AI Data Science - CAIXABANK TECH

Adolfo Artola Madrigal

https://nuwe.io/hackathons

Background

In today's digital economy, banks and financial institutions process massive amounts of transactional data daily, which includes purchases, payments, transfers, and other activities.

This data is not only pivotal for monitoring individual financial behavior but also critical for detecting fraudulent transactions, understanding customer spending patterns, and optimizing personalized financial services.

Objective

The goal of this challenge is to leverage transaction data from various sources, such as APIs and pre-existing datasets—merge and process it for meaningful analysis, and train machine learning models that can predict outcomes like fraud detection.

Task 1: Submit the answers to the following queries:

- **query_1** : The card_id with the latest expiry date and the lowest credit limit amount.

- **query_2** : The client_id that will retire within a year that has the lowest credit score and highest debt.

- **query_3** : The transaction_id of an Online purchase on a 31st of December with the highest absolute amount (either earnings or expenses).

- **query_4** : Which client over the age of 40 made the most transactions with a Visa card in February 2016? Please return the client_id, the card_id involved, and the total number of transactions.

Task 2: Implement the functions

- Function: earnings_and_expenses(data,client_id,start_date,end_date)
- Function: expenses_summary(data,client_id,start_date,end_date)
- Function: cash_flow_summary(data,client_id,start_date,end_date)

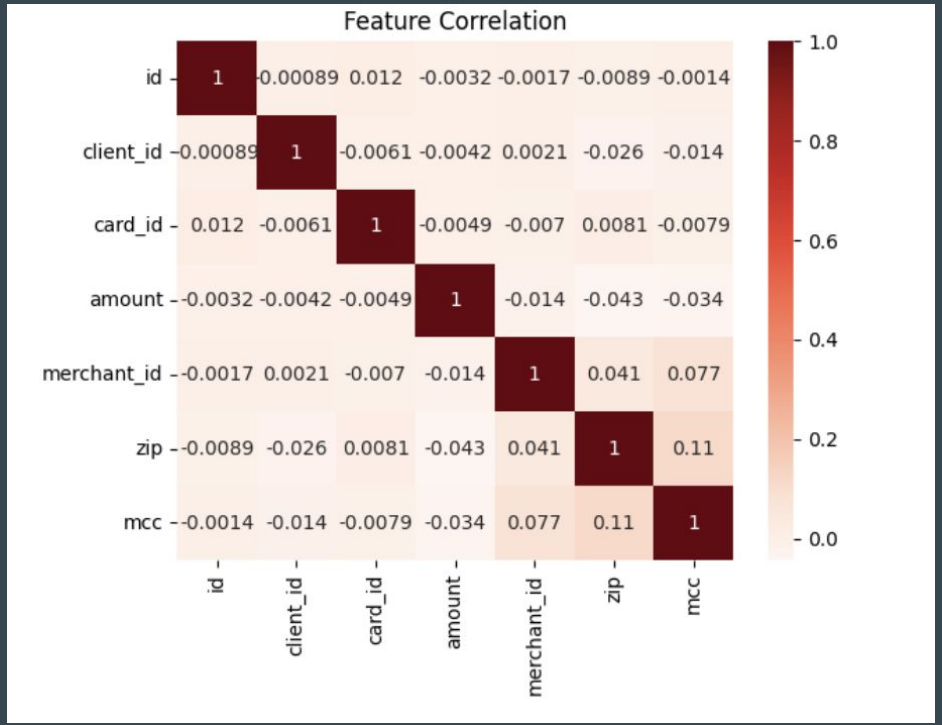
TASK 3: Fraud Detection Model

- This model will classify transactions as either fraudulent or non-fraudulent.

- The goal is to identify suspicious transactions in real-time by analyzing patterns such as transaction frequency, amounts, and locations. Supervised learning techniques, such as logistic regression, decision trees, or ensemble methods, can be employed.

1. Data Cleaning + exploratory data analysis

	cols	null
0	id	0
1	date	0
2	client_id	0
3	card_id	0
4	amount	0
5	use_chip	0
6	merchant_id	0
7	merchant_city	0
8	merchant_state	1563700
9	zip	1652706
10	mcc	0
11	errors	13094522



2. MODEL: XGBOOST

model = XGBClassifier(

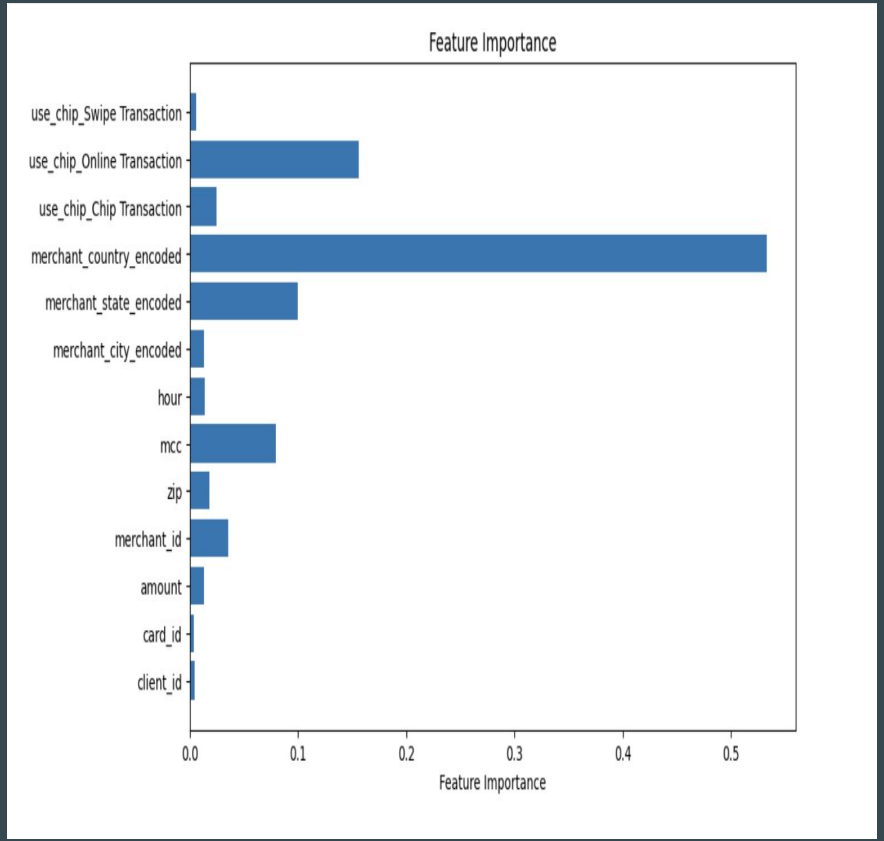
max_depth =8,

learning_rate =0.05,

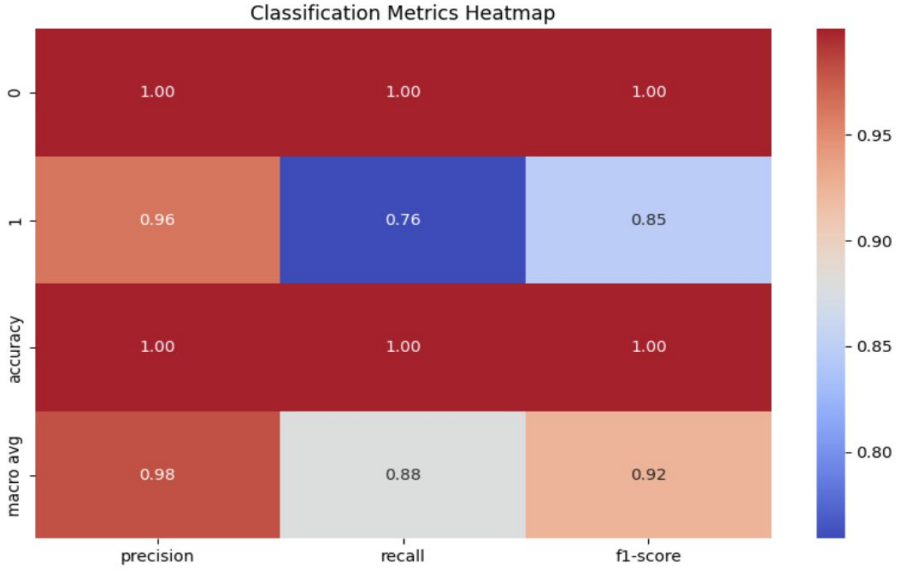
n_estimators =200,

random_state =42,

subsample =0.8)



3. Model Evaluation



Precision: 96% of the instances that the model predicted as fraudulent were actually fraudulent
Recall: 76% of the actual positive instances (fraudulent transactions) were correctly identified by the model
F1: The F1-score is the harmonic mean of precision and recall
The accuracy is 1.00, meaning that 100% of the predictions are correct (this is likely due to the large imbalance in the dataset

4. Improvements

Imbalance Target:

Target	Count
0	8901631
1	13332

Hyperparameter Tunning:

a. Cross Validation for each combination of hyperparameters.

```
model2 = XGBClassifier(random_state=42)

param_grid = {
    "max_depth": [6, 10],
    "learning_rate": [0.01, 0.1],
    "n_estimators": [100, 300],
    "subsample": [0.9, 1.0],
    "colsample_bytree": [0.9, 1.0],
    "scale_pos_weight": [10, 20]
}

grid_search = GridSearchCV(
    estimator=model,
    param_grid=param_grid,
    scoring="f1",
    cv=3,
    verbose=2,
    n_jobs=-1
)
grid_search.fit(X_train, y_train)

print("Best Parameters:", grid_search.best_params_)
print("Best F1 Score:", grid_search.best_score_)

best_model = grid_search.best_estimator_
y_pred = best_model.predict(X_val)

print(classification_report(y_val, y_pred))
```

THANK YOU!!