Analysis of Financial Dataset for Fraud Detection

Nadia Fantello

Problem Statement

- Digital Fraud is, unfortunately, a very common problem that can generate a lot of discomfort.
- Those behind digital fraud are always creating new ways to commit fraud without being detected.
- Detection systems are not perfect: they often mistake non-fraudulent transactions with fraudulent ones, and vice versa.

Project Goals

- Identify key features that can help us understand digital fraud.
- Improve fraud detection accuracy using Machine Learning:
  - Minimize the rate of false positives.
  - Strengthen the detection system.
- Questions to be answered:
  - What is the most fraudulent transaction type?
  - Is there a typical time of day when fraudulent activities are carried out?

Dataset

- Synthetic dataset downloaded from ***Kaggle***.
- Simulates digital money transactions.
- Total of 6,362,620 records.

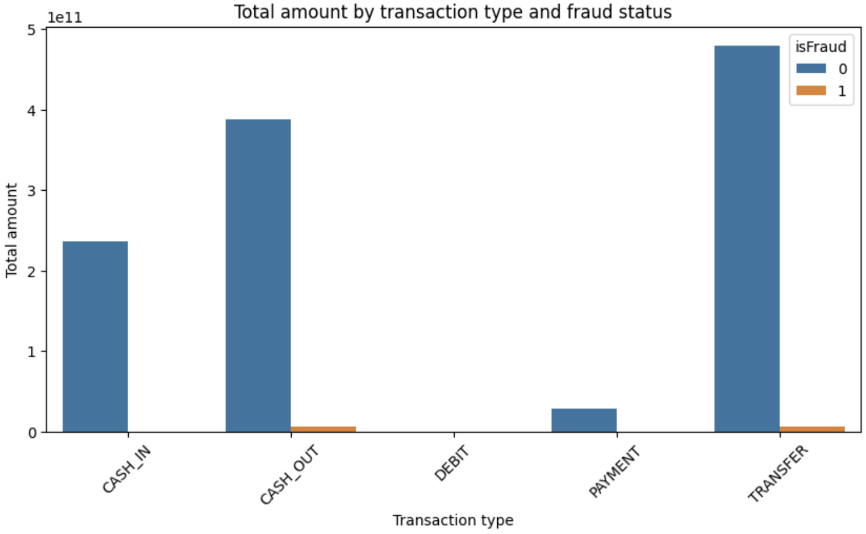
Variable Name	Data Type	Description
step	int64	maps a unit of time in the real world
type	object	CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER
amount	float64	amount of the transaction in local currency
nameOrig	object	customer who started the transaction
oldBalanceOrg	float64	balance before transaction
newBalanceOrig	float64	balance after transaction
nameDest	object	recipient of transaction
oldBalanceDest	float64	balance recipient before transaction
newBalanceDest	float64	balance recipient after transaction
isFraud	int64	identifies a fraudulent transaction (1) and non fraudulent (0)
isFlaggedFraud	int64	an illegal attempt is an attempt to transfer more than 200.000 in a single transaction

Data Cleaning and Exploration

- No null values encountered.
- Analysis of variable ***isFraud:***

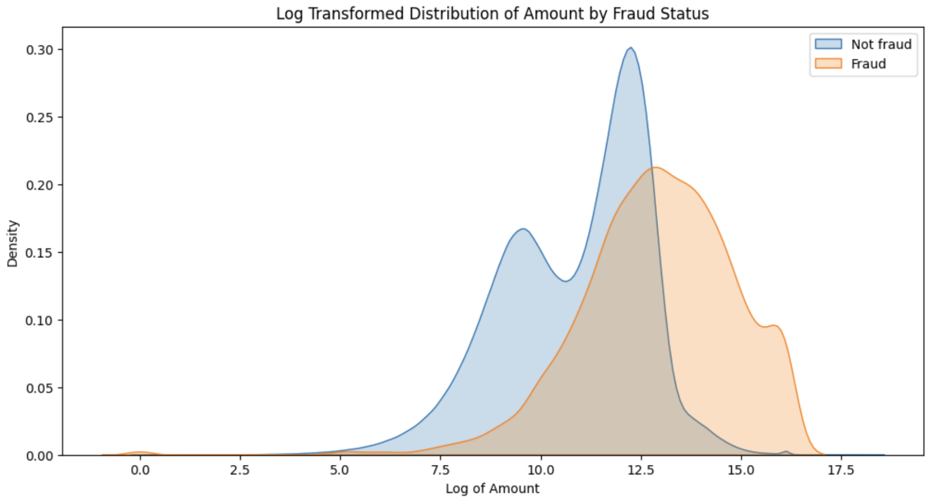
isFraud	mean	median	std
0	1.781970e+05	74684.72	5.962370e+05
1	1.467967e+06	441423.44	2.404253e+06

- Transaction type and fraud status:



Data Cleaning and Exploration

- Data transformation and visualization:
  - Applied log transformation to *transaction amount.*
  - Distribution shows a log amount target area for monitoring and investigation.



Data Cleaning and Exploration

- Class imbalance problem:

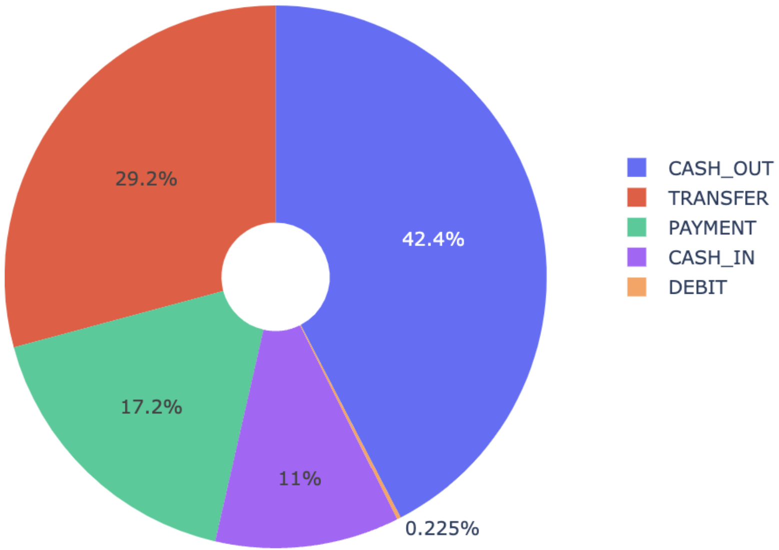
```
isFraud
0    6354407
1       8213
Name: count, dtype: int64
```

- Solution: randomly removed entries from the majority class (non-fraudulent transactions) to match the minority class' count.

```
isFraud
0    8213
1    8213
Name: count, dtype: int64
```

Data Cleaning and Exploration

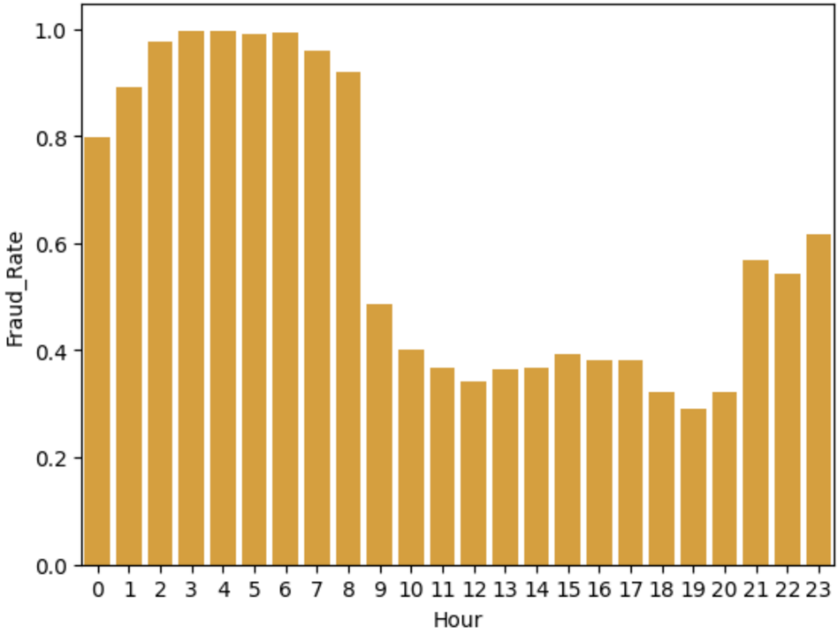
- Variable ***type***:



isFraud	0	1
type		
CASH_IN	1805	0
CASH_OUT	2843	4116
DEBIT	37	0
PAYMENT	2831	0
TRANSFER	697	4097

Data Cleaning and Exploration

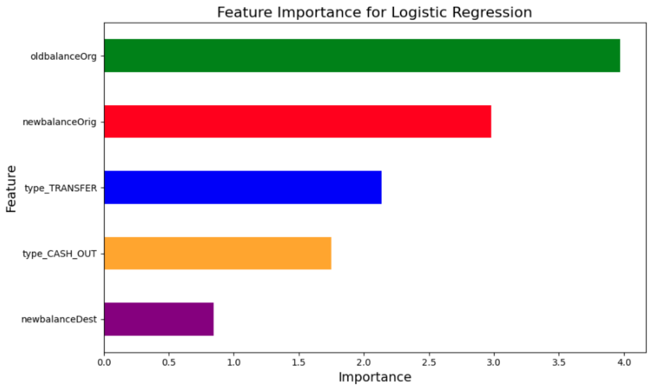
- Variable ***step***:

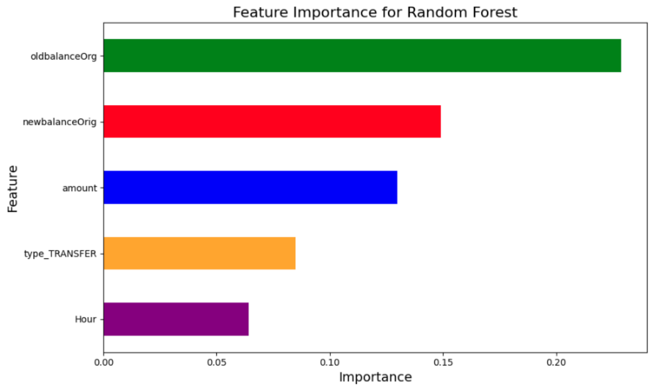


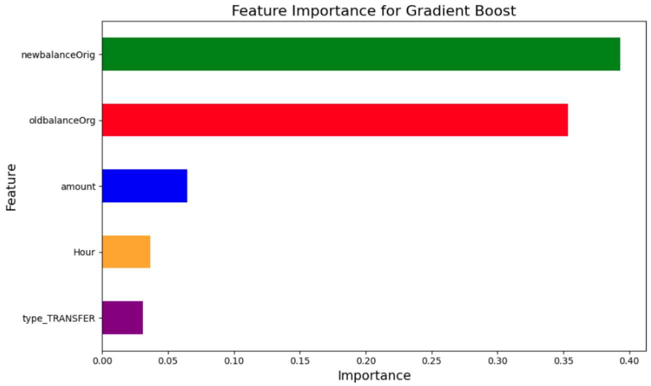
Machine Learning Methodology

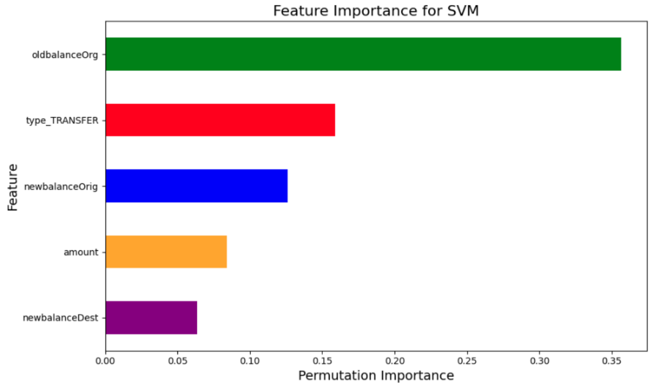
- **Models used:** Logistic Regression, KNN Classifier, Random Forest, Support Vector Machine, and Gradient Boosting.
- **Training and Evaluation Methods:** Training process (80/20), cross-validation (GridSearchCV), ROC, accuracy, and MSE.

Machine Learning Models - Feature Importance









Machine Learning Models - Comparison

Model	Training Accuracy	Test Accuracy	MSE
Logistic Regression	0.8766	0.8609	0.1125
KNN	0.9056	0.8904	0.0937
Random Forest	0.9818	0.9808	0.0143
SVM	0.9469	0.9315	0.0365
Gradient Boosting	0.9831	0.9829	0.0129

Conclusions

- Used ML techniques on a synthetic dataset for fraud detection.
- Random Forest and Gradient Boosting performed well detecting fraud.
- Logistic Regression, SVM, and KNN performed well but not as effective.
- Quantitatively demonstrated high accuracy and strong ability to detect fraudulent activities.

Q&A