CZ4041 Machine Learning

IEEE – CIS Fraud Detection Kaggle Project

Agenda

1. Introduction

2. Exploratory Data Analysis

3. Data Pre-processing

4. Modelling and Results

Team Members

Jason

Kenneth

Yong Wei

Huan Zhang

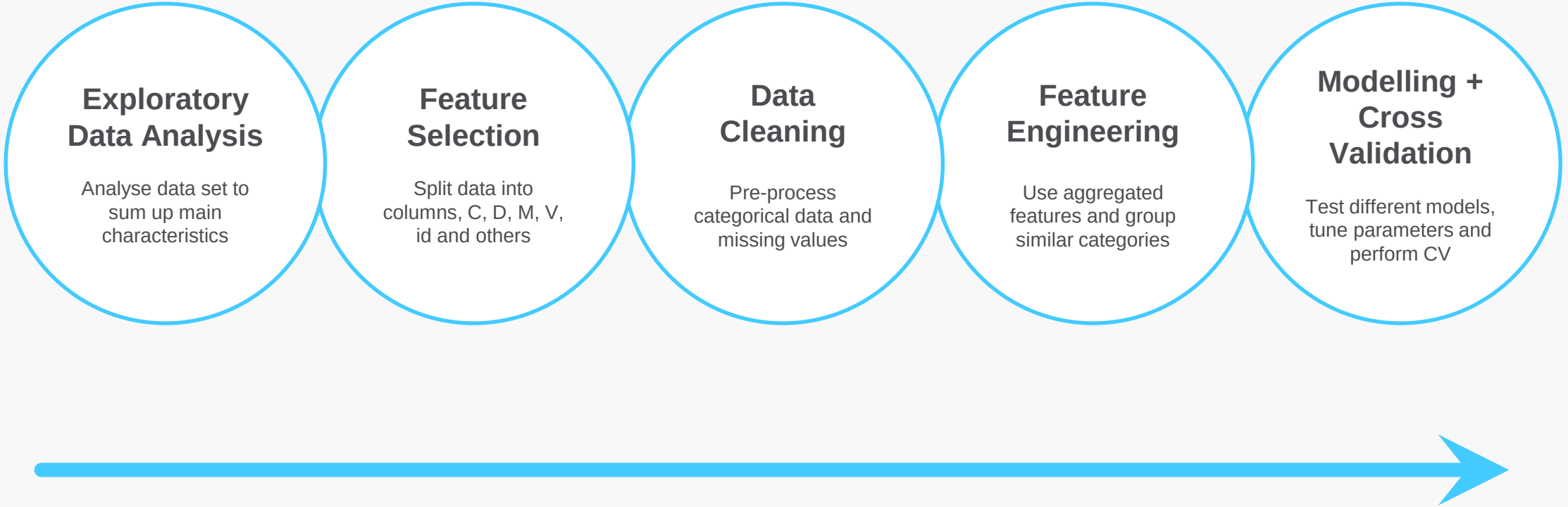
Problem Statement

The aim of the challenge is to benchmark machine learning models on a challenging large-scale dataset to predict if a transaction made is fraudulent.

The dataset of credit card transactions is provided by the Vesta Corporation, described as the world's leading payment service company.

The dataset includes identity and transaction CSV files for both test and train.

Overview


Exploratory Data Analysis
Analyse data set to sum up main characteristics
Feature Selection
Split data into columns, C, D, M, V, id and others
Data Cleaning
Pre-process categorical data and missing values
Feature Engineering
Use aggregated features and group similar categories
Modelling + Cross Validation
Test different models, tune parameters and perform CV

Exploratory Data Analysis

Exploratory Data Analysis

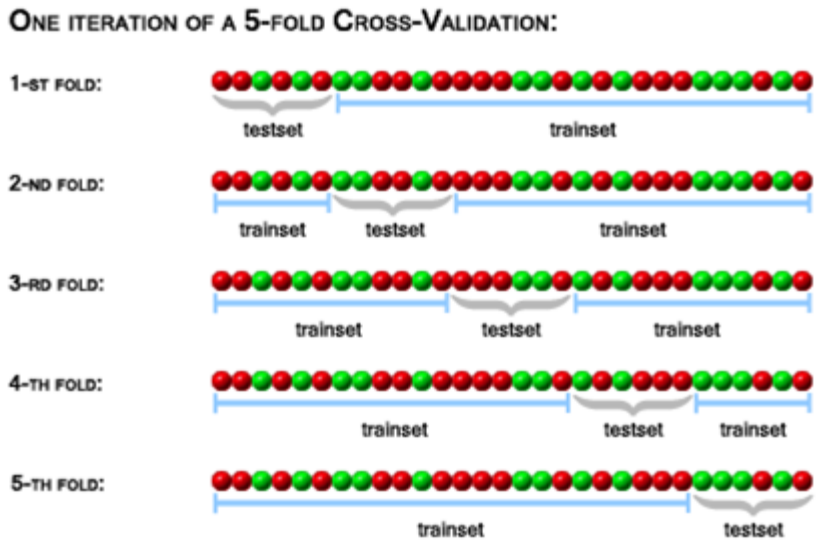
- Train dataset: 590,540 x 434
- Fraud transactions: 20663
- Target variable ‘isFraud’

Explanation of Variables

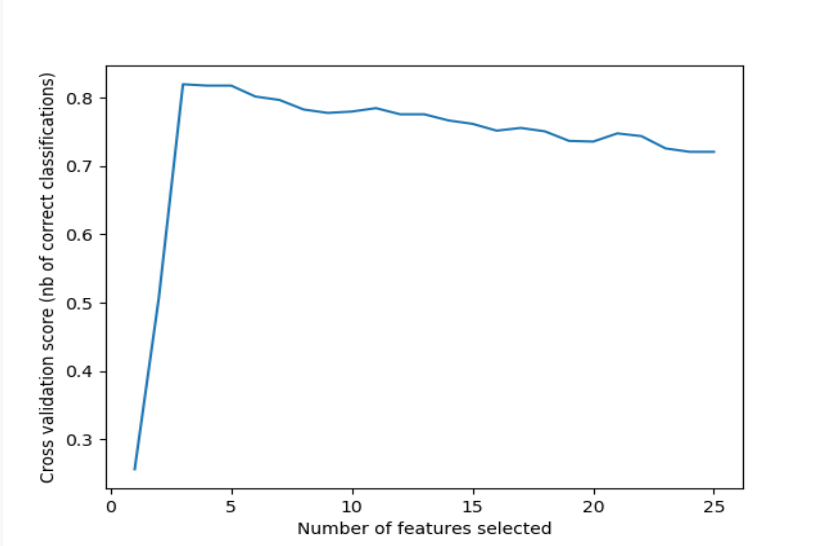
FEATURE	DESCRIPTION
TransactionDT	Timedelta from a given reference datetime (not an actual timestamp)
TransactionAmt	Transaction amount in USD
ProductCD	product code, the product for each transaction
card1-card6	Card used for payment
addr	address
dist	distance
P_ and (R_) email domain	purchaser and recipient email domain
C1-C14	counting, address and other things, actual meaning is masked
D1-D15	timedelta, such as days between previous transactions. etc.
M1-M9	match, such as names on card and address, etc
V1-V339	Vesta engineered rich features, including ranking, counter and other entity relations

Data Pre-processing

Feature Selection method



Cross validation



Recursive feature elimination

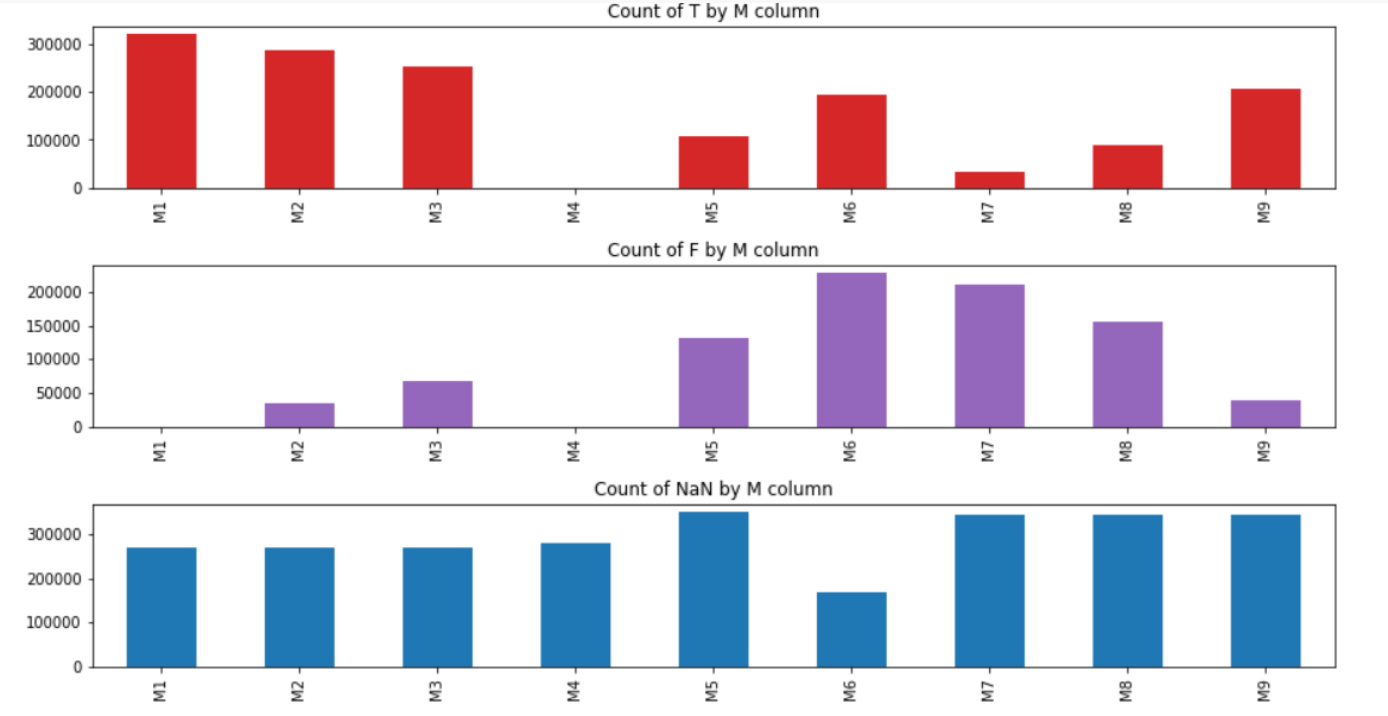
Feature Selection method

Correlation Analysis

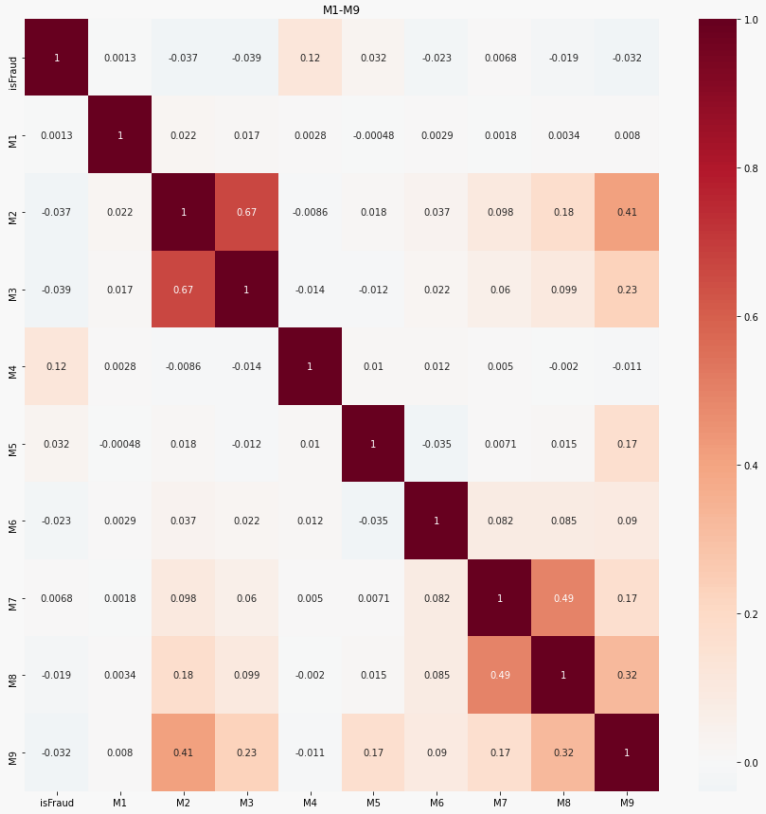
V Feature Selection

- V features have too many columns
- Relationship between features are vague
- Use RFE with cross validation to siphon out V features
- Preprocessing:
  - remove all the columns with high Nan values
- V features of RFE ranking of 1 are selected

M Feature Selection

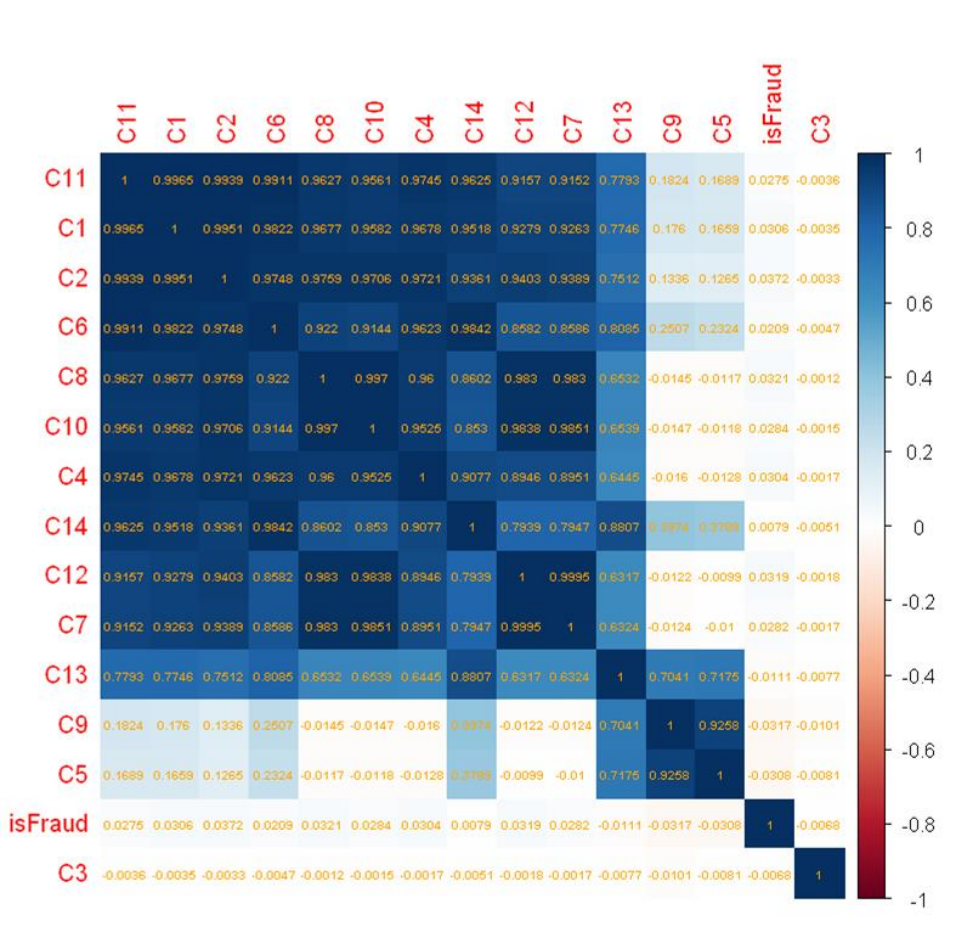


M feature bar chart



Correlation Heatmap for M

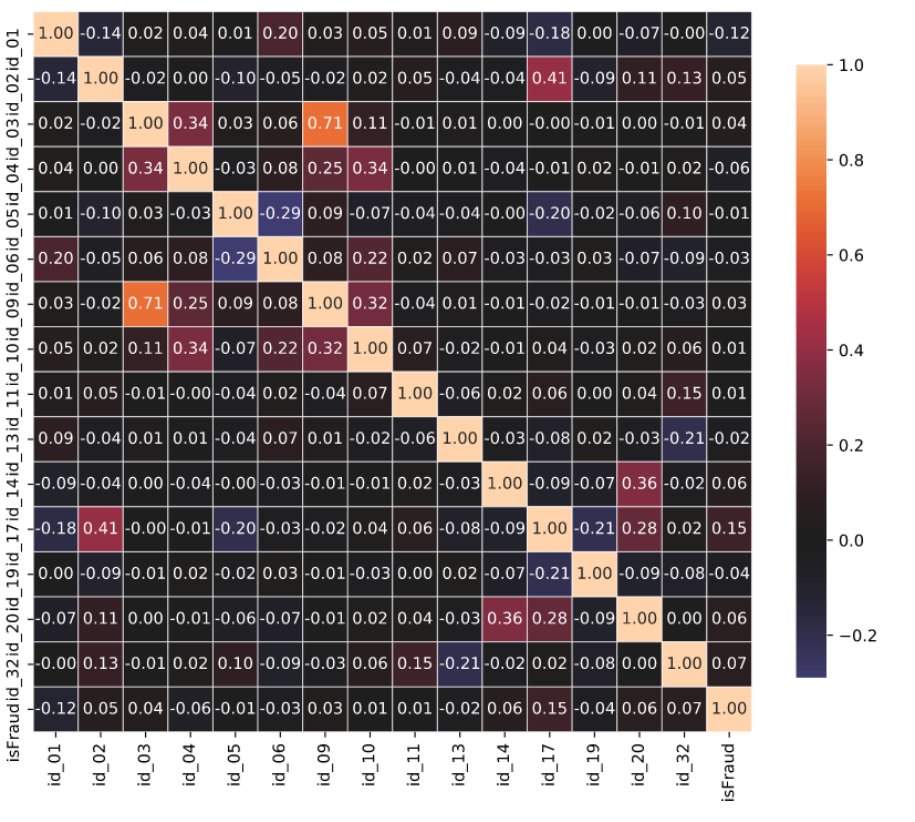
C Feature Selection

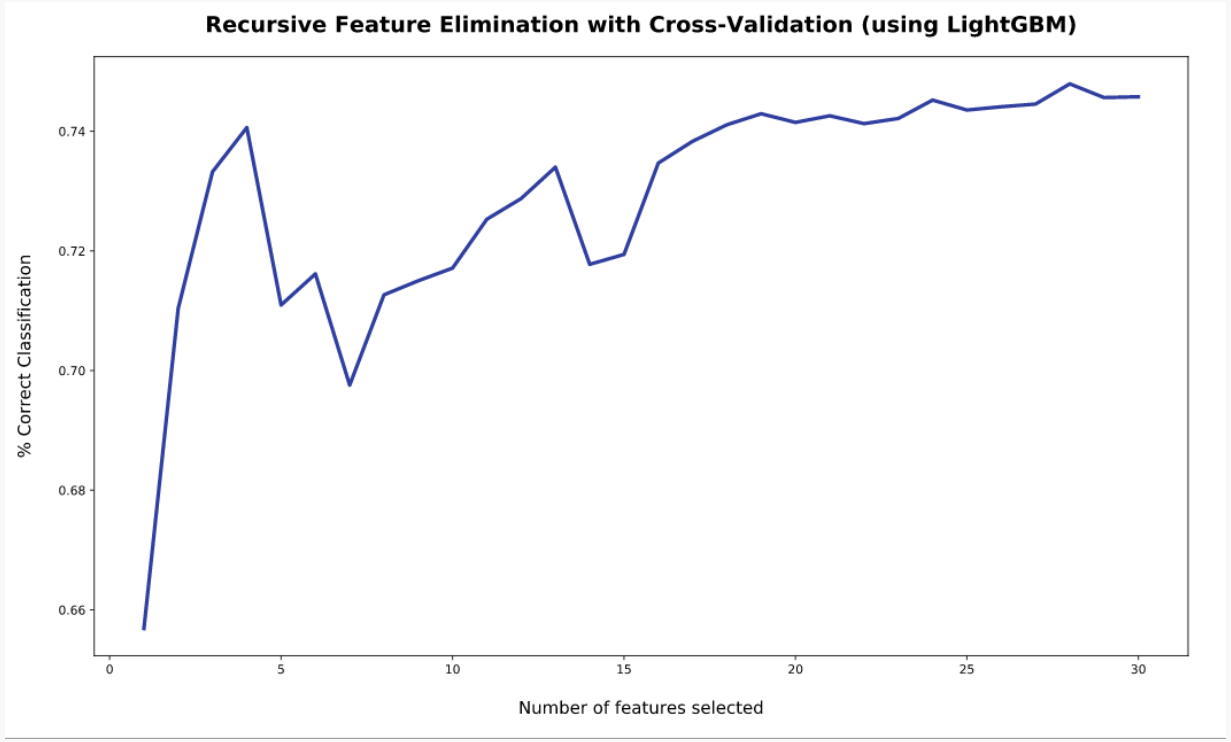


Heatmap for C features

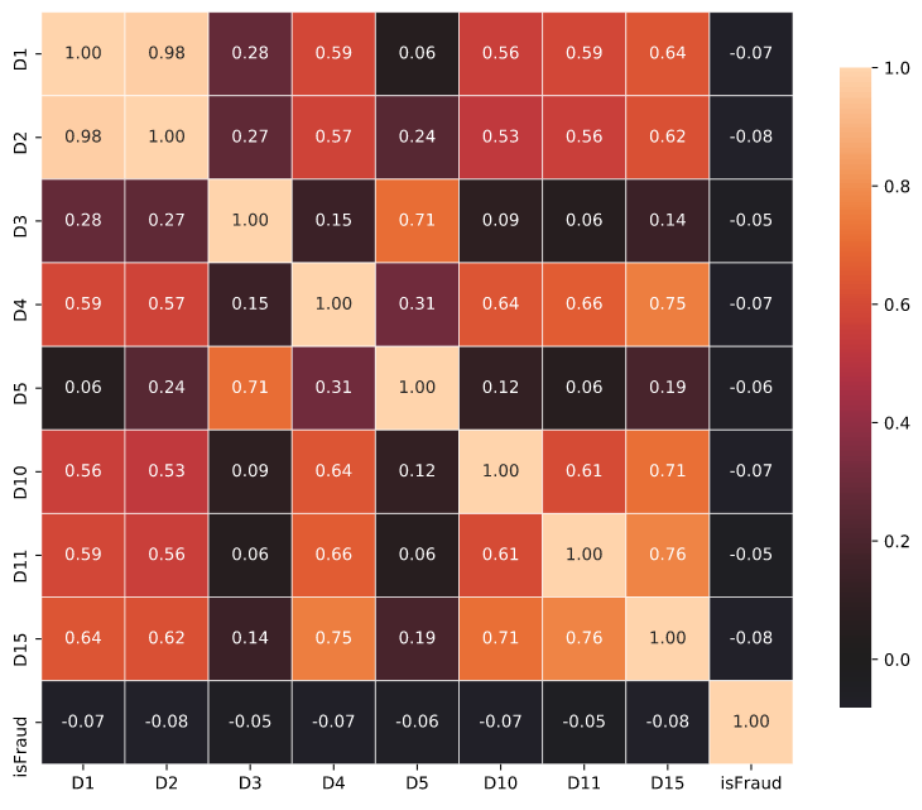
- Just like V, C features are vague
- We can use correlation heatmap to determine relationship of each features

ID Feature Selection





D Feature Selection



Explanation for D features

- Just like V, C features are vague
- We can use correlation heatmap to determine relationship of each features

Pre-processing Features

Categorical Feature Representation

- Treat as categorical features
- Label Encoding
- Mixture of both

Normalisation of Numerical Data

Mean Normalisation
of numerical columns

Handling Missing Values

- Let LightGBM handle
- Replace with value outside feature range (-999)
- Mean Imputation
- Median Imputation
- Mode Imputation

Augment OS and Browser data

- Group OS Type
- Group Browser Type
- Group OS Version & Browser Type

Extract transaction day and time

- Extract Time
- Extract Day
- Extract Day and Time

Feature Aggregation

Use aggregated features

Selected Methods

Categorical Feature Representation	Normalisation of Numerical Data	Handling Missing Values
• Treat as categorical features • Label Encoding • Mixture of both	Mean Normalisation of numerical columns	• Let LightGBM handle • Replace with value outside feature range (-999) • Mean Imputation • Median Imputation • Mode Imputation

Augment OS and Browser data	Extract transaction day and time	Feature Aggregation
• Group OS Type • Group Browser Type • Group OS Version & Browser Type	• Extract Time • Extract Day • Extract Day and Time	Use aggregated features

Modelling and Results

Models Used

Type / Model	LightGBM	XGBoost	Random Forest
Validation Score (AUC)	0.971931	0.975871	0.999903
Kaggle Public Score	0.939755	0.937507	0.898593
Kaggle Private Score	0.911989	0.908053	0.868237

Ensemble Learning

TODO: diagram

Kaggle Score and Rank

CATEGORY	PUBLIC LB SCORE	PRIVATE LB SCORE
Score	0.943850	0.917212
Rank	2744 / 6381	2451 / 6381
Percentile	~43	~38.41

Conclusion