MAKE A COPY

Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

Answer these questions

1. What decisions needs to be made?

A. Here we have to make decision that to send catalog to 250 Customers, is profitable or not. Additionally the profit must be over $10000 and then we can decide to send catalog or not to customer.

2. What data is needed to inform those decisions?

A. For this project we have to make decision that whether we have send catalogs or not to new customers (250) available in mailinglisr.xlsx. Here company's goal is to predict how much profit they can expect by sending these catalogs. For making such decision we have to understand associate cost in sending catalogs and profitable it turns out.

Here we have already given, the cost of catalogs $6.50, Average gross margin on all products sold is 50%.

We have to use p1.customers.xlsx for building model from which we will get purchase history of existing customers and customer demographics and segment information which will help to understand characteristics drive profits with company's existing customer base and allow to build model to predict based on similar data with new customers given in p1-mailinglist.xlsx that new customers will respond to catalogs and make purchase (Score_Yes).

Step 2: Analysis, Modeling, and Validation

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

A. Linear Regression is performed against Average Sale Amount on all variable. Here we can observe p-value of Customer Segment & Average No. of Product is less than 0.05 which depicts statistical significance. Also Adjusted R Square value and R Square values is nearly same 0.84. R Squared is a measure of how much the data are fitted regression line.

Record Report

1

Report for Linear Model Linear_Regression_2

2 *Basic Summary*

3 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)

4 Residuals:

5

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

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 ***
X._Years_as_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 degrees of freedom (DF), p-value < 2.2e-16

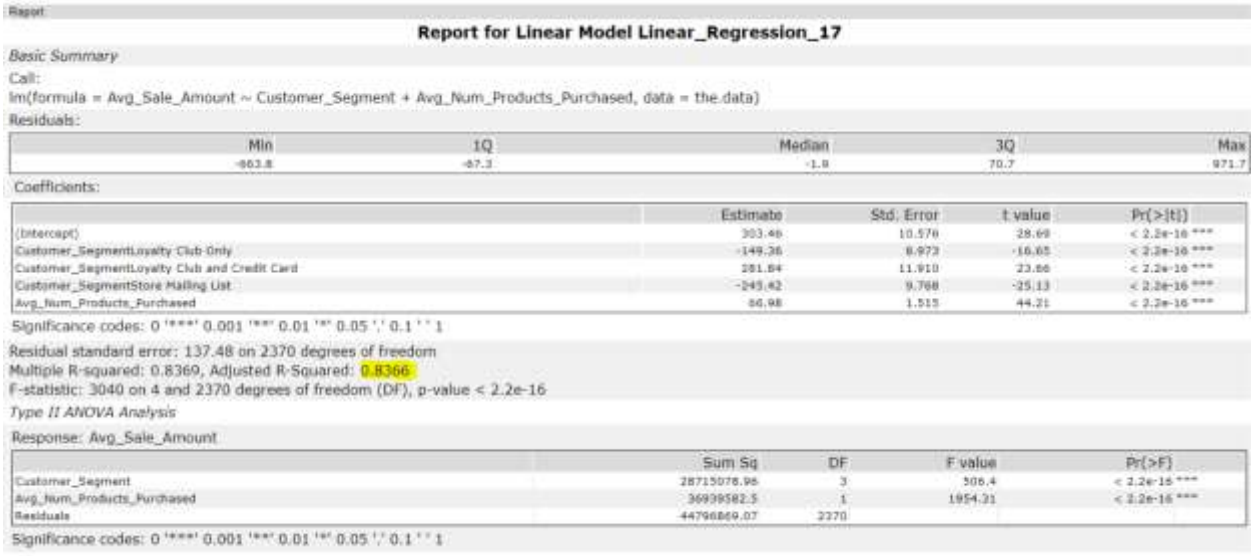
9 *Type II ANOVA Analysis*

Scatterplots of Average no of products purchase versus Average Sales amount shows linearity.

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

A. In linear Regression we calculate and equation that minimize distance between fitted line and all data points. R Squared is percentage of variance in observation that is explained by the model. R Squared lies between 0 and 1. R squared value close to 1 mean that all variance in target variable is explained by model and R Squared value close to 0 means that none of variance in target variable is explained by model. The value of R Squared greater than 0.7 is considered as a strong model. In our model it is 0.8366 which clearly shows that the model is strong enough.

P value decides statistical significance in a hypothesis test. A Hypothesis test is statistical test to find whether we have enough evidence in data sample to anticipate that a condition is true or not for entire population. A hypothesis test observe 2 opposing hypotheses about a population, A Null Hypothesis and An alternative Hypothesis. A Null Hypothesis is a statement of "NO EFFECT". P value is probability that observed results occurred by chance and doesn't have relationship between target variable and predictor. It is probability that the coefficient is zero. P value decides relationship between target variable and predictor. In our results we can clearly observe that for predictor variables (Average number of products purchased and customer segment) that we used in our linear model, p value which is the probability that the coefficient is zero, is less than 0.05. Hence we can say that predictors are significant in deciding target variable.



Report

Report for Linear Model Linear_Regression_17

Basic Summary

Call:

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

Residuals:

Min	1Q	Median	3Q	Max
-863.8	-87.3	-1.8	70.7	971.7

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	303.46	10.570	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	28715076.96	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

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

A. Average Sale Amount = 303.46

+66.98*(Average number of product purchased)

+281.84(If Segment is Loyalty Club & Credit Card)

-149.36*(If Segment is Loyalty Club ONLY)

-245.42*(If Segment is Store Mailing List)

+0*(If Segment is Credit Card Only)

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?

A. Company should send catalogs 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)

A. The expected profit is greater than $10000,Hence it is recommended to send catalogs. Each customer's expected revenue is calculated by multiplying expected sales amount with Score_Yes Value. Gross Margin iof 50%, 50% is deducted from sum of expected revenue before the cost of catalog which is ($6.50) is subtracted to get net profit.

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

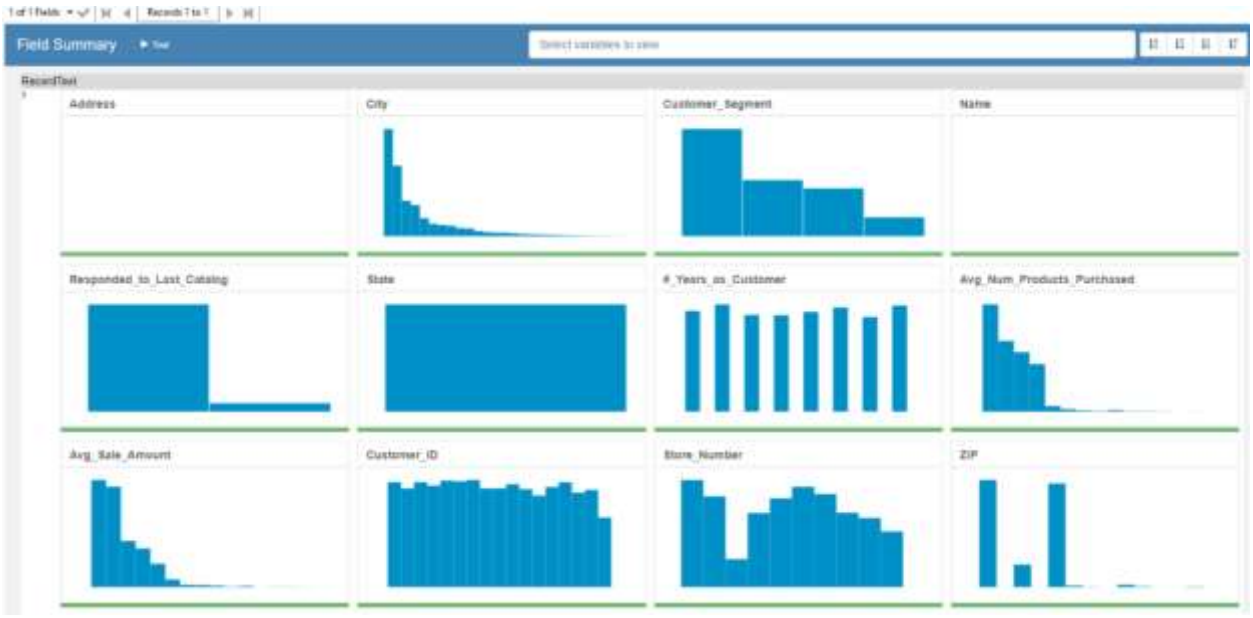
A. Profit Predicated = (Revenue Predicted * Gross Margin) - (Cost of Catalog*No of catalog)

Profit Predicated = (0.5*47224.87) - (6.50*250)
= 23612.435 - 1625
= $21987.44

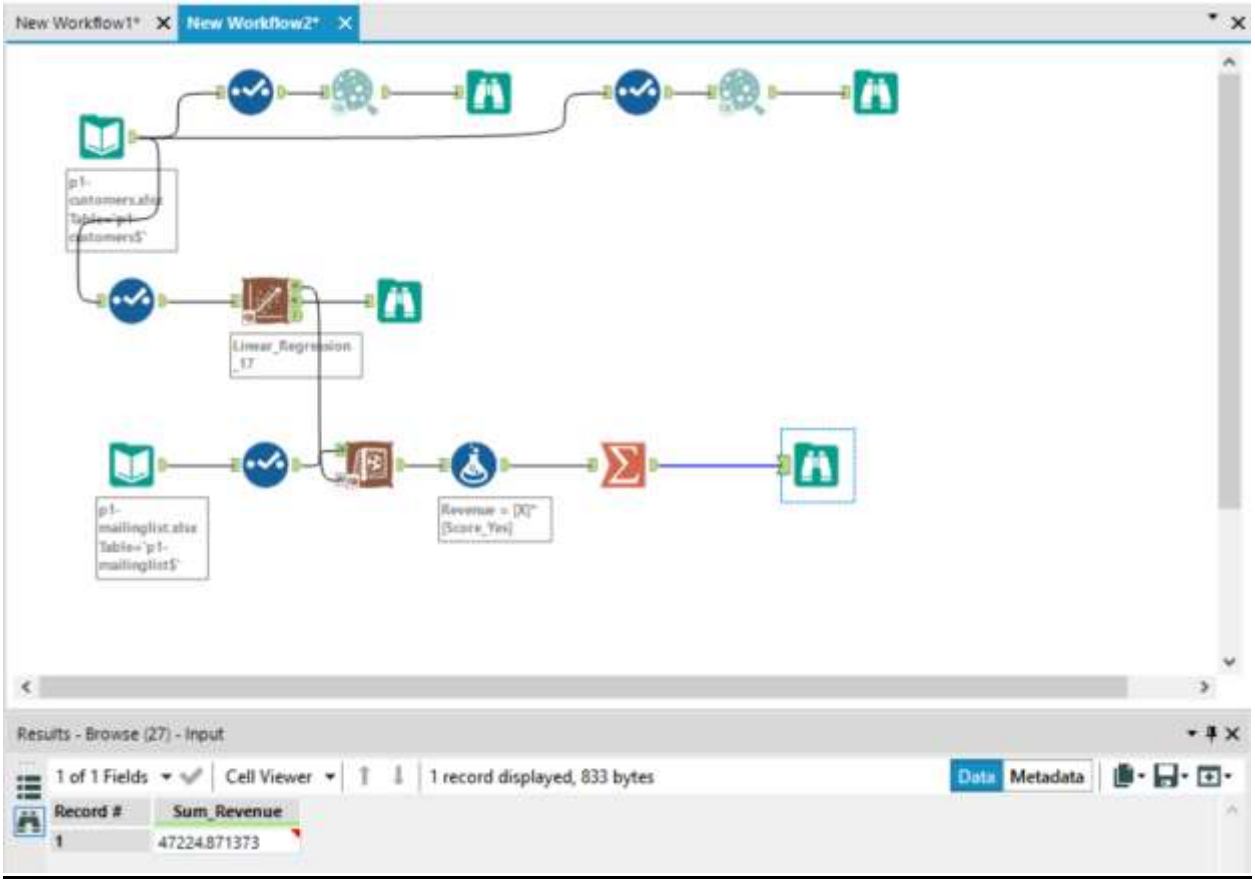
Variable Distribution

Variables like Name, Address, Customer Id, Store Number are not so important for predicting sales.

Data like the items purchased by customers, items turnover duration will be helpful to understand customer behavior and from which we can target to segment our customers and customize catalogs.



Alteryx Workflow



Results - Browse (27) - Input

1 of 1 Fields | Cell Viewer | 1 record displayed, 833 bytes | Data | Metadata

Record #	Sum_Revenue
1	47224.871373