PREDICTING CATALOG DEMAND

IN THIS PROJECT WE ARE GOING TO PREDICT WHETHER WE SHOULD DO A DEAL OR NOT

The Business Problem

I recently started working for a company that manufacturers and sells high-end home goods. Last year the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has **250 new customers** from their mailing list that they want to send the catalog to. My manager has been asked to determine how much profit the company can expect from **sending a catalog** to these customers.

I, the business analyst, was assigned to help my manager run the numbers. While fairly knowledgeable about data analysis, my manager is not very familiar with **predictive models.** I've been asked to **predict the expected profit** from these 250 new customers. Management does not want to send the catalog out to these new customers **unless** the expected profit **contribution exceeds $10,000**.

Details

- The costs of printing and distributing is **$6.50 per catalog**.
- The **average gross margin** (price - cost) on all products sold through the catalog **is 50%.**
- Revenue must be multiplied by the gross margin first before subtracted out the $6.50 cost when calculating profit.

- I must write a short report with recommendations outlining reasons why the company should go with recommendations to my manager.

Business and Data Understanding

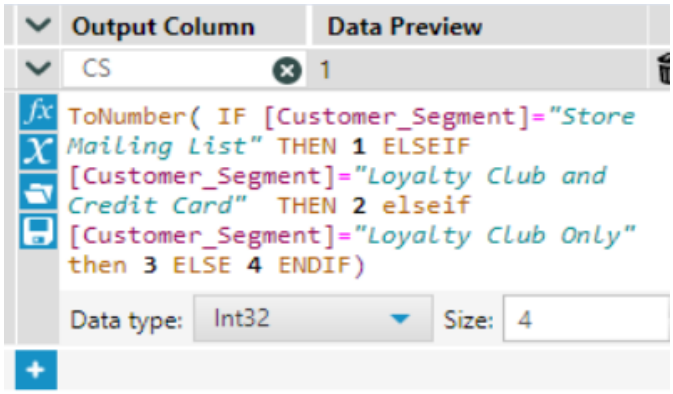
IN this project we are supposed to help our manager in decision making. We have 250 new customers and the decision is to make whether we should send catalogs to them or not and this decision is based on the profit we will make if the predicted profit exceeds $10,000 only then we will make a deal.

In the project we have two data sets . one is the list of **existing customers** on which we are supposed to perform the data modeling and model training for that we have already have lots of variables and we are not going to use all the variables as it may affect the accuracy of the model and create the problem of overfitting which is not good for the model. For that reason we are going to use FIELD SUMMARY and ASSOCIATION ANALYSIS tools to pick out the best variables for a healthy model.

mailinglist contains the data related to 250 customers to whom will be calculated the prediction. The necessary data for the case study prediction was provided.

Steps Used

- In this project for model training i used customer data and loaded it in the alteryx with the help of INPUT DATA tool.
- For variable testing i need to use numeric variables but the customer segment is in categorical data we can't perform analysis directly on it we need to create a dummy variable for that i used formula tool to give each segment a number because this variable has only four distinct values.

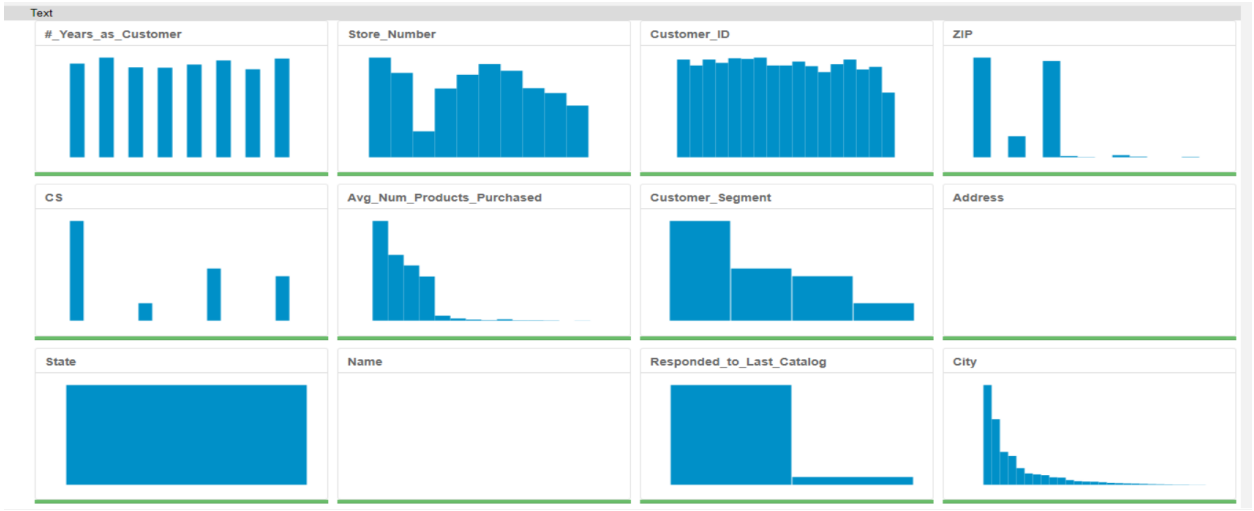


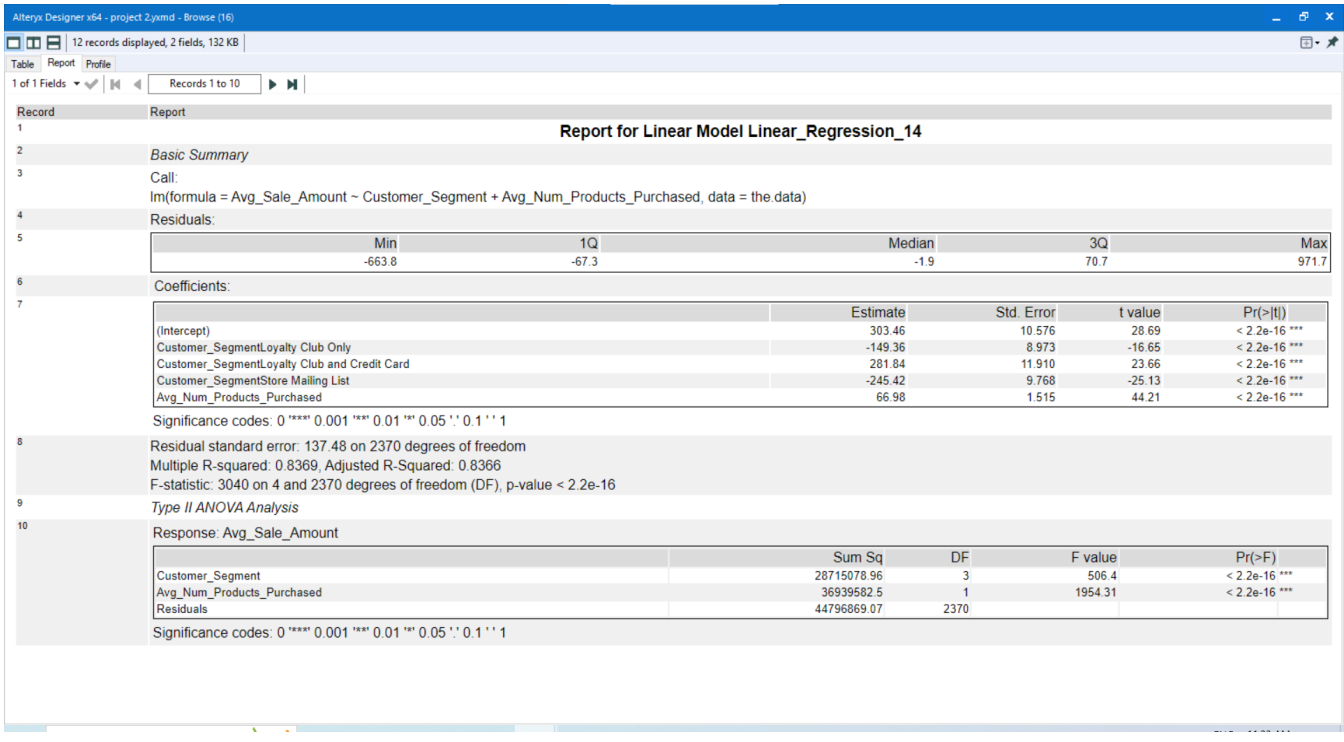
- Converting categorical data into numbers can help us in performing association and field summary analysis on this too.
- Now, I used association analysis on the variables where **target** variable is **AVG_SALES_AMOUNT** and the results shows that only **AVG_NUMBER_PRODUCT_PURCHASES** and **CUSTOMER_SEGMENT is significant** for the analysis but for more details we will also use FIELD SUMMARY TOOL.

Focused Analysis on Field Avg_Sale_Amount

	Association Measure	p-value
Avg_Num_Products_Purchased	0.8557542	0.000000 ***
CS	0.5512116	0.000000 ***
Customer_ID	0.0382352	0.062455 .
X._Years_as_Customer	0.0297819	0.146795
ZIP	0.0079728	0.697758
Store_Number	-0.0079457	0.698734

- AFTER that I used a field summary tool to get the idea why I should not include these variables which are not significant in association analysis . In these visuals we can say that there are few variables which make no sense in including them like CITY being categorical we can include as there are many values NAME and STATES are also insignificant to include. CUSTOMER_ID and ZIP_CODE being numeric values we cant be added as they will not impact also RESPONDED TO LAST CATALOG is categorical and has only two values and most of them are yes so ill not use that.

- 

- So, now I know that the variables that I am going to use are CUSTOMER_SEGMENT and AVG_SALES. Now I can use the **select tool** to select the variables which I am going to use for the prediction.
- now , my data is ready to get processed, now i'll use linear regression model which i'll be connecting the model with mailinglist to predict sales for the consumer.
- 

Alteryx Designer x64 - project 2.yxmd - Browse (16)

12 records displayed, 2 fields, 132 KB

Table | Report | Profile

1 of 1 Fields | Records 1 to 10

Record	Report
1	**Report for Linear Model Linear_Regression_14**
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

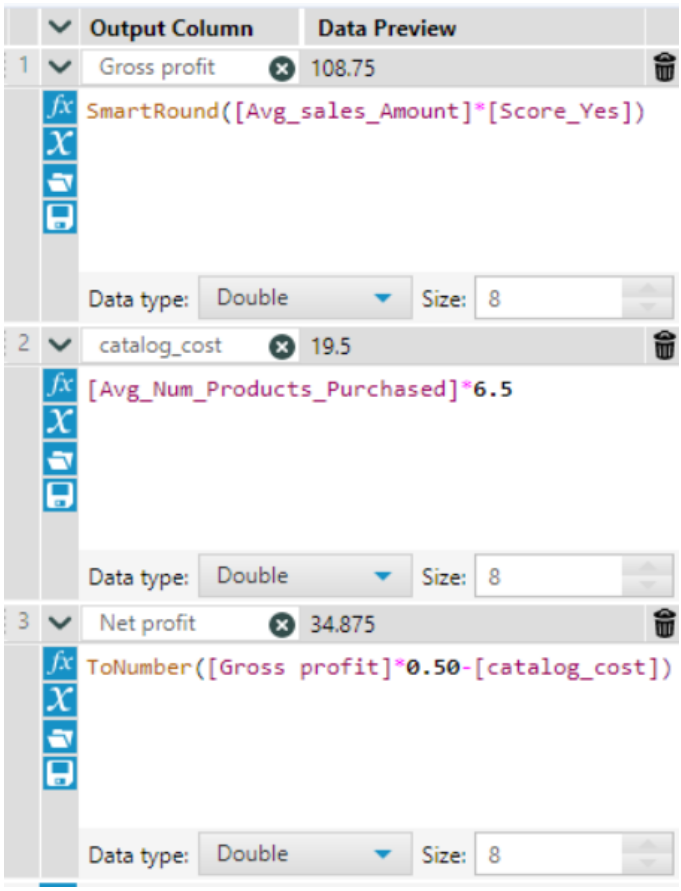
	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28715078.96	3	506.4	< 2.2e-16 ***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16 ***
Residuals	44796869.07	2370		

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

- We can see from the report that adjusted R-squared is 0.8366 which means this model is able to predict 83.66% of variance from the mean which is good.
- All the coefficients are significant having the value lower than 5%.

- From the above report we can drive the equation as

  Avg_predicted_sales = 303.46 + **loyalty club** * -(149.36) + **loyalty club and credit card** * 281.84 + **mailing list** * -(245.42) + **avg_product_purchase** * 66.98

- Now, we will connect this output to the new data set and will connect the output to the SCORE TOOL from the predictive tool pallet to get the avg predicted sales.
- Now we will connect formula tools to get the total profit and make a decision whether we want to make a deal or not.
- First will calculate the sales after multiplying the avg sales with yes score and will calculate total cost by purchase multiply by cost per catalog. Now the last step is to calculate the profit after taking off 50% gross profit margin and deducting the cost. At last we will use summarize tool to get the tool profit.
- 

- *What is your recommendation? Should the company send the catalog to these 250 customers?*

  Yes, the company should send the catalogs to the 250 customers because profit contribution exceeds $ 10,000.

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

  Expected profit is $16016.5 which I get from the last result.