

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?

Answer:

(1)Which parameter we should choose(we should choose parameters which can is positive correlation with the Avg Sale Amount).

(2)If we should send the 250 customers catlogs(we should create the linear model to calculate if the profit is great than 10,000 dollars).

2. What data is needed to inform those decisions?

Answer:

(1)The R^2 parameter.

(2)The profit.

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.

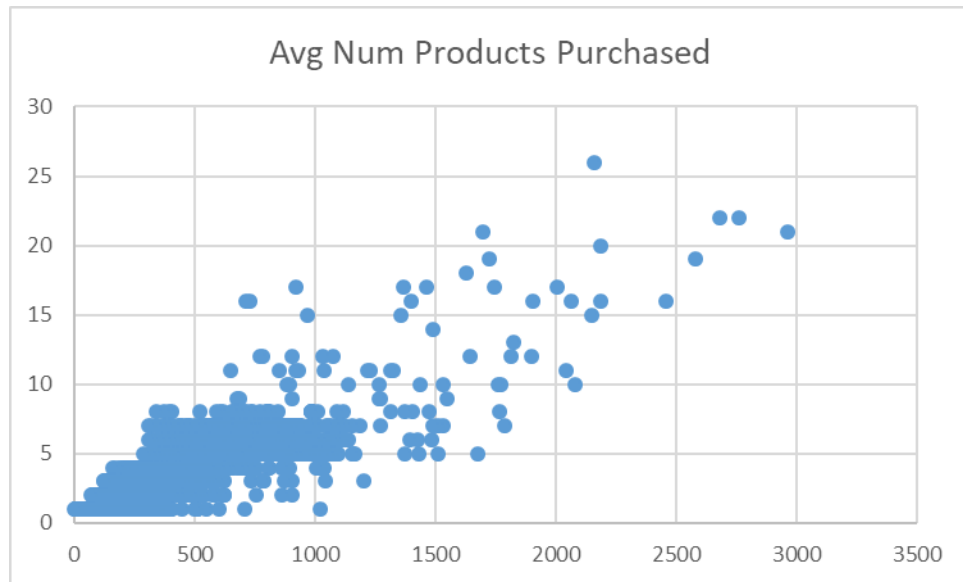
At the minimum, answer these questions:

1. How and why did you select the [predictor variables \(see supplementary text\)](#) 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.

Answer:

In order to choose the valid variables I use scatter plot.And we can learn from the graph below that it's true that there is positive correlation between Avg Num Products Purchased and Avg Sale Amount.



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.

Answer:

Using Excel's tool we can get the result:

SUMMARY OUTPUT								
回归统计								
Multiple F	0.855754							
R Square	0.732315							
Adjusted R	0.732202							
标准误差	176.0071							
观测值	2375							
方差分析								
	df	SS	MS	F	Significance F			
回归分析	1	2.01E+08	2.01E+08	6491.906	0			
残差	2373	73511948	30978.49					
总计	2374	2.75E+08						
Coefficients: 标准误差 t Stat P-value Lower 95% Upper 95% 下限 95.0% 上限 95.0%								
Intercept	44.01516	5.704323	7.716107	1.75E-14	32.82919	55.20114	32.82919	55.20114
X Variable	106.2802	1.319065	80.57237	0	103.6935	108.8668	103.6935	108.8668

When I just using Avg Num Products Purchased and Avg Sale Amount the p-value is 1.75E-14 and R-squared is 0.73.

SUMMARY OUTPUT								
回归统计								
Multiple F	0.91481							
R Square	0.836878							
Adjusted R	0.836602							
标准误差	137.4832							
观测值	2375							
方差分析								
	df	SS	MS	F	Significance F			
回归分析	4	2.3E+08	57456129	3039.744	0			
残差	2370	44796869	18901.63					
总计	2374	2.75E+08						
Coefficients		标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
Intercept	303.4635	10.57571	28.69437	1.1E-155	282.7249	324.2021	282.7249	324.2021
X Variable	66.9762	1.51504	44.20754	0	64.00526	69.94715	64.00526	69.94715
X Variable	-245.418	9.767776	-25.1252	1.1E-123	-264.572	-226.263	-264.572	-226.263
X Variable	-149.356	8.972755	-16.6455	6.35E-59	-166.951	-131.76	-166.951	-131.76
X Variable	281.8388	11.90986	23.66433	2.6E-111	258.4839	305.1936	258.4839	305.1936

As we can see in the figure above that after I add the Customer Segment the R Square update to 0.84 and the p-value get smaller. So I think it is a good fit.

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)

Answer:

The best linear regression equation is :

$$\begin{aligned}
 \text{Avg Sale Amount} = & 303.46 + 66.98 * \text{Avg Num Products Purchased} \\
 & - 245.42 * (\text{If Type: Store Mailing List}) \\
 & - 149.36 * (\text{If Type: Loyalty Club Only}) \\
 & + 281.84 * (\text{If Type: Loyalty Club and Credit Card}) \\
 & + 0 * (\text{If Type: Credit Card Only})
 \end{aligned}$$

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?

Answer:

My suggestion is that we should send the 250 customers catalogs.

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

Answer:

In all, the formula is :

$$\begin{aligned} \text{Avg Sale Amount} = & 303.46 + 66.98 * \text{Avg Num Products Purchased} \\ & - 245.42 * (\text{If Type: Store Mailing List}) \\ & - 149.36 * (\text{If Type: Loyalty Club Only}) \\ & + 281.84 * (\text{If Type: Loyalty Club and Credit Card}) \\ & + 0 * (\text{If Type: Credit Card Only}) \end{aligned}$$

$$\text{Profit} = \sum_{n=1}^{n=250} \text{Avg Sale Amount}(n) * \text{Score_Yes} * 0.5 - 6.5 * 250$$

At first, for each customer I consider x variable, and x variable is a vector which is consisted of $v1: (a, b, c, d)$. For example, the first customer A Giametti can be consider a vector $x1: (3, 0, 1, 0)$ which means he has 3 Avg Num Products Purchased and his segment is Loyalty Club Only.

a: Avg Num Products Purchased,

b: Store Mailing List,

c: Loyalty Club Only

d: Loyalty Club and Credit Card)

Secondly, I use the under formula to calculate the estimated income for each customer.

$$\text{Estimated Income} = \text{Avg Sale Amount} * \text{Score_Yes}$$

Thirdly, I sum up all 250 customers' estimated income.

$$\text{All Estimated Income} = \sum_{n=1}^{250} \text{Estimated Income}(n)$$

Then, calculate Gross Profit.

$$\text{Gross Profit} = \text{All Estimated Income} * 0.5$$

Lastly, calculate Profit.

$$\text{Profit} = \text{Gross Profit} - 6.5 * 250$$

And my result is:

All Estimated Income	47225.91406
GROSS PROFIT	23612.95703
PROFIT	21987.95703

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

Answer:

We can learn from the table above that the profit is 21987.96 dollars, so I hold this point of view that we should send the 250 customers catlogs.