**DSCI 724- Project Report**

**Introduction**

The objective of this project is to create a multiple linear regression model which is a model that is used to make predictions. The model used in this project aims to predict customer loyalty in retail settings based on several relevant marketing concepts used as predictors.

**Business problem description**

The model strives to solve the problem of optimizing customer loyalty in the retail environment. Improving customer loyalty is crucial because it positively contributes to the long-term success of a business by increasing the customer retention rate. Additionally, the model’s prediction aims to help the business formulate effective long-term resource allocation strategies and as a result avert potential harm to its bottom line. The predictive model achieves this by providing insights into the specific predictors that are predicted to significantly impact loyalty, allowing the business to channel its limited resources to focus on optimizing those predictors instead of focusing on irrelevant predictors.

**Data source**

For this project, the dataset is the result of a questionnaire on customer loyalty which was obtained from [data.world](https://data.world/cesarpolo/consumer-loyalty-in-retail). The data consists of 1949 observations and 8 variables. It was examined to validate whether it conformed to the standards of high-quality dataset. The data required minimal cleaning as it had the following issues: missing data and structural error in the form of misspelled variable name. To ensure completeness of the dataset, missing data were removed. Hence, the cleaned data has 1712 observations and 8 variables. The dataset had outliers which were not deleted since they might contain useful information.

**Variable description**

The variables consist of 6 predictors/independent and a response/dependent variable all of which are numeric and continuous except for price which is a discrete value.

The following are the predictors: Price, Quality, Community outreach, Trust, Customer satisfaction, and Negative publicity. The response variable is loyalty.

The dataset contains an identification variable called Customer ID which is excluded from the analysis. These predictors are expected to be relevant for predicting customer loyalty because they generally influence loyalty whether positively or negatively. The expected results are as follows:

* Higher prices are expected to have a lower predicted loyalty level or vice versa.
* Lower quality levels are expected to have a lower predicted loyalty level or vice versa.
* Higher community outreach efforts are expected to have a higher predicted loyalty level or vice versa.
* Higher levels of trust increase predicted loyalty level or vice versa.
* Higher levels of customer satisfaction increase predicted loyalty level or vice versa.
* Higher levels of negative publicity result in a lower predicted loyalty level or vice versa.

**Model specification & building**

The mathematical formula for multiple linear regression is Y = + ++ ··· + *​* + ϵ. To elaborate further:

* Y represents the dependent variable
* is a coefficient that represents the intercept. It is the value of the response variable when the independent variables equal zero.
* *​, ​,…,​* represents the coefficients that is linked to each independent variable. These denote the change in Y for a unit change in the associated independent variable while holding other variables constant.
* *, ​,…,​* represents the predictor variables
* represents the change between the observed and the model’s predicted value for the dependent variable. *Note: denotes the factors that impact the dependent variable but is not captured by the model.*

It should be noted that *, ​,…,​* are not known and can be estimated by fitting the model on the training dataset. The estimated multiple linear regression formula is which can be used to predict the dependent variable.

To fit an effective regression model, there are certain assumptions that the model needs to meet, namely:

* There should be a linear relationship between the response and predictor variables.
* The predictor variables should not be highly correlated with each other.
* The error terms should have a constant variance, be normally distributed, but should not be correlated.

The dataset was randomly split into 70% training and 30% testing dataset which is a crucial step because once the model is trained using the former, a testing dataset will be used to assess the model’s predictions/performance. An initial model was built using the training dataset to explore it, and the results were as follows:

Residuals:

Min 1Q Median 3Q Max

-1.84985 -0.34566 0.01977 0.37657 2.04028

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.22198 0.21833 -10.177 < 2e-16 \*\*\*

Price 0.33459 0.03738 8.952 < 2e-16 \*\*\*

Quality 2.54949 0.20271 12.577 < 2e-16 \*\*\*

`Community Outreach` 0.63607 0.11429 5.566 3.23e-08 \*\*\*

Trust 0.38728 0.04306 8.995 < 2e-16 \*\*\*

`Customer satisfaction` 1.07059 0.14841 7.214 9.65e-13 \*\*\*

`Negative publicity` -0.94149 0.10579 -8.900 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.575 on 1191 degrees of freedom

Multiple R-squared: 0.7505, Adjusted R-squared: 0.7493

F-statistic: 597.1 on 6 and 1191 DF, p-value: < 2.2e-16

The F-statistic for the model is large and its p value is <2.2e-16, indicating that at least one of the predictors is related to the dependent variable (Loyalty). Additionally, since the Pr(>|t|) for all the predictors is less than 0.05, we can conclude that each of these variables are statistically significant, meaning there is a relationship between each predictor and the response variable. Hence, they should be included in the final model. Lastly, the multiple R-squared value suggests that ***75% of the variability*** in Loyalty is explained by the model. The adjusted R-squared shows similar results, making the model appropriate for the data.

Finally, the correlation between the predictors was calculated using the ***VIF*** and the results were ***below 5*** which is the recommended threshold as shown below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Price | Quality | Community Outreach | Trust | Customer  satisfaction` | Negative  publicity |
| 4.486628 | 2.178125 | 1.240229 | 3.915604 | 1.645059 | 1.459119 |

In the next section, the diagnostic plots will assess the validity of the model by examining if the model assumptions are met.

A diagram of a number of values

Description automatically generated with medium confidenceA graph showing the value of a number of individuals

Description automatically generated with medium confidence**Diagnostic plots**

A black and white graph with numbers

Description automatically generatedA diagram of a diagram showing a number of values

Description automatically generated with medium confidence

The ***residuals vs fitted*** plot shows that the model does not perfectly hold the linearity assumption since the line is neither horizontal nor close to it. It shows that the model deviates from having a linear relationship between the predictors and the response variable as there is a bent in the middle of the plot and a curvature. Additionally, this curve suggests that the residuals are not spread constantly.

The ***Q-Q residuals*** i.e., the normal Q-Q plot confirms that the residuals/error terms close to being normally distributed.

The ***scale-location*** plot further confirms that the error terms do not have a constant variance across the predicted values. Since the line is slightly deviating from being horizontal, this suggests heteroscedasticity i.e., unequal variance.

In ***residual vs leverage*** plot, only one of the identified extreme outliers appeared which is 926. While the other two, 982, 474, were substituted by 1122 and 1111, implying that the value 926 requires close inspection. Additionally, the absence of cook’s distance dashed line indicates that all the points are within that distance, no observation has both high leverage and high residual.

***Solution***

Based on the above, *the linearity and equal variance* are violated. Hence, it is recommended to include *non-linear transformation* of the predictors and/or the response variable. For this dataset, since few of the predictors have a slight deviation from depicting a linear relationship with the response variable (as shown below), it was not appropriate to transform the response variable. Moreover, none of the transformations on the response variable improved the model fit. The below scatter plots give insight into the relationship between each predictor with the dependent variable.

A graph with black dots and a blue line

Description automatically generatedA graph showing the difference between price and stock

Description automatically generated

A graph showing a line graph

Description automatically generatedA graph showing a line that is dotted with black dots

Description automatically generated with medium confidence

A graph with black dots and a blue line

Description automatically generatedA graph showing a blue line and black dots

Description automatically generated

*Theoretically*, price and quality have synergy effect and thus it is appropriate to include an *interaction term*, potentially making the relationship between quality and loyalty linear. In general businesses aim to have an optimized combination of quality and price to increase customer loyalty. The impact of price on loyalty is contingent upon quality or vice versa. If quality is low but price is high, loyalty decreases or vice versa. These necessities the use of interaction terms. Furthermore, to confirm the need for interaction term, an interaction plot was utilized (as shown below) and the result shows that as quality increases the impact of price on loyalty increases. When quality is less, price has little effect on loyalty or vice versa, implying the presence of an ***interaction*** or ***synergy*** *effect.* Additionally, since trust variable has slight curve, a quadratic term is added.

A graph showing the quality of a product

Description automatically generatedA graph showing a line with black dots

Description automatically generated

After adding the interaction term, a new model and its summary is as follows:

Residuals:

Min 1Q Median 3Q Max

-1.63545 -0.31318 0.00164 0.34900 1.96802

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 10.79649 1.12948 9.559 < 2e-16 \*\*\*

Price -0.48788 0.12215 -3.994 6.89e-05 \*\*\*

Quality -5.49254 1.29785 -4.232 2.49e-05 \*\*\*

`Community Outreach` 0.62552 0.10923 5.727 1.30e-08 \*\*\*

Trust -1.78957 0.32342 -5.533 3.86e-08 \*\*\*

I(Trust^2) 0.18763 0.02698 6.953 5.87e-12 \*\*\*

`Customer satisfaction` 0.97383 0.14058 6.927 7.01e-12 \*\*\*

`Negative publicity` -0.58934 0.10746 -5.484 5.07e-08 \*\*\*

Price:Quality 0.93219 0.15008 6.211 7.25e-10 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5521 on 1189 degrees of freedom

Multiple R-squared: 0.7656, Adjusted R-squared: 0.764

F-statistic: 485.4 on 8 and 1189 DF, p-value: < 2.2e-16

Both the interaction term and the quadratic term are statistically significant since they had a P value (Pr(>|t|) of less than 0.05, suggesting its appropriateness. Additionally, the adjusted R-squared improved from 0.7493 to 0.764. The adjusted R-squared shows that this model can explain around **76%** of the variability in the response variable compared to the initial model that explained around **75%** of the variability. The multiple r-squared value also rose from 0.7505 to 0.7656. Additionally, the residual standard error was reduced by 0.0229. The below diagrams show the improved diagnostic plots where the residuals show homoscedasticity. The residual vs fitted plot shows a more linear relationship.

A graph showing the value of a product

Description automatically generated with medium confidenceA diagram of a number of values

Description automatically generated with medium confidence

A graph showing the difference between the average and the average

Description automatically generated

A diagram of a scale

Description automatically generated with medium confidence

The new mathematical formula for this final model is as follows:

+ 0.93219 Price x Quality + 0.18763 Trust^2

The coefficient estimate can be interpreted as follows: for the interaction term as Quality increases by one-unit, the effectiveness of Price on impacting Loyalty increases by 0.93219 or vice versa. The coefficient of price, for instance, can be interpreted as follows: While every other variable is held constant, increasing the price by a unit will decreases loyalty by on average. Additionally, Loyalty will equal if all the predictors equal zero.

Additionally, an ***analysis of variance table*** was included to further confirm the superiority of the second model as shown below:

Analysis of Variance Table

Model 1: Loyalty ~ Price \* Quality + `Community Outreach` + Trust + I(Trust^2) + `Customer satisfaction` + `Negative publicity`

Model 2: Loyalty ~ Price + Quality + `Community Outreach` + Trust + `Customer satisfaction` +`Negative publicity`

Res.Df RSS Df Sum of Sq F Pr(>F)

1 1189 362.42

2 1191 402.75 -2 -40.328 66.152 < 2.2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

The p value is less an 0.05 and closer to 0, indicating that the model is statistically significant and superior.

**Predictive results & Summary**

To evaluate the model’s performance and accuracy, adjusted R-squared and mean squared error (MSE) are used. The MSE provides insights into how much the predicted value deviates from the true/observed value. The R-squared value is 0.7020594 while the MSE value is 0.2967258. The calculated MSE value indicates that the squared difference between these two values is considered comparatively small. While the adjusted R-squared value showed a slight reduction, it is not a drastic reduction. Hence, the MSE and the R squared values suggest that the model can generalize well into new unseen data. Additionally, repeated K- fold cross validation was conducted to determine the performance of the model on new data. The R-squared value is similar to the value produced when the model was trained, suggesting the model has good performance.

Linear Regression

1198 samples

6 predictor

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 1078, 1078, 1078, 1078, 1078, 1078, ...

Resampling results:

RMSE Rsquared MAE

0.553442 0.7647635 0.4233664

Tuning parameter 'intercept' was held constant at a value of TRUE

From the model equation, it’s clear that trust and customer satisfaction have the highest impact on loyalty since they have the highest coefficients, impacting the predicted value. Then follows the synergy impact of quality and price. Hence the business should focus on these when optimizing the predictors to improve loyalty.

Appendix

**Work plan- Multiple linear regression (Predicting Customer loyalty)**

|  |  |  |
| --- | --- | --- |
| **Tasks** | **Estimated time** | **Actual time** |
| **Introduction:**   * Describe the project and what the model does. | **Week 11- 10/30**  20 minutes | 14 minutes |
| **Business problem description**   * What problem does the predictive model solve? * Advantages of solving the issue. | **Week 11- 10/30**  30 minutes | 49 minutes |
| **Data source**   * Specify the source of the data. * State the issues with the data and the cleaning methods used. * Provide overview of the data such as number of observations. | **Week 11-10/30**  30 minutes | 1 hour |
| **Variable description**   * List the variables and state their data type. * Comment on the relevance of each variable by providing information on how each variable is generally thought to impact loyalty. | **Week 12- 11/8**  40 minutes | 50 minutes |
| **Model specification & building**   * Define the assumptions. * Explain the mathematical formula. * Variable inflation factor | **Week 12- 11/10**  1 hour | 1 hour and 30 mins |
| **Diagnostic Plots**  **For the following discuss the plots:**   * Residuals vs Fitted * Normal Q-Q * Scale-Location * Residuals vs Leverage   This section included:   * Plot of Individual Relationships between Response and Each Predictor. * Interaction Terms (Indicate whether it’s used or not and explain the rationale). * Non-linear Predictor Transformations (explain inclusion/exclusion like the above). | **Week 13- 11/15**  1 hour and 30 minutes | 4 hours |
| **Predictive results**  **For the following discuss the results:**  Summary of Evaluation Metrics | **Week 14- 11/20**  1 hour and 45 minutes | 1 hour |
| **Summary**   * Summarize the results, highlight key insights, and provide recommendations. | **Week 15- 11/29**  1 hour and 30 minutes | 30 mins |
| **Appendix & Final Review**   * Include estimated and actual time results. * Review and finalize. | **Week 15- 12/1**  2 hours | 1 hour |