Sales Prediction Report

Blackwell Electronics

By Samantha Goodman

Background

For this task I was assigned the job of predicting sales volume for several different types of products, given historical sales data and information about new products. I was also asked to analyze the impact that various reviews have on product sales. To accomplish this I tested three different algorithms, Support Vector Machine, Random Forest, and Gradient Boosting. Once I had a model built that I was happy with, I used it to predict sales volume on a series of new products. I also found which factors were most important for the model, and found that the number of 4 and 5 star reviews had the biggest impact on sales volume.

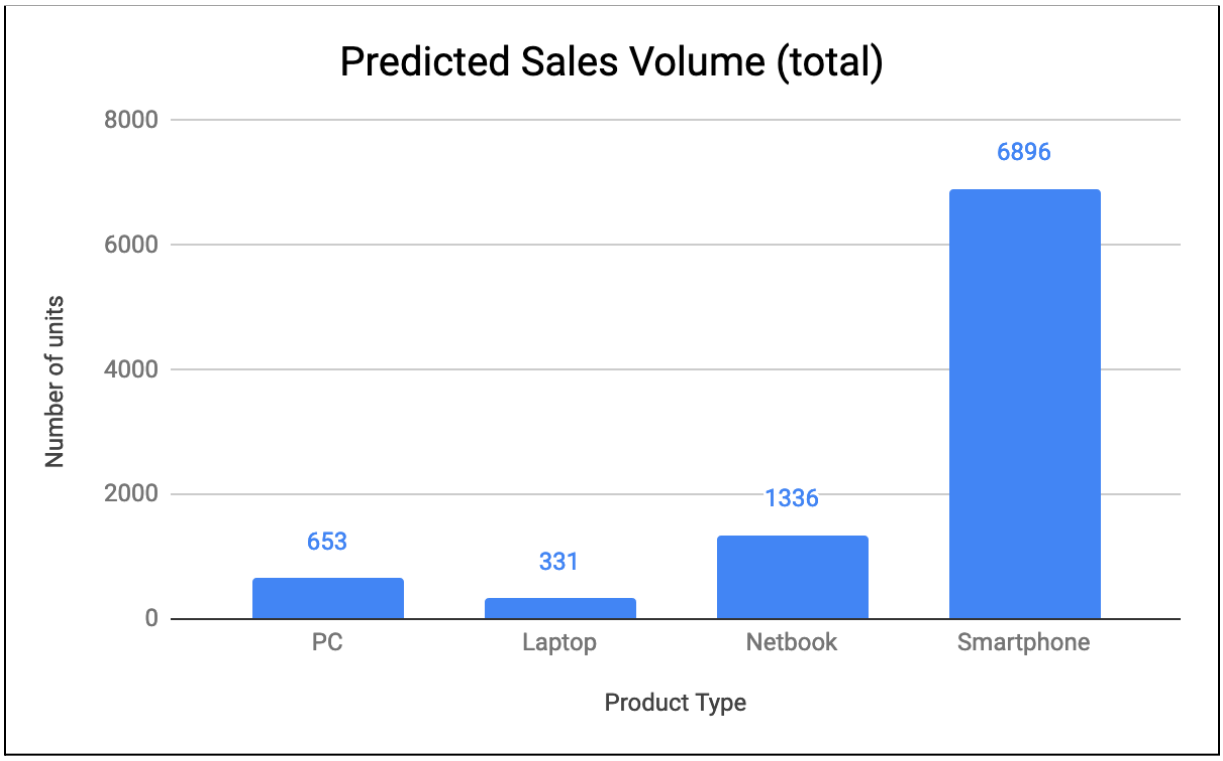
Algorithm Results

I tested three different algorithms to create a model. The results are in the table below. Included are the time it took to build the model, it's estimated accuracy and notes on the model.

Algorithm Name	Time (sec)	Estimated R^2	Notes
SVM	1.4	0.96	Predicted negative sales volume, therefore unusable
Random Forest	6.9	0.97	Model selected
Gradient Boosting	2.5	0.84	Predicted negative sales volume, therefore unusable

I selected the Random Forest model to use for this assignment. It was the only model I tested that didn't predict negative sales volume (this is impossible because a negative number of units cannot be sold). The predicted R^2 was 0.97 but when tested against known sales volume figures, the R^2 was only 0.75. Perhaps with more time, more models could be tested and a better model could be found with a higher actual R^2. Or perhaps a larger dataset of historical sales could be used and that would lead to a more accurate Random Forest model.

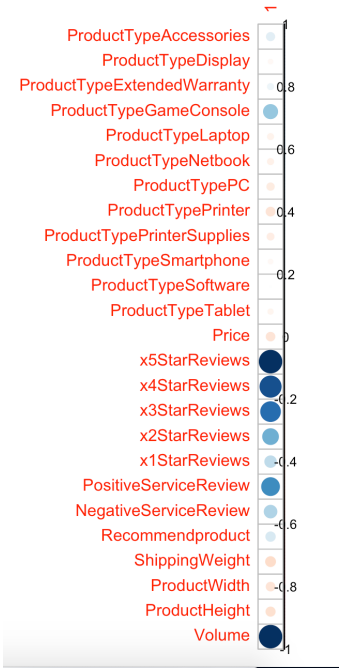
Sales Predictions



This chart shows the total predicted sales volume for the new products, specifically for PCs, laptops, netbooks and smartphones. Smartphones were predicted to have the highest volume of sales, and laptops the lowest. This is all based on the product information that we have, most importantly on 5 and 4 star customer reviews.

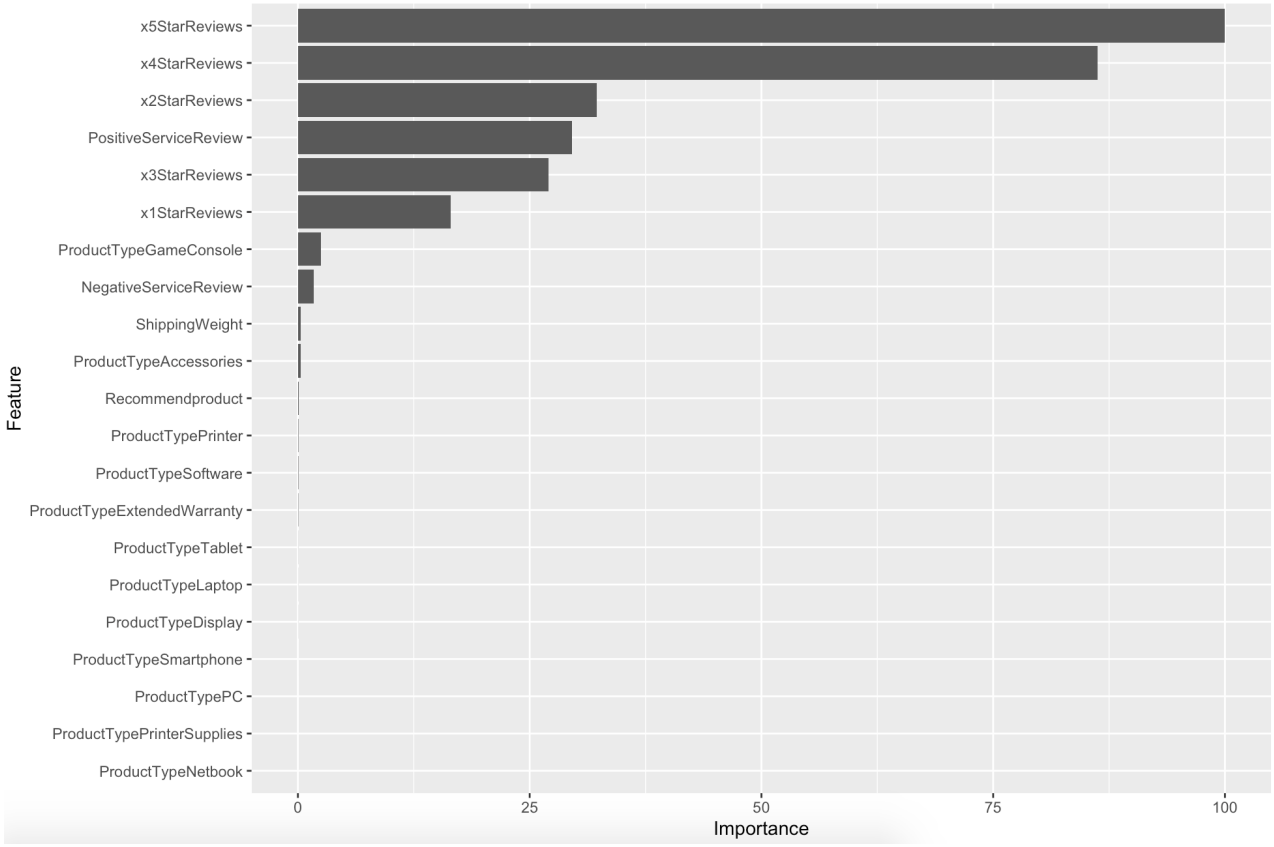
Important Variables

Using the variable importance function and by finding the correlation, I was able to find the most important variables that affect sales volume.



This chart shows the correlation between all of the variables given and the sales volume. A blue circle indicates positive correlation, and a red circle indicates a negative correlation. The larger and darker the circle is, the higher the correlation. From this chart we can tell that the number of 5 star customer reviews (x5StarReviews) and the number of 4 star reviews (x4StarReviews) are very influential.

In the next chart, the most important variables that were used in the model are shown, ranked on a scale of 0 to 100.



As this chart shows, the most important features in the model were the number of 5 star reviews, the number of 4 star reviews, the number of 2 star reviews and the number of positive service reviews. The remaining review types as well as product attributes mattered less.

Conclusion

The volume of sales for new products was predicted, as was requested. Smartphones and netbooks have the highest predicted volumes. These predictions were made using the Random Forest algorithm, as the other types were unusable.

Key takeaways for the sales team is that the number of customer reviews (specifically 5 and 4 star reviews) has a big impact on the number of units sold. Potential customers really value seeing many positive reviews of the product they are considering buying.

An interesting point is that both positive (4 and 5 stars) and negative (1 and 2 stars) reviews both had a positive correlation with sales volume. This indicates that the overall

number of reviews (both good or bad) can help a product sell and thus encouraging customer and service reviews can improve sales. Further analysis of different campaigns to accomplish this is suggested.