

Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

Key Decisions:

Answer these questions

Q1. What decisions need to be made?

Ans. Whether to send the catalog to these 250 people based on the expected profit. The expected profit from these customers should be atleast \$10,000.

Q2. What data is needed to inform those decisions?

Ans. We need data for the prediction of sales and calculations of the profit. In this project I need

These following data sets:

- Customer segment
- Average number of product purchased
- ScoreYes
- ScoreNo
- Margin and cost per catalog
- Average amount of Sale

Step 2: Analysis, Modeling, and Validation

Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

Important: Use the *p1-customers.xlsx* to train your linear model.

At the minimum, answer these questions:

Q1. How and why did you select the [predictor variables \(see supplementary text\)](#) in your model?

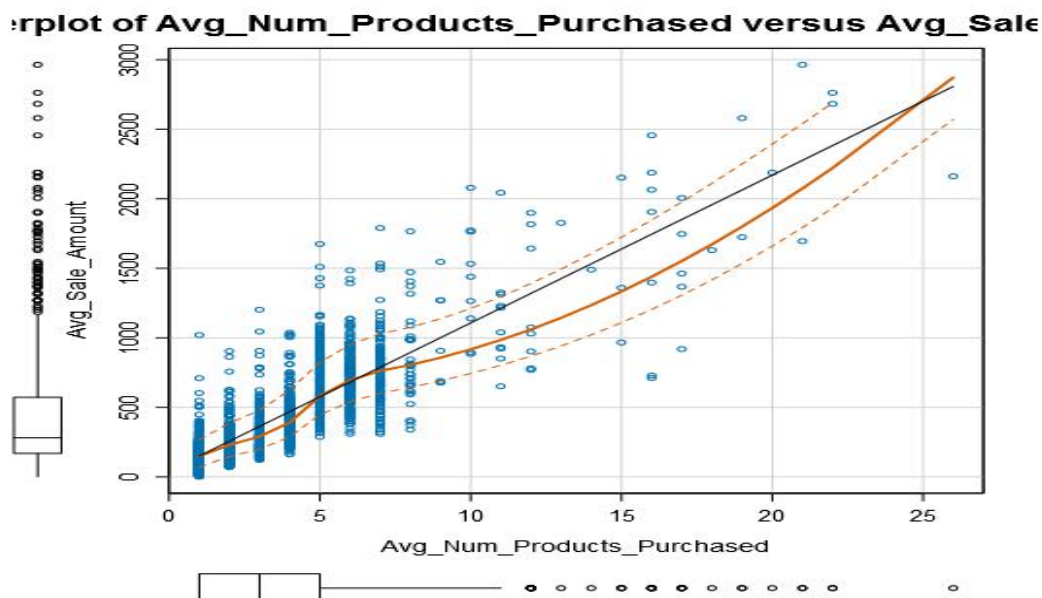
You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this [lesson](#) to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

Ans. In linear regression all the variable against the average number of products. As shown in the Figure(1) (Report for Linear Regression Model Predictor) only the customer segment and

the average value of the product have the p-value is 0.022 which means target variable is statistically significant because p-value is less than 0.05.

| Record Report | | | | | |
|---|------------|------------|----------|---------------|--------|
| Report for Linear Model predictor | | | | | |
| Basic Summary | | | | | |
| Call: | | | | | |
| lm(formula = Avg_Sale_Amount ~ Customer_Segment + Customer_ID + Store_Number + Responded_to_Last_Catalog + Avg_Num_Products_Purchased, data = the.data) | | | | | |
| Residuals: | | | | | |
| | Min | 1Q | Median | 3Q | Max |
| | -665.00 | -66.80 | -2.18 | 70.42 | 968.20 |
| Coefficients: | | | | | |
| | Estimate | Std. Error | t value | Pr(> t) | |
| (Intercept) | 4.275e+02 | 104.98523 | 4.0719 | 5e-05 *** | |
| Customer_SegmentLoyalty Club Only | -1.506e+02 | 8.97807 | -16.7725 | < 2.2e-16 *** | |
| Customer_SegmentLoyalty Club and Credit Card | 2.813e+02 | 11.91635 | 23.6055 | < 2.2e-16 *** | |
| Customer_SegmentStore Mailing List | -2.433e+02 | 9.82366 | -24.7636 | < 2.2e-16 *** | |
| Customer_ID | -1.703e-03 | 0.00294 | -0.5793 | 0.56241 | |
| Store_Number | -1.143e+00 | 0.99496 | -1.1486 | 0.25085 | |
| Responded_to_Last_CatalogYes | -2.849e+01 | 11.26689 | -2.5288 | 0.01151 * | |
| Avg_Num_Products_Purchased | 6.679e+01 | 1.51756 | 44.0085 | < 2.2e-16 *** | |
| Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 | | | | | |
| Residual standard error: 137.34 on 2367 degrees of freedom | | | | | |
| Multiple R-squared: 0.8374, Adjusted R-Squared: 0.8369 | | | | | |
| F-statistic: 1742 on 7 and 2367 DF, p-value: < 2.2e-16 | | | | | |
| Type II ANOVA Analysis | | | | | |

FIGURE(1) Linear Model Predictor



Figure(2) Scattarplot of Average Number Of Productrs Versus Average Sale Amount

Q2. Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

Ans: The Statistical result of the R-squared value is 0.8369 which is highly valued or consider to be a strong and the p-value of the predictor is less than 0.05 this means the target variable is considered to be Statistically Significant. The higher R-squared value explanatory power of the model. This value represents the amount of variation in the target variable explained by the variation predictor variables. Any Model with an adjusted R-squared value above 0.70 is considered to be a strong model.

Q3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

| Report for Linear Model predictor | | | | | |
|--|--------|-------------|------------|---------|---------------|
| Basic Summary | | | | | |
| Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data) | | | | | |
| Residuals: | | | | | |
| | Min | 1Q | Median | 3Q | Max |
| | -663.8 | -67.3 | -1.9 | 70.7 | 971.7 |
| Coefficients: | | | | | |
| | | Estimate | Std. Error | t value | Pr(> t) |
| (Intercept) | | 303.46 | 10.576 | 28.69 | < 2.2e-16 *** |
| Customer_SegmentLoyalty Club Only | | -149.36 | 8.973 | -16.65 | < 2.2e-16 *** |
| Customer_SegmentLoyalty Club and Credit Card | | 281.84 | 11.910 | 23.66 | < 2.2e-16 *** |
| Customer_SegmentStore Mailing List | | -245.42 | 9.768 | -25.13 | < 2.2e-16 *** |
| Avg_Num_Products_Purchased | | 66.98 | 1.515 | 44.21 | < 2.2e-16 *** |
| Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 | | | | | |
| Residual standard error: 137.48 on 2370 degrees of freedom Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366 F-statistic: 3040 on 4 and 2370 DF, p-value: < 2.2e-16 | | | | | |
| Type II ANOVA Analysis | | | | | |
| Response: Avg_Sale_Amount | | | | | |
| | | Sum Sq | DF | F value | Pr(>F) |
| Customer_Segment | | 28715078.96 | 3 | 506.4 | < 2.2e-16 *** |
| Avg_Num_Products_Purchased | | 36939582.5 | 1 | 1954.31 | < 2.2e-16 *** |
| Residuals | | 44796869.07 | 2370 | | |

FIGURE (3) Report of Statistical Result

Important: The regression equation should be in the form:

$$Y = \text{Intercept} + b_1 * \text{Variable}_1 + b_2 * \text{Variable}_2 + b_3 * \text{Variable}_3 \dots\dots\dots$$

Ans.

$$303.46 - 149.36 * (\text{Customer_Segment: Loyalty Club Only}) + 281.84 * (\text{Customer_segment: Loyalty Club and Credit Card}) - 245.42 * (\text{Customer_Segment: Store Mailing List}) + 0 * (\text{Credit Card Only})$$

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

Q1. What is your recommendation? Should the company send the catalog to these 250 customers?

Ans: Company should send the Catalog to these new 250 Customers.

Q2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

Ans. Expected revenue is determined by multiplying sale amount and the Score_Yes and the Gross Margin on all products sold through the catalog is 50%. 50% is deducted from the sum of expected revenue before the cost of the catalog (\$6.50) when calculating your profit.

Q3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

Ans. Expected Profit = (sum of expected revenue * gross_margin) - (250 * cost_of_catalog)
= 23,612.44 - 1,625
= \$ 21,987.44

