***Data Analytics Report***

**PRODUCT LAUNCH – Predicting Sales Volume**

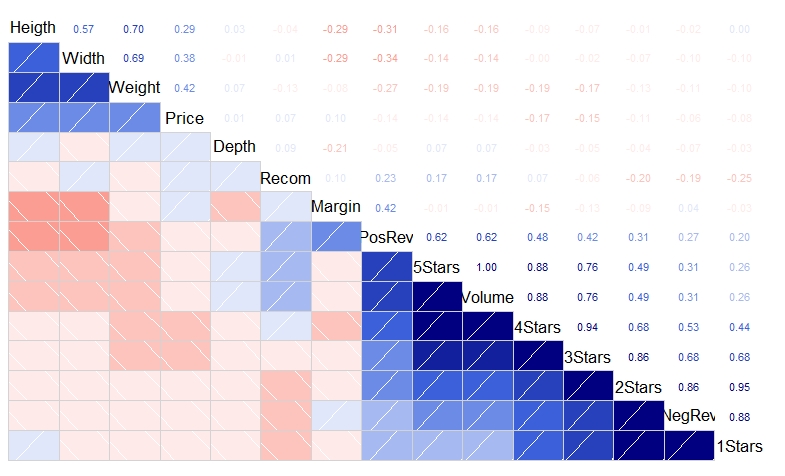
|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date** | **Author** | **Comments** |
| 1.0 | June 19th | Steffen Adolf | Initial write |
|  |  |  |  |
|  |  |  |  |

**1. Explorative Data Analysis**

**1.1 Feature Selection**

Collinearity of features in the data set bears the danger of overfitting the model, that is used for predicting sales volume of our new potential products.

Therefore, removing redundant features on the basis of correlation between the independent variables and the target variable “Volume” is necessary.

**Correlation Matrix**

Features removed:

X5StarsReviews

X3StarsReviews

X1StarReviews

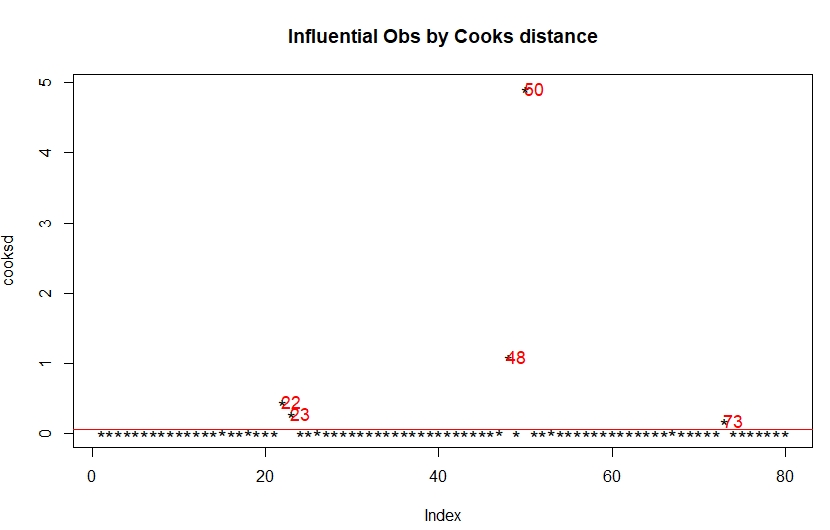
Keeping X4StarsReviews and X2StarsReviews is more intuitively reasonable, because giving 4 or 2 stars seems to be more thoughtful. It also avoids positive and negative motivated attacks that just want to boost or damage a product’s reputation.

**1.2 Outliers Evaluation**

Outliers can have high influence on the design of the prediction model and with it on the prediction results. It has to be decided to what extent they represent the population and if they have to be removed.

“Cook’s Distance” helps to measure the influence of observations on a model design. Those observations can also be interpreted as “Outliers”.

The lm-model used for calculating “Cook’s Distance” has included the following features: Price+x4StarReviews+x2StarReviews+NegativeServiceReview



Removed observations:

|  |  |  |  |
| --- | --- | --- | --- |
| Index | ProductNum | ProductType | Volume |
| 22 | 122 | Software | 1576 |
| 23 | 123 | Software | 2052 |
| 48 | 148 | Accessories | 2140 |
| 50 | 150 | Accessories | 11204 |
| 73 | 198 | GameConsole | 7036 |

The model performance shows that removing outliers improves the performance in this case.

**2.Model Testing and Evaluation**

**2.1 Support Vector Machine**

Support Vector Machines with Linear Kernel

59 samples

23 predictors

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 54, 51, 52, 53, 54, 52, ...

Resampling results across tuning parameters:

cost RMSE Rsquared MAE

0.25 210.5849 0.8974173 134.2265

0.50 208.6749 0.9027956 133.7654

1.00 206.9908 0.9067095 132.5764

2.00 207.4065 0.8943695 132.3012

4.00 208.0678 0.8892041 132.8870

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was cost = 1.

**2.2 Random Forest**

Random Forest

59 samples

23 predictors

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 52, 54, 51, 52, 53, 53, ...

Resampling results across tuning parameters:

mtry RMSE Rsquared MAE

1 344.8673 0.8163945 278.6037

2 253.7391 0.8886410 185.0815

3 224.6205 0.9084589 152.4309

4 201.4261 0.9247081 132.5625

5 196.6853 0.9236354 126.7466

RMSE was used to select the optimal model using the smallest value.

The final value used for the model was mtry = 5.

**2.3 Model Comparison**

RMSE

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

svm 48.10493 78.39624 137.8811 206.9908 315.5316 608.2883 0

rf 16.62152 64.51266 121.4811 129.4576 186.9574 319.1209 0

Rsquared

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

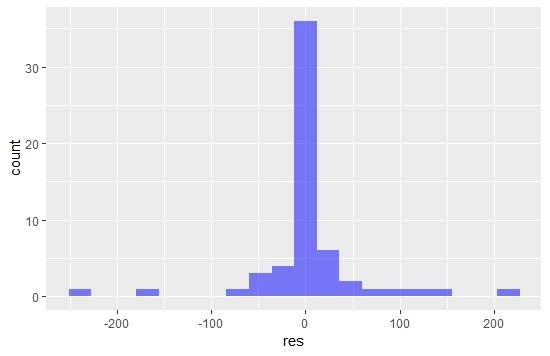
svm 0.7023764 0.8519600 0.9591483 0.9067095 0.9748716 0.9937307 0

rf 0.7283460 0.9209822 0.9770266 0.9472416 0.9876537 0.9997209

The Random Forest Model is performing better on this data set in regards to RMSE and also R^2. Predictions made on the test set show are R^2 value of 0.824. which is slightly worse than for the training set but variance is expectable in such a small data set.

That’s why RandomForesst has been chosen for making predictions in the further process.

The distribution of the residuals of the trained model are also quite normal distributed, which is a good sign for the model quality.



Variable importance for Random Forest model:

rf variable importance

only 20 most important variables shown (out of 23)

Overall

PositiveServiceReview 100.000

x4StarReviews 61.643

Recommendproduct 18.572

ProductDepth 16.813

x2StarReviews 9.320

ShippingWeight 7.796

ProfitMargin 6.512

ProductType.Printer 6.339

ProductType.ExtendedWarranty 5.979

Price 5.092

ProductWidth 4.856

ProductType.Netbook 4.366

NegativeServiceReview 3.701

ProductType.Laptop 3.673

ProductType.Accessories 3.286

ProductType.GameConsole 2.734

ProductType.Software 2.523

ProductType.Smartphone 1.541

ProductHeight 1.358

ProductType.Display 1.352

So Positive Service Reviews and positive ratings are very important feature in regard to predicting volume of sales with this model.

**3. Predicting Sales Volume and Profitability**

The table shows the calculated profit we can expect for each new product on the bases of the “Volume” that has been predicted by the chosen machine learning model “Random Forest”.

The profit has been calculated by multiplying “Price”, ”ProfitMargin” and “Volume”.

The 4 product type categories, “PC”, “Tablet”, “Laptop” and “Smartphone” that we are mostly interested in are colorized.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | ProductType | ProductNum | Price in $ | ProfitMargin | predicted\_Volume | Profit in $ |
| 1. | GameConsole | 307 | 425 | 0.18 | 1296 | 99144 |
| 2. | PC | 171 | 699 | 0.25 | 475 | 83006 |
| 3. | Tablet | 186 | 629 | 0.1 | 1269 | 79820 |
| 4. | Tablet | 187 | 199 | 0.2 | 1251 | 49790 |
| 5. | GameConsole | 199 | 249.99 | 0.09 | 1337 | 30081 |
| 6. | Netbook | 180 | 329 | 0.09 | 998 | 29551 |
| 7. | PC | 172 | 860 | 0.2 | 139 | 23908 |
| 8. | Laptop | 173 | 1199 | 0.1 | 152 | 18225 |
| 9. | Printer | 304 | 199.99 | 0.9 | 75 | 13499 |
| 10. | Smartphone | 193 | 199 | 0.11 | 429 | 9391 |
| 11. | Laptop | 176 | 1999 | 0.23 | 14 | 6437 |
| 12. | Netbook | 181 | 439 | 0.11 | 124 | 5988 |
| 13. | Laptop | 175 | 1199 | 0.15 | 33 | 5935 |
| 14. | Smartphone | 196 | 300 | 0.11 | 152 | 5016 |
| 15. | Smartphone | 194 | 49 | 0.12 | 556 | 3269 |
| 16. | Netbook | 178 | 399.99 | 0.08 | 67 | 2144 |
| 17. | Smartphone | 195 | 149 | 0.15 | 76 | 1699 |
| 18. | Software | 303 | 70.99 | 0.2 | 113 | 1604 |
| 19. | Netbook | 183 | 330 | 0.09 | 28 | 832 |
| 20. | ExtendedWarranty | 306 | 99.99 | 0.4 | 14 | 560 |
| 21. | Display | 201 | 140 | 0.05 | 23 | 161 |
| 22. | PrinterSupplies | 305 | 20.99 | 0.3 | 19 | 120 |
| 23. | Accessories | 302 | 8,5 | 0.1 | 55 | 47 |
| 24. | Accessories | 301 | 20.99 | 0.05 | 33 | 35 |

The results are quite similar but not identical compared to the analysis in the last report. The two Tablets (186,187), the Netbook (180) and the Game Console (199) are among the most profitable products again. The Laptops performing a little bit worse and PC’s better. But this kind of variance is absolutely normal in predictive modelling. There are no 100 % on-point predictions.

**4.Recommendations**

The recommendations in regard to the launch of new products are confirmed with the analysis. Especially types of tables (186,187) and a netbook (180) seem to be promising high profitable candidates. But we should be aware of some substitutional effect that a new product could have on the established product portfolio.

R as a software is much more flexible than RapidMiner. Analysis can be more specific in regard to the task. That’s why the usage of it should be continued.

The model has shown that positive service and product reviews are very important attributes when it comes to explaining high sales volumes of a product.

This doesn’t mean that there is a direct causality, but it could be used as a clear indicator of the performance of a product when a potential launch in a different store is on the agenda.