Xente Fraud Detection

Ivy

Background

- **Xente** is an e-commerce and financial service app serving 10,000+ customers in Uganda

- **Dataset**: 95K transactions, and among them 193 are frauds.

- **Our Goal**: detect fraudulent transactions and save money

- **Evaluation Metric**: F1-score

EDA
&
Feature engineering

Feature engineering

1. direction of transaction (incoming/outgoing)
2. account fraud history (True/False)
3. transaction hour of day & day of week

Number of transactions

	Incoming	Outgoing	Percentage of Incoming
fraud	188	5	97.4%
non-fraud	57285	38184	60%

Feature engineering

1. direction of transaction (incoming/outgoing)
2. account fraud history (True/False)
3. transaction hour of day & day of week

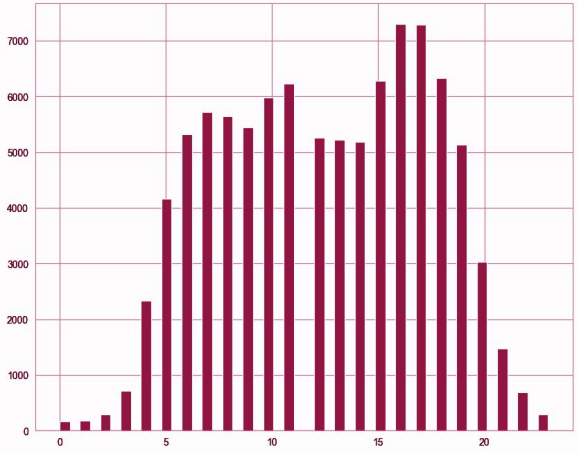
Average number of frauds per fraudulent account id:

3.71

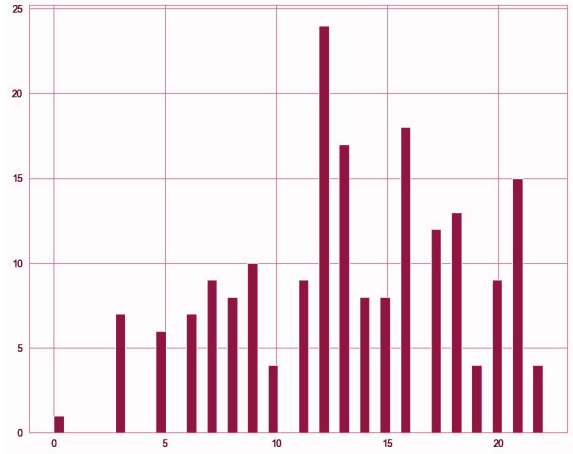
Feature engineering

1. direction of transaction (incoming/outgoing)
2. account fraud history (True/False)
3. transaction hour of day & day of week

non-fraudulent



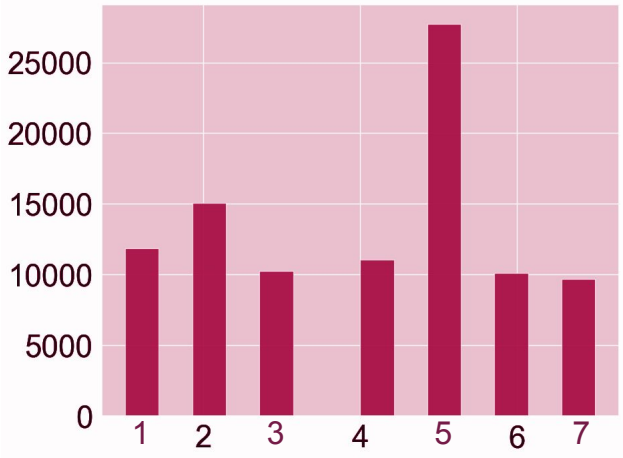
fraudulent



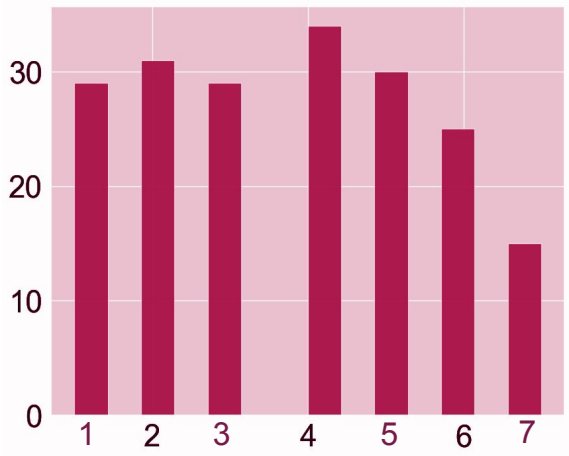
Feature engineering

1. direction of transaction (incoming/outgoing)
2. account fraud history (True/False)
3. transaction hour of day & day of week

non-fraudulent



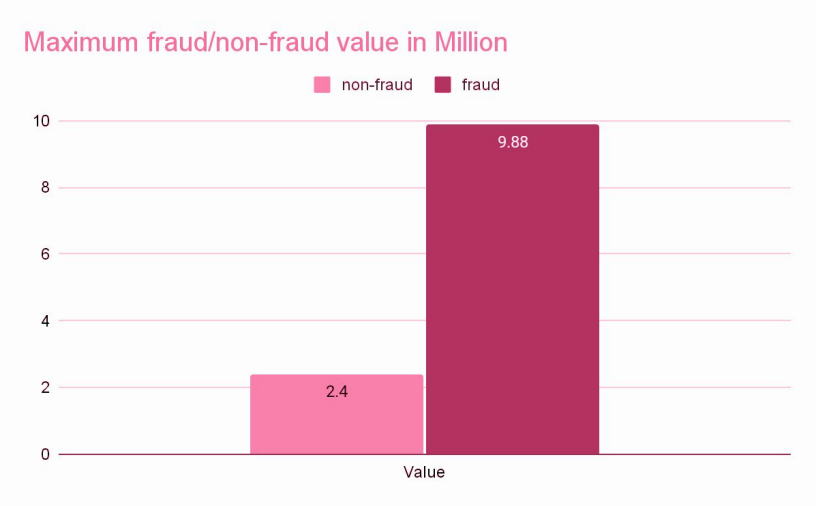
fraudulent

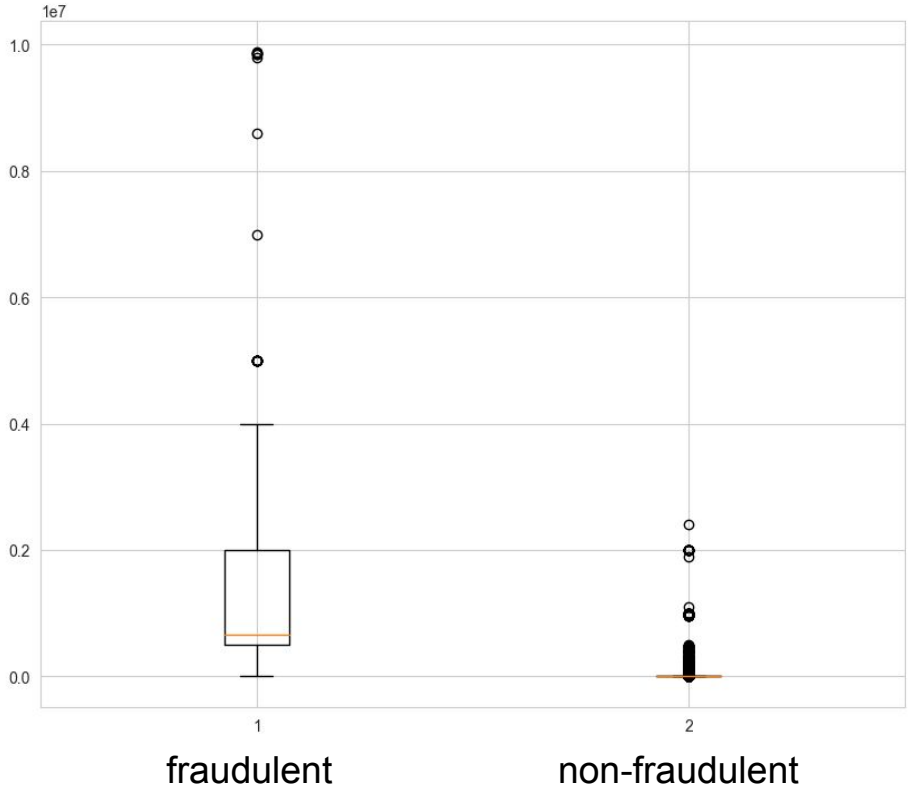


Baseline Model

Baseline model

Value > 2,400,000 → fraud!

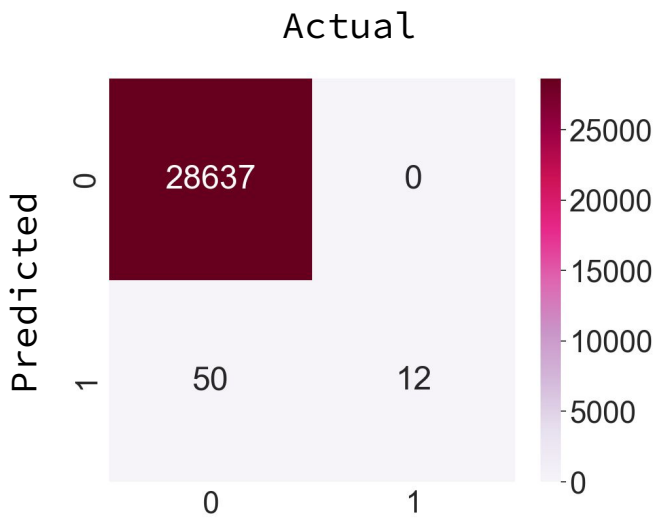




Baseline model

Value > 2,400,000 → fraud!

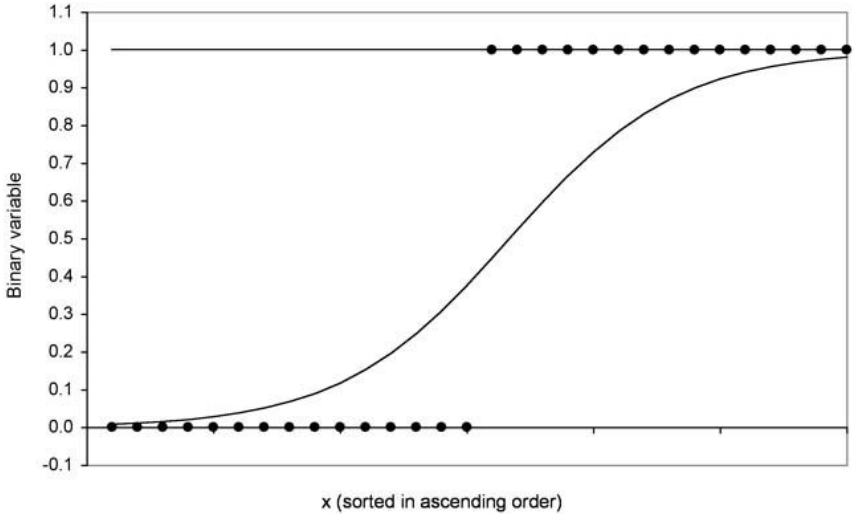
	precision	recall	F1 score
fraud	1.00	0.19	0.32

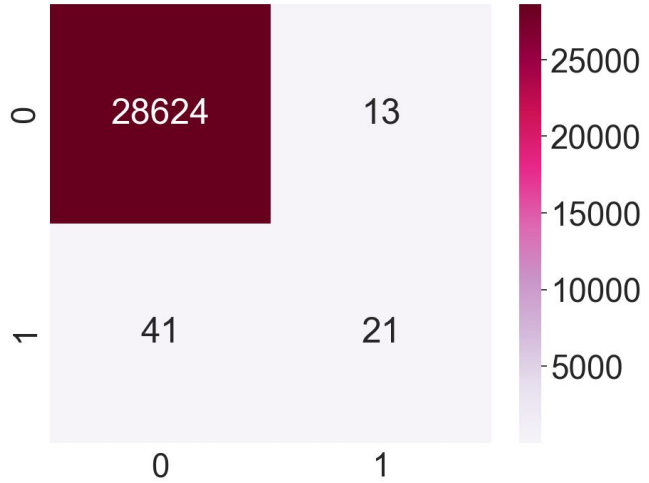


Our Models

Logistic regression

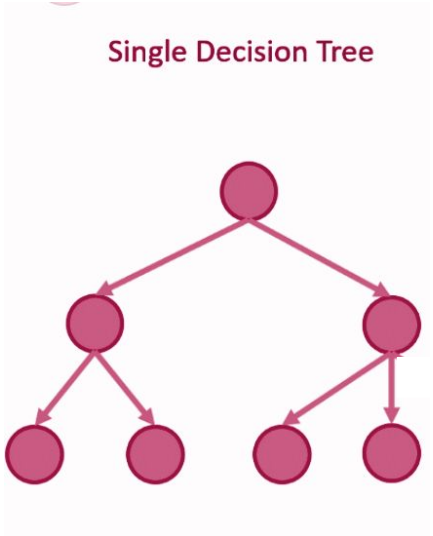
	precision	recall	F1 score
fraud	0.62	0.34	0.44



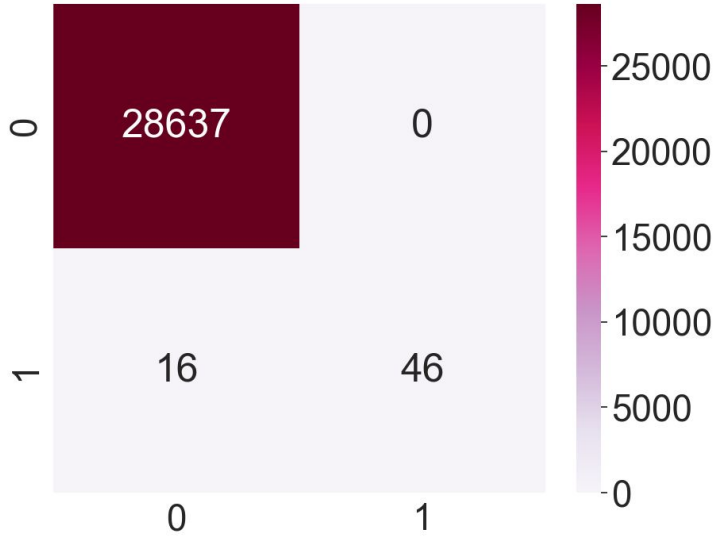


Decision tree

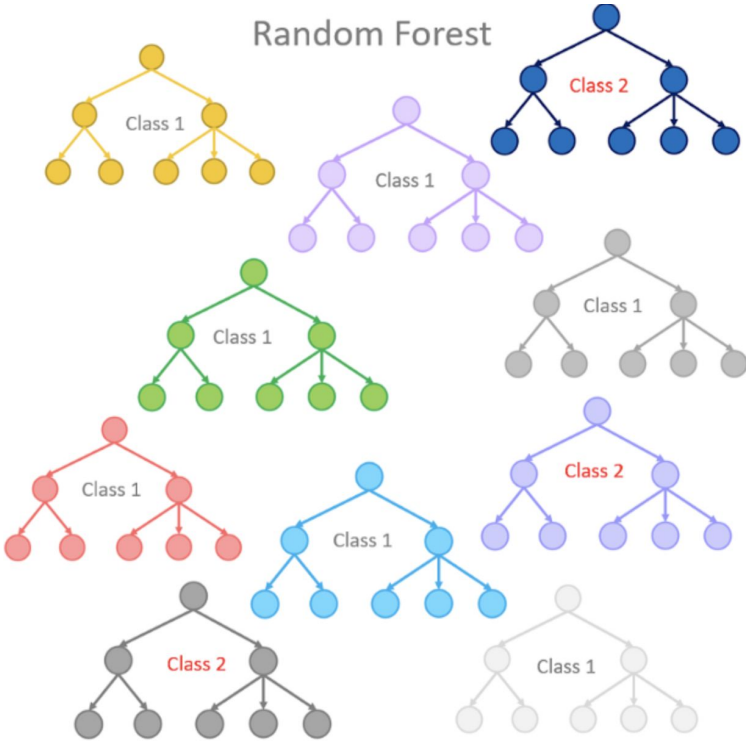
- Gini index
- Entrophy



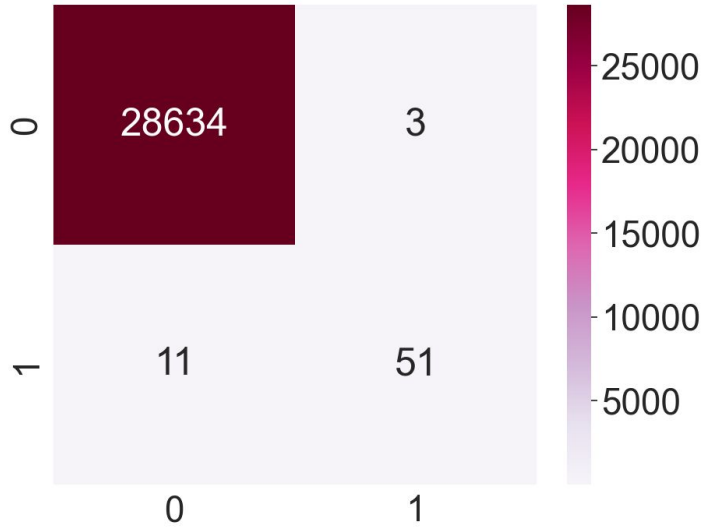
	precision	recall	F1 score
fraud	1.00	0.74	0.85



Random forest

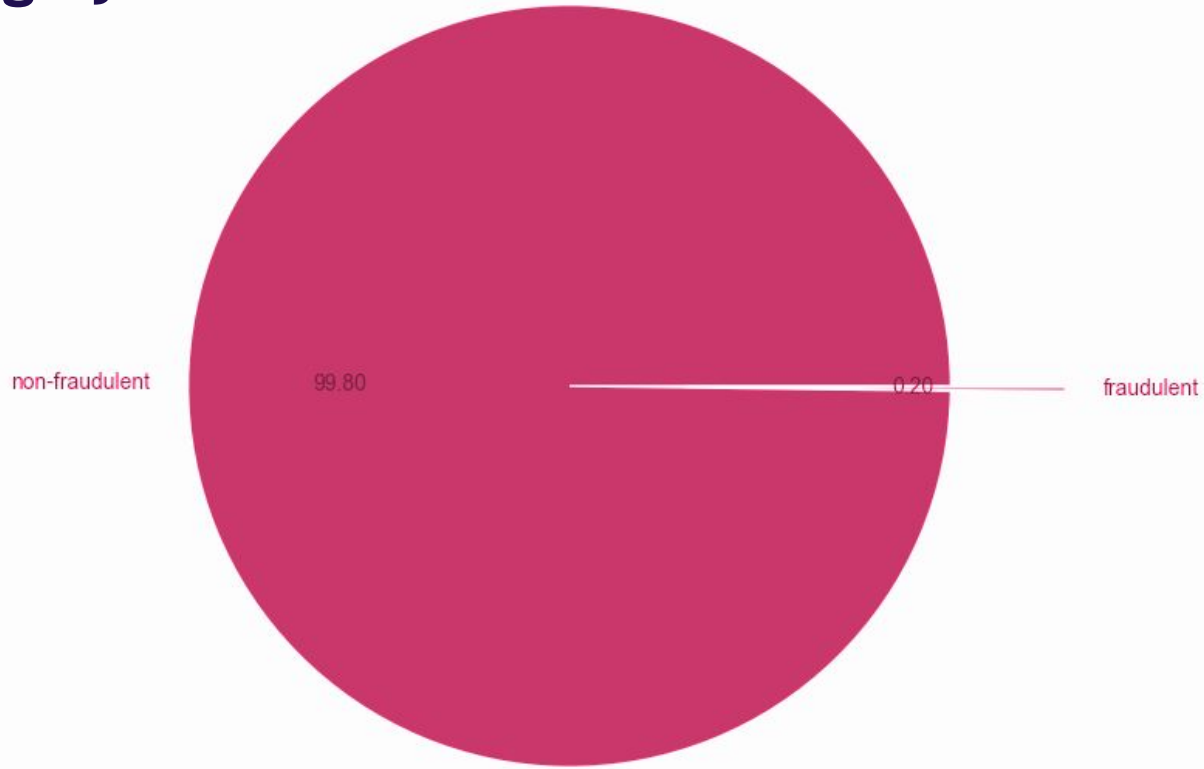


	precision	recall	F1 score
fraud	0.94	0.82	0.88



Dealing with imbalanced dataset

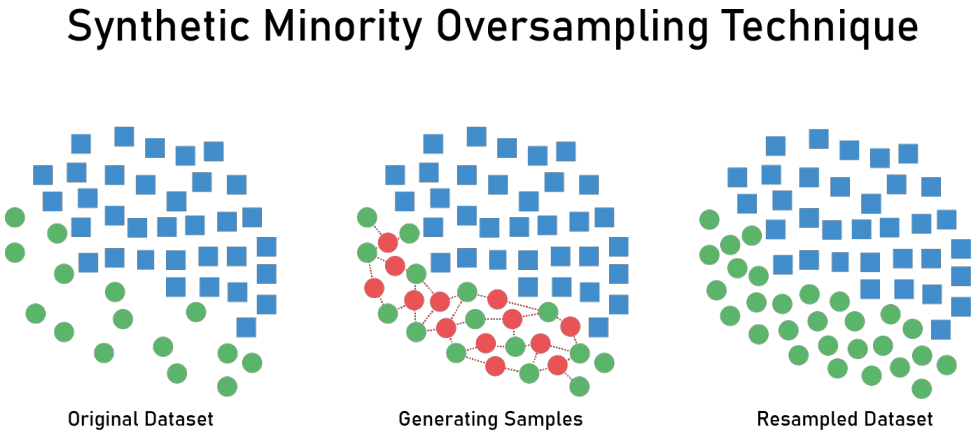
Data is highly imbalanced



SMOTE

Synthetic Minority Oversampling Technique

- randomly pick a point from the minority class
- compute the k-nearest neighbors for this point.
- The synthetic points are added between the chosen point and its neighbors.



SMOTE with random forest

Fraud : Non-fraud = 1 : 2

Before:

0 95469

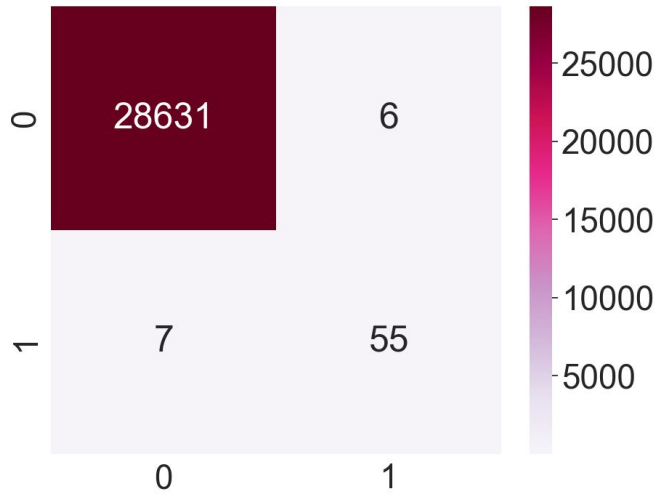
1 193

After:

0 66832

1 33416

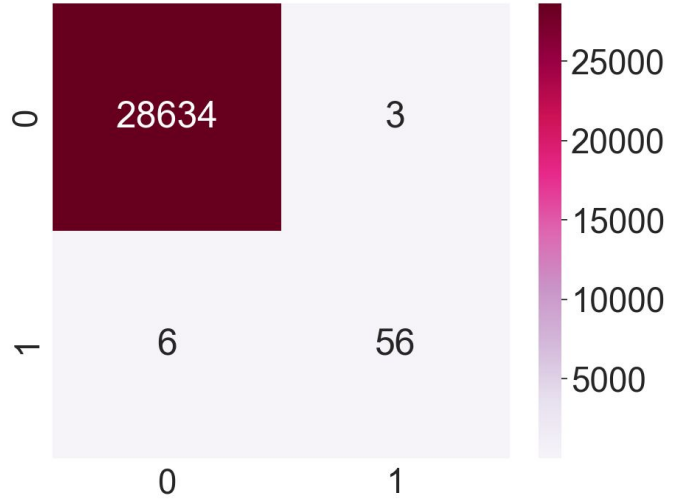
	precision	recall	F1 score
fraud	0.90	0.89	0.89



Ensemble methods

- AdaBoost algorithm: iterative approach to learn from the mistakes of weak classifiers
- SMOTE dataset
- Base estimator: Decision tree

	precision	recall	F1 score
fraud	0.95	0.90	0.93

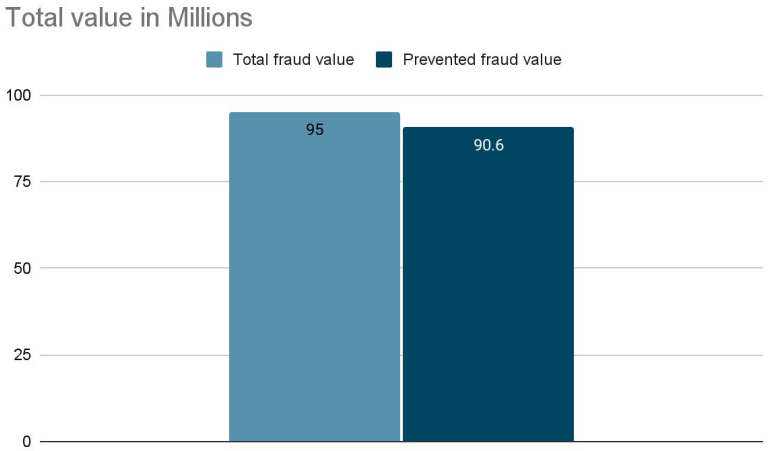


Loss prevented by using our final model

Money saved

We successfully prevented 56 fraud cases out 59. (94.9% of total fraud cases)

We prevented 90.6M loss by detecting fraud. (95.4% of total fraud value)



Error Analysis

Error Analysis

all of the 9 misclassified records are of

- Product Category 2,
- Pricing Strategy 2,
- Value category 0 (incoming),
- Fraud history 1 (has fraud history),

so it is advisable to be extra careful with transactions with these characteristics.

Thank you!