


Motivation

The primary motivation for this project is to attempt to detect passive customers of Starbucks depending on their behavior related to offers provided. Generalized machine learning techniques were used to predict potential passive customers and to classify them into clusters for future offering.

Data Preparation and Labeling

 Data source - Kaggle
Starbucks App Customer Reward Program dataset

Preparation involved data cleansing and joining the offers portfolio with the customer information using the transcript data.

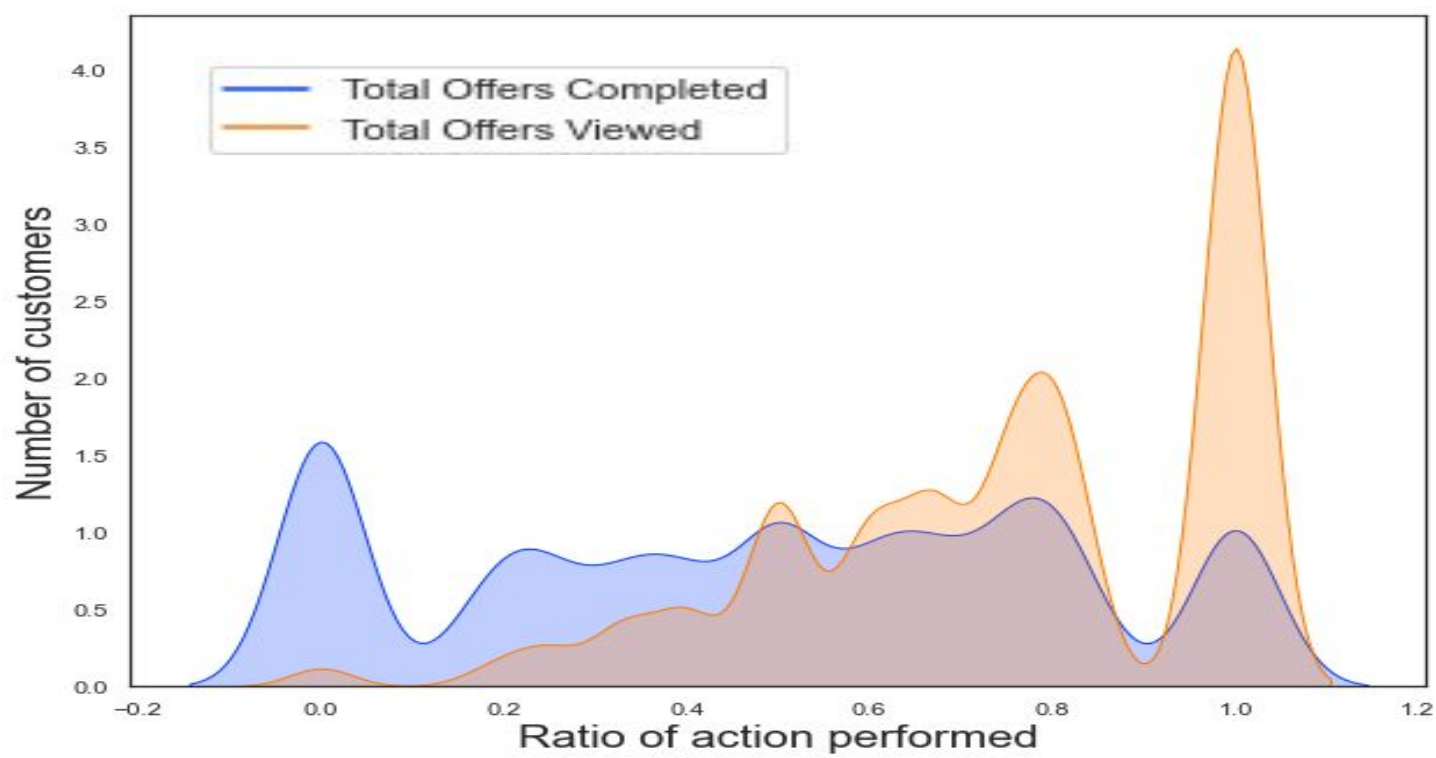
Customers were labeled as Active or Passive based on two criteria:

1. *Total offers viewed ratio* = Total offers viewed / Total offers received
2. *Total offers completed ratio* = Total offers completed / Total offers received

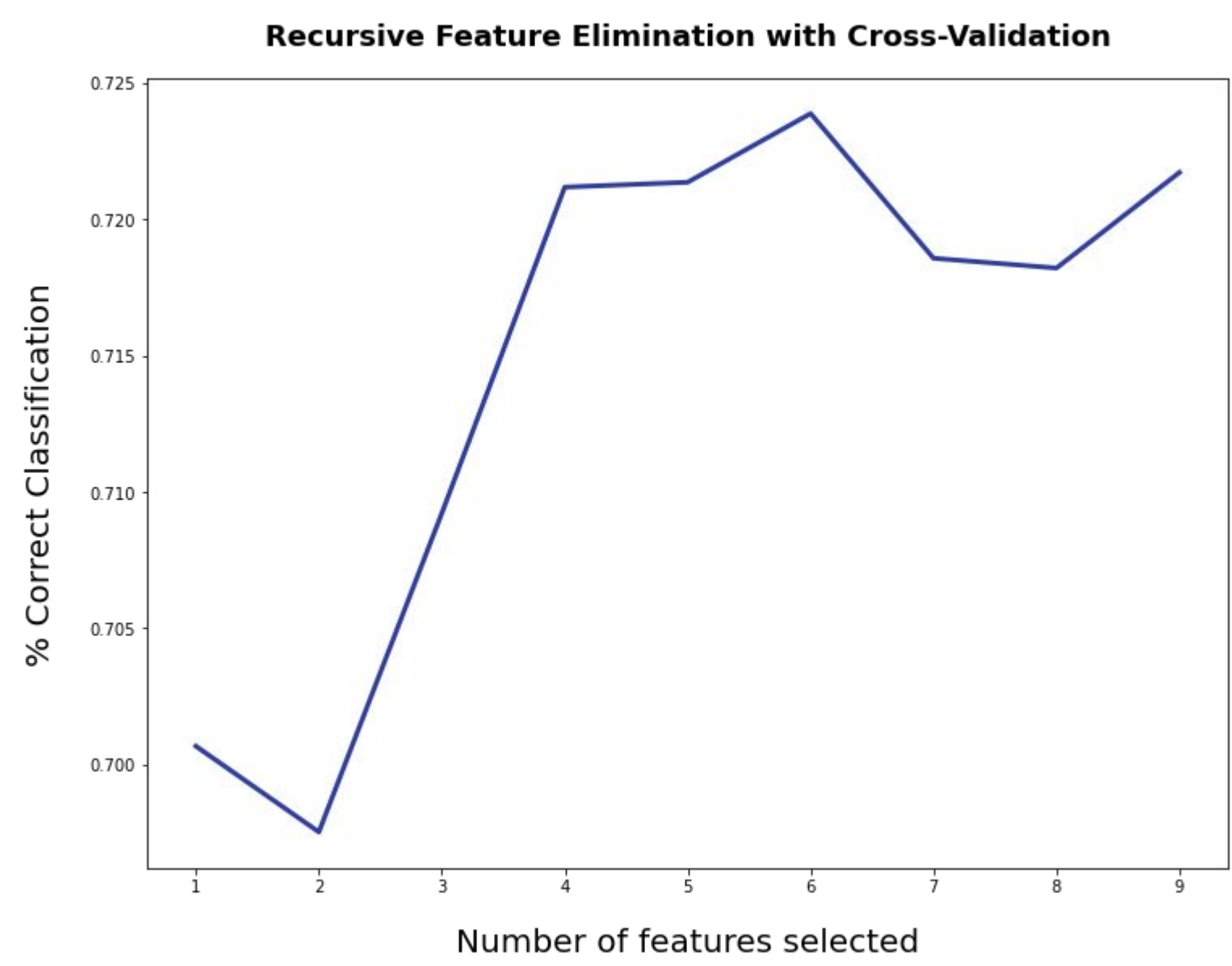
```
if (view_ratio >= 0.6) and (completion_ratio >= 0.2):  
    customer_type = "Active"  
else:  
    customer_type = "Passive"
```

Using this condition, the customer base was split into two types with the following counts:

Active customers: 9637
Passive customers: 5188



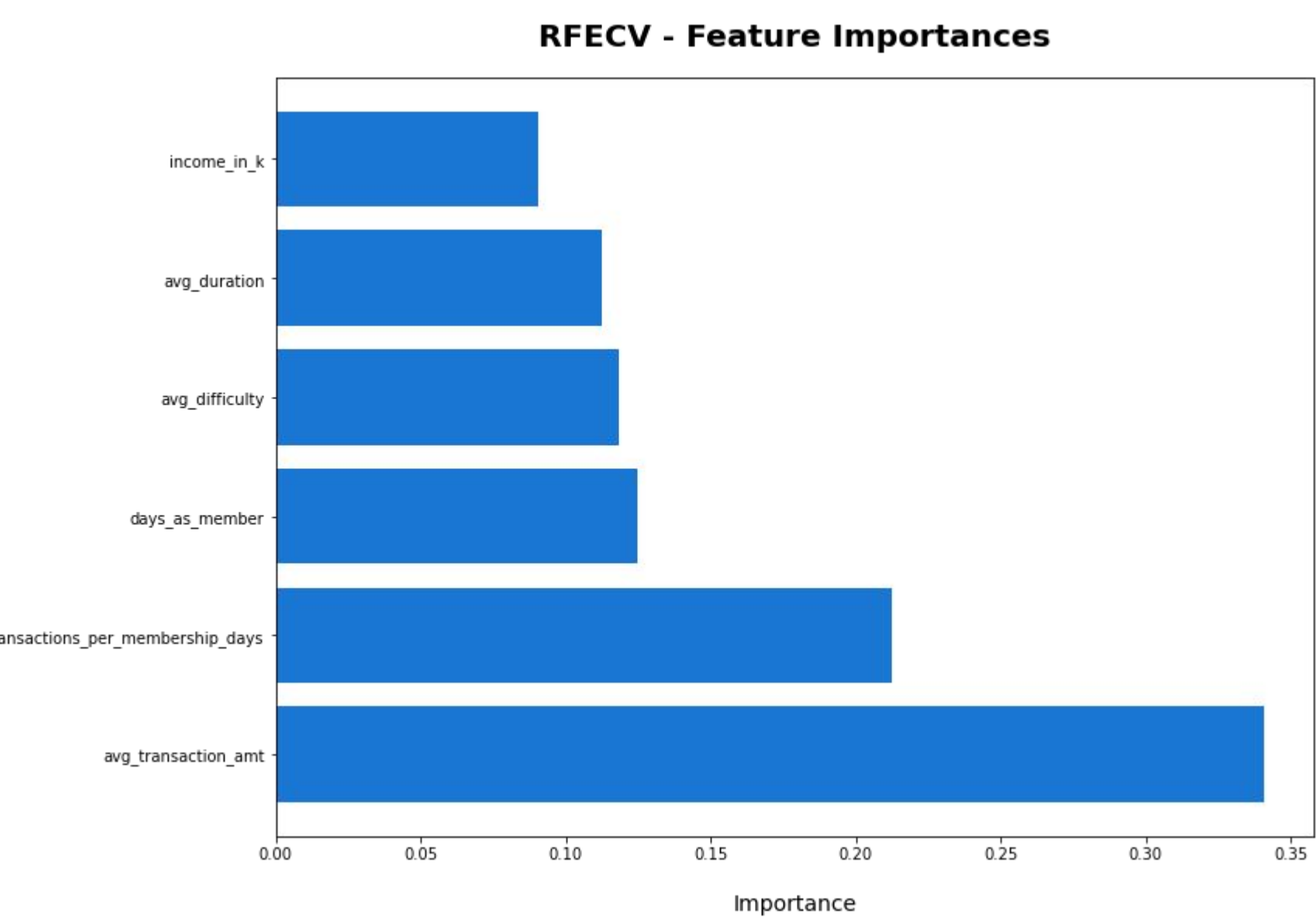
Feature Selection



Upon using a combination general reasoning and correlation matrix method to reduce the potential set of input features, Recursive Feature Elimination with Cross-Validation was then performed to compute the optimum number of features.

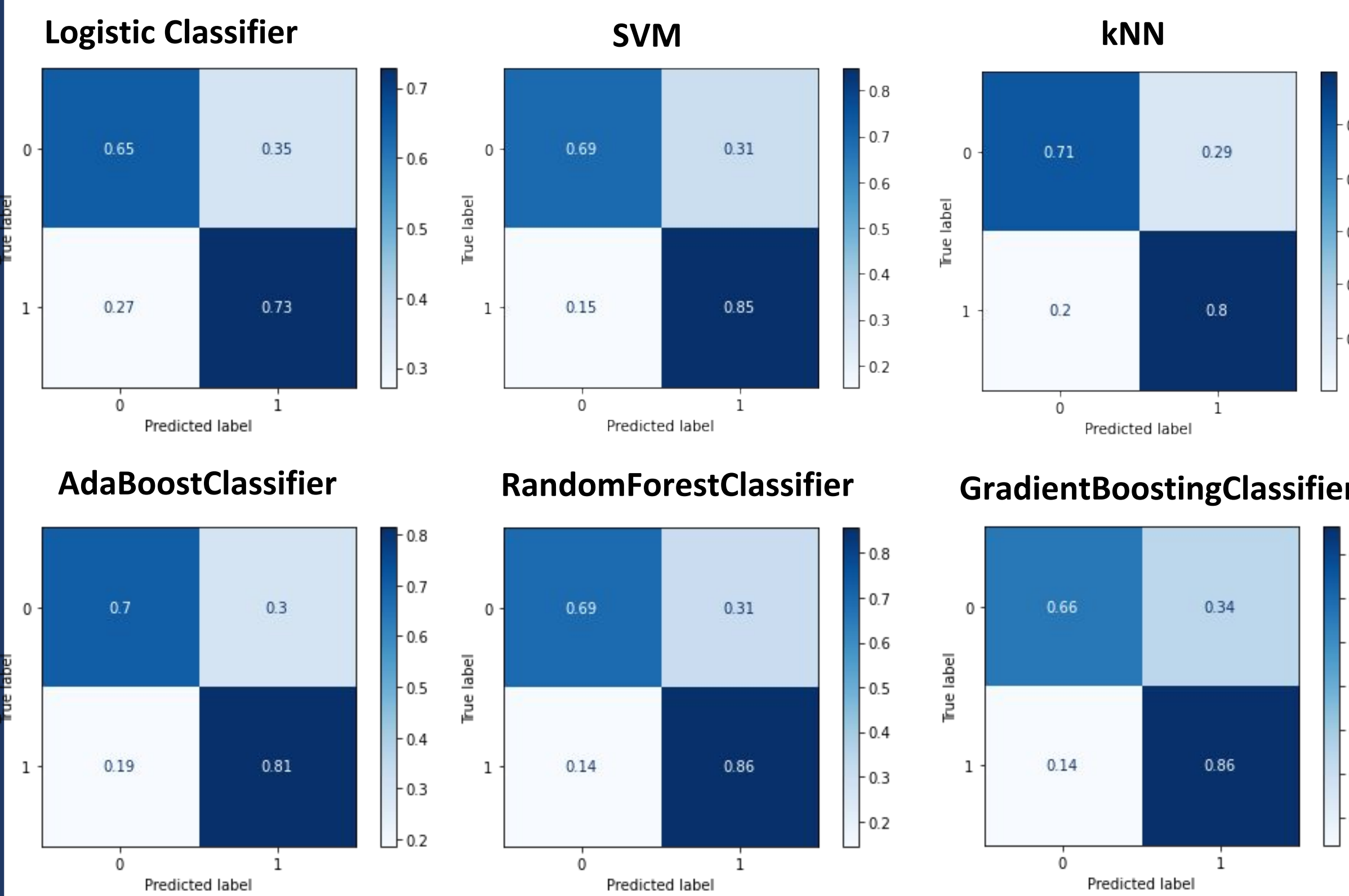
The six features that were selected to be used in all the models are listed below in the increasing order of their feature importance values:

1. *Income*
2. *Avg_duration*
3. *Avg_difficulty*
4. *Days_as_members*
5. *Transactions_per_membership_day*
6. *Avg_transaction_amt*

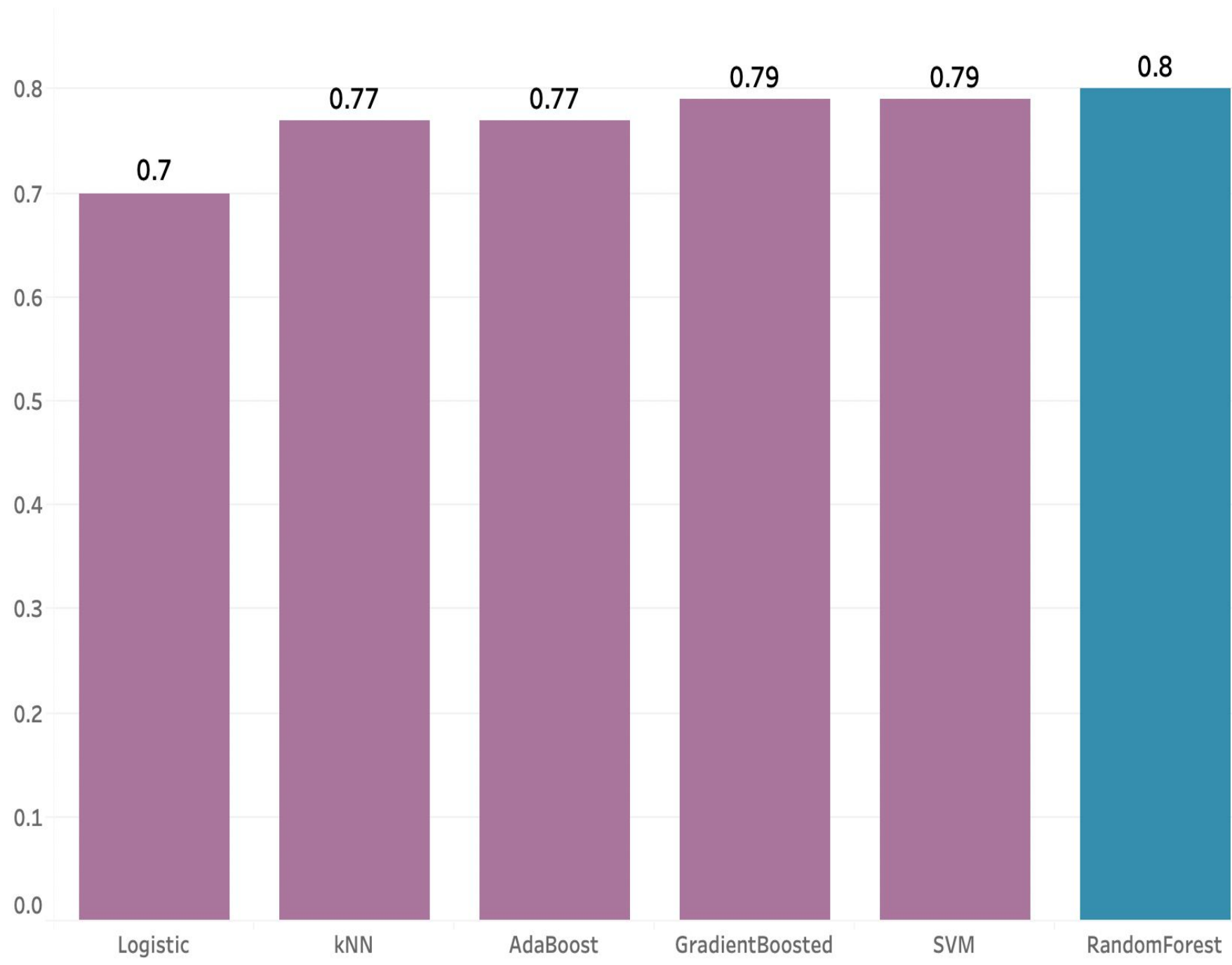


Supervised Learning Models and Prediction

Confusion Matrices for 6 models that were used:



F1 Scores of 6 models used



Conclusions:

From the confusion matrices and F1 scores, we can deduce the following:

- kNN model is the best at classifying active customers.
- Random Forest Classifier produce the best overall prediction outcome.
- Logistic Regression Classifier did not perform as good as other models.

Hyperparameters used for tuning the best model Random Forest Classifier:

*max_depth=25, n_estimators=1200,
min_samples_leaf=1,
min_samples_split=2*

SMOTE Application to address imbalanced target class:

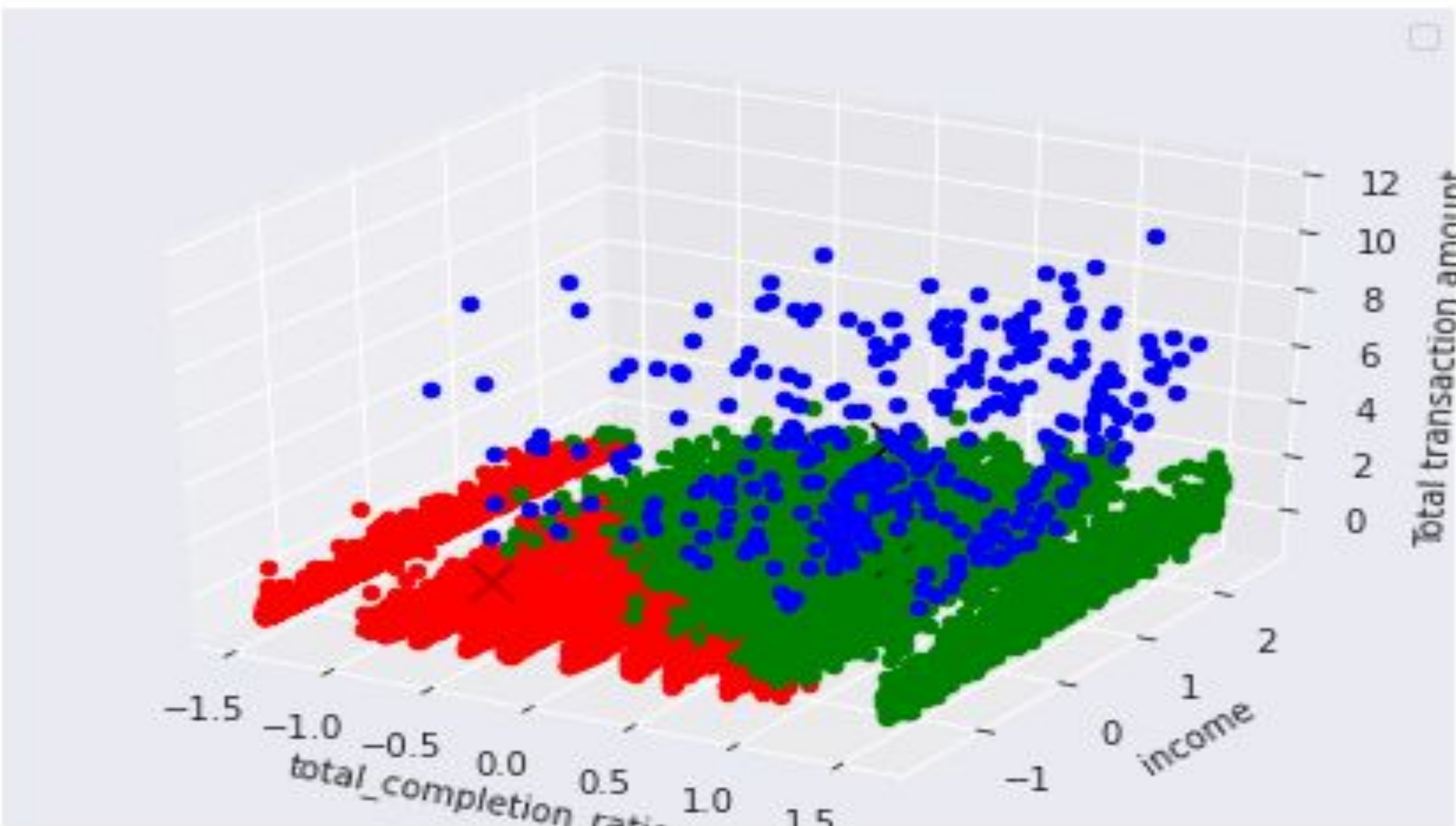
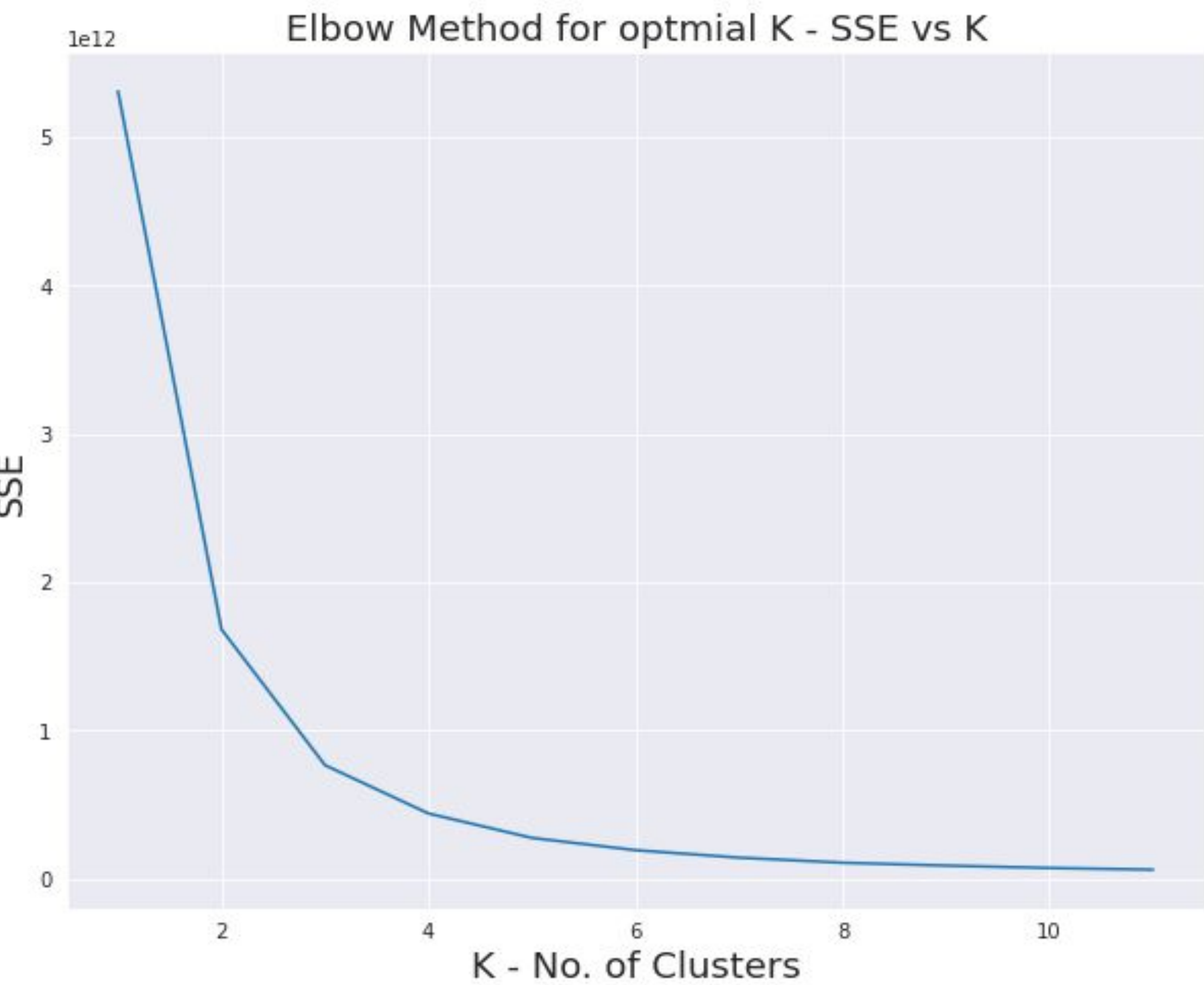
Based on our labeling, the data is imbalanced with about 65% active customers and 35% passive customers. Synthetic Minority Oversampling Technique (SMOTE) is applied to balance the data before applying the Models.

Customer Clustering Based on Offer Type (Future Work)

Using Unsupervised Learning Models to classify Starbucks Customers

K- Means Clustering

For k values ranging from 1 to 12, the Elbow point and Silhouette score gives optimal value at k=5, grouping the Starbucks customers into 5 clusters.



From this clustering graph which uses 3 features and 3 clusters, we can provide a few insights and suggestions:

1. The red group of customers are not sold on our products and offers, most of them belong to low-income group and do not spend much at Starbucks, and are also disconnected from the rewards program. This group will be most difficult to convert into loyal customers.
2. The blue group of customers are willing to spend money at Starbucks as well as completing the offers they receive. We can see that they also earn a decent income. This group of customers are probably loyal to the brand.
3. The green group of customers present opportunities. A lot of them earn good income and are completing their offers, but aren't spending at Starbucks otherwise. We can say that they just come for the offers only. To entice them to spend, we may need to rethink our pricing or customer service strategy.