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:

We need to analyse the data and build a model to predict how much money the company can expect to earn sending the new catalogue to 250 new customers. In order to tackle this, we are provided with 2 different data sets, One for the existing customers with 2375 records and the second one is a list of 250 new customers on which we would like to apply the results of the model to provide an estimated profit from the campaign.

1. The important decision to be made is:

The company needs to decide whether or not, they should send the catalogues to new customers. They have a list of 250 new customers and would like to have an estimate of total profit that could be earned from this campaign if they send the catalogues.

As the management would be happy to run the Direct Mail campaign only if the expected profit contribution exceeds $10,000, we need to predict the expected revenue and profit for each customer as well as the entire list of 250 customers.

2. What data is needed to inform those decisions?
    a. List of existing customers including data on whether they bought something when they received the catalogues, how many items they purchased, how much they spent and what was the total amount they purchased from the catalogues items. Customer segments and the number of years they have been customer would be also helpful.
    b. List of 250 new customers
    c. The Gross margin (50% in this case)
    d. The cost of printing and Distributing the catalogues

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)

First, we import the data into Alteryx and take a careful look into the data in order to ensure the data is clean and does not need any further cleansing.

Second, run the Linear Regression model based on all the possible variables (Costumers segment, Whether they responded to last catalogue, Average number of products purchased, and number of years as customer).

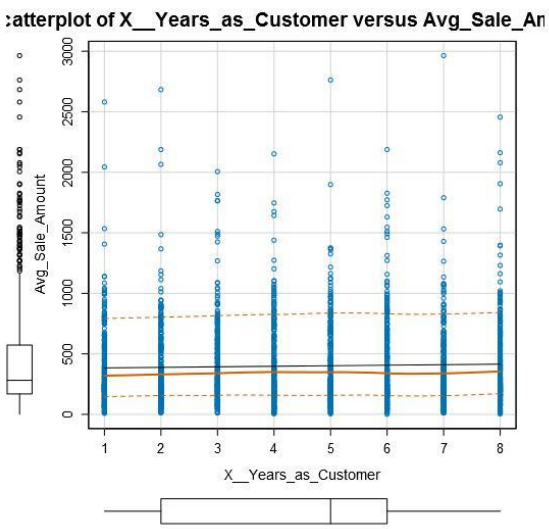
Third, we evaluate the model's results and decide if it produced reliable results using Adjusted R-Squared and P value.

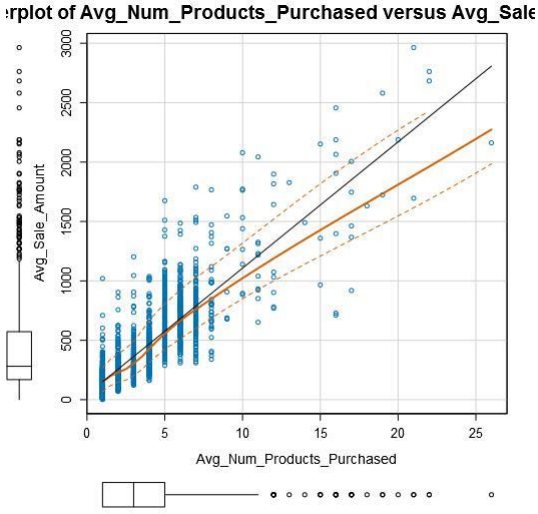
Important: Use the p1-customers.xlsx to train your linear model.

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 first ran the model using all the variables that made logical sense to be included. These values are Customers segment, whether they responded to last catalogue, Average number of products purchased, and number of years as customer. For the 2 numerical values (Average number of product purchased and Number of year as customers) I created a scatter plot against the target value. Below you can see the scatter plots for the 2 values.





As you can see, there is no clear (positive/negative) relationship between "Number of years as customers" and the "average sales amount". However, there is rather strong positive relationship between average number of products purchased and the average sales amount.
In order to determine whether or not to include the "customer_Segment" into the predictive values, we take into account the P-values from the first run (where every possible predictive values were included). Through this trial and error, Customer_Segment which is a categorical variable was chosen as a predictive variable.
When we perform a hypothesis test in statistics, a p-value helps us determine the significance of our results. The p-value is a number between 0 and 1 and interpreted in the following way: A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis. In this case, it shows how significant our predictive variables are. As you can see from the report

below, the P-Values for all the different customer segments are below 0.05, so they are indeed statistically significant to our model.

Record Report

Report for Linear Model Sales_Prediction

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Responded_to_Last_Catalog + Avg_Num_Products_Purchased + X._Years_as_Customer, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-661.90	-68.75	-1.85	70.37	978.20

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	315.165	11.861	26.571	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-149.781	8.963	-16.711	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.467	11.897	23.742	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.842	9.809	-24.756	< 2.2e-16	***
Responded_to_Last_CatalogYes	-27.982	11.254	-2.486	0.01297	*
Avg_Num_Products_Purchased	66.848	1.514	44.147	< 2.2e-16	***
X._Years_as_Customer	-2.313	1.222	-1.893	0.05845	.

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

Residual standard error: 137.26 on 2368 degrees of freedom
Multiple R-squared: 0.8376, Adjusted R-Squared: 0.8371
F-statistic: 2035 on 6 and 2368 DF, p-value: < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	28401236.91	3	502.52	< 2.2e-16	***
Responded_to_Last_Catalog	116471.52	1	6.18	0.01297	*
Avg_Num_Products_Purchased	36716129.06	1	1948.92	< 2.2e-16	***
X._Years_as_Customer	67523.24	1	3.58	0.05845	.
Residuals	44611264.89	2368			

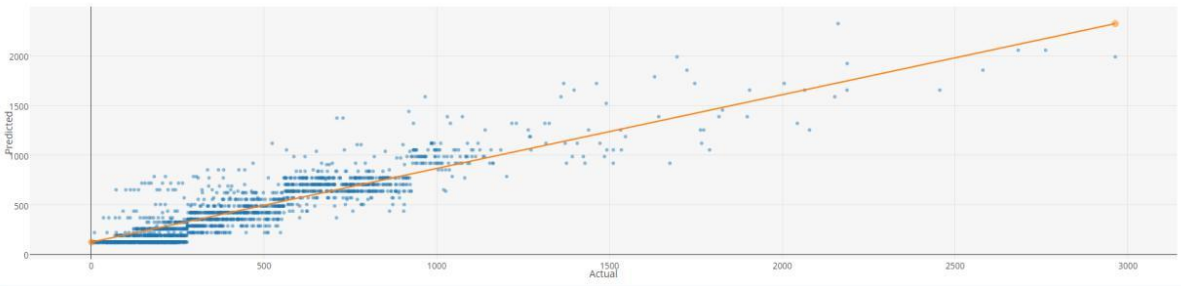
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

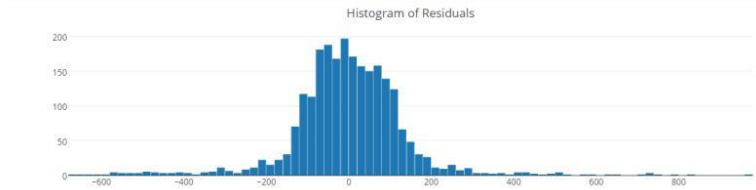
Therefore, the following predictors are included in our model:

- Customer_segment (all 4 different types)
- Avg_num_Products_Purchased

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.

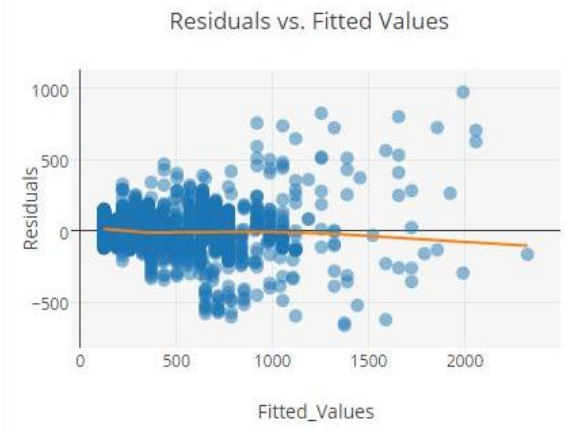
According to the report for linear model we built for the sales prediction, we have the following results:

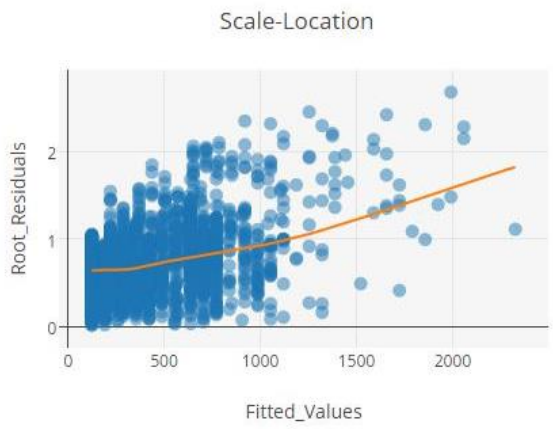


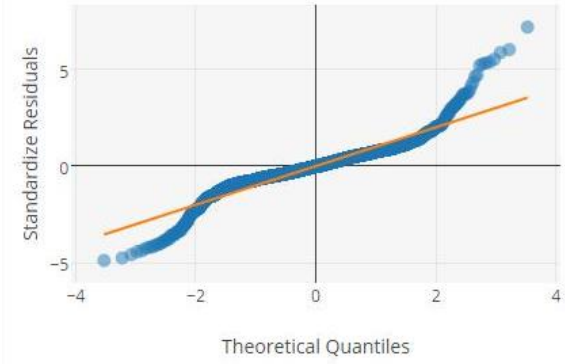


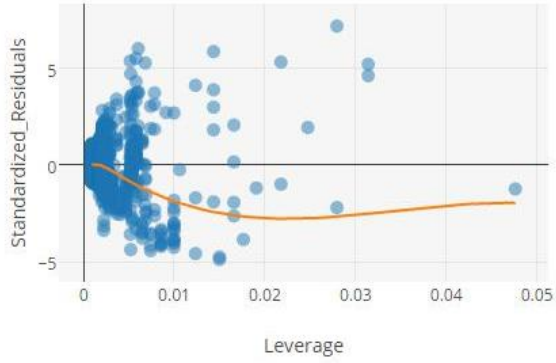
Statistic	Value
Minimum	-663.8
1st Quartile	-67.3
Median	-1.9
Mean	0.0
3rd Quartile	70.7
Maximum	971.7

Regression Diagnostics Plots









Record Report

Report for Linear Model Sales_Prediction

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = inputs$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

As you can see from the charts above, “customer_segment”, and “average_number_of_products_purchased” are statistically significant in our model as the P values are less than 0.05, while the P-Value for “number_of_years_as_customers” is slightly greater than 0.05 so it does not have a significant contribution to our model accuracy.

In multiple linear regression models, we use the Adjusted R-Squared value to indicate how well our data fits the line. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The higher the adjusted R-Squared value, the stronger our model is. In our model, the adjusted p-value show a very high number (0.837). This shows our model has produced reliable results. This is evident in the model performance graphs (presented above) as well.

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……*

Y = 303.46 + 66.98 * Avg_Num_Products_Purchased – 149.36 (If Type: Customer_SegmentLoyalty Club Only) + 281.84 (If Type: Customer_SegmentLoyalty Club and Credit Card) – 245.42 (If Type: Customer_SegmentStore Mailing List) + 0 * (If Type: Customer_SegmentCredit_Card_only)

Note that we have included the 0 coefficient for the type “Credit Card only” as this is the baseline and we need to assign Zero to it.

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?

My model predicted the followings:

- Total sales (predicted) = $138,292.13
- Total expected Revenue = $47,224.87
- Total expected profit = $21,987.43

Considering the above-mentioned numbers and the $10,000 threshold the company had set, I would definitely recommend that the company should send the catalogue to those 250 customers.

3. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
    a. First, I created the model with the target and predictive values discussed earlier.
    b. Second, because I needed to score the model for the list of 250 customers, I brought the p1_mailinglist.xlsx file into the canvas.
    c. I created a separate column to convert the score_no and score_yes into a column with ‘Yes’ or ‘No’ values. IIF([Score_No]>[Score_Yes], ‘No’, ‘Yes’)
    d. Then, I added a Score tool to the canvas and connected the Object end of the model to one side and the second file to the other leg.
    e. The results from the Score tool were input into a Formula tool that calculated “Expected_revenue”, “Revenue_minus_Gross_Margin” using these formulas:
    [Average_Sale_Predicted]*[Score_Yes]
    [Expected_Revenue]*0.5
    [Rev_Minus_Gross_Margin]-6.5

    f. Finally, I used Summarize tool to create the results we required.

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

Based on my calculations, the expected profit from the new catalogue is $21,987.43.

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.