Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?

According to a linear regression model, the expected total profit from sending a catalog to the new 250 customers is $21,987.44.
As it exceeds the company amount criteria which is $10,000.00, the company should decide to send catalogs to these customers.

I reached to above result by applying the following steps:

- I built a linear regression model with two significant predictors:
  Predictor variables: [Customer_Segment],[Avg_Num_Products_Purchased])
  Target variable : [Avg_sale_amount]
- I applied the regression formula to the new data set with the Score tool.
- I multiplied the predicted Revenue per individual by probability that a person will buy the catalog (Field: [Score_Yes])
- I multiplied the resulted value from previous step by the gross margin which is (%50) then subtract out the $6.50 cost to calculate profit per individual.
  Expression: (([Revenue] * 0.50) – 6.50)
- Finally, I summarized the profit value with Summarize tool. the result was 21987.4356865455.

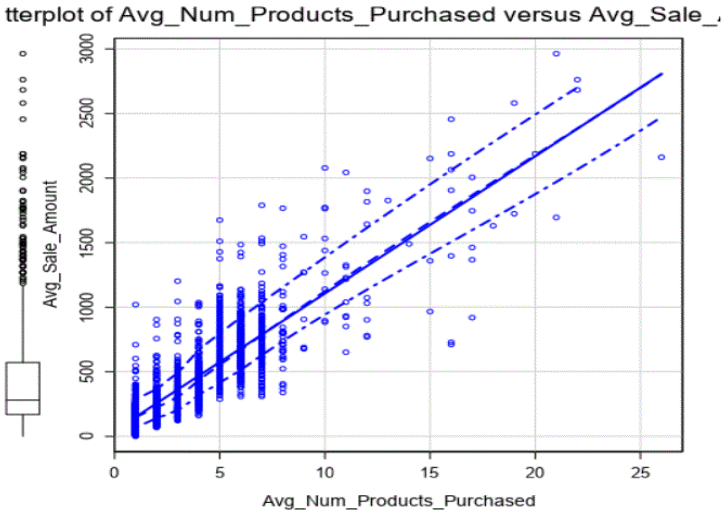
2. What data is needed to inform those decisions?

According to the linear regression and checking the P-value lower than 0.05, I figured out that only two variables are very good candidate to be statistically significant.
one of predictors is categorical variable [Customer_Segment] , and the other variable is a continuous [Avg_Num_Products_Purchased].
In addition, the probability that a person will buy(field [Pay Yes]) is needed to calculate the expected revenue per person.

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.

I draw a scatterplots for each of potential continuous predictors to decide which of them has a positive or negative relationship to the target variable. I found that the Avg_Num_Products_Purchased is positively related to the target variable, as shown below:



For categorical variable to check if there is a correlation, I just added them to linear regression tool and checked the p-value. Since a Customer_Segment variable has p-value <0.05 for all its individual categories, there is an obvious correlation between this categorical variable and the target variable.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.685e+03	2149.8261	-0.7836	0.43334	
Customer_SegmentLoyalty Club Only	-1.503e+02	8.9708	-16.7525	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.825e+02	11.8968	23.7483	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.431e+02	9.8171	-24.7664	< 2.2e-16	***
ZIP	2.627e-02	0.0266	0.9873	0.3236	
Store_Number	-9.991e-01	1.0052	-0.9939	0.32036	
Responded_to_Last_CatalogYes	-2.870e+01	11.2709	-2.5467	0.01094	*
Avg_Num_Products_Purchased	6.679e+01	1.5151	44.0838	< 2.2e-16	***
X._Years_as_Customer	-2.325e+00	1.2216	-1.9033	0.05712	.

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 my model is a good model because the p-value of predictor variables are lower than 0.05, and the R-squared and adjusted R-Squared values are 0.8377, 0.8372 accordingly.

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).

The equation is:

Average sale amount = 303.46 + 66.98 * (Avg_Num_Products_Purchased) + -149.36*(if Loyalty Club Only) + 281.84*(if Loyalty Club and Credit Card) + -245.42*(if Store Mailing List)

Step 3: Presentation/Visualization

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

Since the expected profit for sending catalog to the 250 new customers is $21,987.44, I strictly recommend the company to send catalog to them.

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

I came up with my recommendation by analyzing the past 2375 records of customer data, with the help of linear regression tool, I applied the resulted equation to the new 250 records of new customers with Score tool , the Score tool applied linear equation into each row to produce a predicted revenue per person, I then used Formula tool to multiply each predicted value by both probability the customer will buy, the gross margin of %50, then subtracted the resulted value of $6.50 cost, Finally I used summarize tool to calculate the total expected profit for sending catalog to all new 250 customers.

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.