

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 business decision is whether the company should send catalogs to the 250 new customers?

To answer this question and make the right decision, consider the profit from these 250 new customers. If the profit is over \$10,000, then send the catalogs, otherwise do not send.

2. What data is needed to inform those decisions?

p1-customers.xlsx contains related information of 2,300 customers. It will be used to train the linear regression model.

p1-mailinglist.xlsx contains information of the 250 new customers which will be used to calculate the predicted expected profit to make the business decision.

More specifically, the target variable we're going to predict is Avg_Sale_Amount, i.e.e the revenue generated from the individual customer.

Predictor fields to consider are the followings:

Customer_Segment: categorical, representing source or type of customer, values are Store Mailing List, Loyalty Club Only, Credit Card Only

Geographical info including Address, City, State, Zip

Store Number: categorical, representing different stores

Avg_Num_Products_Purchased: numerical, representing the average number of products the customer purchased

Years_as_Customer: numerical, representing how many years the customer has spent with the store.

In the mailinglist file containing the 250 new customers, we also have Score_No and Score_Yes fields representing the estimated probability the customer will respond to the catalog and make the purchase.

Finally, average gross margin ratio and the cost of printing and distributing each catalog are needed to compute the profits.

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)

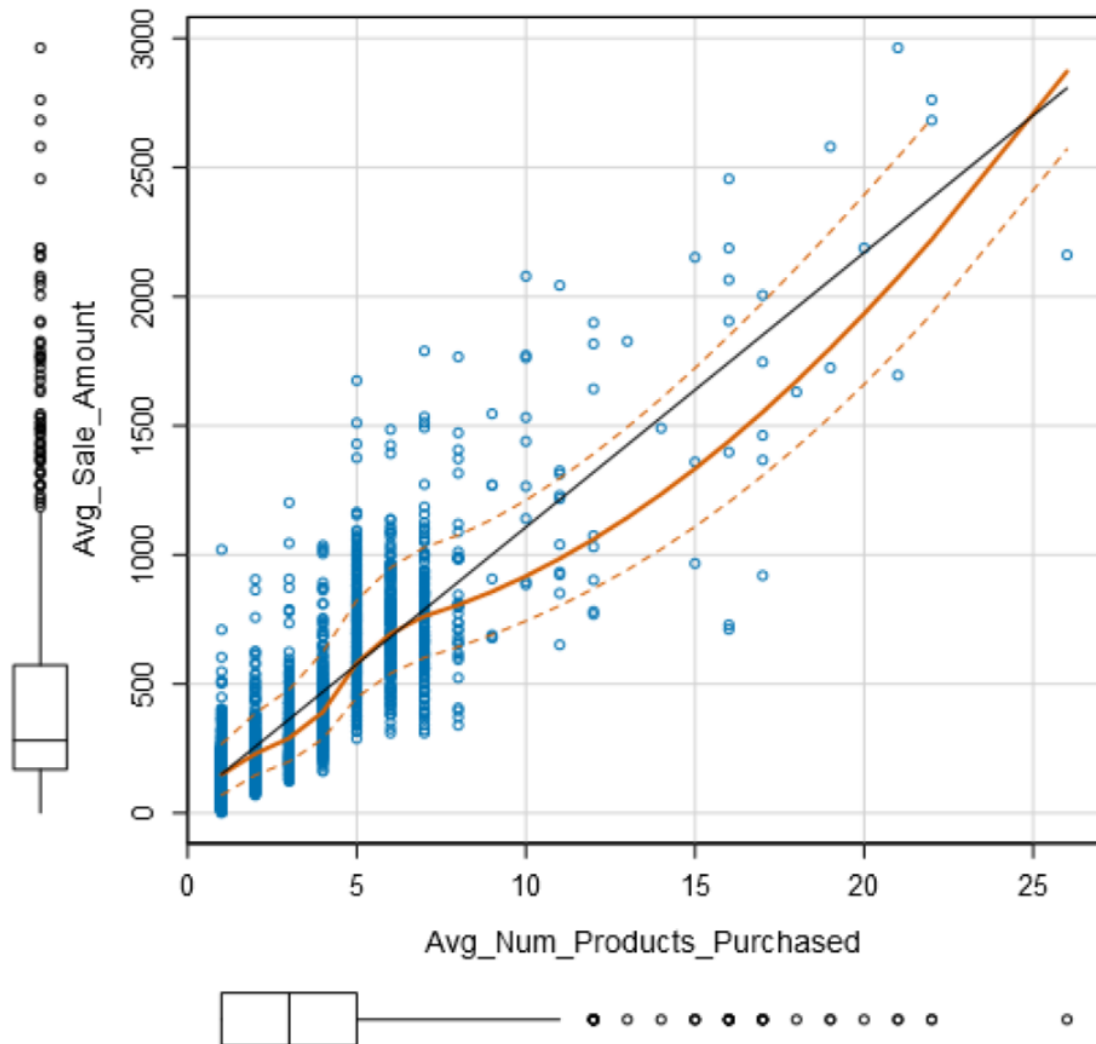
Important: Use the p1-customers.xlsx to train your linear model.

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.

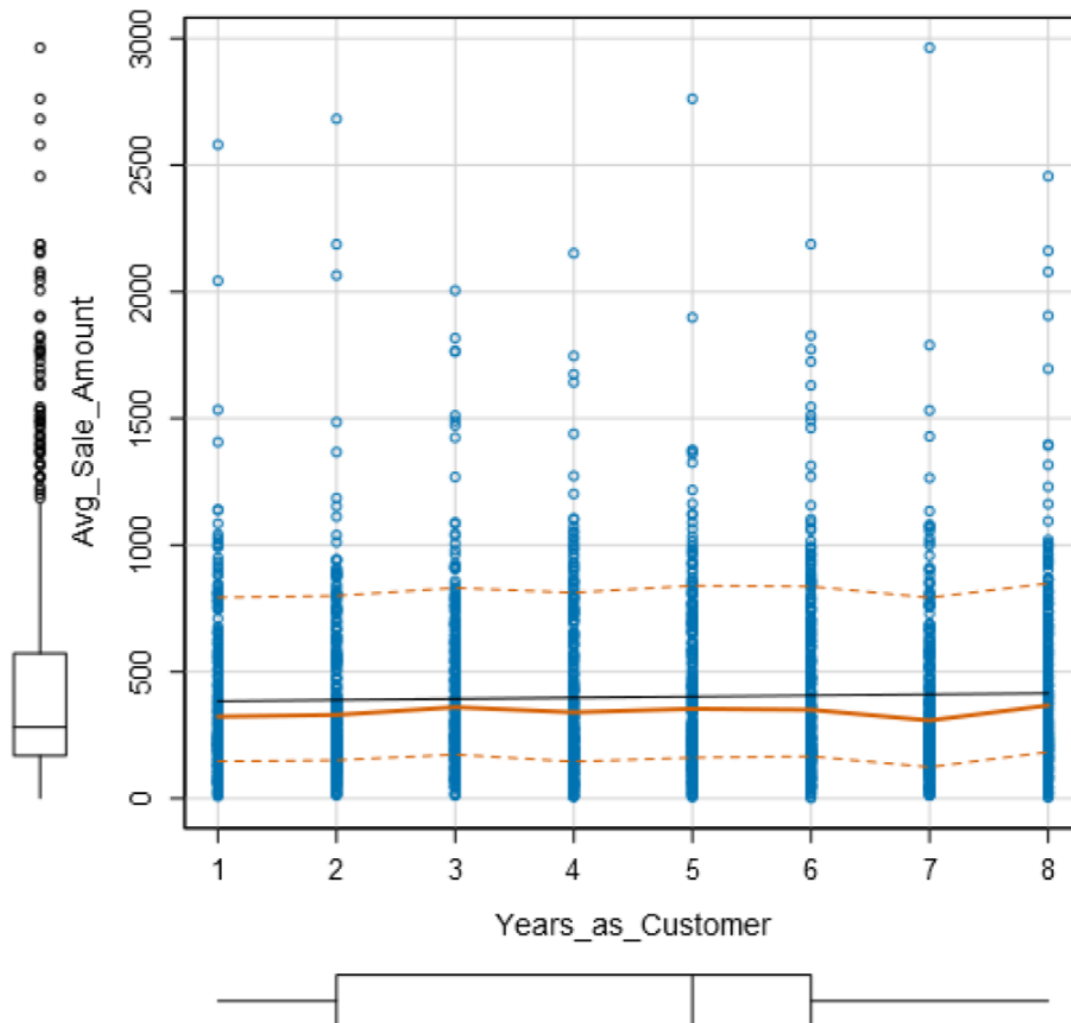
Avg_Num_Products_Purchased (continuous) and Customer_Segment (categorical) are chosen as the predictor variables. The following scatterplot shows the linear relationship between Avg_Num_Products_Purchased and Avg_Sale_Amount(gross revenue). So it is included.

Scatterplot of Avg_Num_Products_Purchased versus Avg_Sale_Amount



Meanwhile, from the following scatterplot, we can see there is no linear relationship between Years_as_Customer and Avg_Sale_Amount, so Years_as_Customer is not included.

Scatterplot of Years_as_Customer versus Avg_Sale_Amc



For all categorical features, we need to select each in the linear regression model for training, and only choose the ones with significant coefficients by looking for those with small (<0.05) p-values. After several tries, only the Customer_Segment feature is significant in the linear regression model with p-value $< 2.2e-16$ for all three segment types.

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 R-squared for the linear model is

Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366.

This means that the two predictor variables explain about 83% of the variation in the Avg_Sale_Amount target variable. So it is a good prediction model.

Moreover, all the predictor variables: intercept, Customer_SegmentLoyalty Club Only, Customer_SegmentLoyalty Club and Credit Card, Customer_SegmentStore Mailing List, Avg_Num_Products_Purchased have p-value < 2.2e-16, so their coefficients are significant.

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)

Score(predicted revenue) = 303.46 - 149.36 x (Customer_SegmentLoyalty Club Only) + 281.84 x (Customer_SegmentLoyalty Club and Credit Card) - 245.42 x (Customer_SegmentStore Mailing List) + 66.98 x (Avg_Num_Products_Purchased).

The three variables Customer_SegmentLoyalty Club Only, Customer_SegmentLoyalty Club and Credit Card, Customer_SegmentStore Mailing List from one-hot encoding of the categorical variable Customer_Segment take values 0 or 1.

Important: The regression equation should be in the form:

$Y = \text{Intercept} + b_1 * \text{Variable_1} + b_2 * \text{Variable_2} + b_3 * \text{Variable_3} \dots$

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.

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 the company to send the catalogs.

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

For each of the 250 new customers, compute the
 $\text{Profit} = \text{Score} * \text{Score_Yes} * 0.5 - 6.5,$

In the above formula, Score represents the estimated revenue from the linear regression model for each new customer. It is then multiplied by Score_Yes field (the estimated probability the customer will respond to the catalog and make the purchase) to get the expected revenue. Then it is multiplied by 0.5 (average gross margin) to get the expected net profit. Finally, subtract the cost 6.5 of the catalog from the net profit. Then sum up all these individual net profits to get the final total profit and compare it with the threshold \$10,000.

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

The total profit is \$21, 987.44 which exceeds \$10,000.

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.