Fraud Detection in Financial Transactions: Methods, Findings & Recommendations

04-02-2024
Pruthvi Raj, Pudi
pp08507n@pace.edu
Practical Data Science
Master’s in Data Science
Seidenberg School of Computer Science, Pace University

Agenda

- Executive Summary
- Project Plan Recap
- Data
- EDA (Exploratory Data Analysis)
- Modeling Methods
- Findings
- Recommendations & Data Science next steps

Executive Summary

- Fraud detection in financial transactions is a critical aspect of maintaining security and trust in financial systems. With the rise of digital transactions, the need for robust fraud detection mechanisms has become paramount. By leveraging advanced technologies such as machine learning and data analytics, financial institutions can detect fraudulent activities efficiently, protecting both themselves and their customers from financial losses and reputational damage.

Models overview refer to appendix

Project Plan Recap

Deliverable	Due Date	Status
Data & EDA	03/19/2024	Completed
Methods, Findings & Recommendations	04/02/2024	Completed
Final Presentation	04/16/2024	Not started

Data

Data Summary

Data Source: Fraud transaction data has taken from data.world
Sample Size: 10127 transactions (this a synthetic dataset generated by PaySim).
Time Period: Data is assumed to be captured in 2018.

- Few rows have been dropped because of "NaN" values from 15 columns ('type', 'amount', 'nameOrig', 'oldbalanceDest', 'nameDest', 'oldbalanceDest', 'newbalanceDest', 'Acct type', 'Date of transaction', 'Time of day', 'isFraud')

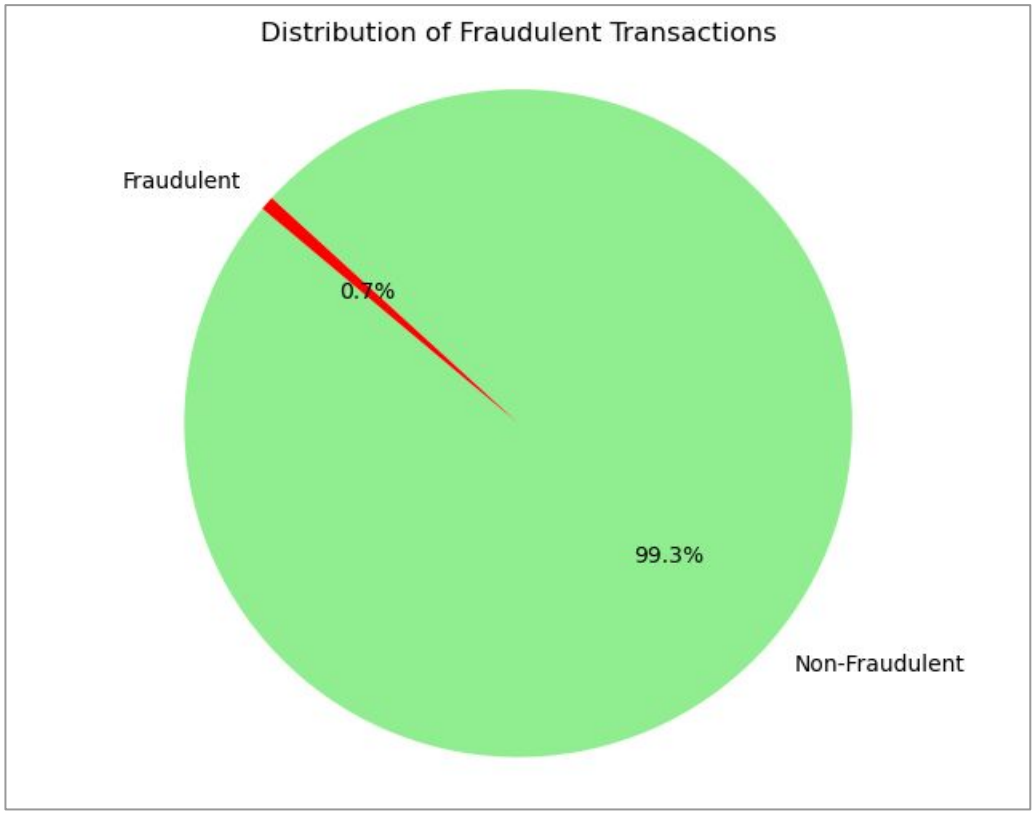
- Converted isFraud and isFlaggedFraud columns to categorical (1:Yes, 0:No)

- For more detailed information on each column refer to appendix

Assumption:
One assumption about the dataset is that it is representative of typical financial transaction data, with a wide range of transaction type, amounts, and other important attributes. Furthermore, it is expected that the dataset accurately captures both fraudulent and legitimate transactions, allowing the creation of effective fraud detection models.

Exploratory Data Analysis

Distribution of transactions



- This pie chart depicts the portion of fraudulent transactions compared to non-fraudulent transactions.
- The small slice **(0.7%)** represents the percentage of transactions identified as fraudulent.
- The large slice **(99.3%)** represents the percentage of transactions classified as legitimate.

Key Point:

- The chart indicates a very low overall fraud rate **(0.7%)**. This is positive indication, suggesting that the system effectively identifies the vast majority of transactions as legitimate.

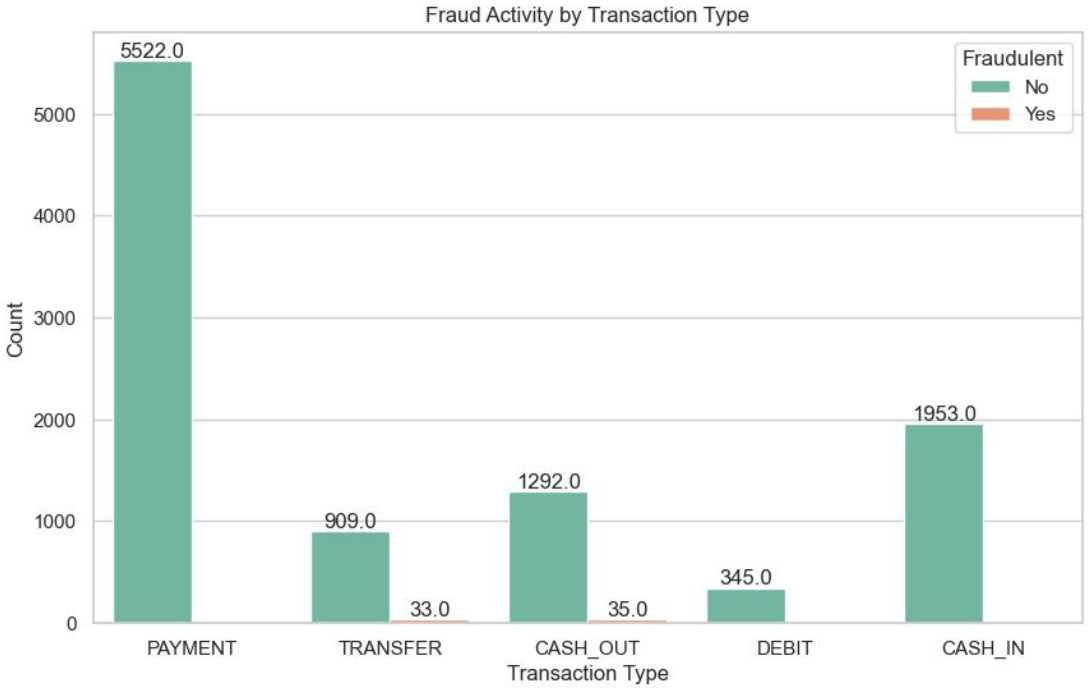
Distribution of flagged fraud

- This pie graph shows the proportion of flagged fraud column (transactions that are flagged above 200.000 per transaction).
- We can assume that the transactions are 100% legitimate.

Key point:

- We can assume that due to limited data, flagged fraud transactions are not captured in the data.

Fraud transaction activity by type



- This graph shows us the total no of transactions along with fraud activity(transaction).

- Out of five transactions types (Payment, Transfer, Cash Out, Debit, Cash In), Majority of transactions took place in Payment and Cash In type.

Key Point:

- Transfer and Cash Out type detected fraud transactions in which majority of fraud transactions happened in transfer type.

Fraud transaction activity by time of day

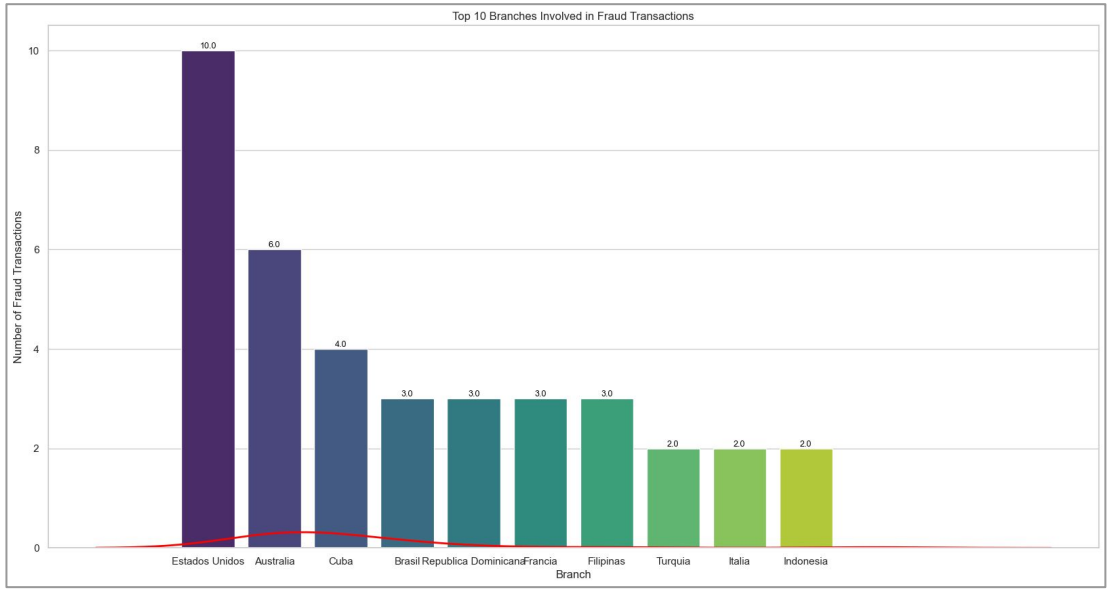
- The chart shows the total number of transactions for each **time period** (day) and it also reveals the number of fraudulent transactions identified within each period.

- There is fraudulent activity throughout the day, with the highest number of fraudulent transactions occurring in the Afternoon (25), Night (24) and, a notable number in the Morning (19).

Key Point:

- This suggests fraudsters may be active across different times, highlighting the need for continuous vigilance.

Top 10 fraud transactions by branch



All branch fraud transactions can be seen in appendix

- A branch in the United States has the most significant number of fraud transactions (10). This could be due to various reasons, including the branches location, the volume of transactions it processes, or security vulnerabilities.

- Several other branches also have a notable number of fraud transactions, ranging from 6 to 2.

Key Point:

- It's important to note that having a high number of fraud transactions could be due to various factors, not necessarily because a particular branch is less secure.

Modeling Methods

Significant features for modeling

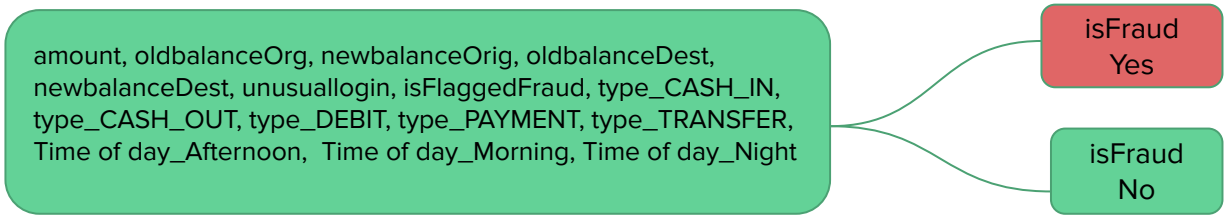
- Now that we have analyzed the data and drew few initial insights, now we'll use them and create a machine learning model.

Outcome Variable (isFruad): Is it a fraudulent transaction?, Here our goal is to predict a financial fraud transaction, Which helps us to maintain our financial systems secure and fraud free transactions.

Features: For this model we consider type of day (Moring, Day, & Night), type of payment modes (Cash In, Cash Out, Debit, Payment, Transfer), account balance (opening and closing balance) and unusual logins.

This features gives us overall insights of how transactions took place and helps us to understand the patterns of the fraud transactions.

What exactly are we predicting?

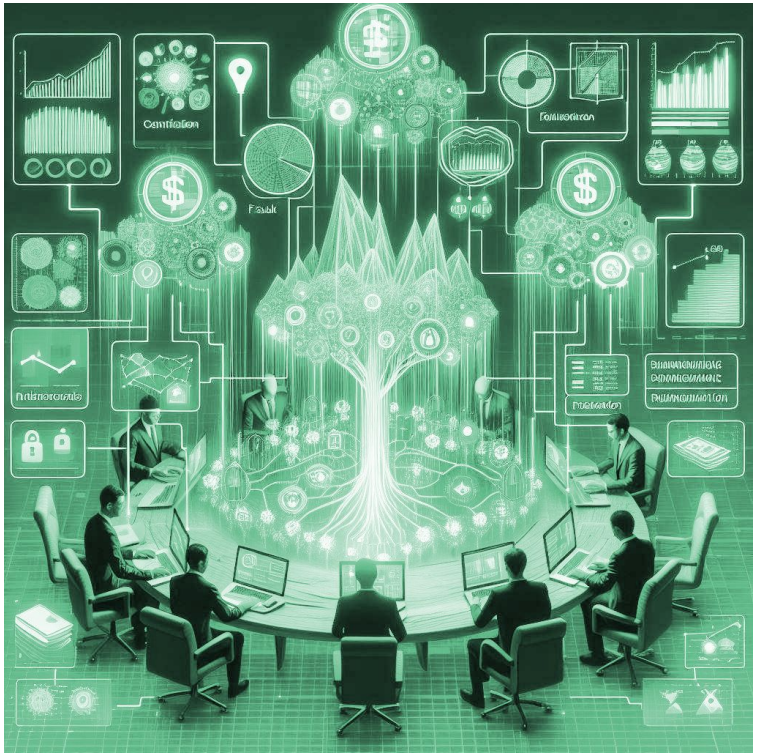


Understanding our fraud detection model & how it works

- A Random Forest Classifier functions similarly to a large squad of skilled security guards. Each guard (known as a decision tree) examines transaction data (amount, time, etc.) using their own experience (splitting rules). Some guards may focus on the amount, flagging transactions that go outside of the regular range. Others may focus on day of transaction to determining at what time of the day transaction took place.

- All the decision trees vote together. If enough trees flag a transaction as suspicious based on their individual rules, the Random Forest Classifier raises the alarm. This combined approach helps catch even cleverly disguised fraud transactions.

- **Why this Model?:** A Random Forest Classifier acts like a team of experts working together to spot fraud in financial transactions, making it a powerful model to detect a fraud transaction.



Understanding our fraud detection model & how it works

Combating Fraud with Balanced Detection:
In finance industry it’s crucial to identify and flag the fraud transactions and maintain the system secure and also it's crucial not to be overly biased by a single factor. Our model naturally avoids getting overly influenced by just one aspect of the data and maintain balanced decision making.

Adaptable Learning for Unique Transactions:
As every transaction is unique in system, Random Forest adapts to varied characteristics of the transactions, like amount, type of day and payment types etc.,

Collective Wisdom from Transaction Patterns:
Each decision tree generates a significant prediction by combining multiple factors and learnings from the transaction patterns.

More detailed (technical version) of our model, refer to (appendix)

Findings

Insights: How well can we detect a fraud transaction?

Accuracy:
Our model achieved a remarkable 94% accuracy in identifying fraudulent transactions. This signifies a very strong performance in differentiating legitimate transactions from fraudulent ones.

Significant benefits for our business:

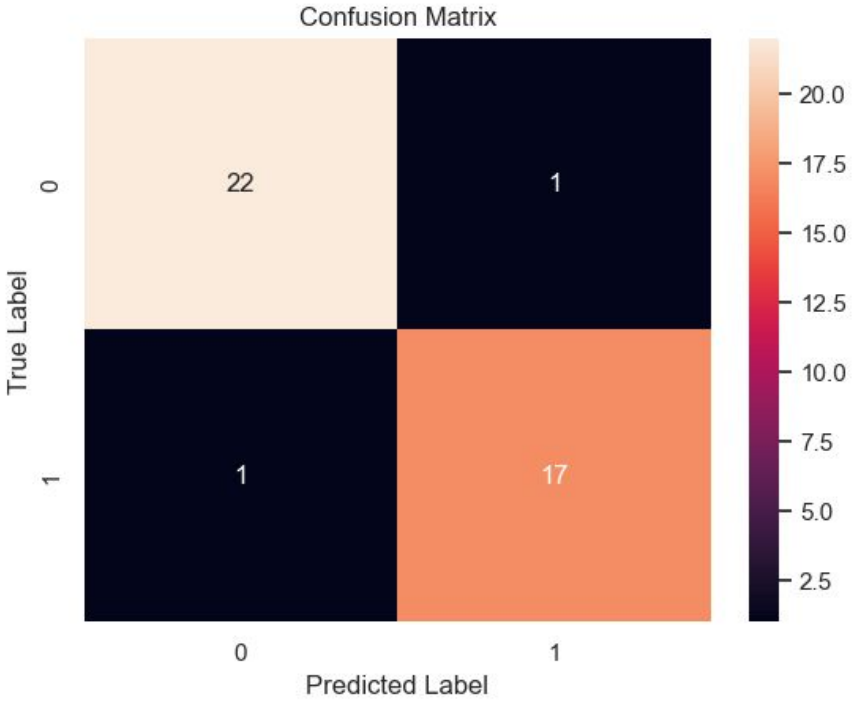
1. **Reduced Financial Losses:** By effectively catching 94% of fraudulent transactions, we minimize financial losses associated with fraudulent activities.

2. **Enhanced Security:** The model strengthens our security posture, making it significantly more difficult for fraudulent transactions to slip through.

3. **Improved Customer Experience:** By proactively identifying and preventing fraud, we ensure a smoother and more secure experience for our customers.

More detailed (technical version) of our model, refer to (appendix)

How does our model predicts fraud transactions?



- Out of 23 not fraud transactions, our model correctly predicted 22 as not fraud and missed 1 transaction and labeled as as fraud.
- For 18 fraud transactions our model predicted 17 transactions as fraud and missed 1 transaction and labeled as not fraud.

Note: Detailed comparison of model's performances can be found in (appendix)

Recommendations & Data Science next steps

Recommendations

Findings:

- Transaction amount and type are vital in spotting fraud transactions.
- Time of day affects transaction patterns, especially during off-peak hours.

Business Insights:

- Detecting and preventing fraudulent transactions is crucial for maintaining system security and customer trust.
- Understanding transaction behavior throughout the day helps identify anomalies and potential fraud.

Actionable Recommendation:

- Set up automatic alerts for large transactions and implement extra verification steps for high-risk transaction types.
- Strengthen monitoring during off-peak hours with real-time anomaly detection and additional security measures.

Technical steps for data science team

Building a more advanced model: To improve the accuracy and reliability of our predictions. Explore more sophisticated machine learning techniques, like ensemble methods Gradient Boosting or neural networks, that could more effectively capture complex patterns in transactions behaviour.

Continuous Model Monitoring & Update data: Regularly update the model with new data and Implement a robust model monitoring system to track model performance over time and detect any degradation or drift in performance.

Handle Class Imbalance: Address the dataset class imbalance with more advanced techniques like SMOTE, K-fold Cross-Validation for better model performance.

Refine Feature Engineering: Continuously improve feature selection to optimize model performance.

Thank you.

Appendix

Data Description

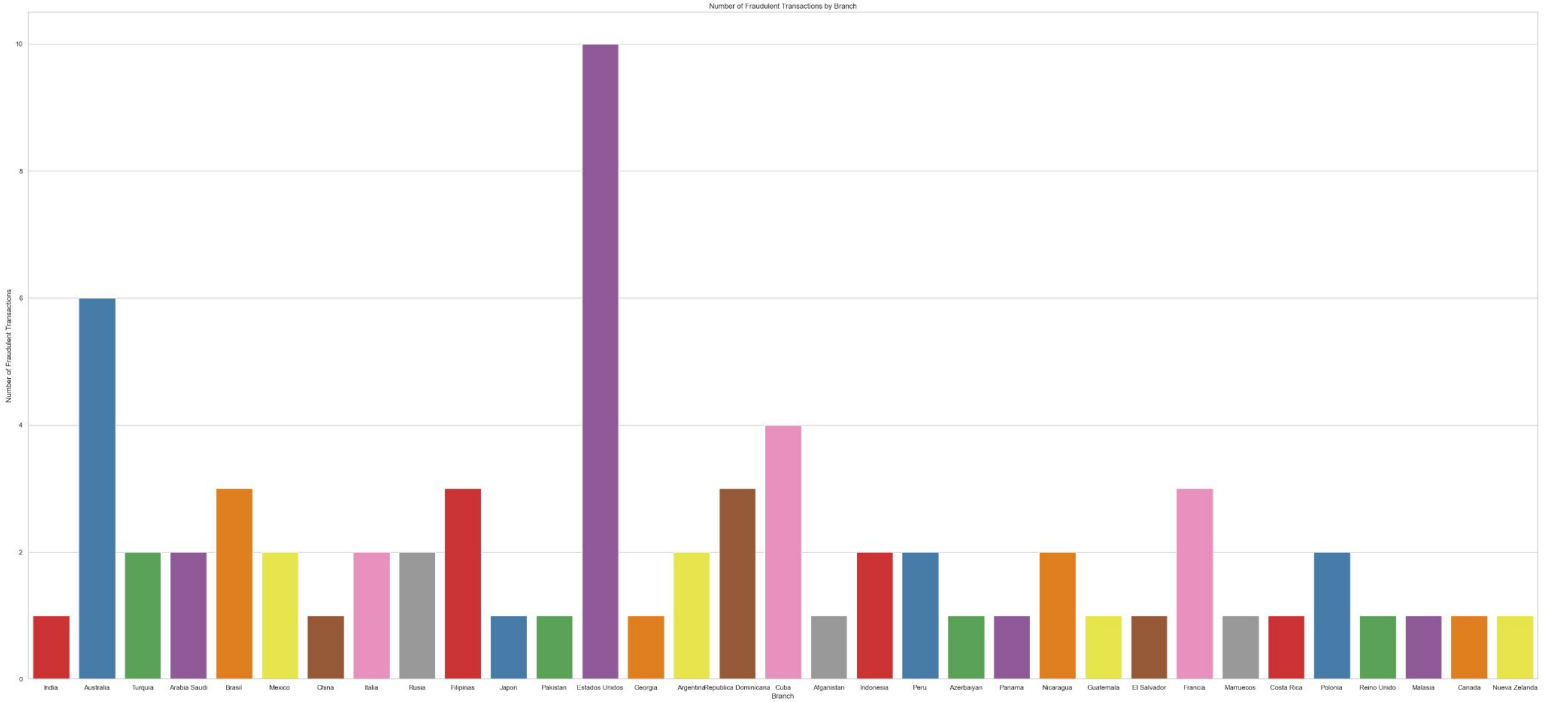
step - maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).
type - CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.
branch - location where the transaction took place.
amount - amount of the transaction in local currency.
nameOrig - customer who started the transaction
oldbalanceOrg - initial balance before the transaction
newbalanceOrig - new balance after the transaction
nameDest - customer who is the recipient of the transaction
oldbalanceDest - initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants)
newbalanceDest - new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).
unusuallogin - suspected or unusual logins made
isFlaggedFraud - The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.
acct_type - The type of account either saving or current
date_of_transcation - date of transaction to place
time_of_day - morning, afternoon, night

isFraud - This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system.

- Financial transaction data is analyzed to detect fraud.
- Data is cleaned, missing values are handled, and exploratory data analysis (EDA) is performed.
- EDA uses visualizations like charts to identify patterns of fraud transactions, peak times of fraudulent activity, and top branches involved.

Click here to return to the previous slide

Fraud Transactions with respect to the location



Click here to return to the previous slide

Technical deep dive

Performance Metrics:

- **Under Sampling:**
  - Precision, recall, and F1 score for non-fraudulent transactions are 0.96, indicating high accuracy.
  - For fraudulent transactions, precision, recall, and F1 score are 0.94, showing good performance but with some room for improvement.
- **Over Sampling:**
  - Achieves perfect precision, recall, and F1 score of 1 for both non-fraudulent and fraudulent transactions, indicating excellent performance.

Model Evaluation:

- Under sampling shows high performance but with slightly lower scores for fraudulent transactions.
- Over sampling effectively addresses class imbalance, resulting in superior performance without any misclassifications.

Class Imbalance Handling:

- Over sampling is more effective in handling class imbalance, achieving perfect scores for both classes without sacrificing performance.

Practical Implications:

- The choice between under sampling and over sampling depends on factors like dataset size and desired performance.
- Over sampling may require more resources but offers better performance in this scenario.

Click here to return to the previous slide

Technical deep dive

Our model is based on the Random Forest Classifier, a robust machine learning algorithm.
Our dataset has imbalanced data, to overcome this problem we have approached with two resampling techniques,

- **Under Sampling:** Scaled down majority class labels ‘0’ using imblearn library, by this approach our model performed far better that the normal approach. This model gave an overall F1 accuracy of 94% with better precision and recall scores.

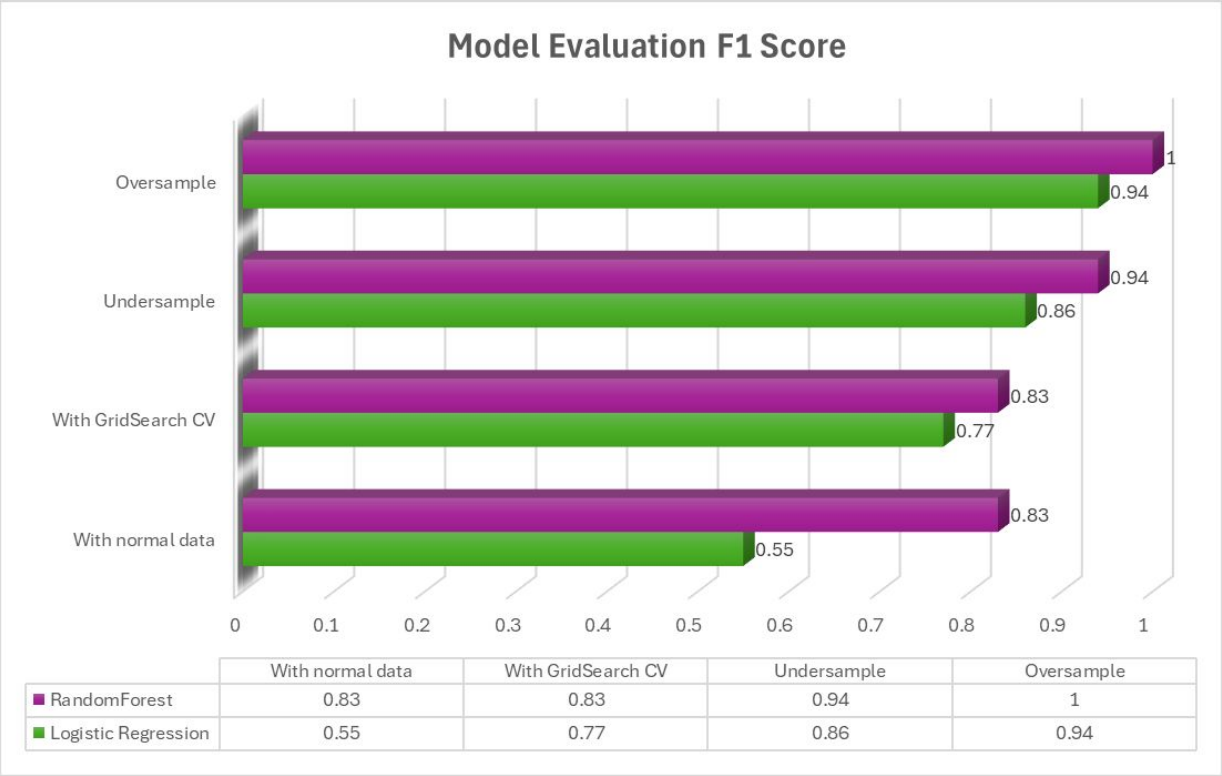
- **Oversampling:** Scaled up minority class label ‘1’ using imblearn random oversampling library, here the model seems to be overfitting which resulted in overall accuracy of 100% followed by precession and recall as 100%.

Random Forest Classifier Model Performance Matrix:

Under Sampling			
Class labels	**Precession**	**Recall**	**F1 Score**
0	0.96	0.96	0.96
1	0.94	0.94	0.94
Over Sampling			
0	1	1	1
1	1	1	1

Click here to return to the previous slide

Which model performed better?



	With normal data	With GridSearch CV	Undersample	Oversample
RandomForest	0.83	0.83	0.94	1
Logistic Regression	0.55	0.77	0.86	0.94

Click here to return to the previous slide

Model overview

Solution Overview:

- We employed machine learning techniques to build a predictive model capable of identifying fraudulent transactions based on transaction characteristics.
- The model was trained on historical transaction data, with a focus on accurately distinguishing between fraudulent and non-fraudulent transactions.
- Our fraud detection solution represents a proactive approach to safeguarding our financial transactions and maintaining trust with our customers.
- By leveraging advanced analytics and machine learning, we are well-equipped to detect and prevent fraudulent activities, ultimately safeguarding our business and ensuring the integrity of our financial transactions.

Click here to return to the beginning