Chenyang Yan cy477
Yutong Wang yw2364
Hongyi Nie hn327

ORIE 5250: Final Report

1. Introduction

For the second dataset "Expedia", we defined two types of customers based on whether the customer wants to make an early or a late reservation, but we hope to divide the customer into more segments by using clustering, to find the optimal assortment to maximize the expected revenue.

2. Description

We plan to further define different types of customers, using K-means techniques, to better provide the optimal subsets of hotels to display. For different clusters of customers, we will use the mixture of MNL models to estimate the probability of each customer choosing hotels and sensitivity parameters for each type of customer using MLE estimation. We will calculate the revenue for each type of customer to decide the optimal subsets to display, assuming that we know the customer type. In the meantime, we will also calculate the revenue under a single MLE model, assuming that we do not know the customer type. Compare the mixture MNL and the single MNL model result.

3. Main

Firstly, we wrote an MLE algorithm to estimate beta based on following model:

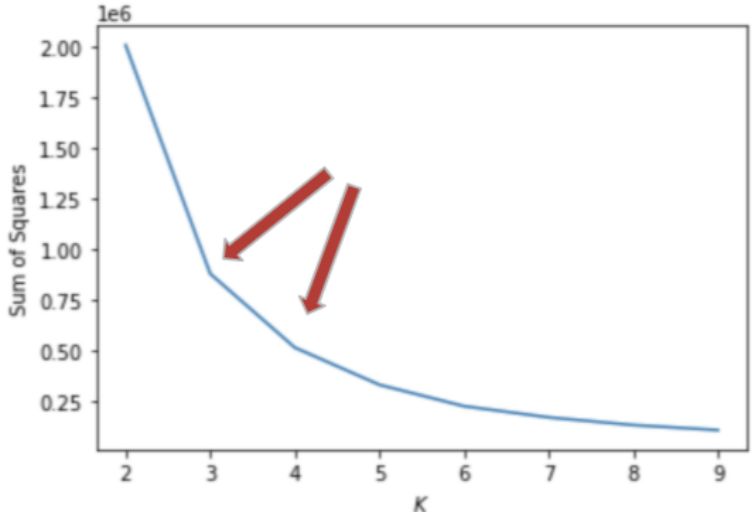
$$Max\ beta \sum_{t=1}^{T} logP(j_t|S_t)$$

$$P(j_t|S_t) = \frac{v_j}{1+\sum_{p\epsilon S} v_p}$$

Now we have the beta_original = [-1.74662743, 0.40812391, 0.10876074, 0.10138235, 0.02202491, 0.04344406, -0.06686938, -1.33110656, 0.15977641].

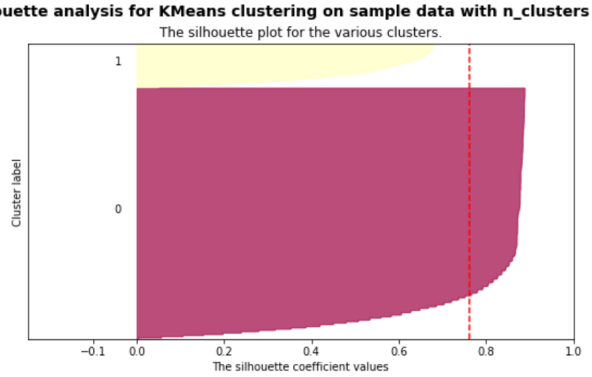
Next, we use the K-mean algorithm (sklearn.cluster.KMeans) to determine which n is the best to use, ranging from 2 to 10. We run the K-mean algorithm using different N, and evaluate the outcome of each model with The Elbow Method and The Silhouette Method.

In The Elbow Method, we look at the sum-of-squares error in each cluster against N(K in the graph.) We compute the distance of each data point to the center of the cluster of its assignment. After plotting a graph (shown below) with N-values on the x-axis and sum-of-squares error score on the y-axis, we found the elbow points are N=3 and N = 4.

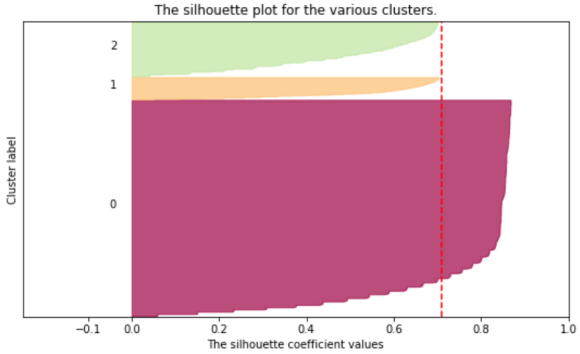


To further determine the value for N, we then evaluate the clusterings using the Silhouette Method, which measures how well each datapoint "fits" its assigned cluster and also how poorly it fits into other clusters. The graph is shown below:

Silhouette analysis for KMeans clustering on sample data with n_clusters = 2

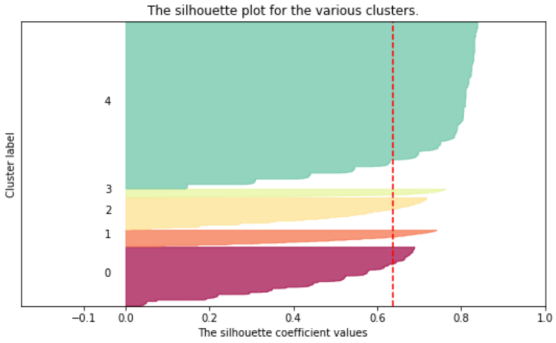


Silhouette analysis for KMeans clustering on sample data with n_clusters = 3

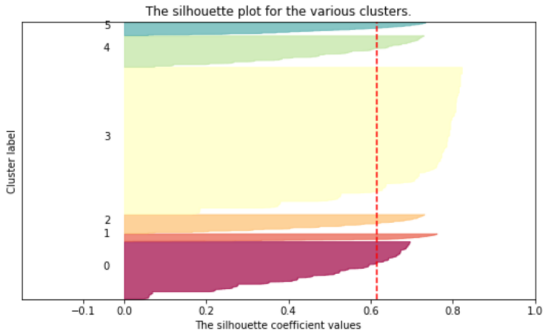


Silhouette analysis for KMeans clustering on sample data with n_clusters = 4

Silhouette analysis for KMeans clustering on sample data with n_clusters = 5

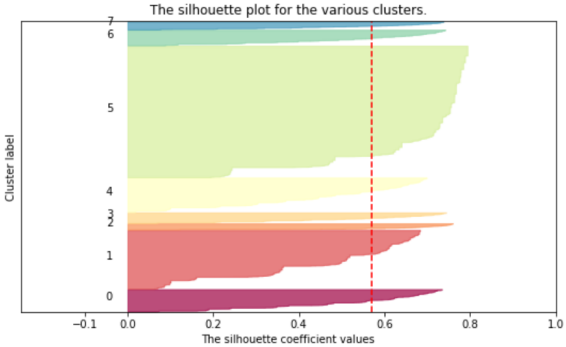


Silhouette analysis for KMeans clustering on sample data with n_clusters = 6

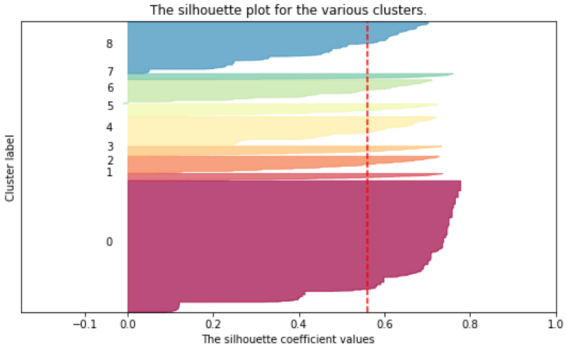


Silhouette analysis for KMeans clustering on sample data with n_clusters = 7

Silhouette analysis for KMeans clustering on sample data with n_clusters = 8



Silhouette analysis for KMeans clustering on sample data with n_clusters = 9



From observing the graphs above, we can see that 4 is the best performing number of clusters. N =2 and N = 3 are relatively bad picks since there are clusters with below average silhouette scores and wide fluctuations in the size of the silhouette plots. N>4 performs better under The Silhouette Score but is not desirable under The Elbow Method. Thus we choose 4 as the number of clusters to run the K-means model.

- **K means Clustering, N = 4**

After we determined n = 4 is the best n for k means, we Let θ_n be the probability that a customer belongs to type n, where n is in {1, 2, 3, 4}.

Let v_{jk} denotes the preference weight of hotel j for customer type k, and

$v_{jk} = e^{u_{jk}}$, $u_{jk} = \beta_{ok} + \sum_{i=1}^{8} \beta_{ik} x_{ji}$, where β_{ik} is the sensitivity of type k customers to feature i.

The probability of customer choosing hotel j given a set of hotels S is:

$$P(j|S) = \sum_{k=1}^{4} \theta_k \frac{v_{jk}}{1+\sum_{p \in S} v_{pk}}$$

Now we have the following results:

Beta = [[-1.59550736,0.46089075, 0.11816455, 0.11046467, 0.0717849, 0.03798292, -0.05805397, -1.03614344, 0.17266994],

[-1.94242923,0.17459783, 0.06103661, 0.01711175, -0.14394501, 0.02369744, -0.14903244, -0.01516122, 0.14749104],

[-2.20624549, 0.338743861, 0.11386014, 0.137717589, -0.00375356203, 0.0376205415, -0.0896334726, -4.35882283, 0.0948561237],

[-2.44854238,0.46847782, 0.13352973, 0.07686038, -0.03170865, 0.06867889, -0.07128673, -4.84973152, 0.13216368]]

Theta = [0.663078642432798,0.04850041500826749,0.09439967583606193,0.1940212667228725]

Then, assuming we don't know the type of an arriving customer, we can calculate the optimal revenue by using integer programming, and get the expected revenue for dataset one to four by using the following model:

$$\text{Max} \sum_{k=1}^{K} \theta_k Z_k$$

$$Z_k + \sum_{j=1}^{n} v_{jk} y_{jk} = \sum_{j=1}^{n} n_j v_{jk} x_j \quad \forall k$$

$$0 \leq y_{jk} \leq M x_j \; x_j \quad \forall k$$

$$-M(1 - x_j) + Z_k \leq y_{jk} \quad \forall j \forall k$$

$$xj \in \{0, 1]\}$$

The results are shown below:

Data1: 107.3396340843958

Data2: 130.5562106874302

Data3: 120.9612284936689

Data4: 97.4088309422418

Suppose we know that the arriving customer is one of the four types, then we can use the beta and theta we just got to calculate the optimal assortment and maximum revenue:

For dataset1:

Revenue for Type 1: 76.09948094662586, choose the assortment of 6 highest price hotels.

Revenue for Type 2: 56.399565981831635, choose the assortment of 4 highest price hotels.

Revenue for Type 3: 17.825977219784924, choose the assortment of 1 highest price hotel.

Revenue for Type 4: 14.40635043647111, choose the assortment of 1 highest price hotel.

For dataset2:

Revenue for Type 1: 95.08681622346974, choose the assortment of 3 highest price hotels.

Revenue for Type 2: 67.26774225862488, choose the assortment of 3 highest price hotels.

Revenue for Type 3: 79.24615991027659, choose the assortment of 3 highest price hotels.

Revenue for Type 4: 73.5583729290145, choose the assortment of 3 highest price hotels.

For dataset3:

Revenue for Type 1: 103.04253226744395, choose the assortment of 9 highest price hotels.

Revenue for Type 2: 50.08825158709839, choose the assortment of 3 highest price hotels.

Revenue for Type 3: 52.0725866801742, choose the assortment of 3 highest price hotels.

Revenue for Type 4: 55.262678799566295, choose the assortment of 4 highest price hotels.

For dataset4:

Revenue for Type 1: 37.166846933375595, choose the assortment of 2 highest price hotels.

Revenue for Type 2: 17.896870188359003, choose the assortment of 1 highest price hotel.

Revenue for Type 3: 23.0126797928701, choose the assortment of 1 highest price hotel.

Revenue for Type 4: 24.956915436056192, choose the assortment of 1 highest price hotel.

- **K means Clustering, N = 3**

Out of curiosity and in order to check the difference between clusters of three and four, we did the same process for n = 3 and got the following results:

Beta = [[-2.08215274, 0.42792156, 0.12208375, 0.11702509, 0.00999331, 0.04675314, -0.06323514, -0.99569225, 0.1493836],
[2.02049388, 0.58806615, 0.12565736, 0.11863238, 0.32397264, 0.0330871, 0.2062005, -1.05314573, 0.18988912],
[1.83949155, 0.28127776, 0.10843109, 0.01705466, 0.16360347, 0.01680751, -0.24775001, -0.36514211, 0.19182447]]

Theta = [0.8274742008640015, 0.1233326144213739, 0.04919318471462463]

Then, assuming we don't know the type of an arriving customer, we can calculate the optimal revenue by using integer programming, and get the expected revenue for dataset one to three. The results are shown below:

Data1: 107.29109444913811

Data2: 131.11344212920363

Data3: 121.01835939563993

Data4: 97.4088309422418

Suppose we know that the arriving customer is one of the three types, then we can use the beta and theta we just got to calculate the optimal assortment and maximum revenue:

Data1:

Revenue for Type 1: 67.21104633429165, choose the assortment of 7 highest price hotels.

Revenue for Type 2: 116.46530195541274, choose the assortment of 1 highest price hotel.

Revenue for Type 3: 118.61326728080468, choose the assortment of 1 highest price hotel.

Data2:

Revenue for Type 1: 71.30801404477035, choose the assortment of 3 highest price hotels.

Revenue for Type 2: 215.10739040432304, choose the assortment of 1 highest price hotel.

Revenue for Type 3: 225.19728779793326, choose the assortment of 1 highest price hotel.

Data3:

Revenue for Type 1: 84.39048093239713, choose the assortment of 9 highest price hotels.

Revenue for Type 2: 146.67162852755308, choose the assortment of 2 highest price hotels.

Revenue for Type 3: 119.96921028047899, choose the assortment of 1 highest price hotel.

Data4:

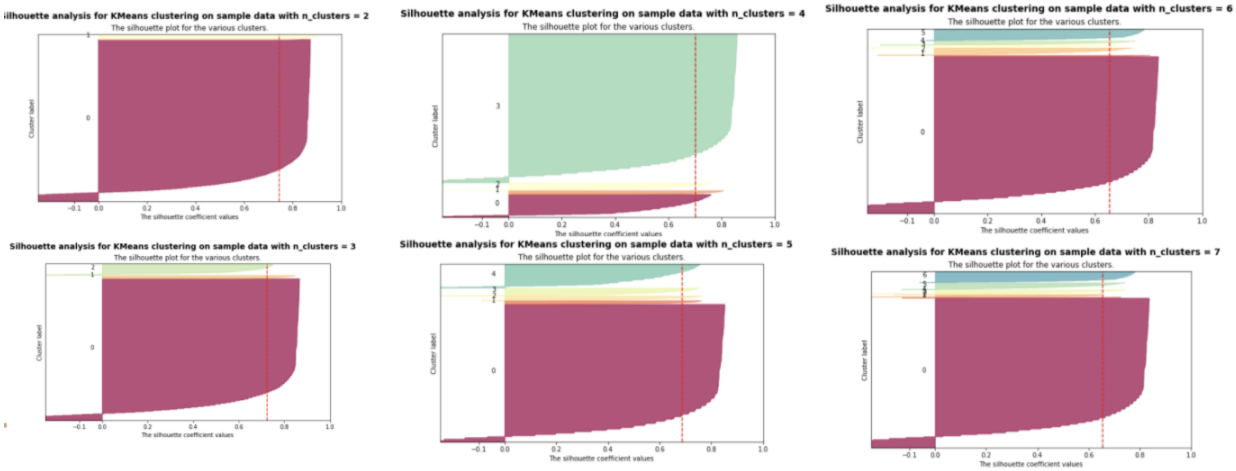
Revenue for Type 1: 27.69700493238335, choose the assortment of 2 highest price hotels.

Revenue for Type 2: 68.44321534637793, choose the assortment of 1 highest price hotel.

Revenue for Type 3: 68.72451466090337, choose the assortment of 1 highest price hotel.

- **Spectral Clustering, N = 4**

We also used Spectral Clustering to define customer types. Since this method does not compute any centers of clusters, we can only use The Silhouette Method to evaluate the performance of models with different numbers of clusters, where we choose 4 as the best k to fit the model.



beta = [[-1.59550736, 0.46089075, 0.11816455, 0.11046467, 0.0717849, 0.03798292, -0.05805397, -1.03614344, 0.17266994],

[-1.94242923, 0.17459783, 0.06103661, 0.01711175, -0.14394501, 0.02369744, -0.14903244, -0.01516122, 0.14749104],

[-23.2809247, 0.191737735, 0.0406044319, 0.0373399985, -0.299671576, 0.00459778762,-0.256169928, -75.0579092, 0.058649329],

[-2.44854238, 0.46847782, 0.13352973, 0.07686038, -0.03170865,0.06867889, -0.07128673, -4.84973152, 0.13216368]]

theta = [0.12256141795580652, 0.018926991222738533, 0.040860341548536364, 0.8176512492729185]

The revenues for unknown customer types is shown below:

data 1 = 107.3396340843958

data 2 = 130.5562106874302

data 3 = 120.9612284936689

data 4 = 97.4088309422418

And for each type of customer in dataset one to four the result is following:

data1:

Revenue for Type 1: 1.50.22221267036282 top 3

Revenue for Type 2: 2.21.190809084136927 top 1

Revenue for Type 3: 3.3.048 top 1

Revenue for Type 4: 4.60.73677000846684 top 8

data2:

Revenue for Type 1: 1.95.2280514959225 top 3

Revenue for Type 2: 2.72.2625063412781 top 3

Revenue for Type 3: 3.4.717 top 3

Revenue for Type 4: 4.80.9750431152258 top 7

data3:

Revenue for Type 1: 1.102.53211015470683 top 8

Revenue for Type 2: 2.39.73228698385515 top 2

Revenue for Type 3: 3.8.2158780216051 top 3

Revenue for Type 4: 4.63.03778328790421 top 12

data4:

Revenue for Type 1: 1.36.688574026859406 top 1

Revenue for Type 2: 2.19.756013117429823 top 1

Revenue for Type 3: 3.0.009354698538806955 top 1

Revenue for Type 4: 4.57.48420476680186 top 3

From the result above, we can see that it is very close to the result of k means with N =4. But it costs too much time to run, it's a heritage cons for such a method, especially for such a large dataset. Thus, we find the k means is a better method to use in this case.

4. Conclusion

Vertically comparing two different methods k means and Spectral Clustering when N = 4, we can see that it is very close to the result of k means with N =4. But Spectral Clustering costs too much time to run, it's a heritage cons for such a method, especially for such a large dataset. Thus, we find the k means is a better method to use in this case.

Horizontally comparing the N values, we compared the k means Method with N = 4 and N = 3. Generally N = 3 will bring a higher expected revenue for all datasets when unknown the customer type. Although in The Elbow Method, we look at the sum-of-squares error in each cluster against K, N = 4 appears to be the best match (with the least sum-of-squares error), but it will not necessarily bring the largest expected revenue, because the minimal sum-of-squares error model is just statistically correct but not considered any actual customer behavior, and that showed our algorithms fit the customers actual need better. For example, for N=3, data1 is for type three customers, data2 is for type two and tree, data3 is for type 2 customers, and lastly data4 is for type two and three. N =4, on the other hand, all datasets are for customers two and three.