Project: Predicting Catalog Demand

Step 1: Understanding the Business problem and Data.

The company has 250 new customers from their mailing list that they want to send the catalog to. The company wants to decide whether to send those customers a catalog. To do that they want to make sure that, the expected profit contribution of those customers exceeds $10,000. Therefore, I need to predict the expected profit from those 250 new customers. I created a regression model based on the already existing data of the past year. After creating this regression model, I calculated the expected sales of each new customer and then the potential total profit in case the company decides to send them a catalog.

First, I split the data into training data and validation data. In these data sets, I identified the target variable. In this case, it is the average sale amount for each former customer. I also identified the potential predictor variables like

Customer_Segment,
City,
Responded_to_Last_Catalog,
Avg_Num_Products_Purchased and
Years_as_Customer.

Step 2: Analysis, Modeling, and Validation

In the first step, I checked the data for data cleansing/data featuring and verifying that the data meets all the assumptions of a linear regression procedure. The data were good to go. Based on my knowledge of the field and variables I ran some initial models to see how well certain predictors predict the target variable and based on the results of these initial analyses I decided which variables to include and which not. In my initial model, I included all the variables I mentioned above and later I removed the ones which did not seem to be significant predictors. I decided to go with customer_se
gment (categorical variable), and the average number of the products purchased, because they were the significant variables (p-value <= 0.05) and the explanatory power of the model was pretty good. (Both R Squared and adjusted R squared values are .83).

Browse (12) - Configuration

12 records displayed, 2 fields, 168 KB

Report | Profile

1 of 1 Fields | Records 1 to 10

Record	Report
1	**Report for Linear Model Linear_Regression_Project_1**
2	*Basic Summary*
3	Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)
4	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

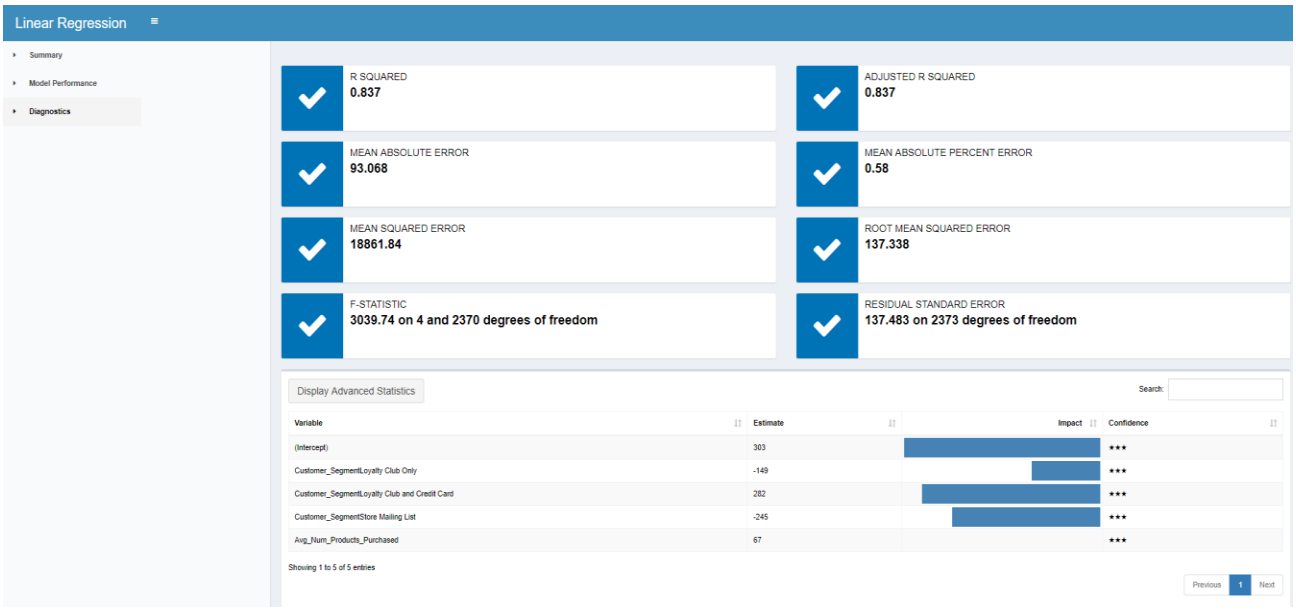
Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 degrees of freedom (DF), p-value < 2.2e-16

Type II ANOVA Analysis

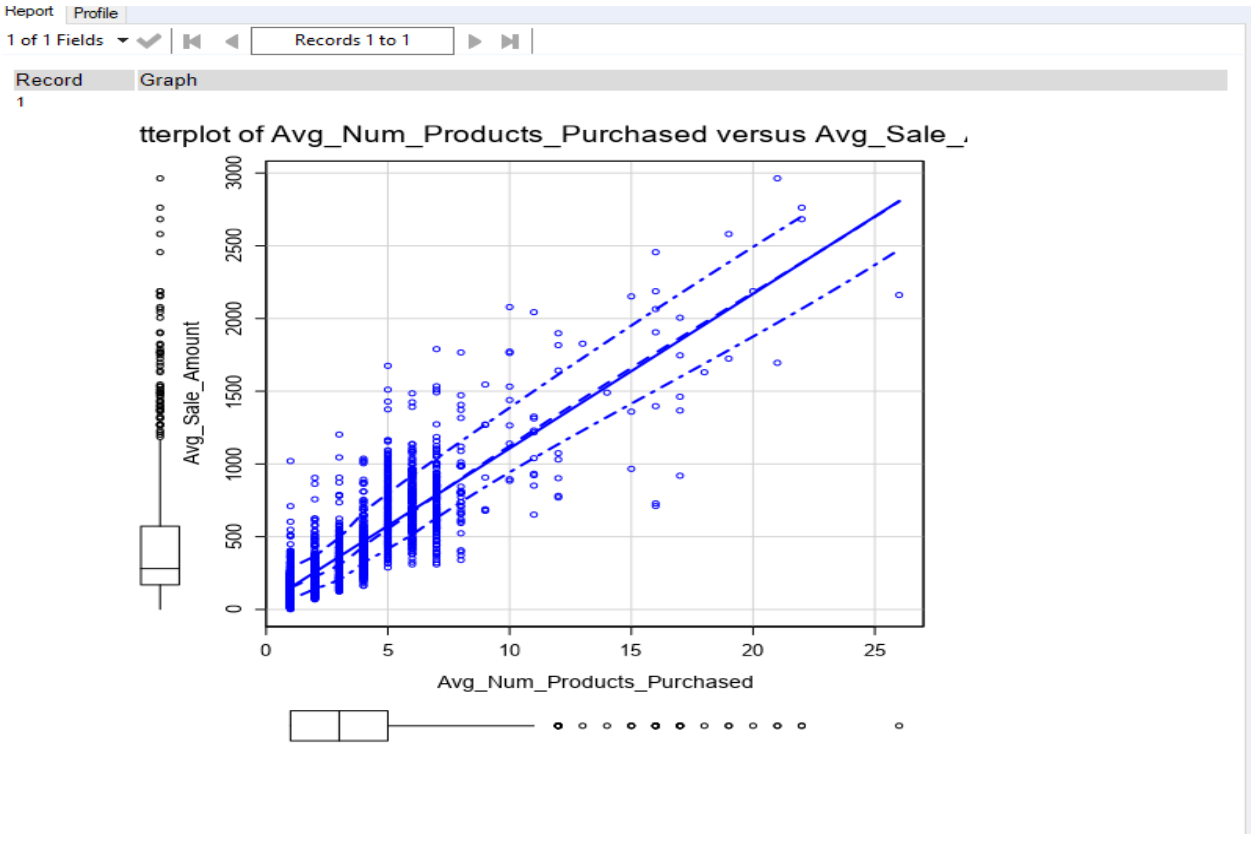
Response: Avg_Sale_Amount

	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



As can be seen from the scatterplot below, there is a linear relationship between the average number of product purchased and the Average sale amount. Since one of my predictor variables “Customer_Segment” is a categorical variable scatterplot between this and the target variable is not available. I plotted the relationship between Avg_Num_Products_Purchased and Avg_Sale. There seems to be a linear relationship between the two variables.



Overall, the model explains .83 of the variation in the target variable which is pretty much good. Both R Squared and adjusted R squared values are .83 and the predictor variables that I used have all statistically significant relationship with the target variable.

The Ideal Linear Regression Equation

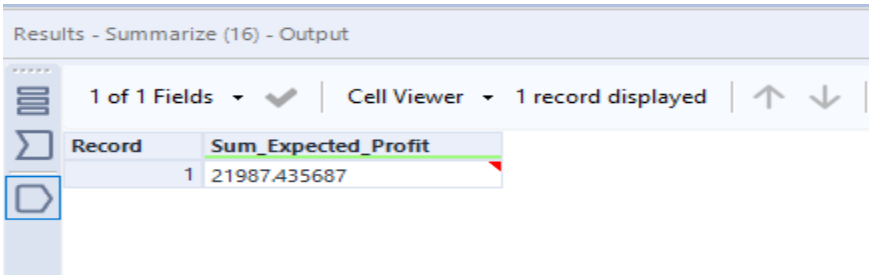
Here is the ideal linear regression equation I designed.

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

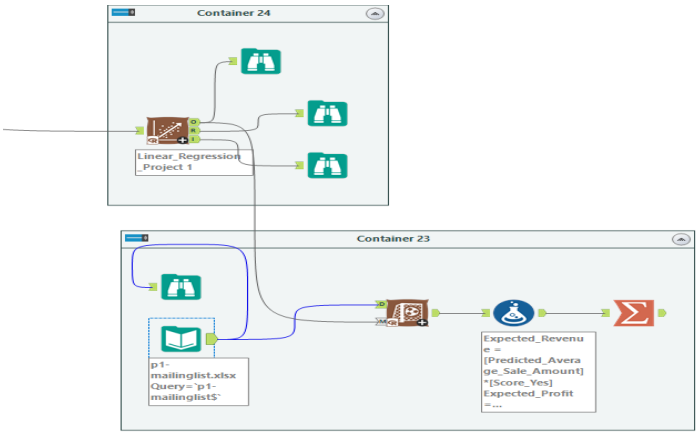
Important: I included the 0 coefficient for the type Cash.

Step 3: Presentation/Visualization

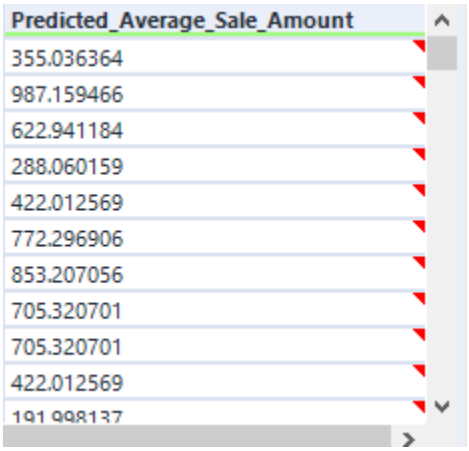
Based on my analysis my recommendation is that the company should send the catalog to those 250 customers, *because as can be seen from the results the expected profit amount from those customers if the catalogs are sent) is $21, 987.43. The company sees $10.000 profit projection enough to send the catalogs to the new customers. The projected profit in this case more than double therefore the company should send the catalogs.*



To come up with my recommendation I first created the ideal regression equation as I explained above. The results of the regression analysis and the coefficients can be seen above. After creating the ideal regression equation, I applied this to the new data with the 'Tool Score' in Alteryx.

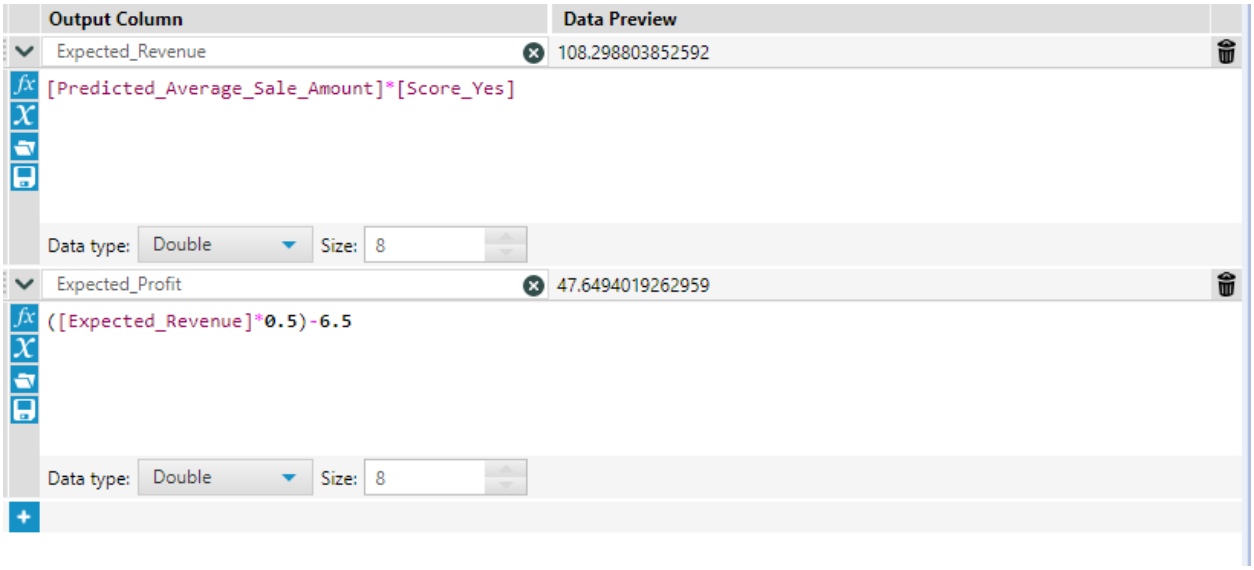


Via tool score I applied the model to the new dataset and created the predicted average sale amount for every single new customer.



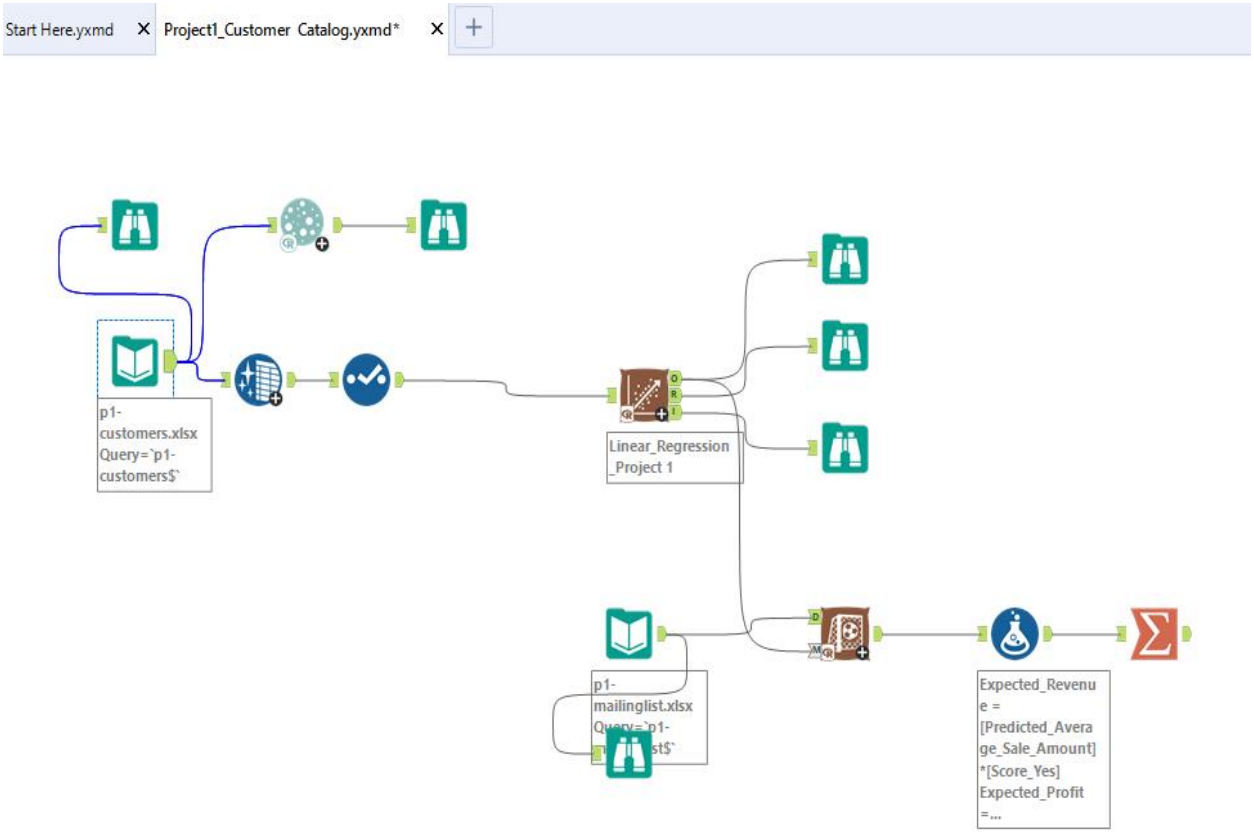
Predicted_Average_Sale_Amount
355.036364
987.159466
622.941184
288.060159
422.012569
772.296906
853.207056
705.320701
705.320701
422.012569
191.998137

Then with the formula tool I calculated the expected revenue for each customer; multiplying the predicted sale score with the probability of the customers to buy the products (yes score). Then multiplied it with the gross margin to calculate the profit and then subtracted the cost of the catalog. Then I added up all the values and found the total expected profit which was $21,987.43



The expected profit from the new catalog (assuming the catalog is sent to these 250 customers is $21,987.43

Below is the Alteryx workflow for the analysis.



Prepared by **Kadir AKYUZ**