Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

1. What decisions needs to be made?
   1.1. We have a business problem which involves prediction. Hence we have to undertake predictive analytics.
   1.2. We have a lot of data in this problem. Hence, it will require data-rich analytical techniques.
   1.3. The outcome we seek in this problem is of numeric nature. Hence numeric data analysis may be applied.
   1.4. The numeric data outcome is continuous is nature. Hence linear/stepwise regression model will be used to build the model.
   1.5. Once the appropriate equation is ready, it is used to predict sales for the individual people in the mailing list.
   1.6. These sales figures are used to generate expected profit and then whether the company should send the catalog or not is to be determined.

2. What data is needed to inform those decisions?
   2.1. Data on the Customer ID and #of years as customer of the list of customers receiving the mail catalog the previous year.
   2.2. Response data from the exercise the previous year.
   2.3. Data on the Customer ID and #of years as customer of the list of customers receiving the mail catalog this year.
   2.4. Purchase and sales data from the previous year customers.
   2.5. Expected profit from each catalog {determined as [Revenue i.e. (Price-margin)}-Costs)]

Step 2: Analysis, Modeling, and Validation

Steps used:

1) The file p1-customers.xlsx was chosen as the base data file to set up the linear regression model.
2) Only Customer segment and Avg items purchased were chosen as predictor variables. This was because of the following reasons:
   a) Responded to last catalog not relevant as new customers sought
   b) Rest are address determiners and have no impact on new purchases
3) Target variable was Average sales.
4) The Linear regression tool output was used to score the p1-mailinglist.xlsx. Score _yes representing probability of purchases was the X-axis and score values at Y-axis

Results of Regression analysis and scatterplot

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

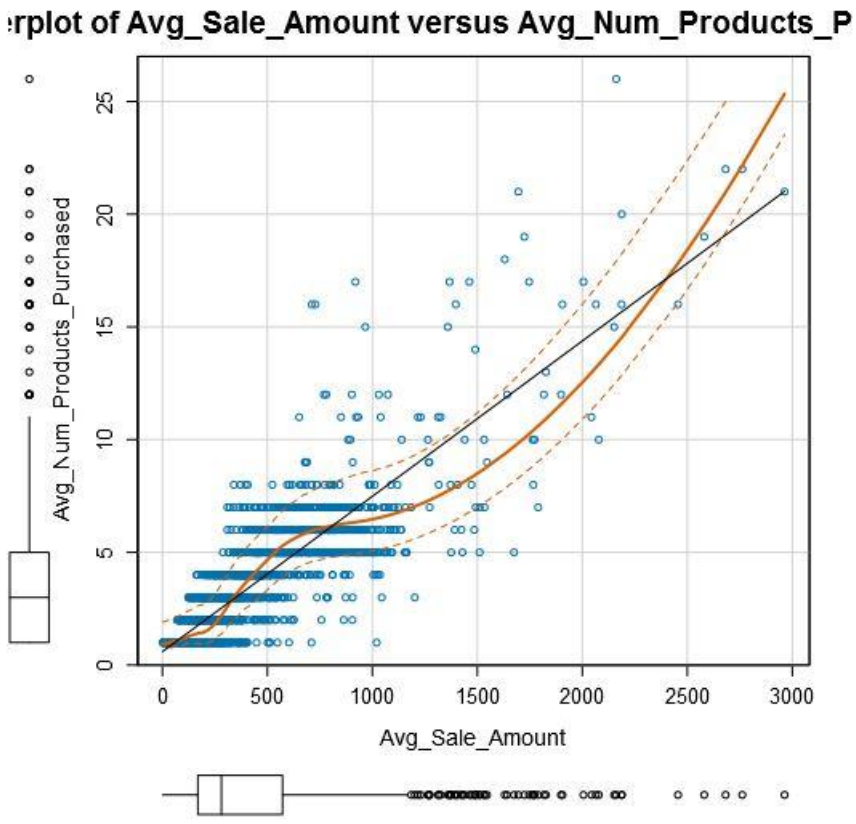
Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 DF, p-value: < 2.2e-16

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



Linear Regression equation:

Y = 303.46 + (281.84 X Customer_SegmentLoyalty Club and Credit Card) + (-149.36 X Customer_SegmentLoyalty Club Card Only) + (-245.42 X Customer_SegmentStore Mailing List) + (66.98 * Avg_Num_Products_Purchased) + Credit Card * 0

The linear model developed is a good model because:

- Multiple R-squared: 0.8369 and adjusted R-squared: 0.8366 values are high and the predictor variables are highly significant as per p-values.

Step 3: Presentation/Visualization*:*

1) What is your recommendation? Should the company send the catalog to these 250 customers?
   Yes, if the profit is greater than $10,000.

2) How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
   a) The Score field gives the predicted sales amount for the 250 customers.
   b) Then it is multiplied with score_yes which is the probability of buying products.
   c) This individual values are totaled.
   d) This is multiplied with gross margin of 50% ie 0.5 on products sold via catalog
   e) The total of (250* $6.50) is then subtracted from it
   f) $6.50 is the cost of printing one catalog and 250 is the total number of people in it
   g) The profit is greater than $10000 and hence profitable to send catalog to the new customers

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

Record #	Sum_Expected Revenue	Margin	cost	Final profit
1	47224.871373	23612.435687	1625	21987.435687