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?

Answer: Whether new product lists should be sent to 250 new customers.

2. What data is needed to inform those decisions?

Answer:

I will use historical customer data from the company to build model to predict the revenue got from prospective customers. The data must include the profit and the costs to do so, including printing and distribution values. If the final profit is greater than $10,000, then the lists should be sent to 250 new customers.I will work on the average number of items each customer bought from the company and the average customer expenses when previously ordering from the past catalogues. These values can be used to calculate revenue from sending out the catalogs.

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.

   I draw a scatterplot and find that there seems to be some correlation between Avg Num Products Purchased and Avg Sale Amount.

Scatterplot of Avg_Num_Products_purchased versus Avg_Sale_Amount

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.

   The p-value of Avg.Num.Products.Purchased and Customer.Segment are 2.2e-16. I should only use the attribute with a p-value below 0.05. So I choose Customer.Segment and Avg.Num.Products.Purchased to build model.
   The R-Squared value of my linear regression model is 0.8369 and the adjusted R-Squared value is 0.8366.

: Suggestion: We found out that the Adjusted R-squared is about 84% with all variables being significant at p-values less than 0.05. It would be nice to include some discussion of what these percentages mean. For example, we could say that - "The model can account for 84% of the actual sales amounts in my training set."

This is a [nice article](http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit) on how to interpret R-squared.

Record Report

Report for Linear Model X

Basic Summary

Call:
lm(formula = Avg.Sale.Amount ~ Customer.Segment + Avg.Num.Products.Purchased, data = the.data)

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 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

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)

   Regression equation:
   Y = 303.46 – 245.42 * Customer.Segment_Store Mailing List + 281.84 * Customer.Segment _Loyalty Club and Credit Card – 149.36 * Customer.Segment_Loyalty Club Only + 0 * Customer.Segment_ 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?

   I recommend to send new product lists to 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 predicted profit is much higher than $10,000.
3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

   The gross profit is 47224.87. The cost of printing and shipping 250 catalogs is 1625. The predicted profit is $21987.44 (47224.87 * 0.5 -1625=21987.44)

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.

: Suggestion:
We could provide a little bit more detailed answer here. For example:
- We need to use the linear model that was created to calculate the predicted revenue for each of the 250 customers. From there we need to multiply the predicted revenue by Score_Yes to get the expected revenue for each customer.

- Then for each expected revenue, we need to multiply the expected revenue by 50% and deduct $6.50 to get a final expected profit for each customer.

- Then we need to add up all of the expected profits from each customer to get the total expected profit.I hope this makes things clearer.

: Awesome: Excellent!