C3T3 Report

Please see the ***Final Charts*** worksheet in the *C3T3\_output.csv file*.

This document will outline how I went about predicting possible volume quantities for new products Blackwell Electronics is interested in selling. I will restate the scope of the problem and explain what techniques were used to generate my conclusions.

In its attempt to improve sales, Blackwell Electronics is trying to redesign its sales prediction analysis. For this, they want to include “product type” information in their sales prediction analysis and determine if it’s possible to predict the volumes of new products.

To this end, my analysis will involve two datasets. One dataset is historical sales data which contains information regarding the product type, customer review, price, physical properties (size and weight), and sales volume of various products. This dataset will be used to train models -- of which the best performing model will then be used to predict the possible sales volumes of new products using the second dataset.

The second dataset includes information about new products that Blackwell Electronics is interested in eventually selling. However, this second dataset lacks any information regarding sales volumes, as products have not yet been sold. This dataset will be used as the testing dataset in order to generate possible predictions regarding sales volumes of said products.

Ultimately, Blackwell Electronics is interested in predicting the possible sales volumes of the following product types: PCs, Laptops, Netbooks and Smartphones. In addition, the company is interested to see how might customer reviews relate to sales volume.

An important consideration to make, which will be reinforced later in this report, is that the historical dataset only has 80 observations. Why this matters is because, generally speaking, the quality of predictions are greatly affected by the size and scope of data. Small data may not be robust enough to yield reliable predictions, as such data may not contain enough information by which models may be capable of “discerning” consistent patterns or relationships.

For my analysis, I experimented using seven models. These models were Linear Regression, Support Vector Machine, Random Forest and Gradient Boosting Machine. I used additional versions of the LR, SVM and RF models using a modified training set from the historical data set. My analysis is informed by using the technique of Multiple Regression in order to generate predictions regarding sales volume given predictor variables from the historical dataset (price, customer review, etc.).

Having said that, I will now say more about each of the models used before giving the final performance results and then using the best performing model to predict possible sales volumes using the second dataset.

In total, I created seven models. The first set of models (LM, SVM, RF, GBM) used 25 predictors, which involved every variable of the historical dataset except the following: the product number (unique identifier), the best seller rank, and profit margin. This is because these variables did provide insightful information regarding the scope of the problem.

As for the second set of models (SVM, RF, GBM) used 24 predictors. The only difference between these models and the aforementioned models is that these secondary models were not trained using the “5 Star Reviews” variable. This variable had a perfect correlation with the dependent variable of volume. I decided to remove this variable from the training set of the secondary models because the inclusion of this variable heavily skewed the results of the primary models. Moreover, it is generally very unlikely to have a predictor variable that is 100% correlated with the dependent variable. Ultimately, I believe the secondary models yield more realistic and accurate results as compared to the primary models.

Given the images in the following page, I went ahead and decided to use the secondary Random Forest model in order to predict the possible volume of new products given the aforementioned second dataset. These images show the performance results of each model and their results predicting volume.

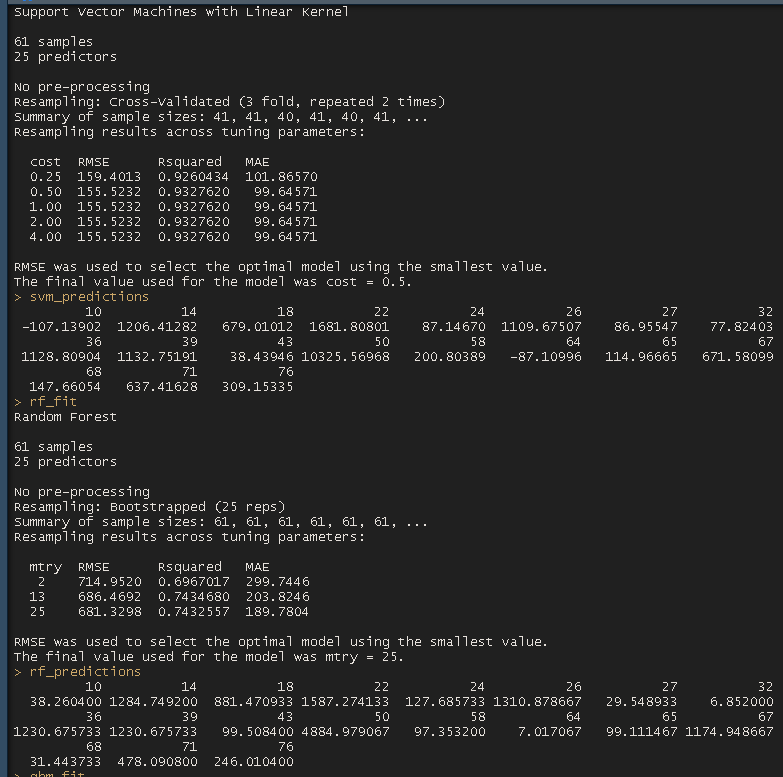
In evaluating the performance of the models, the two metrics of importance were the RMSE and R-Squared values.

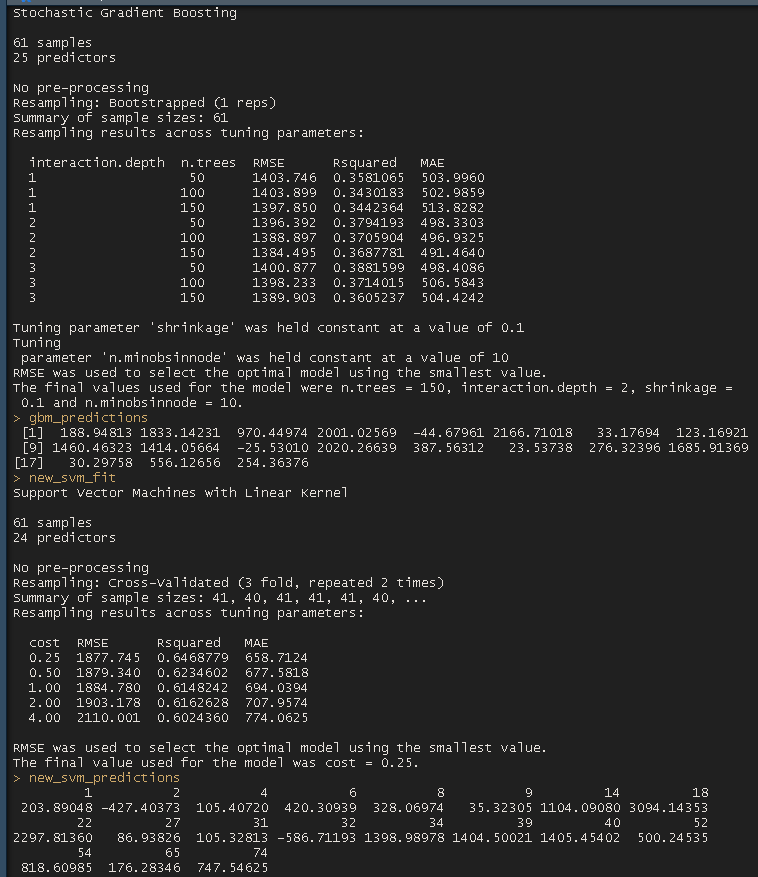
In order to avoid over-fitting, it was important to disregard models that had metric scores that were too consistent, which clearly corresponded to the first SVM model. This SVM model also had unreasonably high R-Squared scores, which suggests that the model over-performs given the dataset, which may hinder its ability to generalize when interacting with other data. Finally, this model had some negative prediction values, which shouldn’t possible. Simply put, it is impossible to sell a negative amount of a product. Every prediction value should at least start at an amount greater than 0.

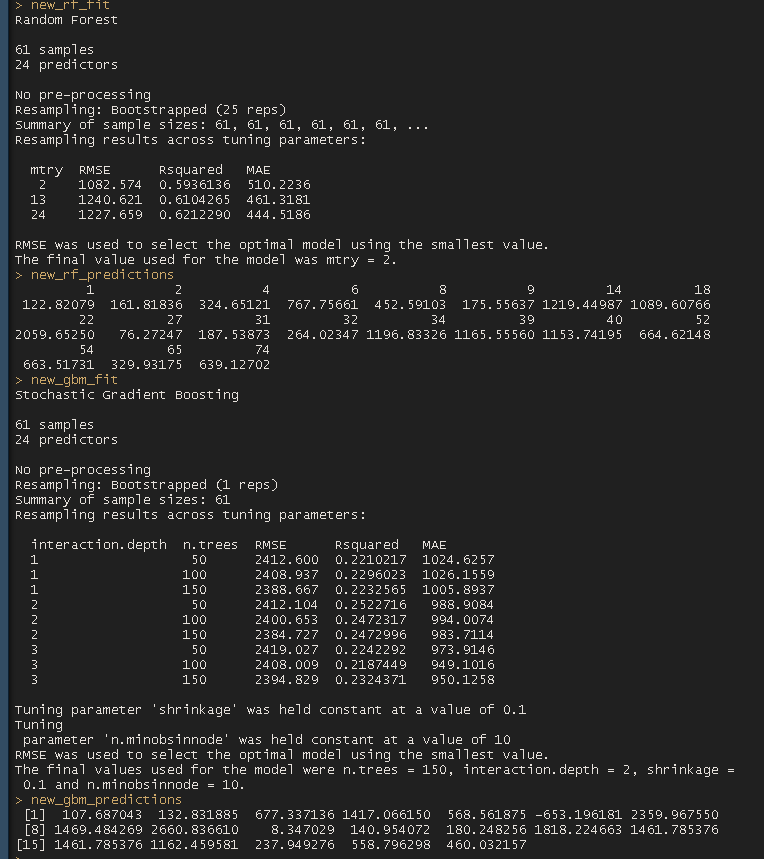
As mentioned earlier, I ended up choosing the second Random Forest model. This is because its results were not as skewed as the primary models. If you compare the second RF model with the first RF model, it is clear that the second RF model yielded more realistic prediction values than the first RF model. That is, in the first RF model, some of the prediction scores were between 6 and 38, which seem like very small amounts in the context of a retail electronics store.

**Chart of Predicted Sales Volumes**

|  |  |  |
| --- | --- | --- |
| Product Type | Volume | Percentage |
|  |  |  |
| Tablet | 5309 | 31% |
| GameConsole | 4532 | 27% |
| Smartphone | 2263 | 13% |
| Netbook | 2060 | 12% |
| PC | 805 | 5% |
| Laptop | 555 | 3% |
| Accessories | 538 | 3% |
| Software | 283 | 2% |
| ExtendedWarranty | 261 | 1% |
| PrinterSupplies | 197 | 1% |
| Printer | 136 | 1% |
| Display | 124 | 1% |
| TOTAL | 17603 | 100% |







**Conclusion (recommendations, what I learned, etc.)**

In order for more reliable and realistic predictions to be generated, it is necessary for the historical dataset to be much larger. A major issue that this kind of analysis could encounter is the problem of over-fitting, which is in large part due to the small size of the historical dataset. So, my major recommendation would be to encourage Blackwell Electronics to gather more sales data.

An important issue throughout this analysis is how models perform given the composition of the data. In the earlier paragraph, I mention the issue of small data. However, another issue involves feature selection. In particular, when selecting features, it is important that such features are not excessively correlated with the dependent (response) variable. In the case of this project, the 5 Star Review variable was entirely correlated with the Volume variable. In fact, the relationship between these variables were 4:1. That is, for every 5-star reviews a product had, its volume would increase by 1. This sort of situation is extremely unlikely in real-world situation, especially as it relates to the context of retail stores. The dynamics of supply and demand can’t be so easily and neatly expressed in clean, even ratios such as what this relationship conveys. As such, I made the decision to choose a model whose predictive capacity would not be influenced by 5-star review variable. Ultimately, this project helped me to better understand why and how feature selection affects the performance of models.