Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

Sending this year's catalog out to 250 new customers from the mailing list if the expected profit contribution exceeds $10,000 or not sending out the catalogs if it does not exceeds.

2. What data is needed to inform those decisions?

First, we need previous customer data so we could decide on the target and predictor variable. In this case, we need data about the average sale amount, whether the customer responded to the last catalog, average number of products purchased, and number of years as customer. We also need the same data about the new customer on predictor variable values so we could apply the model to validate the sample. Then we need data for the costs of printing and distributing per catalog and the average gross margin on all products sold through the catalog to calculate the expected revenue.

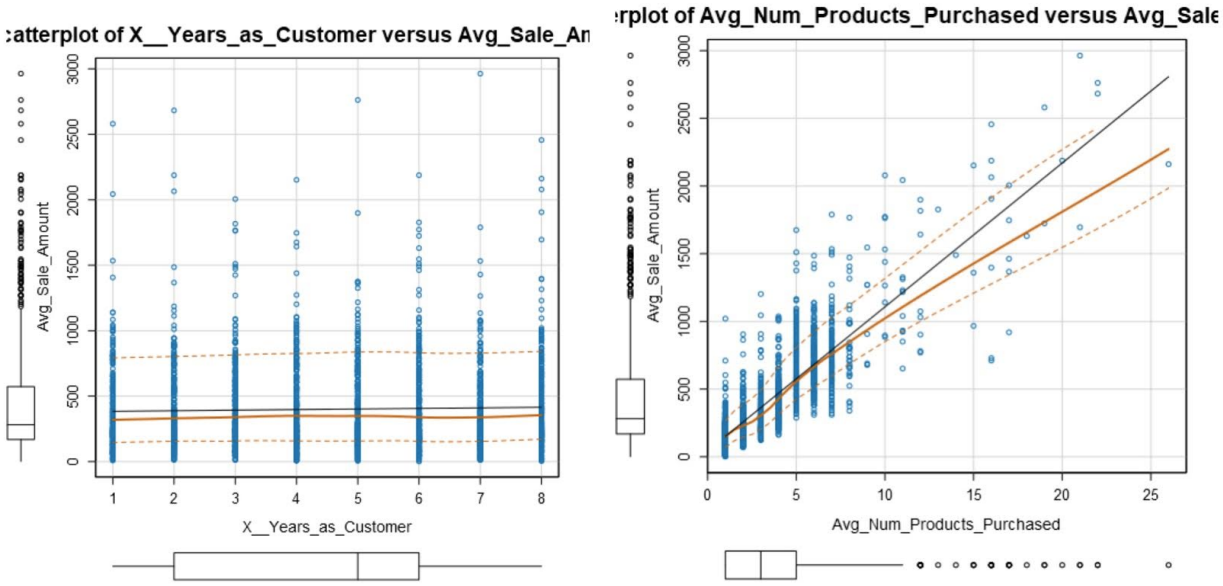
Step 2: Analysis, Modeling, and Validation

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.

1. 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.

I chose Customer_Segment as the non-numeric predictor variable. Plus, I initially chose Numeber_Years_as_Customer and Avg_Num_Products_Purchased as the predictor variables because both have continuous numeric values and are likely to help predict sales. I created a scatter-plot between the variable Numeber_Years_as_Customer and Avg_Num_Products_Purchased



The scatter plot graph for Numeber_Years_as_Customer vs. Avg_Sale_Amount shows hardly any slop therefore there is not much of a relationship between the predictor and target variable. On the other hand, as Avg_Num_Products_Purchased increases the Avg_Sale_Amount also increases. There is a slope to the line and this indicates this is a good predictor variable for the target variable. In conclusion, I will only use Customer_Segment and Avg_Num_Products_Purchased as my predictor 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.

	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

The linear model that I have created is a good model because all of the p-values are less than 2.2e^-16 which the relationship between predictor and target variable is considered statistically significant. Also my model has a R-squared value of 0.8369 and the adjusted R-squared value of 0.8366 which has high explanatory power of the model for the amount of variation in the target variable explained by the variation in the predictor variables.

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)

Exp_Avg_Sale_Amount = 303.46 – 149.36 * (Customer_SegmentLoyalty Club Only) + 281.84 * (Customer_SegmentLoyalty Club and Credit Card) – 245.42 * (Customer_SegmentLoyalty Mailing List) + 66.98 * (Avg_Num_Products_Purchased)

Expected_Profit = Σ[Exp_Avg_Sale_Amount * Score_Yes] * 0.5 – (6.5 * 250)

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 would recommend to sending out the catalog to 250 new customers from the mailing list since the expected profit contribution exceeds $10,000.

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

I multiply the Exp_Avg_Sale_Amount by Score_yes to figure out the expected revenue from each of the 250 customers on the mailing list. I sum up all of the expected revenue for all 250 customers. I then multiply by the gross margin which is 50% and subtract the total costs of printing and distributing which is (6.5 * 250).

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

The expected profit from the new catalog is $21,987.44