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?

Answer. The main decision to be made is to send out catalogues to 250 new customers. But the company doesn't want to send the catalogues if the expected profit is less than $10,000.

2. What data is needed to inform those decisions?

Answer. We need the Customer Segment, Average Sale Amount, Average number of products purchased, cost of each catalogue ($6.50), gross margin (50%), Response to the last catelogue, the total amount a customer spent ordering from the catalogue and Score_Yes data to predict the sales.

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.

Answer. We use Linear Regression to calculate the Average Sales Amount. As we can see the p-value for Customer Segment and Average Number of Products is less than 0.05, which means they are of most significance. The scatterplot graphs for the same are shown below-

12 records displayed, 2 fields, , 152 KB

Report

1 of 1 Fields | Records 1 to 10

Record | Report

1

Report for Linear Model Predicting_Catalog_Demand

2 *Basic Summary*

3 Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Store_Number + Responded_to_Last_Catalog + Avg_Num_Products_Purchased + no_of_years_as_a_customer, data = the.data)

4 Residuals:

5

Min	1Q	Median	3Q	Max
-665.20	-67.82	-2.17	70.42	975.30

6 Coefficients:

7

	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	***
no_of_years_as_a_customer	-2.326	1.222	-1.904	0.05707	.

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

8 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 DF, p-value: < 2.2e-16

9 *Type II ANOVA Analysis*

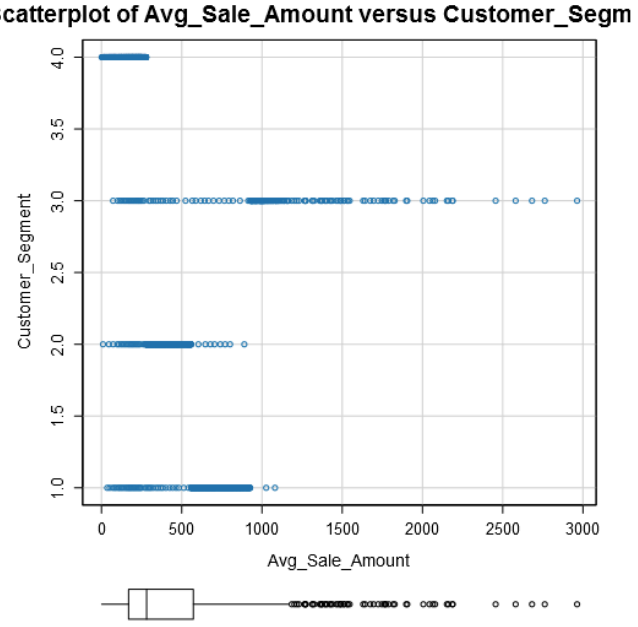
1 record displayed, 1 field, , 40 KB

Report

1 of 1 Fields | Records 1 to 1

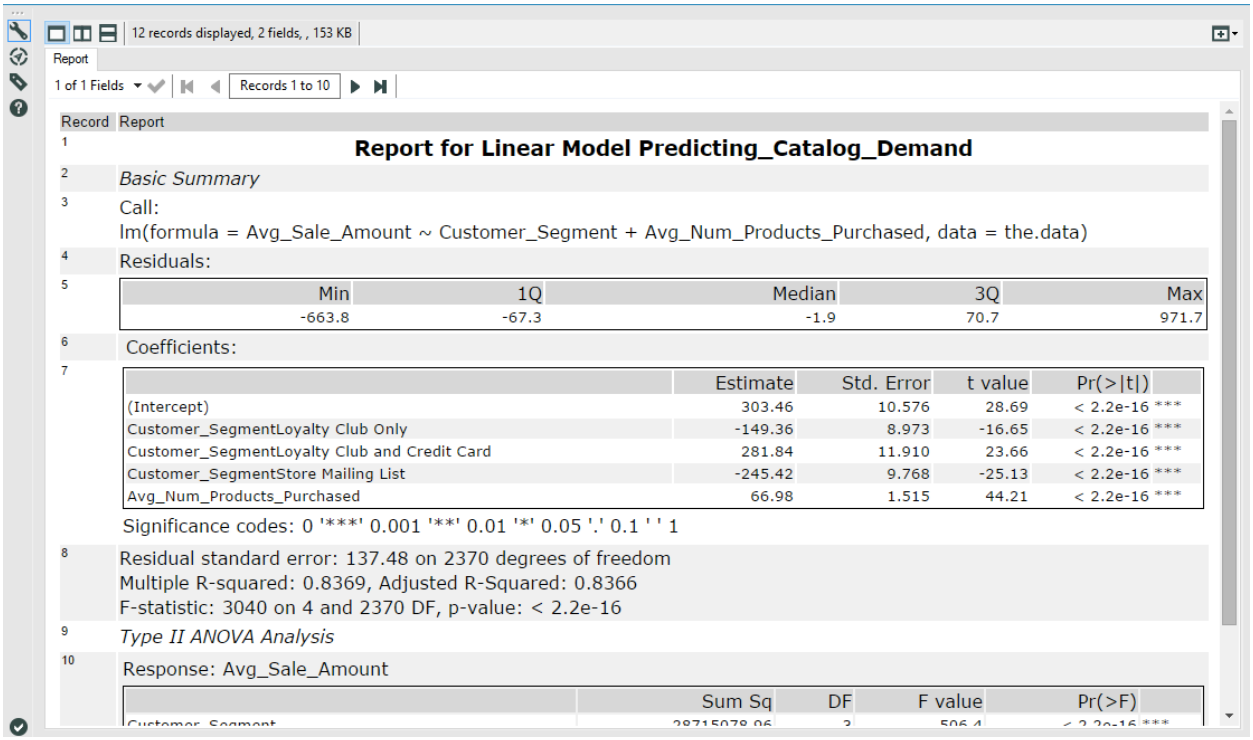
Record | Graph

1



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.

Answer. As shown above, the Multiple R-squared value (0.8369) and Adjusted R-squared value (0.8366) are both greater than 0.7 which is considered to be good. Also, the p-value for Customer Segment and Average Number of Products is less than 0.05 which is of high significance. Thus, the model is considered to be a good 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)

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.
Answer.
Avg_Sale_Amount= 303.46-149.36*(If Type=Loyalty Club Only)+ 281.84*(If Type=Loyalty Club and Credit Club)-245.42*(If Type=Store Mailing List)+ 0*(If Type=Credit Card Only) + 66.98 * (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?

Answer. Yes, the company should send the catalogues to 250 customers.

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

Answer. Firstly, we need to calculate the Score of each customer which is given by: [Score] * [Score_Yes]. Now, we take the summation of all the scores which will give us the Total Score i.e [Sum_Score]. Also, its given in the question that the gross margin is 50% of the Total Score which is given by [Sum_Score]*0.5. Finally, we subtract expenses of sending the catalogues to 250 clients i.e we subtract (6.5*250) from [Sum_Score]*0.5. In this way we obtain the expected profit.

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

Answer. Expected Profit= (Sum of total revenue * Gross margin) – (cost of each catalogue * 250) = (47224.87 * 0.5) – (6.50 * 250) = 23612.43 – 1625 = 21987.43 (in dollars)

Workflow Diagram