

Project: Predicting Catalog Demand

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made.

The Business Problem

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.

Details

- The costs of printing and distributing is \$6.50 per catalog.
- The average gross margin (price - cost) on all products sold through the catalog is 50%.
- Make sure to multiply your revenue by the gross margin first before you subtract out the \$6.50 cost when calculating your profit

1. What decisions need to be made?

The decision to be made is whether to launch a catalog to 250 new customers for their high-end home goods product service. This decision is based on (1) expected profit calculated from these 250 new customers and (2) the company's ability to make a 50% margin and generate an expected profit exceeding its own \$10,000 contribution in order for catalog campaign to be successful.

2. What data is needed to inform those decisions?

Data needed to inform sales predictions and expected profits are: *Customer_Segment, Avg_Num_Products_Purchased, Store_Number, #_Years_as_Customer, Avg_Sale_Amount, Gross Margin, Cost of catalog printing & distribution to 250 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.

Important: Use the *p1-customers.xlsx* to train your linear model.

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 scatter plots to search for linear relationships. You must include scatterplots in your answer.

A linear regression study is performed on all variables against Avg_Sale_Amount, the target variable. The linear regression report shows that Avg_Num_Products_Purchased and Customer_Segment have a p-value of less than 0.05 which shows statistical significance. Therefore, these two were chosen as predictor variables. Variables such as address or geographical locations can be intuitively omitted because they don't impact future purchases.

Scatterplots of Avg_Num_Products_Purchased vs Avg_Sale_Amount and Customer_Segment vs Avg_Sale_Amount are detailed below.

Report					
Report for Linear Model Predictor_Variables					
Basic Summary					
Call: lm(formula = Avg_Sale_Amount ~ Customer_Segment + Store_Number + Responded_to_Last_Catalog + Avg_Num_Products_Purchased + X_Years_as_Customer, data = the.data)					
Residuals:					
	Min	1Q	Median	3Q	Max
	-665.19	-67.82	-2.17	70.42	975.25
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	435.318	104.854	4.152	3e-05	***
Customer_SegmentLoyalty Club Only	-150.224	8.971	-16.746	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.455	11.897	23.743	< 2.2e-16	***
Customer_SegmentStore Mailing List	-243.279	9.816	-24.784	< 2.2e-16	***
Store_Number	-1.146	0.994	-1.153	0.2489	
Responded_to_Last_CatalogYes	-28.085	11.253	-2.496	0.01264	*
Avg_Num_Products_Purchased	66.787	1.515	44.082	< 2.2e-16	***
X_Years_as_Customer	-2.326	1.222	-1.904	0.05707	.
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
Residual standard error: 137.25 on 2367 degrees of freedom					
Multiple R-squared: 0.8376, Adjusted R-Squared: 0.8372					
F-statistic: 1745 on 7 and 2367 degrees of freedom (DF), p-value < 2.2e-16					
Type II ANOVA Analysis					

Figure 1: Report for Linear Model Predictor_Variables

Scatterplot of Avg_Num_Products_Purchased versus Avg_Sale_Amount

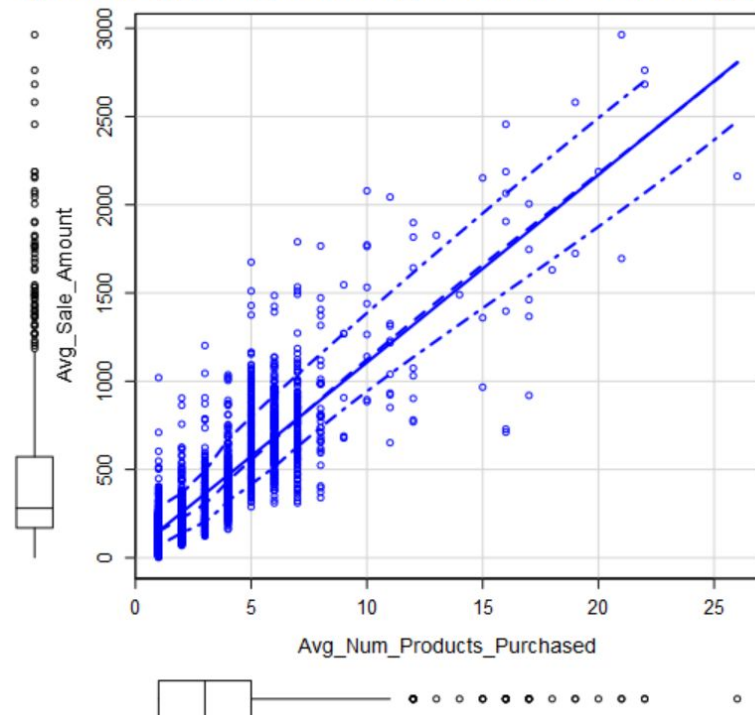


Figure 2: Scatterplots of Avg Number of Products Purchased vs Avg Sale Amount

Scatterplot of Avg_Sale_Amount versus Customer_Segm

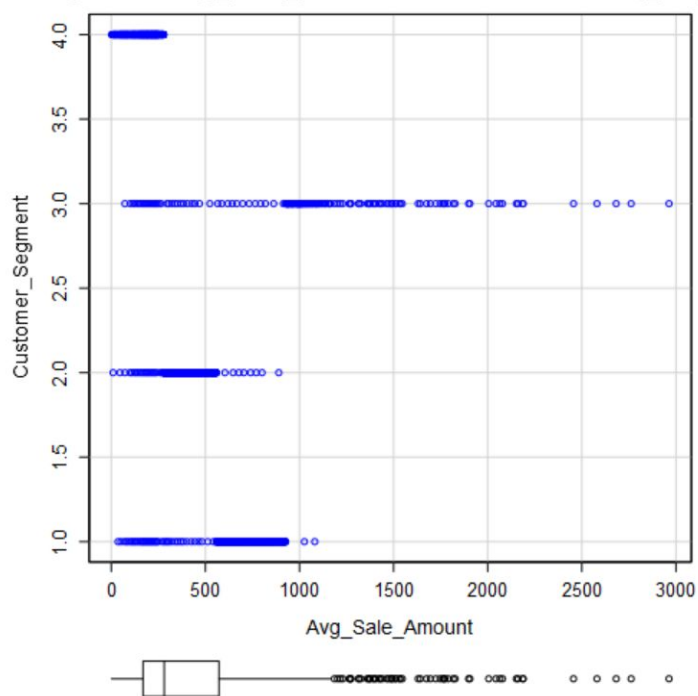


Figure 3: Scatterplot of Avg Sale Amount vs Customer Segment

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.

The linear model is a good model because our statistical result shows a multiple R-squared value of 0.8369 and adjusted R-squared value of 0.8366---both of which are high value and greater than 0.70. A high r-squared (0 to 1) indicates a greater explanatory power of the model and represents the amount of variation in the target variable associated with the variation in the predictor variables. Thus, a 0.8366 adjusted R-squared shows that 83.66% of the variance can be explained by the model.

It's also a good model since the p-value for Customer Segment and Average Number of Products is less than 0.05 showing their statistical significance. P-value indicates the probability that the coefficient is zero.

Report for Linear Model Sales_Predictor				
Basic Summary				
Call:				
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)				
Residuals:				
	Min	1Q	Median	3Q
	-663.8	-67.3	-1.9	70.7
				Max
				971.7
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				
Type II ANOVA Analysis				

Figure 4: Report for Linear Model Sales_Predictor

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 - 149.36 x (If Type: Customer_Segment Loyalty Club Only) + 281.84 x (If Type: Customer_Segment Loyalty Club and Credit Card) - 245.42 x (If Type: Customer_Segment Store Mailing List) + 0 x (If Type: Customer_Segment Credit Card Only) + 66.98 x (Avg_Num_Products_Purchased)

Step 3: Presentation/Visualization

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

The company should send the catalog to these 250 customers since the catalog is predicted to generate a profit of **\$21,987.44** after factoring the costs and margins---and exceeds the \$10,000 expected profit contribution. See calculation in question 3.

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

Note: Costs of printing and distributing is \$6.50 per catalog.

The average gross margin (price - cost) on all products sold through the catalog is 50%.

The expected revenue from each customer is calculated by multiplying the X field (predicted sales amount) with their respective Score_Yes value (probability of purchasing products). Summing the expected revenues for all 250 customers generates a total of \$47,224.87. Next, the sum of expected revenue for all 250 customers is multiplied by 0.50 to account for the 50% gross margin on all products sold through the catalog. Finally, the expected profit is calculated by subtracting (6.50*250) from the (summed expected revenue*gross margin). The expected profit of \$21,987.44 exceeds the company's \$10,000 contribution and thus, the company should send the catalog to the 250 customers.

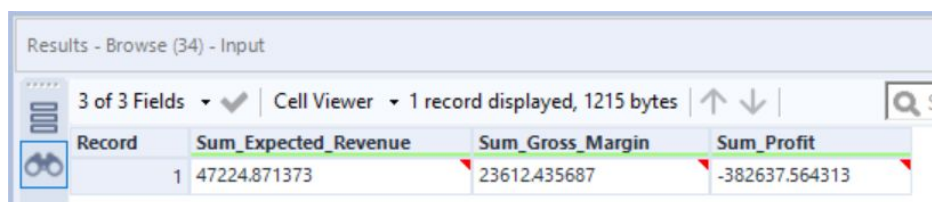
3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?

Expected Profit = (Sum of expected revenue x Gross Margin) – (Cost of Catalog x 250)

$$= (47,224.87 \times 0.5) - (6.50 \times 250)$$

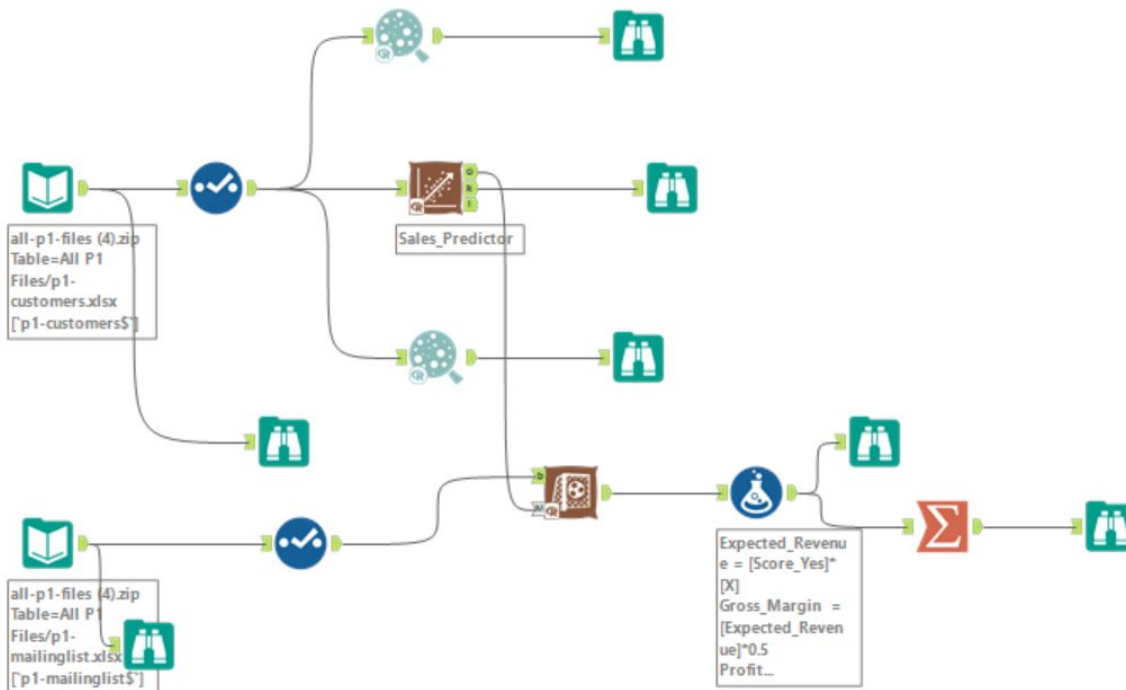
$$= 23,612.44 - 1,625$$

$$= \mathbf{\$21,987.44}$$



Record	Sum_Expected_Revenue	Sum_Gross_Margin	Sum_Profit
1	47224.871373	23612.435687	-382637.564313

Alteryx Workflow



Variable Distribution

Show the distributions for each variable in the Customer List dataset. How would these distributions affect your analysis? Would you need to go back to your manager and try to get more data?

Intuitively, variables such as name, state, customer ID, ZIP, and address are not great predictor variables since they do not have a tangible impact on predicting catalog sales.

These distributions can affect the analysis negatively since each variable mentioned above contains a value unique to each category or name---skewing the consistency of the data. Also, the p-values in our linear regression model have confirmed to be over 0.05 and would produce a bad model.

Data like customer behavior from other business deals can be obtained to analyze catalog sales.

