MAKE A COPY

Project 1: Predicting Catalog Demand

Complete each section. When you are ready, save your file as a PDF document and submit it here:
https://classroom.udacity.com/nanodegrees/nd008/parts/c0b53068-1239-4f01-82bf-24886872f48e/project

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?

The company needs to predict how much money they can expect to earn from sending out a catalog to new customers and whether it is profitable to do this.For this work they need to find out the profit. A decision needs to be made to help the company determine, whether or not it should send its catalogs to new customers on the mailing list. The catalogs will only be sent if the profits exceed $ 10,000.

2. What data is needed to inform those decisions?

We need to find out the average profit the company will make by sending out the catalog. By looking at the entire data set and for carrying out the analysis, we will need the following data sets:
List of customers with: customers id, customers segment, average number of products brought from the last catalog, number of years as customers.
Customer segment data includes store mailing list, loyalty club and credit card members.

Average of sales amount for all the customers with different customers segment (store mailing list, loyalty club and credit card members).

Score_Yes data and Score_No data in the mailing list will give us the value of customers who bought the products and the sale amount whereas Score_No data will give the sale amount data of customers who did not buy from the last catalogue.
.
By looking at the customer's data and mailing list data, we can see that we need to build a linear regression model in order to predict out the average 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.

Looking at the p1-customer data file, we can use customer segment (as categorical variable) and average number of products purchased as the numeric variable. Since there is no linear relation between # years of customer data which we can determine by looking at the Fig. 1 and Fig. 2, we will not use this data for our linear model.

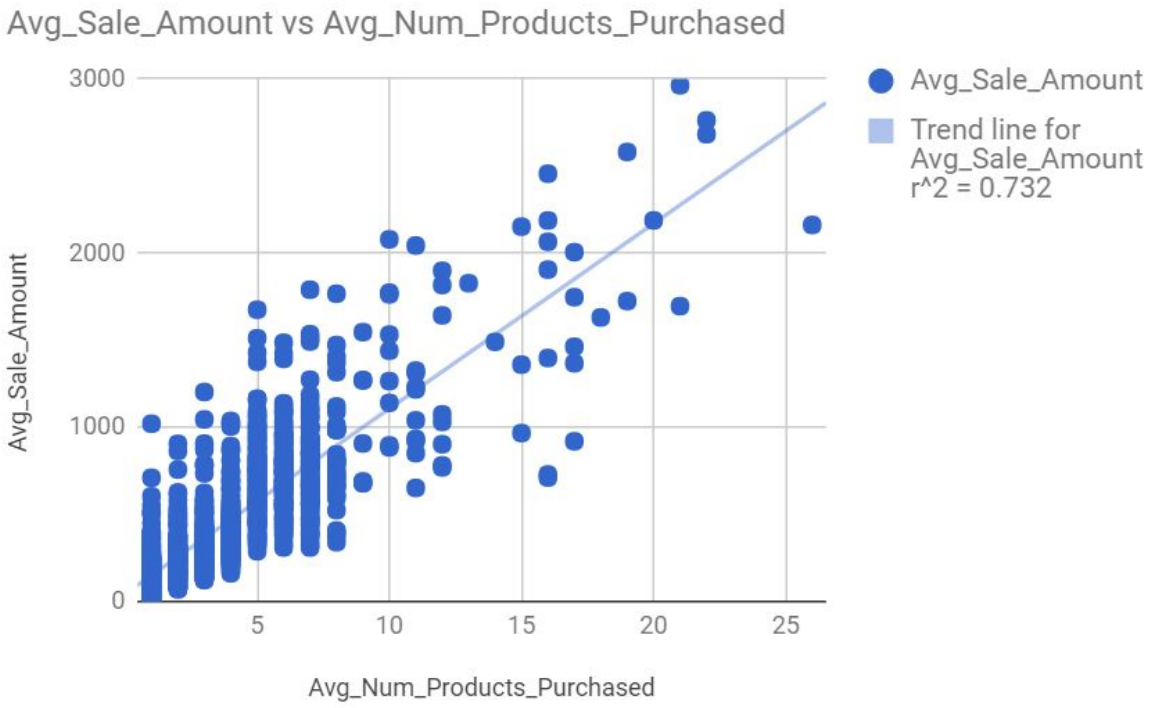

Fig. 1. Average sale amount vs Average number of products purchased

For categorical variable, # years of customer vs average sale amount, we can see the regression model and find out what are the values of regression coefficient. The p-value is less than 0.05 with three '*' and the multiple R squared value is also 0.84 as per Fig. 2. Since the relationship is linear we get a significant value for p and multiple R squared value and therefore # years of customers can be used to build the model.

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.

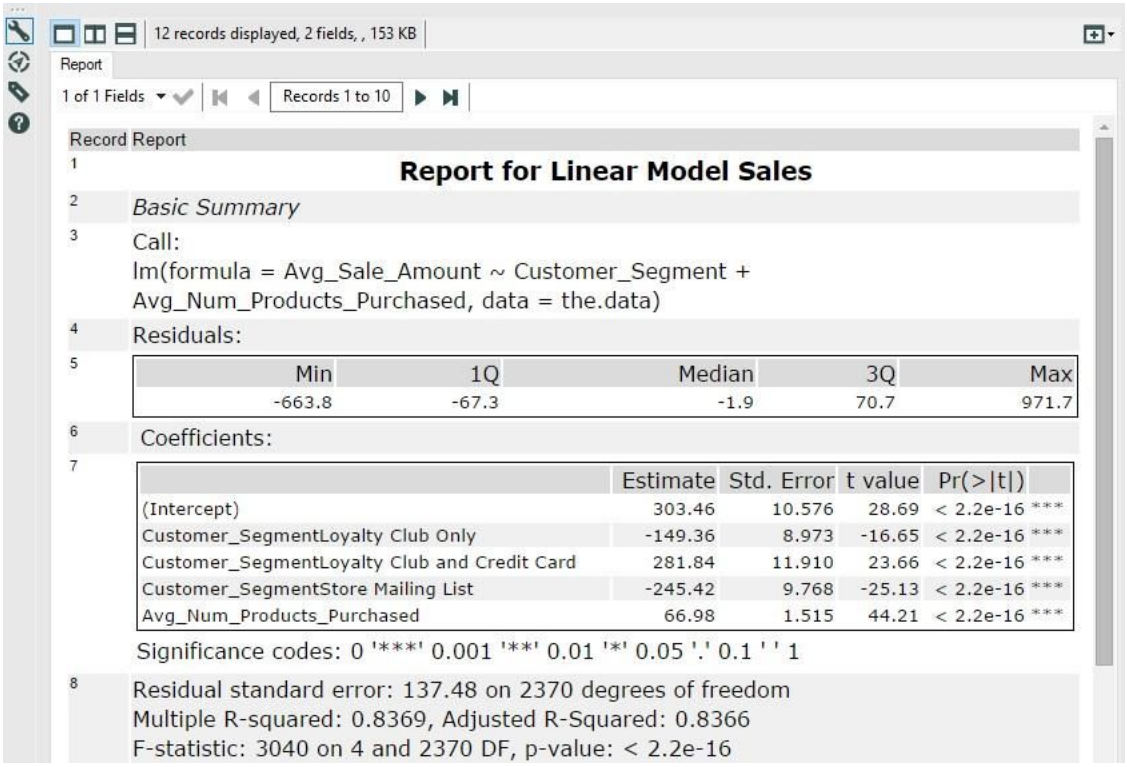

Fig. 2 Report for Linear Model

From the above report we can conclude that the model is good. The R-squared value as well as the p-value justifies it.The p-value is the probability of a more extreme test statistic (a convenient summary of the data) than the one observed, and this probability is evaluated under a given statistical model. From the report we can see that p-vale is very less 0.05, which is significant as well as R square and adjusted R square value is approximately the same 0.84.R-squared is a statistical measure of how close the data are to the fitted regression line

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.

*Y=303.46 + (-149.36*customer_segment Loyalty club)+(281.84*Customer_segment Loyalty club and credit card) +(-245.42*customer_segment store mailing list) + (66.98*avg number of 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 these 250 customers.

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

Considering the data, we can see that there is a linear relation between the predictor variable and target variable. Each predictor variable has a p value less that 0.5. By using the model developed the profit calculated is above $10000, which is clear indication that the developed model is good.
First using the regression equation we find the average sale value of the mailing list.
Secondly, using the average sale value and the value of score_yes (The probability that the customer will respond to the catalog and make a purchase), we calculated the expected spend.
Third, profit is calculated by multiplying the average gross margin and average sales value and substracting the cost of printing and distribution.

4. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
   Expected Profit = [Score]*[Score_Yes*0.5]-6.50
   The Expected Profit comes out to $21987.44