Project 1: Predicting Catalog Demand

OLUWATOSIN OBALANA

Project Overview:

In this project, I'll analyze a business problem in the mail-order catalog business. I have been tasked to predict how much money the company is expected to earn from sending out catalog to new customers. This task will involve me building a model and applying the results in order to provide recommendations to the management.

The Business Problem:

I have recently started working for a company that manufactures and sells high end home goods. Last year the company sent out its first print catalog and is preparing to send out this year's catalogue in the coming months. The company has 250 new customers from their mailing list that they want to send their catalogue to.

My manager has been asked to determine how much profit the company can expect from sending a catalogue to these customers. I have been assigned to help my manager run the numbers.

I've been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalogue out to these new customers unless the expected profit contribution exceeds $10000.

Details

- The cost of printing and distributing is $6.50 per catalogue
- The average gross margin (price – cost) on all products sold through the catalog is 50%
- When calculating the profit, the revenue must be multiplied by the gross margin first before subtracting it from the $6.50 cost.

Business and Data Understanding:

1. What are the key decisions that needs to be made?

The key decisions to be made is

- to determine how much profit the company can expect from sending a catalogue to these new customers.
- also, to determine if the expected profit contribution exceeds $10000, only then can the catalogue be sent to the new customers.

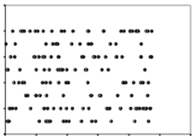
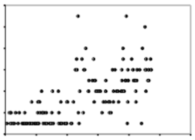
2. What data is needed to inform those decisions?

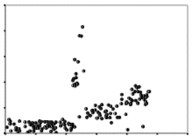
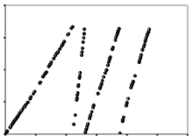
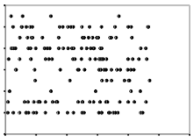
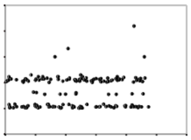
First, we need to build a linear regression model to determine the strength of the predictors, to do that we would need to use the p1-customer.xlsx file. This dataset contains information about 2300 customers and we would be focusing on relationship between variables like the avg_sales_amount and avg_number_of_product_purchased

Secondly, since we need to predict the number of sales, we would need to use the p1-malinglist.xlsx file. This is the list of customers that the company would send out the catalog to. It includes all the fields from p1_customer.xlsx except for Responded_to_Last_Catalogue. We would be focusing on two variables the Score_No and the Score_Yes, they show the probability that a customer will or will not respond to the catalogue and make a purchase.

Step 2: Analysis. Modeling. and Validation

The target variable is the variable we want to predict and that is the **avg_sale_amount**. So when selecting predictor variables, the first thing you do is to use logic to determine what might be a good predictive variable in this scenario we can say that the customer segment, city, and average number of products purchased might be good predictive variables but before I moved forward I used the field summary to explore the data more.

Name	Plot	% Missi	Unique Values	Min	Mean	Median	Max	Std Dev	Remarks
#_Years_as_Customer		.0%	8	1.000	4.501	5.000	8.000	2.310	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
Avg_Num_Products_Purchased		.0%	23	1.000	3.347	3.000	26.000	2.739	

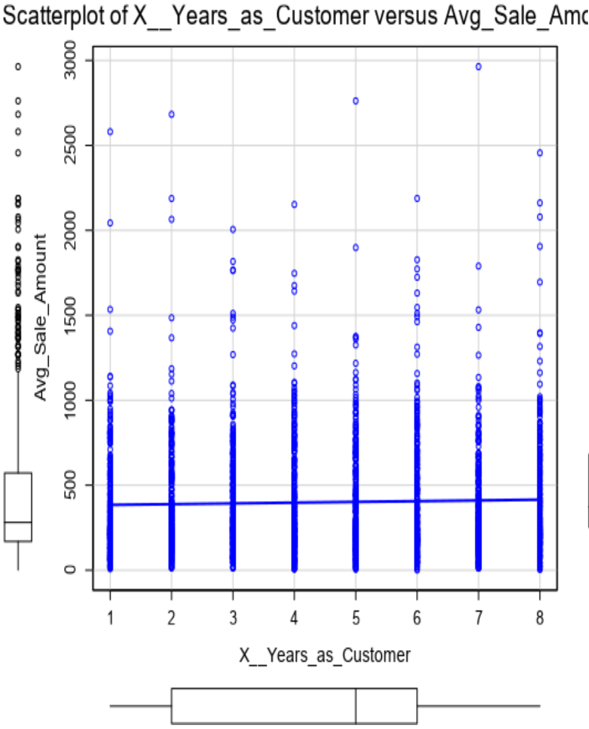
Avg_Sale_Amount		.0%	2,345	1.220	399.774	281.320	2,963.490	340.116	
Customer_ID		.0%	2,375	2.000	1,647.845	1,629.000	3,335.000	962.728	
Store_Number		.0%	10	100.000	104.298	105.000	109.000	2.837	This field has a small number of unique values, and appears to be a categorical field. Consider changing the field data type to "string".
ZIP		.0%	86	80,002.000	80,123.333	80,123.000	80,640.000	107.256	

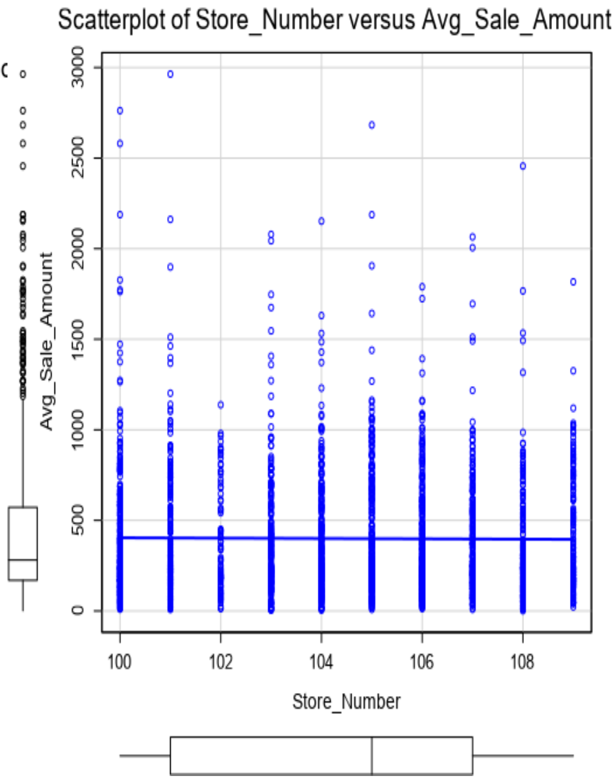
Name	% Missin	Unique Values	Shortest Value	Longest Value	Min Value Count	Max Value Count	Remarks
Address	0.0%	2,321	60 Ivy St	5250 E Cherry Creek South Dr	1	6	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
City	0.0%	27	Denver	Greenwood Village	1	750	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.
Customer_Segment	0.0%	4	Credit Card Only	Loyalty Club and Credit Card	194	1,108	
Name	0.0%	2,366	J Ritz	Angela Edington-Molyneux	1	2	Some values of this field have a small number of value counts. If Appropriate, consider combining some value levels together.

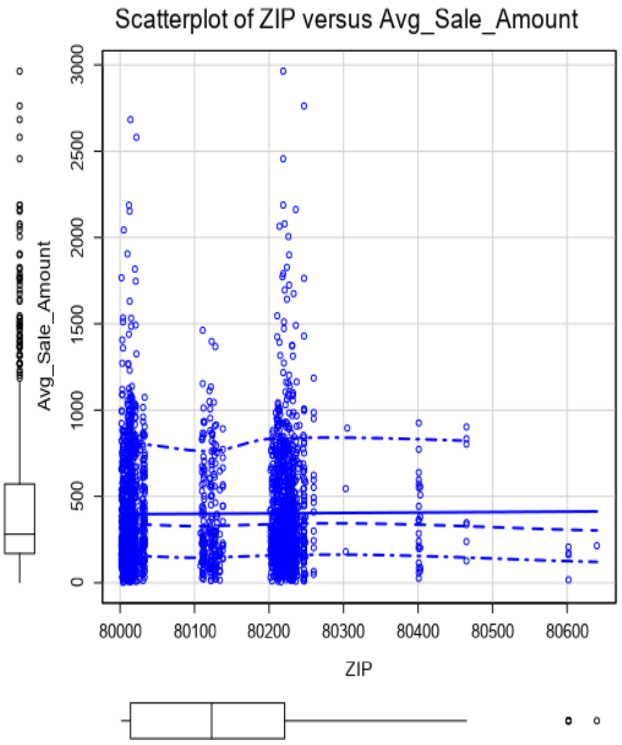
Responded_to_ Last_Catalog	0.0%	2	No	Yes		171	2,204	
State	0.0%	1	CO	CO		2,375	2,375	

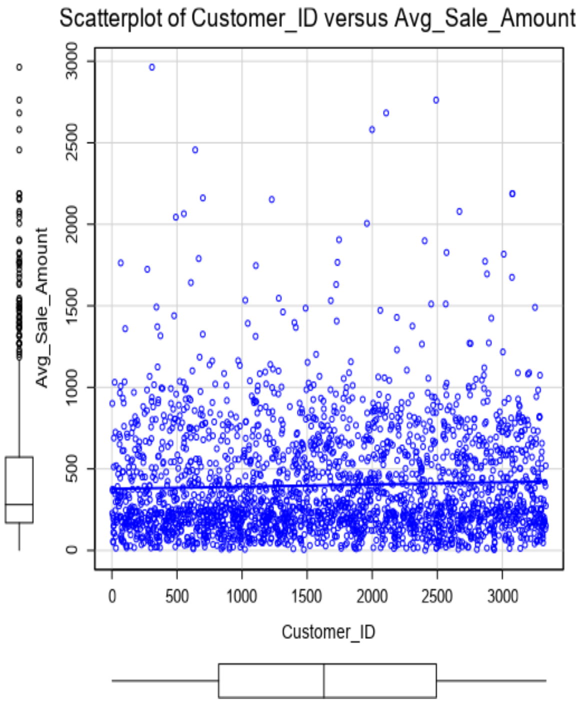
From the report, It can be seen that in the string field, the **address** and **name** has a lot of unique values so when using linear regression this would create a lot of dummy variables and we would have a lot of predictors and that won't be good for our model, so we have to exclude them when selecting the predictors. The state has only one unique variable, it won't be enough to create a dummy variable so we exclude this too.

Next, we use the scatter plot tool to check the correlation between the numeric fields and the target variable to see if there's a strong linear relationship between the two variables.



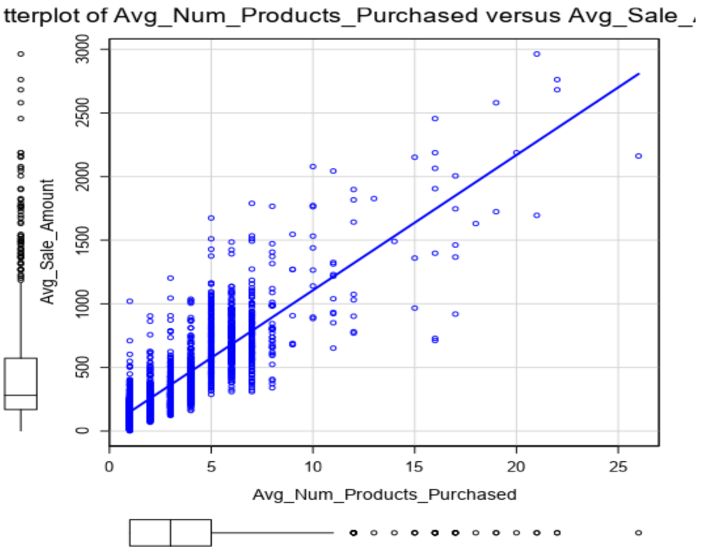



Scatterplot of ZIP versus Avg_Sale_Amount


Scatterplot of Customer_ID versus Avg_Sale_Amount

From the graphs above we can see that there is no linear relationship between these numeric fields [ZIP, Customer_ID, X_Years_as_a_Customer,Store Number] and the target variable[Avg_sale_price] so we can discard these variables when choosing the predictors.

The only variable that shows a strong positive linear relationship is the Average_Product_Purchased so we can use that when selecting the predictor variables


tterplot of Avg_Num_Products_Purchased versus Avg_Sale_

For the categorical variables remaining we can use the linear regression tool to check if they are statistically significant [when the p-value is sufficiently small (e.g. 5% or less)]

<table>
<tr><td>FitStats</td><td>Residual standard error: 137.61 on 2344 degrees of freedom
Multiple R-squared: 0.8384, Adjusted R-Squared: 0.8363
F-statistic: 405.3 on 30 and 2344 degrees of freedom (DF), p-value < 2.2e-16</td></tr>
<tr><td>ANOVAAnalysisTitle</td><td><i>Type II ANOVA Analysis</i></td></tr>
<tr><td>ANOVAAnalysis</td><td>Response: Avg_Sale_Amount
<table>
<tr><th></th><th>Sum Sq</th><th>DF</th><th>F value</th><th>Pr(&gt;F)</th><th></th></tr>
<tr><td>Customer_Segment</td><td>28312992.26</td><td>3</td><td>498.38</td><td>&lt; 2.2e-16</td><td>***</td></tr>
<tr><td>City</td><td>409663.12</td><td>26</td><td>0.83</td><td>0.70799</td><td></td></tr>
<tr><td>Avg_Num_Products_Purchased</td><td>36579739.66</td><td>1</td><td>1931.7</td><td>&lt; 2.2e-16</td><td>***</td></tr>
<tr><td>Residuals</td><td>44387205.95</td><td>2344</td><td></td><td></td><td></td></tr>
</table>
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1</td></tr>
<tr><td>BasicDiagnosticPlotsTitle</td><td><i>Basic Diagnostic Plots</i></td></tr>
</table>

We can see that only avg_num_products and Customer Segments are the only statistically significant variables, so that’s going to be the predictor variable we are going to use for the model.

1. 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

The model is a good model because there’s a strong linear relationship between the variables. This is shown by the r-squared and the adjusted r-squared below. The model showed that the p value of Customer_Segment was less than 2.2e-16 and Avg_Num_Products_Purchased was less than 2.2e-16 whenever the p-value is less 0.05. It shows the predictors used are statistically significant hence the model is a good model because it is caused by something other than chance.

Desc Report

Title

Report for Linear Model Linear_Regression_28

SummaryTitle

Basic Summary

Call

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)

ResidualsTitle

Residuals:

Residuals

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

CoefficientsTitle

Coefficients:

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

FitStats

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

ANOVAAnalysisTitle

Type II ANOVA Analysis

ANOVAAnalysis

Response: Avg_Sale_Amount

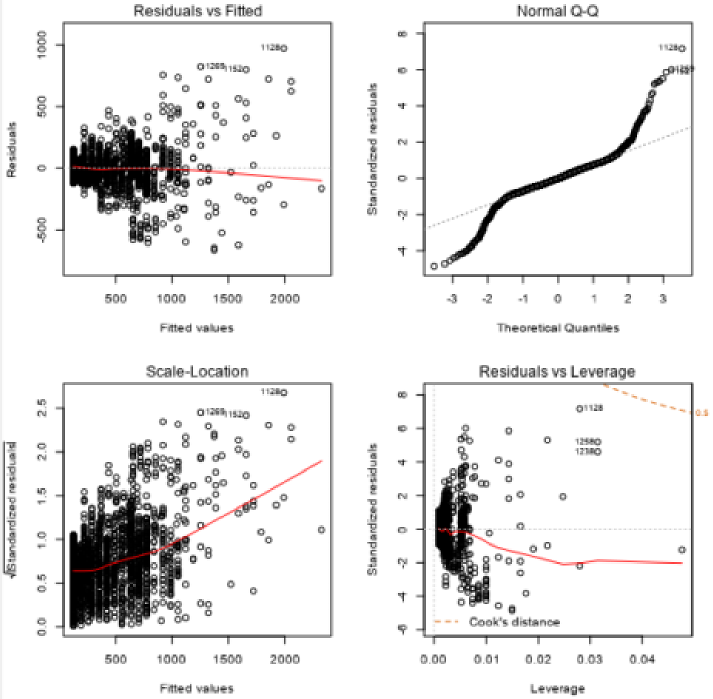
	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	28715078.96	3	506.4	< 2.2e-16 ***
Avg_Num_Products_Purchased	36939582.5	1	1954.31	< 2.2e-16 ***
Residuals	44796869.07	2370		

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

BasicDiagnosticPlotsTitle

Basic Diagnostic Plots

BasicDiagnosticPlots



2. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal

Y = 303.46 +(-149.36 * customer_segmentLoyalty_Club_Only) + (281.84 * customer_segmentLoyalty_Club_and_Credit_Card) + (-245.42 * customer_segmentStore_Mailing_List) + (66.98 * avg_num_products_purchased) + 0

Step 3: Presentation/Visualization

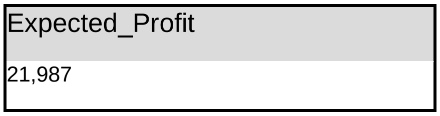
1. What is your recommendation? Should the company send the catalog to these 250 customers?

If we look at the analysis, we can see the predicted profit from sending catalogues to new customers is about $21,987.43 that's is more than the goal set of $10,000 by management. Therefore, the company should send the catalogs to new customers.

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

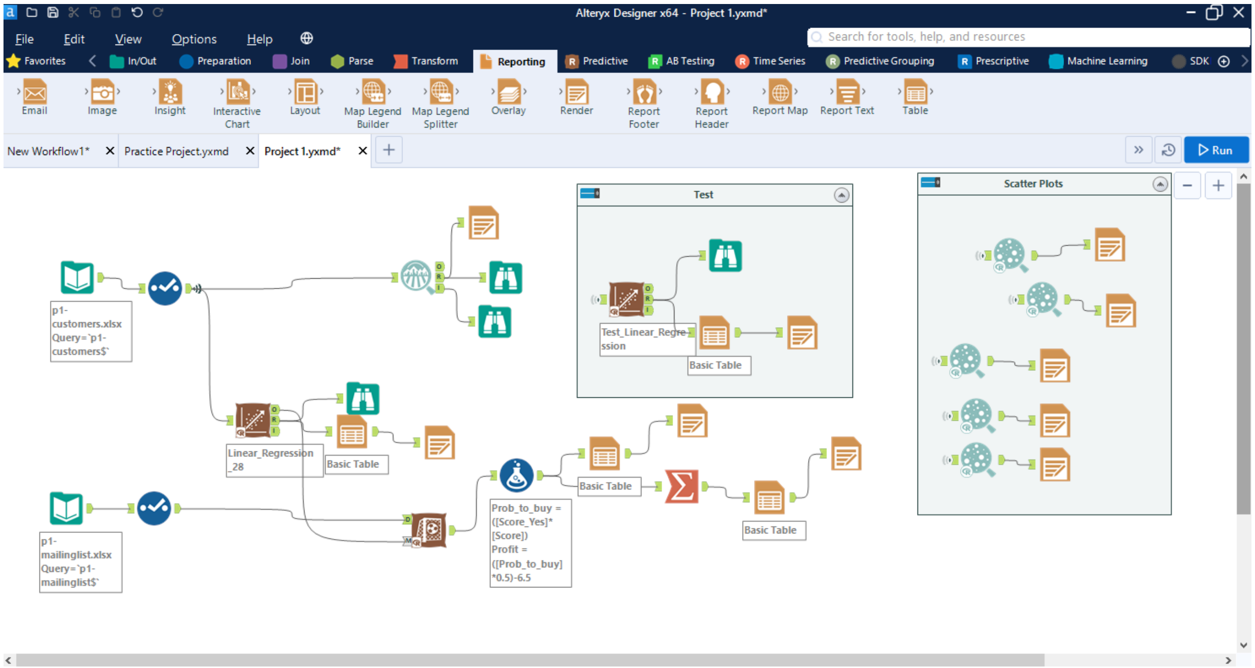
I used the linear regression equation from the linear regression model to predict sales for the new customers. Then, I multiplied the predicted sales by the probability [score_yes] that a customer would make a purchase if they get the catalogue, Then I multiplied this by 50% that's the average gross margin(price-cost) on all products sold through the catalog. Finally, I subtracted $6.50 cost to get the final predicted profit for the new customers.

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
   The expected profit will be **$21,987.43**



Expected_Profit
21,987

Alteryx Workflow shown below



```
Prob_to_buy =
([Score_Yes]*
[Score])
Profit =
([Prob_to_buy]
*0.5)-6.5
```