Gillian Tatreau
DSC680
Professor Iranitalab
4 February 2023

10 Questions from Audience

1. Why should a retailer care about doing this analysis if they have the money to run marketing campaigns and can put out more ads?

   While implementing more ads will generally also increase more revenue, it makes more sense to spend some time tailoring ads to specific customer types. Instead of spending money on generating ads that will not reach the target audience or reaches too broadly without bringing in any revenue, these individualized ads are able to target specific audience types, bringing in more revenue for less overall money spent on ads.

2. What is the purpose of creating these customer profiles that pigeonholes our customers when we should be treating each one as an individual?

   These customer profiles serve as just the basis of future marketing campaigns. They help generalize the customer base so we can look at the customers on several different levels when creating marketing campaigns and make operational decisions. For example, we can make decisions that affect all customers that uses data that spans all transactions or make decisions regarding customers within one cluster and use data just from those transactions, or even zoom in to individual customers within each cluster to see trends and variations within that cluster and how each one might respond or be affected by these decisions.

3. Why were rows with NAs removed when this removed a lot of data?

   Rows with NAs were removed because in general, missing values create problems within machine learning models. Even though deleting these rows appears to remove a lot of data, there is still more than enough that remains. It is better to work with a slightly smaller amount of complete data than a larger amount of data that does not contain the same amount of information for every entry. Any analysis or modeling that is performed with missing data requires the model to either make a guess, or some machine learning model algorithms will ignore that entry anyways.

4. If you took the time to remove some outliers, why were some left in?

   Some of the outliers represent the natural variation in the data- some products sold by the retailer could actually be more expensive than the majority, and these values represent natural outliers in the data. The outliers removed represent errors in processing or entry and are not representative in any way of the data.

5. How do you know that you picked the correct number of clusters?

   The correct number of clusters is determined visually using the elbow method, in which the number of clusters is plotted versus the inertia. The inflection point, or elbow, represents the number of clusters that best divides the data. The model is then retrained with the number of clusters indicated by the elbow method, and the silhouette score is calculated as further validation of the appropriateness of the number of clusters.

6. What purpose does the transaction segmentation serve for the retailer?

   The transaction segmentation gives insight into the different types of transactions that exist within their sales, which could help them make decisions about product pricing.

7. Since the silhouette score for the transaction segmentation is higher than that for the customer segmentation, does that mean that the transaction segmentation is better?

   The silhouette score for the transaction segmentation is slightly higher than that for the customer segmentation, but this means that the 5 clusters of the transaction segmentation are slightly more defined, more decisively different, than those of the customer segmentation; however, the difference is small, and the customer segmentation silhouette score still validates the appropriateness of the 4 clusters in the customer segmentation.

8. If there are four types of customers, how do we determine what kind of customer someone will be?

   k-means clustering is not a predictive machine learning model- it explains patterns that already exist in the data. To determine what kind of customer someone might be, we could build a predictive model using more data, if it becomes available, in future projects.

9. Why shouldn't we allocate most of the marketing budget towards just one customer type, like the 'Big Spenders' of cluster 3?

   Even though 'Big Spenders' represent a large monetary value, they do not shop very often and have not shopped very recently. These customers are likely people that have made a single very expensive purchase. While these customers bring value, their value is not constant or predictable. Meanwhile, the 'Regulars' may make purchases of lower monetary value, but they do so more frequently and recently which provides a more consistent source of revenue. It is not good practice to prioritize any one group and neglect the others.

10. If this is our data from our site why do we have to be concerned about ethical considerations? Why can't we exploit all of the information to our advantage if the data belongs to us?

    Even though we created this data, and it is data from transactions on our website, we do not necessarily own the data- it is still the customers' data that we have collected; therefore, we must make sure that the customers are aware that we are collecting their data and how we intend on using it. We also have an obligation to our customers to protect and respect their data if they give their consent for us to use it. Also, we must be mindful of how we are using the data to make decisions because any decisions made that disproportionately affect, or discriminate against, some groups more than others would be an unethical use of the data.