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)

The company needs to determine if it is worth sending this year's catalog to the new clients. A decision is to be made and it is deemed that the expected profit should be greater than $10,000 to opt for sending out the catalog.
Given the average profit margin and costs for printing our catalogs we can calculate the expected profit as follows:

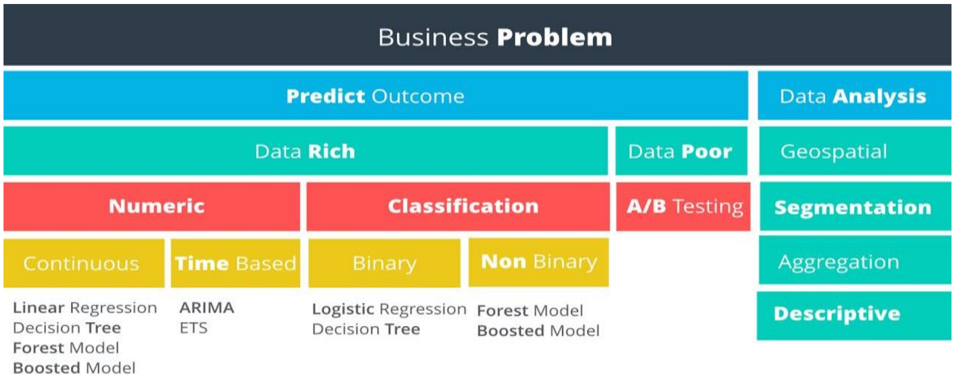
Profit = Rev. from Catalogs * Avg. Gross Margin – Printing Costs =>
= Rev. from Catalogs *0,5 - $6.50 * 250 = Rev. from Catalogs *0,5 - $1,625

The only unknown quantity from the above formula is the expected revenue from sending out the catalogs to the new customers which we have to predict somehow. Since this is a predictive exercise we need data from past experiences with clients to inform our analysis. The company has provided the *p1-customers* and *p1-mailing list* spreadsheets (data rich business problem), including information on customers (state, city, address, no. stores) their consumer behaviour (no. items purchased, total value of items bought, and if responded to last year's catalog). The above is the data we would hope to have to investigate the business problem.

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)

The type of analysis to be employed was decided by following the methodology map below:



In our case, the business problem requires predictive analysis and is data rich with numeric variables, such as average sales amount. Revenue is a continuous variable and as such we can build a linear regression model to predict sales. For this purpose, Alteryx was our chosen tool for the analysis.

Building the predictive model requires to select our target and predictor variables. The target variable is the Avg_Sale_Amount, but we are not sure about predictor variables. I chose to run an initial predictive analysis with many variables as a way to rule out those who do not show a relationship with the target variable and are not statistically significant (please see below in red).

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	493.3116	122.893	4.01416	6e-05	***
Customer_SegmentLoyalty Club Only	-150.6401	9.014	-16.71251	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.7801	11.956	23.65203	< 2.2e-16	***
Customer_SegmentStore Mailing List	-242.8152	9.888	-24.55721	< 2.2e-16	***
CityAurora	-20.5819	11.086	-1.85660	0.06349	.
CityBoulder	-41.1805	80.029	-0.51457	0.6069	
CityBrighton	-59.4890	97.639	-0.60927	0.5424	
CityBroomfield	-4.3414	15.124	-0.28705	0.7741	
CityCastle Pines	-93.0347	97.642	-0.95282	0.34078	
CityCentennial	-9.5731	18.158	-0.52721	0.59809	
CityCommerce City	-33.2255	44.454	-0.74742	0.45489	
CityDenver	0.2317	10.551	0.02196	0.98248	
CityEdgewater	27.9712	40.612	0.68875	0.49105	
CityEnglewood	6.0143	20.737	0.29002	0.77183	
CityGolden	-11.4221	32.719	-0.34910	0.72705	
CityGreenwood Village	-44.4576	38.059	-1.16812	0.24288	
CityHenderson	-285.8339	137.847	-2.07357	0.03823	*
CityHighlands Ranch	-28.1976	30.420	-0.92694	0.35405	
CityLafayette	-43.7104	62.140	-0.70342	0.48186	
CityLakewood	-7.3541	12.858	-0.57195	0.56741	
CityLittleton	-28.7184	18.967	-1.51412	0.13013	
CityLone Tree	77.3956	137.769	0.56178	0.57432	
CityLouisville	-30.5955	69.266	-0.44171	0.65874	
CityMorrison	-18.6190	52.789	-0.35271	0.72434	
CityNorthglenn	-14.7157	29.393	-0.50066	0.61666	
CityParker	-6.0965	28.177	-0.21636	0.82873	
CitySuperior	-56.1322	46.681	-1.20245	0.22931	
CityThornton	29.0992	24.814	1.17271	0.24103	
CityWestminster	-6.6966	17.284	-0.38745	0.69846	
CityWheat Ridge	8.9128	20.673	0.43114	0.66641	
Store_Number	-1.6365	1.146	-1.42779	0.15348	
Responded_to_Last_CatalogYes	-29.5786	11.335	-2.60943	0.00913	**
Avg_Num_Products_Purchased	66.9147	1.527	43.81327	< 2.2e-16	***

X._Years_as_Customer	-2.3411	1.231	-1.90197	0.0573	.

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

Hence, the model is reconstructed to predict Avg_Sales_Amount by using only the following predictor variables: *Customer Segment, Avg_Num_Products_Purchased*. Please note that although Responded_to_Last_Catalog is statistically significant it is omitted from the model. This is because the new customers have never received a catalog before and so the relative product within the regression equation will be 0. Re-running the model:

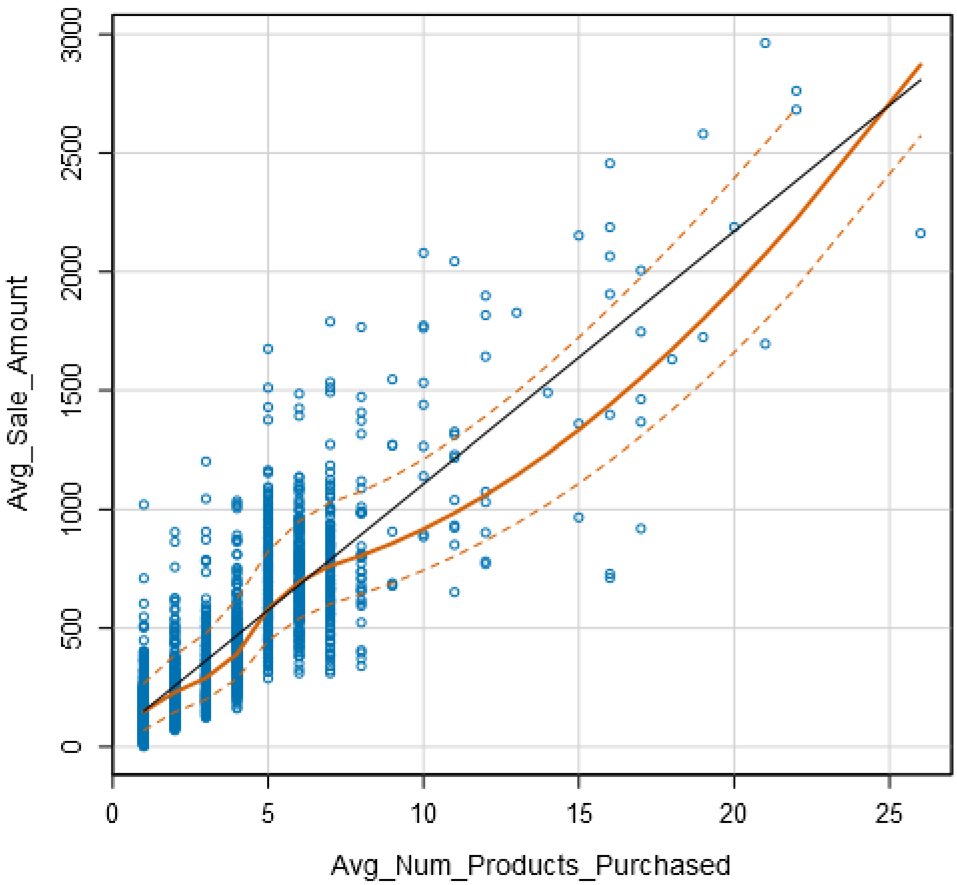
Coefficients:

	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
```

All predictor variables show p-values<<0.05, are statistically significant and the Adjusted R-squared value is high – 0.8366. However, we need to make sure that the all continuous predictor variables have a linear relationship with the target variable, i.e. create a Avg_Num_Products_Purchased vs Avg_Sales_Amount scatterplot.



Indeed, there is a linear relationship between the two variables (fitted curve/straight in black).

The derived linear regression equation based on the available data is shown below:

Avg_Sales_Amount = 303.46 – 149.36**(If: Loyalty Club Only)* + 281.84**(If: Loyalty Club and Credit Card)* – 245.42 * *(If: Store Mailing List)* + 66.98 * *Avg_Num_Products_Purchased*

Given that all new customers are under the store mailing list category the equation takes a more simplistic form:

Avg_Sales_Amount = 303.46 – 245.42 * 1 + 66.98 * *Avg_Num_Products_Purchased* ➔

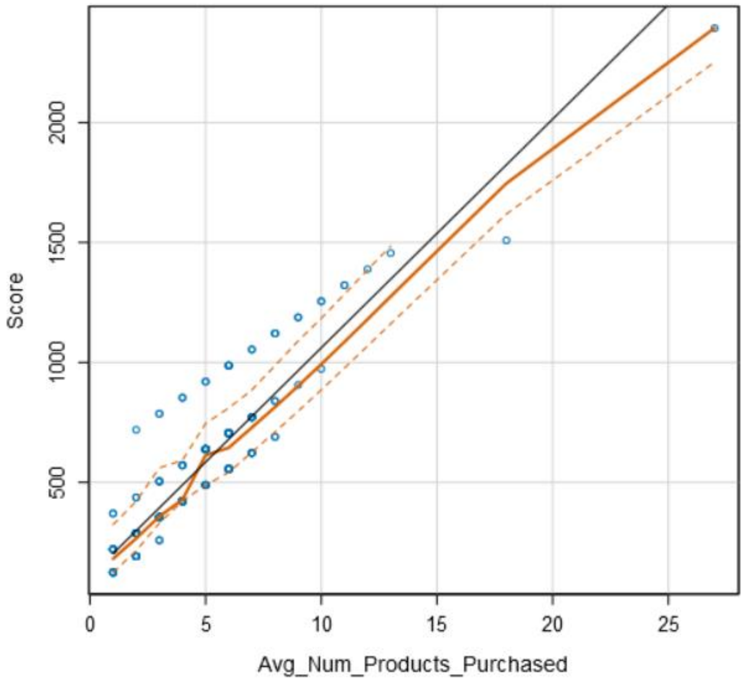
Avg_Sales_Amount = 58.04 + 66.98 * *Avg_Num_Products_Purchased*

Step 3: Presentation/Visualization

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

To make a recommendation we first need to apply the regression equation to the p1-mailing list customers (Score). Then, having evaluated the probability that a new customer will make a purchase (Score_Yes) we multiply by the projected average sales amount for each customer.

The following scatterplot illustrates the linear relationship between Scored Sales and Average Number of Products Purchased.

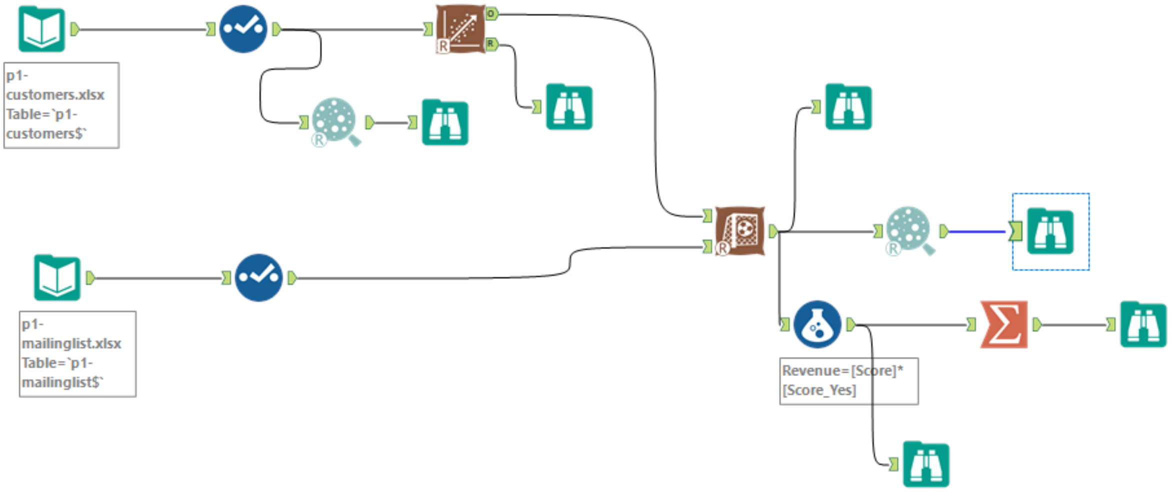


Summarizing all these yields the total anticipated revenues from sending out the catalogs, which is $47,224.87. The predicted profit is:

Profit = Rev. from Catalogs *0,5 - $1,625 = $47,224.87 * 0,5 - $1,625 = $21,987.44

<u>Our recommendation is that the company should send the catalog to these 250 new customers since the anticipated profit exceeds the $10,000 threshold.</u>

<u>Appendix 1 – Alteryx Model</u>


p1-customers.xlsx
Table=`p1-customers$`
p1-mailinglist.xlsx
Table=`p1-mailinglist$`
Revenue=[Score]*
[Score_Yes]