Multiple Regression in R

We have been tasked to predict the sales of 4 different product types while assessing the effects service and customer reviews had on sales. Moreover, we utilized three algorithms to assist in this task, and chose the most optimized model to support my finalized predictions. We will go into detail explaining my results and proceed to provide recommendations to the sales team.

Exploratory Analysis:

- The provided data set had 80 rows (observations) and 18 columns (variables)

- The structure of the data set was all Numeric but Product Type which was labeled as a character.

- There was 15 NA's in the data set.

- Product Number is an integer attribute, also known as unique Identifier.

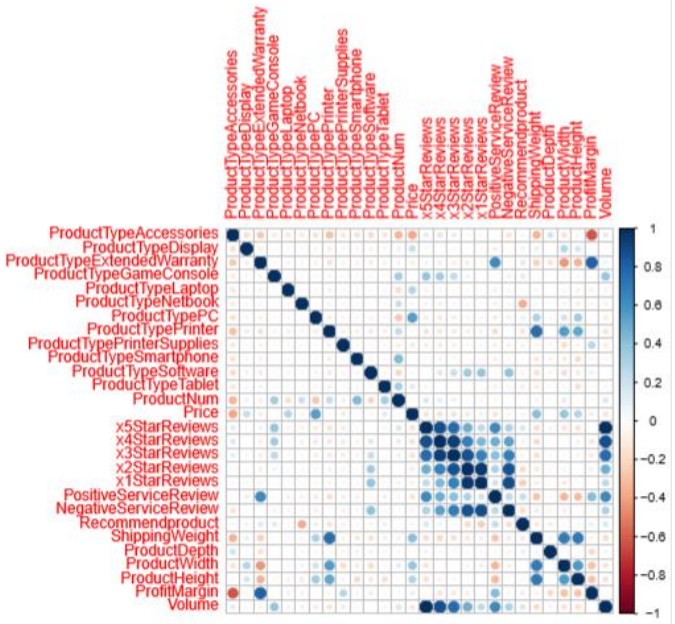
Preprocessing Data:

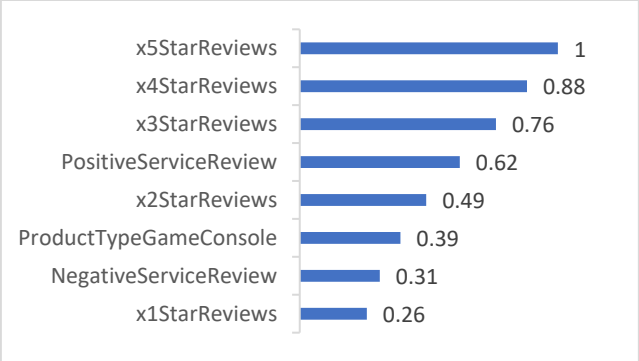
A) Converting Non-Numeric Data to Numeric Data:
Because we were tasked to utilize regression models to predict sales, we cannot perform regression analysis on nun-numeric variables. Consequently, we needed to dummify the data. Dummifying the data converts categorical variables to binary variables.

B) Removing NA's:
As earlier indicated, there was 15 NA's in the data set. We were instructed to remove missing data.

Feature Selection:

A) Correlation:
Correlation is a measure of the relationship between two or more features/variables. In order to measure the correlation between variables, we first had to ensure that all relevant data did not contain nominal data types. Upon completing this, we started our analysis by visualizing the correlation between the relevant independent variables and the dependent variable (Volume). The visualization to the left (correlation heat map) can be explained by the darker and larger the circle, the more correlated the independent variable is to the dependent variables. Visually, we can see the customer service reviews *look* highly correlated.





As you can see in the figure to the left, we have the top 8 most correlated variables with Volume sorted from highest to lowest. 5 Star Reviews is perfectly correlated with Volume. This will cause any model to overfit. This variable will be addressed later part C of this section.

Moreover, any correlation between the absolute values of .5-1. are considered strong, .3-.5 medium, and .1-.3 small.

B) Collinearity:

Collinearity is checking moderate to high intercorrelations among independent variables. The problem with collinearity is if 2 or more independent variables contain essentially the same information to a large extent, one gains little by using both in the regression model. This leads to unstable estimates as it tends to increase the variances of regression coefficients. Solution is to keep only one of the two independent variables that are highly correlated in the regression model. In the correlation heat map, we can visibly see some strong correlations primarily between customer service reviews. Below is the correlation matrix between the relevant independent variables.

	x4StarReviews	x3StarReviews	x2StarReviews	x1StarReviews
x4StarReviews	1.0000000	0.9372142	0.6790056	0.4449417
x3StarReviews	0.9372142	1.0000000	0.8614800	0.6792762
x2StarReviews	0.6790056	0.8614800	1.0000000	0.9519130
x1StarReviews	0.4449417	0.6792762	0.9519130	1.0000000

X3StarReviews and X4StarReviews are highly correlated. Also, X2StarReviews and X1StarReviews. Collinearity will be addressed in the next of this section C.

C) Variable Selection

To optimize models, we had to remove a few columns:

1) Product Number – Unique Identifier and does not add any added value for analysis (i.e. reduce “noise”)
2) X5StarReviews – Perfectly correlated with Volume. Need to remove due to overfit.
3) X3StarReviews – Collinearity with X4StarReviews. Did not remove X4StarReviews due to X4StarReviews having a stronger relationship to Volume.
4) X1StarReviews - Collinearity with X2StarReviews. Did not remove X2StarReviews due to X2StarReviews having a stronger relationship to Volume.

Models Training

With the data set preprocessed and cleaned, we were ready to build our models and choose the best performer. But before performing each model, we first had to set seed to our model so that the results could be reproduced. Also, modified data set was split into 75% training data and 25% testing. The models were trained on the 75% training data and the predictions were made on the 25% test data.

Lastly, we used 10-fold cross validation to evaluate each model. Below are the best results of the models trained.

Model	RMSE	Rsquared	MAE
Support Vector Machine	227.07	0.94	177.30
Random Forest	630.35	0.96	285.24
Gradient Boosting	993.61	0.87	580.62

Support Vector Machine performed the best in our trained model. However, upon testing the models through predictions via the test set, we found that only Random Forest did not produce negative predicted values for Volume. Consequently, we choose this model to further carry out our task.

Predictions:

Upon predicting volume, we needed to first perform the same steps performed with the earlier dataset with the current data set (dummy variables, remove NA's, Correlation and collinearity). After completing those steps and utilizing the Random Forest model on the new data set, predictions were computed. Below are the predicted values that my model performed for volume (for the Product types Danielle wanted):

Product Type	Predictions	Profit
Netbook	1767	$ 164.81
Smartphone	1240	$ 145.94
PC	690	$ 162.48
Laptop	313	$ 36.30

Per Danielle's request, to the left are the 4 different product types she wanted to see along with the profit. It is recommended that the sales team primarily focus on PC & Notebook since they both had the highest profit of the 4 product types that were of interest.