Project 1: Predicting Catalog Demand

Business and Data Understanding

1. What decisions needs to be made?

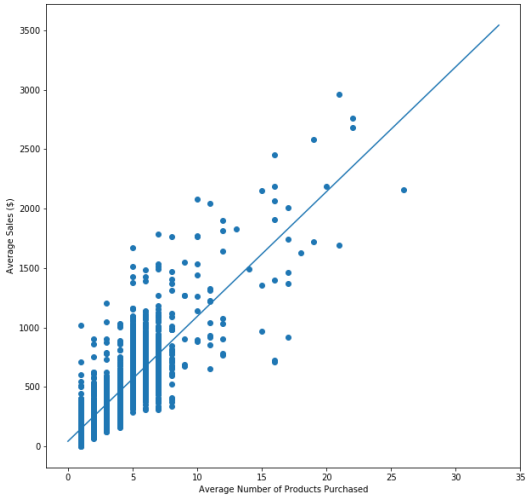
Should the company send the catalog out to the new customers?

2. What data is needed to inform those decisions?

The expected profit contribution as a result of sending the catalogs to the new customers.

Analysis, Modeling, and Validation

We chose two predictor variables: '`Avg_Num_Products_Purchased`' & 'Customer_Segment'. Below is a scatter plot showing the relationship between Average Sales Per Customer and Average Number of Products.



We have used a dummy variable (1 hot encoding) to transform the categorical data from the 'Customer_Segment' column. To be more precise; we looked Customer_Segment unique values which are: '`Credit Card Only`', '`Loyalty Club Only`', '`Loyalty Club and Credit Card`', and `Store Mailing List`, and transformed them to 3 numerical columns instead of 1 categorical column. Each one of the three columns represents a unique category with either a 1 or a 0, and together they represent the remaining (fourth) category (in this case 'Store Mailing List'). For example: looking at the sample of the predictor variables in the table below, each case represents a customer belonging to one 'Customer_Segment' category. The reason why did not add a specific column for the '`Store Mailing List`' is that we and more importantly the linear regression model can infer the value of Store Mailing List as having the rest customer segment columns be 0.

Index	Avg_Num_Products_Purchased	Credit Card Only	Loyalty Club Only	Loyalty Club and Credit Card
0	4	0	0	1
1	7	0	1	0
2	6	1	0	0
3	2	0	0	0

By looking at the tables below, we can tell that the R-squared score is fairly good. A score of 0.837 tells us that about %84 of the variance is explained by explained by our function (coefficients and predictor variables).

Furthermore, the P-Values for each predictor variable are very close to zero and much smaller than 0.05. Which is the value of P usually used to indicate %95 confidence.

R-squared	0.837

Predictor Variable	**P-Value**
Avg_Num_Products_Purchased	7.989877e-312
Credit Card Only	1.050298e-123
Loyalty Club Only	3.487456e-34
Loyalty Club and Credit Card	1.497675e-247

Average Sales prediction function

Y = 58.05 + Avg_Num_Products_Purchased * 66.98 + Credit Card Only * 245.42 + Loyalty Club Only * 96.06 + Loyalty Club and Credit Card * 527.26

Presentation/Visualization

We strongly recommend sending the catalogs to our new customers because printing and distributing the catalog for our 250 customers is predicted to bring $21,987 in profit. Based on our data from over 2000 customers that we have sent catalogs to in the past.

By knowing their customer segment and the average number of products they bought, the linear regression model could predict the average sales for our new customers.

We calculated the predicted average sales for each new customer and took into account the cost of printing and distribution ($6.5) of the catalogs and our gross profit margin (%50) in order to reach the conclusion it could bring $21,987 in profit.