SC1015 Mini Project

FCE3 Group 3:

- Damien (U2322620H)
- Yik Sheng (U2322759F)
- Jung Kit (U2322047C)

Content

Introduction

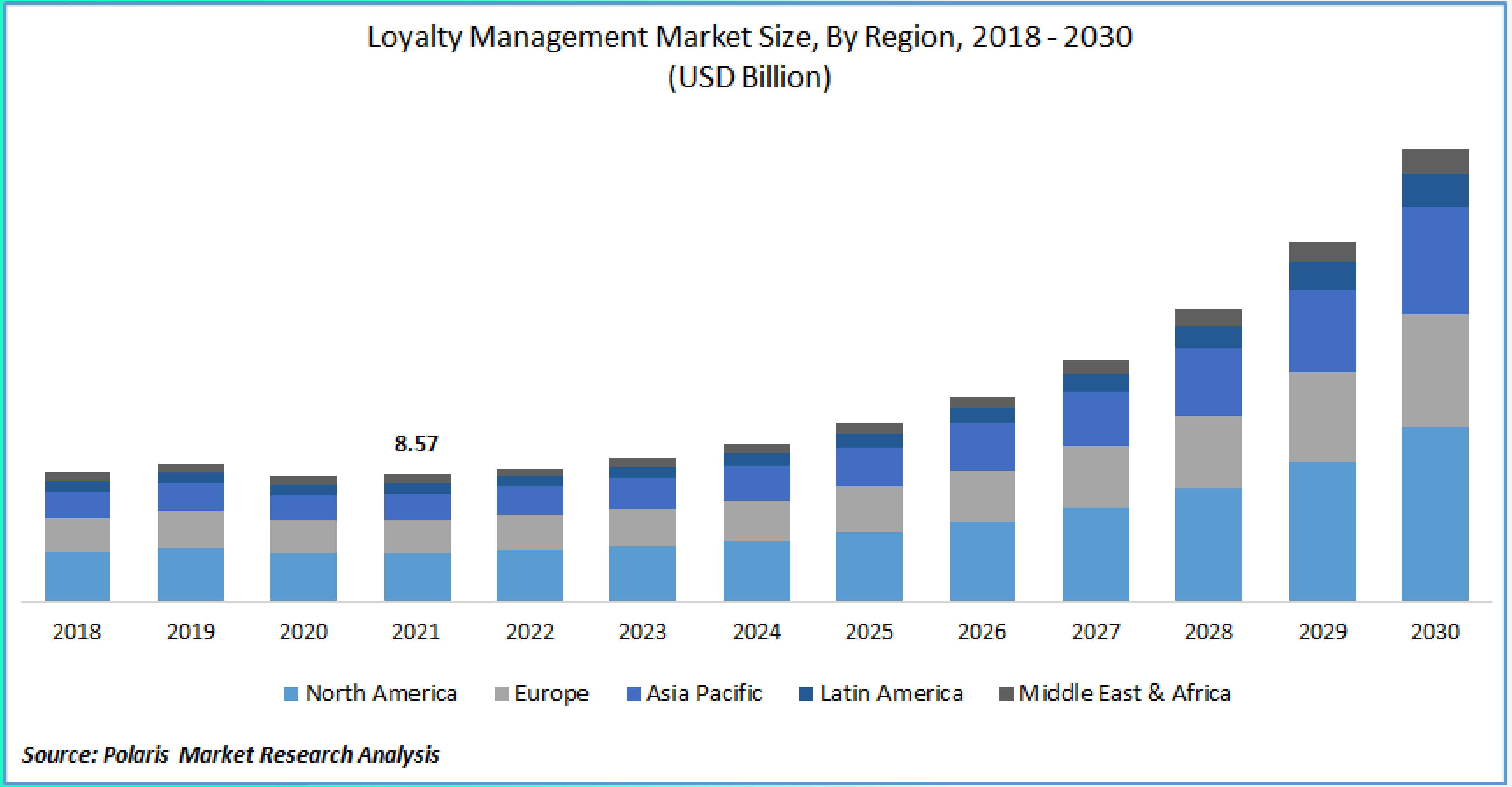
Problem statement

Exploratory Data Analysis

Machine learning

Insights and conclusion

Loyalty Management



Loyalty Management Market Size, By Region, 2018 - 2030 (USD Billion)

Source: Polaris Market Research Analysis

Global Market Size

Valued at **$8.57 billion** in 2021

Is expected to grow at Compound Annual Growth Rate (CAGR) of **16.5%**

Meaning it will reach a Market Size of **$33.8 billion!!!**

Loyalty Management

Recurring Revenue

“84% of consumers say they're more likely to stick with a brand that offers a customer loyalty scheme.”

Lower effort for sales

“You have a 60-70% chance of selling to an existing customer, versus a 5-20% chance of selling to a new prospect.”

Problem Statement

1. How can businesses increase the chances of consumers enrolling in their loyalty program?

2. What are the factors that affect a consumer decision on enrolling in a loyalty program?

Exploratory Data Analysis

Customer Shopping Preferences Dataset offers insights into consumer behaviour and purchasing patterns mainly in United States of America.

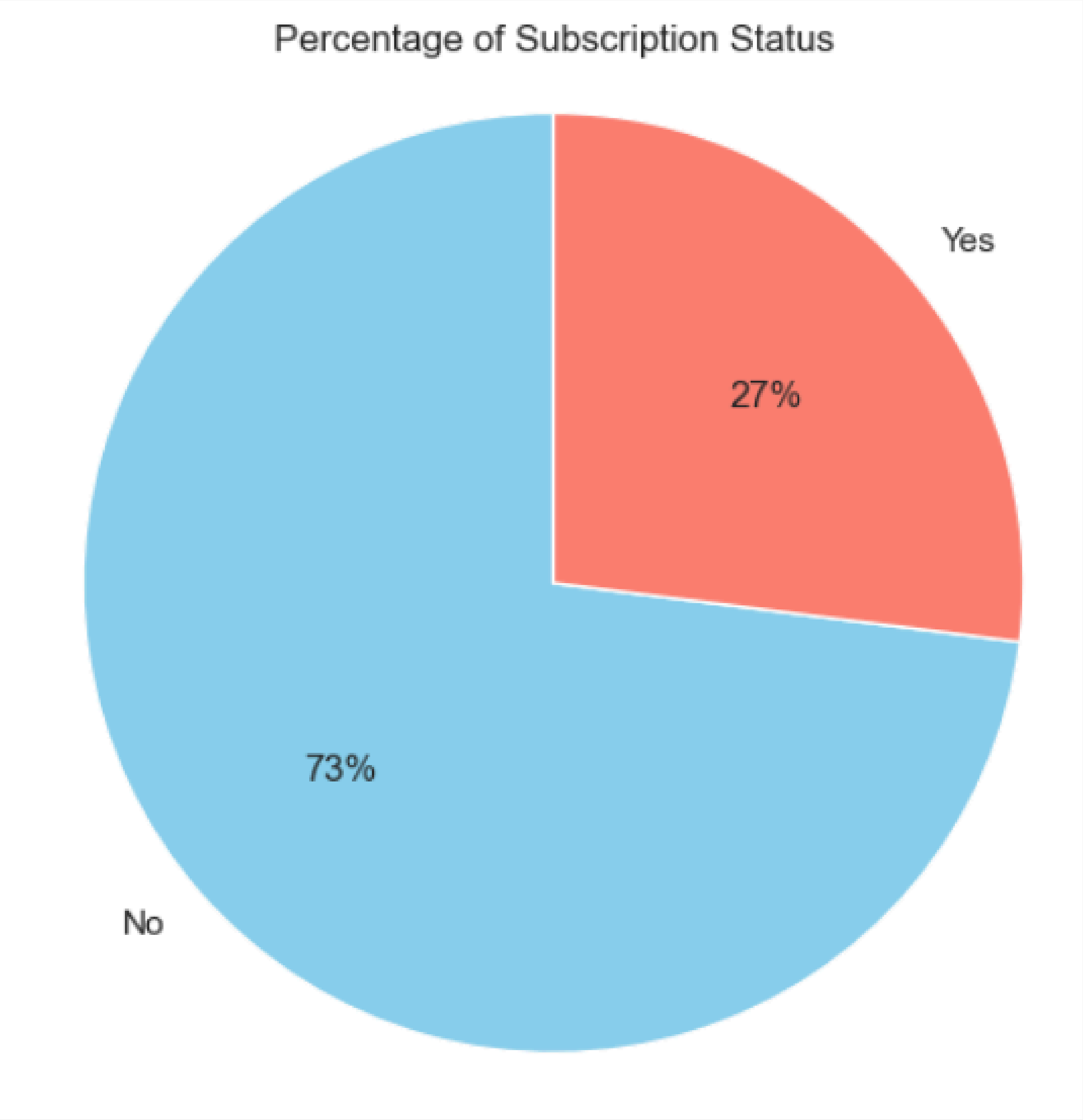
Features

- Age
- Discount applied
- Frequency of Purchases
- Pervious purchase
- Review ratings
- Payment Method
- Items purchased
- size
- Purchase amount (USD)

Subscription Status

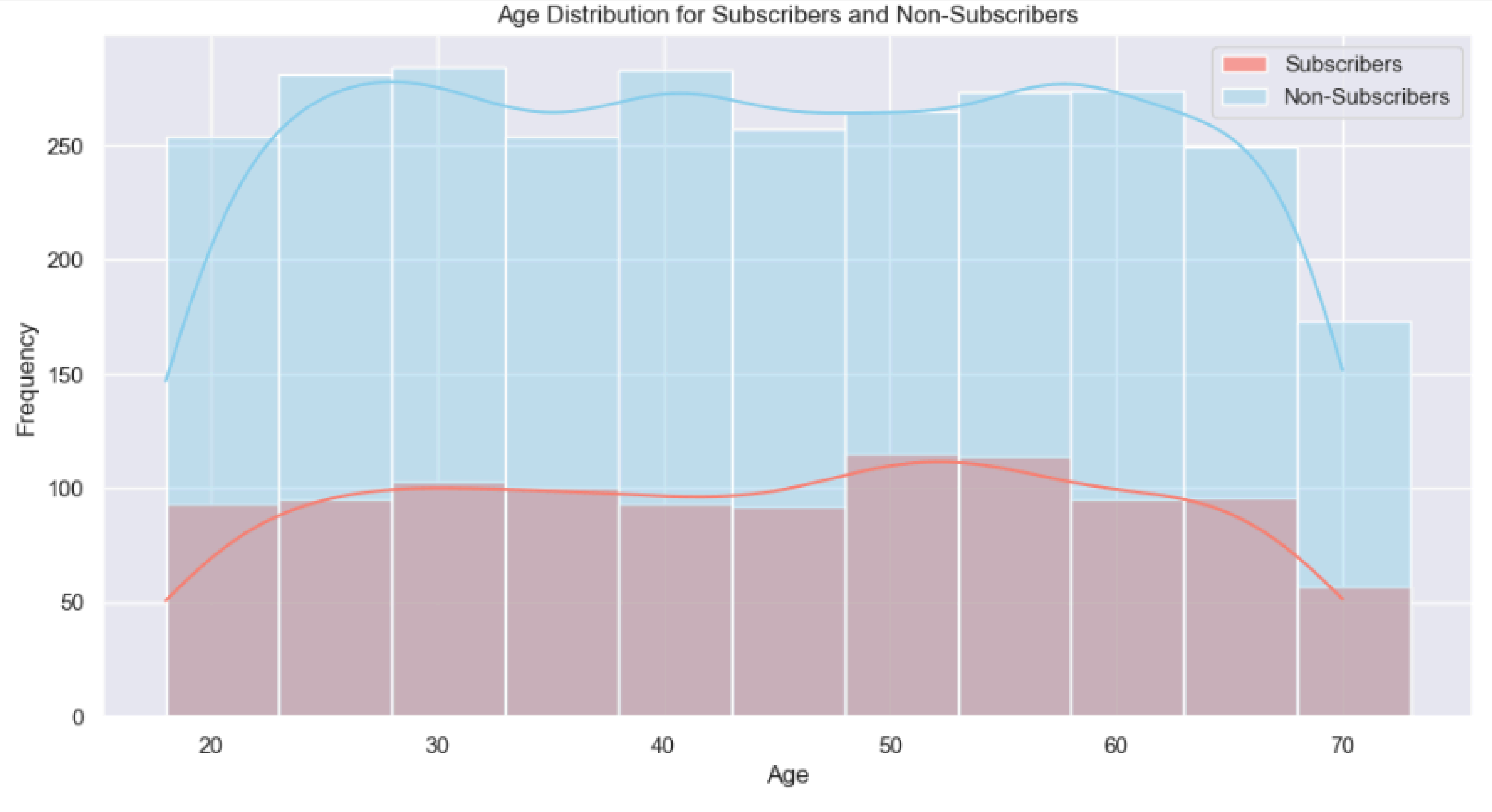
Whether the consumer is currently subscribed to a loyalty program

Subscription Status



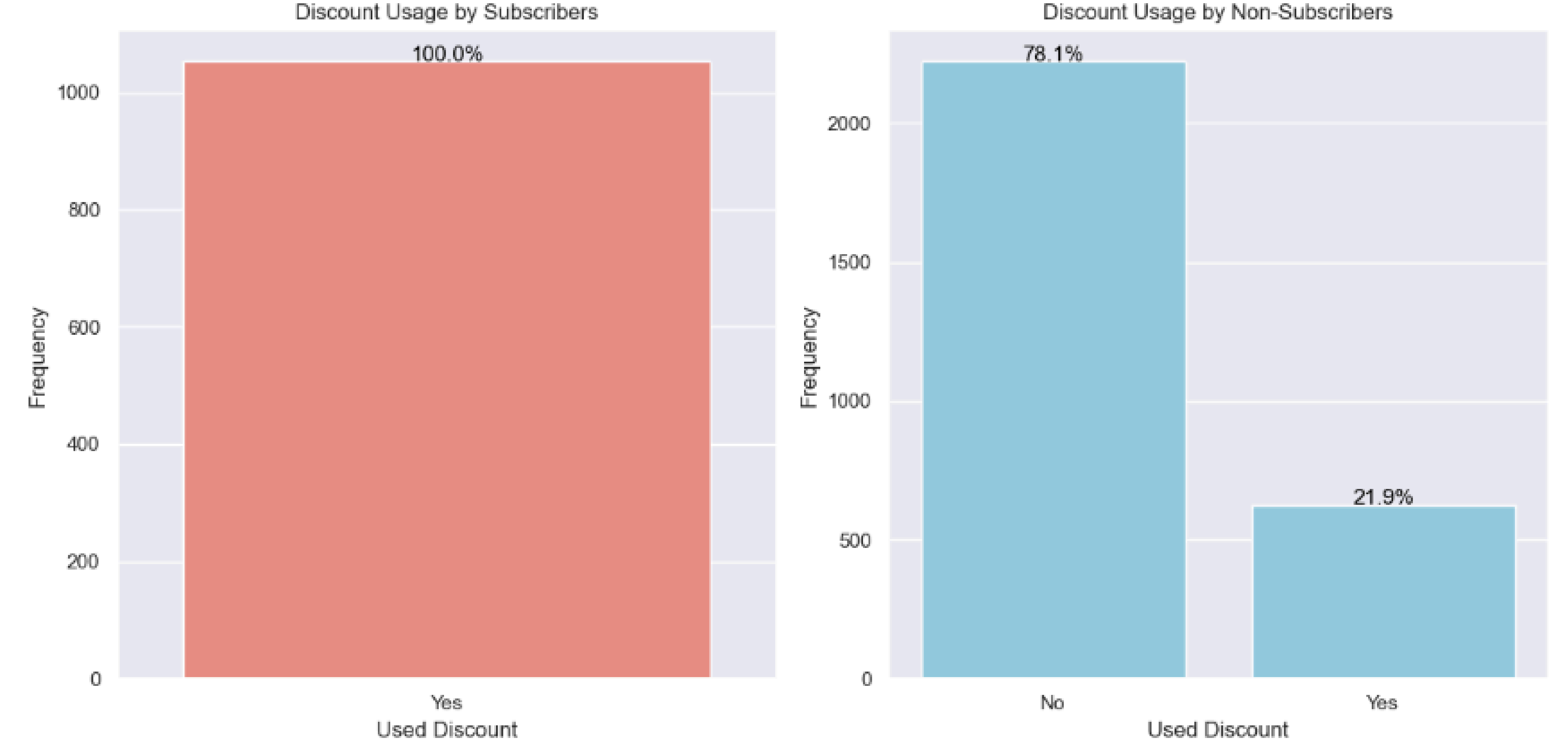
More than two-third of our customers are not subscribed to our loyalty program. Hence, this is an area to be addressed

Age



There is no significant trend between Age groups and subscriber count as it stayed leveled throughout except the start and end which dipped

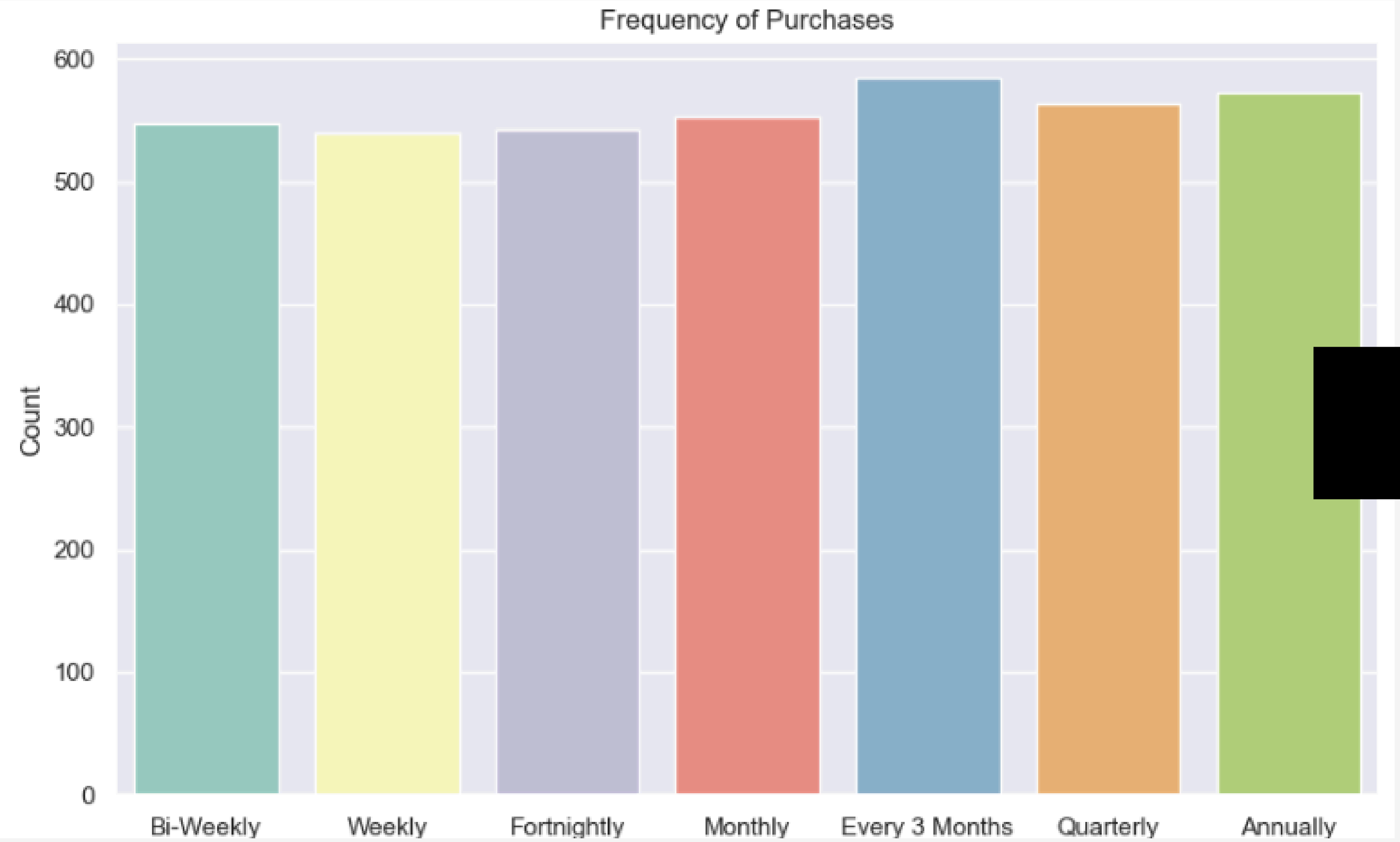
Discount Applied

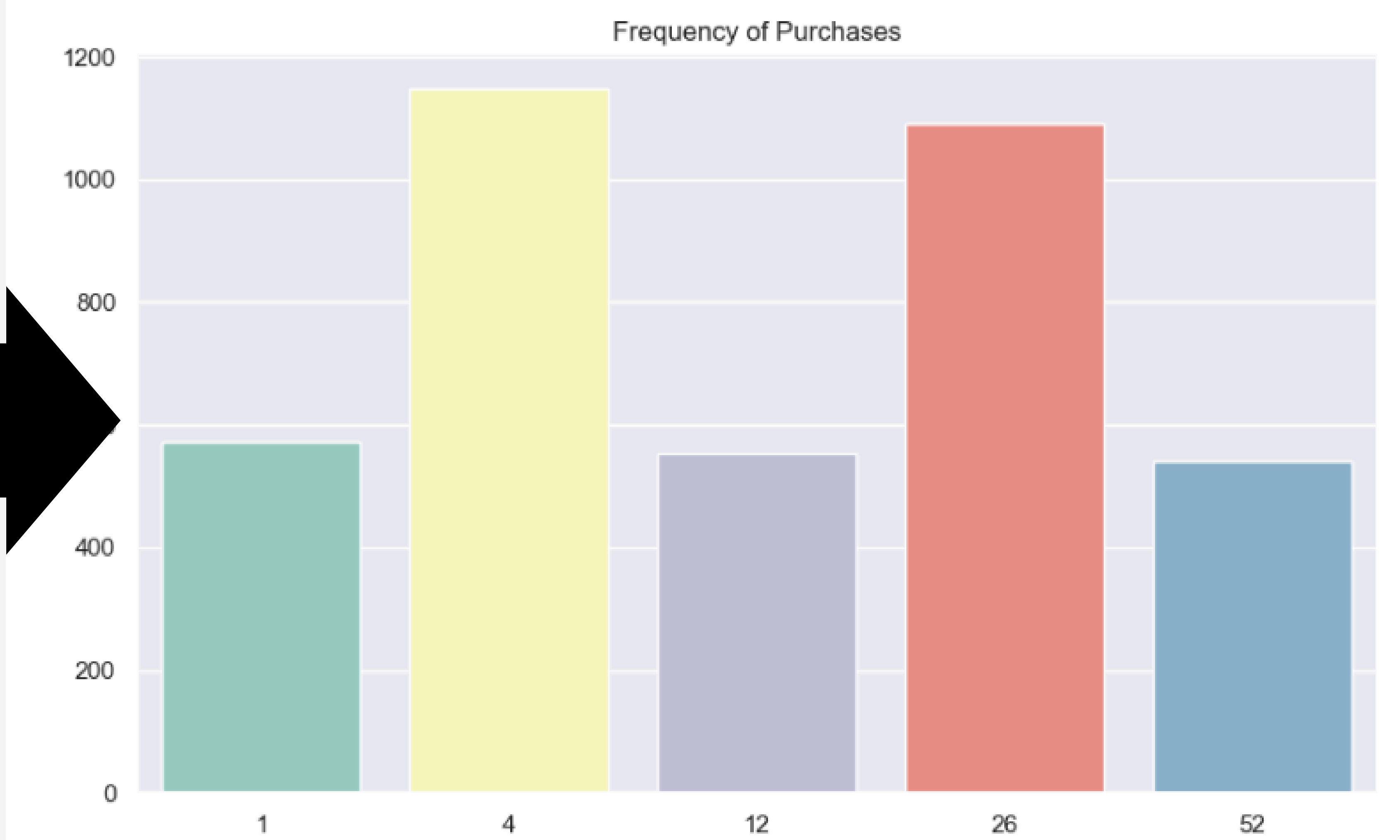


Subscribers fully utilised the discount and vouchers given compare to only 21.9% of non-subscribers used.

Discounts and vouchers may affect a consumer decision of a purchase and being in loyalty program.

Frequency of Purchase





Realised there are duplicate representation of data in frequency of purchase

Combined Bi-weekly & Fortnightly
Combined 3 months and quarterly
Converted it to numerical

Machine Learning

Naives Bayes Algorithm

- Classification technique grounded in Bayes' Theorem, with the "naive" aspect stemming from its assumption of feature independence within classes.

Random Forest + GridSearch

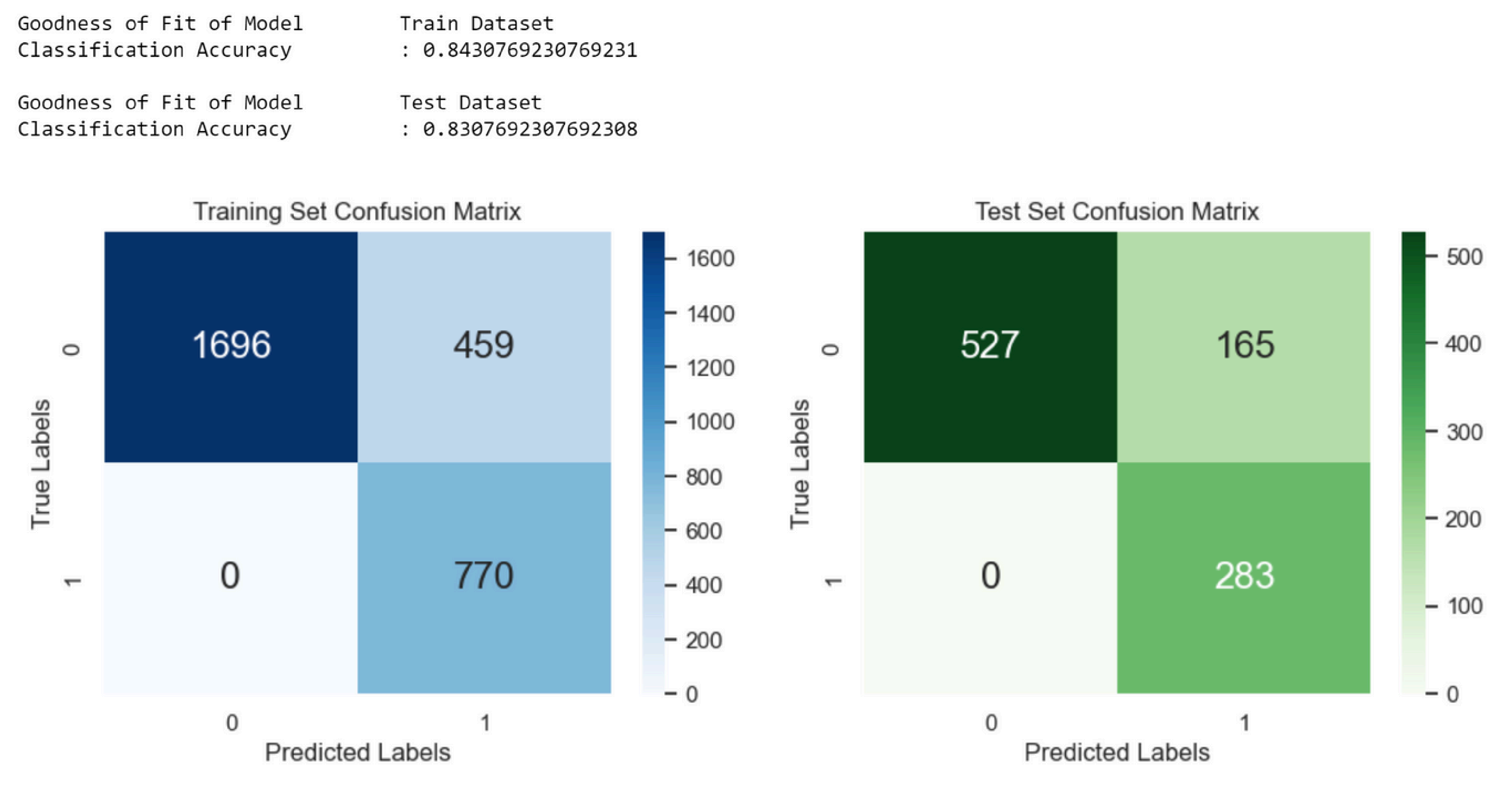
- Ensemble learning technique that constructs multiple decision trees during training, combining their predictions to enhance accuracy and reduce overfitting.
- It randomly selects subsets of data and features for each tree, making it robust for classification and regression tasks

- Determine the Best hyperparameters to train Random Forest Model (e.g. max depth, min_sample_split, etc)

Naive Bayes

(Initial Dataset)

- We used the features mentioned earlier for training, with a total of 9 Features
- Results shows that our Naive Bayes model is performing well, with the high accuracy.
- Difference between Train and Test accuracy are negligible



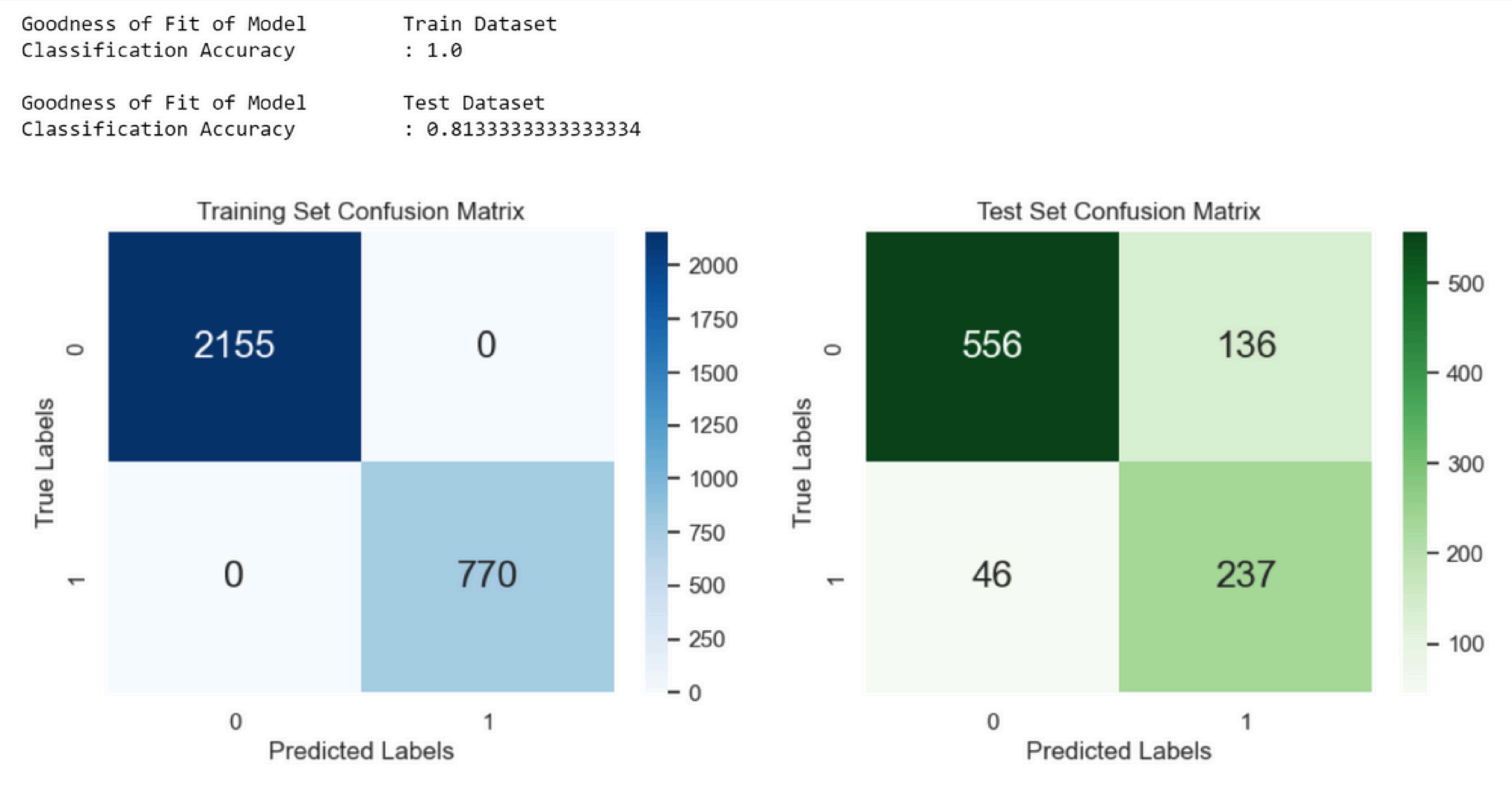
```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.8430769230769231

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.8307692307692308
```

Random Forest

(Initial Dataset)

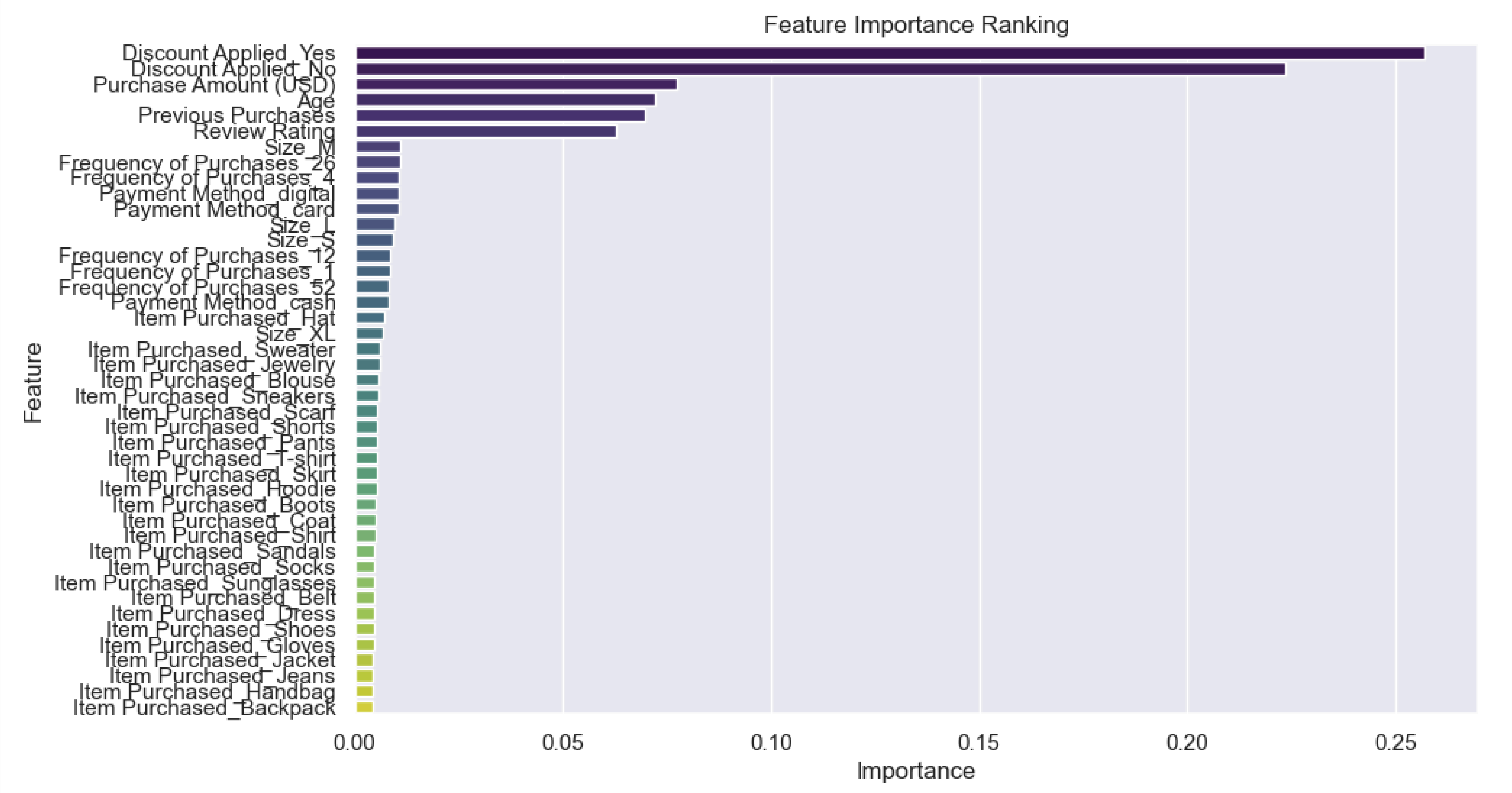
- We used the features as mentioned earlier for training, with a total of 9 Features
- Results shows that our Random Forest model is overfitted as the seen in the very high train accuracy
- This shows that our Random Forest Model is capturing a lot of noise from our dataset



```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 1.0

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.8133333333333334
```

Feature Importance

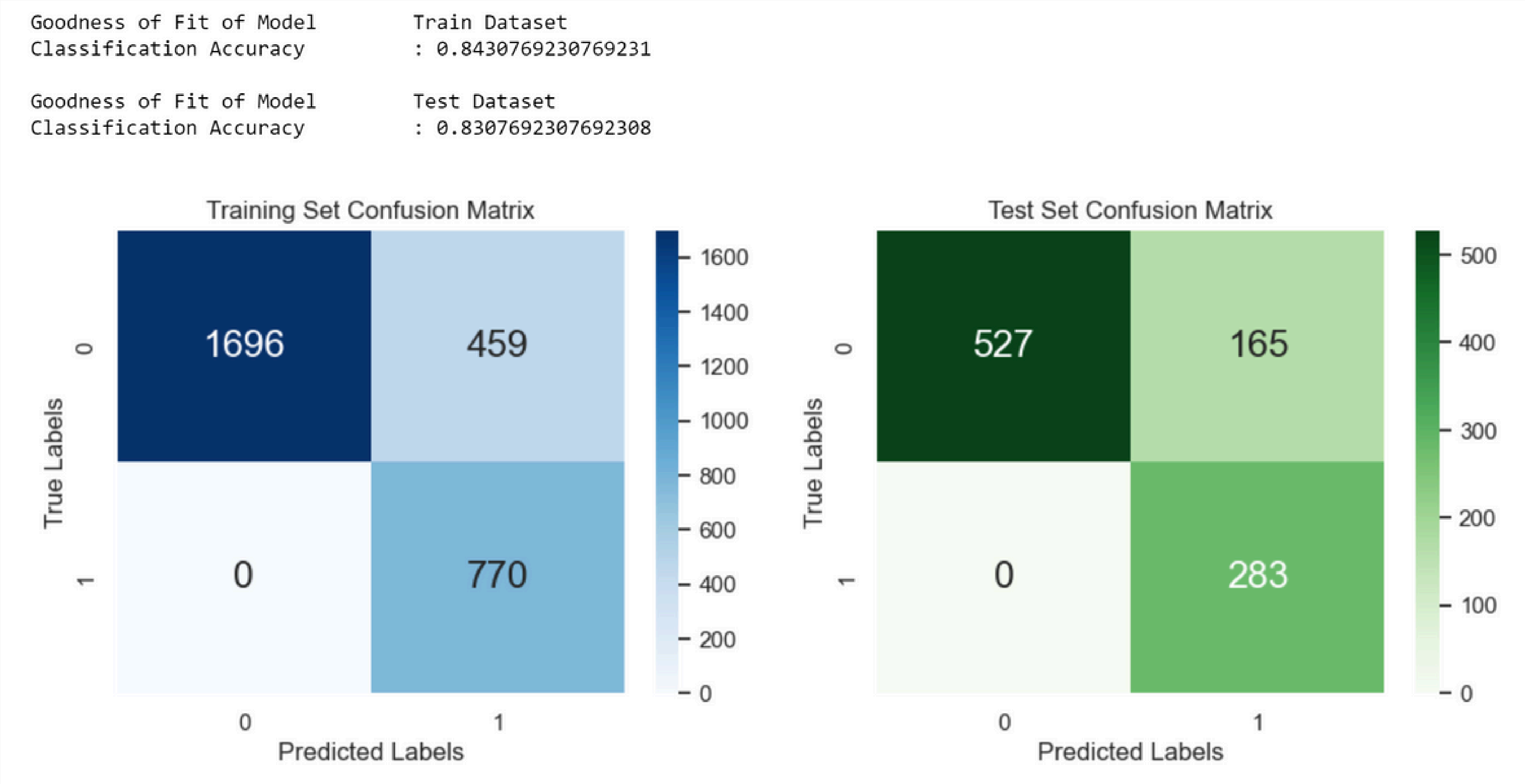


We plot the feature importance of our Random Forest and decided to reduce our dataset to only the top 2 features (Discount Applied & Purchase Amount (USD))

Naive Bayes

(Reduced Dataset)

- Performance remains the same despite reducing our dataset
- This shows that the other features are probably irrelevant to our model and the feature independence within classes



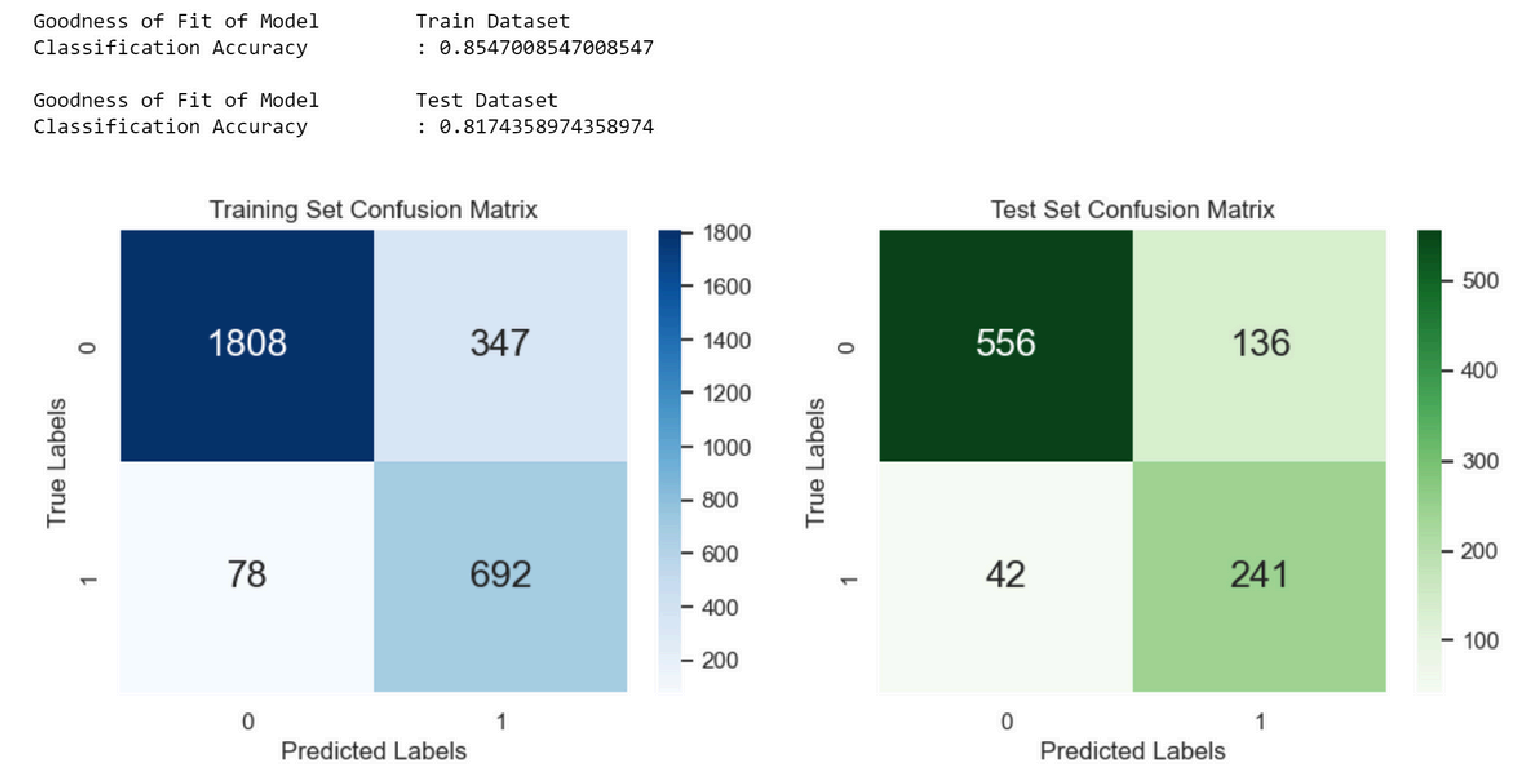
```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.8430769230769231

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.8307692307692308
```

Random Forest

(Reduced Dataset)

- Our Random Forest model has significantly improved
- Even though both accuracy has dropped, both accuracy are now closer to each other, showing that our model is not as overfitted as before
- This shows that our previous Random Forest model was capturing a lot of noise in our original dataset



```
Goodness of Fit of Model        Train Dataset
Classification Accuracy         : 0.8547008547008547

Goodness of Fit of Model        Test Dataset
Classification Accuracy         : 0.8174358974358974
```

Training Set Confusion Matrix

True Labels \ Predicted Labels	0	1
0	1808	347
1	78	692

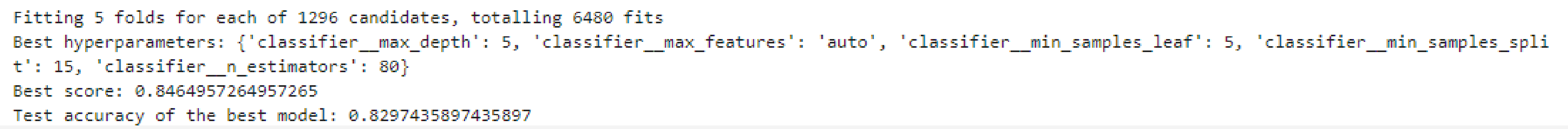
Test Set Confusion Matrix

True Labels \ Predicted Labels	0	1
0	556	136
1	42	241

Random Forest + GridSearch

(Reduced Dataset)

- We used GridSearch to get the best hyperparameters for our Random Forest model
- Test accuracy improved slightly while train accuracy dropped slightly
- This shows that by getting the best hyperparameters, our model improved in terms of reducing overfitting



```
Fitting 5 folds for each of 1296 candidates, totalling 6480 fits
Best hyperparameters: {'classifier__max_depth': 5, 'classifier__max_features': 'auto', 'classifier__min_samples_leaf': 5, 'classifier__min_samples_spli
t': 15, 'classifier__n_estimators': 80}
Best score: 0.8464957264957265
Test accuracy of the best model: 0.8297435897435897
```

Model Performance Summary

	Train Accuracy	Test Accuracy
Naive Bayes (Initial Datasset)	84.3%	83.0%
Random Forest (Initial Dataset)	100%	81.3%
Naive Bayes (Reduced Dataset)	84.3%	83.0%
Random Forest (Reduced Dataset)	85.4%	81.7%
Random Forest + GridSearch	84.6%	82.9%

What have we learnt?

Machine Learning Function

- Naive Bayes
- Random Forest
- GridSearch

Conclusion

- Discounts / Vouchers is a very important factor to get more customers to subscribe to us
- Customers that subscribed to us tend to spend more in our store compared to non-subscribers