Credit Card Fraud Detection

...

Agenda

- Overview/Approach the Problem
- Background
- Key Insights
- Business Impact : Cost Benefit Analysis
- Appendix : Data Methodology

Overview

- This capstone project aims to address the increasing number of credit card frauds faced by Finex, a leading financial service provider based in Florida, US. With the rise of digital payment channels, fraudulent transactions have become more prevalent, leading to significant revenue and profitability crises for the bank. The current situation necessitates the development of a fraud detection system using machine learning techniques to identify and prevent unauthorized transactions effectively.

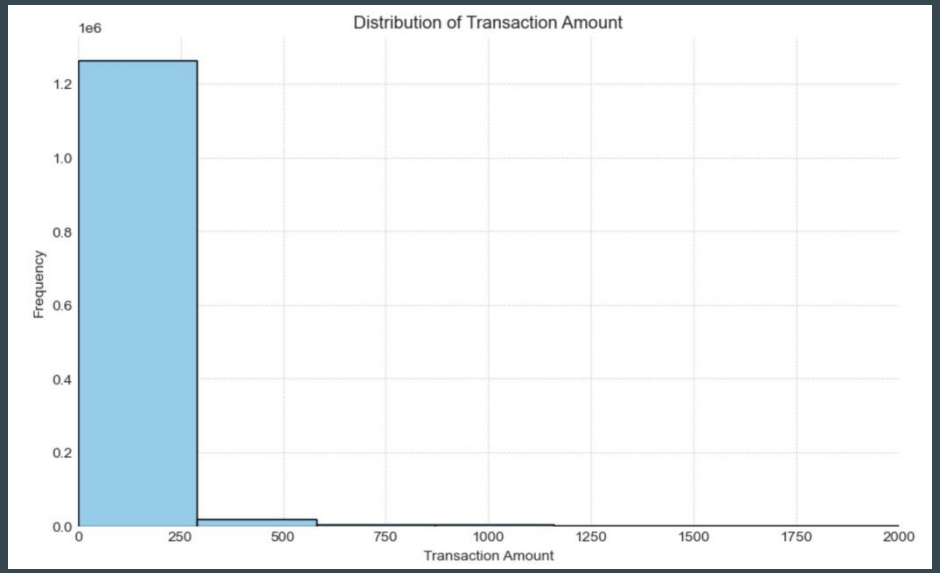
Steps to approach the problem:

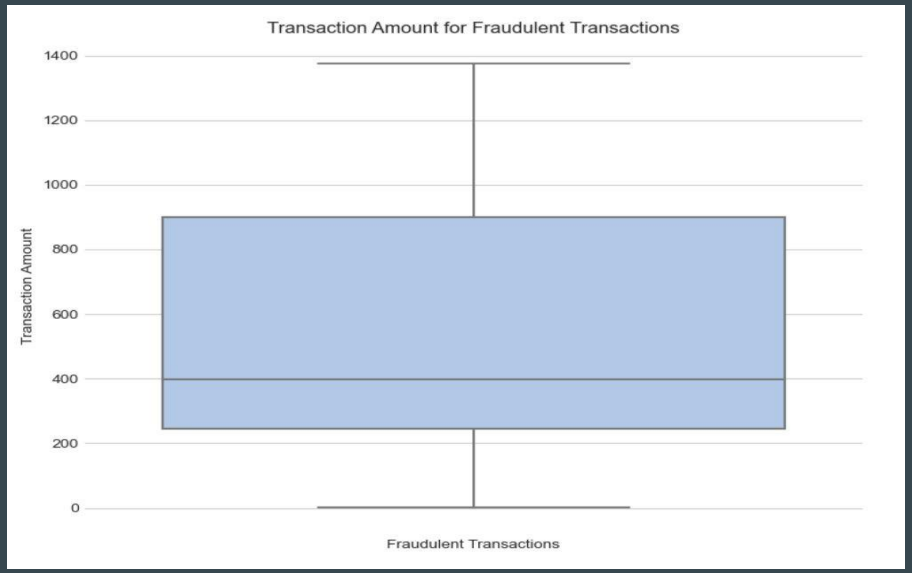
- As part of the analytics team tasked with developing a fraud detection model, we aim to leverage historical transactional data to detect fraudulent activities among customers and merchants.
- Our objective involves analyzing the business repercussions of fraudulent transactions and proposing strategies for the bank to minimize fraud risks effectively.
- Implementation of proactive monitoring and fraud prevention measures is paramount in our approach.
- Machine learning plays a crucial role in streamlining processes, diminishing manual review efforts, mitigating chargeback expenses, and minimizing the rejection of legitimate transactions.

Background

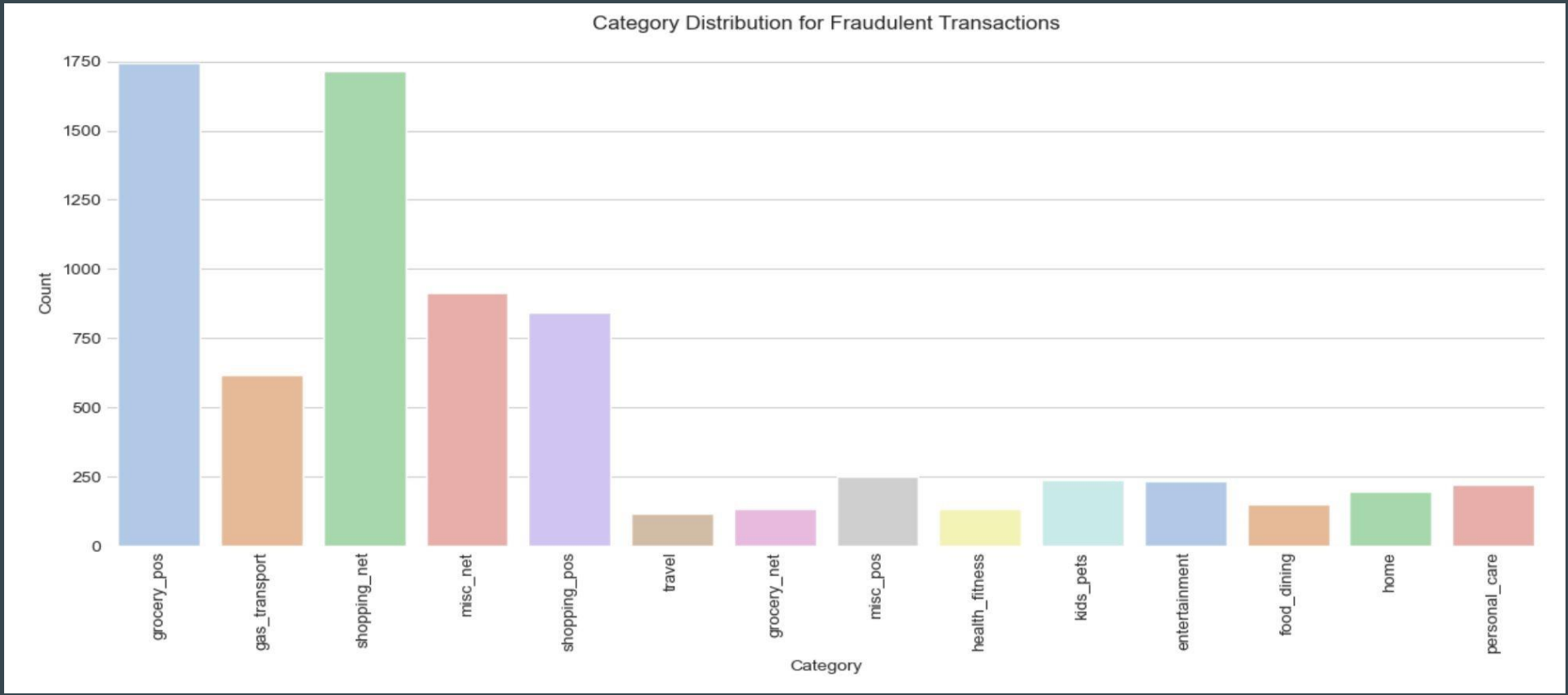
Problem Identification: Recognizing the surge in credit card frauds and the impact on Finex's revenue and profitability.
Root Cause Analysis: Understanding the factors contributing to the increase in fraudulent transactions, including stolen/lost cards, data breaches, and ATM skimming.
Proposed Solution: Building a fraud detection model using machine learning techniques to identify fraudulent activities in real-time.
Data Collection: Gathering a comprehensive dataset of credit card transactions, including relevant features such as transaction amount, timestamp, and merchant information.
Data Preprocessing: Cleaning the dataset, handling missing values, removing duplicates, and performing feature engineering to extract valuable insights.
Exploratory Data Analysis (EDA): Analyzing the distribution of fraudulent transactions, visualizing transaction patterns, and identifying key features.
Feature Selection: Selecting the most relevant features to enhance the predictive power of the model.
Model Selection: Experimenting with various machine learning algorithms suitable for fraud detection, such as **Logistic Regression, Decision Tree, Random Forest, and XGBoost.**
Addressing Class Imbalance: Utilized SMOTE and ADASYN techniques.
Hyperparameter Tuning and Final Model Selection: Utilized GridSearchCV for hyperparameter tuning, and selected XGBoost (ADASYN) as the optimal model.

Key Insights



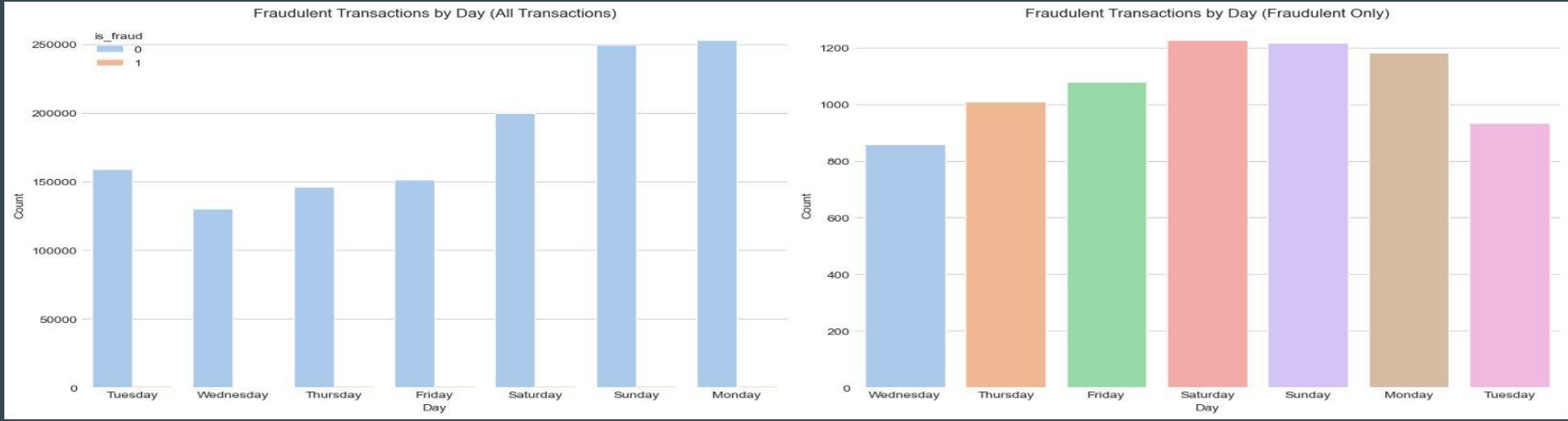


- The majority of transactions made by customers fall within the $300 bracket, indicating typical spending patterns.
- When focusing specifically on fraudulent activities, fraudulent transactions are concentrated primarily between $200 and $1000.
- The concentration of fraudulent transactions between $200 and $1000 indicates a risk threshold. Transactions within this range require heightened scrutiny for effective fraud risk mitigation.



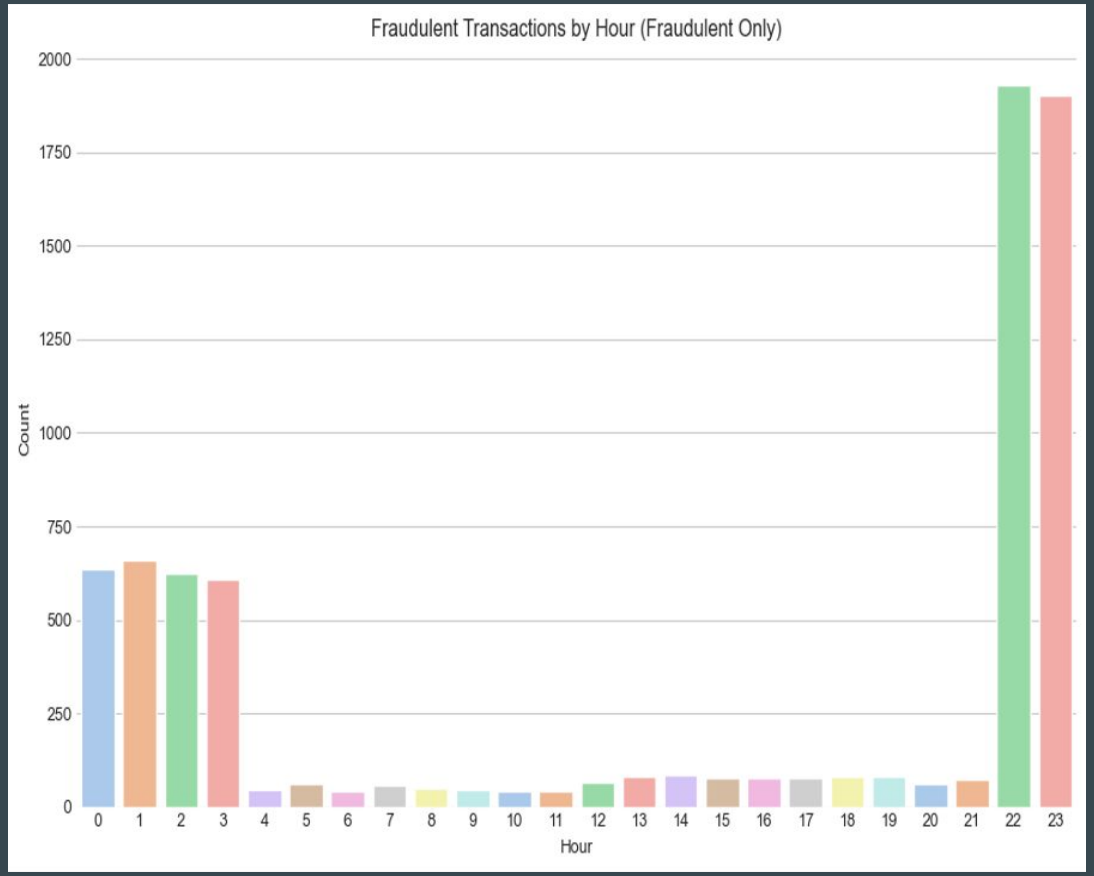
Most fraudulent transactions originate from categories like grocery_pos, shopping_net, misc_net, shopping_pos, and gas_transport.

THE HIGHEST OCCURRENCE OF FRAUDULENT TRANSACTIONS IS OBSERVED ON FRIDAY, SATURDAY, SUNDAY, AND MONDAY

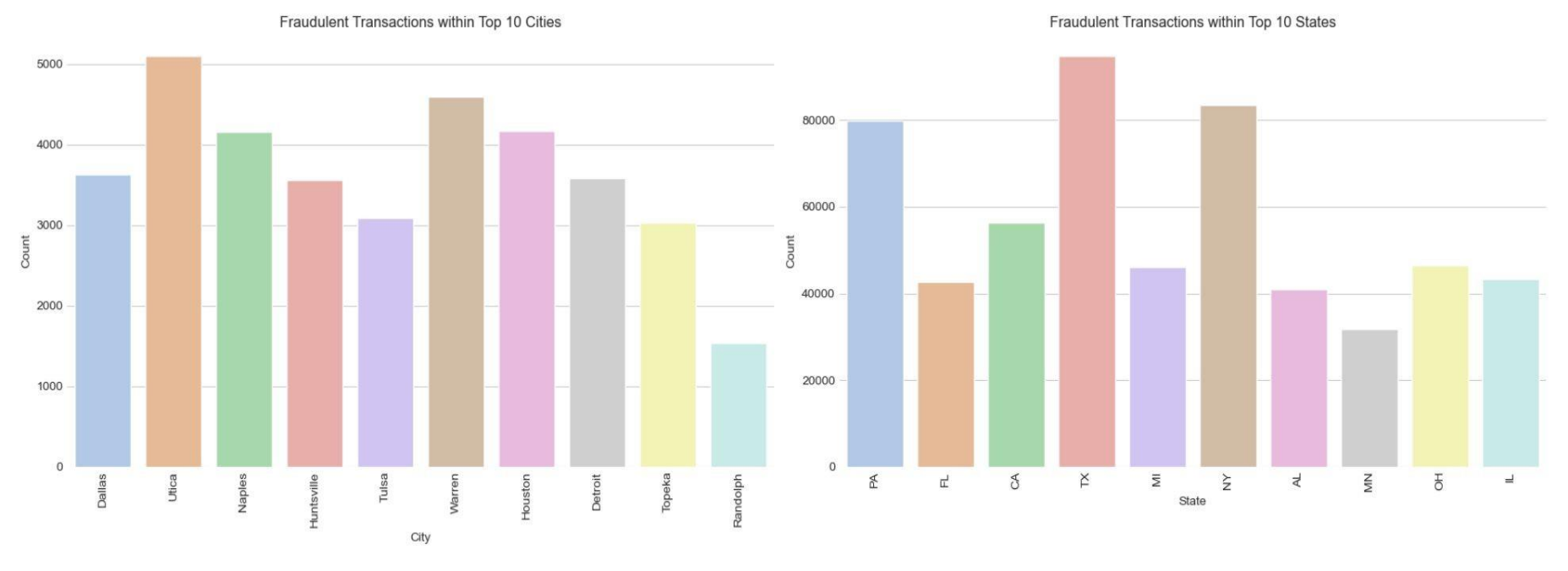


- Saturday, Sunday, and Monday emerge as the busiest days for transactions, indicating typical consumer spending behaviours over the weekend and the start of the week.
- More fraud occurs on Fridays through Mondays, possibly linked to high transaction volumes on those days.
- Fraudulent transactions are more common on weekends and at the beginning of the week, underscoring the importance of reinforcing security measures during busy periods.

THE MAJORITY OF FRAUDULENT TRANSACTIONS ARE CLUSTERED INTO TWO MAIN TIME PERIODS - 22:00-23:00 AND 0:00-3:00



- A substantial number of fraudulent transactions occur during the late-night hours, particularly between 22:00 and 23:00.
- Another significant surge in fraudulent transactions is observed in the early hours of the morning, specifically between 0:00 and 3:00.
- By focusing on these two distinct time periods, 22:00-23:00 and 0:00-3:00, security measures can be targeted effectively to mitigate the risk of fraudulent activities during peak hours.



- Among cities, Utica, Houston, Warren, and Naples have the highest number of fraudulent transactions.
- Among states, Texas (TX), New York (NY), and Pennsylvania (PA) lead in fraudulent transactions.

Business Impact : Cost Benefit Analysis

Cost Benefit Analysis - Part I			
a	Average number of transactions per month	$	77,183.08
b	Average number of fraudulent transaction per month	$	402.13
c	Average amount per fraud transaction	$	530.66

Cost Benefit Analysis - Part II			
1	Cost incurred per month before the model was deployed (b*c)	$	2,13,392.22
2	Average number of transactions per month detected as fraudulent by the model (TF)	$	1,411.71
3	Cost of providing customer executive support per fraudulent transaction detected by the model	$	1.50
4	Total cost of providing customer support per month for fraudulent transactions detected by the model (TF*$1.5)	$	2,117.57
5	Average number of transactions per month that are fraudulent but not detected by the model (FN)	$	22.43
6	Cost incurred due to fraudulent transactions left undetected by the model (FN*c)	$	11,901.98
7	Cost incurred per month after the model is built and deployed (4+6)	$	14,019.55
8	**Final savings = Cost incurred before - Cost incurred after(1-7)**	**$**	**1,99,372.67**

- After the model deployment, the monthly cost decreased significantly, from $213,392.22 to $14,019.55, which is 93.43%.
- This translates to substantial monthly savings of $199,372.67.
- This substantial decrease underscores the model's effectiveness in mitigating fraudulent transaction losses and enhancing the bank's financial security.

Appendix : Data Methodology

	Train	Test
Model	Logistic Regression (SMOTE)	Logistic Regression (SMOTE)
Accuracy	0.929380	0.926774
Precision	0.918751	0.021808
Recall	0.942074	0.409790
F1 Score	0.930265	0.041412
Model	Logistic Regression (AdaSyn)	Logistic Regression (AdaSyn)
Accuracy	0.917154	0.901117
Precision	0.896838	0.015683
Recall	0.942717	0.398601
F1 Score	0.919205	0.030179
Model	Decision Tree (SMOTE)	Decision Tree (SMOTE)
Accuracy	0.952210	0.971496
Precision	0.970316	0.103756
Recall	0.932961	0.835897
F1 Score	0.951272	0.184598
Model	Decision Tree (AdaSyn)	Decision Tree (AdaSyn)
Accuracy	0.925626	0.899007
Precision	0.911885	0.034527
Recall	0.943064	0.933333
F1 Score	0.926861	0.066591
Model	Random Forest (SMOTE)	Random Forest (SMOTE)
Accuracy	0.911594	0.918302
Precision	0.911263	0.033538
Recall	0.911995	0.724942
F1 Score	0.911629	0.064109
Model	Random Forest (ADASYN)	Random Forest (ADASYN)
Accuracy	0.906615	0.882394
Precision	0.877733	0.024021
Recall	0.944813	0.743590
F1 Score	0.910035	0.046538
Model	XGBoost (SMOTE)	XGBoost (SMOTE)
Accuracy	0.996160	0.996300
Precision	0.997248	0.513522
Recall	0.995066	0.787879
F1 Score	0.996156	0.621781
Model	XGBoost (ADASYN)	XGBoost (ADASYN)
Accuracy	0.995732	0.996774
Precision	0.997352	0.558163
Recall	0.994101	0.787413
F1 Score	0.995724	0.653259

Top Performing Models Before Hyperparameter Tuning:

Model 4: Decision Tree (AdaSyn)
Model 3: Decision Tree (SMOTE)
XGBoost (SMOTE)
XGBoost (AdaSyn)

Evaluation metrics of the hyperparameter tuned models on test data:

	Model	Accuracy	Precision	Recall	F1 Score
0	Model 1 – Decision Tree (SMOTE)	0.978610	0.143307	0.912354	0.247706
1	Model 2 – Decision Tree (ADASYN)	0.983333	0.178458	0.920746	0.298971
2	Model 3 – XGBoost (SMOTE)	0.997855	0.698129	0.782751	0.738022
3	Model 4 – XGBoost (ADASYN)	0.997936	0.712521	0.779953	0.744714
4	Model 5 – XGBoost (ADASYN)_0.40	0.997416	0.628582	0.807925	0.707058
5	Model 6 – XGBoost (ADASYN)_0.15	0.986826	0.215979	0.917483	0.349649
6	Model 7 – XGBoost (ADASYN)_Penalty	0.985512	0.201174	0.926807	0.330590

Optimal Model After Hyperparameter Tuning:

Model 7 - XGBoost (ADASYN)_Penalty has been selected as the optimal choice for addressing credit card fraud detection as it demonstrates superior recall along with corresponding precision, establishing it as the optimal choice for addressing this credit card fraud detection challenge.

Thank You

Biplab Mondal