Project 02

Predicting Catalog Demand

The Business Problem

I recently started working for a company that manufacturers 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.

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

I'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.

Details

- The costs of printing and distributing is $6.50 per catalog.
- The average gross margin (price - cost) on all products sold through the catalog is 50%.
- Revenue must be multiplied by the gross margin first before subtracted out the $6.50 cost when calculating profit.
- I must write a short report with recommendations outlining reasons why the company should go with recommendations to my manager.

Step 1, Business and Data Understanding

A description of key business decisions that need to be made. (500 word limit)

Key Decisions

Answer these questions

1. What decisions needs to be made?

   The key decision that needs to be made is whether or not to send annual catalog to 250 customers. Catalogs will be sent only if profit exceeds $10.000. Profit should be then predicted.

2. What data is needed to inform those decisions?

   There are two data sets. Both data sets contain Name, Customer_Segment, Customer_ID, Address, City, State, ZIP, Store_Number, Avg_Num_Products_Purchased, #_Years_as_Customer.

   - p1-customers.xlsx contains the data used to build the regression model based on 2.300 customer history.
     Avg_Sale_Amount and Avg_Num_Products_Purchased are specific variables from this dataset.
   - p1-mailinglist.xlsx contains the data related to 250 customers to whom will be calculated the prediction. The necessary data for the case study prediction was provided.
     Score_No and Score_Yes are specific variables from this dataset.

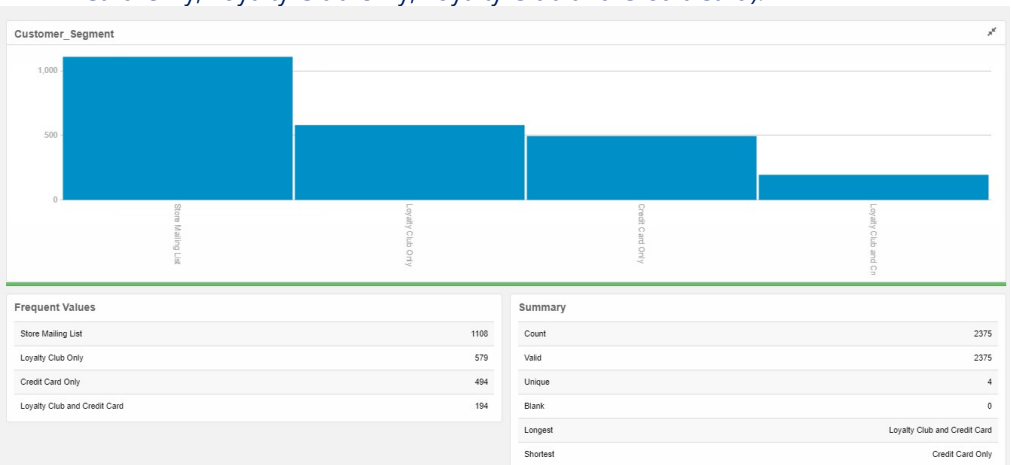
Step 2, Analysis, Modeling and Validation

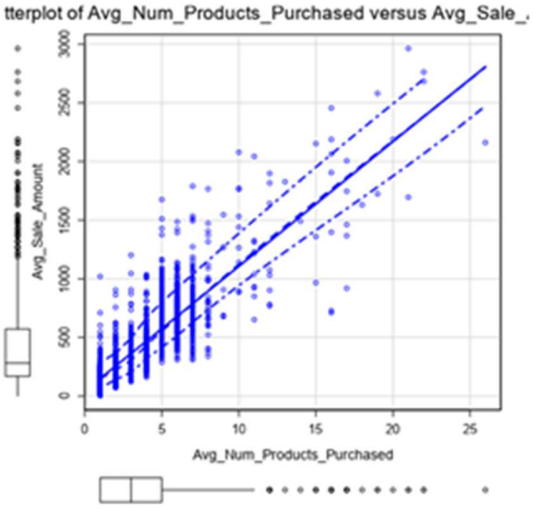
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)
Important: Use the p1-customers.xlsx to train your linear model.
At the minimum, answer these questions:

1. How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

- Once I downloaded the data set I analyzed them using alteryx, verified data types preparing to Multiple Linear Regression. And understood relationship between predictor variables and *target variable* = *Avg_Sale_Amount* .
- "Due to the nature of categorical variables, we cannot use a scatterplot or any other graph to see whether a linear relationship exists for categorical variables."
- Using *Field Summary* tool report I reinforcing the above concept I decided not to use some variables on prediction.
  - all the registers are from the same *State*;
  - *Address* and *Name*, each one has its own;
  - *Customer ID and ZIP*, despite numeric doesn´t make sense;
  - *City*; "deselect any categorical, non-numeric variable that contain more than four categories";
  - *Store_Number;*
  - *Years_as_customer,* I renamed the variable because of the "#";
  - *Responded_to_Last_Catalog*, is categorical and now I decided not to focus on this one.

- Considered the following as predictor variables. Avg_Num_Products_Purchased ↑, Avg_Sale_Amount ↑ increases too.
  - *Avg_Num_Products_Purchased*;
  - *Customer _segment*, considered despite categorical = 4 categories (Store Mailing List, Credit Card Only, Loyalty Club Only, Loyalty Club and Credit Card).





- Using *linear regression* tool I got equation and obtained multiple R-squared.

Report

Report for Linear Model LR_Catalog

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

Type II ANOVA Analysis

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		

2. 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.
   - It is a good model, adjusted R-Squared is **0.8369**. Values close to 1 means that variance in target variable is highly explained by the model.
   - Variables selected have good fit. P-values for all predictors variables are way less than 0.05. A p-value less than 0.05 (typically ≤ 0.05) is statistically significant. The smaller the p-value, the stronger the evidence that you should reject the null hypothesis.

3. What is the best linear regression equation based on the available data? Each coefficient should have no more than 2 digits after the decimal (ex: 1.28)

   Important: The regression equation should be in the form:

   *Y = Intercept + b1 * Variable_1 + b2 * Variable_2 + b3 * Variable_3……*

   For example: Y = 482.24 + 28.83 * Loan_Status – 159 * Income + 49 (If Type: Credit Card) – 90 (If Type: Mortgage) + 0 (If Type: Cash)

   Note that we **must** include the 0 coefficient for the type Cash.

   Note: For students using software other than Alteryx, if you decide to use Customer Segment as one of your predictor variables, please set the base case to Credit Card Only.

   Linear regression equation is

   Y = 303.46
   *– 149.36 * (Customer_SegmentLoyalty_Club_Only)*
   *+ 281.84 * Customer_SegmentLoyalty_Club_and_Credit_Card)*
   *– 245.42 * (Customer_Segment_Mailing_List)*
   *+ 66.98 * (Avg_Num_Products_Purchased)*

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 ***

Step 3, Presentation/Visualization

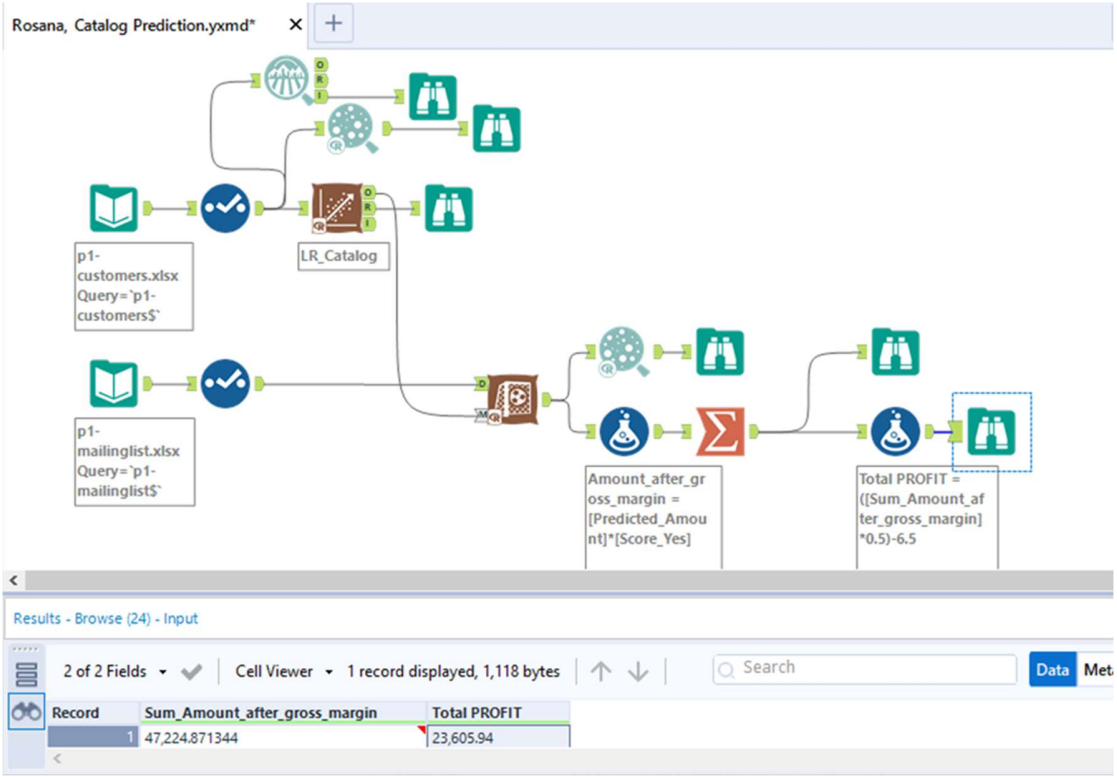
Use your model results to provide a recommendation. (500 word limit)
At the minimum, answer these questions:

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

   Yes, the company should send the catalogs to the 250 customers because profit contribution exceeds $ 10,000.

2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)
   First I understood business and data. Then I prepared data by through alteryx. The next step was analysis and modeling using linear regression tool. I obtained regression equation. Using score tool, multiple R-Squared I validated the model and here I am making it visual and presenting.

3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)?
   Expected profit is $ 23,605,94
   The following image refers to Alteryx workflow I developed.



Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.