Cluster Assignment

Cade Muntz

**Overview**

We have been tasked with helping a local shoe retailer (ShoesClues) be more efficient with their marketing communications. Essentially, ShoesClues would like to be able to group their customers into relevant segments. This will allow ShoesClues to get a better grasp of how to target their marketing message. To help ShoesClues achieve this goal, we suggested that we run some cluster analyses and a regression model to see whether or not their data is useful to inform decision making. A cluster analysis was chosen as their main purpose is to segment their customers into relevant categories. Regression analysis will then be used to set up a predictive model to see if any descriptive and demographic variables may be useful in predicting cluster membership for ShoesClues.

**Initial Insights**

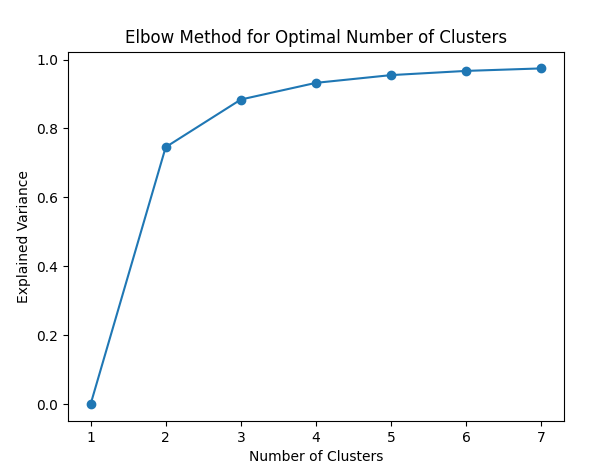
We started off by running a series of cluster analyses on their data to get a feel of what segments their customers may fall into. The variables used for the cluster analysis is as follows:

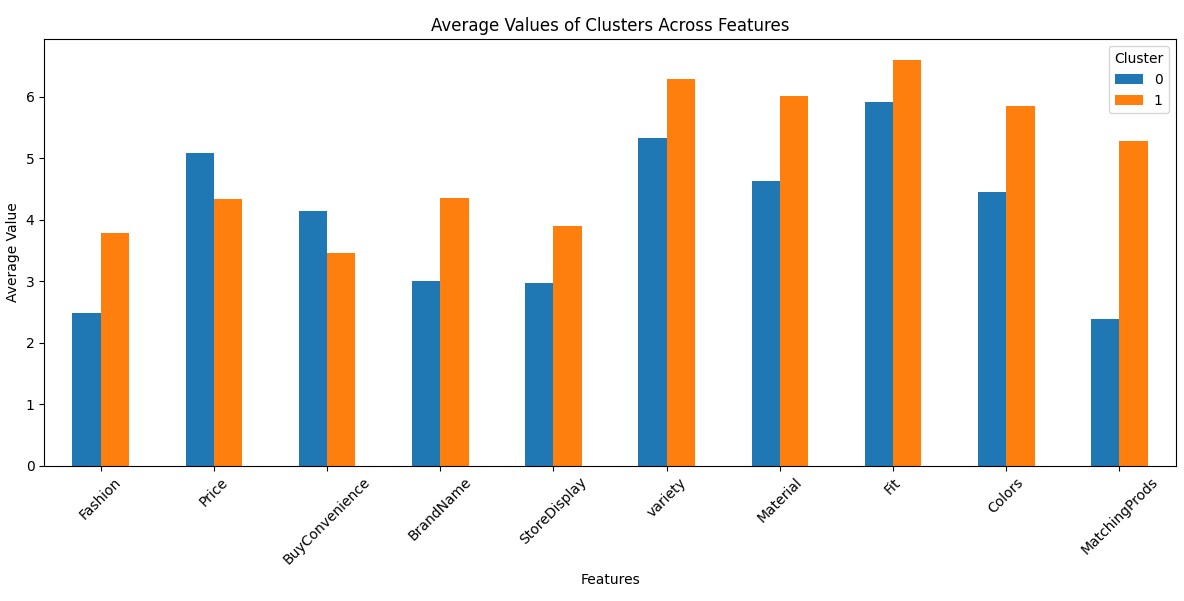
* Fashion (How important fashion is for the customer)
* Price (How important price is for the customer)
* BuyConvience (How important convenience is when shoe shopping for the customer)
* BrandName (How important a shoes brand name is for the customer)
* StoreDisplay (How important the attractiveness of the store’s shoe displays are)
* Variety (How important a variety of shoe choices are to the customer)
* Material (How important the material a shoe is made of is for the customer)
* Fit (How important the shoes fit on their foot is for the customer)
* Colors (How important the color of a shoe is for the customer)
* MatchingProds (How important it is for the shoe to offer products that match the customer's shoes)

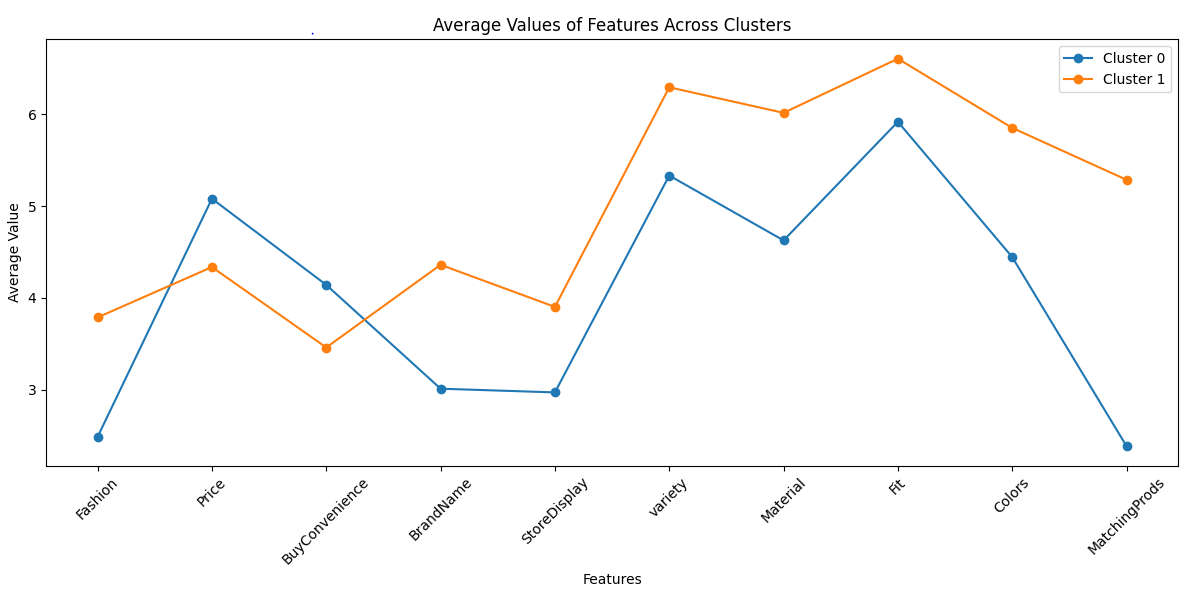
On the first run, we decided to start with 2 clusters and go from there to see how many clusters would be the best for ShoesClues customer base.

**Cluster Analysis:**

Upon running a couple different cluster models we figured out that 2 clusters was the optimal amount of clusters. This is for a couple of different reasons.

* *Reason 1:* Information loss - As you can see in the graph below, at 2 clusters the line begins to slightly flatten out meaning that less information is being lost once 2 clusters are achieved.
* 
* *Reason 2:* Cluster size - When running more than 2 clusters, for example 3 clusters, the 3rd cluster becomes a very small number of customers at 13. This amount of customers would only account for 5.8% of ShoesClues customer base which is a very insignificant amount of customers to target.
* *Reason 3:* Differentiability - When running the cluster analysis with 2 clusters there is a good amount of differentiation between the clusters with only small overlap in the price and convenience categories. When running more than 2 clusters the differentiability is lessened and the clusters get very convoluted within one another. Which means that it will be harder to have target market segmentation with low differentiation. In the graphs below you can see the differentiation between the 2 clusters. Since the lines and bars vary quite greatly, we can easily segment these clusters into different customer niches.





**Cluster Outcomes**

Upon running this 2 cluster analysis we get 2 defined customer segments. We will call these segments the Fashionistas(Cluster 1) and the Convenience shoppers(Cluster 0). Let's get into what both of these segments entail:

* *Fashionistas*
  + When looking at the clusters it is noted that these customers are much more interested in fashion
  + Price and Convenience are not as big of a deal to these customers
  + Brand Name, Fit, and store display are also factors that influence these customers
  + However, these customers are very heavily influenced by Variety, material, colors, and matching products.
* *Convenience Shoppers*
  + These customers are not interested in fashion at all
  + These customers are much more heavily influenced by factors such as price and convenience.
  + These customers do not care too much about the brand name, fit, and store display
  + These customers especially do not care about the variety, colors, materials, and matching products

Essentially, these 2 customers segments are complete opposites from each other, and they provide very good differentiation. Upon setting up our optimal cluster analysis, and diving into what each cluster actually represents, we set up a logistic regression model. Logistic regression is the best model to use because there are only 2 outcomes, cluster 0 or cluster 1.

**Predictive Insights**

When considering a predictive model for ShoesClues it is important to consider variables that have good predictive weight such as information search, psychographic and demographic variables. These variables give more personal information into the customer, and informative information such as where their information of shoes they might purchase come from. Using these predictive variables, we will be able to set up an ideal model for ShoesClues to help them predict which target segment a customer may belong in. This will help ShoesClues target their marketing message much more effectively.

**Initial Model:**

In this model a customer’s cluster can be based on 3 different types of variables, these are - Information Search, Pschographic, and Demographic variables.

*Information Search Variables*

* InfoNet - How likely is a customer to use the internet as a source of information when shopping for shoes
* InfoFriends - How likely is a customer to consult a friend as a source of information when shopping for shoes
* InfoMags - How likely is a customer to to use a magazine as a source of information when shopping for shoes
* InfoCat - How likely is a customer to use a catalogue as a source of information when shopping for shoes
* InfoSocial - How likely is a customer to use social media as a source of information when shopping for shoes

*Psychographic Variables* - These were questions asked to customers on a scale of 1-7 (1 being does not describe them and 7 being describes them very well. The question being “ How well do the following terms describe you?”

* Thrifty
* Athletic
* Cool
* Rational
* Fun

*Demographic Variables*

* PplAskMe - Similar to the psychographic variables with the question being “to what extent do other people ask for your opinion when buying shoes?”
* Age - 1 = 18-25 , 2 = 26-32, 3 = 33-40, 4 = 41-50, 5 = 50+
* Income - Annual Income in $’s - 1 < 40,000 , 2 = 41,000-60,000, 3 = 61,000-90,000, 4 = 90,000-150,000, 5 = 151,000+
* Edu - Education Level - 1 = highschool diploma, 1.81 = Some college, 2 = Associates degree, 3 = bachelors degree, 4 = post-graduate

Upon running this model, there are 5 variables that are good predictors of cluster membership. These variables are ImAthletic, InfoFriends, InfoNet, InfoMags, and InfoCat. Therefore, for future runs of this model we will only include these variables.

**Model Fit:**

Based on the set of significant predictor variables from the initial model, our model was able to predict cluster membership with an accuracy of 69%. This model was much better for predicting cluster 1 membership (Fashionistas) over cluster 0 membership (Convenience shoppers.) The model was able to predict Fashionistas with an accuracy of 82% while it was only able to predict convenience shoppers with an accuracy of 61%. However, both of these are still better than the 50% accuracy ShoesClues would have with their current method of guessing.

|  | True 0 | True 1 | Class Precision |
| --- | --- | --- | --- |
| Predicted 0 | 17 | 11 | 0.56 |
| Predicted 1 | 3 | 14 | 0.85 |
| Class Recall | 0.61 | 0.82 |  |

**Significant Predictors:**

Our most significant predictors were InfoFriends and ImAthletic with p values of 0.001 and 0.002 respectively. InfoNet was also significant having a p value of 0.007 while InfoCat and InfoMags were a bit less predictive with p values of 0.028 and 0.035 respectively, which still makes them significant.

**Insignificant predictors:**

Every other variable used for this model was insignificant however, PplAskMe was very close with a p-value of 0.059. It was a bit surprising that income and age were not significant predictors in terms of shoe shopping, however intuitively speaking, income and age do not necessarily correlate with a person's shopping habits. As some people with low incomes are still in love with nicer shoes, while some people with high incomes are unconcerned with the shoes they wear.

**Interpretations of predictors:**

* InfoNet - This variable has a coefficient of 0.506. Since this variable is positive, customers who tend to get their information on shoes from the internet are more likely to be a fashionista.
* InfoFriends - This variable has a coefficient of 0.680. Since this variable is positive, customers who tend to get their information on shoes from their friends are more likely to be a fashionista.
* InfoMags - This variable has a coefficient of -0.519. Since this variable is negative, customers who tend to get their information on shoes from magazines are more likely to be a convenience shopper. Since magazines are almost the “poorer” version of the internet, this makes sense, as these customers are typically less interested in price.
* InfoCat - This variable has a coefficient of 0.518. Since this variable is positive, customers who tend to get their information from catalogs are more likely to be fashionistas. This makes sense, as catalogs typically feature higher end brands, more geared toward this market segment.
* ImAthletic - This variable has a coefficient of 0.575. Since this variable is positive, customers who tend to see themselves as athletic in response to the survey will more likely be fashionistas.

Based on the interpretations of the predictors, it makes sense that the model has a higher accuracy of predicting fashionistas than convenience shoppers as 4 out of the 5 significant predictors are predictors geared towards the fashionista cluster.

**Recommendations:**

Overall, based on the cluster analysis and predictive model, we can conclude that this model is best fit for predicting fashionistas membership. However, it still has a better than 50% accuracy of predicting convenience shoppers, which will still be helpful to them. Based on the significant predictors, ShoesClues should target their marketing communications using these variables. For example, since fashionistas are typically getting their information from the internet, they should target their ads on the internet towards fashionistas rather than convenience shoppers. On the other hand, they should be using magazines to target their ads towards convenience shoppers.

In terms of the psychographic and demographic variables, these might want to be used less since they are not great at predicting customer market segments. However, since the ImAthletic variable seems to be a good predictor of fashionistas, it might be smart for ShoesClues to target the athletic market with more fashionable products. It could be a good move to try and collaborate with athletic shoe brands, which may attract more customers to their market.