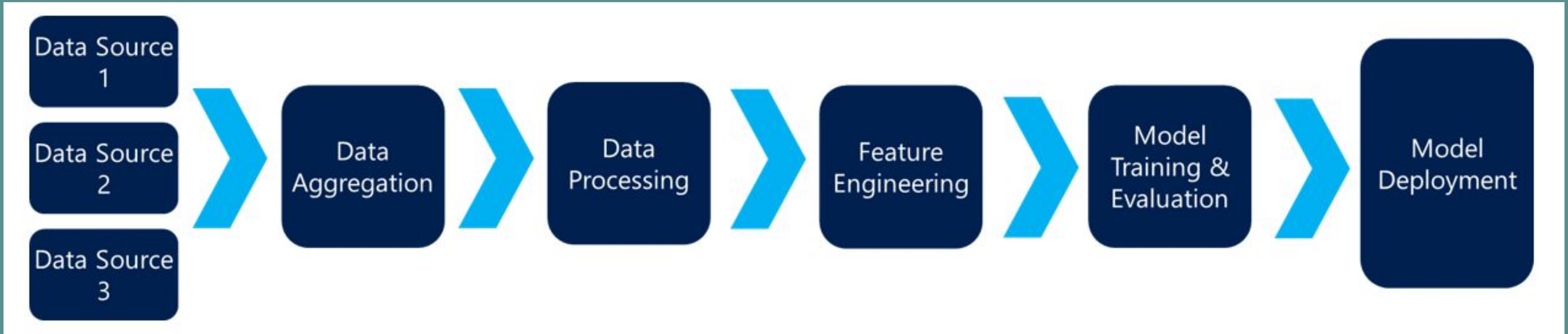
Online Transaction Fraud Prediction with Machine Learning

By: Samantha Lee

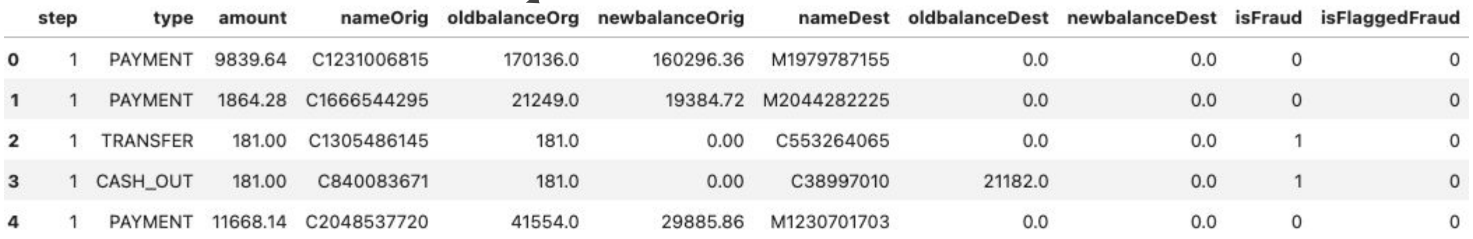
Introduction

- Financial fraud in online transaction poses significant risks to institutions and consumers.
- Detecting and preventing fraud is important for maintaining system integrity and preventing financial losses.
- Machine learning offers promising solutions by leveraging data-driven insights and predictive algorithms.
- Project aims to enhance fraud detection using machine learning techniques in financial transactions to identify the most effective approach for fraud detection.
- Machine Learning algorithms used Linear Regression, Random Forest, XGBoost and LightGBM.



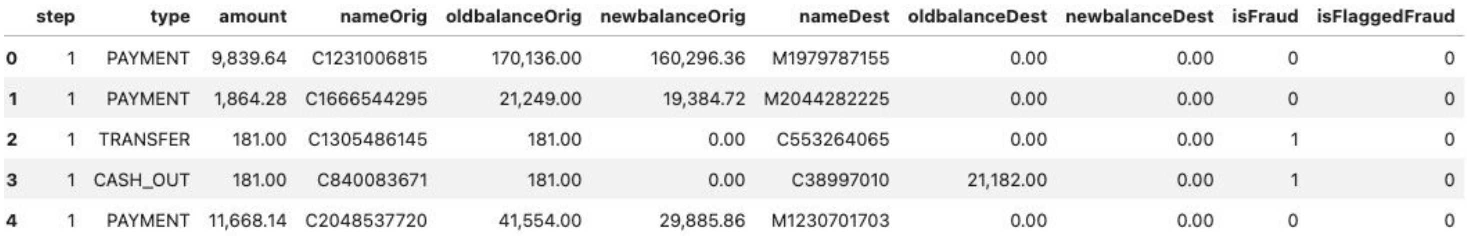
Data Pre-Processing

Before



	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1	0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0

After



	step	type	amount	nameOrig	oldbalanceOrig	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9,839.64	C1231006815	170,136.00	160,296.36	M1979787155	0.00	0.00	0	0
1	1	PAYMENT	1,864.28	C1666544295	21,249.00	19,384.72	M2044282225	0.00	0.00	0	0
2	1	TRANSFER	181.00	C1305486145	181.00	0.00	C553264065	0.00	0.00	1	0
3	1	CASH_OUT	181.00	C840083671	181.00	0.00	C38997010	21,182.00	0.00	1	0
4	1	PAYMENT	11,668.14	C2048537720	41,554.00	29,885.86	M1230701703	0.00	0.00	0	0

Summary Statistics:

- Offer insights into the distribution and characteristics of numeric features.
- Highlight key statistics such as range, mean, and standard deviation for each attribute.

Column Renaming:

- Changed the column 'oldbalanceOrg' to 'oldbalanceOrig' for consistency within the dataset.

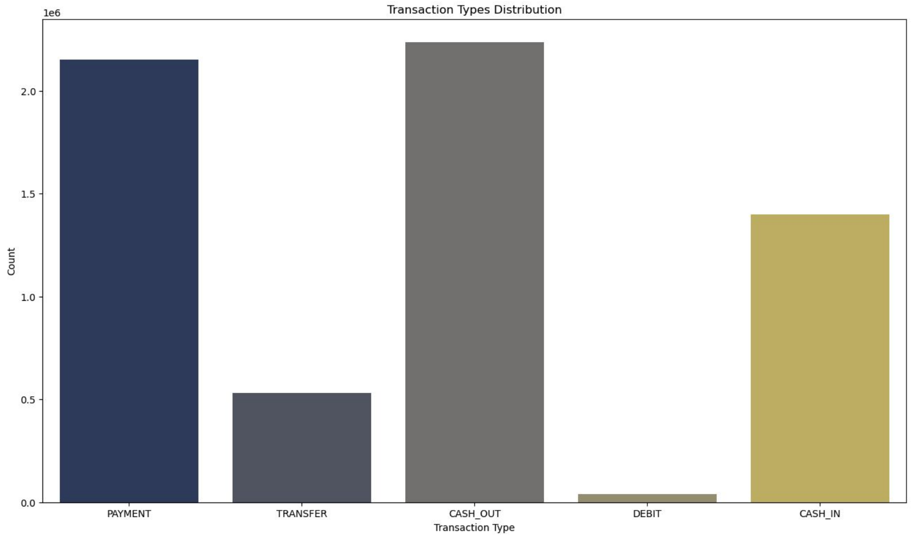
Dataset Composition:

- After renaming, the dataset now comprises columns like step, type, amount, nameOrig, oldbalanceOrig, newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest, isFraud, and isFlaggedFraud.

EDA: Transaction By Type

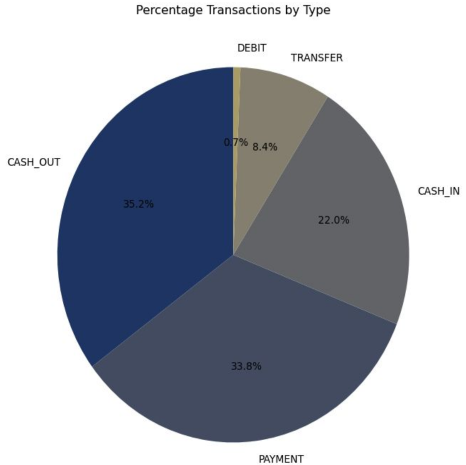
Visualization of Transaction Types Distribution:

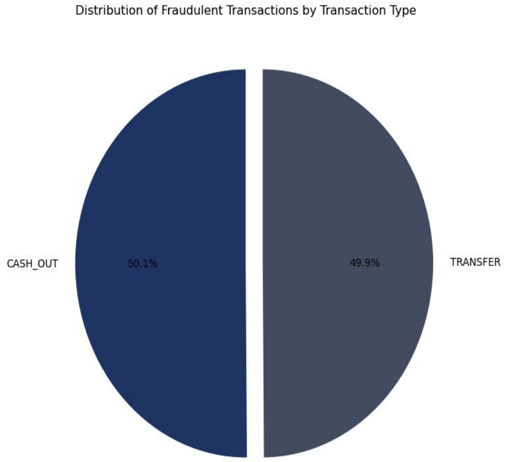
- Plotted a countplot to visualize the distribution of transaction types.
- Analyzed the transaction type in financial datasets.
- Plotted a countplot to visualize the distribution of transaction types using Seaborn.
- PAYMENT: 33.81%, CASH_OUT: 35.17%, CASH_IN: 21.99%, TRANSFER: 8.38%, DEBIT: 0.65%


Transaction Types Distribution

Distribution of Fraudulent Transaction by Type:

- "CASH_OUT" type constitutes 50.1% of the total.
- "TRANSFER" type accounts for 49.9%.
- This distribution suggests vulnerability to fraud activities, particularly in "CASH_OUT" and "TRANSFER" transaction types.


Percentage Transactions by Type


Distribution of Fraudulent Transactions by Transaction Type

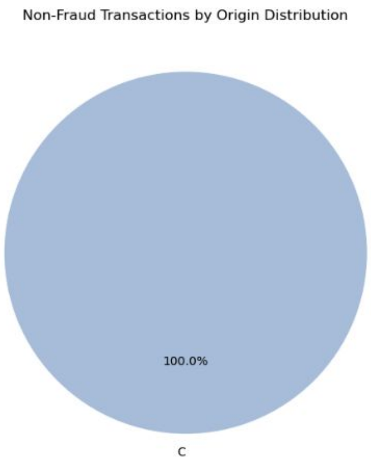
EDA: Non-fraud and Fraud by Origin & Destination

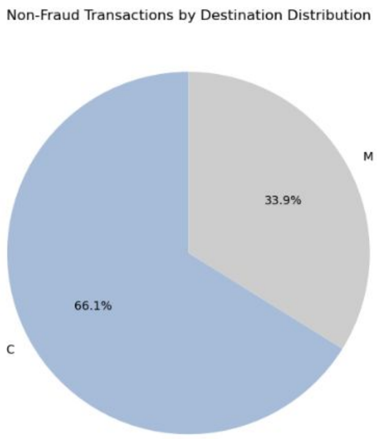
Non-Fraud Transactions by Origin and Destination Distribution:

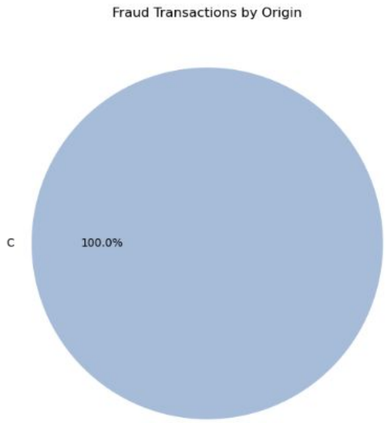
- Extracted first letter of transaction origin and destination, illustrating non-fraudulent transaction distribution via pie charts.
- Analyzed the percentages of non-fraudulent transactions originating from different categories.
- Found that 100% of non-fraudulent transactions originate from accounts classified as customers ('C').
- Highlighted that while 66.1% of non-fraudulent transactions are directed towards customer accounts, 33.9% involve merchant accounts.

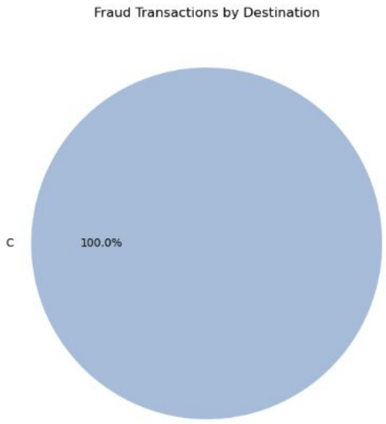
Fraudulent Transactions by Origin:

- Discovered that 100% of fraudulent transactions originate from customer accounts ('C').
- Highlighted the prevalence of fraudulent activities within or involving customer accounts.


Non-Fraud Transactions by Origin Distribution


Non-Fraud Transactions by Destination Distribution


Fraud Transactions by Origin


Fraud Transactions by Destination

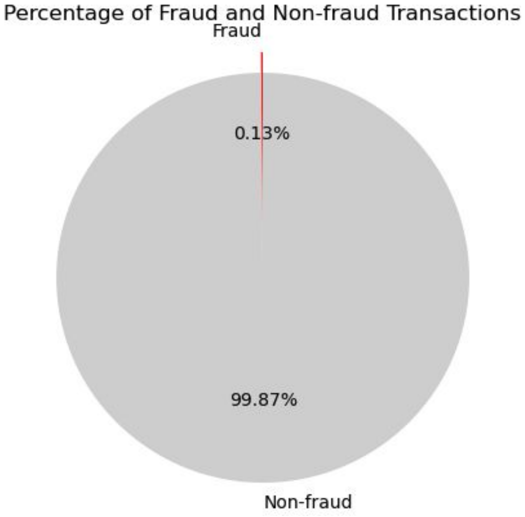
EDA

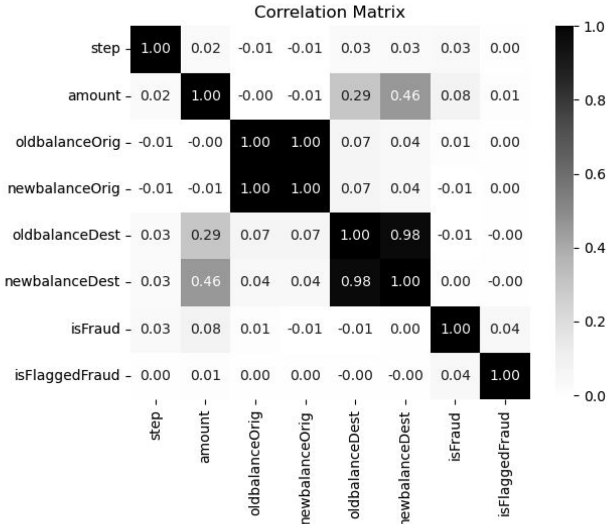
Percentage of Fraud and Non-Fraud Transactions:

- In the dataset, there are 8,213 fraud transactions and 6,354,407 non-fraud transactions.
- Fraud transactions represent 0.13% of the total, while non-fraud transactions constitute 99.87%.

Correlation Matrix:

- Observed a strong positive correlation of 0.98 between new balances and old balances in destination accounts.
- **Expected behavior:** Accounts with higher old balances tend to have higher new balances in subsequent transactions.
- Noted that the heatmap lacks strong correlations with fraud (isFraud).
- While individual features may not strongly predict fraud (correlation with isFraud of 0.01), exploring feature relationships can refine feature engineering and enhance fraud detection algorithms.





	step	amount	oldbalanceOrig	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
step	1.00	0.02	-0.01	-0.01	0.03	0.03	0.03	0.00
amount	0.02	1.00	-0.00	-0.01	0.29	0.46	0.08	0.01
oldbalanceOrig	-0.01	-0.00	1.00	1.00	0.07	0.04	0.01	0.00
newbalanceOrig	-0.01	-0.01	1.00	1.00	0.07	0.04	-0.01	0.00
oldbalanceDest	0.03	0.29	0.07	0.07	1.00	0.98	-0.01	-0.00
newbalanceDest	0.03	0.46	0.04	0.04	0.98	1.00	0.00	-0.00
isFraud	0.03	0.08	0.01	-0.01	-0.01	0.00	1.00	0.04
isFlaggedFraud	0.00	0.01	0.00	0.00	-0.00	-0.00	0.04	1.00

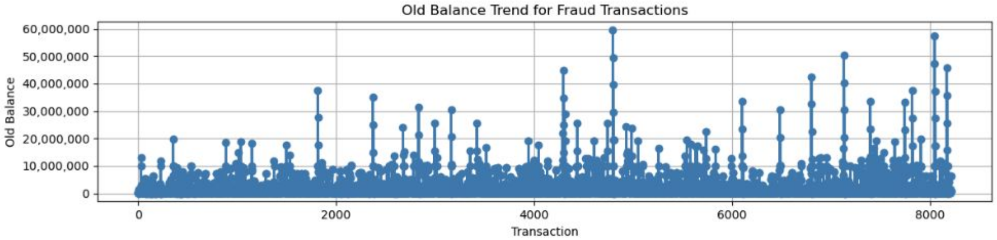
EDA: Old Balance & New Balance Trend

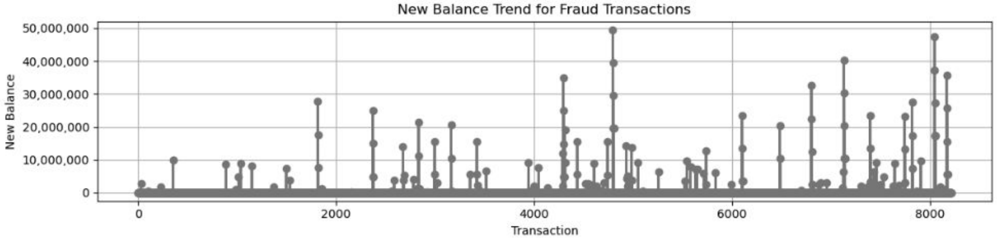
Fraudulent Transactions

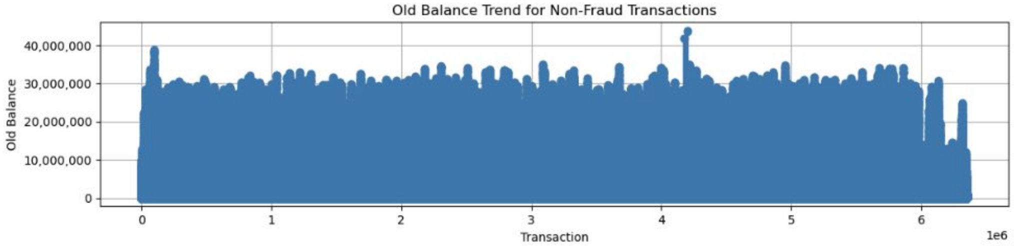
- **Old balance trend:** Initial balances typically range from 0 to 10,000,000.
- **New balance trend:** Many accounts with high initial balances drop to zero or significantly reduce post-transaction, indicating potential fraudulent activity.

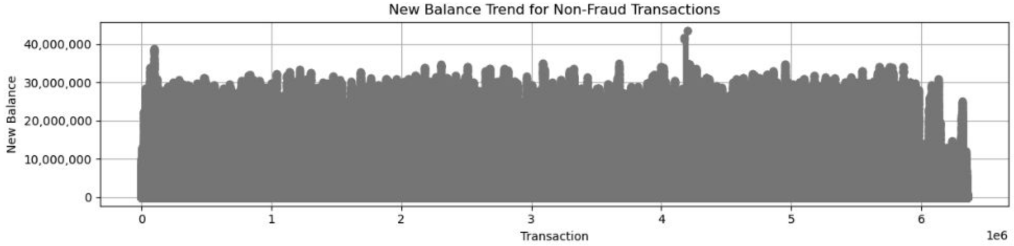
Non-Fraudulent Transactions

- **Old and new balance trends:** Show diverse patterns with varying balances and outcomes after transactions. Reflect normal financial activities within the dataset. Stays the same.









EDA

Monthly Fraudulent Transactions Histogram (Left):

- Stable frequency across months.
- No distinct monthly patterns.

Monthly Non-Fraud Transactions Histogram (Right):

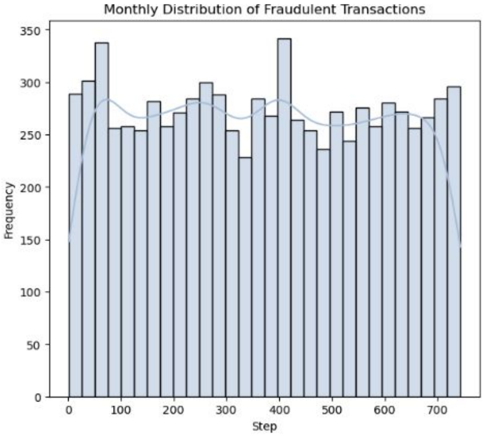
- Fluctuating frequency, indicating temporal patterns influenced by external factors.

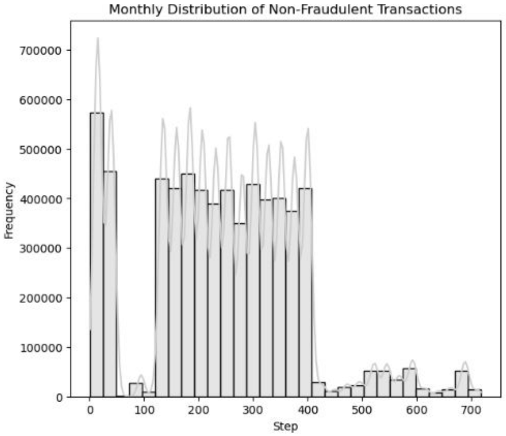
Histogram Fraudulent Transactions:

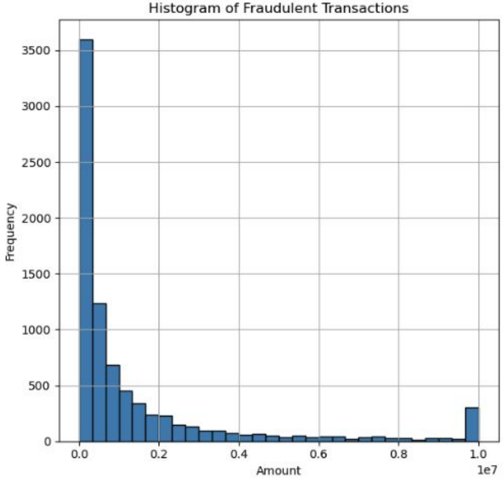
- Histogram shows varied transaction amounts for fraud, ranging from small to potentially large sums.
- Fraudulent transactions range from $181.00 to $6,311,409.28, indicating diverse fraudulent behaviors.

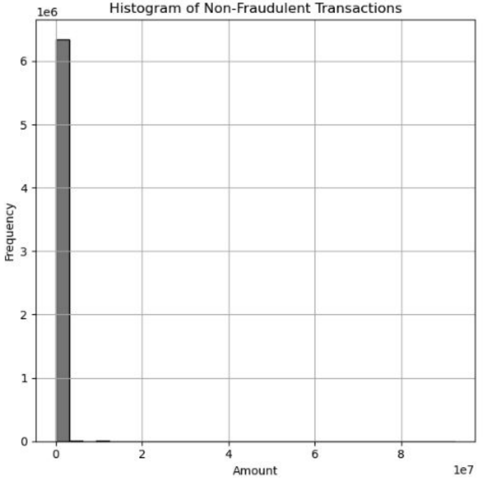
Histogram Non-Fraudulent Transactions:

- Histogram for non-fraudulent transactions displays consistent amounts, mostly falling within a single range.
- Non-fraudulent transactions span from $1,864.24 to $6,311,409.28, clustering mostly within lower to moderate amounts.
- Fraudulent transactions exhibit broader transaction amounts compared to non-fraudulent ones, aiding in fraud detection and analysis.









Data Processing (Dynamic Calculation):

Data Selection:

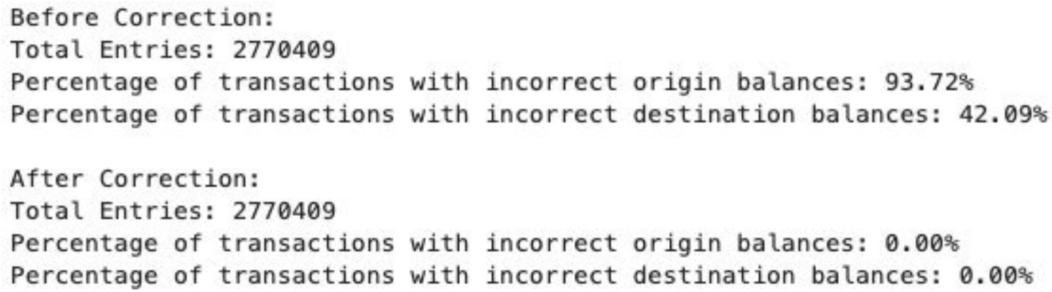
- Removed unnecessary columns like 'isFlaggedFraud', 'nameDest', and 'nameOrig' to streamline the dataset.
- Excluded transaction types such as 'CASH_IN', 'DEBIT', and 'PAYMENT' to focus on specific types.

Calculation Process

- Calculates percentages of transactions with incorrect origin and destination balances before any correction.
- Identifies and corrects transactions where old balances do not align with transaction amounts.
- Recalculates percentages of transactions with incorrect balances after correction.

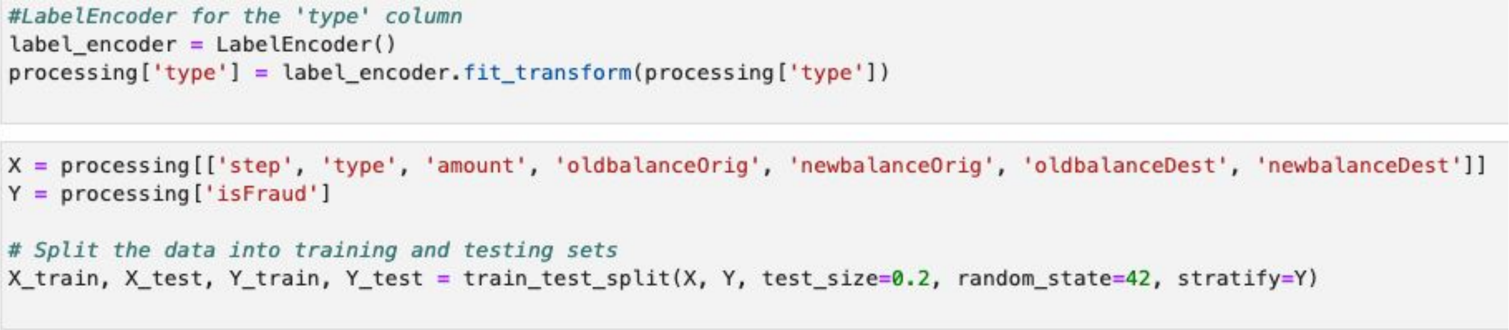
Data Correction:

- Before correction, a significant percentage of transactions exhibit inconsistencies in balances.
- After correction, percentages of transactions with incorrect balances reduce to zero.



```
Before Correction:
Total Entries: 2770409
Percentage of transactions with incorrect origin balances: 93.72%
Percentage of transactions with incorrect destination balances: 42.09%

After Correction:
Total Entries: 2770409
Percentage of transactions with incorrect origin balances: 0.00%
Percentage of transactions with incorrect destination balances: 0.00%
```

Feature Engineering



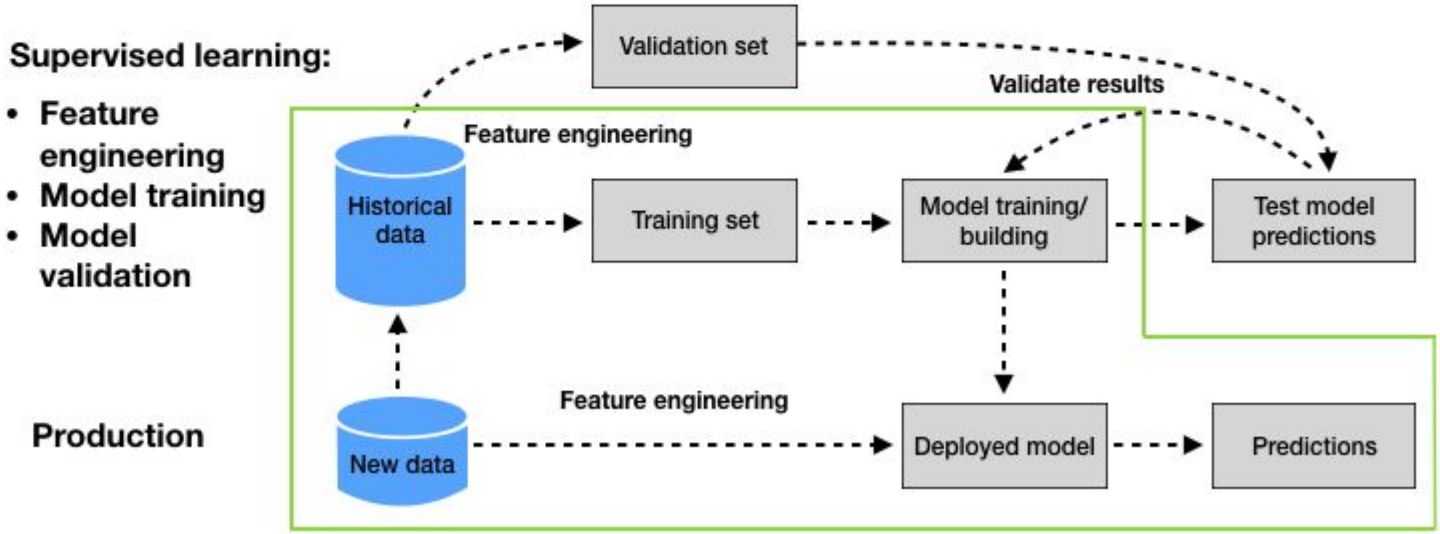
```
#LabelEncoder for the 'type' column
label_encoder = LabelEncoder()
processing['type'] = label_encoder.fit_transform(processing['type'])
```

```
X = processing[['step', 'type', 'amount', 'oldbalanceOrig', 'newbalanceOrig', 'oldbalanceDest', 'newbalanceDest']]
Y = processing['isFraud']

# Split the data into training and testing sets
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42, stratify=Y)
```

- Encoding the categorical feature 'type' into numerical values using LabelEncoder.
- Selecting relevant features like 'step', 'amount', 'oldbalanceOrig', 'newbalanceOrig', 'oldbalanceDest', and 'newbalanceDest'.
- Defining the target variable 'isFraud'.
- Split the dataset into 80% for training and 20% for testing to ensure the model learns from diverse transaction types and predicts fraud accurately.

Modeling

- In this project various algorithms were utilized including logistic regression, random forest, XGBoost, and LightGBM for the modeling process.
- Addressed class imbalance using SMOTE and undersampling techniques.
- Employed hyperparameter tuning with grid search and randomized search CV.



Logistic Regression

SMOTE Model:

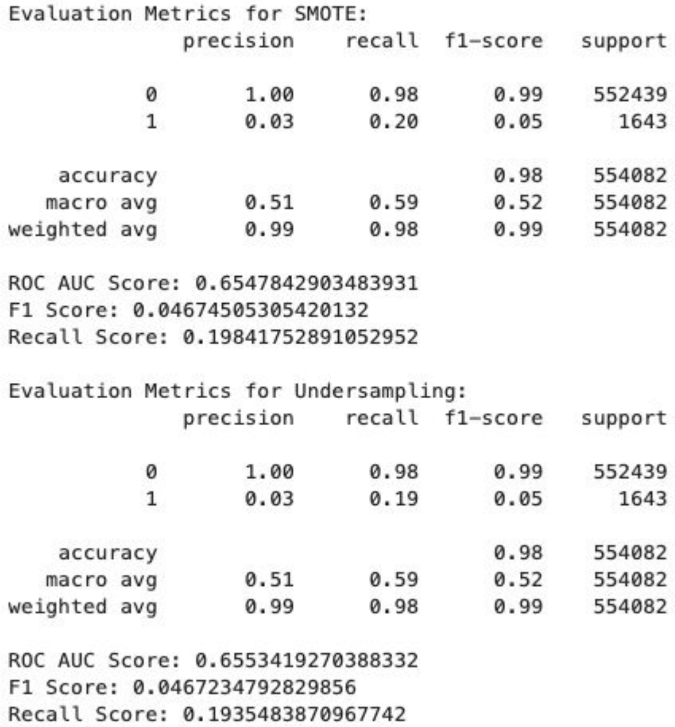
- Demonstrated a low precision of 3% with a recall of 20%.

Undersampling Model:

- Slightly better recall than SMOTE
- Still struggles to identify fraudulent transactions effectively.

Performance Metrics:

- Both logistic regression models faced challenges.
- Comparable accuracy, F1 score, and macro-average metrics.
- Highlighting complexity in addressing class imbalance.
- Needs more advanced techniques in fraud detection.



```
Evaluation Metrics for SMOTE:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99    552439
           1       0.03      0.20      0.05      1643

    accuracy                           0.98    554082
   macro avg       0.51      0.59      0.52    554082
weighted avg       0.99      0.98      0.99    554082

ROC AUC Score: 0.6547842903483931
F1 Score: 0.04674505305420132
Recall Score: 0.19841752891052952

Evaluation Metrics for Undersampling:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99    552439
           1       0.03      0.19      0.05      1643

    accuracy                           0.98    554082
   macro avg       0.51      0.59      0.52    554082
weighted avg       0.99      0.98      0.99    554082

ROC AUC Score: 0.6553419270388332
F1 Score: 0.0467234792829856
Recall Score: 0.1935483870967742
```

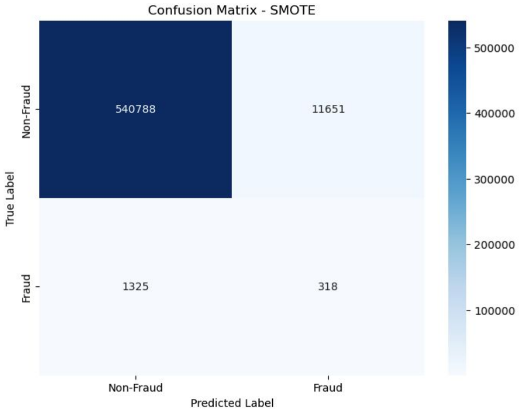
Logistic Regression: Confusion Matrix & ROC Curve

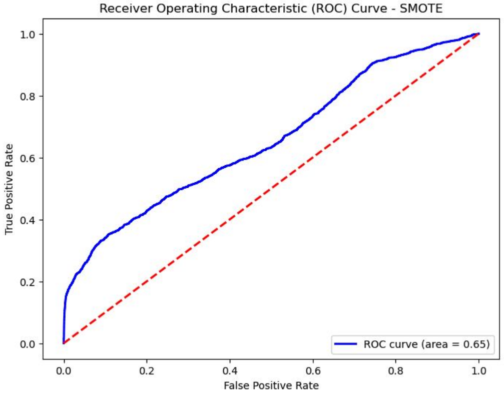
SMOTE Oversampling:

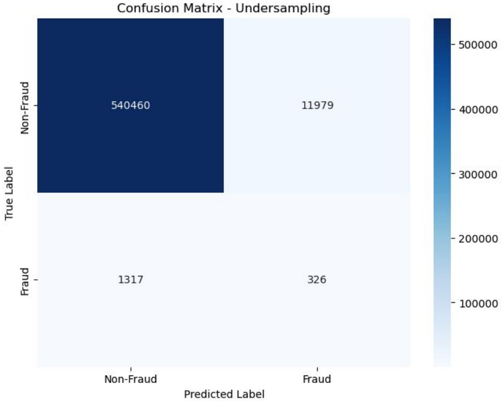
- Achieved 540,788 true negatives, 11,651 false positives, 1,325 false negatives, and 318 true positives.
- Resulted in an AUC of 0.654.

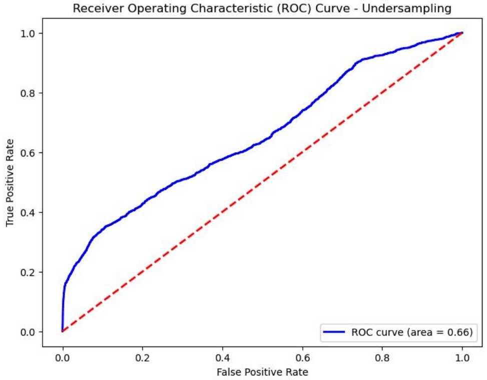
Undersampling:

- Achieved 540,460 true negatives, 11,979 false positives, 1,317 false negatives, and 326 true positives.
- Demonstrated a slightly higher AUC of 0.655.
- These findings suggest that both models exhibit similar performance, with Undersampling showing a slight edge in the area under the curve.


Confusion Matrix - SMOTE


Receiver Operating Characteristic (ROC) Curve - SMOTE


Confusion Matrix - Undersampling


Receiver Operating Characteristic (ROC) Curve - Undersampling

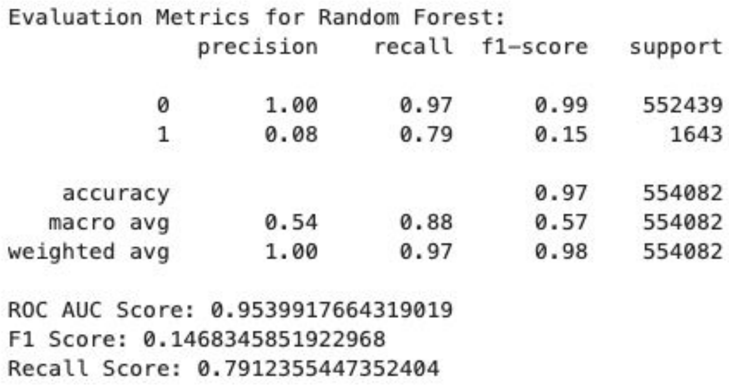
Random Forest

Random Forest Model Evaluation:

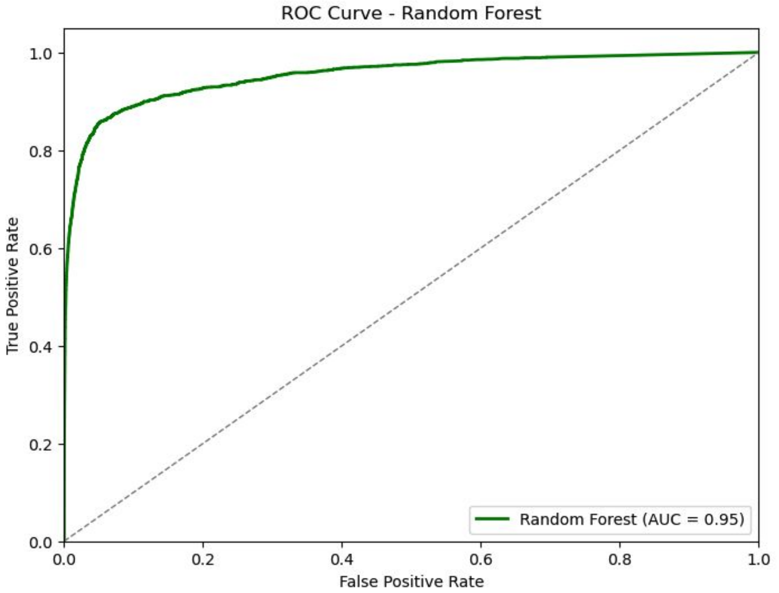
- The Random Forest model achieved a 79% recall rate and an F1 score of 0.15, outperforming logistic regression in fraud detection.

Overall Performance:

- **Accuracy:** 97%, showcasing the model's effectiveness in classifying transactions.
- **ROC AUC Score:** 95%, demonstrating the model's capability to distinguish between fraudulent and non-fraudulent transactions.



```
Evaluation Metrics for Random Forest:
              precision    recall  f1-score   support

           0       1.00      0.97      0.99    552439
           1       0.08      0.79      0.15      1643

    accuracy                           0.97    554082
   macro avg       0.54      0.88      0.57    554082
weighted avg       1.00      0.97      0.98    554082

ROC AUC Score: 0.9539917664319019
F1 Score: 0.1468345851922968
Recall Score: 0.7912355447352404
```



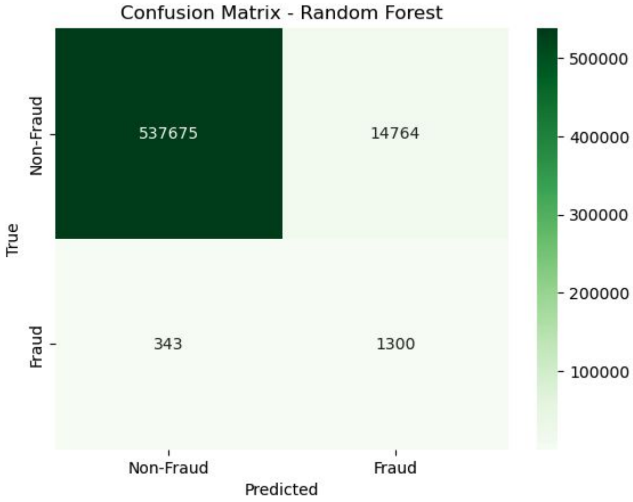
Random Forest Metrics

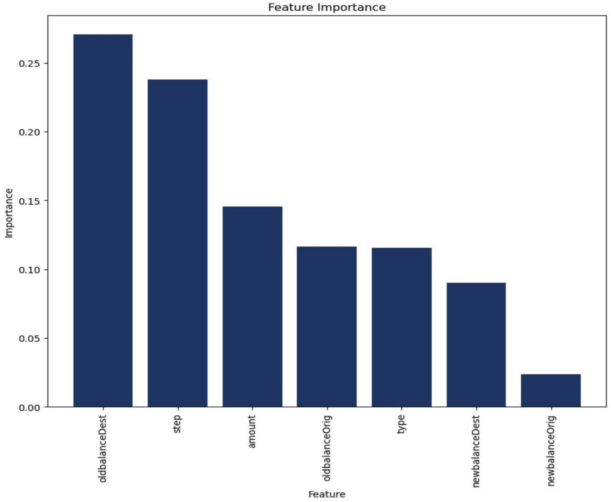
Confusion Matrix:

- Effectively distinguishes between fraudulent and non-fraudulent transactions.
- Identifies 537,675 non-fraudulent and 1,300 fraudulent transactions.

Key features:

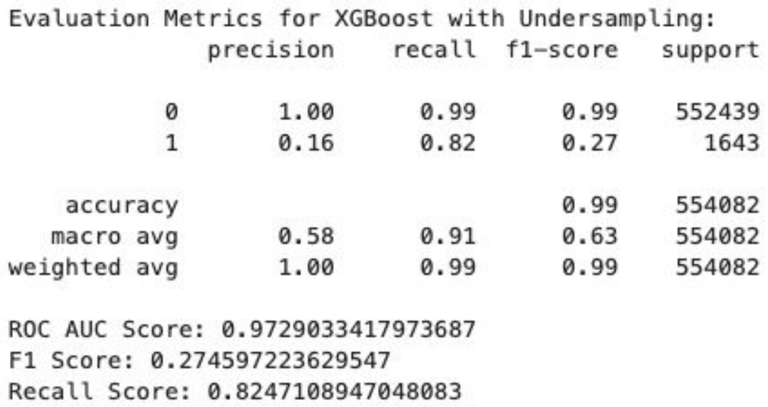
- oldbalanceDest, step, and transaction amount.
- Highlights 14,764 false positives and 343 false negatives.
- Attributes significant importance to features.
- Robust approach to fraud detection, integrating various transaction attributes for accuracy.





XGBoost

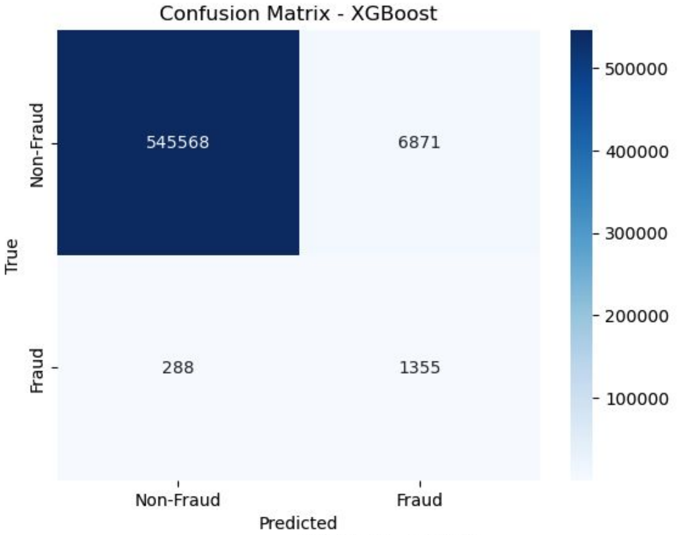
- **XGBoost Performance:**
- Highest recall of 82%, indicating its ability to capture a large portion of actual positive instances.
- F1 score of 0.27, representing a balanced performance between precision and recall.
- Accuracy of 0.99, reflecting its overall correctness in predictions.

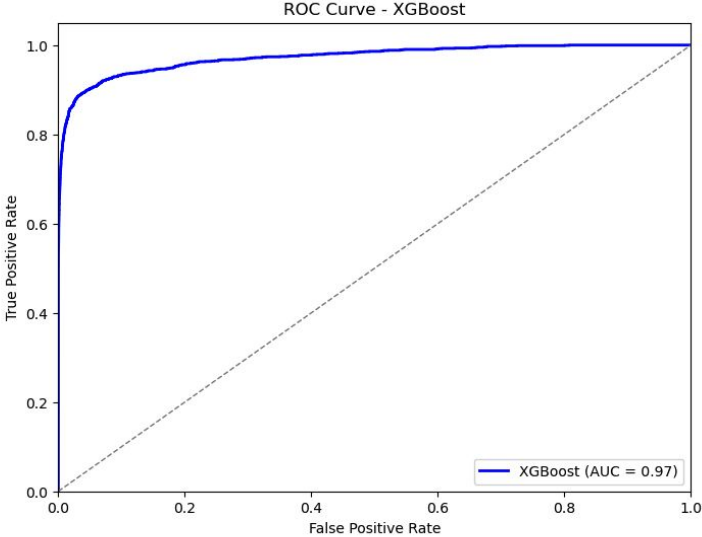


```
Evaluation Metrics for XGBoost with Undersampling:
              precision    recall  f1-score   support

           0       1.00      0.99      0.99    552439
           1       0.16      0.82      0.27      1643

    accuracy                           0.99    554082
   macro avg       0.58      0.91      0.63    554082
weighted avg       1.00      0.99      0.99    554082

ROC AUC Score: 0.9729033417973687
F1 Score: 0.274597223629547
Recall Score: 0.8247108947048083
```

XGBoost Metrics

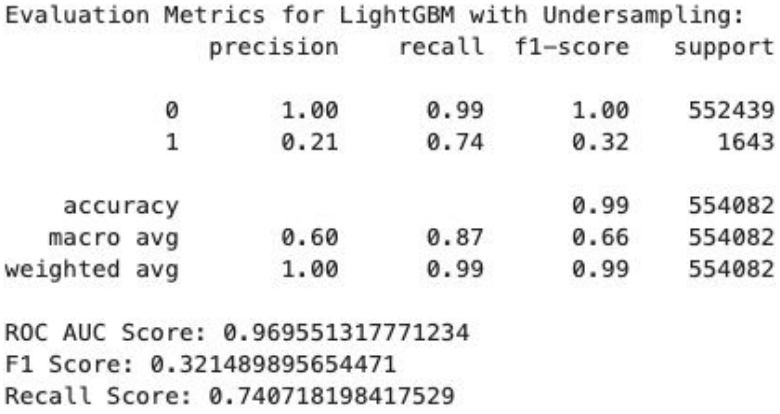
- **Confusion Matrix:** Revealed few false negatives (288), indicating effective fraud detection.
- **ROC AUC Score:** XGBoost achieved a high ROC AUC of 0.97, showing excellent performance in classifying fraudulent and non-fraudulent transactions.
- **Model Comparison:** XGBoost accuracy and ability to handle complex data structures position it as the preferred model for fraud detection tasks, surpassing both Logistic Regression and Random Forest.





LightGBM

LightGBM Metrics: LightGBM achieved a highest F1 score of 0.32 and a recall score of 74%, indicating its effectiveness in accurately identifying fraudulent transactions.



```
Evaluation Metrics for LightGBM with Undersampling:
              precision    recall  f1-score   support

           0       1.00      0.99      1.00    552439
           1       0.21      0.74      0.32      1643

    accuracy                           0.99    554082
   macro avg       0.60      0.87      0.66    554082
weighted avg       1.00      0.99      0.99    554082

ROC AUC Score: 0.969551317771234
F1 Score: 0.321489895654471
Recall Score: 0.740718198417529
```

LightGBM Metrics

Confusion Matrix:

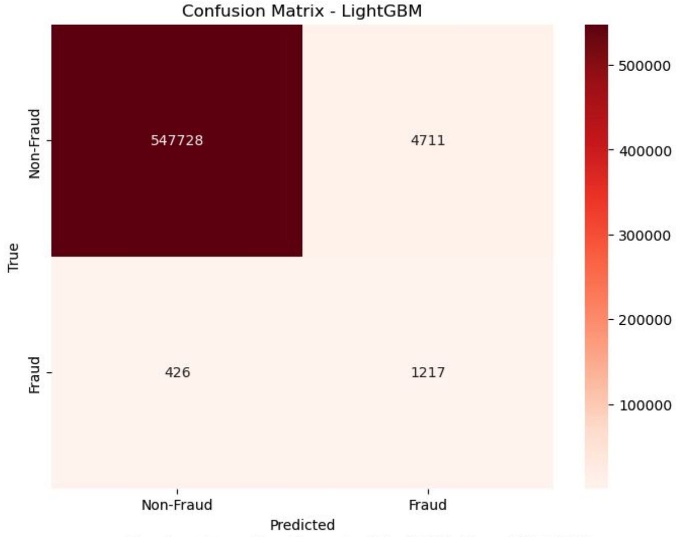
- LightGBM had 4,711 false positives, much lower than XGBoost's 6,871 false positives.
- Indicates better performance in avoiding misclassification of non-fraudulent transactions.

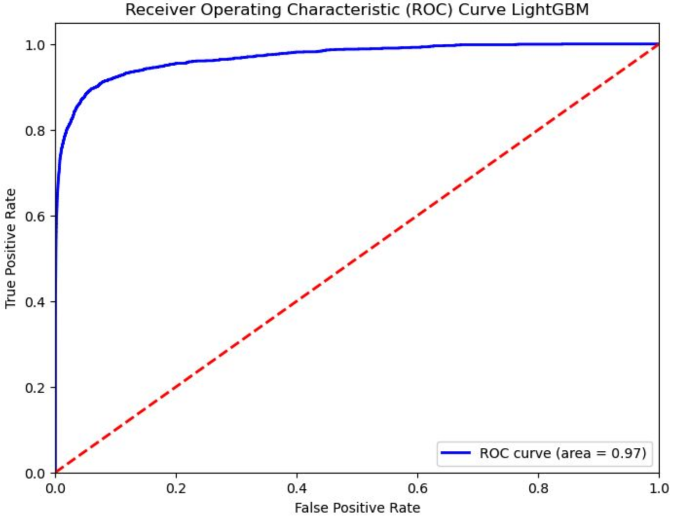
ROC AUC Score:

- LightGBM achieved an impressive ROC AUC score of 0.97, showcasing its strong capability in effectively distinguishing between fraudulent and non-fraudulent transactions.

Model Comparison:

- While both LightGBM and XGBoost show strengths and weaknesses in different aspects of fraud detection, LightGBM's lower false positive rate makes it a favorable choice.


Confusion Matrix - LightGBM


Receiver Operating Characteristic (ROC) Curve LightGBM

Conclusion

- The project aimed to address fraud detection in financial transactions using machine learning models, including Logistic Regression, Random Forest, XGBoost, and LightGBM. Each model underwent evaluation based on metrics like precision, recall, F1-score, and ROC AUC score.

- **Logistic Regression** provided a baseline, but it struggled with feature relationships and imbalanced datasets.

- **Random Forest,** trained on undersampled data, showed a high recall score but also had increased false positives.

- **XGBoost** exhibited efficiency in identifying fraudulent transactions, with a low false negative rate, signifying its effectiveness in capturing actual fraud cases.

- **LightGBM** surpassed XGBoost with higher F1 and precision score, indicating its potential for accurate fraud detection.

- Future directions involve exploring ensemble methods, advanced feature engineering, hyperparameter tuning, anomaly detection techniques, real-time monitoring, and improving model interpretability. These strategies can enhance fraud detection systems, adapting them to dynamic environments and ensuring continuous accuracy and efficiency in financial transaction security.