Food Delivery Case Study

FINAL PROJECT

MIS64060: Fundamentals of Machine Learning

Professor:Murali Shanker

Kent State University

Sai Venkata Aravind Bandhanadam

Abstract:

In this case study,the customer data of a food delivery company is examined.The goal is to extract insights from the data using machine learning techniques.

Contents:

1.Introduction:

To focus on the early consumer identification and awareness of what terms that are crucial to them when placing an order, so that they may simply target customers based on their ordering preferences. In this food delivery case study the order frequency, value and the average delivery time of the customers is calculated by using the k-means clustering algorithm method. I'm predicting the need of average delivery time when orders are being placed by the customers based on the timeline of 4 weeks.

This dataset is extracted from the Kaggle website.
https://www.kaggle.com/asaumya/k-means-clustering-food-delivery/data

The first restaurant in the United States that started food delivery service is World Wide waiter. It was started in 1995 and is still operating today as waiter.com, Apart from that DoorDash,Grubhub,and UberEats are the top three food delivery services in the United states that account for around 80% of the sector's sales.

As a result of the pandemic,numerous restaurants have stopped serving clients in store in order to comply with mandatory safety procedures and have shifted their focus to food delivery.As a result food delivery companies experienced a boom in demand, While the demand for food delivery services has increased the issues that come with it have increased as well.

According to this case study the online food delivery systems set up a food menu online and customers can easily place the order as per they like.Also they can easily track the orders.The management maintains the customer database,to improve the food delivery service,
The main problem of this case study is to Increase the Customers based on their order frequency and also predict the active customers and their preferences.

2.Data Exploration:

As indicated in the table below ,the customers dataset has 10000 observations and 11 variables.we can tell that the majority of the variables are numeric using the "str" function.The % sign on the other hand ,causes some percentage variables to be classes as a character.I"ll provide a solution to this problem in the next section.I renamed the customers data in to (overview_of_customer_orders_raw)

```
# To find out how many rows and columns there are in total.
 dim(overview_of_customer_orders_raw)
```

```
[1]   10000  11
```

```
'data.frame':   10000 obs. of  11 variables:
 $ Cust_Id                              : int  1269647 167631 301524 1268254 357161 1294857 387095 785080 1288527 1111111 ...
 $ Time_for_the_first_order             : chr  "6/29/15 10:57" "7/4/15 15:39" "6/26/15 9:56" "7/1/15 1:51" ...
 $ Frequently_Order_DateTime            : chr  "12/10/15 2:18" "12/15/15 14:42" "12/9/15 20:45" "12/14/15 1:43" ...
 $ All_of_the_orders                    : int  212 211 189 184 182 171 168 160 160 158 ...
 $ Last_7_Days_orders                   : int  6 8 9 6 4 8 13 NA 7 1 ...
 $ In_the_Last_4_weeks_orders           : int  43 19 33 37 23 27 43 25 40 28 ...
 $ Total_Amount                         : int  138808 56404 36020 32489 85150 55597 19055 39588 4343 15279 ...
 $ Amount_in_the_Last_7_days            : int  4291 1925 1772 975 1738 1710 1231 0 215 94 ...
 $ Amount_during_theLast_4_weeks        : int  26853 4177 6404 7110 9958 8436 4014 6705 1060 3336 ...
 $ Distance_Fromthe_Resturant_on_Average: num  1.6 2.2 2.5 3.1 2.4 1.6 2.1 1.8 2.1 2.1 ...
 $ Typically_DeliveryTime               : int  51 42 57 55 36 31 48 16 49 54 ...
```

By seeing the data I Am assuming that the average distance from the restaurant and the average delivery time apply to all of the customer's orders ,and also we observed that DateTime is in Factor format. we will convert it to Date format in the following steps.

Now we have to look into the summary of the data to determine how the data is distributed and see if it has any missing values.

#To find out the summary we use
Summary(overview_of_customer_orders_raw)

```
   Cust_Id        Time_for_the_first_order Frequently_Order_DateTime
Min.   :     28   Length:10000            Length:10000
1st Qu.: 336515   Class :character        Class :character
Median : 668340   Mode  :character        Mode  :character
Mean   : 671402
3rd Qu.:1005002
Max.   :1355445

All_of_the_orders Last_7_Days_orders In_the_Last_4_weeks_orders  Total_Amount
Min.   :  1.000   Min.   : 1.000     Min.   : 1.000             Min.   :     1
1st Qu.:  1.000   1st Qu.: 1.000     1st Qu.: 1.000             1st Qu.:   279
Median :  2.000   Median : 1.000     Median : 2.000             Median :   688
Mean   :  7.006   Mean   : 1.735     Mean   : 3.198             Mean   :  2253
3rd Qu.:  7.000   3rd Qu.: 2.000     3rd Qu.: 4.000             3rd Qu.:  2040
Max.   :212.000   Max.   :14.000     Max.   :46.000             Max.   :138808
                  NA's   :8077       NA's   :5659
Amount_in_the_Last_7_days Amount_during_theLast_4_weeks
Min.   :    0.0           Min.   :    0.0
1st Qu.:    0.0           1st Qu.:    0.0
Median :    0.0           Median :    0.0
Mean   :  109.5           Mean   :  455.5
3rd Qu.:    0.0           3rd Qu.:  398.0
Max.   :10150.0           Max.   :26853.0

Distance_Fromthe_Resturant_on_Average Typically_DeliveryTime
Min.   :-0.800                        Min.   :15.00
1st Qu.: 1.700                        1st Qu.:26.00
Median : 2.400                        Median :36.50
Mean   : 2.356                        Mean   :36.91
3rd Qu.: 3.025                        3rd Qu.:47.00
Max.   : 5.900                        Max.   :83.00
```

From the above data we observed three issues arises they are At first many users,in the last 7 days and 4 weeks orders are missing.We will dig deeper in to it in the future section.The second the Average distance from the restaurant is not good the next one is All users who have placed at least one order are included in the data set

II. Data cleaning and creation of new columns:

I modified the date format, and because we don't have details for all the orders,I eliminated the time data for the first and last order. Then over the last 7 days and 4 weeks I filtered by cases where the order value was NA and from those users,I removed the users with the shortest time since their last order.

```
[1] "For Users who had NA value in last 7 Days orders , the minimum value for Recent Order placed is 9Days"
[1] "For users who had NA value in last 4 Week orders,the minimum value for Recent Order placed is30Days"
```

From the above output we can observe that for recent orders,the minimum days are more than 7 days and 28 days.As a result,we may assume that the NA values are not missing but are ,in fact zero.As a result, I'm going to replace them with 0 .

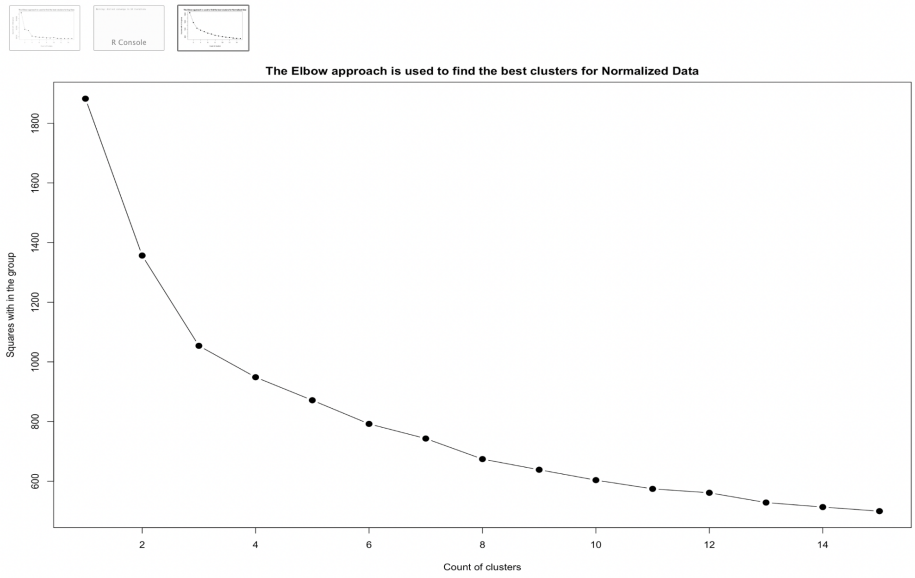
And also in 44 situations,the average distance from the restaurant is negative; for better understanding I refer to them as 0.In addition I established an Average order value (Av)Column,which will be used in place of the overall order value.

III. K-Means Clustering Algorithm

K-means clustering is used for high-dimensional data and DBSCAN cannot be used for my dataset. Since, it contains high dimensional data and hierarchical clustering is not used because it incorporates computational data.

K means clustering is an unsupervised technique for organizing enormous amount of retail data and generating competitive insights for our company..It used to find the assumptions in the business to find the unidentified clusters in large data sets Here we consider the food delivery was considered to be business-firm so i consider k means algorithm for analyzing the data.

The K-means algorithm is used in this data set to construct a clustering model in order to see whether we can separate users into distinct buckets.The summary data has variables at different scales , so that i decided to use elbow method to calculate the minimum error and to decide how many the data is divided into



We can see from the charts above that using the normal distribution of data make more sense.Because the total of squares within a group is on a considerably smaller size for it.A side from that ,it appears that 3 or 4 clusters is the ideal amount, as the incremental gain after that is modest I will start with three clusters now.

Created the three clusters with k means to denormalize the data and also

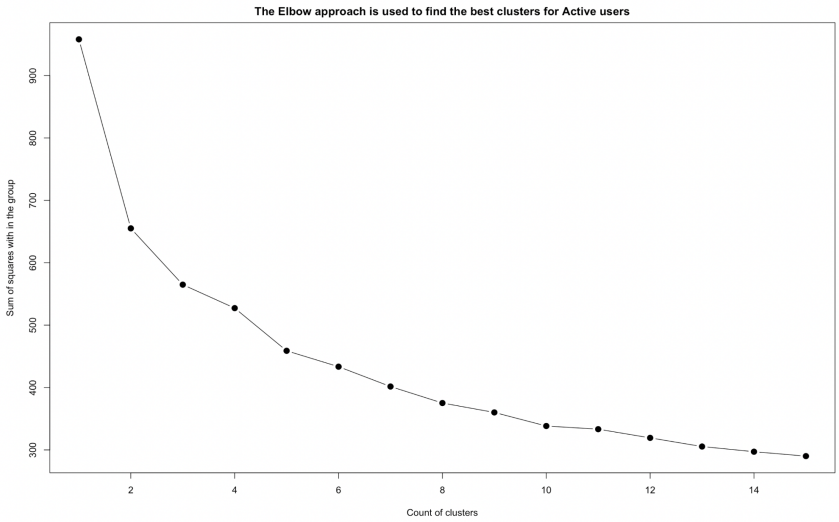
To find out the means of each variable of every cluster.

```
[1] "Below is the mean value of all variables in each cluster"
  All_of_the_orders Last_7_Days_orders In_the_Last_4_weeks_orders
1          2.797698          0.3056266                   1.204859
2          2.513654          0.0000000                   0.000000
3         17.050763          0.7103517                   3.042468
  Distance_Fromthe_Resturant_on_Average Typically_DeliveryTime
1                             2.355601               36.06829
2                             2.378023               38.51625
3                             2.338056               36.37691
  countdown_to_the_Last_Order Days_since_Initial_Order   Av_All Av_Last_7_Days
1                    56.89335                 74.20307 377.2760       73.37749
2                   142.50163                159.91125 360.9083        0.00000
3                    47.51924                173.80425 336.6768      124.55076
  Av_Last_4_Weeks
1        200.5054
2          0.0000
3        255.7472
[1] "The following table shows the number of customers in each cluster"
[1] 3910 3076 3014
```

We can observe from the mean values of each variable in each of the three clusters that transaction frequency and AV are the most variable, whereas typical delivery time and distance on average are not.

Classification of customers:
1.Over the last 7 days and 4 weeks customers are active but with Low frequency and Low Av.
2.Over the last 7 days and 4 weeks ,customers are active with high frequency and High Av.
3.Customers are not in active since last 7 days and over 4 weeks

Although fifty seven percent of users have not transacted in the last 4 weeks,the size of cluster 3 does not support that. So in order to improve clustering for the active users i created a cluster for the active users by excluding the customers who are not interacted in last 4 weeks



In the 3 and 4 appear to be the best options, so moved forward with 4 to find any improvement in the user segments

Make a closer look at the outcomes.

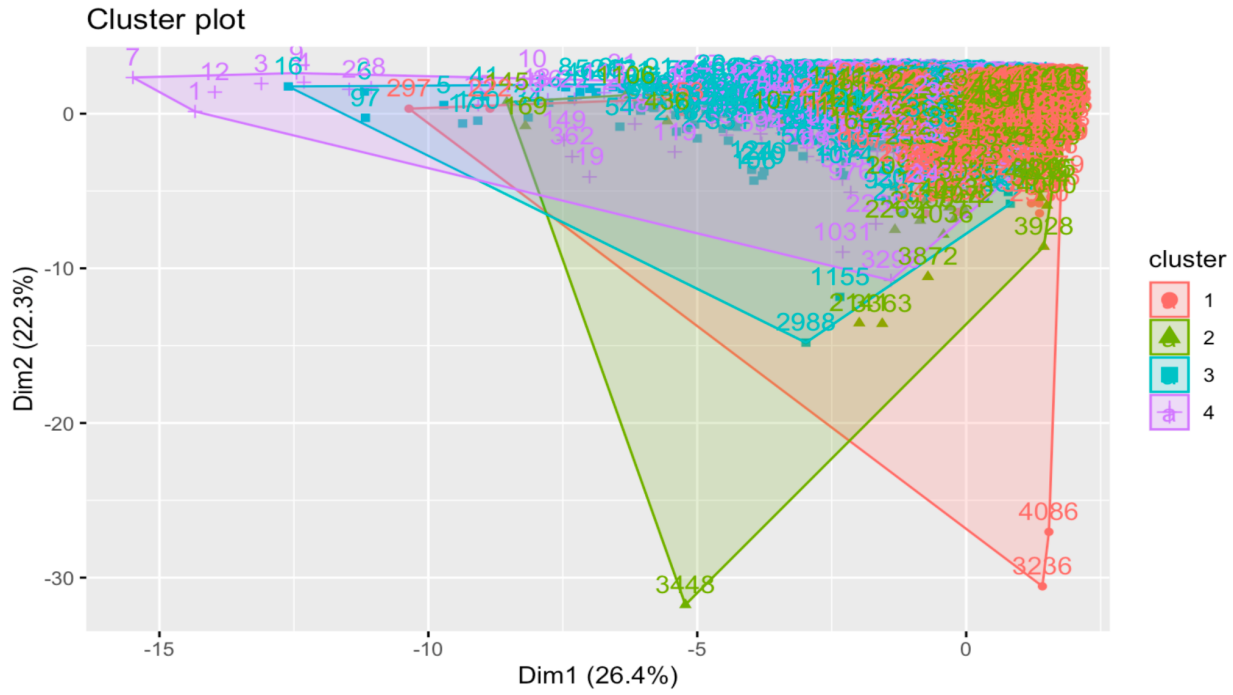
```
  All_of_the_orders Last_7_Days_orders In_the_Last_4_weeks_orders
1         19.563269          0.9521090                   4.152451
2          3.474092          0.5661017                   2.247953
  Distance_Fromthe_Resturant_on_Average Typically_DeliveryTime
1                             2.338752               36.80888
2                             2.385908               35.70654
  countdown_to_the_Last_Order Days_since_Initial_Order   Av_All Av_Last_7_Days
1                    38.74868                173.07865 338.9868       167.4262
2                    38.88475                 61.14092 366.1361       136.1937
  Av_Last_4_Weeks
1        345.9350
2        371.6465
[1] 2276 2065
[1] "Below is the mean value of all variables in each cluster"
  All_of_the_orders Last_7_Days_orders In_the_Last_4_weeks_orders
1          3.257091          0.5141812                   2.139128
2          3.699584          0.6122661                   2.314438
3         19.138718          1.0131694                   4.219237
4         19.862249          0.8988666                   4.117466
  Distance_Fromthe_Resturant_on_Average Typically_DeliveryTime
1                             2.198994               25.93779
2                             2.587838               47.22349
3                             2.302107               25.57243
4                             2.384307               47.60680
  countdown_to_the_Last_Order Days_since_Initial_Order   Av_All Av_Last_7_Days
1                    40.36597                 61.06587 361.3715       119.7832
2                    37.48649                 60.74428 367.8285       154.9158
3                    37.75417                170.86392 344.2511       172.5979
4                    39.49869                174.70619 337.1168       161.9538
  Av_Last_4_Weeks
1        365.2315
2        377.2214
3        349.2529
4        344.3017
[1] "The following table shows the number of customers in each cluster"
[1] 1093  962 1139 1147
```

1.cluster(1) has a low frequency , a high Average order value ,a long delivery time,and a higher distance from the restaurant.
2.cluster(2) features a high frequency of users,a poor Average order value and long delivery time and distance from the restaurant is lower
3.cluster(3) has a low frequency ,a low Average order value particularly in the recent 7 days,and a low delivery time, and a lower distance from the restaurant
4.cluster (4) has high frequency , a high Average order value and a low delivery time

So in addition to order frequency and average order value we are seeing delivery time are to a lesser extent, distance from the restaurant emerge as important variables in creating clusters. And that delivery time and restaurant distance are not directly proportional,with more data we can find so this is the case .

At last consider the 4 cluster plot.

[1] “The visualization of clusters on 4 Clusters is shown in the graph below”



From the above graph it is difficult to understand the information,but we can see how each cluster has its own different boundaries on the x and y axis, which are distinguishing qualities that divide the users into different groups.

As a consequence of my investigation ,i found that when customers place an order they think about delivery time and options for restaurants that offer quicker delivery .Explanation of the above -mentioned statement in detail the cluster 1 and cluster 2 that the average delivery time is not important to the customers and i can also display the restaurants which are at long in distance,and from the cluster 3 and cluster 4 i should provide more choices for quicker delivery.

Conclusion:

I looked at Food Delivery customer data in the previous analysis .I noticed that the order frequency and values as crucial indicators to cluster users when i looked at all users together, but because of a large portion of them not being active,i only had a look at those who transacted in the previous four weeks to better understand them. Apart from the above sentences I also discovered that Average delivery time was another important element for the future users.This investigation can be used to work on early consumer identification and understanding the importance to them while placing the orders which help to better target the customers. For better explanation i can consider the cluster 1 and cluster 2 i find that the average delivery time is not important to the customers and i can also display the restaurants which are at long in distance,and from the cluster 3 and cluster 4 i should provide more choices for quicker delivery.