The first and foremost step to approach any Data Science problem is to understand the business requirement and data. Let's understand the business requirement before moving further.

Business and Data Understanding

1. What is the business requirement?
   Management wants to decide whether to send out this year's catalogue to the new 250 customers the business has got. If the predicted profit exceeds $10,000, the firm is planning to send out the catalogues.

2. What data is needed to make a prediction?
   To predict the expected profit, we need a dataset of sales and demographics info from our existing customers to train the model and a similar set for our 250 new customers that we'll apply the model to get our predictions.
   Additionally, we need the gross margin for each product and the cost of printing the catalogue.

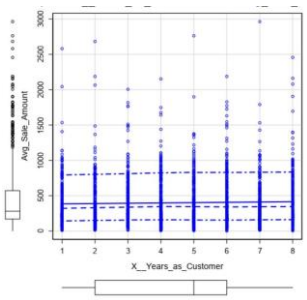
Next, we should analyze the data and find the optimal variables for modeling.

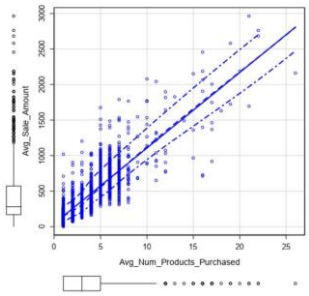
Analysis

Analyzing the data before modeling can give us many insights about the data set and the problem we are trying to solve. We can find the relationship between predictor variables, and also we can check for insights like the location that yielded a high profit to the business or city with an excellent response to the catalogue in the past. But, we are trying to predict only the total profit by sending out the catalogues. So, let us find the best predictor variables for predicting the average sales amount for each new customer.
Note: In a real business scenario, we will be analyzing the relationship between variables like city and response, city and average sales amount and present the findings to the management.

Few variables like Name, Customer ID, and Address won't provide much impact on our analysis. Sometimes address will be a significant variable to predict customer spending because the customer will often be purchasing from the store near his location. But we need Latitude and Longitude info the store and address to use them on our analysis. Since we don't have them, we can eliminate the address from our study. A state can have a good impact on customer behaviour, but all of our training data is from the same state, so that it won't be useful for our analysis.

The best way to study the impact of the numerical variable is to visualize them using a scatter plot. Let us analyze the average number of products purchased and years as a customer with our predictor variable, i.e., average sales amount. From the plot, we can see that years as a customer has no substantial relationship with the target variable, while the average number of products purchased has a strong linear relationship.





The impact of categorical variables can be checked using p-value, which shows whether the variable is statistically significant or not. From the below results, we can see that the Customer Segment has a low p-value, while City, Zip, and Store number has high p-value. Hence, we will be not be using City, Zip, and Store number in our model. Zip code and Store number will look like a numerical variable, but they are categorical variables. Also, from these results, we can see that the Average num of the product purchased has low p-value, and it supports the assumption we made using a scatterplot.

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	27413020.63	3	480.37	< 2.2e-16 ***
City	290021.73	18	0.85	0.6448
ZIP	1290040.03	77	0.88	0.76054
Store_Number	194978.46	9	1.14	0.3312
Avg_Num_Products_Purchased	35308379.63	1	1856.17	< 2.2e-16 ***
Residuals	42952075.43	2258		

3. What are the best predictor variables for modeling?

   From the above analysis, we are taking only the Customer Segment and Average Number of products purchased as the predictor variable for finding the average sales.

Modeling and Validation

We have identified the optimal predictor variables, and the next step is to build the model. Usually, we will perform steps like scaling and one hot encoding before modeling, but Alteyrx takes care of that. So let us create a Linear regression model with these predictor variables.

After building a model, it is crucial to validate the model. Upon validating the model, we can see that the model has an adjusted R-squared value of 0.8366. It's important to note that the r-squared value represents the variation in the target variable that the model captures. Generally, we consider models with an adjusted r-squared >0.7 to be acceptable. In this case, about 84% of the variation is accounted for, leaving 16% to be explained by variables outside the model. This leaves room in the future for the marketing team to work with the analysts in identifying other relevant variables that they may not be monitoring yet.

Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366

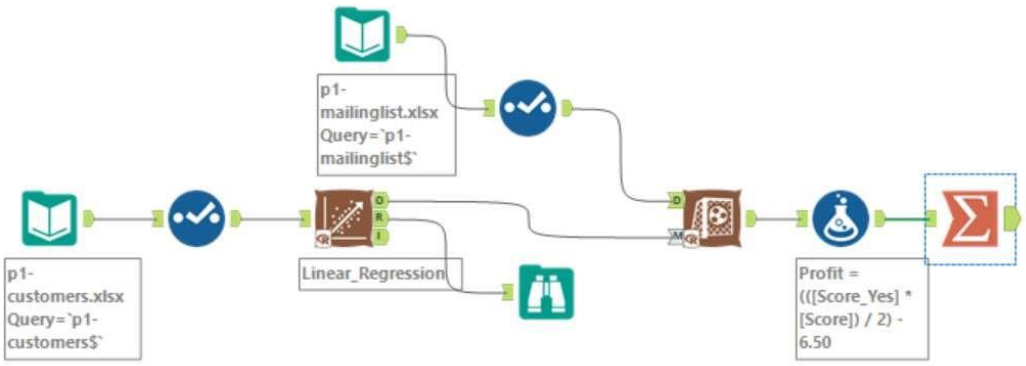
From the model, we can get the below regression equation:

Avg_Sale_Amount = 303.46 - 149.36 * Customer_Segment(Loyalty Club Only) + 281.84 * Customer_Segment(Loyalty Club and Credit Card) - 245.42 * Customer_Segment(Store Mailing List) + 0 * Customer_Segment(Credit Card only) + 66.98 * Avg_Num_Products_Purchased

Results

To make a recommendation, we must estimate the average sales for each customer on the new list. We can use the model we build to predict the average sales for each new customer. One of our team already predicted the probability of the customer responding to the catalogue, and that probability can be multiplied with the expected average sales to give the estimated sales from each customer. We should account for the gross margin, i.e., 50% and the cost of printing catalogue, i.e., $6.50, to calculate the profit for each new customer. By performing all of the above steps, we can get the profit for each customer, adding the overall profit gives $21,987.44.

Alteryx Workflow



Recommendation

With detailed analysis, we build a model that can account for an 85% variance of the average sales from a customer. Customer segment and Average products purchased are the only variables that have a good impact on the sales, and we should work with our marketing team to collect more relevant variables to build a better model in future.

We have estimated that sending out the catalogues will yield a profit of $21,987.44 to the business. Since the estimated profit is much higher than the threshold set by management, we strongly recommend sending out the catalogues.