MAKE A COPY

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?

The company has new customers on the mailing list. The decision is whether to mail the new customers the new catalog based on expected profits.

2. What data is needed to inform those decisions?

In order to make that decision, data from the already existing customers and the profit/revenue received by mailing them the catalog will be used to predict the profit/revenue for the new customers.

## **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.

The target variable was the average sale amount (double). In my initial model, I chose customer segment (string), store number (string), average number of products purchased (double), and number of years as a customer (double) as my predictor variables. The reason why I chose store number and not zip code nor state is because I believe store number provides more specific information about location. Stores within different neighborhoods can show different customer purchasing characteristics. Unfortunately, store number was not significant upon conducting the regression, so I chose to use zip code. Zip code is more refined than state.

In the second model, of the 86 zip codes (string), only two were deemed to have statistical significance and the F-score yielded a probability of 0.83617. This means the variable is not significant to be used in the model. I will now attempt to use state.

In the third model, I received errors. There is only one state in this data. I believe city and address are too refined to be used as geospatial data.

I will stick with customer segment, average number of products purchased, and number of years as a customer as the predictor variables in the fourth model. From assessing the p-values of the linear regression, customer segment and average number of products purchased are statistically significant predictors (P-value<2.2e-16). The number years as a customer predictor yields a p-value of 0.0558, slightly above the 0.05 significance level, so it will not remain in the model. The F-test also shows the same p-values, so it is reasonable to assume minimal correlation between predictors will affect the model. A look at the scatter plots and plot of means will show a correlation between the target variable and the predictor variables. I added the plots CustomerSegmentVersusAvgSaleAmount, AvgProductsPurchasedVersusAvgSaleAmount, and YearsAsCustomerVersusAvgSaleAmount to the submission as pdfs.

I will stick with customer segment and average number of products purchased as the predictor variables in the fifth model. From assessing the p-values of the linear regression, customer segment and average number of products purchased are statistically significant predictors (P-value<2.2e-16). The F-test also shows the same p-values, so it is reasonable to assume minimal correlation between predictors will affect the model.

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.

I believed I responded to this in question number 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)

With Y= Predicted Sale Amount,

CS\_LCO= Customer Segment Loyalty Club Only,

CS\_LCC= Customer Segment Loyalty Club and Credit Card,

CS\_SML= Customer Segment Store Mailing List,

CS\_CCO= Customer Segment Credit Card Only, and

PP= Average Number of Products Purchased

**Important: The regression equation should be in the form:**

*Y = Intercept + b1 \* Variable\_1 + b2 \* Variable\_2 + b3 \* Variable\_3……*

**For example:** Y = 482.24 + 28.83 \* Loan\_Status – 159 \* Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (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?

According to the model, the expected profit to be gained is 21,987.44 dollars. This exceeds 10,000 dollars, so I would recommend sending the catalog.

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

After I predicted the sale amount from the linear regression, I predicted the revenue generated from each customer by multiplying the predicted sale amount by the Score\_Yes variable. To predict the profit generated from each customer, I halved the revenue, then subtracted the mailing cost (6.5). I then summed all the predicted profits to obtain 21,987.44 dollars, which is more than the 10,000 dollars wanted in profits to recommend mailing the catalog to clients.

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

The expected profit is 21,987.44 dollars.

Show the distributions for each variable in the Customer List dataset. How would these distributions affect your analysis? Would you need to go back to your manager and try to get more data?

I provided pdfs of plots: CustomerSegmentVersusAvgSaleAmount, AvgProductsPurchasedVersusAvgSaleAmount, and YearsAsCustomerVersusAvgSaleAmount.

YearsAsCustomerVersusAvgSaleAmount did not show significance in the scatterplot, but was almost significance in the regression model. I would recommend more data to observe if the Years as a Customer produces significance. It was so close to being significant that I would request for more data.

Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/186/view) here. Reviewers will use this rubric to grade your project.