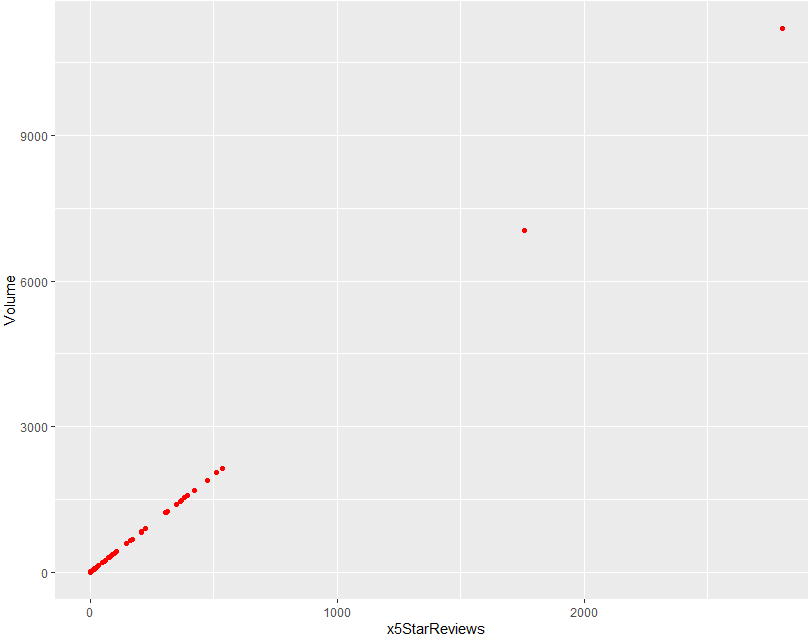
Danielle,

Hi again. I can’t believe we are still dealing with this Covid19 virus. What are your kids doing for school this fall? So far, in Leander, they are giving parents the choice of 100% online or 100% in person. It is not feasible for the district to offer a 50/50 option. Personally, as a teacher, I think all of this is going to be incredibly difficult no matter what the district decides. We just have to be flexible each and every day.

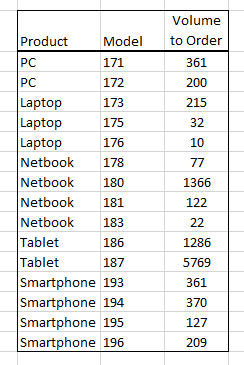
However, let’s get down to business. You asked me to create a model that would predict the sales of 4 target products: PCs, Laptops, Netbooks, and Smartphones. I was given one dataset with 80 rows of historical data that included features such star reviews, customer reviews, product dimensions and the volume of product ordered/sold. The other dataset was significantly smaller with only 24 rows of data. This dataset did not have the volume of units that were going to be sold. For this specific task, you asked me to build a model that would predict the number or units sold and asked if star ratings and customer reviews have any bearing on sales.

I selected to use the Caret package available in R. The algorithms I chose were linear regression, support vector machines with radial basis function kernel, extreme gradient boosting tree and extreme gradient boosting linear.

During the exploration of the data, one thing I noticed right away was the strong correlation between the number of positive ratings and the volume ordered. This makes sense, if there are many positive star reviews then there was a lot of the product ordered. For example, you could not possibly receive 100 5-star reviews if there were only 20 products sold. You would have to at least sell 100 products and have every customer do the review.



This chart shows the very linear correlation between the star reviews and the volume of the product ordered. I combined the 5-star, 4-star, and 3 -star ratings together.

Here are the predicted number of units I believe you should order.

Below you’ll see the all of the numbers are really good based on my testing set.

RMSE Rsquared MAE

37.132 0.997 20.717

During my explorations, I tried several of the models mentioned above. In the end, I chose Random Forest for the model. I wanted to chose Linear Regression but, the model gave me some negative predictions, so I went with Random Forest.

In the end, I really considered the correlation between the 3,4, and 5-star reviews and I used that feature along with Random Forest to generate these predictions. I made predictions with several of my models and the results were all very close. But, I believe these are the most accurate.

Please let me know if you have any questions.

Thank you.

Sherri Koski