PROJECT 1: PREDICTING CATALOG DEMAND

Chandan Mishra (chandan.dce07@gmail.com)

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?

The company which manufactures and sells high-end home goods wants to determine if sending out the printed catalog to its 250 customers from the mailing list would result in profit more than $10,000 i.e. the expected profit from these 250 new customers should be more than $10,000 so that management can say yes to sending out the printed catalog.

2. What data is needed to inform those decisions?

The available dataset for customers contains information on about 2300 customers. This customer dataset contains following information about customers:

Name	V_String
Customer_Segment	V_String
Customer_ID	V_String
Address	V_String
City	V_String
State	V_String
ZIP	V_String
Avg_Sale_Amount	Double
Store_Number	V_String
Responded_to_Last_Catalog	V_String
Avg_Num_Products_Purchased	Double
#_Years_as_Customer	V_String

Since Responded_to_Last_Catalog variable is not present in the Mailing List data, this variable is not used in building the regression model. We have to calculate the Profit in sending out the catalogs to 250 people.

Profit = (Predicted_Avg_Sale_Amount*Score_Yes*Avg_Gross_Margin)- (Cost of Printing and distributing a catalog*# of New Customers in the mailing list)

To calculate profit, following info is already provided:
Avg_Gross_Margin = 50% or 0.5

Cost of Printing and distributing a catalog = 6.5
# of New Customers in the mailing list = 250

Since Predicted_Avg_Sale_Amount would be decided based on the Avg_Sale_Amount from the customers data so we would need to calculate this target variable. For predictor variables, we need to check which variables are statistically significant or have a linear relationship with the target variable to include them in the models. Following variables are not considered to build the model since these variables won't make any sense to determine the Avg_Sale_Amount:
Name, Customer_ID, Address, State (since this has only one value 'CO'). For other categorical variables, we will do a trial and error method to see which variables are statistically significant and then include only the significant ones in building the model.

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.

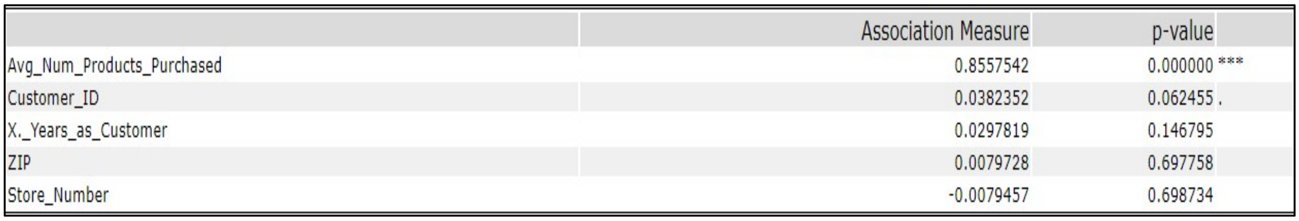
At the minimum, answer these questions:

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.

   Initially, the predictor variables were divided into two groups:
   a) Numeric Predictor Variables
      - Avg_Num_Product_Purchased
   b) Categorical Predictor Variables
      - Customer_Segment
      - City
      - Zip
      - Store_Number
      - #_Years_as_Customers

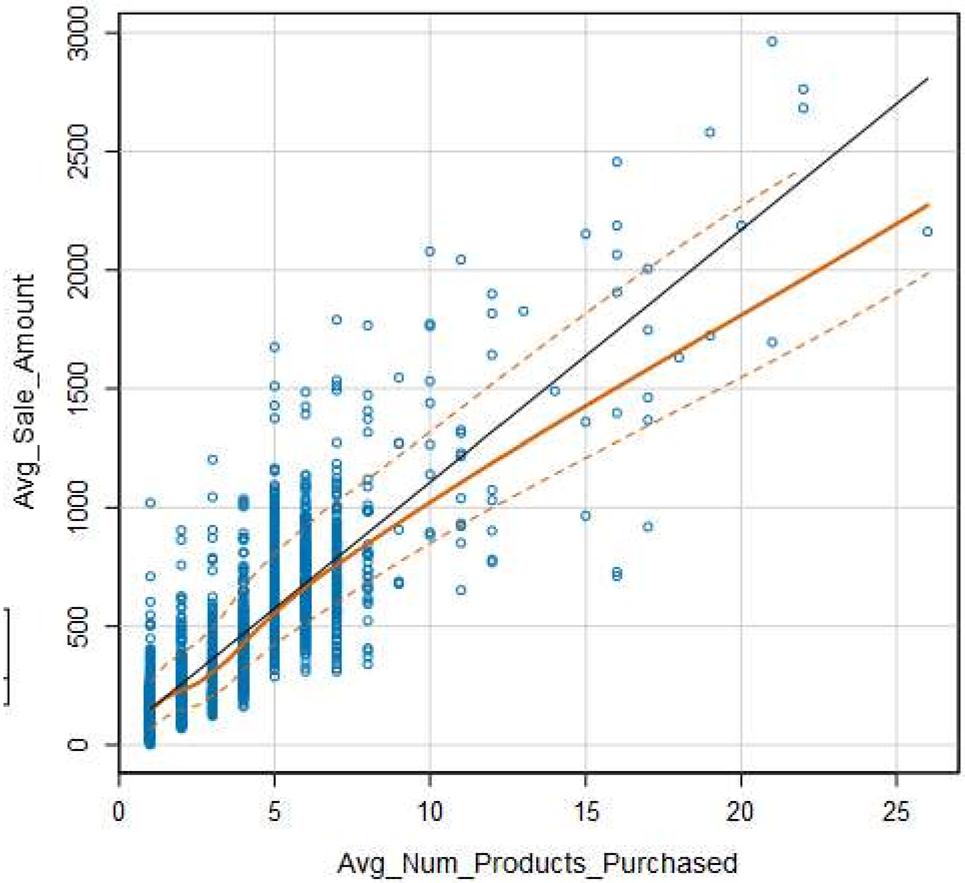
Variables such as Zip, Store_Number or # of Years as Customers were converted into String from Double since we don't need to add the values of these variables.

When association analysis was run on the original dataset keeping Avg_Sale_Amount as Target variable, following results were noticed

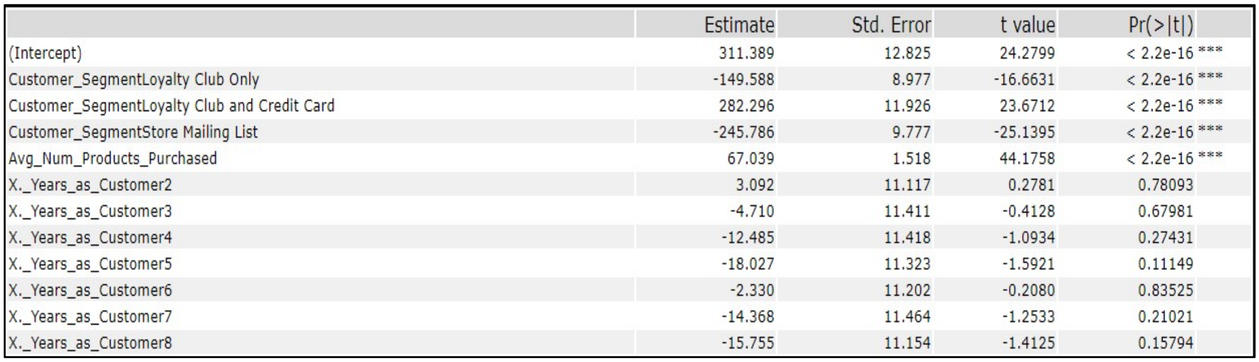


	Association Measure	p-value
Avg_Num_Products_Purchased	0.8557542	0.000000 ***
Customer_ID	0.0382352	0.062455 .
X._Years_as_Customer	0.0297819	0.146795
ZIP	0.0079728	0.697758
Store_Number	-0.0079457	0.698734

As this table shows, only Avg_Num_Products_Purchased is statistically significant. Checking the scatter plot between Target Variable and Predictor variable shows there is a linear relationship and hence this variable is selected as a predictor variable.



For categorical variable customer_segment and #_years_as_customers, we run the trial and error method and include them in the linear regression model to check whether they are significant or not.



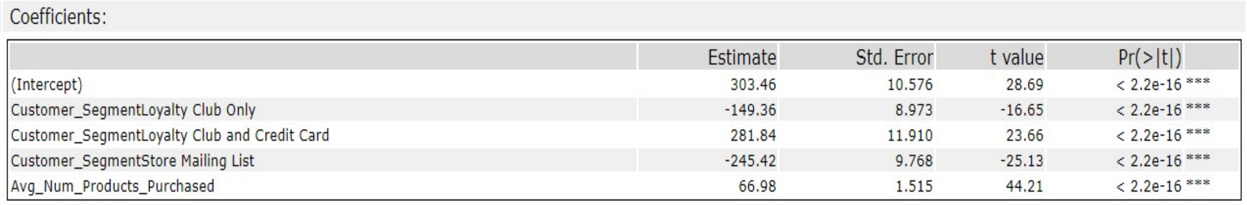
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	311.389	12.825	24.2799	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.588	8.977	-16.6631	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.296	11.926	23.6712	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.786	9.777	-25.1395	< 2.2e-16 ***
Avg_Num_Products_Purchased	67.039	1.518	44.1758	< 2.2e-16 ***
X._Years_as_Customer2	3.092	11.117	0.2781	0.78093
X._Years_as_Customer3	-4.710	11.411	-0.4128	0.67981
X._Years_as_Customer4	-12.485	11.418	-1.0934	0.27431
X._Years_as_Customer5	-18.027	11.323	-1.5921	0.11149
X._Years_as_Customer6	-2.330	11.202	-0.2080	0.83525
X._Years_as_Customer7	-14.368	11.464	-1.2533	0.21021
X._Years_as_Customer8	-15.755	11.154	-1.4125	0.15794

This table shows only Customer_Segment is significant. Hence, the final two statistically significant variables are selected in building the model:

- Avg_Num_Products_Purchased
- Customer_Segment

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.

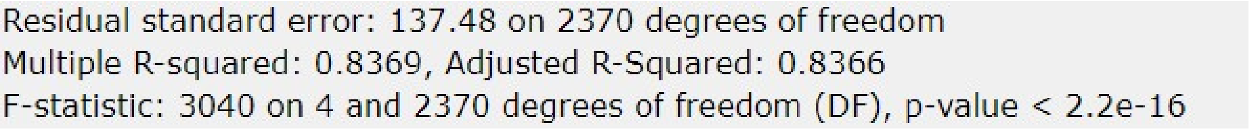
After selection of final two predictor variables, the model was run and produced this result:



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 ***

The customer segment which is a categorical variable and numeric predictor variable Avg_Num_Products_Purchased shows the p-value of 2.2e-16 which is very less than 0.05 and hence is statistically significant.



```
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
```

The adjusted R-Squared shows value of 0.836 which is greater than 0.7 for the model to be considered good. This linear model is a good model because of the selection of only statistically significant variables having low p-values in building the model and having

Adjusted R-Squared value greater than 0.7. Low p-values and high Adjusted R -squared value means model is highly predictive.

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 regression equation should be in the form:

Avg_Sale_Amount = 303.46 + –149.36* (If Type: Loyalty Club Only) +281.84*(If Type: Loyalty Club and Credit Card) + -245.42* (If Type: Store Mailing List) + 0*(If Type: Credit Card Only) + 66.98 * Avg_Num_Products_Purchased

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

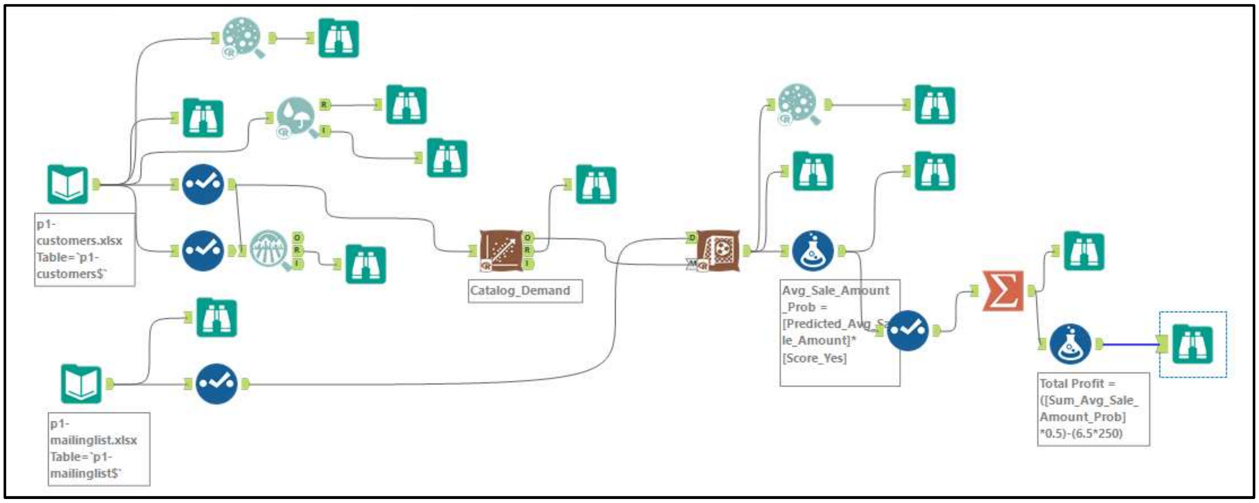
At the minimum, answer these questions:

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

The overall predicted profit is $21,987.45. Since the criteria was to have profit more than $10,000 to send out the catalogs to these 250 new customers, and the predicted profit exceeds that, hence the catalogs should be sent.

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

The model looks like this:


p1-customers.xlsx
Table=`p1-customers$`
Catalog_Demand
Avg_Sale_Amount_Prob = [Predicted_Avg_Sale_Amount]* [Score_Yes]
Total Profit = ([Sum_Avg_Sale_Amount_Prob] *0.5)-(6.5*250)
p1-mailinglist.xlsx
Table=`p1-mailinglist$`

First the model was built using available data for 2300 customers and only two predictor variables Avg_Num_Products_Purchased (numeric) and Customer_Segment (categorical) were selected to build the model because of being statistically significant. After checking that low p-values (<0.05) and high Adjusted R-squared (0.83), the model is used to predict the Predicted_Avg_Sale_Amount for 250 new customers whom the company wants to send out the catalogs to. The score tool in Alteryx is used to calculate the predicted amount for each customer.

Avg_Sale_Amount_Prob = Predicted_Avg_Sale_Amount * Score_Yes

Score_Yes probability that the customer will buy. Then, a summarize tool is selected to sum the Avg_Sale_Amount_Prob. To calculate profit, the sum must be multiplied by 0.5 (because of 50% gross margin) and out of that total $6.5 for 250 customers need to be subtracted because of cost of printing and distribution. The final Profit equation comes out to be:

Total_Profit = (Sum_Avg_Sale_Amount_Prob*0.5) – (6.5*250)

Then a browse report tool is added to check the total profit which comes out to be $21,987.45(>$10,000). Hence, a recommendation to send out the catalogs to these 250 new customers is provided.

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

$21,987.45