Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

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

Key Decisions:

Answer these questions

1. What decisions needs to be made?

Predicting the expected profit if the catalog was sent to new customers and then, on the basis of profit, determining whether or not the catalog should be sent to new customers.

2. What data is needed to inform those decisions?

Following Data is required

- Customer Type with the number of products purchased and its average sale in the last year.
- Probability of the new customer buying
- New customers type and their shopping habbit
- Profit Margin (Given 50%)
- Cost for catalog for calculation of profit

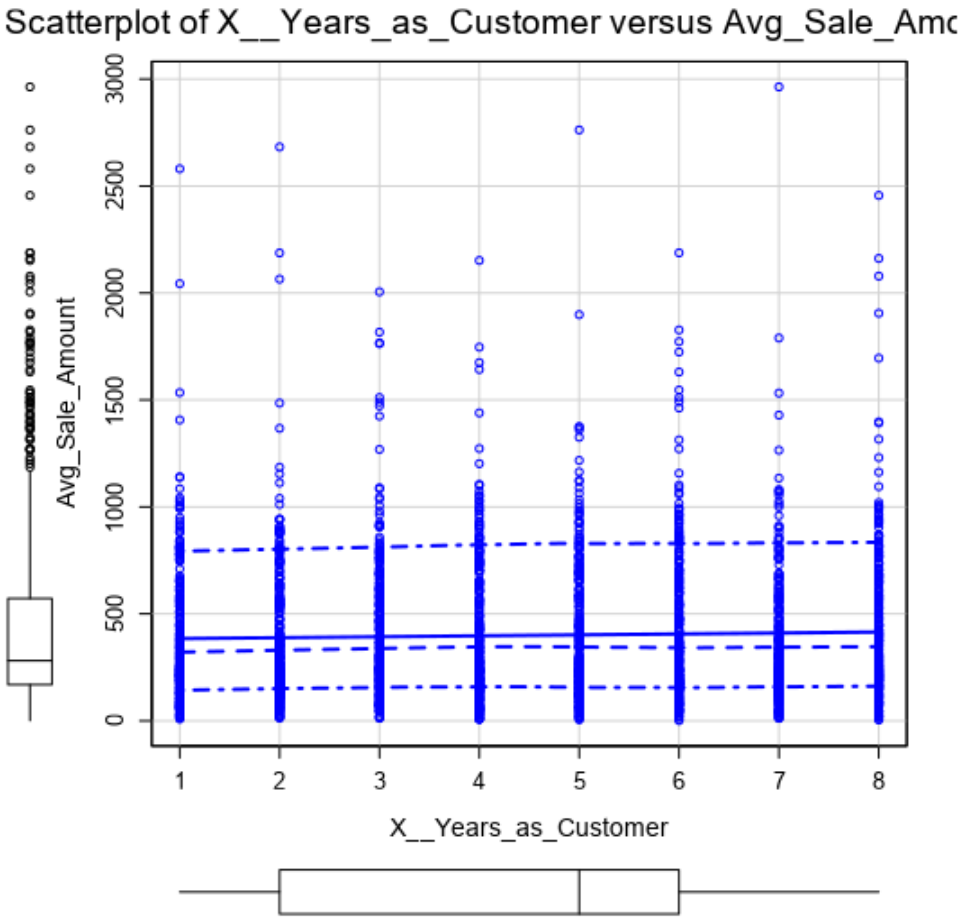
Step 2: Analysis, Modeling, and Validation

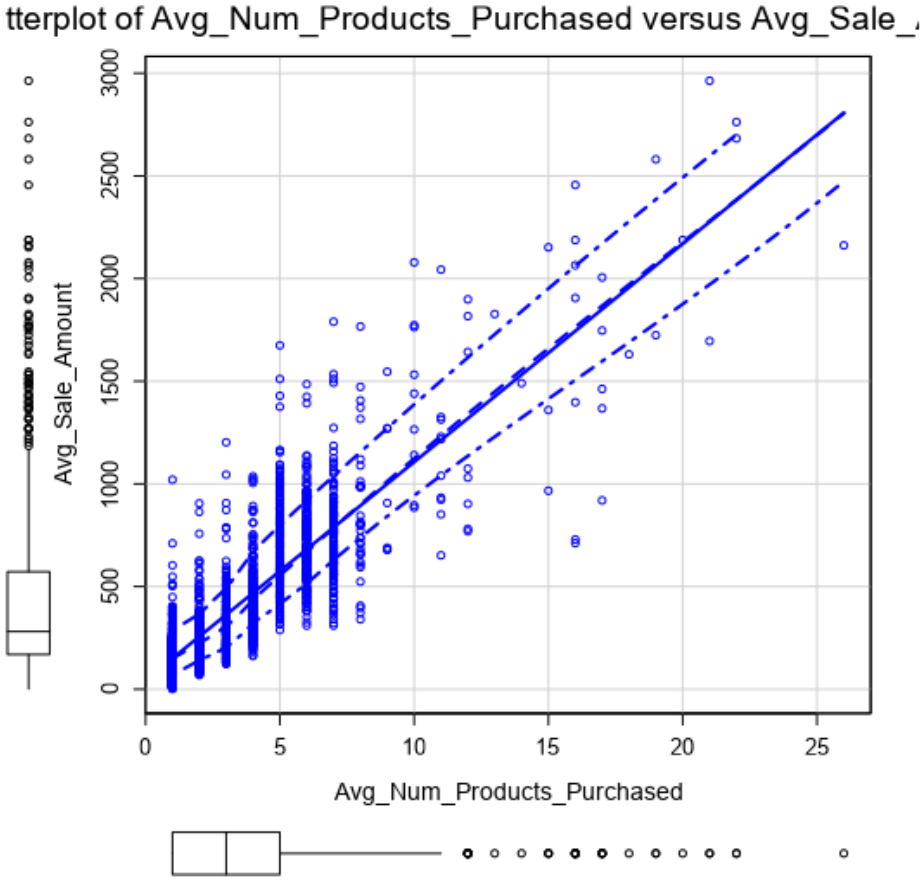
Important: Use the p1-customers.xlsx to train your linear model.*:*

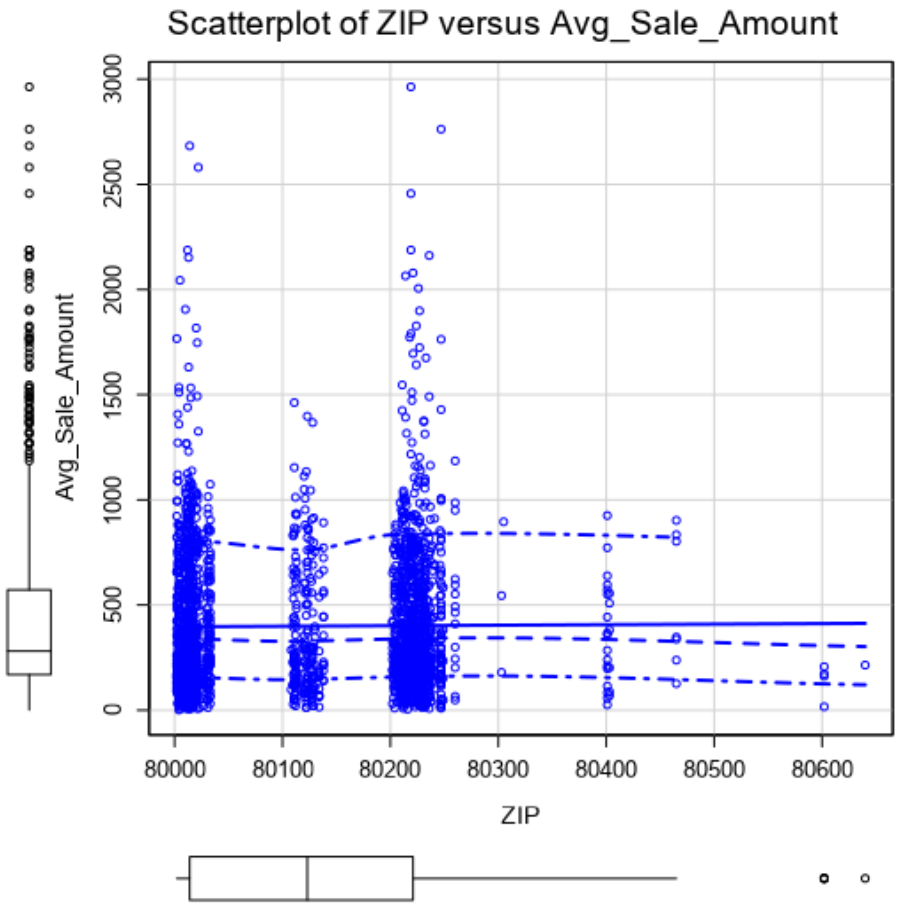
1. How and why did you select the predictor variables in your model?

A linear regression analysis is conducted on all Average Sale Amount variables.
Out of all the linear variables, only Avg_Number_Of_Products_Sold was linearly related with target variable. Then tried “Avg_Number_Of_Products_Sold”, “Customer_Segment” and “City” to build linear model. However, as per the model created “City” was not significant therefore removed city and again created a model using only two. Out of which only the Average Product Number and the Consumer Segment have a p-value of less than 0.05, which means statistical significance. Scatterplots of Average Product Number and Consumer Segment versus Average Selling Sum are also plotted to test linearity.
Analyzed the scatterplot between numerical predictors and target variable.

1) Avg Sales – Avg Number of Products Purchased (liner)
2) Avg Sales – Customers loyalty (Not linear)
3) Avg Sales – ZIP (Not linear)


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


tterplot 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



Didn't tested on the following because:
Name: Sales don't depend on the name of the customer.
Customer Id: It is unique ID assigned to the customer again doesn't affect sales of avg.
Address: It's too specific, instead used other variables like city or zip however they didn't had any relation too.

2. Explain why you believe your linear model is a good model.

Report

Report for Linear Model Linear_Regression_5

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

The linear regression function of Alteryx is used to determine the strength of the linear regression and the statistical result shows an adjusted R-square value of 0.8366, which is a high value. Consumer segment and average number of products have a p-value less than 0.05, which indicates statistical significance. Therefore, the model is considered to be a strong one.

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)

Avg Sale Amount = 303.46 – 149.36 x (If Type: Loyalty Club Only) + 281.84 x (If Type: Loyalty Club and Credit Card) – 245.42 x (If Type: Store Mailing List) + 0 x (If Type: Credit Card Only) + 66.98 x (Avg Num Products Purchased)

Step 3: Presentation/Visualization

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

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

Yes, the company should send the catalog to the new 250 customers. The condition of income $10,000 is being fulfilled.

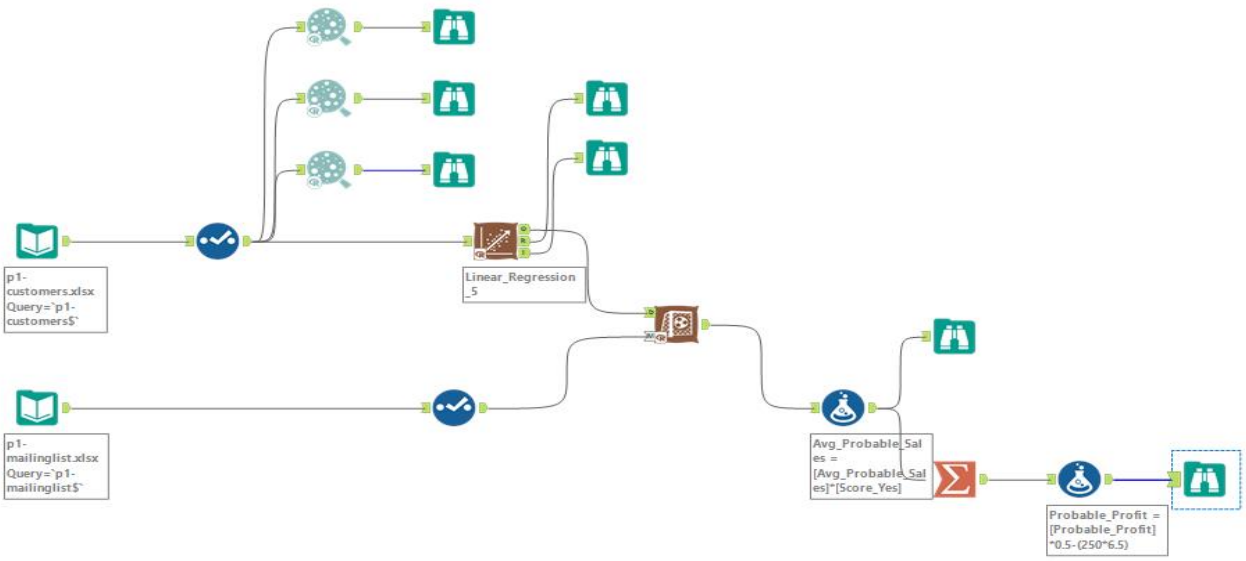
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

Using a linear regression model, the estimated revenue from each customer is calculated by multiplying the actual sales sum with the Probability of the new customer buying(score_yes)
Then multiplying with the profit margin of 50%, the expense of the product ($6.50) with the number of new customers is subtracted for net income.
Avg_Probable_Sales = Avg_Sales * Score_Yes
Profit = (Avg_Probable_Sales * 0.5) – (250*6.5)

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

Expected Profit: 21987.435686545

Alteryx Workflow



Bipin Kumar Sultania