Project: Predicting Catalog Demand

# Step 1: Business and Data Understanding

## Key Decisions:

1. What decisions needs to be made?

We consider a company that manufactures and sells high-end home goods. A decision needs to be made to help the company determine, whether or not it should send its catalogs to the new customers on the mailing list. Catalogs will only be sent if the profits exceed $10,000. The question is how much profit the company can expect from sending a catalog to these customers. To answer this question, the central point is accurate sales forecasting. That is, what is the average amount of sales we make to each customer? The available data are the sources of information to help with the sales forecast. The handling of this information in a way that makes it possible to estimate what sales will be. To construct a model of forecasting, it is necessary to make assumptions about some of the relationships and then investigate this assumption. The model includes factors and expresses mathematically the relevant relationships between the factor to be forecast and other factors. The model generates forecasts based on historical data that take past data and project it forward into the future.

2. What data is needed to inform those decisions?

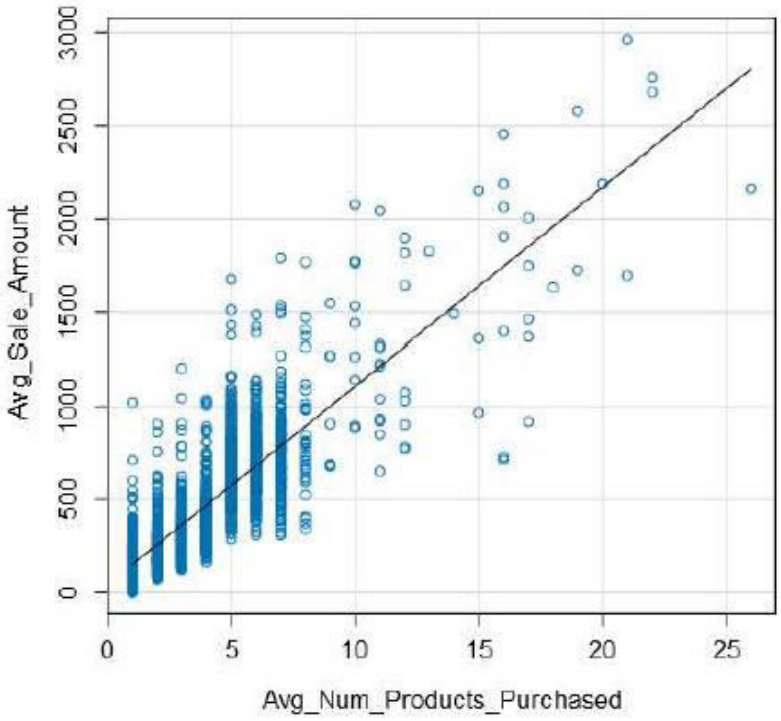
- Last year's sales are the starting point for sales forecast. Therefore, we look at the level of sales, revenue, the dollar amount a company makes and sales volume, items that a business sells.
- Other inputs for sales forecast is to identify the characteristics that can describe customers and develop a customer profile based on some type of shared characteristics is
- Estimate the accumulated total of all costs used to create a product, which has been sold.
- The chance of the sale given that the expected profit is calculated by multiplying each of the possible outcomes by the likelihood each outcome will occur.

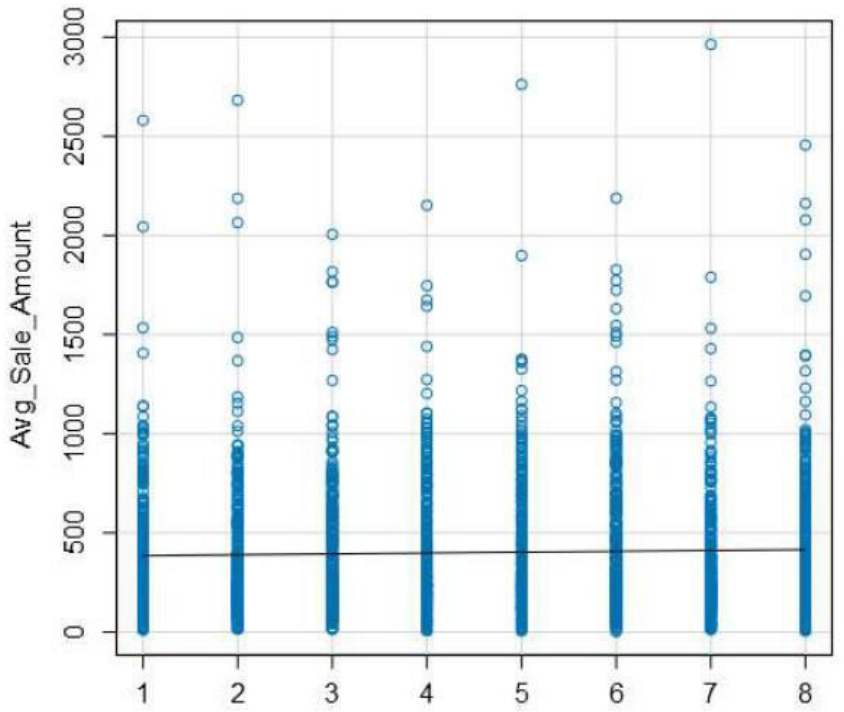
# Step 2: Analysis, Modeling, and Validation

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. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

The dataset available includes the information on about 2,300 customers. Avg_Sale_Amount is the target variable of the model that has been identified to sales forecasting. Among the variables in the dataset, those that we hypothesis been relevant to predict outcome: Avg_Num_Products_Purchased, Years_as_Customer. Only the first variable has a strong linear relationship with target variable. To confirm this conclusion, it is shown below the matrix correlation and scatter plot.

| Record # | FieldName | Avg_Sale_Amount | Avg_Num_Products_Purchased | Years_as_Customer |
|---|---|---|---|---|
| 1 | Avg_Sale_Amount | 1 | 0.855754 | 0.029782 |
| 2 | Avg_Num_Products_Purchased | 0.855754 | 1 | 0.043346 |
| 3 | Years_as_Customer | 0.029782 | 0.043346 | 1 |






The most interesting variable in the dataset that divide or segment of consumers in sub-groups is Customer_Segment. This variable is found to be highly statistically-significant when it has been added to Avg_Num_Products_Purchased in the regression linear model. Furthermore, R squared is increased from 0.7323 to 0.8369.

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 Regression that has been built is a good model, it is shown below the report of model “Sales Forecasting”.

**Report for Linear Model Sales_Forecasting**

*Basic Summary*

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = inputs$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

The factors include are all highly statistical significance at the 0.01 level. The R squared of is 0.8369, and then is a high proportion of the variance in the dependent variable that is predictable from the independent 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)

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

# Step 3: Presentation/Visualization

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

The decision, based on available information, is to send the catalog out to 250 customers given that the expected profit is greater than the established threshold of profit from management.

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

First, we have to calculate the Score of each customer. For this, we multiply [Score] * [Score_yes], being (Score-Yes) the probability that each customer makes a purchase.
Then, to obtain the total sales for the 250 clients, we will make a summation of the Total Score that we have obtained [Sum_Score].
To get the gross margin we have to multiply [Sum_Score] *.5, and to this result, we have to subtract the total expenses of sending the catalogs to the 250 clients [- (6.5 * 250)]. In this way, we will obtain the benefits, which must be > $10,000.

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

The expected profit is equal to $ 21987.44