Project 1: Predicting the Profit Made through Investigating the Catalog Demand using Linear Regression Modelling

By Ivy Huong Nguyen
The workflow for this analysis is demonstrated in Figure 3

Step 1: Business and Data Understanding

A high-end home goods manufacturer started sending out their catalog to their customers with the aim to increase the company's profit. This company has a total of 250 new clients who just got added to their mailing list. The manager of the company needs to decide whether it is worth sending out their catalog to the new 250 clients. In order to make this decision, the manager needs to know if the predicted profit from their 250 clients would be exceeded $10000, which is the baseline for them to make the final decision. In this data analysis, we will have to build a multiple variable linear regression model using a given dataset with the historical information of all customers. We then apply this model to a new dataset which includes the information of the new 250 clients. In order to calculate the profit, we will need to multiply the predicted average amount of purchase from each new client by the average gross margin (50%) before subtracting the cost of printing and distributing a catalog. We then sum all the profit of the new 250 clients to determine if the total profit is above $10,000.

Key Decisions:

1. What decisions needs to be made?
   - Determine if the total profit of the new 250 clients is above $10,000
   - Decide whether it is worthwhile for the home goods company to send out its catalog to the new 250 clients based on the total profit made from these new clients.
2. What data is needed to inform those decisions?
   - The actual profit from each new client, which is determined by:
     - Which customer segment the customer belongs to? (Customer_Segment)
     - How long has the customer been with the company? (#_Year_as_Customer)
     - The probability the new customer will buy from the catalog (Score_yes)
     - Cost of printing and distributing a catalog

Step 2: Analysis, Modeling, and Validation

1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

The predictor variables were first chosen by using human logic. After looking through all the fields in the given dataset, I paid very close attention to three different columns: the number of years as customer, the customer segment, and the average number of products purchased. I think the chance that the customer would buy the company's products is strongly dependent on which customer segment they belong to and the average number of products purchased previously by them. The number of years that a customer stays with the company is also potentially another key feature which could determine the probability they would buy the product. I then started my analysis by generating two different scatter plots:

- Scatter plot 1 (Figure 1) shows the relationship between the average sale amount (y axis) and the number of years as customer (x axis). The plot is grouped by customer segment.
- Scatter plot 2 (Figure 2) demonstrates the relationship between the average sale amount (y axis) and the average number of products purchase (x axis). This plot is also grouped by customer segment.

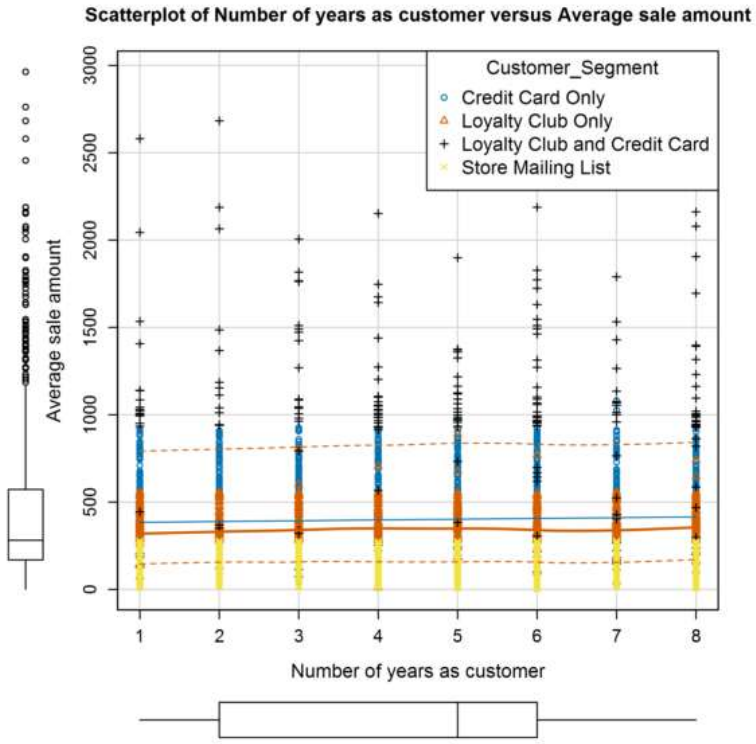

Figure 1. Scatter plot shows the relationship between the average sale amount and the number of years as customer, grouped by customer segment.

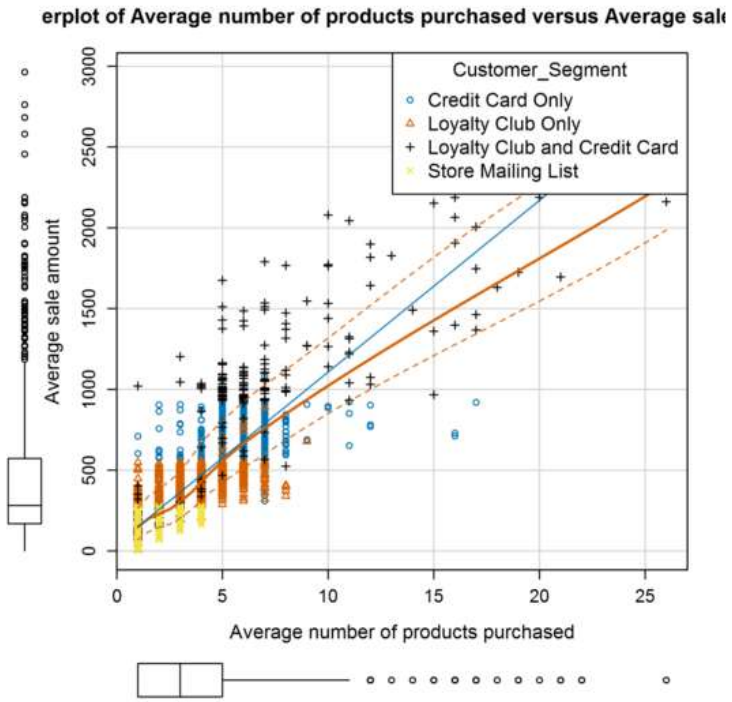

Figure 2. Scatter plot shows the relationship between the average number of products purchase and the average sale amount

While Figure 2 showed a somewhere linear pattern with high variation between the target variable (average sale amount) and the predictor variable (the average number of products purchased), Figure 1 did not quite point out the same linearity between the chosen predictor (the number of years as customer) variable and the target variable. However, one thing that should be noted here is that the larger average sale amount tends to belong to the 'credit card only' group. These two scatter plots made me think that: whereas the number of years did not show any significant linear relationship to the target variable, the two features: customer_segment and the average number of product purchased must have some influential significance the average sale amount. To fully determine whether or not I should eliminate the 'number of years as customer' variable in my model, I decided to generate two multivariable linear regression models and then compare how well they fit using adjusted R^2 and p-values.

- Predict_revenue_1 regression model would use average number of products purchased, customer segment and the number of years as customer as predictor variables.
- Predict_revenue_2 regression model would only use average number of products purchase and the customer segment as predictor variables.

After generating the two models using Alteryx, I noticed the two models showed that all the chosen parameters are statistically significant based on their p-values (Table 1). However, there is slightly or no difference in the adjusted R^2 values between the two models (Table 1). I thus decided to use

the less complicated model (predict_revenue_2) to predict the average sale amount of the new clients.

Table 1. Obtained adjusted R^2 values and p-values from the two linear regression models.

	p_values			Adjusted R^2
	Avg_product_purchased	Customer_Segment	#_years_as_customer	
Predict_revenue_1	<2.2E-16	<2.2E-16	<2.2E-16	0.8368
Predict_revenue_2	<2.2E-16	<2.2E-16	N/A	0.8366

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.

If the p-value of each predicted variable is less than 0.5, then there is a statistical significance between that variable and the target variable. Since the p-value is less than 0.5 (Table 1), the chosen two predictors for the predict_revenue_2 model appear to have a significant statistical relationship with the target variable. In addition, I also took a look at the value of the adjusted coefficient of determination, which is 0.8366 and greater than 0.7. This value indicates the target variable is well-explained by the chosen predicted variables.

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)

Avg_Sale_Amount = 303.46 -149.36 (if Loyalty Club Only) + 281.84 (if Loyalty Club and Credit Card) - 245.42(if Store Mailing List) + 0 (if Credit Card only) + 66.98*Avg_Num_Products_Purchased

Step 3: Presentation/Visualization

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

I highly recommend the company to send out its new catalog to the new 250 clients since the predicted total profit from these client is $21987.4, which is exceeded the $10,000 baseline.

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

I came up with the recommendation by comparing the baseline profit that the company set at the beginning to the calculated total profit from the new 250 clients that I obtained from my model. The total profit is determined by the following steps:

- Construct the multivariable linear regression model using customer segment and the average number of products purchased as two key features.
- Validate the model by looking at the p-value for each predicted variable (whether the p-value < 0.5 or not) and the adjust R^2 value.

- Use the obtained model to predict the average sale amount for the new 250 clients by applying the constructed model to the p1_mailinglist.xlsx file.
- The profit for each new client is then calculated by subtracting the cost of printing and distributing a catalog from the product of the average sale amount and the average gross margin.
- The total profit is the sum of the calculated profit from the new 250 clients. This sum is then compared to the baseline of $10,000 to make the final recommendation.

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

The expected profit from the new catalog would be $21987.4

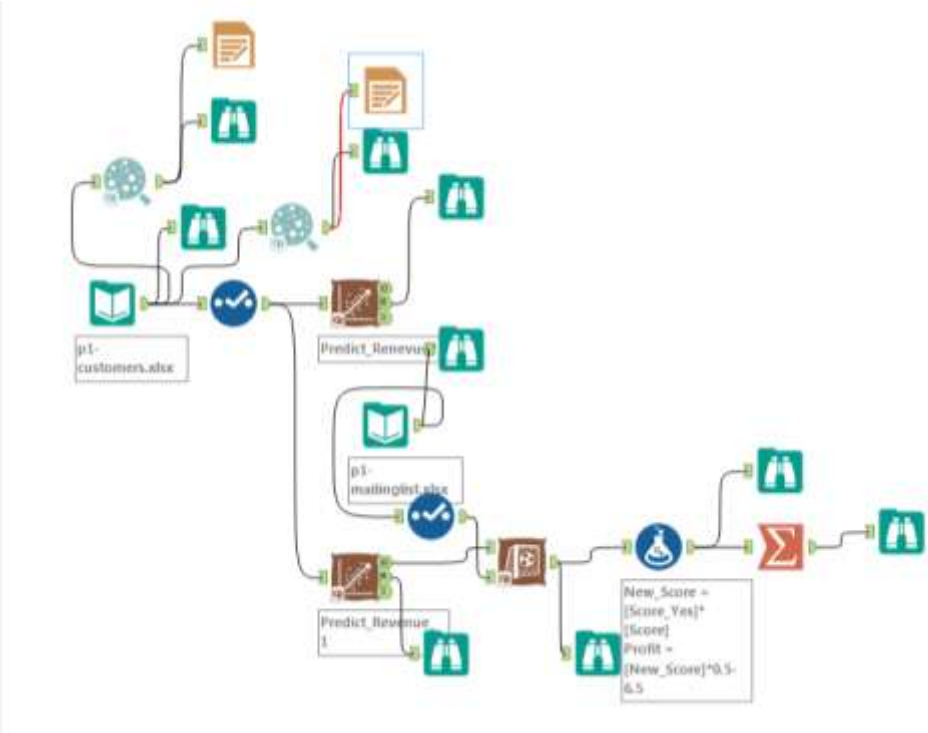

Figure 3. Workflow from Alteryx