**BANA 200: Foundations of Business Analytics**

**Starbucks Corporation Analysis**

**Assignment #3**

**Submitted by**

**Hari Sellappan (38562796)**

**INTRODUCTION:**

The study is about the ratings of customers at Starbucks Corporation conducted in Orange County. This study would be useful to conduct some insightful analysis to help senior management understand more about how to improve customer engagement and profitability. In this study, as a first step, the dataset is prepared for analysis by cleaning it. The cleaned Excel file contains survey data on a random sample of 6,121 Starbucks Coffee customers that include zip code-wise ratings, recommend, income, and profits at each customer level. With the sampled data, it is understood that on average, a Starbucks customer gives a profit of $100.99 per month with the median satis100 rating of 55. The median income of the customers who visit Starbucks is little more than $100k. The median likely recommendation rating by the Starbucks customer is 6 out of 10 scales which shows that a lot of the customers have given more than a neutral rating. It is also noticed that the customers who have given recommend a rating of 7 out of 10 scales have given higher profits to Starbucks on monthly basis. Variables X1 – X22 are all measured on a 5-point scale (1 = terrible, 2 = poor, 3 = average, 4 = good, 5 = excellent. The dataset that contains 6121 customers is split to train and test to run a regression model to check the highly significant variables that contribute to the prediction of recommend. A Forward Variable Selection technique was employed to identify the important rating variables. Later, the customers were segmented by clustering technique as Most Satisfied and All other customers. A sample regression was done to understand the difference in each segment and the results are discussed in the later chapters. Starbucks is very interested in drivers that may affect a customer’s willingness to recommend Starbucks to others. Based on the preliminary market research, management believes that it can increase each customer’s ratings in the “All Other” segment by one point for X1, X2, X7, X8, and X10. This study also reveals how much the average willingness to recommend for the “All Other” segment increases if each customer’s survey ratings increase by 1 point which could serve as a basis for its plans to invest in a series of advertisements.

The summary of the dataset is given below:

Table

Description automatically generated

**Q1) a. First, divide the data into a training and test sample.**

**Solution:**

Split the dataset into Train and Test using R:

Training data contains 5000 customers and test data contains the remaining 11121 customers.



**Q1) b. Run a multiple regression on the training sample using “recommend” as the dependent variable and X1 – X22 as the 22 independent variables.** **How many of the 22 predictor variables are significant at the 5% level (have a p-value less than 0.05)? Report the R2 value on the training sample and comment.**

**Solution:**

Dependent variable = Y = recommend

Independent Variables = X = X1:X22 ratings

Table

Description automatically generated

There are about **17 predictor variables** that are significant at the 5% level (have a p-value less than 0.05). The significant variables are **X1, X2, X3, X4, X5, X7, X8, X9, X10, X12, X13, X14, X15, X16, X19, X20 and X22.**

The following are the ratings that are **not significant** at the 5% level of confidence. **X6** – Variety, **X11**- Presentation, **X17**- Personal Treatment, **X18** – Polite, and **X21**- Providing Prompt Service

R-Squared gives a measure of what percentage of the variability of the data is explained by a linear model. The R-squared value is low for the training sample here. The value of 0.**3547** explains only a moderate correlation.

**Q1) c. Using your regression model estimated from part b) above, calculate the out-of-sample R2 value for the 1121 observations in the test sample and report it below. Compare the R2 value from the training sample to the R2 value you calculated in the test sample. What can you conclude about the model’s ability to predict “recommend” in the test sample? How much of a difference is there in the R2 values between the training and test samples?**

**Solution:**

R-squared value from the training and testing sample:



The trained model does not fit well with the test data. The R-squared value has got lowered even more which shows there is poor predictability of the model. There is a difference of **0.06** in the R-squared value between the train and test samples.

**Q2) Variable selection**

**Using only the training sample, perform a forward variable selection procedure by using “recommend” as the dependent variable and X1 – X22 as the 22 predictor variables. Paste the results of your regression results based on the final variables selected below. Which variables were dropped? What is the R2 of the forward selection model? When you compare the R2 of the full model (with all 22 variables) and the R2 of the model using forward selection, by how much did the R2 go down by? What can you conclude about how much those dropped variables really matter?**

**Solution:**

The data is prepared in such a way that rows are observations (individuals) and columns are variables. And there is no missing value in the data that must be removed or estimated before clustering. The data must be standardized (i.e., scaled) to make variables comparable. But in this case, there is no need for standardization as the variables are on the same scale.

A multiple linear regression model is fit using *recommend*as our dependent variable and all the other 22 variables from X1 to X22 in the dataset as potential predictors variables to perform a forward stepwise selection. At first, an intercept-only model was fit. This model has provided an AIC of 8741.599.

Next, we fit every possible one-predictor model. The model that produced the lowest AIC and has a statistically significant reduction in AIC compared to the intercept-only model used the predictor X1. This model has provided an AIC of 7656.418. The same procedure was implemented multiple times until it turned out that none of the models could produce a significant reduction in AIC.

Table

Description automatically generated

Table

Description automatically generated

The final model turns out to be:

Recommend ~ -3.50 + 0.63\*X1 + 0.44\*X16 + 0.4\*X7 + 0.38\*X13 + 0.32\*X20 + 0.22\*X5 + 0.35\*X10 - 0.16\*X12 - 0.27\*X19 + 0.25\*X22 + 0.11\*X15 + 0.2\*X2 - 0.15\*X14 + 0.12\*X8 - 0.12\*X9 + 0.09\*X21 - 0.10\*X4 + 0.09\*X3 - 0.08\*X11

There are 3 variables **X6** – Variety, **X17**- Personal Treatment, and **X18** – Polite are dropped from the list of the original 22 variables. The R-squared value of the forward selection model is **0.3543.** The R-Squared of the full model (with all 22 variables) is **0.3547**. The difference in R-squared values is **0.0004** which is very negligible. This difference shows that these 3 variables do not contribute much to the prediction of recommend.

**Q3) a. Using all the data (all 6121 observations), create a data matrix called “X” which includes the 22 predictor variables: X1, X2, …, X24, X25. Your data matrix X should have 6121 rows and 22 columns.**

**Solution:**

Created a Matrix called X using the first 22 columns of the Starbucks data.

X <- as.matrix(star\_data[,1:22])

**Q3) b. Once you have created your data matrix, use the NbClust procedure discussed in class to determine the optimal number of segments (clusters) for X. Based on the analysis performed, what are the optimal number of clusters for X? Use the “majority rule” to determine the optimal number of clusters. Paste the bar chart you obtained from the analysis in R below and report the optimal number of clusters to use.**

**Solution:**

Chart, bar chart, histogram

Description automatically generated

Table

Description automatically generated with low confidence

Chart, line chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Using Euclidean distance as the distance measure, the minimum number of clusters to test = 2, the maximum number of clusters to test = 10, and method = “kmeans”, the analysis is performed and using the majority rule, **the** **optimal number of clusters for X = 2**

**Q3) c. Using the optimal number of clusters, you found in part Q3b above, run a k-means cluster analysis on the X matrix (perform a** **k-means cluster analysis on the X matrix using X1 – X22). Set “centers =” to the optimal number of clusters you found in step Q3b and set the iter.max = 1000 and nstart=100. Report below how many customers are in Cluster 1 and how many customers are in Cluster 2.**

**Solution:**

This k-means cluster analysis was performed on the dataset X matrix using the variables X1 to X22. Centers is set to the optimal number of clusters which is 2, the iter.max = 1000 and nstart=100.

The customers in cluster 1 and cluster 2 are reported as follows:



**Q3) d. Executive management has asked you to identify the “most satisfied” segment of customers. Examine the cluster centers from your k-means analysis and identify the segment of customers that seem the most satisfied. Hint: The most satisfied segment should be the one that generally has the highest average ratings (the highest cluster center values for X1 – X22). Once you have identified the most satisfied segment of customers, flag this segment and set them aside. Report below the cluster center values for X1, X2, X3, X4, and X5 (rounded to two decimal places) for this most satisfied segment of customers.**

**Solution:**

After examining the average cluster center values given below for both the clusters, **Cluster 1** is the most satisfied segment that has the highest average ratings (the highest cluster center values for X1 – X22).

The average rating for Cluster 1: **4.13**

The average rating for Cluster 2: **3.28**

The cluster center values for X1, X2, X3, X4, and X5 (rounded to two decimal places) for the most satisfied segment of customers are reported as below:

Text

Description automatically generated

**Q3e) 1.** **Split your overall data sample into two groups: “Most Satisfied” and “All Other”. The “most satisfied” group of customers should consist of the one segment that is most satisfied based on Step 3d above, and “All Other” customers will include all other customers that are not in the “most satisfied” segment.**

The overall data sample is split into two groups Most Satisfied and All Other. The customers belonging to **cluster 1** are identified as **Most Satisfied** and Customers belonging to **cluster 2** are identified as **All Other** customers.

**Q3e) 2.** **Next, run two separate regression analyses for each group. Use “recommend” as the dependent variable for both regressions and X1 – X22 as the 22 predictor variables.**

**Regression results for Most Satisfied Customers:**

A screenshot of a computer

Description automatically generated with low confidence

**Regression results for All Other Customers:**

A screenshot of a computer

Description automatically generated with medium confidence

**Q3e) 3. For each one of the regressions, report the average predicted values. That is, extract the two sets of predicted values from the two lm objects by using the “fitted. Values” function, and for each regression, take the average of these fitted values and report these two averages below. By how much more (in terms of average predicted 4 “recommend”) is the “Most Satisfied” segment likely to recommend Starbucks? Round all answers to two decimal places**

The average predicted value for the “Most Satisfied” segment is **7.33** whereas for the “All Other” segment is **5.19**. On average, the most satisfied segment of customers is likely to recommend **2.14 ratings more** than all other customers.





**Q4) a. Management wants to figure out by how much more it can increase the “All Other” segment’s willingness to recommend Starbucks to others. It has conducted some market research and plans to invest in a series of advertisements. Based on the preliminary market research, management believes that it can increase each customer’s ratings in the “All Other” segment by one point for X1, X2, X7, X8, and X10. Starbucks has asked you to recalculate the average willingness to recommend for the “All Other” segment if each customer’s survey ratings increase by 1 point for X1, X2, X7, X8, and X10 in that segment.**

For the “All Other” segment only, increased X1, X2, X7, X8, and X10 by one point as depicted in the below code:

Text

Description automatically generated

For example, if Customer 1 has X1 = 3, It is set as X1 = 4 for this customer. However, if Customer 1 has X1 = 5, the rating is not changed.

**Q4) b. Once you have changed the ratings for X1, X2, X7, X8, and X10 for the “All Other” segment, only use your existing regression model results from Q3 to recalculate the predicted “recommend” for the “All Other” segment.**

Management wants to figure out how much more it can increase the “All Other” segment’s willingness to recommend Starbucks to others by one point increase in a few variables as listed above. Predicted recommend using the existing regression model of “All Other” segments. It is made sure thatnew values for X1, X2, X7, X8, and X10 are used as the basis for these predictions. The results are discussed subsequently.

**Q4) c. Once you have recalculated the predicted values for all customers in the “All Other” segment (based on the updated values for X1, X2, X7, X8, and X10), take the average of these new predicted values and report this average below, rounded to two decimal places. By how much is the willingness to recommend expected to increase by if Starbucks can get the “All Other” customer segment to be one point more satisfied for X1, X2, X7, X8, and X10? Does this seem like a worthwhile thing to do? Comment on whether the change seems significant or not.**

The average of the new predicted values of All Other segment customers when X1, X2, X7, X8, and X10 are increased by one point is **7.05.**

The average of the predicted values of the original data of All Other Segmentsis **5.19.**

The difference between these values comes out to be **1.86**

The **willingness to recommend is expected to** **increase by 1.86** if Starbucks can get the “All Other” customer segment to be one point more satisfied for X1, X2, X7, X8, and X10.

If each customer’s survey ratings increase by 1 point for X1, X2, X7, X8, and X10 in All Other segments, the ratings of All Other Customers will improve closer to the Most Satisfied Customers. The change seems to be highly significant because 1 point increase in a few variables leads to an increase in nearly 2 ratings in recommend which further could drive profit keeping customer experience at first.

It is worthwhile for Starbucks to invest in increasing the recommend rating of All Other Customers.