Kacper Ksieski

# Project 1: Predicting Catalog Demand

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/c0b53068-1239-4f01-82bf-24886872f48e/project

## 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?
   The company needs to decide whether to send a print catalogue to 250 new customers or not. The company will only undertake the investment if the predicted profit is $10 000. This is the amount after COGS and the cost of the print catalogue. Each catalogue costs $6.50 * 250 customers = $1625. The profit of $10 000 + $1625 cost of the catalogues = $11 625. $11 625 / 50% gross margin = $23 250. The company will make this investment if the gross value is greater than or equal to $23 250.

2. What data is needed to inform those decisions?
   The decision rests on the amount of predicted profit this endeavour will yield. The predicted profit is the sum of the expected value per customer. The expected value is the amount each customer will buy multiplied by the probability that the customer will buy.

## 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:*

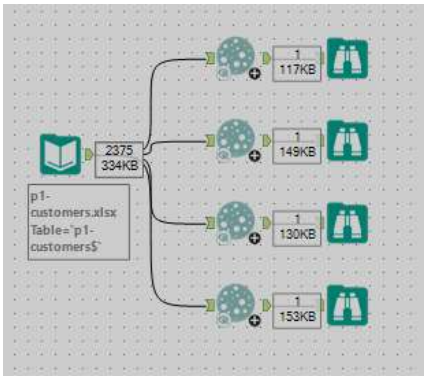
1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.
   I used Alteryx to chart the target variable, Average Sale Amount, with other potential predictor variables using the Scatterplot tool.

: Awesome: Indeed, that is the main decision that the company needs to make.

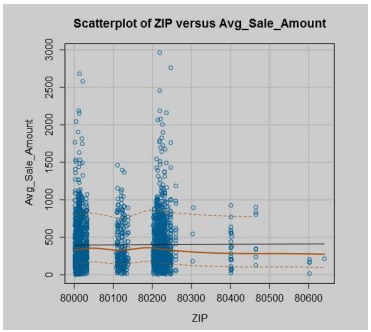
: Suggestion: Nice work! Indeed we need all of the data listed here. This answer could be improved a little bit. Besides the data listed here we also need data on all the cost of printing the catalog. We could also expand this answer to include data on all of the customers and any data that can tell us whether they've bought something in the catalogue in the past, including not limited to:
a. Bought an item from a past catalogue
b. Average amount of items the customer buys from the company
c. The total dollar amount that the customer spent ordering from our catalogues

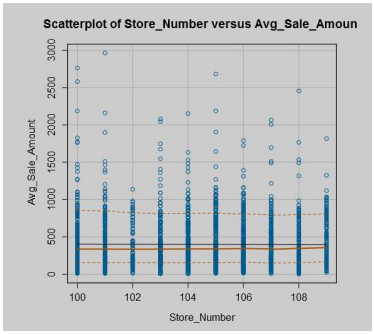
I am leaving this just as a suggestion, and I am not going to make you resubmit  for this since I think you have a very good grasp of the main concepts presented in the project and you did a great job in general.

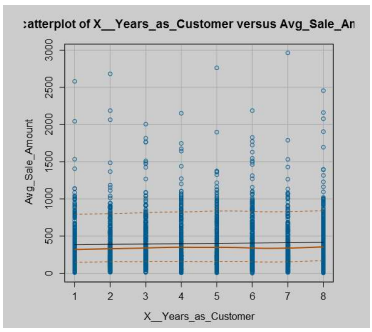
Kacper Ksieski



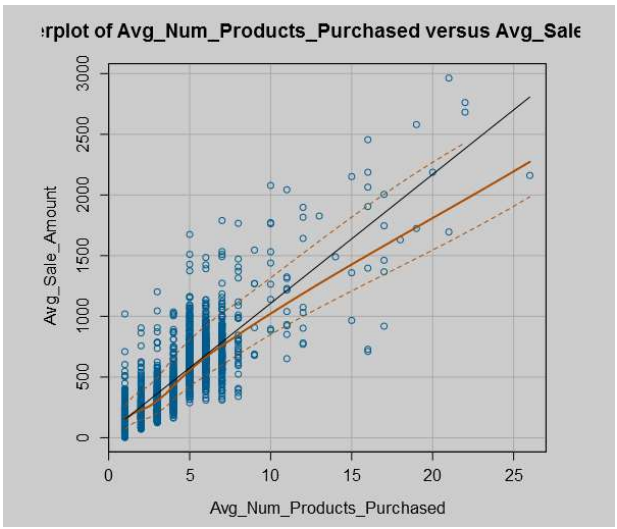

I examined the plots to look for a correlation. The first three that I chose, ZIP, Store Number, and Number of Years as a customer, yielded the following charts which clearly show no correlation.




Only the last variable, Avg.Num.Products.Purchased, had a correlation with the target variable.




Then I explored the Customer Segment categorical variable to check if it was appropriate to include in my model. Because it is not a binary categorical variable, I couldn't chart it. Instead, I checked this variable in the Linear Regression tool and examined its significance code.

Kacper Ksieski

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.
Alteryx assigns a significance code based upon where a P value falls from 0 to 1. The Customer Segment variable earned ***, a P value between 0 and 0.001 . With Customer Segment and Avg.Num.Products.Purchased included in the model, the R squared value is 0.837. This contrasts with the 0.732 when Customer Segment is excluded from the model. As evidenced in the prior exercise, Avg.Num.Products.Purchased shows a correlation with the target variable. It earns ***, a P value between 0 and 0.001.



The model is a good fit of the available data because the R squared is higher than 0.75 and the P values of each of the predictor variables are significant.

| | 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 | *** |

I checked the distribution of the data across each customer segment. Loyalty club and credit card doesn't have as much data as store mailing list. I would like to ask the manager for more data. More data may confirm my model's accuracy, change the coefficients of the predictor variables, or even cause me to reevaluate my model completely.

| Customer Segment | Count |
|---|---|
| Credit Card Only | 494 |
| Loyalty Club Only | 579 |
| Loyalty Club and Credit Card | 194 |
| Store Mailing List | 1108 |

: Suggestion: We found out that the Adjusted R-squared is about 84% with all variables being significant at p-values less than 0.05. It would be nice to include some discussion of what these percentages mean. For example, we could say that - "The model can account for 84% of the actual sales amounts in my training set."

This is a [nice article](http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit) on how to interpret R-squared.

: Awesome: Excellent work!

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)

**Important: The regression equation should be in the form:**

*Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

**For example:** Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

Note that we **must** include the 0 coefficient for the type Cash.

Y = 303.46 + 66.98 Avg.Num.Products.Purchased – 149.36 (If type: Loyalty Club Only)
+ 281.84 (If type: Loyalty Club and Credit Card) – 245.42 (If type: Store Mailing List)
+ 0 (If type: Credit Card Only)

: Awesome: Excellent! The regression model is correct!

Coefficients:

| | Estimate |
|---|---|
| (Intercept) | 303.46 |
| Customer.SegmentLoyalty Club Only | -149.36 |
| Customer.SegmentLoyalty Club and Credit Card | 281.84 |
| Customer.SegmentStore Mailing List | -245.42 |
| Avg.Num.Products.Purchased | 66.98 |

**Note**: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

## Step 3: Presentation/Visualization

*Use your model results to provide a recommendation. (500 word limit)*

*At the minimum, answer these questions:*

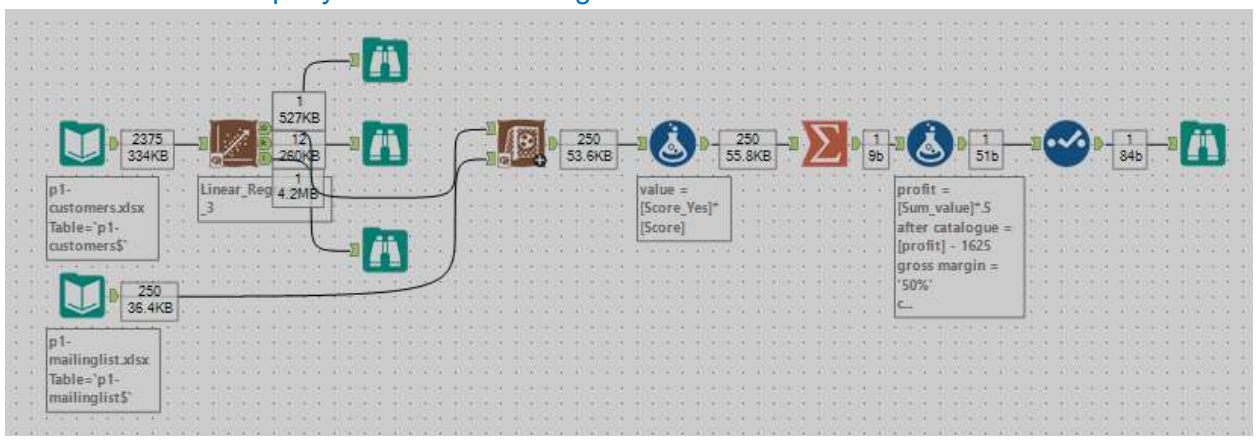
1. What is your recommendation? Should the company send the catalog to these 250 customers?
   I recommend the company sends the catalogues to these 250 customers.

   : Awesome: The recommendation made is correct - well done!

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
   After I constructed my linear regression model, I used the output as an input to the Score tool. The Score tool's right input was the output of the 250 customers files. The Score tool yielded a score, the average sales amount. I used a Formula tool to multiply this score by the score_yes field, which is the probability that the customer will buy. This product was the expected value per customer. I used the Summarise tool to sum the expected values and the result was $47 224.87. I used another Formula tool to multiply the sum by 50% to get the gross profit, $23 612.44. Since each catalogue costs $6.50 and there are 250 catalogues, one per customer, that is a total cost of $1625. Gross

   : Awesome: Very thorough explanation of the process - great job!

profit of $23 612.44 less $1625 is $21 987.44. This is above the $10 000 threshold to recommend the company to send the catalogues.




| revenue | gross margin | gross profit | catalogue cost | profit |
|---|---|---|---|---|
| 47224.87 | 50% | 23612.44 | 1625 | 21987.44 |

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
   The expected profit is $21 987.44.

: Awesome: The expected profit is correct - well done!

| revenue | gross margin | gross profit | catalogue cost | profit |
|---|---|---|---|---|
| 47224.87 | 50% | 23612.44 | 1625 | 21987.44 |

## Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here.
Reviewers will use this rubric to grade your project.