Project 1: Predicting Catalog Demand

https://classroom.udacity.com/nanodegrees/nd008/parts/c0b53068-1239-4f01-82bf-24886872f48e/project

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

In the case of this problem statement we have 250 new customers to whom we must send the catalog to. There is a probability that the customer will have likely to buy from us and use our products which generates revenue. The decisions that need to be made at this point is to maximize revenue.
For this we will have to first select the related variables which will eventually correspond to calculating the revenue and finally, the profit for the catalogs sold.
We will also have to take the probability of the goods purchased by the targeted customers - we will have to consider when we calculate revenue as not all customers will require anything they see in the catalog. The company will send the catalogs to these customers only if the profit earned from them is above 10,000 USD.

2. What data is needed to inform those decisions?
   - The previous year's data that includes about 2300 customers and what type of service/product they have ordered, how many of them have they placed an order for, the cost for each being a customer.
   - We will Need the 'customer (avg sale amount)', so we can use it to compute how much we expect to earn from these new 250 catalogs.
   - We should also know the probability that each customer will buy from the catalog (score_yes variable in this dataset).

Step 2: Analysis, Modeling, and Validation

1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

At first, I selected all the variables that could have been potential predictor variables using average salary as my target variable. On closer inspection, the Pr|t| value of most of the variables like store number, city, customer_ID and #_of_years_customer from which i inferred

they were not statistically significant and can be removed from the model.

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Customer_ID + Store_Number + Responded_to_Last_Catalog + Avg_Num_Products_Purchased + X._Years_as_Customer, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-664.40	-67.90	-2.62	70.52	972.60

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.393e+02	1.051e+02	4.179	3e-05	***
Customer_SegmentLoyalty Club Only	-1.503e+02	8.974e+00	-16.752	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.821e+02	1.192e+01	23.670	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.434e+02	9.818e+00	-24.786	< 2.2e-16	***
Customer_ID	-1.622e-03	2.939e-03	-0.552	0.58102	
Store_Number	-1.159e+00	9.944e-01	-1.166	0.24388	
Responded_to_Last_CatalogYes	-2.829e+01	1.126e+01	-2.512	0.01206	*
Avg_Num_Products_Purchased	6.683e+01	1.517e+00	44.055	< 2.2e-16	***
X._Years_as_Customer	-2.316e+00	1.222e+00	-1.895	0.05819	.

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.27 on 2366 degrees of freedom
Multiple R-squared: 0.8377, Adjusted R-Squared: 0.8371
F-statistic: 1526 on 8 and 2366 DF, p-value: < 2.2e-16

Only average number of products purchased and the responded to last catalog are the variables that have a below threshold Pr(>F) value. So, we can include them into our model.

1 of 1 Fields | Records 1 to 10

Record Report

Report for Linear Model catalog_demand2300

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Responded_to_Last_Catalog + Avg_Num_Products_Purchased, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-662.60	-67.17	-2.96	69.88	973.90

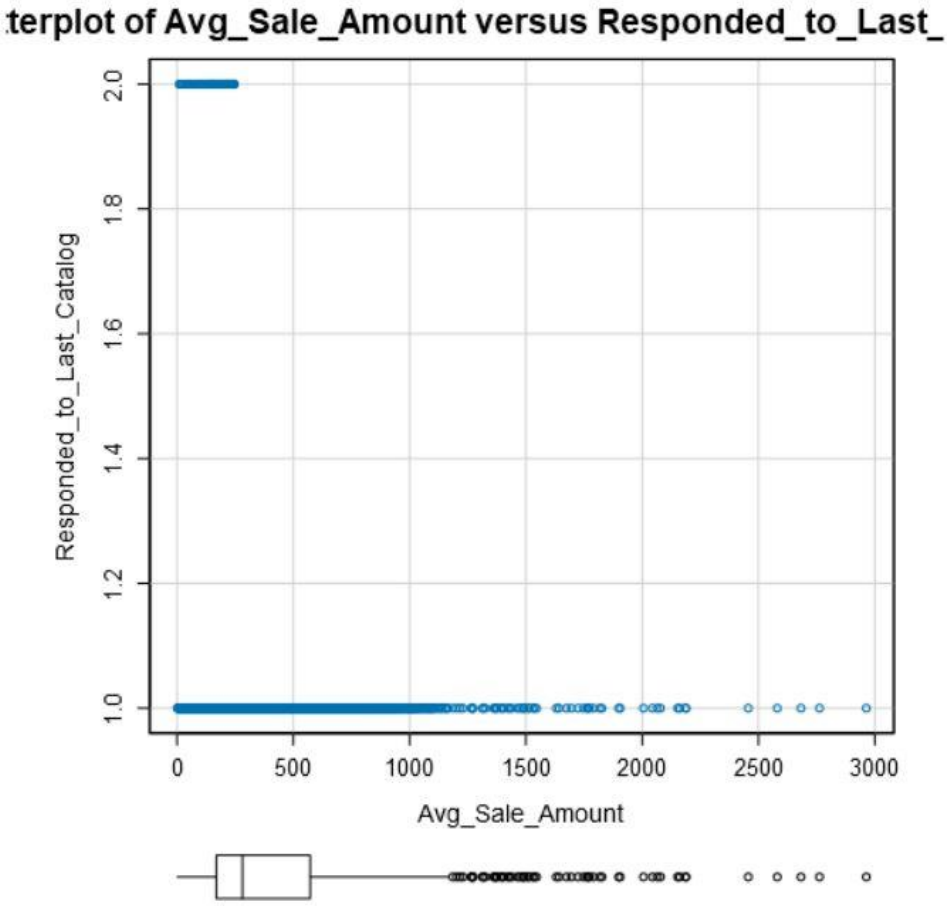
Coefficients:

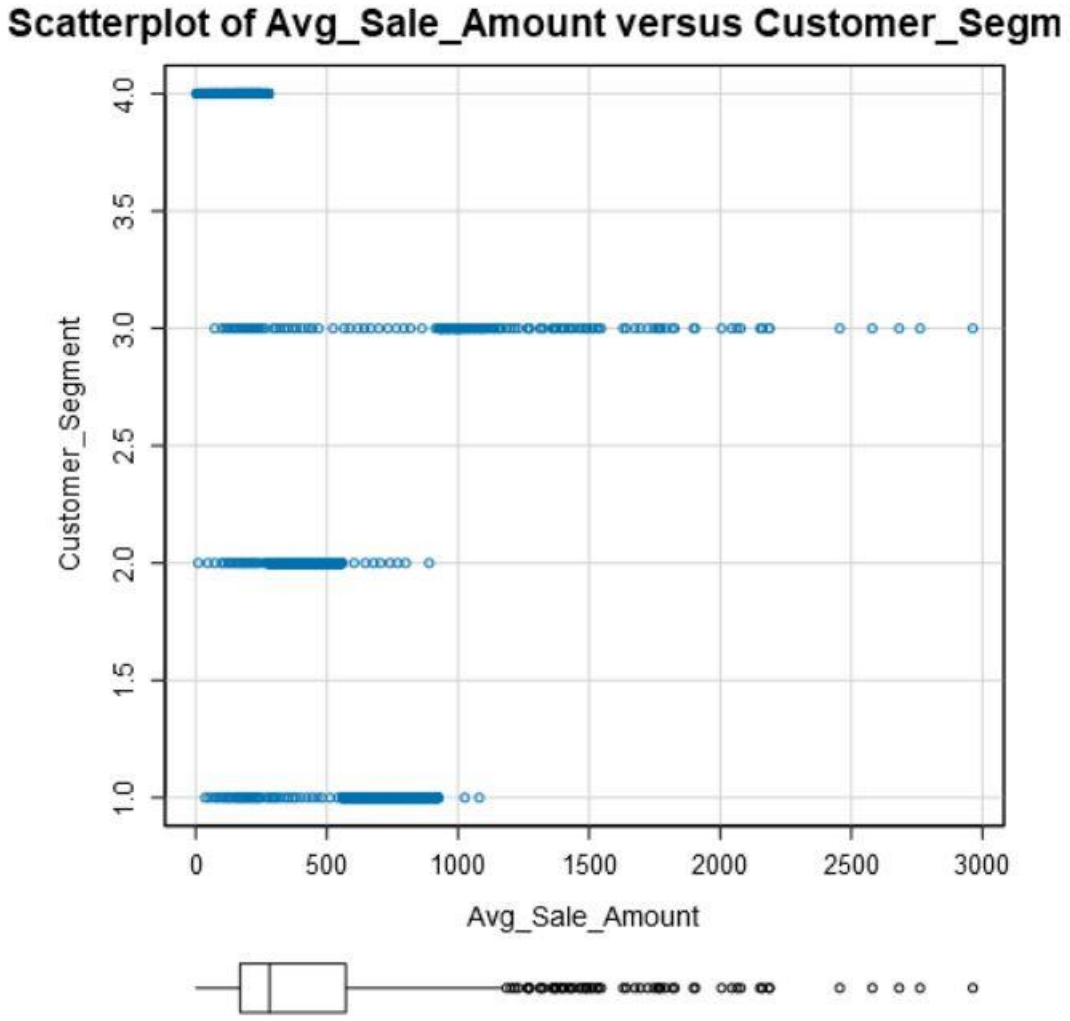
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	305.00	10.582	28.823	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-150.03	8.967	-16.732	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.69	11.897	23.678	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.76	9.815	-24.734	< 2.2e-16	***
Responded_to_Last_CatalogYes	-28.17	11.259	-2.502	0.01241	*
Avg_Num_Products_Purchased	66.81	1.515	44.099	< 2.2e-16	***

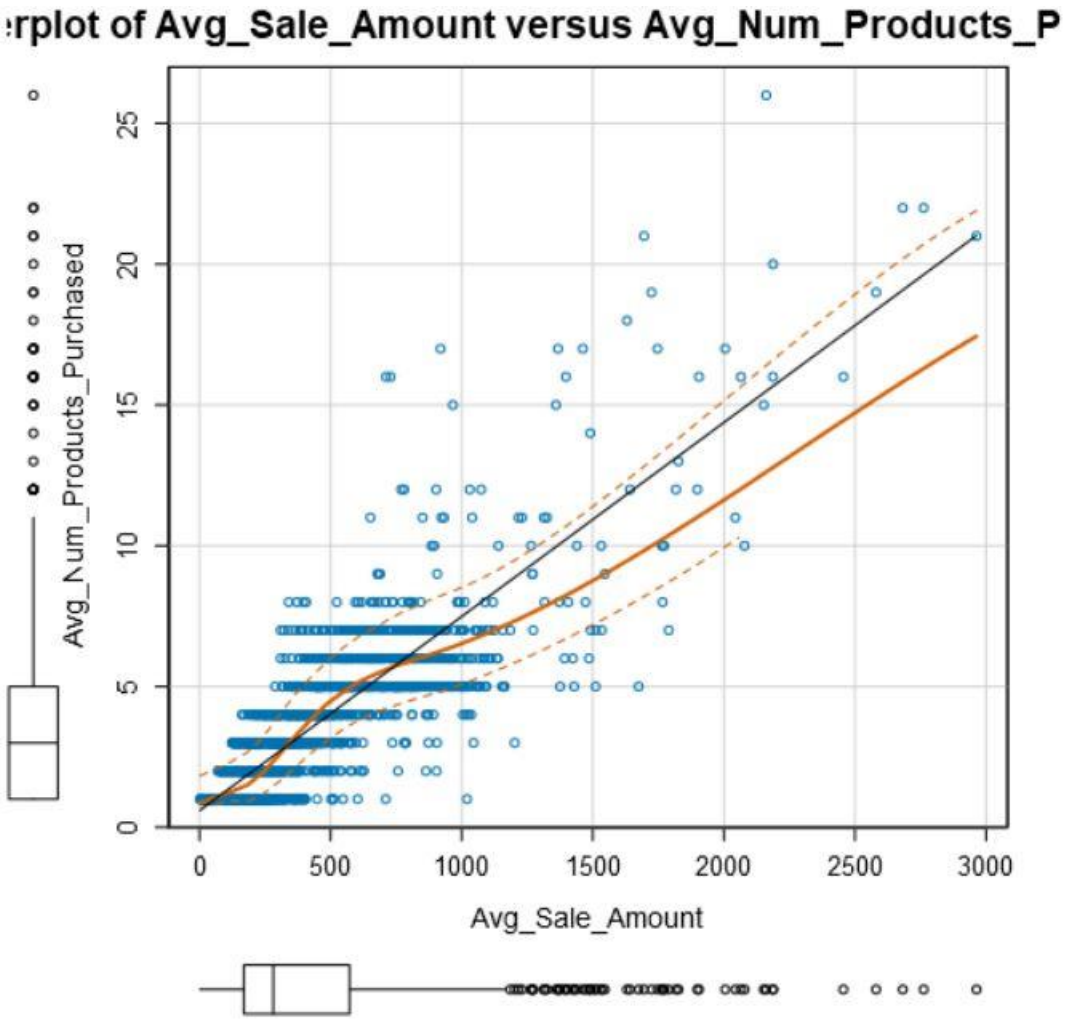
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.33 on 2369 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.837
F-statistic: 2438 on 5 and 2369 DF, p-value: < 2.2e-16

Upon further investigation using scatterplots to find a linear relationship with the target variable we can see that –




Scatterplot of Avg_Sale_Amount versus Customer_Segm
Customer_Segment
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0
500
1000
1500
2000
2500
3000
Avg_Sale_Amount



The above three scatterplots show the relationship of the predictor variables and the target variable.

The variable “responded to last catalog yes” is also omitted.

The reason is, after running the whole model in alteryx and upon using the data set p1_mailinglist with score function, the above variable is a missing predictor variable in the data set and therefore we cannot use it in our model to predict the catalog sales.

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)

Report for Linear Model catalog_demand2300

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

Coefficients:

	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	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 DF, p-value: < 2.2e-16

The target variable Y in our case is = average salary amount
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.

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.

As we can see above, all the predictor variables chosen have p-values smaller than 0.05 which means we have chosen the statistically significant variables with respect to the target variable. The adjusted R-squared of 0.8366 means that about 84% of the target variable can be explained by the model. Normally, if the Adjuster R-squared value is above 0.7, it is generally considered as a good model.

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

So, the average of my predicted sales amount which is the ‘Score’ = $138292.13
and The average gross margin (price - cost) on all products sold through the catalog is 50%.
So the average gross margin = score*score_yes*0.5 = 23555.860

For 250 customers, the price of catalogs sold at the rate of 6.5$ = $1625

Additionally, We want to calculate the expected revenue from these 250 people in order to get expected profit. This means we need to multiply the probability that a person will buy our catalog as well.

So, the average number of customers that have responded yes to the catalog = 34% = 0.34

Now we can calculate the expected profit = $(23555.860 – 1625)
Since Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000, from the above calculated data we can make a recommendation to send the catalogs to the 250 customers and have an expected profit of $21,930.86 which exceeds the 10,000$ mark.

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

The expected profit is calculated to be $21,930.86.