CUSTOMER SEGMENTATION

Using Machine Learning

PROJECT EXHIBITION - 2

FINAL REVIEW

Team Members & Guide

Krishan Kumar Gupta
19BAI10114

Himesh Sharma
19BAI10125

Harsh Pal
19BAI10127

Adwitiya Dubey
19BAI10176

Dr. Nilamadhab Mishra
100133

Introduction

- In this project, We look at how consumers use the Starbucks rewards app on their phones. Customers receive discounts a few days after signing up for the app. The aim is to find out which consumers are most affected by promotional deals and what types of offers to give them so that revenue is maximized.

- Each offer is valid for a certain number of days before it expires. Discounts and BOGOs have varying degrees of complexity, depending on how much the consumer must pay to qualify for the deal. Promotions are disseminated through several outlets (social, web, email, mobile).

- All transactions made through the app are tracked automatically. The app also records information about which offers have been sent, which have been viewed and which have been completed, and when these three events happened.

Existing Work

- Companies may classify multiple segments of consumers using clustering strategies, enabling them to reach the potential consumer base.

- We will use PCA & K-means clustering in this machine learning project, which is the most important algorithm for clustering unlabeled datasets.

Limitations

- Instead of simply imputing missing age, income, and gender values, one option for improving clustering results is to forecast them.

- In this case, we will use supervised machine learning and experiment with various models such as Random Forest, Ada Boost, and others.

- Another useful enhancement would be to use the clustering results as labels in supervised modeling to estimate the likelihood of the customer fulfilling the deal.

- This would be a useful practical implementation that would allow for the expansion of existing customers and thus aid in the potential execution of a successful marketing campaign.

Proposed Work & Methodology

- The idea is to categorize app users into four groups: those who prefer deals, those who prefer BOGOs, and those who don't care for promotions at all. The number of groups was determined later, based on the final dataset's patterns.

- The optimum number of segments to use in a customer segmentation task is a vital decision. The issue with unsupervised machine learning is that it lacks clearly defined benchmark metrics for evaluating model performance on par with supervised machine learning (e.g. accuracy score, f1-score, AUC, etc.).

Novelty of the Project

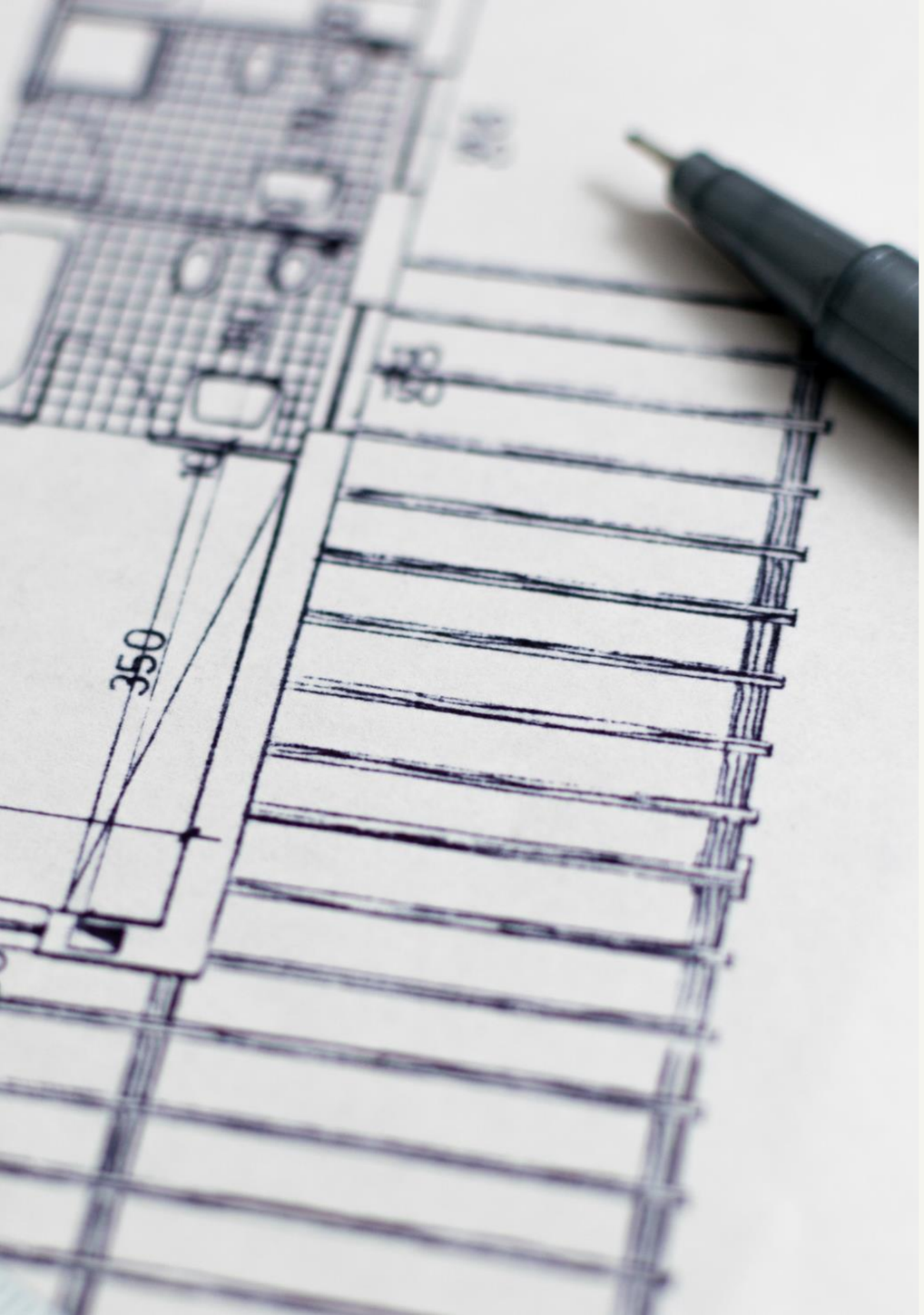
- GUI based web app.
- With Better accuracy.
- Works on the existing customer segmentation model.
- Requires minimal setup from the customer’s side.
- Portable, hence can be taken anywhere.
- Also, downloadable through internet.

Real Time Usage

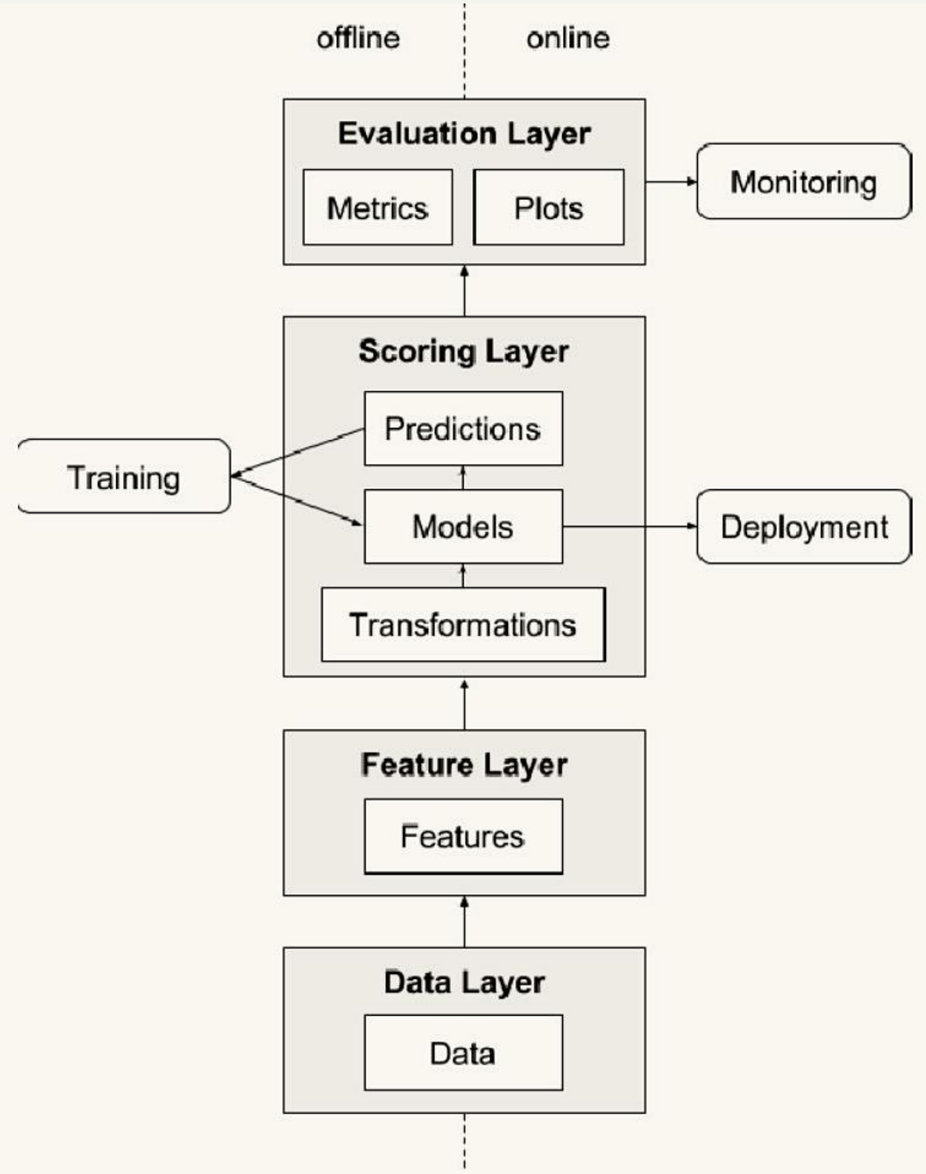
- Businesses can use segmentation to make better use of their marketing budgets, gain a competitive advantage over competitors, and, most importantly, demonstrate a better understanding of their customers' needs and wants.
- Better matching of customer needs.
- Better opportunities for growth.
- Target marketing communications.
- Enhanced profits for the business.

Hardware & Software Requirements

- A laptop with a web browser
- Dataset to work on
- GPU
- Software:
  - Python3
    - libraries
      - pandas
      - NumPy
      - seaborn
      - Matplotlib
      - TensorFlow
      - flask
      - Keras



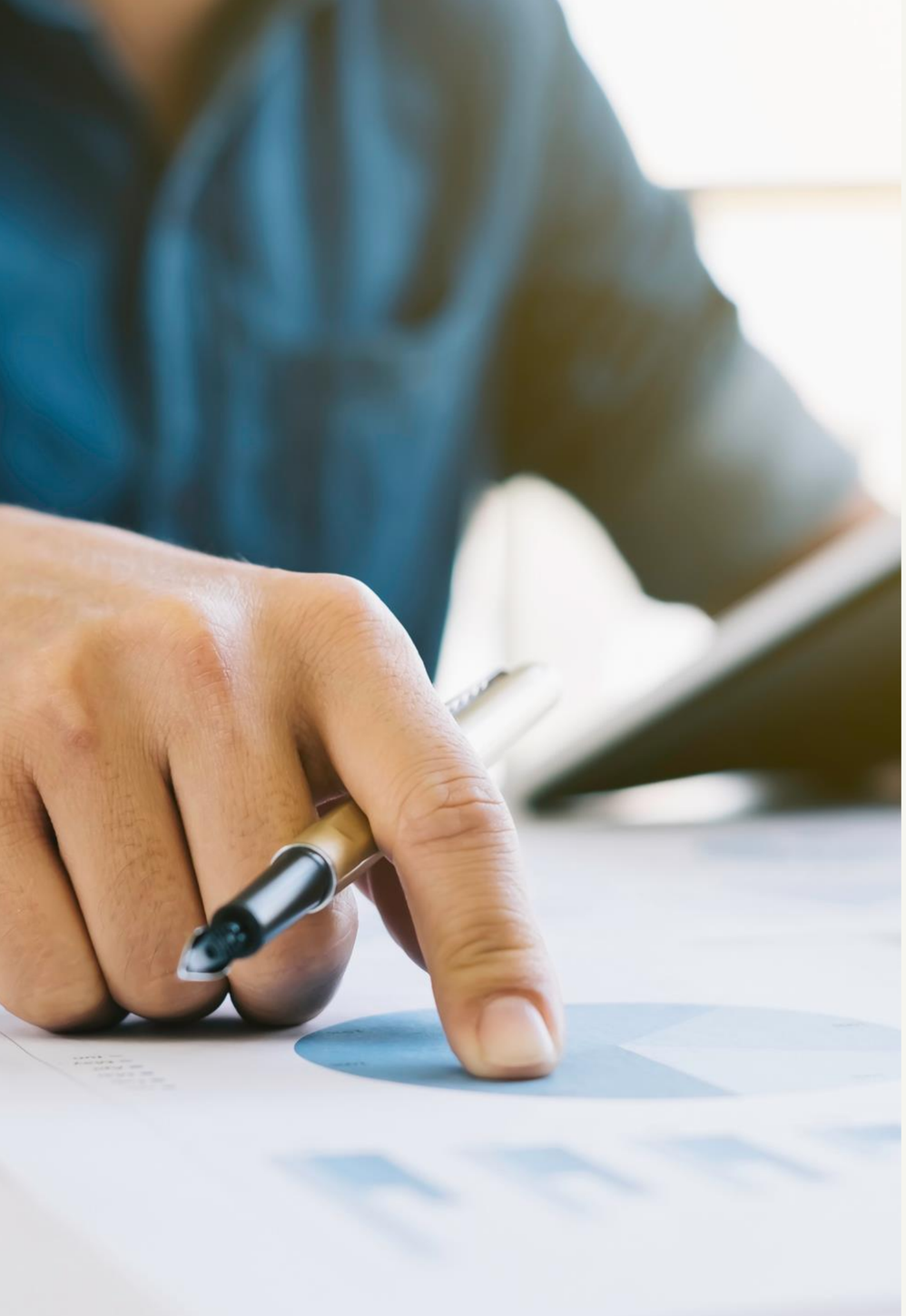
System Architecture



Literature Review

Customer Classification

- Over the years, the commercial world has become more competitive, as organizations such as these have to meet the needs and desires of their customers, attract new customers, and thus improve their businesses.
- The task of identifying and meeting the needs and requirements of every customer in the business is very difficult. This is because customers can vary according to their needs, wants, demographics, size, taste and taste, features etc.



Literature Review

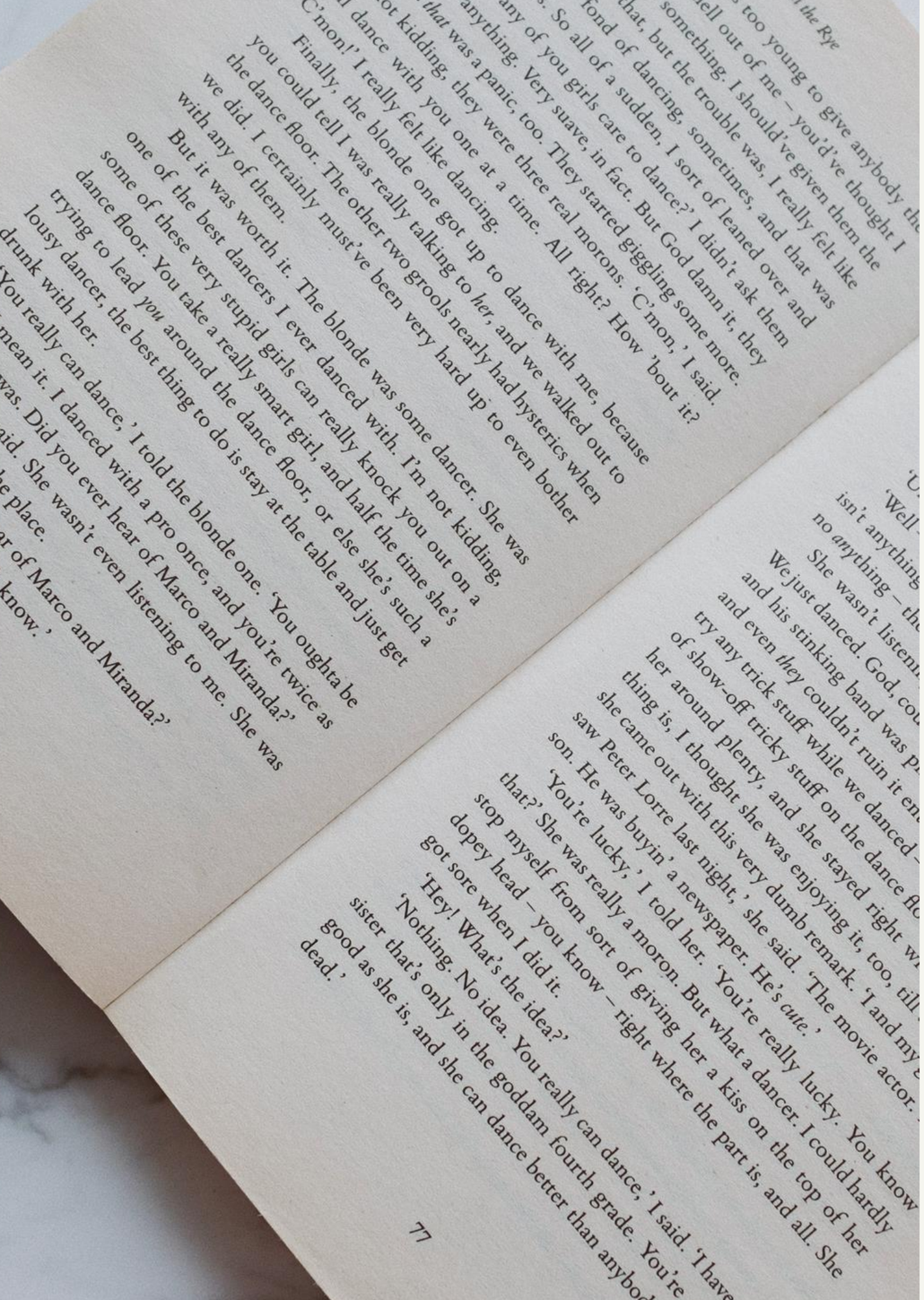
Big Data

- Big Data analysis has recently gained traction. Companies have billions of data about their clients, suppliers, and activities, and millions of internally linked sensors send sensing, production, and communications data to the real world on devices like cell phones and vehicles.
- Ability to improve forecasting, save money, improve performance, and improve a variety of areas including traffic control, weather forecasting, disaster management, banking, fraud control, business transactions, national security, education, and healthcare.

Literature Review

Data Collection

- Data collection is the process of gathering and analyzing information in response to specific changes in an existing structure, allowing one to answer pertinent questions and assess the outcomes.
- Data collection is part of research in all fields of study including physical and social sciences, humanities and business.
- The purpose of all data collection is to obtain quality evidence that leads the analysis to construct concrete and misleading answers to the questions presented.



Literature Review

Clustering Data

- Clustering is the method of categorizing data into groups based on commonalities.
- There are a number of algorithms that can be used on datasets based on the given condition. However, since there is no universal clustering algorithm, it is critical to choose the necessary clustering techniques.

Literature Review

K-Means Clustering

- An algorithm with a K value is one of the most widely used classification algorithms. This clustering algorithm is based on centro, which places each data point in one of the overlapping clusters that have been pre-sorted using the K-algorithm.
- Clusters are formed that correspond to hidden trends in the data, providing the required information to assist in the decision-making process. There are a variety of ways to put together K-means, but we'll use the elbow type.

Module Description & Work Flow

Step 1
Gathering data from various sources

Step 2
Cleaning data to have homogeneity

Step 3
Model Building-
K Means Clustering

Step 4
Gaining insights from the model's results

Step 5
Data Visualization-
Transforming results into visuals graphs

Dataset Overview

The data is organized in three files:

- portfolio.json (10 offers x 6 fields) - offer types sent during 30-day test period
- profile.json (17000 users x 5 fields) - demographic profile of app users
- transcript.json (306648 events x 4 fields) - event log on transactions, offers received, offers viewed, and offers completed

Here is the schema and explanation of each variable in the files:

portfolio.json

- id (string) - offer id
- offer_type (string) - type of offer ie BOGO, discount, informational
- difficulty (int) - minimum required spend to complete an offer
- reward (int) - reward given for completing an offer
- duration (int) - time for offer to be open, in days
- channels (list of strings)

Dataset Overview

profile.json

- age (int) - age of the customer
- became_member_on (int) - date when customer created an app account
- gender (str) - gender of the customer
- id (str) - customer id
- income (float) - customer’s income

transcript.json

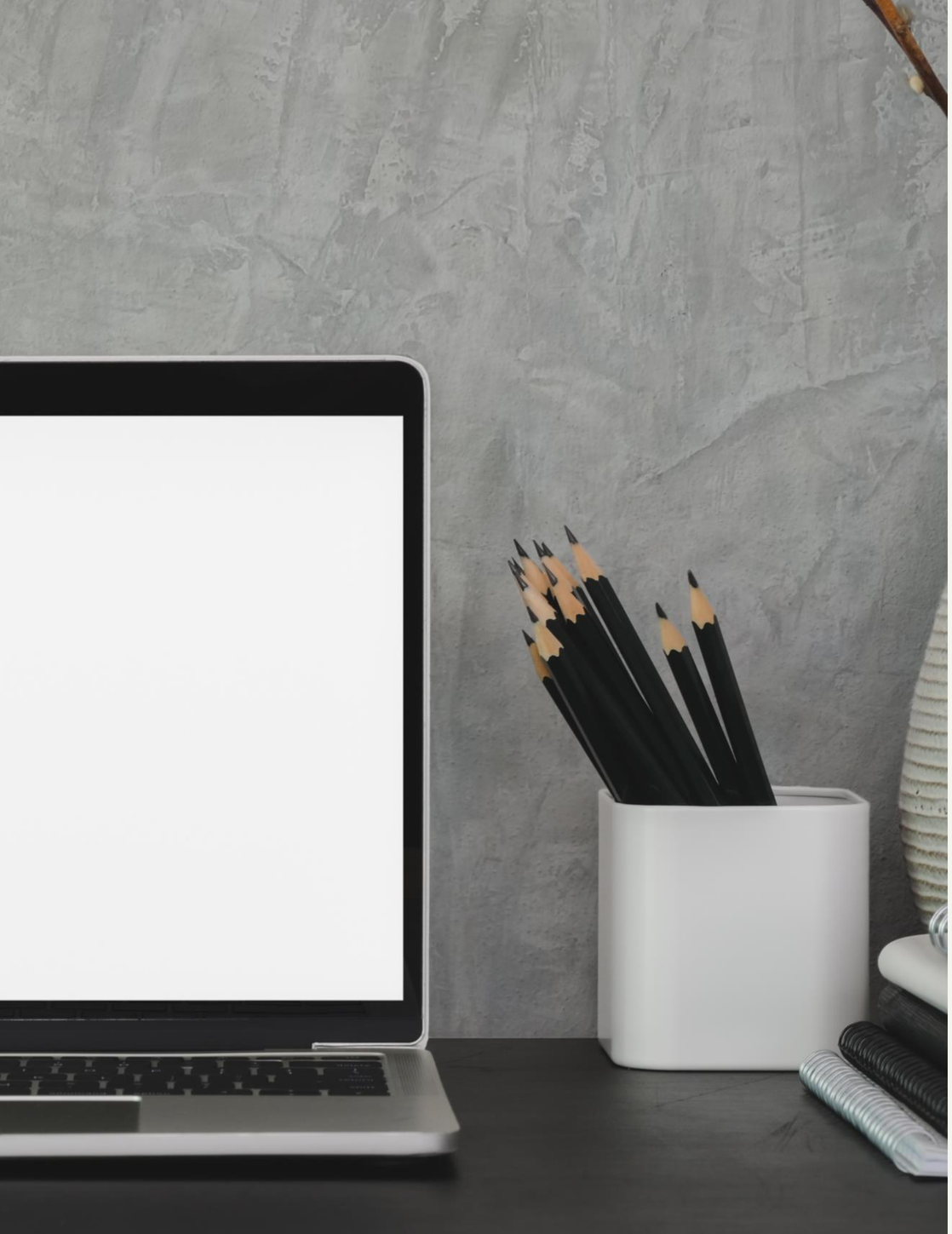
- event (str) - record description (ie transaction, offer received, offer viewed, etc.)
- person (str) - customer id
- time (int) - time in hours since start of test. The data begins at time t=0
- value - (dict of strings) - either an offer id or transaction amount depending on the record

Data Cleaning

The first move was to amalgamate the three datasets (portfolio, profile, and transactions) into a single final dataset.

Since the purpose of the project was to define consumer segments based on their interaction with promotional offers, we chose to aggregate data at the customer level.

However, before we could do so, We needed to reorganise the transaction details, which was very disorganised.



K-Means Clustering

Since linear unsupervised machine learning and K-means clustering depend on distances as a measure of similarity, they are susceptible to data scaling.

To address this problem, we transformed the data before clustering by:

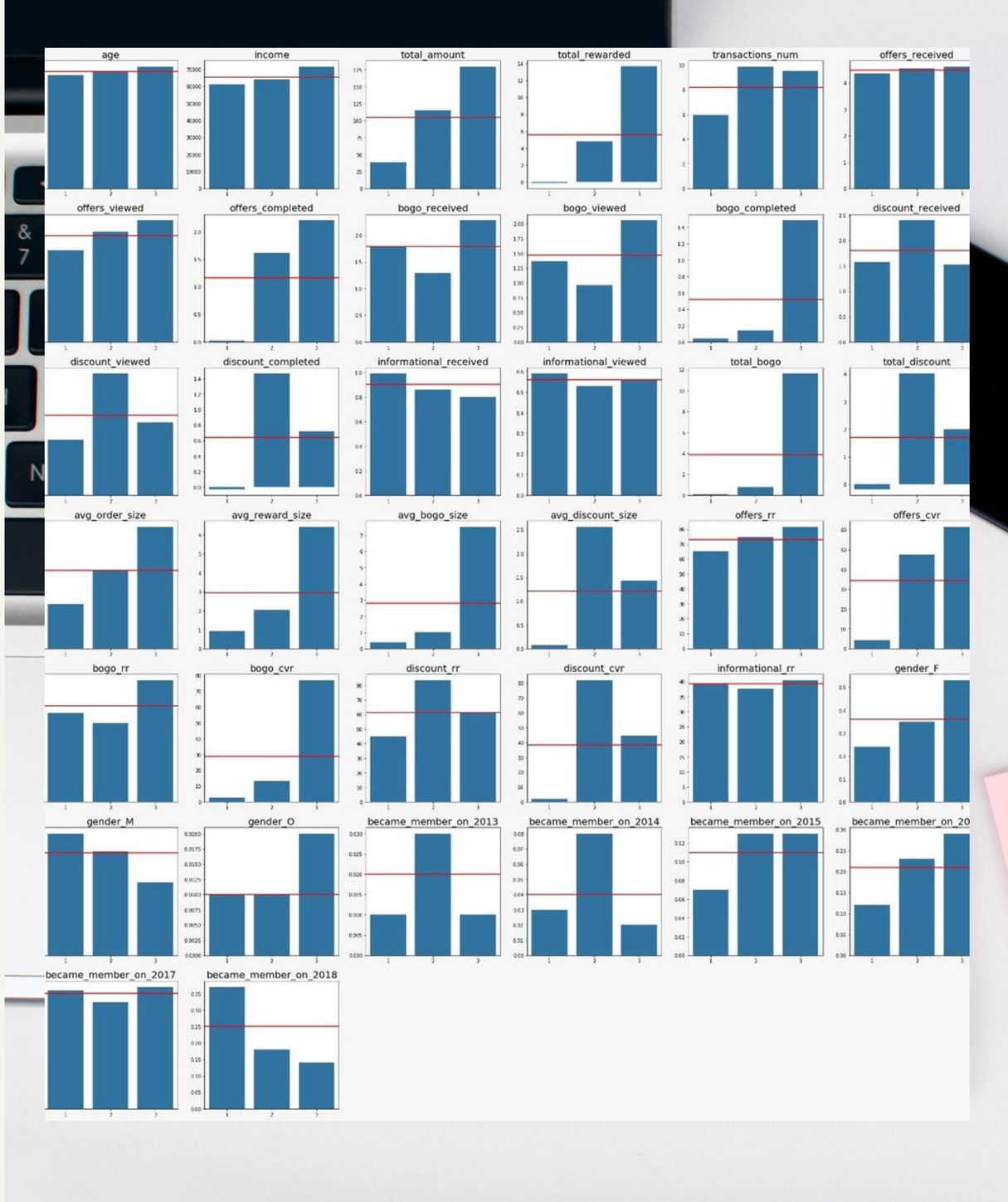
- one-hot encoding categorical features,
- scaling the features,
- performing dimensionality reduction.

Model Results

While the density-based algorithms tend to perform better in the charts above, the k-means clusters have more distinct characteristics that make sense in our business context.

The key issue tends to be that DBSCAN and OPTICS overestimate customer gender and membership year because those variables are more tightly clustered.

The resulting data shows minimal variance in our view rate and conversion rate metrics, and is therefore not actionable in our marketing.



Implementation and Coding

K-means Clustering

The aim of clustering is to find groups of data with values that are close to one another. K-means is a well-known algorithm for locating data points that are nearest to the cluster centroid. In technical terms, it attempts to reduce intra-cluster variance, which is calculated as the amount of each data point's squared distance to the mean of the allocated cluster.

You can access our executable source codes and files via the below mentioned link.

https://bit.ly/2QmAjxE

https://github.com/Krishan101/CustomerSegmentation

Evaluation and validation

Although the density-based algorithms may appear to perform better in the charts above, the clusters generated using k-means show more distinct characteristics that make sense in our business context.

The main problem appears to be that DBSCAN and OPTICS overemphasize the gender and membership year of the customers, as those variables are more densely clustered.

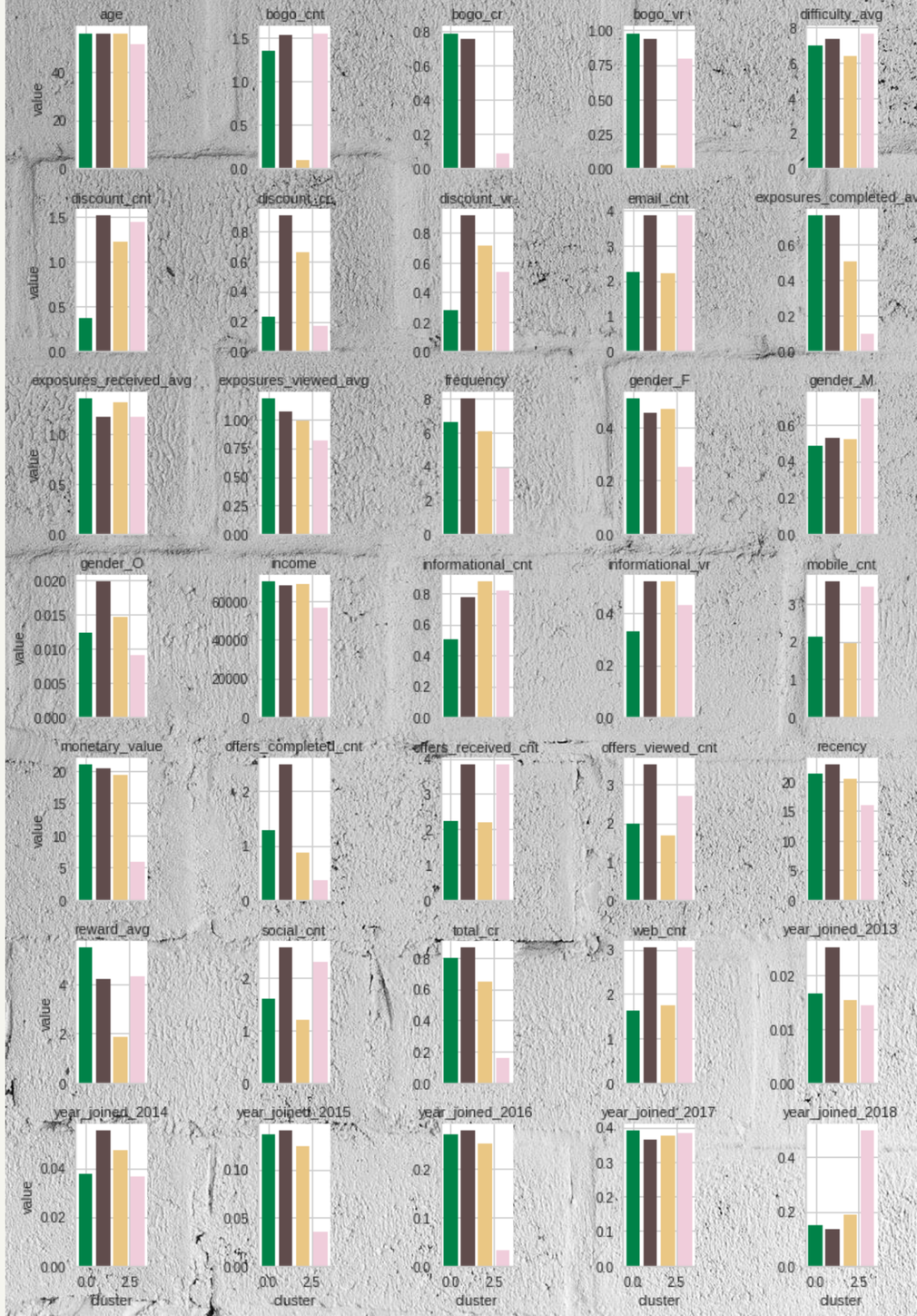
Below, we can immediately identify four distinct segments with clear business implications:

Segment 1

Customers in this segment receive regular BOGO offers, and practically no discount offers. These BOGO offers involve more valuable rewards than for customers in other segments.

Segment 2

Customers in this segment receive a higher than average number of offers, and convert really well for both BOGOs and discounts. Demographically, a higher than average share of these customers selected their gender as Other.



Segment 3

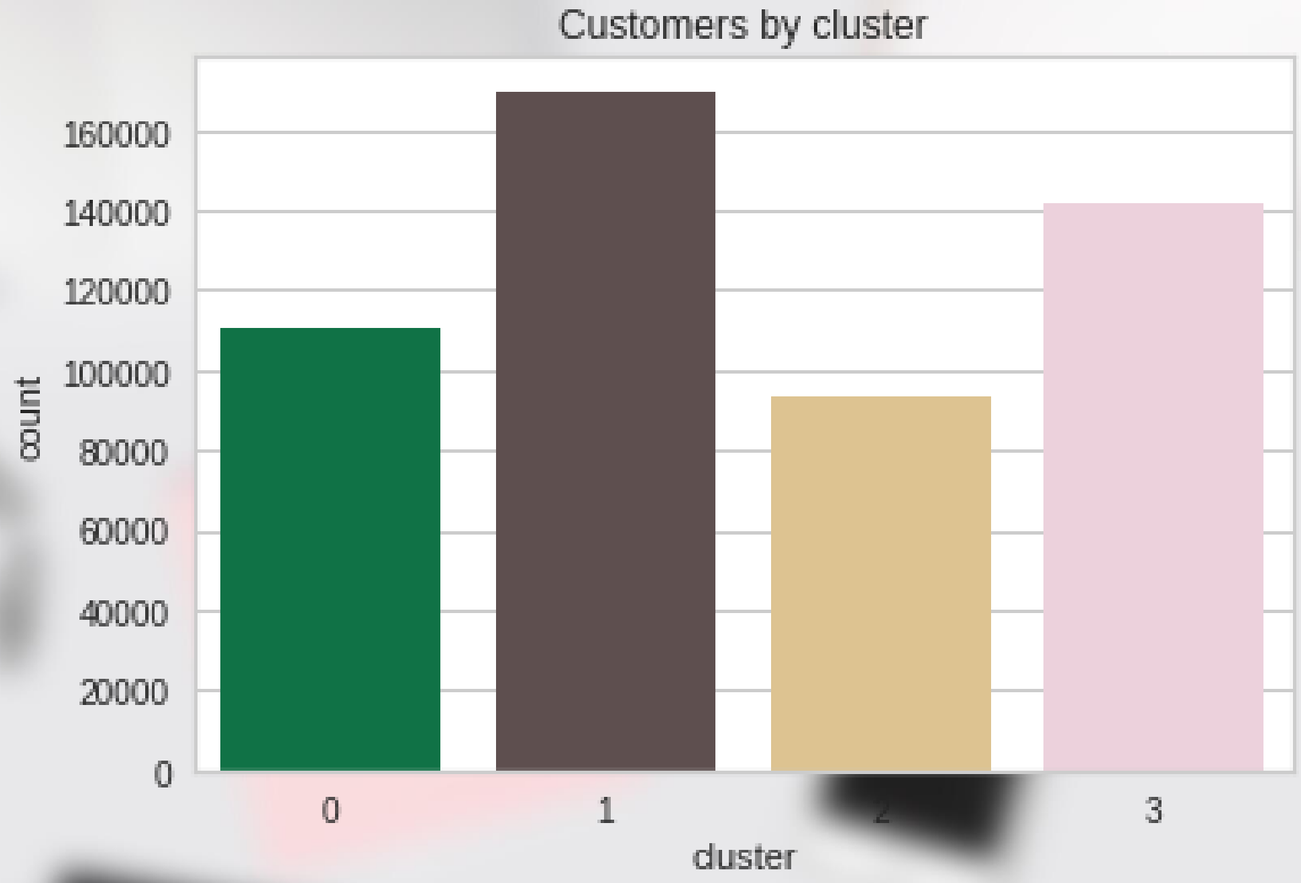
Customers in this segment receive no BOGO offers. They do get occasional discount offers, on which they convert about average, as well as slightly more informational messages than other customers.

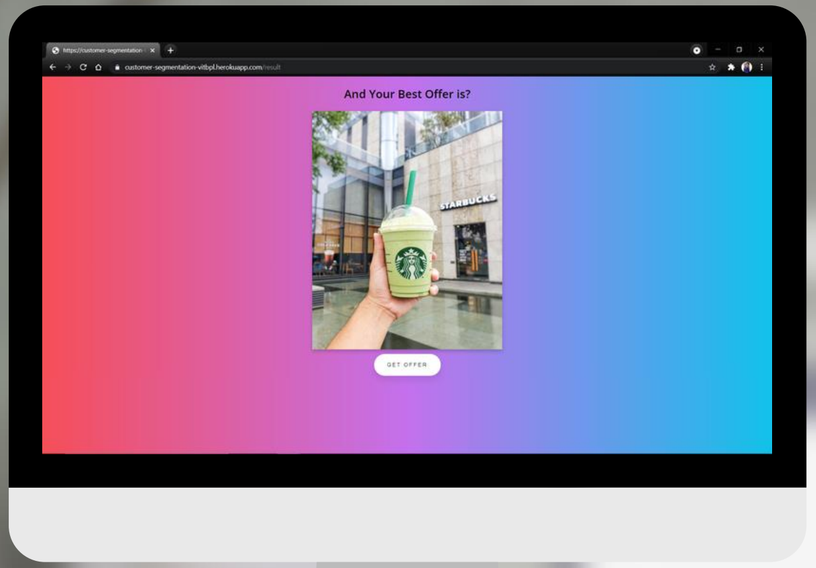
Segment 4

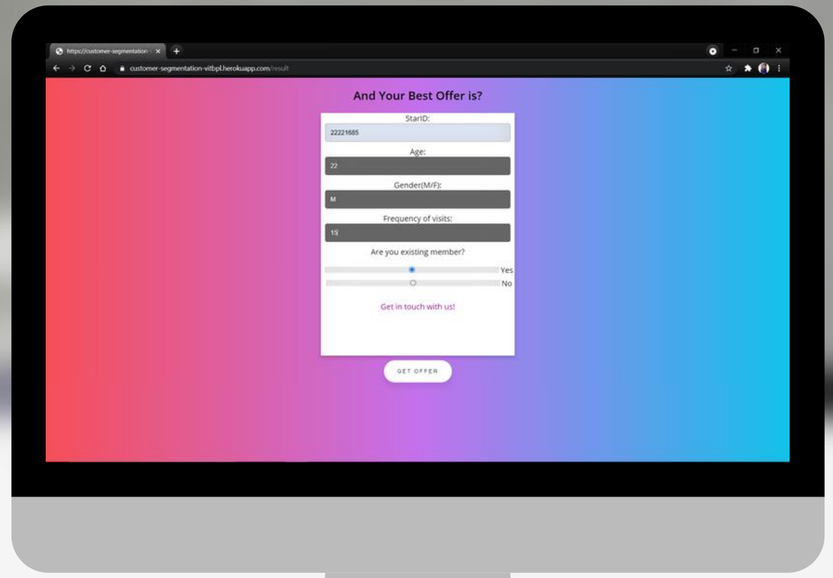
Customers in this segment receive regular offers, which they open, but never convert.

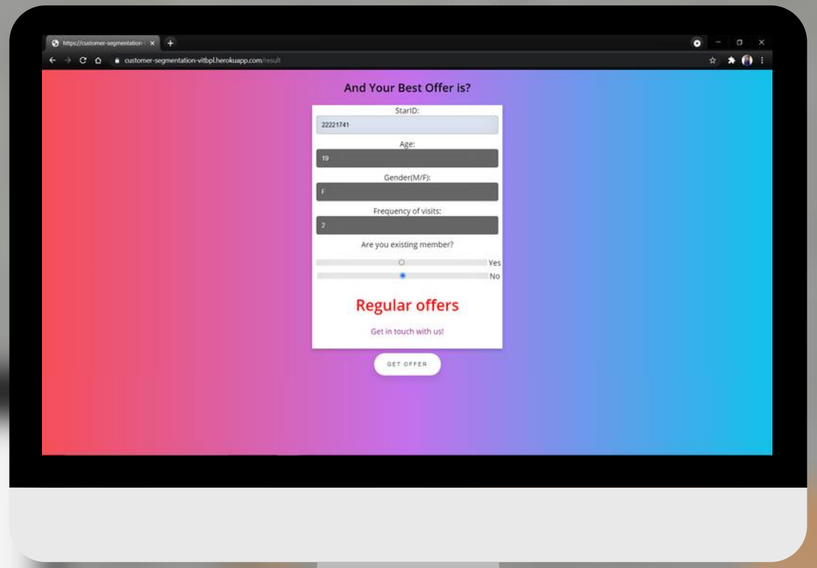
Demographically, they are predominantly male, and lower than average income. They also visit Starbucks less frequently, and make smaller average purchases.

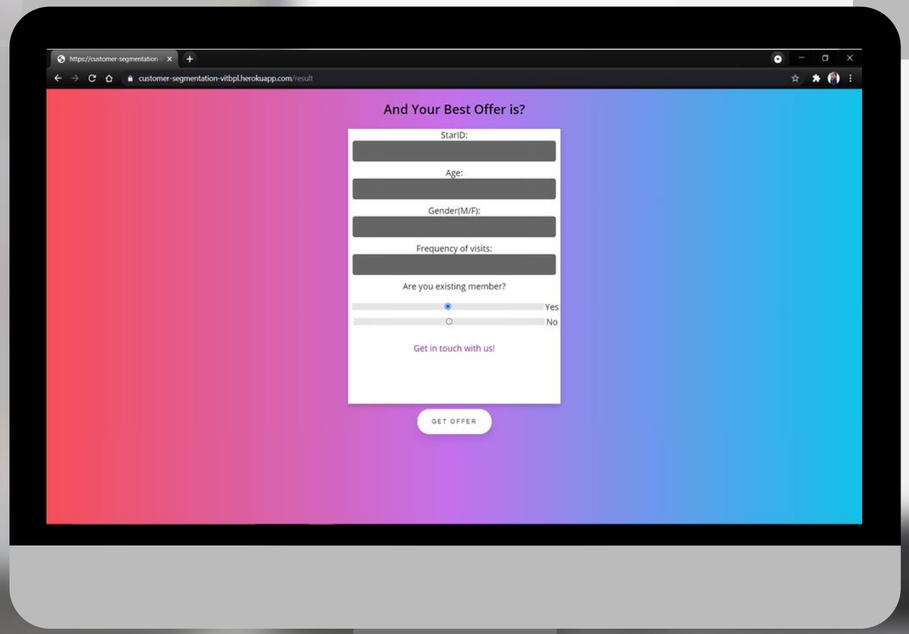
Snapshots

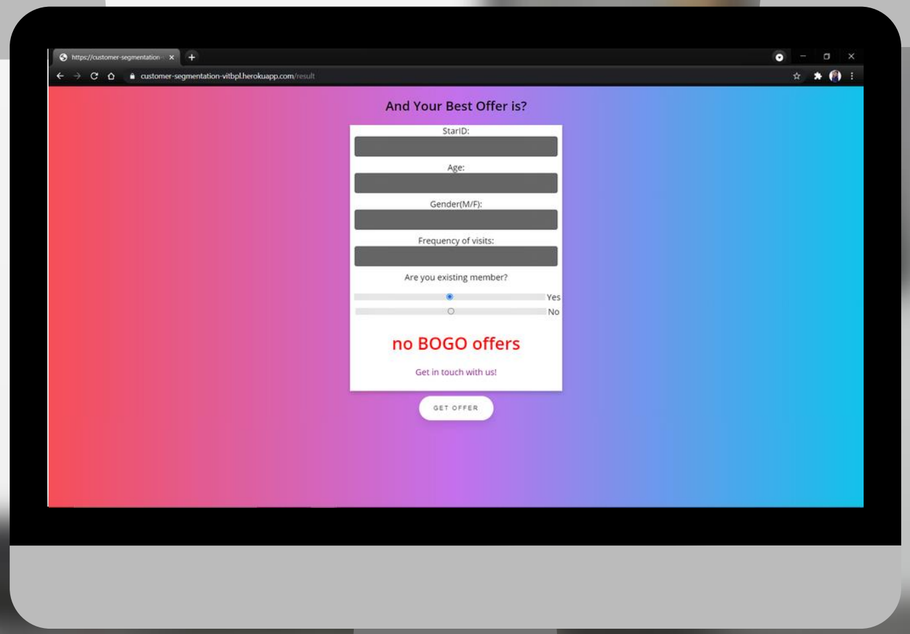

Customers by cluster
160000
140000
120000
100000
80000
60000
40000
20000
0
count
0
1
2
3
cluster


customer-segmentation-vitbpl.herokuapp.com/result
And Your Best Offer is?
STARBUCKS
GET OFFER

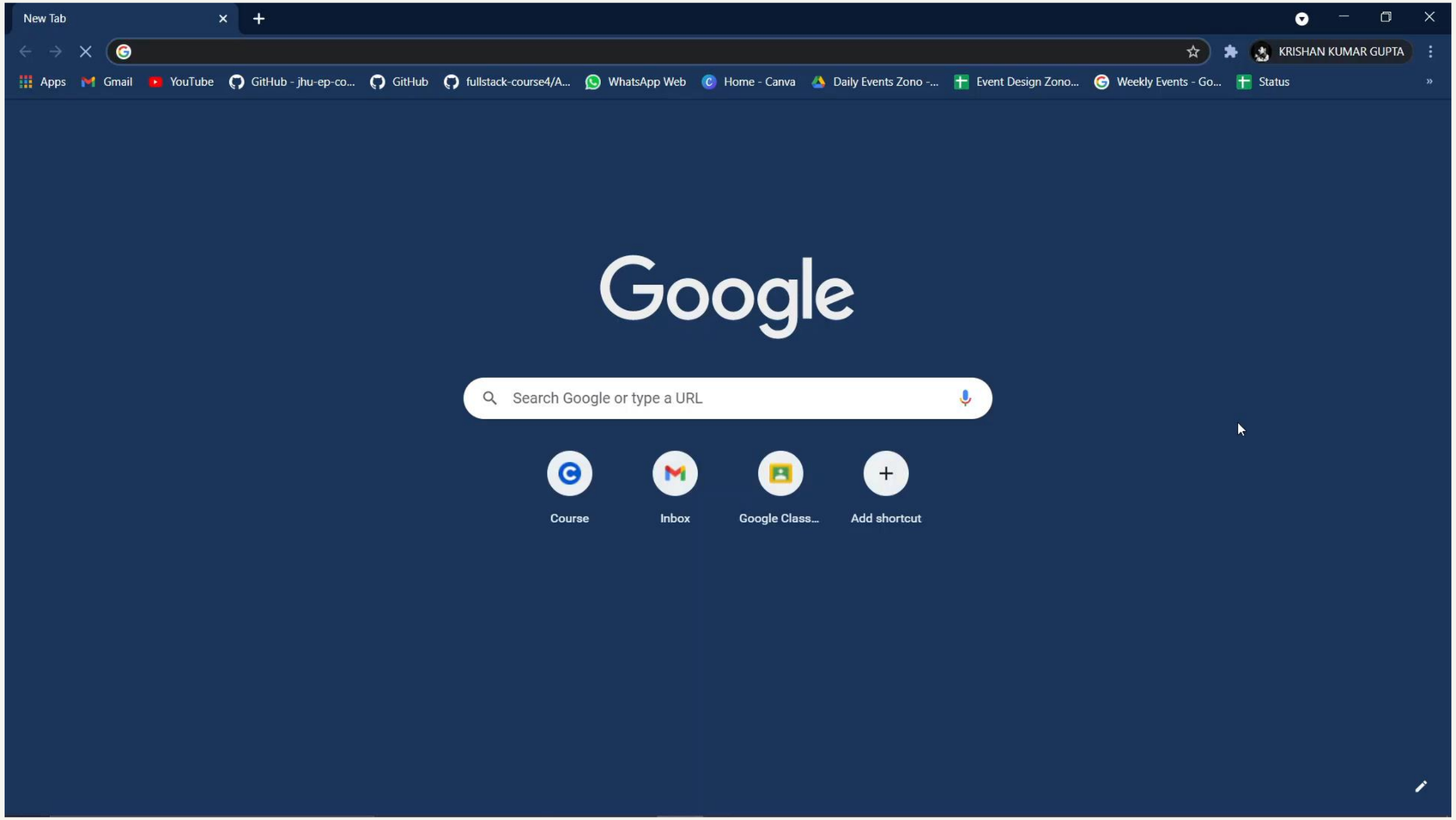

customer-segmentation-vitbpl.herokuapp.com/result
And Your Best Offer is?
StarID:
22221685
Age:
22
Gender(M/F):
M
Frequency of visits:
15
Are you existing member?
Yes
No
Get in touch with us!
GET OFFER


customer-segmentation-vitbpl.herokuapp.com/result
And Your Best Offer is?
StarID:
22221741
Age:
19
Gender(M/F):
F
Frequency of visits:
2
Are you existing member?
Yes
No
Regular offers
Get in touch with us!
GET OFFER


customer-segmentation-vitbpl.herokuapp.com/result
And Your Best Offer is?
StarID:
Age:
Gender(M/F):
Frequency of visits:
Are you existing member?
Yes
No
Get in touch with us!
GET OFFER


customer-segmentation-vitbpl.herokuapp.com/result
And Your Best Offer is?
StarID:
Age:
Gender(M/F):
Frequency of visits:
Are you existing member?
Yes
No
no BOGO offers
Get in touch with us!
GET OFFER

Demo Video



You can view our web app through this link:
https://customer-segmentation-vitbpl.herokuapp.com/result

Conclusion

Reflection

Having segmented data into 4 segments, we now have better insight into Starbucks rewards user base. The clustering results would allow the company to better target audience with tailored offers in the next marketing campaign.

To arrive at this solution, we performed the full analysis cycle - cleaning and preprocessing the data, dealing with missing values, feature engineering, feature scaling, one hot encoding, dimensionality reduction and clustering.

We also wrote a number of functions to generate the clustering results automatically, which allowed some quick experimentation.

While working on this project, We found that data preprocessing consumed a lot of time and was particularly challenging in this case because the event log didn’t account for particular marketing needs.

As a marketer, for instance, We would not like when my response rates would be contaminated by offers viewed after the offer expiration date.

Similarly, we would not like to waste marketing budget on customers that would buy the product even without promotional offer or earn rewards even when not viewing the offers.

By imposing conditions on when offers should be considered as properly “viewed” or “completed”, we managed to fix this problem with original records.

In the future, however, it would be advisable for Starbucks to implement a different tracking strategy as, for instance, by putting a reference code in the campaign message and asking the consumers to mention the code in order to receive the reward.

Improvement

One of the possible ways to improve the clustering results is to predict the missing age, income and gender values instead of simply imputing them.

In this case, we would use the supervised machine learning, experimenting with different models like RandomForest, AdaBoost, etc.

Another useful improvement would be to use the clustering results as labels in supervised modeling to actually predict customer's probability of completing the offer.

This would be a nice practical application that would allow extension on new customers and so would assist in executing a successful marketing campaign in the future.

Thank You!


TIDINGS OF
COFFEE & JOY
CÉLÉBRONS
LE CAFÉ ET LA JOIE
STARBUCKS