PROJECT 1: PREDICTING CATALOG DEMAND

Step 1: Business and Data Understanding

1. **What decisions need to be made?**

Answer

The business is analyzing the expected profit gained by sending out this year's catalog to its 250 new customers. The decision needs to be made is whether to send out or not. The threshold to decide to send out the catalog is when the expected profit contribution exceeds $10,000.

2. **What data is needed to inform those decisions?**

Answer

Data that is needed

We need to calculate the total expected profit that the business can gain by sending out this new catalog to new customers. To do that, we need to study past sales data to identify which variables or factors that contributed to past revenue. Based on that, we will need data about new customers with variables that the model based on past data can be applied to in order to find the expected revenue.

Finally, to arrive at the expected profit, on top of the expected revenue, we will need data average gross margin on products sold, and costs of printing and distributing the catalog to customers.

Determining the right analytical approach

We use the following Methodology Map to decide to use a linear regression model.

First, we want to predict the profit gained from sending the catalog to 250 new customers. Therefore, we are looking to predict an outcome.

Second, since we have past data on the variables we are trying to predict, this problem can be classified as a data-rich problem.

Third, our target outcome that we're trying to predict is a number representing the expected profit from 250 customers. Therefore, we should use a Numeric Model.

Lastly, since we are not looking to forecast profits for a specific calendar week or a month, we will use a continuous model, specifically a linear regression model, to solve our problem.

Step 2: Analysis, Modeling, and Validation

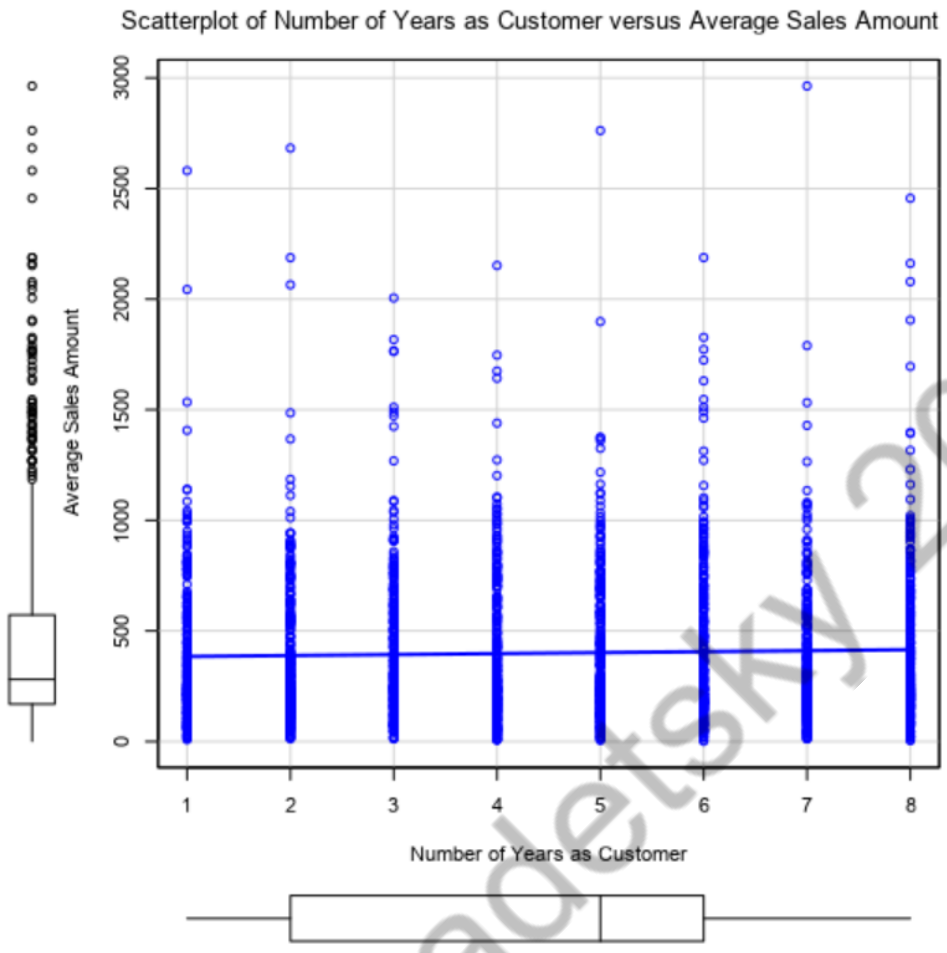
1. **How and why did you select the predictor variables in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer back to the "Multiple Linear Regression with Excel" lesson to help you explore your data and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

Answer:

For numeric variables, I use scatterplots between an individual variable and the target variable **average sales amount** to see if a variable might be a good candidate for a predictor variable.

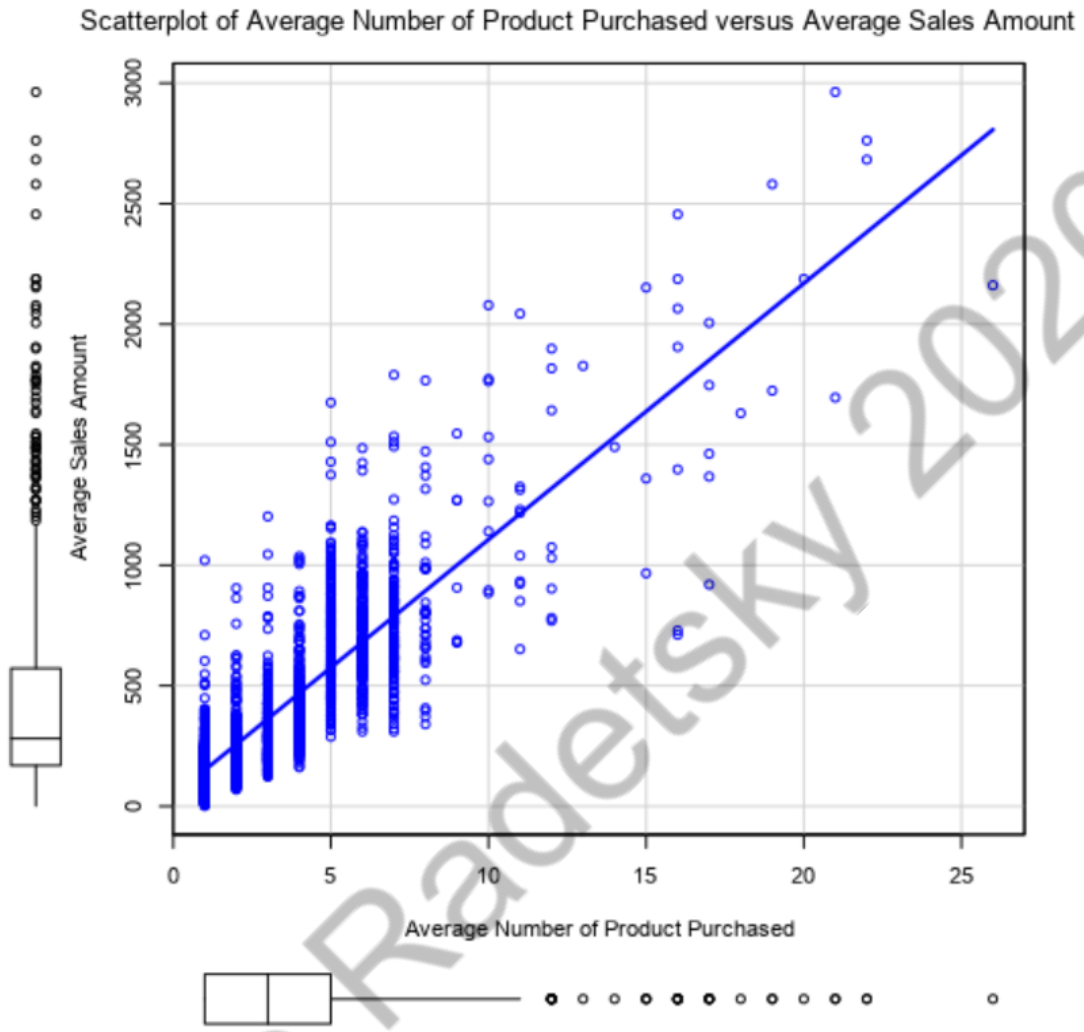
Two numeric variables that I will investigate are **Average Number of Products Purchased** and **Number of Years as Customers.**

For **Number of Years as Customers:**



It can be seen from the above scatterplot that the best-fitted line is slightly upward but almost flat. That means there is very little association between the variable **Number of Years as Customers** and **Average Sales Amount.** Therefore, I will not use **Number of Years as Customers** as the predictor variable.

For **Average Number of Product Purchased:**



It is evident from the above scatter plot that there is a positive linear relationship between **Average Number of Product Purchased** and **Average Sales Amount**. As the number of products purchased increases, the average sales amount also increases. Therefore, the variable **Average Number of Product Purchased** has a strong potential to be a predictor variable. Therefore, I will use this variable for our linear regression model.

For categorical variables, I will mainly use the trial and error method. However, I will eliminate unfit variables.

I eliminate the variable **Name**, **Customer ID**, **Address**, **ZIP**, and **Responded_to_Last_Catalog** because it is impossible to segment those variables. In addition, those variables are unique to the old data set and cannot be applied to the new data set.

I skip the **State** variable because there is only one state in question, which is Colorado (CO).

Therefore, the only two categorical variables left for investigation are **Customer Segment** and **Store Number.**

Plugging the variable **Store Number** into the linear regression model leads to the table below:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	314.1144	12.839	24.46628	< 2.2e-16	***
Customer_SegmentLoyalty Club Only	-149.6083	8.993	-16.63599	< 2.2e-16	***
Customer_SegmentLoyalty Club and Credit Card	282.9202	11.930	23.71494	< 2.2e-16	***
Customer_SegmentStore Mailing List	-245.6574	9.786	-25.10208	< 2.2e-16	***
Store_Number101	-4.5178	11.250	-0.40157	0.68803	
Store_Number102	-6.0276	16.751	-0.35983	0.71901	
Store_Number103	-4.1290	11.920	-0.34638	0.72909	
Store_Number104	-20.2725	11.318	-1.79124	0.07338	.
Store_Number105	-20.1168	10.968	-1.83418	0.06675	.
Store_Number106	-16.7502	11.186	-1.49740	0.13442	
Store_Number107	-12.9769	11.908	-1.08979	0.27592	
Store_Number108	-11.4332	12.178	-0.93885	0.3479	
Store_Number109	-0.1682	13.047	-0.01289	0.98971	
Avg_Num_Products_Purchased	66.8889	1.518	44.06305	< 2.2e-16	***

It is clear that the P-values of all dummy variables for store numbers are very large (> 0.05). The P-values for Store 102, 103, and 109 in particular are unacceptably large, being 0.72, 0.73, and 0.99, respectably. That means that the estimates for their corresponding coefficients are not statistically significant and may happen merely by chance. Therefore, I will not use **Store Number** as a predictor variable.

Plugging the variable **Customer Segment** into the linear regression model results in the below table:

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	***

The above linear regression result confirms that Customer Segment is a good predictor as the P-Values are extremely small.

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.

Answer

After plugging the two variables - **Customer Segment** and **Average Number of Product Purchased,** I obtained the following result:

Record | Report

1

Report for Linear Model Predict_average_sales

2 *Basic Summary*

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

4 Residuals:

5

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

6 Coefficients:

7

	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

8 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

9 *Type II ANOVA Analysis*

10 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

The model is great to predict the average sales amount because of a large value of R-squared (0.8369). That means the model explains more than 83% of the linear relationship between the two predictor variables and the target variable. The P-values for all predictor variables are extremely small, <2.2e – 16, significantly smaller than 0.05. That means that the estimates of coefficients are statistically significant and are very reliable to use as predictors.

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)

Answer:

Average Sales Amount = 303.46 + 66.98 * Avg_Num_Products_Purchased – 149.36 (If Customer Segment is Loyalty Club Only) + 281.84 (If Customer Segment is Loyalty Club and Credit Card) - 245.42 (If Customer Segment is Store Mailing List) + 0 (If Customer Segment is Credit Card Only)

Step 3: Presentation/Visualization

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

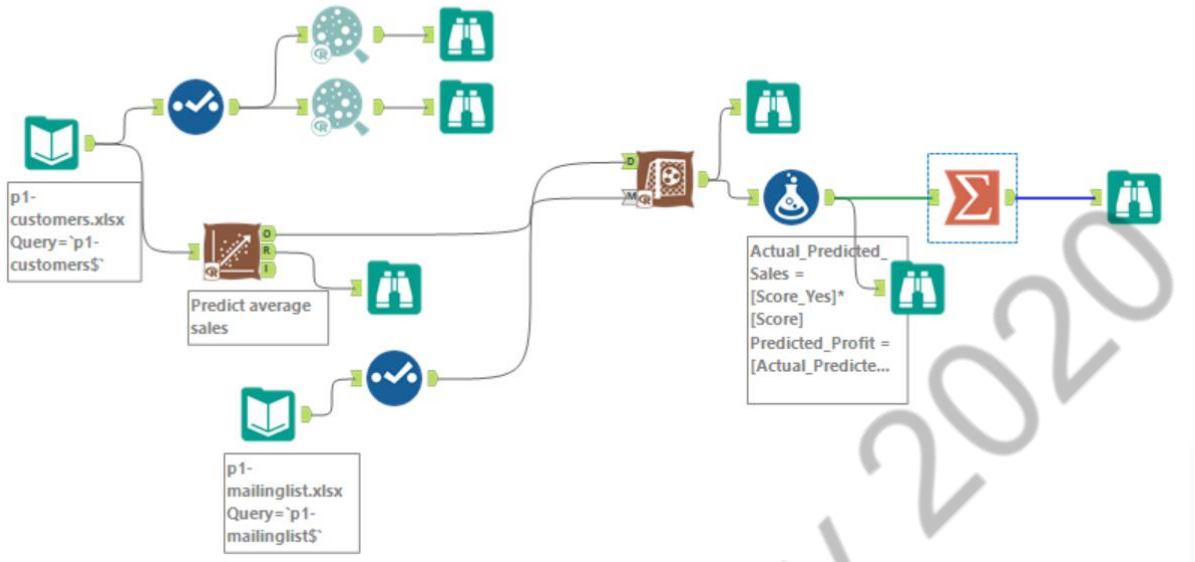
Answer

I recommend the company send the catalog to the 250 new customers because of the high expected profit.

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

Answer

I used Alteryx to build the workflow, including Linear Regression Model, Scoring the Model, and Calculating the expected profit.



Linear regression

First, I add the Input Data tool and load the "p1-customers.xlsx" file. It is the file of past data that is used to train the linear model.

Next, I add the Linear Regression tool. In its configuration pane, I plug two predictor variables that we have identified in Step 2 (which are **Customer_Segment** and the **Avg_Num_Products_Purchased**) and the target variable (which is **Avg_Sale_Amount**). For the categorical variable **Customer_Segment,** Alteryx by default codes it into three dummy variables, which are **Customer_SegmentLoyalty Club Only**, **Customer_SegmentLoyalty Club, and Credit Card**, and **Customer_SegmentStore Mailing List** while keeping **CustomerSegmentCredit Card Only** as the baseline variable and not included in the model.

Because the model has a good predictive ability with statistically significant, I will use this model to apply to our new customer list.

Score the model

After I have come up with the model and validate its predictive ability, I add the Score tool with its M node connected to the linear regression model (O node). The new mailing list, which is loaded into the Input Data tool from the "p1-mailinglist.xlsx" file, is connected to the D node of the Score tool. This step will result in a new column called "Score" to the data table; each record shows the predicted sales amount for each new customer.

However, we may not gain those whole predicted sales amounts because customers may choose to buy or may not. Therefore, I multiply each score with the probability that a customer will respond to the catalog and make a purchase, represented by the column Score_Yes. To do that, I used the **Formula Tool** and added the formula [Score_Yes] * [Score]. This will result in a column called **Actual_Predicted_Sales** that contains actual predicted sales gained from all 250 new customers.

To calculate the predicted profit from each new customer, I need to factor in all the costs. As such, I add the formula [**Actual_Predicted_Sales**] * 0.5 – 6.5. The number 0.5 represents the average gross margin on all products sold and the number 6.5 represents the cost of printing and distributing each catalog. This process results in a new column called Predicted_Profit that contains predicted profit amounts from all 250 new customers.

Calculating the expected profit

To calculate the total expected profit, I aggregated all values in the column Predicted_Profit by using the Summarize tool.

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

Answer

If the catalog is sent to these 250 new customers, the expected profit is $21,987.44, which is more than double the minimum that management requires.