Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

We need to decide whether if sending new year catalogs to our new 250 customers worth printing costs or not. In our case, our decision of approval depending on the expected profit resulting from sending the catalogs. In particular, if the expected profit exceeds $10,000, we recommend catalogs printing, otherwise, we don’t recommend that.

2. What data is needed to inform those decisions?

We need to use both current and new customers data. We could construct a model using current customers data in which, the model represents the average sales per customer (the target variable) as a function of some joint variables between the two datasets. Consequently, we should apply the generated model on new 250 customers data in order to predict the new customers’ average sales.

Step 2: Analysis, Modeling, and Validation

We have two datasets, one for current customers data stored in ‘p1-customers.xlsx’ file and another for new customers data stored in ‘p1-mailinglist.xlsx’ file.

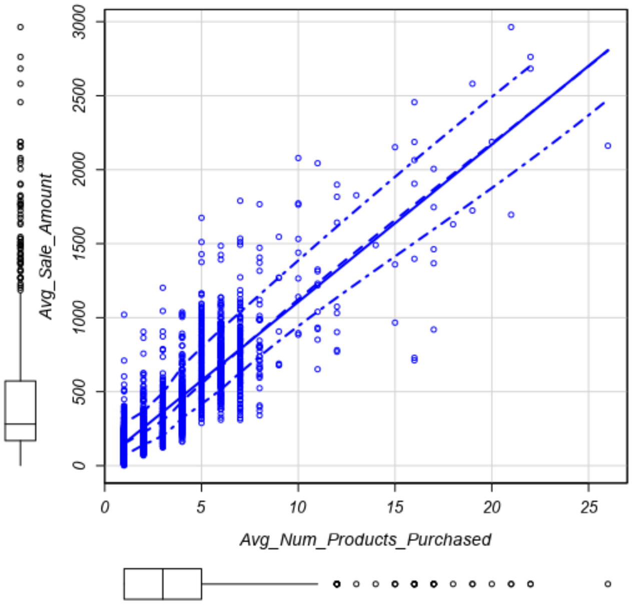
Variables for the two datasets

	‘p1-customers.xlsx’	*‘p1-mailinglist.xlsx’*
Joint Variables	Name	Name
	Customer_Segment	Customer_Segment
	Customer_ID	Customer_ID
	Address	Address
	City	City
	State	State
	ZIP	ZIP
	Store_Number	Store_Number
	Avg_Num_Products_Purchased	Avg_Num_Products_Purchased
	#_Years_as_Customer	#_Years_as_Customer
	Avg_Sale_Amount	Score_No
	Responded_to_Last_Catalog	Score_Yes

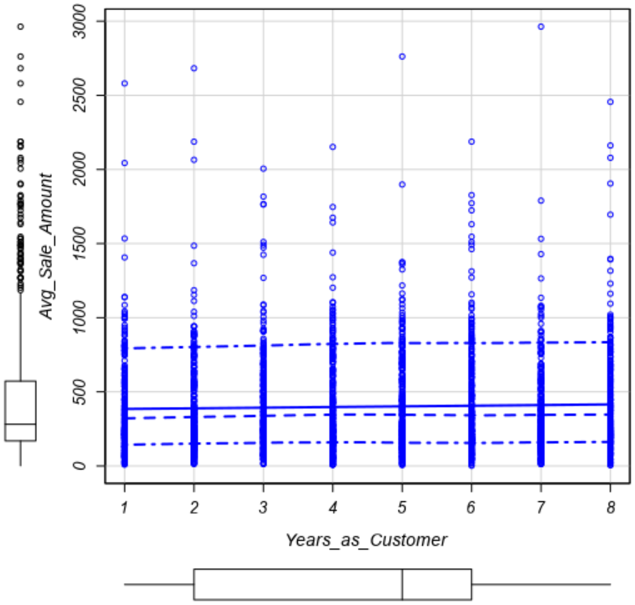
1. How and why did we select the predictor variables in our model? explaining how our chosen continuous predictor variables have a linear relationship with the target variable using scatterplots.

Since we aim to predict the Avg_Sale_Amount variable, I used scatterplots between the joint quantitative variables and the target variable to see if a variable might be a good candidate for predictor variable. The following plots was made in Alteryx using Scatterplot tool. The target *variable (Avg_Sale_Amount) has put as my Y and the numeric predictor variable as X (Avg_Num_Products_Purchased, Years_as_Customer).*

Scatter Plot of Avg_Sale_Amount vs. Avg_Num_Products_Purchased

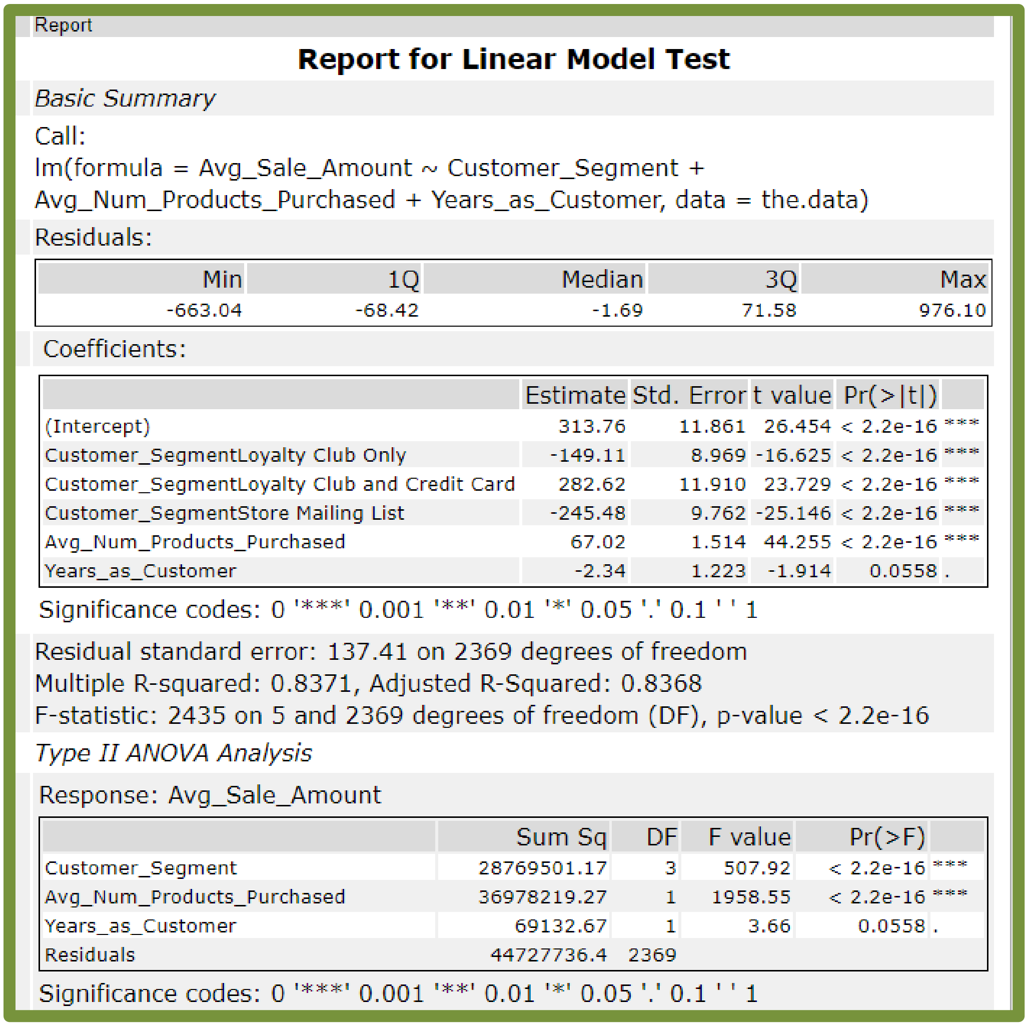


Scatter Plot of Avg_Sale_Amount vs. Years_as_Customer



It is clear that both Avg_Sale_Amount and Avg_Num_Products_Purchased have a strong positive correlation while both Avg_Sale_Amount and Years_as_Customer don't have a linear correlation. Thus I think that Avg_Num_Products_Purchased is a good predictor for our model but Years_as_Customer is a bad one.

2. Why we believe our linear model is a good model using the statistical results that our regression model created. For each selected variable, we justify how each variable is a good fit for our model by using the p-values and R-squared values that our model produced.



Report

Report for Linear Model Test

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment +
Avg_Num_Products_Purchased + Years_as_Customer, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-663.04	-68.42	-1.69	71.58	976.10

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	313.76	11.861	26.454	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-149.11	8.969	-16.625	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.62	11.910	23.729	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.48	9.762	-25.146	< 2.2e-16	***
Avg_Num_Products_Purchased	67.02	1.514	44.255	< 2.2e-16	***
Years_as_Customer	-2.34	1.223	-1.914	0.0558	.

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

Residual standard error: 137.41 on 2369 degrees of freedom
Multiple R-squared: 0.8371, Adjusted R-Squared: 0.8368
F-statistic: 2435 on 5 and 2369 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	28769501.17	3	507.92	< 2.2e-16	***
Avg_Num_Products_Purchased	36978219.27	1	1958.55	< 2.2e-16	***
Years_as_Customer	69132.67	1	3.66	0.0558	.
Residuals	44727736.4	2369			

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

Notes:

We notice the following:

- *Adjusted R-Squared equal 0.84 which means that the model explains 84% of the fitted data.*
- *Each of Customer_Segment and Avg_Num_Products_Purchased has P-value less than 0.05 which means that both of them are significantly affecting Avg_Sale_Amount.*
- *Years_as_Customer has P-value greater than 0.05 which means that it is not significantly affecting Avg_Sale_Amount.*

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

The best regression equation should be in the form:

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

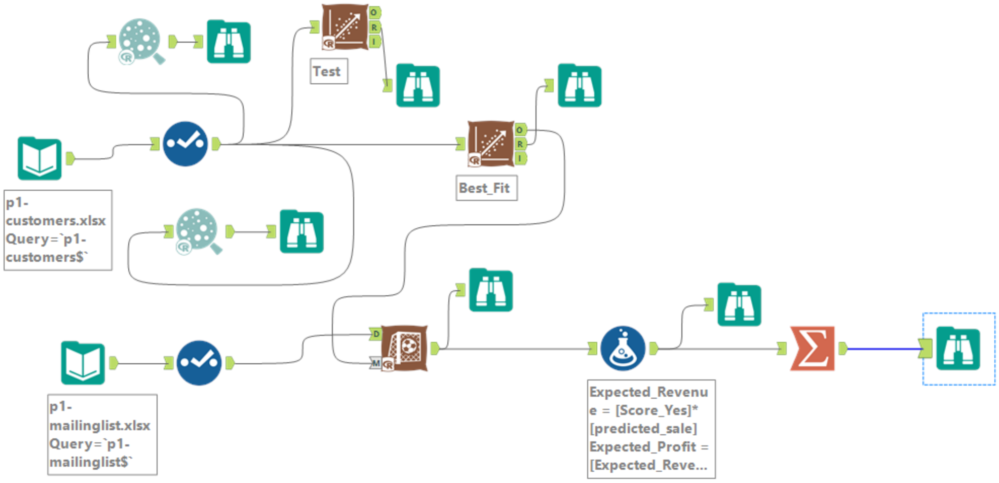
Step 3: Presentation/Visualization

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

It is recommended to send the catalog to the new 250 customers.

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

I have applied the best fit model on the new customers dataset to find scores of predicted_sales per customer. After that I have multiplied the predicted_sales by the Score_yes probability to obtain the Expected_Revenue per customer. Then I have multiplied the Expected_Revenue by 0.5 and subtract $6.5 from each Expected_Revenue to obtain the Expected_Profit per customer. Finally, I have obtained the sum of all Expected_profit column. All of these steps are done using the following Alteryx flowchart:



O
R
I
Test
O
R
I
Best_Fit
p1-customers.xlsx Query=`p1-customers$`
D
M
p1-mailinglist.xlsx Query=`p1-mailinglist$`
Expected_Revenue = [Score_Yes]* [predicted_sale] Expected_Profit = [Expected_Reve...

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

The Expected_profit form all 250 customers are equal to $21,987.44 which exceeds our limit of $10,000.