Project 1.2: Predicting Catalog Demand

The Business Problem

You 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 catalog in the coming months. The company has 250 new customers from their mailing list that they want to send the catalog to.

Your manager has been asked to determine how much profit the company can expect from sending a catalog to these customers. You, the business analyst, are assigned to help your manager run the numbers. While fairly knowledgeable about data analysis, your manager is not very familiar with predictive models.

You've been asked to predict the expected profit from these 250 new customers. Management does not want to send the catalog out to these new customers unless the expected profit contribution exceeds $10,000.

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?

The decision to be made is whether to send the catalogs to the 250 customer list based on expected profit.

2. What data is needed to inform those decisions?

To predict sales and expected profit we need following data

Type of customer
Average products purchased
Gross Margin
Cost of sending catalog to 250 customers

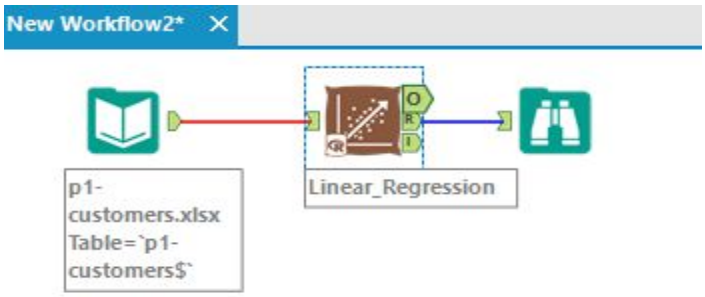
Step 2: Analysis, Modeling, and Validation

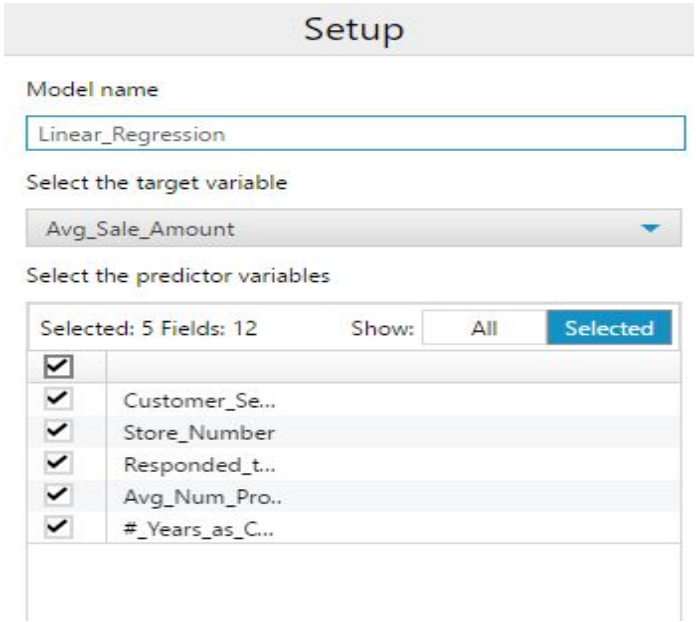
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

Important: Use the p1-customers.xlsx to train your linear model.

1. How and why did you select the predictor variables (see supplementary text) in your model?

A linear regression study is performed against Average Sale Amount.





Record Report

Report for Linear Model Linear_Regression

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Store_Number + Responded_to_Last_Catalog + Avg_Num_Products_Purchased + X._Years_as_Customer, data = inputs$the.data)

Residuals:

Min	1Q	Median	3Q	Max
-665.20	-67.82	-2.17	70.42	975.30

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	435.318	104.854	4.152	3e-05 ***
Customer_SegmentLoyalty Club Only	-150.224	8.971	-16.746	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.455	11.897	23.743	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-243.279	9.816	-24.784	< 2.2e-16 ***
Store_Number	-1.146	0.994	-1.153	0.2489
Responded_to_Last_CatalogYes	-28.085	11.253	-2.496	0.01264 *
Avg_Num_Products_Purchased	66.787	1.515	44.082	< 2.2e-16 ***
X._Years_as_Customer	-2.326	1.222	-1.904	0.05707 .

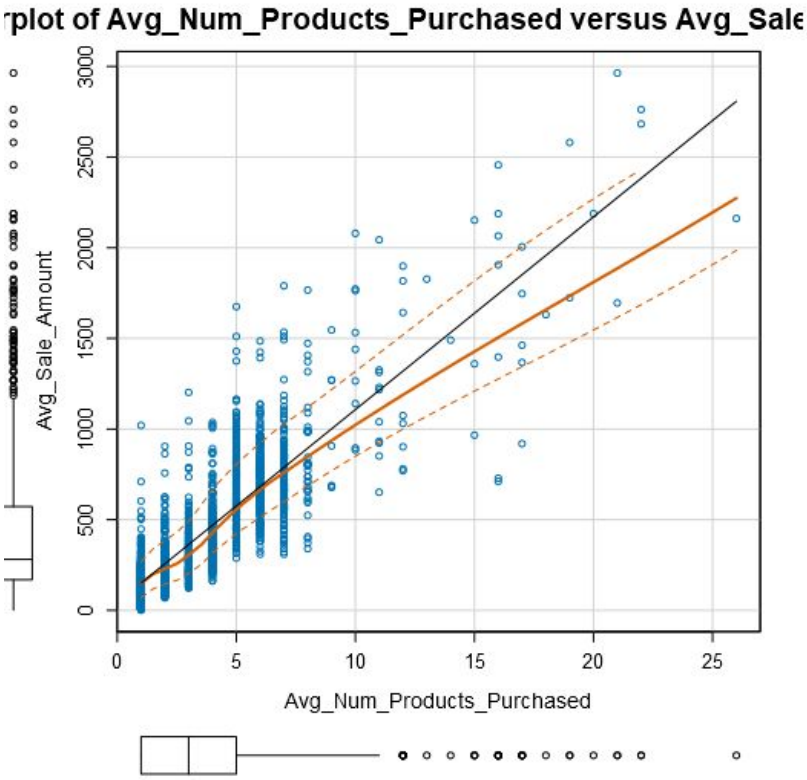
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 137.25 on 2367 degrees of freedom
Multiple R-squared: 0.8376, Adjusted R-Squared: 0.8372
F-statistic: 1745 on 7 and 2367 DF, p-value: < 2.2e-16

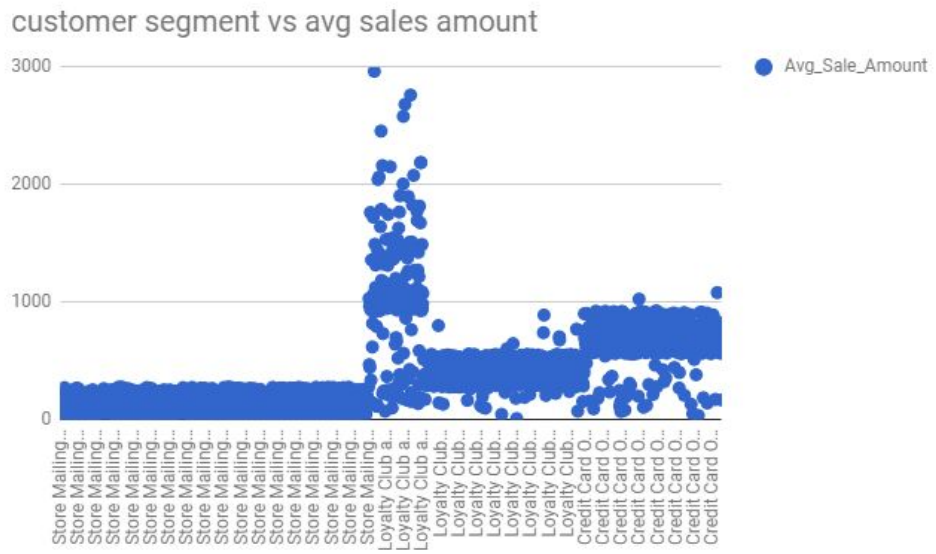
Type II ANOVA Analysis

Average Number of Product Purchased and Customer Segment have a p-value of less 0.05 which shows statistical significance. Scatterplots of Average Number of Product Purchased and Customer Segment versus Average Sale Amount are below.

This scatterplot is plotted in Alteryx



The scatterplot below is plotted in Google Sheets



2.. Explain why you believe your linear model is a good model.

Adjusted R-squared value of 0.8366 , which is a high value. Customer Segment and Average Number of Products Purchased have a p-value lower than 0.05, implying their statistical significance. The model is considered a good one for the above reasons.

Report

Report for Linear Model Linear_Regression

Basic Summary

Call:

lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = inputs$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

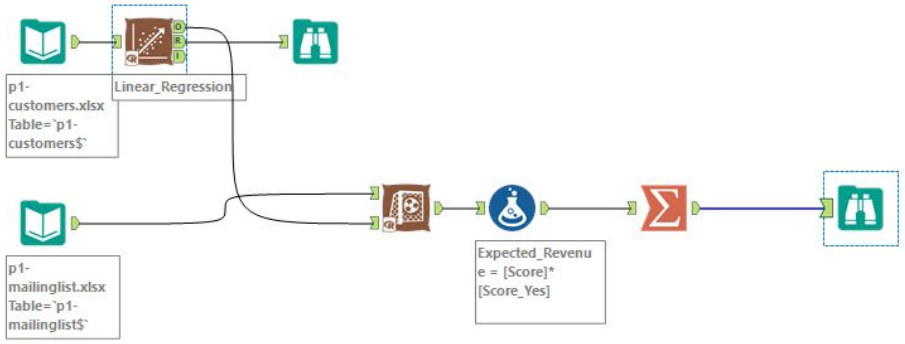
Residual standard error: 137.48 on 2370 degrees of freedom
Multiple R-squared: 0.8369, Adjusted R-Squared: 0.8366
F-statistic: 3040 on 4 and 2370 DF, p-value: < 2.2e-16

Type II ANOVA Analysis

3. What is the best linear regression equation based on the available data?

Avg_Sale_Amount = 303.46 – 149.36 x (If Type: Loyalty Club Only) + 281.84 x (If Type: Loyalty Club and Credit Card) – 245.42 x (If Type: Store Mailing List) + 0 x (If Type: Credit Card Only) + 66.98 x (Avg_Num_Products_Purchased)

Step 3: Presentation/Visualization



At the minimum, answer these questions:

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

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
Expected Profit >$10,000, so it is a good idea to send the catalogs.
The expected revenue from each customer is determined by multiplying expected sale amount with Score_Yes value.
With a gross margin of 50%, 50% is deducted from the sum of expected revenue before the cost of catalog ($6.50) is subtracted to obtain net profit.

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

Expected Profit = *(Sum of expected revenue x Gross Margin) – (Cost of Catalog x 250)*

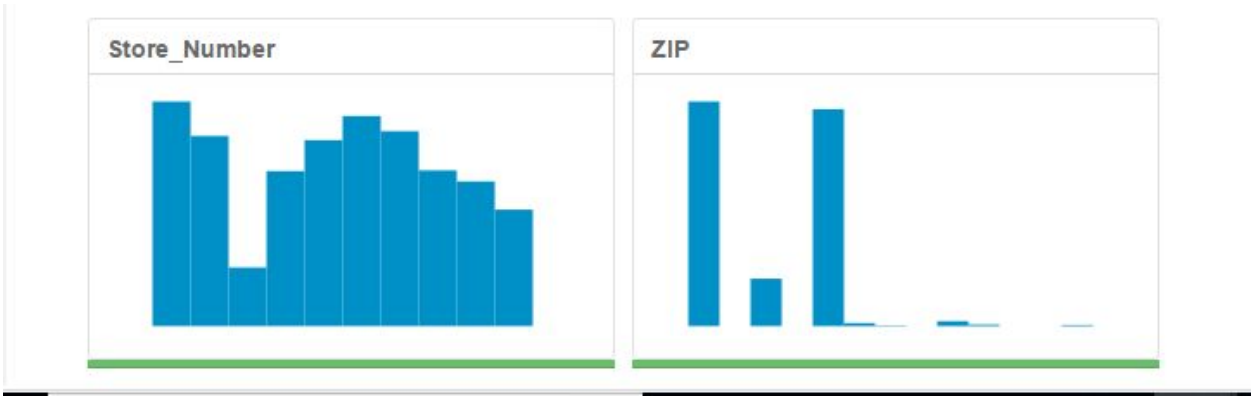
= *(47,225.87 x 0.5) – (6.50 x 250)*

= *23,612.44 – 1,625*

= *$21,987.44*

Variable Distribution

More data like which items were purchased by customers will be helpful in customizing the catalogs.


Store_Number
ZIP