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

1. What decisions needs to be made?
   - A decision needs to be made whether or not to send a catalog to 250 new mailing list customers. The decision is based upon a prediction of additional profit. The recommendation can be made to send catalogs out to the new customers if predicted profit meets the minimum profit target of $10,000.

2. What data is needed to inform those decisions?
   - Two data sets are required to make a data-based decision. A training set with known outcomes with a variety of potential predictor variables and a known numerical target variable is required to build the model. A test set containing the same predictor variables in the predictive model for the new customers are required to calculate a predicted profit. In addition to historic and new customer data, we need the gross margin (50%), the cost of printing/sending the catalogs ($6.50/customer) and the probability of a new customer will place an order (score_yes).

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:

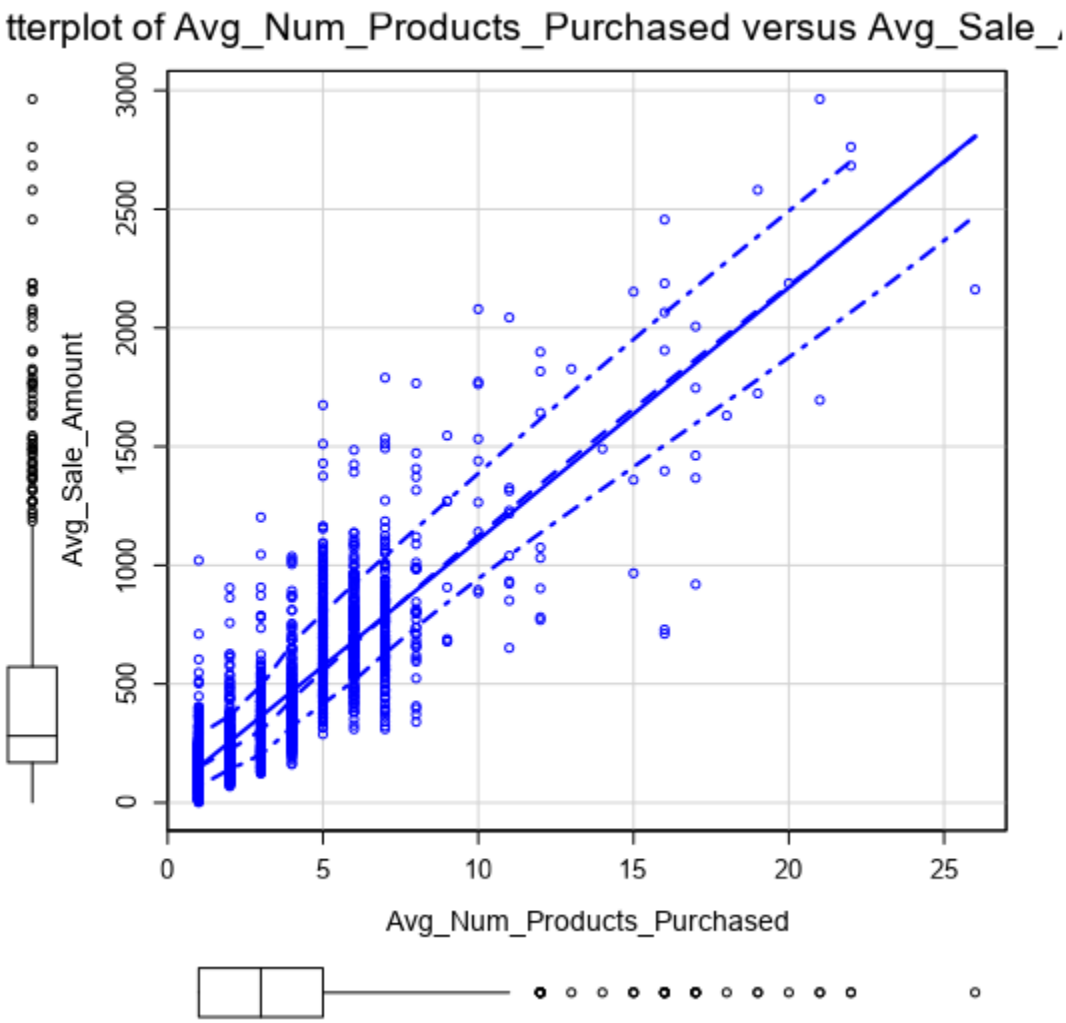
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use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

   - A scatter plot was created for each numerical variable in the training set. The requirements for predictor variables require a p-value for each coefficient no higher than 0.05 and an R-squared value close to one. The "Avg_Num_products_Purchased"

variable met this criterion. Visually you can get a good feel if a variable has a chance to become a predictor variable if you see a large slope in either direction.

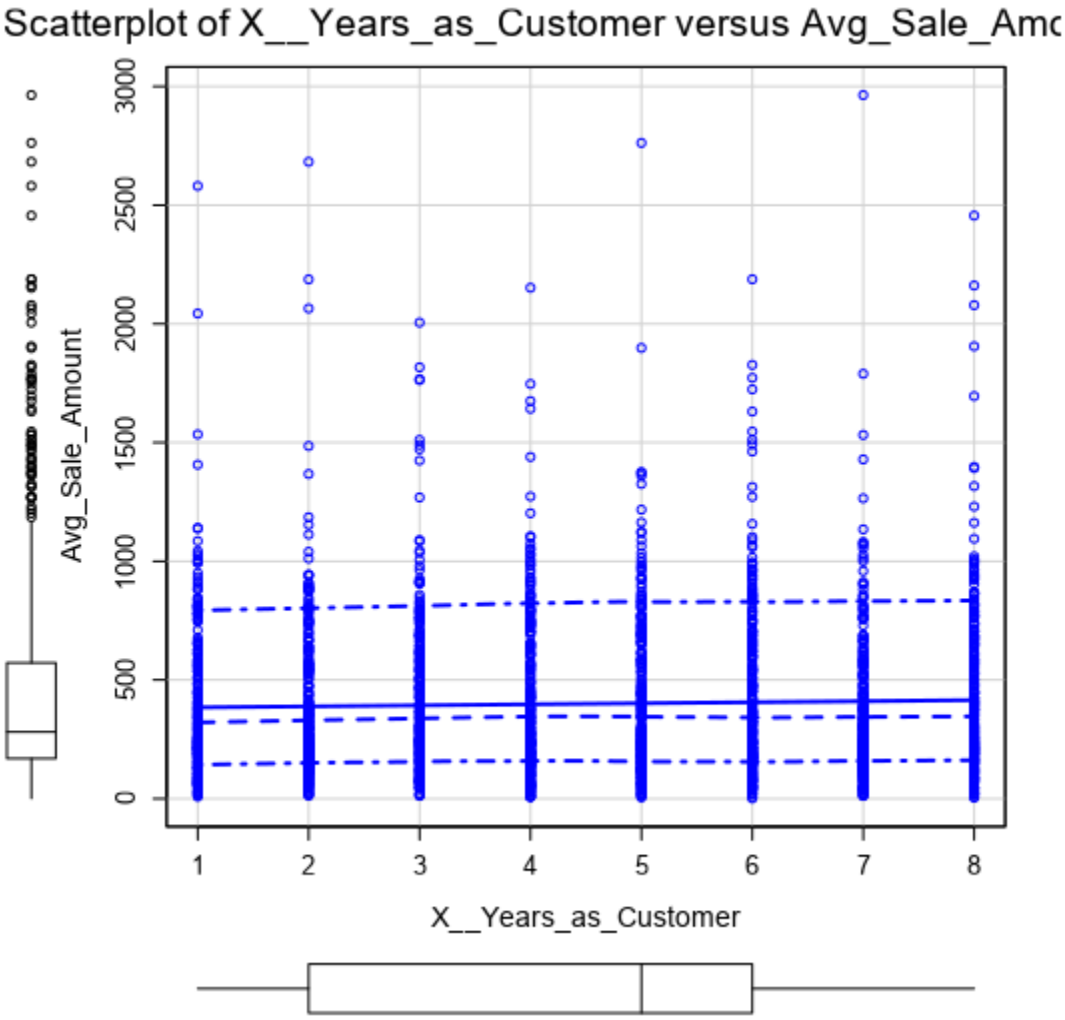
- Categorical predictor variables were selected by adherence to statistical standards for a good model. If the test statistics in a model did not meet the standard, a predictor variable was dropped until the standard was met.

Avg_Num_Products_Purchased vs Avg_Sale_Amount



Intercept p-value = $1.75 * 10^{-14}$
Avg_Num_Products_Purchased Coefficient p-value = $2.2 * 10^{-16}$
$R^2 = 0.7323$
This variable is a candidate predictor variable for the linear regression model.

#_Years_as_Customer vs Avg_Sale_Amount



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.

Two predictor variables are used in the model: Avg_Num_Products_Purchased and Customer_Segment. The p-values for each coefficient were less than 0.05, R^2 value is 0.8369, and adjust R^2 value is 0.8366.

Coefficient	p-value
(Intercept)	< 2.2e-16
Customer_SegmentLoyalty Club Only	< 2.2e-16
Customer_SegmentLoyalty Club and Credit Card	< 2.2e-16
Customer_SegmentStore Mailing List	< 2.2e-16
Avg_Num_Products_Purchased	< 2.2e-16

The R^2 value is an indicator of how much variation in the dependent variable (Avg_Sale_Amount) can be explained by the predictor variables (aka independent variables). This model explains 83.7% of the variation. An alpha of 0.05 was selected for the study. The alpha tells us that assuming the null hypothesis (coefficient = 0) is true, you can expect values to fall outside of the acceptable range of values 5% of the time. A p-value is the probability of a sample having **at least** the same effect in the sample assuming the null hypothesis is true. In other words, it's a test for compatibility between the null hypothesis model (at a given alpha) and the test sample. The two are considered equivalent if the p-value is greater than alpha. Otherwise the null hypothesis is rejected. All p-values in the table are below alpha. Therefore, the null hypothesis (coefficient = 0) is rejected and conclude each coefficient is not equal to zero.

Additional tests can be performed to test for strength in model such as checking for normal distribution of residuals, testing for consistent variance between test data and predicted values, and plotting independent variables against residuals. This analysis is not included in the report, but it is worth mentioning.

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)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

Avg_Sale_Amount = 303.46 + 66.98 * Avg_Num_Products_Purchased
-149.36 (If Customer_Segment: Loyalty Club Only)
+281.84 (If Customer_Segment: Loyalty Club and Credit Card)
-245.42 (If Customer_Segment: Store Mailing List)
+ 0 (If Customer_Segment: Credit Card Only)

Step 3: Presentation/Visualization

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

At the minimum, answer these questions:

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

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
I used the formula tool after building the model to multiply the “Score_Yes” by “Score”. The result is the probable sales for each of the new customers. I then used the summary tool to sum the new result for all customers (Total_Predicted_Sales) and count the number of customers. I followed up with another formula tool to create “Predicted_Profit” and used the following formula:

Round(([Total_Predicted_Sales] * 0.5) – (6.50 * [Customer_Count]) , 0.01).

The predicted profit was greater than $10,000. Therefore, the minimum target was met, and I could recommend sending the catalogs to the new customers.

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
The predicted profit is $21,987.44.

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.