Sales Predictions

From: Usman Iftikhar

The task requires to predict sales of four specific product types (PC, Laptops, Netbooks and Smartphones), and the impact service & customer reviews have on sales of different product types. Two sets of data are provided: 1) data for the existing products, and 2) data for the new products to make predictions on. Both sets of data contain identical attributes. Any preprocessing done to the existing products dataset is also done to the new products dataset.

To begin the analysis, obvious features, like Product Number, are removed because these do not add any value to the analysis. In addition, attributes with missing values are also omitted from the analysis. To be able to use regression analysis, all the remaining attributes are converted from text to numerical. A dataset is formed from this to generate regression models for prediction. Another dataset is generated with additional attributes removed based on feature selection using the Filtering method. Independent variables that were highly correlated (> 0.8) were removed to not have collinearity. Both datasets, original and the one with certain attributes removed based on feature selection, are used in the regression analysis, and the dataset with better performance metrics is chosen as the dataset to predict. The better performing dataset in this case was the original dataset, that was preprocessed to remove the product number and the best seller rank attribute due to missing values.

The preprocessed data was split into a training set with 75% of the data from the existing products and the remaining 25% of the data were used to test the model’s output to gauge performance metrics. The training set data was trained on different algorithms, shown in Table 1. The top-ranking algorithm, eXtreme Gradient Boosting highlighted in Table 1, was then used in postResample to predict the volume from the new products dataset with the following performance metrics: R2 = 0.631 and RMSE = 2388.



Table 1: Algorithms tried on the training set data with the selected algorithm highlighted.

The predicted volume of the new product’s dataset was then used to calculate profitability. Among the four products of interest, Figure 1 below shows the PC product to have the highest profitability, followed by Laptops, Netbooks, and finally Smartphones. Figure 2 shows the impact of customer and sales reviews on the total sales volume. The data shows a positive correlation between the total number of reviews and total sales volume.

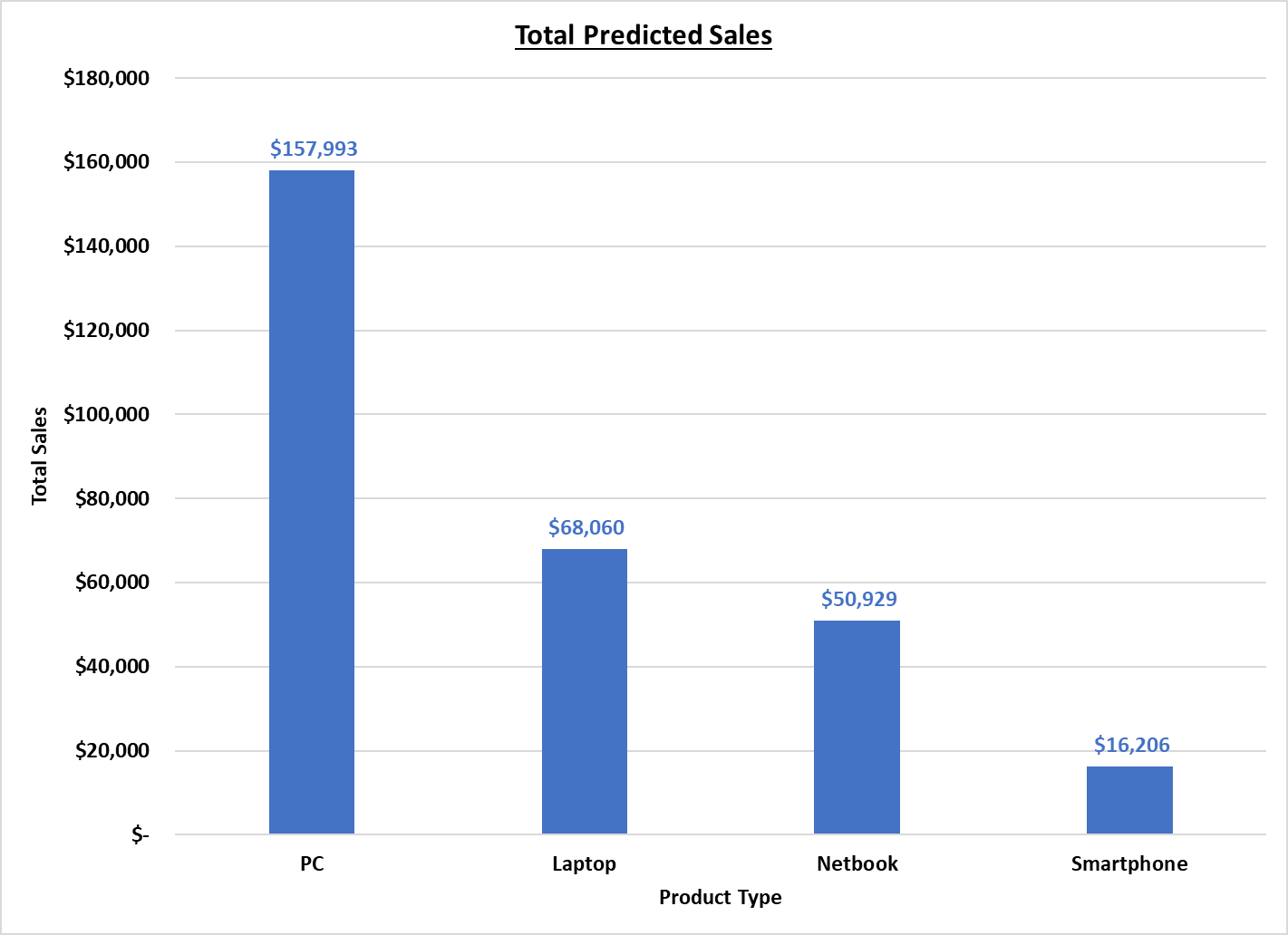


Figure 1: Predicted sales of the new products.

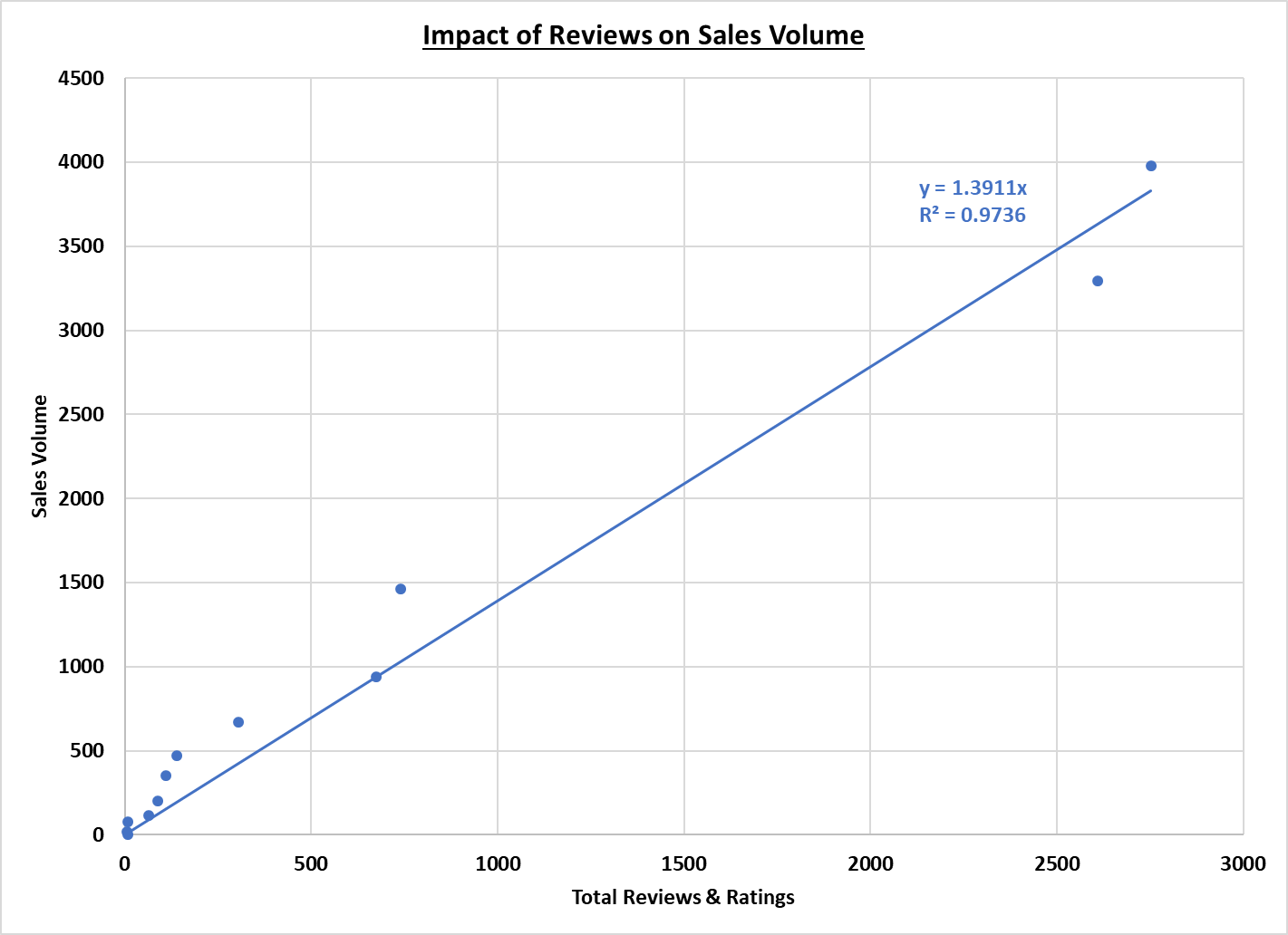


Figure 2: Impact of customer & services reviews on sales volume.