Brian Berns Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

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

Key Decisions:

Answer these questions

1. What decisions needs to be made?

   We need to decide if it is worth it to send out a catalog to 250 new customers based on the expected profit.

2. What data is needed to inform those decisions?

   Some of the data that we need (that is also given to us) includes the following:

   Historical customer data (response data, how many average purchases, location, # of years a customer).
   They told us about the 50% profit margin and the cost of a catalog $6.50.
   We need to also use the probability that any customer the company sends the catalog to will use it a buy things out of it. (Score_Yes).
   All of this can help predict the expected profit.
   Using the historical customer data/past sales we can predict current year sales. Once we have these sales numbers we can multiply by the "Score_Yes" (probability a person chosen will make a purchase out of the catalog) and then we get the current totals. We then can use these totals to make a decision of whether to send these catalogs out.

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 (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.

Using all variables at first I ran a couple scatter plots for my data. I noticed that I was only getting a linear relationship by using Avg_Num_Products_Purchaed versus Avg_Sales_Amount. I tried making other relationships such as Avg_Sales_Amount versus X_years_as_customer, and versus Store_Number but nothing was linear.

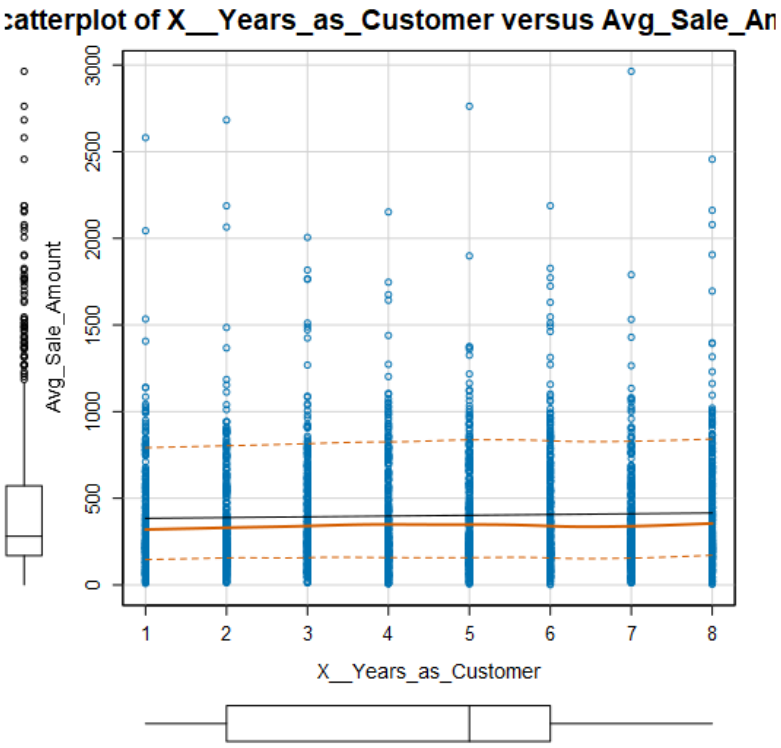
Some items weren't needed, such as:

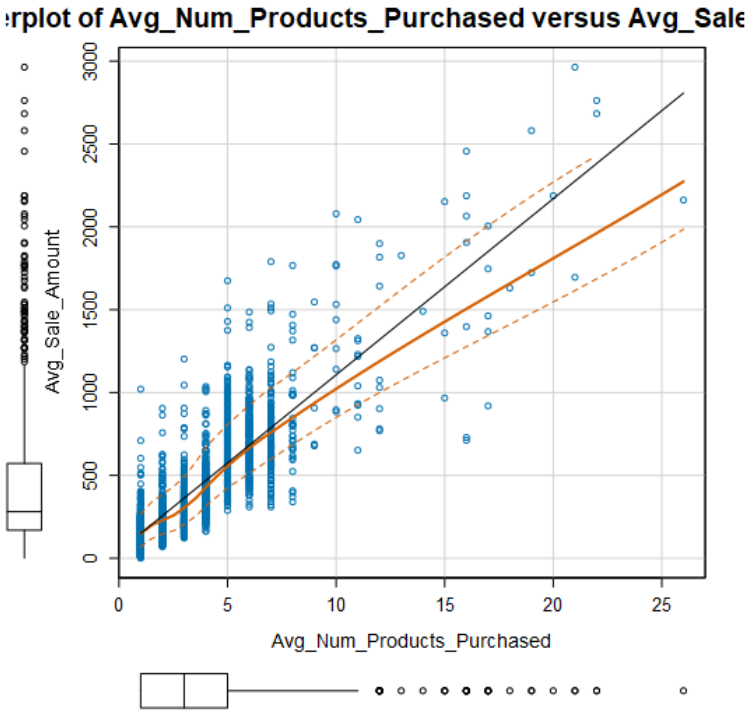
Addresses: because it just seems a little bit much including names and numbers. I do guess that if you want to try it by a type of location you can try by zip code. On this set of data though the correlation is not linear.

Names: because it doesn't matter what a person's name is to buy an item.

Customer_ID : same type of thought process as the name (above)

this is why I used Avg_Sales_Amount as my target variable. Customer Segment and Avg_Num_Products_Purchased were my predictor variables.


catterplot of X__Years_as_Customer versus Avg_Sale_An
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
1
2
3
4
5
6
7
8
X__Years_as_Customer


rplot of Avg_Num_Products_Purchased versus Avg_Sale
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
25
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.

Report for Linear Model Linear_Regression_4

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = 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 degrees of freedom (DF), p-value < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28715078.96	3	506.4	< 2.2e-16	***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16	***

- **The first important aspect I looked for were p-values lower than .05, which were Customer Segment and Avg_Num_Products_Purchased. This means that they are statistically significant which means that we can throw out a hypothesis that this is due to a anomaly or chance. We also notice that the adjusted R squared value of 0.8366, which is high strength, so we can tell that this is a strong model.**

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)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

Y= 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) + 0*(Credit Card)

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?

   The company should send out the 250 catalogs because the profit exceeds $10,000.

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

   First I calculated avg_sales (score field) then I multiplied it by the score_yes field, which means the probability of buying a product. Then you have to multiply it by .5 (50% gross margin). Then the cost of the catalog ($6.50*250) is also subtracted, for 250 people. This will get us our expected profit.

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

Expected Profit = (Sum of expected * gross margin) – (Cost of 1 Catalog * 250 ppl)

Predicted sum of expected = $47,225.87

PLUG IN……..($47,225.87 * .5) – (6.50 * 250)

= $21,987.44

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.