CREDIT CARD FRAUD DETECTION

THE IMPORTANCE OF MACHINE LEARNING IN CREDIT CARD FRAUD DETECTION

PROBLEM STATEMENT

In the banking industry, **credit card fraud** is a growing concern. Fraudulent transactions cause significant losses for both banks and cardholders.

Banks need proactive systems to detect and prevent fraud quickly, before substantial damage is done.

MACHINE LEARNING AS A SOLUTION

- Machine learning helps banks automate fraud detection, reducing the need for manual reviews and improving detection speed.
- By leveraging historical transaction data, ML models can predict fraudulent behavior and flag suspicious transactions in real-time.

BENEFITS FOR BANKS

- Cost Reduction: Reduce costs related to chargebacks, fraud investigation, and customer support.
- Improved Customer Experience: Avoid unnecessary denials of legitimate transactions, ensuring smooth service for customers.
- Risk Mitigation: Detect fraud early to minimize financial losses.

THE IMPORTANCE OF MACHINE LEARNING IN CREDIT CARD FRAUD DETECTION

DEFINITION

Credit card fraud is unauthorized access to a credit card account with the intent of making fraudulent transactions for financial gain.

IMPACT OF FRAUD

- Significant financial losses for banks and cardholders.
- Trust erosion with customers if fraud is not managed effectively.

COMMON FRAUD TECHNIQUES

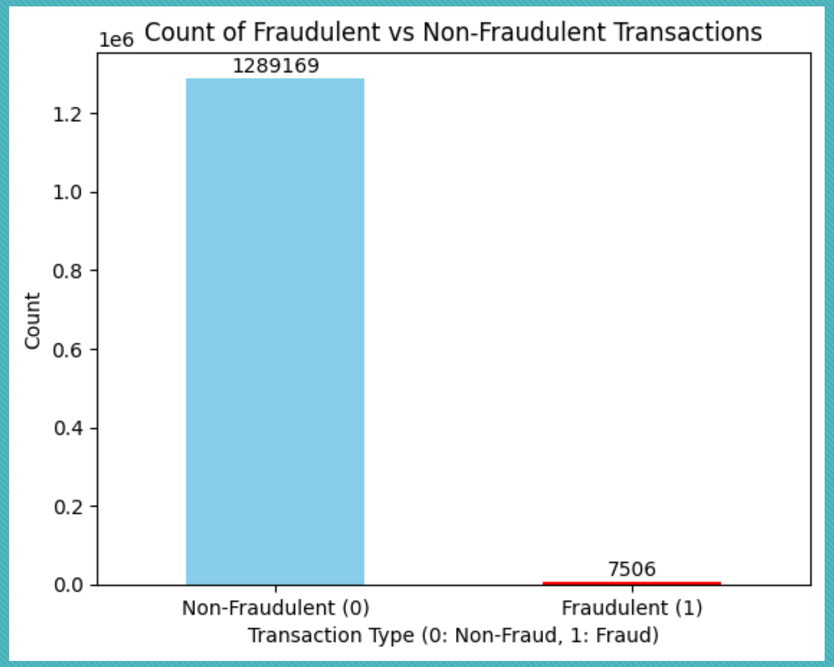
1. **Skimming**: The most common type of fraud where the magnetic stripe of a card is copied and used for fraudulent transactions.
2. **Card Manipulation**: Altering or tampering with legitimate cards to bypass security checks.
3. **Counterfeit Cards:** Creating fake cards using stolen information.
4. **Stolen/Lost Cards**: Fraudulent use of credit cards after being lost or stolen.
5. **Fraudulent Telemarketing**: Scammers obtaining credit card details through deceptive phone calls or other means.

THE IMPORTANCE OF MACHINE LEARNING IN CREDIT CARD FRAUD DETECTION

DATA OVERVIEW

The dataset contains 1,852,394 transactions from 1,000 cardholders across 800 merchants.

Time Period: January 1, 2019 to December 31, 2020.



TRANSACTION TYPES

- Both legitimate and fraudulent transactions are included.
- The dataset includes features such as transaction amount, merchant details, time, and customer ID.
- Fraudulent Transactions - Out of the total transactions, 9,651 are fraudulent.
- Imbalance: Fraudulent transactions account for only 0.52% of the total dataset.
- Challenge: The high class imbalance (fraud vs. legitimate) requires special handling techniques (e.g., resampling, SMOTE, or class weights) to build an effective fraud detection model.

THE IMPORTANCE OF MACHINE LEARNING IN CREDIT CARD FRAUD DETECTION

"DATA LOAD → DATE CONVERSION → AGE CALCULATION → FEATURE ENGINEERING"

DATA LOADING

The dataset is imported from Google Drive into **Google Colab** for analysis and model building.

The data is sourced from a Kaggle dataset containing transactions for 1,000 cardholders and 800 merchants.

DATA CLEANING & PREPARATION

Date Conversion

trans_date_trans_time converted to pd.to_datetime to extract year and month for feature engineering.

dob (Date of Birth) converted to pd.to_datetime to calculate the age of the cardholder at the time of transaction.

Age Calculation:

The age of the individual at the time of the transaction is derived by subtracting dob from trans_date.

Dropping Irrelevant Columns:

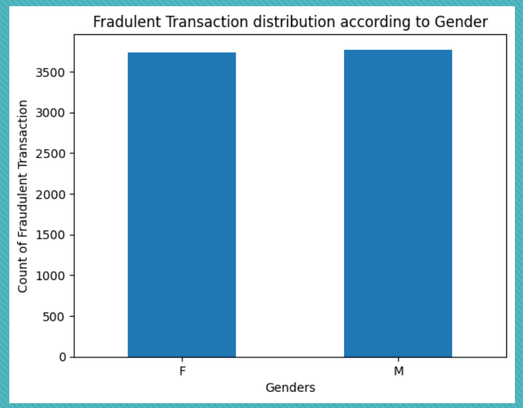
Unnecessary columns such as Unnamed: 0, cc_num, first, last, street, zip, trans_num, and unix_time are dropped.

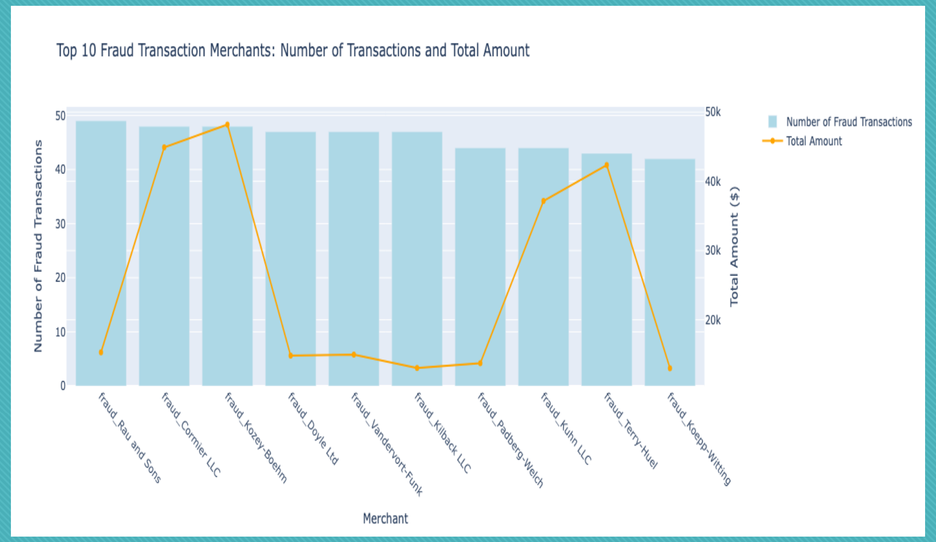
FEATURE ENGINEERING

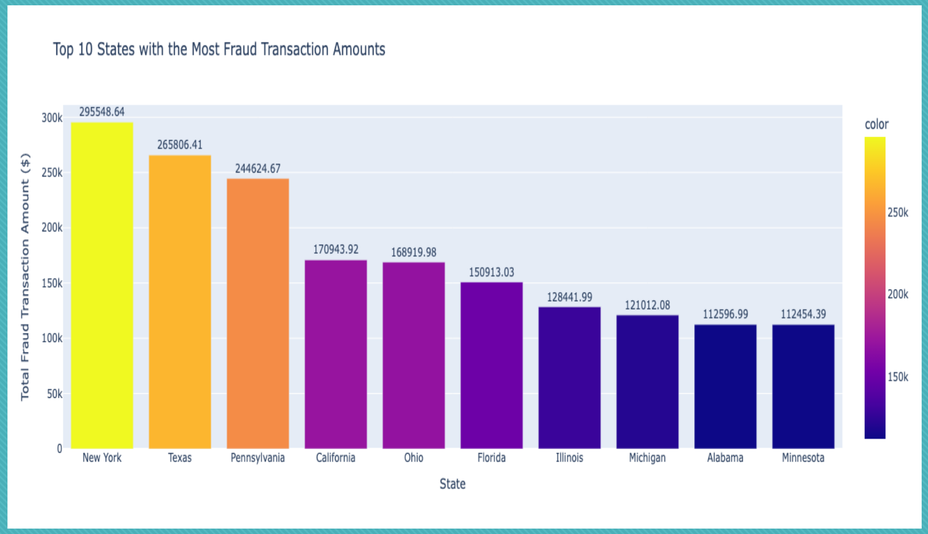
A custom function, calculate_distance(), computes the Euclidean distance between the cardholder's location (latitude and longitude) and the merchant's location.

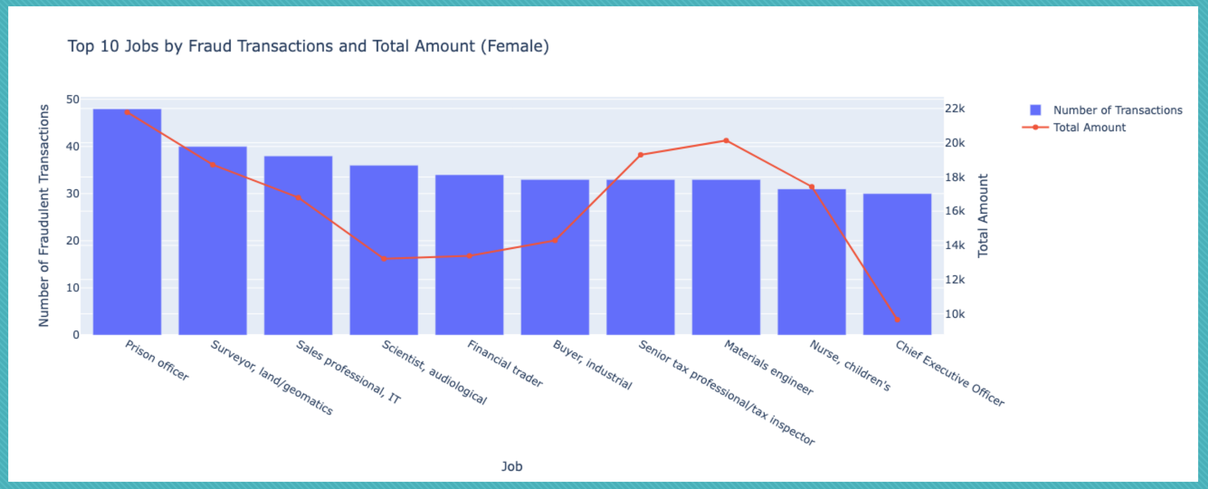
Latitude and longitude are first converted from degrees to radians, and then the distance is calculated using the Haversine formula (approximated for simplicity here).

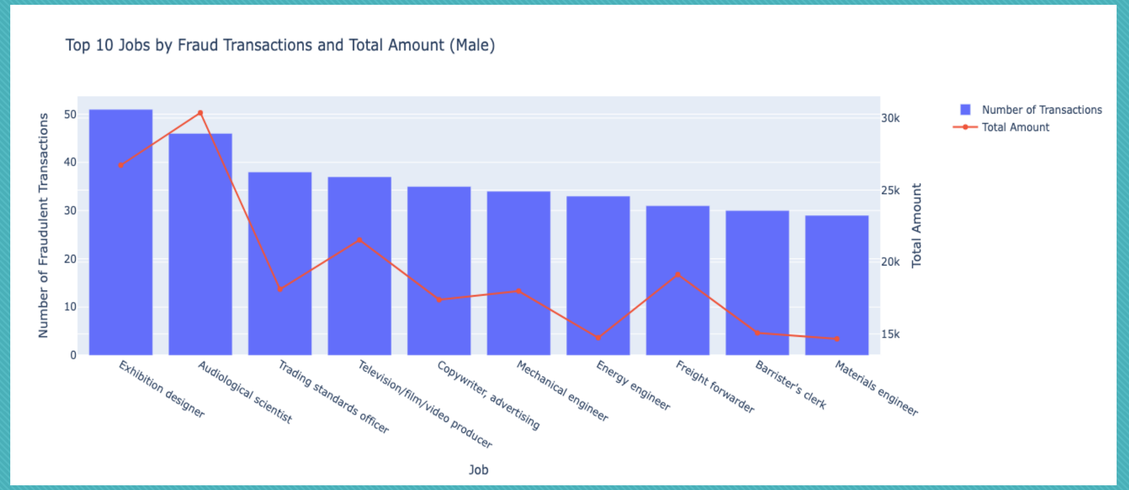
FRAUD DATA ANALYSIS











THE IMPORTANCE OF MACHINE LEARNING IN CREDIT CARD FRAUD DETECTION

"DROP COLUMNS → ENCODE GENDER → CREATE DUMMY VARIABLES → TRAIN-TEST SPLIT → POWER TRANSFORMATION"

FEATURE SELECTION & DATA PREPROCESSING

Dropping Irrelevant Columns:

Removed columns such as trans_date_trans_time, merchant, city, state, lat, long, job, dob, and merch_lat to focus on relevant features.

Gender Encoding:

The gender column is mapped to numerical values: 'M' → 1 and 'F' → 0 for model compatibility.

Dummy Variables for Categorical Data:

category column is converted into dummy variables using pd.get_dummies() to handle categorical data.

TRAIN-TEST SPLIT

Features and Target Variables:

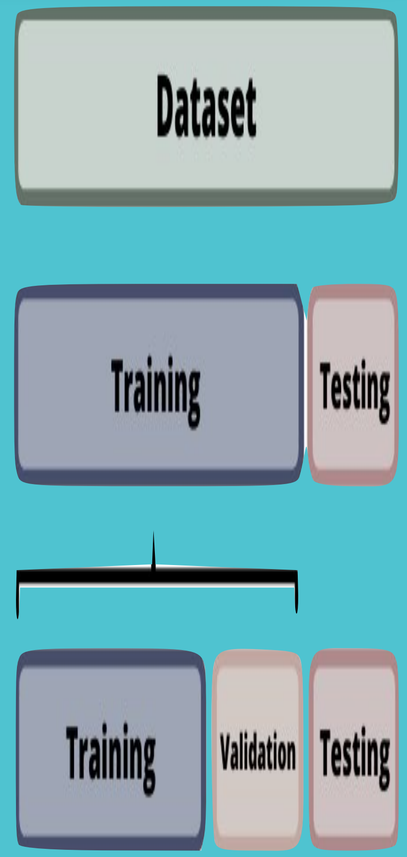
X_train and X_test are created by removing the target variable is_fraud.

y_train and y_test represent the target variable indicating whether the transaction is fraudulent (1) or not (0).

Handling Skewness:

Power Transformation:

PowerTransformer is used to handle skewed data and normalize features, ensuring all variables have a more Gaussian distribution.



THE IMPORTANCE OF MACHINE LEARNING IN CREDIT CARD FRAUD DETECTION

RANDOM FOREST > XGBOOST MODEL

ROC Score:

- Random Forest: 0.9469
- XGBoost: 0.9986

Recall Score:

- Random Forest: 0.9490
- XGBoost: 0.95 (for class 1)

Accuracy:

- Random Forest: 0.95
- XGBoost: 0.99

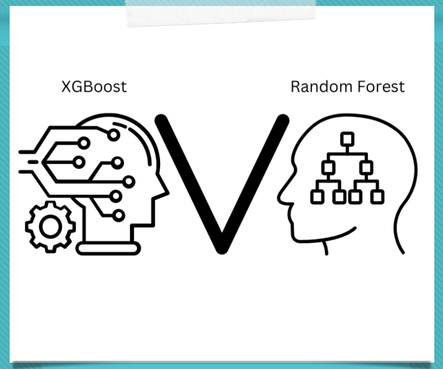
F1-Score:

- Random Forest: 0.95 (for both classes)
- XGBoost: 0.63 (for class 1), 1.00 (for class 0)

SUMMARY AND CONCLUSION

XGBoost outperforms Random Forest in terms of ROC score (0.9986 vs. 0.9469) and accuracy (0.99 vs. 0.95). It has a higher recall for class 1 (fraud) and avoids false positives (FP = 0).

However, **Random Forest has a much better precision for class 1 (0.95) and a much higher F1-score (0.95 for both classes)**. This makes Random Forest the better model in terms of practical performance, especially for fraud detection, because it achieves a good balance of high recall and high precision for detecting fraudulent transactions.



BEST MODEL FOR FRAUD DETECTION

Random Forest is likely the better model for fraud detection overall due to its high precision and balanced performance, especially if the cost of false positives (misclassifying non-fraudulent transactions as fraud) is important.

XGBoost could be considered if avoiding false positives is the highest priority, but you will need to address the low precision for fraud detection.

Thus, **Random Forest is the recommended choice for your fraud detection problem, especially when both recall and precision are important**

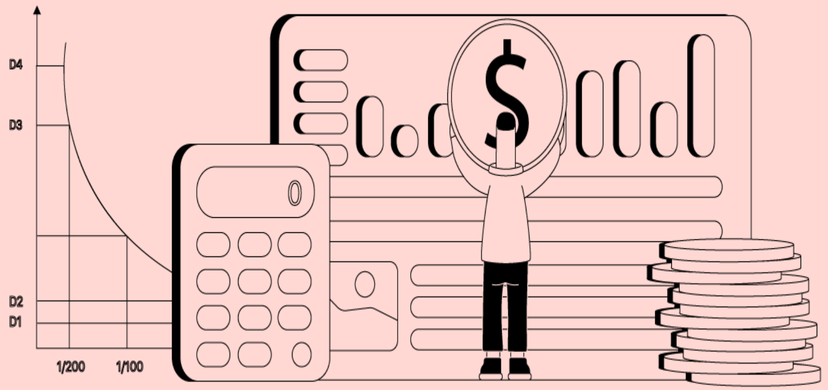
THE IMPORTANCE OF MACHINE LEARNING IN CREDIT CARD FRAUD DETECTION

LOSS

$ 213462

SAVING

$ 197802



Cost Benefit Analysis		
S. No	Questions	Answer
a	Average number of transactions per month	77183
b	Average number of fraudulent transaction per month	402
c	Average amount per fraud transaction	$ 531.00

Cost Benefit Analysis		
S. No	Questions	Answer
1	Cost incurred per month before the model was deployed (b*c)	$ 213462.00
2	Average number of transactions per month detected as fraudulent by the model (TF)	4068
3	Cost of providing customer executive support per fraudulent transaction detected by the model	$1.5
4	Total cost of providing customer support per month for fraudulent transactions detected by the model (TF*$1.5)	$ 6102.00
5	Average number of transactions per month that are fraudulent but not detected by the model (FN)	18
6	Cost incurred due to fraudulent transactions left undetected by the model (FN*c)	$ 9558.00
7	Cost incurred per month after the model is built and deployed (4+6)	$ 15660.00
8	Final savings = Cost incurred before - Cost incurred after(1-7)	$ 197802.00

THE IMPORTANCE OF MACHINE LEARNING IN CREDIT CARD FRAUD DETECTION

Shashank H P

Kumar Mayank

Abbigel Sadhu

THE IMPORTANCE OF MACHINE LEARNING IN CREDIT CARD FRAUD DETECTION

VEDIO AND PROJECT CODE IS UPDATED IN DRIVE LINK BELOW

https://drive.google.com/drive/folders/1tranbo-h4yezDXoq5Nl8Erg0lvEn3srZ?usp=sharing