Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?

The primary decision that needs to be made is that will there be enough profit if the company sends out the catalog. This decision leads to us predicting the total profit that will be generate if the catalogs are sent out. The final profit value will determine the decision that will be made. Prediction of the profit can be done using the data about the customers. (Explained in the next answer)

2. What data is needed to inform those decisions?

The data needed to inform these decisions is about the existing customers, their average sale amount, the number of products purchased and other related data. Another important piece of data is the probability that the customer will buy the product if the catalog is sent. Using the above data we can calculate the amount that will be purchased by each customer and thus enabling us to calculate the gross profit (after subtracting the cost to send the catalog).

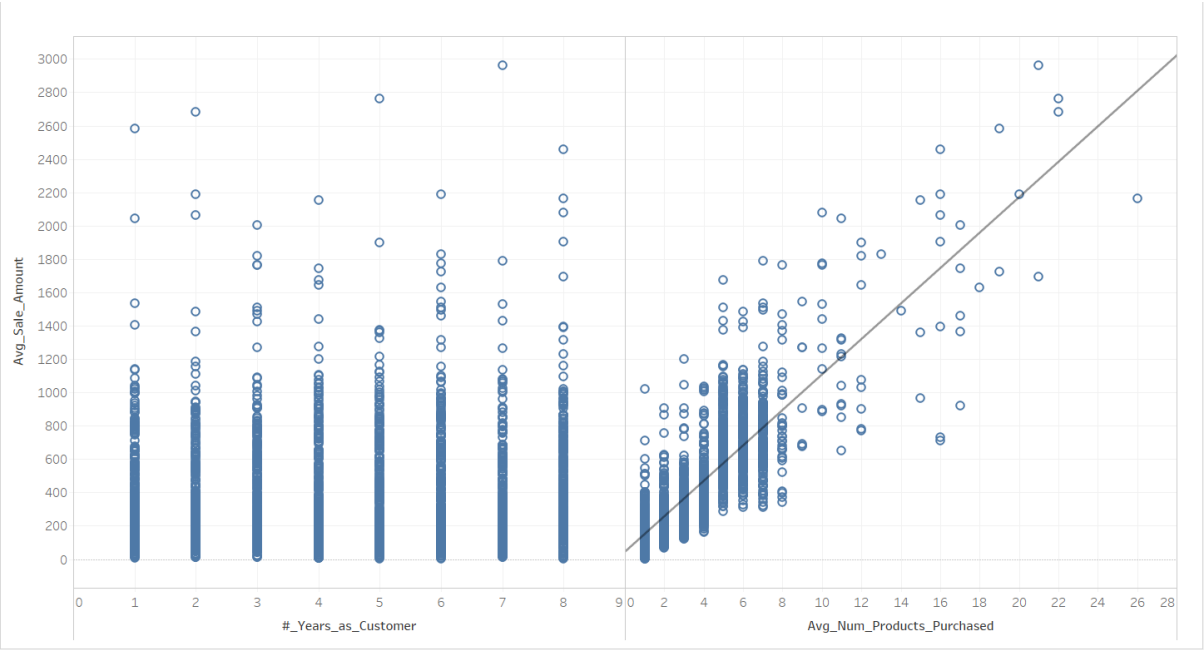
Step 2: Analysis, Modeling, and Validation

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.

There are two continuous variables namely:

a. Average number of products purchased
b. Number of years as a customer

We make scatter plot for both of these variables with respect to Average Sale Amount.



In the above plot we can clearly see that the Number of Years as a customer has no linear relation to Average Sale Amount. Due to this fact we do not include it in the regression model. This is further justified as the p value for No of Years in the regression model exceeds 0.05.
Next we can see there is a clear linear relation between the Number of Products purchased and Average Sale Amount. The trend line denotes the same. Hence we include this variable in our model.

As for other variables Customer Name and ID can make no contribution to the model and hence are not considered.

City, State can be condensed into ZIP code, however it has a very high p value in the model and hence is not statistically significant and is discarded.

Customer Segment is a categorical variable and shows a good p value in the analysis and is thus considered in the final model.

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.

The Adjusted R-squares value comes out to be 0.8366. This is a very good R-squared values that suggests that model is highly predictive. Thus we can conclude that the linear model is a good model.

	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 ***

The p value for Avg Number of Products Purchased is close to 0 making it highly likely that this variable is related to Average Sale Amount.
Similarly for Customer segment the p value is close to 0 and hence a good variable to include in the model.

Thus based on the above values we selected the two variables in our model.

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 = 303.46 - 149.36*(Loyalty Club Only) + 281.84*(Loyalty Club and Credit Card) - 245.42*(Store Mailing List) + 0*(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?

Yes, the company should 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)

The recommendation was arrived at by the following process:

a. The decisions to be made were scrutinized.
b. The data was analyzed.
c. Scatter plots were made for continuous variables to check for their significance in the model.
d. Trial and error was followed for categorical variables using Alteryx. Based on the p values the relevant variable was selected.
e. The regression model was trained using the training set with the identified (significant) variables using Alteryx.
f. The model was used to determine the Average sale amount for the 250 customers in the test data.
g. This predicted output was multiplied with the score_yes and cost of shipping catalog was subtracted.
h. Total Profit was determined for the 250 customers.

The total profit comes out to be around $22,000 which is way above the $10,000 margin set by the company. Hence the sending of catalog should be done.

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

The expected profit from this new catalog is **21,987.44$**.
This was calculated by multiplying the probability that the customer will respond to the catalog with the Average Sale Amount, then multiplying with 0.5 as the gross margin of the company and also subtracting 6.50$ which is the cost of shipping.
[(Score_Yes x Predicted_Sale_Amount x 0.5) – 6.5]
This is done for every customer and then the total is calculated for 250 customers.