TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

IEEE-CIS FRAUD DETECTION

Aadithya Anandaraj
Chetna Singhal
Matt Viteri
Shikha Singh
Sindhu Patnam

Agenda

INTRODUCTION - BUSINESS PROBLEM

EXPLORATORY DATA ANALYSIS

DATA PRE-PROCESSING

FEATURE SELECTION & ENGINEERING

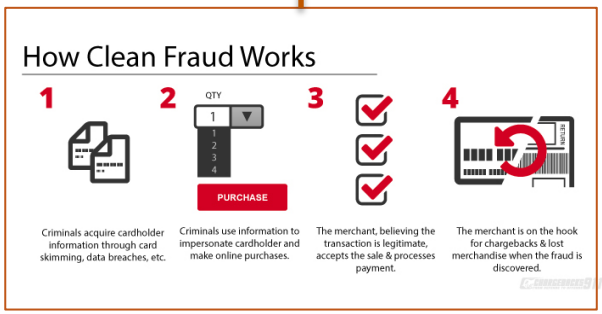
MODEL PERFORMANCE COMPARISON

DEMO & FOOD FOR THOUGHT

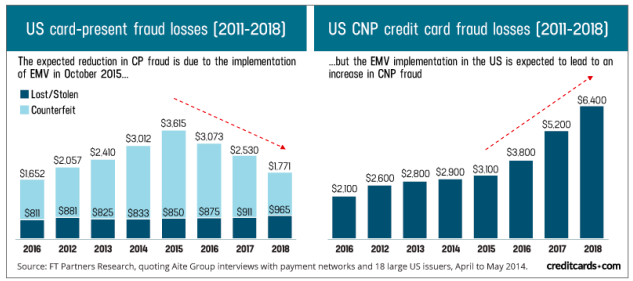
Introduction – Business Problem

Detect **fraudulent** e-commerce payment transactions

Prevention is **cheaper** than cure



Challenge: card-not-present (CNP)



Improve **customer experience**

BUSINESS

DATA SCIENCE

Supervised Classification:
Predict probability of fraud for each incoming transaction

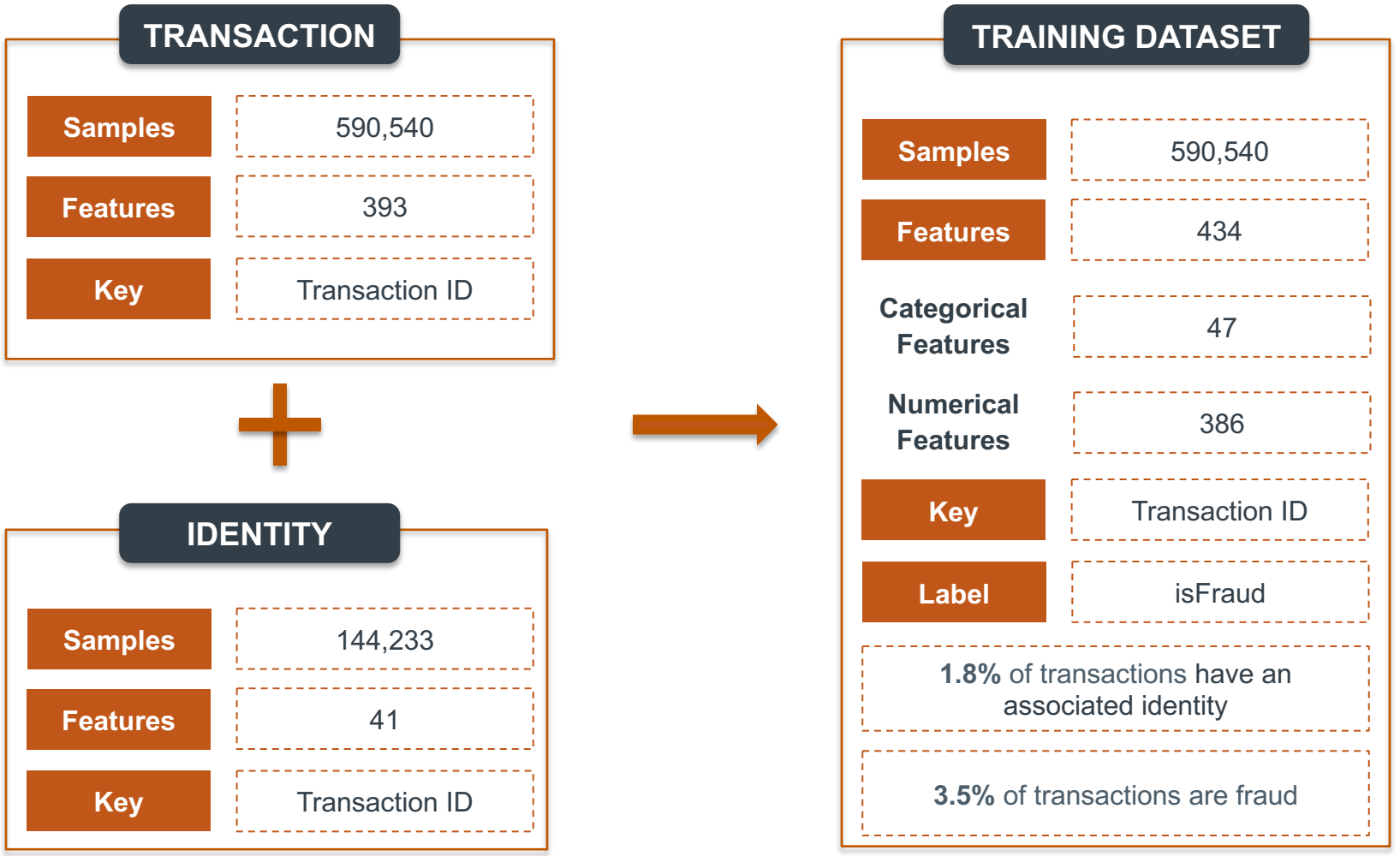
Higher **accuracy** fraud detection

Reduce **false positive** alarms

Source: printful.com

Data Description

Source : Kaggle competition co-hosted by IEEE-CIS & Vesta Corporation

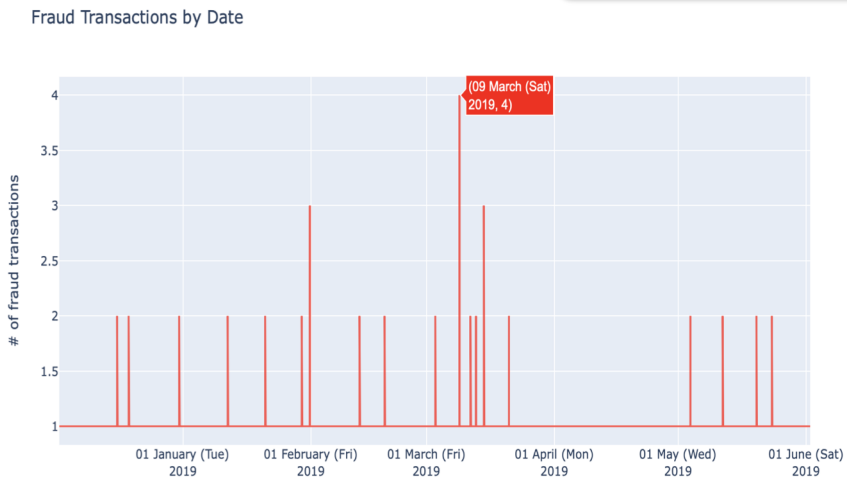


Exploratory Data Analysis - I

CATEGORICAL

NUMERICAL

TRANSACTION TIMEDELTA



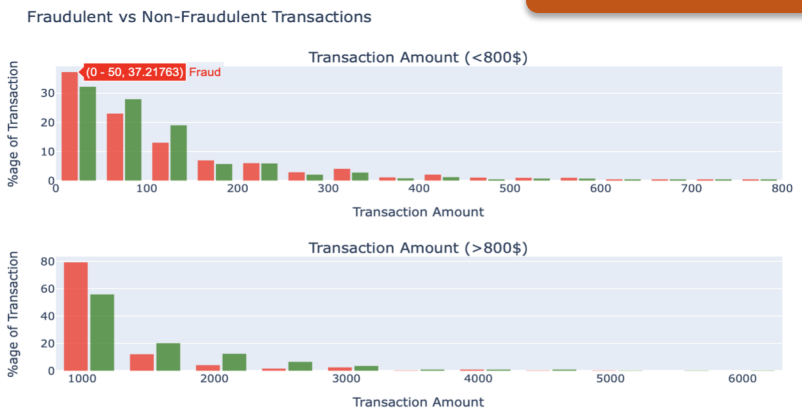
DESCRIPTION

Time delta from a given reference datetime
Transaction dataset provided for 6-month period

OBSERVATION

Multiple fraudulent transactions occurring
in a short period of time

TRANSACTION AMT



DESCRIPTION

Transaction payment amount in USD

OBSERVATION

Mean transaction amount for fraud is $149.2
Mean transaction amount for non-fraud is $134.5

Exploratory Data Analysis - II

CATEGORICAL

NUMERICAL

PRODUCT CODE

DESCRIPTION

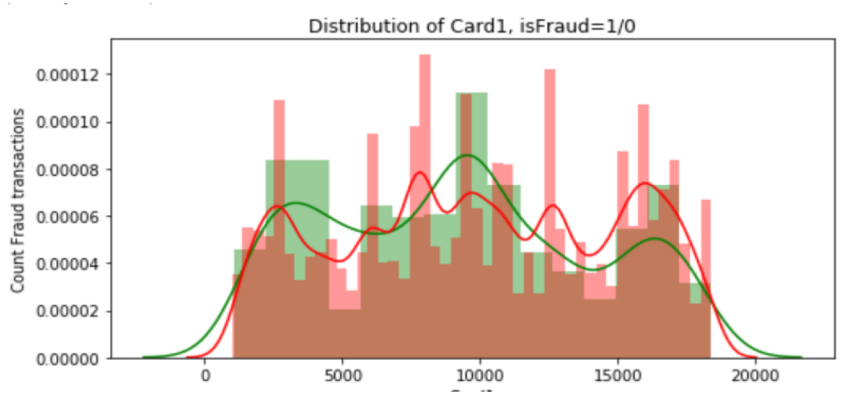
Product code, the product for each transaction

OBSERVATION

Product C has the most fraud >13%

Product W has the least fraud ~2%

CARD INFORMATION



DESCRIPTION

Payment card information, such as card type, card category, issue bank, country, etc.

OBSERVATION

Card 1 - high cardinality categorical variable

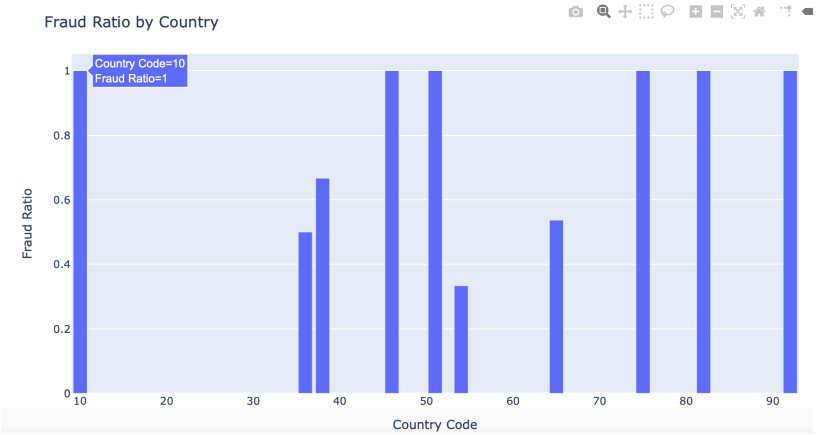
Might need to convert to numerical values

Exploratory Data Analysis - III

CATEGORICAL

NUMERICAL

ADDRESS



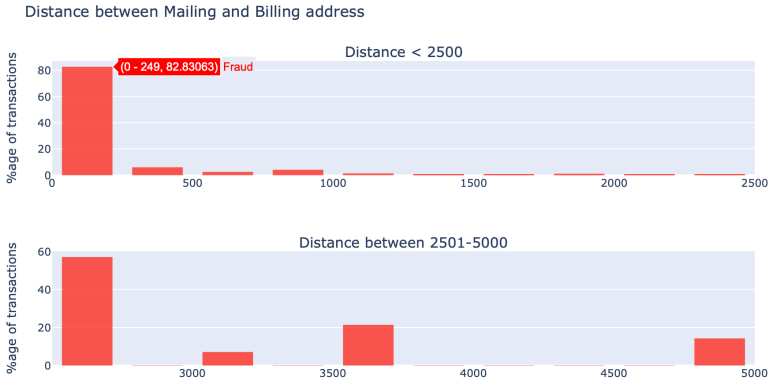
DESCRIPTION

Address of purchaser (billing region & country)

OBSERVATION

Fraudulent transactions concentrated in few country codes [10, 51, 75, 82, 92 and 46]

DISTANCE



DESCRIPTION

Distances between (not limited) billing address, mailing address, zip code, IP address, phone area, etc.

OBSERVATION

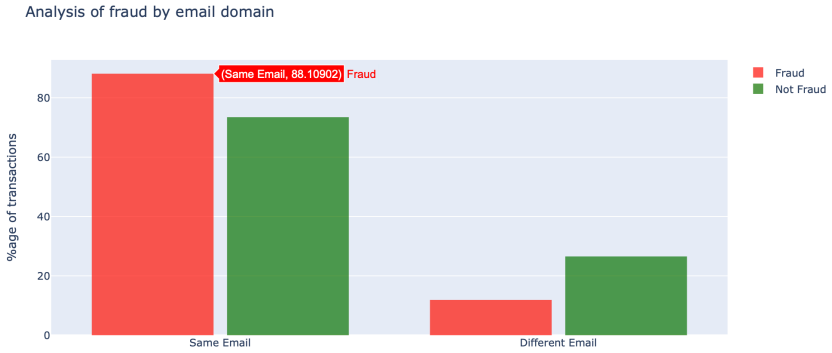
Distance between mailing and billing address is mostly low for fraudulent transactions

Exploratory Data Analysis - IV

CATEGORICAL

NUMERICAL

EMAIL DOMAIN



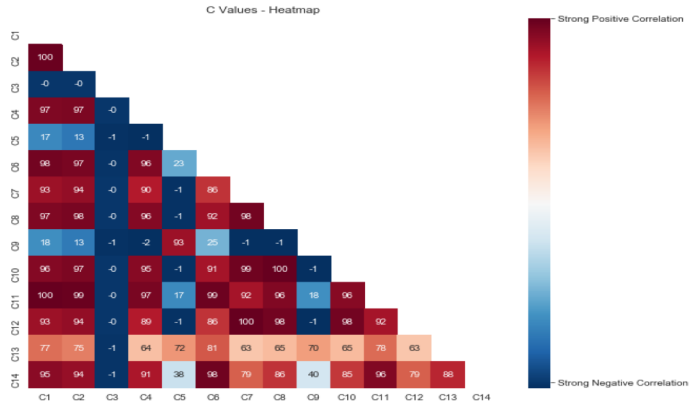
DESCRIPTION

Purchaser and Recipient e-mail domain

OBSERVATION

Same e-mail id for purchaser and recipient – indicative of fraud transactions

C1 – C14



DESCRIPTION

Vesta engineered feature : Count of addresses associated with the payment card, etc.

OBSERVATION

Many of these features are highly correlated with each other
Feature selection important

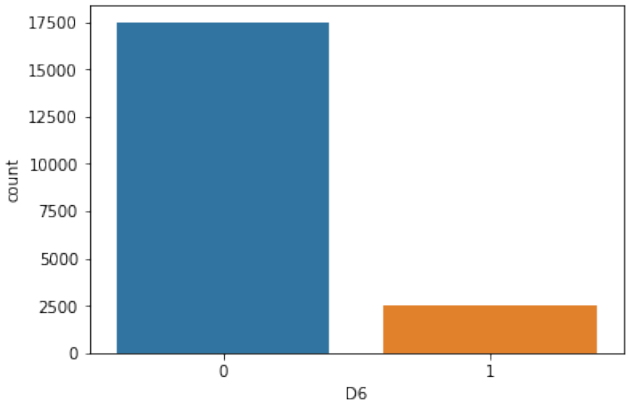
Exploratory Data Analysis - V

CATEGORICAL

NUMERICAL

D1 – D15

% Missing
D1 - 0.0%
D2 - 48.0%
D3 - 45.0%
D4 - 29.0%
D5 - 52.0%
D6 - 88.0%
D7 - 93.0%
D8 - 87.0%
D9 - 87.0%
D10 - 13.0%
D11 - 47.0%
D12 - 89.0%
D13 - 90.0%
D14 - 89.0%
D15 - 15.0%



DESCRIPTION

Vesta engineered feature

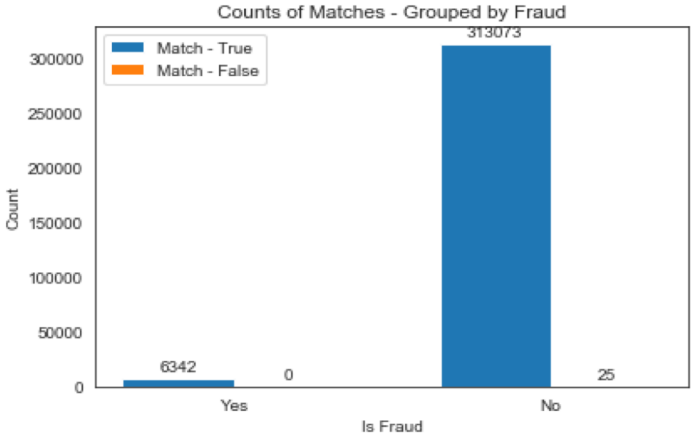
Time delta - days between previous transaction

OBSERVATION

A lot of these columns contain missing data

Feature selection important

M1 – M9



DESCRIPTION

Vesta engineered feature

Match with names on card and address, etc.

OBSERVATION

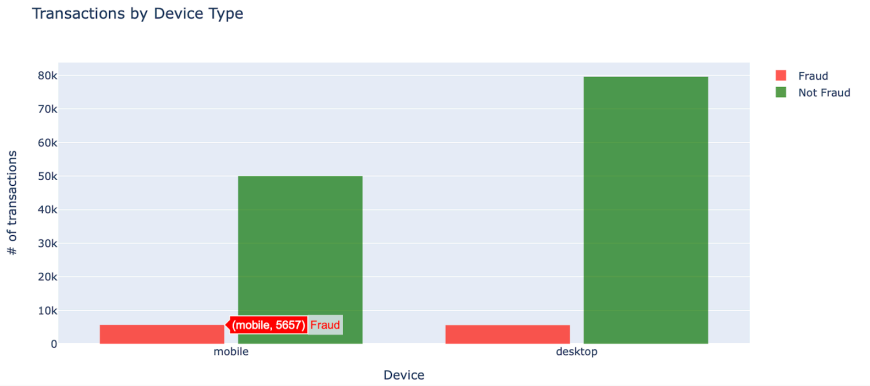
Fraudsters have all the matching information!!

Exploratory Data Analysis - VI

CATEGORICAL

NUMERICAL

DEVICE TYPE



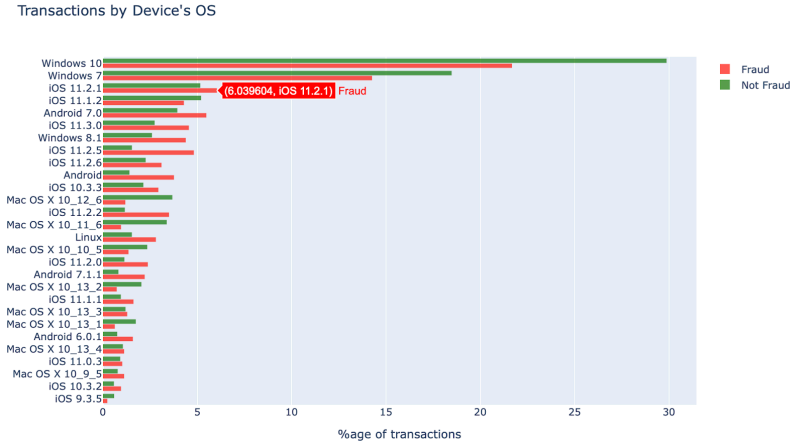
DESCRIPTION

Device used to make transaction

OBSERVATION

Mobile witnessed higher % of fraud transactions than desktop

DEVICE INFO



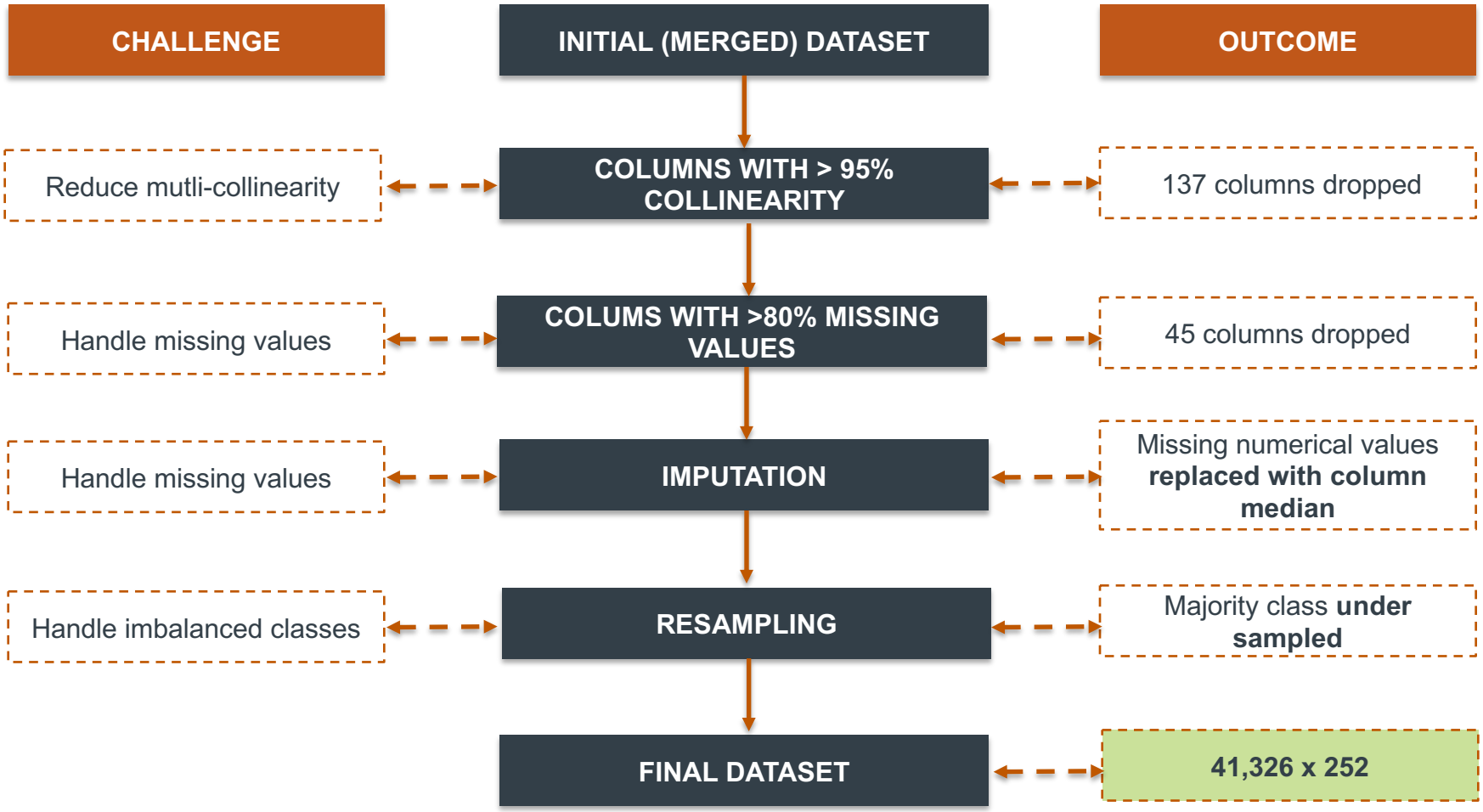
DESCRIPTION

Device OS version used to make the transaction

OBSERVATION

Higher fraud observed on Windows than Mac

Data Pre-Processing



Feature Selection & Engineering

FEATURE SELECTION — GOAL - Reduce number of features to avoid “The Curse of Dimensionality”

- Recursive Feature Elimination **(RFE)** from sklearn
- **Chi2 & F_classif** from sklearn
- **Forward Selection** Sequential Feature Selector

FEATURE ENGINEERING — GOAL - To capture more information and improve performance of the model

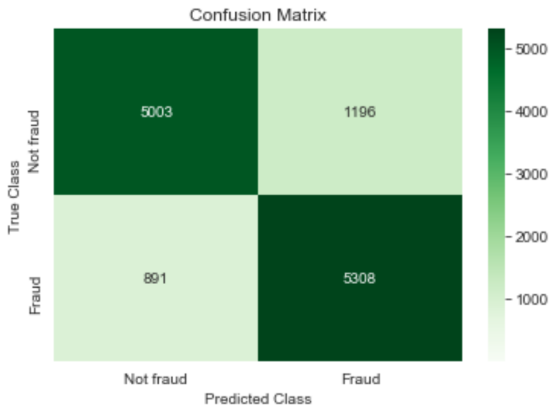
- **>200** engineered features provided by Vesta
- **Interaction terms** card and address columns
- Split device **OS type and version** to create 2 new features

Performance Comparison - ML Models

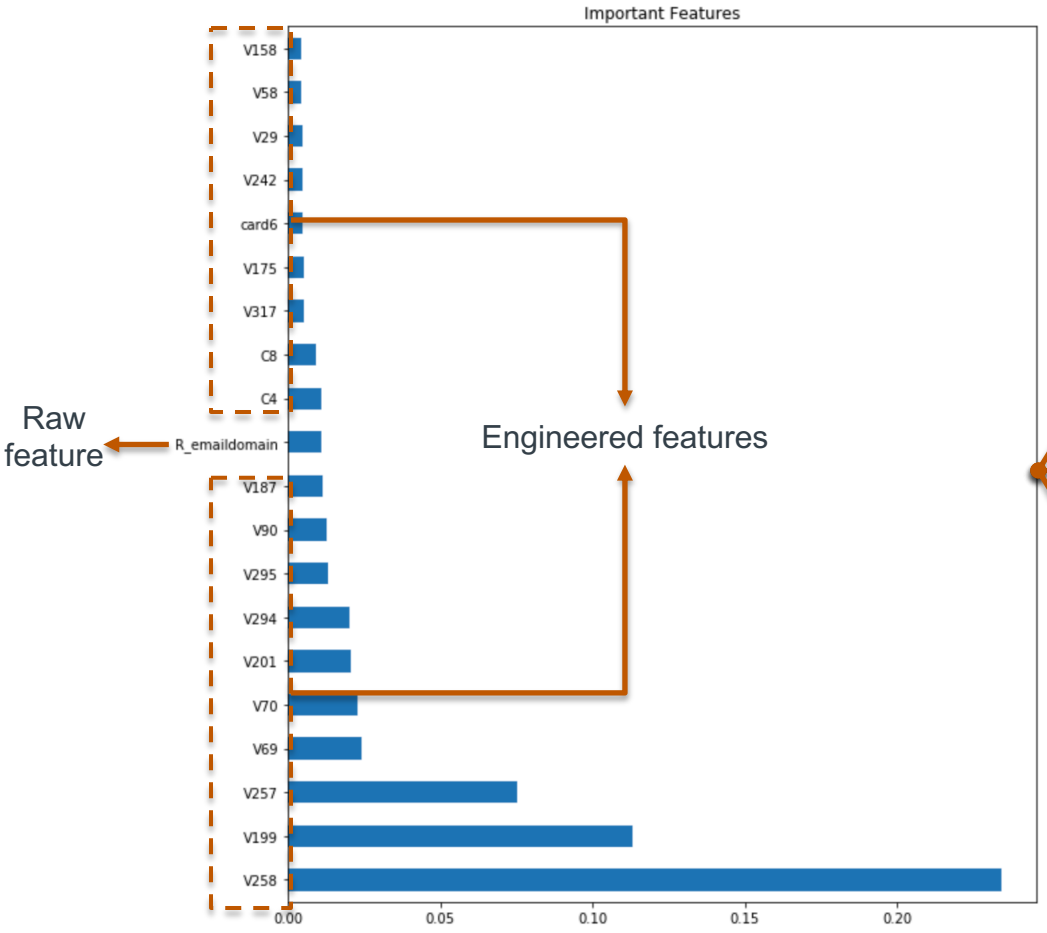
	LOGISTIC REGRESSION	RANDOM FOREST	LIGHTGBM	LIGHTGBM + XGBOOST
HYPERPARAMETER TUNING	L2 regularization alpha = 0.001	N_estimators Max_features Min_samples_leaf	Max_depth Num_leaves Learning_rate	Max_depth Learning_rate Min_child_weight
KAGGLE SCORE (ROC AUC)	0.87	0.91	0.92	0.93

OBSERVATIONS

- **Light GBM** proved **computationally efficient** with large data sets
- **XGBoost** learnt slowly but performed the **best**!
- Here, **Recall – a better score** than AUC as misclassifying a fraud Tx costs us more, so we want to increase the TPR and decrease the FPR
- Adjusting the threshold for prediction, we increased the recall score to **correctly capture more fraud cases** (0.86 recall at threshold = 0.4)



Feature Importance



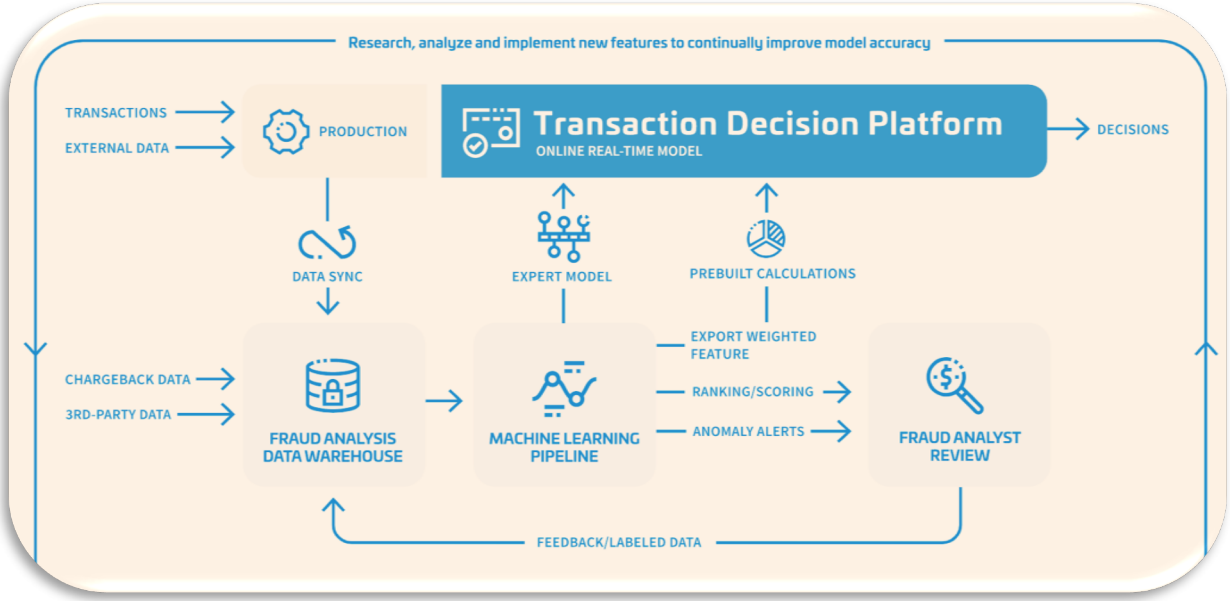
Gives how useful or valuable each feature was in solving the classification problem

Calculated from the performance improvement derived from each attribute split
(in our case Gini index)

It is weighted by the number of observations that the node is responsible for

Demo & Food for Thought

ARCHITECTURE



Source: trustvesta.com

DEMO LINK

https://fraud-score.appspot.com

NEXT EPOCH

Address **"Algorithmic Prison"**


thank you