horizontal line

Predicting Customer Preferences: Multiple Regression in R

# OVERVIEW

Blackwell Electronics has asked us to analyze historical sales data and make sales volume predictions on a new product dataset. The overall goal is to give the sales team an idea of the sales performance of different product types. The products that they are most interested in are PCs, Laptops, Netbooks, and Smartphones--although there are other product types that we will be analyzing as well. The company is also interested in understanding the impact of customer reviews and service reviews on sale volume of different product types.

# MODELS

Three different models were evaluated:

1. Support Vector Machine (SVM)
2. Random Forest (RF)
3. Gradient Boosting (GBM)

A basic linear regression model was evaluated initially, and it resulted in a perfect R-squared value of 1, indicating that this is not a reliable algorithm to use in this case, and that the data is non-parametric. Therefore the linear regression model was scrapped and models that are used for non-parametric data (SVM, RF, and GBM) were used for testing.

# TESTING RESULTS

**SVM Model**

Results:

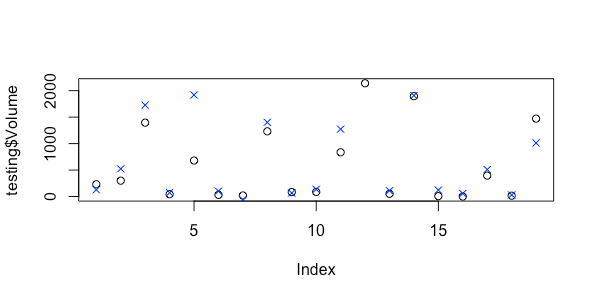
The first SVM model was built using all of the features of the dataset (except for BestSellersRank, which was removed due to having several NA values). The results of the training were:

|  |  |  |  |
| --- | --- | --- | --- |
| C | RMSE | Rsqared | MAE |
| 0.1 | 177.7844 | 0.9546032 | 124.9061 |

Although the Rsquared score is good, the RMSE is quite high, and the model appears to be overfitting. When the model was used to make predictions on the held-out test set, and Post-Resampling was used to evaluate those predictions, scores were not as good:

|  |  |  |
| --- | --- | --- |
| RMSE | Rsqared | MAE |
| 547.1429196 | 0.8060408 | 286.8606898 |

A graph illustrating the difference between the predicted values (blue x) and actual values (black circle) in the test set is shown below:



Next, variable importance for the model was evaluated. The five values with the highest importance are illustrated below:

|  |  |
| --- | --- |
|  | Importance |
| fivestar | 100 |
| fourstar | 93.5551 |
| PosReview | 82.7288 |
| threestar | 69.6304 |
| twostar | 67.6755 |

This indicates that customer and service reviews were very highly influential in the impact of sales volume, with Five Star Reviews being the most important.

When analyzing the correlation between the features, it was noted that the reviews (Five Star, Four Star, etc) were highly correlated with one another, and with Positive and Negative Reviews. Another SVM model was built eliminating some of those features, and the best results were found when Four Star Reviews were deleted. The results are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| C | RMSE | Rsqared | MAE |
| 0.25 | 143.2764 | 0.9618955 | 119.4248 |

Although results appear to be slightly better, the model is still overfit. The post resample results were similar to the previous model.

**Random Forest Model**

Next, random forest was trained using the same test set. The best results for random forest are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| mtry | RMSE | Rsqared | MAE |
| 26 | 703.2548 | 0.9600554 | 305.6153 |

Random Forest had a good Rsquared value, but a very high RMSE value. Because of this, random forest was not used as the final model.

**Gradient Boosting Model**

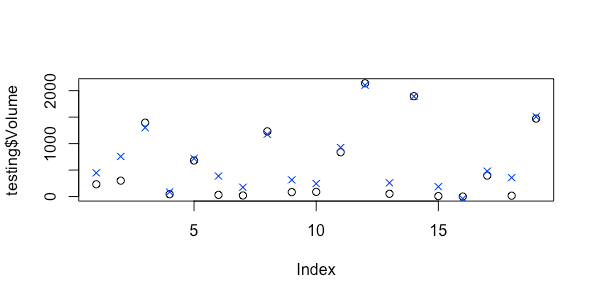
Gradient boosting was then trained using the same test set. The best results are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| mstop | RMSE | Rsqared | MAE |
| 150 | 173.1271 | 0.9434552 | 133.1898 |

Post resampling results are as follows:

|  |  |  |
| --- | --- | --- |
| RMSE | Rsquared | MAE |
| 193.0971028 | 0.9634827 | 148.3221924 |

A graph illustrating the difference between the predicted values (blue x) and actual values (black circle) in the test set is shown below:



Although all of the models appear to have overfit, the gradient boosting model had the best scores (low RMSE, high Rsquared) and also had better post resampling values. Additionally, the predictions that were made were all positive values, unlike the other models. Therefore, the gradient boosting model was used to make predictions on the new products dataset.

# PREDICTION RESULTS

The prediction results were added to the new products dataframe and have been attached as a .csv file. The filename is ‘newproductspreds.csv’

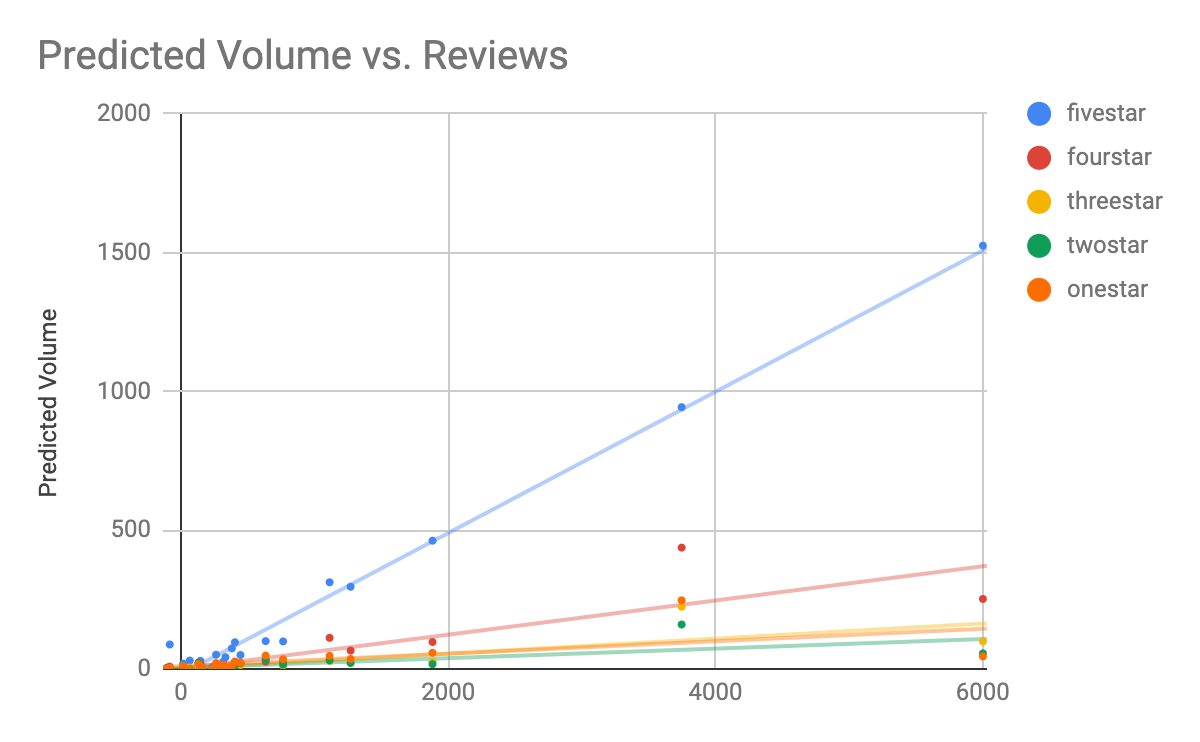
The prediction results for the product types of interest are as follows:

|  |  |  |
| --- | --- | --- |
| Product Type | ProductNum | Predicted Volume |
| PC | 171 | 404.0477486 |
| PC | 172 | 262.9845627 |
| Laptop | 173 | 381.1831329 |
| Laptop | 175 | 237.2632863 |
| Laptop | 176 | 365.0115922 |
| Netbook | 178 | 16.56294937 |
| Netbook | 180 | 1112.681558 |
| Netbook | 181 | 132.3494048 |
| Netbook | 183 | -104.3840862 |
| SmartPhone | 193 | 764.7149722 |
| SmartPhone | 194 | 634.0648932 |
| SmartPhone | 195 | 333.3919871 |
| SmartPhone | 196 | 445.5164639 |

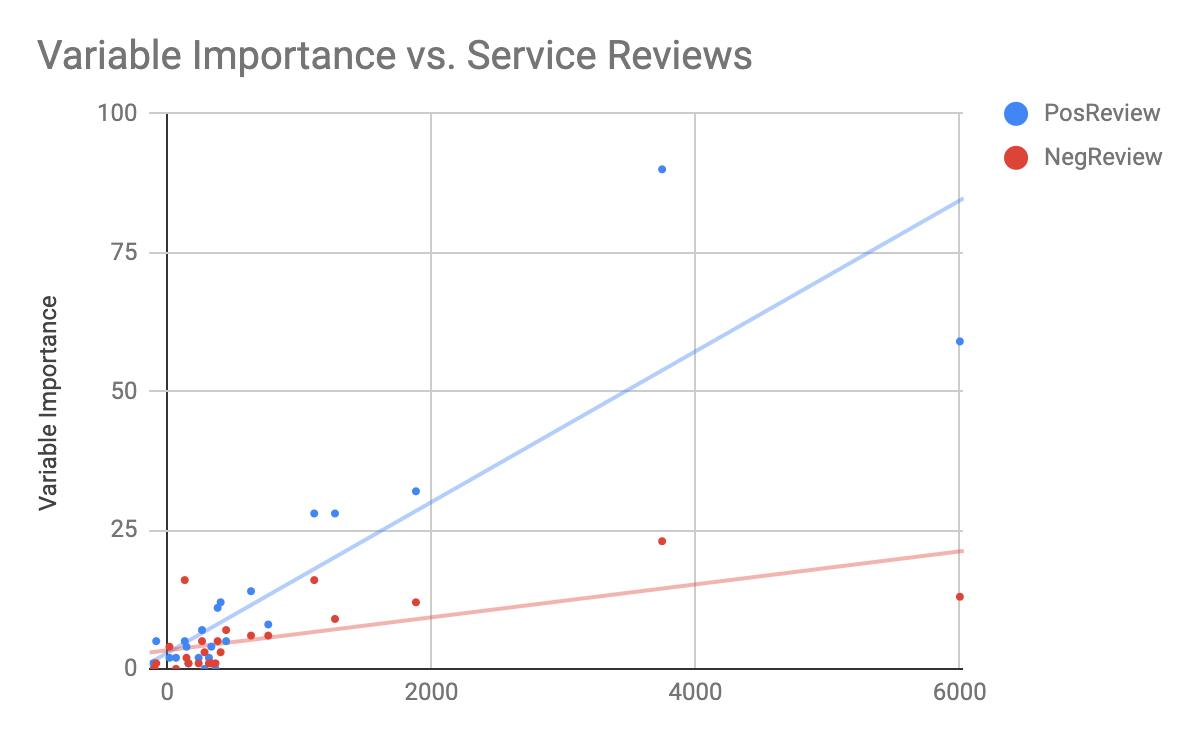
Additionally, when looking at variable importance for our model, we can see that customer and service reviews were the most important features for predicting sales volume, as illustrated by the chart below:

|  |  |
| --- | --- |
|  | Variable Importance |
| fivestar | 100 |
| fourstar | 93.5551 |
| PosReview | 82.7288 |
| threestar | 69.6304 |
| twostar | 67.6755 |
| NegReview | 34.2345 |
| onestar | 25.1087 |

When graphing customer reviews vs predicted volume, we can see that five star reviews are most influential when it comes to increasing sale volume, as is also indicated by the variable importance. Sale volume tends to be less with reviews with fewer stars.



We can also see below that service reviews are impactful in determining sales volume. Positive reviews tend to increase volume, while negative reviews don’t increase volume as much.



# RECOMMENDATIONS

While the results do seem promising, it is difficult to put a lot of faith in the predictions, as all of the models overfit the data, even the final model that was chosen. The only way to combat this would be to have additional and stronger data. As it is now, there are very few data points, so when there are outliers in the data, it impacts the models performance much more significantly. Therefore, we recommend collecting more data for different products and product types.

However, the trends of increase in sales volume with five star customer reviews and positive service reviews are promising. I would recommend that the sales department encourage satisfied customers to write positive reviews about products that they enjoyed, in order to increase sales volume.