

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 need to be made?
 - The company wants to send catalog to its 200 new customers in the mailing list
 - The company needs to know the expected profit from the new customers before sending the catalogs
 - The company will send the catalogs only if the expected profit exceeds \$10,000 dollars
2. What data is needed to inform those decisions?

To predict the expected profit from the new customers, a Linear Regression Model needs to be built for Prediction.

The data needed are:

The data from Previous Customers' used to train the model

The data from the new customers (mailing list) to predict the expected profit from the new customers

We use the Customers' Segment and the Average Number of Products purchased to predict the Average Sale Amount for Each Customer.

Step 2: Analysis, Modeling, and Validation

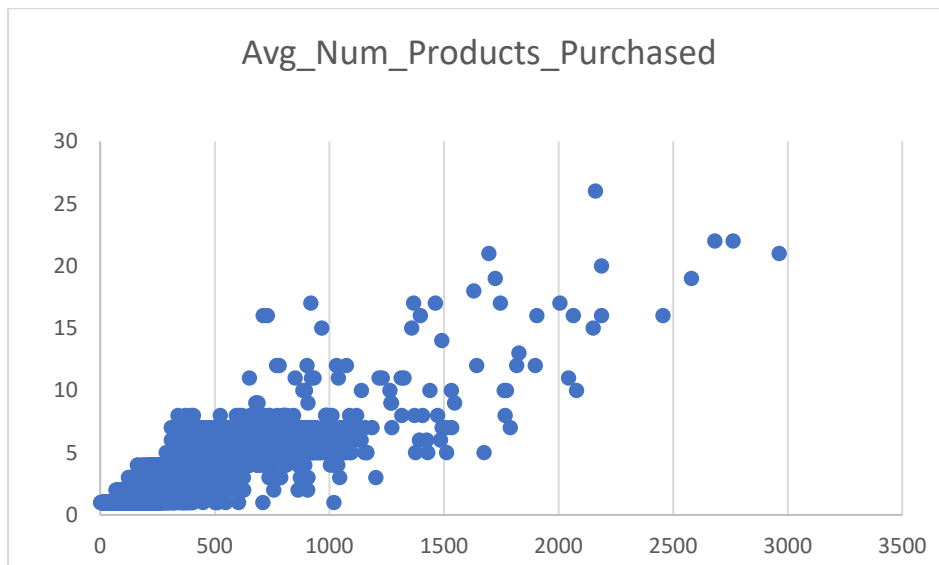
1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

The most important predictor variables that affect the Average Sale Amount are The Average Number of Products Purchased and Customer Segment.

Based on the Model Summary Below, all the continuous variables in the data except Avg_Num_Products_Purchased have large and insignificant P Values.

OLS Regression Results

Dep. Variable:	Avg_Sale_Amount		R-squared (uncentered):		0.888			
Model:	OLS		Adj. R-squared (uncentered):		0.887			
Method:	Least Squares		F-statistic:		3748.			
Date:	Sun, 18 Apr 2021		Prob (F-statistic):		0.00			
Time:	15:10:16		Log-Likelihood:		-15648.			
No. Observations:	2375		AIC:		3.131e+04			
Df Residuals:	2370		BIC:		3.133e+04			
Df Model:	5							
Covariance Type:	nonrobust							
			coef	std err	t	P> t	[0.025	0.975]
Customer_ID			-0.0047	0.004	-1.237	0.216	-0.012	0.003
ZIP			0.0005	0.002	0.304	0.761	-0.003	0.004
Store_Number			0.1507	1.263	0.119	0.905	-2.326	2.628
Avg_Num_Products_Purchased			106.4176	1.323	80.438	0.000	103.823	109.012
#_Years_as_Customer			-1.0550	1.566	-0.674	0.500	-4.125	2.015
Omnibus:	581.693	Durbin-Watson:	1.509					
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3666.291					
Skew:	1.002	Prob(JB):	0.00					



The Graphical Relationship between Avg_Num_Products_Purchased and Avg_Sale Amount in a Scatter Diagram.

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.

Dep. Variable:	Avg_Sale_Amount	R-squared (uncentered):	0.908
Model:	OLS	Adj. R-squared (uncentered):	0.908
Method:	Least Squares	F-statistic:	5832.
Date:	Sun, 18 Apr 2021	Prob (F-statistic):	0.00
Time:	14:41:34	Log-Likelihood:	-15415.
No. Observations:	2375	AIC:	3.084e+04
Df Residuals:	2371	BIC:	3.086e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Avg_Num_Products_Purchased	102.2377	1.028	99.415	0.000	100.221	104.254
Customer_Segment_Loyalty Club Only	26.5817	7.603	3.496	0.000	11.673	41.490
Customer_Segment_Loyalty Club and Credit Card	327.9296	13.696	23.944	0.000	301.073	354.786
Customer_Segment_Store Mailing List	5.7581	5.030	1.145	0.252	-4.106	15.622

Omnibus:	258.834	Durbin-Watson:	1.776
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2043.923
Skew:	-0.167	Prob(JB):	0.00
Kurtosis:	7.532	Cond. No.	18.5

The P Values for the 2 Predictor Variables chosen are small and less than 0.05 (Except Customer_Segment_Store Mailing List) and the R Square and Adjusted R Square values are 0.908 which means that our model explains the data pretty well.

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)

Avg_Sale_Amount=303.46 + 66.98*Avg_Num_Products_Purchased -
149.36*Customer_Segment_Loyalty Club Only + 281.84*Customer_Segment_Loyalty Club
and Credit Card - 245.42*Customer_Segment_Store Mailing List

Step 3: Presentation/Visualization

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

Based on the results from the Predicted Profits from the new 250 customers, the Company can send the catalogs to the 250 customers.

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

First of all, we calculate the expected revenue from the 250 customers to get the expected profit. This involves multiplying the probability that a person will buy our catalog. For Example, if there is a 30% chance that a person will buy a product from us, we would have an expected revenue of $\$450 \times 30\% = \135 .

The average gross margin is multiplied to the expected revenue and then the cost of printing and distribution (\$6.50) is subtracted from each profit price.

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

The expected profit from all customers is gotten by adding up the sum of the profit from all customers and it is given as approximately \$21987 dollars, which is more than the \$10,000 dollars indicated by the company. Hence, the catalog can be sent to the new customers.