Project 2: Predicting Catalog Demand

Step 1: Business and Data Understanding

We have to analyze a business problem in the mail-order catalog business, were we have to predicting how much money the company can expect to earn from sending out a catalog to 250 new customers and provide a recommendation to management based on the expected profit contribution.

Key Decisions:

1. What decisions needs to be made?

The decision that needs to be made is whether the company should send out a catalog to 250 new customers from their mailing list, or not, based on the requirement to meet the profit minimum specified by management, which is $10,000.

2. What data is needed to inform those decisions?

In order to inform the decision, and predict the average gross margin, we require the following data:

1. To build the model and predict the net sales, we require the followings:
    a. Data regarding past customers behavior on older catalog shipments, such as: transactional information, for example: sales amount, the customer segment and the number of products each one purchased;
    b. Additionally we also need new customer's data and factors such as yes_score and margins;
2. Afterwards, we also need data regarding the cost of goods sold, such as:
    a. Costs of labor and materials/ manufacturing – as specified in Part. 2, project, concept 3 – the average gross margin is 50%, meaning that this costs represent 50%;
    b. Catalog cost per unit.

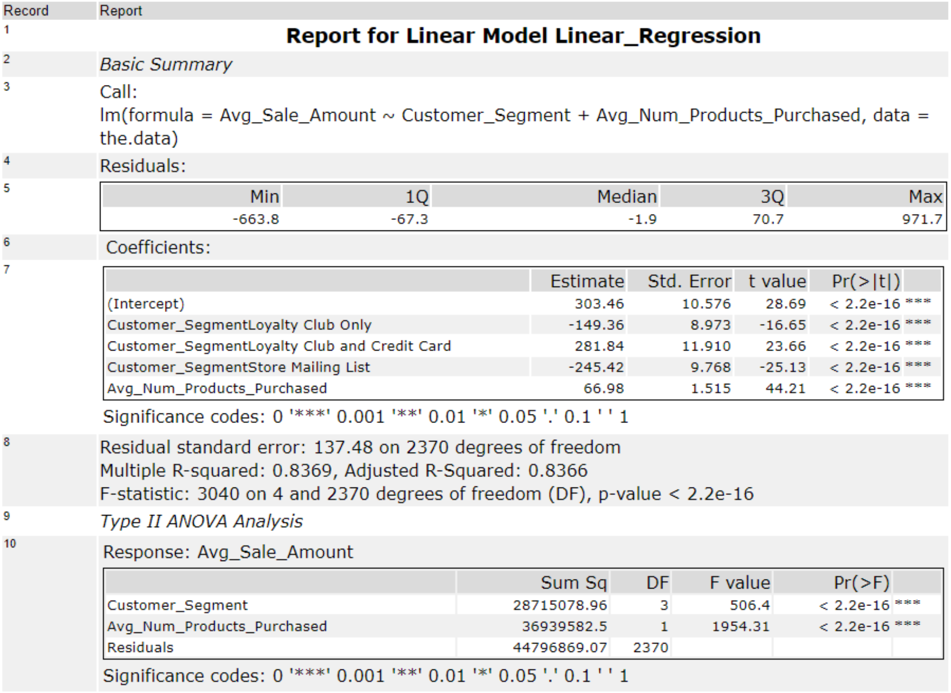
After getting the data and running it in our model, we simply subtract the cost of goods sold from the predicted net sales, thus getting the average gross margin/ predicted profit.

Step 2: Analysis, Modeling, and Validation

Using Alteryx and the p1-customers.xlsx file, I have setup a linear regression using the following variables:

- **Target variables:** Avg_Sale_Amount – I have used this as the target variable because the decision that needs to be made depends on the expected profit contribution;
- **Predictor variables:**
    - Customer_Segment: I considered this categorical variable as significant variable for the model based on the P-values, which is <= 0.05 (< 2.2e-16), also, because the customer segment might impact the average number of products purchased;
    - Avg_Num_Products_Purchased: I have used this as a predictor variable for the model based on the correlation with the target variable (0.8558), also, because a client may buy multiple goods at a time; thus, have an impact on the expected profit.

My decision to use these variables is based on the individual reasons mentioned above and because of the coefficient of determination of this model, were the adjusted R-Squared is 0.8366; the results of the model are as follows bellow in Fig. 1.



Record Report

1 **Report for Linear Model Linear_Regression**

2 *Basic Summary*

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

4 Residuals:

5

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

6 Coefficients:

7

	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

8 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

9 *Type II ANOVA Analysis*

10 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

Fig. 2.1. - Report of the linear model.

1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you have 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.

I have decided to use Avg_Num_Products_Purchased as a predictor variable based on the correlation with the target variable (0.8558) and by visualizing the plot presented in Fig. 2. we can see that we almost have a high positive correlation in cases where the average products purchased is less than 10, but it can be off on more.

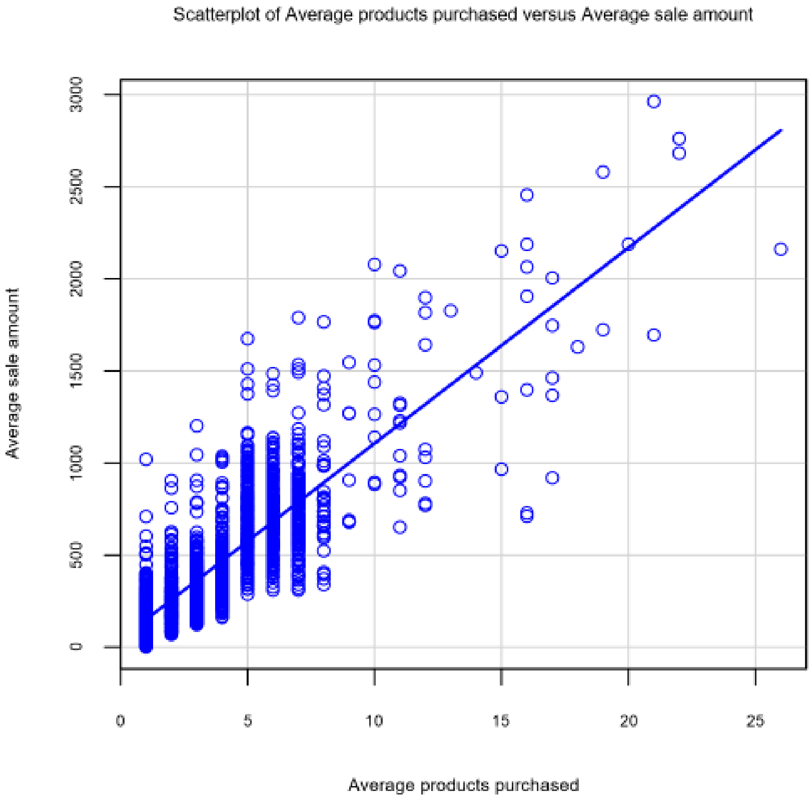


Fig. 1.2. Predictor variables scatterplot.

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.

I consider the Avg_Num_Products_Purchased and Customer_Segment variables a good fit for the model based on the statistical results created by the regression model, which are as follows:

- The Adjusted R-Squared value of 0.8369 is very close to 1, indicating that the variation in the target variable explained by the variation in the predictor variable is very small;
- The p-value is less than 0.05 and more than 2 stars (**), which indicates that this is a good model, where the probability of a relationship between the target variable and the predictor variables is high.

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)

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

Step 3: Presentation/Visualization

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

Based on my model the average gross margin/ predicted profit is $21,987.44, which exceeds management’s minimum expected profit of $10,000, therefore, my recommendation is to send out a catalog to the 250 new customers from their mailing list.

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

I came up with my recommendation based on the following process:

- By using the historical data I have first identified the variables that should be used for my model based on the correlation with the target variable and coefficient of determination, thus, determining the regression model and equation;

- I then used the new customers data and passed it though the linear regression and a score tool in order to predict the net sales of individual people in the mailing list;
- Afterwards, I have summed the predicted net sales of individual people in order to get the revenue/ total net sales;
- In the end, in order to get the average gross margin (price – cost) I have multiplied the revenue by 50% and finally subtracted the cost of printing and distribution of the catalog ($6.50 per catalog).

$$Average\ gross\ margin = Net\ sales - Cost\ of\ goods\ sold$$

$$Cost\ of\ goods\ sold = 50\%\ of\ net\ sale + (\$6.5 * nr.new\ customers)$$

$$Net\ sales = \$47{,}224.87$$

$$Cost\ of\ goods\ sold = (\$47{,}224.87 * 0.5) + (\$6.5 * 250) = \$25{,}237.44$$

$$Average\ gross\ margin\ (profit) = \$47{,}224.87 - \$25{,}237.44 = \$21{,}987.44$$

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

Assuming the catalog is sent to the 250 new customers from their mailing list the expected profit from the catalog is $21,987.44.