Predicting Catalog Demand

Christopher Giler
Udacity Business Analyst Nanodegree

August 23, 2017
Project #1-2

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

Key Decisions:

For this project, the key decision to make is whether or not sending a catalog to 250 new customers on a mailing list will be profitable. More specifically, the company is interested in predicting how much profit they can expect from sending these catalogs out. In order to make these decisions, we must first have an understanding of the associated costs in sending these catalogs, and how profitability is defined. In this case, we know that:

- The cost of printing and distributing is $6.50 per catalog
- The average gross margin on all products sold through the catalog is 50%

We will also need purchase history of existing customers as well as some basic customer demographics and segment information. This will allow us to determine which characteristics drive profits with the company's current customer base, and will allow us to build a model to predict profit based on similar data with 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)

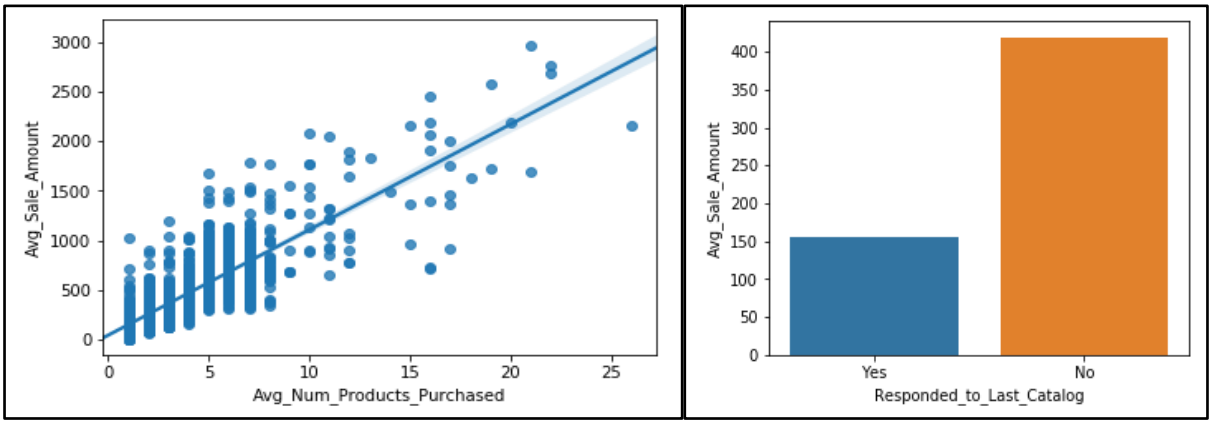
The customer data contains the following information:

- Name & Customer ID
- Customer Segment
- Location (Address, City, State, and Zip Code)
- Store Number
- Responded to Last Catalog
- Average Number of Products
- # Years as Customer
- Average Sale Amount **(This will be our target variable)**

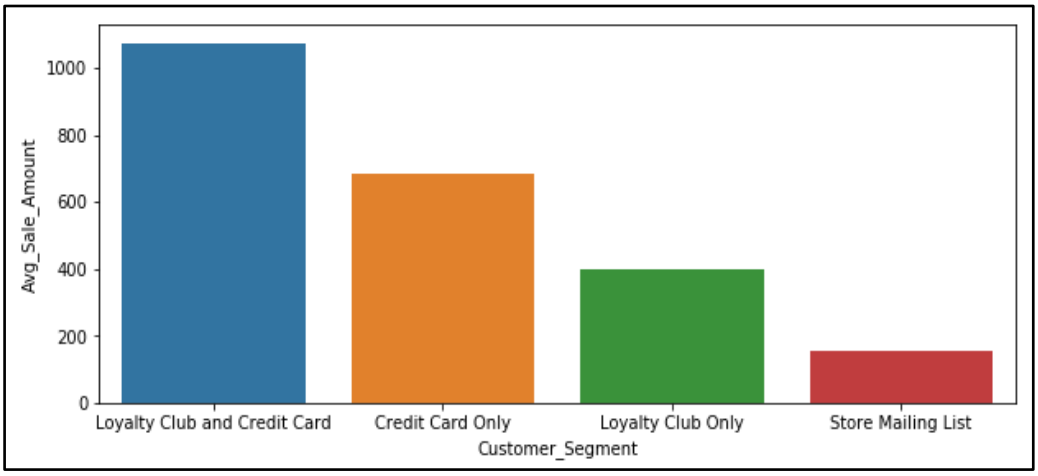
Selecting predictors for the linear model involves exploring the relationship between each variable available against average sale amount using bivariate analysis of the dataset. If a

metric shows a somewhat linear relationship with the target variable, we can assume it would work well as a linear regression input.

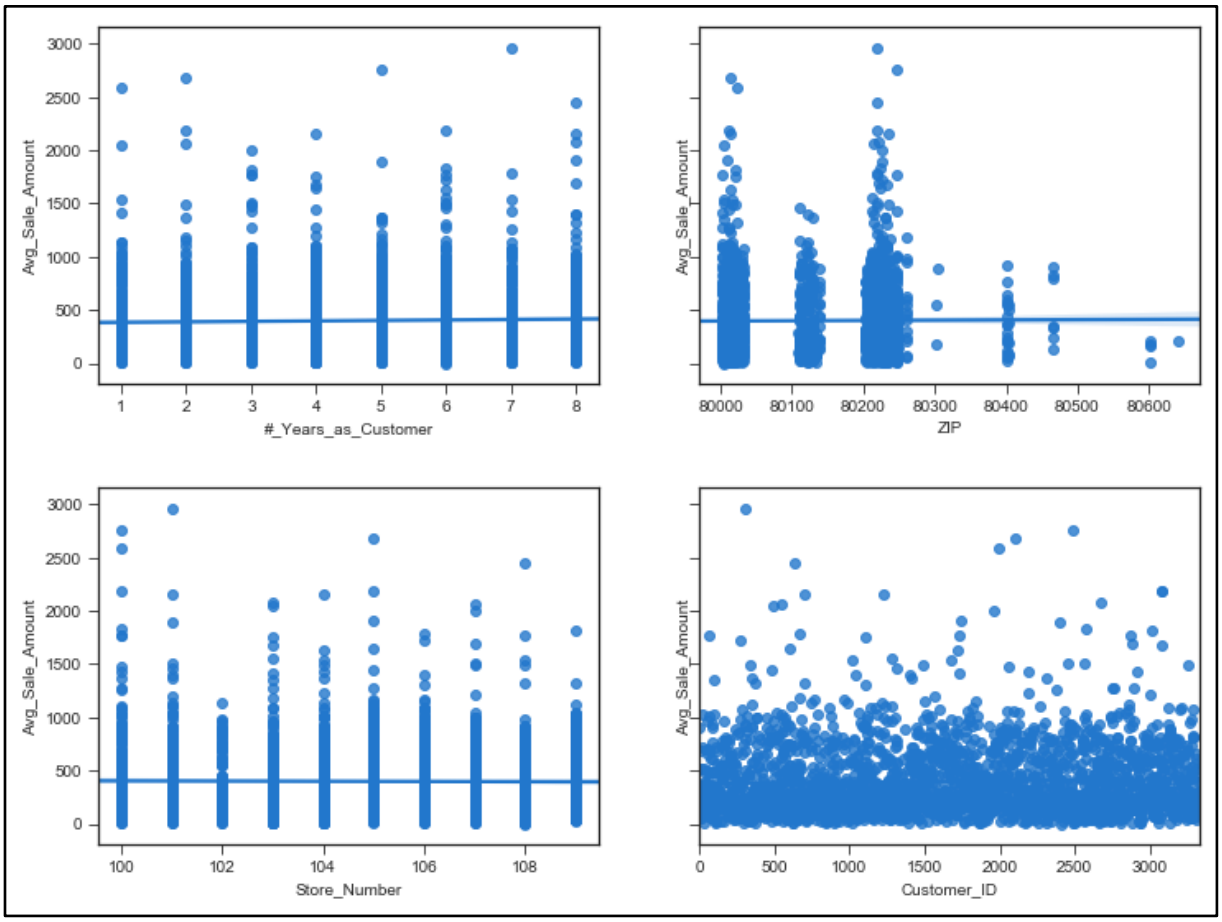
The strongest linear relationship in this dataset is that between the average sale amount and average number of products purchased. There is also a relationship between average sale amount and customer's response to the last catalog.



There is also a strong relationship between average sale amount and customer segment. Customers who own a credit card and are a member of the loyalty club tend to have the highest average sale amount, while customers only on the store mailing list tend to have the lowest.



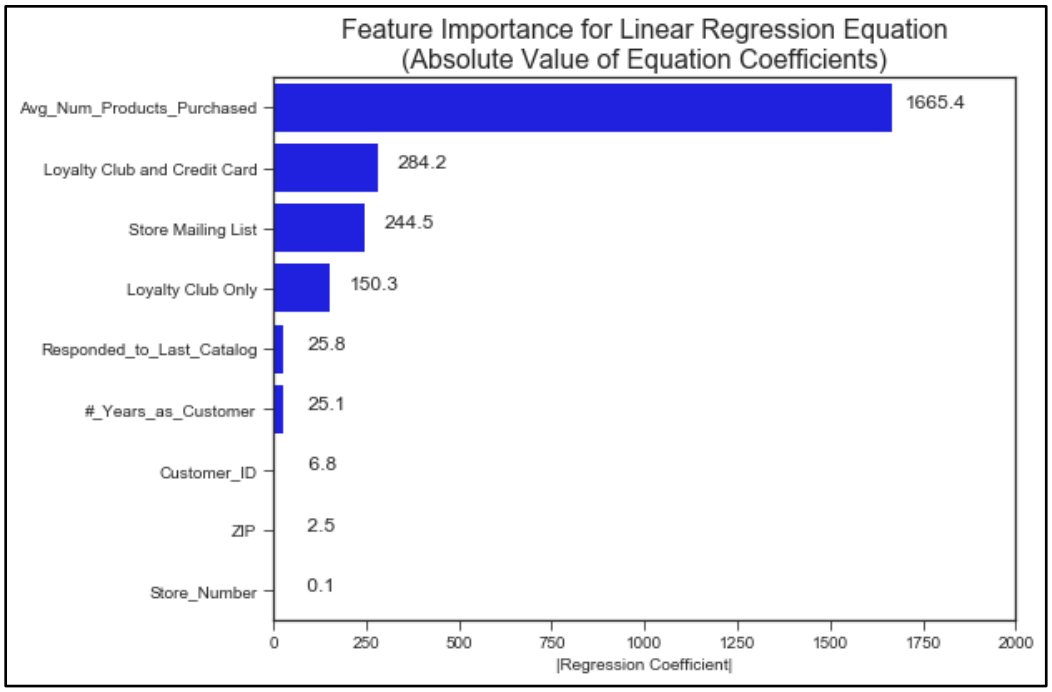
The remaining metrics, which include the number of years as a customer, zip code, store number, and customer ID number, do not appear to have much influence on the average sale amount, so they will not be used as predictor variables for the regression model.



The “Responded_to_Last_Catalog” variable is coded to binary values to work with the linear regression model (0 = No; 1 = Yes). Customer Segment is converted to a set of dummy variables, with one variable for each category within the customer segment.

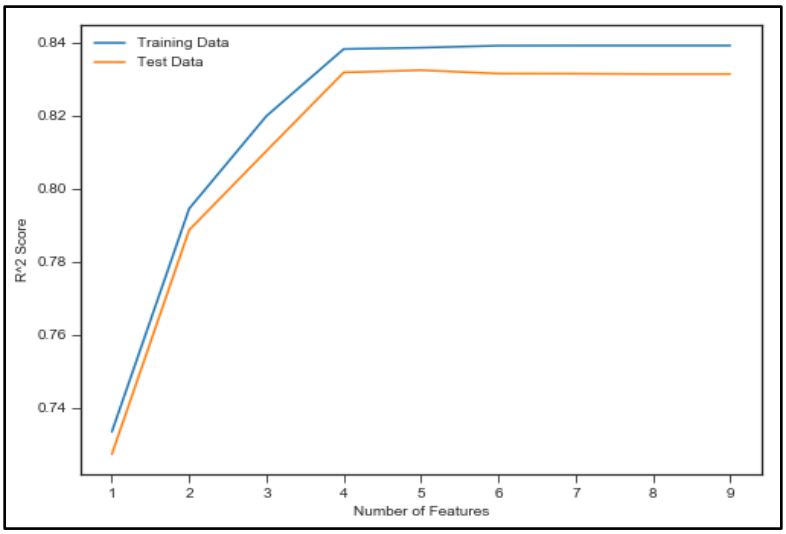
Customer_Segment	Credit Card Only	Loyalty Club Only	Loyalty Club and Credit Card	Store Mailing List
Store Mailing List	0	0	0	1
Loyalty Club and Credit Card	0	0	1	0
Loyalty Club Only	0	1	0	0
Credit Card Only	1	0	0	0

As a first step in selecting the most effective features for the linear model, a linear regression of all customer data is fit to all available numeric features, including those converted to binary values and dummy variables. Entries in the customer segment “Credit Card Only” are taken as the base case. All data is normalized using a Min-Max Scaler, and the resulting equation coefficients are considered to determine the influence of each feature on determining average sales.



The average number of products purchased seems to have the most importance in determining average sale amount. Customer segment is also an important contributor. The features labeled “Responded_to_Last_Catalog” and “#_Years_as_Customer” have little impact in the linear model, but more work is needed to make the final feature selection.

To select model features, a new linear regression model is built and is trained on a randomized subset of 70% of the customer data, without any scaling applied. The remaining 30% of data is used to test and validate the resulting model. This model is built and fit for varying sets of top predictors, and the R^2 score is compared for both data splits to determine the minimum number of predictors needed. Results from this test are shown in the figure below.



From this test, we determine that it is only necessary to use the top 4 predictors to build an effective linear model. Therefore, the selected features are:

- Avg_Num_Products_Purchased
- Loyalty Club and Credit Card
- Store Mailing List
- Loyalty Club Only

Building the linear regression model based on these four features, and training on the full set of customer data, results in a linear equation given as:

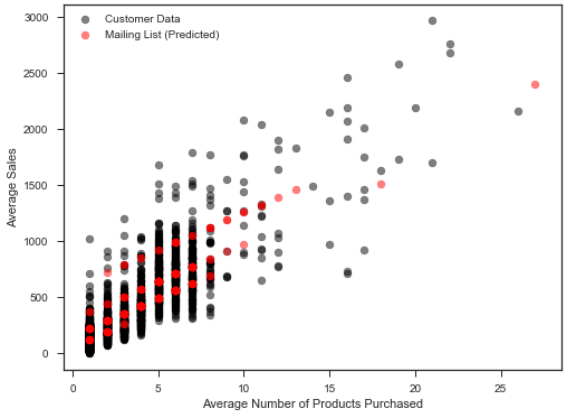
$$\begin{aligned} Average\ Sale\ Amount = 303.46 + \ & 66.98 \times (Average\ Number\ Products\ Purchased) \\ & + 281.84\ (If\ Segment: Loyalty\ Club\ and\ Credit\ Card) \\ & - 245.42\ (If\ Segment: Store\ Mailing\ List) \\ & - 149.36\ (If\ Segment: Loyalty\ Club\ Only) \\ & + 0\ (If\ Segment: Credit\ Card\ Only) \end{aligned}$$

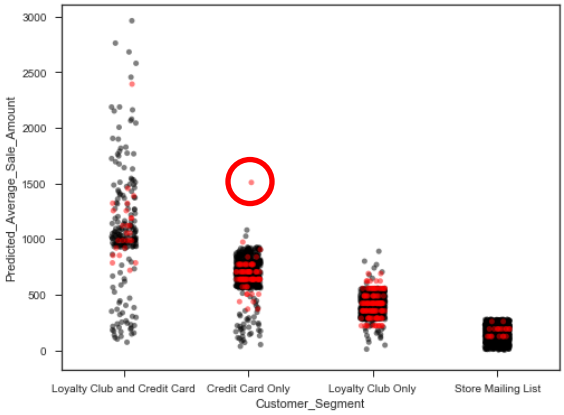
The R^2 score for this model is approximately 0.84 on the overall data set.

Step 3: Presentation/Visualization

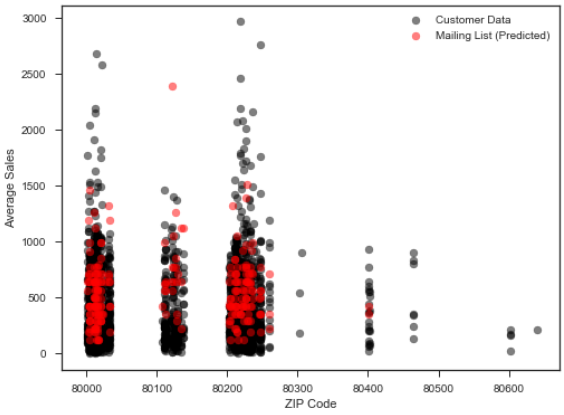
Use your model results to provide a recommendation. (500 word limit)

The linear model built in the previous step is used to predict average sale amount for each customer on a mailing list of 250 potential customers, based on their customer segment and average number of products purchased. Predicted sales are shown in the figures below against the two predictor variables.





The majority of predicted results fall within the ranges found for the feature values, with only one exception in the customer segment of “Credit Card Only.” The results also fall within the general ranges observed in the customer dataset for each zip code, despite this metric not being used as a predictor variable.



The predicted revenue from sending these 250 catalogs is the sum of the product of predicted sale amount and the probability that the customer will respond and make a purchase. That is:

$$Revenue_{pred} = Sales_{pred} \times P(Customer\ will\ make\ purchase)$$

Additionally, the predicted profits must also take into account the average gross margin of all profits sold through the catalog and the costs of printing and distributing each catalog. Thus, the expected profit from each customer is calculated by:

$$Profit_{pred} = \left(Revenue_{pred} \times Gross\ Margin\right) + Cost\ of\ Catalog$$

For an average gross margin of 50% and a cost of $6.50 per catalog, this equation reduces to:

$$Profit_{pred} = 6.50 + \left(0.5 \times Revenue_{pred}\right)$$

These metrics are calculated for each customer.

	Customer_ID	Name	Customer_Segment	Predicted_Average_Sale_Amount	Score_Yes	Predicted_Revenue	Predicted_Profit
0	2213	A Giametti	Loyalty Club Only	355.036364	0.305036	108.298804	47.649402
1	2785	Abby Pierson	Loyalty Club and Credit Card	987.159466	0.472725	466.654501	226.827251
2	2931	Adele Hallman	Loyalty Club Only	622.941184	0.578882	360.609345	173.804672
3	2231	Alejandra Baird	Loyalty Club Only	288.060159	0.305138	87.898046	37.449023
4	2530	Alice Dewitt	Loyalty Club Only	422.012569	0.387706	163.616744	75.308372

Summing these calculations over all 250 customers on the mailing list yields the following results.

Predicted Average Sales	$ 138,292.13
Predicted Revenue	$ 47,224.87
Predicted Profit	**$ 21,987.44**

Factoring in the accuracy of the linear regression model developed in Step 2, as well as the positive profit margin calculated for the new mailing list, we can expect a profitable outcome from sending catalogs to the 250 customers on the mailing list.