Project 1: Predicting Catalog Demand

## **Step 1: Business and Data Understanding**

### **Key Decisions:**

*Answer these questions*

1. What decisions needs to be made?

The decision is whether to send out catalog to 250 customers. This will depend on whether they will bring in profit of above $10,000 or not.

2. What data is needed to inform those decisions?

We need data on the revenue of each customer that we aim to target. This should be the average revenue times and the probability of 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.***

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.

First, I have graphed the relationship between the numerical variables and the dependent variable. I have found that there is a linear relationship between average number of products purchased and average sale, but there isn’t a linear relationship between years as customer and dependent variable.

I have also excluded irrelevant variables: name, customer\_ID, address, state (as they were all in one state),and zip. I have found that store\_number, and most cities were insignificant.

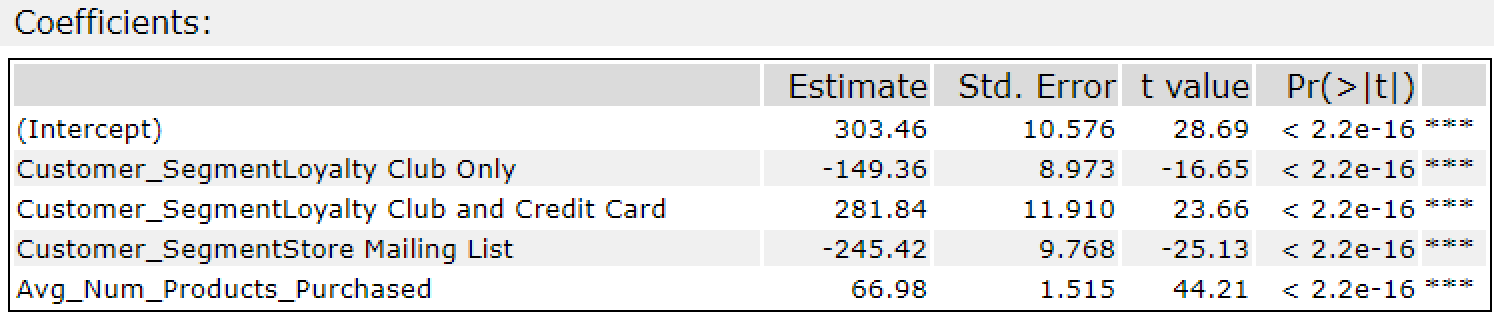
Chart, scatter chart

Description automatically generatedA picture containing chart

Description automatically generated

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.

I believe the model is a good model because it has a high adjusted R2 (0.8366), which measures how much the variations to real values are explained by the model. All of the predictor variables have P value significantly lower than the threshold (0.05), which means that they are significant.



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)

*Y = Intercept + b1 \* Variable\_1 + b2 \* Variable\_2 + b3 \* Variable\_3……*

The best linear regression is:

Avg\_Sale\_Amount = 303.46 -149.36 \* Customer\_SegmentLoyalty Club Only + 281.84\* Customer\_SegmentLoyalty Club and Credit Card - 245.42\* Customer\_SegmentStore Mailing List + 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?

The company should send the catalog as expected profit of $21,987 is above $10,000 that is required.

2. How did you come up with your recommendation?

I have calculated the profit per customer by multiplying the probability of yes with predicted average sale amount in order to get expected revenue per customer, then multiplied by 0.5 (gross margin 50%), and then subtracted $6.5 of costs of printing and distributing for each customer. Finally, I summed the net profit for each customer to get $21,987. This value is greater than $10,000 required.

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

The expected profit is $21,987