10/31/2018

Predicting catalog demand

Predictive Analytics for Business Project

Udacity Nanodegree

Tools: Alteryx

Katerina Bosko, PhD
WWW.CROSS-VALIDATED.COM

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%.

Write a short report with your recommendations outlining your reasons why the company should go with your recommendations to your manager.

Step 1: Business and Data Understanding

Key Decisions:

1. The company needs to decide whether to print and send the catalog to the 250 new customers in the mailing list. This decision to do so will be made only if the expected profits exceed $10,000.

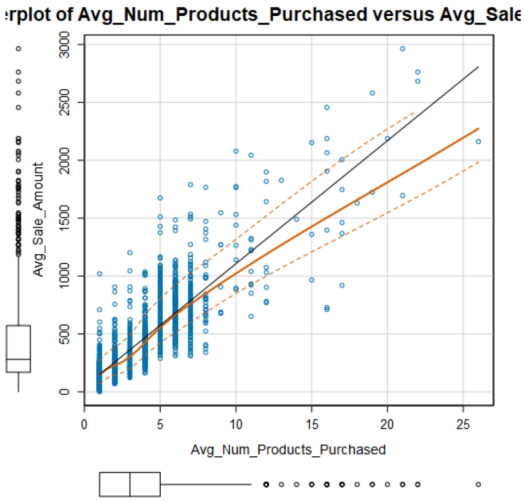
2. In order to inform the decision, we need data on previous sales with the same variables as in the current mailing list. We also need to know what are the costs to print the catalog and the average gross margin. The information from the previous year catalog sale would be also useful to estimate the chances of actually buying from the catalog.

Step 2: Analysis, Modeling, and Validation

1. The linear regression model was created in Alteryx and trained using data from p1-customers.xlsx. The target variable was set to **Avg_Sale_Amount.** The predictor variables were chosen based on the following procedure.

In the first step, the model was run with several potential predictors in order to select the most statistically significant variables. Such variables as Name, Customer_ID, Address, State were excluded from the analysis, because it either makes no sense to test them or they have not enough variability like State (there's only one state in the database).

This initial analysis showed that there are only two highly statistically significant predictors of the sales – **Customer_Segment** and **Avg_Num_Products_Purchased**. Because Customer_Segment is a categorical variable, it wasn't checked for the appropriateness of the use in a linear regression model. However, Avg_Num_Products_Purchased is continuous numerical variable, so it was checked to see whether there is a linear relationship with the target variable (see the scatterplot below). Since the sales amount tend to increase together with the increase of the number of products purchased, there is a linear relationship between the two variables.



2. According to the results reported in Alteryx, the **R-squared** value is 0.8369 and the **adjusted R-squared** value is 0.8366, indicating that it is a good model (>0.5) and the chosen predictors have high explanatory power.

The used predictors are both highly statistically significant with **p-values** < 2.2e-16, meaning that it is highly probable that the relationship between the target variable and the predictors actually exists.

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	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

3. The final linear regression model that was applied to forecast sales of the 250 customers is as follows:

*Predicted sales = 303.46 + 66.98 * Avg_Num_Products_Purchased – 149.36 (If Customer_Segment: Loyalty Club Only) + 281.84 * (If Customer_Segment: Loyalty Club and Credit Card) – 245.42 * (If Customer_Segment: Store Mailing List) + 0 * (If Customer_Segement: Credit Card Only)*

Step 3: Presentation/Visualization

1. I recommend sending the catalog to the chosen 250 customers.
2. The recommendation is based upon results of the analysis done according to specifications provided by the managers.

   First, the sales that we predicted using the linear regression model were multiplied by the probability that the customer buys the catalog

   the Alteryx formula: [sales_predicted][Score_Yes]*

   Second, the expected revenues were then summed up and multiplied by the gross margin specified at 50%. Third, the expected profits were calculated by subtracting the cost to print the catalog for the 250 customers.

   *the Alteryx formula: [Sum_exp_revenue]*0.5-(6.5*250)*

3. The expected profit from sending the catalog to the 250 customers is about $22,000, which is twice higher than the sum specified by managers in order to make this decision ($10,000). Hence, the chances to make good profits are high even if the sales turn out to be worse than prognosed by the analysis.