

Project 1: Predicting Catalog Demand

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?

Management wants to decide whether or not to send catalog out to 250 new customers depending on the expected profit contribution exceeds \$10,000.

... ...: Correct

2. What data is needed to inform those decisions?

- The amount of sales from already existing customers that made purchases in order to train the model.
- Data such as average number of products purchased and customer segment to use as predictor variablesi in the linear regression model.

... ...: Required :: Excellent start ! You are absolutely correct highlighting all these data points. We will need all this data.

Additionally , we will also need data regarding "input costs" and "margins" without which we will not be able to compute the expected profit. Kindly add this point to your answer.

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.

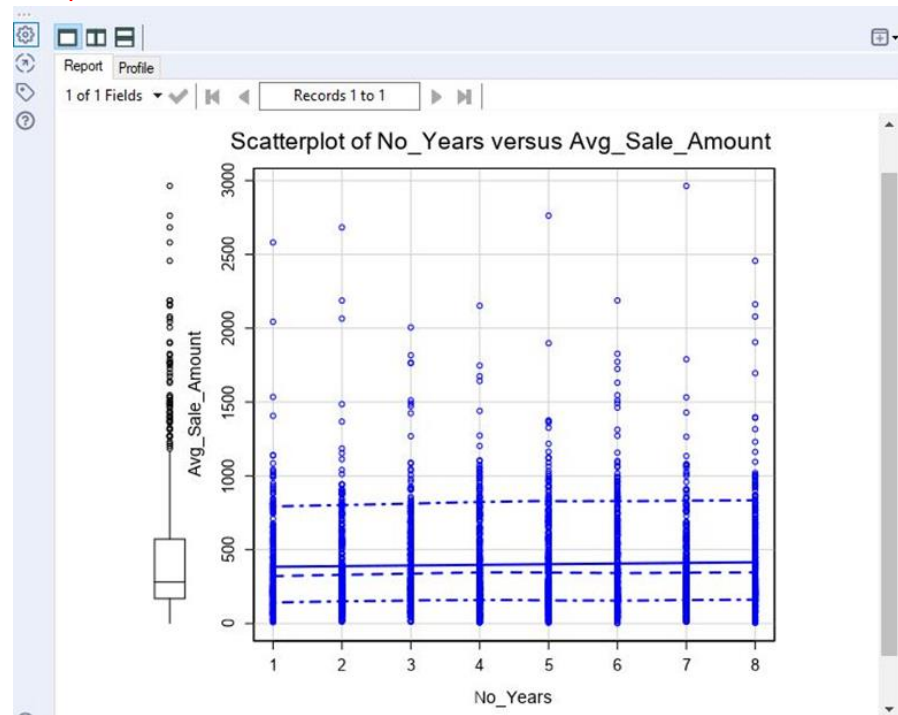
At the minimum, answer these questions:

1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you’ve chosen have a linear relationship with the target variable. Please refer back to the “Multiple Linear Regression with Excel” lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

First, I had to investigate the relationships between numerical predictor variable such as average number of products and number of years as a customer and the target variable average sales amount.

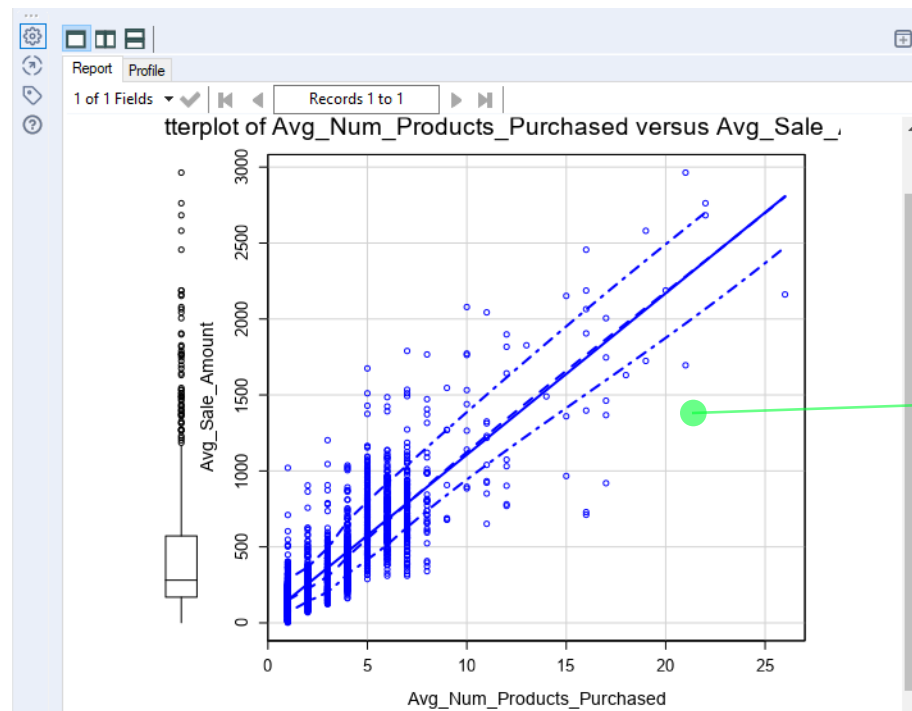
- A. I made a scatter plot of No of years as a customer and average sales amount using the scatter plot tool on Alteryx using p1 customer (train data). From the graph below, I

noticed there was no relationship between these variables, so I didn't use the No of years as a predictor variable



B. I also made a plot with the 2<sup>nd</sup> numerical variable, average number of products purchased and average sales amount. There was a positive relationship between them, so I made it a predictor variable.

... ..: Comment :: Along with the positive relationship , please also highlight that its p-value was less than 0.05



... ..: This visual looks great!

- C. There was a categorical variable known as responded to last catalog which would have been excellent for the model. But there was no record of it in the p1-mailing (test data). So I didn't make use of it.
- D. I made use of customer segment after getting the model, I inspected the p value and star, and found it statistical significant.

... ..: Awesome :: The correct predictor variables have been shortlisted

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.

Below is a picture of my linear regression output

Record

Report

1

**Report for Linear Model Predict\_Average\_Sales\_Amount**

2

*Basic Summary*

3

Call:

Im(formula = Avg\_Sale\_Amount ~ Customer\_Segment + Avg\_Num\_Products\_Purchased, data = the.data)

4

Residuals:

5

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

6

Coefficients:

7

	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

8

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

... ..: This report is absolutely correct

Both the average number of products purchased and customer segments had a very small p-value, less than 0.05 (<0.05), so that both of these predictor variables are statistically significant in predicting the average sales per customer. I got an adjusted r-squared of 0.8366 which is close 1, so in general model showed that these variables can be used to predict the average sales per customer.

... ..: The goodness of the model is well justified.

... ..: Suggestion :: To further enhance the quality of your report , I would advise you to further elaborate your explanation about the adjusted r-squared value.

Make sure that you communicate the threshold to

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:**

$$Y = \text{Intercept} + b1 * \text{Variable}_1 + b2 * \text{Variable}_2 + b3 * \text{Variable}_3 \dots$$

**For example:**  $Y = 482.24 + 28.83 * \text{Loan\_Status} - 159 * \text{Income} + 49 (\text{If Type: Credit Card}) - 90 (\text{If Type: Mortgage}) + 0 (\text{If Type: Cash})$

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.

$$Y = 303.46 - 149.36 * \text{customer\_segment}(\text{loyalty\_club\_only}) + 281.84 * \text{customer\_segment}(\text{loyalty\_club\_and\_credit\_card}) - 245.42 * \text{customer\_segment}(\text{mailing\_list}) + 66.98 * \text{avg\_num\_products\_purchased}$$

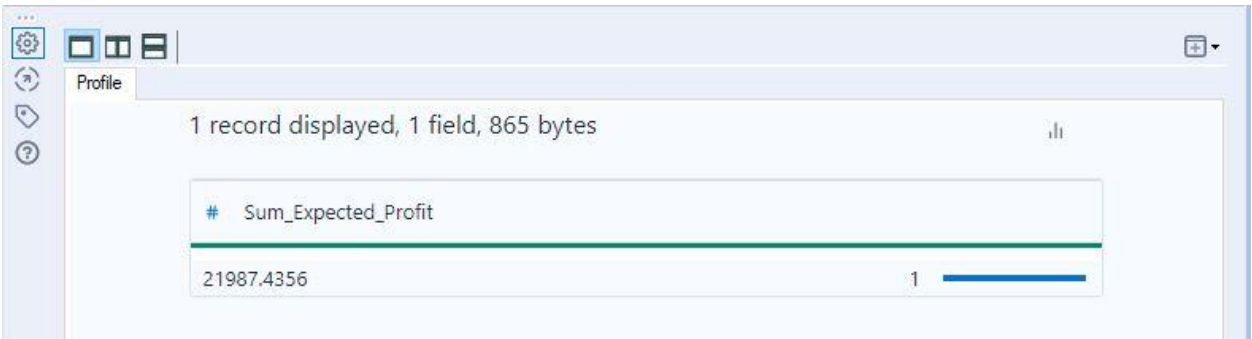
### Step 3: Presentation/Visualization

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?

Yes!



Management should send out the catalog to the 250 customers because based on the Predicted Sales gotten from the model, the company would generate \$21,987.44 in profit which is far more compared to the \$10,000 minimum expected profit set by management.

which you are comparing the adjusted r-squared value of the model.

For example ::

"The higher the adjusted r squared value, the higher the explanatory power of the model. This value represents the amount of variation in the target variable explained by the variation in the predictor variables. Any model with an adjusted r-square value above 0.70 is considered to be a strong model. Our present linear model has a value of 0.8366, hence it is a good model."

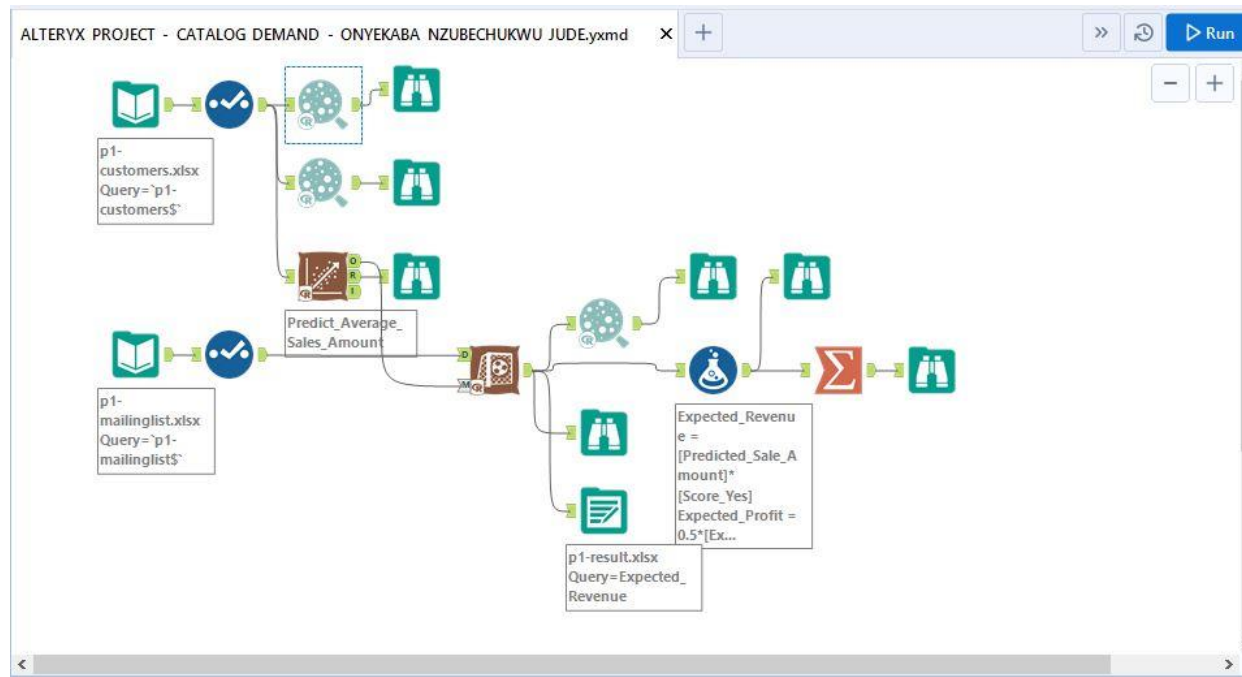
... ..: Nice work!

... ..: Suggestion: The 4th category of Customer segment "Credit Card Only" which is the baseline should be presented like this  $0 * (\text{Customer Segment Credit Card Only})$ . It is a good practice to be explicit when presenting the Linear Regression formula. The reason is that if we give this formula to someone else, he/she may not know what to do exactly for Credit Card Only data points. We cannot assume everyone in the business world will innately understand that "Credit Card Only" is the baseline.

... ..: This is the correct recommendation!

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

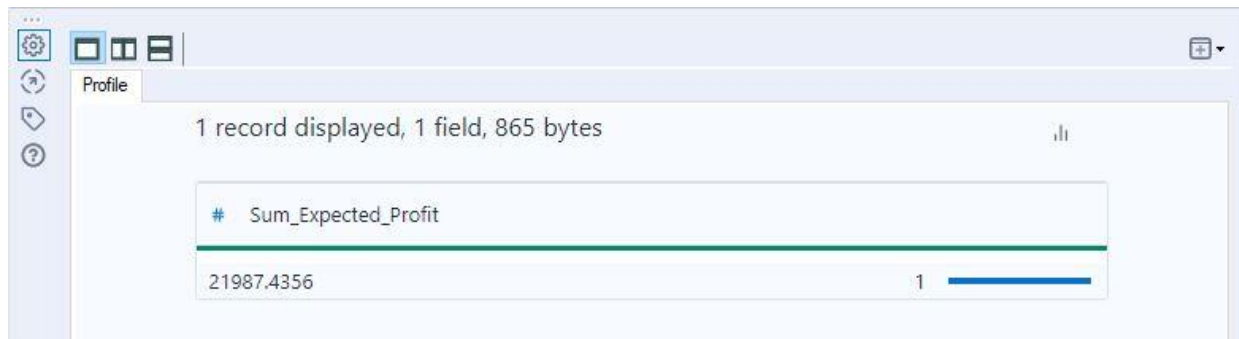
Here is a picture of my Alteryx workflow



- I selected average number of products purchased and customer segment as predictor variables and average sale amount as the target variable then ran a linear regression model in Alteryx from predictive tool.
- I used a score tool to test the data in the p1-mailing sheet
- I calculated using formula tool in Alteryx, the  $[Expected\ Revenue] = Predicted\ Sales * [Score\_Yes]$ . This showed how much predicted sales the customer would generate based on the probability the customer would respond to the catalog. I added another formula for  $[Predicted\_Profit]$  which would be the  $[Expected\ Revenue] * 0.5 - 6.5$  cost per catalog.
- I used the output data tool to save the Predicted Sales in an excel sheet known as p1-result

... ..: Very well explained.

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



The screenshot shows a software interface with a sidebar on the left containing icons for settings, a table view, a chart view, and a help icon. The main area is titled 'Profile' and displays the text '1 record displayed, 1 field, 865 bytes'. Below this, a table with a blue header row is shown. The header row has a column labeled '# Sum\_Expected\_Profit'. The data row shows the value '21987.4356' in the first column and '1' in the second column. A blue progress bar is visible at the bottom right of the table area.

# Sum_Expected_Profit	
21987.4356	1

The expected profit was 21,987.44 dollars

... ..: Fantastic!

### 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.