Customer Segmentation Using Machine Learning in Python

A PROJECT REPORT

Submitted by

Krishan Kumar Gupta (19BAI10114)
Himesh Sharma (19BAI10125)
Harsh Pal **(**19BAI10127**)**
Adwitiya Dubey **(**19BAI10176**)**

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Specialization in

Artificial intelligence and machine learning

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

VIT BHOPAL UNIVERSITY

KOTHRIKALAN, SEHORE
MADHYA PRADESH – 466114

MAY 2021

VIT BHOPAL UNIVERSITY, KOTHRIKALAN, SEHORE MADHYA PRADESH – 466114

BONAFIDE CERTIFICATE

Certified that this project report titled **"Customer Segmentation Using Machine Learning in Python"** is the bonafide work of "**Krishan Kumar Gupta (19BAI10114), Himesh Sharma (19BAI10125), Harsh Pal (19BAI10127), Adwitiya Dubey (19BAI10176)",** who carried out the project work under my supervision. Certified further that to the best of my knowledge the work reported here does not form part of any other project / research work on the basis of which a degree or award was conferred on an earlier occasion on this or any other candidate.

PROGRAM CHAIR
Dr S Sountharrajan, Programme Chair
B.Tech CSE Spl in Artificial Intelligence and Machine Learning
School of AI &ML division
VIT BHOPAL UNIVERSITY

PROJECT GUIDE
Nilamadhab Mishra
School of AI &ML division
VIT BHOPAL UNIVERSITY

The Project Viva-Voce Examination is held on 05/05/2021

INTERNAL EXAMINER

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

First and foremost, I would like to thank the Lord Almighty for His presence and immense blessings throughout the project work.

I wish to express my heartfelt gratitude to **Dr. Manas Kumar Mishra**, Head of the Department, School of Computer Science and Engineering for much of his valuable support encouragement in carrying out this work.

I would like to thank my internal guide **Mr. Nilamadhab Mishra**, for continually guiding and actively participating in my project, giving valuable suggestions to complete the project work.

I would like to thank all the technical and teaching staff of the School of Computer Science and Engineering, who extended directly or indirectly all support. Last, but not the least, I am deeply indebted to my parents who have been the greatest support while I worked day and night for the project to make it a success.

TABLE OF CONTENTS

ABSTRACT

We members have decided to create a web-based app which will be helpful in segregating customers according to various parameters and providing them with offers that the customer can avail.

The company has to keep making sure it targets the right customers at the right time through their customer journey. It has a diverse customer base and understanding patterns of customer behavior is a crucial need for the business.

Starbucks uses demographic segmentation (markets by age, gender, income, ethnic background, and family life cycle) as well as geographic segmentation (markets by region of a country or the world, market size, market density, or climate).

A small startup can afford to target users based on broad-stroke rules and rough demographics.
Once a company grows to the size of Starbucks, with millions of daily customers, and $1.6B in credit stored on loyalty cards, they have got to graduate to a more sophisticated method to target their marketing.

One such approach, cluster analysis, uses mathematical models to discover groups of similar customers based on variations in their demographics, purchasing habits, and other characteristics.

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).

Cluster analysis, uses mathematical models to discover groups of similar customers based on variations in their demographics, purchasing habits, and other characteristics. We will use k-means unsupervised Machine Learning algorithm to group these customers into clusters that can be used to automate an effective outreach campaign.

LITERATURE SURVEY

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.

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.

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.

Clustering Data

- Clustering is the method of categorizing data into groups based on commonalities.

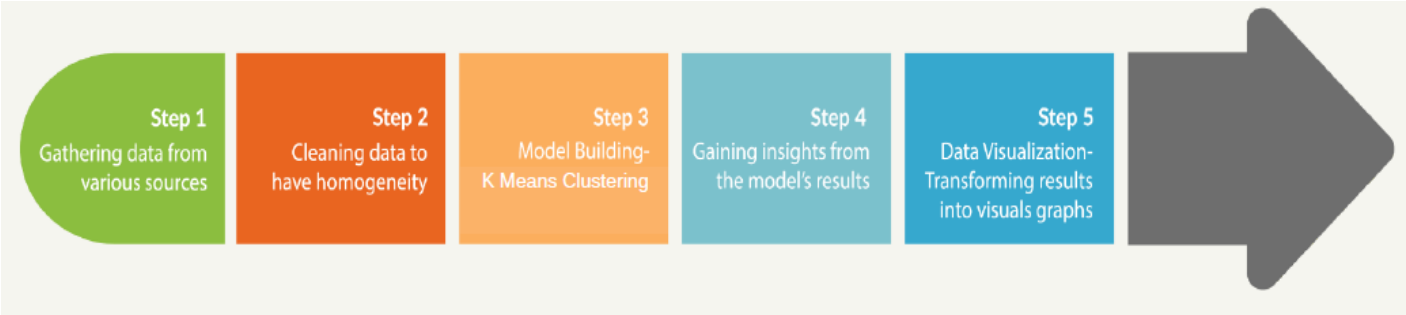
- There are a number of algorithms that can be used on datasets based on the given condition. However, since there is no universal clustering algorithm, it is critical to choose the necessary clustering techniques.

K-Means Clustering

- An algorithm with a K value is one of the most widely used classification algorithms. This clustering algorithm is based on centroid, 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



Step 1: Gathering data from various sources.

Step 2: Cleaning data and preparing the data for our model.

Step 3: Building of the K-Means clustering model.

Step 4: Gathering insights from the model results.

Step 5: Visualizing the data into visual graphs.

PROJECT PROCEDURE

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 i.e., 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)

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 reorganize the transaction details, which was very disorganized.

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.

WORK DONE

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.

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.

LIMITATIONS

- Cannot predict missing values in the data.
- Cannot predict if the customer will avail the offer or not.
- New offers can be introduced to give a variety of offers to the user, hence gaining more attraction.

IMPROVEMENTS

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, the supervised machine learning is used, experimenting with different models like Random Forest, 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.

OBSERVATION

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", I 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.

CONCLUSION

The web app has a minimal interface where the user has the option to enter his/her details. After clicking the get offer button the offer that the user in entitled to will be displayed on the screen. Also, there is a help link given so that if the user faces any problem, he/she can get the problems rectified.

REFRENCES

- https://www.google.com/imghp?hl=EN
- https://www.wikipedia.org/
- github.com
- https://www.youtube.com/
- heroku.com