## SALES PREDICTION REPORT

#### BLACKWELL ELECTRONICS

Nastaran Mahmoodzadeh Fall 2020 Historical Data Analysis

### FREQUENCY OF PRODUCT TYPES

As we see, <u>Accessories</u> have the highest frequency (32.5%) amongst existing product types.

The four new products that we want the sales prediction for are:

- PC (5%)
- Laptops (3.75%)
- Smartphone (5%)
- Netbook (2.5%)

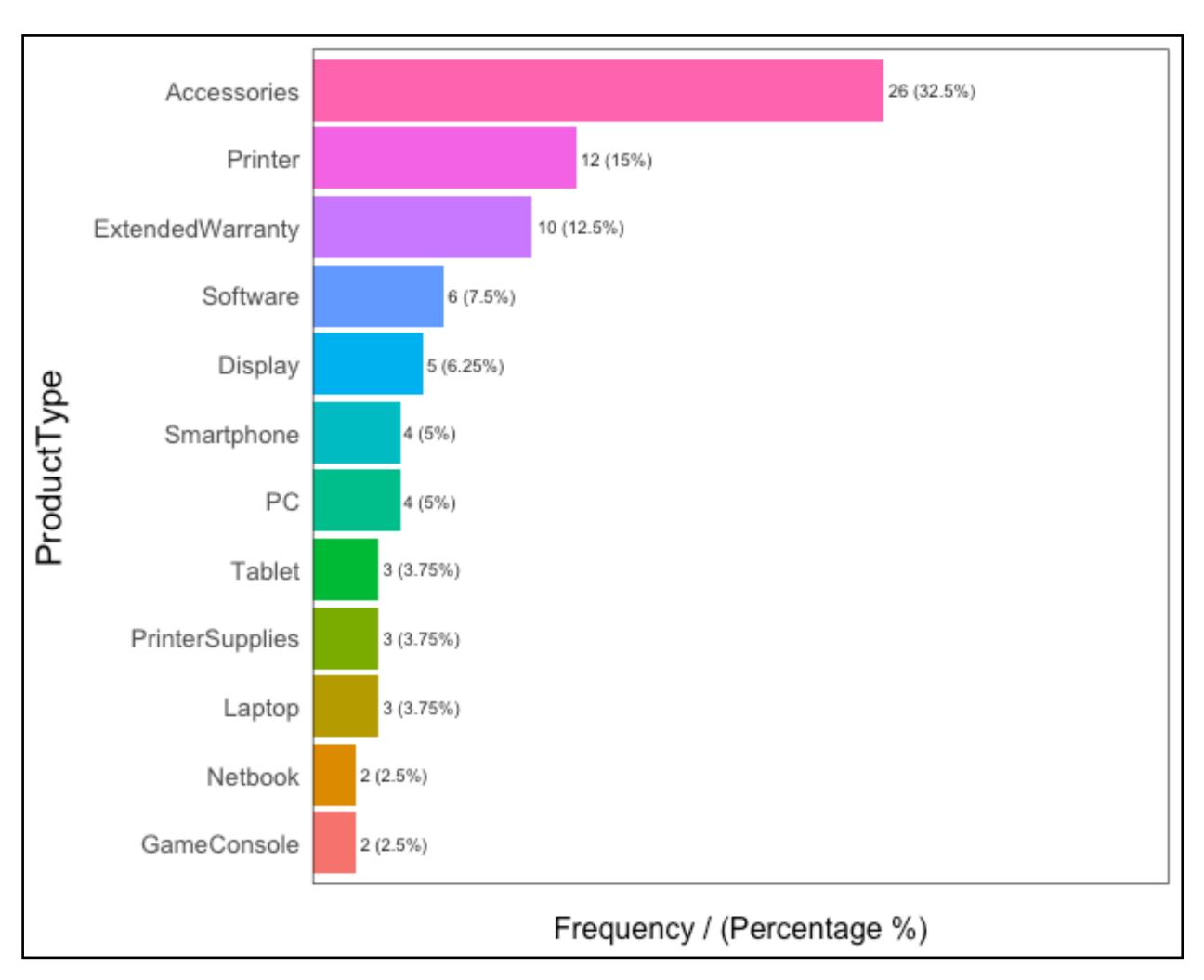


Figure 1

#### INDIVIDUAL VALUE PLOTS

We see the value distribution of each feature, separately.

- Most of Prices are under \$1000

- Left Skewed distribution in RecommendedProduct

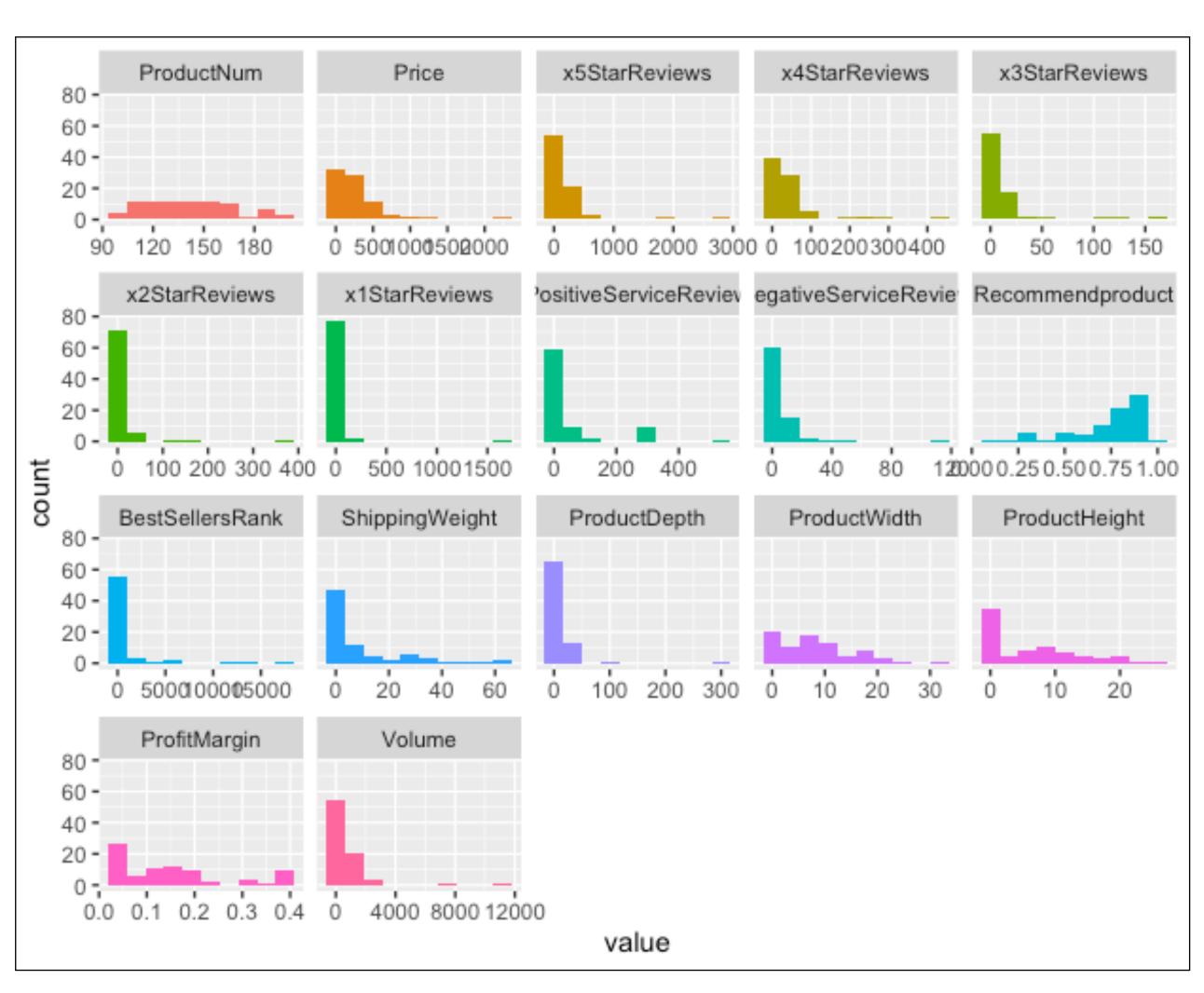


Figure 2

As we see in Figure 3, most of the prices we have are under \$1000. The highest price average belongs to Laptop and PC. Also, the least price distribution is for Netbook. Finally, PC has the most outliers in price distribution in compare to other products.

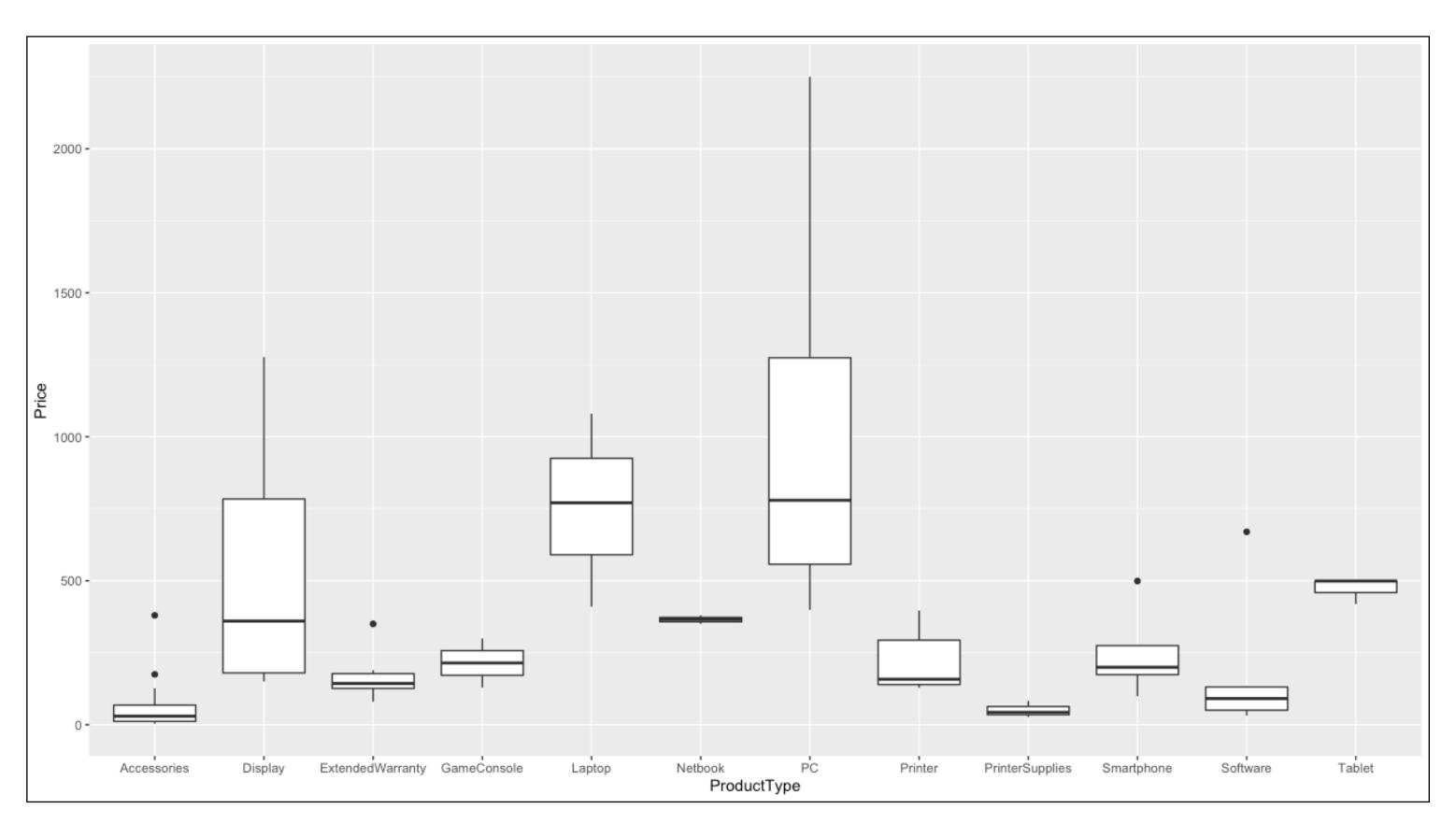


Figure 3

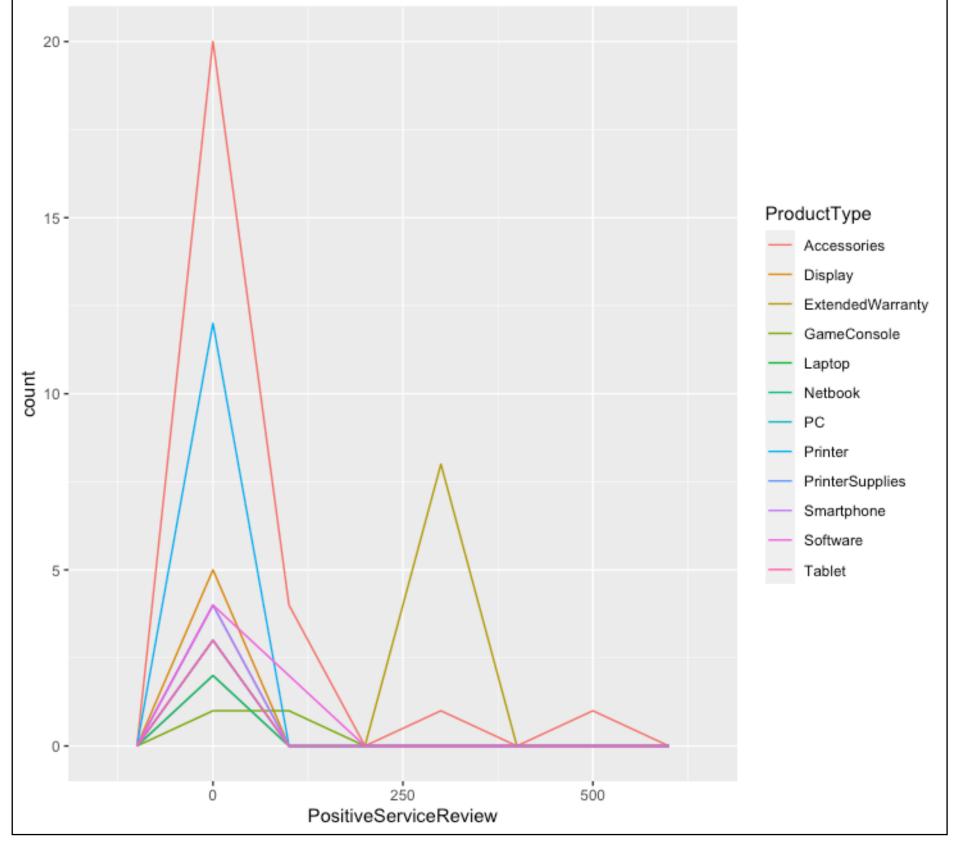


Figure 4

-100 200 NegativeServiceReview Figure 5 Looking at Figure 4, Accessory, Extended Warranty, Tablet, and Software have gotten the highest positive reviews from customers. However, in Figure 5, we see that they mostly got under 100 negative reviews.

20 -ProductType — Accessories ExtendedWarranty GameConsole count 10 -PrinterSupplies Smartphone Software Tablet 0 -

In the correlation matrix, we can see the relationship between predictors (features) themselves, and the relationship they have with the ground truth which is Volume.

Multicollinearity: <u>5StarReviews</u> has a really high correlation with Volume which is concerning. We tried *Combolnfo* function, and it too warned us about 5StarReviews variable. Hence, we decided to omit it from the analysis process.

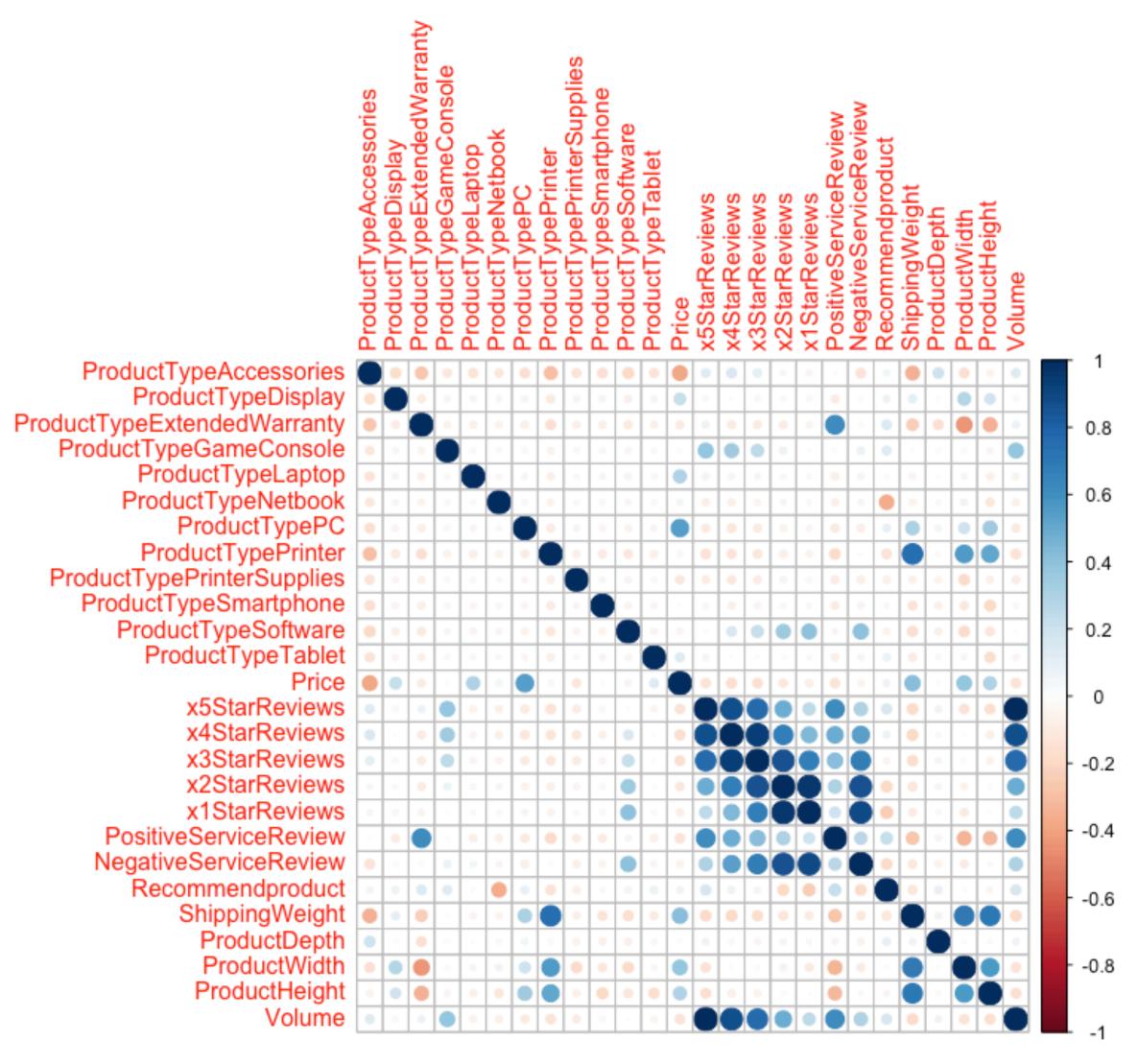


Figure 6. Correlation Heatmap

## Methods

#### LINEAR REGRESSION MODELS

- Linear Model (Version 1)
   We used the whole dataset as predictors for the Volume, except for ProductNum, ProfitMargin, and BestSellerRank.
  - We ran the model and got I for both R-squared and Adjusted R-squared which means the model is overfitted, hence not trustworthy to use.
  - Linear Model (Version 2)
    Since there are multicollinearity between volume and 5XReview, we omitted the 5XReview from the dataset and ran the linear model. This time we got a much better model with 0.93 for R-squared and 0.89 for Adjusted R-squared.

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 545.8 on 41 degrees of freedom

Multiple R-squared: 0.93, Adjusted R-squared: 0.8908

F-statistic: 23.7 on 23 and 41 DF, p-value: < 2.2e-16
```

### SUPPORT VECTOR MACHINE (SVM)

After running the best tune function, we found out that cost 0.25 with RMSE of 1650 and R-Squared of 0.75 is the best one.

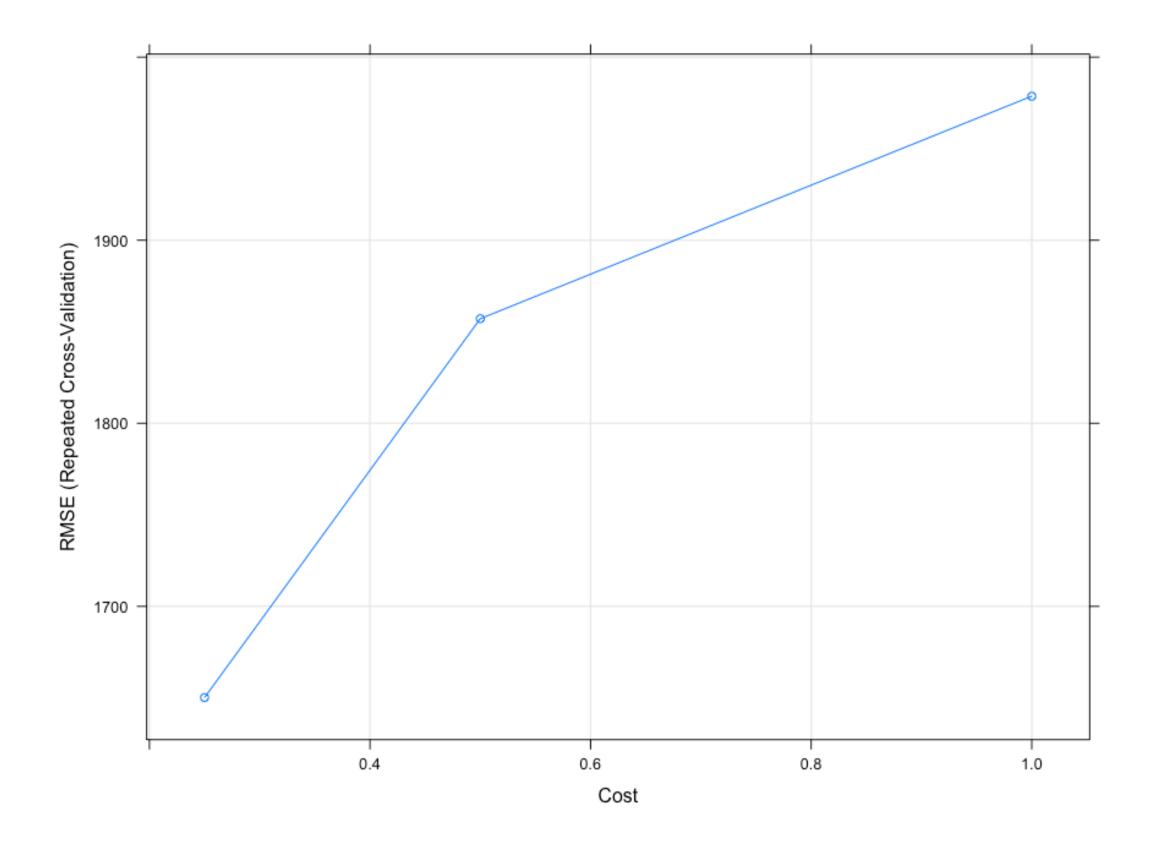
```
Support Vector Machines with Linear Kernel

65 samples
24 predictors

Pre-processing: centered (24), scaled (24)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 58, 58, 58, 59, 59, 59, ...
Resampling results across tuning parameters:

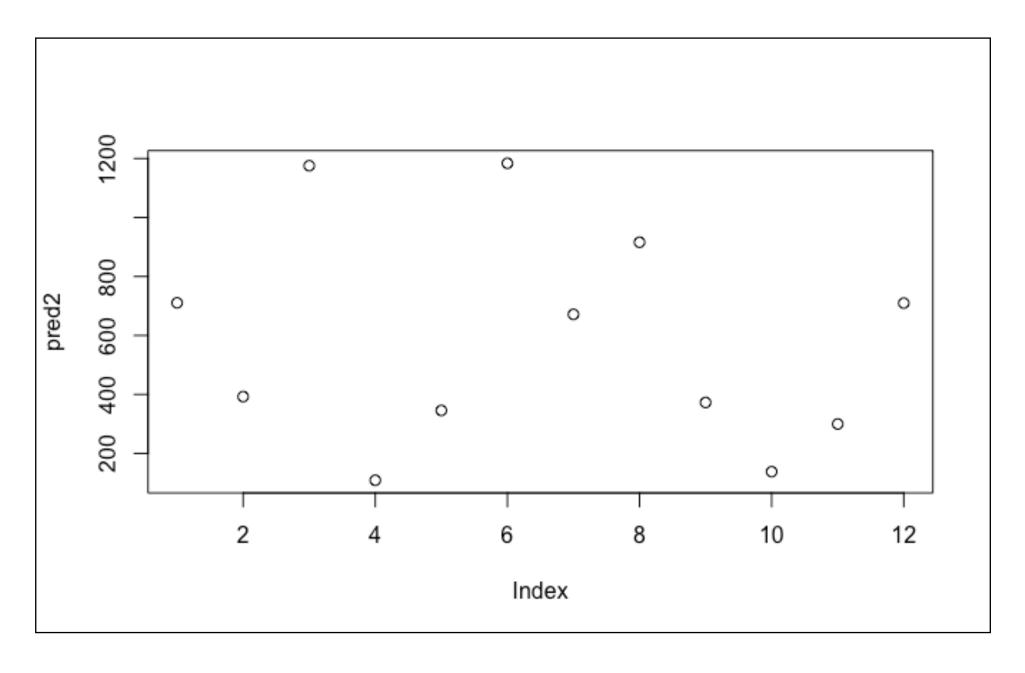
cost RMSE Rsquared MAE
0.25 1650.186 0.7504885 853.0078
0.50 1857.138 0.7319222 938.5436
1.00 1978.664 0.7267168 995.7443

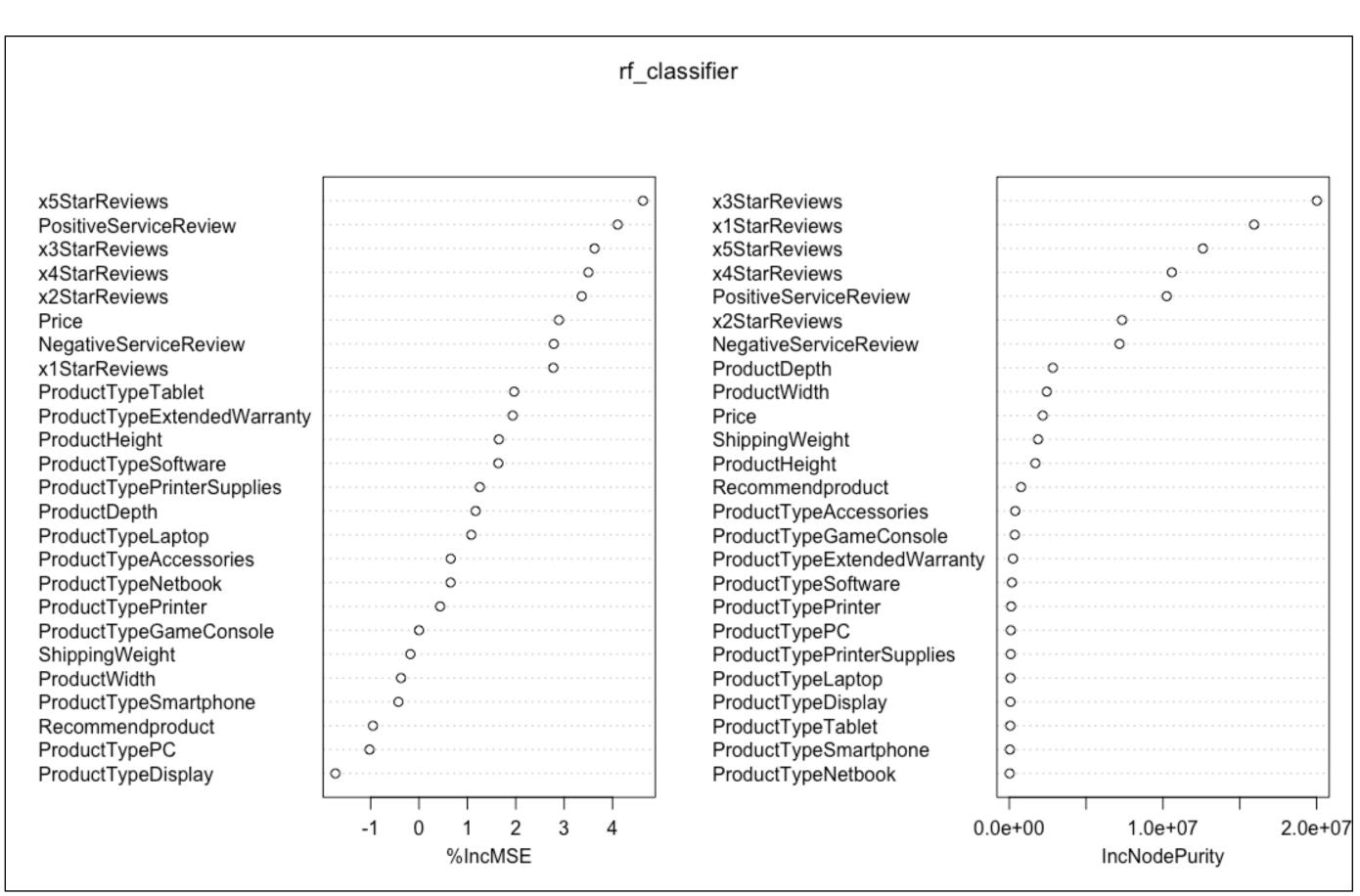
RMSE was used to select the optimal model using the smallest value.
The final value used for the model was cost = 0.25.
```



#### RANDOM FOREST CLASSIFIER

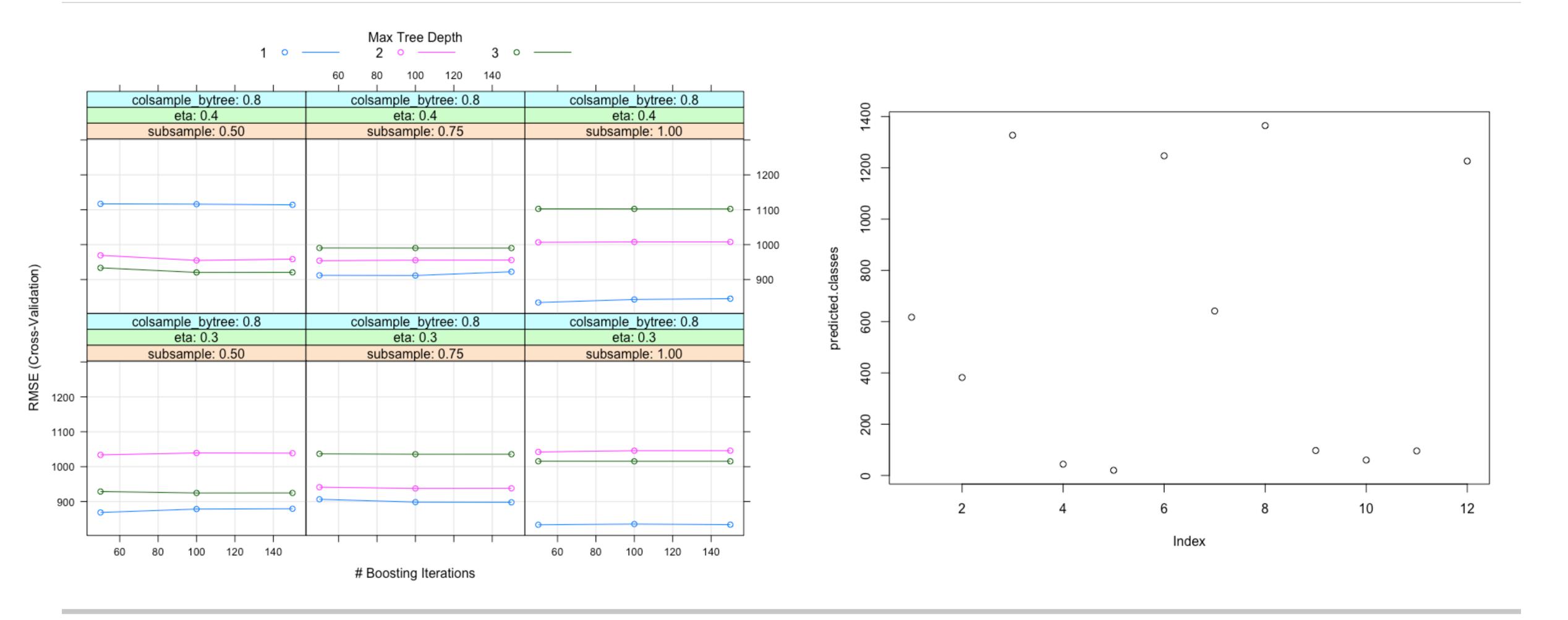
By running postResample on the test set, we got 297 for the RMSE and 0.76 for the R-Squared for this model.





#### GRADIENT BOOSTING

By running postResample on the test set, we got 278 for the RMSE and 0.80 for the R-Squared for this model.



# Results

Comparing the R-Squared value, the second linear model has the highest accuracy; hence, we used that model in the predict() function in R to predict Volume in the New Products dataset.

**Did you learn anything of potential business value from this analysis?** I learned mostly about the importance of big data in final prediction. This dataset is small which make analysis process quick, however affects the overall accuracy of the analysis results.

Was it straightforward to rerun your projections of sales volume using both models? I struggles in this very step! I know I definably need more practice here.

What are the main lessons you've learned from this experience? To make sure the new data does not over right the old one. And, to search R errors in a smart way.

What recommendations would you give to the sales department regarding your findings relating to the different types of reviews? They should prepare the operation department and their inventory plantings based on the volume predictions so they will not face any trouble in the future.