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?
   Ans: In order to first understand the data, we first need to find out what to predict. As from the project details, we can easily find out how much profit can the company earn by sending catalog to the new customers. The decision we need to make is if sending printed catalog to the new customers is profitable for the company. As we need to find expected profit from these 250 customers, we also need to multiply total predicted revenue with the probability of the customers will actually buy our catalog. There are some other question need to be answered in order to reach our decision: How will we predict the potential profit of sending out the catalog? Considering the data we have, how it will be useful for us to coming up with this prediction?

2. What data is needed to inform those decisions?
Ans: We will need Avg_Sale_Amount, Customer_Segment, Avg_Number_Products_Purchase and Score_Yes, Predicted_revenue(X) in order to inform these decisions.

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.

Ans: A sloped line in a scatter plot indicates that as the X increases or decreases the Y increase or decreases indicating the two variables are related. So if there is a slope to the line then this might indicate that this is a good predictor variable for this target variable.
Figure: 1 indicates Avg_Num_Products_Purchase is a good predictor variable for the target variable Avg_Sale_Amount.

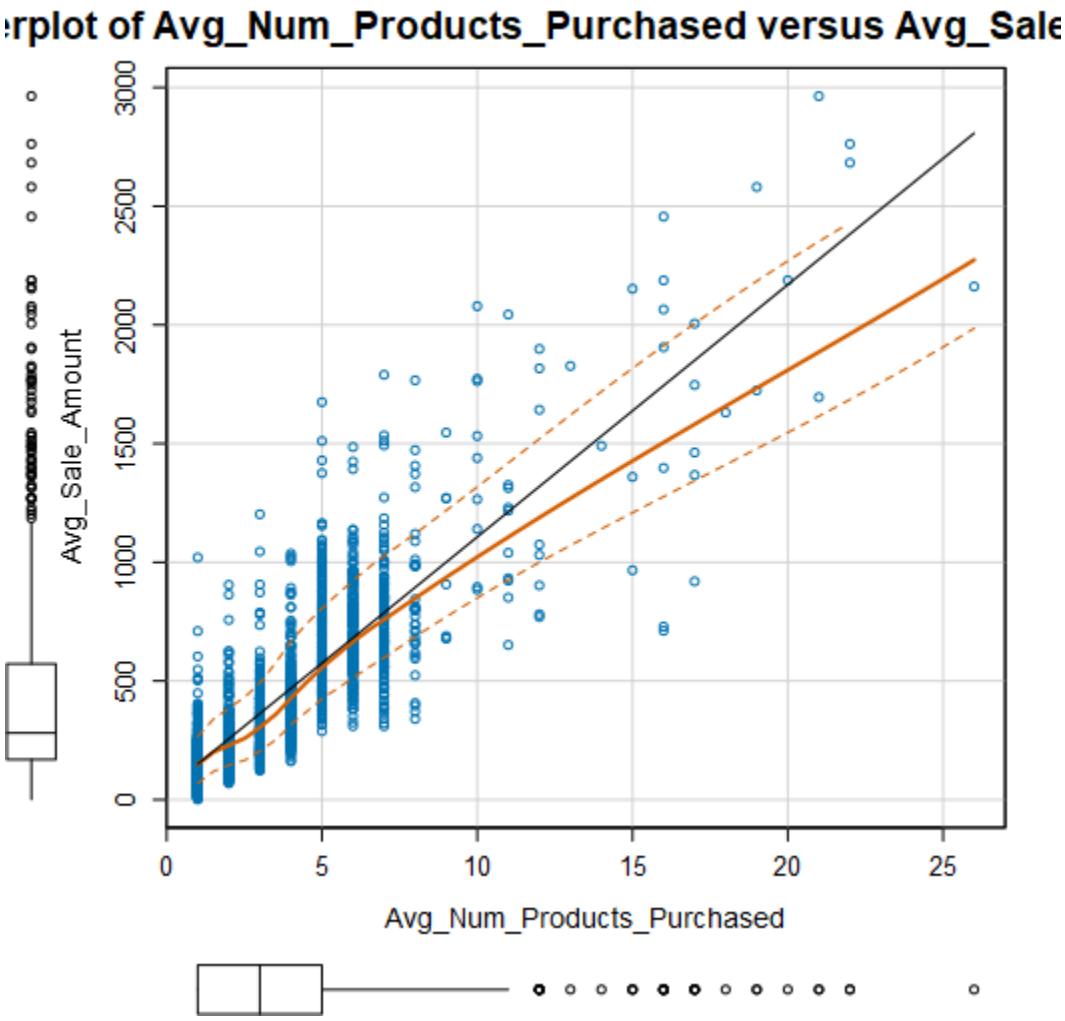

Figure: 1- Avg_Num_Products_Purchased vs Avg._Sale_Amount

Figure: 2- Customer_ID vs Avg_Sale_Amount

Figure-2 indicates there is a linear relationship between potential predictor variable (Customer_ID) and target variable Avg_Sale_Amount. In case of categorical variable, we can use trial and error to see statistical significance. By interpreting the p-value of linear regression report, we can easily interpret this.
I have included a screenshot from the R output of the Linear Regression tool in Alteryx which is the Linear Regression Report below to illustrate where to check for these values.

From figure:3 we can see all the p-values including categorical variables are well below 0.05. It means these all are very much statistically significant for our model.

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

Figure 3: Linear Regression Report

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.

Ans: Low P-values and a high R-squared suggest the model is highly predictive, Low P-values means it is highly unlikely that the two variables are not related. Low R squared means the model is not very fit. From figure-3 we can see, the p-value is very much lower to 0.05 which suggest a very much significant value to go on with the prediction. Our Adjusted R-Squared value as per figure-3 is 0.8366. An R-squared value close to 1 would mean that nearly all variance in the target variable is explained by the model. In this case it indicates high rate of successful predictive nature of the data. The above reasons are the base behind my believe, the linear model I designed is a good 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……*

Target variable = 303.46 -149.36 * (If Type: Customer_Segment_LoyaltyClubOnly) + 281.84 * (If Type: Customer_Segment_LoyaltyClub and Credit Card) – 245.42 * (If Type: Customer_Segment_Store Mailing List) + 0 * (If Type: Customer_Segment_Cash) + 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?

Ans: The company should send the catalog to these 250 customers as it indicates a hefty profit.

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

Ans: Firstly, I have developed a linear regression model by using *p1-customers.xlsx* where more than 2300 customer data where available. From this regression model I have found the p-value which is much lower than 0.05 and adjusted R-Squared value of 0.8366 which indicates the dataset is significant enough and fit for predictive modeling. Then I have used *p1-mailinglist.xlsx* which contains 250 customer data for whom we need to make the prediction. I have connected this table with regression model with a score table in order to get the predicted value. From the score table, I have used formula to calculate individual profit. Score_Yes is the purchase probability for individual customers, therefore, we need to do the calculation individually, i.e. for each customer, we use the predicted sales Score(X) to multiply Score_Yes to arrive at the expected sales, then multiplying 0.5 we can get the margin and deduct the costs to get individual profits and finally sum up all the profits to get the total predicted profit amount.

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

Ans: The expected profit from the new Catalog is 21987.44.

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.