Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

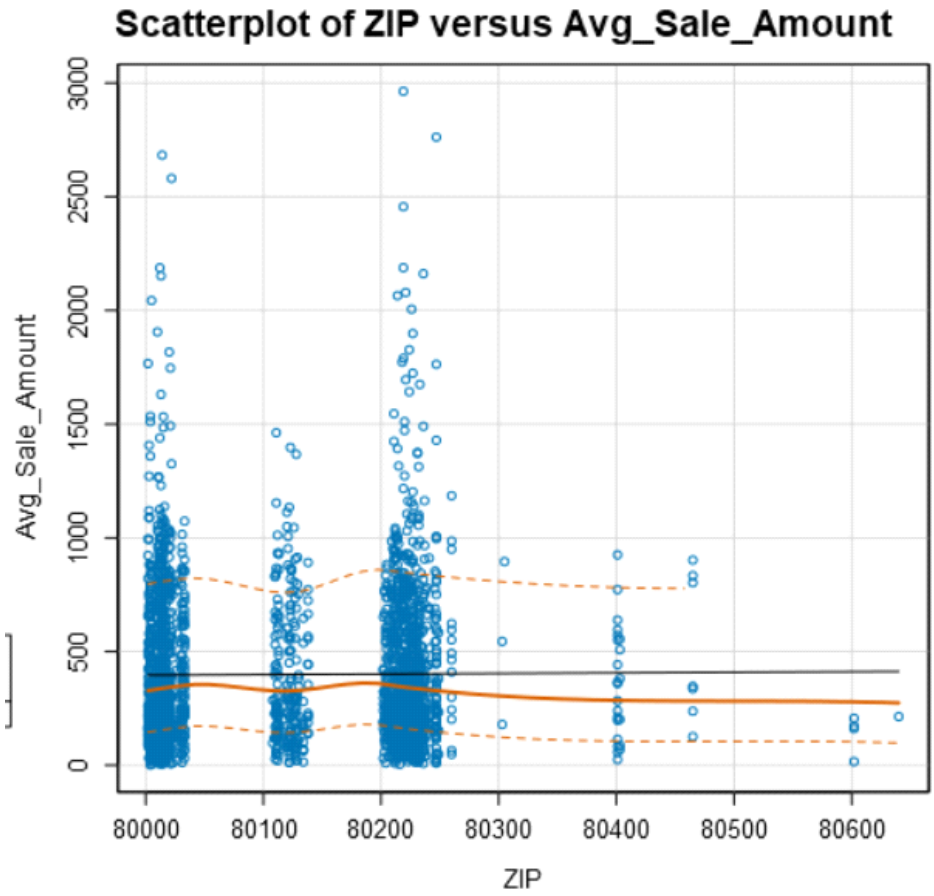
As a Business Analyst in the company, I'm required to predict or estimate the profit that the company will get from the new 250 customers. If the predicted profit exceeds $10,000 then the company will go ahead and send catalogue to these new 250 customers otherwise they'll drop the idea.

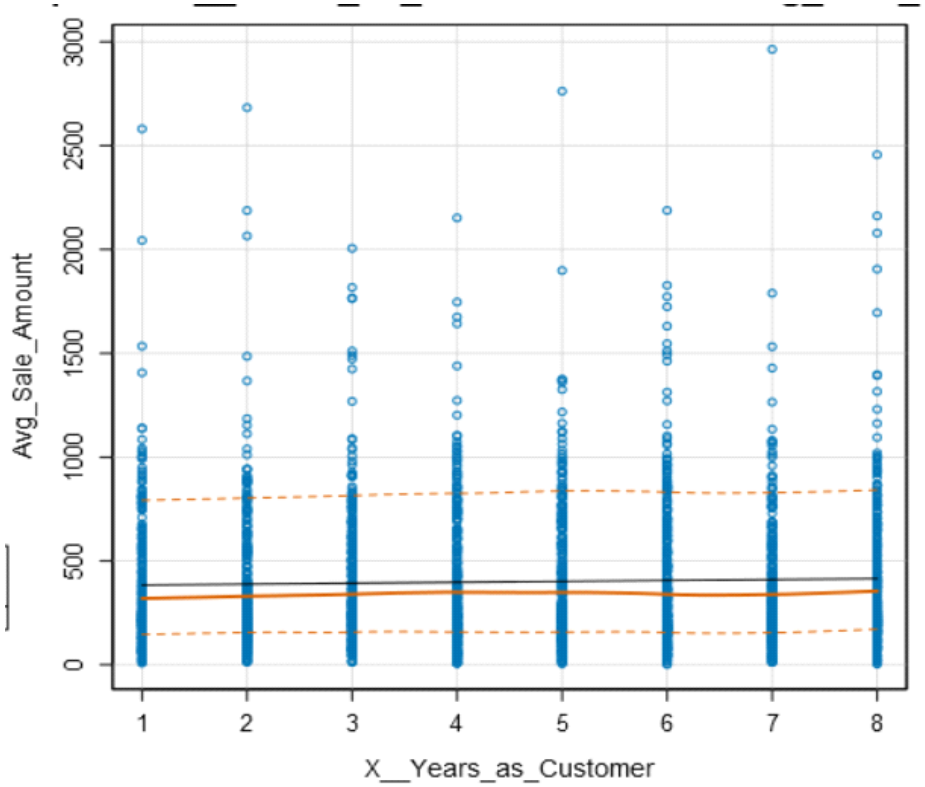
In order to predict the profit for these new customers, it is essential to have:
a) Data collected from existing customers (to calculate expected Sales/Revenue),
b) Gross Margin and Cost per catalogue (to calculate Profit).

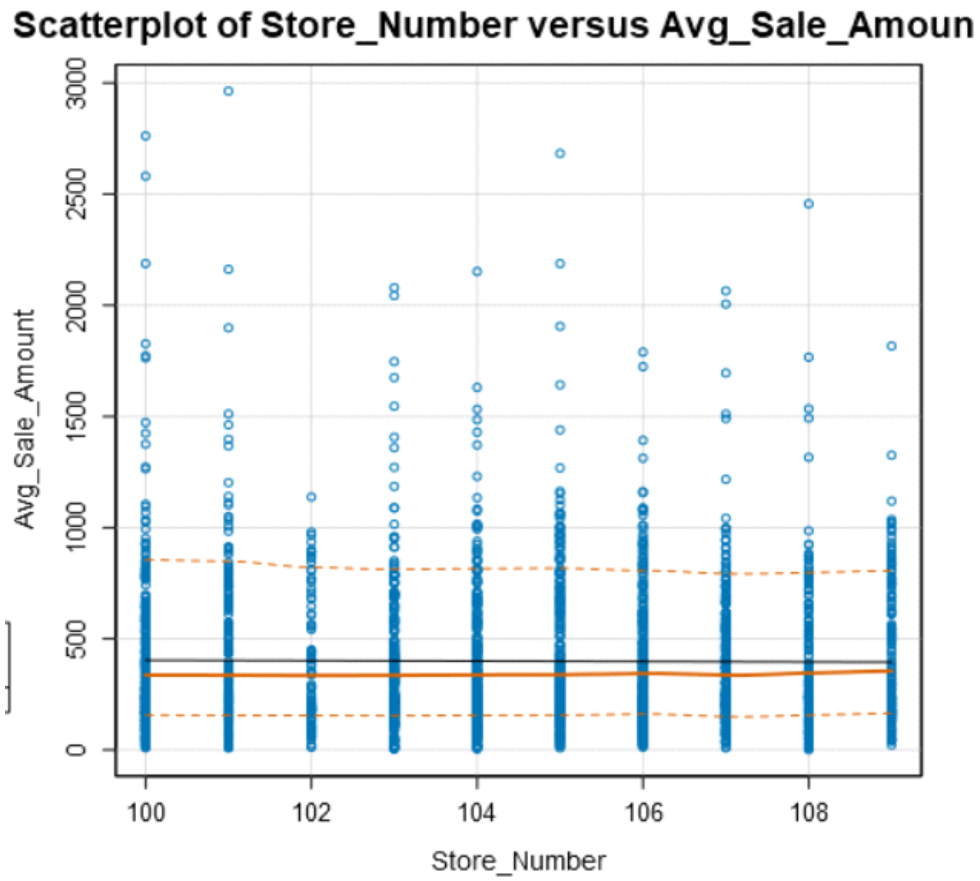
Step 2: Analysis, Modeling, and Validation

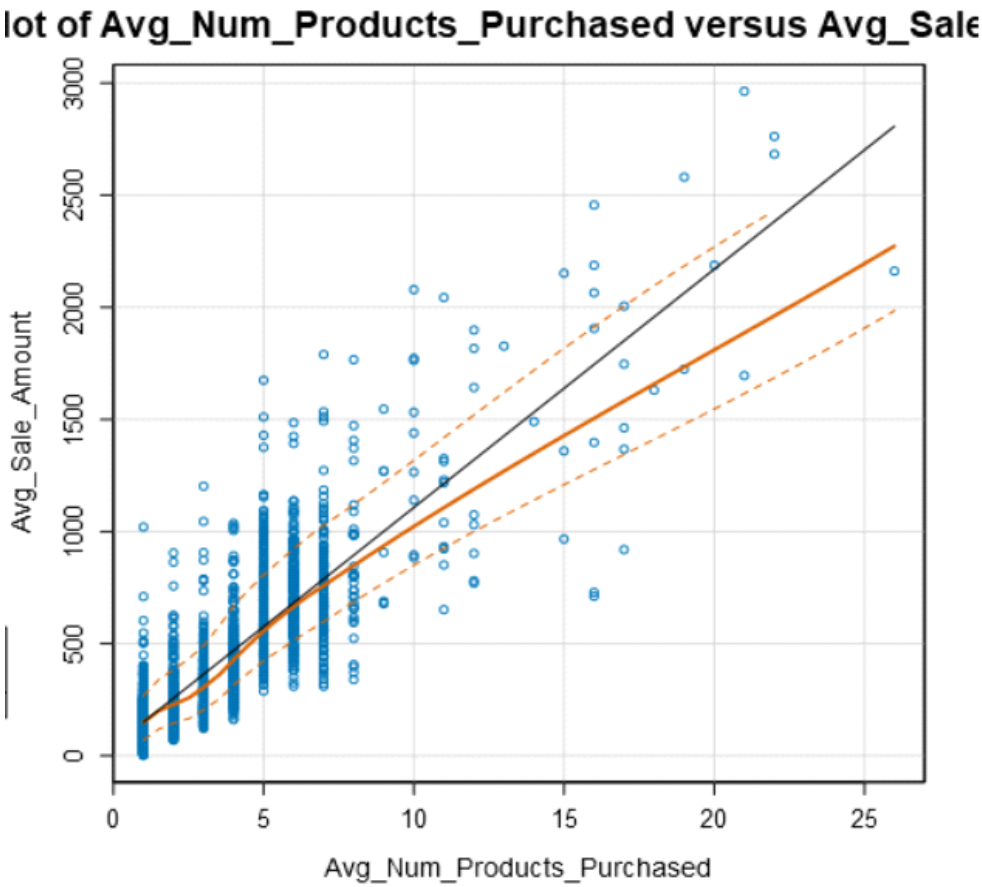
As mentioned in the course, it is generally preferred to check linear relationship of the predictor variable with the variable to be predicted. Hence, I plotted scatter plots of **Avg Sale Amount vs Zip, # of years as customer, Store Number and Avg Number of Products Purchased**. There was no need of plotting Avg Sale Amount against Customer ID as it is unique for every customer.
Other fields provided were strings.
As I plotted the scatter plots, I noticed that only 'Average Number of Products Purchased' has linear relationship with 'Average Sales Amount'. So, I took this variable in consideration.

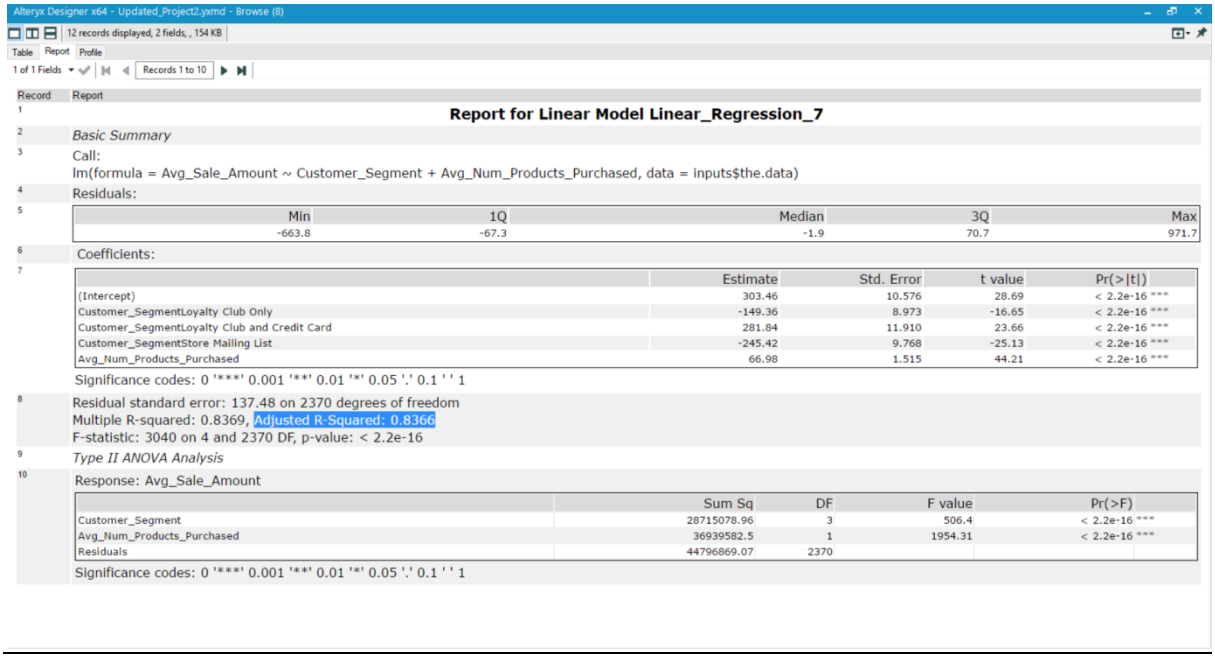
After this, I used Linear Regression Model and used 'Average Number of Products Purchased' in consideration with other string variables. I noticed the 'p' value in the report after running the model along with the 'Adjusted R squared value'. The variable which had 'p' value less than 0.05 was 'Customer Segment', so it was also selected as a predictor variable. Other variables had 'p' value greater than 0.05 which indicates that they don't have any strong relationship with the predicted variable.


Scatterplot of ZIP versus Avg_Sale_Amount
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
80000
80100
80200
80300
80400
80500
80600
ZIP


Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
1
2
3
4
5
6
7
8
X__Years_as_Customer


Scatterplot of Store_Number versus Avg_Sale_Amoun
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
100
102
104
106
108
Store_Number


lot of Avg_Num_Products_Purchased versus Avg_Sale
Avg_Sale_Amount
0
500
1000
1500
2000
2500
3000
0
5
10
15
20
25
Avg_Num_Products_Purchased



Alteryx Designer x64 - Updated_Project2.yxmd - Browse (8)

12 records displayed, 2 fields, , 154 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 10

Record | Report

Report for Linear Model Linear_Regression_7

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

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

A Linear Regression model is said to be reliable/strong if the 'Adjusted R squared value' is greater than 0.7, as we can see (in the picture above) our model has 'Adjusted R squared value' of 0.8366.
For the predictor variables, we can see that the 'p' value for all of them is less than 0.05 which indicates that there exists some relationship between the predicted variables and the predicting variable.
Hence, our model is a good model.

The best linear regression equation based on the available data is:

Avg_Sale_Amount = **303.46** – [**149.36***(Customer_SegmentLoyalty Club Only)] + [**281.84** *(Customer_SegmentLoyalty Club and Credit Card)] – [**245.42***(Customer_SegmentStore Mailing List)] + [**0***(Customer_SegmentCredit Card Only) + [**66.98***(Avg_Num_Products_Purchased)]

Step 3: Presentation/Visualization

Yes, the company should send the catalogue to these new 250 customers. The predicted profit exceeds $10,000 which was the condition put by the company on the future sales to these new potential customers. In fact, the predicted profit is more than double of the conditional amount ($21,988).

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

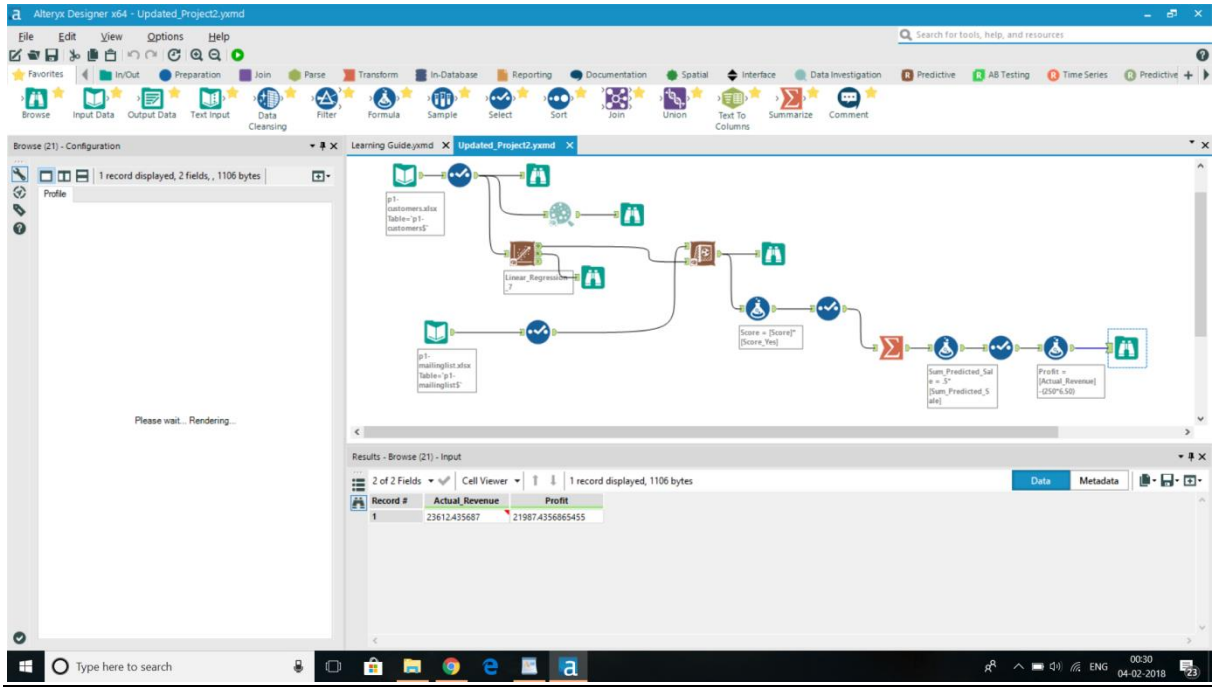
I was able to find solution in few steps:

1) I plotted scatter plots with y-variable as ‘Average Sales Amount’ and x-variable as different numeric variables and identified the best suitable predictor variable (which had direct linear relationship). For string variables, I had to use Linear Regression Model and try different combinations to identify which combination of variables provided better ‘Adjusted R-squared value’ and have ‘p’ value less than 0.05. Using this process, I identified ‘Customer Segment’ and ‘Average Number of Products Purchased’ as strong and reliable predictor variables.

2) Using these variables, I made the best possible Linear Regression Model. I scored the new 250 customer’s data to predict the average sales amount.

3) After getting the predicted average sales amount, I multiplied it with the probability that the customer will buy the product and then performed summation over the entire field to find out the predicted average revenue.

4) Multiplied it with gross margin (50%) and after that subtracted cost of sending catalogue to these new 250 customers ($6.50*250) to get Profit.



Expected Profit from the new catalogue is **$21,988**.