Project 1: Predicting Catalog Demand

By Nirupama Puthur Venkataraman
(for Project Reviewer application)

Step 1: Business and Data Understanding

For this problem, we need to analyze historical customer data of a high-end home goods manufacturer. The company had launched a sales catalog in the previous year and is planning to repeat the process in the current year as well. They also want to send their catalog to newly acquired customers.

However, the management wishes to base this decision on financial feasibility. Accordingly, the business goal is to analyze and make predictions about the anticipated profits from these 250 new mailing list customers.

We do not have data for the new customer segment, but we do possess data about existing customers and their purchase patterns based on the catalog sent out the previous year. So, we can make predictions upon analyzing this data. As we are forecasting outcome, we classify this a 'Predict Outcome'. Since, we do have data available, it is also a 'Data Rich' business problem. The data is also numeric and continuous. By following the flow of our Methodology Map, we choose Linear Regression as the correct methodology.

Predict Outcome > Data Rich > Numeric > Continuous > Linear Regression.

Key Decisions:

1. What decisions needs to be made?
   Predict how much money the home goods company can expect to earn from sending out a catalog to new customers.
   One of the decisive metrics is to ascertain if the total expected profit from these new customers exceeds $10000. This amount will determine if the catalog should be mailed or not to the new customers.

2. What data is needed to inform those decisions?
   We need data about the new customers purchase decisions. Since, this is missing, we will use the historical data from existing customers to forecast how profitable the new mailing list segment will be and base the decision of sending catalogs to them accordingly.

Step 2: Analysis, Modeling, and Validation

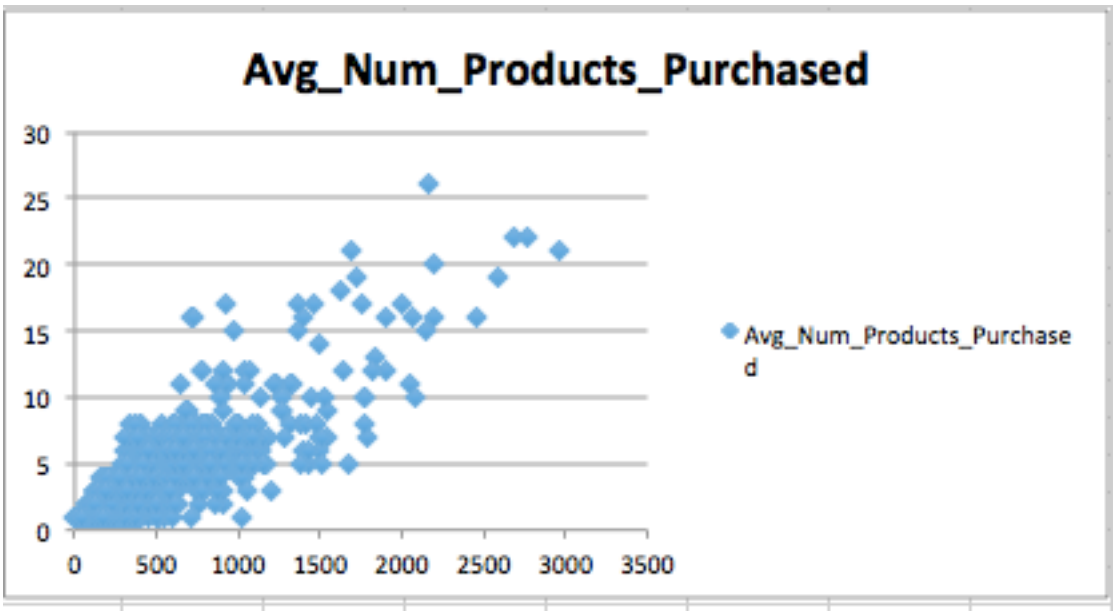
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

Notes:
Used the p1-customers.xlsx to train my linear model.
Selected the Linear Regression model based on Methodology Map

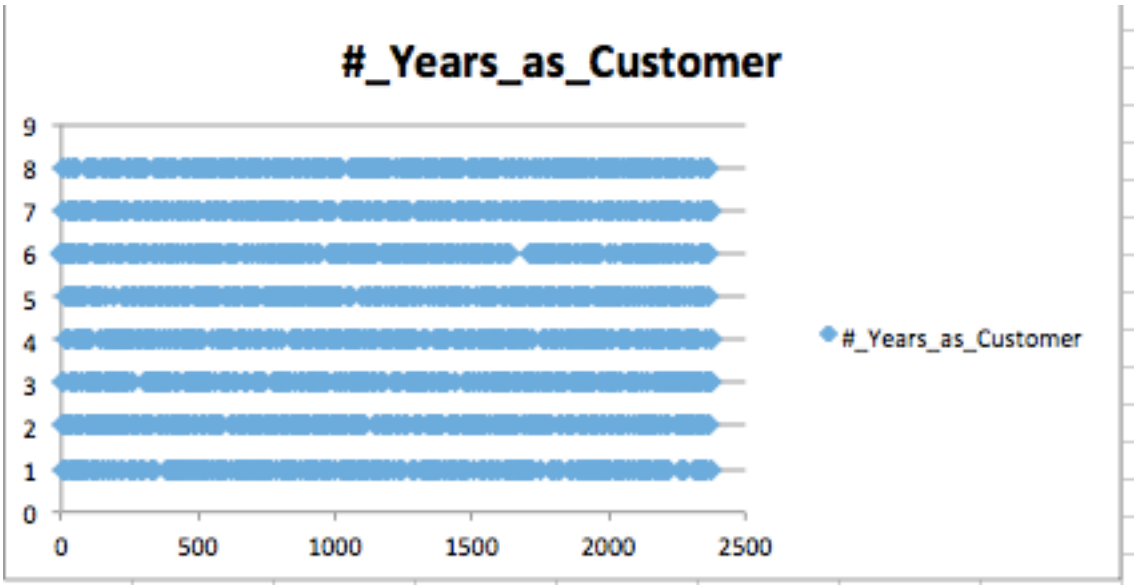
At the minimum, answer these questions:

1. How and why did you select the predictor variables (see supplementary text) in your model?

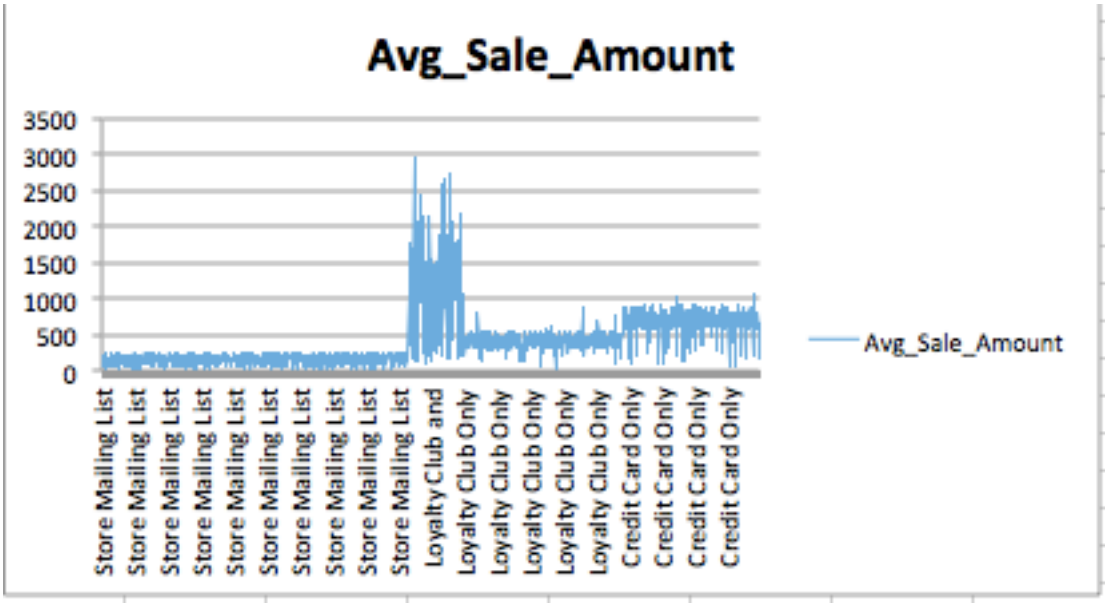
I first selected the variable “Avg_Num_Products_Purchased” and created a scatter plot with the target variable “Avg_Sale_Amount”. The plot showed a linear trend and hence, high potential to be a predictor variable.



Next, I first selected the variable “#_Years_as_Customer” and created a scatter plot with the target variable “Avg_Sale_Amount”. The plot did not show any sort of linear relationship and hence, was discarded as not being a good fit for analysis as a predictor variable in linear regression model.



Further I considered the “Customer_Segment” variable. Being a categorical variable, the scatterplot did not show any significant linear trend overall. But it does show distinct trends for each type of customer segment. This looked promising and hence, was chosen as the categorical variable for linear regression.
As per the project instructions, I set the Credit Card only category as the base and created dummy variables for the others. This can be viewed in the sheet titled “p1-customers-analysis” on tab named “p1-customers_dummy”.



3. 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.
   The adjusted R-square and p values from the linear regression analysis are as follows:

Multiple Linear Regression:
Adjusted R-square: 0.837

Source	Value	Standard error	t	Pr > |t|	Lower bound (95%)	Upper bound (95%)
Avg_Num_Products_Purchased	0.539	0.012	44.208	**< 0.0001**	0.515	0.563
Loyalty Club and Credit Card	0.227	0.010	23.664	**< 0.0001**	0.208	0.246
Loyalty Club Only	-0.189	0.011	-16.645	**< 0.0001**	-0.211	-0.166
Store Mailing List	-0.360	0.014	-25.125	**< 0.0001**	-0.388	-0.332

We consider that any value of adjusted R-square greater than 0.7 and P-values less than 0.05 is a great fit. Here, we see that the Adjusted R-square is 0.837 and P-values are less than 0.0001. Hence, the linear model for these predictor variables is statistically appropriate.

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 equations for multiple linear regression is:

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

The equations for linear regression for Customer Segment is:

Avg_Sale_Amount = 682.68
+391.48*(If Type: Loyalty Club and Credit Card)
-286.35*(If Type: Loyalty Club Only)
-525.32*(If Type: Store Mailing List)
+0 *(If Type: Credit Card Only)

The equations for linear regression for Average Number of Products Purchased is:

Avg_Sale_Amount = 44.015+
106.28*Avg_Num_Products_Purchased

Note: Here, as we are using Customer Segment as one of the predictor variables, we set the base case to Credit Card Only.

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?
   Yes, the company should send the catalog to the mailing list customers, as the predicted profit is over $20000, well above the threshold of $10000 as determined by the management.
   Even assuming some errors and deviations in actual and expected profits, the margin is sufficiently large to still meet the $10000 mark. Hence, it is safe to say that the catalogs can be sent 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)
- Create scatter plots to ascertain if the variable is suited to be a predictor variabl using the linear regression analysis.
- Set average number of products as numerical predictor variable.
- Set customer segment as categorical predictor variable.
- Conduct linear regression for each of the two predictor variables.
- Based on the admissible p-values and adjusted r-square values in these, perform multiple linear regression.
- Observe the p-values and adjusted r-square values. Being within the permissible limits, select the regression equation.
- Apply the linear model to the mailing-order customer list and compute the average sales values for each customer, given the data for average number of products.
- Calculate the revenue based on predicted (supplied in dataset) values of yes_score.
- Calculate the predicted profit by multiply the revenue by the gross margin (50%) and then subtracting the catalog costs ($6.50).
- Calulate the sum total of profit from all the 250 customers.
- As this value is over $20000 and provides ample room for any deviations in expected and actual profits, make a recommendation to go ahead with sending the catalog to these new 250 customers.

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
Based on the Yes_scores and $6.50 catalog costs and on assuming 50% average gross margin, the total profit works out to be **$20754** [rounding up **$20753.45**].