Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

The main outcome is to predict the expected profit from sending a catalogue to 250 new customers. This prediction will be used to justify if there is any value in sending this new catalogue out.

1. What decisions needs to be made?

    The decision whether or not to launch the catalog to the new customer set. The amount of profit must exceed $10,000 to be worthwhile doing.

    We have the data from the current set of mailing list customers, so we can use this data to predict the average sale amount a new catalogue would generate to them. After predicting average sales to our new customer set, we need to calculate the profit to be made. This profit needs to be in excess of $10,000.

2. What data is needed to inform those decisions?

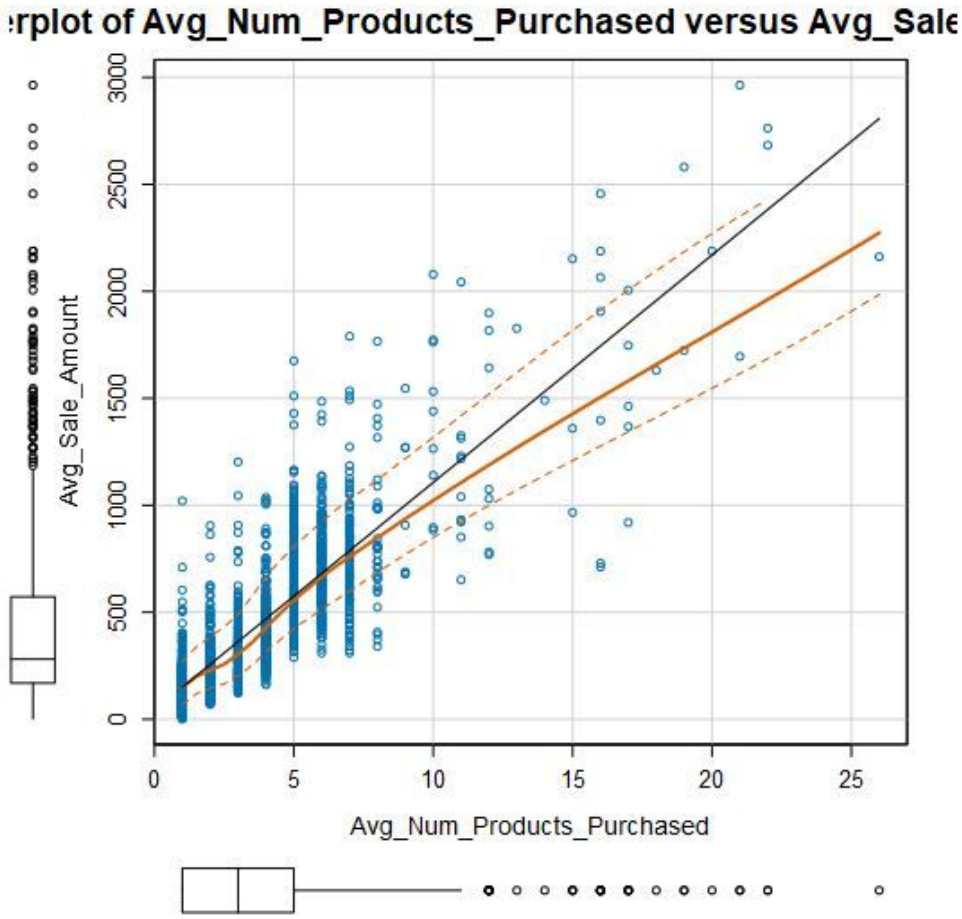
    Av_Sale_Amount
    Av_Num_Of_Products_Purchased
    Customer_Segment

Step 2: Analysis, Modeling, and Validation

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.

    The first thing I did was select the average sales and the average number of products purchased and plotted them on a scatter graph to look for correlation.

    I selected average sales as that will predict our 'expected' revenue. Average number of products was selected as it is the only continuous numeric predictor value.



The average sale amount increases with the average number of products purchased. The trend line is linear and positive but there is a lot of variation. Additional predictor variables will help explain some of this variation.

Customer segment is a categorical variable and can be tested to see if there is any association. I added the customer segment and predictor variables to the linear regression model and looked at the p-value on the results table (below).

Report for Linear Model Linear_Regression_5

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	***
Residuals	44796869.07	2370			

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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.

For this model all the predictor variables are below 0.05, so are statistically significant. All of the variables are of a very low p-value (<2.2e-16) so we can be confident that there exists a relationship between the predictor and target variable.

The adjusted R-squared is above 0.8 which is considered a strong model and can explain a lot of variation.

From the above results we should use Average_Num_Products_Purchased and each of the Customer_Segments as the predictor variables.

3. What is the best linear regression equation based on the available data?

Y = 303.46 (intercept) + (-149.36*Customer_SegmentLoyalty Club Only) +
(281.84*Customer_SegmentLoyalty Club and Credit card) +
(-245.42*Customer_SegmentStore Mailing List) +
66.98*Avg_Number_Products_Purchased) + 0(Customer_Segment Credit Card Only)

Step 3: Presentation/Visualization

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

My recommendation is the company should send the catalogue.

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

Using the predicted sales value for the new customers we need to calculate the probability that they will buy.
Predicted av sales (Y) * Score_Yes = Expected Revenue.
Expected profit = (Expected revenue * 0.5 [*50% av gross margin*]) - 6.50 [costs]

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

Expected profit = sum of profit / customer
= $21,987.44 which is above the $10,000 required.