Erin Howland
DSC680
Project I: Customer Segmentation
25 September 2022

Project I: Customer Segmentation Q & A

Q1: What is customer segmentation?
A1: Customer segmentation is the categorization of a customer base by demographic, geographic, psychographic, and behavioural patterning. This allows a company to see whether we're reaching our target market and whether or how we might be able to expand into a wider market or deeper into our current market. In short, by analyzing our customers, we're able to see where and how we may have opportunity to increase profitability.

Q2: How does customer segmentation work?
A2: Customer segmentation can be done by analyzing customer data, which can be collected during surveys, purchase documents, and similar methods. How, when, and why we collect data should be made transparent to our customers. The disclaimers needed should be provided by our legal department. Once we have data, we can start building an algorithm that allows us to understand the data in a way that we can comprehend and at a rate we are not able to achieve on our own. At its core, segmentation is a method of classification. There are many different classification algorithms we could use, so we've actually tested out a few of them so we can make a recommendation here.

Q3: What algorithm should we use if we want to implement this?
A3: Unsupervised learning classification algorithms are most commonly used for customer segmentation. Two of the six tested algorithms I used are considered to be unsupervised – KMeans and Gaussian Mixture. I would recommend we proceed with KMeans.

Q4: You mentioned supervised and unsupervised learning. What does that mean?
A4: Supervised learning means we have data that we can train our algorithms on that is labeled. That is, we have an initial batch of data that we've already gone through and checked and classified. We're able to measure its success in ways that we more naturally understand, like being able to measure how accurate or precise it is. Unsupervised learning does not require labeling. In this case, instead of labels, we actually create clusters. It's not an intuitive to understand but because it's able to look at data and see patterns we might not normally see, it tends to be very good for problems like customer segmentation. We can evaluate these algorithms using what's called a silhouette score, which tells us how well the algorithm has clustered the data.

Q5: How/where did you obtain your data?
A5: The data we used was obtained from our marketing department by collecting information from purchase documents.

Q6: What risks are there to using the model you recommend?
A6: Risks of unsupervised learning algorithms, regardless of the specific model chosen, are similar. They are less easy to intuitively understand, which could be a problem for some groups of stockholders. It's also a possibility that they show us a pattern where there really isn't one, so there's always a risk that

we end up implementing a project that won't be as successful as we thought it would be. Depending on the project, that can equate to significant financial risk.

Q7: What is the benefit to using the model you recommend?
A7: Unsupervised learning is definitely a double-edged sword. Because those algorithms are so good at finding patterns that we would not otherwise be able to visualize, we could end up finding opportunity in a market section we didn't think we have. We could be missing out on huge growth opportunities. Also, should we want to obtain market data externally, it's often easier to obtain unlabeled data, which would make unsupervised learning a better choice.

Q8: What should we do if we decide we're not comfortable with an unsupervised model?
A8: If you're uncomfortable with an unsupervised model, I would recommend Decision Tree Classifier or KNN. These are supervised learning algorithms and they scored similarly. DTC did marginally better than KNN in this case, but it's similar enough that there's not likely to be a significant difference in results.

Q9: What additional resources will you need to implement this and what timeline do you think you would need?
A9: For the initial implementation, I would recommend we increase the amount of data we analyze internally and combine it with externally sourced data. I would think this could be done in 3 months or less, assuming we already have contracts in place with vendors who can give us access to appropriate data. If we need to create a contract, that could add time to the process. After we run the analysis, we'll want to present our findings to determine next steps, which would necessitate input from multiple departments, including our engineering, marketing, finance, and legal teams at various stages. The length of time needed at that stage will depend on what we find as a result of segmentation.

Q10: Is this project ethical?
A10: The project is currently ethical. We would need to take additional considerations were we to obtain additional market data externally to ensure that data is also collected in an ethical, unbiased manner.