Project 1: Predicting Catalog Demand

São Paulo, 22 February of 2019
Felipe Mahlmeister

Step 1: Business and Data Understanding

Key Decisions:

1. **What decisions needs to be made?**

   The company that I'm working for needs to know if worth sending out print catalogs to their 250 new customers and what will be the expected profit.

   Is important to highlight that the expected profit should be at least $ 10,000.00, otherwise, they'll not make this decision.

2. **What data is needed to inform those decisions?**

   To the business analyst be able to inform the expected profit, the company must provide the registration data of the current and the new customers.

   Both registrations must have the customer segments, the average number of the products purchased and for the current customers must also have the average sale amount spent with their products, so we could predict the amount of money the new customers will spend, and thereafter the profit.

Step 2: Analysis, Modeling, and Validation

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.**

   After importing raw data of the current customers, *Select* tool was applied to guarantee that the data types has correct and also *Data Cleaning* tool was applied to remove nulls.

   For Numeric variables I used scatterplots between the individual variable and the target variable to see if a variable might be good candidate for predictor variable. The following plots was made in Alteryx using Scatterplot tool.
   The target variable (Avg Sale Amount) has put as my Y and the numeric predictor variable as X (Customer ID, ZIP, Store_Number, Avg_Num_Products_Purchased, Years_as_Customer).

   Please see below 5 Scatterplots with the numeric predictor variables candidates and a brief explanation about its relationship with target variable.

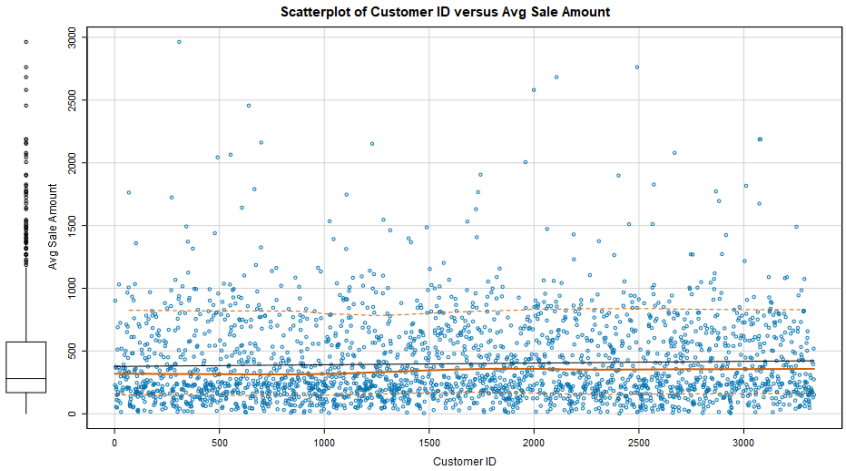

Figure 1 - Scatterplot of Customer ID versus Avg Sale Amount

Figures 1, 2, 3 and 5 indicates by its absence of the slope, that the X variable (Customer ID, ZIP, Store Number, Years as Customer) hasn't shown any relationship with the target variable, which excludes it as a potential candidate to be a predictor variable for the target variable in this model.

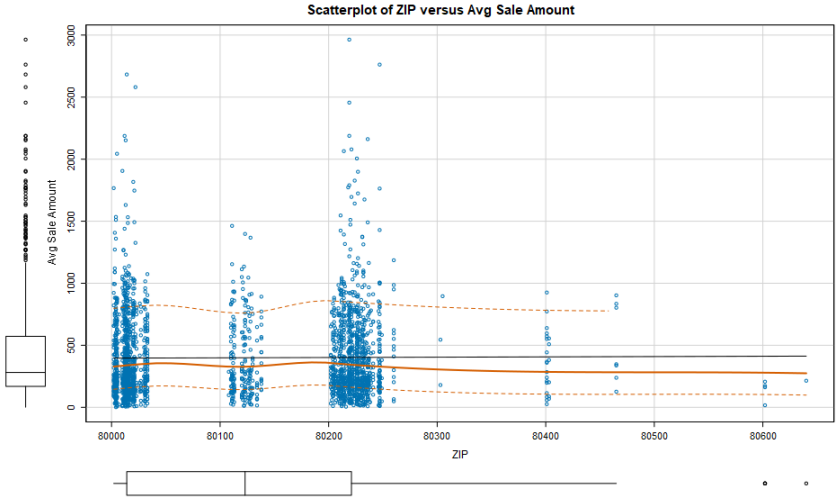

Figure 2 - Scatterplot of ZIP versus Avg Sale Amount

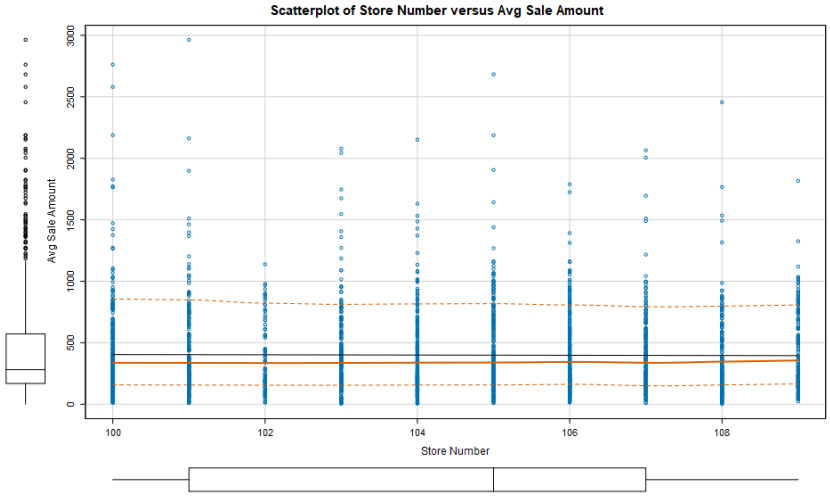

Figure 3 - Scatterplot of Store Number versus Avg Sale Amount

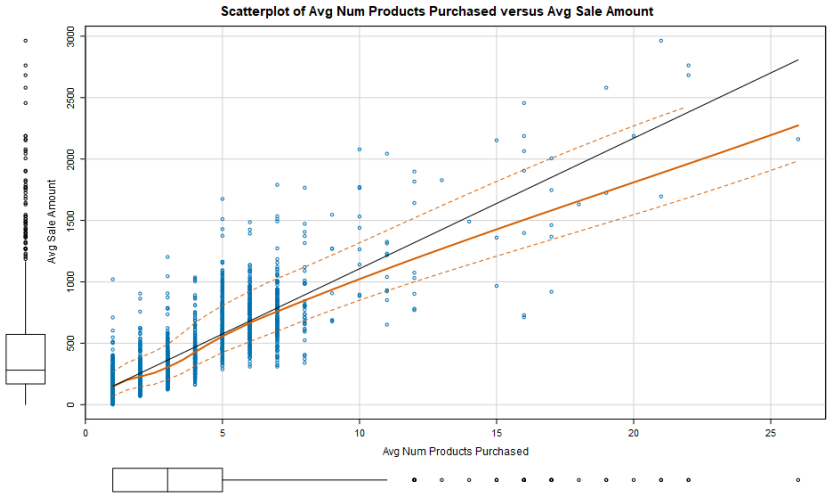

Figure 4 - Scatterplot of Avg Num Products Purchased versus Avg Sale Amount

A sloped line could be seen in Figure 4, which might indicate that this Numeric variable (Average Number of Products) is a good predictor variable for the target variable (Average Sale Amount). This sloped line in this scatter plot indicates a positive trend, which means that as the X values increases, the Y values also increase, indicating its relationship between both variables.

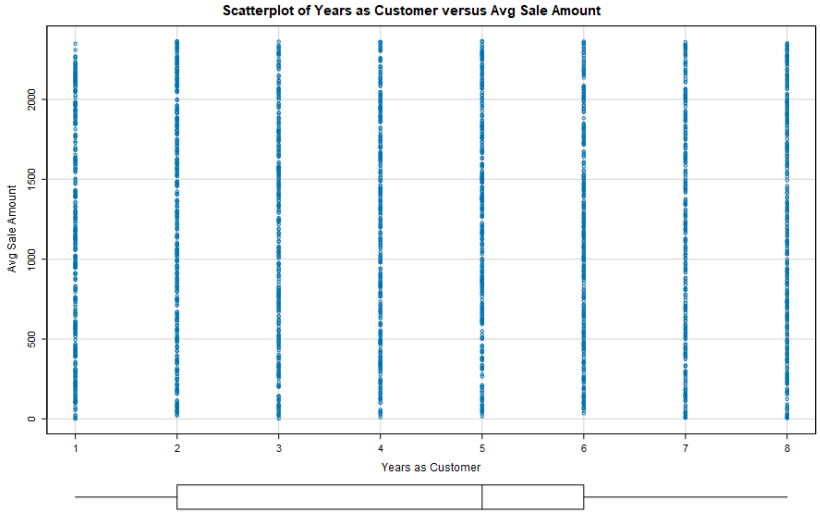

Figure 5 - Scatterplot of Years as Customer versus Avg Sale Amount

For categorical variables I had to use trial and error to see which are statistically significant.
But after making some assumptions wasn't necessary to look for all categorical variables:

- The variable couldn't have a hundred of unique values, which would categorize it as a dispensable variable because of its large amount of possibilities and fewer relevancy of each ones – Variables Name and Address was classified in this category;
- The variable needs to have more than one value – Variable State was classified in this category;

Leaving us with these categorical variables with potential to be predictor variables:

- Customer Segment;
- City;

The following table (Table 1) shows the Linear Regression Report from Alteryx and illustrate a simulation with the Numeric variables (already chosen) and with both Categorical variables (see above) which could be potential predictor variables:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	490.9834	123.042	3.99038	7e-05	***
Customer_SegmentLoyalty Club Only	-149,8780	9,020	-16,61610	< 2,2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.9160	11.971	23.63433	< 2,2e-16	***
Customer_SegmentStore Mailing List	-245,5024	9,846	-24,93384	< 2,2e-16	***
CityAurora	-19.9592	11.097	-1.79862	0.07221	.
CityBoulder	-40,3075	80,127	-0,50304	0,61498	
CityBrighton	-70,4394	97,670	-0,72120	0,47086	
CityBroomfield	-3.8360	15.142	-0.25334	0.80003	
CityCastle Pines	-91,0506	97,760	-0,93137	0,35176	
CityCentennial	-9,9043	18,180	-0,54479	0,58595	
CityCommerce City	-35.5183	44.500	-0.79816	0.42486	
CityDenver	-0.4258	10.561	-0.04032	0.96784	
CityEdgewater	29,6501	40,657	0,72928	0,46591	
CityEnglewood	5,0515	20,760	0,24333	0,80777	
CityGolden	-12.8415	32.755	-0.39205	0.69506	
CityGreenwood Village	-50,0735	38,045	-1,31616	0,18825	
CityHenderson	-284,9016	138,017	-2,06425	0,0391	*
CityHighlands Ranch	-27.7328	30.457	-0.91055	0.36263	
CityLafayette	-47.4442	62.200	-0.76277	0.44568	
CityLakewood	-7,9770	12,872	-0,61974	0,53549	
CityLittleton	-28,8015	18.991	-1.51663	0.1295	
CityLone Tree	77.5542	137.939	0.56223	0.57401	
CityLouisville	-28,6488	69,348	-0,41312	0,67956	
CityMorrison	-17,2654	52,851	-0,32668	0,74394	
CityNorthglenn	-15.2884	29.428	-0.51952	0.60345	
CityParker	-6,0407	28,212	-0,21411	0,83048	
CitySuperior	-53,5206	46,728	-1,14535	0,25218	
CityThornton	28.5171	24.843	1.14787	0.25114	
CityWestminster	-6.8925	17.305	-0.39829	0.69045	
CityWheat Ridge	7,1755	20,687	0,34685	0,72873	
Store_Number	-1.6275	1.148	-1.41823	0.15626	
Avg_Num_Products_Purchased	67.1063	1.527	43.93513	< 2.2e-16	***
Years_as_Customer	-2,3582	1,232	-1,91353	0,0558	.

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

Residual standard error: 137,5 on 2342 degrees of freedom
Multiple R-squared: 0.8388, Adjusted R-Squared: 0.8366

Table 1 - Linear Regression Report

As we are looking for P-values at least statistically significant (P-value less or equal to 0.05) of the categorical variables, City was discarted as a potencial to be predictor variable.

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.**

   To consider this model as a good model, we are looking for statistical significance, P-value less or equal to 0.05 and high R-square values are desirable.

   All the variable chosen are statically significant and the R-squared value is close, but below 0.9, which is a good sign that this model isn't very fit and could be improved, but for this case, we will accept it and go forward.

	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. **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 * (If CS_Loyalty_Club_Only) + 281.84 * (If CS _Loyalty_Club_and_Credit_Card) – 245.42 * (If CS _Store_Mailling_List) + 0 * (If CS _Credit_Card_Only) + 66.98 * Avg_Num_Products_Purchased

Step 3: Presentation/Visualization

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

   Based on all the steps taken, I recommend the sending of these 250 catalogs to the new customers.

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

   After modeling the linear regression (statistically relevant) with the data of current customers, we could predict what would be the predicted revenue, as well as the profit of sending the catalogs to these 250 new customers, as could be seen on the equation below (applied for each new customer):

   Profit = (([Predicted_Value]*[Score_Yes])*0.5) - 6.5

   Applying this equation to the 250 new customers we get a total profit of $ 21.987,44, which is 55% greater than expected by the management.

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

   Applying this equation to the 250 new customers we get a total profit of $ 21.987,44, which is 55% greater than expected by the management.