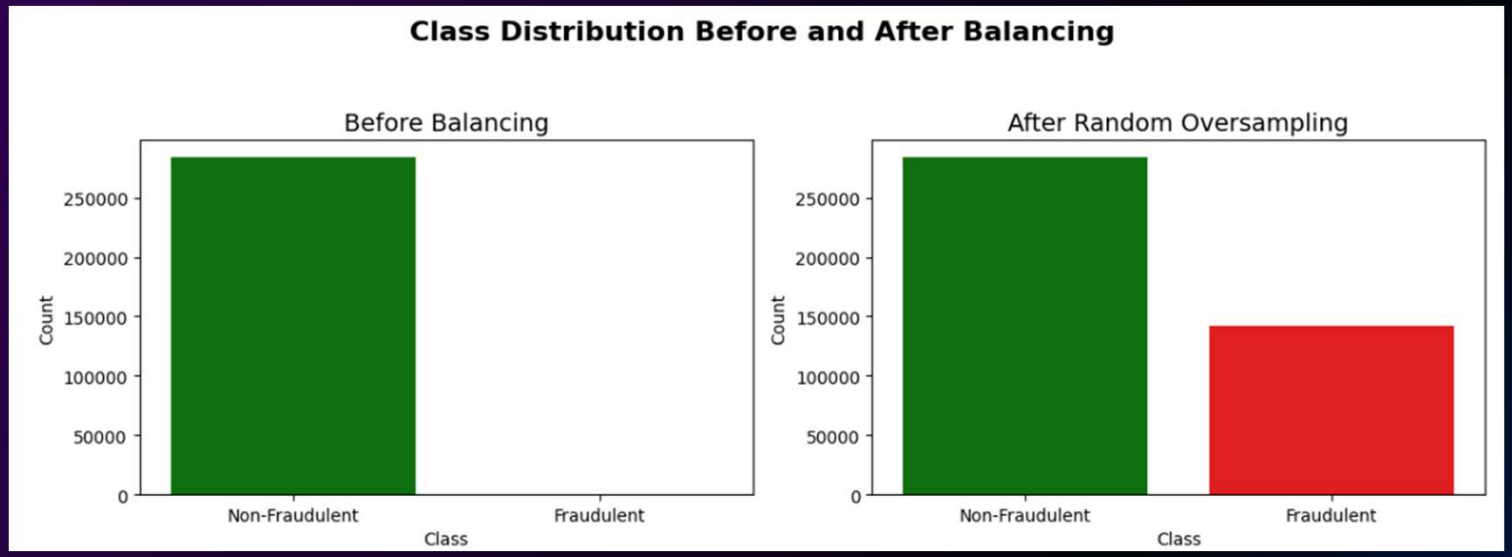


ROC: TRUE POSITIVES / FALSE POSITIVES



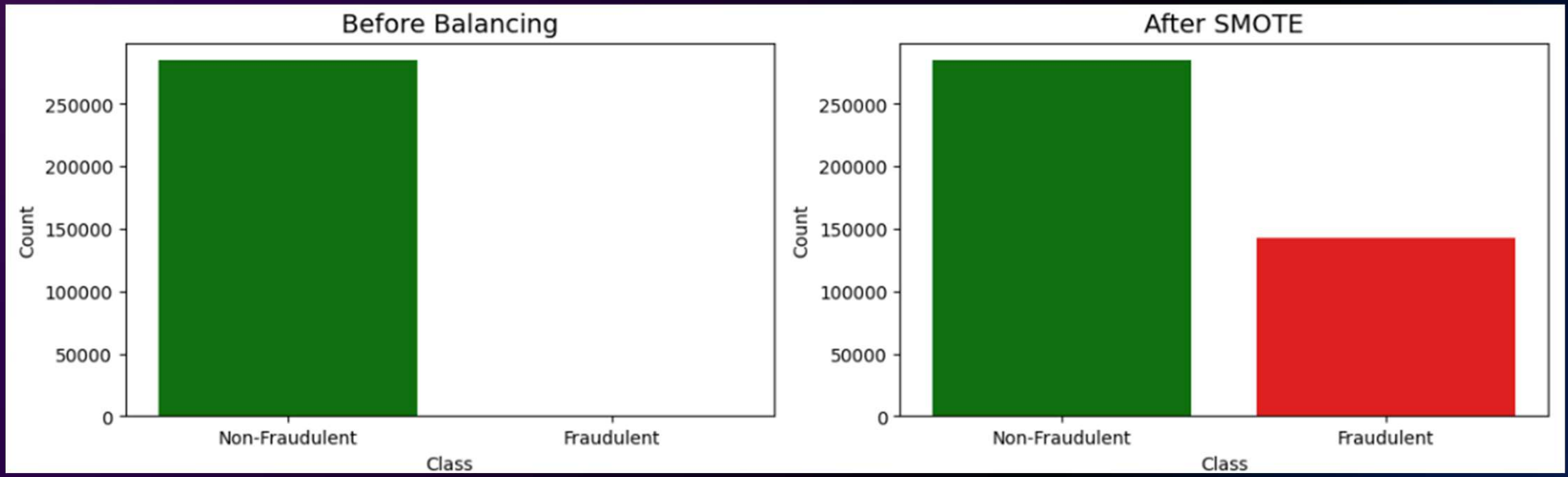
CLASS BALANCING

RANDOM OVERSAMPLING



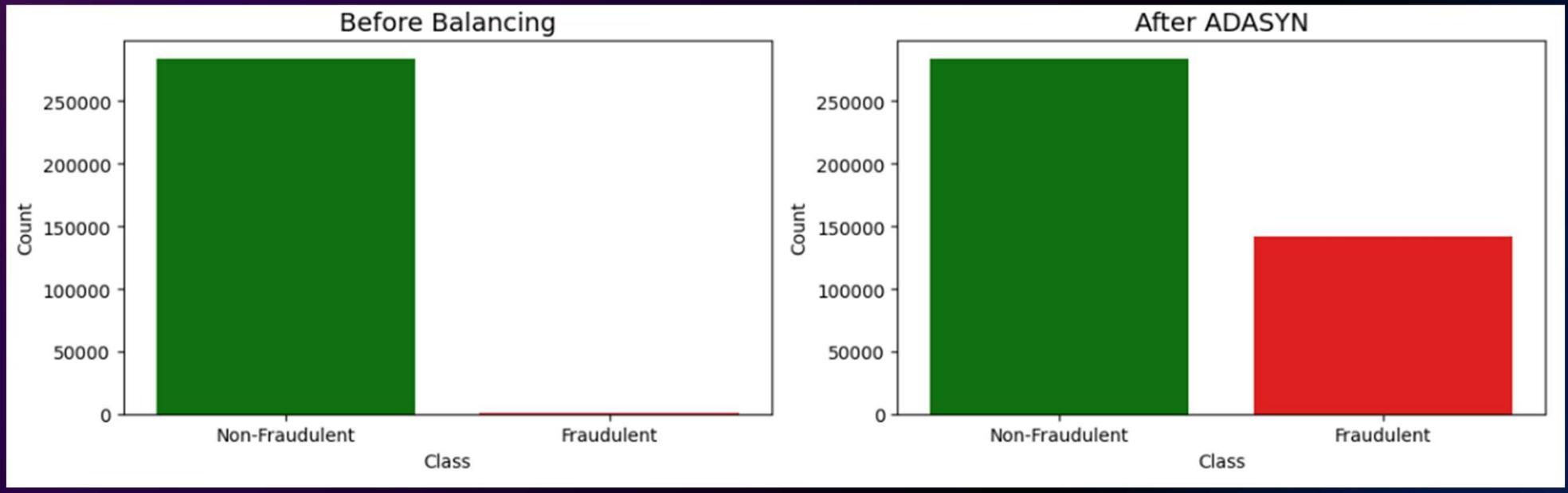
CLASS BALANCING

USING - SMOTE



CLASS BALANCING

USING - ASASYN

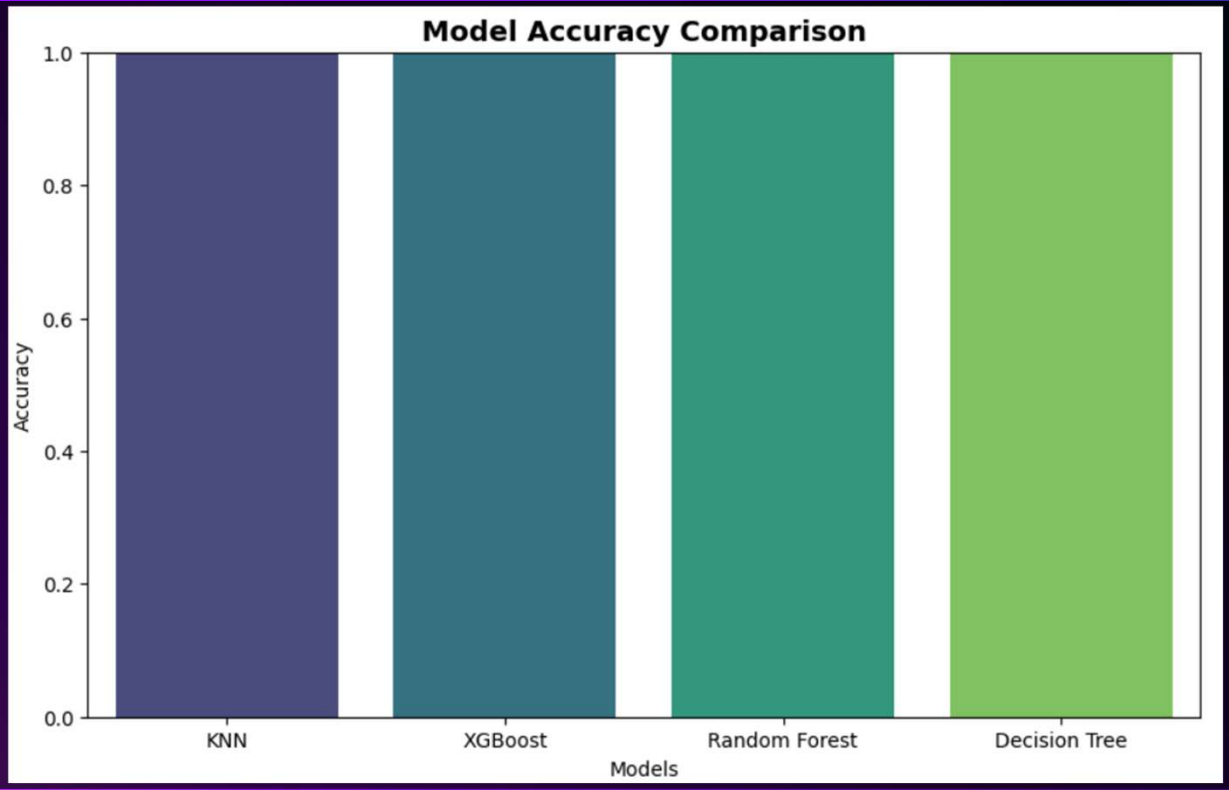


MODEL COMPARISON

Model Performance Comparison:

	Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
0	KNN	0.998183	0.994594	1.000000	0.997290	0.999604
1	XGBoost	0.999801	0.999404	1.000000	0.999702	0.999991
2	Random Forest	0.999859	0.999720	0.999860	0.999790	0.999996
3	Decision Tree	0.997561	0.995103	0.997616	0.996358	0.997575

SPEAKING ENGAGEMENT METRICS



CONCLUSION:

LOGISTIC REGRESSION:

RANDOM FOREST:

XGBOOST:

```
Conclusion:
- Logistic Regression provides a baseline model but may not handle complex patterns well.
- Random Forest improves performance by handling non-linearity and interactions.
- XGBoost often outperforms other models due to its advanced boosting technique and feature selection capabilities.
- Based on ROC-AUC scores, the model with the highest AUC is the best at distinguishing fraudulent transactions.
```

THANK YOU!!

Narayana Isanaka