Credit Card Fraud Transactions

By: Raka P. N.

Project Workflow

1. Business Understanding
2. Data Understanding
3. Data Exploration
4. Data Visualization
5. Data Insight

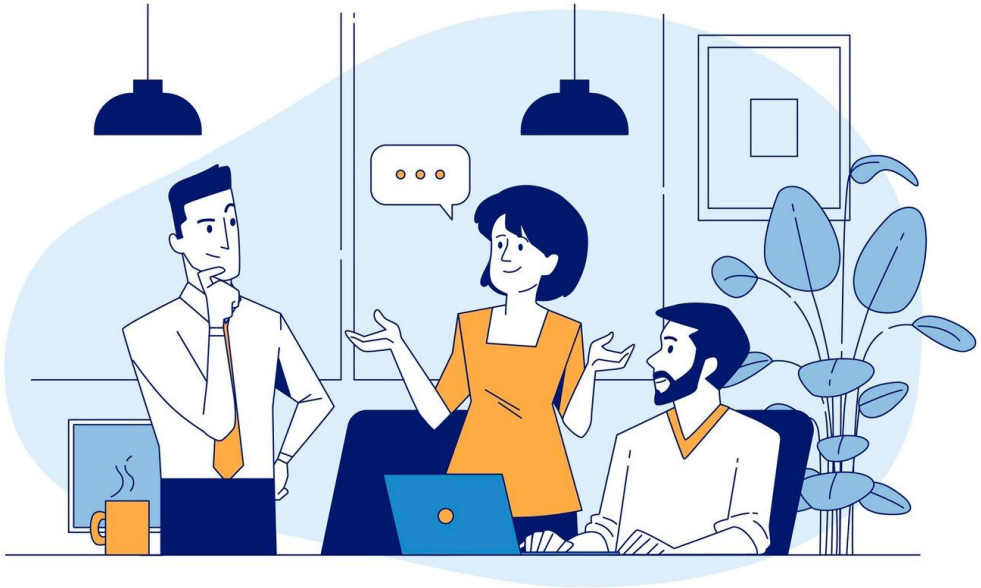
Business Understanding

Company Background

Company ABC, faces challenges with their existing fraud detection system. The current system exhibits slow responsiveness in recognizing new patterns of fraud, leading to significant financial losses. To address this issue, they have contracted us to design and implement an algorithm that can efficiently identify and flag potentially fraudulent transactions for further investigation.

Problem Statement

1. What are the key indicators used to determine whether a transaction is fraudulent or not?
2. In which regions do fraudulent transactions occur?
3. How accurate and efficient is the model in distinguishing fraudulent and non-fraudulent transactions, and what metrics are used to evaluate its performance?



Data Understanding

Origin of Data

The data used comes from ABC company in .csv format, consisting of two main files:

1. cc_info.csv, which contains general information about credit cards and cardholders.
2. transactions.csv, which records details of credit card transactions from August 1 to October 30.

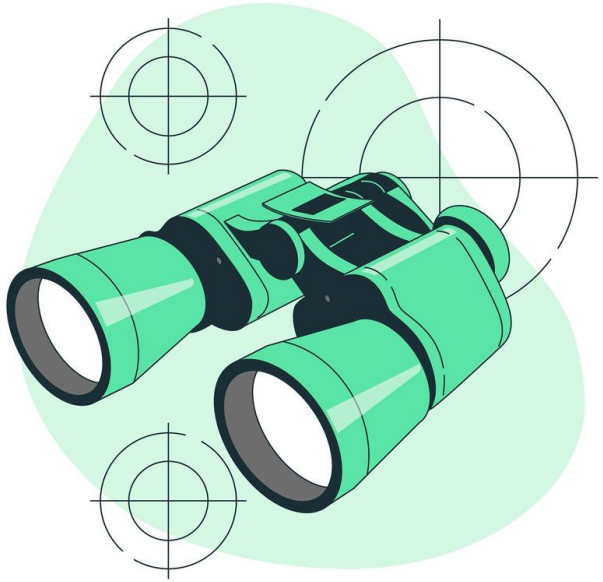
cc_info.csv

Total Row: 984, Column: 5

No	Column Name	Description	Data Type
1	credit_card	List of Credit Card ID	Integer
2	city	Credit Card Town Location	String
3	state	Credit Card State Location	String
4	zip code	Credit Card Zip Code Location	Integer
5	credit_card_limit	Credit Card Limit Transactions	Integer

transactions.csv

Total Row: 294588, Column: 5



No	Column Name	Description	Data Type
1	credit_card	List of Credit Card ID	Integer
2	datetime	Transactions Timestamp	Date
3	transaction_dollar_amount	Total Transactions	Float
4	long	Geography Longitude Location	Float
5	lat	Geography Latitude Location	Float

Data Exploration

Working Project Tools

Packages Used

pandas

NumPy

statsmodels

seaborn

scikit learn

matplotlib

SciPy

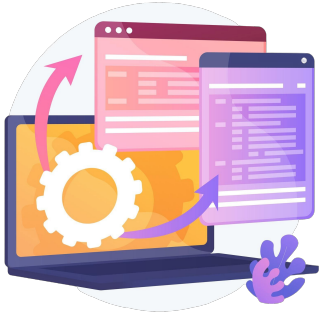
GeoPy

Programming Language

Visualization

Data Workflow

Origin / Raw Data



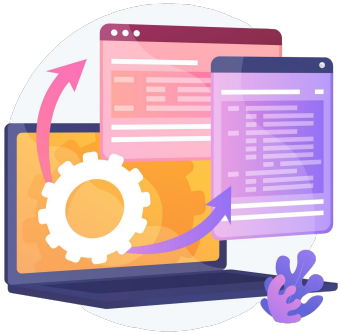
Processing Data

Data Mart

Data Modeling

Data Processing

Data Processing



1. Combine Dataset
2. Null Checking
3. Duplicate Checking
4. Data Enrichment

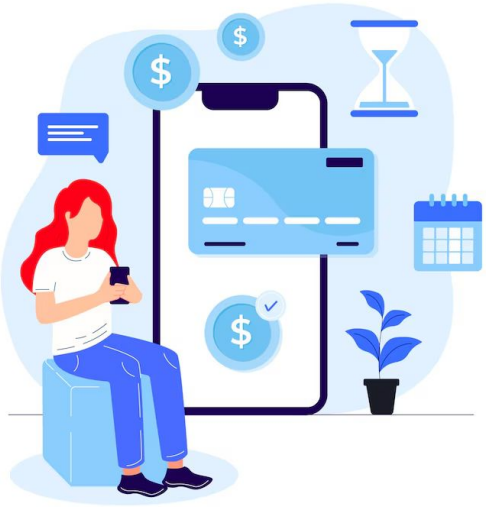
5. Data Filtering
6. Data Structuring
7. Outlier Checking
8. Split Data:
   1. Datamart
   2. Modeling

Data Mart

Credit Card Info

No	Column Name	Description	Data Type
1	credit_card	List of Credit Card ID	Integer
2	city	Credit Card Town Location	String
3	state	Credit Card State Location	String
4	zip code	Credit Card Zip Code Location	Integer
5	credit_card_limit	Credit Card Limit Transactions	Integer

Credit Card Transactions



No	Column Name	Description	Data Type
1	credit_card	List of Credit Card ID	Integer
2	date	Transactions Timestamp	Date
3	transactions_dollar_amount	Total Transactions	Float
4	Long	Geography Longitude Location	Float
5	Latitude	Geography Latitude Location	Float

Data Schema

No	Column Name	Description	Data Type
1	credit_card	List of Credit Card ID	Integer
2	date	Transactions Timestamp	Date
3	transactions_dollar_amount	Total Transactions	Float
4	Long	Geography Longitude Location	Float
5	Latitude	Geography Latitude Location	Float

No	Column Name	Description	Data Type
1	credit_card	List of Credit Card ID	Integer
2	city	Credit Card Town Location	String
3	state	Credit Card State Location	String
4	zip code	Credit Card Zip Code Location	Integer
5	credit_card_limit	Credit Card Limit Transactions	Integer

Data Modeling

Data Modeling

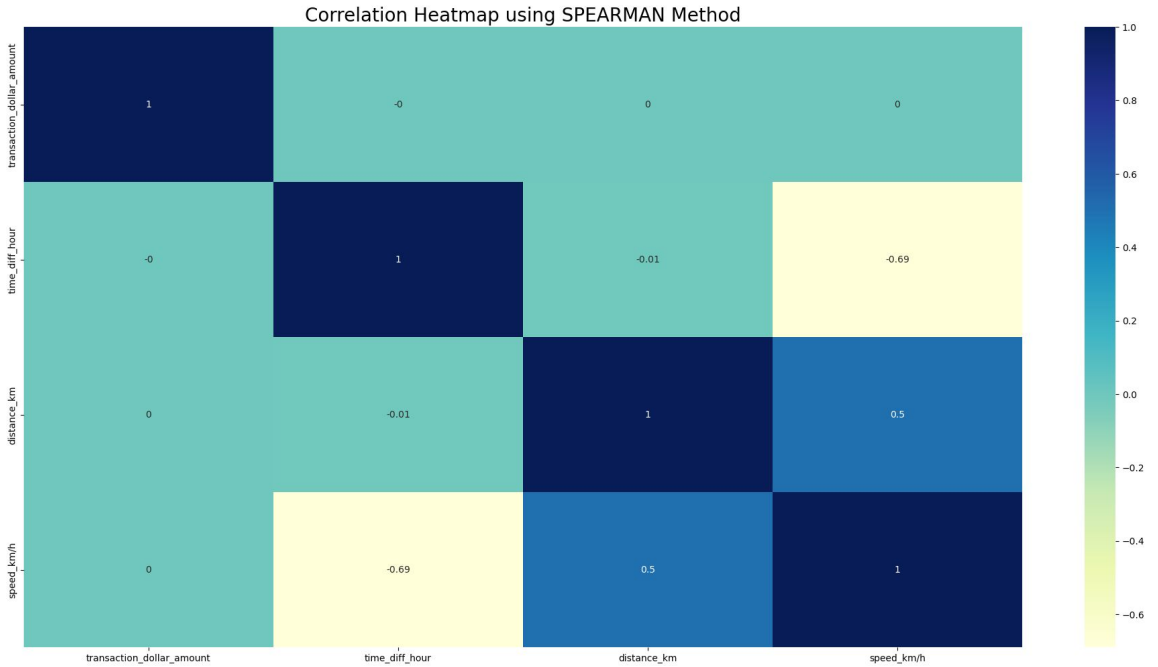
1. Check Imbalanced Data
2. Check Noise Data
3. Split Data
4. Data Resampling

5. Check Data Leak
6. Pipeline Blueprint
7. Model Tuning
8. Model Visualization

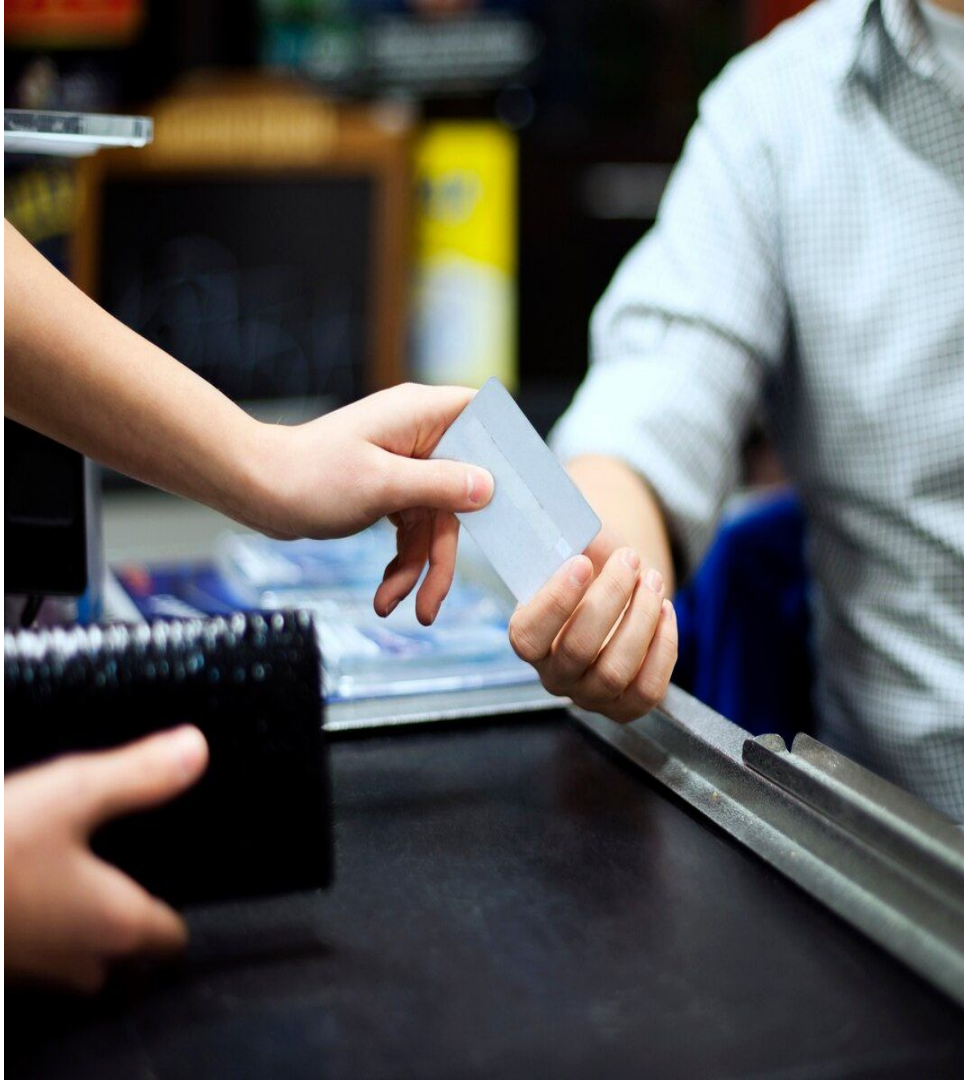
Data Overview

The correlation analysis reveals a strong negative relationship between time_diff_hour and speed_km/h tends to decrease. This could indicate that longer transaction gaps are associated with slower movements, possibly due to waiting times or extended idle periods.

There is a moderate positive correlation between distance_km and speed_km/h, meaning that longer travel distances are generally associated with higher average speeds, which could reflect more direct or faster routes for longer journeys.



Data Visualization



Fraud Key Indicators

Fraud Key Indicators

A transaction is considered suspicious if its **value exceeds 80%** of the credit card limit, or **5 times greater than the average** user transaction. In addition, the **location** of the transaction is also a risk factor, especially if it occurs in a high-risk zone. Some other factors that are taken into account are:

- **High transaction frequency**: **If two transactions occur in less than an hour**, this could be an indication of unusual activity.
- **Unusual travel speed**: If the movement of the transaction location **shows a speed of more than 50 km/h**, the transaction has the potential to be fraudulent.

Regions Often Do Fraud

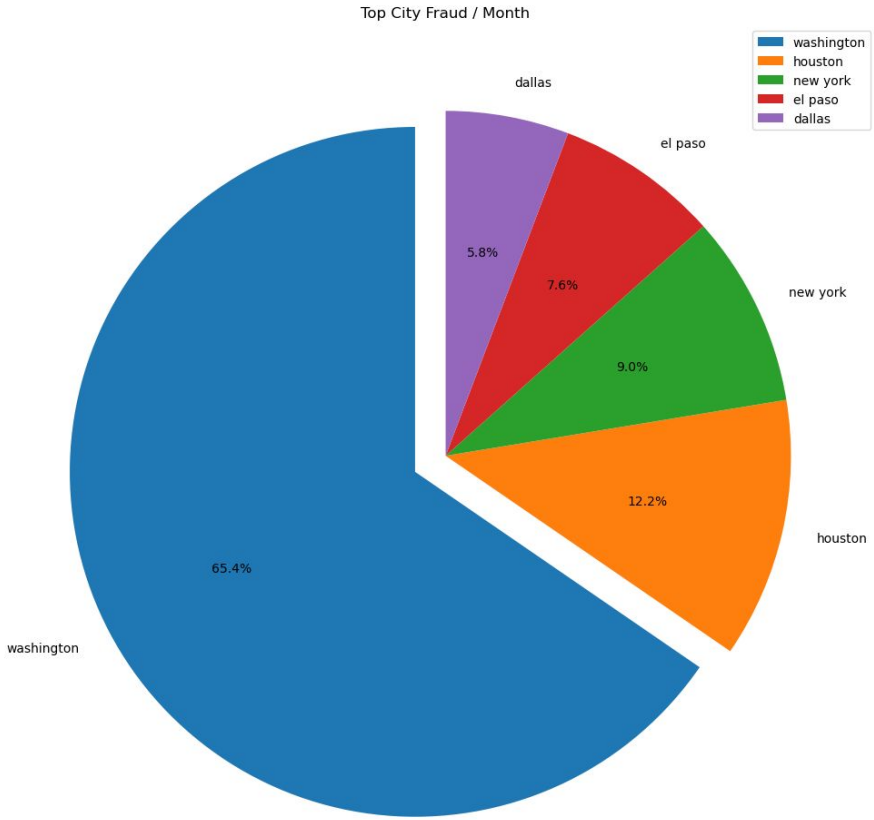
Regions Often Do Fraud

Washington was the leading center for fraud transactions with 65.4% of the total cases, significantly higher than Houston (12.2%) and New York (9.0%).

This spike **indicates a recurring fraud pattern or security gap** that needs to be addressed immediately.

Further investigation is needed to identify triggers, such as the **type of transaction and the characteristics of the perpetrator**.

With such a high level of fraud, **Washington should be a top priority in implementing prevention strategies**, including the use of machine learning for early detection and more effective mitigation.



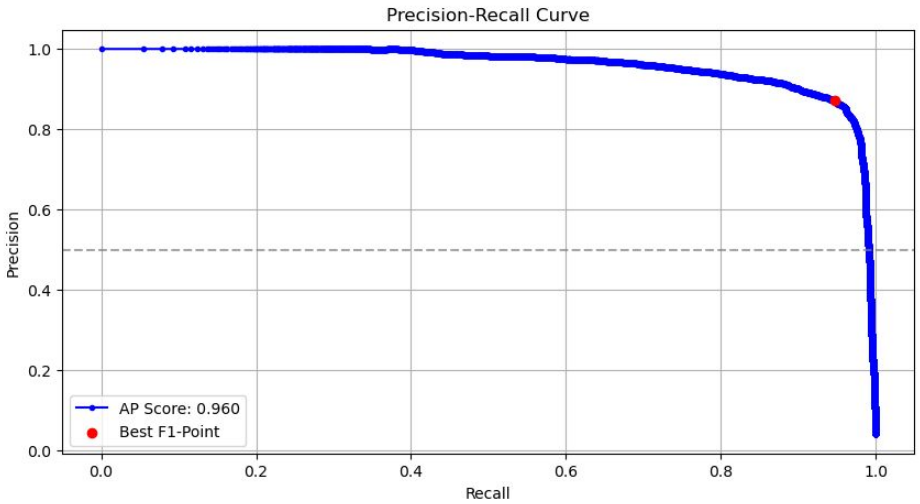
Model Evaluation

Evaluation by Precission Recall Curve

The Precision-Recall curve shows that the **model performs very well in detecting fraud**, with an **AP Score of 0.960** indicating a **high balance between precision and recall**.

Precision remains stable at various levels of recall, **indicating the reliability of the model in identifying fraudulent transactions** without many false positives.

The best point **(Best F1-Point) shows the optimal balance between precision and recall**. However, as recall approaches 1.0, precision drops drastically, indicating an increase in false positives. **Further optimization can be done by adjusting the threshold or applying cost-sensitive learning according to business needs.**

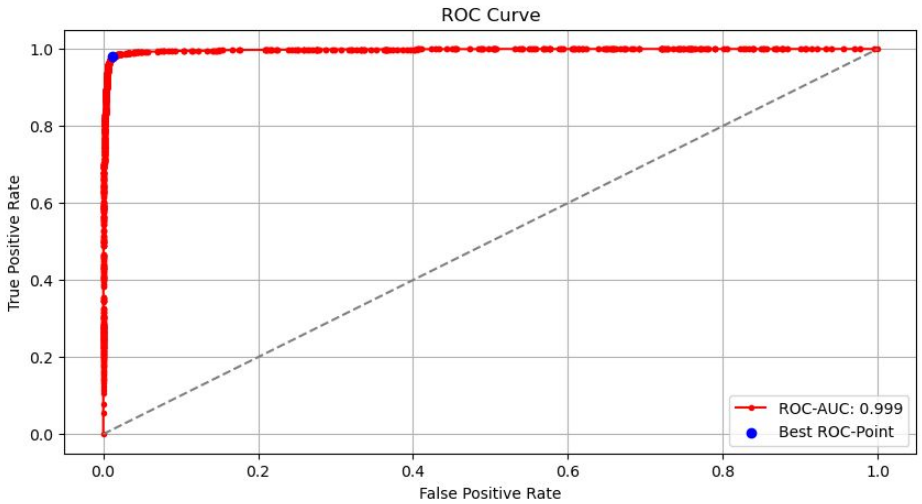


Evaluation by ROC Curve

ROC plot shows a **very high performance of the classification model**, with an **AUC of 0.999**, which is **almost perfect in distinguishing positive and negative classes**.

The curve that is close to the upper left corner indicates a **high True Positive Rate (TPR)** and a **low False Positive Rate (FPR)**, **indicating accurate detection with minimal errors.**

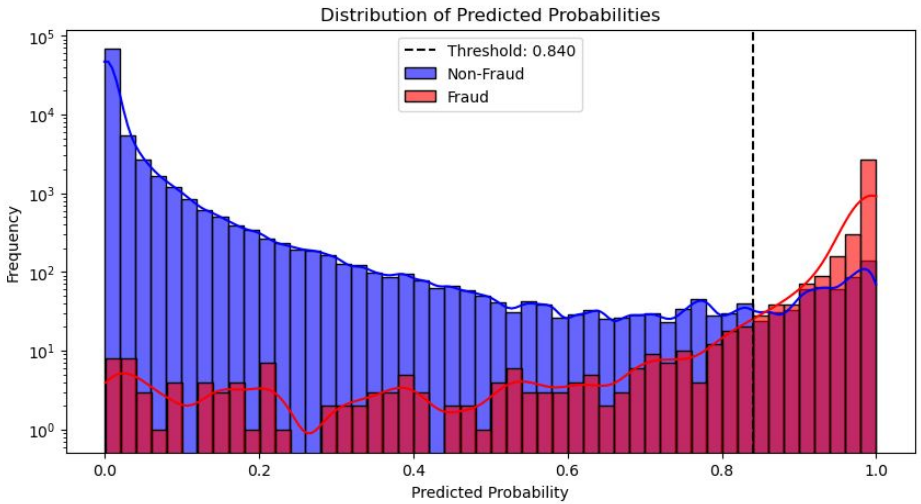
The **best ROC-point (blue dot) marks the optimal threshold for balancing sensitivity and specificity.** The distribution of red dots along the curve **shows stable model performance at various thresholds**, well above the **random baseline (AUC = 0,5).**



Evaluation by distribution model's predictions

The **majority of non-fraud (blue) has a low probability, while fraud (red) is concentrated at high values.**

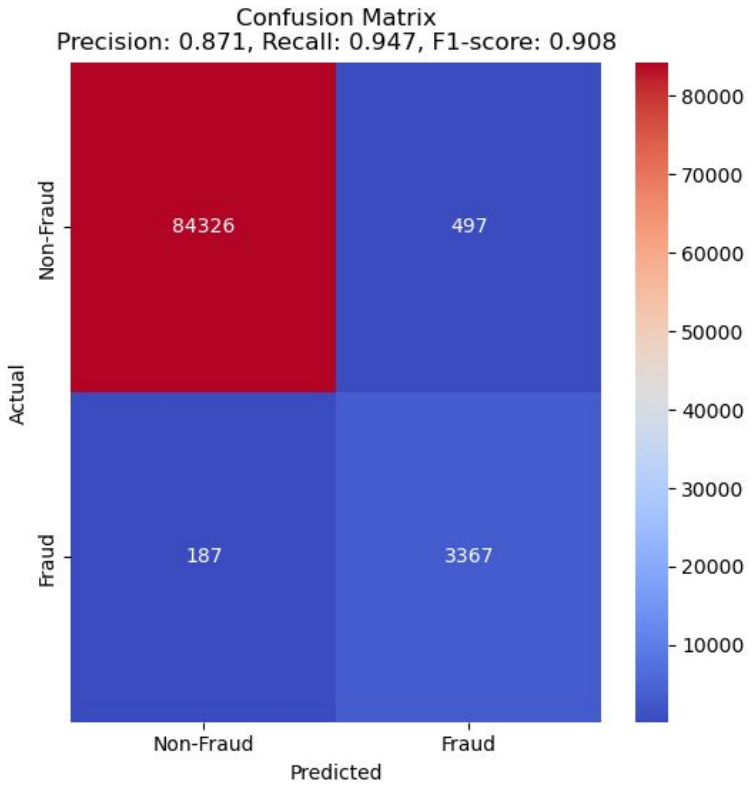
The threshold of 0.840 (dashed line) is the model's decision limit in classifying fraud. The **model is able to separate the two classes, but there is still overlap, which can lead to false positives or false negatives.**



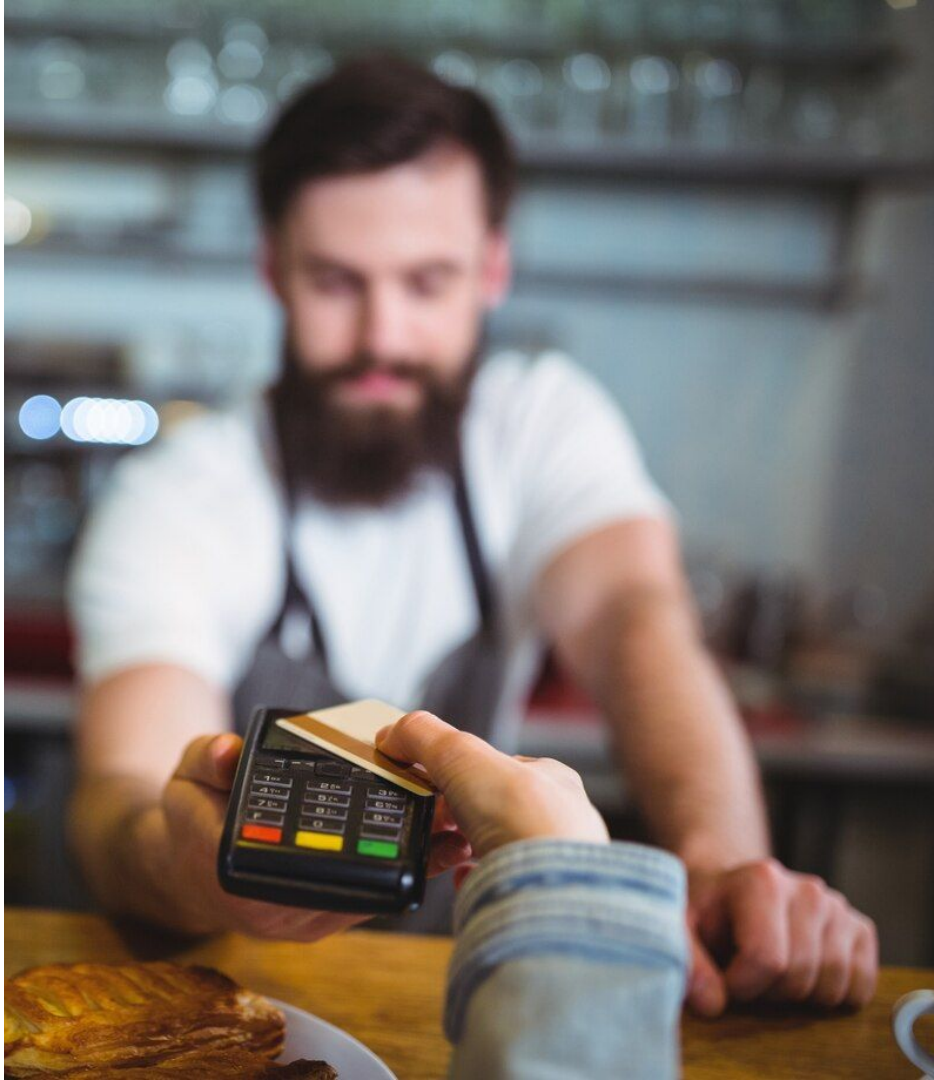
Evaluation by Confusion Matrix

The **model successfully classified 83,426 non-frauds and 3,367 frauds correctly.** However, **there were still 497 false positives (non-frauds that were incorrectly classified as fraud) and 187 false negatives (frauds that were missed).**

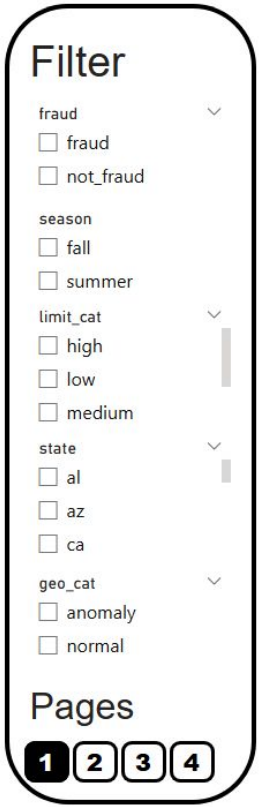
With a **precision of 0.871, a recall of 0.947, and an F1-score of 0.908, the model has a good balance between accuracy and fraud detection.**

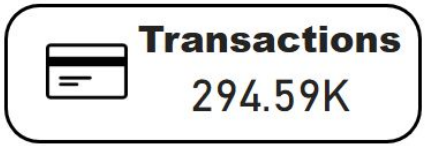


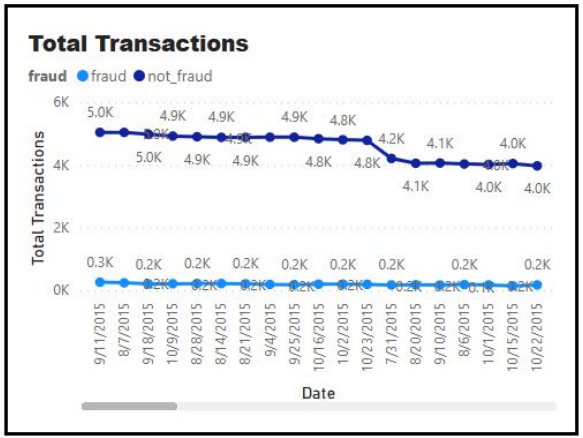
Data Visualization

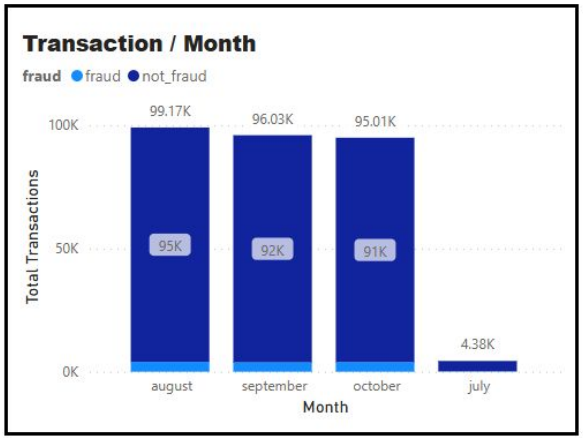


Cab Dashboard



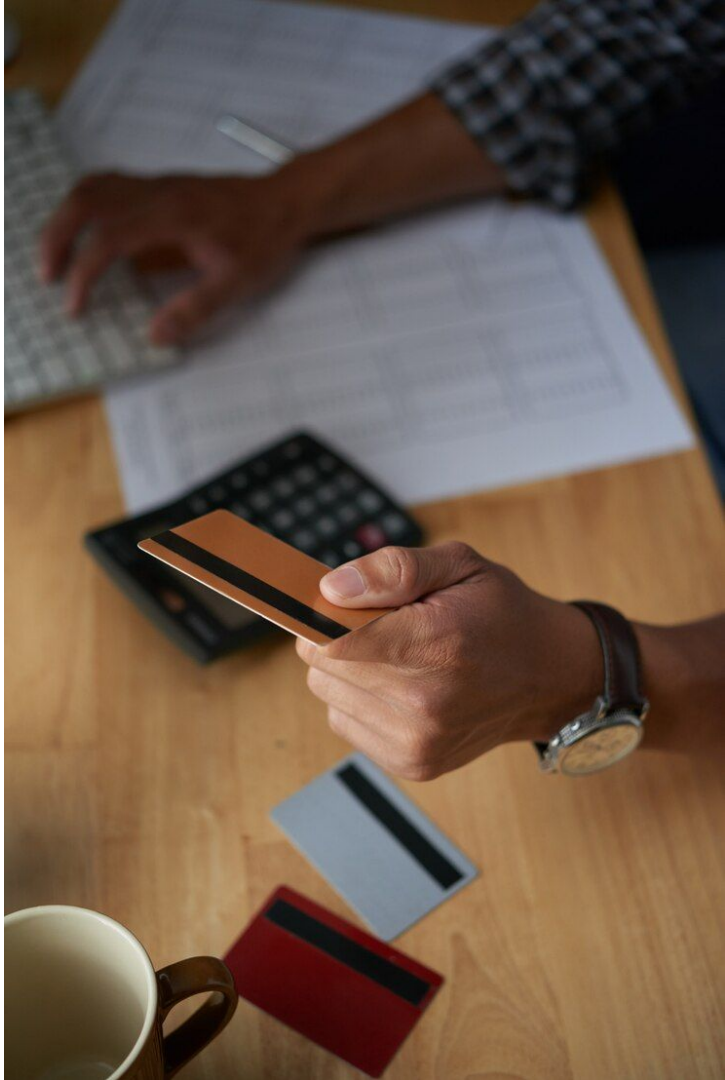






datetime	month	week_cat	city	long	lat	distance_km	Sum of amount_$	Sum of time_diff_hour
10/4/2015 9:39:52 PM	october	weekend	san diego	-984711122943232	27680057057169	9550635609856384	$3,133	120486
8/9/2015 10:43:57 PM	august	weekend	san diego	-984706792819681	276760631936531	10138204633914398	$10,419	114902
10/12/2015 7:23:51 PM	october	weekday	san diego	-984702012401144	276707105672263	4652115412645666	$3,326	81385
9/7/2015 7:32:08 PM	september	weekday	san diego	-984698608783737	276908903992548	7855518454941424	$602	10108
Total							**$2,303,626,023**	**19232518866**

Data Insight



Based on Key Indicators

Factors that determine whether a transaction is fraudulent include:

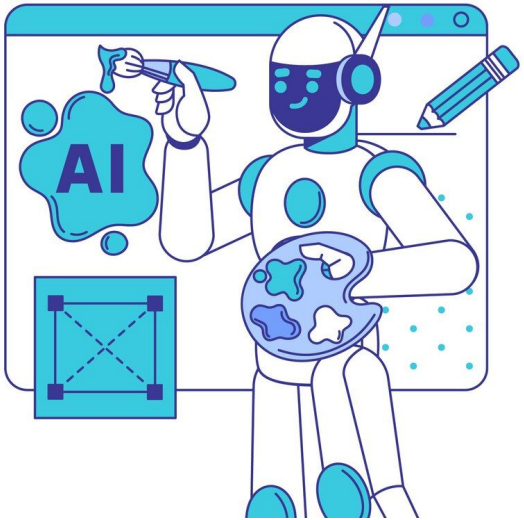
- **Number of duplicate transactions**: The more duplicate transactions there are, the more likely it is that the transaction is occurring.
- **Transaction duration**: Unusual time taken to send a transaction can be an indication of suspicious activity.
- **Geographic location of the transaction**: Significant differences in location from normal transaction patterns can be a sign of potential fraud.
- **Speed of transaction**: Transactions that occur in a very short time compared to normal patterns can be toxic.
- **Large transaction amounts**: Transactions with values that are much larger than average can be an indication of fraud.

Based on Region

Areas with **high population density** often have **large transaction volumes**, including credit card usage, which is an **opportunity for fraudsters** to commit fraud. The density and number of transactions **create loopholes for fraudulent practices** that are difficult to detect.

Therefore, **optimal steps are needed to detect fraudulent transactions.** The manual process is time-consuming, so an **automated system that analyzes transactions** in real-time is needed to detect suspicious activity more efficiently.

Based on Models



Fraud detection model performed very well in distinguishing fraudulent and non-fraudulent transactions, as indicated by an **AP Score of 0.960 and an AUC of 0.999**. The **model showed an optimal balance between precision and recall**, with stable detection capabilities across thresholds.

Although the **model successfully classified most transactions correctly**, there were **still false positives and false negatives that needed to be minimized**. To further improve accuracy, **optimization can be carried out by adjusting the threshold or applying cost-sensitive learning to better suit the model's** business needs and operational risks.

If you want to collaborate

Profile (Clickable)

Email (Clickable)

Project (Clickable)


THANK
YOU