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 wants to decide whether to send the catalogue to the 250 customers or not.
In order to do that we have to predict how much the catalogue will generate revenue and profit if we really sending them out to the 250 customers,
If the total profit exceeds $10.000,00 then we can send the catalogue
If the total profit is less than $10.000,00 then we should not send the catalogue

(profit is revenue generated from mailing list * margin contribution – printing cost)

2. What data is needed to inform those decisions?

In order to inform this decision, we need the existing data of customers and the information about the 250 mailing list that we targeted.

The data we need:
1. the data of the 250 mailing list customer and its profile (have they bought anything from our company yet, how much do they purchase product on average, whether they sent and respond a marketing campaign before this campaign, how long are the become customer, maybe gender, customer segment, etc.)
2. The data of our current customers and its profile
3. The data of all cost associated with printing and mailing the catalogue

: Awesome: Good job updating this section.

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.

In order to find out what variables that relates with the variable "Average Sale Amount" itself. I conducted multiple regression analysis from customers dataset, trying to predict the average sale amount using many of the predictor variables,

We also found that "customer segment", "whether customer responded to last catalogue", and "average number of product purchased" are significant.

After I did the regression, from the initial prediction report below, we can see that the variable of "store number", "X years as customer" are not significant, so we will exclude in the next regression.

While customer responded to last catalogue were significant, I tried to match the mailing list data with the customer data to find the variables, but can't find any. It is probable that the mailing list data is different from customer data and doesn't include the responded to last catalogue information, therefore I exclude the "customer responded the last catalogue" variable. I the company have this information, I would have use the variable to predict the average sale.

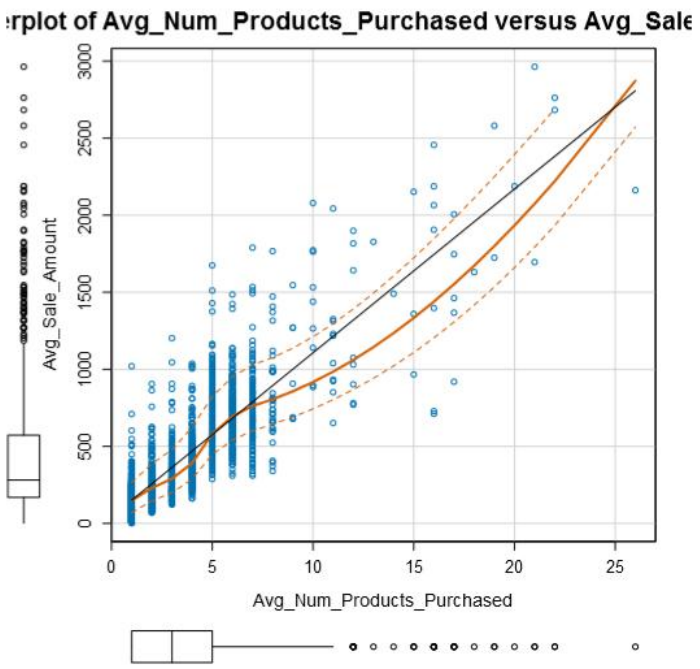
As summary, predictor that I selected in the end were

1. "Customer segment",
2. "Average number of product purchased" ,

Excluded variable were:

1. "Store number"
2. "X years" as customer
3. "responded to last catalogue"

From the above selected variable, the only continuous variable that we can plot to average sales variable was "average number of product purchased". Beside the significance, we can see the relationship between average sale and average number of purchased product from this scatter plot



The scatter plot shows the positive correlation between average number of product to average sale. The higher the average number of product purchased, leads to higher average sale in general.

The predictor “customer segment” was a categorical variable, which cannot descripted using scatter plot, and number of years as a customer was a discrete variable that has less significance so we cannot see the variable relationship well using the scatter plot also, but both significance can be shown from initial prediction regression analysis report.

: Awesome: The predictor variables and explanation is right this time.

Report for Linear Model Initial_prediction

Basic Summary

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)

Residuals:

Min	1Q	Median	3Q	Max
-666.30	-67.15	-2.53	71.12	973.30

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	327.0135	13.976	23.39859	< 2.2e-16 ***
Customer.SegmentLoyalty Club Only	-150.0280	8.982	-16.70353	< 2.2e-16 ***
Customer.SegmentLoyalty Club and Credit Card	283.6100	11.916	23.80030	< 2.2e-16 ***
Customer.SegmentStore Mailing List	-242.9831	9.827	-24.72667	< 2.2e-16 ***
Store.Number101	-5.5294	11.235	-0.49214	0.62267
Store.Number102	-8.6893	16.743	-0.51897	0.60383
Store.Number103	-4.4862	11.903	-0.37688	0.70629
Store.Number104	-21.2748	11.303	-1.88229	0.05992 .
Store.Number105	-20.9124	10.951	-1.90956	0.05631 .
Store.Number106	-18.1956	11.175	-1.62823	0.10361
Store.Number107	-14.7112	11.899	-1.23631	0.21647
Store.Number108	-12.0088	12.158	-0.98773	0.32339
Store.Number109	-0.1426	13.024	-0.01095	0.99127
Responded.to.Last.CatalogYes	-29.1449	11.277	-2.58455	0.00981 **
Avg.Num.Products.Purchased	66.7485	1.517	43.99951	< 2.2e-16 ***
X..Years.as.Customer	-2.3737	1.224	-1.93886	0.05264 .

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

Residual standard error: 137.28 on 2359 degrees of freedom

Multiple R-squared: 0.8381, Adjusted R-Squared: 0.8371

F-statistic: 814.2 on 15 and 2359 DF, p-value: < 2.2e-16

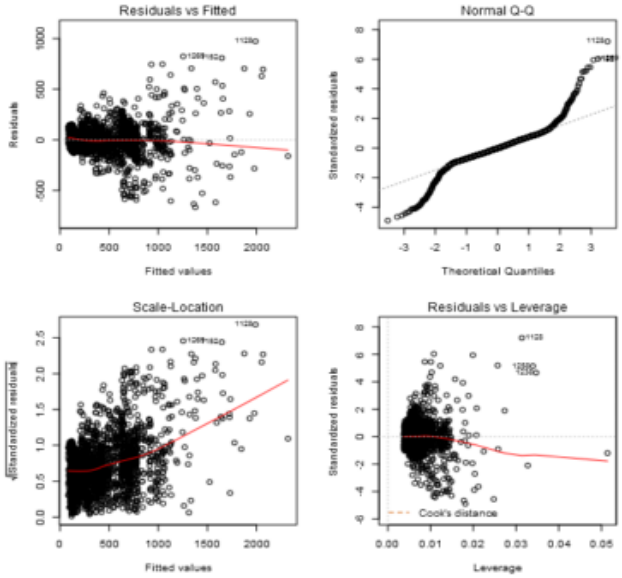
Type II ANOVA Analysis

Response: Avg.Sale.Amount

	Sum Sq	DF	F value	Pr(>F)
Customer.Segment	28414788.81	3	502.58	< 2.2e-16 ***
Store.Number	153560.19	9	0.91	0.51948
Responded.to.Last.Catalog	125889.2	1	6.68	0.00981 **
Avg.Num.Products.Purchased	36485035.54	1	1935.96	< 2.2e-16 ***
X..Years.as.Customer	70845.8	1	3.76	0.05264 .
Residuals	44457704.7	2359		

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

Basic Diagnostic Plots



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.

This is the new linear regression

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

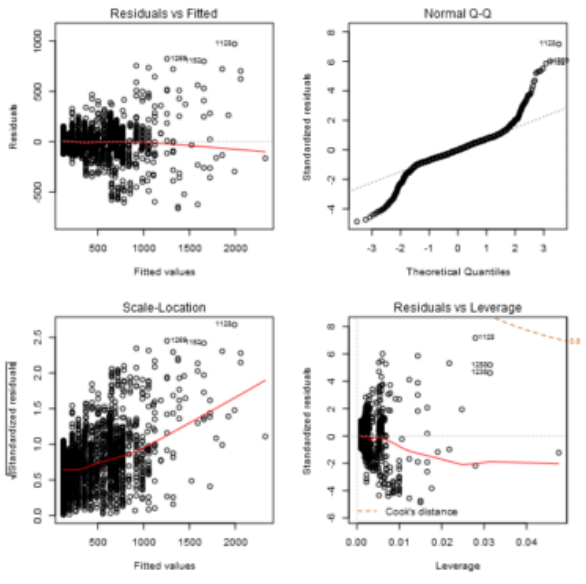
Type II ANOVA Analysis

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

Basic Diagnostic Plots



This regression is a good model because all the variable has low p-value , all below 0.05, except X years as customer which is still around 0.05. And the R-square of the model is 0.8369, which is good, but can be better if it's get around 0.9. but as long as its above 0.7 is still useable

: Awesome: Good job using the R-squared and p-values to justify the model.

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 = 327.01 + (*66.98 * Avg Num Product Purchased)* + (0 if customer segment: Credit Card Only) – (149.36 if Customer Segment: Loyalty Club only) + (281.84 If Customer Segment : Loyalty Card and Credit Card) – (245.42 If Customer Segment: Store Mailing List)

: Requirement: The constant term is incorrect. It was correct in the previous submission (313.76). The other terms are correct.

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?

The Recommendation is to send the catalog to these 250 customers

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

First:
I change the variable type into appropriate types using auto-field and select tools

Second :
I checked the cleanliness of the data using field summary tools

Third:
I made multiple regression analysis using many predictor variables, and check the ones that have significant p-value. I used the customers list to create the model

Fourth:
I use score tools to find the predicted average sale amount of the 250 customers using the model I made, and applied that model to mailing list data

Fifth:
then I multiply all the predicted avg sales with the yes score

Sixth: I sum all the result, and found sum of predicted overall sales were $ 47224.87

Seventh: I Multiply by 50% for profit contribution, and minus it with the cost (250 * 6.50 = $1625)

The result were $ 47224.87* 50% - (250*6.50) = $21.987,44

: Suggestion: The profit should be rewritten to $21,987.44 instead of the current format.

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

The expected profit was $21.987,44

: Awesome: The profit calculation is correct this time.

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.