Project 1: Predicting Catalog Demand

Business and Data Understanding

The main outcome is to predict whether a new catalogue release will generate enough profit to justify releasing the catalogue. We need to predict the amount of sales the catalogue would generate when it is sent out to the company's mailing list. Moreover, the company needs to make a 50% margin and to also generate sales greater than the business' own requirement of $10,000.

What decisions needs to be made?

The decision here is to decide whether to go with a new catalogue launch for their high-end home goods product range. There is a cost to producing these catalogues and an expected profit margin. The sum of the profits after costs and margin needs to also be greater than $10,000 in order for this catalogue to be launched successfully.

What data is needed to inform those decisions?

- Average sales amount
- Average number of products purchased
- Number of years as customer
- Store number
- Customer segment

Analysis, Modeling, and Validation

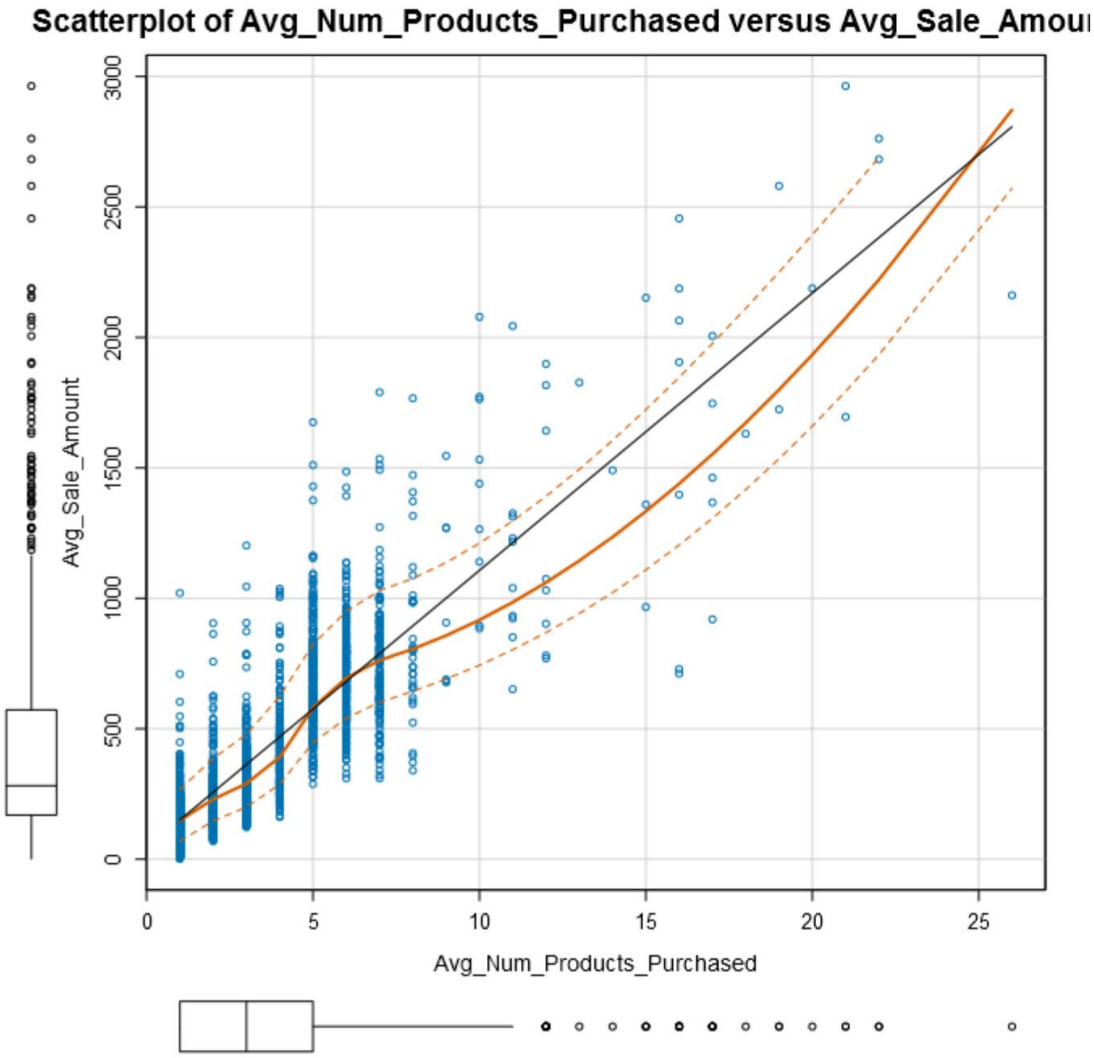
The Pearson's Correlation between all the predictor variables against the target variable (Average Sales Amount):

Pearson Correlation Analysis

Focused Analysis on Field Avg_Sale_Amount

	Association Measure	p-value
Avg_Num_Products_Purchased	0.8557542	0.000000***
Customer_ID	0.0382352	0.062455.
Years_As_Customer	0.0297819	0.146795
Store_Number	-0.0079457	0.698734

Clearly, Avg_Num_Products_Purchased is the best variable with the highest association measure, p-value of 0, and with three stars.



Customer Segment is a categorical variable found in our data and a variable that needs to be tested to see if there is any correlation with the target variable. Using Alteryx and running a linear regression using Average Number of Products Purchased and Customer Segment, we get the below summary:

Report for Linear Model multi_linear_reg

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 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		

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

Evidently, both variables are highly useful in predicting the average sales amount as shown by the adjusted R-Squared value and the three stars across the board.

Therefore, the linear equation is:

Predicted Sales = 303.46 + 66.98(Avg_Num_Products_Purchased) - 149.36(Customer_Segment Loyalty Club Only) + 281.84(Customer_Segment Loyalty Club and Credit Card) – 245.42(Customer_Segment Store Mailing List) + 0(Customer_Segment Credit Card Only)

Presentation/Visualization

To obtain the actual profit, for each customer, we first multiply the predicted sales with the probability of the customer buying the products (the “Score_Yes” variable), and then

multiply that value with 0.5 (which accounts for the margin), and then finally subtract $6.50. After summing all the individual profits, we get the total profit as $21,987.44.

Since this value is greater than $10,000 that was the baseline for the company, I would recommend that the company go ahead with the catalogue launch. This recommendation can be trusted as the model built above is very robust and accurate.