Predictive Analytics for Business Nandegree

Project 1: Predicting Catalog Demand

Latifa M. Alyaeesh

Misk Academy & Udacity

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

The Business Problem

You recently started working for a company that manufactures and sells high-end home goods. Last year the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.

Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers. You, the business analyst, are assigned to help your manager run the numbers. While fairly knowledgeable about data analysis, your manager is not very familiar with predictive models.

You’ve been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Key Decisions:

Answer these questions

1. What decisions needs to be made?

The decision needs made to determine how much profit the company can expect from sending catalog to these 250 new customers. Because of Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

2. What data is needed to inform those decisions?

We need to calculate the average sales amount by using both average number of products purchased and the customer segment. Then, we need to multiply it with the probability that a customer will buy, in order to get the predicted sales per customer. Once we have this information, we aggregate the predicted sales of each customer to get a total of predicted sales, which eventually we can use to predict the expected profit of sending out these catalogs.

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 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.**

Let me know to explain how I can create a linear regression model by using Altryx tool. First, we need to understanding the relationship between the predictor variables and target variables. First, I started by using an input tool to bring data from p1_customer Excel spreadsheet. Next, I used scattoeplot in order to know which can be an excellent predictor variable for a target variables. Then, I used linear regression tool in order to build linear model. After that, I used Score tool in order to reach out to the target variable.

For Numeric variables, we can use scatterplots between an individual variable and the target variable to see if a variable might be a good candidate for a predictor variable or not so, I chose the predictor variables by using scattorplot tool between the target variable and all the numeric individual variables. Each predictor variable should be significant (p-value <= 0.05). By doing this, I found out Avg_Num_Products_Purchased , Customer_Segment are the good predictor variables for the target variables which is the Avg_Sale_Amount.

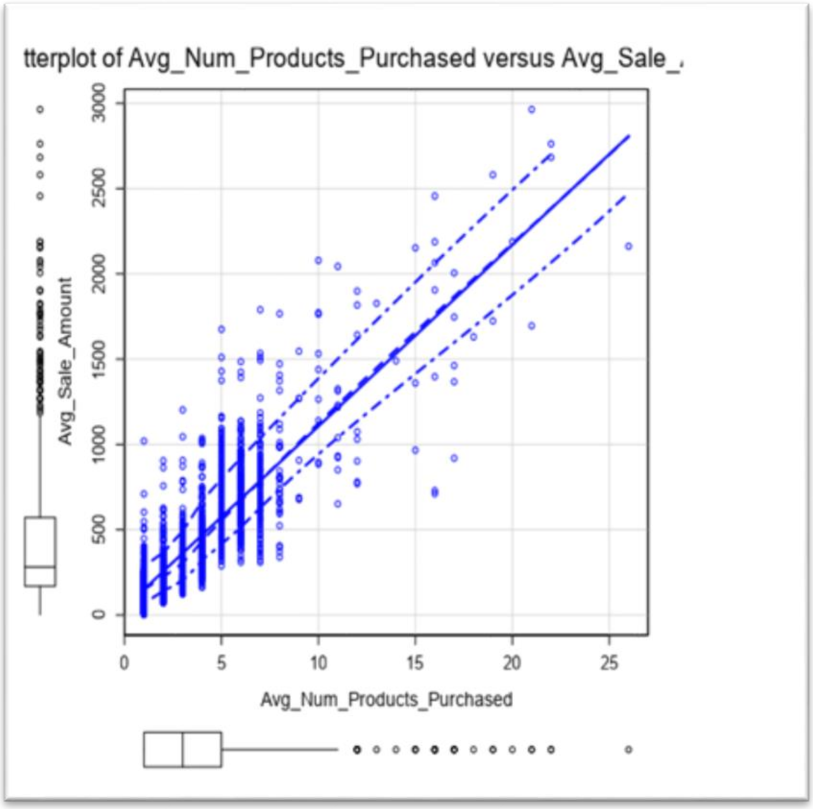
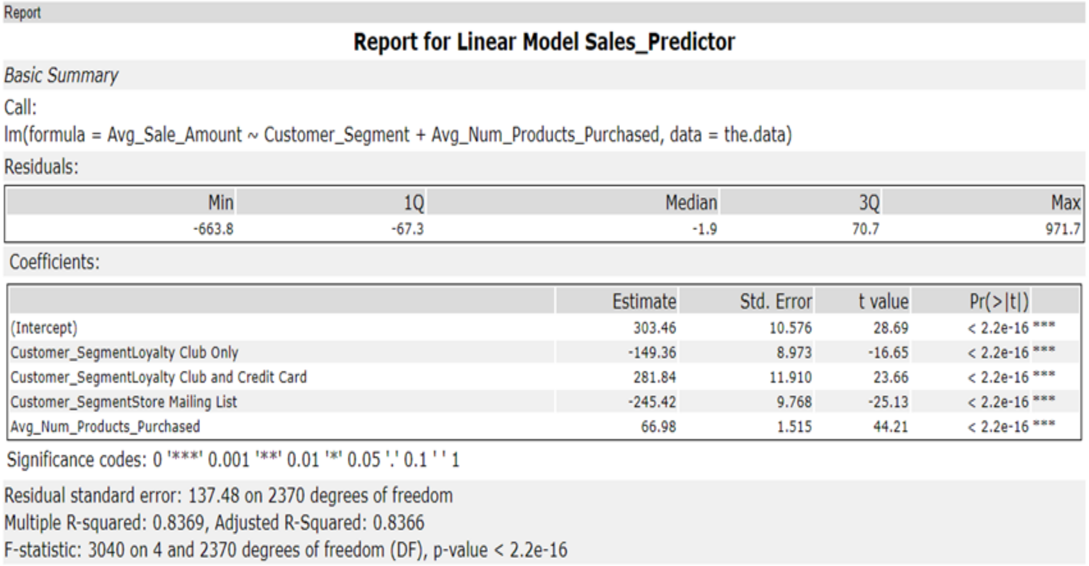


Figure 1: Scatterplots of customer segment vs. Average sales Amount

The Scatterplot of Avg_Num_Products_Purchased (Predictor Variable) and Avg_Sale_Amount (Target Variable), there is a linear relationship between them which indicates that it is a good predictor variable for this target variable.



Report

Report for Linear Model Sales_Predictor

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

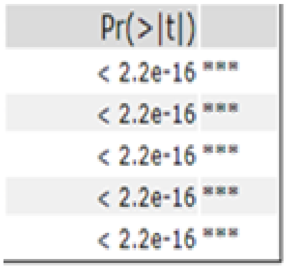
Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 degrees of freedom (DF), p-value < 2.2e-16

Figure 2: Report for Linear Model Sales Predictor

P_Value of Customer segment categorical variable, as shown above, is less than 0.05 and since it has ***, it is statistically variable. Therefore, it is a good candidate for a predictor variable.

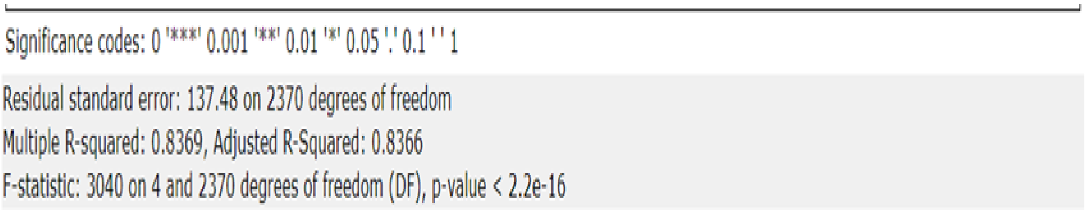
2- Explain why you believe your linear model is a good model. You must justify your reasoning using the statistical results that your regression model created. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

p-values and R-squared values are used to justify how well the linear model works. As you can see in the output below, P_Values of the selected predictor variables (Avg_Num_Products_Purchased , Customer_Segment) in this report of Linear model for sales predictor is less than 0.05. So, it represents that there is a relationship between the predictor variables and the target variable and it is statically significant since they have more than *



```
Pr(>|t|)
< 2.2e-16 ***
< 2.2e-16 ***
< 2.2e-16 ***
< 2.2e-16 ***
< 2.2e-16 ***
```

As you can see in the output below, Multiple R_squared: 0.8369, Adjusted R_squared: 0.8366 so the values range from 0.8369 to 0.8366, which shows the amount of variation that represents a high explanatory of the model. Therefore, low P_values and high R_squared values indicated that the model is highly predictive.



```
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
```

Hint: Low P-values and a high R-squared suggest the model is highly predictive, Low P-values means it is highly unlikely that the two variables are not related. Low R squared means the model is not very fit.

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)

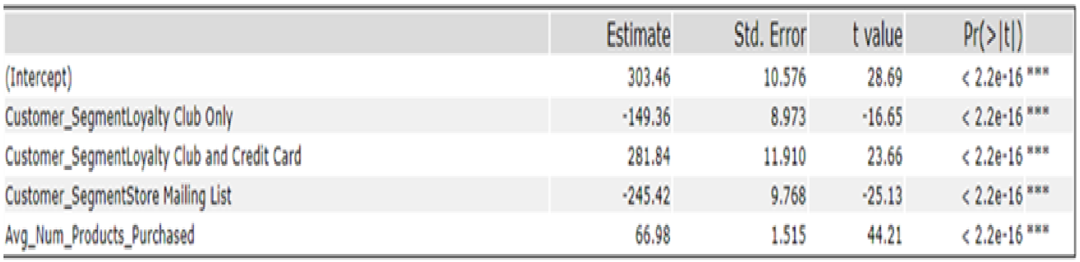
Important: The regression equation should be in the form:

Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we must include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.



	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 ***

Y= 303.46+ -149.36* Customer_SegmentLoyalty Club Only + 281.4 * CustomerSegmentLoyalty Club and Credit Card + -245.42* Customer_SegmentStore Mailing List+66.98* Avg_Num_Products_Purchased + 0 * Customer Credit Card only

Step 3: Presentation/Visualization

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

At the minimum, answer these questions:

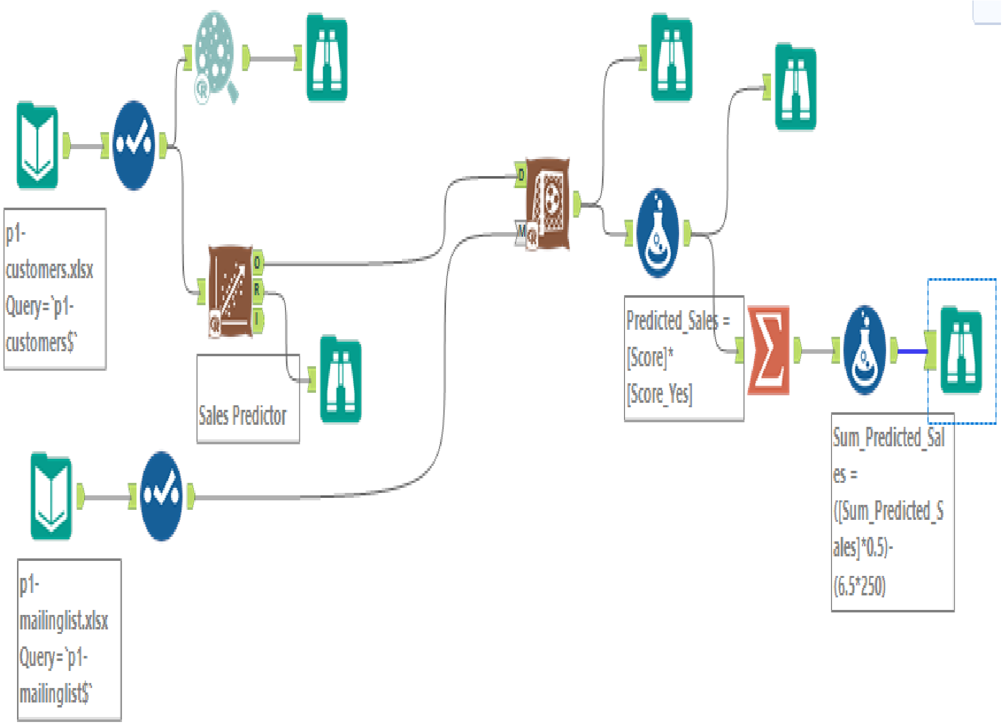
1. **What is your recommendation? Should the company send the catalog to these 250 customers?**

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

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

First, I need to calculate the Predicted_Sales for every customer by multiplying [Score] by [Score_yes], where [Score_yes] is the probability that every customer will buy. Then, in order to get the total predicted sales for all 250 new customers, I have to make a summation of the Predicted_Sales. Next, in order to get the gross margin, I multiply the [Sum_Predicted_Sales] by 0.5. After that, I subtract the total expense of the catalogs for 250 new customers ([Sum_Predicted_Sales * 0.5] – (6.5 * 250)) by doing the Total Predicted Profit, which is 21,987.4356865455. Therefore, I highly recommended the management to send the catalogs to the 250 new customers, since the predicted profit will be more than $10.000.

Alteryx’s workflow:



p1-customers.xlsx Query=`p1-customers$`

Sales Predictor

p1-mailinglist.xlsx Query=`p1-mailinglist$`

Predicted_Sales = [Score]* [Score_Yes]

Sum_Predicted_Sales = ([Sum_Predicted_Sales]*0.5)-(6.5*250)