Hybrid Model

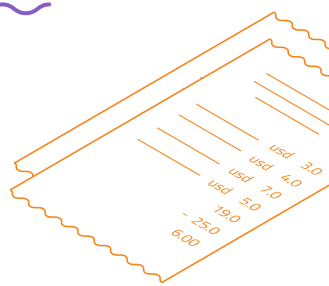
For Fraud Detection *in* Online Transaction

Under Supervision

Dr. Pritish Kumar Varadwaj

Md Ahmad Jami
M.Tech (DSA)

Table of contents



Introduction

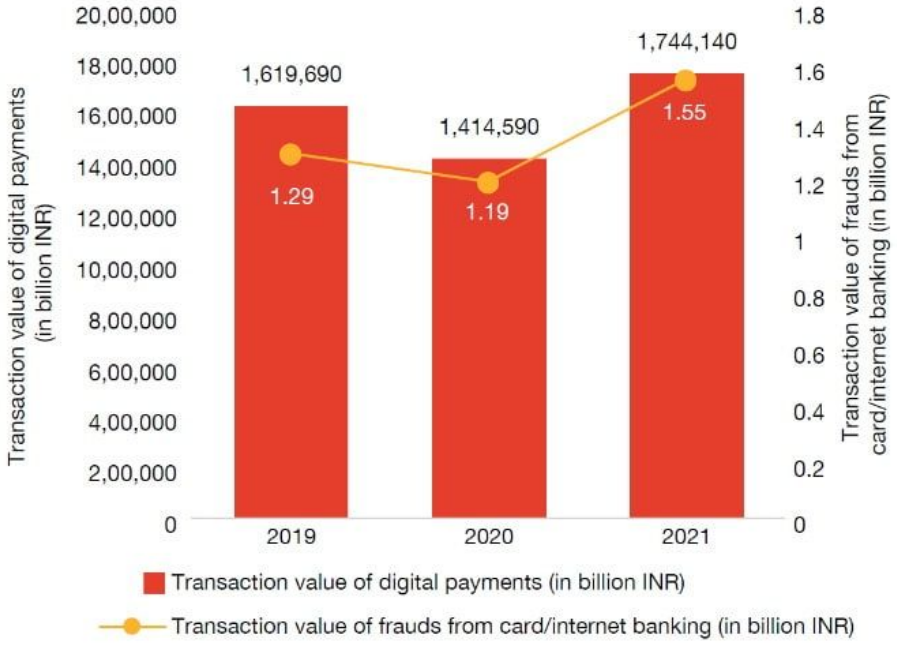
- Rapid development of technology has digitized customers' payment behavior towards a cashless society.
- People with bad intentions have come up with new ways to deceive the common masses and make them fall into the fraud traps.
- From the time when COVID hit there has been an exponential increase in digital payments fraud.

As per the Reserve Bank of India's (RBI) Annual Report 2021–22:

- The volume of frauds reported by financial institutions using cards and internet banking was 34% higher at 3,596 in 2021–22 as against 2,677 frauds in 2019–20.

- The value of fraudulent transactions in 2021–22 was INR 1.55 billion – 20% more than that in 2019–20 (INR 1.29 billion)

Value of frauds vis-a-vis the digital payments transactions (cards and internet banking)



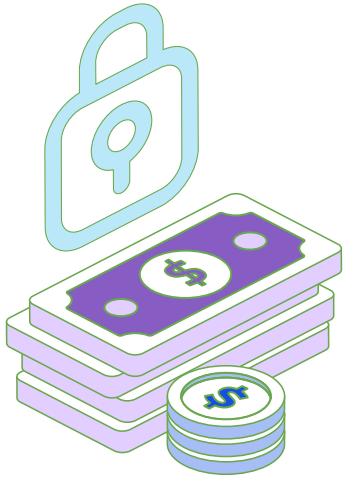
Hence, fraud detection systems are the need of the hour.

Source: PwC India

Problem Statement

The main objective of finding fraud transactions can be following:

- If a fraud transaction is found out, the company should immediately block that card.
- We should be able to predict the probability of fraud transaction.
- We should not predict fraud transactions as non fraud. Also the vice versa. So precision and recall should be taken care of.



Literature Review

Author	Title	Dataset	Findings
K Kamusweke et.al (2019)	Data mining for fraud detection in large scale financial transactions	Bank	Discovered hidden patterns using data mining
Yeming Chen et.al (2021)	CatBoost for Fraud Detection in Financial Transactions	IEEE-CIS	Apply feature engineering into CatBoost
S. O. Arik and T. Pfister (2021)	Tabnet: Attentive interpretable tabular learning	Multiple dataset	Deep learning architecture for tabular learning.

Dataset

The dataset provided by **Vesta Corporation** and it is available in the Kaggle by researchers from **IEEE Computational Intelligence Society (IEEE-CIS)**.

Dataset consists of two files **transaction** and **identity** joined by **TransactionID**.

Identity (categorial features)

Information related to the identity of a purchaser such as device type, device information,network connection information and digital signatures etc.

Transaction (Numerical+categorical features)

Contains transactional information, product code, address, email domains of purchaser and recipient, and special features with hidden meaning engineered by Vesta.

Dataset consists:

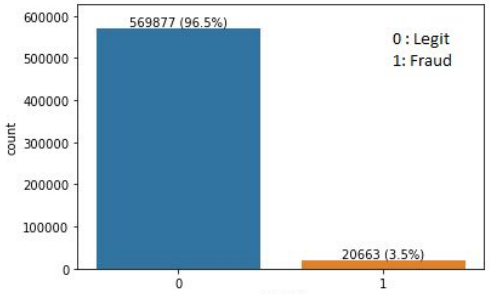
0.5 Million transactions

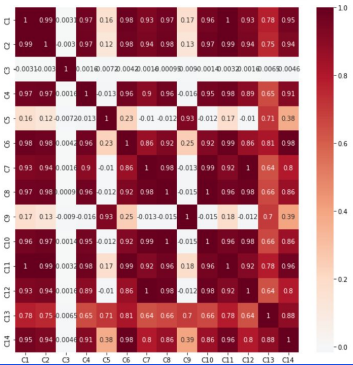
432 features and **a target feature** for each transaction.

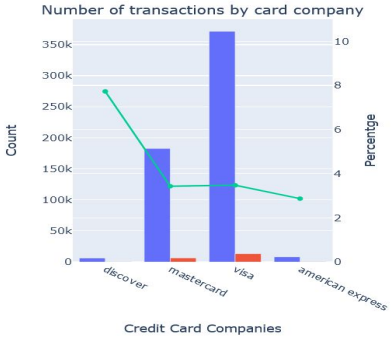
Columns:

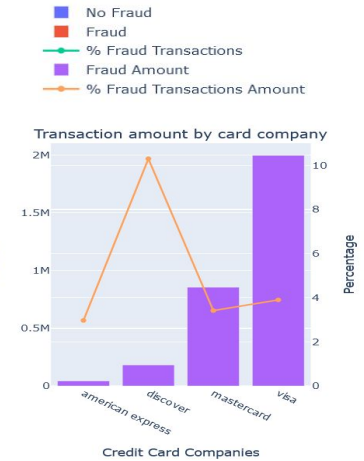
- **TransactionDT**: timedelta from a given reference datetime (not an actual timestamp)
- **TransactionAMT:** transaction payment amount in USD
- **ProductCD**: product code, the product for each transaction
- **card1 - card6**: payment card information, such as card type, card category, issue bank, country, etc.
- **addr**: address
- **dist**: distance
- **P_ and (R__) emaildomain**: purchaser and recipient email domain
- **C1-C14**: counting, such as how many addresses are found to be associated with the payment card, etc. The actual meaning is masked.
- **D1-D15**: timedelta, such as days between previous transaction, etc.
- **M1-M9**: match, such as names on card and address, etc.
- **V1-V339**: Vesta engineered rich features, including ranking, counting, and other entity relations, it's masked
- **DeviceType** : Mobile/Desktop
- **DeviceInfo** : Windows/Mac iOS etc
- **id_12 - id_38** : network connection information and digital signatures etc. it's masked

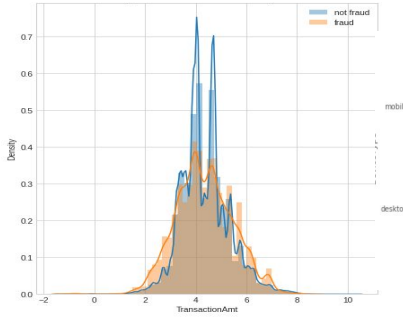
EDA

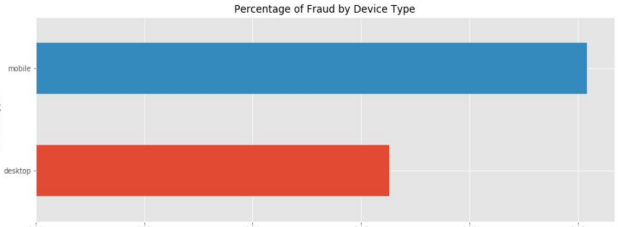






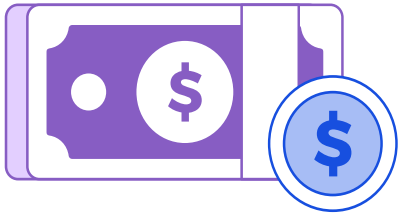




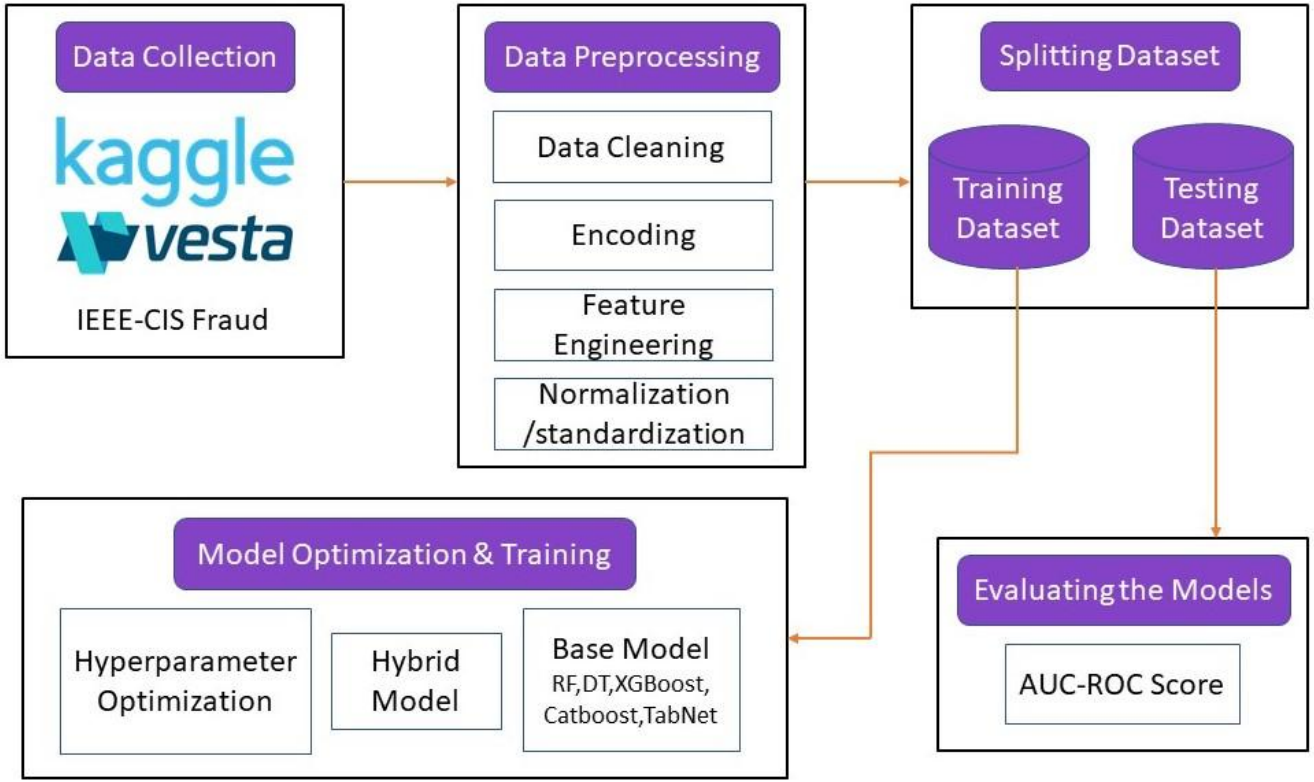


	Missing Values	% of Total Values
id_24	585793	99.200000
id_25	585408	99.100000
id_07	585385	99.100000
id_08	585385	99.100000
id_21	585381	99.100000
id_26	585377	99.100000
id_27	585371	99.100000
id_23	585371	99.100000
id_22	585371	99.100000
dist2	552913	93.600000
D7	551623	93.400000
id_18	545427	92.400000
D13	528588	89.500000
D14	528353	89.500000
D12	525823	89.000000

Proposed Methodology



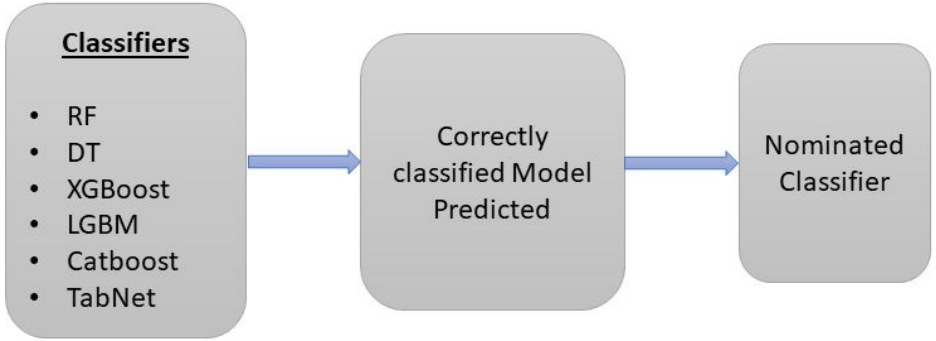
Workflow


Data Collection
kaggle
vesta
IEEE-CIS Fraud
Data Preprocessing
Data Cleaning
Encoding
Feature Engineering
Normalization /standardization
Splitting Dataset
Training Dataset
Testing Dataset
Model Optimization & Training
Hyperparameter Optimization
Hybrid Model
Base Model
RF,DT,XGBoost, Catboost,TabNet
Evaluating the Models
AUC-ROC Score

- In first phase we applied different machine learning classification techniques LR, NB, RF, XGBOOST, LGBM and Catboost applied to detect fraudulent transactions and their performance investigated.
- The algorithm with the best performance based on the highest Area Under the Receiver Operating Characteristic (AUROC) metric served as a baseline model. Tree based models shown high performance.
- In second phase, did basic feature engineering which need to be extracted.

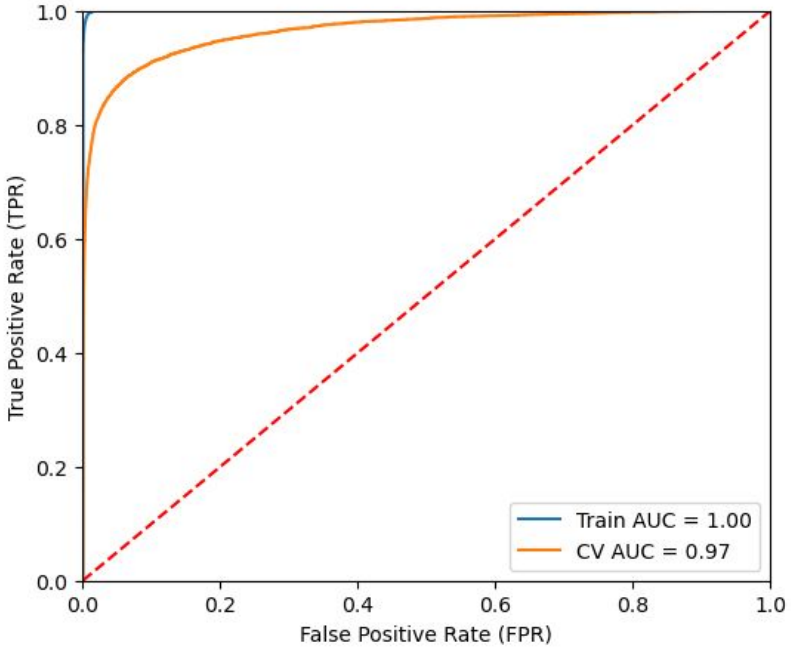
  ➔ Card, DeviceType
  ➔ yearly/monthly/weekly transaction
  ➔ Statistics feature (mean,median,max, percentiles etc)

- Train each model on featured engineered dataset.
- Build Hybrid Model with all classifiers and final output decided by Voting Classifier.



Results and Discussion

AUC-ROC curve of Hybrid Model.



Comparison table of various ML models and the proposed Hybrid models.

Models	Precision	Recall	F1-Score	AUC-ROC
base model				
DT	0.55	0.55	0.55	0.7676
RF	0.95	0.42	0.58	0.7085
XGBoost	0.29	0.82	0.43	0.8746
LGBM	0.21	0.82	0.34	0.8544
Catboost	0.38	0.84	0.52	0.8956
feature engg.				
DT	0.6	0.59	0.59	0.7862
RF	0.96	0.56	0.7	0.7776
XGBoost	0.32	0.84	0.47	0.8897
LGBM	0.23	0.82	0.36	0.8614
Catboost	0.43	0.86	0.57	0.9087
TabNet	0.17	0.86	0.28	0.8545
Hybrid Model	0.69	0.76	**0.73**	**0.9661**

References

[1] Sercan O. Arik and Tomas Pfister, "Tabnet: Attentive interpretable tabular learning", arXiv, 2020.

[2] Kamusweke K, Nyirenda M, Kabemba M.,”Data mining for fraud detection in large scale financial transactions”. EasyChair, 2019

[3] Yeming Chen, Xinyuan Han, “CatBoost for Fraud Detection in Financial Transactions”,IEEE (ICCECE), 2021.

[4] Qizhi Cai; Jiaxin He Credit Payment Fraud detection model based on TabNet and Xgboot. 2022

[5] . Vengatesan et al., "Credit card fraud detection using data analytic techniques", Advances in Mathematics: Scientific Journal, vol. 9, no. 3, pp. 1185-1196, 2020.

Thanks !