Enrique Franco

Project 1: Predicting Catalog Demand

Step 1: Business and Data Understanding

Provide an explanation of the key decisions that need to be made. (500 word limit)

Key Decisions:

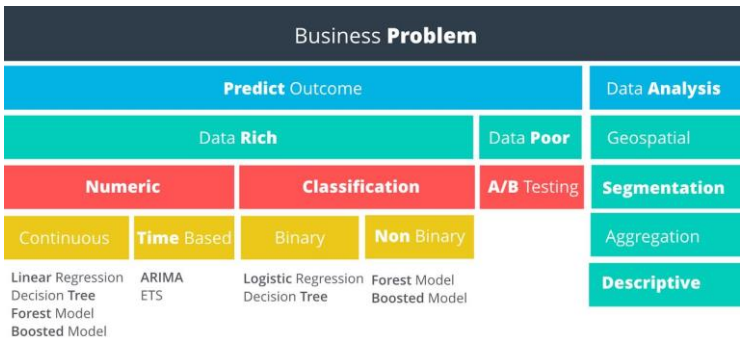
Answer these questions

1. What decisions needs to be made?

There needs to be an analysis on which decisions must be made. We use the business problem chart tool in order to understand better our decisions.

Our business problem requires to predict outcome and it is data rich, so we have to decide which type of model we are going to use. Since the data given and the outcome are both numeric, where there is no time-dependent relevant value, then it is obvious that our model must be a continuous Linear Regression predictive model, in order to predict the possible profit for the company.

The company must thrive to construct this model with data relevant to build the lineal regression model, in order to predict the profit that is going to generate through the new catalog.



The final decision as to be made, if the profit reaches 10,000$ or more, the project is viable. If the prediction is less than this amount, then it is not viable.

Figure 1: Business Problem Flow Chart
Source: Udacity.com (2018)

2. What data is needed to inform those decisions?

We need data from the past, where there is sufficient numeric relevance in order to construct our predictive linear regression model. Our target variable is profit, but one

must know which predictor variables affect our target variable through the data collected. With this data we construct our model and apply it to a new set of data, that will give a profit prediction outcome. We conclude the following:

1. We need data from past profit in order to construct our model.
2. We need present data that lets us apply the model and predict profit.

One shall know that the profit is given when there are expenses taken in account. There is a 50% gross margin which also takes account the price/cost on products sold through the catalog and taking in account that each catalog costs around 6.50$ each.

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.

In order to choose the predictor variables, one must find a relationship with possible predictor variables with the target variable. The predictor variable however, as stated before, is profit, but there is no variable named in the data set that is named profit, but average sale amount. If one operates the average sale amount at the end, one must get the profit, so our target variable in the predictive model is now average sale amount.

In this case, one shall elaborate a scatter plot graph with each possible predictor variable with the target variable in order to see if there is any linear relationship between them. The graph must be done with each one of the possible predictor variable vs. profit, but a faster and simple way using alteryx is provided. While producing a linear regression model there is the "p" factor that shows how the predictor variable and the target variable may have a relationship. In order to show less graphs in this document, one will choose one predictor vs. target variable that shows statistical significance and one that does not. We insert the past data into the "linear regression" function and obtain the following statistical significance values:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	4.945e+02	1.233e+02	4.0111	6e-05	***
Customer_SegmentLoyalty Club Only	-1.500e+02	9.024e+00	-16.6199	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	2.826e+02	1.199e+01	23.5683	< 2.2e-16	***
Customer_SegmentStore Mailing List	-2.456e+02	9.850e+00	-24.9344	< 2.2e-16	***
Customer_ID	-1.429e-03	2.957e-03	-0.4830	0.62911	
CityAurora	-2.000e+01	1.110e+01	-1.8018	0.0717	.
CityBoulder	-4.024e+01	8.014e+01	-0.5022	0.61559	
CityBrighton	-7.004e+01	9.769e+01	-0.7170	0.47344	
CityBroomfield	-3.687e+00	1.515e+01	-0.2434	0.8077	
CityCastle Pines	-9.243e+01	9.782e+01	-0.9449	0.34481	
CityCentennial	-9.752e+00	1.819e+01	-0.5362	0.59186	
CityCommerce City	-3.542e+01	4.451e+01	-0.7959	0.42618	
CityDenver	-3.866e-01	1.056e+01	-0.0366	0.97081	
CityEdgewater	2.999e+01	4.067e+01	0.7375	0.46088	
CityEnglewood	5.162e+00	2.076e+01	0.2486	0.80368	
CityGolden	-1.254e+01	3.277e+01	-0.3827	0.70197	
CityGreenwood Village	-5.019e+01	3.805e+01	-1.3191	0.18727	
CityHenderson	-2.848e+02	1.380e+02	-2.0630	0.03922	*
CityHighlands Ranch	-2.729e+01	3.048e+01	-0.8954	0.37067	
CityLafayette	-4.581e+01	6.230e+01	-0.7353	0.46222	
CityLakewood	-8.086e+00	1.288e+01	-0.6280	0.53008	
CityLakewood	-8.086e+00	1.288e+01	-0.6280	0.53008	
CityLittleton	-2.874e+01	1.899e+01	-1.5132	0.13036	
CityLone Tree	7.800e+01	1.380e+02	0.5654	0.57188	
CityLouisville	-2.755e+01	6.940e+01	-0.3970	0.69144	
CityMorrison	-1.777e+01	5.287e+01	-0.3361	0.73681	
CityNorthglenn	-1.519e+01	2.943e+01	-0.5160	0.60588	
CityParker	-6.111e+00	2.822e+01	-0.2166	0.82858	
CitySuperior	-5.317e+01	4.674e+01	-1.1374	0.25547	
CityThornton	2.856e+01	2.485e+01	1.1492	0.25058	
CityWestminster	-7.015e+00	1.731e+01	-0.4053	0.68532	
CityWheat Ridge	7.179e+00	2.069e+01	0.3470	0.72864	
Store_Number	-1.639e+00	1.148e+00	-1.4277	0.15351	
Avg_Num_Products_Purchased	6.714e+01	1.529e+00	43.9030	< 2.2e-16	***
X._Years_as_Customer	-2.349e+00	1.233e+00	-1.9058	0.0568	.

Table 1: Coefficients of Linear Regression to show P factor
Source: My Own + Alteryx Tool (2018)

If one observes all the data provided, only customer segment, intercept and average number of products purchased show statistical significance with the target variable. The numeric variable average number of products vs. average sale amount is the only plot obtainable in the data, since intercept is a constant and customer segment is a categorical variable. One shall observe the plot relationship with statistical significance and other with no significance to compare how the “p” factor analysis works.

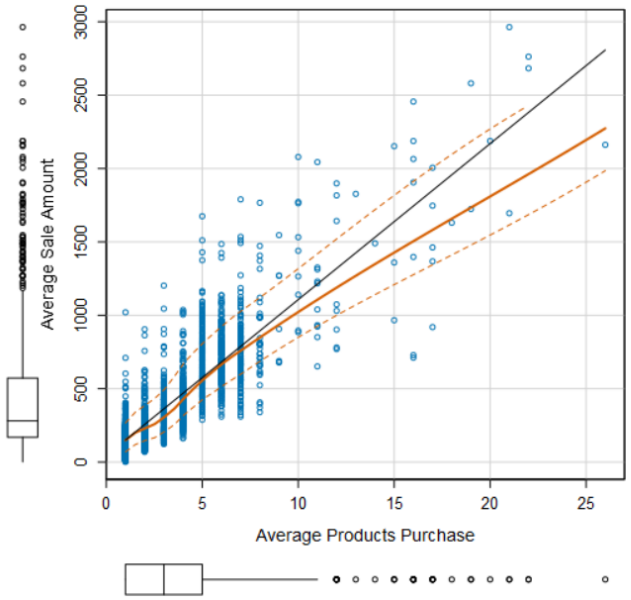


Figure 2: Plot Average Product Purchase vs. Average Sale Amount
Source: My Own + Alteryx Tool (2018)

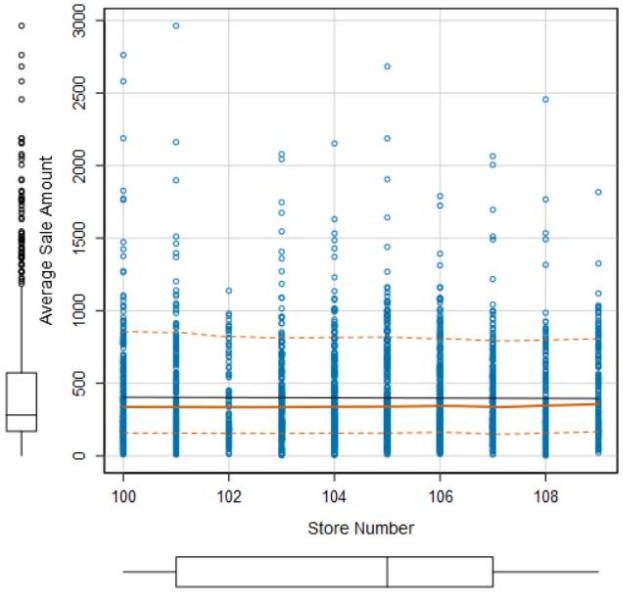


Figure 3: Plot Store Number vs. Average Sale Amount
Source: My own + Alteryx Tool (2018)

So, the target variables are:

Customer Segment and average sale amount.

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.

Previously, one evaluated all the possible predictor variables in the data and filtered them with the p-value, by stating that those values had a statistically relevant and a relationship with the target variable. We shall now only include these variables and observe their R squared values in order to test if it really generates a good linear model. In the following table, one shall observe the new results:

Record | Report

Report for Linear Model Linear_Regression_4

Basic Summary

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

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

Table 2: Report for Lineal Regression Model
Source: My Own + Alteryx Tool (2018)

The R squared values are greater than 0.8, which makes this a solid linear regression model (above 0.75 indicates that it is a very trustable model, as an indicator of trust, while less than 0.5 are indicated as very poor predictive model). The p values reflect statistical significance since it is almost “0” with each predictor variable (it reflects how there is statistical relevance in each variable, where p values lower than 0.001 are highly relevant, 0.001 are relevant and 0.05 are still relevant but in a lower category, anything higher is not relevant). One must state that these values reflect a solid linear model.

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)

The best regression equation for the given model and predictor variables is:

Y = 303.46 + 66.98(Avg. Num. Product Purchase) – 149.36 (If Type: Loyalty Club Only) + 281.84(if Type: Loyalty Club and Credit Card) - 245.42(if Type: Mailing List) + 0 (If Type: 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?

Yes, they should totally send these catalogs to the registered costumers.

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

I used alteryx to build a linear regression model, using the “Linear Regression” function from alteryx. In the input of the model I plugged the past data to build the predictive sales. After that, I used the function “score” to build the predictive model, using the Linear Regression and the data used to make predictions. On the “score” output, I multiplied the Score_Yes row with the predicted sales output in order to adjust the answer according to the clients probability of buying the product, using the “formula” function. In the same function I divided by two (2) which represents de 50% gross margin. After this, I used another “formula” function to substract 6.50 to each and every profit calculated per person, since it is the cost per catalog. At last, I summed the row with every cost calculated with the “summarize” function and landed with the final value (21987.43$). The alteryx simulation had the following look:

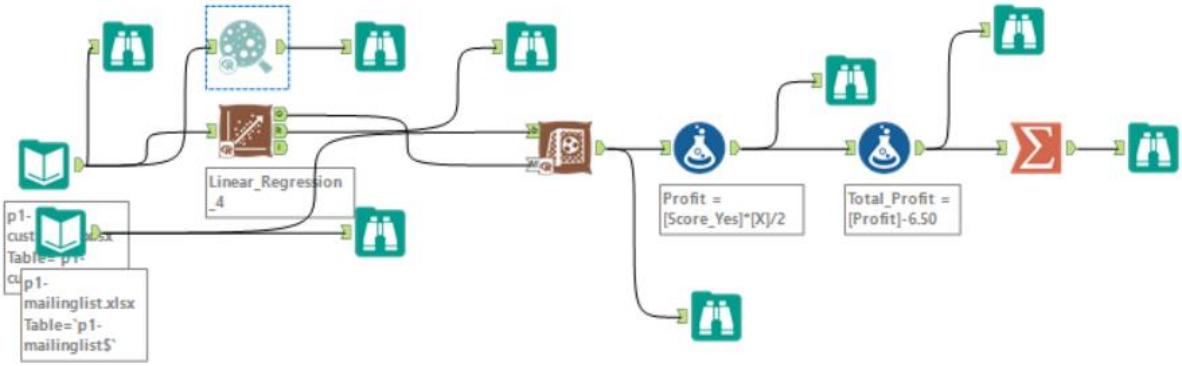


Figure 4. Alteryx Scheme for Solving the Problem

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

The expected profit from the new catalog is 21987.43 $.