PROJECT 1

Business and Data Understanding

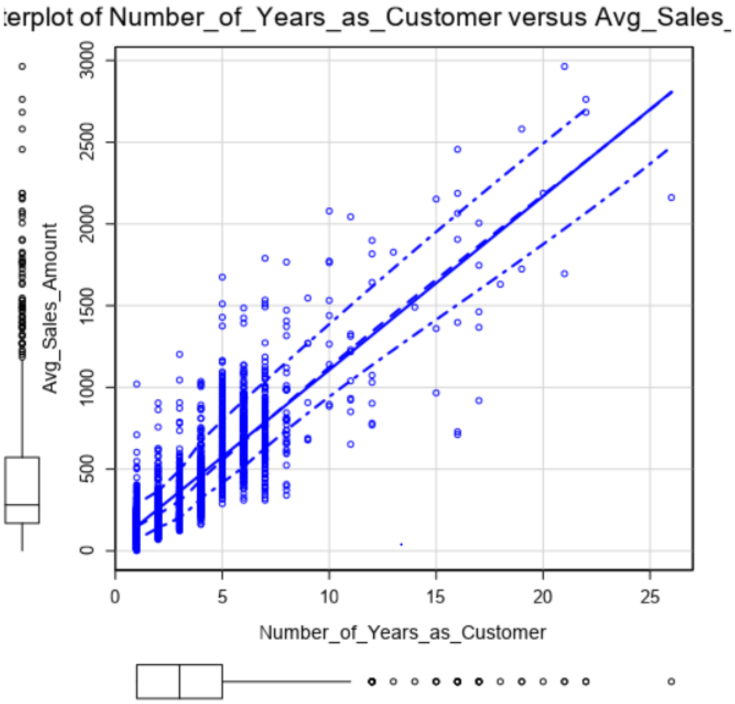
Business Objective :- A high-end goods company is preparing to launch a catalog in the coming months. The company currently has 250 new customers from their mailing list who can be potential customers. Also, the Business Analyst needs to predict the expected profit of these 250 new customers. The new customers will be sent the catalog, if their overall **expected profit** sum crosses $10,000.

Data Understanding :- Data contains of an old mailing list(previous customers) and a new mailing list(potential customers). Both the tables contain the same metrics, except the old mailing list contains the average sales amount for customers and the new list does not. The linear model is trained on the old mailing list by considering "average sales amount" as the target variable. Using this model, the predicted average sales amount for the new customers is calculated. To achieve the expected price, we multiply the predicted sales amount by the gross margin and the probability the person will respond to the last catalog launched i.e. P(Yes). Since profit = (price – cost), we finally subtract $6.50 from price to predict profit. After getting this value we sum the profits for all 250 customers.

Data Preparation :- For this dataset, the data is cleaned. Considering the data present, to predict the average sales amount we consider the explanatory variables as "number of years as customer"(numerical) since it has a high positive correlation with the target variable and Customer Segment[Store Mailing List, Loyalty Club, Credit Card) since Alteryx can handle dummy variables without the need to binary encode the first. The rest of the numerical variables are not linear to the target variable. Hence, these variables are not considered in the linear model.

Analysis, Modeling, and Validation

***Analysis* :-** While analyzing the data, we used scatter plots to determine the predictor variables for the target variable in the linear model. Categorical data was determined by trial and error method based on P-value (P-value <= 0.05 is statistically significant for the model). The target variable in the old mailing list(avg. Sales amount) has a minimum value of -663.8 and a maximum value of 971.9.



Predicted_Avg_Sales_Amount = 303.46 + (-149.36 * Customer_Segment Loyalty Club Only) +

(281.84 * Customer_SegmentLoyalty Club and Credit Card) + (-245.42 * Customer_SegmentStore Mailing List) + (66.98* Avg_Num_Products_Purchased)

Modeling :-The above eqn. is the linear equation produced by the linear model. The first value is the intercept, it also represents the predicted avg. sales amount if the customer segment is Credit Card, provided all other variables are constant. The following are the predictor variables.

Numerical – Avg_Num_Products_Purchased
Ordinal – Customer Segment

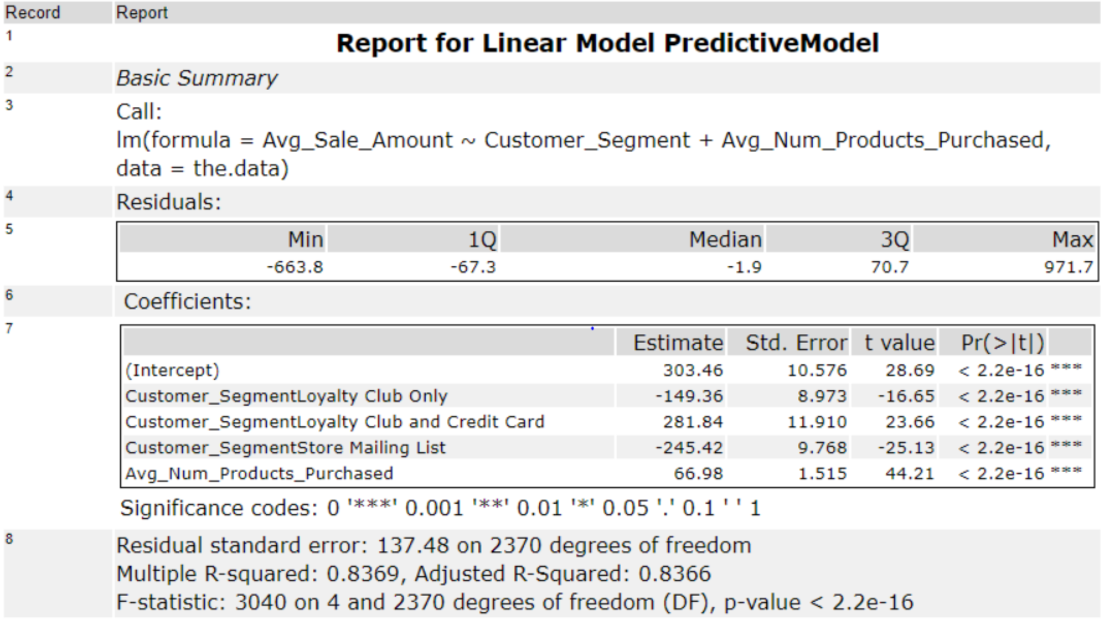
The coefficients of the categorical variable tells us:-

1) Avg_Sales amount for loyalty club only is $149.36 less than that of credit card only, when all other variables are constant
2) Avg_Sales amount for loyalty club and credit is $281.84 more than that of credit card only, when all other variables are constant.
3) Avg_Sales amount for mailing list is $245.42 less than that of credit card only, when all other variables are constant.

The coefficients of the numerical variables tell us,

1) For one unit increase in the average number of products purchased, the average sales amount increases by $66.98. Keeping all other variables constant.

Validation **:-** The p-value for all the coefficients are less than 0.05 and are statistically significant to the target variable. Also, suggesting a relationship between the target and predictor variables. The adjusted R-squared value is 0.8366. A higher adjusted R-square value (close to 1) suggests the predictor variables can explain the variability in the target variable with more clarity.



Record Report

1 **Report for Linear Model PredictiveModel**

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:

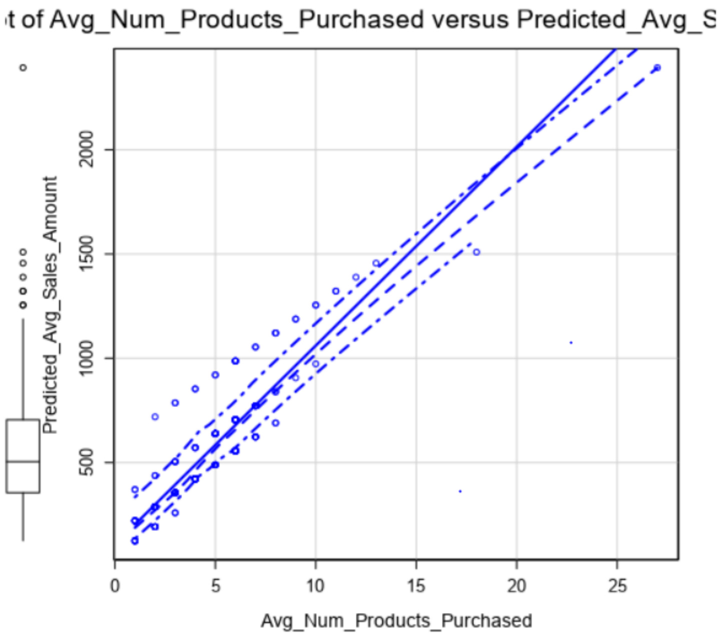
7

	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

Therefore, the above model has predictor variables with low p-value and a high r-square value, telling us that the model is good for predictions.



The above figure shows that the data is more compact as compared to the model built with the training set. The variation in the training model shows that the confidence intervals are more spread throughout the distribution, while in the test model the confidence interval range is smaller.

Presentation/Visualization

A seven step process representing the ***Problem Solving Framework*** is followed :-

1) Understanding Business Issue/Needs
2) Data Understanding
3) Data Preparation
4) Data Analysis
5) Data Modelling
6) Data Validation
7) Presentation/Visualization

All questions have been resolved. Using the Score Tool in Alteryx, the average sales amount is predicted. To predict the final profit, we need to consider the following 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%.
- Multiple revenue by the gross margin first before subtracting the $6.50 cost when calculating profit.

The following equation is applied to calculate profit for one customer-

([Predicted_Avg_Sales_Amount]*[Score_Yes] * 0.5) - 6.5

After we predict the sum price for all customers, we get a value of $21,987.44. This fulfills the criteria of crossing $10,000 and hence, these it is profitable for the company to send out this year’s catalog.

Recommendation

The linear model allowed us to reach a solution using metrics from a dataset. These metrics are first analyzed to be valid in predicting a continuous target variable. Once these variables are valid, they can be applied to the linear model. A linear model, with high R-square value is needed for a model to be a good predictor and low P-values for the predictor variables is needed for a model to explain the relationship of the coefficients with the target variable.

This model can be used to predict the sales amount of a large number of customers since the data is more compact.