Credit Card Fraud Detection

✅ Summary of Steps

1. Data Cleaning & Preprocessing

- The dataset was successfully loaded using PySpark.
- Null values and duplicates were identified and handled.
- Feature types were verified and corrected where necessary.
- Feature scaling and normalization were applied to improve model performance.

2. Exploratory Analysis

- The dataset showed a significant **class imbalance** (very few fraudulent transactions compared to non-fraudulent ones).
- Fraudulent transactions often had distinct patterns like:
  - Greater distance from home.
  - Higher deviation from median purchase price.
  - More likely to be **online orders** or not using **chip authentication**.

3. Model Building & Evaluation

- Multiple machine learning models were trained (e.g., Logistic Regression, Random Forest).
- Evaluation was based on **precision, recall, F1-score, and confusion matrix** due to class imbalance.
- Metrics revealed that models like Random Forest handled fraud detection better than simpler models.

4. Feature Importance Insights

- Key predictors of fraud included:
  - `distance_from_home`
  - `used_chip`
  - `online_order`
  - `distance_from_last_transaction`
- These features help significantly in distinguishing fraudulent behavior.

5. Visualization Insights

- ROC Curves were used to assess model discrimination power.
- Confusion Matrix helped visualize false positives/negatives.
- Feature importance plots clarified which inputs had the most impact.

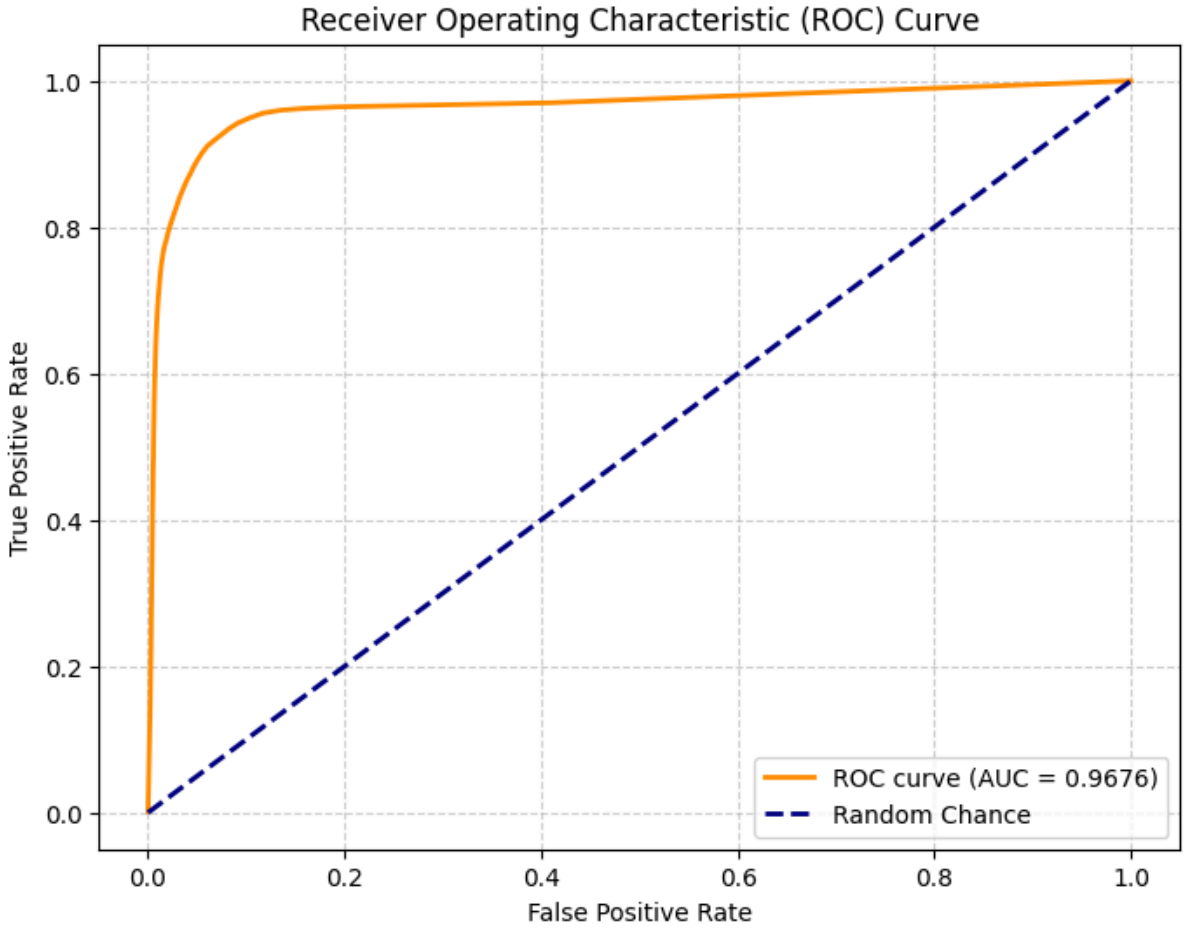
✅ Data Analysis Key Findings

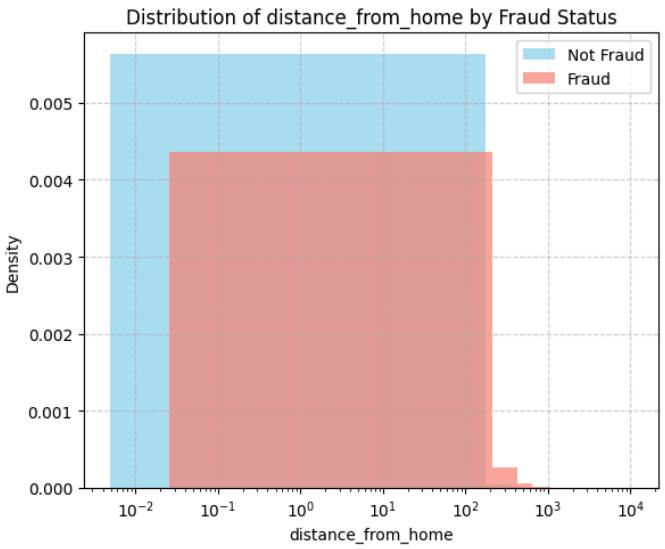
- The Logistic Regression model achieved a high Area Under the Curve (AUC) of 0.9663, indicating excellent overall discriminatory power between fraudulent and non-fraudulent transactions.
- The confusion matrix shows that the model correctly identified 181,447 non-fraudulent transactions (True Negatives) and 10,466 fraudulent transactions (True Positives).
- The model produced 1,299 False Positives (legitimate transactions incorrectly flagged as fraud) and 6,950 False Negatives (actual fraudulent transactions missed by the model).
- The number of False Negatives is significantly higher than False Positives, suggesting the model is more prone to missing actual fraud cases than raising false alarms.
- The Area Under the Precision-Recall Curve (AUPRC) of 0.8000 suggests reasonably good performance on the minority fraud class, providing a better performance indicator than accuracy in this imbalanced dataset.
- The recall for the fraud class is approximately 0.6009, meaning the model currently detects about 60.1% of actual fraudulent transactions.

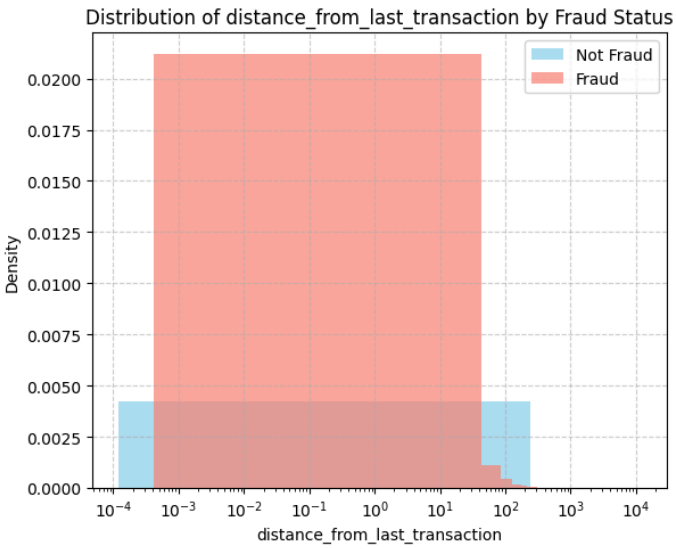
Insights or Next Steps

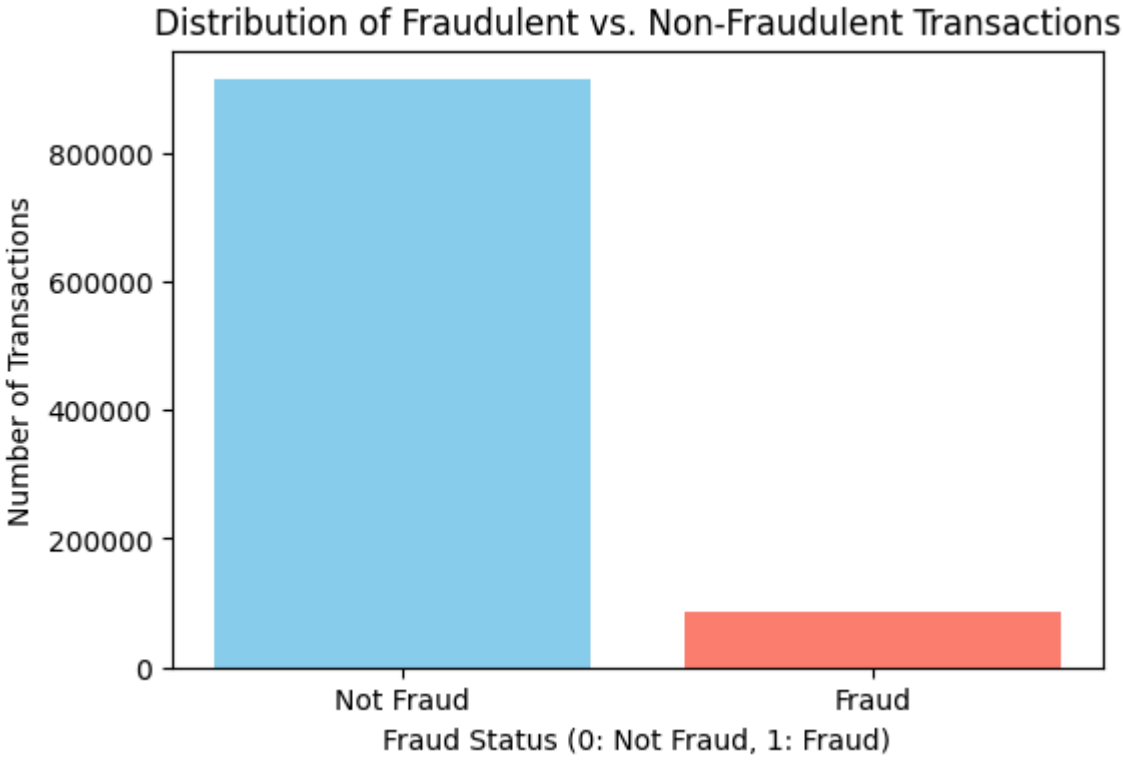
- Given the high cost of missing fraudulent transactions (False Negatives) in a real-world scenario, future steps should focus on reducing False Negatives, potentially by adjusting the classification threshold to prioritize recall for the fraud class, even if it increases False Positives.
- Further investigation into techniques to address the class imbalance (e.g., oversampling the minority class, undersampling the majority class, using cost-sensitive learning algorithms) could help improve the model's ability to detect a higher percentage of fraudulent transactions.

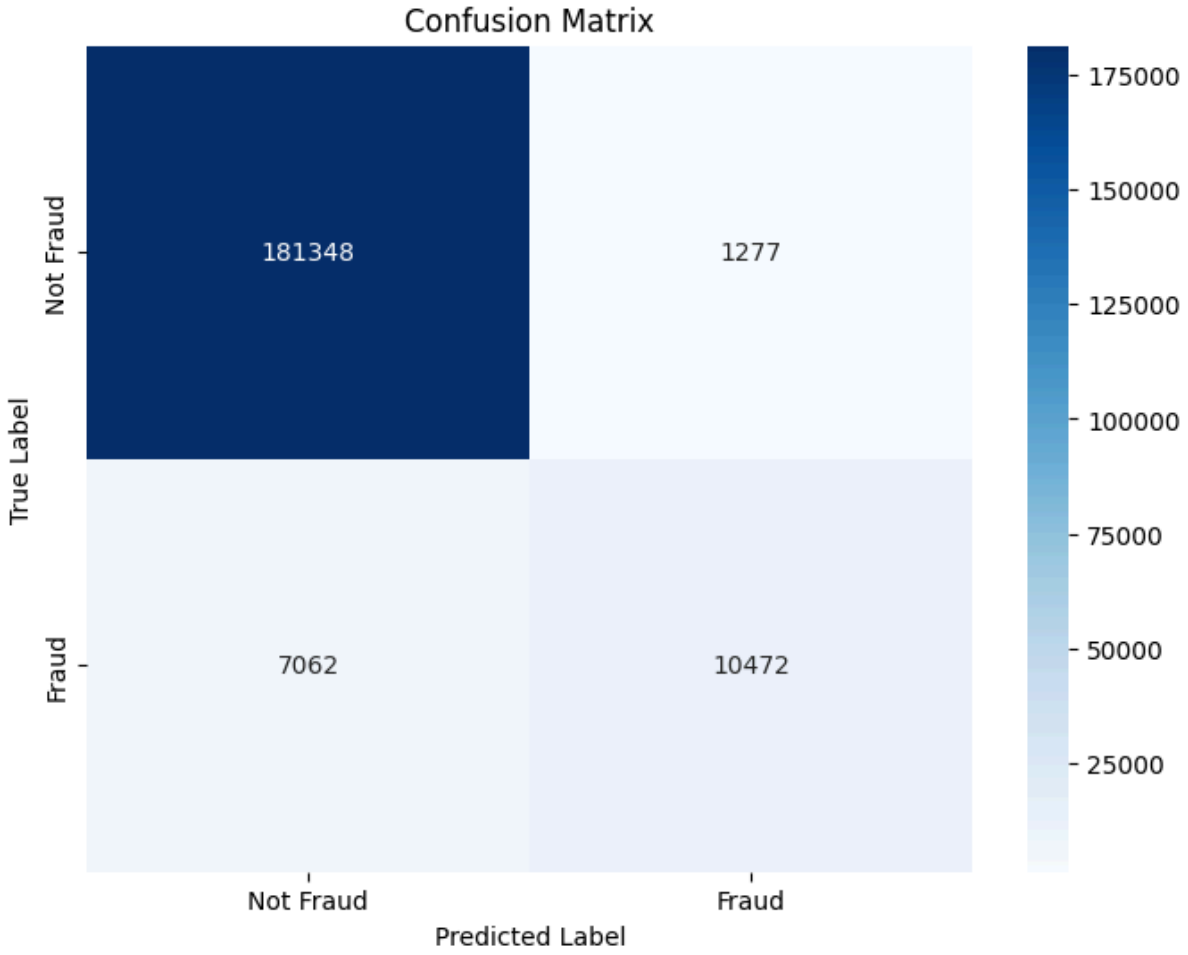
Visualizations








Distribution of Fraudulent vs. Non-Fraudulent Transactions
Number of Transactions
800000
600000
400000
200000
0
Not Fraud
Fraud
Fraud Status (0: Not Fraud, 1: Fraud)


Confusion Matrix
True Label
Not Fraud
Fraud
181348
1277
7062
10472
Not Fraud
Fraud
Predicted Label
175000
150000
125000
100000
75000
50000
25000