Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

The decision is: would the company send the catalog out to the new 250 customers? Or wouldn't? That decision depends on predicting if the expected profit contribution of these customers would exceed $10,000.

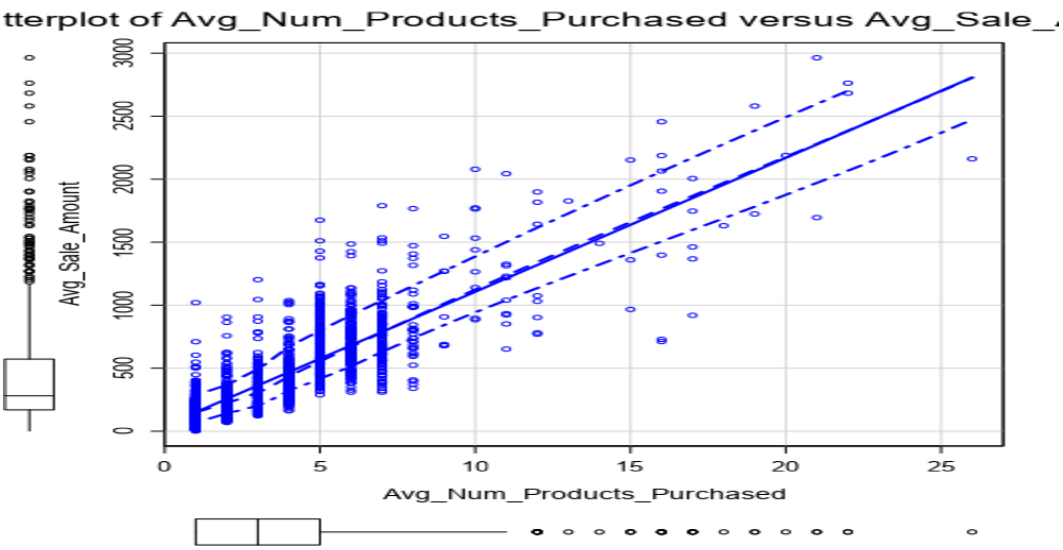
2. What data is needed to inform those decisions?

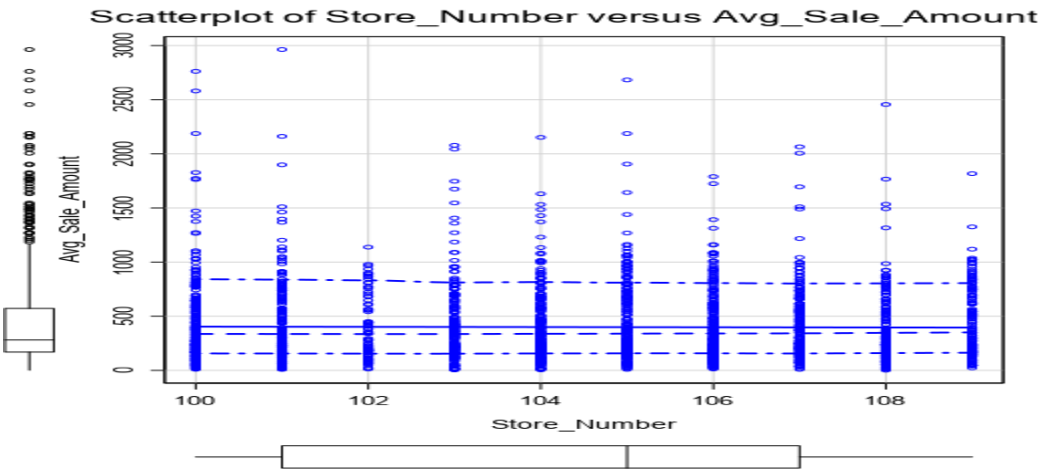
We need to predict the profit of the new 250 customers.
We will need to have previous data for old customers purchasing (given) and any other helpful data about their behaviors to help us on finding a good regression model for the purchasing amount of new customers.
W also need to know the probability that the customer would buy if he receives the catalog or not (given).
We are given the cost of catalog distribution, so, we can find the expected revenue of each customer.

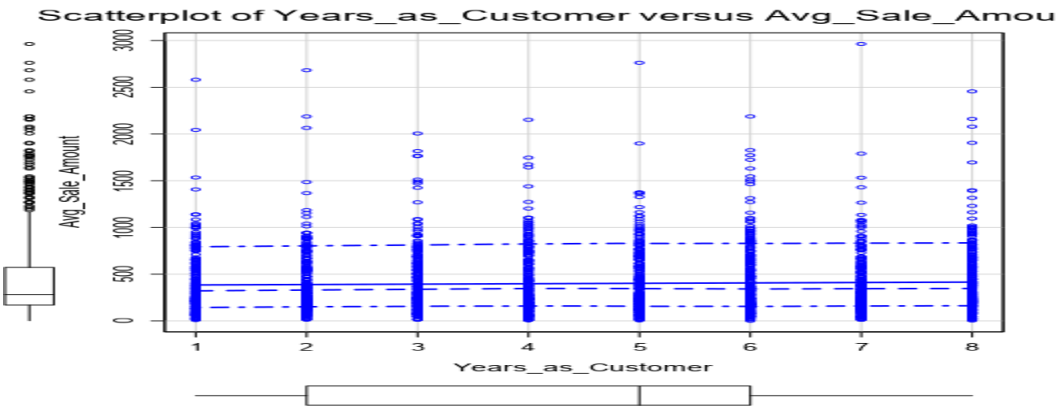
Step 2: Analysis, Modeling, and Validation

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.

For my regression model, looking into the data, I can directly exclude 'Name', 'Customer_ID' and 'Address' because it cannot be related to our prediction logically.
Also, 'Respond_to_Last_Catalog' because it is not in the new customers data.
So, we have some numeric and categorial fields to try.
For numeric fields such 'Years_as_Customer', 'Avg_Number_Products_Purchased' and 'Store_Number', we can use Scatter Plot to check if there is any linear relationship with 'Avg_Sales_Amount', down are the figures:







As noticed, only AVG_Number_Product_Purchased has a linear relationship with Avg_Sales_Amount.

For categorical fields, I will use the linear regression report to check which of (Customer_Segment, City, State and the chosen numeric 'Avg_Number_Products_Purchased') will fit in our model:
Note: I could not put city or state as predictor variables in the regression Model, since there is not enough variation of these variables, error shows up.

Report

1

Report for Linear Model SalesAmountReg

2 *Basic Summary*

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

4 Residuals:

5

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

6 Coefficients:

7

	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

8 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 degrees of freedom (DF), p-value < 2.2e-16

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 linear model is a good model because as in the report, we can see a high Adjusted R-Square value, about 0.84 (the model is good) and very low P values (<< 0.05, prediction variables are perfect).

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 + 66.98 * Avg_Num_Products_Purchased *- 149.36 (if Segment = club only) + 281.84 (if Segment = club and credit card) – 245.42 (if Segment = store mailing list) + 0 (if Segment = credit card only)*

Step 3: Presentation/Visualization

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

I would recommend the company to send the catalog to the new 250 customers.

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

I applied my regression model to the new 250 customers to predict each customer sales amounts. Then, I calculated the expected revenue from each customer by multiplying the predicted sales amount by the probability that the customer will buy (given).
Then, I calculated the profit of each customer knowing that the gross margin on all products sold through the catalog is 50% (multiply expected revenue by 0.5) and subtract the catalog cost (6.5 $). Summing that up will give 21,987.44 $ > 10,000 $ (the amount the company put as a condition to or not to send the catalog).

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

21,987.44 $