Predicting Catalog Demand

Should we send catalogs out to our 250 new customers?

Contents

Reporting for Udacity Support Team:

Ebrahim Tayara

03/05/2018

Overview

A company that manufactures and sells high-end home goods sent out its first print catalog last year 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.

My manager has been asked to determine how much profit the company can expect from sending a catalog to these customers.

Problem

We need to predict the expected profit from these new customers. Management does not want to send the catalog out unless the expected profit contribution exceeds $10,000.

Approach

What we know:

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%.

We're going to analyze two datasets to calculate expected profits using a predictive model. Clean historical data was provided for this project, so no data preparation was necessary. When running the calculation, I'm going to multiply our expected profit by the gross margin before subtracting out the $6.50 cost to get our actual revenue.

Data Understanding

Looking at our dataset, we want to figure out if there's a relation between discrete numeric variables like the average number of products or number of years as a customer and our only continuous numeric variable (average amount in $ spent per customer). We can also take categorical variables like a customer's segment into consideration. It would make sense that a customer who has a credit card or is part of the loyalty club would spend more than a customer who doesn't or isn't. Name, address and customer ID are identifiers and wouldn't factor in.

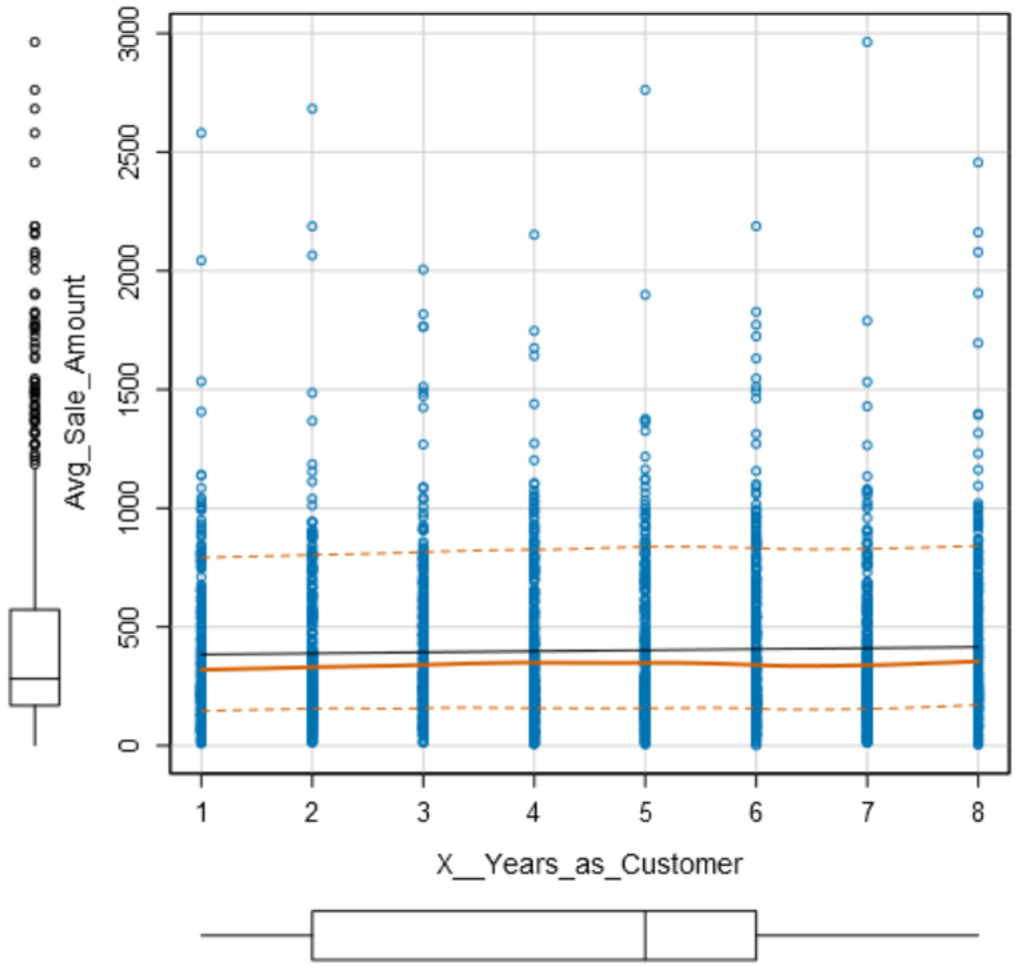
Data Analysis

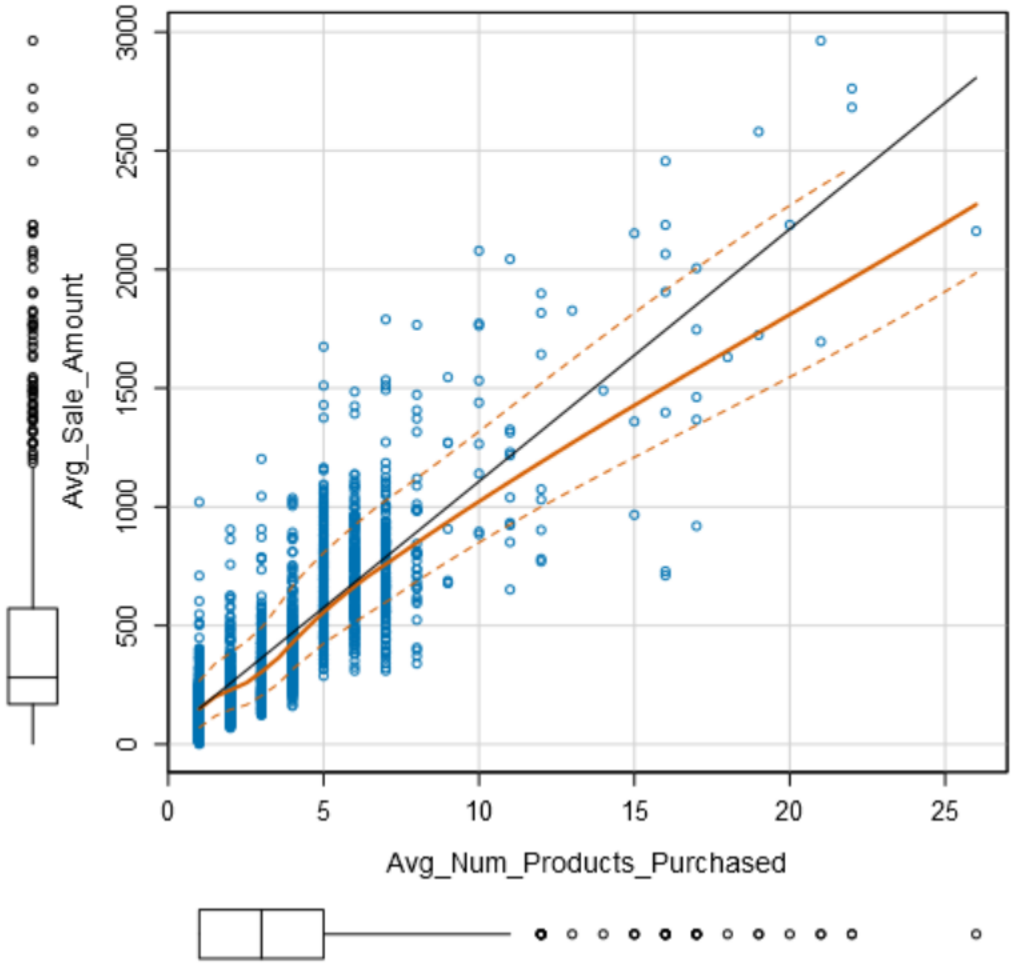
We'll use Alteryx to help us determine if our predictor variables have a linear relationship with our target variable. We want to build our initial model by using the customer data before scoring the mailing list and applying our linear regression formula.

Note: Alteryx automatically transforms our categorical variables like the customer segment into dummy variables. Although it's not a numeric value, this allows us to use that data as an additional predictor.

Data Modeling

1) Tested whether there's a linear relationship between my x (predictor) and y (target) variables. Please see a couple examples below:





2) Accounted for Customer_Segment by changing the data type to a forced V string. Connected a linear regression tool, ran the report and found four of my predictors to be statistically significant.

```
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data =
inputs$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 DF, p-value: < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28715078.96	3	506.4	< 2.2e-16 ***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16 ***
Residuals	44796869.07	2370		

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Equation (abbreviated x values):

Expected Profit = 303.46-149.36*(CSLC)+281.84*(CSLC&CC)-245.42*(CSML)+66.98*(ANPP)

Data Validation

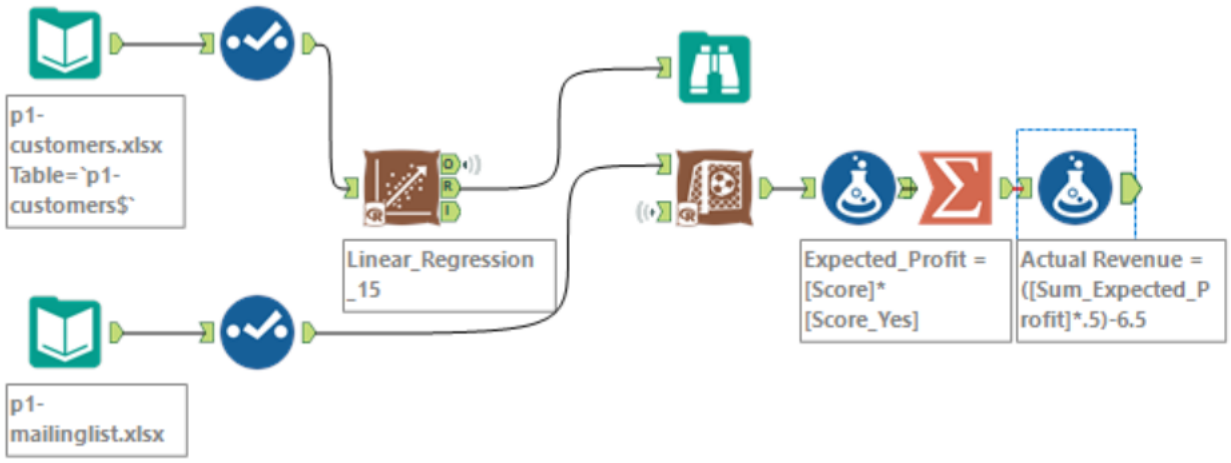
We know our model is good because our adjusted R-squared value is 0.84. R-squared values can range from -1 to 1. The closer the number is to 1, the better the model.

Another way to validate our model would be to look at our p-value. The lower it is, the better the model. Our p-value is sitting at a pretty < .00000000000000022, which is basically 0.

Finally, Alteryx has a unique way of telling us whether the variables used to create our equation are significant with the * symbol. As we can see, we have *** next to ours. Our model is awesome.

Workflow

Scored the model and applied the formula discussed in our approach to calculate our expected profit.



Results - Formula (41) - Output

2 of 2 Fields | Cell Viewer | 1 record displayed

Record #	Sum_Expected_Profit	Actual Revenue
1	47224.871373	23605.9356865455

My Recommendation

I would recommend sending out the catalogs because we'd be making an estimated $23,605.94.

Sources:

Compiled from data/hints obtained from Udacity and tips from my mentor Karan.

https://discussions.udacity.com/c/nd008-bizand-problem-solving/project-1-2

https://udacity-pand.slack.com/messages/C1P2J8QF3/

https://community.alteryx.com/

Update History:

1. Created – 05-March-2018.
2. Updated – 06-March-2018.
3. Updated – 07-March-2018.