DEAKIN UNIVERSITY

Descriptive Analytics and Visualisation

MIS 771

ASSIGNMENT TWO

Analysis of B-Hive



Submitted by: Ayush Pradhananga

Student ID: 221417078

**Introduction**

In this report, analysis on sales forecast, prediction on quantity ordered and recommending the B-Hive Honey is done. B-Hive has had a great sales growth over the past years. However, their recent sales have not been that great. This report aims to provide in depth analysis regarding the characteristics of the customers as to what factors do they consider when ordering the products and under what circumstances do they recommend B-Hive Honey to others. Also, sales forecast for the next four quarter is presented in this report.

**Body**

**Task 1: Linear Regression Model**

**1.1 Correlation and Scatter Plot**

Firstly, the relationship of all independent variables (excluding recommend) with order quantity has been analysed. Scatter plot and correlation matrix has been used for this purpose. I have excluded recommend as independent variable because I will be using this variable as a dependant variable in another model.

**Loyalty and Order Quantity**

The relationship is positive and linear meaning as the number of years of customer buying from B-Hive increases, the order quantity also increases. The strength of the relationship is moderate with a value of 0.41.

**Customer Type and Order Quantity**

The relationship between the two variables is very weak having a correlation value of -0.05. Customer type’s influence on order quantity seems to be negligible. Also, it is categorical variable which is not the best fit for linear regression model.

**Region and Order Quantity**

Their relationship is very weak with low correlation of -0.12. It is also a categorical variable which is not a best fit for a linear model.

**Distribution Channel and Order Quantity**

They have a moderate positive relationship. The correlation value among them is 0.40.

**Quality and Order Quantity**

It is seen that better the quality of B-Hive products, higher is the order quantity. Their relationship is linear, positive and the strength is moderate with a correlation value of 0.43.

**SM\_Presence and Order Quantity**

Social Media presence has a strong influence on order quantity. Higher the social media presence more is the quantity ordered. The relationship is positive and linear.

**Advert and Order Quantity**

They have a weak positive and linear relationship with 0.24 correlation value.

**Brand image and Order quantity**

They have weak to moderate relationship with 0.34 correlation value and their relationship is linear. Their relation is positive indicating that when brand image gets better, more quantity is ordered.

**Competitive Pricing and Order Quantity**

We see a linear and negative relation between the two variables indicating that the customers do not like to order more quantity when the prices are very competitive. However, the strength is weak with a correlation value of -0.22.

**Order Fulfilment and Order Quantity**

We can see a positive and linear relationship. Their relation is weak to moderate with 0.31 correlation value.

**Flex Price and Order Quantity**

Price flexibility has no influence in order quantity. **Hence, this variable is not considered for the model.**

**Shipping Speed and Order Quantity**

A moderate, positive and linear relationship is observed. The correlation value is 0.43.

**Shipping Cost and Order Quantity**

A strong, positive and linear relationship is observed. As the shipping cost gets more affordable, more quantity is ordered. Their correlation value is 0.5.

After analysing the relationship of all variables, it is seen that all variables have linear relationship and hence linear regression is a suitable model.

It is important to get rid of multi-collinearity issue in the dataset. This issue occur when two independent variables are highly correlated. As a general rule of thumb, I have considered a limit of 0.80. It is observed from the correlation matrix table that there is a high correlation between shipping speed and shipping cost. One of the variables must be eliminated from the model. Both statistically and logically it is right to remove shipping speed. Statistically shipping cost has a higher influence on order quantity. Logically, every delivery would have shipping cost and faster delivery isn’t necessary for every item. However, lower shipping cost is desirable by every customer. Hence, shipping speed is eliminated from the model.

**1.2 Model Building**

After pre-processing of the data, 11 variables have been considered that can influence the order quantity. They are Loyalty, Customer Type, Region, Distribution Channel, Quality, Social Media Presence, Advert, Brand Image, Competitive Pricing, Order Fulfilment and Shipping Cost.

In the model building process, the least statistically significant variable will be eliminated one by one. The variables are considered not significant if their p-value is greater than alpha 0.05.

|  |  |
| --- | --- |
| Iteration | Description |
| 1st | Advert eliminated p value = 0.86 |
| 2nd | Distribution Channel eliminated p value = 0.79 |
| 3rd | Customer Type eliminated p value = 0.29 |
| 4th | Region eliminated p value = 0.09 |
| 5th | Brand Image eliminated p value = 0.21 |

The final regression model included Loyalty, Quality, Social Media Presence, Competitive Pricing, Order Fulfilment and Shipping Cost as predictors. All of the predictors are statistically significant.

**Final Regression Model is:**

**Order Quantity= 6.38+ (0.07\*Loyalty) + (0.21\*Quality) + (0.54\*Sm\_Presence) - (0.10\*Competitive Pricing) - (0.20\*Order Fulfilment) + (0.37\*Shipping Cost)**

Assuming all independent variables constant, the order quantity will be 6.38 kg of Honey Buckets.

Everything else remaining the same, a unit increase in Loyalty will result in the order quantity increase by 0.07 times on average.

Everything else remaining the same, a unit increase in Quality will result in the order quantity increase by 0.21 times on average.

Everything else remaining the same, a unit increase in Social media presence will result in the order quantity increase by 0.54 times on average.

Everything else remaining the same, a unit increase in competitive pricing will result in the order quantity decrease by 0.10 times on average.

Everything else remaining the same, a unit increase in order fulfilment will result in the order quantity decrease by 0.20 times on average.

Everything else remaining the same, a unit increase in shipping cost will result in the order quantity increase by 0.37 times on average.

R-squared value is 64% which means 64% of the variation in the model is explained by our model. The model has some predictive power. Approximately, 36% of the variation in quantity ordered is explained by other factors not included in the model.

The data has 10 potential outliers. The LINE assumption is met and model is valid.

**Task 2: Interaction Effect**

The model has some predictive power but it is very low. All the predictors are statistically significant with p-values lower than 0.05.

It is seen from the graph that the two variables do not intersect depicting no interaction effect. For high brand image the quantity ordered was slightly more for high quality than low brand image. However, when quality was low, ordered quantity was significantly high for high brand image. Regardless of the brand image, as the quality improved, the ordered quantity also increased. Moreover, regardless of the quality the quantity ordered for high brand image was always higher than that of low brand image.

**The model:**

**Order Quantity= 0.75+ (1.30\*Brand Image) + (1.04\*Quality) - (-0.10\*Brand Image\*Quality)**

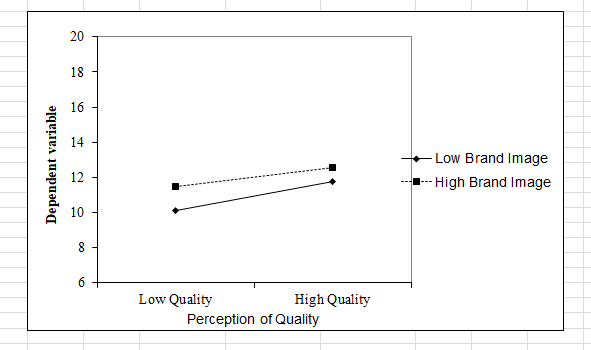
Assuming all independent variables constant, the order quantity will be 0.75 kg of Honey Buckets.

Everything else remaining the same, a unit increase in Brand Image will result in the order quantity increase by 1.30 times on average.

Everything else remaining the same, a unit increase in Quality will result in the order quantity increase by 1.04 times on average.

Everything else remaining the same, a unit increase in interaction term will result in the order quantity decrease by 0.10 times on average.

The r-squared value is 0.35 which is very low. About 35% of the variation in the order quantity is explained by our model. The rest 65% of the variation would be explained by other factors not included in our model.



**Task 3: Logistic Modelling**

In this model I have excluded order quantity as independent variable. Also, shipping cost (due to multi-collinearity), price flexibility and customer type has been eliminated due to low correlation. The dependent variable is Recommending B-Hive Honey.

This model goes through multiple iterations. The least significant variable is removed one by one. In the baseline model only distribution channel and quality is significant. In the final model, only the statistically significant variables are included that is distribution channel, quality, brand image and shipping speed.

**The final model:**

**Recommend = -12.53 +0.93\* Distribution Channel + 0.61\* Quality + 0.57\*Brand Image + 1.12\*Shipping speed**

The likelihood (odds) of recommending B-Hive Honey to others is 154.24 per cent greater for direct sales than through a distribution network.

A unit of increase in perception of quality increases the likelihood (odds) of recommending B-Hive Honey to others by 83.61 %.

A unit of increase in perception of brand image increases the likelihood (odds) of recommending B-Hive Honey to others by 77.51%.

A unit of increase in perception of quality shipping speed increases the likelihood (odds) of recommending B-Hive Honey to others by 206.39 %.

As per the Hosmer and Lemeshow r-squared, 29% of the variation in recommending B-Hive Honey can be explained this model. However, the Cox and Snell r-squared and Nagelkerke r-squared are 33% and 45% respectively. The model does not have high predictive power.

The hit ratio was 74% meaning the 74% of the customers were correctly classified.

Out of 100 customers who recommended B-Hive Honey 74 were correctly classified and 26 was incorrectly classified.

Out of 100 customers who did not recommend B-Hive Honey 74 were correctly classified and 26 was incorrectly classified. This is a balanced dataset and hence, cut-off of 50% can be used. Since the accuracy is more than 50%, our model is significant.

In the ROC curve, we can see that Area under the curve (AUC) is close to 1, approximately 0.84, so the model fits the data.

**Task 4: Visualising and Interpreting Predicted Probability**

**4a and 4b**

The likelihood (odds) of recommending B-Hive Honey to others is 188.16 per cent greater for direct sales than through a distribution network.

A unit of increase in perception of quality increases the likelihood (odds) of recommending B-Hive Honey to others by 74.60 %.

A unit of increase in perception of brand image increases the likelihood (odds) of recommending B-Hive Honey to others by 94.65%.

As per the Hosmer and Lemeshow r-squared, 23% of the variation in recommending B-Hive Honey can be explained this model. However, the Cox and Snell r-squared and Nagelkerke r-squared are 27% and 36% respectively. The model does not have high predictive power.

The hit ratio was 70.5% meaning the 70.5% of the customers were correctly classified.

Out of 100 customers who recommended B-Hive Honey 68 were correctly classified and 32 was incorrectly classified.

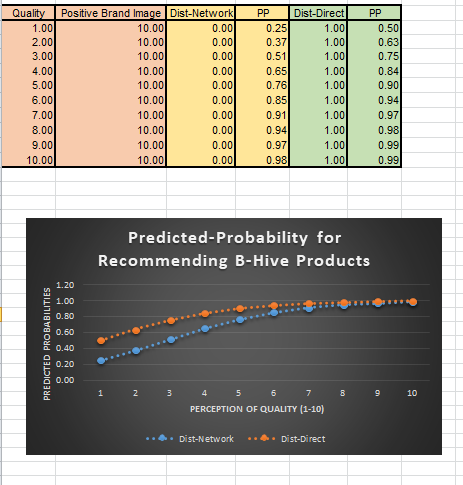
Out of 100 customers who did not recommend B-Hive Honey 73 were correctly classified and 27 was incorrectly classified. This is a balanced dataset and hence, cut-off of 50% can be used. Since the accuracy is more than 50%, our model is significant.

In the ROC curve, we can see that Area under the curve (AUC) is close to 1, approximately 0.80, so the model fits the data very well.

There are two distribution channels:

1. O meaning through a network
2. 1 meaning direct distribution

Overall, for the higher quality both the direct and distribution network have similar recommending probabilities. When, the quality is lower, probability of recommending B-Hive honey is significantly higher for direct distribution than through a network. In general, as the quality increases, the probability of recommending B-Hive Honey increases for both the distribution channel. The maximum probability of recommending B-Hive Honey when buying directly is 99% and lowest is 50% whereas when buying through a network, the maximum probability is 98% whereas lowest is 25%.



**Task 5 – Forecasting Sales**

When the sales data is plotted in a line graph, it is observed that there is a seasonality effect. It is seen that the sales increase from 1st quarter and peaks during 3rd quarter and sales starts declining in the final quarter. Overall, a linear growth is seen in sales over the years. However, in the year 2020, the sales have declined and in 2021 the growth isn’t seen as such. This might be due to the covid pandemic.

I have used a multiplicative time series model to forecast the sales for the next 4 quarters. This involves the following process:

1. Seasonal index calculation
2. Data de-seasonalising
3. Non- seasonal forecasting
4. Forecasting by re-seasonalising

After following the mentioned steps, the sales forecast for the next four quarter are as follows:

|  |  |
| --- | --- |
| 2022 Q2 | 2155.62 kgs |
| 2022 Q3 | 2654.86 kgs |
| 2022 Q4 | 1997.49 kgs |
| 2023 Q1 | 2212.78 kgs |

We can see that the seasonal effect is observed in the forecast as well. The stagnant growth in 2020 and 2021 seems to be recovering in the near future. The forecasting error needs to be analysed as well. The Mean Absolute Percentage Error (MAPE) is 15.71% which means that our forecast in the past is off by 15.71%. This may be similar in the future.

**Conclusion**

This report has addressed the entire task assigned in the minute. Overall, the order quantity is influenced by Loyalty, Quality, Social Media Presence, Competitive Pricing, Order Fulfilment and Shipping Cost. Distribution channel, quality, brand image and shipping speed influenced whether or not a customer will recommend B-Hive Honey. The next four quarter seems to have a good run with the same pattern repeating that is sales peaking in Q3 and declining in Q4. This analysis has limitations due to many factors such as larger dataset would have been better including other attributes like honey type, taste, packaging, google ratings. Also, irregularities play significant role in forecasting. Minimizing such noises could have improved the analysis.