**Assessment 2**

**MIS771 Descriptive Statistics**

****

**Name: Yash Ahuja**

**Student ID: 219608443**

**Introduction**

The purpose of this report is to provide analysis of the large-scale survey of the TassPaperMill (TMP) customers. In 2019, the company produced tonnes of paper products and marketed it to local as well as global areas. The purpose of this analysis is to get an overview of the ordered quantity of TPM products, the likelihood of customers signing a contract with TPM and to forecast the turnover for the next four quarters. In order to estimate the ordered quantity of TMP products, various factors such as loyalty, product quality, brand image, shipping cost are analyzed. Moreover, we have tried to investigate if the effect of product quality on ordered quantity depends upon brand image.

**Main Body**

**To:** **Hugo Barra**

**Topic:** **TPM Research Project – Analytics Details**

**Q1)** **Ordered Quantity**: Out of the sample provided to us, the mean ordered quantity is 7.665 tonnes. By applying descriptive statistics, we found that the distribution of ordered quantity variable is negatively skewed (left). Also, the interquartile range of the variable is 1.1, no extreme case was detected as all the values are approximately in between upper and lower fence limits. Moreover, it is evident from the figure that around 67.5% of the values of ordered quantity variable is less than or equal to 8 tonnes (see figure 1.1).

**Contract with TPM**: A categorical variable which is best visualized by cross tabulations or pie charts. As we can see from the figure, the proportion of surveyed people having contract with TPM is approximately equal as compared to those who did not have a contract (see figure 1.2).

**Q2) 2.1)** On performing the correlation analysis between the dependent variable Ordered Quantity and other independent variables, it has been found that Ordered quantity is most significantly influenced by factors such as Loyalty, Distribution Channel, Quality, Product line, Shipping Speed, Shipping Cost, Contract with TPM. All the above variables have positive relationship with Ordered quantity and the strength of their relationship with ordered quantity is medium. Other variables such as social media presence, advertising, brand image, order fulfillment, flexibility in price also have positive relationship with ordered quantity but the strength of their relationship is low with flexibility in price having weakest positive relationship (correlation value almost 0). Contract with TPM has the strongest positive relationship with ordered quantity out of all the independent variables as its coefficient of correlation with ordered quantity is 0.51.

Moreover, there are certain independent variables who have negative relationship with ordered quantity. These include customer type, region, competitive pricing. The strength of their relationship with ordered quantity is low. Competitive pricing has the most significant negative relationship with the ordered quantity as its coefficient of correlation is -0.22.

**2.2)** **Removing Multi-collinearity**: After performing the correlation analysis, we found the strength and direction of relationship of independent variables with the ordered quantity. We also found that Shipping Speed and Shipping Cost are highly correlated as their correlation coefficient is 0.84 and it violates the assumption of regression model that all the selected variables are independent. Therefore, we need to eliminate one out of Shipping Cost and Shipping Speed variables before building our regression model.

As Shipping Speed has low correlation value (0.43) with ordered quantity as compared to the Shipping Cost (0.50), we will eliminate Shipping Speed variable from our model.

Now, we will start building our model with the all the available variables excluding Shipping Speed.

After selecting the available 14 variables we start building the model which required nine iterations to develop statistically significant model at the last.

**Reason for Iteration**:

After analyzing the results for the first time, we found that the p-value of region variable is 0.913 (not less than 5% level of significance). Hence, the region variable is statistically insignificant, and we need to eliminate this variable from our model and perform regression again.

**Second Model**:

After performing the first iteration, we found that the variable we found that the p-value of customer type variable is 0.92 (not less than 5% level of significance). Hence, the customer type variable is statistically insignificant, and we need to eliminate this variable from our model and perform regression again.

**Third Model**:

After performing the second iteration, we found that the p-value of product line variable is 0.72 (not less than 5% level of significance). Hence, the product line variable is statistically insignificant, and we need to eliminate this variable from our model and perform regression again.

**Fourth Model:**

After performing the third iteration, we found that the p-value of distribution channel variable is 0.62 (not less than 5% level of significance). Hence, the distribution channel variable is statistically insignificant, and we need to eliminate this variable from our model and perform regression again.

**Fifth Model:**

After performing the fourth iteration, we found that the p-value of advertising variable is 0.49 (not less than 5% level of significance). Hence, the advertising variable is statistically insignificant, and we need to eliminate this variable from our model and perform regression again.

**Sixth Model:**

After performing the fifth iteration, we found that the p-value of price flexibility variable is 0.53 (not less than 5% level of significance). Hence, the price flexibility variable is statistically insignificant, and we need to eliminate this variable from our model and perform regression again.

**Seventh Model:**

After performing the sixth iteration, we found that the p-value of social media presence variable is 0.25 (not less than 5% level of significance). Hence, the social media presence variable is statistically insignificant, and we need to eliminate this variable from our model and perform regression again.

**Eighth Model:**

After performing the seventh iteration, we found that the p-value of order fulfillment variable is 0.08 (not less than 5% level of significance). Hence, the order fulfillment variable is statistically insignificant, and we need to eliminate this variable from our model and perform regression again.

**Ninth Model:**

After performing the eighth iteration, we found that the p-value of competitive pricing variable is 0.09 (not less than 5% level of significance). Hence, the competitive pricing variable is statistically insignificant, and we need to eliminate this variable from our model and perform regression again.

**Final Model (Tenth Model):**

After performing the ninth iteration, we found that the p-value of all the independent variables is less than 0.05 (5% level of significance). Hence, all the independent variables namely Loyalty, Quality, Brand Image, Shipping Cost and Contract in the final model are statistically significant, and we do not need to perform regression again.

**Results of Final Model:**

**p-value:**

The p-value of the model is approximately 0.00 (less than 5% level of significance). Hence, the model has some predictive power.

**R square:** 54%

Approximately 54% of the variation in ordered quantity can be explained by the regression model (i.e. the variables included in the regression model). And, approximately 46% of the variation in ordered quantity would be explained by other factors which are not included in the model. The value of R square is 54%, hence it has moderate predictive power.

**Adjusted R square:** 53%

Approximately 53% of the variation in ordered quantity can be explained by the regression model (i.e. the variables included in the regression model) considering the sample size of 200 examples and the number of independent variables (5). And, approximately 47% of the variation in ordered quantity would be explained by other factors which are not included in the model. The value of adjusted R square is 53%, thus the model has medium predictive power.

**Standard Error:** 0.61

The standard error of our model is 0.61. The standard represents the average distance the observed values fall from the regression line. The standard error of the model is low and hence acceptable.

**Coefficient Interpretation:**

Loyalty variable: For one unit increase in loyalty i.e. the length of time a customer has been buying from TPM, the average value of ordered quantity will increase by 0.05 tonnes, assuming no change in other variables.

Quality variable: For one unit increase in quality, the average value of ordered quantity will increase by 0.21 tonnes, assuming no change in other variables.

Brand Image variable: For one unit increase in brand image variable, the average value of ordered quantity will increase by 0.13 tonnes, assuming no change in other variables.

Shipping Cost variable: For one unit increase in Shipping Cost variable, the average value of ordered quantity will increase by 0.22 tonnes, assuming no change in other variables.

Contract variable: The average value of ordered quantity will increase by 0.34 tonnes if the order is placed by customer having a contract with TPM as opposed to the one who is not having a contract with TPM, assuming no change in other variables.

**Extreme Cases:**

Also, we found ten residuals with high values by conditional formatting of residuals (greater than 2 and less than -2). These are extreme cases and needs further investigation.

**Q3)** **Interaction Effect:** We conclude that at 5% level of significance, interaction does exist in the regression model as p-value for interaction variable is 0.04. In other words, effect of product quality on ordered quantity does depend upon the brand image. Moreover, it can be seen from figure 3 that Hugo’s claim is correct as the strength of the relationship between product quality and ordered quantity is stronger for people who have a favorable perception of TPM brand (High Brand Image) as compared to people with less favorable perception of TPM brand (Low Brand Image).

Figure 3 Interaction between Quality & Brand Image

**Performance Metrics of Model:**

**R square:** 35%

Approximately 35% of the variation in ordered quantity can be explained by the regression model (i.e. the variables included in the regression model). And, approximately 65% of the variation in ordered quantity would be explained by other factors which are not included in the model. The value of R square is 35% which has dropped significantly compared to the regression model built earlier and hence it has low predictive power.

**Adjusted R square:** 35%

Approximately 35% of the variation in ordered quantity can be explained by the regression model (i.e. the variables included in the regression model) considering the sample size of 200 examples and the number of independent variables. And, approximately 65% of the variation in ordered quantity would be explained by other factors which are not included in the model. The value of adjusted R square is 35% which has dropped significantly compared to the regression model built earlier and hence it has low predictive power.

**Standard Error:** 0.72

The standard error of our model is 0.72 which has increased from the regression model built earlier. The standard represents the average distance the observed values fall from the regression line. The standard error of the model is still low and hence acceptable.

**Coefficient Interpretation:**

Loyalty variable: For one unit increase in loyalty i.e. the length of time a customer has been buying from TPM, the average value of ordered quantity will increase by 0.05 tonnes, assuming no change in other variables.

Quality variable: For one unit increase in quality, the average value of ordered quantity will increase by 0.69 tonnes, assuming no change in other variables.

Brand Image variable: For one unit increase in brand image variable, the average value of ordered quantity will increase by 0.86 tonnes, assuming no change in other variables.

Interaction variable (Quality\*Brand Image): For one unit increase in interaction variable Quality\*Brand Image variable, the average value of ordered quantity will decrease by 0.07 tonnes, assuming no change in other variables.

**Extreme Cases:**

Also, we found 9 residuals with high values by conditional formatting of residuals (greater than 2 and less than -2). These are extreme cases and needs further investigation.

**Q4) 4.1)**

After performing the logistic regression, we found that the p-value of competitive pricing is 0.77 which is statistically insignificant (not less than 5%). Hence, we removed the competitive pricing variable and performed the logistic regression again.

**Final Logistic Regression Model Results:**

**Assessment of Statistical Significance**:

“Pseudo” R square measures:

|  |  |  |
| --- | --- | --- |
| R-square (L) McFadden | R-square (CS) Cox and Snell | R square (N) Nagelkerke |
| 29% | 33% | 45% |

According to **R-square(L)**, 29% of the variation in the dependent variable i.e. ordered quantity can be explained by the regression model. Around 71% of the variation in the dependent variable would be explained by other factors which are not included in the model.

As per **R square (CS)**, 33% of the variation in the dependent variable i.e. ordered quantity can be explained by the regression model. Around 67% of the variation in the dependent variable would be explained by other factors which are not included in the model.

As per **R-square (N)**, 45% of the variation in the dependent variable i.e. ordered quantity can be explained by the regression model. Around 55% of the variation in the dependent variable would be explained by other factors which are not included in the model.

**LL0:** -138.62

**LL1:** -97.87

Compared to the baseline model LL0, the final logistic regression model LL1 has reduced LL (Log Likelihood) value significantly. Hence it is evident that the final model has statistical significance.

**p-Value: 0.00**

As the p-Value for the final model is less than 5% level of significance. Hence, the model possesses statistical significance.

**ROC Curve:**

As we can see from figure 4.1 that the area of the ROC curve is near to 1, indicating a good measure of separability. Thus, model has ability to differentiate between classes.

**Assessment of Practical Significance:**

For unequal group sizes as it is in this case with 101 having a contract with TPM and 99 not having a contract with TPM, Proportional Chance Criterion (PCC) would be used.

**Proportional Chance Criterion (PCC) Hit Ratio** = 50%

50% of the customers were accurately classified by the regression model. The remaining miss-classification would have been captured accurately if other relevant independent variables were also included in the model.

**Standard Hit Ratio** = 63%

Since, Accuracy rate of 76.5 % is greater than Proportional Chance Criterion Hit Ratio and Standard Hit Ratio, it provides evidence for the practical significance of the model.

**Assessing the Contribution of Predictors:**

Quality: For one unit increase in quality i.e. the perceived level of quality of TPM's products, the likelihood (odds) of the customer signing the contract with TPM increases by 125.12%.

Product line: For one unit increase in product line i.e. Depth and breadth of TPM product line to meet customer needs, the likelihood (odds) of the customer signing the contract with TPM increases by 84.79%.

Brand Image: For one unit increase in brand image variable, the likelihood (odds) of the customer signing the contract with TPM increases by 114.32%.

Price Flexibility: For one unit increase in price flexibility variable, the likelihood (odds) of the customer signing the contract with TPM increases by 90.07%.

**4.2)** The customers analysed here have neutral feelings towards brand image and product line. As we can see from the figure 4.2 that for high- and low-quality products, Price Flexibility does not have significant impact in affecting the likelihood of customers signing a contract with TPM. But for medium quality products, it is evident that price flexibility has significant impact in affecting the likelihood of customers signing a contract with TPM. For high price flexibility level (10), the predicted probability is high as compared to the 5-price flexibility level. The predicted probability for customers signing a contract with TPM for 5 price flexibility level is low for medium quality products.

Figure 4.2 Predicted Probability Plot

**Q5)** **Forecasting Turnover**

Method used for forecasting turnover is trend-based forecasting for seasonal component as the data given to us has seasonal components i.e. it is quarterly data and we would to like to predict the turnover for next four quarters (seasonal data).

The forecasted turnover is following the seasonal trend and the turnover values for the next four quarters are shown in figure 5.

It is evident from the model that predicted turnover for next year’s Q1, Q3 and Q4 will be approximately 4-5% higher than last year’s respective quarters. Moreover, the highest increase is forecasted for Q2 as turnover for next year’s second quarter will be approximately 9.5% higher than last year’s second quarter.

**MAPE**: The Mean Absolute Percentage Error of our model is approximately 2%. It is one of the most common measure of fit used in time series forecasting.

**Conclusion**

Thus, we can conclude by saying that the factors such as loyalty, product quality, brand image, shipping cost and contract with TPM are statistically significant in estimation of ordered quantity of products. Also, the strength of the relationship between product quality and ordered quantity is stronger for people who have a favorable perception of TPM brand as compared to people with less favorable perception of TPM brand. Moreover, for high- and low-quality products, Price Flexibility does not have significant impact in affecting the likelihood of customers signing a contract with TPM. But for medium quality products, it is evident that price flexibility has significant impact in affecting the likelihood of customers signing a contract with TPM. Furthermore, the forecasted turnover for 3 out of the 4 quarters of next year will be approximately 5% higher than last year.

**Appendices**

Figure 1.1 Distribution of ordered quantity

Figure 1.2 Percentage of customer having contract with TPM

Figure 4.1 ROC Chart

|  |  |
| --- | --- |
| Quarter 2020 | Approximated Forecasted Turnover (in Tonnes) |
| Q1 | 1723 |
| Q2 | 2133 |
| Q3 | 1601 |
| Q4 | 1768 |

Figure 5 Forecasted Turnovers