Course 3, Task 3: Report to Blackwell

Sales Prediction Report

Anders Dowd - Spring 2022 Data Analytics Certification Program

Overview

Blackwell is expanding its product catalog and needs sales projections on its new items; the potential sales of PCs, laptops, netbooks, and smartphones are of particular interest. Using sales data, customer reviews, and specifications from items currently offered, we can inform a model that will give reasonable predictions of sales figures for the new products, thus allowing Blackwell to make efficient purchasing decisions and maintain a tight inventory. I tested seven different algorithms and chose the one with the best balance of RMSE (closest to 0 being ideal) and R^2 (closest to 1 being ideal) scores to create the most reliable model.

Methods

I initiated my process with feature selection. By examining feature correlation to the target variable (in this case, sales volume), I eliminated unnecessary features from the model. Any variable that had less than a 25% correlation to volume was thrown out, leaving us with eight predictive features. The most highly correlated variables were user reviews; quite simply, the better reviews a product received, the more that product sold.

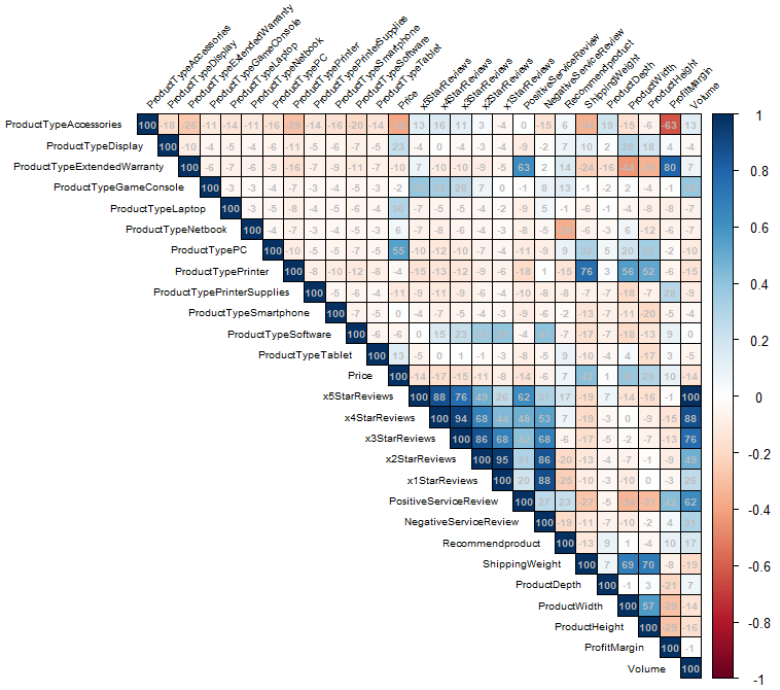

Figure 1: Correlation Heat Map of Features

I tested seven algorithms to find the most appropriate one for the data: two gradient boosting, two random forest, and three support vector machines.

I began by first running the models out-of-the-box with no parameter tuning to get a baseline performance measure, then narrowed down the number of algorithms and tuned parameters to get the best fit.

Gradient Boosting

I used a Stochastic Gradient Boosting and eXtreme Gradient Boosting (XGB) models. Out of the box, the eXtreme Gradient Boosting model performed the better of the two with an RMSE of 935.3323 and an R^2 value of 0.8409913. After tuning the XGB model, I was able to achieve RMSE=766.7632 and R^2=.9190734

on the training set. It performed even better on the test set, returning RMSE=150.4327999 and R^2=0.9727781.

Random Forest

I also ran a basic Random Forest as well as Parallel Random Forest algorithms. These performed very similarly with no tuning, but were not as effective as the eXtreme Gradient Boosting model. After tuning I was able to achieve RMSE=1106.668 and R^2=0.7637524 with Random Forest. Like XGB, it performed better on the test set with RMSE=566.4709958 and R^2=0.8278531.

Support Vector Machines

There is a plethora of SVM algorithms to choose from, so I experimented with three for this task. Of the three, SVM with Polynomial Kernel performed the best on the training set (RMSE=232.7992 and R^2=0.9584967), but showed signs of overfit when employed on the test set. Not only did it return poorer scores (RMSE=394.077703 and R^2=.0.888723), but it also predicted negative values.

Making Predictions

The eXtreme Gradient Boosting model was the most effective and reliable of all models tested. There are 13 items on the new product list that fall into the categories of interest. When employed on the new product data, the tuned XGB model predicts the following sales forecast for these items.

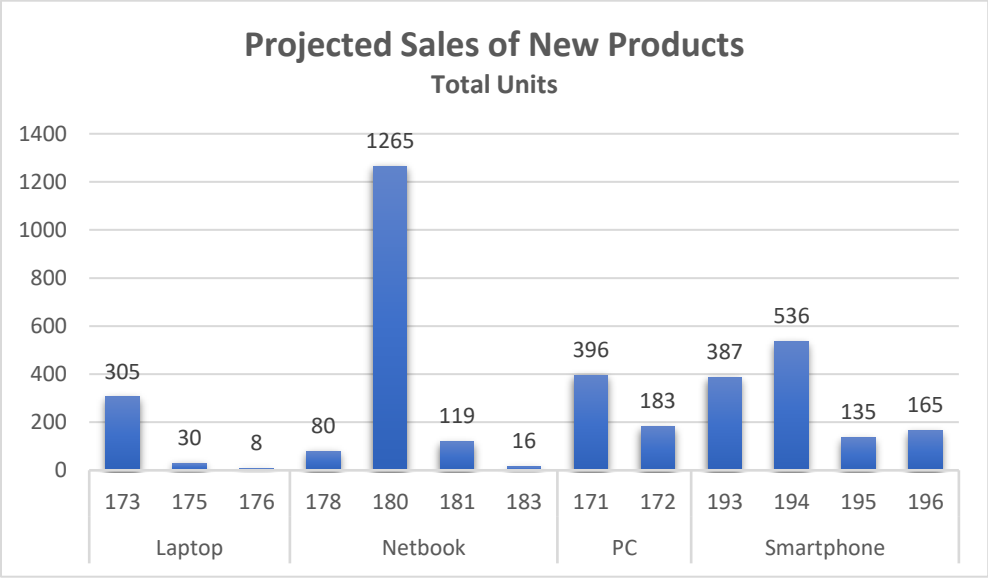

Figure 2: Sales Projections

Final Thoughts

There is an exceptionally strong correlation between customer reviews and sales volumes. If Blackwell wants to drive sales, it is imperative that customer feedback is sought on every sale made. Furthermore, 5-star reviews have the highest impact on driving sales, but as is indicated in Figure 3, even poor reviews show a positive correlation toward sales figures, indicating that feedback of any kind is likely better than no feedback at all.

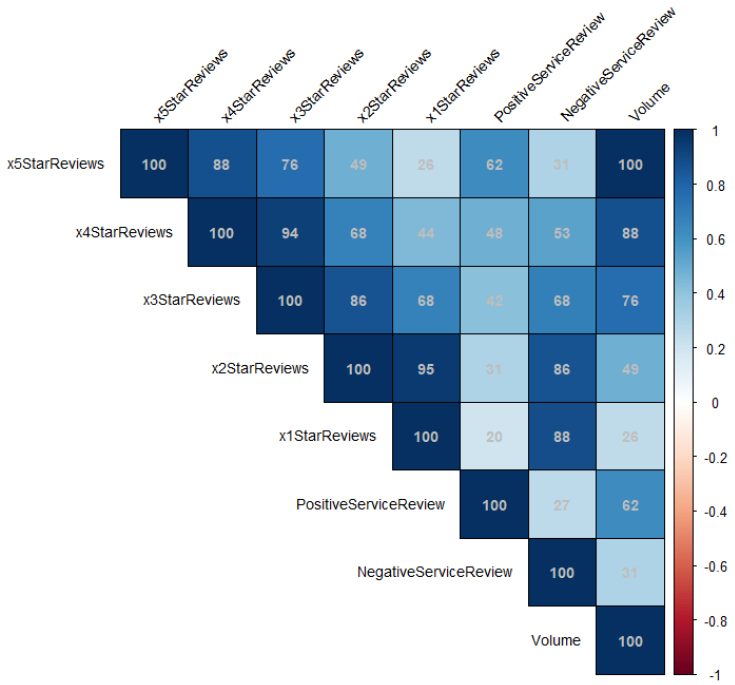

Figure 3: Correlation Heat Map of Customer Reviews to Sales Volume