WATCHDOG MODEL IMPROVEMENT

BY COMBINGING EMAILAGE DATA

OUTLINE

- Background
- Feature Selection and Engineering
- Model Training
- Future Improvements
- Production Pipeline

BACKGROUND WATCHDOG

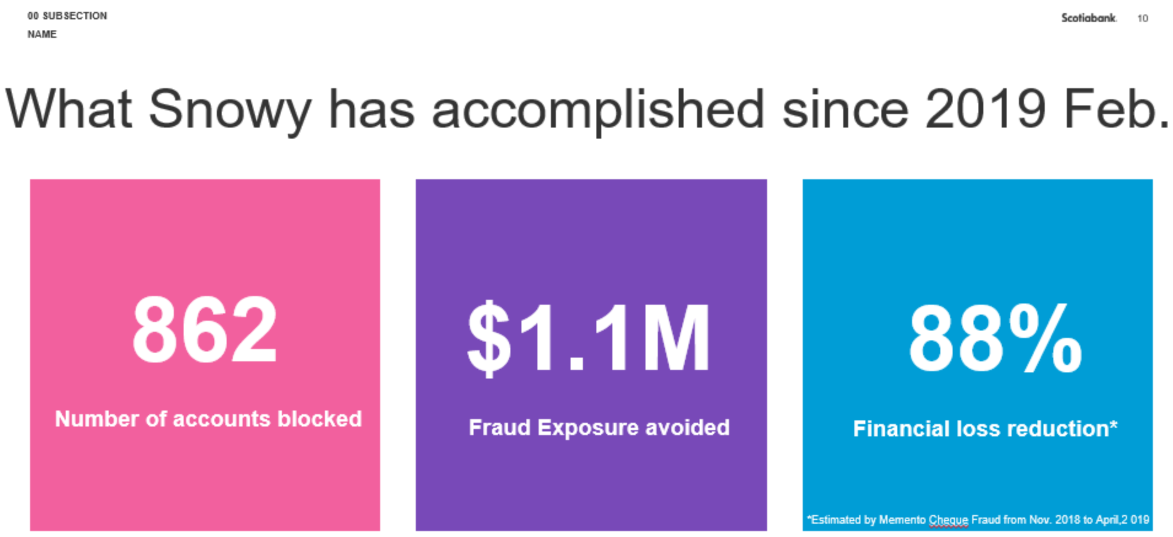
Cited from Edison's Digital Watchdog presentation

- Watchdog Vision:

A machine learning engine that unifies data sources across customer digital onboarding journey to support various business decisions on increasing digital revenue, and reducing operational risk and cost.

- Snowy:

Part of the pipeline that is specialized in detecting suspicious behaviors during digital onboarding journey



BACKGROUND WATCHDOG

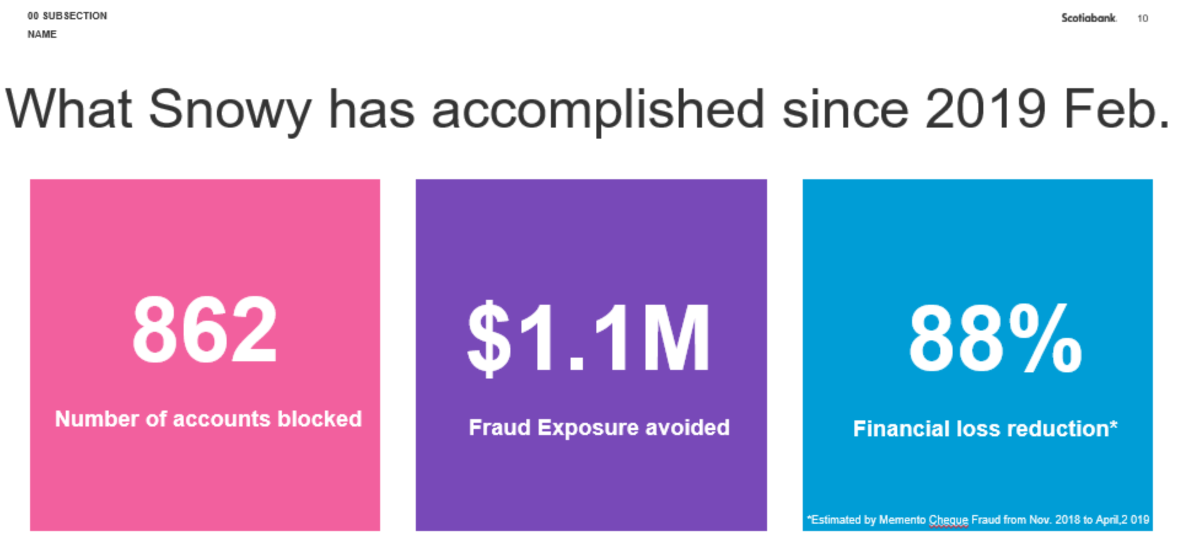
Cited from Edison’s Digital Watchdog presentation

- Watchdog Vision:

A machine learning engine that unifies data sources across customer digital onboarding journey to support various business decisions on increasing digital revenue, and reducing operational risk and cost.

- Snowy:

Part of the pipeline that is specialized in detecting suspicious behaviors during digital onboarding journey



BACKGROUND

Existing Watchdog Model

- Data Source : Pega, clickstream data
  - Information from application process
    - Number of clicks per page
    - Log in device
    - Cookies
    - …
- Training Period : 2018.11.01 to 2019.03.31

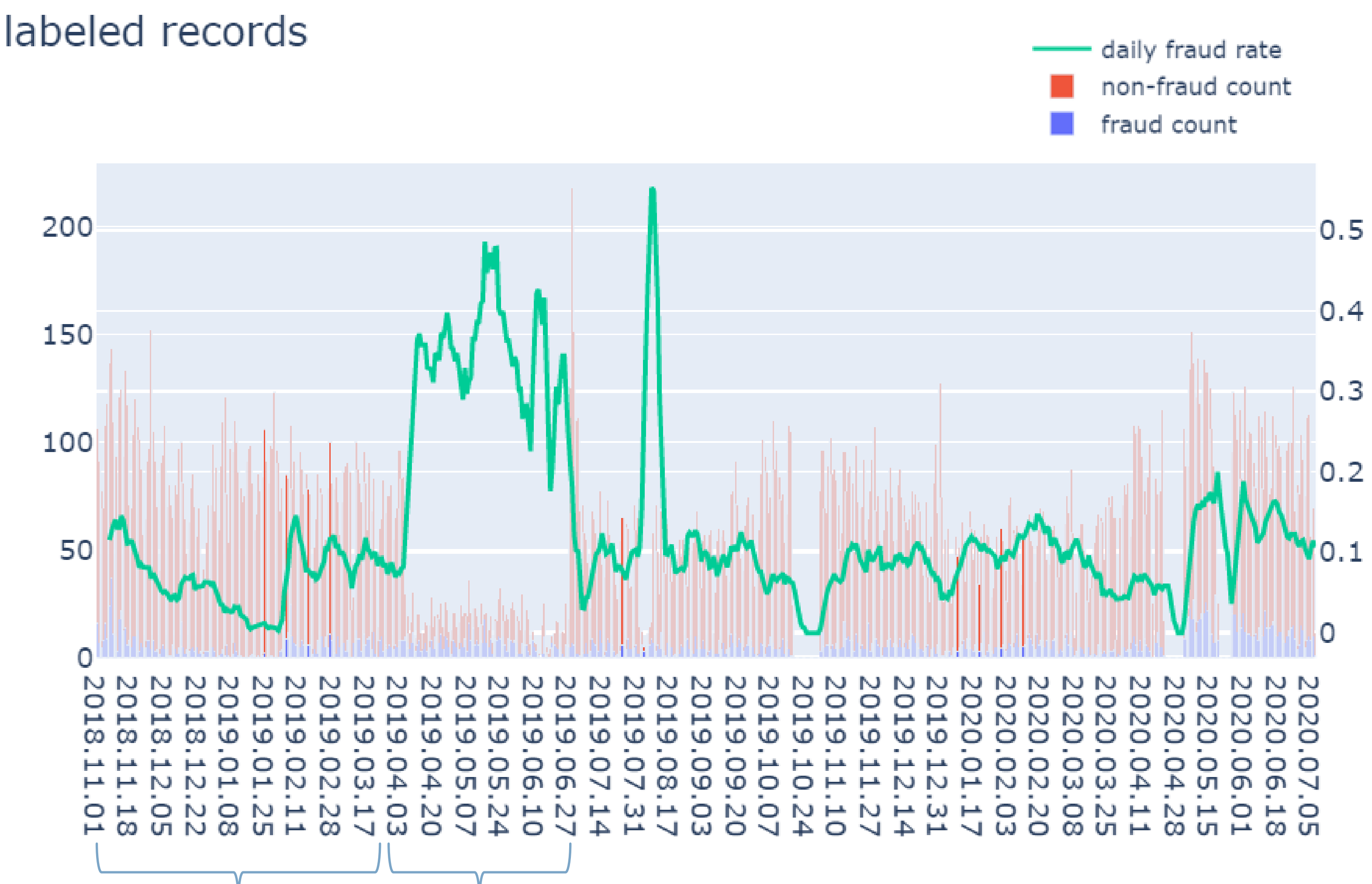
New Iteration

- New data source : EmailAge data
  - information associated with risks of applicants' email
    - IP_city
    - DomainCompany
    - FirstVerificationDate
    - …
- Training Period : 2018.11.01 to 2020.03.31

FEATURE SELECTION AND ENGINEERING :

LABEL AVAILABILITY

- The features are from EmailAge data source from 2018.11.1 to 2020.04.20.
- Only the records that are reviewed (with fraud label) are used for feature exploration.
- There are 31521 records with label, # fraud is 2918 and % fraud is 9.26%



2018.11.01
-2019.03.31

Snowy_v2
Training period

2019.04.01
-2019.06.26

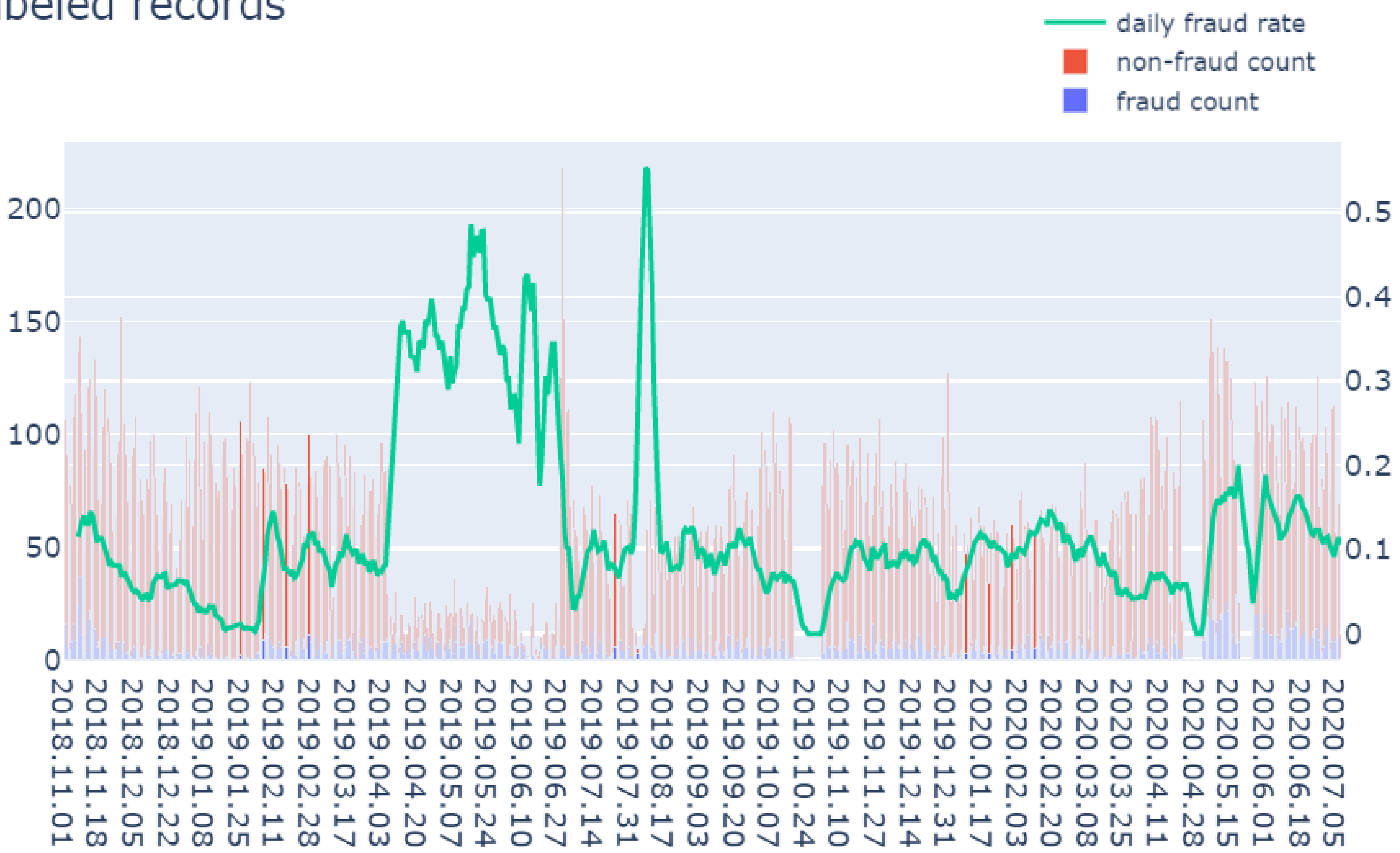
Snowy_v2
Test period

FEATURE SELECTION AND ENGINEERING :

LABEL AVAILABILITY

- The features are from EmailAge data source from 2018.11.1 to 2020.04.20.
- Only the records that are reviewed (with fraud label) are used for feature exploration.
- There are 31521 records with label, # fraud is 2918 and % fraud is 9.26%

labeled records



2018.11.01-2019.12.31

2020.01.01-2020.03.31

New
Training period

New
Test period

FEATURE SELECTION AND ENGINEERING : EARLY FEATURE SELECTION OVERVIEW

- There are 99 features available in Emailage data
- Many of them have string and object data type
- Measures are taken to filter out low quality or unusable features:
  1. Features with high NA rate
  2. Features that are redundant
  3. Features that are not necessary
  4. Features that have almost single value

Dtype	Count
Object	65
Float	23
Int	12

FEATURE SELECTION AND ENGINEERING :

EARLY FEATURE SELECTION

- Overall stability / NA rate
  - Feature “ename” has notable difference in missing rate between stratified fraud groups
  - Other high missing rate variables are dropped

Features	Missing Rate Overall	Missing Rate Fraud=1 Cases	Missing Rate Fraud=0 Cases
ename	0.750996	0.905502	0.722411
gender	0.779071	0.916268	0.753688
location	0.885769	0.960526	0.871939
company	0.955677	0.988437	0.949616
source_industry	0.955926	0.953748	0.956329
lastflaggedon	0.955926	0.953748	0.956329
phone_status	0.970120	0.902313	0.982665
title	0.971551	0.993620	0.967468
emailage	0.974228	0.988836	0.971526
fraud_type	0.979768	0.962520	0.982960
dob	0.989915	0.993620	0.989230
shipforward	1.000000		
shipcitypostalmatch	1.000000		
responsestatus.description	1.000000		
citypostalmatch	1.000000		
ipdistancemil	1.000000		
ipcountrymatch	1.000000		
ipaccuracyradius	1.000000		
ip_riskscore	1.000000		
ipriskcountry	1.000000		
ipdistancekm	1.000000		

FEATURE SELECTION AND ENGINEERING : EARLY FEATURE SELECTION

- Features that are redundant

fraudrisk, cariskband and cariskbandid are dropped because they are identical to cascore

fraudrisk	cascore	cariskband	cariskbandid
906 Very High	906	Fraud Score 900 to 999	6
089 Very Low	89	Fraud Score 1 to 100	1
129 Low	129	Fraud Score 101 to 300	2

- Variables with almost one value :
  - Ip_netspeedcell: 16055 out of 16064 values are "broadband"

- Variables that are unnecessary :
  - Ipaddress, ip_postalcode: values are not accurate while cardinality is too high

FEATURE SELECTION AND ENGINEERING : EARLY FEATURE SELECTION

- Features that are redundant

fraudrisk :

fraudrisk, cariskband and cariskbandid are dropped because they are identical to cascore

fraudrisk	cascore	cariskband	cariskbandid
906 Very High	906	Fraud Score 900 to 999	6
089 Very Low	89	Fraud Score 1 to 100	1
129 Low	129	Fraud Score 101 to 300	2

- Variables with almost one value :
  - Ip_netspeedcell: 16055 out of 16064 values are "broadband"
- Variables that are unnecessary :
  - Ipaddress, ip_postalcode: values are not accurate while cardinality is too high

FEATURE SELECTION AND ENGINEERING : EARLY FEATURE SELECTION

- Features that are redundant

fraudrisk :

fraudrisk, cariskband and cariskbandid are dropped because they are identical to cascore

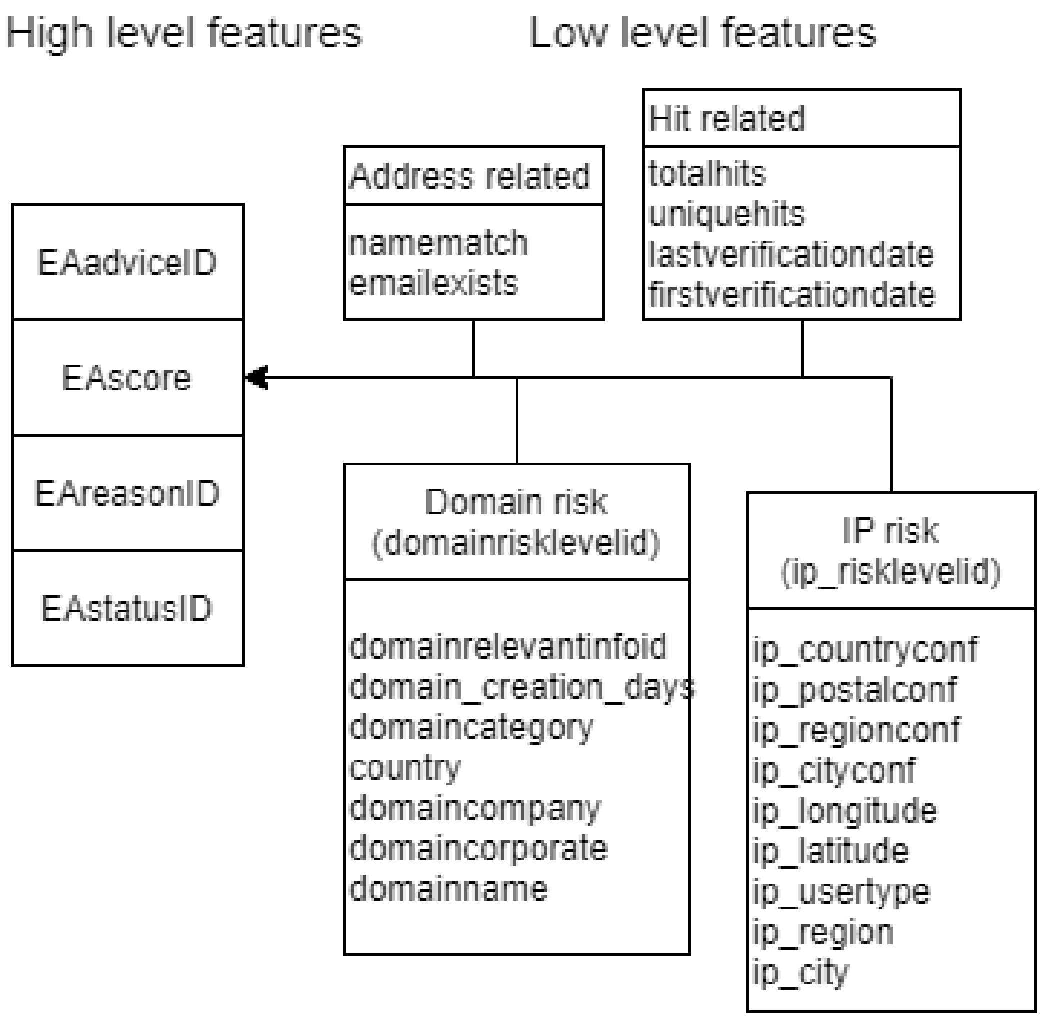
fraudrisk	cascore	cariskband	cariskbandid
906 Very High	906	Fraud Score 900 to 999	6
089 Very Low	89	Fraud Score 1 to 100	1
129 Low	129	Fraud Score 101 to 300	2

- Variables with almost one value :
  - Ip_netspeedcell: 16055 out of 16064 values are "broadband"
- Variables that are unnecessary :
  - Ipaddress, ip_postalcode: values are not accurate while cardinality is too high

FEATURE SELECTION AND ENGINEERING :
ANALYSIS OF FEASIBLE FEATURES

- Logic structure of remaining variables:
  - There are 33 potentially feasible variables left after the elimination process

- A availability / stability check is conducted to validate the use of these variables in model



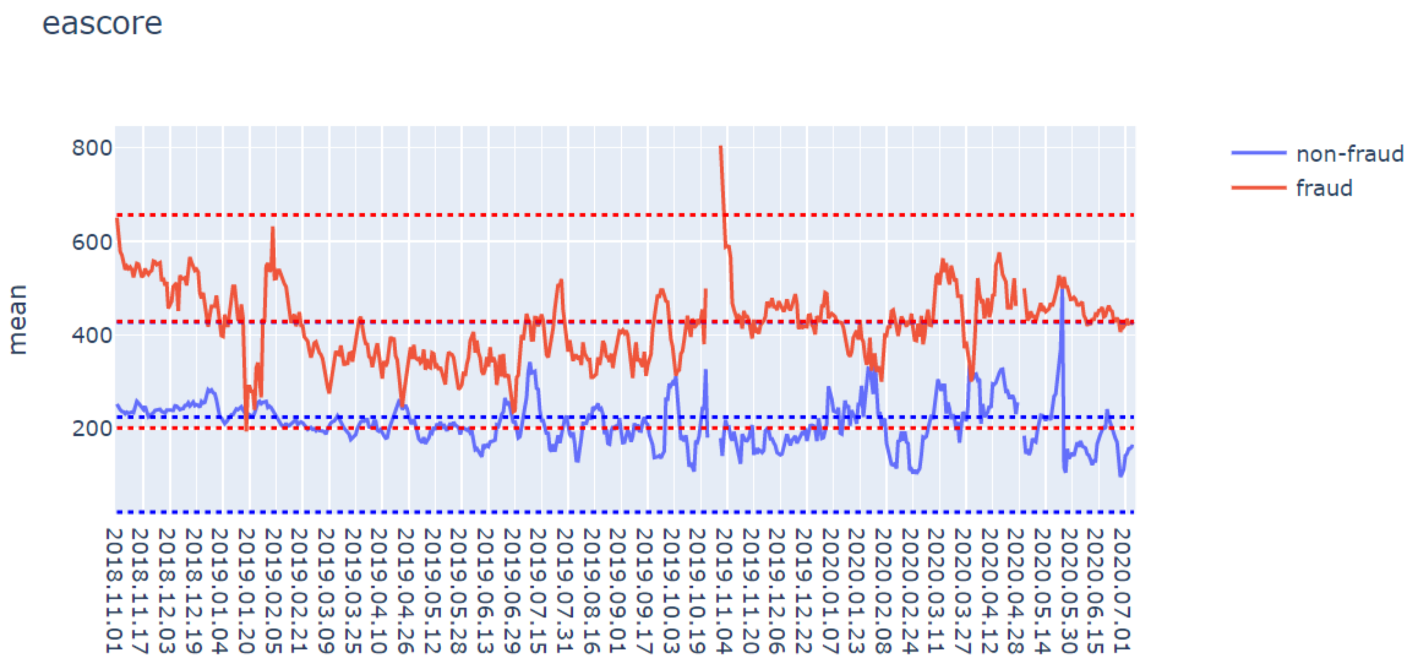
FEATURE SELECTION AND ENGINEERING :

ANALYSIS OF FEASIBLE FEATURES

- For integer variables::

  Mean value per day for fraud and non-fraud groups are plotted

  Mean +/- 1* SD is also plotted to indicate stability



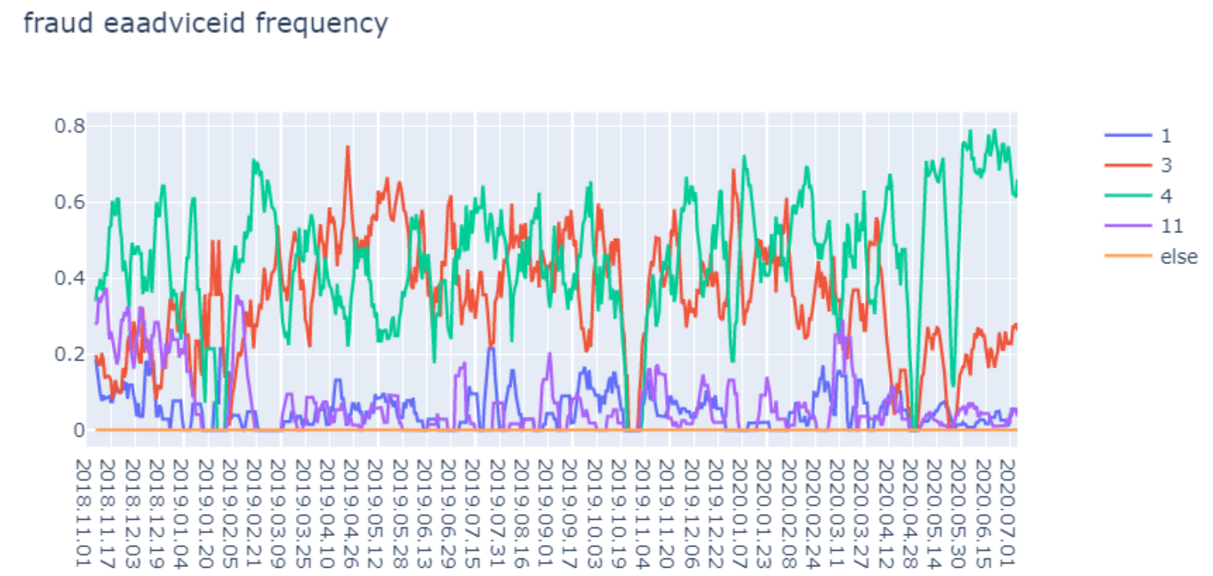
FEATURE SELECTION AND ENGINEERING :

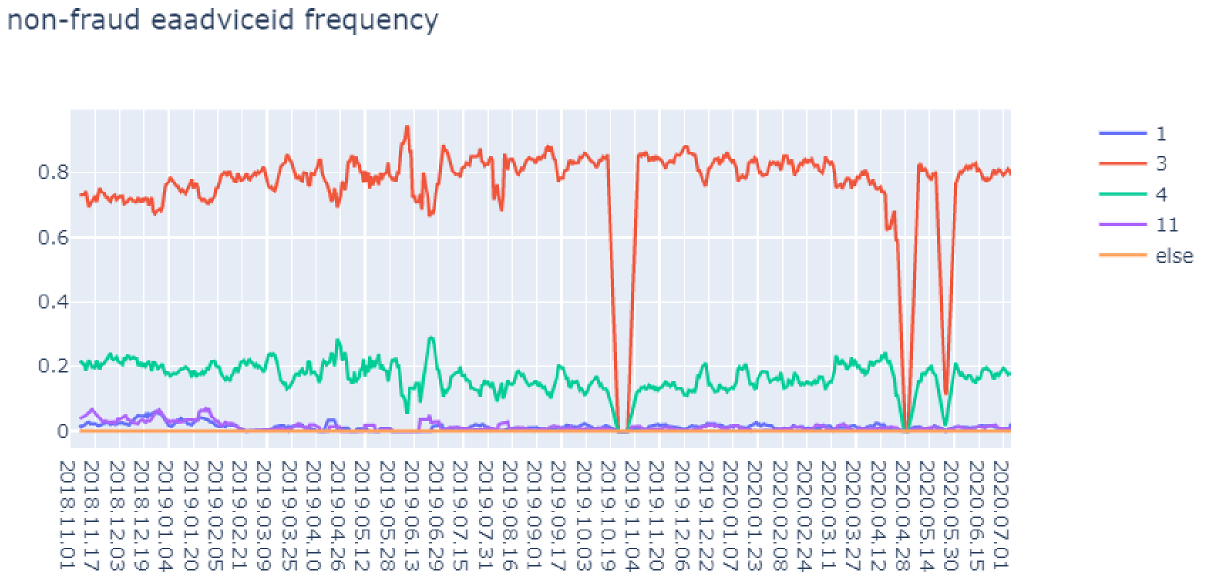
ANALYSIS OF FEASIBLE FEATURES

- For categorial variables::

  Frequency by percentage per day for fraud and non-fraud groups are plotted

- This feature can be safely turned into dummy variables using One-Hot encoding
- However, …





FEATURE SELECTION AND ENGINEERING : CHALLENGES

- High dimension :
  - Some categorical variables have very high cardinality : ip_city, eareasonid …
- Solution :
  - Grouping low frequency values as one “other” category
    - less likely to overfit (not sensible to label),
    - stable categories
    - feasible for day to day pipeline (new/lost categories)
- Risk:
  - XGBoost tends to select features with continuous values, making binary features underrepresented

FEATURE SELECTION AND ENGINEERING : FEATURE ENGINEERING

- Grouping for high cardinality
  Ordinal features

For example: DomainRelevantInfoID

- Assigned with value 2:
  - 524 - VeryLowRiskEmailDomainforCompany
  - 525 - VeryLowRiskEmailDomainforIndustry
  - 526 - VeryLowRiskEmailDomainforNetwork
- Assigned with value 3:
  - 521 -LowRiskEmailDomainforCompany
  - 522 - LowRiskEmailDomainforIndustry
  - 523 – LowRiskEmailDomainforNetwork

new values	#observations	# fraud	% fraud
0	0		
1	153	13	0.084967
2	0		
3	15746	2434	0.154579
4	165	61	0.369697
5	0		

FEATURE SELECTION AND ENGINEERING : FEATURE ENGINEERING

- Grouping for high cardinality
  Ordinal features

For example: DomainRelevantInfoID

- Assigned with value 2:
  - 524 - VeryLowRiskEmailDomainforCompany
  - 525 - VeryLowRiskEmailDomainforIndustry
  - 526 - VeryLowRiskEmailDomainforNetwork
- Assigned with value 3:
  - 521 -LowRiskEmailDomainforCompany
  - 522 - LowRiskEmailDomainforIndustry
  - 523 – LowRiskEmailDomainforNetwork

new values	#observations	# fraud	% fraud
0	0		
1	153	13	0.084967
2	0		
3	15746	2434	0.154579
4	165	61	0.369697
5	0		

* %fraud = #fraud / #obs

%fraud is not used for grouping but supports the grouping decision

FEATURE SELECTION AND ENGINEERING :

FEATURE ENGINEERING

- Dummy Variables for Nominal Features
  - New variables are created according to their frequency
  - Low frequency values are grouped as "other" category

eareasonid:

eareasonid is broken into 9 dummy variables, 8 of which are categories with highest frequency and one consisting all other variables.

original value	explanation	# observations	# fraud	% fraud
14	Email Created at least X Years Ago	10153	826	0.0814
8	Limited History for Email	2988	885	0.2962
28	Valid Email From X Country Domain	742	249	0.3356
2	Email does not exist	480	234	0.4857
11	Good Level X	384	23	0.0599
13	Email Created X Years Ago	354	20	0.0565
1	Fraud Level X	299	85	0.2843
4	Risky Domain	72	39	0.5417
other	other	664	116	0.2937

MODEL TRAINING: DATA SOURCE

- Overall data source :
  - Features : EmailAge features and snowy_v2 features
  - Labels : Label consists of reviewer feedback labels, cheque fraud labels, and cerb fraud labels
- Training set : positive label rate : 9.38%
  - Date : 2018-11-01 to 2019-12-31
- Test set 1: positive label rate : 9.48%
  - Date : 2020-01-01 to 2020-03-31
  - Is supposed to be have the same distribution as training data
- Test set 2:
  - Date : 2020-04-01 to 2020-07-05
  - Is supposed to be different from the training data with effect of pandemic
  - There is also strategy change in reviewing increasing the total

MODEL TRAINING: HYPERPARAMETER SEARCH

- Use Hyperopt to tune parameters
  - A Bayesian probabilistic model based approach for finding the minimum of loss function
  - Search path in parameter space is based on previous evidence
  - More efficient than random/grid search
- Metric: Average Precision
  - Loss function : 1- AP
  - Weighted precision according to increase in recall
  - Evidence shows AP performs better on small positive class

$$\text{AP} = \sum_n (R_n - R_{n-1}) P_n$$

MODEL TRAINING: PERFORMANCE ON TEST

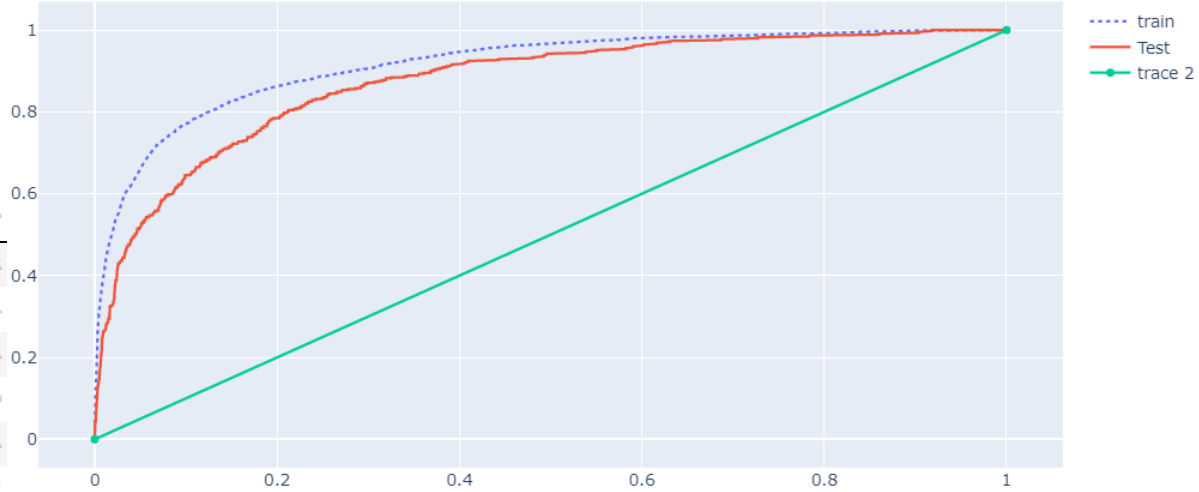
Train cv AP:0.5846

	precision	recall	review_perc	threshold	FP/TP
0	0.093754	1.000000	1.000000	0.00	9.666243
1	0.223214	0.921220	0.386928	0.05	3.480000
2	0.348217	0.835663	0.224993	0.10	1.871769
3	0.463952	0.757730	0.153119	0.15	1.155394
4	0.551456	0.689962	0.117301	0.20	0.813382
5	0.609836	0.630241	0.096891	0.25	0.639785
6	0.663776	0.583651	0.082437	0.30	0.506531
7	0.705199	0.540025	0.071794	0.35	0.418039
8	0.742058	0.494706	0.062502	0.40	0.347603
9	0.777293	0.452351	0.054561	0.45	0.286517
10	0.809564	0.401525	0.046500	0.50	0.235232
11	0.842583	0.353664	0.039352	0.55	0.186826
12	0.874222	0.297332	0.031887	0.60	0.143875
13	0.895570	0.239729	0.025096	0.65	0.116608
14	0.898520	0.180008	0.018783	0.70	0.112941
15	0.915625	0.124100	0.012707	0.75	0.092150
16	0.917526	0.075392	0.007704	0.80	0.089888
17	0.898990	0.037696	0.003931	0.85	0.112360
18	0.920000	0.009742	0.000993	0.90	0.086957
19	0.000000	0.000000	0.000000	0.95	inf

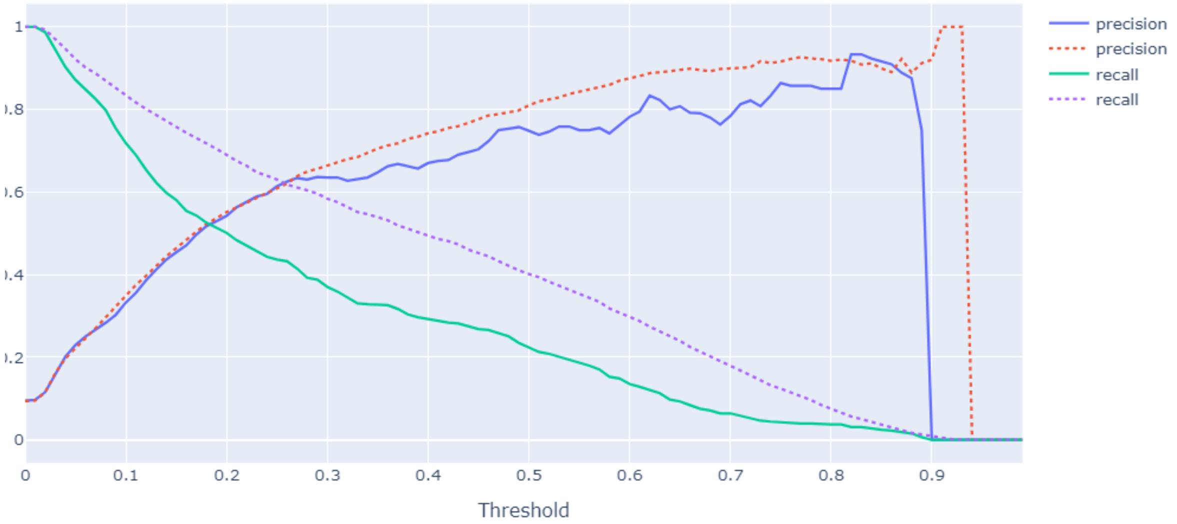
Test cv AP:0.4935

	precision	recall	review_perc	threshold	FP/TP
0	0.094768	1.000000	1.000000	0.00	9.552106
1	0.229825	0.871397	0.359319	0.05	3.351145
2	0.332649	0.718404	0.204665	0.10	2.006173
3	0.454073	0.580931	0.121244	0.15	1.202290
4	0.541966	0.501109	0.087623	0.20	0.845133
5	0.613707	0.436807	0.067451	0.25	0.629442
6	0.634981	0.370288	0.055264	0.30	0.574850
7	0.647577	0.325942	0.047699	0.35	0.544218
8	0.670051	0.292683	0.041395	0.40	0.492424
9	0.703488	0.268293	0.036142	0.45	0.421488
10	0.748148	0.223947	0.028367	0.50	0.336634
11	0.750000	0.186253	0.023534	0.55	0.333333
12	0.782051	0.135255	0.016390	0.60	0.278689
13	0.807692	0.093126	0.010927	0.65	0.238095
14	0.783784	0.064302	0.007775	0.70	0.275862
15	0.863636	0.042129	0.004623	0.75	0.157895
16	0.850000	0.037694	0.004203	0.80	0.176471
17	0.916667	0.024390	0.002522	0.85	0.090909

- ROC-AUC on train:0.8959
  test : 0.8571

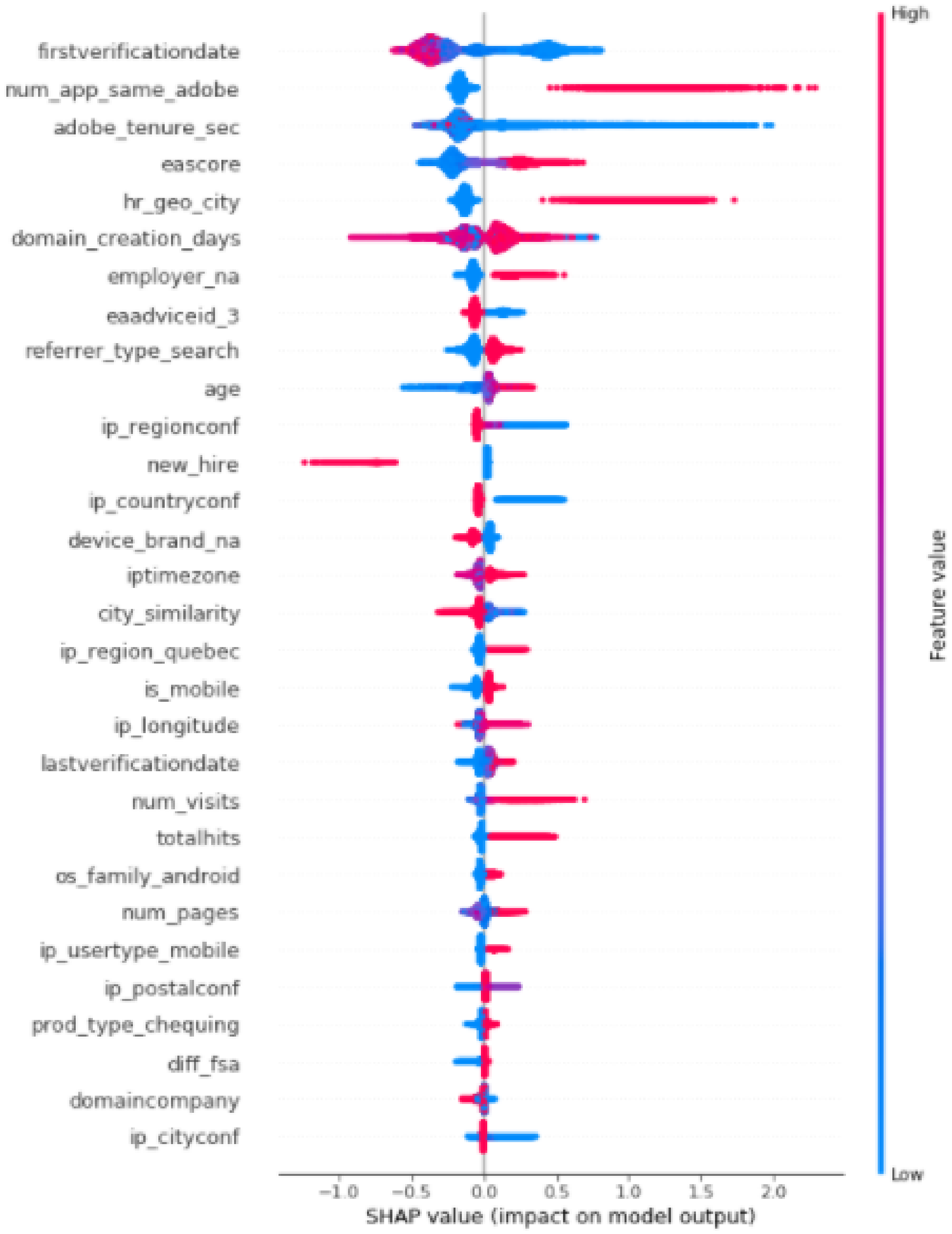


- Precision & Recall over threshold on test :



MODEL TRAINING: VARIABLE IMPORTANCE (SHAP VALUE)

- SHAP value for the 30 most important features
- Top emailage variables:
  - FirstVerificationDate
  - EAscore
  - Domain_creation_days



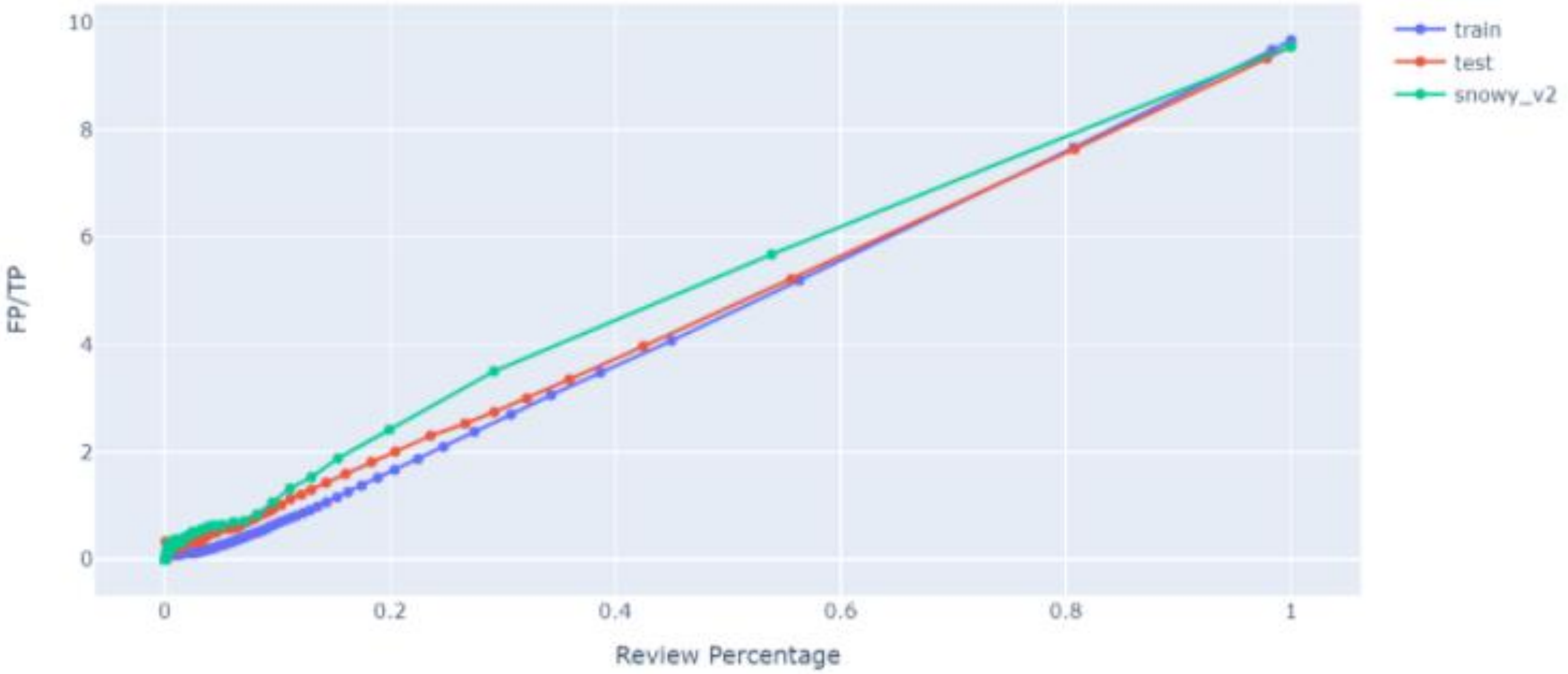
MODEL TRAINING: COMPARISON WITH SNOWY_V2

- Comparison of some metrics on test set:
  - AUC : Snowy_v2 : 0.8377
    New Model: 0.8571
  - AP : Snowy_v2 : 0.4827
    New Model: 0.4935

FP/TP : # non fraud records encountered per true fraud records

Review Percentage : # predicted fraud / # total labels

Both negatively related to threshold



MODEL TRAINING: COMPARISON WITH SNOWY_V2

- Comparison of some metrics on test set:
  - AUC : Snowy_v2 : 0.8377

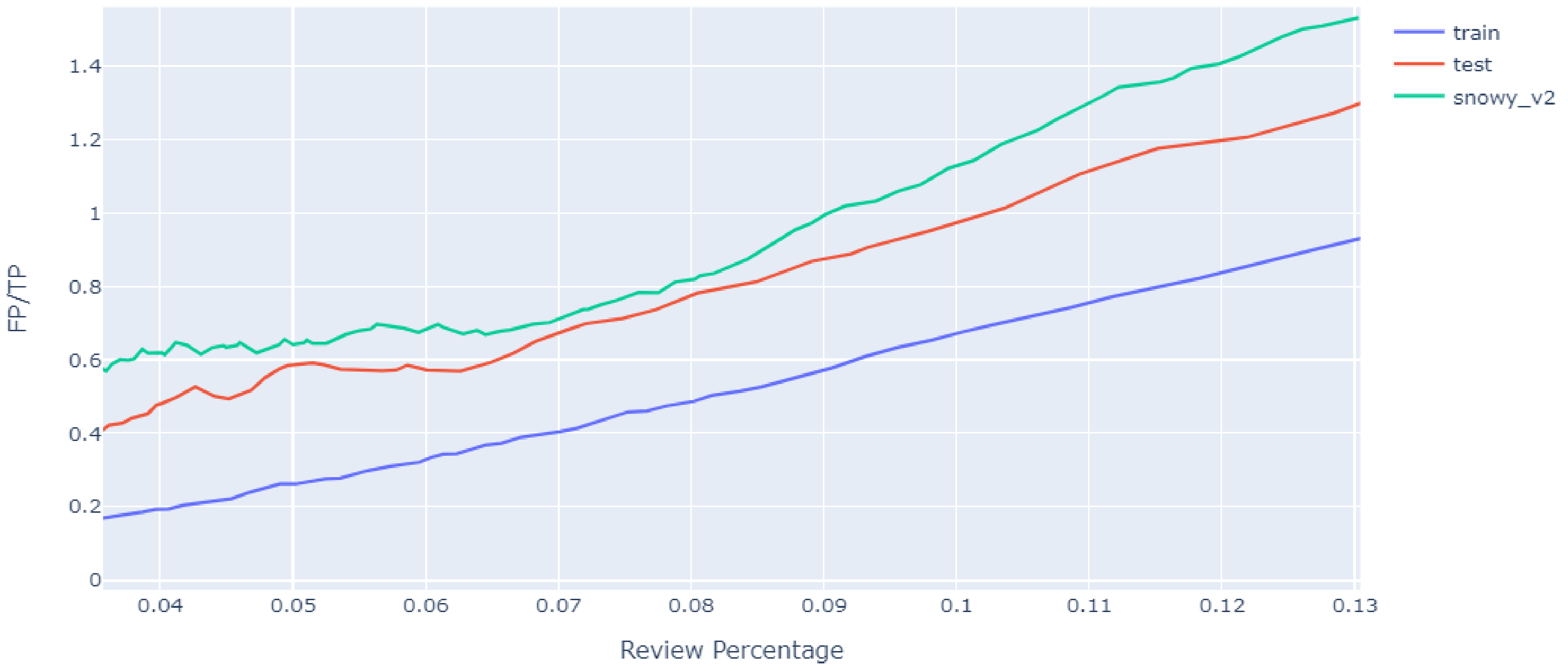
    New Model: 0.8571
  - AP : Snowy_v2 : 0.4827

    New Model: 0.4935

FP/TP : # non fraud records encountered per true fraud records

Review Percentage : # predicted fraud / # total labels

Both negatively related to threshold



FUTURE IMPROVEMENTS

- Next Steps:
  - Finalize model and threshold
  - Test different label performance
  - Improve performance
- Model Performance:
  - Find a sophisticated way to deal with high cardinality categorical variables
  - Develop new features by referencing different data sources
  - Improve model performance by fine tuning

PRODUCTION PIPELINE INTEGRATION AND TESTING

- Data Preparation :
  - Change data source of EmailAge data from API to EDL;
  - Add code for parsing day to day and credit card EmailAge data
- Feature Engineering :
  - Add functions dealing with exceptions resulting pipeline breakdown
  - Add feature engineering functions for EmailAge data for day to day needs
- Model Generation :
  - Update model training code (not tested)

WHAT I LEARN

- Difference between academic and work environment
- Unique experience
- Teamwork
- Communication

THANKS