Megan Cusey

Unit 5 Mid Term

Summer 2019

Exercise 1: Descriptive Statistics and Visualization

I began to calculate the frequencies for each variable in the data set and realized that the only variable that does not have equal representation is the Rating variable.

The below figure describes the frequencies of the ratings.

The proportion for each rating is as follows:

* Unacceptable: 70%
* Acceptable: 22%
* Good: 4%
* Very Good: 4%

This tells us that most of rows of the data set have a rating of unacceptable. Very few have a rating of Good or Very Good.

A screenshot of a cell phone

Description automatically generated

Exercise 2: Association Rules

The following are the top 3 association rules with two & three frequent itemsets with the strongest support.

|  |  |  |  |
| --- | --- | --- | --- |
| Premises | Conclusion | Support | Confidence |
| Unacceptable Rating | Low Safety | .333 | .476 |
| Unacceptable Rating | 2 Person Vehicle | .333 | .476 |
| Low Safety | Unacceptable Rating | .333 | 1 |
| Unacceptable, Small Luggage | Low Safety | .111 | .427 |
| Unacceptable, Small Luggage | 2 Person Vehicle | .111 | .427 |
| Unacceptable, Medium Luggage | Low Safety | .111 | .490 |

In order to put these association rules into perspective I made the follow statements:

1. 33% of all data under analysis (1728 rows) shows vehicles with an unacceptable rating and low safety. 47.6% of vehicles with unacceptable ratings have low safety.
2. 33% of all data under analysis (1728 rows) shows vehicles with an unacceptable rating and 2-person vehicle. 47.6% of vehicles with unacceptable ratings are 2 person vehicles.
3. 33% of all data under analysis (1728 rows) show vehicles with low safety and an unacceptable rating. All of the vehicles with low safety have an unacceptable rating.
4. 11.1% of all data under analysis show vehicles with unacceptable ratings and small luggage also have low safety. 42.7% of vehicles with an unacceptable rating and small luggage also have the low safety characteristic.
5. 11.1% of all data under analysis show vehicles with unacceptable ratings and small luggage to also be a 2-person vehicle. 42.7% of vehicles with an unacceptable rating and small luggage also have the characteristic of a 2-person vehicle.
6. 11.1% of all data under analysis show vehicles with an unacceptable rating and medium luggage to also have low safety. 49% of vehicles with an unacceptable rating and medium luggage also have the characteristic of low safety.

I believe some of these results are interesting. My thoughts before this exercise was that I believe the minimum support and minimum confidence should be greater, at least over 50%, in order to produce any interesting and actionable results. However, given the nature of the data, I believe that these results should be given consideration.

Quick side bar: It’s interesting to me that there isn’t any frequent itemsets in the top 3 for both 2-itemsets and 3-itemsets that contain a characteristic of a rating above unacceptable. There must not be too many people boasting about how great their car purchase was, or people’s expectations are high regardless of the price, size, etc. of the vehicle.

If I was a car manufacturer, I would be more concerned with the 2-itemsets that have a higher level of support and confident. The 3-itemsets do not make as much sense to me. For instance, an unacceptable rating, small luggage, and low safety seems like a reiteration of the 2-itemset for unacceptable rating and low safety. It may just be a coincidence that small luggage appeared in the date enough times for it appear as a frequent itemset, even though the support it very low. I would be more concerned with the 2-itemset Unacceptable Rating and Low Safety. The other two most frequent 3-itemset shows similar patterns where their support is low and 2 out of the three items are more valuable 2-itemsets.

While the most frequent 2-itemsets don’t necessarily give actionable information, I believe it’s important for a car manufacturer to know that customers who would categorize their vehicle as “Unacceptable” also would categorize their car as “Low Safety” or a “Two Person” vehicle. It’s very interesting that 100% of customers who have indicated their vehicle as a “Low Safety” vehicle given a rating of “Unacceptable”.

While causation can’t be clear, if I was a car manufacturing company, I would interpret the results that feeling safe in a vehicle impacts the rating on a vehicle specifically not feeling safe and giving a bad rating.

Exercise 3: K- Means Cluster

When K=5, clusters 2 and 3 have the highest weights for the “Acceptable” rating. Clusters 0, 1, and 4 have the highest weights for “Unacceptable”. There doesn’t seem to be very high rates for other ratings (good and very good). What is interesting about clusters 2 and 3 VS 0 and 4 is that the purchase price for 2 & 3 have a higher rating for the medium and low purchase price while the other clusters appear have a “Very High” purchase price. You would think that the more you pay for a vehicle; the happier customers would be with their purchase because I would assume it’s “higher quality”. I was also interested to see that in clusters with an unacceptable rating, the safety ratings ranged from low-high and didn’t appear to specifically impact the rating. From a business perspective, I would be most interested in the common characteristics between 2 and 3. For instance, it appears more consumers appreciate bigger vehicles that can hold more people. The cars have a decent space to store luggage. The maintenance and purchase price on the vehicles aren’t extreme. Producing more of these vehicles or incorporating more of these characteristics into other vehicles may help raise the ratings of the vehicles.

Below is a brief description of each cluster:

Cluster 0 – Characteristics with the largest weights include: 2 person vehicles, unacceptable rating, high safety, very high purchase, very high maintenance, luggage varies about evenly.

Cluster 1 – Characteristics with the largest weights include: low safety, unacceptable rating, 2-5+ persons, luggage varies evenly, 2-5+ doors vary evenly. The purchase price categories appear to be evenly weighted.

Cluster 2 – Characteristics with the largest weights include: high safety, acceptable rating, 4-person, more than 4 persons, medium-big luggage, low-medium maintenance, low-medium purchase price

Cluster 3 – Characteristics with the largest weights include: medium safety, acceptable rating, 4-person, 4+ persons, big-medium luggage, low purchase, medium-low maintenance, medium purchase

Cluster 4 – Characteristics with the largest weights include: medium safety, unacceptable, 2-person vehicle, small luggage, very high maintenance, very high purchase, medium luggage.

Exercise 4: K-NN Predictive Model

When k=5, my model had an overall accuracy of 79.73%. In general, I am pleased with the accuracy of the model. I believe if the accuracy was higher, I’d be more concerned with over fitting the model to fit my training data. For instance, if I increased k to 10, the accuracy increases to 87.64%. While it may make somebody more confident to say their model is 87.65% accurate versus 79.73% accurate, in reality, this is due to the analyst making the model fit to the sample data instead of making decisions that make more sense for the data in general. When k is lower, the clusters of data are more general in nature and less “jagged” towards specific data points that may be outliers or just awkward instances in the data.

The model predicted the unacceptable rating 100% of the time when the row actually had a rating of “Unacceptable”. The other categories didn’t do as well however, there is a bit more information on rows with a rating of “Unacceptable” versus the other categories since it makes up 70% of the data. In fact, the accuracy of the model is 79.73% so the model essentially does a really good job at predicting if the rating was “Unacceptable” or not. There isn’t as much representation from the other categories or else the model may have done better. In situations where there is less information to go off of, it may be better to insert some business knowledge in order for the model to make better decisions on which rating to predict.