Team 3
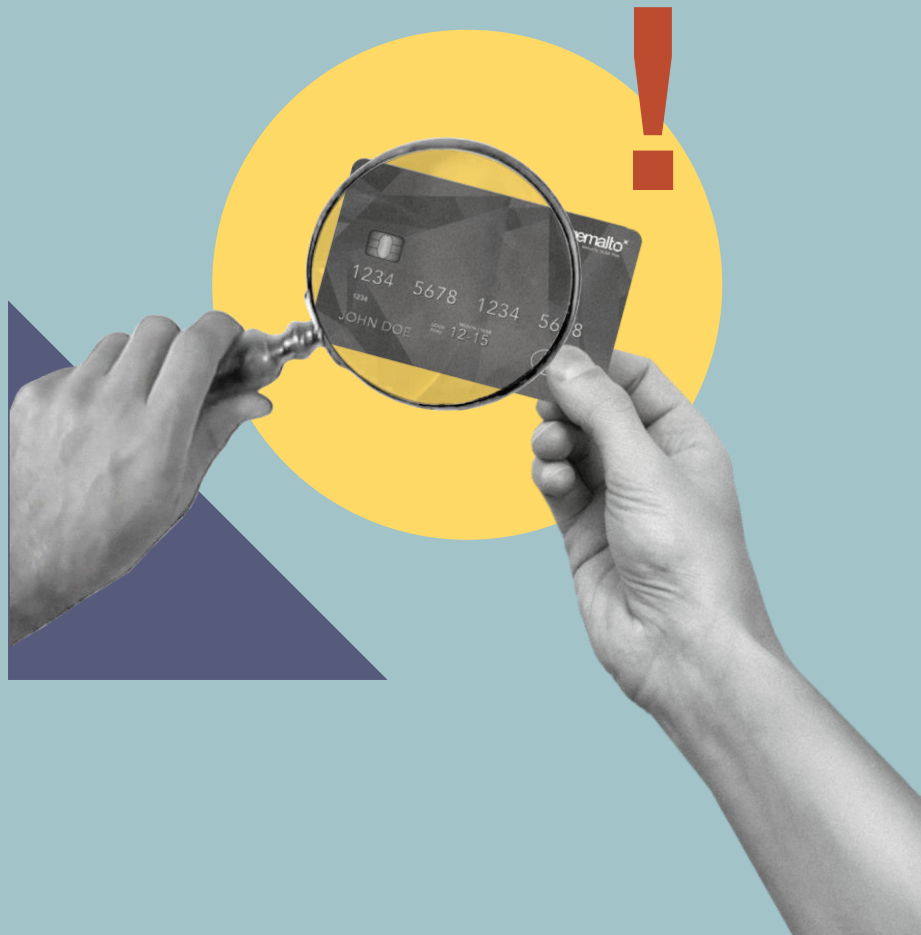
Detecting
Credit Card Fraud
Using
Machine Learning

Members

Edward Apostol
Eric Magno
Jay Isidro
Phoemela Ballaran

Mentor

Ric Alindayu

What is credit card fraud?

According to the FBI, credit card fraud "*... is the unauthorized use of a credit or debit card, or similar payment tool to fraudulently obtain money or property.*"

What are the effects?

Aside from the loss of the card holder, fraud can cause companies to suffer loss of brand value which impacts their growth plans, leading to erosion in shareholder value.

Features Used

Personal

Gender
Age

Location

City
City population
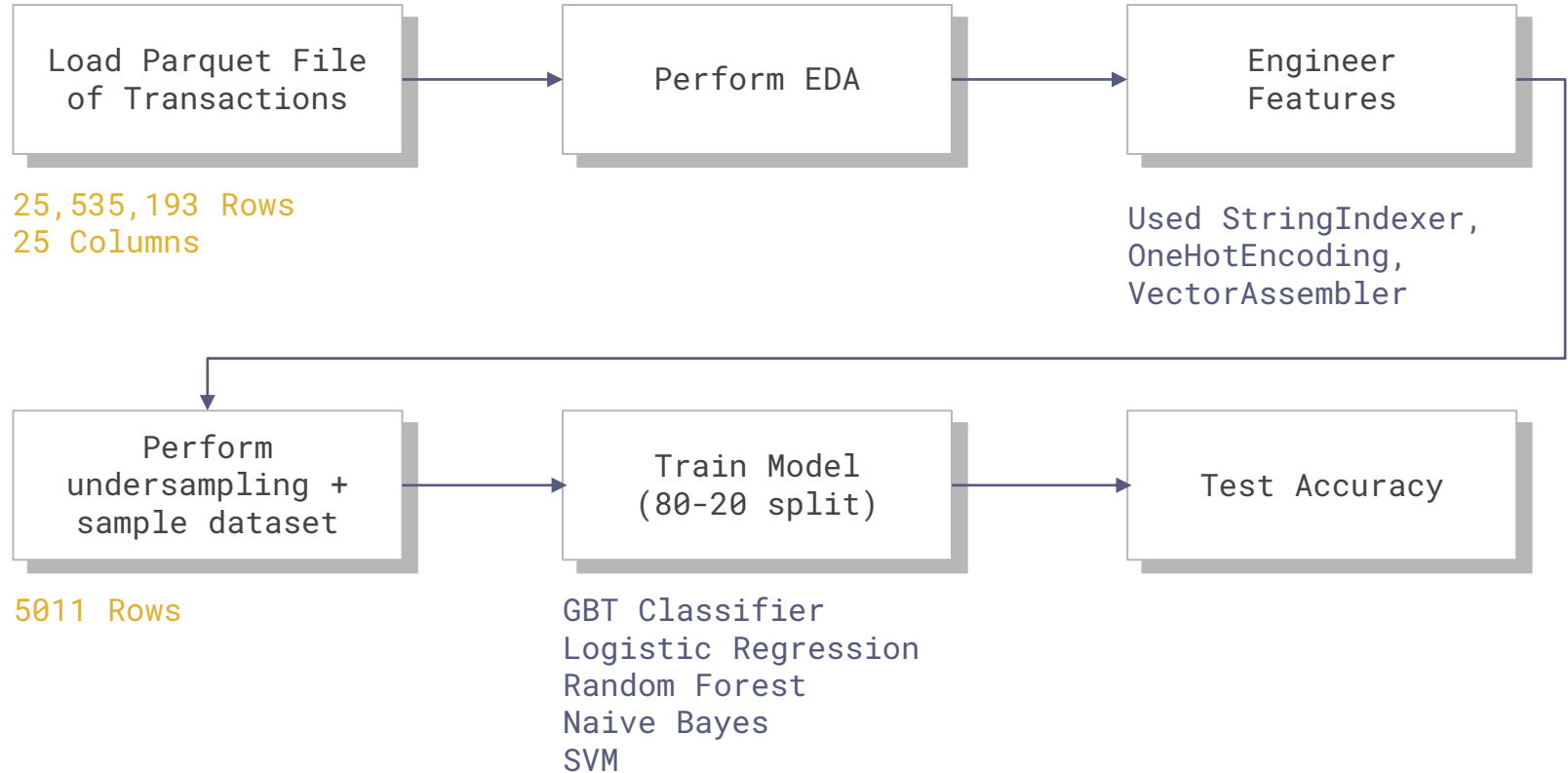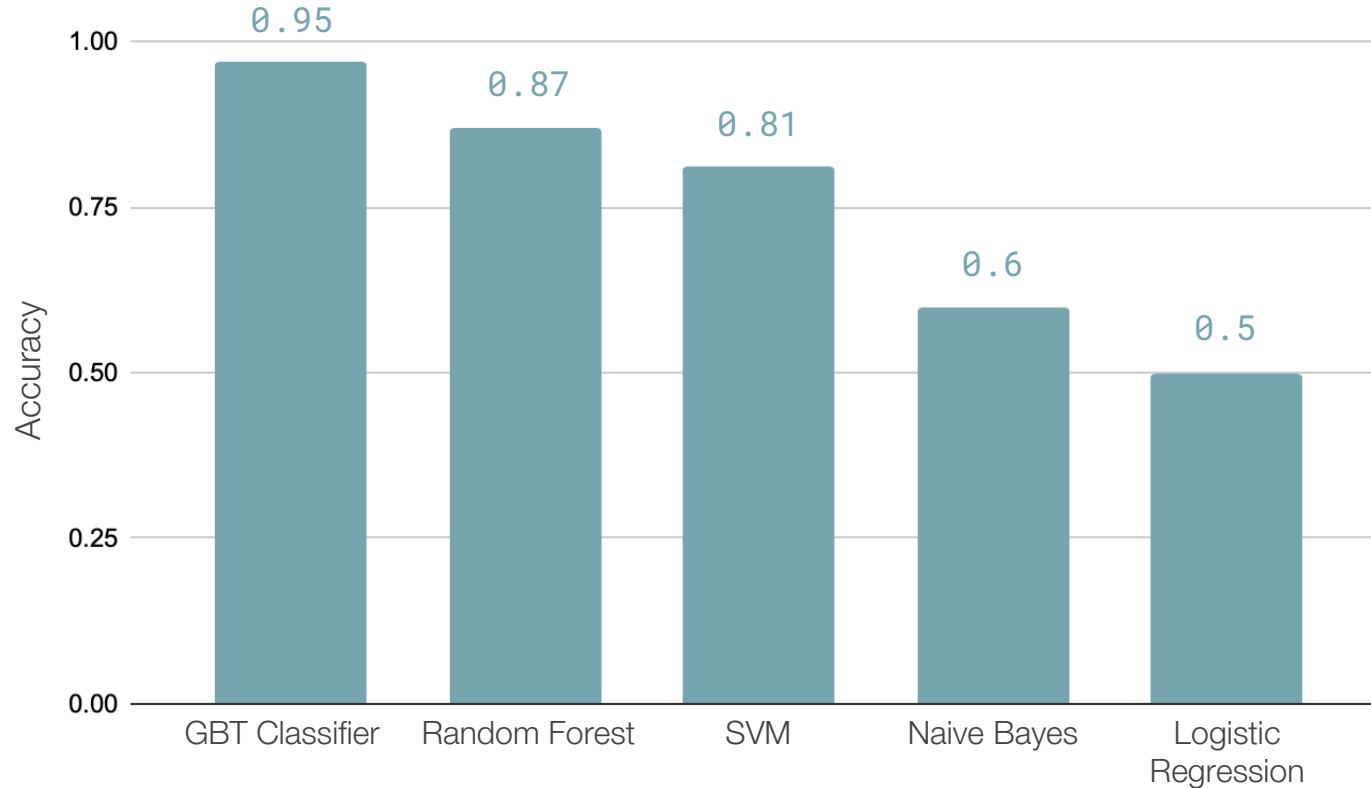Merchant distance

Transaction

Category
Type
Amount

Time

Year
Month
Day of week
Hour

Methodology

Load Parquet File of Transactions

25,535,193 Rows
25 Columns

Perform EDA

Engineer Features

Used StringIndexer,
OneHotEncoding,
VectorAssembler

Perform undersampling + sample dataset

5011 Rows

Train Model
(80-20 split)

GBT Classifier
Logistic Regression
Random Forest
Naive Bayes
SVM

Test Accuracy

How did the models perform?

Accuracies for 5 runs of the **GBT Classifier** model using different samples:

1st run	0.97360
2nd run	0.957248
3rd run	0.94820
4th run	0.961662
5th run	0.941618
avg	**0.956466**

Taking a closer look at the GBT Classifier model

Confusion Matrix

Label	Prediction	Count
Fraud	Fraud	518
Fraud	Not fraud	9
Not fraud	Fraud	18
Not fraud	Not Fraud	486

Classification Report

	Precision	Recall	F1-score	Support
0 (not fraud)	**0.98**	0.96	0.97	504
1 (fraud)	0.97	**0.98**	0.97	527
accuracy			0.97	1031
macro avg	0.97	0.97	0.97	1031
weighted avg	0.97	0.97	0.97	1031

● 2% of genuine transactions are labeled as fraud.
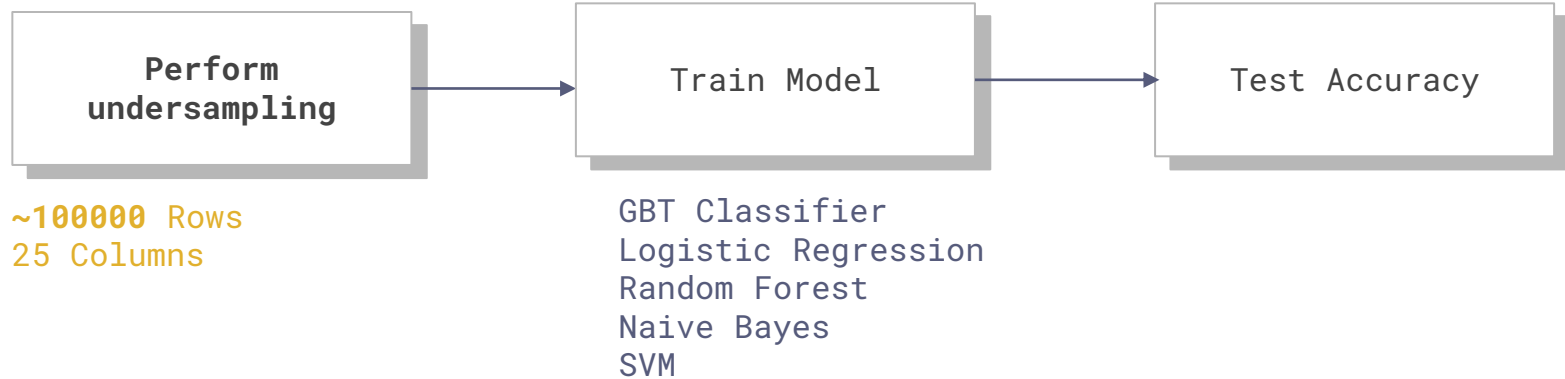
(False positive)

● 98% of fraudulent transactions are detected. 2% are not detected.
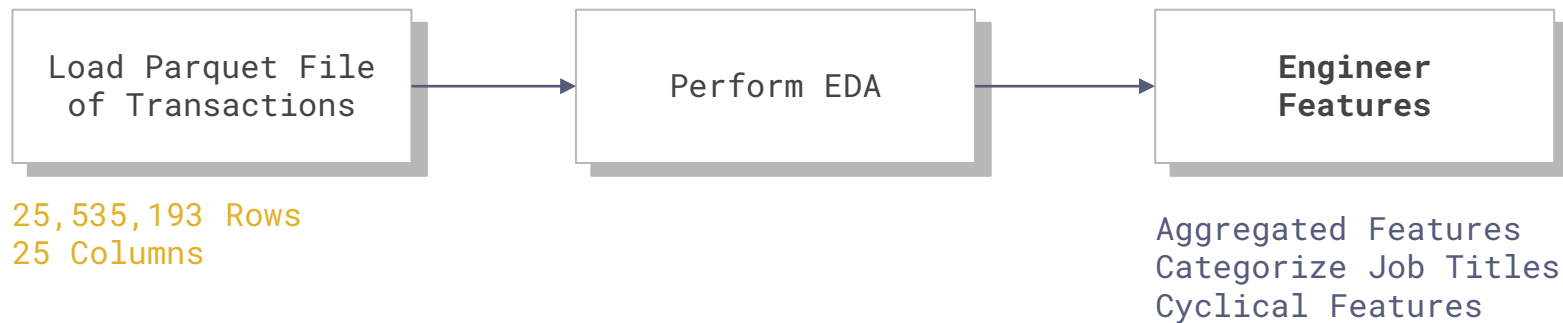(False negative)

Recommendations

1   Use the entire undersampled balanced dataset with GBT Classifier

Perform undersampling	→	Train Model	→	Test Accuracy

~100000 Rows
25 Columns

GBT Classifier
Logistic Regression
Random Forest
Naive Bayes
SVM

In the interest of **computational efficiency**, a smaller sample size of the dataset was used. It's suggested to use the entire dataset with undersampling in the next iteration of this project.

Recommendations

2   Additional feature engineering

Load Parquet File of Transactions	→	Perform EDA	→	**Engineer Features**

25,535,193 Rows
25 Columns

Aggregated Features
Categorize Job Titles
Cyclical Features

There are more features that can be extracted from the raw dataset such as:

- **aggregated features** like average amount spent per credit card number in a given timespan
- **categorizing job titles** using Natural Language Processing
- converting specific columns into **cyclical features** such as hours, day of week, month of year, etc.

Recommendations

3 Model experimentation & hyperparameter optimization

GBT Classifier
Logistic Regression
Random Forest
Naive Bayes
SVM
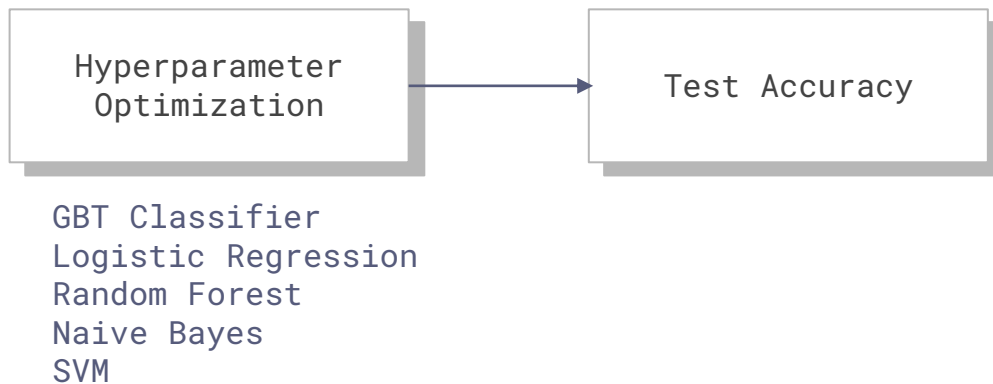
- **Different models** could have been tested without the existing time constraints.

- **Hyperparameter optimization** could have also been applied to the models that displayed high accuracy.

The End

Thank You