Machine Learning Engineer Nanodegree
Project Report

Udacity Starbucks Capstone Challenge

Sharvari Dhote

April 20, 2020

Contents

1 Introduction

1.1 Background

The businesses want to target right customers throughout the customer journey which helps to reduce the marketing budget, and increase customer satisfaction with profits. Machine learning (ML) and artificial intelligence (AI) technologies are helping to automate the digital marketing with real time decision-making such as selecting right and optimal messaging platform, time and offer to customers by integrating data from different platforms [1, 3, 2, 4] as depicted in figure 1.

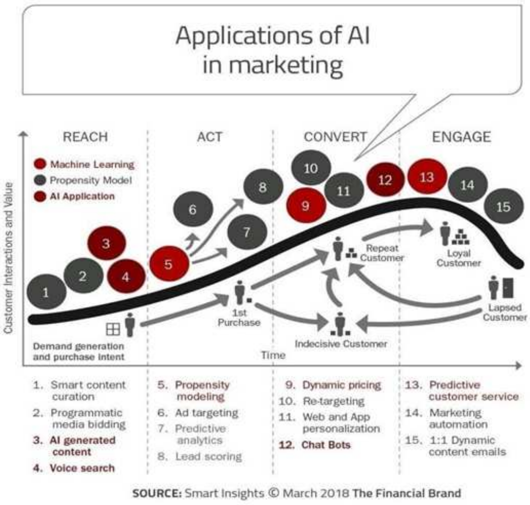

Figure 1: Application of AI in digital marketing [1].

Udacity's Starbucks Capstone challenge project contains simulated data that mimics customer behaviors on the Starbucks rewards mobile app. ML techniques will be used to develop an effective marketing strategy for the Starbucks business using the available data.

1.2 Problem Statement

Starbucks sends out an offer to users of the mobile app every few days. An offer can be merely be an advertisement for a drink or an actual offer such as a discount or BOGO (buy one get one free). A given data set contains demographic, transaction and offer related information. This project is divided into two parts:

- **Customer Segmentation Analysis:** It is important to know your customers in order to increase business. First, customer segmentation using unsupervised ML technique, K-means clustering is carried out to understand customer characteristics and demographics [5, 6].
- **Effective Offer Prediction Model:** Second the best performing supervised ML model is developed by training and comparing Logistic classifier and random forest using performance based on confusion matrix, receiver operating curve (ROC) to predict if customer will respond to an offer or not for better targeted marketing [7].

1.3 Project Design Steps

The project is carried out in the following steps:

- *Data Cleaning and Exploratory Data Analysis (EDA)*- First step is to understand the data. Started with the data statistic, count missing values, and distribution of variables to get insights into the data. Basic cleaning steps such as filling missing values, checking outliers, dropping unimportant and duplicate columns and encoding categorical variables.

- *Feature Engineering* - Created new features from the raw features and prepared dataframe for customer segmentation and effective offer prediction model.

- *Customer Segmentation* - to understand Starbucks customer's characteristics.

- *Offer Prediction Model* - Two different models are implemented and the best model is selected based on the aurea under the curve (AUC) and other metrics.

2 Data Cleaning and Exploratory Data Analysis

This section presents basic data set statistics, data cleaning, exploration and visualization to understand characteristics of the Starbuck's dataset.

2.1 Data Cleaning

2.1.1 Portfolio Dataset

It contains information about offer type and duration sent on the Starbucks app during promotional period. Summary of the portfolio dataframe

- Encoded offer type column - three types of offers sent buy one get one free (BOGO), Discount (discount with purchase) and Informational (information about products).
- No direct reward for the information type offer, so we will need a strategy to check its offer effectiveness.
- Channel feature does not look much helpful. We can see that all the offers have been sent through mobile and email with an exception. We will not use channel feature in our model or segmentation. However, if there was additional data showing through which channel offer was completed. It may help to target the channel for sending the effective offers.
- Offer id corresponds to 4 types of bogo, 4 types of discount and 2 types of informational offers sent.
- Offer duration - defines the validity of the offer during that period. We can check the offer duration if it's too long or too short to decide its effectiveness. It will be helpful to separate offer and transaction dataframes for further analysis.
- Difficulty - it's minimum amount spent to be eligible for the offer. We need to see average amount spend and compare difficulty to see what's effective difficulty level.

	difficulty	duration	offer_id	offer_type	reward_offered	email	mobile	social	web	bogo	discount
0	10	7	ae264e3637204a6fb9bb56bc8210ddfd	bogo	10	1	1	1	0	1	0
1	10	5	4d5c57ea9a6940dd891ad53e9dbe8da0	bogo	10	1	1	1	1	1	0
2	0	4	3f207df678b143eea3cee63160fa8bed	informational	0	1	1	0	1	0	0
3	5	7	9b98b8c7a33c4b65b9aebfe6a799e6d9	bogo	5	1	1	0	1	1	0
4	20	10	0b1e1539f2cc45b7b9fa7c272da2e1d7	discount	5	1	0	0	1	0	1
5	7	7	2298d6c36e964ae4a3e7e9706d1fb8c2	discount	3	1	1	1	1	0	1
6	10	10	fafdcd668e3743c1bb461111dcafc2a4	discount	2	1	1	1	1	0	1
7	0	3	5a8bc65990b245e5a138643cd4eb9837	informational	0	1	1	1	0	0	0
8	5	5	f19421c1d4aa40978ebb69ca19b0e20d	bogo	5	1	1	1	1	1	0
9	10	7	2906b810c7d4411798c6938adc9daaa5	discount	2	1	1	0	1	0	1

Figure 2: Portfolio Dataframe

2.1.2 Profile Dataset

- Important dataset to understand the customer demographic.
- Encode age, gender categorical features later.
- become member - changed to membership years.
- age maximum value : 118 - looks like an outlier.
- income and gender columns have missing values - 2175 corresponds to age outlier. We will check overall missing values in the combined dataset to decided if we need to impute or delete missing values.

	age	gender	customer_id	income	membership_years	year
0	118	None	68be06ca386d4c31939f3a4f0e3dd783	NaN	3.0	2017
1	55	F	0610b486422d4921ae7d2bf64640c50b	112000.0	3.0	2017
2	118	None	38fe809add3b4fcf9315a9694bb96ff5	NaN	2.0	2018
3	75	F	78afa995795e4d85b5d9ceeca43f5fef	100000.0	3.0	2017
4	118	None	a03223e636434f42ac4c3df47e8bac43	NaN	3.0	2017

Figure 3: Profile Dataframe

2.1.3 Transcript Dataset

Transcript data set with no missing values - records for transactions, offer types : received, viewed and completed.

- Value column separated in to offer id, reward and amount columns.
- Event column separated to offer received, offer completed, offer viewed and transaction time will be converted to days.

	event	customer_id	time	amount	offer_id	reward_given	offer completed	offer received	vi
0	offer received	78afa995795e4d85b5d9ceeca43f5fef	0	NaN	9b98b8c7a33c4b65b9aebfe6a799e6d9	NaN	0	1	
1	offer received	a03223e636434f42ac4c3df47e8bac43	0	NaN	0b1e1539f2cc45b7b9fa7c272da2e1d7	NaN	0	1	
2	offer received	e2127556f4f64592b11af22de27a7932	0	NaN	2906b810c7d4411798c6938adc9daaa5	NaN	0	1	

Figure 4: Transcript Dataframe

2.1.4 Missing Data Analysis

If there is less than 5% of missing data and features are not important then we can delete the data[12]. Almost 12.79% data is missing in gender and income columns. To further understand the missing data, we had a closer look at the transcript and profiled combined dataframe. We can see that with 10.79% missing customer transaction data and 3.38 % of missing gender data. It is an important feature. It can overall affect the segmentation analysis and offer predictions modeling. Thus, we will impute income with mean values and gender with unknown *U*. Age has an outlier as seen from the histogram. We will impute above 100 age value with the mean value.

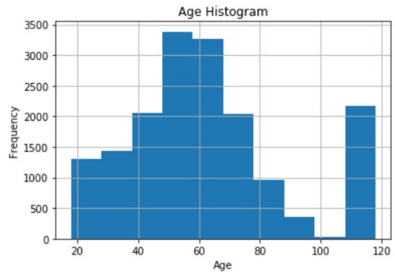

Figure 5: Age histogram

2.2 Exploratory Data Analysis

2.2.1 Offer Type and Event Analysis

- Bogo offer: More bogo offers were present as compared to discount offers. Bogo offer view rate is above 80 % and completion rate is around 50 %. Almost same number of bogo and discount offers were sent to the customers.
- Discount offer: More discount offers were completed. Discount offer view rate is around 70% and offer completion rate is 58%.
- Informational offer: Offer view rate is around 70%. We need to count informational offer related counts.
- Offer bogo and discount completion rate is 45% and offer view rate is 76%. Offer completed transaction count 24.17 % from total transactions and many customers received more than 4 offers.
- We can see that 12774 customers completed the bogo and discount offers. No offer was sent to 6 customers.

2.2.2 Gender, Starbucks Membership Years and Joining Year

- More members have completed 2 -3 years of membership period. Maximum number of new customers joined in 2017. Starbucks has 60% male member out of 17000 (2175 missing customer data).

2.2.3 Age, Gender and Income

.
- Starbuck has more male customer than female
- More younger male customers as compared to female.
- Male customers have high income than female
- More members are older with the age above 50.

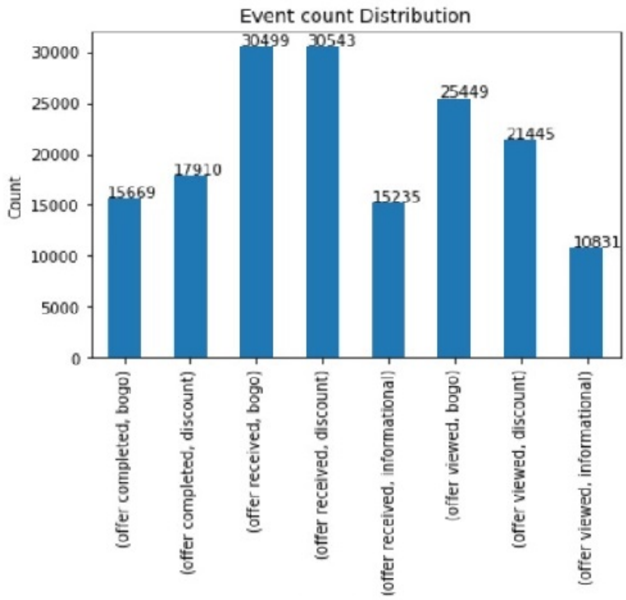

Figure 6: Event count distribution

Figure 7: Gender distribution of (a)membership years and (b)joining year

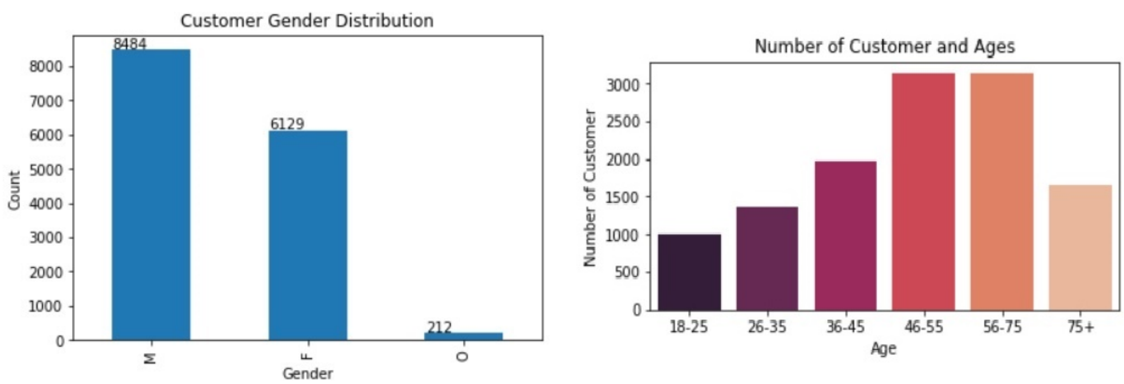

Figure 8: Starbucks customer (a) gender and (b) age distribution

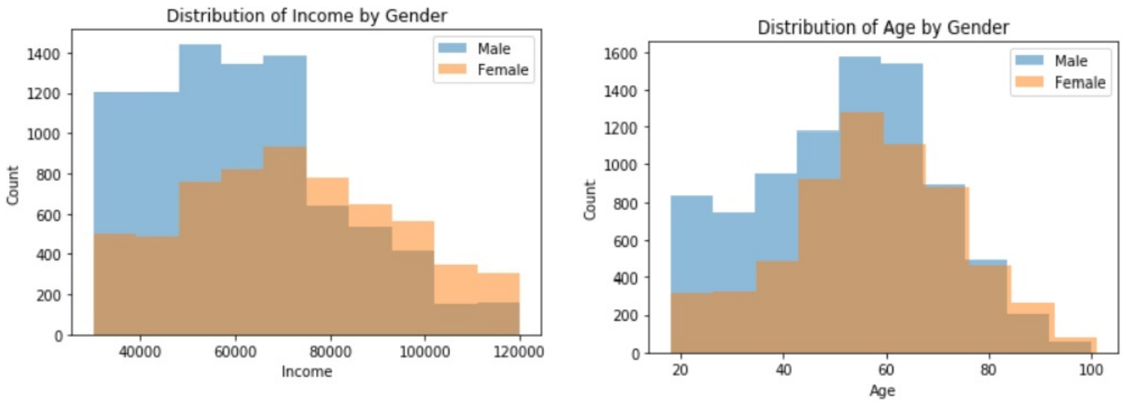

Figure 9: Starbucks customer gender (a) income and (b) age distribution

3 Feature Engineering and Offer DataFrame

3.1 Offer DataFrame

We are going to prepare a labelled data frame for the supervised machine learning model. Thus, we need to define successful and failed offers and create raw and the engineered features. Following challenges were mentioned in the Starbucks dataset.

- Not all users receive the same offer.
- Different validity period for the offer type and informational offer to influence customer.
- Customer might make a purchase through the app without having received an offer or viewing an offer.
- A user can receive an offer, never actually view the offer, and still complete the offer

In order to define effectiveness of the offer fail/success of an offer. Let's have a closer look at a single customer transaction history from the merged transcript, profile and portfolio dataframe. We can see following characteristic. Let's define our binary classifier. offer fail =0 (customer was not influenced by the offer) and offer success=1 (customer did purchase due to an offer).

Based on the customer purchase history, we define an effective offer. Which offer sent were effective?. To answer this question, we filter the effective offer using *pandas.core.groupby.DataFrameGroupBy.shift* method. For each customer id and offer id and offer type i.e. discount or bogo and bogo with discount, check if the event in the sequence *offer received* → *offer viewed* → *offer completed* within an offer time duration, then assign offer success as 1. We filter the bogo, discount and informational offer in the similar way.

- bogo and discount offer success =1: offer received → offer viewed → offer completed → transaction (within offer validity duration).
- informational offer success = 1: offer received → offer viewed → transaction (within offer validity duration).
- bogo and discount offer fail =0: offer received → offer viewed → transaction (not within offer validity duration).

	customer_id	offer_id	time	event	amount	reward_given	difficulty	duration	offer_t
59352	0020ccbbb6d84e358d3414a3ff76cffd	2298d6c36e964ae4a3e7e9706d1fb8c2	7	offer received	NaN	NaN	7.0	7.0	di
67584	0020ccbbb6d84e358d3414a3ff76cffd	2298d6c36e964ae4a3e7e9706d1fb8c2	7	offer viewed	NaN	NaN	7.0	7.0	di
88010	0020ccbbb6d84e358d3414a3ff76cffd	2298d6c36e964ae4a3e7e9706d1fb8c2	9	offer completed	NaN	3.0	7.0	7.0	di
156808	0020ccbbb6d84e358d3414a3ff76cffd	5a8bc65990b245e5a138643cd4eb9837	17	offer received	NaN	NaN	0.0	3.0	inform
165442	0020ccbbb6d84e358d3414a3ff76cffd	5a8bc65990b245e5a138643cd4eb9837	17	offer viewed	NaN	NaN	0.0	3.0	inform

Figure 10: Merged Portfolio, Profile and Transcript dataframe

	customer_id	offer_id	time	event	amount	reward_given	difficulty	duration	offer_type	reward_offered
16379	eb540099db834cf59001f83a4561aef3	NaN	0	transaction	4.74	NaN	NaN	NaN	NaN	NaN
108268	eb540099db834cf59001f83a4561aef3	NaN	13	transaction	5.09	NaN	NaN	NaN	NaN	NaN
228700	eb540099db834cf59001f83a4561aef3	NaN	22	transaction	7.40	NaN	NaN	NaN	NaN	NaN

Figure 11: Single customer transaction data

- bogo and discount offer fail =0: offer received → transaction → offer viewed (within offer validity duration).
- bogo and discount offer fail =0: offer received → offer viewed and no transaction.
- We saw above that only transaction without offer sent to customer when offer id is *NaN* (We will filter informational offer successful transaction later).

3.2 Feature Engineering

Feature engineering is important aspect to obtain customer demographic DataFrame is grouped by customer. There are 17000 customers. Following are the 40 raw and engineered features for each customer. Pandas *groupby* feature is used to obtain the features.

- Types of offers received, viewed and completed : bogo, discount and informational (9 Features)
- Total offers received, viewed and completed : all types of offers (3 features)
- Total transaction count, average transaction amount total transaction amount - number of transactions, average amount
- spent for each transaction and total amount spent (3 features)
- Total reward given to customer for completing offer and average reward given (2 features)
- Offer difficulty - average offer difficulty amount need to spent after completing offer. This feature can show how level of offer difficulty impacts the offer completion.
- Offer completion rate [13] - Number of completed offers / Number of offer's sent (all offers sum) (1 feature)
- Offer view rate - Number of offer viewed/ Number of offers received. (1 feature)
- Similarly, we obtained bogo, discount and informational offer completion rate (3 features)

- Demographic raw features - age, income, gender, membership years, year (5 features)
- We will encode gender features. (3 Features)
- Total offer success -number successful offers completed.
- Output binary class feature offer success- if customer has completed offer successfully 1 and failed offer 0.

	customer_id	offer_received_bogo	offer_received_discount	offer_received_informational	total_offer_received	offer_viewec
0	0009655768c64bdeb2e877511632db8f	1.0	2.0	2.0	5.0	
1	00116118485d4dfda04fdbaba9a87b5c	2.0	0.0	0.0	2.0	
2	0011e0d4e6b944f998e987f904e8c1e5	1.0	2.0	2.0	5.0	
3	0020c2b971eb4e9188eac86d93036a77	2.0	2.0	1.0	5.0	
4	0020ccbbb6d84e358d3414a3ff76cffd	2.0	1.0	1.0	4.0	

average_reward_given	offer_completed_informational	average_difficulty	offer_success_bogo	offer_success_discount	total_success_completed
0.450000	2	8.333333	0	0	
0.000000	0	0.000000	0	0	
0.722222	2	10.666667	1	2	
0.736842	0	10.000000	1	1	
0.565217	1	5.666667	2	1	

Figure 12: Offer Dataframe

4 Customer Segmentation Analysis

It is important to know your customers in order to do a better targeted marketing. In this section, we present the KMeans unsupervised machine learning algorithm to cluster customers into groups based on similar behaviors. All engineered features are used to find subgroups of customers with similar purchase behavior and offer response.

4.1 K-Means Clustering

We want to know the customers who will respond if Starbucks sends the offer and separate the customers group who may not respond to the offer. We apply standard scalar to scale the variables except age. Metrics used is within-Cluster-Sum-of-Squares (WCSS), the summation of observations from their cluster centroids. the Elbow method is used to determine the number of clusters. The WCCS plot with number of clusters helps to decide the number of clusters. The k value is picked at the spot where WCSS forms an elbow before it flattens.

4.2 Cluster Exploratory Data Analysis

- Cluster 0 - male dominated, low income customers and low response to offer sent. Almost 50% customers are not influenced by offers.
- Cluster 1 - includes high income groups and consists of both comparable male and female customers. Customers interested in both bogo and discount offers. Some customers are interested in informational offer. Regular customers with respond well to the offer.
- Cluster 2 - medium to high income group and again male dominated cluster. They respond well mostly to the informational offer.

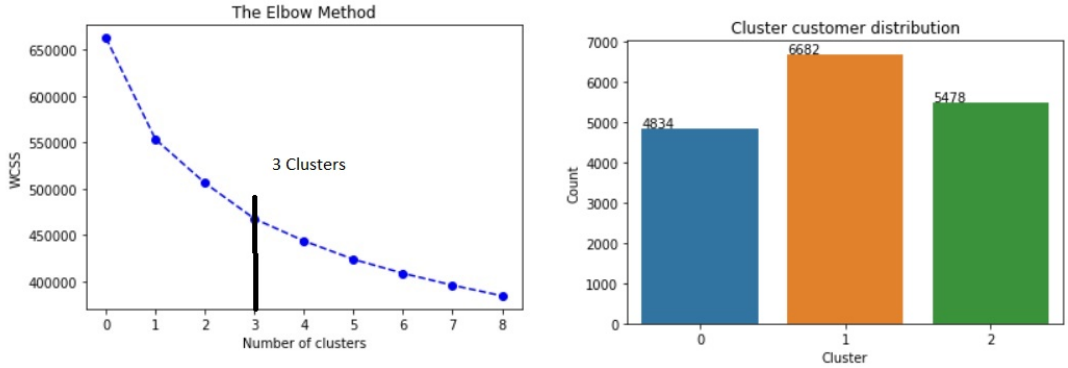

Figure 13: (a)The elbow method (b) cluster customer distribution

Figure 14: Distribution of (a) offer completion (b) offer type completion within cluster segments.

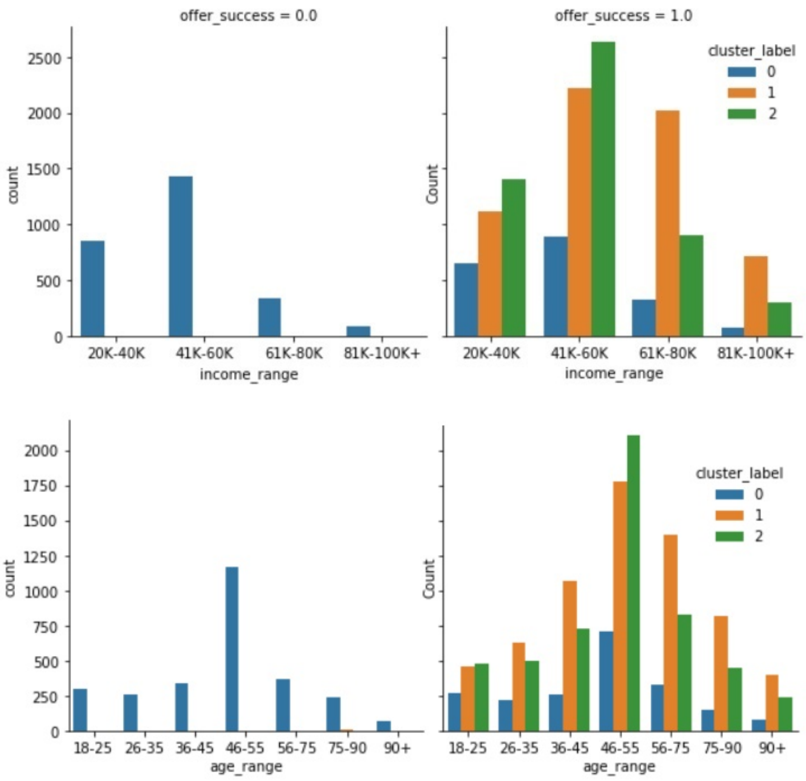

Figure 15: Effect of (a) income (b) age on the offer success

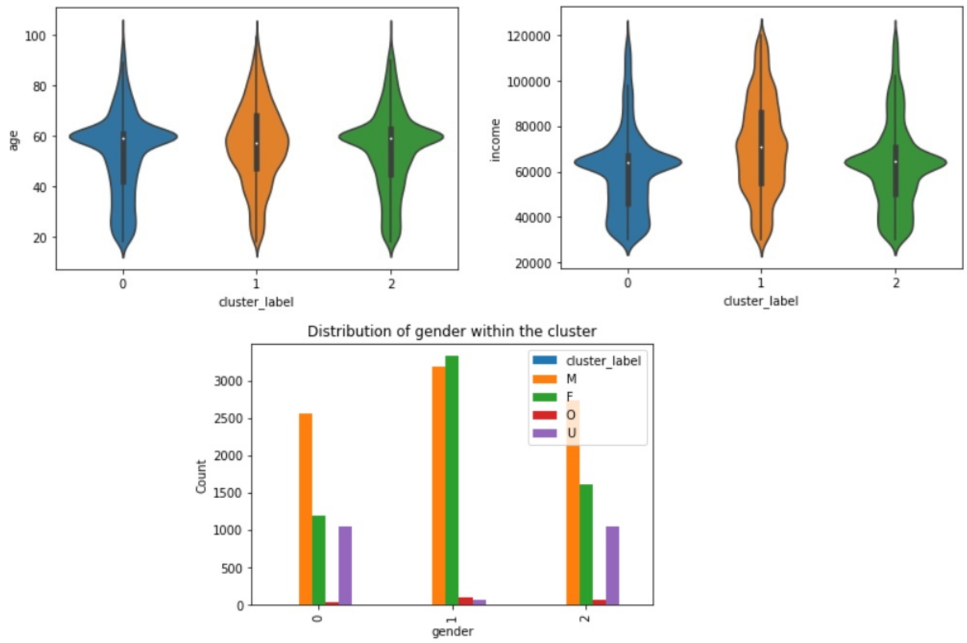

Figure 16: (a) age (b) income (c) gender distribution within the cluster.

4.3 Segmentation Analysis Results

The project aim is to provide with a strategy for the targeted promotional offers to high valued customers of Starbucks. Customer segmentation analysis showed the following distinct group of customers characteristic and demographics.

- **Group 1:** Very less response to the offer. It consists of young male dominated customers with low to medium income. They do not respond to offer frequently. Mostly above average age and older infrequent customers respond to offer. They are not regular customers. Starbucks should consider before sending offers to them.
- **Group 2:** This group is very responsive to all types of discount, informational and BOGO offers. It includes both male and female customers with high average income. These are high spending customers with or without offer. These are Starbuck's regulars customers to be targeted with the all types of offers.
- **Group 3:** This group is also interested in Starbuck's promotions mostly informational type offer. Again male dominated group with medium to high salary range. Mostly above average age customers.

First group is less influenced by offers so Starbucks can send selected individuals occasional bogo or discount offers. Starbucks can target second group with bogo and discount offer. Last group customers could be targeted with informational offers for their future promotions.

5 Effective Offer Prediction Model

5.1 Supervised Learning Model

Feature Engineering determines the performance of developed model. We have created a labeled dataset for developing the supervised learning offer prediction model. A target output is binary which predicts if customer will respond to offer or not. Benchmark model is Logistic regression since the business objective for Starbucks is to develop effective customer targeting. In order develop a best performing model, grid search is used to compare performance of all trained models namely, Logistic regression, Random Forest are compared based on AUC score.

Evaluation Metrics : Metrics are used for evaluating the model performance.

- Accuracy - The ratio of correctly predicted examples by the total examples. It shows often is the classifier correct. Accuracy may not be right metric always.
- Receiver operating characteristic (ROC) curve - more visual way to measure the performance of a binary classifier. It is created by plotting the true positive rate (TPR) (or recall) against the false positive rate (FPR),
- AUC: relation between true positive rate and false positive rate - AUC stands for Area under the ROC Curve. It provides an aggregate measure of performance across all possible classification thresholds.
- Confusion Matrix - table below showing calculated correct predictions and types of incorrect predictions.

Table 1: Confusion Matrix

Actual Value	Predicted class		
		Positive - Class 1	Negative - Class 2
	Positive - Class 1	True Positive (TP) (Right)	False Negative (FN) (Wrong)
	Negative - Class 2	False positive (FP) (Wrong)	True Negative (TN) (Right)

5.2 Model Settings

Feature Selection There are total 39 features. We have engineered new features. Based on the correlation values, we will remove offer received, viewed and completed features. We have engineered features representing them. The bogo and discount by offer view and completion rate were removed. We will not use the gender (using dummies) feature. We will use the average reward given instead of the total reward given. The year feature was also removed.
Grid Search I have implemented the model pipeline and used grid search to tune the hyperparameters. In order to improve the model prediction, hyperparameter tuning is an important step. In this project, I carried out an exhaustive grid search. I have set three levels of hyperparameters and it tries every single possible combination of the hyperparameters as well as cross-validations. Grid search is a good way to determine the best hyperparameter values in this case since our customer dataset is not very large.

Model Selection I have selected two models. Linear model for simplicity to use as a benchmark. Second model is the Random Forest a tree based algorithms.

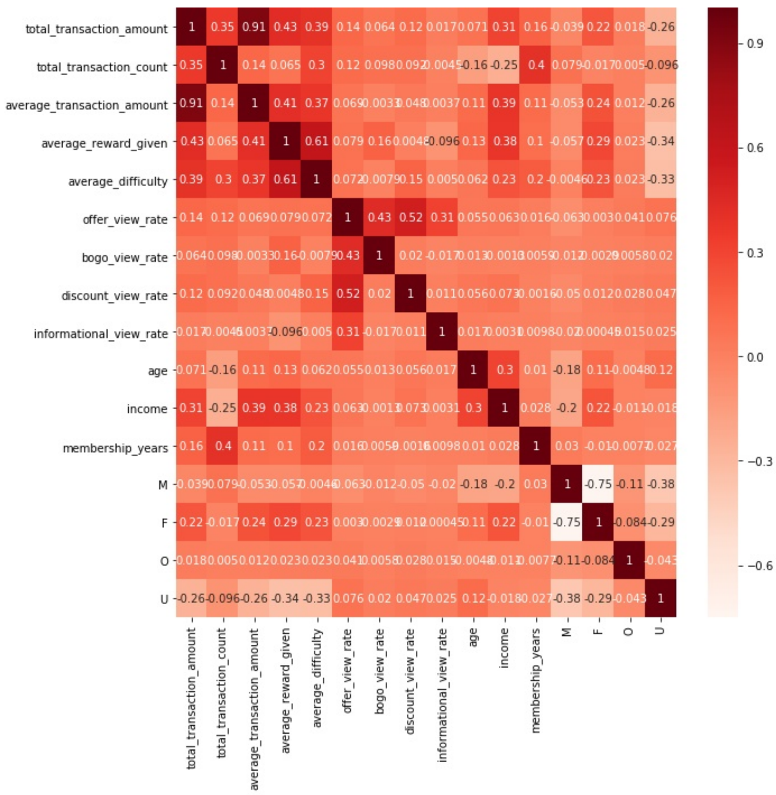

Figure 17: Correlation heat map

- A logistic regression model : benchmark model - constructs a linear decision boundary to separate successful and unsuccessful offers.
- Random forest (RF) [15]- RF model is a combination of multiple decision trees. RFs train each tree independently, using a random sample of the data. This randomness helps to make the model more robust than a single decision tree, and less likely to over fit on the training data.

5.3 Model Evaluation and Results

Logistic regression benchmark model AUC score was 0.8711. Grid search helped to improve the random forest model AUC score to 0.9723. Training and test accuracy for the random forest is comparable. When we fit the model to all the features, it was over fitting. Based on correlation heatmap, we removed few features. It helped to reduce the over-fitting. Feature importance metric helps to understand the significance of different features to the model fitting. The top 5 features are information view rate, offer difficulty, average reward given, total transaction amount and offer view rate are important features. In other words, if offer is viewed then more likelyhood of completion. When customer is spending lot of money, he is regular Starbucks customer so interested in offers too. Obviously offer difficulty amount showed that regular customers care not affected by offer difficulty who are high spending Starbucks customers. Income, age are at the bottom of feature importance metric. We have seen from EDA that high income people spend more amount. and older people are Starbucks main customer base. Overall, I am surprised that demographic features are not at the top. Gender did not play important role.

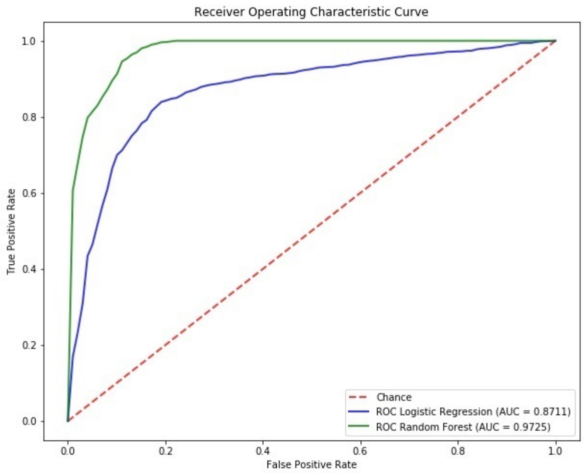

Figure 18: Receiver operating characteristics curve

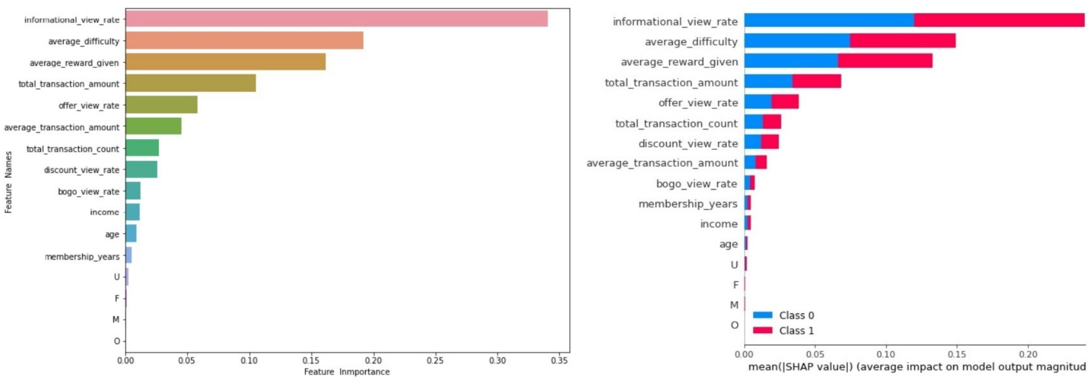

Figure 19: (a) Feature importance (b) Shap values for the developed model

Classifier Name	AUC	Train Accuracy	Test Accuracy
Logistic Regression	0.87	0.86	0.86
Radom Forest	0.97	0.99	0.97

Figure 20: Model Results

Model Prediction Result

customer_id	Target	prob_0	prob_1
169ad543fdaa433a820d84502c32a826	1	0.167106	0.832894
139ba634b97a48699224973289e3b484	1	0.009556	0.990444
c6c1259f6fbc4e6c82a7b9dd3c83f636	1	0.145582	0.854418
2afd6f11beba4470bdb8cc4cca19cec2	0	0.608845	0.391155

6 Reflection and Future Improvements

Market segmentation, targeting, and positioning (STP) process influence a company's strategy for pricing, communication, and customer management [11]. Thus, this capstone project was interesting to get hand on experience to solve an important business problem. The project goal was to develop which offer is effective for the Starbucks customer using the provided dataset. Data cleaning, feature engineering and model development were straight forward tasks. As mentioned in the Starbucks challenge, the challenging part was to separate the effective offer and informational offer based on the merged customer profile and transactions dataset. Overall, the customer segmentation gave insight about who are real Starbucks customer. There are three distinct groups of customers. The customer segmentation analysis provided direction for the targeted offer promotion. The effective offer prediction model performance on the test dataset provided 97% AUC score. The offer prediction test results confirmed that model is not over-fitting.

- Bogo, Discount and Informational individual offer model can be developed to check the effectiveness of specific offers to be sent to the customer.
- Missing features were imputed. Implementing machine learning model based missing value prediction.
- Using different clustering algorithm for the customer segmentation.
- Recommendation system to suggest right offers to right customers.

References

[1] Application of AI in Marketing

[2] What Are The Technological Advancements In Marketing?

[3] How AI is Changing Digital Marketing

[4] Here's What You Need to Know About Propensity Modeling

[5] A gentle introduction to Customer segmentation

[6] The Most Important Data Science Tool for Market and Customer Segmentation

[7] Credit Card Fraud Imbalanced Dataset: Review, Metrics, Resampling techniques, Data Exploration and Comparing Machine Learning Models

[8] 11 Important Analytical Steps for your Data Science project

[9] A Data Science Workflow

[10] Comprehensive Guide to Data Exploration

[11] Market Segmentation and Targeting identify potential customers, choose target customers, and create value

[12] Handling Missing Values in Machine Learning

[13] Difference between response rate and completion rate

[14] Feature selection algorithms every data scientist need to know

[15] Optimizing Hyperparameters in random forest classification