Christopher W. Hong

Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

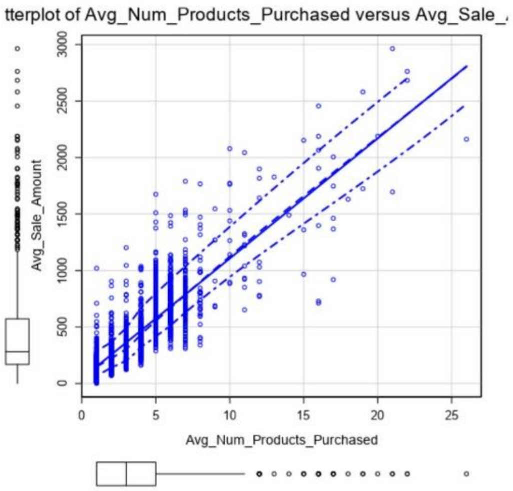
- Before sending out the catalog to the 250 new customers from their mailing list, the manager needs to determine whether the expected profit the company can expect from these customers exceeds $10,000 or not.

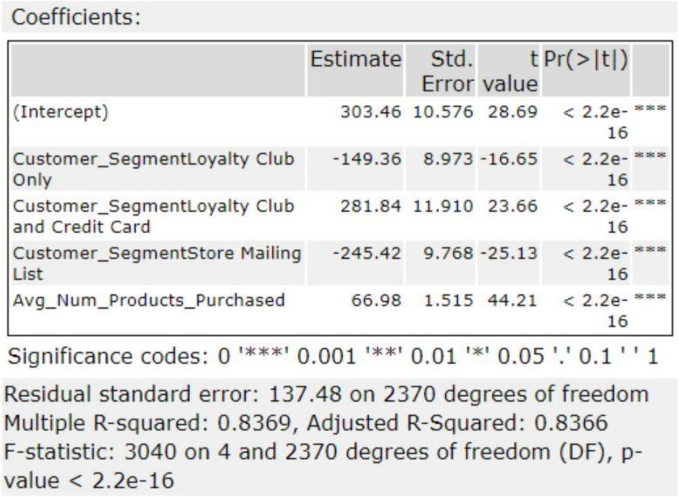
2. What data is needed to inform those decisions?

- Meta data of the 250 new customers such as the customer segment and the likelihood that these customers will respond to the catalog and make a purchase.
- Historical data of the company's existing customers such as their segments and sale amount so that a model can be built to predict these new customers' sale amount.

Step 2: Analysis, Modeling, and Validation:

1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.




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 degrees of freedom (DF), p-value < 2.2e-16

The scatter plot shows the individual variable Avg_Num_Products_Purchased and the target variable Avg_Sale_Amount. There is a line with positive slope in the plot, which means that these two variables are correlated and the individual variable is a good predictor for this target variable. Plus, the p-value of this individual variable is statistically significant at the level of 0.05.

Through constructing a linear regression model between the categorical variable Customer_Segment and target variable Avg_Sale_Amount, p-values of the individual categories of the categorical variable shown above are all statistically significant at the level of 0.05. Thus, the categorical variable is a good predictor for the target variable.

2. 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.

Since p-values of all the predictors shown above are statistically significant at the level of 0.05 and the Adjust R-Squared (0.8366) is high, the linear model is suggested as highly predictive.

3. 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)

Avg Sale Amount = 303.46 - 149.36 * (If Type: Loyalty Club Only) + 281.84 * (If Type: Loyalty Club and Credit Card) - 245.42 * (If Type: Store Mailing List) + 0 * (If Type: Credit Card Only) + 66.98 * Avg Num Products Purchased

Step 3: Presentation/Visualization

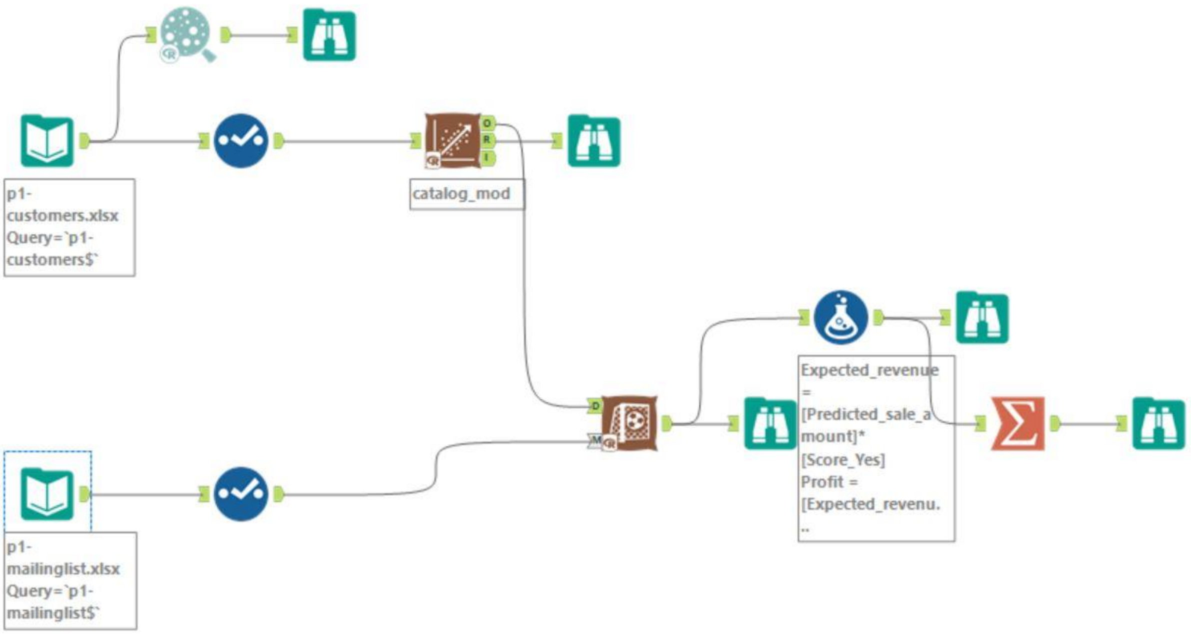
1. What is your recommendation? Should the company send the catalog to these 250 customers?

Since the predicted expected profit (around $21,987.44) is much greater than the threshold of $10,000, the company should definitely send the catalog to these 250 customers.

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

I arrived at the aforementioned number by using a formula from the regression model provided that was trained on historical customer data and applied it to the new customer data to predict the average sales amount. Then, I calculated the expected revenue of each new customer through multiplying the predicted average sales amount by the probability that a new customer will buy the catalog. Next, I calculated the expected profit of each customer through multiplying the expected revenue by average gross margin (0.50) and subtracted out that amount by the costs of printing and distributing ($6.50). Finally, I summed up the expected profit to get the amount $21,987.43.

Below is the screen shot of the workflow using Alteryx Designer:



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

The expected profit is about $21,987.44.