Task 6

Report: Predicting Profitability with R

Barcelona, October 2018

Student:
Javier E. Villasmil
Helena Bonan

ÍNDEX

1. SCOPE 1
2. EXECUTIVE SUMMARY 1
3. METHODOLOGY 2
4. ANALYSIS 2
4.1 PRE-PROCESSING 2
4.2 FEATURE SELECTION 3
4.2.1 CORRELATION MATRIX 3
4.2.2 DECISION TREE 4
4.2.3 OUTLIERS 4
4.3 PREDICTIVE MODELS 5
4.3.1 MULTIPLE LINEAR REGRESION (MLR) 6
4.3.2 RANDOM FOREST (RF) 6
4.3.3 SUPPORT VECTOR MACHINE (SVM) 6
4.3.4 GRADIENT BOOSTED TREES (GBT) 6
5. PREDICTIONS 6

LIST OF FIGURES

Fig 1. Product Type against Volume. 2
Fig 2. Correlation Matrix. 3
Fig 3.Decision Tree results: Attributes importance against the label (sales volume) 4
Fig 4. Scatter plot of sales volume against the independent variables. 4
Fig 5. Distribution of Training and Testing subsets. 5

LIST OF TABLES

Table 1. Multiple Linear Regresion Performance Metrics. 6
Table 2. Random Forest Performance Metrics. 6
Table 3. Support Vector Machine Performance Metrics. 6
Table 4. Gradient Boosted Trees Performance Metrics items. 6
Table 5. Selected Models Performance Metrics. 6
Table 6. Predicted volumes for each products using RF, SVM and Boosted Trees. 7

1. SCOPE

Analyze using data mining and modeling methods the dataset supplied by the CTO and head of Blackwell's eCommerce Team – the goal is to predict ***sales volume*** and ***profitably*** of a list of new products.

This dataset contains information about different product types that Blackwell sells, as well as important attributes for their classification (price, reviews, product size, margin, etc.)

2. EXECUTIVE SUMMARY

RESULTS

In the following table, you can see the predicted Volume for the potential new products. The products are summarized in descending order regarding the predicted Sales. We have highlighted in green the 'best' new product for every category we have considered: PC, Netbook, Laptop and Smartphone.

Product Type	Product Number	Volume (predicted)	Sales Profit (predicted)
PC	**171**	**489,66**	**$ 85.568**
Netbook	**180**	**1215,20**	**$ 35.982**
PC	172	128,77	$ 22.149
Laptop	**173**	**175,99**	**$ 21.102**
Smartphone	**193**	**444,62**	**$ 9.732**
Laptop	175	38,93	$ 7.002
Laptop	176	14,86	$ 6.836
Netbook	181	136,45	$ 6.589
Smartphone	196	157,48	$ 5.196
Smartphone	194	684,75	$ 4.026
Smartphone	195	87,18	$ 1.948
Netbook	178	60,02	$ 1.920
Netbook	183	24,34	$ 722

Selected products by type and sales.

MODEL USED

To predict we have use the ***random forest algorithm***. The independent variables we have considered are ***Four Star Reviews***, ***Two Star Reviews*** and ***Positive Service Reviews***.

CONCLUSION

- To implement our model we have used only products with Volume at most equal to *2140 units*. For this, reason this reason we can predict accurately low and mid-range volumes.
- It was concluded that the services reviews and costumer's reviews have a high impact on the volume sales of the different product types. This was verified using a correlation matrix and a decision tree.

SUGGESTION FOR FURTHER ANALYSIS

The data set was very small and with products that are completely different one from each other. This limits the quality of the prediction. To make a better analysis we need more data for every type of object. In this way, we could consider also other features of the products (like the price) and not just the reviews of the costumers.

In addition, there is a too strong relation between the 5 Star Rating Variable and Sales Volume. This could be due to an error during the acquisition of data. If it were not the case, we would need more samples in order to study the 'correct' relation between these two attributes.

3. METHODOLOGY

The approach used to assess the dataset was to apply basics methods of data mining, descriptive statistics and simple charting to observe the distribution and correlation between variables (features) in our dataset.

In addition, R / Rstudio helped with the evaluation of histograms, scatter plots, correlation matrixes, and with the cleaning / transformation of the data.

On the other hand, for the predictive models, the algorithms used were ***Linear Regression (LR)***, ***Random Forest*** *(rf)*, ***Support Vector Machine*** *(SVM)* and ***Gradient Boosted Tree (GBT)***, these algorithms were tuned according to each own parameters.

To decide the best-fitted model, error metrics such as Root mean squared error (RMSE), Mean absolute error (MAE) and R-squares were used.

4. ANALYSIS

4.1 PRE-PROCESSING

The attribute named *"BestSellerRank"* had missing values in twenty percent of the samples and since there were not indicatives of how the company determined this feature and how it related to the other columns; the missing values could not be filled.

Leaving the column as provided would affect our model so the best option was to remove the feature.

On the other hand, columns *"ProductType"* and *"ProductNumber"* were also removed due to the nature of each; *"ProductNumber"* is just an ID without any data value and *"ProductType"* is a categorical variable and since there are only eighty (80) records in the dataset the twelve (12) products are not statistically representative for a predictive model. See fig 1.

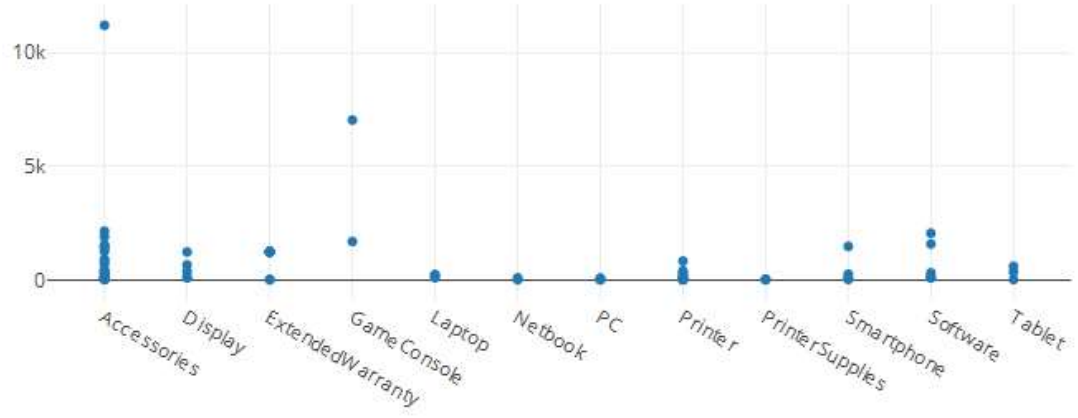

Fig 1. Product Type against Volume.

4.2 FEATURE SELECTION

Selecting which columns are representative for our model to predict the sales volume is a task where the analyst needs to compare different relationships between your dependent variable and predictors.

To do this, three tools were used: Scatter charts, correlation matrixes and decision tree.

4.2.1 CORRELATION MATRIX

A correlation matrix shows the correlation coefficients between sets of variables. This allowed us to see which pairs of attributes have the highest linear correlation. See Fig 1.

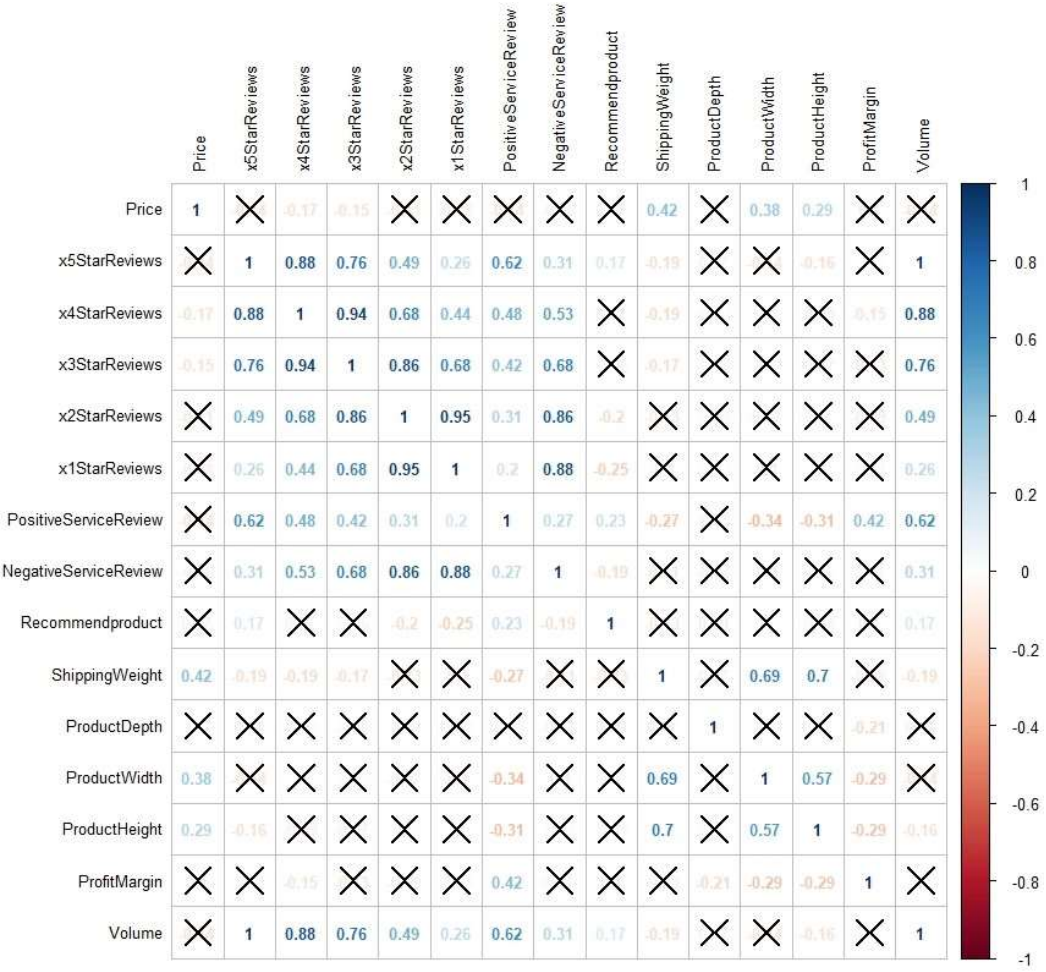

Fig 2. Correlation Matrix.

In this case, the feature *"FiveStarRank"* was removed because it is correlated with a coefficient of one (perfect fit) against volume - leaving this column would cause an over fit in our model.

Additionally, it is important to compare collinearity between independent variables. If one, have a correlation coefficient of +-0.85 or higher with any of the other independent variables, one of the features must be removed to address the issue.

- "FourStarRank" is highly correlated to "ThreeStarRank"
- "TwoStarRank" is highly correlated to "OneStarRank"
- "TwoStarRank" is highly correlated to "NegativeServiceReview"

4.2.2 DECISION TREE

The first approach was to check the relationship of all the variables (predictors) against the label (dependent variable) and their respective importance. The parameter used to tune the decision tree was ***complexity number (cp): 0.02***.

Attributes	Importance
Four Star Reviews	47 %
Two Star Reviews	34 %
Positive Service Review	10 %
Profit margin	3 %
Product Depth	3 %
Product Width	3 %
Price	0 %
Shipping Weight	0 %
Profit margin	0 %
Would recommend	0 %
Product Depth	0 %
Product Width	0 %
Product Height	0 %
Product Type	0 %

Fig 3.Decision Tree results: Attributes importance against the label (sales volume)

As seen in Fig 3, the best variables to use in our modelling according to decision tree are *Four Star Reviews*, *Two Star Reviews* and *Positive Service Review*.

All Attributes with importance zero can be removed without affecting our predictive model.

We decided not to consider *Profit Margin*, *Product Depth* and *Product Width* because their relevance is lower than the first three and the common sense dictates in this case that these variables are not related to the volume.

It appears that costumers are buying products by their rating and reviews instead of price and dimensions.

NOTE: In the decision three (fig.2) one can see only Four Star Reviews and Positive Service Review. We assume that the reason why we do not see Two Star Reviews (also if we increase the complexity) is that Two Star Reviews is high correlated to Four Star Reviews.

4.2.3 OUTLIERS

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In this case, we detect several interesting values after plotting the remaining independent variables against volume. See fig 4.

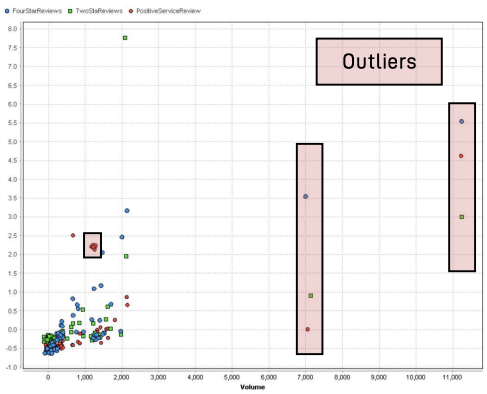

Fig 4. Scatter plot of sales volume against the independent variables.

A selection of ***nine (9) registers*** were removed after analyzing their impact over the predictive model.

- *Two (2) records with over 7.000 volume in sales* – this will improve our model by making more conservative estimates when predicting lower volumes.
- *Seven (7) records with the same value in all the features* – it appears that there was an error in the supplied dataset, where one record repeated itself eight times. The approach to fix this problem was to remove seven records and leave one inside our data sample.

4.3 PREDICTIVE MODELS

The algorithms used to make models that could predict our new products volumes are ***Linear Regression*** *(LR)*, ***Random Forest*** *(rf)*, ***Support Vector Machine*** *(SVM)* and ***Gradient Boosted Tree*** *(GBT)*.

For each algorithm, three models were tested.

- ***Model 1*** – Using one predictor: Four Stars Reviews.
- ***Model 2*** – Using two predictors: Four Stars Reviews and Two Stars Reviews.
- ***Model 3*** – Using three predictors: Four Stars Reviews, Two Stars Reviews and Positive Service Reviews.

In addition, we checked that the distribution of training and testing are similar by using bar plots. See fig 5.

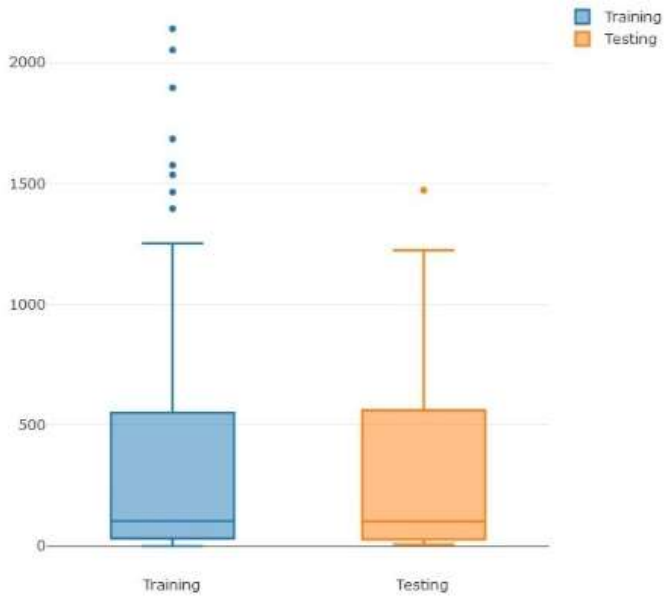

Fig 5. Distribution of Training and Testing subsets.

On the other hand, to validate and train our models the method used was a *ten-fold cross validation* with *three (3) repeats.*

In the following tables, it is presented the performance metrics (errors) of each model against the testing set.

It is important to note that the models using the ***three main variables*** are ***the most accurate*** against the test subset**.**

4.3.1 MULTIPLE LINEAR REGRESION (MLR)

	MLR – Model 1	MLR – Model 2	MLR – Model 3
RMSE	336.74	336.57	331.35
R-Squared	0.54	0.53	0.56
MAE	193.71	190.76	182.27

Table 1. Multiple Linear Regresion Performance Metrics.

4.3.2 RANDOM FOREST (RF)

	RF – Model 1	RF – Model 2	RF – Model 3
RMSE	144.08	142.07	86.35
R-Squared	0.92	0.93	0.97
MAE	67.40	72.16.76	41.89

Table 2. Random Forest Performance Metrics.

4.3.3 SUPPORT VECTOR MACHINE (SVM)

	SVM – Model 1	SVM – Model 2	SVM – Model 3
RMSE	313.59	299.61	52.46
R-Squared	0.62	0.65	0.99
MAE	135.33	130.18	48.86

Table 3. Support Vector Machine Performance Metrics.

4.3.4 GRADIENT BOOSTED TREES (GBT)

	GBT – Model 1	GBT – Model 2	GBT – Model 3
RMSE	247.85	256.64	88.83
R-Squared	0.76	0.74	0.97
MAE	103.81	143.49	52.60

Table 4. Gradient Boosted Trees Performance Metrics items.

5. PREDICTIONS

We selected the best model for every algorithms and compared them against each other. See table 5.

	MLR – Model 3	RF – Model 3	SVM – Model 3	GBT – Model 3
RMSE	331.35	86.35	52.46	88.83
R-Squared	0.56	0.97	0.99	0.97
MAE	182.27	41.89	48.86	52.60

Table 5. Selected Models Performance Metrics.

The prediction of the new products were done using the ***RF***, ***SVM*** and ***GBT*** models.

The MLR model was not included because its performance was not as good as the others.

Product Type	Product Number	Prediction RF	Prediction SVM	Prediction GBT
PC	171	489,66	600,08	535,42
PC	172	128,77	661,65	86,59
Laptop	173	175,99	689,72	181,97
Laptop	175	38,93	-83,58	49,49
Laptop	176	14,86	120,97	35,89
Netbook	178	60,02	44,96	73,90
Netbook	180	1215,20	827,8	1167,82
Netbook	181	136,45	827,72	54,47
Netbook	183	24,3410	-50,60	35,89
Tablet	186	1214,25	827,83	1305,90
Tablet	187	1897,51	827,87	2052,16
Smartphone	193	444,62	697,82	427,35
Smartphone	194	684,75	827,87	650,47
Smartphone	195	87,18	207,54	73,90
Smartphone	196	157,48	826,32	70,80
GameConsole	199	1309,93	834,25	1166,14
Display	201	18,57	42,51	35,89
Accessories	301	29,82	-10,09	49,49
Accessories	302	60,80	142,28	49,49
Software	303	127,72	174,05	73,90
Printer	304	83,06	117,58	73,90
PrinterSupplies	305	18,57	42,51	35,89
ExtendedWarranty	306	4,03	1,06	35,89
GameConsole	307	1600,98	827,87	1931,23

Table 6. Predicted volumes for each products using RF, SVM and Boosted Trees.

After checking the predictions, the SVM model was discarded because it predicted negatives values and the volumes predicted differed in range from the other two models. See Table 6.

Between the RF and GBT models, the ***RF was selected*** for having the best performance metrics.

In conclusion, the model selected to make the new-products predictions was the ***Random Forest with three variables.***