MLND – Starbucks – Capstone

Amit Kumar

Dec’1 2020

Definition

Project Overview

Starbucks, as a marketing strategy, provides discounts or free services on its offerings to users. This helps in increasing sales and creating customer awareness of new offering. Starbucks has a diverse user base. Hence not all offers are rolled out to each customer. Starbucks gives different offers to different customers based on their demographics and user profile. In a simple words Starbucks' offering are personalised. Business lean towards propensity model to reduce the expenditure in advertising, as model give them more specific target customer.

Problem Statement

Since we have varied groups of users with varied preference or liking, we need to be smart in rolling out offers to them. Based on the historical data that is available to Starbucks, we can make a data backed decision whether to rollout the offer will result in a successful business value. So, in this project we will try to build an ML model which predicts whether a user or group of users will respond to an advertisement or offer (propensity modelling). With this known beforehand with some accuracy we can be extremely specific while we do advertisement targeting

Metrics

FI -Score of the machine learning models is used as the performance metrics.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

This is harmonic mean of Precision (True positive/ Predicted Positive) and Recall (True Positive/Actual Positive).

Analysis

Data Exploration

Portfolio Dataset:

	reward	channels	difficulty	duration	offer_type	id
0	10	[email, mobile, social]	10	7	bogo	ae264e3637204a6fb9bb56bc8210ddfd
1	10	[web, email, mobile, social]	10	5	bogo	4d5c57ea9a6940dd891ad53e9dbe8da0
2	0	[web, email, mobile]	0	4	informational	3f207df678b143eea3cee63160fa8bed
3	5	[web, email, mobile]	5	7	bogo	9b98b8c7a33c4b65b9aebfe6a799e6d9
4	5	[web, email]	20	10	discount	0b1e1539f2cc45b7b9fa7c272da2e1d7

Channels -> web, email, mobile, social
Offer_type -> bogo, informational, discount

Profile Dataset:

	gender	age	id	became_member_on	income
0	None	118	68be06ca386d4c31939f3a4f0e3dd783	20170212	NaN
1	F	55	0610b486422d4921ae7d2bf64640c50b	20170715	112000.0
2	None	118	38fe809add3b4fcf9315a9694bb96ff5	20180712	NaN
3	F	75	78afa995795e4d85b5d9ceeca43f5fef	20170509	100000.0
4	None	118	a03223e636434f42ac4c3df47e8bac43	20170804	NaN

Gender -> F, M, O
Age -> 18-118 [118 was outlier]
Income -> 30-120K
Became member on -> 2017-2018

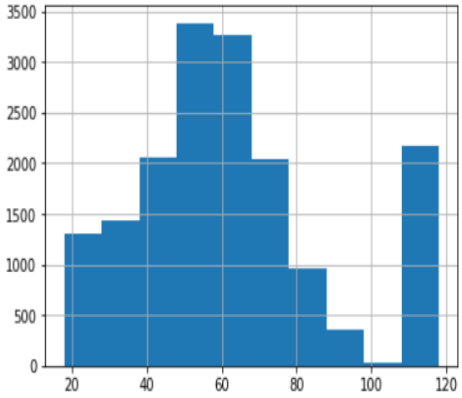
Transcript Dataset:

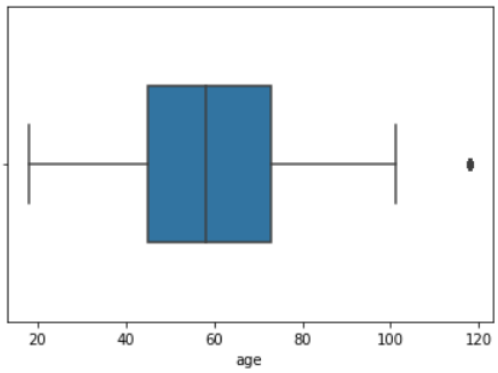
	person	event	value	time
0	78afa995795e4d85b5d9ceeca43f5fef	offer received	{'offer id': '9b98b8c7a33c4b65b9aebfe6a799e6d9'}	0
1	a03223e636434f42ac4c3df47e8bac43	offer received	{'offer id': '0b1e1539f2cc45b7b9fa7c272da2e1d7'}	0
2	e2127556f4f64592b11af22de27a7932	offer received	{'offer id': '2906b810c7d4411798c6938adc9daaa5'}	0
3	8ec6ce2a7e7949b1bf142def7d0e0586	offer received	{'offer id': 'fafdcd668e3743c1bb461111dcafc2a4'}	0
4	68617ca6246f4fbc85e91a2a49552598	offer received	{'offer id': '4d5c57ea9a6940dd891ad53e9dbe8da0'}	0

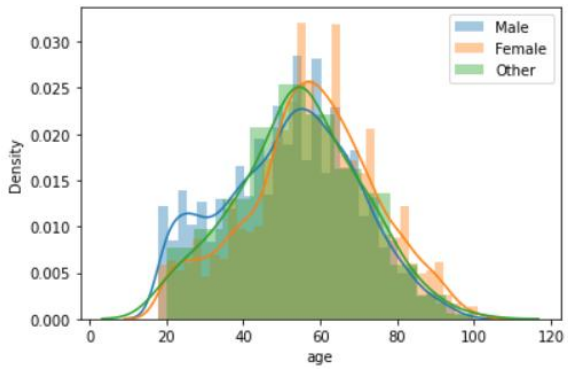
Event -> Offer received, offer viewed, offer completed, transcation
Value -> offer_id, reward, amount

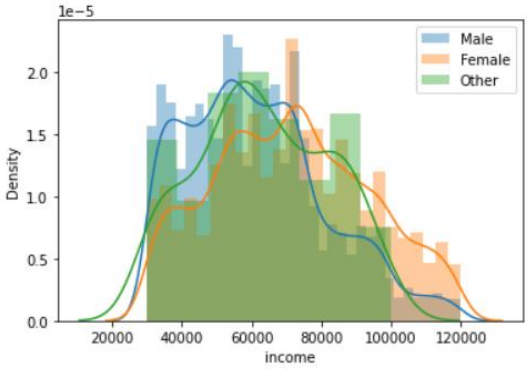
Visualization

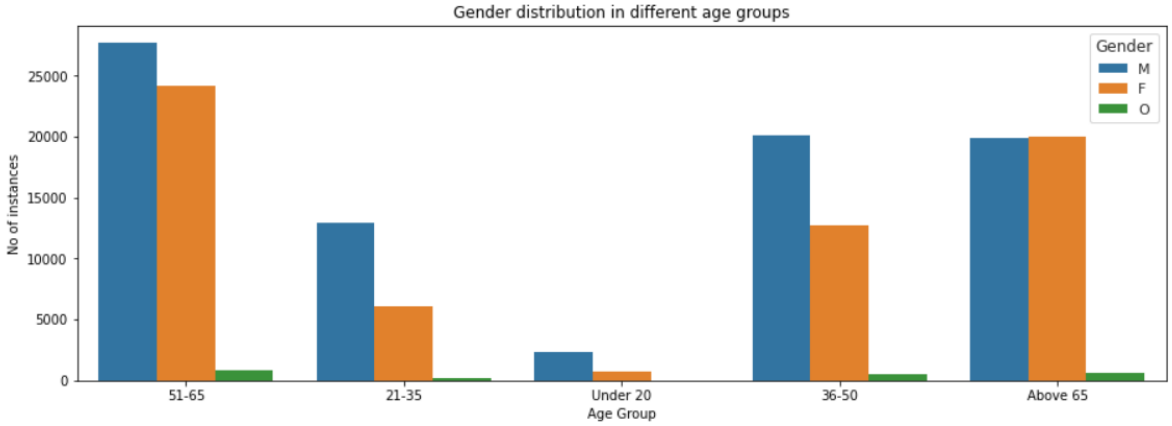
We tried to explore that dataset visually. We used barplot, histogram, boxplot etc.

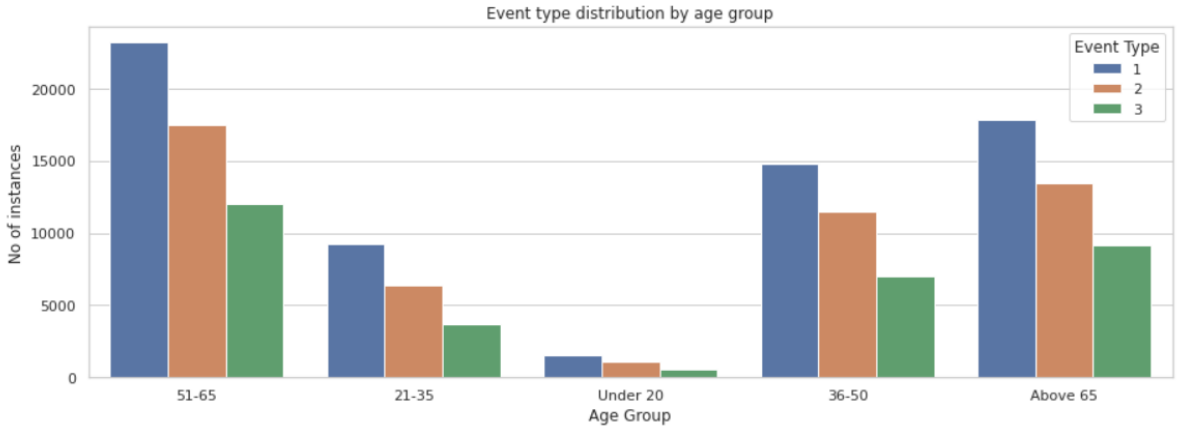


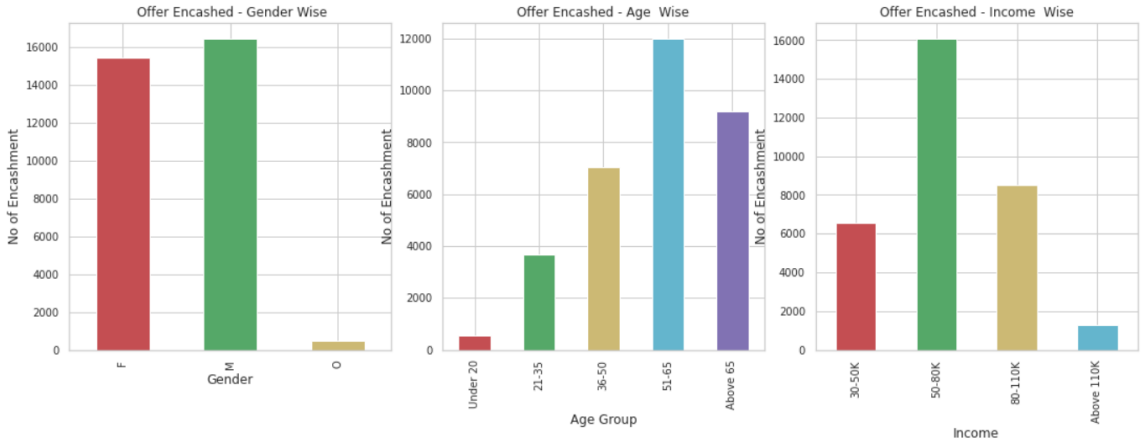












Algorithm & Techniques

In this process we tested below algorithms

- Logistic Regression
- Decision Tree
- Random Forest
- KNN

Benchmark

For our study we chose Logistic Regression as Benchmark.

Methodology

Data Processing

Dropped all null rows.

$$profile = profile.loc[profile['gender'].isnull() == False]$$

We clubbed Age into Age groups. As similar age group's people's behvior is likely to be similar.

In line with that we also grouped people based on their income group. Below table green box depicts how a person of age 62 falls in **51-65 Age group**. And same person has income in access of 110k so he has been placed into **Above 110K income group**.

	gender	customer_id	Age_group	Income_group	member_since_days
1	F	0610b486422d4921ae7d2bf64640c50b	51-65	Above 110K	376
3	F	78afa995795e4d85b5d9ceeca43f5fef	Above 65	80-110K	443
5	M	e2127556f4f64592b11af22de27a7932	Above 65	50-80K	91
8	M	389bc3fa690240e798340f5a15918d5c	51-65	50-80K	167
12	M	2eeac8d8feae4a8cad5a6af0499a211d	51-65	50-80K	257

In the transcript dataset we have ana attribute with a nested values(map/dictionary). I have splitted that attribute into three different columns (attribute keys) where each column contains corresponding value for the key.

	customer_id	event	time	offer_id	reward	amount
0	78afa995795e4d85b5d9ceeca43f5fef	offer-received	0	9b98b8c7a33c4b65b9aebfe6a799e6d9	0.0	0.0
1	a03223e636434f42ac4c3df47e8bac43	offer-received	0	0b1e1539f2cc45b7b9fa7c272da2e1d7	0.0	0.0
2	e2127556f4f64592b11af22de27a7932	offer-received	0	2906b810c7d4411798c6938adc9daaa5	0.0	0.0
3	8ec6ce2a7e7949b1bf142def7d0e0586	offer-received	0	fafdcd668e3743c1bb461111dcafc2a4	0.0	0.0
4	68617ca6246f4fbc85e91a2a49552598	offer-received	0	4d5c57ea9a6940dd891ad53e9dbe8da0	0.0	0.0

Implementation

Transcript dataset and Portfolio dataset were combined together based on the offer_id key. This was a left join as transaction record were not having offer_id. This combined datframe was joined with profile dataset on customer id. Then merged dataset had 24 attributes. We also generated Target attribute, based on whether the Customer id and Offer Id pair were completed or not. If completed it's a success else it's a failure.

Model were supplied with `52300 Training records & 22415 Test records.`

Results

Model Evaluation results

Once we analysed our models on the given dataset we had the following F1- score table. We can see that RandomForest performed better than others including Benchmark.

	Classifier	**F1-Score(x100)**
0	LogisticRegression(Benchmark)	**64.903**
1	RandomForest	**75.497**
2	DecisionTree	**67.497**
3	KNN	**69.451**

After doing Optimization (Hyperparameter Tuning).

	Classifier	**F1-Score(x100)**
1	RandomForest	**81.231**
2	DecisionTree	**76.441**
3	KNN	**79.103**

Justification

From the above table we see that Random Forrest classifier was better from the Benchmark without any optimization. With optimization its F1 score went further up. Since we do not have some critical application we can be fairly confident with .81 F1 score.

Conclusion

Reflection

From the analysis we found that Men in the age group 51-65 and having income in range 50-80K are very receptive to the offers and their conversion rate is high. We can target these to boost sale.

Improvement

We can add session information, to create a customer journey to see iit transition through various stages and then rebuild the model to take into account of chain of events.

Offer received -> Offer Viewed -> Offer Completed -> Transaction

We then can map a timeline. Also we can explore other segmentation model to club similar customers into one.