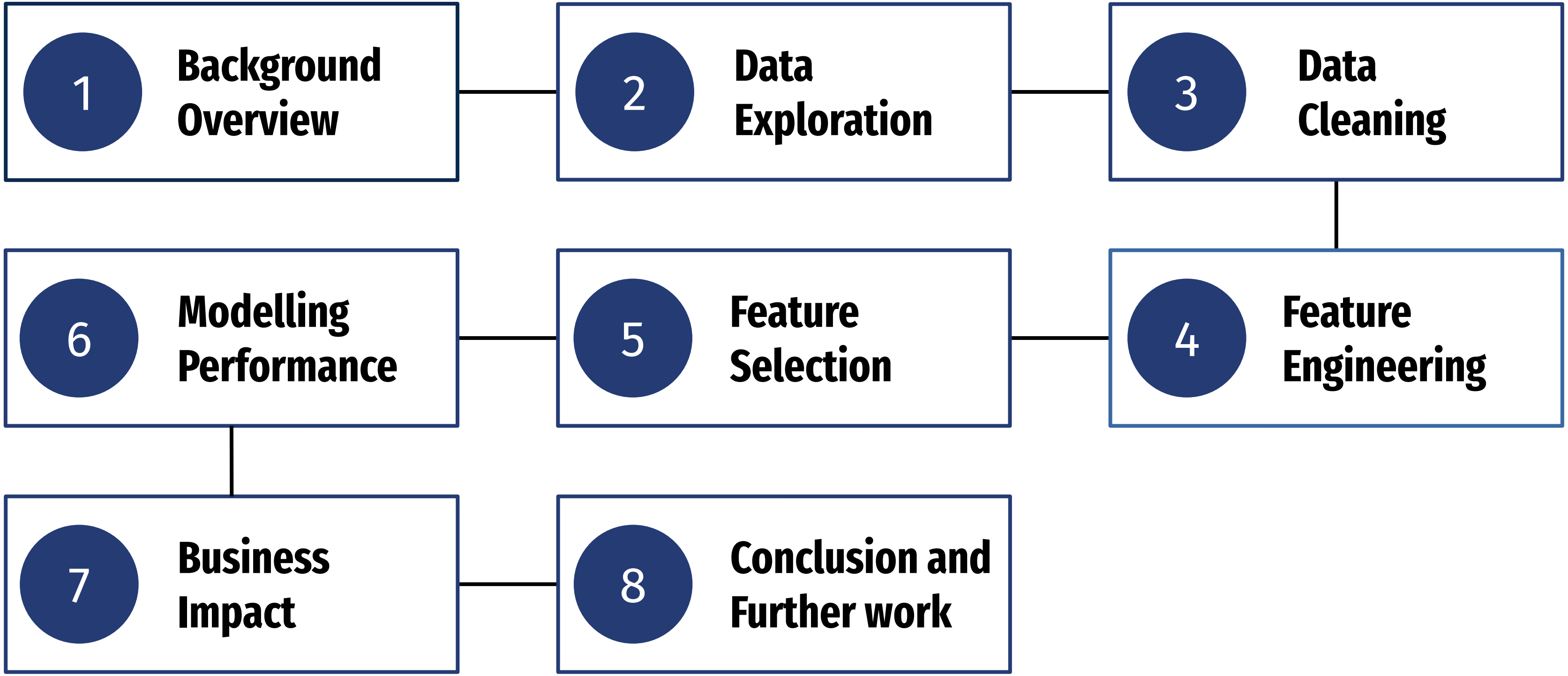
Credit Card Transactions Fraud Detection Analysis

Group 109

ChinKai Huang | Sih-Yu Huang
Shih-Ting Liu | Yi-Ching Lin

Agenda



1 Background Overview

2 Data Exploration

3 Data Cleaning

4 Feature Engineering

5 Feature Selection

6 Modelling Performance

7 Business Impact

8 Conclusion and Further work

Background Overview

Rising Card Fraud Losses

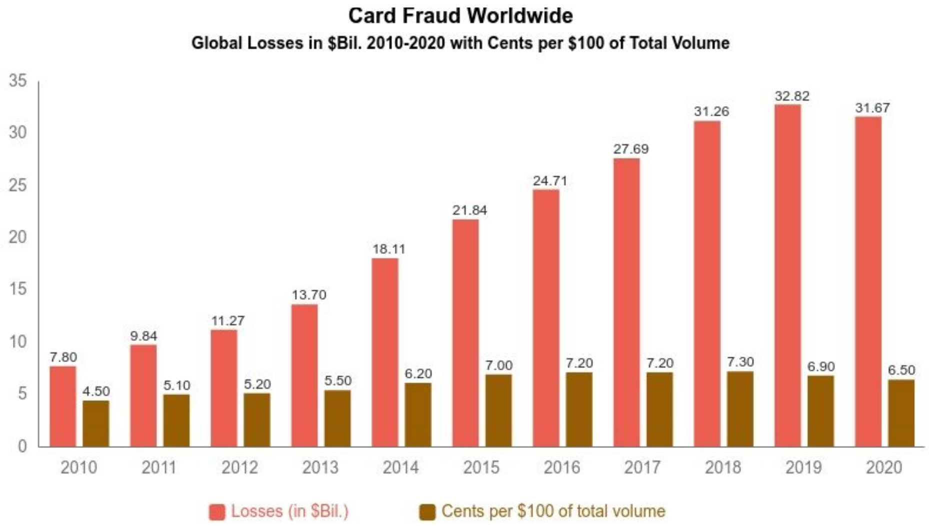
From $7.8B to almost $35B in one decade

More Crooks In Pandemic

Fraudsters do not let an opportunity to slide especially when the country is shutdown

Cheap Credit Card Data

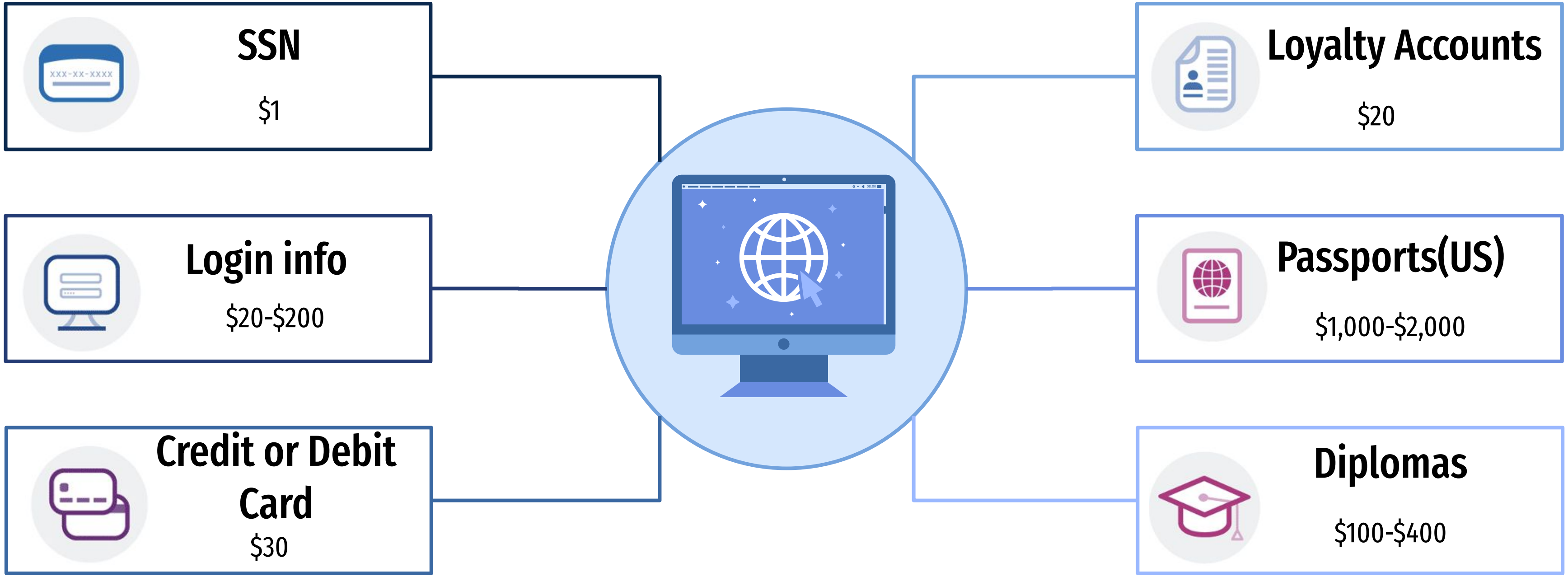
Crooks can steal thousands of dollars before detected



Card fraud global losses keep rising.

Till 2019, losses rocket up to $32.82B.

Background Overview



SSN

$1

Login info

$20-$200

Credit or Debit Card

$30

Loyalty Accounts

$20

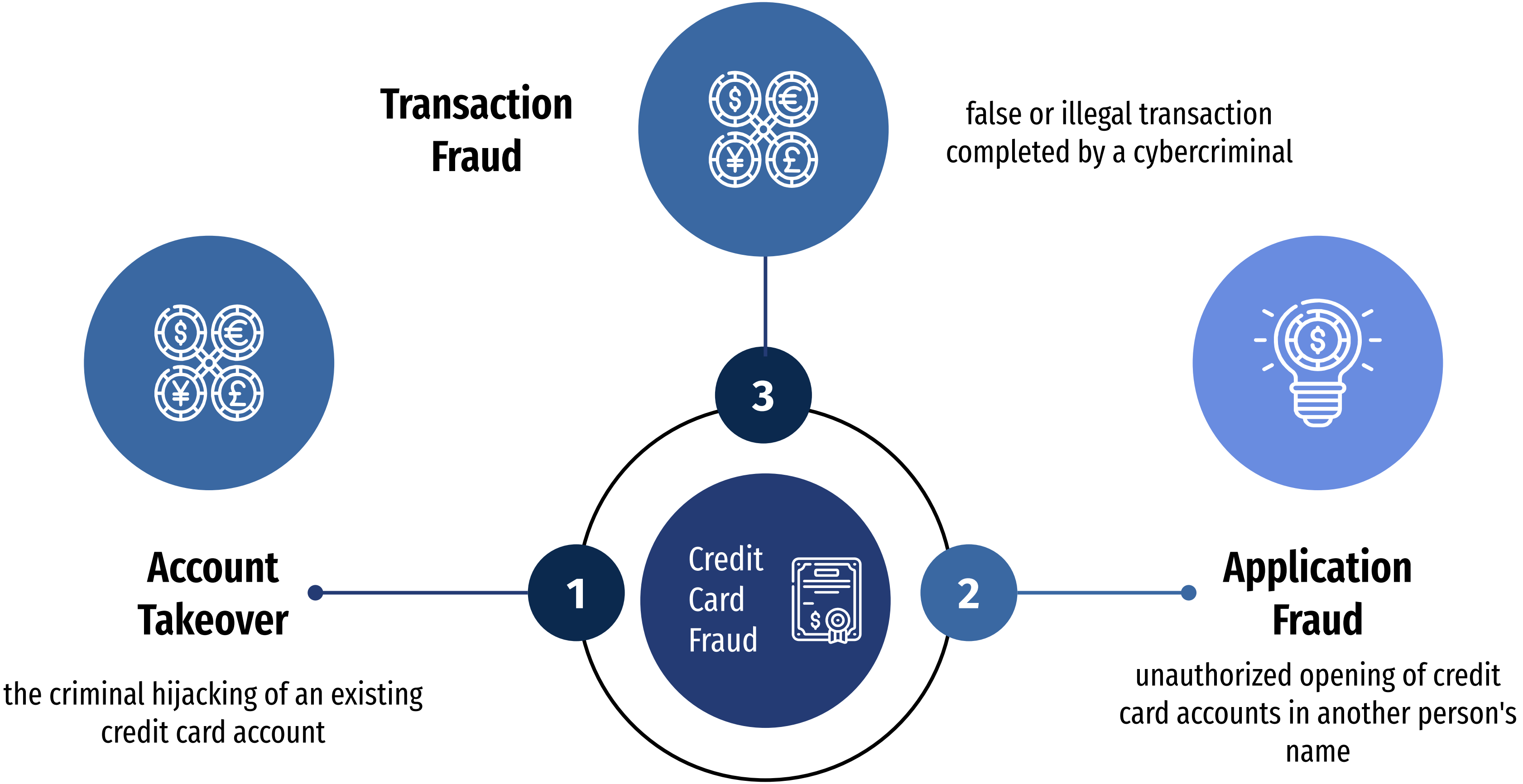
Passports(US)

$1,000-$2,000

Diplomas

$100-$400

Credit Card Frauds



Credit Card Fraud

1

Account Takeover

the criminal hijacking of an existing credit card account

2

Application Fraud

unauthorized opening of credit card accounts in another person's name

3

Transaction Fraud

false or illegal transaction completed by a cybercriminal

Motivations

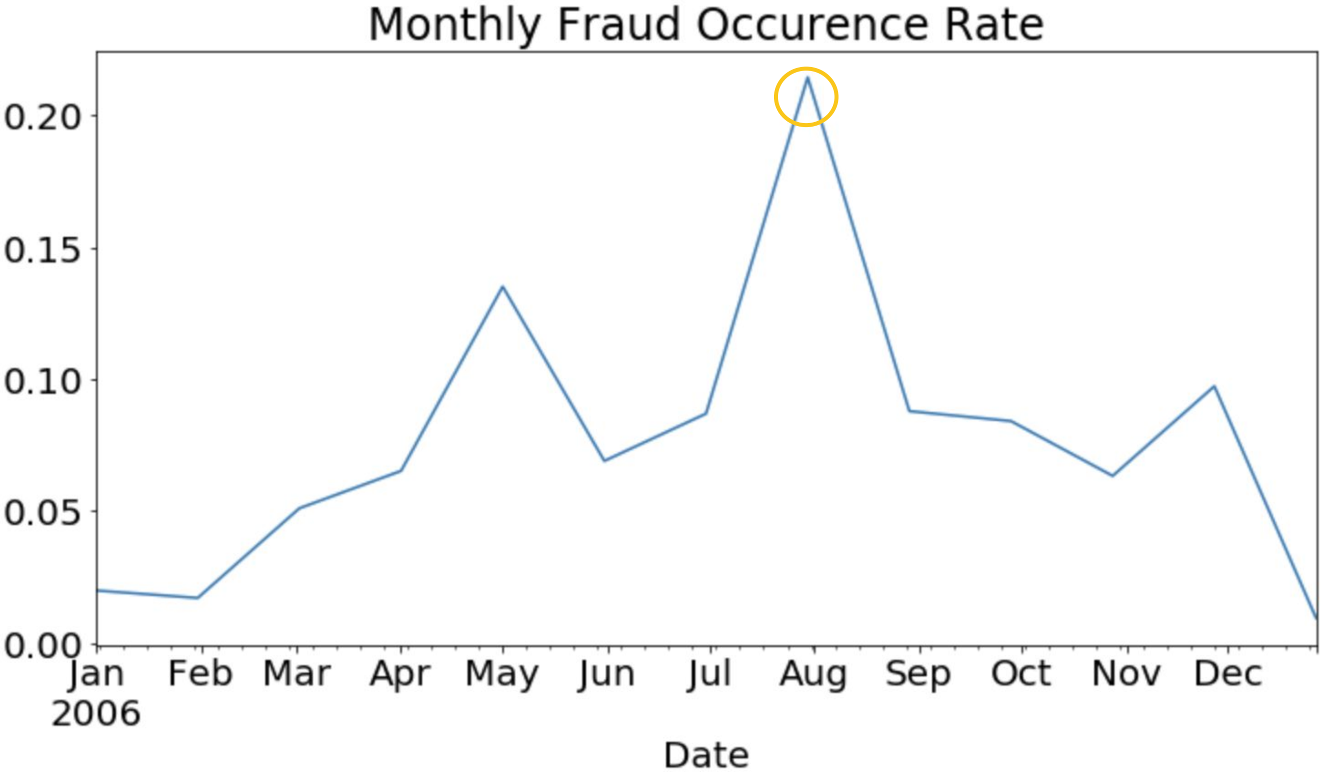
> “ Eliminate 59.4% of the fraud by declining 3% of the transactions ”

Motivations

> “Eliminate 59.4% of the fraud by declining 3% of the transactions”

save **$1.2 M** annualy

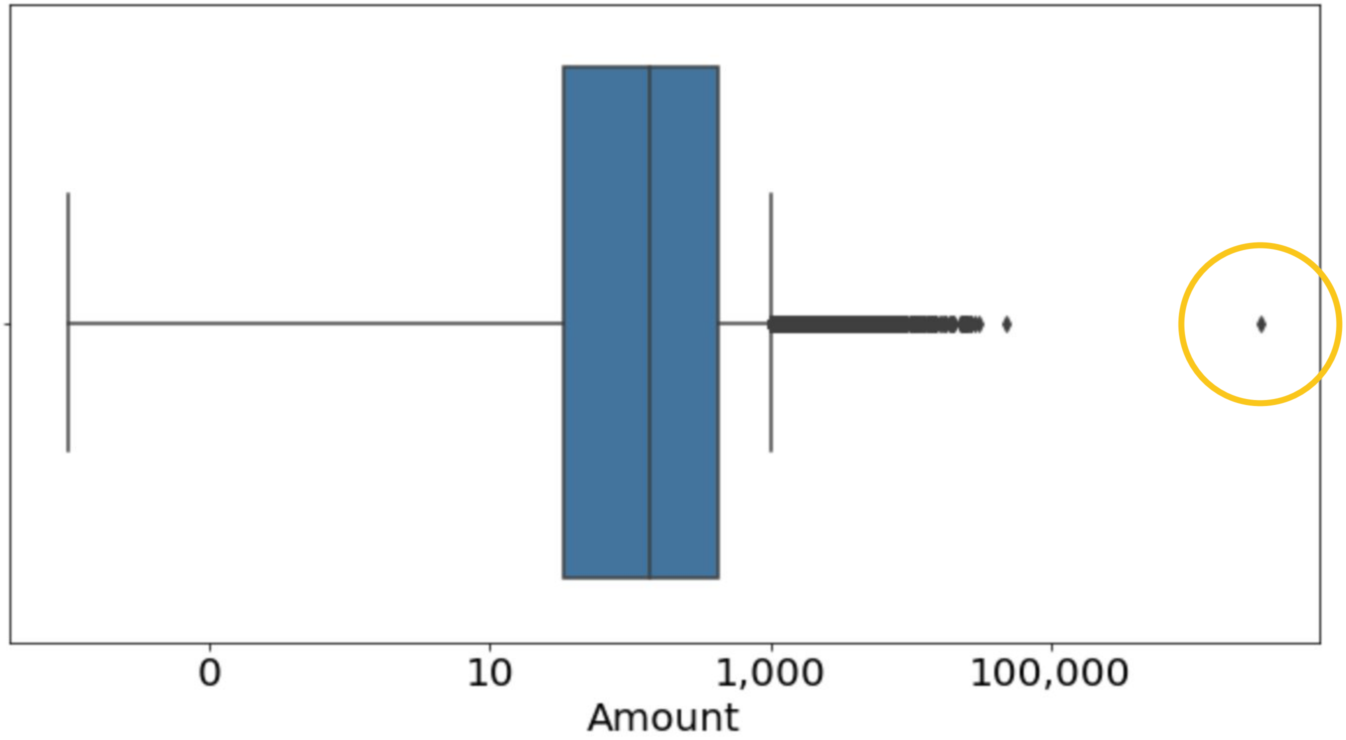
Data Exploration



Data Exploration— Numerical

Field Name	% Populated	Min	Max	Mean	Stdev	%Zero
Date	100%	2006-01-01	2006-12-31	N/A	N/A	0
Amount	100%	0.01	3,102,045.53	427.89	10,006.14	0

Data Exploration— Numerical



Data Exploration— Categorical

Field Name	% Populated	# Unique Values	Most Common Values
Recnum	100%	96,753	N/A
Cardnum	100%	1,645	5142148452
Merch description	100%	13,126	GSA-FSS-ADV
Transtype	100%	4	P
Fraud	100%	2	0
Merchnum	**96.51%**	13,091	930090121224
Merch state	**98.76%**	227	TN
Merch zip	**95.19%**	4,567	38118

Data Cleaning

Filtering

- Kept only the transaction type with “P”

- Removed the single high purchase record outlier

Filling in Merchnum

- Replaced 0 with NaN

- Mapped with the “Merch description”

- Filled in the left values with “Unknown”

Filling in Merch State

- Mapped with “Merch zip”, “Merchnum” and “Merch description”

- Filled in the left values with “Unknown”

Filling in Merch Zip

- Mapped with the mode of “Merchnum”, “Merch description”

- Filled in the left values with “Unknown”

Feature Creation

1 **New Entities**

{'card_merch', ‘card_state', 'card_zip', 'card_amount', 'merch_amount', 'merch_zip', 'merch_state'}

2 **Days-since**

Days since a transaction with the entity has been seen

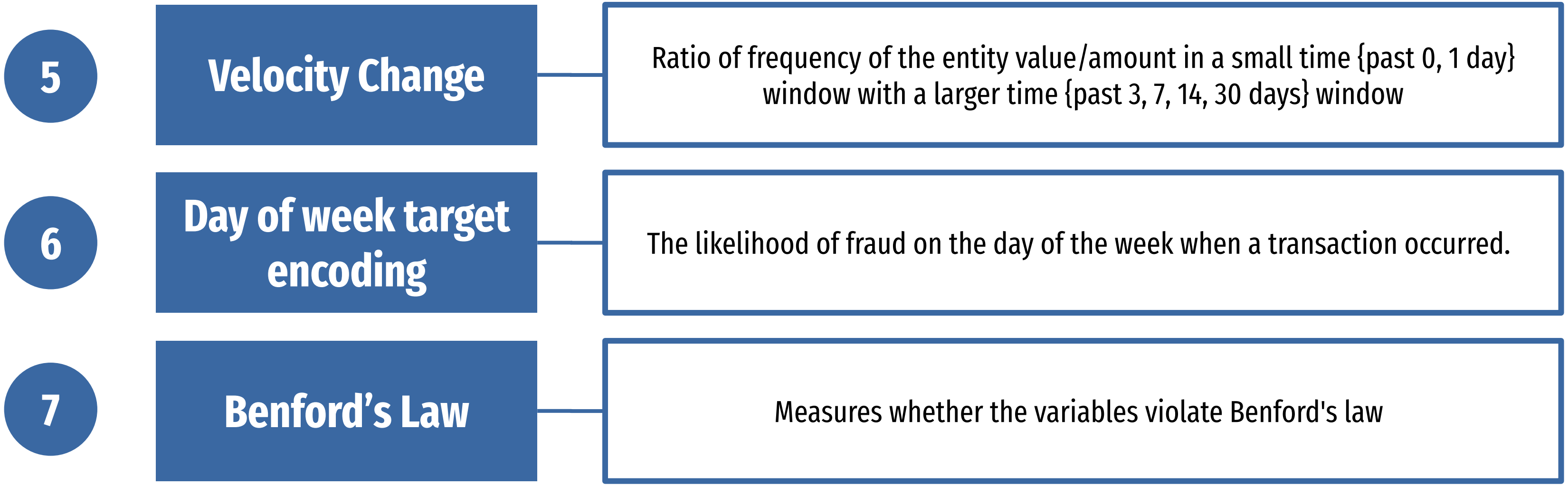
3 **Frequency**

Transaction at the entity over the past {0, 1, 3, 7, 14, 30} days

4 **Amount**

{Average, Maximum, Median, Total, Actual/average, Actual/maximum, Actual/median, Actual/total} amount at the entity over the past {0, 1, 3, 7, 14, 30} days.

Feature Creation



5 **Velocity Change**

Ratio of frequency of the entity value/amount in a small time {past 0, 1 day} window with a larger time {past 3, 7, 14, 30 days} window

6 **Day of week target encoding**

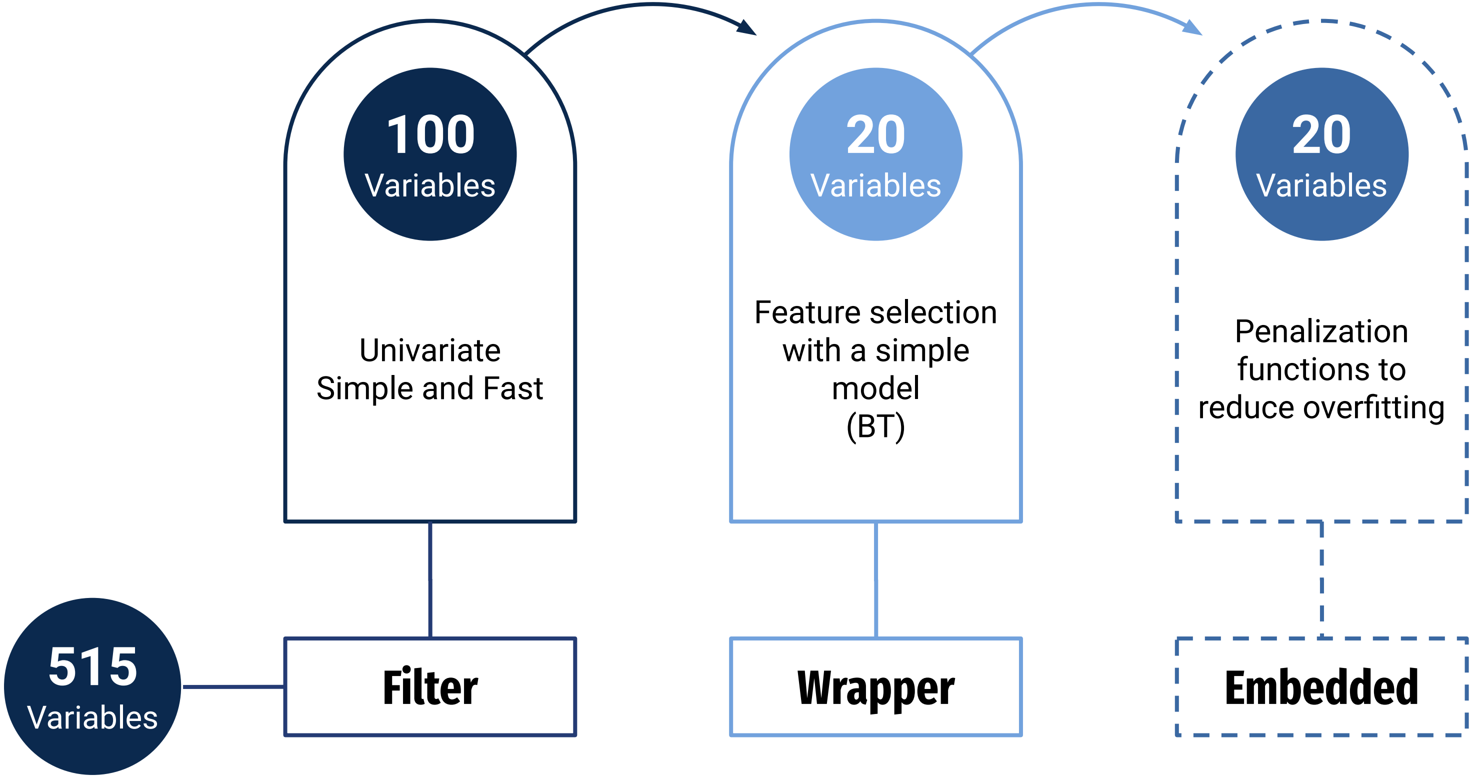
The likelihood of fraud on the day of the week when a transaction occurred.

7 **Benford's Law**

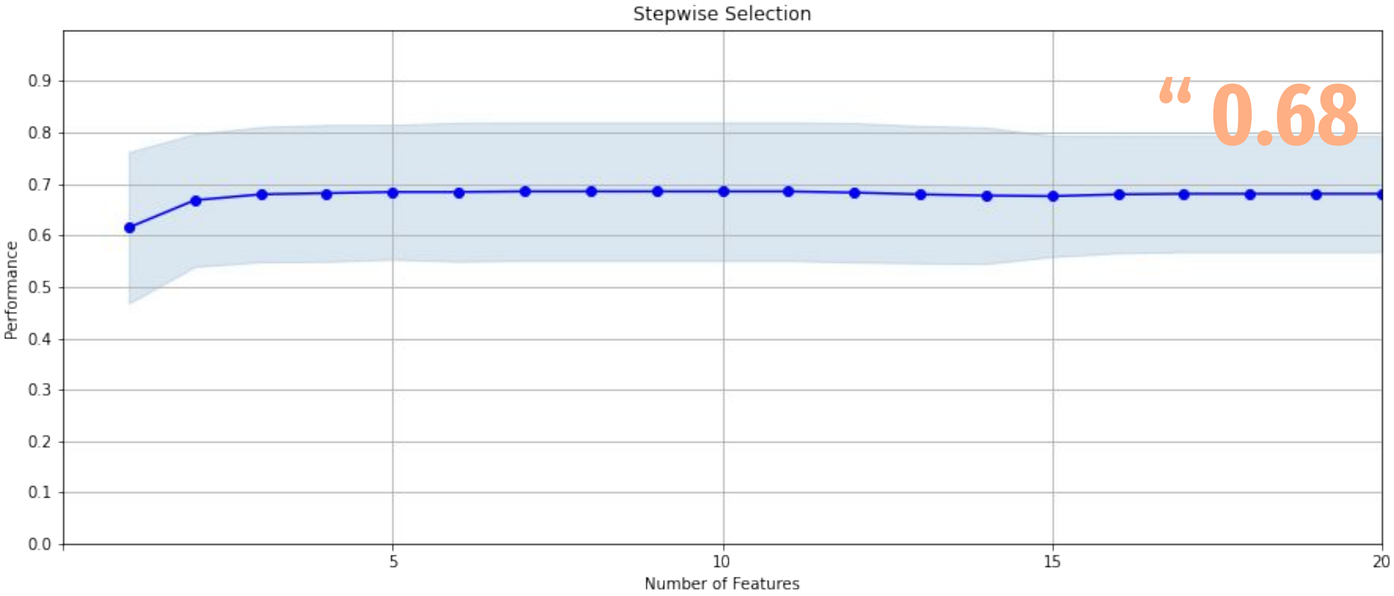
Measures whether the variables violate Benford's law

505 Variables Created in Total

Feature Selection – Process



Feature Selection — Performance



“ 0.68

Feature Selection — Alternative Wrapper

Select K best

select the features according to the 100 highest score

# of Features: 20

Backward

begins with a full model and eliminates variables at each step

# of Features: 20

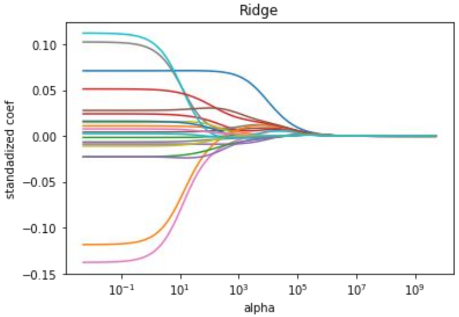
Feature Selection — Regularization

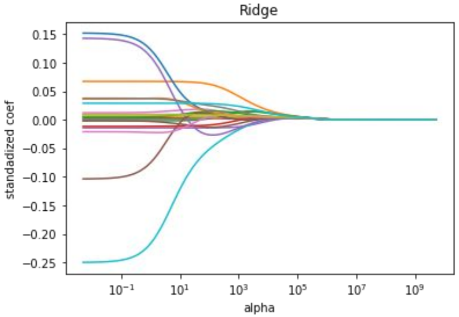
Forward

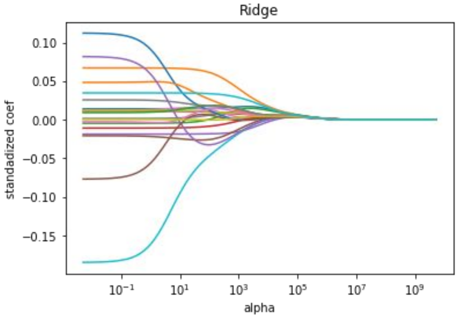
Select K Best

Backward

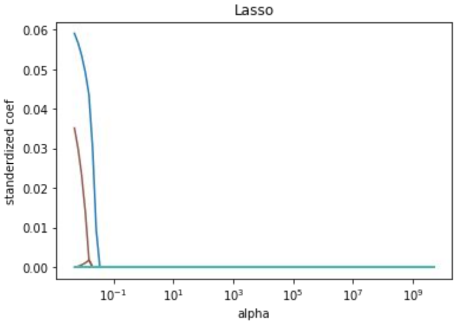
Ridge

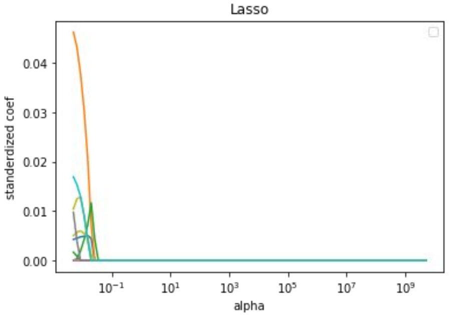


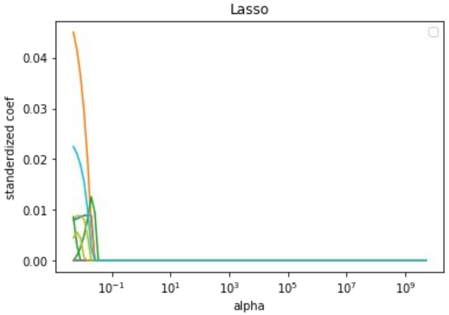




Lasso



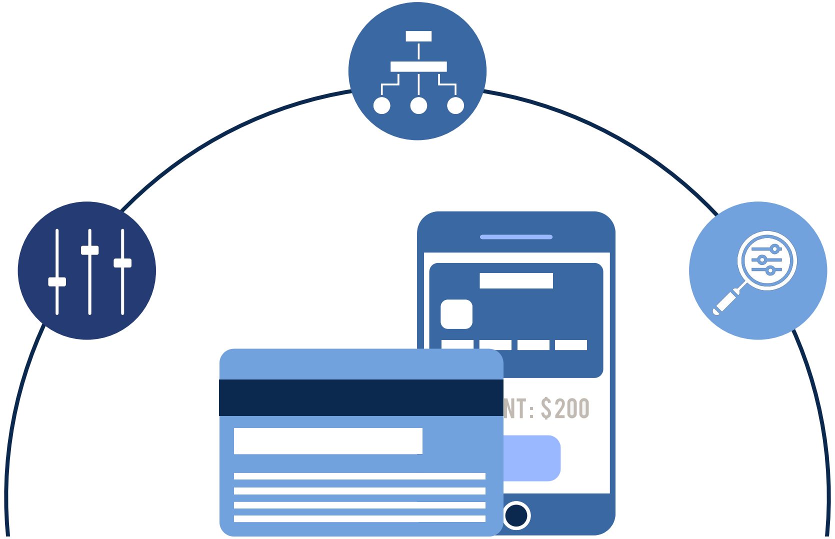




Model — Process

Defualt Models

Try 10 classification algorithms for
different variable combinations
Separate data into trn / tst / oot



Variables

Use different variable sets selected from different feature selection methods

Hyperparameter Tuning

Choose a set of optimal hyperparameters for our best model

Model — Performance Comparison

% Fraud Detected at *3% FDR*

	Logistic Regression	Decision Tree	Random Forest	Boosted Tree	Gradient Boosting Tree	Cat Boost	XGBoost	KNN	SVM	NN
TRN	63.3	76.6	100	83.3	89.3	82.8	85.2	82.6	88.2	77.9
TST	63.3	69.6	79.1	74.8	73.4	77.7	76.9	75.7	71.7	76.1
OOT	36.3	42.1	57.3	52.4	53.8	56.3	57.4	54.9	58.3	59.4

Model — Performance Comparison

% Fraud Detected at *3% FDR*

	Logistic Regression	Decision Tree	Random Forest	Boosted Tree	Gradient Boosting Tree	Cat Boost	XGBoost	KNN	SVM	NN
TRN	63.3	76.6	100	83.3	89.3	82.8	85.2	82.6	88.2	77.9
TST	63.3	69.6	79.1	74.8	73.4	77.7	76.9	75.7	71.7	76.1
OOT	36.3	42.1	57.3	52.4	53.8	56.3	57.4	54.9	58.3	59.4

Model — Performance Comparison

% Fraud Detected at *3% FDR*

	Logistic Regression	Decision Tree	Random Forest	Boosted Tree	Gradient Boosting Tree	Cat Boost	XGBoost	KNN	SVM	NN
TRN	63.3	76.6	100	83.3	89.3	82.8	85.2	82.6	88.2	77.9
TST	63.3	69.6	79.1	74.8	73.4	77.7	76.9	75.7	71.7	76.1
OOT	36.3	42.1	57.3	52.4	53.8	56.3	57.4	54.9	58.3	59.4

Model — Performance Comparison

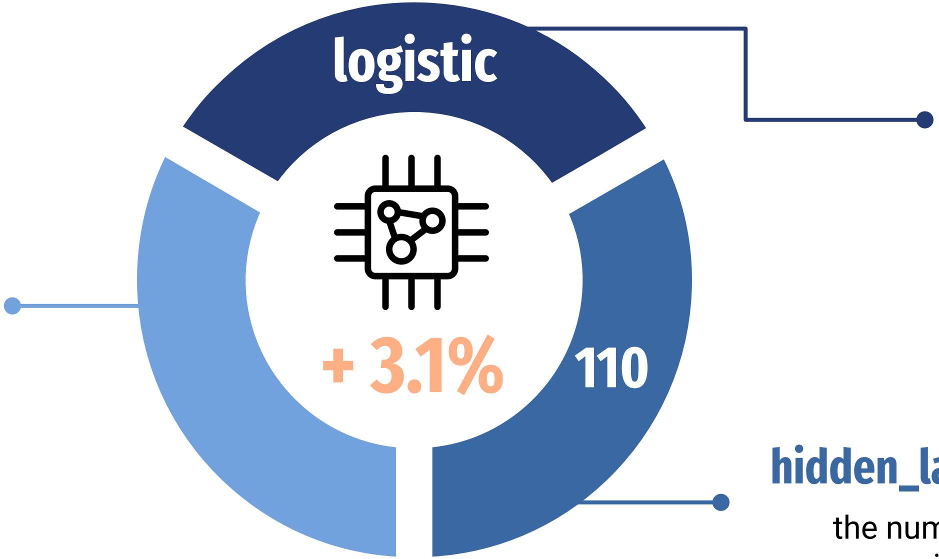
% Fraud Detected at *3% FDR*

	Logistic Regression	Decision Tree	Random Forest	Boosted Tree	Gradient Boosting Tree	Cat Boost	XGBoost	KNN	SVM	NN
TRN	63.3	76.6	100	83.3	89.3	82.8	85.2	82.6	88.2	77.9
TST	63.3	69.6	79.1	74.8	73.4	77.7	76.9	75.7	71.7	76.1
OOT	36.3	42.1	57.3	52.4	53.8	56.3	57.4	54.9	58.3	59.4

Model — Optimization

56.3% → 59.4%

defult → tuned



logistic

+ 3.1%

110

Neural Network Model

activation

activation function for the hidden layer

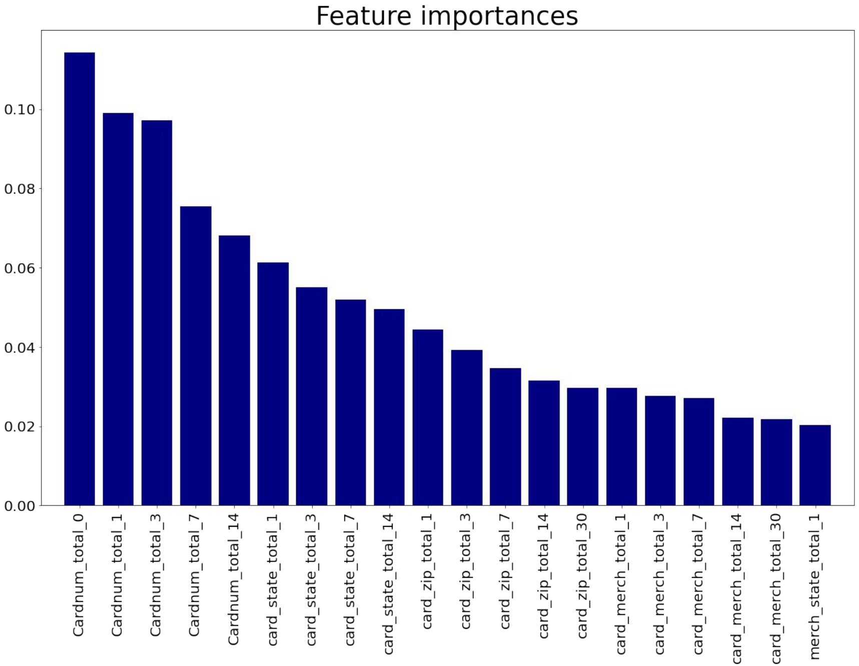
Stratify

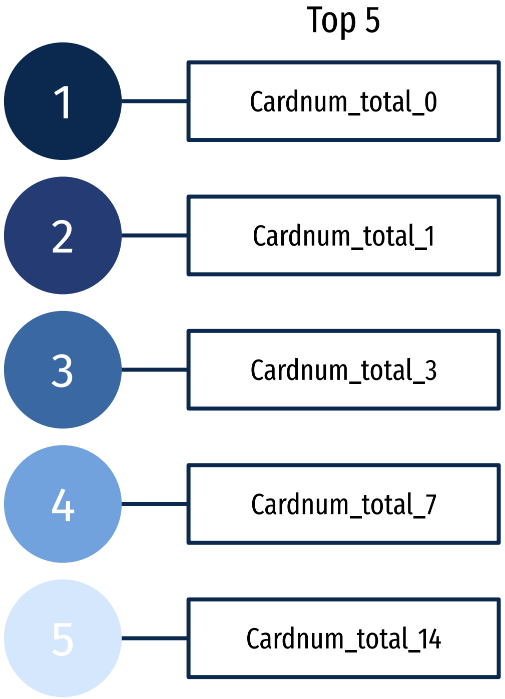
stratify training data with dependent variable (training)

hidden_layer_sizes

the number of neurons in the ith hidden layer

Top Variables



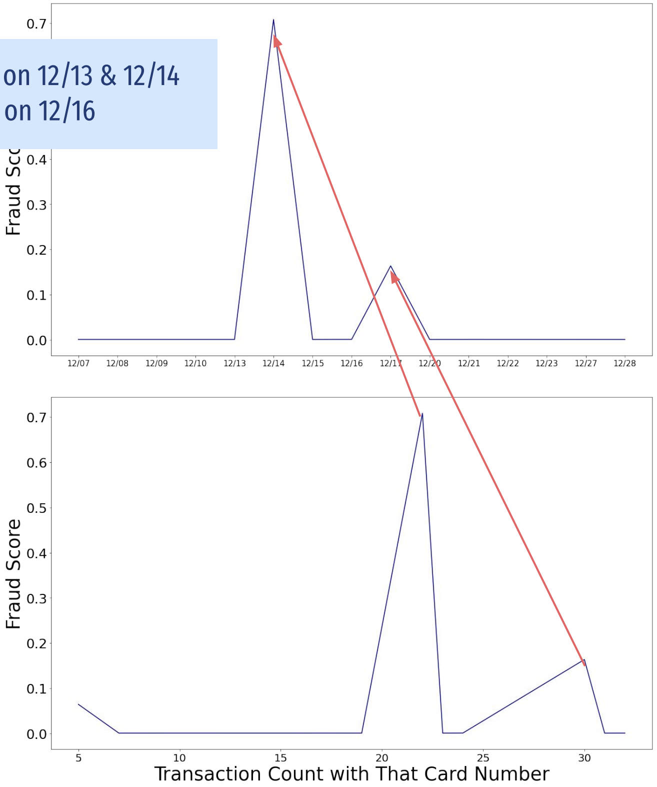


Top 5

1. Cardnum_total_0
2. Cardnum_total_1
3. Cardnum_total_3
4. Cardnum_total_7
5. Cardnum_total_14

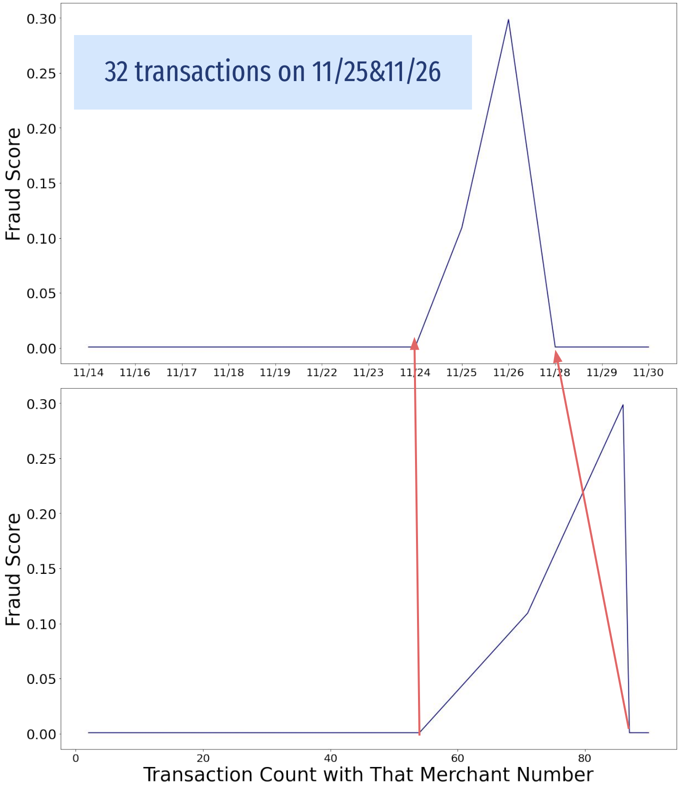
Time Dependence of Fraud Score

Cardnum = 5142134797



- 5 transactions on 12/13 & 12/14
- 6 transactions on 12/16

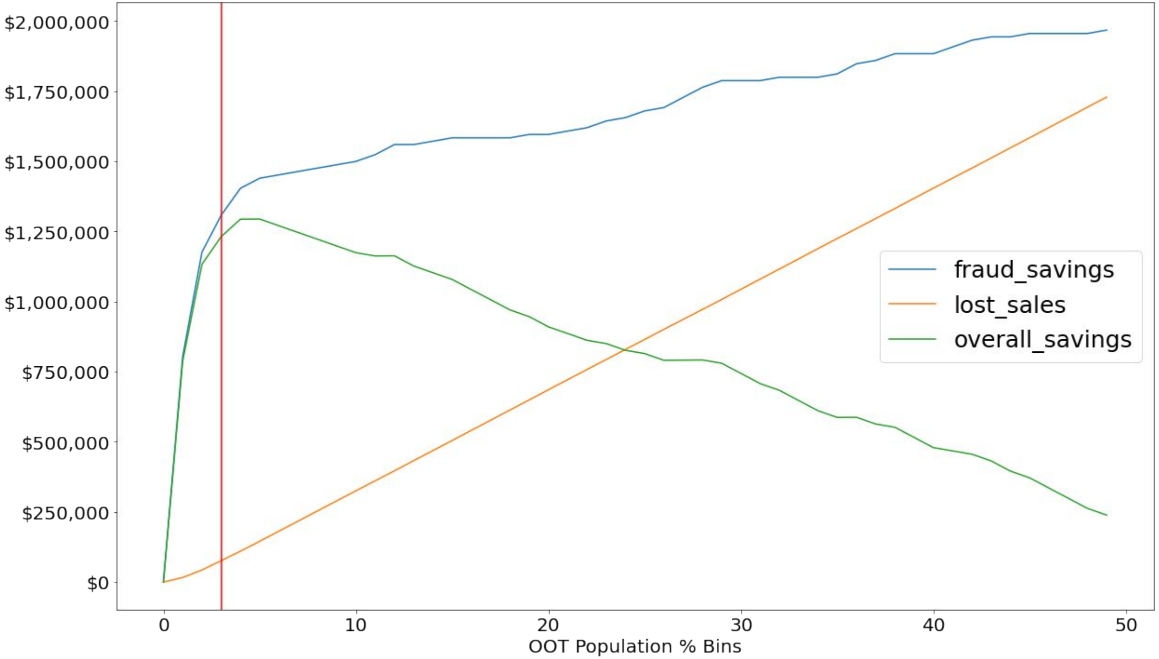
Merchnum = 4353000719908



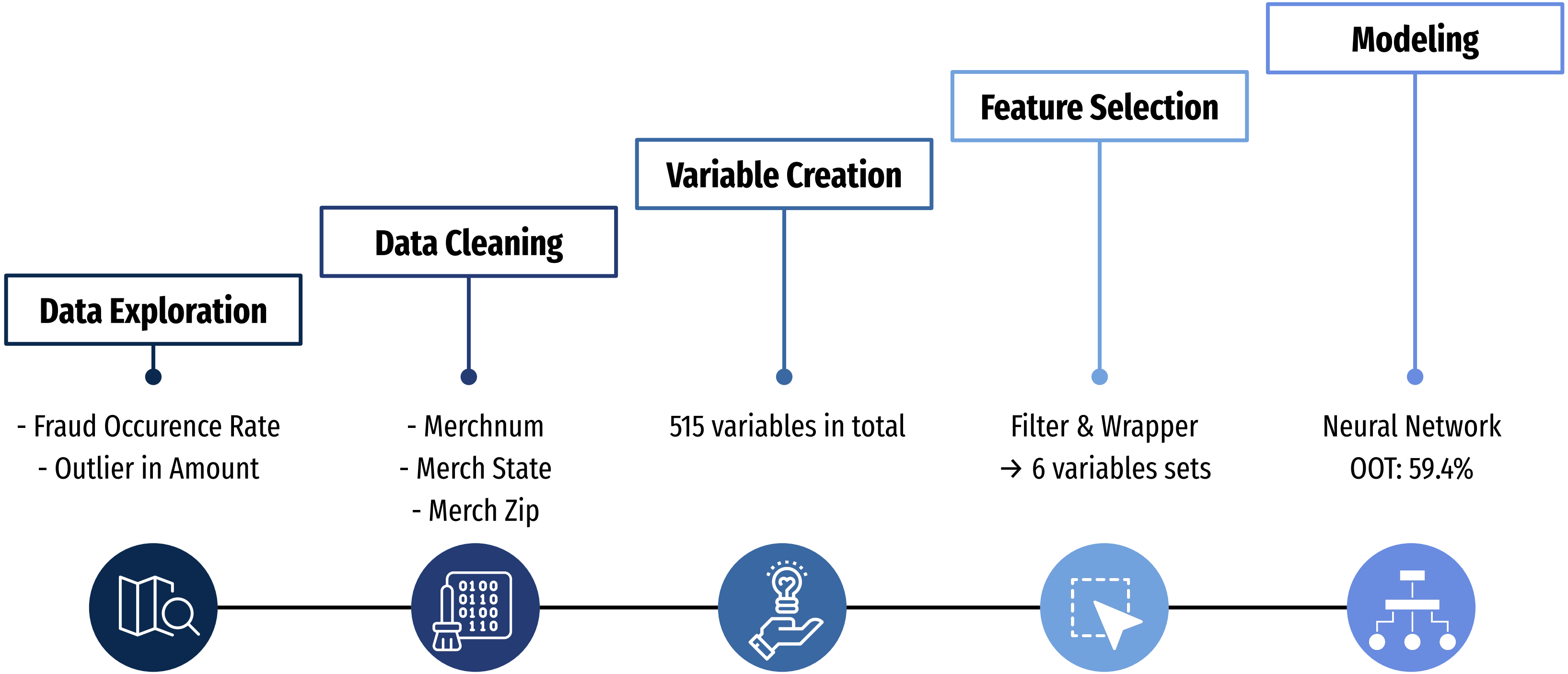
32 transactions on 11/25&11/26

Financial Cutoff

Cutoff at 3%, expected annual savings = $ 1.2 million



Conclusion



Data Exploration

- Fraud Occurence Rate
- Outlier in Amount

Data Cleaning

- Merchnum
- Merch State
- Merch Zip

Variable Creation

515 variables in total

Feature Selection

Filter & Wrapper
→ 6 variables sets

Modeling

Neural Network
OOT: 59.4%

Areas to Improve

Domain Knowledge

Consult with domain experts

Imbalanced Data

Try undersampling or oversampling method (e,g, SMOTE)

Feature Selection

Try different wrappers and algotithms

Machine Learning Algorithm

Unsupervised learning

Thank you