Project 1

Predicting Catalog Demand

Structure

1. Business and Data Understanding
    - The Business Problem
    - Details
    - Key Decision
2. Analysis, Modeling, and Validation
    - Analysis,
    - Modeling
    - Validation
3. Presentation/Visualization
4. Conclusion

1. Business and Data Understanding

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

The Business Problem

- ✓ A company manufactures and sells high-end home goods and preparing to send out this year's catalog
- ✓ 250 new customers from their mailing list that they want to send the catalog to.
- ✓ Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.
- ✓ The costs of printing and distributing are $6.50 per catalog. The average gross margin (price - cost) on all products sold through the catalog is 50%.

Details

- ✓ There are two datasets:
  - o p1-customers.xlsx contains the data used to build the regression model.
  - o p1-mailinglist.xlsx contains the data of the 250 new customers the company want to send catalog.
- ✓ Both datasets contain information of 'Customer_Segment', 'Address', 'City', 'ZIP', 'Avg_Sale_Amount', 'Store_Number', 'Avg_Num_Products_Purchased', 'Nb_of_Years_as_Customer'.
- ✓ Only p1-customers.xlsx contains Responded_to_Last_Catalog and 'Avg_Sale_Amount'. We'll predict 'Avg_Sale_Amount' (Revenue) from new customers of the p1-mailinglist.xlsx dataset.
- ✓ 'p1-mailinglist.xlsx' contains the probability of Responded_to_Last_Catalog. So based on probability of response to catalog (Score_Yes, Score_No), we can calculate the expected revenue of 250 customer.

Key Decisions

- ✓ What decisions needs to be made?
  - o Determining how much profit the company can expect from sending a catalog to these customers.
  - o Deciding to send the catalog out to new customers only if the expected profit exceeds $10,000.
- ✓ What data is needed to inform those decisions?
  - o Avg_Sale_Amount
  - o Responded_to_Last_Catalog
  - o Avg_Num_Products_Purchased
  - o Score_No, Score_Yes
  - o Avg_Num_Products_Purchased
  - o Nb_Years_as_Customer

2. Analysis, Modeling, and Validation

Analysis

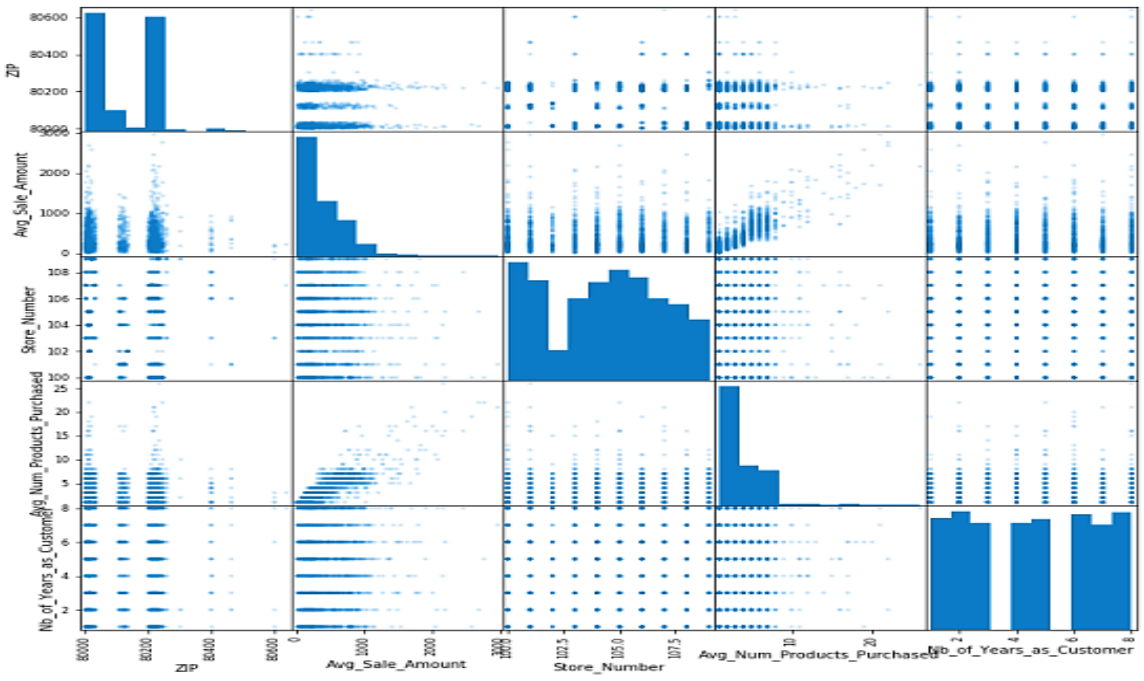


Figure 1: Plot of correlation between involved factors

Here we need to calculate the expected profit from new customers and we can send the catalog out only if profit contribution exceeds $10,000. To calculate the profit, we need at first to see how this factor is influenced by other factors. With the Figure 1, we have a quick look about the relationship between revenue and other numeric factors. And we can see that the average sale amount has a clear relationship with the number of product purchased (Avg_Num_Products_Purchased). It is easy to understand that the

more products are purchased, the larger the amount of sale the company will get. For another, continuous variable we see unclearly relation between amounts of sales with other factors.

With the headmap we can see it clearlier that the average sale amount has a strong positive relationship with the number of purchased products. With the others, it is no clue to see a relationship.

Figure 2: Heatmap of correlation between factors

After having some views with numerical factors, we give some views for the category variables. Here, with the Figure 3, we can see that in some cities, the total number of sales is very higher than another city. However, the revenue line in chart does not show any clear relation with the city or the number of sales making at these.

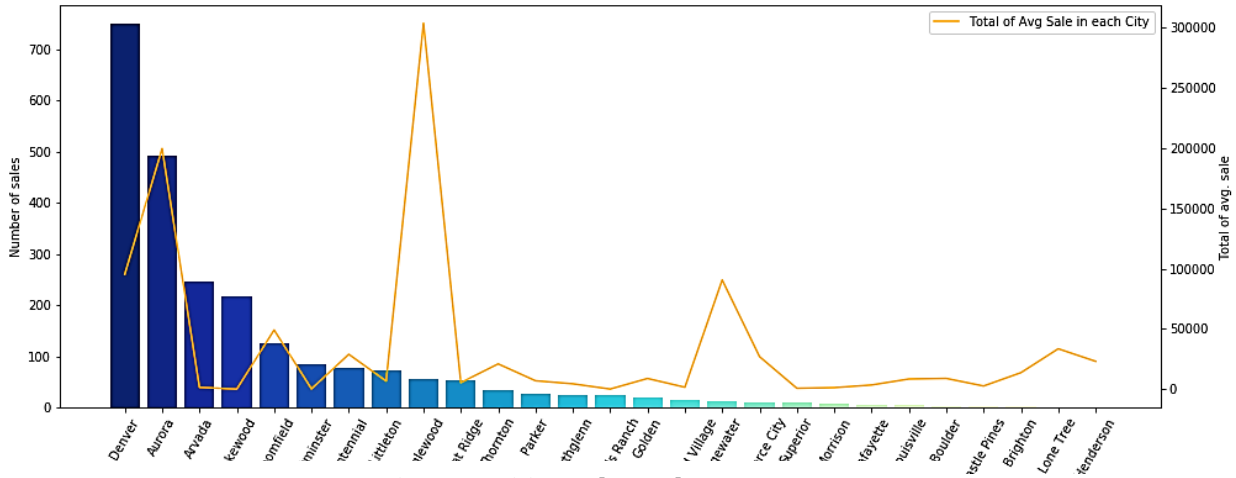


Figure 3: Cities and actual revenue

In order to understand the relationship between revenue and other category variables such as the customer segment and the respond to catalog, we use the violin chart. With the Figure 4 we can see that there are an obvious relation betweet the average amount of sale with the customer segments. The average amount of sale in the segment 'Loyal Club and Credit Card' is obviously higher than others, following is 'The credit card only' segment. With the violin chart, we can see shows the full distribution of the data in each variable in the category. So it is clear that customer segment plays an important role in the predicted sale amount of product.

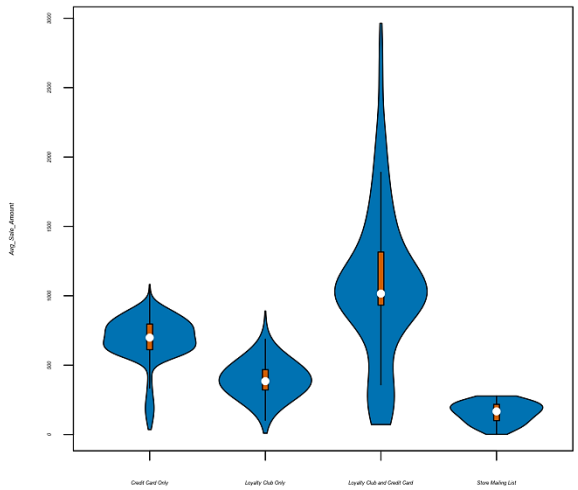


Figure 4: Sale Amount and Customer Segment

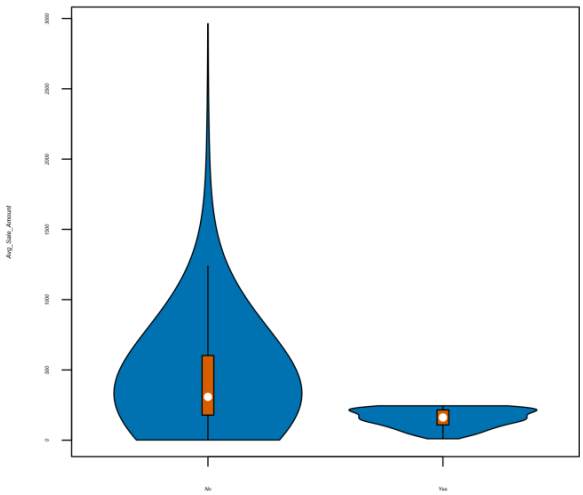


Figure 5: Sale Amount and Response to Catalog

As the same, in the Figure 4, we can see that larger group has no response to catalog and it has a wide rage of amount to purchase products, however 95% of the average amount of sale is not far higher than the group having response to cataloge. On contrast, also the group having response to cataloge is quite smaller, the number purchased no product accounts only a very small part of the total numbers having response yes to catalog. The variable having yes response to cataloge shows that almost customers will purchase products at the average amount. So sending catalog to new customers seems to be a good idea.

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)	
Customer_Segment	27972982.1	3	493.62	< 2.2e-16	***
City	420585.49	26	0.86	0.67363	
Responded_to_Last_Catalog	1290 o03.32	1	6.83	0.00902	**
Avg_Num_Products_Purchased	36288117.67	1	1921.07	< 2.2e-16	***
Residuals	44258202.63	2343			

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

To have an overview about how good is the prediction we use the linear regression with the *Type II ANOVA Analysis*. With the result we can see that, the variable of city plays unimportant role in the influence on revenue. Out of them, the Customer_Segment and Avg_Num_Products_Purchased play significant roles in predicting such amount of sales with p_value < 2.2e-16. Still a significant variable but a little bit less significant is Responded_to_Last_Catalog (p_value = 0.00902).

However with the data for the number of new customers, other words in dataset 'p1-mailinglist.xlsx', we don't have variable 'Responded_to_Last_Catalog', but only have the probability of response yes or not. So in the regression model, we will remove this variable from the regression model and only calculate the Customer_Segment and Avg_Num_Products_Purchased in the predicted model. After that we will process the regression model with the probability of response to calculate the expected revenue and then the expected profit.

Modeling

Linear regression

We will use linear regression tool in Alteryx. The variable Avg_Sale_Amount is defined as the target variable that we want to predict. With the analysis above, the control variables are Customer_Segment and Avg_Num_Products_Purchased.

Score the model

After having a model, we will apply it to the dataset with 250 new customers to make a prediction about the revenue gaining from sending catalog to them. With Alteryx, we can do this with a tool called a Score tool. After having predicted revenue under the consideration of two control variables Customer_Segment and Avg_Num_Products_Purchased, we need to calculate the revenue with the influence if there is a yes response to the catalog. Because sending catalog is still a plan, we donot have the information if a customer will respond or not, so we need to make score, other words calculate the probability wheather a customer responds no or yes and this data is given is dataset of 250 new customers.

Validation

Report for Linear Model Linear_Regression

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

With the result of regression (run by Alteryx) we can see that the variables in Customer_Segment and Avg_Num_Products_Purchased are significant with p-value < 2.2e-16. The Adjusted R-Squared: 0.8366 is also very good and shows that the model is good in oder to predict expected revenue. Here we can see that in Customer_Segment misses 1 variable which is Credit Card Only because the variable Credit Card Only in the regresssion model is the base variable in the Customer_Segment.

So we have the linear regression equation based on the available data:

$Y =$ 303.46
+ 66.98 * Avg_Num_Products_Purchased
+ 0.00 * If Credit_Card_Only
- 149.36 * If Customer_SegmentLoyalty Club Only
+ 281.84 * If Customer_SegmentLoyalty Club and Credit Card
- 245.42 * If Customer_SegmentStore Mailing List

3. Presentation/Visualization

With the visualization of actual and predicted revenue, we can see the predicted revenue is calculated with the model from the old one with the average expected revenue of 4 customer segments far lower than the average real revenue. The same is in the relationship between revenue and purchased products. Those are due to the tighter dataset with most of new customers purchased products with avg revenue presenting with orange bar.

Figure 6: Avg Sale Amount and Customer Segments Actual and

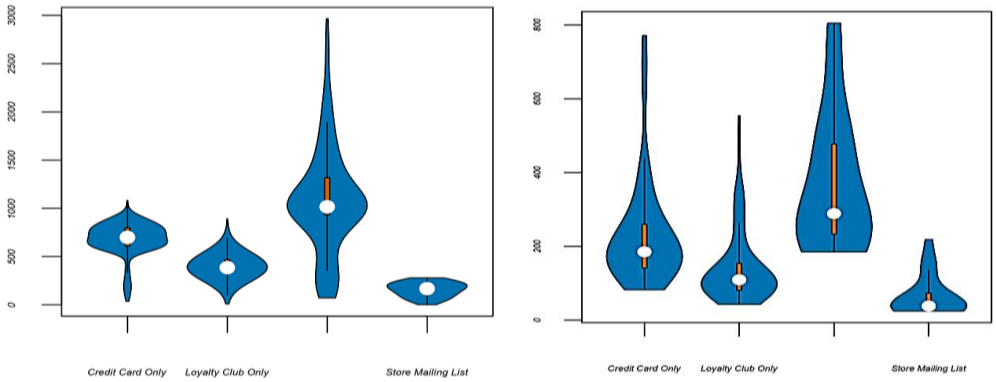
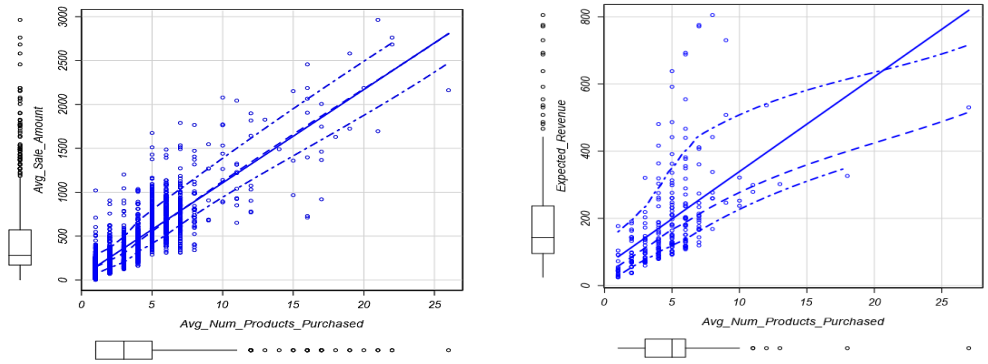


Figure 7: Avg.Sale Amount and Purchased Products Actual and Predicted



4. Conclusion

- ✓ **Recommendation**

With all information above, I recommend that the company should send the catalog to 250 new customers if the condition is fulfilled because the model seems to be quite exact with the significant control variables and in a good logical relation.

- ✓ **Eexpected profit from the new catalog (assuming the catalog is sent to these 250 customers)**

The expected profit = Total Expected Sale*0.5-6.50*250

= Sum (Expected Sale*Score Yes) *0.5-6.50*250

= 21987.435687

Alteryx Workflow