Alhanoof Alyabes
15/3 /2018

Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

1. What decisions needs to be made?

Make decision about wither the 250 new customers will purchased from the catalog based on previous sales in order to get expected profit

2. What data is needed to inform those decisions?
The data will be use like Predicted variables are

- customer segments
- Avg_Num_Products_Purchased

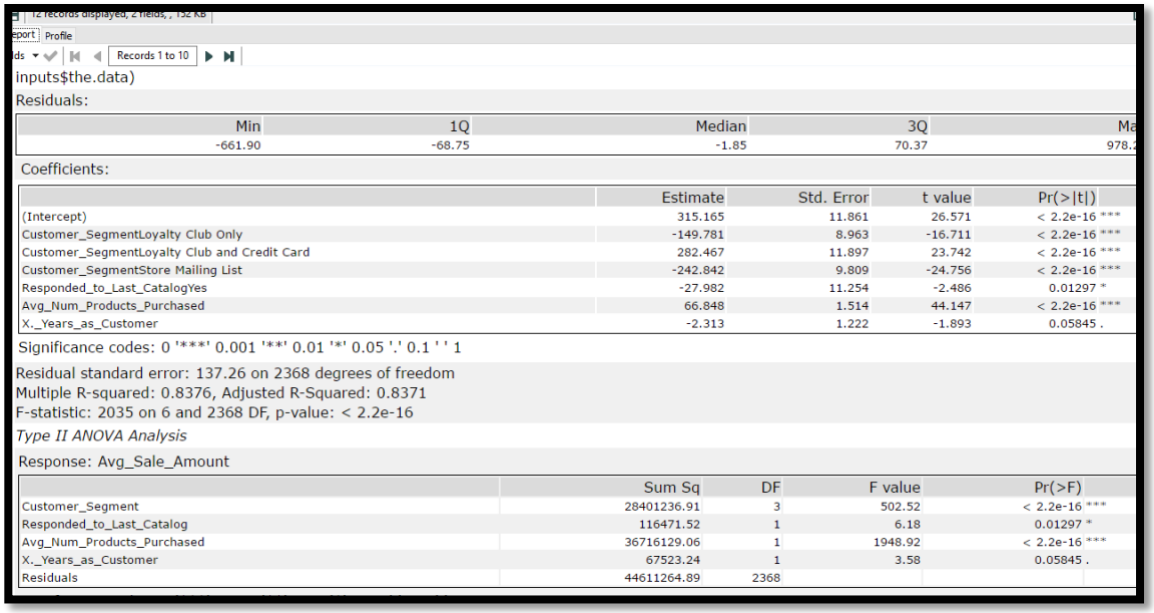
and the data from

- score_yes, and
- margin –cost of catalogs

Step 2: Analysis, Modeling, and Validation

1. How and why did you select the predictor variables (see supplementary text) in your model?

- By using linear regression in altreyx , we determined the predictor variables , as it shown below , the customer segment and average product purchased with p-value below .05 and are statistically significant ,while the other are not.
- I did not select the other variables like state, zip ..etc, because either it is unique value and have no impact on new customer purchase or not statistically significant



inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Ma
-661.90	-68.75	-1.85	70.37	978.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	315.165	11.861	26.571	< 2.2e-16 ***
Customer_SegmentLoyalty Club Only	-149.781	8.963	-16.711	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.467	11.897	23.742	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-242.842	9.809	-24.756	< 2.2e-16 ***
Responded_to_Last_CatalogYes	-27.982	11.254	-2.486	0.01297 *
Avg_Num_Products_Purchased	66.848	1.514	44.147	< 2.2e-16 ***
X._Years_as_Customer	-2.313	1.222	-1.893	0.05845 .

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

Residual standard error: 137.26 on 2368 degrees of freedom
Multiple R-squared: 0.8376, Adjusted R-Squared: 0.8371
F-statistic: 2035 on 6 and 2368 DF, p-value: < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28401236.91	3	502.52	< 2.2e-16 ***
Responded_to_Last_Catalog	116471.52	1	6.18	0.01297 *
Avg_Num_Products_Purchased	36716129.06	1	1948.92	< 2.2e-16 ***
X._Years_as_Customer	67523.24	1	3.58	0.05845 .
Residuals	44611264.89	2368		

figure (1) : linear regression report of different variables

By sing Scatterplot Tool, the average num product purchased are good as predicted variables .slop line indicate that the predict and target variables are related

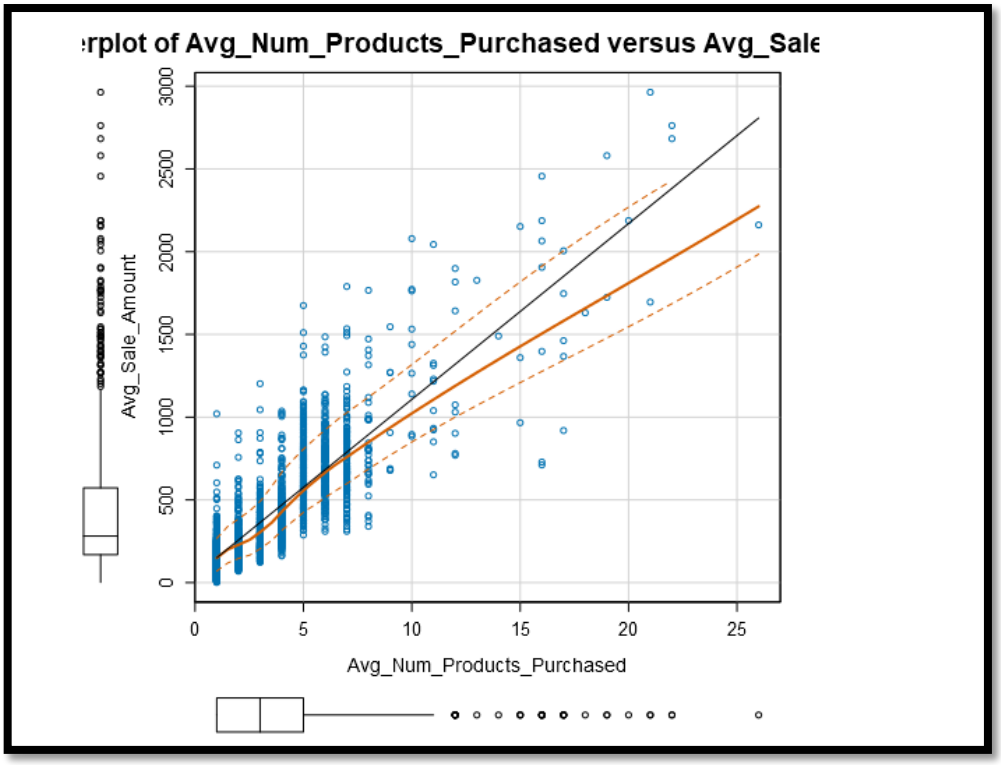
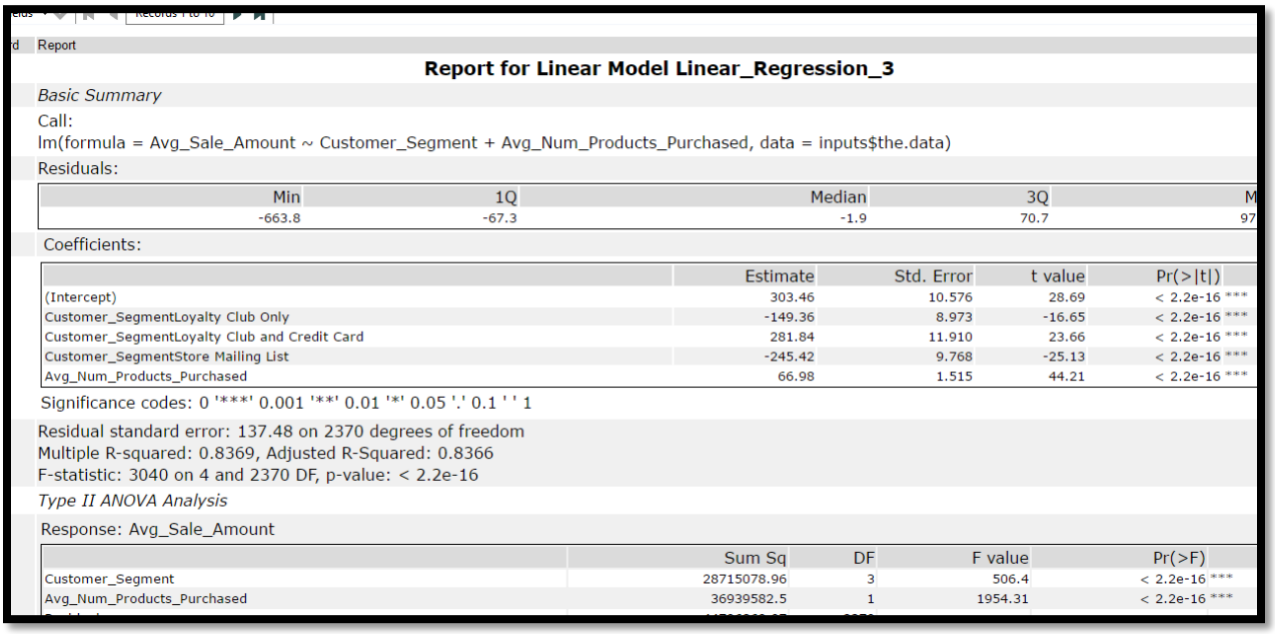


figure (2) : scatterplot of numeric variables (average purchased and average sale)

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.

By using linear regression model the result show that R-squared .8376 which is good , and the p-value of customer segments and average products purchased lower than .05 significant .so the model is good model



Report

Report for Linear Model Linear_Regression_3

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	M
-663.8	-67.3	-1.9	70.7	97

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	***

figure (3) : report for liner regression of predicted variables

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)

Average of sale amount = 303.46 -149.36*(if customer_segmentL: oyalty Club only)+ 281.84* (if customer_segment :loyalty Club and credit card) – 25.42 (if customer_ segment :Store mailing list + 0 *(if customer segment: credit card only)+ 66.96 *Avg_Num_Products_Purchased

Step 3: Presentation/Visualization

1. What is your recommendation? Should the company send the catalog to these 250 customers?
Yes, the company can send the catalog for these 250 customers, total predicted profit more than $10000 .

2. How did you come up with your recommendation?

After using linear regression, the expected revenue can get it by multiplying average sale with score_yes . Then, the margin 50% are used before sub the $6.50 to get the profit.

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

Expected profit = (total expected revenue * margin) - (cost *250 customers)

=(47224.87* .5)- (6.50*250)
=$ 21987.43

Distributions for each variable in the Customer List dataset

- By using linear regression in Alteryx , we determined the predictor variables , as it shown below , the customer segment and average product purchased with p-value below .05 and are statistically significant ,while the other are not.
- Values like city , #years of customer and respond to last catalog are not statistically significant as shown on linear regression model
- I did not select the other variables like state, zip , name ,customer ID ,address and store number , because it is unique value and have no impact on new customer purchase .

However, there is need for more fields or data category to predict precisely of customer behaviors.
I used field summary histogram with visual indicator of data quality dashboard to provide visual and distributed profile of each variables in the customer dataset

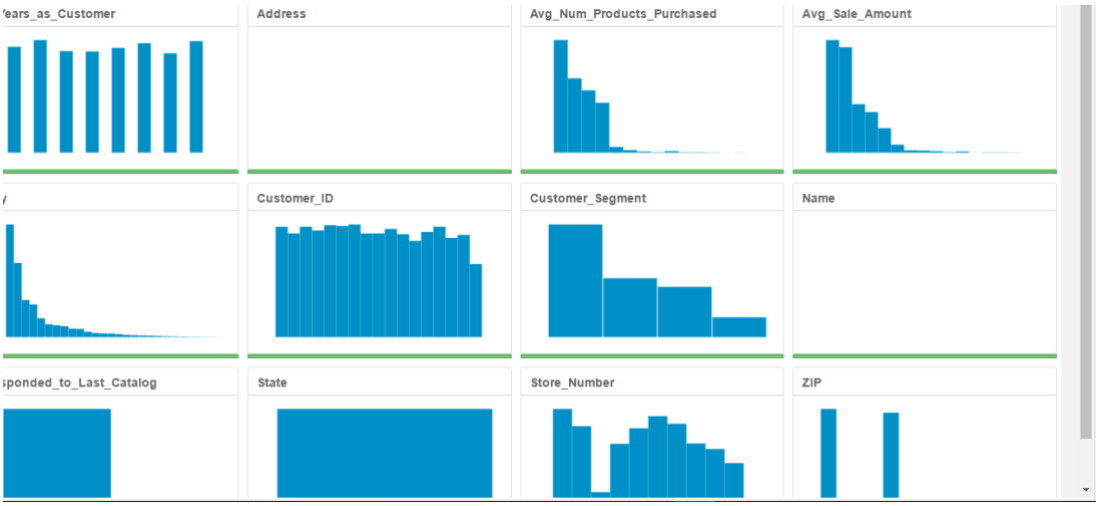


figure (4): distribution for each variable

Alteryx workflow

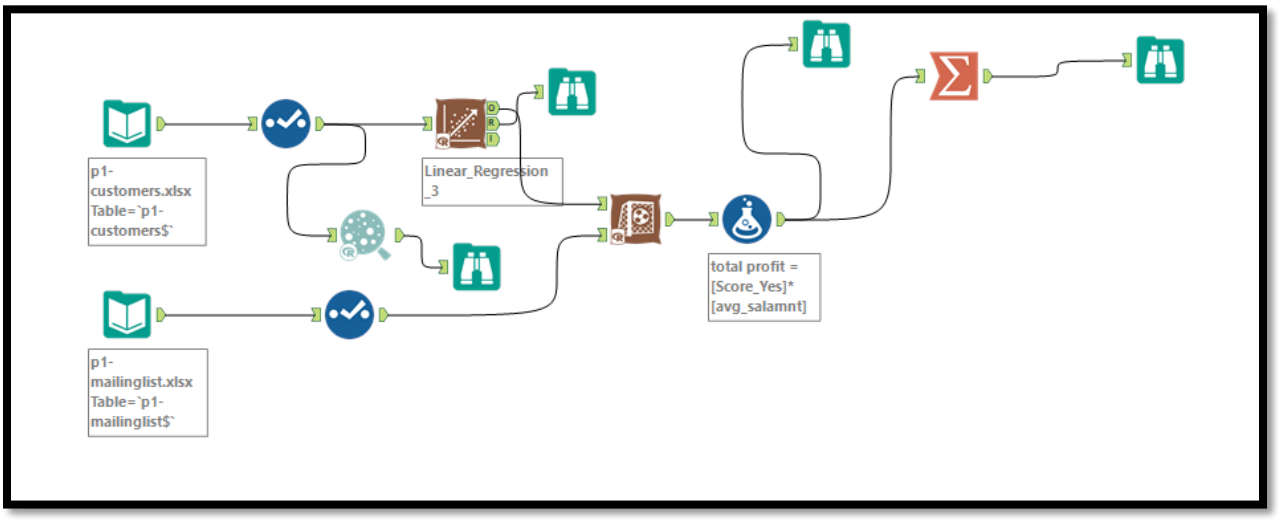


figure (5) : workflow of the entire process