**Predicting Sales Report**

**Executive summary**

This report predicts the sales of four product types:

* PCs
* Laptops
* Netbooks
* Smartphones

It also looks at the effect that service and customer reviews have on sales.

The process for modelling the data is in *Appendix 1: The Process*.

**Recommendation**

***Service and customer reviews are key to sales volumes***

Positive service reviews, four-star reviews and product recommendations attribute to sales volume. Increases in the frequency of these, increases the number of products sold.

***Issues with the predictions***

The dataset used to make the predictions has only 80 observations. This is too small a sample to make reliable predictions. However, the model does work well when the predicted volumes are small. When volumes are predicted to be above 500, the error rates suggest the model is not working well. It is recommended that this exercise is repeated with a larger dataset.

**The predictions**

Based on the supplied data set, the predicted sales volumes for the four product types are shown below.

|  |  |
| --- | --- |
| Product type | Sales volume |
| PC | 438 |
| Laptop | 207 |
| Netbook | 1257 |
| Smartphone | 888 |

**Appendix: The process**

The stages in the analysis were:

1. Data preparation and pre-processing
2. Set up base models and assess them with different variables
3. Assessing the models
4. The selected model
5. Investigation of the model’s errors
6. Make the predictions
7. **Data preparation and pre-processing**

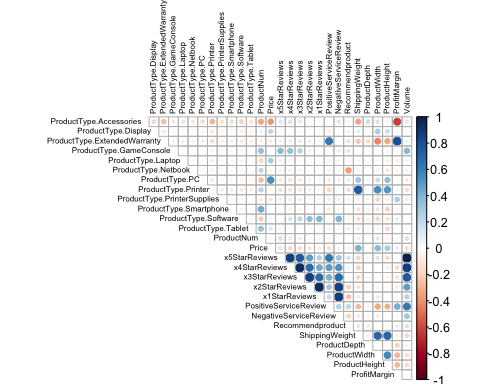
***First steps***

* Removed the Best Sellers Rank from the dataset because it has many missing values.
* Removed Product Number because it is just an identifier.
* Dummified the variables that are categorical so that they are recognized as factors.
* Plotted the independent variables with the dependent variable, volume.

***Correlation between variables***

A heatmap of the correlation gives an overview of the relationships between the variables and a function was written to find the pairs of variables with a correlation of over 0.9. This established the highly correlated pairs and which of the pair needs to be removed before modelling.

Figure 1: Correlation heatmap



***Removal of variables and observations***

In summary, the following were removed:

* 1-star reviews
* 3-star reviews
* 5-star reviews
* Best-selling rank
* Product number
* All product types
* Observations with volumes of above 3,000

Plots reveal the variables with outliers that may compromise the models. Figures 2 and 3 for volume suggest volume above 3,000 should be removed from the dataset. While Figures 4-8 also show outliers, these observations haven’t been removed. The dataset is very small with only 80 observations so have only removed the most extreme so the dataset isn’t too depleted.

Figure 2: histogram of volume

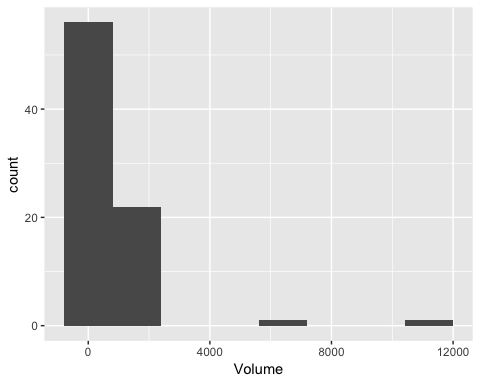


Figure 3: boxplot of volume

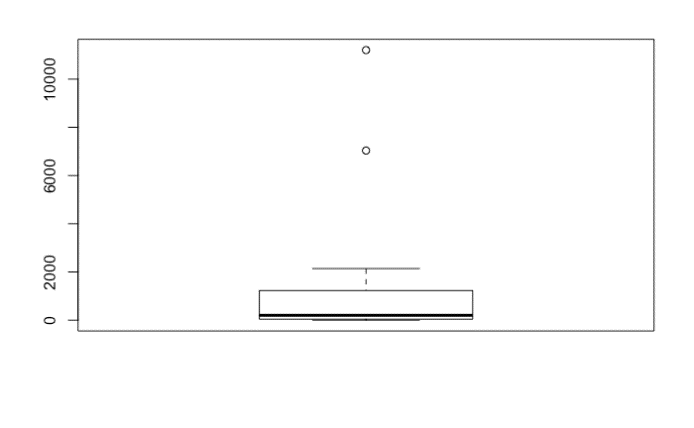


Figure 4: histogram of positive service reviews

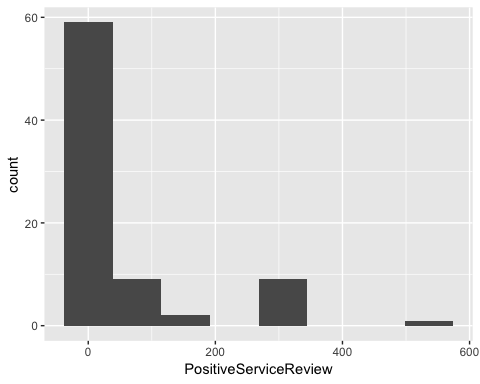


Figure 5: boxplot of positive service reviews

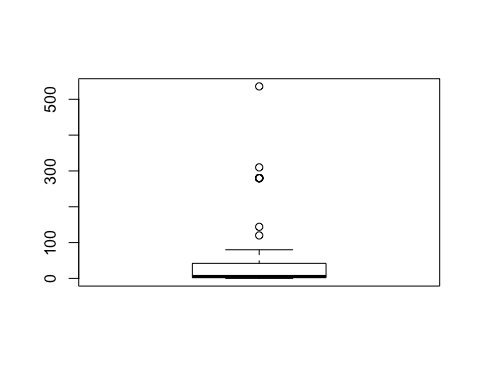


Figure 6: histogram of negative service reviews

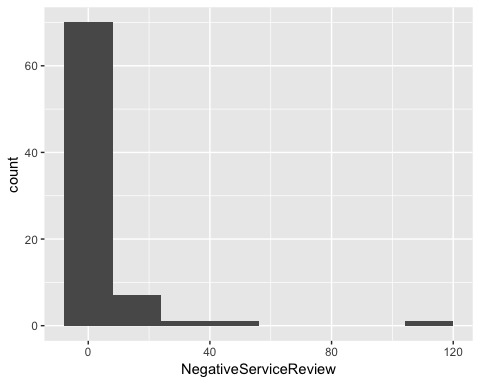


Figure 7: histogram of 2-star reviews

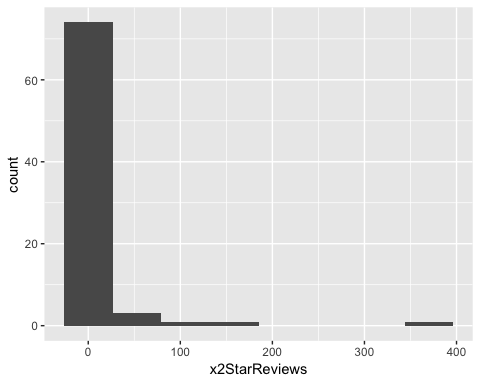
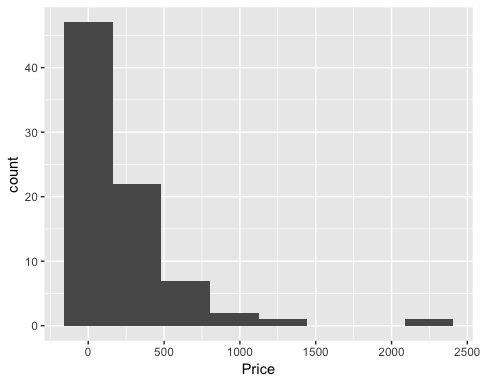


Figure 8: histogram of price



***Relationships between variables***

It would appear that there is not a strong relationship between volume and price, see Figure 9. Figures 10 and 11 show a clear relationship between volume and positive service reviews, and volume and 4-star reviews. These figures zoom in on the relationship and the outliers have not been plotted.

Figure 9: scatterplot of price and volume

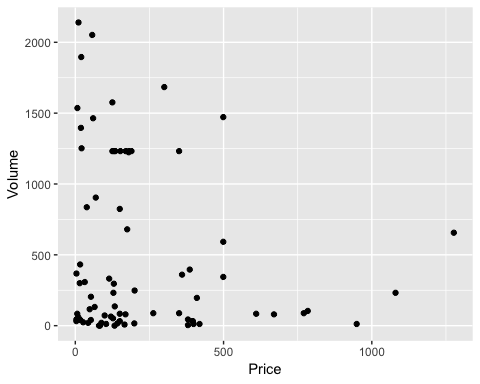


Figure 10: scatterplot of positive service reviews and volume

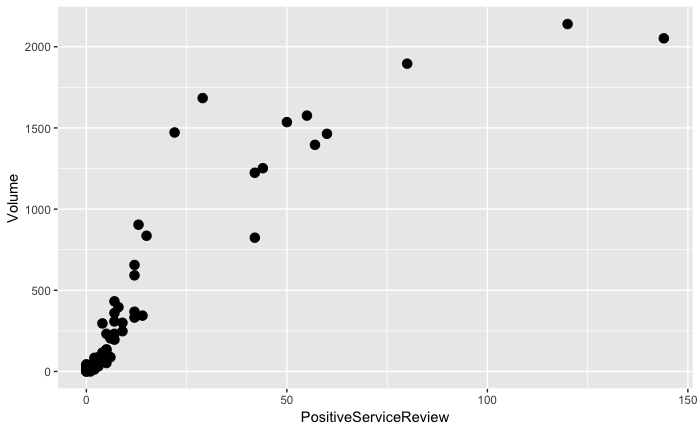
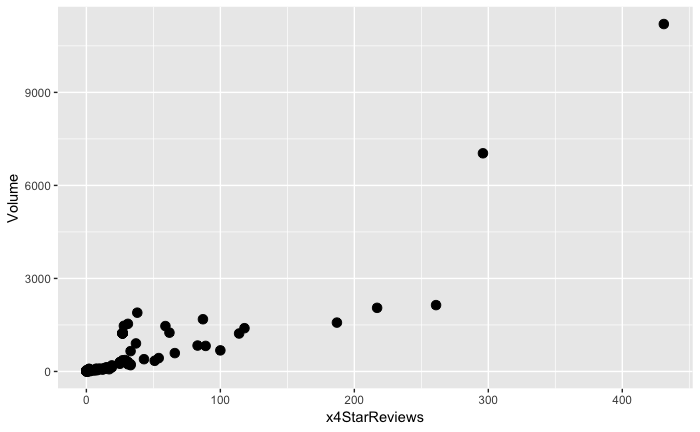


Figure 11: scatterplot of 4-star service reviews and volume



1. **Set up base models and assess them with different variables**

***A linear model***A linear model using all the variables appears to perform well with an R-squared value of 0.922. However, the RMSE is very high at 169.7.

A plot of the residuals, Figure 12, shows a non-linear relationship and Figure 13 shows the residuals are not equally and randomly spread. The conclusion is that the relationship between Volume and the independent variables is not linear.

Figure 12: plot of the residuals v fitted values

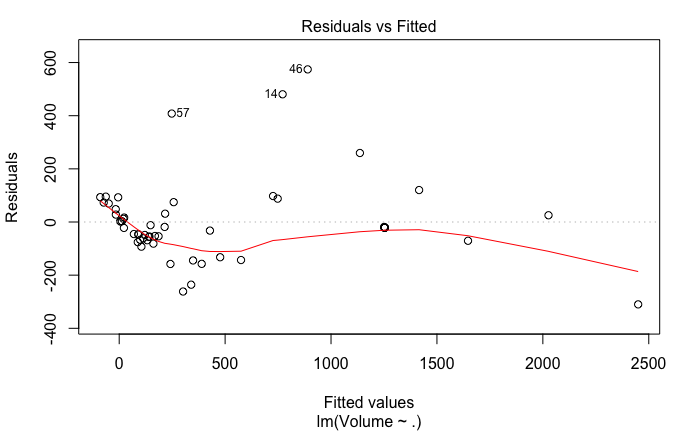
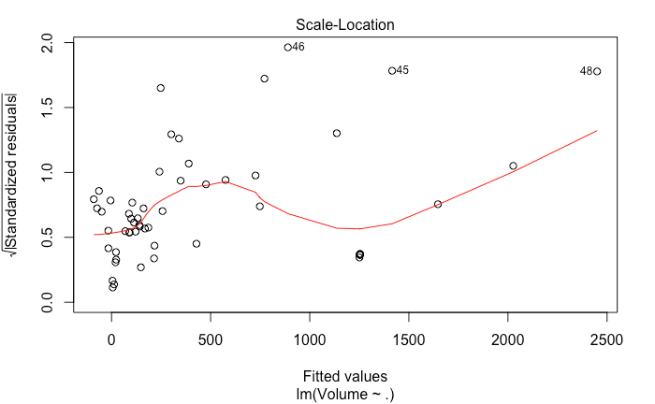


Figure 13: scale-location



***Random forest, SVM and KNN models***Created a training set with 70% of the existing product data and a testing set with the remaining 30%. The seed was set at 998.

Three models were used:

* Random forest with 5-fold cross validation and manually tune 5 different mtry values.
* A linear SVM model with the variables normalized.
* KNN with the variables normalized.

1. **Assessing the models**

Each model was trained using all independent variables and then the accuracy of making predictions was measured by making the predictions for the known volumes in the test set. The models were then assessed with different variable combinations based on those variables identified as the most important.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Random Forest** | **Training** |  |  | **Testing** |  |  |
| Variables | RMSE | R² | MAE | RMSE | R² | MAE |
| PositiveServiceReview | 173.713800 | 0.897191 | 99.651450 | 230.632607 | 0.908185 | 137.180672 |
| PositiveServiceReview + x4StartReviews | 171.736600 | 0.905076 | 87.175710 | 301.696059 | 0.858554 | 172.929052 |
| PositiveServiceReview + x4StartReviews + x2StartReviews | 176.938000 | 0.899743 | 89.580380 | 316.491682 | 0.850896 | 178.505302 |
| PositiveServiceReview + x4StartReviews + x2StartReviews + Recommendproduct | 169.331100 | 0.915394 | 87.142440 | 332.682706 | 0.852218 | 184.249057 |
| PositiveServiceReview + x4StartReviews + Recommendproduct | 173.978300 | 0.909678 | 89.882890 | 306.555665 | 0.873487 | 172.589475 |
| **K-NN** | **Training** |  |  | **Testing** |  |  |
| Variables | RMSE | R² | MAE | RMSE | R² | MAE |
| PositiveServiceReview | 174.653000 | 0.888912 | 94.351760 | 296.553735 | 0.848350 | 165.395982 |
| PositiveServiceReview + x4StartReviews | 250.228500 | 0.806859 | 122.668400 | 409.123107 | 0.739978 | 222.181818 |
| PositiveServiceReview + x4StartReviews + x2StartReviews | 273.495900 | 0.754649 | 137.933900 | 440.422048 | 0.700921 | 234.176623 |
| PositiveServiceReview + x4StartReviews + x2StartReviews + NegativeServiceReview | 291.658800 | 0.727015 | 152.252600 | 438.635589 | 0.692375 | 235.903030 |
| **SVM** | **Training** |  |  | **Testing** |  |  |
| Variables | RMSE | R² | MAE | RMSE | R² | MAE |
| PositiveServiceReview | 226.054300 | 0.872190 | 137.202300 | 376.232917 | 0.812409 | 208.988798 |
| PositiveServiceReview + x4StartReviews | 160.903800 | 0.919553 | 91.192380 | 479.956757 | 0.549818 | 276.275239 |
| PositiveServiceReview + x4StartReviews + x2StartReviews | 227.177500 | 0.861743 | 141.502300 | 474.557441 | 0.708684 | 258.899298 |
| PositiveServiceReview + x4StartReviews + x2StartReviews + NegativeServiceReview | 278.283300 | 0.764347 | 179.867200 | 530.232607 | 0.481103 | 319.896878 |

1. **The selected model**

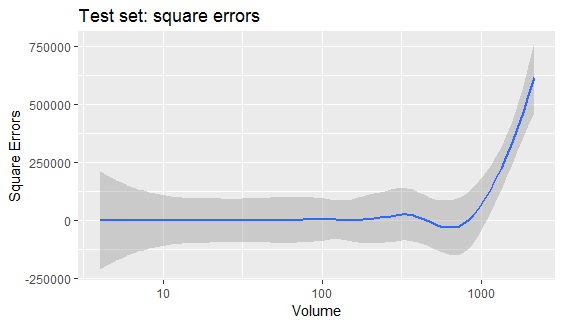
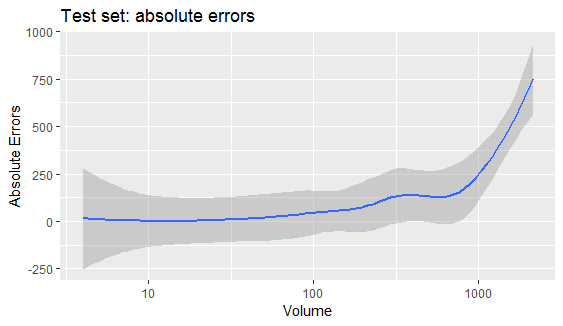
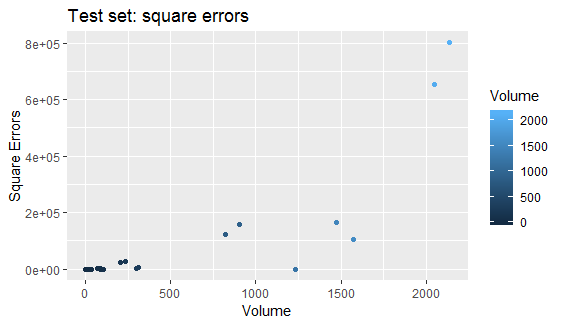
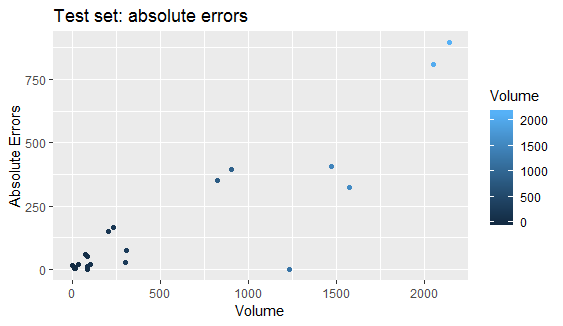
As can be seen in the highlighted row in the table in Section 3, the random forest model with the following independent variables had the best performance results:

* Positive Service Review
* 4-Star Reviews
* Recommend products

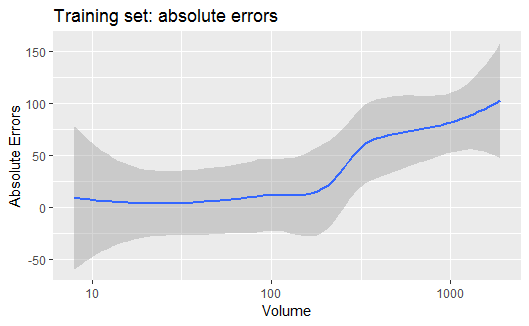
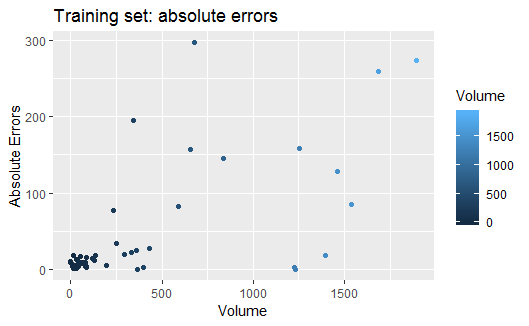
1. **Investigation of the model’s errors**

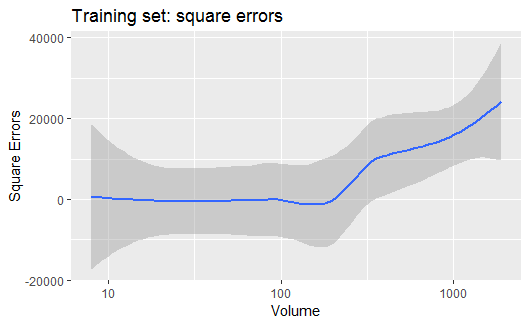
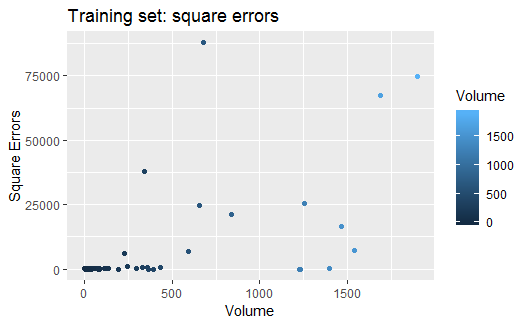
Plots of the absolute and squared errors for the testing set indicate that the model works best at predicting volumes of below 500. See Figure 14 below. This conclusion was investigated further by looking at the absolute and squared errors when the model is applied to the training set. Figures 15 show that the same conclusion can be made.

Figure 14: plots of the absolute errors and squared-errors for the test set



Figures 15: plots of the absolute errors and squared-errors for the training set





1. **Make the predictions**

The table below shows the predictions for all the new products in the dataset.

|  |  |
| --- | --- |
| ProductType | Volume |
| PC | 304.312 |
| PC | 133.6812 |
| Laptop | 157.485333 |
| Laptop | 34.3001333 |
| Laptop | 15.6581714 |
| Netbook | 49.6374667 |
| Netbook | 1051.01453 |
| Netbook | 127.383733 |
| Netbook | 29.4072 |
| Tablet | 1187.35347 |
| Tablet | 1349.47 |
| Smartphone | 276.870533 |
| Smartphone | 387.8512 |
| Smartphone | 82.6678667 |
| Smartphone | 140.610933 |
| GameConsole | 1270.3912 |
| Display | 15.7169244 |
| Accessories | 29.6793905 |
| Accessories | 33.2789333 |
| Software | 131.053333 |
| Printer | 84.3570667 |
| PrinterSupplies | 16.1070577 |
| ExtendedWarranty | 10.2691048 |
| GameConsole | 1377.48091 |