Jason Grenig
Udacity Business Analyst Nanodegree

Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Key Decisions:

1. What decisions needs to be made?

Management needs to know whether or not to send catalogs to the 250 customers on the mailing list. If the expected profit of what the customers will buy is less than $10,000, then management will not send out the catalogs.

2. What data is needed to inform those decisions?

In order to make this decision, we need to determine how much each customer will spend who receives the catalog. Building an appropriate linear model and validating it will inform us whether the value proposition of sending the catalogs is justified. If it doesn't merit sending, we could determine the threshold needed (in avg customer spending) to reach the $10,000 tipping point in expected profit.

We also need to know some given information:

- The costs of printing and distributing is $6.50 per catalog.
- Avg gross margin (price - cost) on all products sold through the catalog is 50%.

Step 2: Analysis, Modeling, and Validation

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.

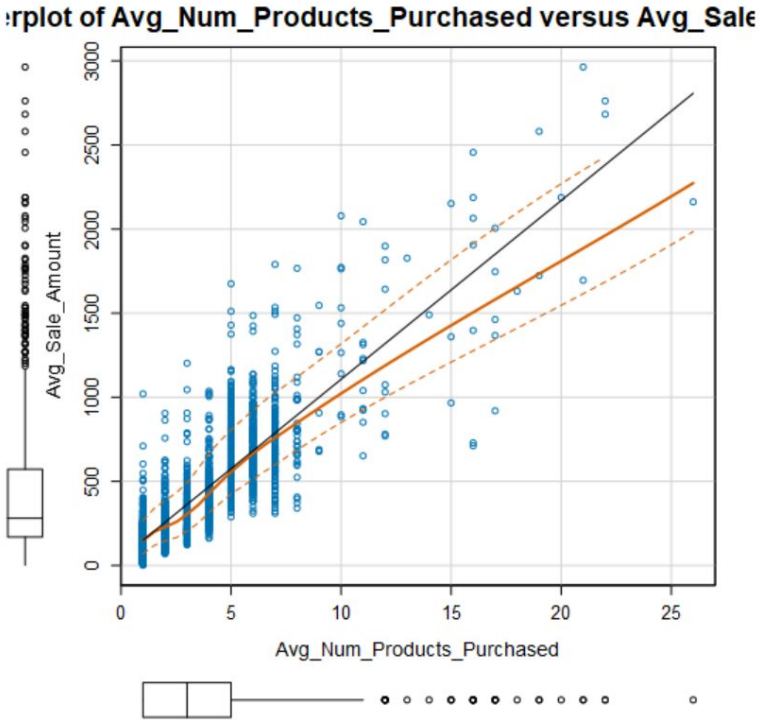
Since we need to predict the amount of expected profit, our target variable for the model is Avg_Sale_Amount. And since we're using a linear model, we're looking for predictor variables that have a strong correlation with our target.

The variables that I hypothesized to have the biggest impact on avg sales were:

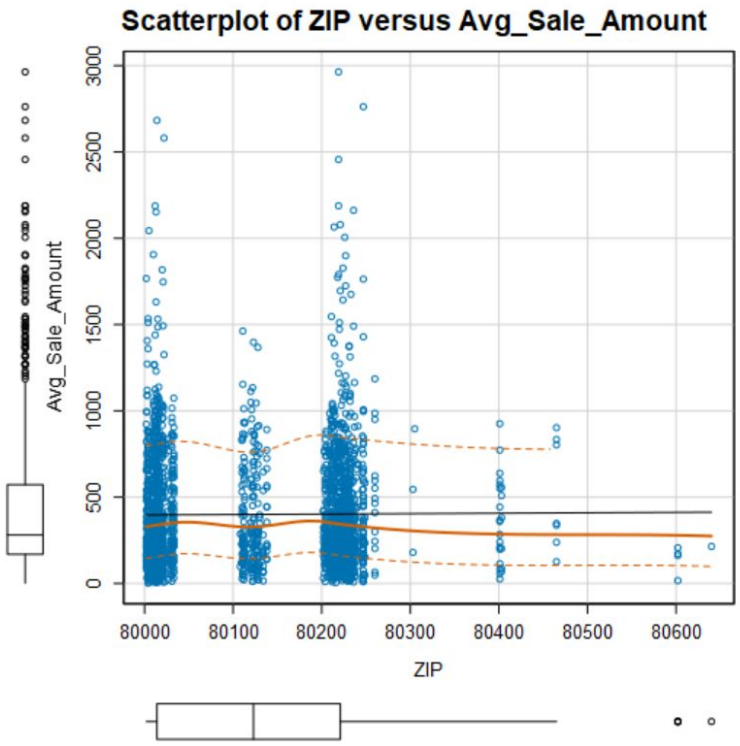
- **Avg_Num_Products_Purchased -** avg products purchased should have a direct relationship with avg sales.

- **Customer_Segment** - I expected each segment to be correlated to sales, but with different coefficients. For ex. Customers with credit cards have push/pull incentives to use the card to earn rewards and if there's an APY, to get enough value from it to merit that cost. This would be in contrast to customers only on the mailing list who may only be interested in the products.
- **ZIP** - zipcodes with wealthier customers (more purchasing power) could be spending more on products.
- **#_Years_as_Customer -** I didn't expect the duration a customer has been on file to have a strong relationship to avg sales, but I wanted to rule it out. It has a very weak correlation to avg sales is similar to the correlation in the below ZIP vs Avg_Sale_Amount scatterplot.

I tested the numeric predictor variables using scatterplots. For Avg_Num_Products_Purchased variables, the scatterplot shows a strong positive linear relationship. In the below figure, we see a trend that the more products a customer buys, the higher the sales amount will be on avg.



I tested these variable relationships vs the target using a scatterplot. In the below figure for zipcode, we can see that the data don't support a strong correlation between avg sales and ZIP. The avg sales amount is pretty consistent across each zipcode.



Since Customer_Segment is a categorical variable, we need to use the R-Squared value to test it. Below we can see the adjusted R-Squared value is 0.702, meaning the coefficients (customer segments) are pretty highly correlated to the target variable.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	682.7	8.354	81.72	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-286.3	11.372	-25.18	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	391.5	15.732	24.89	< 2.2e-16	***
Customer_SegmentStore Mailing List	-525.3	10.045	-52.30	< 2.2e-16	***

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

Residual standard error: 185.67 on 2371 degrees of freedom
Multiple R-squared: 0.7024, Adjusted R-Squared: 0.702
F-statistic: 1865 on 3 and 2371 degrees of freedom (DF), p-value < 2.2e-16

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.

As per the below model results, the model was improved to an Adjusted R-Squared value of 0.8366 - this is based on using only the predictor variables with strong correlations to the target variable in the model: Avg_Num_Products_Purchased (strong positive relationship on the scatterplot) and Customer_Segment (acceptably high r-squared value of 0.7).

Using these 2 predictor variables in the model gives a 0.8366 r-squared value. Raising the value closer to 1.0 and increasing our confidence in the model's accuracy.
Each p-value should also be less than 0.05, which is our threshold to be considered statistically significant. All of the p-values are < 2.2e-16, which tells us the coefficients are significant predictors of our target variable.

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 degrees of freedom (DF), p-value < 2.2e-16

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……*

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.

The best linear regression equation is as below, where the Intercept, 303.46, represents the Credit_Card_Only segment:

Y = 303.46 - 149.36 * Loyalty_Club_Only + 281.84 * Loyalty_and_Credit_Card - 245.42 * Store_Mailing_List + 66.98 * Avg_Num_Products_Purchased

Step 3: Presentation/Visualization

1. What is your recommendation? Should the company send the catalog to these 250 customers?

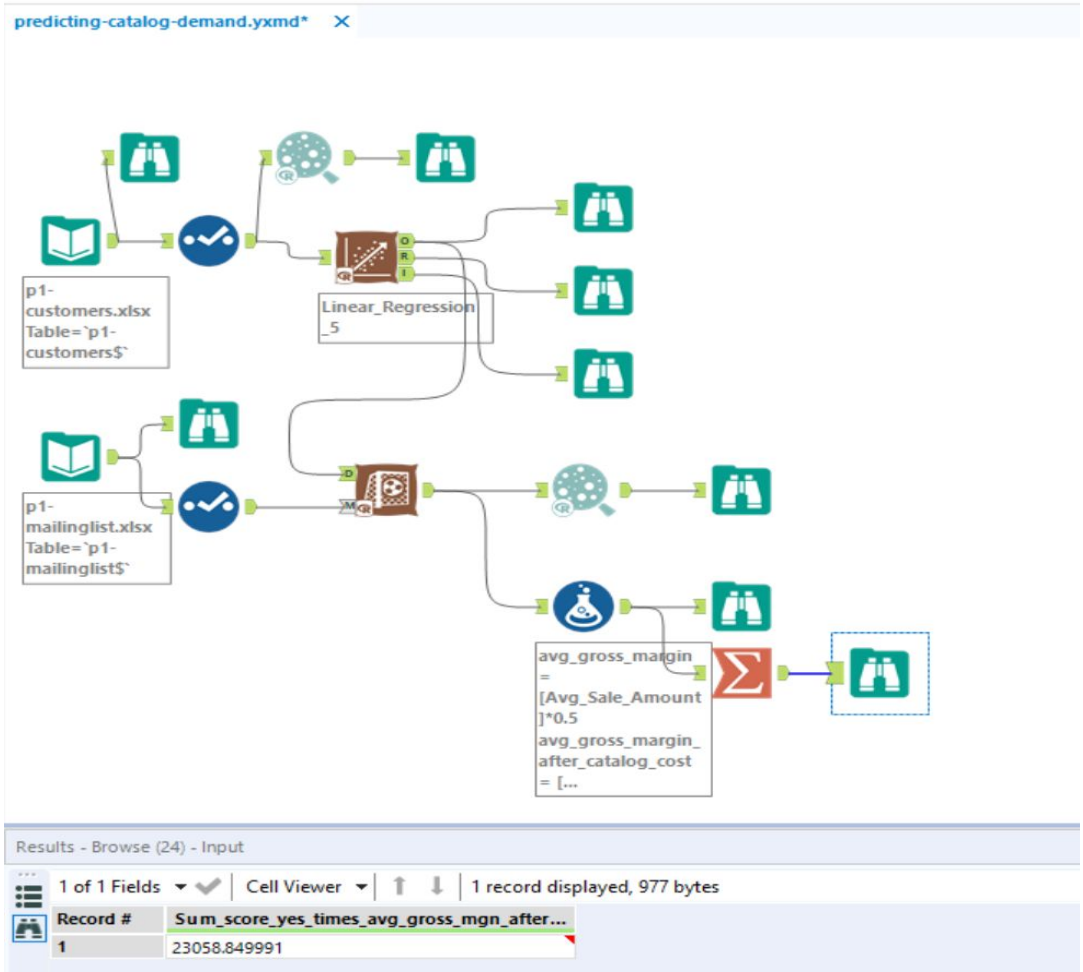
Management should send the catalogs out to these 250 customers. The minimum expected profit needed to break even is $10,000, and the expected profit based on the linear regression model is $23,059.

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

After building the linear regression model, I used the Score Tool to apply the model results to the dataset. From there, I used the Formula Tool to create a couple formulas:

- [avg_sale_amount] * 0.5 (the gross margin) = [avg_gross_margin]
- [Avg_gross_margin] - 6.5 (the cost to produce the catalog) = [avg_gross_margin_after_catalog_cost]
- [Score_yes] * [avg_gross_margin_after_catalog_cost] = [score_yes_times_avg_gross_mgn_after_catalog]

Then I summed the total [score_yes_times_avg_gross_mgn_after_catalog] to get the expected profit of $23,059.
See complete workflow below.



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

The expected profit is $23,059.