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?
   - A decision needs to be made to determine whether or not to send the catalog to the 250 new customers. Catalogs will only be sent if the expected profit exceeds $10,000.

2. What data is needed to inform those decisions?
   - In order to make these informed decisions, an analyst would need data such as:
     - Customer segment
     - Avg_num_Products_purchased
     - Score_yes
     - Catalog price ($6.50)
     - Avg gross margin

Step 2: Analysis, Modeling, and Validation

Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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

At the minimum, answer these questions:

1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

- Since we are trying to predict the expected profit, the target variable is in linear regression is (avg_sale_amount) and the predictor variables are (customer_segment and avg_num_products_purchased). The reason these two are the predictor variables is because their p-value has the stars (*) that indicate statistical significance. "Statistical significance is a result that is not likely to occur randomly, but rather is likely to be attributable to a specific cause." (Investopedia). From the picture below, both the customer_segment and avg_num_products_purchased are significant.

Alteryx Designer x64 - Browse (6)

12 records displayed, 2 fields, , 158 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 10

Record | Report

Report for Linear Model Linear_Regression

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Store_Number + Responded_to_Last_Catalog + Avg_Num_Products_Purchased + X._Years_as_Customer, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-665.19	-67.82	-2.17	70.42	975.25

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	435.318	104.854	4.152	3e-05 ***
Customer_SegmentLoyalty Club Only	-150.224	8.971	-16.746	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.455	11.897	23.743	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-243.279	9.816	-24.784	< 2.2e-16 ***
Store_Number	-1.146	0.994	-1.153	0.2489
Responded_to_Last_CatalogYes	-28.085	11.253	-2.496	0.01264 *
Avg_Num_Products_Purchased	66.787	1.515	44.082	< 2.2e-16 ***
X._Years_as_Customer	-2.326	1.222	-1.904	0.05707 .

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

Residual standard error: 137.25 on 2367 degrees of freedom
Multiple R-squared: 0.8376, Adjusted R-Squared: 0.8372
F-statistic: 1745 on 7 and 2367 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	28425698.84	3	503.02	< 2.2e-16 ***
Store_Number	25054.95	1	1.33	0.2489
Responded_to_Last_Catalog	117319.82	1	6.23	0.01264 *
Avg_Num_Products_Purchased	36603783.15	1	1943.23	< 2.2e-16 ***
X._Years_as_Customer	68263.47	1	3.62	0.05707 .
Residuals	44586209.93	2367		

17:16
6/25/2018

- Relationship between avg_sales_amount vs avg_num_products_purchased

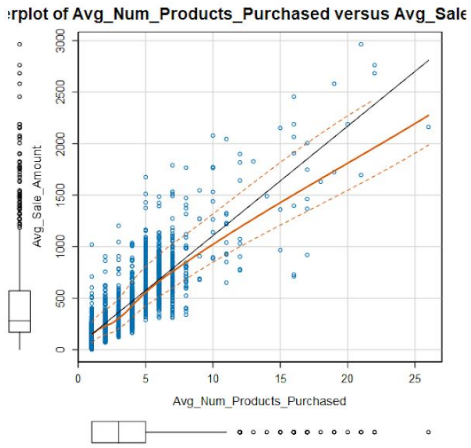
Alteryx Designer x64 - Browse (4)

1 record displayed, 1 field, , 95 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 1

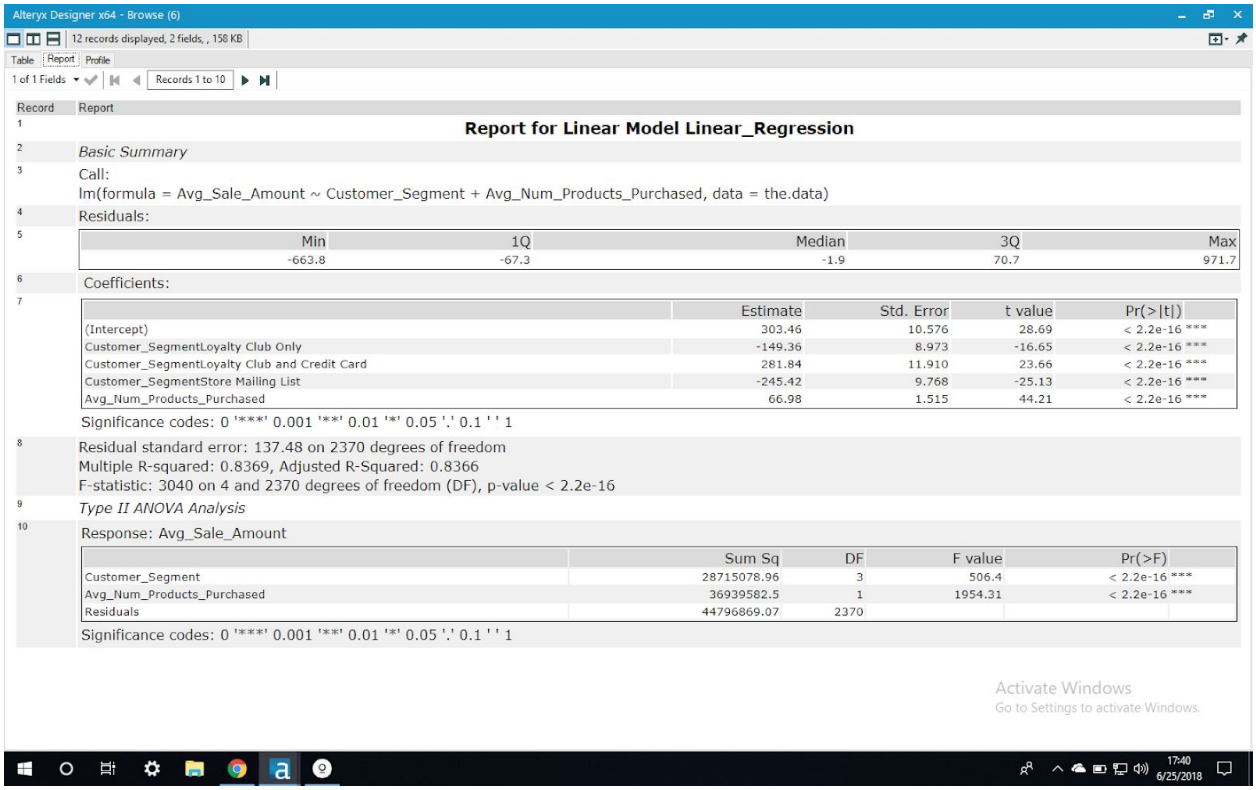
Record | Graph



17:24
6/25/2018

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. For each variable you selected, please justify how each variable is a good fit for your model by using the p-values and R-squared values that your model produced.

- After running the linear regression tool on Alteryx, from the statistical results, the adjusted r-squared value came out as 0.8366. Usually, if the r-squared value is above 0.7, it is considered a good model. In this case, I believe the linear model is good.



Alteryx Designer x64 - Browse (6)

12 records displayed, 2 fields, , 158 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 10

Record	Report
1	**Report for Linear Model Linear_Regression**
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

Activate Windows
Go to Settings to activate Windows.

17:40
6/25/2018

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)

Important: The regression equation should be in the form:

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

- Avg_sale_amount = 303.46 - 149.36 * (if type: loyalty club only) + 281.84 * (If type: loyalty club and credit card) - 245.42 * (if type: mailing list) + 66.98 * (if type: avg_num_products_purchased)

Step 3: Presentation/Visualization

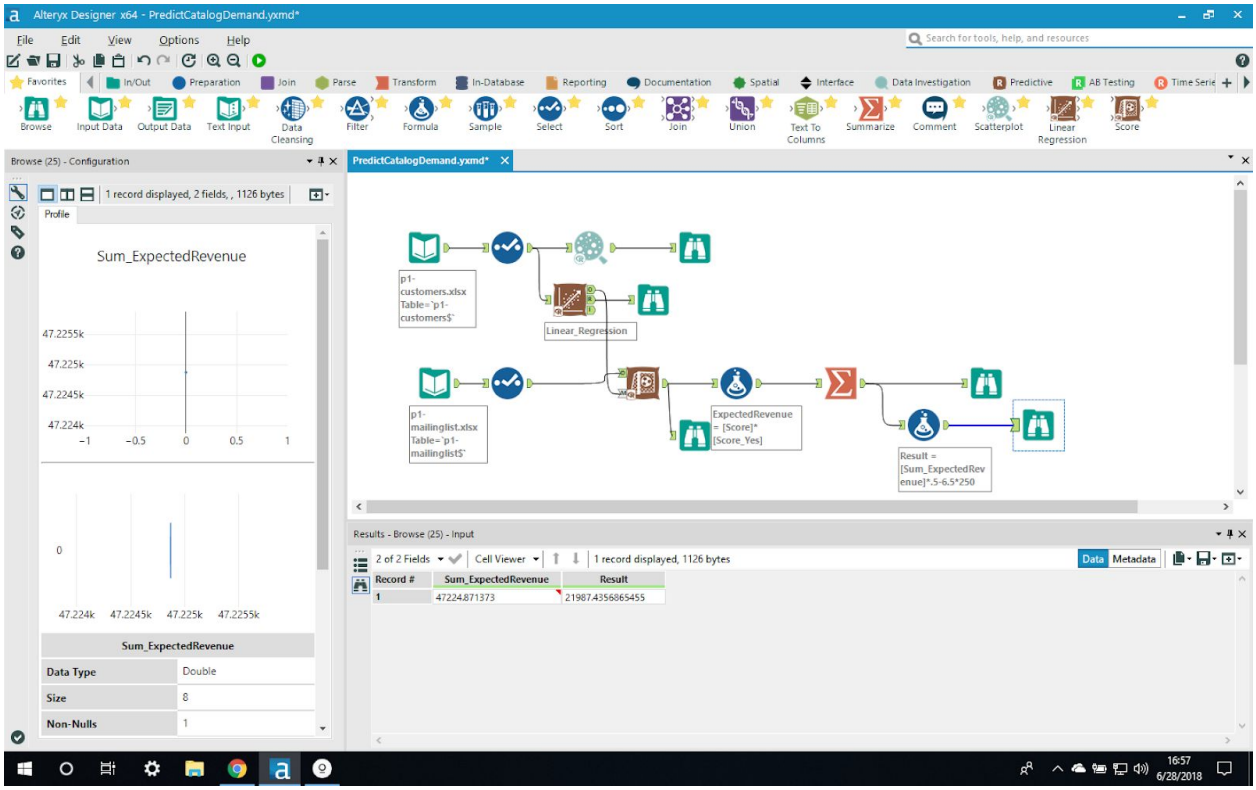
Use your model results to provide a recommendation. (500 word limit)

At the minimum, answer these questions:

1. What is your recommendation? Should the company send the catalog to these 250 customers?
   - Yes, the company should send the catalog to the 250 new customers.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
   - The first step is to multiply [Score]*[Score_Yes]. Then, get the summarize tool and obtain the total sum for the ExpectedRevenue. In my case, ExpectedRevenue = [Score]*[Score_Yes].
   - The second step is to get the gross profit, which is 50 % or .5 and multiply with ExpectedRevenue, and to this result, we have to subtract the price of the catolog ($6.50) times 250 new customers.

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
   - ExpectedRevenue * 0.5 - 6.5 * 250
   - 47,224.871373 * 0.5 - 6.5 * 250 = 21,987.4356865455


Alteryx Designer x64 - PredictCatalogDemand.yxmd*
File Edit View Options Help
Search for tools, help, and resources
Favorites In/Out Preparation Join Parse Transform In-Database Reporting Documentation Spatial Interface Data Investigation Predictive AB Testing Time Serie
Browse Input Data Output Data Text Input Data Cleansing Filter Formula Sample Select Sort Join Union Text To Columns Summarize Comment Scatterplot Linear Regression Score
Browse (25) - Configuration
1 record displayed, 2 fields, , 1126 bytes
Profile
Sum_ExpectedRevenue
47.2255k
47.225k
47.2245k
47.224k
−1 −0.5 0 0.5 1
0
47.224k 47.2245k 47.225k 47.2255k
Sum_ExpectedRevenue
Data Type Double
Size 8
Non-Nulls 1
PredictCatalogDemand.yxmd*
p1-customers.xlsx Table='p1-customers$'
Linear_Regression
p1-mailinglist.xlsx Table='p1-mailinglist$'
ExpectedRevenue = [Score]* [Score_Yes]
Result = [Sum_ExpectedRevenue]*.5-6.5*250
Results - Browse (25) - Input
2 of 2 Fields Cell Viewer 1 record displayed, 1126 bytes
Data Metadata
Record # Sum_ExpectedRevenue Result
1 47224.871373 21987.4356865455
16:57
6/28/2018