Eskwebanks:
Real Time Fraud Detection

What is Fraud?

"Any attempt of criminals to achieve financial gain at the expense of legitimate customers of financial institutions."

First-Party Fraud

Legitimate customer betrays the bank

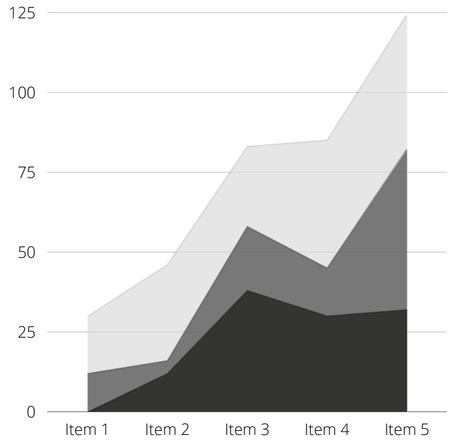
Third-Party Fraud

Customer becomes a victim of criminals who steal identities and use lost or stolen cards

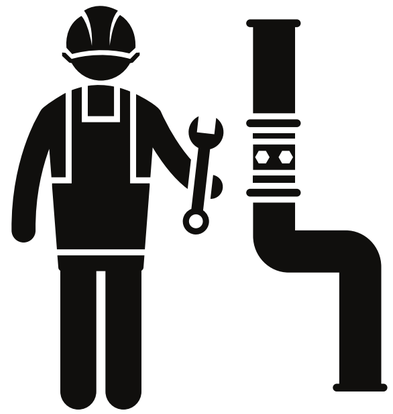
Source: Arvid Hoffman and Cornelia Birnbrich, "The Impact of Fraud Prevention on Bank-Customer Relationships," *www.emeraldinsight.com*

PREPARING THE DATA

Drop sensitive information
Feature engineering (age, distance, datetime)



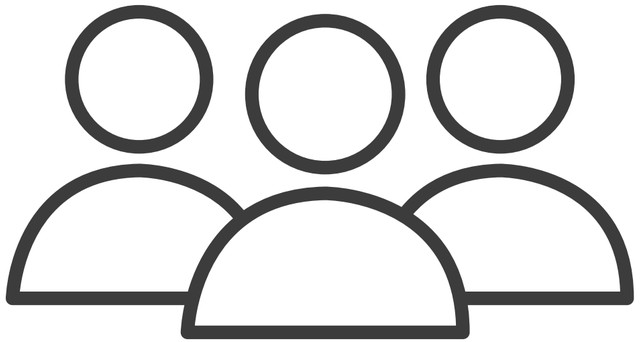
DATA EXPLORATION



BUILDING THE PIPELINE

Indexer
One-hot encoder
Vector Assembler
Random Forest Classifier
Cross Validator

The Process



CREDIT CARD HOLDER

Negative emotional impact

Time spent to block, close, and re-open account

BANKS

Monetary Losses - refunds

Negative Impact on Bank's Reputation

Loss of customer's trust

MERCHANTS

Negative Impact on Store's Reputation

Additional costs - chargebacks, fees, and merchandising replacement

Impact of Fraud

Impact of Fraud on Banks

90,000

Fraudulent transactions were detected out of 25,000,000 transactions from 2018 - 2020

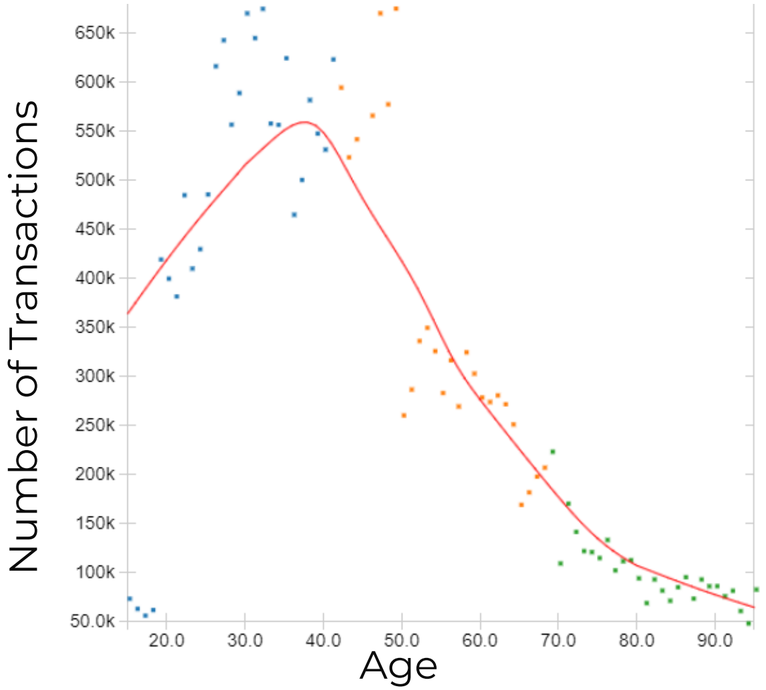
P15,000,000

is lost annually on the average due to fraud

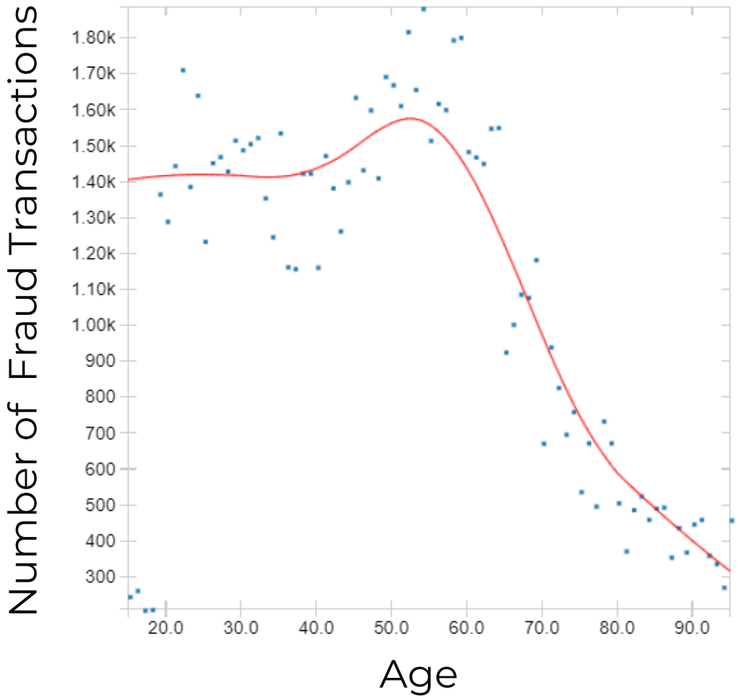
Typical Fraud Patterns

Distribution of Total Transactions across Ages

All Transactions



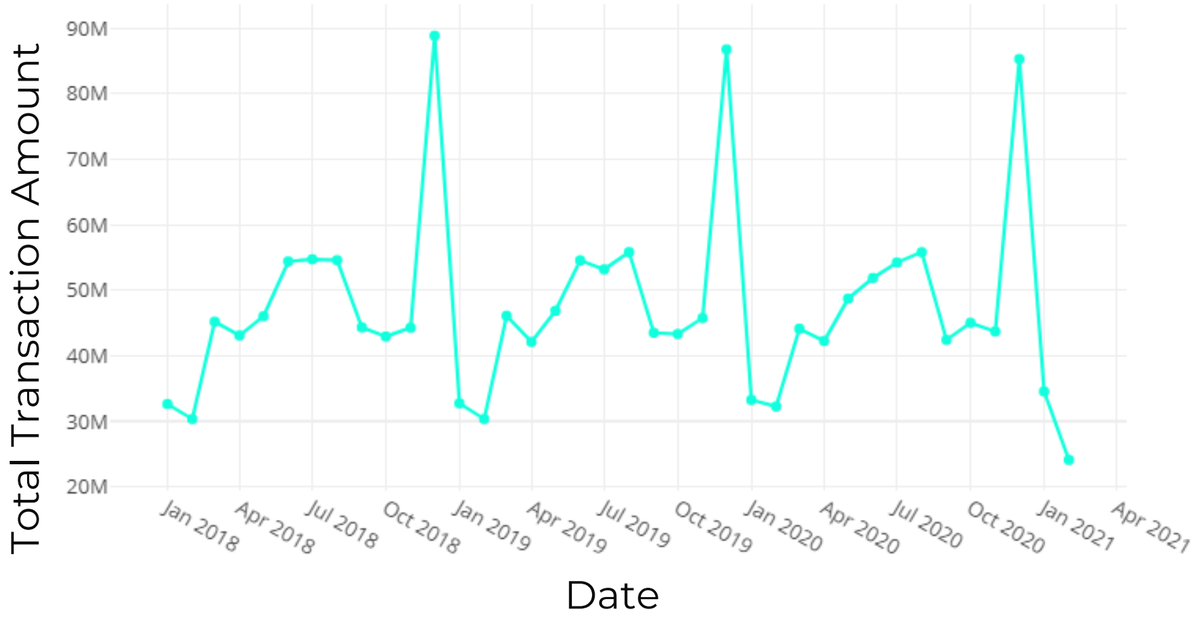
Fraudulent Transactions



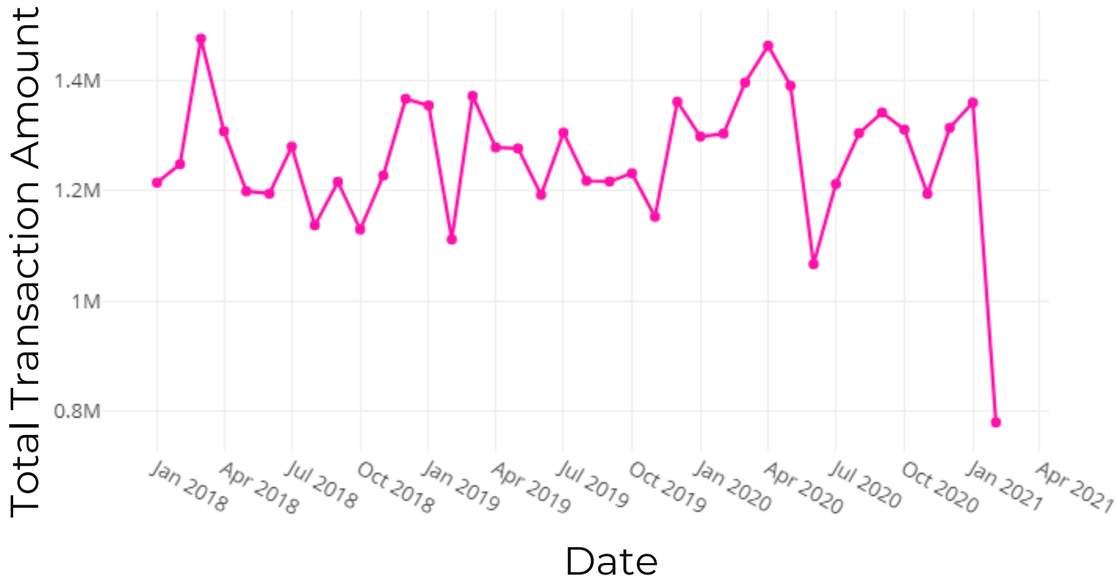
People aged 45 - 65 years old are more susceptible to fraud.

Typical Fraud Patterns

All Transactions



Fraudulent Transactions



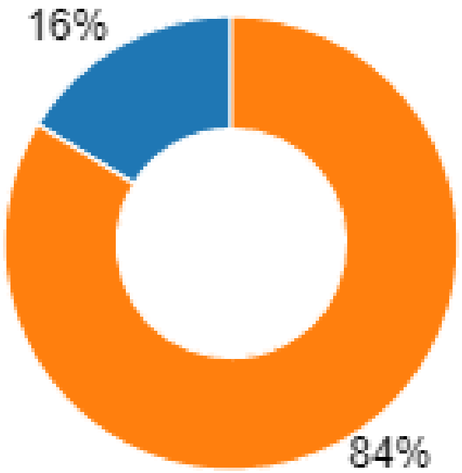
Total transactions tend to increase during December.

Total fraudulent transactions exhibit no obvious pattern throughout the years.

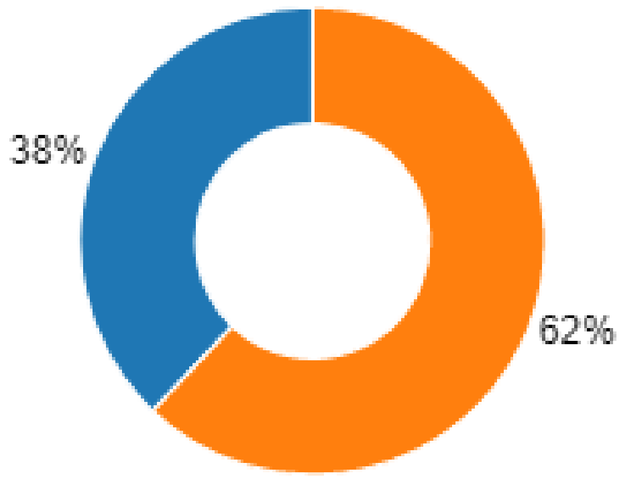
Typical Fraud Patterns

Physical Transactions
Online Transactions

Distribution of Physical and Online Transactions



All Transactions

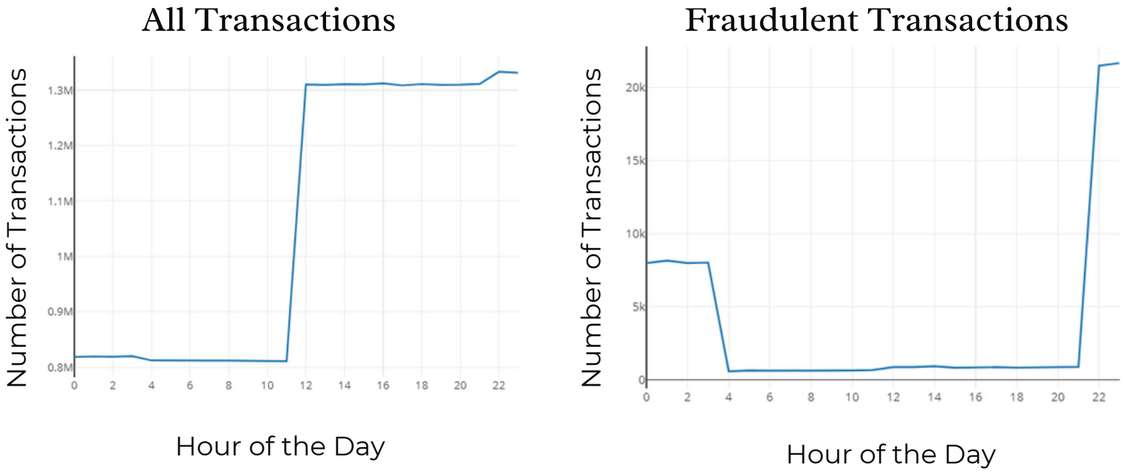


Fraudulent Transactions

Average Transaction Amount

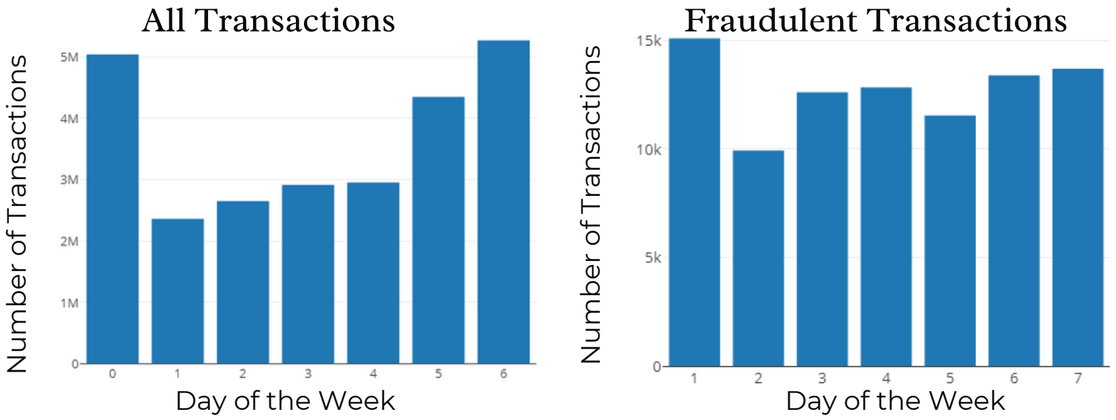
- Online transactions comprise a larger portion of fraudulent transactions.
- Fraud transactions tend to be larger than non-fraud transactions.

Typical Fraud Patterns



No. of Transactions per Hour of the Day

Fraudulent transactions tend to occur during late nights to early mornings.



No. of Transactions per Day of the Week

There are usually more transactions during the weekends, but this pattern is not as evident in the fraud transactions.

Fraud Detection Modelling

Modelling Process

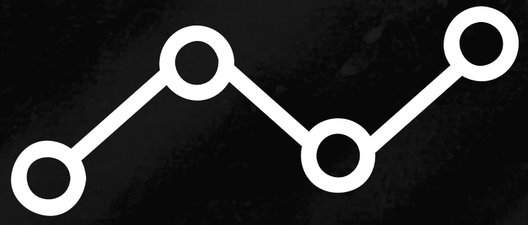
1

Undersampling of imbalanced dataset.

2

Train model on balanced dataset.

3



Manage ML builds using MLflow.

4

Test model on 'real-time' transactions.

Undersampling

25M → 178k

Used only 0.71% of the entire dataset

Features: Age, Transcation Amount, Category. Location, Hour of the Day, Day of the Week, Day of the Month, City Population, Distance of Trans. from Home Address

To address the issue of imbalance within the dataset

And faster too!

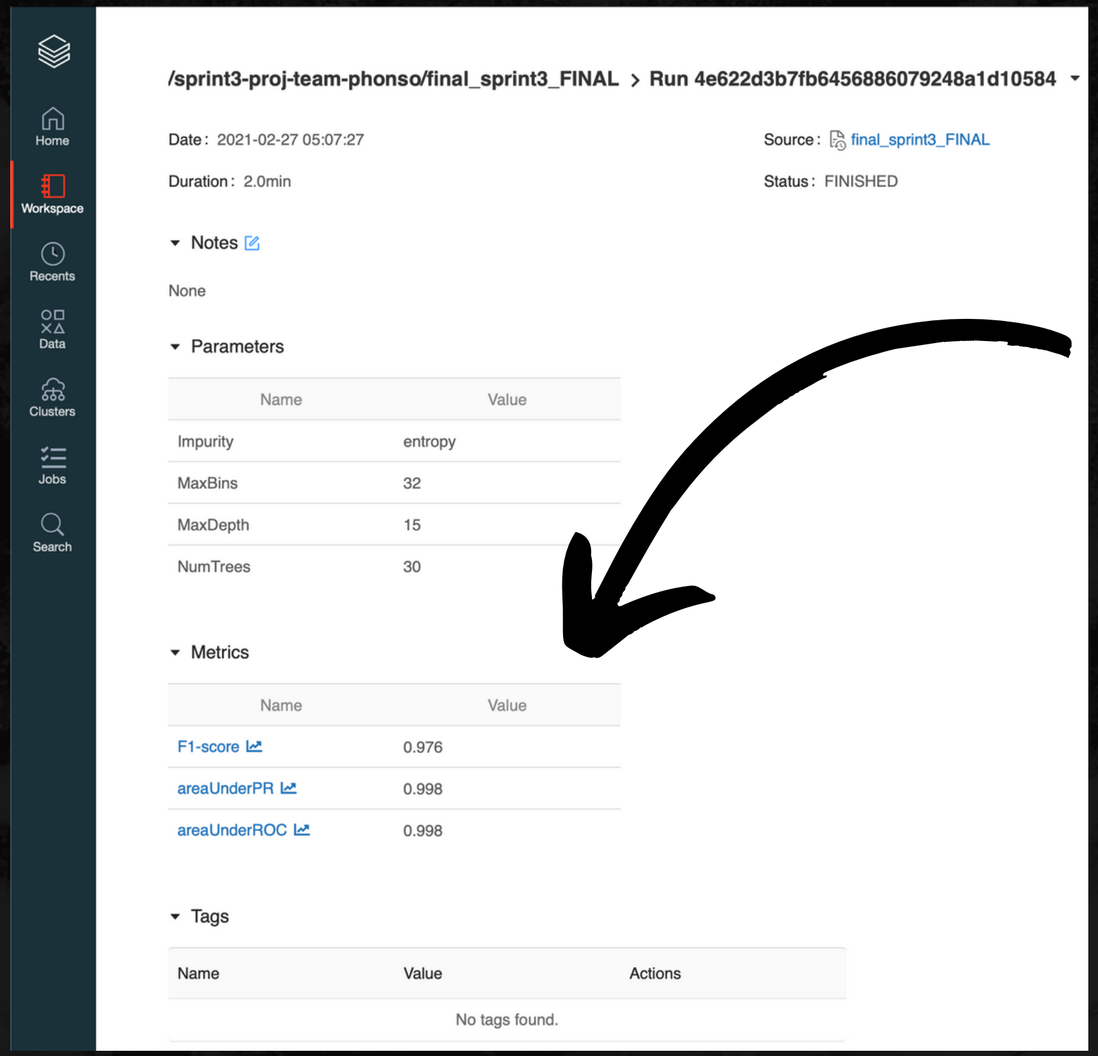
Model Results (5-fold CV)

Model	F1 Score
Random Forest	97.79%
Decision Tree	94.95%
Gradient Boosting	94.57%
Logistic Regression	80.76%

Hyperparameter	Values
Max Depth	15
Impurity	Entropy
No. of Trees	30

Feature	Importance
Amount	0.60
Hours	0.18
Gas Transport	0.03
Shopping Net	0.03
Age	0.02
Grocery POS	0.02
Shopping POS	0.02

Model Results (Test set)



F1 - score	97.6%
AUC-ROC score	99.8%
Area under PR	99.8%

*Test set contains 328 records only

Real Time
Fraud Detection

Conclusions

With realtime fraud detection system, our beloved client EskweBank, would greatly improve its processes *(e.g. time-spent to block, refund, close, open, and tag accounts and transactions)*. Therefore, EskweBank would have a competitive edge in its mission of fighting fraud.

EskweBank's clients *(i.e. merchants and CC holders)* would have better customer experience in terms of trust and security.

Postive impact on EskweBank's reputation.

Recommendations

Upon detection of fraud transactions, it is recommended that Eskwebanks **flag consumers on a potential threat** through call or text. This is due to the **low false negative** and **higher false positive** detection of the model.

Oversampling using SMOTE algorithm gave us promising baseline ML model metrics. With that, we recommend to further **improve the fraud detection** model by **experimenting on other algorithms**.

We also recommend using **a larger, balanced dataset in modeling**.

Meet the Team

MASTER PHONSO

PySpark Chairman and CEO

ANDREI

PySpark Undersampler

NICO

PySpark SQL Master

RODS

PySpark Lead Engineer

KAYE

PySpark Goddess