Multiple Regression Informal Report

### Overall Impression

In this multiple regression task, first I followed the plan of attack to convert all factor data to binary features that contain ‘0’ and ‘1’ classes, a process called ‘Dummify’. I then used the summary() function to check for the missing value and removed BestSellersRank attribute from the original existing products dataset.

Next, I used the packages “corrplot” to find out the correlation (heat map) between these features. From here, I did not remove any more features from the existing products dataset. Instead, I took the master dataset that contains 28 features as the source, created 5 different sub datasets based on my own understand and judgement of the data. These 5 sub datasets are then used to test out 5 models. This allows me to try out different feature combinations against models to select the most optimal result.

At the end, primarily rely on the values of RMSE and R Square for each model and sub dataset combination, I chose Random Forest to generate my final prediction. I have used several commonly used R commands to train and test models including trainControl(), train(), predict(), and postResample().

**Sub Datasets Summary**

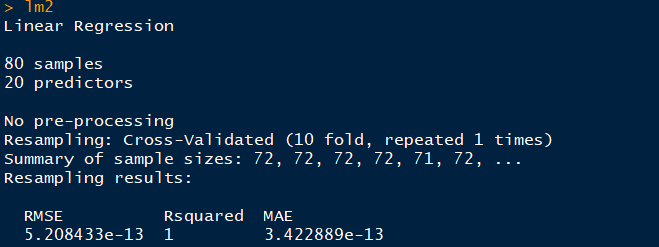
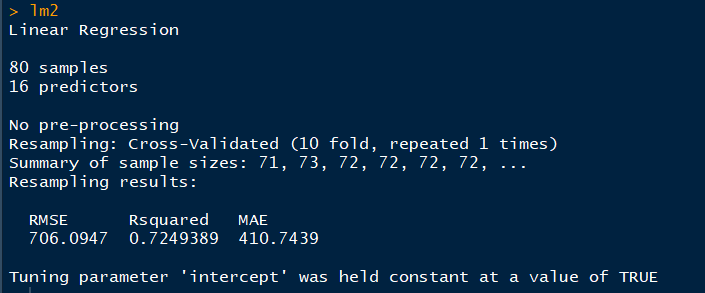
|  |  |
| --- | --- |
| Sub Dataset | Features |
| newData1 | All 12 dummies Product Types, and all other features except the BestSellersRank |
| newData2 | All 12 dummies Product Types, price, 4 Star Reviews, 2 Star Reviews, Positive Service Reviews, Negative Service Reviews, Recommend Product, Shipping Weight, and Volume |
| newData3 | All 12 dummies Product Types, price, **5 Star Reviews**, 4 Star Reviews, 2 Star Reviews, Positive Service Reviews, Negative Service Reviews, Recommended Product, Shipping Weight, and Volume |
| newData4 | All 12 dummies Product Types, 4 Star Reviews, 2 Star Reviews, Positive Service Reviews, Negative Service Reviews, Shipping Weight, and Volume |
| newData5 | All 12 dummies Product Types, 4 Star Reviews, 2 Star Reviews, Positive Service Reviews, Negative Service Reviews, Volume |

*In order to keep this report concise, only screenshots using feature selections described in* ***newData3*** *and* ***newData5*** *are included in the answers and discussions below. The full results using other sub datasets are in the Excel file attached separately.*

### The algorithms you tried and the results of each model you constructed, exported from R.

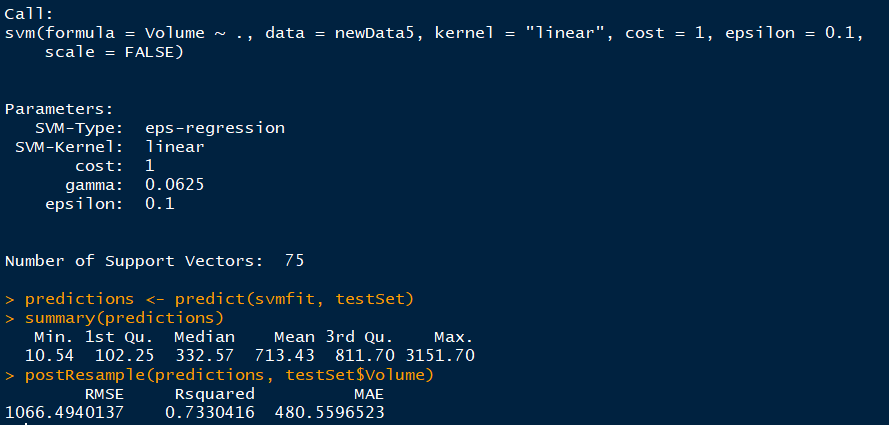
I have applied carets and other algorithm methods on Linear Model, Support Vector Machines (2 ways of building the model), Random Forest, and Gradient Boosting Machines (see below screenshot for each)

I have exported the R script, newproductforcase.Rproj, and C2.T3output1.csv file in the attached zip file.

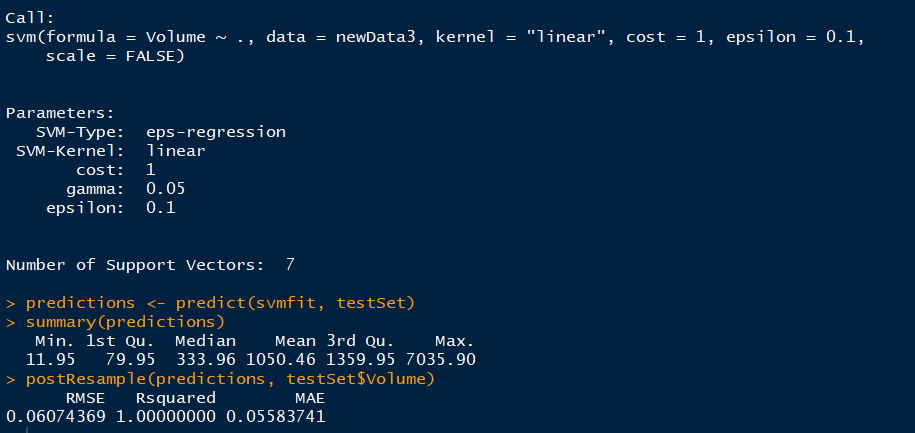
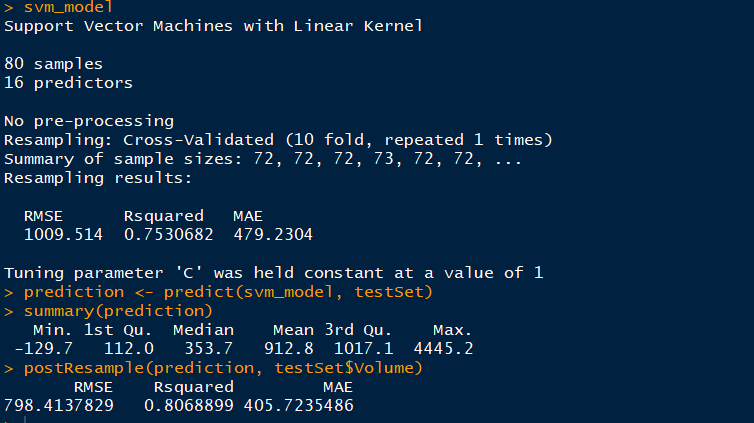
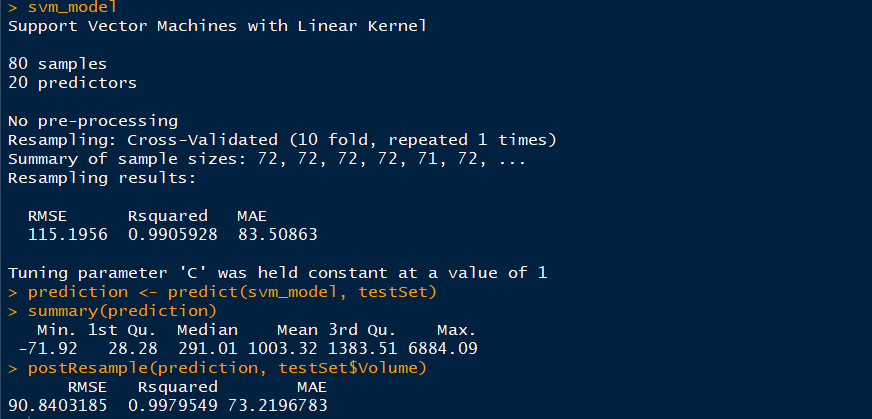


Linear Model newData5 (without 5star reviews)

Linear Model newData3—include 5star reviews



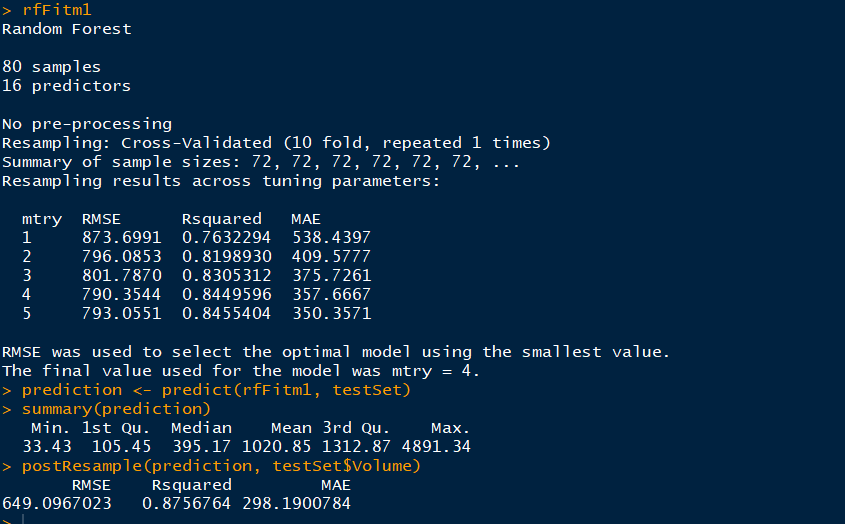
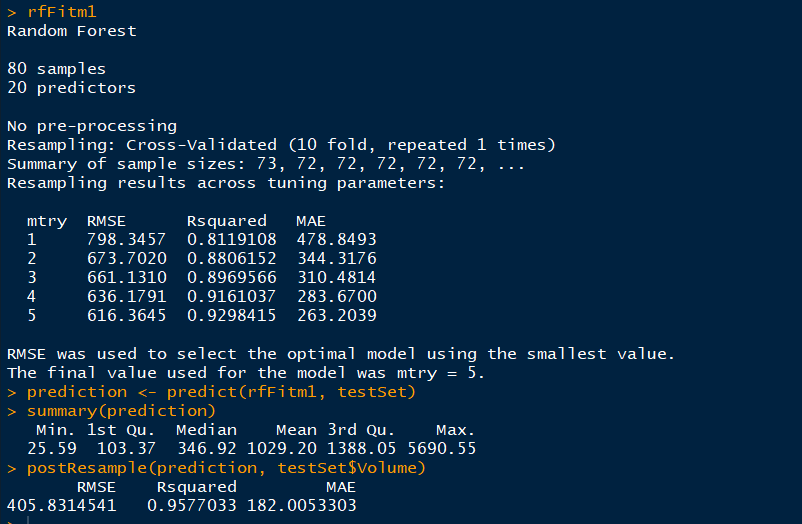
SVM kernel linear newData5 (without 5star reviews)



SVM Model using caret newData5 (without 5star reviews)

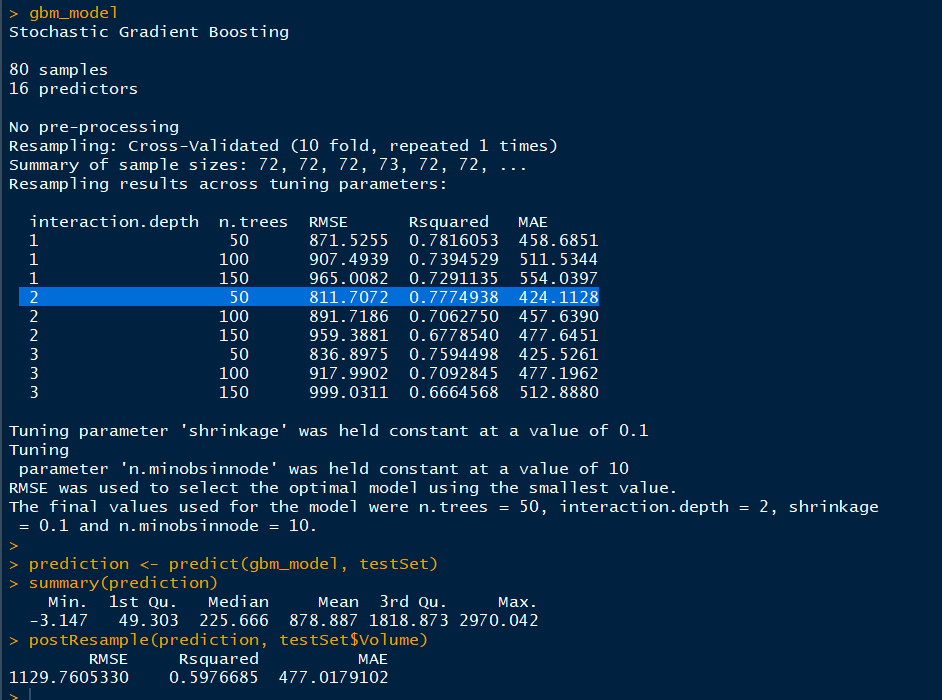
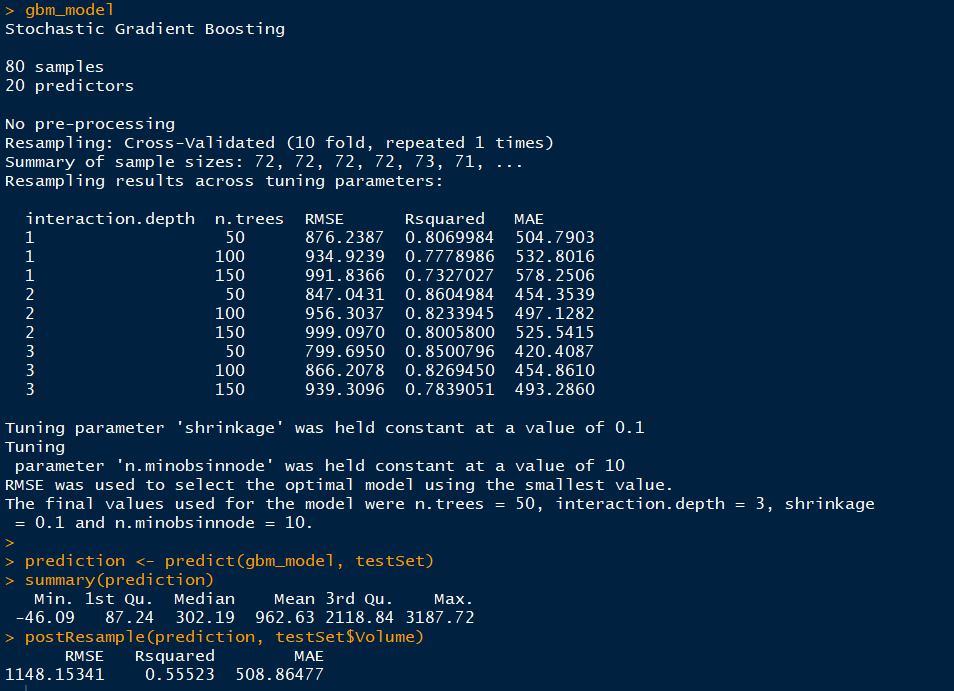
SVM Model using caret newData3 (include 5star reviews)

SVM kernel linear newData3 (include 5star reviews)



Random Forest Model newData3 (include 5star reviews)

Random Forest Model newData5 (without 5star reviews)



Gradient Boosting newData3 (Include 5star reviews)

Gradient Boosting newData5 (without 5star reviews)

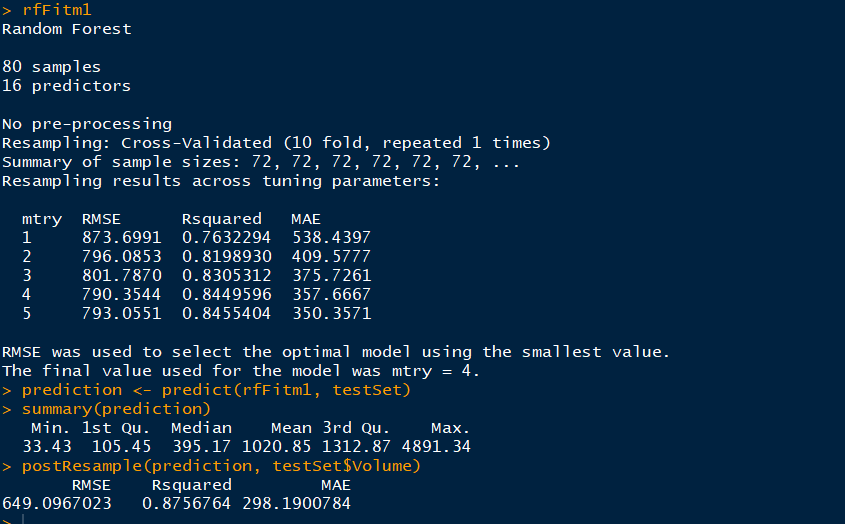
### The algorithm you selected to make the predictions, including a rationale for selecting the method you did and the level of confidence in the predictions.

There are total 5 algorithms tried: Linear Regression, Gradient Boosting Machines, Support Vector Machines (2 ways of building the model), and Random Forest. I used a 2-step selection process described below to locate the best algorithm.

First, I run all the algorithms against the various sub datasets to get a general feeling of their results in solving this problem. When viewing the results, it shows that best RMSE and R Square results of other algorithms are worse when comparing with Random Forest. Therefore, I removed the other algorithms from the follow-up tuning and focused my effort on the Random Forest only.

In the second step I focused on tuning the Random Forest result with different sub dataset and mtry values. At the end, using the sub dataset newData5 with mtry = 4, I have the best result where the RMSE value is 790.3544 and R square value is 0.8449596. The prediction on the testSet are RMSE: 649.0967023 and R square: 0.8756764.

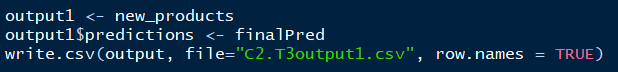
As we can see from the screenshots included earlier, models that includes 5StarReviews attribute are overfitting. Their RMSE values are close to 0 or very low (below 300) and R square value are close to or equal to 1. This is a strong sign of overfitting. Therefore, I did not include the 5StarReviews feature in my final training.

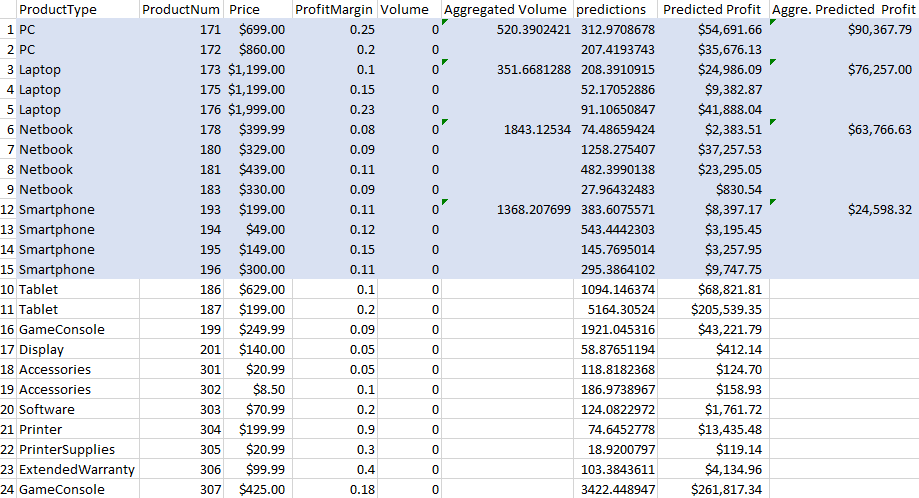


Random Forest Model newData5 (without 5star reviews)

### Your sales predictions for four target product types found in the new product attributes data set.

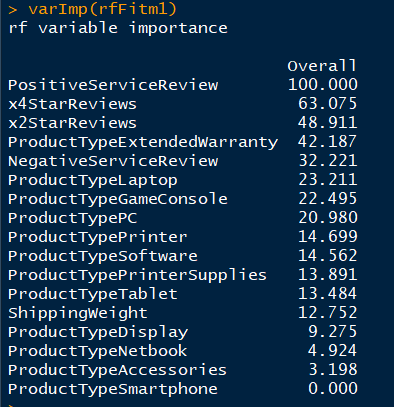
In order to output my data set and predictions from RStudio, I added my predictions to the new products data and then create a csv file. Below screenshot was written directly to my hard drive.





### A chart that displays the impact of customer and service reviews have on sales volume.

Here I ran the varImp () function to find out the importance for each attribute on the sales volume. The Positive Service Review is 100% importance on the sales volume, and the Negative Service Review is 32.221% importance on the sales volume. The importance for 4 Star Reviews and 2 Star Reviews are 63.075% and 48.911%, respectively.



### Did you learn anything of potential business value from this analysis?

I learnt how to perform deep analysis on the existing product dataset to forecast the sales volume on the new products in order to provide directions for the sales and marketing team.

Excel Analysis

There are so many products a company can carry, and it can be quite time-consuming to decide which product(s) or item(s) we need to invest more money or salesforce in. Traditionally, people use Excel to forecast the future marketing trend or customer purchase behaviors, however, due to the limitation of Excel, the forecast function is for sure not as powerful as the big data modelling analysis.

Big Data Analysis

Data science can be used to support decision making, optimize processes, and improve products and service across a wide range of business activities. We load the original dataset into RStudio, instead of manually input formulas, R can help us select and train the best winning model, we can then validate it against the testSet (i.e. new product) where real business value (i.e. forecast on volume) can be created.

### Was it straightforward to return your projections of sales volume using both models?

It literally took me lots of time to figure out the output dataset in the Global Environment panel, as there are many newly generated vectors, models, and processed datasets compared with task 1 and 2. It is easy once I found it, as I just clicking into the output dataset and the predicted volume was shown on the last column of the output1(for newData5 sub dataset) under the script panel in the RStudio.

In addition, it was also easy to locate the C2.T3output1.csv in my hard drive under newproductforecast folder. When I opened this .csv file, the predictions column was generated automatically.

### What are the main lessons you've learned from this experience?

**Warning Messages**

Don’t get scared by the warning messages. You can google them to find out the causes, but do not let the warning message prevent you from proceeding using the commands. Just keep tuning and testing the model until you get the results you like, and do not think too much at this moment.

The whole procedure is a good learning experience, try to find the answer yourself, and be positive when trying 😊

**Always try caret packages first**

At the beginning of the task, I forgot to use the caret package. Instead I tried writing the linear model and SVM linear model directly (I had to install the e1071 package). Later, I was told that e1071 is an older model before caret was invented. Although it is a good experiment to try and compare key testing results among various models, it is a little overwhelming to remember all the different syntaxes of various algorithms.

The beauty of caret is its simplicity as it provides an abstraction layer on top of different models to make it extremely easy to work with all the models. In many cases I can simply change the method name in the train() function to switch to a different model. The caret package contains 238 different models which covers all frequently used algorithms. Therefore, it’s always a good idea to try caret first when applying various machine learning models, as chances are, the algorithm is included in the caret package already.

**Model results could be very different**

It might also confuse a R beginner that different algorithms could lead to dramatically different results even if we chose the same attributes.

### What recommendations would you give to the sales department regarding your findings relating to the different types of reviews?

**Focus on negative and low star reviews**

I would suggest that the sales department should focus on addressing the negative service reviews as well as 1-3 stars customer reviews. They should track these reviews down individually whenever possible, resolve customer’s concern and hopefully lead to more positive service reviews and high star reviews.

The rationale behind this recommendation is the earlier finding where positive service reviews and high star customer reviews have very high importance on the sales volume. Thus, reduce negative reviews and increase positive reviews would likely translates to higher sales volume.

**Analyze positive and high star reviews**

In addition, marketing and sales team should analyze the positive customer review and 4-5-star reviews. It provides a better indication on which product# we need to invest more effort or more money on the advertisements, TV commercials, or promotions, as they are our cash cow. We must keep them.

**Other areas of further investigations**

We also need to ask ourselves some questions like:

1. Is there a correlation that low star review also leads to the low service review on a single product?
2. How can customers change their reviews? Do we have any historical data showing the improvement on product quality eventually leads to an increase the level of satisfaction?
3. Are there any substitutes on the market? What are their price strategies?