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?**

A decision has to be made regarding a promotional activity in which a catalogue is being sought to be sent to a database of 250 new customers. The decision would be based on expected profit from these 250 customers.

2. **What data is needed to inform those decisions?**

To predict profit, we can follow the below approach:

1. Previous year data of sales from catalogue activity- To build a predictive model
2. Data of present target customers
3. Probability of buying
4. Profit Margin on each sale
5. Other expenses like catalogue printing costs
6. Threshold of Net Margin over which we would go for the catalogue activity. This is $10,000

Basis our predictive model, we can predict current sales. This multiplied by the Probability of buying and the profit margin would give us the expected gross margin. This would be subtracted by the catalogue printing costs to give the net margin. The net margin would help us with the decision

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.

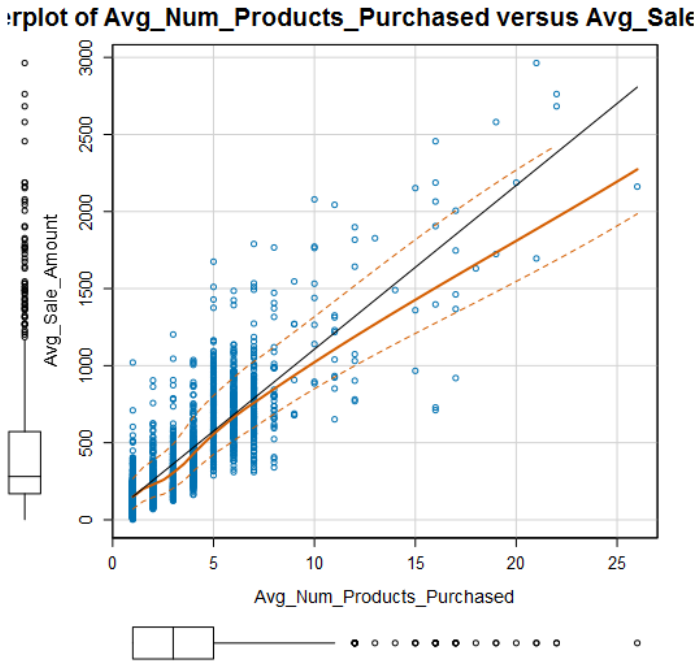
At the minimum, answer these questions:

1. **How and why did you select the predictor variables 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.**

Factors like Name, Customer ID and Address are unique to individual customers and hence am not including them in the analysis.

As a step 1, I did scatter plots of avg_sale_amount with all the numeric variables

a) Avg_Sale_Amount has a linear relationship with Avg_Num_Products



As step 2, I did a Linear Regression with the Avg_Num_Products and the Categorical variables which could have an effect on the sales. The model is being predicted for the new customers only so the Responded_to_Last_Catalog is not valid. Hence, finally considering two variables:

a) Customer Segment
b) Avg_Num_Products_Purchased

Report

Report for Linear Model Linear_Regression_2

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			

We can pick up factors which have a lower p value than 0.05 such that we can reject the null hypothesis. Changes in these predictor variables would lead to a change in the predicted variables which is the Avg_Sale_Amount.

Finally, we have the below factors which impact the total sales:

1. Avg_Num_Product
2. Customer_SegmentLoyalty Club and Credit Card
3. Customer_SegmentLoyalty Club Only
4. Customer_SegmentStore Mailing List

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 linear model has Multiple R-squared of 0.8369 and Adjusted R-squared of 0.8366. These corresponds to a high strength model. Further, the p values of the predictor values are less than 0.05. These indicate that this is a good prediction model for the total sales.

R^2

The R-squared represents the variation in the target variable which can be explained by the variation in the predictor variables. The values are between 0 and 1. The greater the value, the greater amount of variation can be explained by the predictor variables.

In this case, the values of R-squared are more than 0.8- indicated a high strength moodel.

p-value

A p-value of below 0.05 indicates that the relationship between the predicted and the target variables is not by chance. In other words, it indicates that the probability of a relationship between predicted value and predictor variable is high. In this case, the p-values of the predictor values are less than 0.05, indicating a high probability of a relationship between predicted and predictor variables.

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 Linear Model basis the present data and the significant variables is as below:

The final equation is as below:

Avg_Sales_amount = 303.46 +
(281.84 X Customer_SegmentLoyaltyClub and Credit Card) +
(-149.36 X Customer_SegmentLoyalty Club Card Only) +
(-252.42 X Customer_SegmentStore Mailing List) +
(66.98 * Avg_Num_Products_Purchased) +
Credit Card * 0

Step 3: Presentation/Visualization

Use your model results to provide a recommendation. (500 word limit)

We have developed a linear model to predict sales and hence profits. As shared above, the predictor variables are as below:

1. Avg_Num_Product
2. Customer_SegmentLoyalty Club and Credit Card
3. Customer_SegmentLoyalty Club Only
4. Customer_SegmentStore Mailing List

The relation to the target variable has been found to be statistically significant. The *R-squared values of the linear model are more than 0.8.*

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

The recommendation is to go ahead with sending the catalogue to the 250 customers.

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

a) Basis the linear model, we predicted expected sales from the 250 customers
b) This was multiplied by the probability of buying for every customer to get the predicted revenue
c) After multiplying this by the gross margin, we get the expected gross margin
d) The Gross margin subtracted by the cost of printing a catalogue gives us the net margin
e) The summation of this for all 250 customers gives us the total expected net margin
f) This number is more is the 2X the threshold

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

The expected profit is $21987.