Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

This project attempts to model the expected profit from the 250 new customers.

Key Decisions:

1. What decisions needs to be made?

- The management needs a fair estimate on the revenue that can be generated by sending the catalogue to 250 new customers. The decision to send the catalogue to the potential customers can only be made if the expected profit contribution of above $10,000 is met.

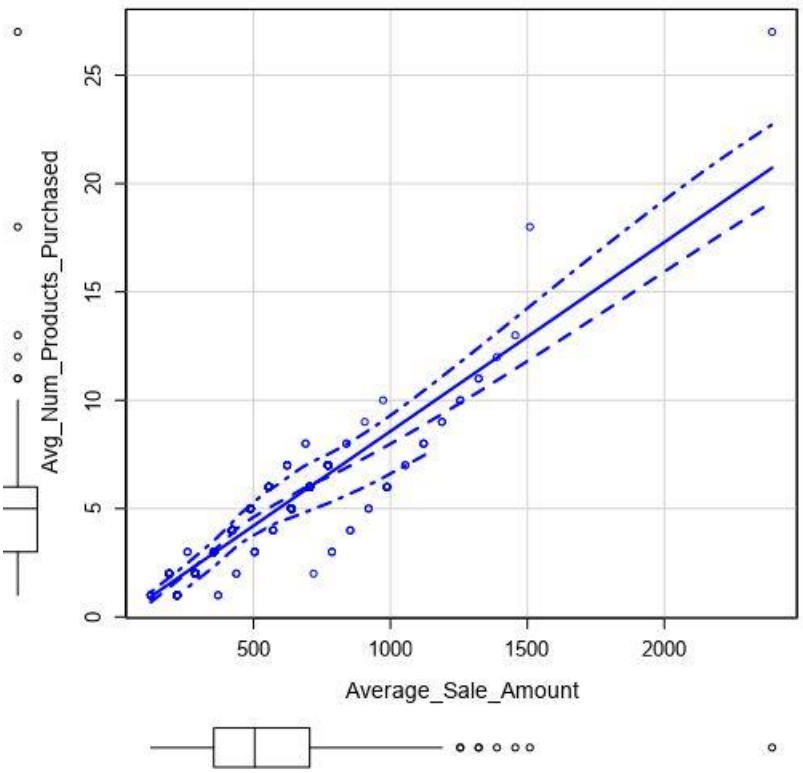
2. What data is needed to inform those decisions?

- The historic data on the last year customers will help evaluate the expected rate of return. This data must hold the appropriate metrics that would significantly affect the revenue generated from a single customer.

Step 2: Analysis, Modeling, and Validation

1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

- In order to build the model, we utilize the past historic data that is provided in the file named p1-customers.xlsx. The target variable chosen for the Multiple linear regression is the Average_sale_amount. The predictor variable such as Customer_Segment, Responded_to_Last_Catalogue and Average_Num_Products_Purchased are seen to be statistically significant. The scatterplot shows a linear slope between the dependent and independent variable.
- The Average Sale amount from the people who Responded yes to our catalog was $156 and the Average Sale amount from the people who did not respond to our catalog was $419, which shows that, after responding positively to our catalog the changes of people buying products from our company is very low.



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.

- The Multiple-Linear-Regression model is the best fit for estimating the expected Revenue from the new customer by studying the past historic data. The model scores a R-Squared value of 0.8373 and an Adjusted R-Squared value of 0.837 which shows to tell the model performs well and there is a significant correlation between dependent and independent variables. All the categorical and numerical variables receive a P-value less than 0.05 which makes them statistically significant.

Report

Report for Linear Model Linear_Regression_3

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment +
Responded_to_Last_Catalog + Avg_Num_Products_Purchased, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-662.58	-67.17	-2.96	69.88	973.88

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	305.00	10.582	28.823	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-150.03	8.967	-16.732	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	281.69	11.897	23.678	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.76	9.815	-24.734	< 2.2e-16	***
Responded_to_Last_CatalogYes	-28.17	11.259	-2.502	0.01241	*
Avg_Num_Products_Purchased	66.81	1.515	44.099	< 2.2e-16	***

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

Residual standard error: 137.33 on 2369 degrees of freedom
Multiple R-squared: 0.8373, Adjusted R-Squared: 0.837
F-statistic: 2438 on 5 and 2369 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	28347499.57	3	501.02	< 2.2e-16	***
Responded_to_Last_Catalog	118080.95	1	6.26	0.01241	*
Avg_Num_Products_Purchased	36677336.71	1	1944.74	< 2.2e-16	***
Residuals	44678788.12	2369			

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

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)

Average_Sale_Amount = 305 – 150.03 x *Customer_Segment* (Loyalty if Yes : 1) + 281.69 x *Customer_Segment* (Loyalty Club or Credit Card if Yes:1) – 242.76 x *Customer_Segment* (Store Mailing List if Yes :1) – 28.17 x (Responded to Last Catalogue if Yes: 1) + 66.81 x (*Average_no_of_prod_purchased*)

Step 3: Presentation/Visualization

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

- As the expected rate of return (profit) - $21,987 is double the threshold value $10,000 set by the management, it is safe to say that the company should sent out these catalogs to 250 new customers given in the mailing list.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

- After evaluating the past data and modeling a Multiple-Linear-Regression to assume a line of best fit, we feed in the new data to get the estimated Revenue earned from each customer. After multiplying the revenue with the probability that the person would buy from us, we get our share of revenue. This revenue amount is summed up and multiplied with our gross margin of 50% to get $23,612.43. The cost of printing and distributing the catalog is $6.50/customer, which has to be factored in to calculate the profit for these 250 new customers.

- Cost of Catalog Distribution: $6.50 x 250 (Customers) = $1625
- Expected Profit: $23612 - $1625 = $21,987