Data Report – Predicting Catalog Demand

by Travis Gillespie

Table of Contents

Introduction

You recently started working for a company that manufactures and sells high-end home goods. Last year the company sent out its first print catalog, and is preparing to send out this year's catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.

Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers. You, the business analyst, are assigned to help your manager run the numbers. While fairly knowledgeable about data analysis, your manager is not very familiar with predictive models.

You've been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

Business and Data Understanding

Key Decisions

Question 1

What decisions needs to be made?

Whether or not to send this year's company catalog to new customers; dependant on the profit exceeding $10,000.

Question 2

What data is needed to inform those decisions?

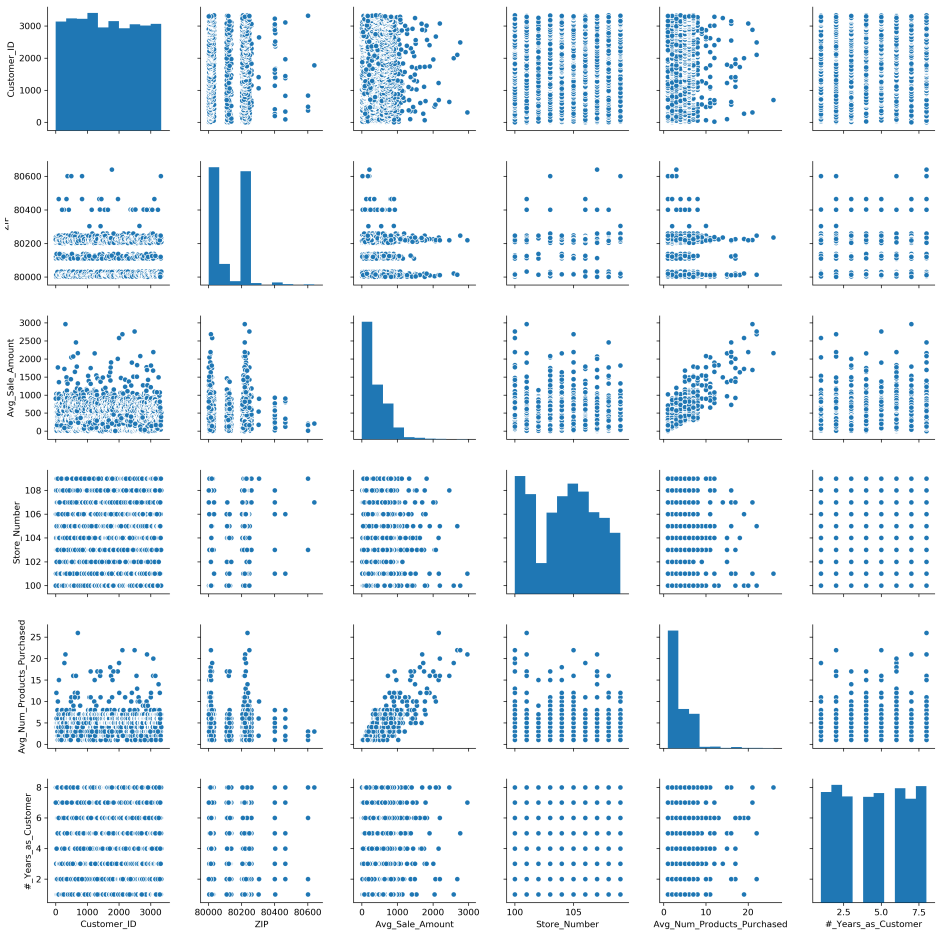
- The expected revenue from 250 new customers.
- The probability a custome will buy the catalog.
- Number of catalogs purchased.
- Categorical varibles converted to dummy variables

Analysis, Modeling, and Validation

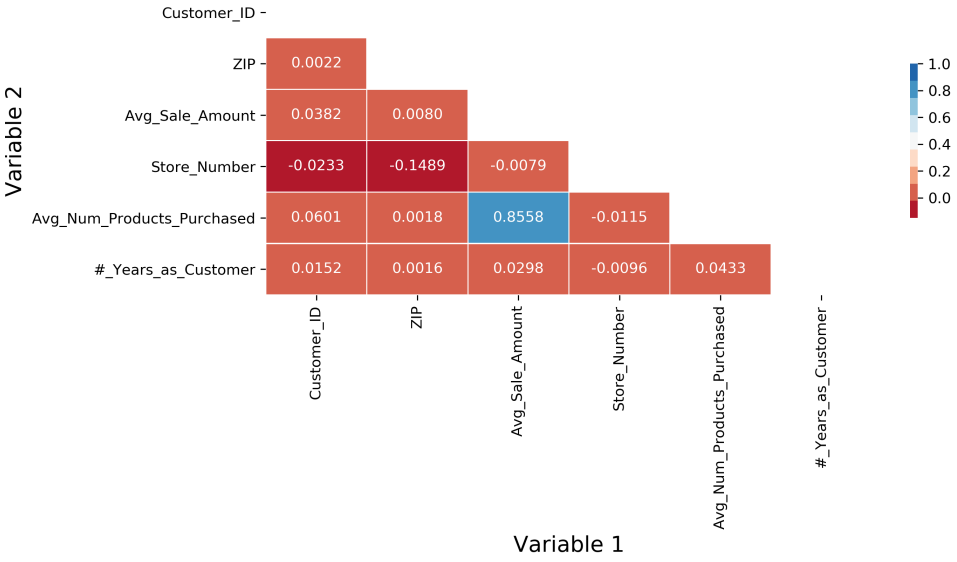
Question 1

How and why did you select the predictor variables in your model?

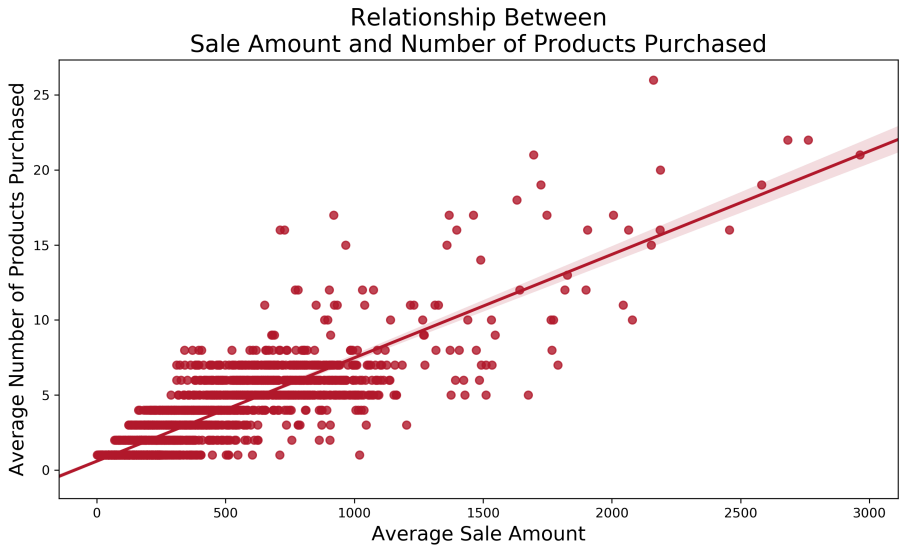
Using linear regression models I was able to assess which predictor variables have the strogest correlation.



Heatmap



The pair plot and Pearson Correlation matrix (above) suggests *Avg_Sale_Amount* and *Avg_Num_Products_Purchased* have a strong positive correlation of approximately 0.8558.



The scatter plot provides a cleaner display of the relationship between *Avg_Sale_Amount* and *Avg_Num_Products_Purchased* as a positive correlation.

Question 2

Explain why you believe your linear model is a good model.

OLS Regression Results

Dep. Variable:	Avg_Sale_Amount	**R-squared:**	0.837
Model:	OLS	**Adj. R-squared:**	0.837
Method:	Least Squares	**F-statistic:**	3040.
Date:	Thu, 03 Jan 2019	**Prob (F-statistic):**	0.00
Time:	21:22:38	**Log-Likelihood:**	-15061.
No. Observations:	2375	**AIC:**	3.013e+04
Df Residuals:	2370	**BIC:**	3.016e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	303.4635	10.576	28.694	0.000	282.725	324.202
Avg_Num_Products_Purchased	66.9762	1.515	44.208	0.000	64.005	69.947
Customer_Segment_Loyalty_Club_Only	-149.3557	8.973	-16.645	0.000	-166.951	-131.760
Customer_Segment_Loyalty_Club_and_Credit_Card	281.8388	11.910	23.664	0.000	258.484	305.194
Customer_Segment_Store_Mailing_List	-245.4177	9.768	-25.125	0.000	-264.572	-226.263

Omnibus:	359.638	**Durbin-Watson:**	2.045
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4770.580
Skew:	0.232	**Prob(JB):**	0.00
Kurtosis:	9.928	**Cond. No.**	25.0

First categorical variables were converted to dummy variables rather than assign random numbers to each category for the model. Thus preventing an erroneous relationship between the target variabel and the category variable(s) due to arbitrarily assigned value(s).

From there a simple regression analysis was conducted using the dummy variables. The Adjusted R-squared value is 0.837, indicating a strong positive correlation between my predictor variables (listed below):

- Avg_Num_Products_Purchased
- Customer_Segment_Loyalty_Club_Only
- Customer_Segment_Loyalty_Club_and_Credit_Card
- Customer_Segment_Store_Mailing_List

Note: The p-value is truncated after the third decimal place. Although, I cannot determine if p-value is exactly equal zero, the data suggests this a very small p-value (e.g. $p < 0.05$, and $p < 0.001$) which is statistically significant.

Question 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).

$$\begin{aligned} PredictedAverageSaleAmount \;=\; & 303.46 \;+ \\ & (66.98 \;\times\; AvgNumProductsPurchased\;) \;+ \\ & (\text{-}149.36 \;\times\; CustomerSegmentLoyaltyClubOnly\;) \;+ \\ & (281.84 \;\times\; CustomerSegmentLoyaltyClubAndCreditCard\;) \;+ \\ & (\text{-}245.42 \;\times\; CustomerSegmentStoreMailingList\;) \end{aligned}$$

Presentation/Visualization

Question 1

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

Yes. The company should send the catalog out to the 250 new customers in their mailing list.

Question 2

How did you come up with your recommendation?

First I used my multiple linear regression model to calculate the *Predicted_Average_Sale_Amount*, then I calculated *Predicted_Revenue*, and finally calculated *Predicted_Profit*. I placed these values for these variables in corresponding columns in the dataset labeled df_mailingList_dummies

(./assets/data/df_mailingList_dummies.csv). To find the overall value for these variables I calulcated the sum of their corresponding column within the dataset. The overall values were finally written to a csv titled df_overallValues (./assets/data/df_overallValues.csv).

Further information on how each vairable was calcualted is listed below, while the actual calculations reside in the Data Wrangling (./Data%20Wrangling.ipynb#three) file.

- *Predicted_Average_Sale_Amount* is calculated by following the *PredictedAverageSaleAmount* formula above. Substitute the formula's variables with corresponding column values for each of the 250 customers in the mailing list dataset. Finally sum the *Predicted_Average_Sale_Amount* values. Example formula below:
    - ```
      PredictedAverageSaleAmount = 303.46 +
                                   (66.98 × AvgNumProductsPurchased)
      +
                                   (- 149.36 × CustomerSegmentLoyalty
      ClubOnly) +
                                   (281.84 × CustomerSegmentLoyaltyCl
      ubAndCreditCard) +
                                   (- 245.42 × CustomerSegmentStoreMa
      ilingList)
      ```
- *Predicted_Revenue* is calculated by finding the product of *Average_Sale_Amount* and *Score_Yes* (the probability a customer will respond and make a purchase), then taking the sum of all those values. Example formula below:
    - ```
      Predicted_Revenue = Predicted_Average_Sale_Amount * Score_Yes
      ```
- *Predicted_Profit* is calculated by subtracting the catalog cost (given $6.50) from the product of *Predicted_Revenue* and average gross margin (which is a given value of 50%). Example formula below:
    - ```
      Predicted_Profit = (0.5 * Predicted_Revenue) - 6.5
      ```

Question 3

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

As shown in the image below, the *Overall Predicted Profit* is approximately `$21,987.96` . This is more than double the $10,000 breaking point.

	Variable Name	Values
0	Overall Predicted Average Sale Amount	$138,295.16
1	Overall Predicted Revenue	$47,225.91
2	Overall Predicted Profit	$21,987.96
      ```

In [ ]: