Fraud Policy Analysis

Yufei (Olivia) Wu

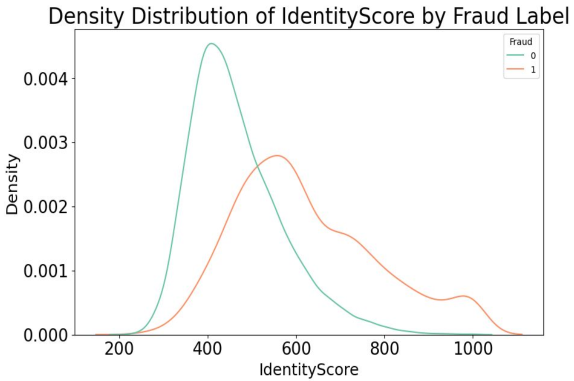
Case Part 1

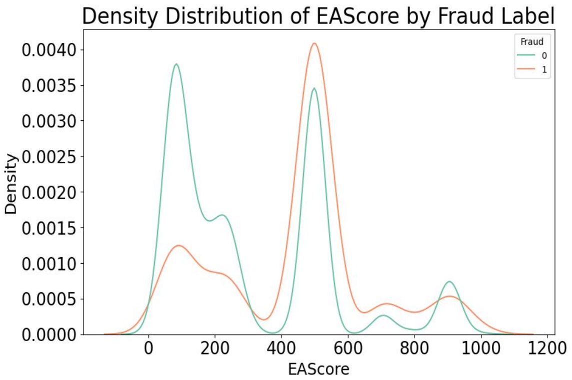
Whether or not we would engage with a new vendor provided information regarding phone number ownership and risk factors?

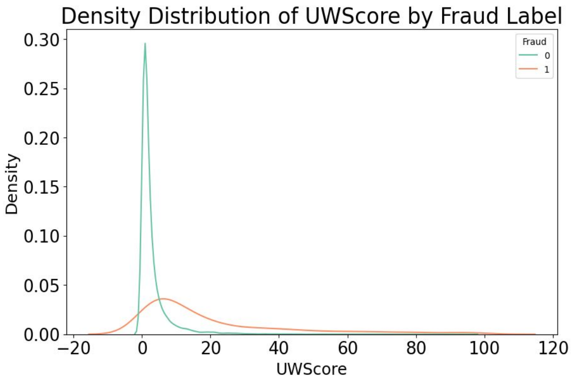
Step 1

design a strategy (a set of rules) to decision an application using the existing internal scores

Check data distribution by fraud label

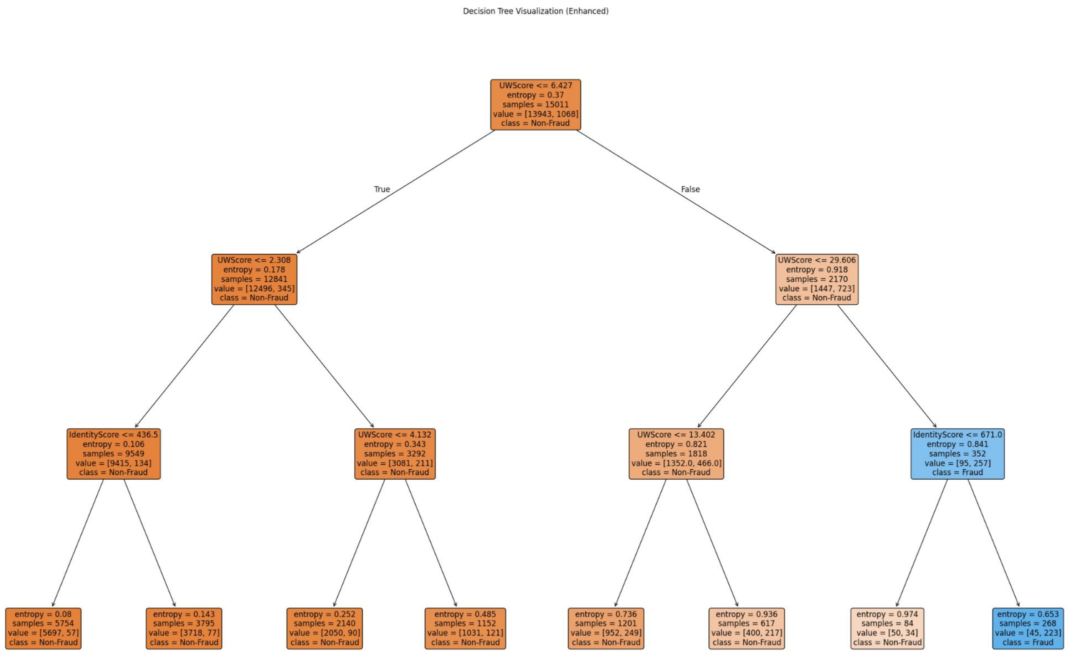






Detection rule and detection results on all transactions

Rule: SMOTE + Decision Tree Visualization

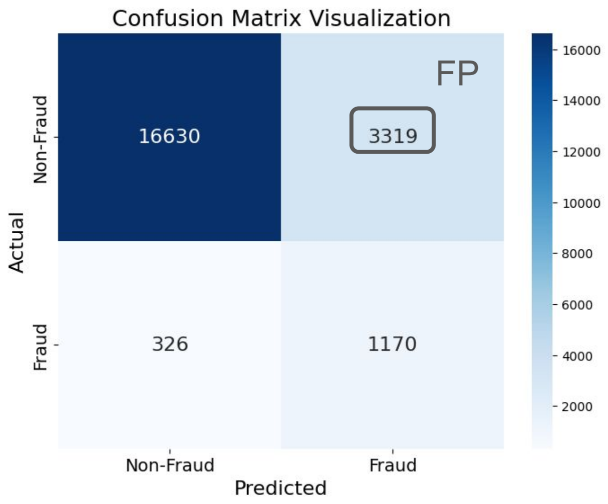


Detection Results

```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.83      0.90     19949
           1       0.26      0.78      0.39      1496

    accuracy                           0.83     21445
   macro avg       0.62      0.81      0.65     21445
weighted avg       0.93      0.83      0.87     21445
```



Step 2

vendor data analysis and evidence if the vendor data can enhance the above strategy

Qualitative analysis

Factors like:

- Identity completeness
- Service Discontinued Indicator
- voice over IP
- Business_Phone_Indicator

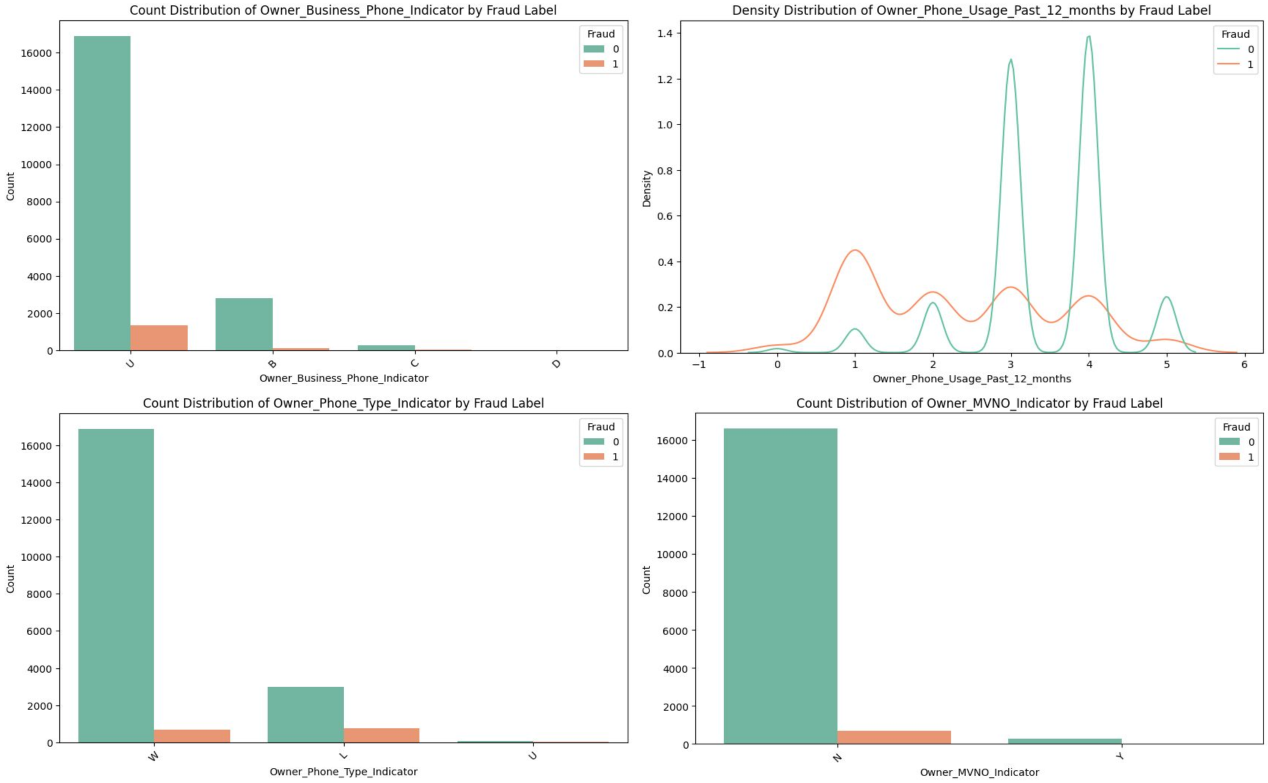
These are factors seem correlated with fraudulent behavior.

Quantitative analysis

Four different parallel methods:

- Multivariate analysis
- Correlation analysis
- Chi-square test on numerical new vendor data
- Modeling accuracy by two different datasets

1, Multivariate Analysis on new vendor data

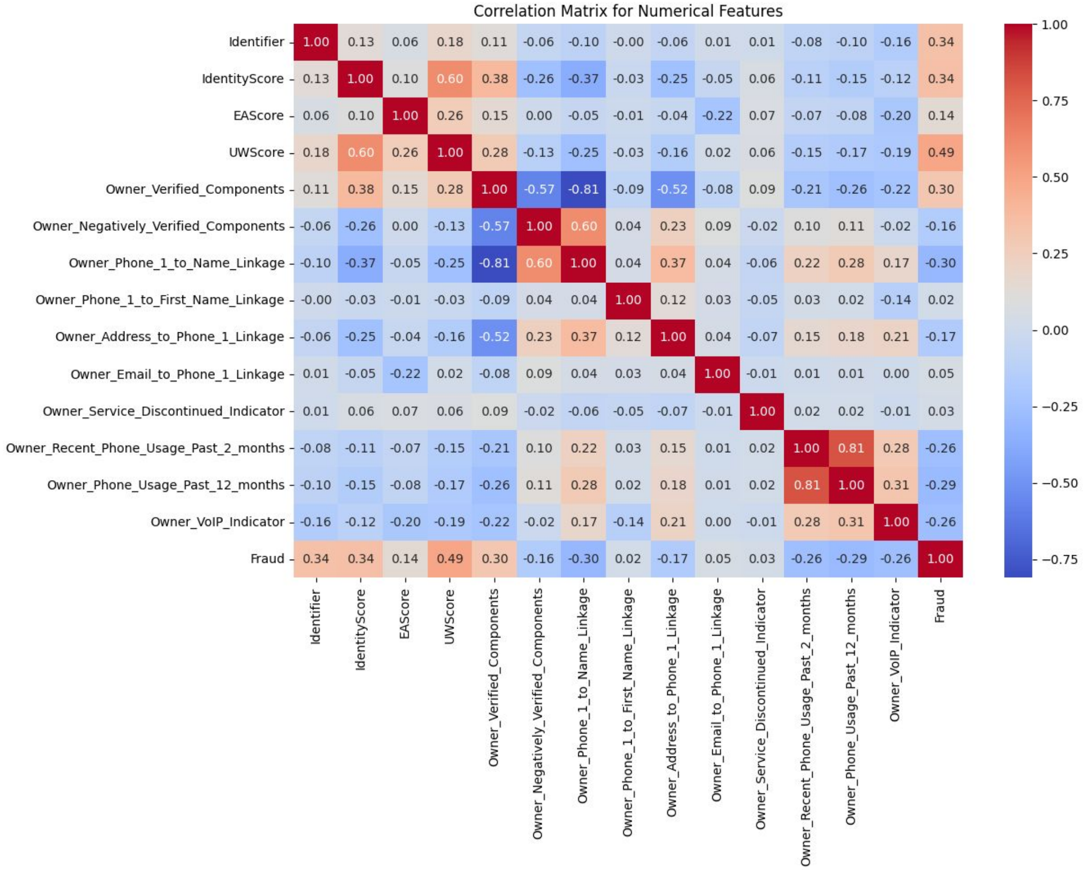


For many features in the vendor data, the distribution of fraud between different feature values is significantly different, which indicates that these features are correlated with the fraud label.

2, Correlation Analysis

Relation between fraud and new variables：

Relation between old variables and new variables：



3, chi-square test on numerical new vendor data

All p-value are small enough, which means these new features have correlations on fraud label

	Feature	Chi2	P-value
0	Owner_Verified_Components	2492.452896	0.000000e+00
2	Owner_Phone_1_to_Name_Linkage	2363.018612	0.000000e+00
8	Owner_Recent_Phone_Usage_Past_2_months	2582.785342	0.000000e+00
9	Owner_Phone_Usage_Past_12_months	3124.049705	0.000000e+00
10	Owner_Phone_Carrier	3474.317076	0.000000e+00
11	Owner_Parent_Phone_Carrier	3046.856800	0.000000e+00
12	Owner_Technology_Indicator	1323.678255	3.689088e-288
6	Owner_Phone_Type_Indicator	1280.409597	9.174632e-279
1	Owner_Negatively_Verified_Components	863.387671	3.860373e-182
4	Owner_Address_to_Phone_1_Linkage	750.621976	3.802382e-161
3	Owner_Phone_1_to_First_Name_Linkage	695.439761	3.385723e-149
5	Owner_Email_to_Phone_1_Linkage	412.928540	4.473527e-88
7	Owner_Service_Discontinued_Indicator	260.996482	1.262132e-52
13	Owner_MVNO_Indicator	12.153252	4.900234e-04

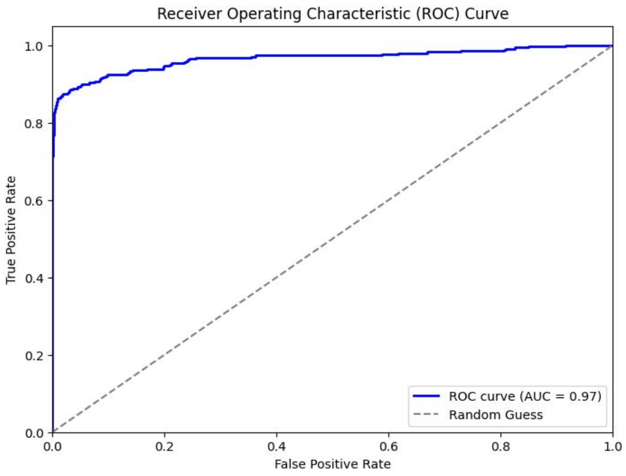
4, modeling by two different datasets - procedure

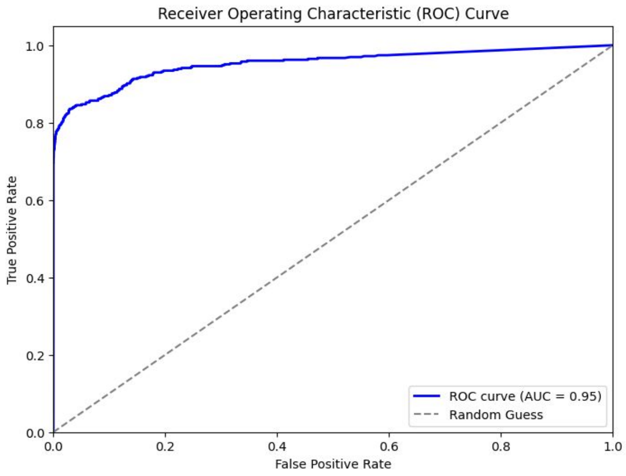
Preprocessing	Training	Evaluation
Encoding • One-Hot Encoding on the 'Owner_MVNO_Indicator_category' feature • Perform Target Encoding on other categorical features	• SMOTE • Grid Search	• Confusion matrix • Classification report • ROC-AUC

4, modeling by two different datasets - model

	ROC-AUC	CI
Vendor data & Internal score	0.97	[0.9670, 0.9730]
Internal score	0.95	[0.9380, 0.9620]

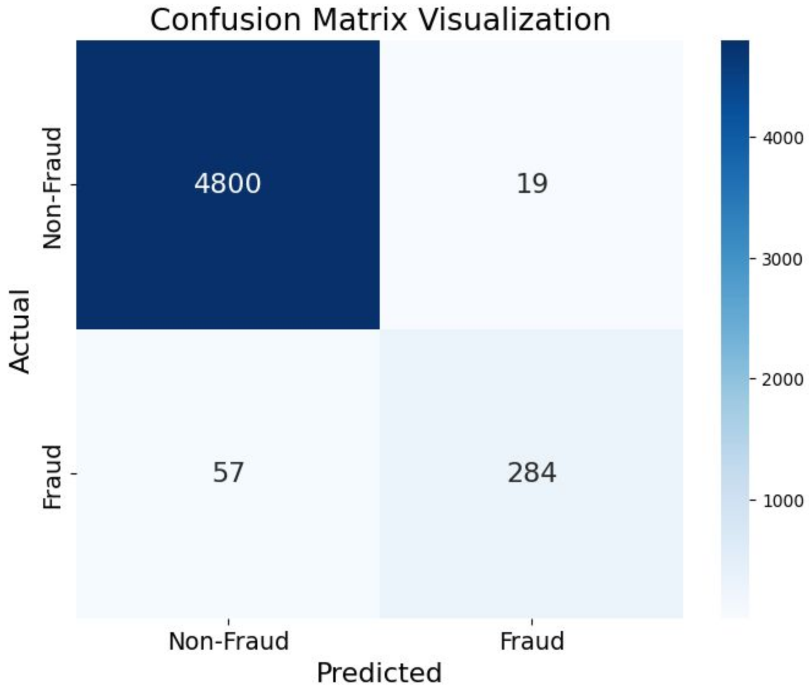
Built XGBoost model by datasets with only internal score and whole data sets with new vendor features, the latter's ROC-AUC is 2% greater than the former





5, Detection Results by new ML model by all data sources

ML model



All performance metrics of the new ML model are better than the rule-based method (decision tree) metrics

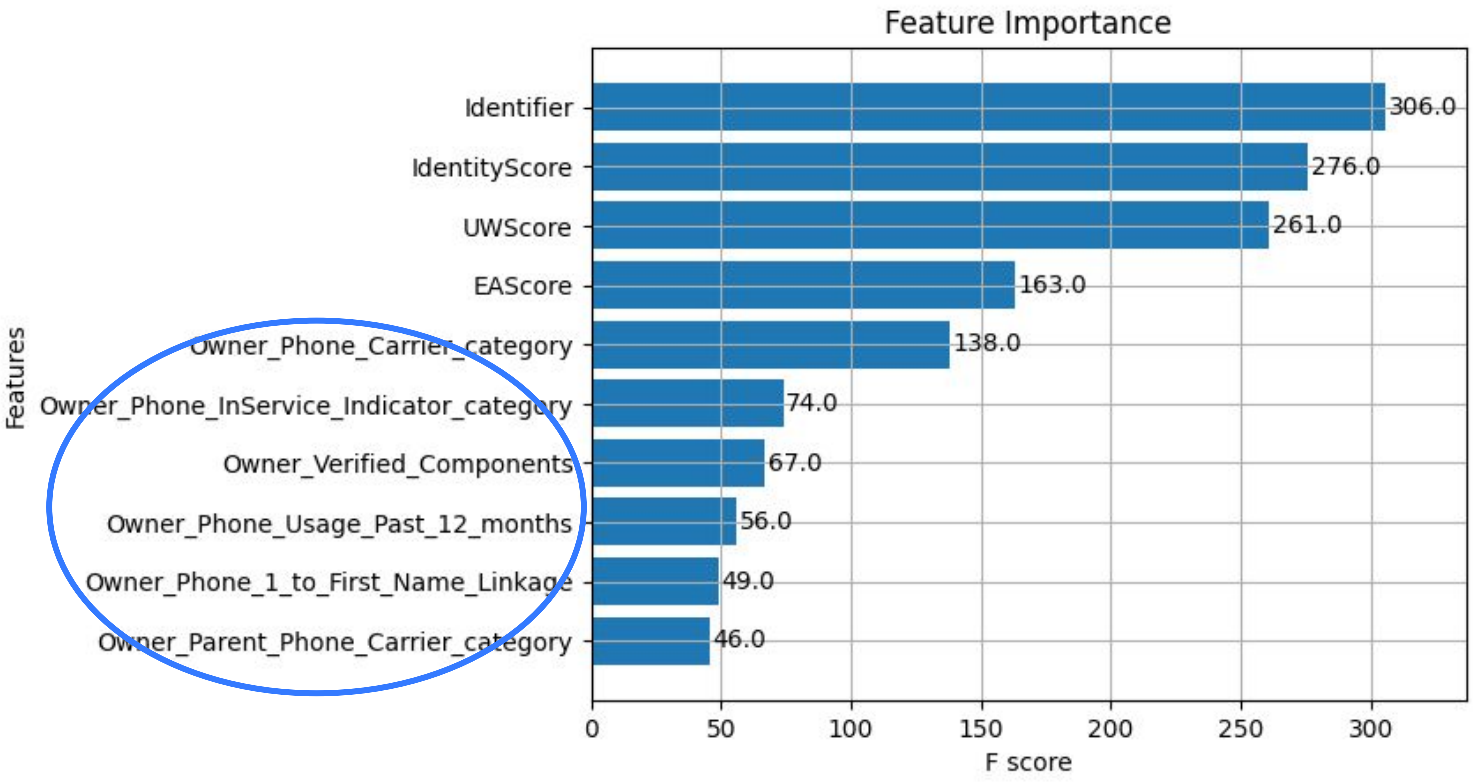
ML model

	precision	recall	f1-score	support
0	0.99	1.00	0.99	4819
1	0.94	0.83	0.88	341
accuracy			0.99	5160
macro avg	0.96	0.91	0.94	5160
weighted avg	0.98	0.99	0.98	5160

Rule-based

	precision	recall	f1-score	support
0	0.98	0.83	0.90	6001
1	0.26	0.77	0.38	449
accuracy			0.83	6450
macro avg	0.62	0.80	0.64	6450
weighted avg	0.93	0.83	0.86	6450

4, modeling by two data sources - feature importance



Methods

Step 1: rule-based, approve negative cases, reject positive cases

Step 2: review all the approved cases from the last step again, put these transactions into detection ML model

Step 3: manual review all the cases who are labeled as fraud by ML model in step 2 again

<u>Reason:</u>
1, Based on the experience of domain experts, the rule model can define some simple conditions to initially determine whether it is fraudulent behavior.
2, Machine learning model are always computing intensive, they can detect more complex fraud patterns that are difficult to define through rules

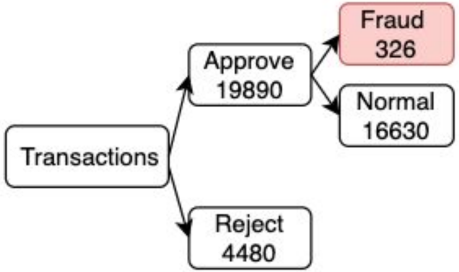
Step 3

ROI Calculation for the next 12 months

ROI with and without new vendor data

- Assume all manual cases are correct

ROI without vendor data



ROI1= revenue/cost-1

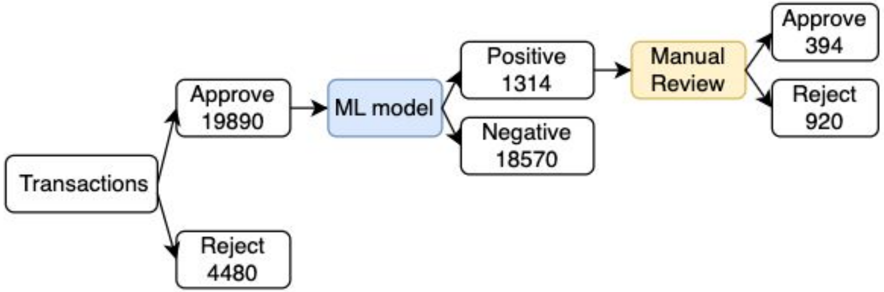
$= \dfrac{\text{revenue by approval}}{\text{fraud cost}} -1$

$= 16950*40*(1+...+12)/(326*500*12)$

$= 26-1$

$=$ **25**

ROI with both vendor data and internal data



ROI2= revenue/cost-1

$= \dfrac{\text{revenue by approval-}}{\text{manual cost+fraud cost+vender call cost}}$

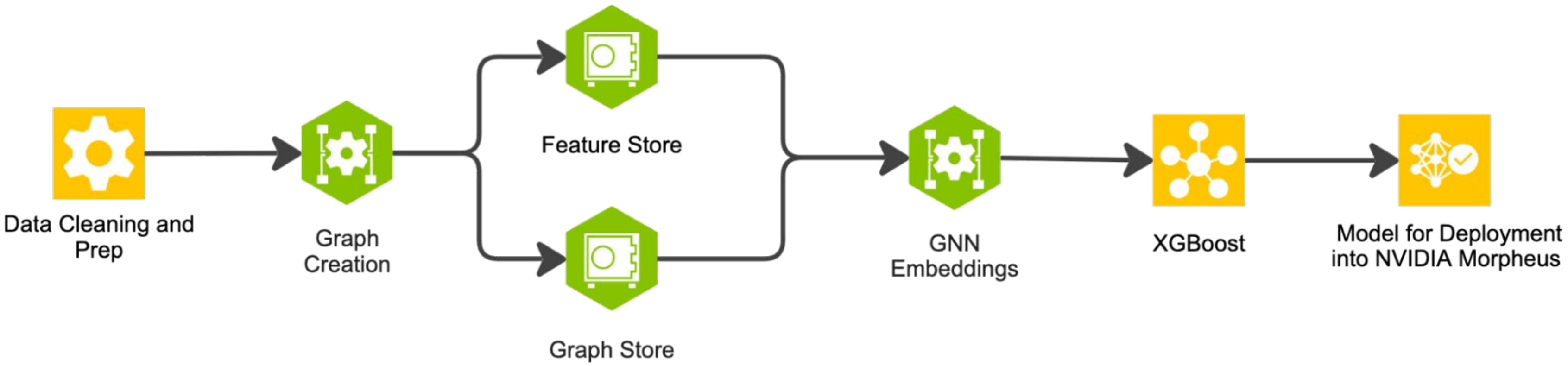
$= 18970*40*(1+...+12)/(1314*50*12+19890*0.5*12)-1$

$= 65-1$

$=$ **64**

Next steps

- Further improve the accuracy of the detection algorithm(more advance model, different threshold)
- Deploy on cloud to check online performance of the model, test the effect with real-time data
- Continuously monitor the long-term performance and iterate the detection algorithms



A discussion of other factors I want to analyze

a. Account Level Risk Data:

- Historical transaction patterns (e.g., frequency, amounts, and geolocation of transactions).
- Demographic data of merchants (e.g., industry, location, business size, and registration date).
- Relationship between merchant's account creation date and the first fraudulent activity (e.g., fraudsters may target newly created accounts).
- Time of day or day of the week the account was created or payments were made.
- Distance between billing address, IP address, and phone number location.
- High-risk regions (e.g., certain countries or states with high fraud rates).
- Textual data (NLP)

b. External Data Sources:

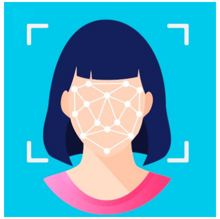
- Historical blacklists: Whether the email, phone number, or IP address is associated with known fraud.

c. Use AI/more fancy models as reference

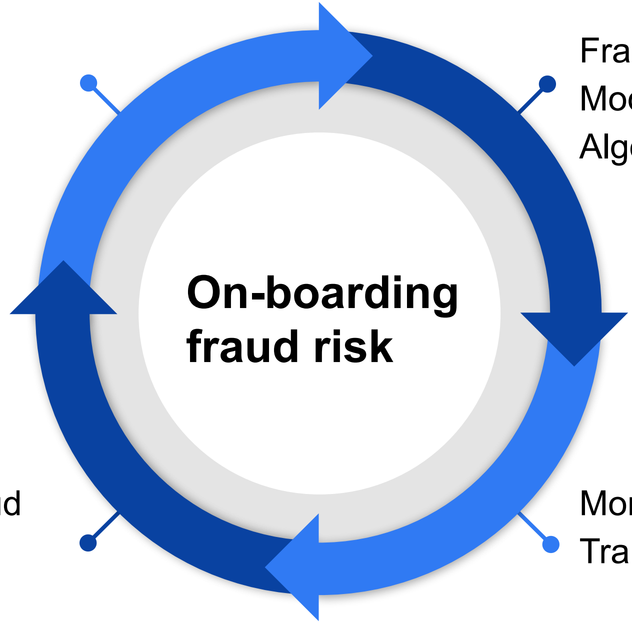
Case Part 2

How to manage fraud risk at new seller on-boarding while balancing growth?

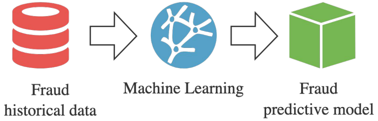
Strategy on managing fraud risk at new seller on-boarding



Strengthen Identity Verification



Fraud Detection Models and Algorithms



Proactive Fraud Education

Monitor Early Transactions

Key metrics system to monitor performance

Balance Growth and Fraud Prevention:

- 1, Try to minimize false positive rate;
- 2, Simplify the onboarding process, especially for low-risk sellers;
- 3, More user-friendly UI and UX to decrease dropout during onboarding process
- 4, Monitor new sellers dropout rate during the onboarding process, establish feedback system by rating and comments

🌟North Star: Fraud Loss Amount per week & New sellers per week
🧱Guadrial: Onboarding Time increase, latency time due to complex onboarding
⚡Driver: Onboarding dropout rate, Revenue created by new merchants, …

- Marketing, Finance, Operation, Product, Engineering

Communication with multifunctional team

Similarity: They are very hard to persuade because new fraud strategy will influence their own objectives and add extra work. Therefore, a common strategy is to emphasize how our new policy fraud and risk strategy can improve and enhance their work.

Difference: Different team have different objectives and communication habit, we should tailor our languages and strategy when handel different stakeholders.

- **Product Team**: They are focused on user experience and integration of fraud management tools without interrupting onboarding. The integration of fraud strategy may affect user experience.
- **Engineering Team**: Engineering cares about technical feasibility and system performance. We should discuss fraud strategy integration and scalability into our platform.
- **Marketing Team**: Marketing is focused on customer acquisition and engagement. However, the stricter onboarding fraud detection may cause decrease in user engagement.
- **Operations Team**: They are concerned with smooth operations and cost efficiency. The new onboarding fraud detection will cause extra work on vendor call or information collection, also more operation cost.
- **Finance Team**: Finance is focused on maximizing revenue. The new onboarding fraud detection will decrease potential new sellers due to longer audit procedures.