Predictive Analytics for Business Nanodegree

Project: Predicting Catalog Demand

Laila Hussain Alqawain

_____

Step 1: Business and Data Understanding

We have datasets provided as two files**: *p1-mailinglist.xlsx* and *p1-customers.xlsx***

Key Decisions:

1. What decisions needs to be made?

   The decision to need to be made is that we should be sending the catalog to new 250 customers if the profit will exceed over 10.000$.

2. What data is needed to inform those decisions?

   We need to calculate the expected revenue from new 250 customers for getting the expected profit.
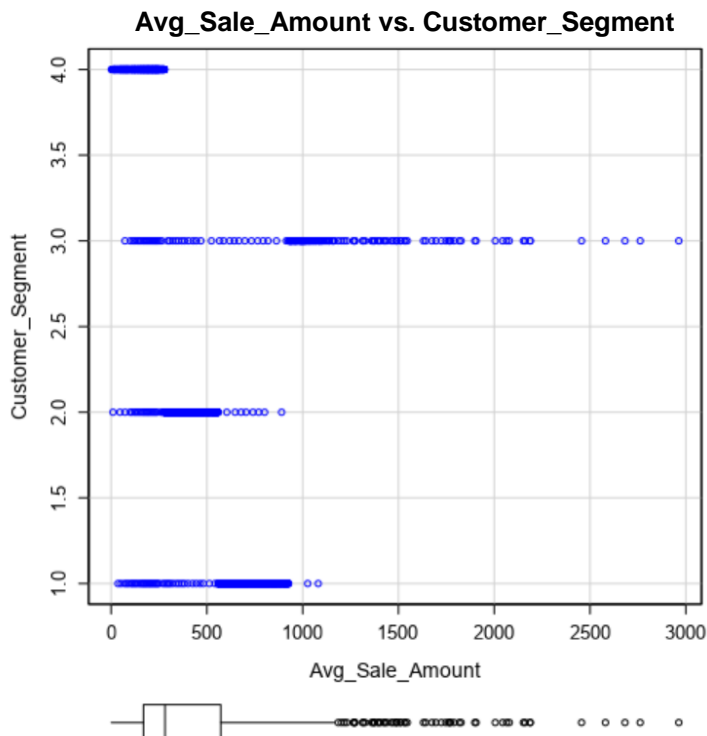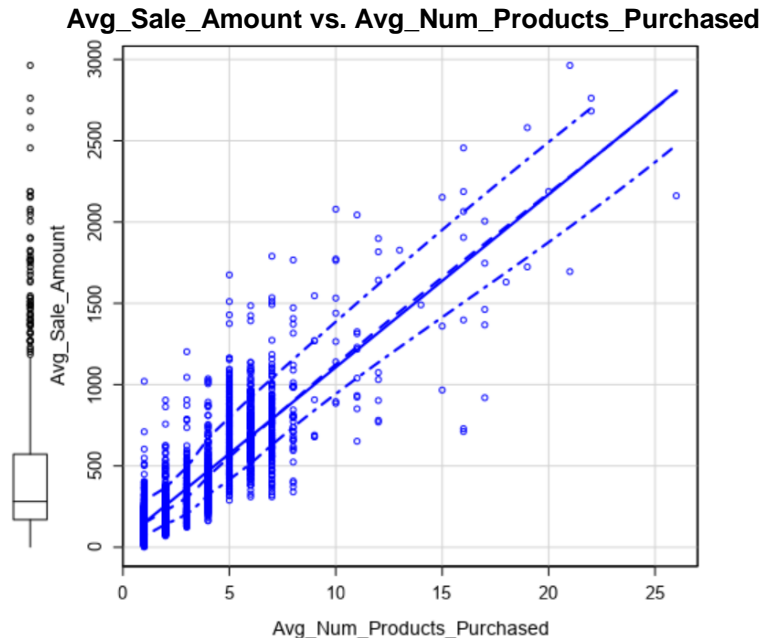   So, we will multiply the predicted sales amount (Score) by probability to buy (Score_Yes) for getting the predicted sales.
   As provided, we will multiply the predicted sales by 50% then subtract out the $6.50 cost for getting the predicted profit.

Step 2: Analysis, Modeling, and Validation

1. How and why did you select the predictor variables in your model?

   I selected Avg_Num_Products_Purchased and Customer_Segment as and predictor variables. Due to the strong correlation between Customer_Segment, Avg_Num_Products_Purchased and Avg_Sale_Amount and it's 0.8369 as shown in the scatterplots below.

Avg_Sale_Amount vs. Avg_Num_Products_Purchased

Avg_Sale_Amount vs. Customer_Segment

2. Why you believe your linear model is a good model?

Report for Linear Model linear

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		

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

From Report for Linear Model linear above, there is a strong correlation between the preductors variables which I selected and it shows in R-squared: 0.8369 and Adjusted R-Squared: 0.8366. All these are above 0.7. That means a strong relationship. Also, P-value less than 0.05. All these factors indicate that the model is very good.

3. What is the best linear regression equation based on the available data?

Avg_Sale_Amount = 303.46

– 149.36 * (Customer_Segment: Loyalty Club Only)

+ 281.84 * (Customer_Segment: Loyalty Club and Credit Card)

– 245.42 * (Customer_Segment: Store Mailing List)

+ 66.98 * (Avg_Num_Products_Purchased)

+ 0 * (Customer_Segment: Credit Card Only).

Step 3: Presentation/Visualization

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

   Yes, it should. I recommend the company send the catalog to these new 250 customers. Due to the results reached recently.

2. How did you come up with your recommendation?

   I came up with my recommendation by using Alteryx program. I built a Linear Regression Model. After calculating the Predicted Profit, I found the sum of Predicted Profit exceed 10,000$.

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

   Sum of Predicted Sales = $47,224.87
   Costs of Printing and Distributing = $6.5
   Average Gross Margin = $0.5
   Predicted Profit = (47,224.87 * 0.5) − (6.5 * 250) = **$21,987.44**

For more details, click **here**.