Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 words limit)

Key Decisions:

Answer these questions

1. What decisions needs to be made?

My company wants to send out catalogue of our new products to a mailing list of our 250 new customers. I have been tasked to make an estimate of the profit to be generated from the new customers if the catalogue is sent out to the customers. The catalogue should be sent out only if the profit expected from the customers' purchase exceeds $10,000. This profit will be determined using the data available for 2,300 old customers and their response to previous catalogue sent to them.
The profit expected from the 250 new customers should exclude the cost of printing and distributing each catalogue to each customer and should also take into account customers who will not respond to the catalogue that is, customers who will not make a purchase.
From the result of the analysis, the company will be able to decide whether to send the catalogue of products to the 250 new customers or not.

2. What data is needed to inform those decisions?

The data that needs to inform this decision is customers response to catalogue sent out to previous customers and the expected profit to be made after analysis.

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.

At the minimum, answer these questions:

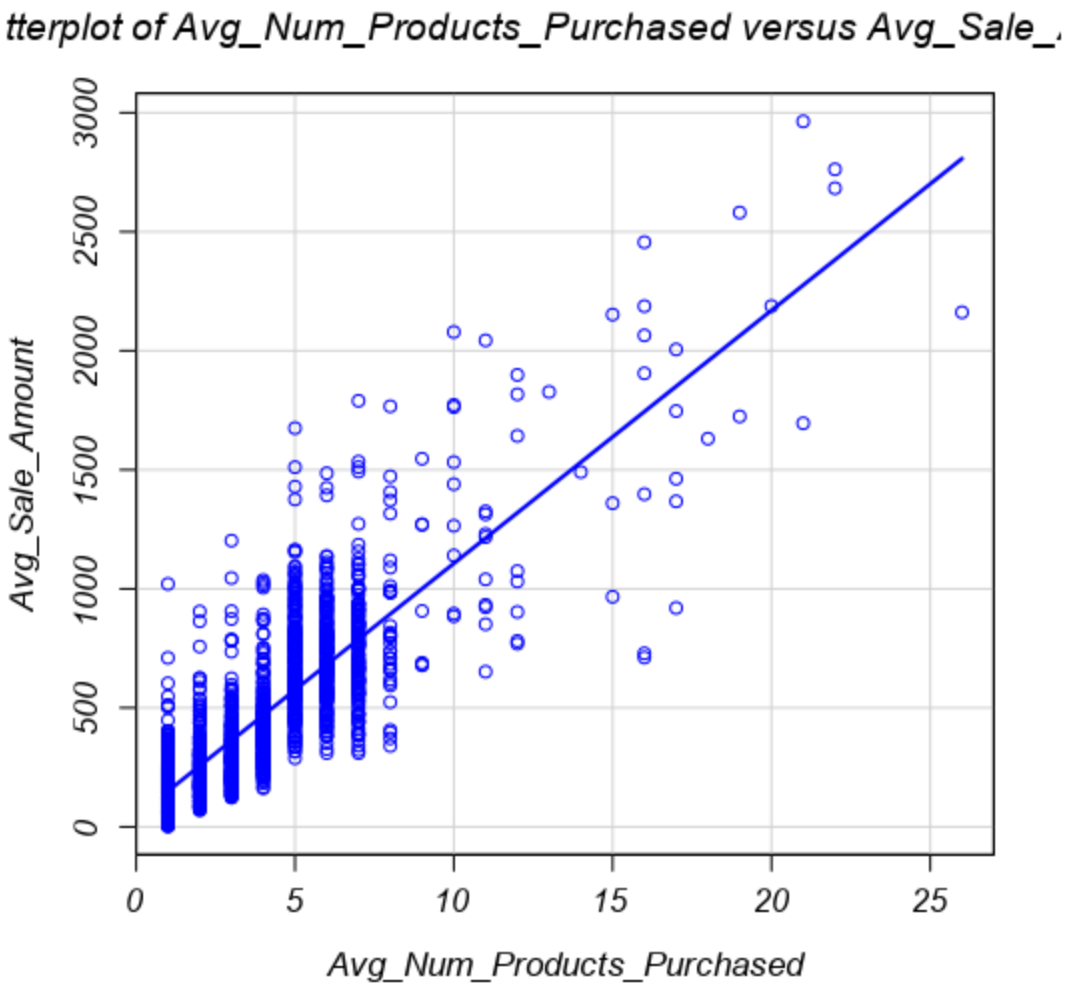
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

To ensure that I chose the right predictor variables, I ran a linear regression of all the possible predictor variables (Customer_Segment, Store_Number, Avg_Num_Products_Purchased and #_Years_as_Customer) against the target variable (Average_sales_amount) except Responded_to_Last_Catalog because it is not available in the new customer dataset and as such cannot be used in the model. Only Customer_segment and

Avg_Num_Products_Purchased variables shows a statistical significance to the target variable with a P value less than 0.05.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	431.852	104.9602	4.114	4e-05	***
Customer_SegmentLoyalty Club Only	-149.540	8.9763	-16.659	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.610	11.9095	23.730	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.922	9.7695	-25.173	< 2.2e-16	***
Store_Number	-1.127	0.9951	-1.132	0.25759	
Avg_Num_Products_Purchased	66.959	1.5152	44.192	< 2.2e-16	***
X._Years_as_Customer	-2.353	1.2229	-1.924	0.05449	.

I dropped the variables with no statistical significance to the target variable then ran a scatter plot of Avg_Num_Products_Purchased against Avg_Sales_Amount and the plot shows a positive linear relationship between the target variable and the predictor variable.



I could not run a scatter plot of Avg_Num_Products_Purchased against Customer_Segment because Customer_Segment is a categorical variable and cannot be represented on a scatter plot.

The predictor variables I used are Avg_Num_Products_Purchased and Customer_Segment and the target variable is Avg_Num_Products.

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.

From the linear regression analysis, the predictor variables show a good significance to the target variable with each variable having a P value less than 0.05 which is considered to show a good relationship between the predictor variable and the target variable.

The model is a good model because the R squared value is 0.8369 and adjusted R squared value is 0.8366. This is considered to be a good model because linear regression model with multiple predictor variables with an adjusted R squared closer to -1 or 1 is consider good to be used.

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 best linear equation based on the available data is
Y = 303.46 + 281.84(customer_segment: Loyalty Club and Credit Card) - 149.36(customer_segment: Loyalty Club Only) - 245.42(customer_segment: Store Mailing List) + 66.98(Avg_Num_Products_Purchased)

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?

From my analysis, having carefully studied the data available, selected the variables that are needed for the analysis and then analyzing the data available, I recommend that the company should send the catalog to the 250 new customers because the predicted response of the new customers towards the catalog shows a profit above the company's expected profit.

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

Business problem:
The company wants to send out a catalog of its high-end furniture to 250 new customers it recently got their contact. The company is skeptical on sending the catalogue to the new customers because it wants to send the catalog to the customers only if the expected profit from the customers exceeds $10,000.

Available data:

The data available for the analysis is the data of the previous sales made to the company's 3000 customers. The data includes customer name, customer id, customer address, customer city, customer zip code, customer city, store number, the average number of products purchased, number of years as a customer, whether or not the customer responded to the last catalog sent to him and customer segment which is either loyalty club only, credit card only, loyalty club and credit card or store mailing list. This is the data I used to train the prediction model after which I applied it to the data of the new customers which has customer name, customer id, customer address, customer city, customer zip code, customer city, store number, the average number of products purchased, number of years as a customer, probability of the customer to make a purchase and probability of the customer not to make a purchase.

Training the model:
To ensure that I chose the right predictor variables, I ran a linear regression of all the possible predictor variables (Customer_Segment, Store_Number, Avg_Num_Products_Purchased and #_Years_as_Customer) against the target variable (Average_sales_amount) except Responded_to_Last_Catalog because it is not available in the new customer dataset and as such cannot be used in the model. Only Customer_segment and Avg_Num_Products_Purchased variables shows a statistical significance to the target variable with a P value less than 0.05.
I dropped the variables with no statistical significance to the target variable then ran a scatter plot of Avg_Num_Products_Purchased against Avg_Sales_Amount and the plot shows a positive linear relationship between the target variable and the predictor variable.
I could not run a scatter plot of Avg_Num_Products_Purchased against Customer_Segment because Customer_Segment is a categorical variable and cannot be represented on a scatter plot.
Then I used Avg_Num_Products_Purchased and Customer_Segment as the predictor variable and ran a linear regression against the target variable which is Avg_Num_Products. After training the model, I then applied the model to the dataset containing the 250 new customers.

Profit calculation:
After applying the trained model to the new dataset, I got the predicted amount each customer will spend assuming they all have to make a purchase. I then multiplied the probability of each customer to make a purchase with the amount each customer will spend to get the actual that amount the company will get from each customer that receives the catalog.
The average gross margin on all products sold through the catalog is 50% and the cost of printing and distributing each catalog is $6.50. Having gotten the amount from each customer, I multiplied it by the average gross margin which is 50% then subtracted the $6.50 from the amount the company will get from each customer. Lastly, I summed the amount from each customer to get the expected profit from distributing the catalog to each of the 250 new 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 about $21,987 assuming the catalog is sent to these 250 customers