Data Challenge

Brandon Hadfield

# Initial Exploration

The data was first explored to inform strategies for the future tasks. It was noted that there were 5 categorical variables:

* Gender – Male, Female
* City – LA, Miami, NY, Houston, SF, Chicago
* Membership Type – Gold, Silver, Bronze
* Discount Applied – True, False
* Satisfaction Level – Satisfied, Neutral, Unsatisfied, (missing)

And 4 numerical / continuous variables:

* Age – range 26 – 43, avg 33.8
* Items purchased – range 7 – 21, avg 12.5
* Average Rating – range 3.0 – 4.9, avg 4.0
* Days Since Last Purchase – range 9 – 62, avg 27.2

Since the goal was to predict the gross amount spent, it seemed appropriate that this would be highly related to the number of items bought. As can be seen in the figure below that does seem reasonably accurate, with more items purchased typically leading to more money spent.

A graph with blue dots and numbers

Description automatically generated

However, there is still spread, and it is expected that some of that can be explained by the different categorical variables. Some immediate validation of this theory can be found by replotting the previous figure and coloring by city.

A graph with different colored lines

Description automatically generated

Clearly, there are very stark differences in both the typical shopping spree size and spend in each of these cities, with averages of:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| LA | New York | Miami | Houston | Chicago | SF |
| $805.46 | $1165.85 | $691.50 | $446.77 | $500. 20 | $1463.10 |

Already, these would provide very good starting points for predictive clusters for total spend. It would also make sense for total amount spent to be related to whether or not you were paying under a discount, or potentially your gender. So again the figure is coloured by gender and discount status and we receive the following plot:

A graph of a graph with different colored lines

Description automatically generated with medium confidence

Although this grouping is also very stark, it can be seen that we don’t actually receive any new information – outside of one woman shopping in Miami, a given city either contains exclusively men or women shoppers. However, if I were managing this business I may notice that in Chicago, Miami, and New York shoppers are always shopping with a discount. After generating similar plots for membership tiers and satisfaction we can again see that cities are largely homogeneous in their categorical characteristics. There may be a little more information in satisfaction rating. However, as a business owner I would probably note that customers in the lower cap areas (Houston, Chicago, Miami) are very consistently unsatisfied or neutrally satisfied. It may be by design to treat higher paying customers better but having 65% of customers unsatisfied is not sustainable.

A graph with red green and yellow crosses

Description automatically generatedA graph with numbers and symbols

Description automatically generated

Finally, by coloring by the continuous variables we can see a little more differentiation within a city center. Folks who spend more are on average seem to be more happy, whereas potentially in certain cities, folks who have waited longer will spend more, although this is less clear a relationship within a city.

A graph of a graph with different colored lines

Description automatically generated with medium confidenceA graph with different colored lines

Description automatically generated

Finally, since the goal is to predict on unseen data, a quick comparison to the test set was made below, where one can see that for the observations which showed greatest predictive power (qualitatively), the distribution of values is reasonably comparable between training and test. This is important to confidently apply regression and clustering techniques by training exclusively on the training set.

# Data Transformation

Since a regression tree can utilize both categorical and continuous variables, no transformation was applied for the first part.

In the second part the DBSCAN algorithm was implemented for clustering which does require a Euclidean distance to be computed. Thus, for all continuous variables, the mean was removed and values were scaled by the standard deviation. For categorical variables an integer value was used to correspond to each possible value. The neighborhood radius in the DBSCAN implementation was used to tune how easy it was to group different categorical variables together.

# Decision Tree Regression

In my implementation, I decided which variable to split upon by seeing which would provide a decision node which would minimize the sum of squares error. I managed the complexity by using a very simple hard-coded number of nodes. If I had more time, I would have done more intelligent pruning by using some kind of complexity penalty. I would propose tuning this complexity penalty by doing some cross-validation within the training set.

Once applied to the test set, predictions look reasonable compared to that which I saw in the exploration phase.

A graph with blue and white lines

Description automatically generated

# Clustering Algorithm

I implemented the DBSCAN algorithm. Since this puts things into ranges which will have some spread I believe it is worthwhile knowing how confident you are in your predictions, thus I would want to return the range associated with the predicted spend.

In my implementation the DBSCAN algorithm is fit to the training dataset and then when predicting the cluster for a new observation, it simply chooses the closest observation which is a member of a cluster and assigns that cluster to the new observation.

In my implementation there are many naïve aspects, such as the fact that all categorical variables are weighted equally. As I saw previously certain variables carry much more information than others. In future implementations I would have tried to weight categorical variables more accurately given the observed trends as well as tuned the neighborhood distance to provide more detailed visibility. As it stands my clustering algorithm was essentially able to identify which city a given person in the training set comes from as seen below.

A graph of a graph

Description automatically generated with medium confidence

Since this is the major driver of total spend that should still provide reasonable predictions for a range of spend.