**Eva Lipska**

**Sales Prediction Report**

task 3, course 3

**Goal of the project:**

Performing sales volume predictions for a list of new product types on a basis of historical sales data analysis including:

- predicting sales of four different product types: PC, Laptops, Netbooks and Smartphones;

- assessing the impact service reviews and customer reviews have on sales of different product types.

Data: historical sales data and new product data set.

**Selection of the prediction model**

*Data preparation:*

Categorical variables have been converted to binary values (“dummy variables”) so they can be used directly as predictor in a multiple regression model.

Dealing with non-available data:

- removing BestSellersRank

Addressing overfitting:

- removing x5StarReviews: as correlation between independent variable (x5StarReviews) and dependent variable (Volume) is perfect.

Three algorithms appropriate for non-parametric models have been used during the analysis:

- Support Vector Machine

- Random Forest

- Gradient Boosting

Best performance – highest RSquared, lowest RMSE – was obtain through the use of *Random Forest:*

|  |  |  |
| --- | --- | --- |
| mtry | RMSE | Rsquared |
| 26 | 195.6918 | 0.9183356 |

Other algorithms presented following results:

*Support Vector Machine*

|  |  |  |
| --- | --- | --- |
| cost | RMSE | Rsquared |
| 1.00 | 320.2370 | 0.8589161 |

*Gradient Boosting*

|  |  |  |
| --- | --- | --- |
| n. trees | RMSE | Rsquared |
| 100 | 224.7400 | 0.9046193 |

**Prediction of the sales volume based on Random Forest**

|  |  |
| --- | --- |
| **ProductType** | **Volume** |
| PC 1 | 470.3112 |
| PC 2 | 135.2131 |
| Laptop 1 | 171.3948 |
| Laptop 2 | 20.12893 |
| Laptop 3 | 11.9064 |
| Netbook 1 | 76.0364 |
| Netbook 2 | 1490.092 |
| Netbook 3 | 146.5695 |
| Netbook 4 | 18.8176 |
| Smartphone 1 | 383.0955 |
| Smartphone 2 | 685.9441 |
| Smartphone 3 | 78.3248 |
| Smartphone 4 | 166.3227 |

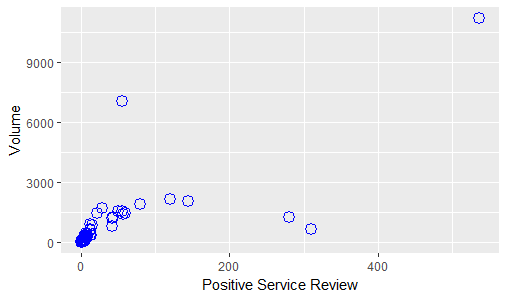
**Business recommendations based on existing product attributes dataset**

Activities should be focused on achieving positive service reviews as well as positive customer reviews.

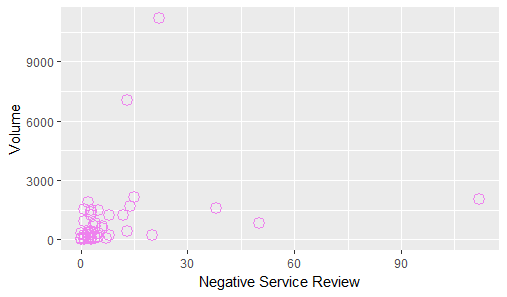
In both cases positive reviews seem to influence sales volume in greater scale than negative reviews.

Product type is not a good predictor for sales. In general 45% of existing product sales volume are in the category of accessories (including 26 different products), 18 % - extended warranties (10 different products) and 15 % - game consoles (2 products)

**A chart displaying the impact of service reviews have on sales volume.**

****

General trend concerning positive service review indicates that positive service reviews influence sales volume.

****

There is no significant trend indicating relation between negative service reviews and sales volume. Strong negative service reviews don’t lead to lower sales value. Moreover there are a number of outliers which should be a subject to a further investigation.

**Lessons learned:**

- models should be tried before addressing collinearity;

- “dummy variables” as a solution for using categorical data in predictive modeling;

- there shouldn’t be any nominal data, as well as NAs in the dataset during verification of the correlation;

- linear regression is not an appropriate model for non-parametric data.