Date: 18 Dec 2018

Multiple Regression in R

Authors: Maria Farrés & Florian Unger

Study Objectives

The following report is intended to help the sales team better understand how types of products might impact sales across the enterprise by:

- Predicting sales and profitability of four different product types that are bringing up sales concerns in one of the stores of Blackwell Electronics; these product categories that require further study are PCs, Laptops, Netbooks and Smartphones.
- Assessing the impact service and customer reviews have on sales of different product types.

1. Executive Summary

The hereby report outlines the process of modelling and selecting new predictive models to understand the role of *product types* in our company's sales. Therefore, the models raise sales and profitability predictions for the concerning product categories in our store, and help us study further the variables that affect product sales in Blackwell Electronics.

The study proves that '*Product Type'* is actually unrelated to 'V*olume'*. However, as explained in the last data analysis regarding product profitability in Blackwell Electronics, the new prediction results also show a really promising scenario for some of the categories that concern our store**.**

In terms of sales volume:

- The categories ***Netbooks, Smartphones, PCs and Laptops*** **are in the Top 6 of the predicted most demanded products**, only following *tablets* and *Game Consoles* which happen to be the predicted most sold categories.

- The predicted Top 5 best sellers among the concerning categories is: Netbook Acer 180, Smartphone Samsung 194, PC Dell 171, Smartphone Motorola 193 & Laptop Apple 173.

In terms of profitability:

- The categories ***PCs, Netbooks, Laptops and Smartphones*** **are in the Top 6 of the predicted most profitable products**, only following *Game Consoles* and *Tablets* which happen to be the predicted most profitable categories.

- The predicted Top 5 most profitable products among the concerning categories is: PC Dell 171, Netbook Acer 180, PC Dell 172, Laptop Apple 173 and Laptop Toshiba 175.

This results have been collected by a *RandomForest* model including 3 features, *4 star reviews, Positive Service Reviews & Product Depth* (***See Feature and Model Selection in the Technical analysis***). Therefore, the study proves that, whether Negative Service Reviews is not affecting relevantly to sales volume, Positive Service Reviews has a significant impact on it.

If concern still arises after this report among the store's top management, the data analysis team suggests a meeting to discuss further and solve any concerns regarding the company's portfolio direction.

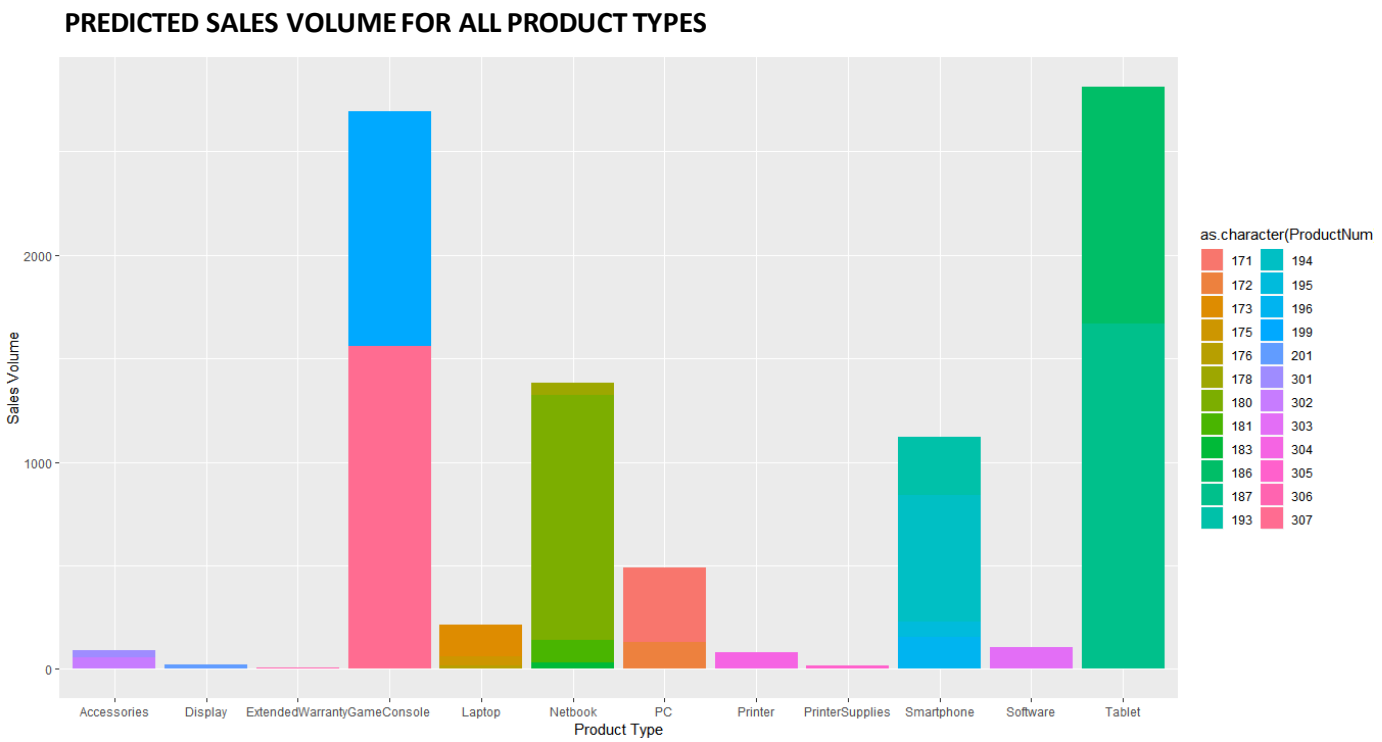
2. Results

The results presented gather the predicted sales and the consequent profitability for all the product types, and especially for those categories that are raising concerns in one of our stores, and their consequent profitability for Blackwell Electronics.

Next on, the final table containing the sales volume predictions raised by the chosen model (Random Forest) and the calculated profitability, is presented:

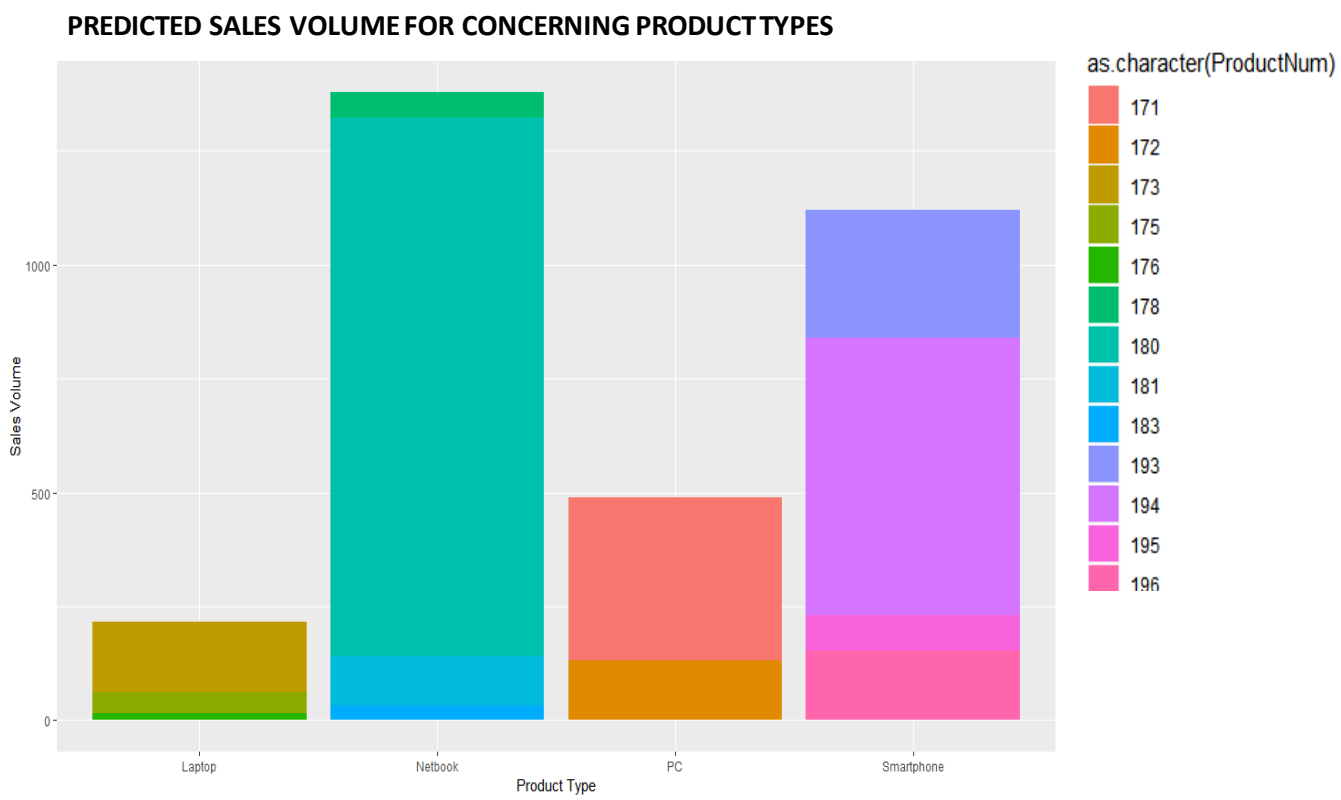
ProductType	ProductNum	Price	ProfitMargin	PredictedVolumeRF	Profitability
PC	171	699.0	0.25	358.40	62630.8
PC	172	860.0	0.20	130.77	22492.3
Laptop	173	1199.0	0.10	153.23	18372.3
Laptop	175	1199.0	0.15	45.95	8264.9
Laptop	176	1999.0	0.23	15.94	7330.9
Netbook	178	400.0	0.08	56.60	1811.2
Netbook	180	329.0	0.09	1182.94	35026.8
Netbook	181	439.0	0.11	107.49	5190.7
Netbook	183	330.0	0.09	33.18	985.5
Tablet	186	629.0	0.10	1146.87	72138.1
Tablet	187	199.0	0.20	1666.74	66336.3
Smartphone	193	199.0	0.11	279.82	6125.3
Smartphone	194	49.0	0.12	610.47	3589.6
Smartphone	195	149.0	0.15	77.23	1726.0
Smartphone	196	300.0	0.11	153.47	5064.5
GameConsole	199	250.0	0.09	1134.69	25529.6
Display	201	140.0	0.05	22.35	156.4
Accessories	301	21.0	0.05	34.57	36.3
Accessories	302	8.5	0.10	54.78	46.6
Software	303	71.0	0.20	108.04	1533.9
Printer	304	200.0	0.90	80.99	14576.9
PrinterSupplies	305	21.0	0.30	15.33	96.5
ExtendedWarranty	306	100.0	0.40	7.39	295.6
GameConsole	307	425.0	0.18	1559.66	119313.8

Let's give it a more visual approach by checking the plots below:



From the graph above we may extract that the categories *Netbooks, Smartphones, PCs and Laptops* are in the Top 6 of the predicted most demanded products, only following *tablets* and *Game Consoles* which happen to be the predicted most sold categories.

Moreover, if we study the sales volume in these categories further, we discover the following:

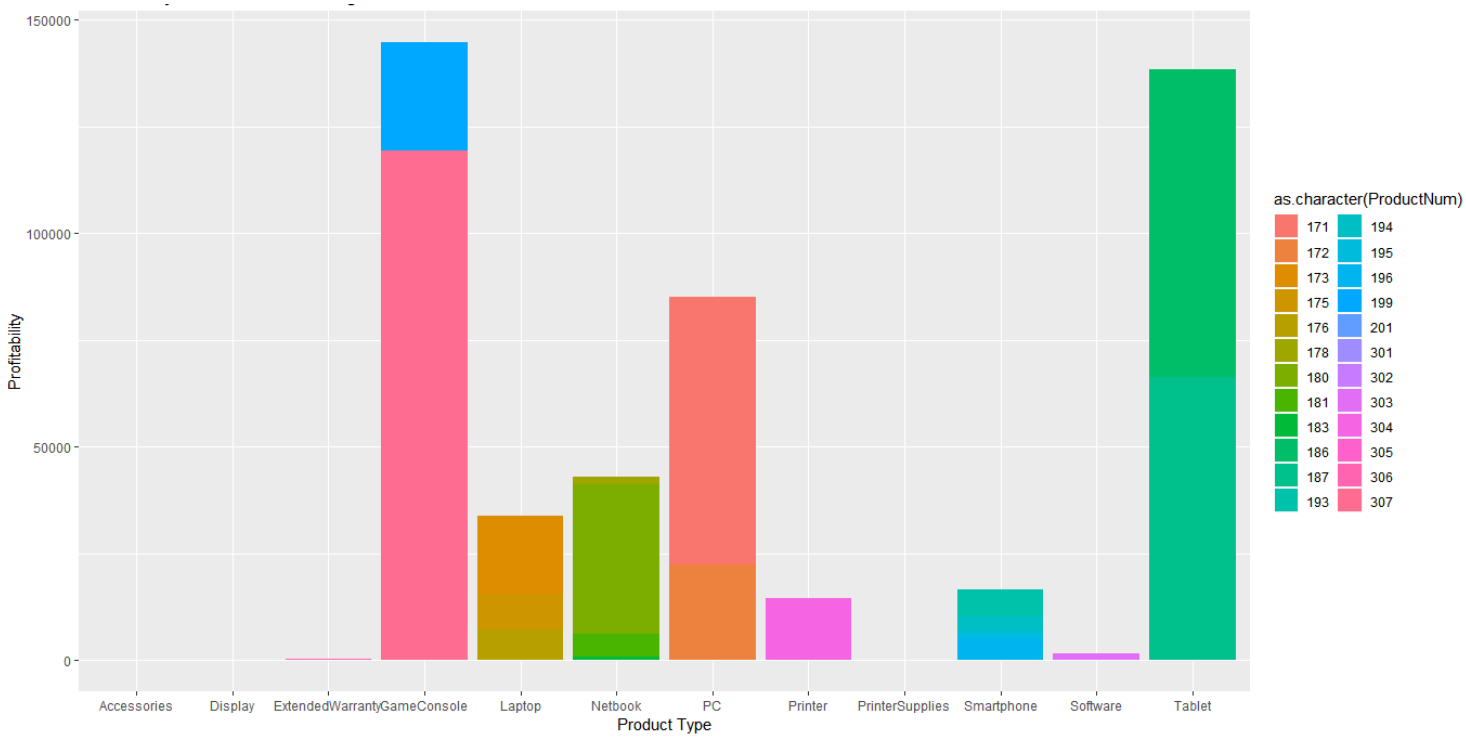


Looking at the four concerning product types (PCs, Netbooks, Laptops, Smartphones), our model predicts that **Notebooks are meant to become the category with a higher sales volume among the ones selected, followed by Smartphones, PCs and Laptops.**

Furthermore, we also get that the predicted Top 5 best sellers among the concerning categories is: Netbook Acer 180, Smartphone Samsung 194, PC Dell 171, Smartphone Motorola 193 & Laptop Apple 173.

Nevertheless, this results are not enough as sales need to be transformed to profitability results, so that the store knows. not only those products that will be sold the most, but also the ones that will generate more profit to their subsidiary. So, let's check the same bar graphs changing the y axis to '*Profitability'*:

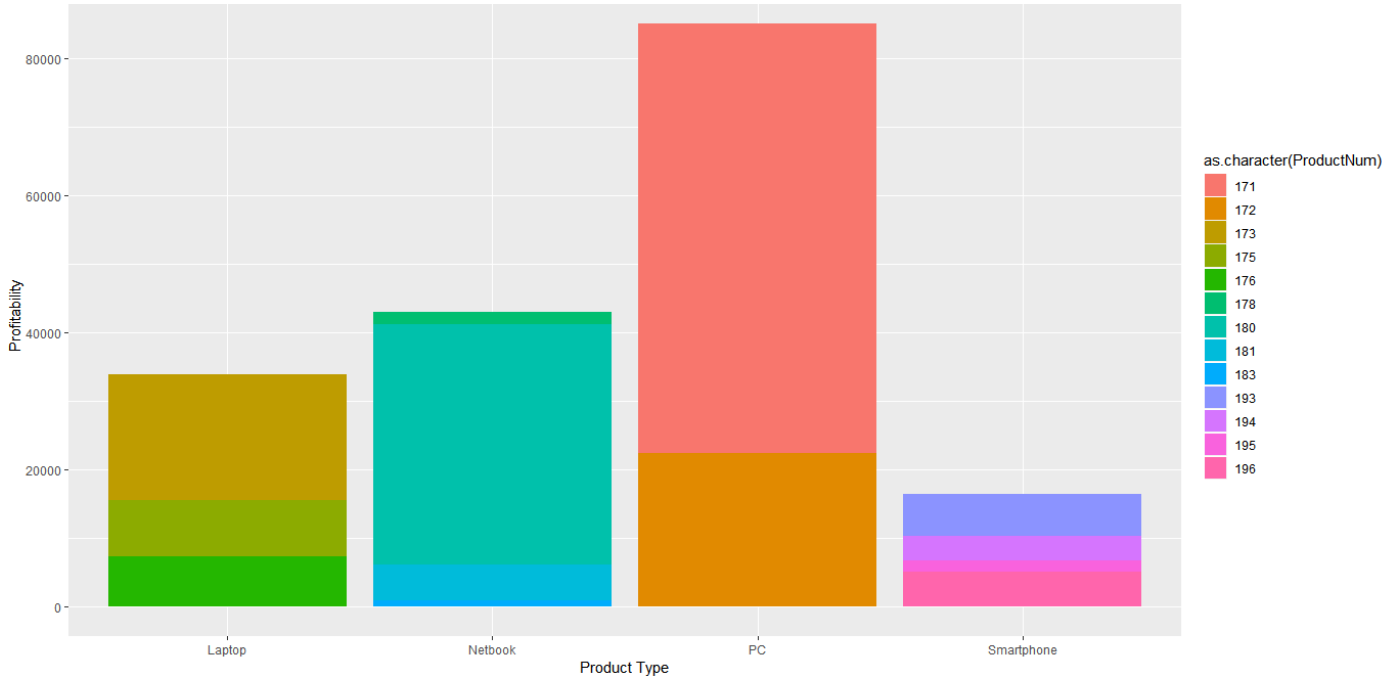
PREDICTED PROFITABILITY FOR ALL PRODUCT TYPES



The categories *PCs, Netbooks, Laptops and Smartphones* are in the Top 6 of the predicted most demanded products, only following *Game Consoles* and *Tablets* which happen to be the predicted most profitable categories.

Moreover, if we study the profitability in these categories further, we can also get the predicted Top 5 most profitable products among the concerning categories is: PC Dell 171, Netbook Acer 180, PC Dell 172, Laptop Apple 173 and Laptop Toshiba 175.

PREDICTED PROFITABILITY FOR CONCERNING PRODUCT TYPES



3. Limitations

The data provided counts with a really low amount of observations, so the predictions have huge limitations because the models do not have access to enough data to create reliable patterns.

If more data was provided, the stores could count with a deeper sales analysis, which would help them not only make informed decisions on what products to include or remove from their portfolio according to our customers' needs, but also to detect market trends, improve the supply chain efficiency, etc.

It is then advised, to collect more data from current sales not only in the store raising the issue, but from all of them. This, would help the headquarters establish regional buying patterns too, which could be really useful for the marketing department to run regional campaigns.

Annex - Technical Analysis

Overview

The hereby analysis has been carried out taking as a reference two provided data sets:

- **Existingproductattributes2017.csv**→ it provides sales information from Blackwell Electronics' existing products. The data set presents 80 observations in a normal distribution and is divided into 18 variables, including *Product type, Product Number, Price, 1 to 5 Star Reviews, Service reviews, Product Recommendations, Best seller ranks, Shipping weight, Product depth-width-height, Profit margin and Volume* each with 80 observations.
- **Newproductattributes2017.csv**→ it contains a list of the potential new products Blackwell Electronics is considering to add to its portfolio. It counts with 24 potential new products as single observations, each with information for all the attributes mentioned in existingproductattributes207.csv but for *Volume*, which becomes our dependent variable to further extract each product's profitability.

4. 1 Pre-processing

Looking at the histograms of the various variables, we observe that all variables, are mainly normally distributed. However, their distribution is heavily skewed.

4. 1. 1 Reclassifying and dummyfying variables

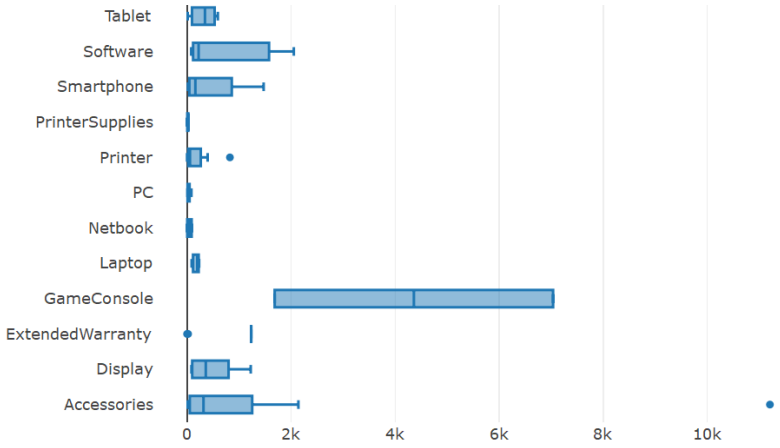
Firstly, we need to be aware that the task requires our team to run a multiple regression, as we need to derive the value of a criterion (Volume) from several other independent variables. Regression models cannot deal with 'chr' classes, therefore all our categorical variables need to be converted to binary features. Consequently, 'ProductType' is divided into nine dummy variables, according to each product type the data set contains.

```
#Reclassify variables

sapply(existing, class)
id = 2:18
existing[id] = data.matrix(existing[id])
sapply(existing, class)

#Dummy variables for producttypes

existing.dummified <- dummyVars(" ~ .", data = existing)
str(existing.dummified)
existing.final <- data.frame(predict(existing.dummified,
                                     newdata = existing)
```

4.1.2 Missing values, Outliers and Duplicates detection and treatment

In order to keep the results reliable, it is crucial to check for NA's, outliers and duplicates. The data set counts with 15 missing values located in the attribute 'BestSellerRank'. It has been advised to remove this attribute from the study to solve the missing values issue.

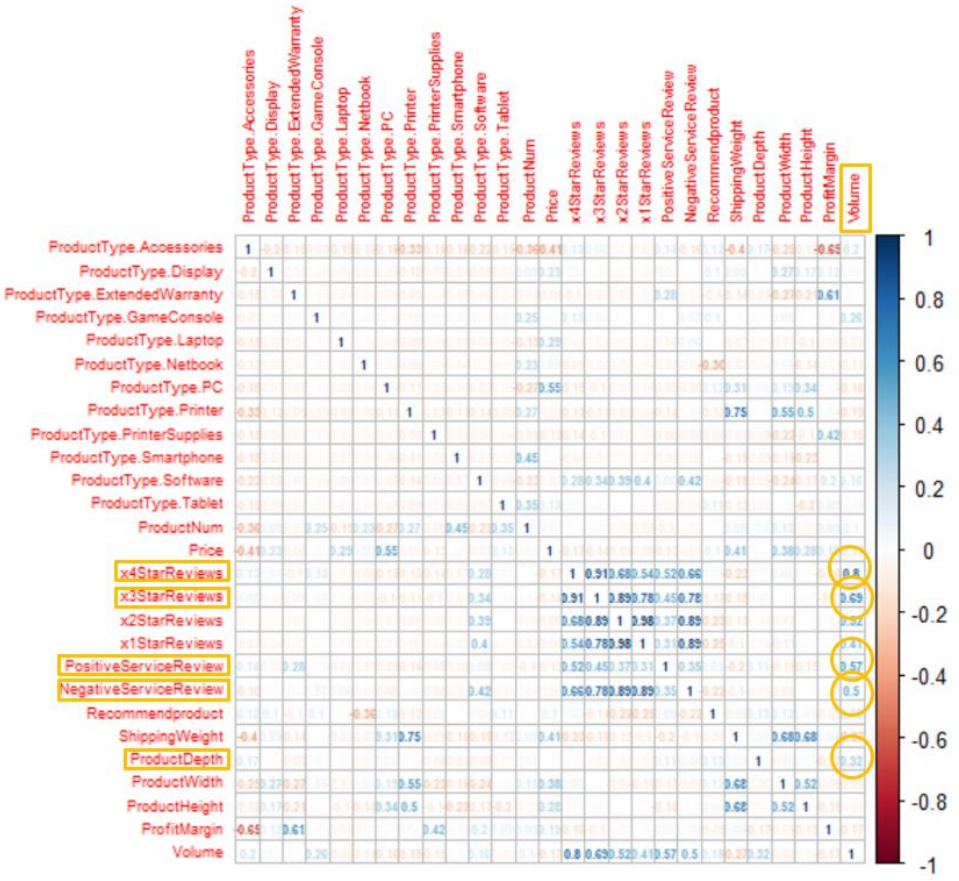
Regarding the outliers, although most of the variables present them, we will center our attention in the ones placed in our dependent variable 'Volume', as the provided data set is extremely short and it is not advisable to remove observations. So, two outliers are detected in 'Volume', an Accessory and a Game Console, and have been removed from the study.



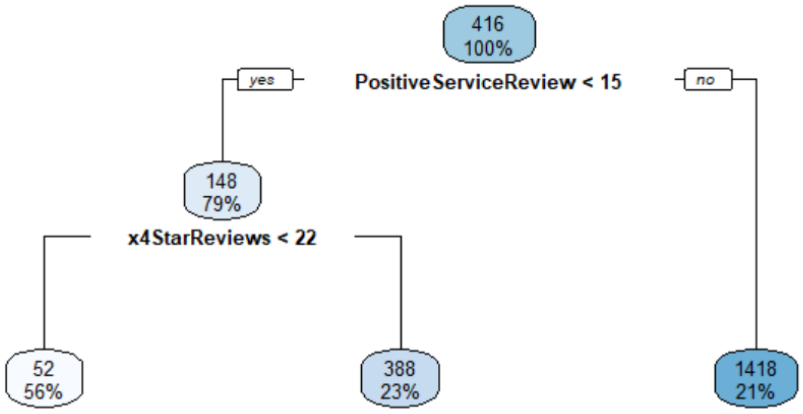
Besides, even though the data set presents no duplicates, we detect a possible issue in the product category 'Extended Warranty', as while its weight in the existing products data set is heavy (9 references with pretty similar values for all attributes but for price), in the new products one, the category's presence is almost nonexistent. This could affect our predictions, so we treat this category to lower its weight in the predictions and not let the numerous and duplicate values affect them.

4. 2 Attribute Selection

As our independent and dependent variables are of numerical nature we will refrain back to our correlation matrix, so we understand the linear relationships between our variables (see below). Thereby we can see that *x4StarReview* (0.8), *x3StarReview* (0.69), *PositiveServiceReview* (0.57), *NegativeServiceReview* (0.5) and *ProductDepth* (0.32) are the highest correlated with our dependent variable *Volume*. Due to the multicollinearity of 0.91 between *x4StarReview* and *x3StarReview*, we decided to exclude *x3StarReview*.



As a correlation matrix solely covers the linear relationships between the variables and dummy categories, we further conduct a decision tree analysis to understand the non-linear relationships between the variables. After conducting two decision trees analyses with two different packages in R leads us to the same conclusion, namely that *x4StarReviews* and *PositiveServiceReviews* are the most significant variables in determining our dependent variable *Volume*.



Furthermore, we run several multiple linear regressions to understand the linear significance (p-value) in determining Volume.

Linear Model #1: Volume ~

Within this model *x4StarReviews* and *ProductDepth* are significant in determining *Volume*, both with a p-value below 0.005, thereby contradicting the Null Hypothesis of insignificance. Henceforth, there is a 0.5% probability that the two variables above are insignificant. Furthermore, we can see that the Adj. R^2 is 78% while the p-value of the model itself lies below 0.05%, thereby making it significant.

```
Residual standard error: 268 on 45 degrees of freedom
Multiple R-squared:  0.863,     Adjusted R-squared:  0.787
F-statistic: 11.3 on 25 and 45 DF,  p-value: 2.99e-12
```

Linear Model #2: Volume ~ PositiveServiceReview + NegativeServiceReview + x4StarReviews + ProductDepth + 0

Within this model, we have removed the intercept, which has been suggested by previous models. In this model, we can see the clear significance of *PositiveServiceReview*, *x4StarReview* and *ProductDepth*. However, although suggested by the research team, *NegativeServiceReviews* is not significant in determining sales volume.

```
                     Estimate Std. Error t value Pr(>|t|)
PositiveServiceReview    1.638      0.742    2.21    0.031 *
NegativeServiceReview   -1.457      2.999   -0.49    0.629
x4StarReviews            9.033      0.928    9.73  1.9e-14 ***
ProductDepth             5.119      0.871    5.88  1.4e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 282 on 67 degrees of freedom
Multiple R-squared:  0.852,     Adjusted R-squared:  0.843
F-statistic: 96.2 on 4 and 67 DF,  p-value: <2e-16
```

Linear Model #3: Volume ~ PositiveServiceReview + x4StarReviews + ProductDepth + 0

Within this model, we have again removed the intercept, which has been suggested by previous models. In this model, we can see the clear significance of *ProductDepth*, *x4StarReviews* and *PositiveServiceReview*. Overall, compared to the previous two models, LM#3 is highly significant with a F-stats of 130 and an extremely small p-value.

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
PositiveServiceReview    1.632      0.737    2.21     0.03 *
x4StarReviews            8.748      0.716   12.21  < 2e-16 ***
ProductDepth             5.154      0.863    5.97  9.5e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 280 on 68 degrees of freedom
Multiple R-squared:  0.851,     Adjusted R-squared:  0.845
F-statistic:  130 on 3 and 68 DF,  p-value: <2e-16
```

After running the correlation matrix, the decision tree for non-linear relationships and several linear model (with standardized, non-standardized values, transformed and non-transformed variables), we decide to further progress with the following variables: ***PositiveServiceReview, x4StarReview* and *ProductDepth*.**

4. 3 Model Selection

Before we progress to applying the models, we have split the 80 observations dataset into a 75% (training) - 25% (testing) mix. However, this is quite a problem as the *ProductType* categories are very small, such as seen in the game console, which only has one single observation in the training set after the mix. This again highlights the difficulty to split up the data set according to the different categories which is the reason why we are using dummy variables. As the dummy variables fit into the Multiple Regression, their coefficient in the equation is only utilized for the specific dummy/product category (eg.: *Volume* goes up by x, when we have specific product). However, if we have only one example of a specific product, we need to assume that our estimation is lacking in accuracy.

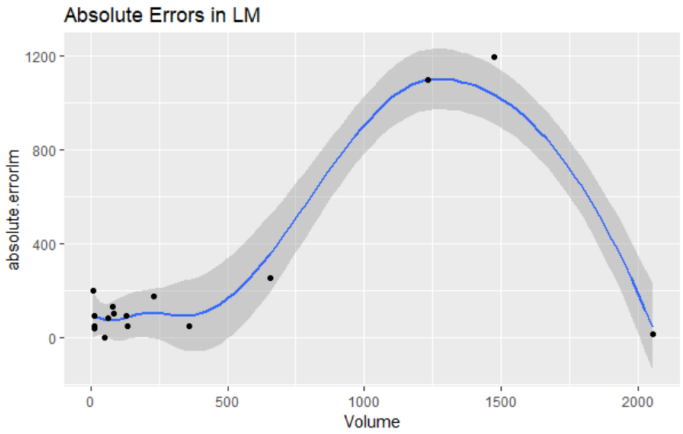
4. 3. 1 Linear Model

Following the general guideline (Statistics by Jim, 2018), we first use the linear model to understand whether we can fit the particular curve on our data. However, since we are now in the multidimensional space, we will not be able to show a graphical representation on our fitted regression lines.

```
Call:
lm(formula = Volume ~ PositiveServiceReview + x4StarReviews +
    ProductDepth + 0, data = training)

Residuals:
   Min     1Q Median     3Q    Max
-400.1 -120.3  -59.9    5.7  874.2

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
PositiveServiceReview   -0.447      0.912   -0.49     0.63
x4StarReviews            9.686      0.745   13.01  < 2e-16 ***
ProductDepth             5.717      0.785    7.28  1.8e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 246 on 52 degrees of freedom
Multiple R-squared:  0.885,     Adjusted R-squared:  0.878
F-statistic:  133 on 3 and 52 DF,  p-value: <2e-16
```

Within our **training set**, we manage to achieve a R^2 of 87%, with a p-value of below 0.05, henceforth making our model significant.

However, when we applied the model on the **test set**, our R^2 dropped to 54%, which is clearly a **sign of overfitting**. This need not be a sign of a wrongful modelling, but furthermore a sign that the dataset is too small, henceforth artificially creating overfitting. The misusage of a linear model has been further highlighted by the graph below (where we plot the errors against volume). This clearly shows us two things. Firstly, it is **extremely inaccurate**, which can especially be seen in the lower volume regions. Secondly, one can clearly see a **quadratic trend** in the volume regions around 1300-1500 *volume*. This trend has **not been captured** by the linear model, which is the main reason why we disregard it.



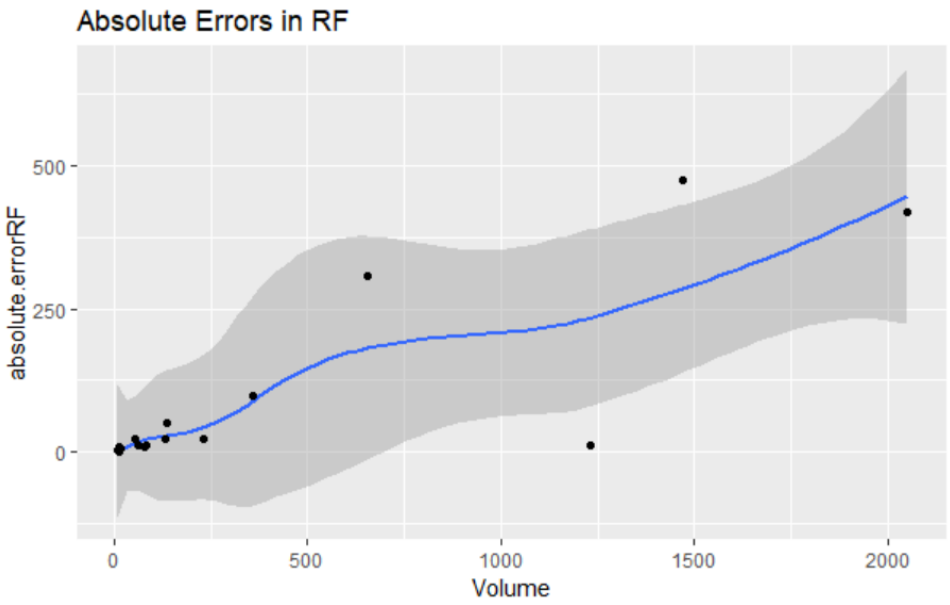
4. 3. 2 Random Forest

After conducting a repeated cross validation, we run our random forest model with the attributes selected above. Thereby, the random forest once again suggests, that the three variables selected are relevant (achieving the best results with mtry=3)

Within our **training set**, we manage to achieve a R^2 of 92%, RMSE of 224 and an absolute error of 127.

When we applied the model on the **test set**, our R^2 slightly increased to 96%, which is clearly a **sign against overfitting and underfitting**. Additionally, our RMSE of 178 remained very similar as well as

our absolute error of 93. This can be clearly seen that Random Forest is a model that works very well with the given dataset.



Looking absolute errors against the volume again shows us that the predicted values are very similar to the real volumes. However, this changes as the model underpredicts the volume in the upper regions. Henceforth, we need to be wary of the estimations in the higher volume regions.

4. 3. 3 GBT

After conducting a repeated cross validation, we run our GBT model with the attributes selected above. Thereby, the GBT determines that the final values used for the model were n.trees = 150, interaction.depth = 3, shrinkage = 0.1 and n.minobsinnode = 10.

Using these parameters within our **training set**, we manage to achieve a R^2 of 81%, RMSE of 308 and an absolute error of 227.

However, we applied the model on the **test set**, our R^2 decreases to 66%, which is clearly a **sign of overfitting**. Additionally, our RMSE increases to 397 while our MAE remains fairly stable at 258. Henceforth, we conclude that this model overfits the training model and underfits thereby the testing model. This makes sense as the Gradient Boosted Tree is consequently learning from the error output. However, this can be changed by the weightings of different trees.

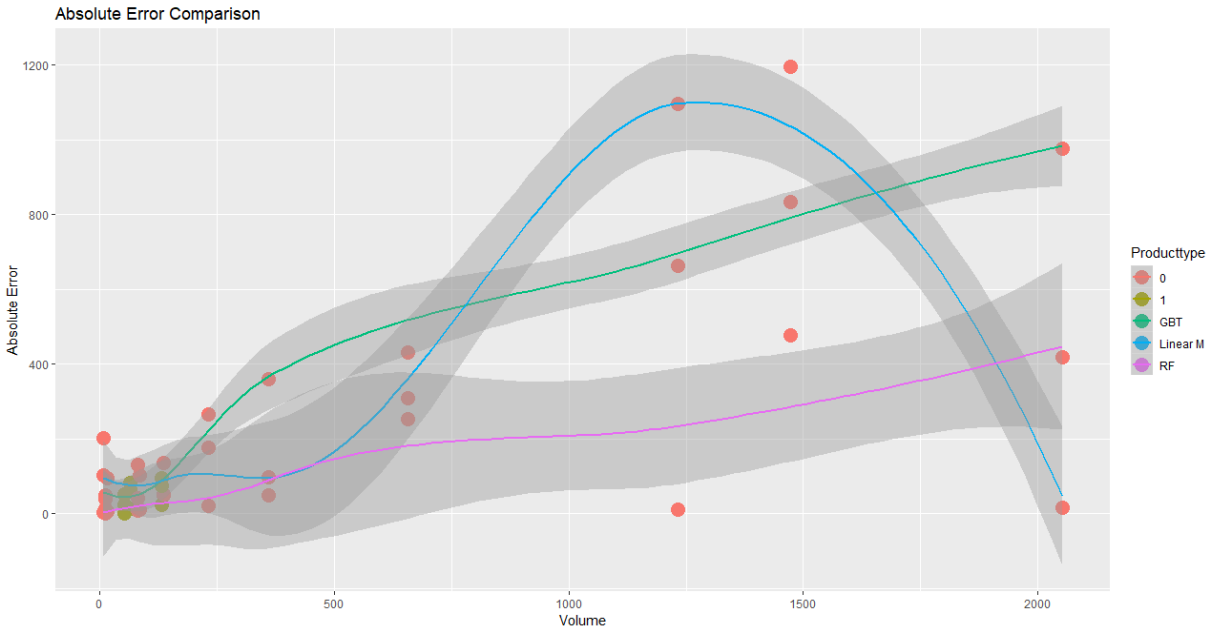
Looking at the errors, we can clearly see that the model predicts volumes as too high, which negatively impacts our predictions.

4.4 Model selection

Looking at the residual plots above, it becomes apparent that the **Random Forest model is the best to predict volume based on *PositiveServiceReview*, *x4StarReview* and *ProductDepth*.** This has been further supported by the Model metrics (see below).

	Metrics.LM	Metrics.GBT	Metrics.RF
RMSE	420.188	397.668	178.614
Rsquared	0.544	0.662	0.966
MAE	226.588	258.071	92.135

This becomes further apparent, when we merge the 3 Residual plots and highlight those product categories, which are essential for our analysis with Blackwell Electronics (colored in brown). **Throughout the whole graph, we can see that the Random Forest model consistently provides us with the lowest errors.**



4. 5 Running the model

After uploading the file with the new, unreleased products we run again through the pre-processing to ensure the accurate handling of the data as such. As we are not conducting classification tasks, it is fairly hard for us to display our confidence in the results (amount of classifications could be compared between the test, training, new data). However, we do believe that the Random Forest is the best model to predict profitability of new products.

ProductType	ProductNum	Price	ProfitMargin	PredictedVolumeRF	Profitability
PC	171	699.0	0.25	358.40	62630.8
PC	172	860.0	0.20	130.77	22492.3
Laptop	173	1199.0	0.10	153.23	18372.3
Laptop	175	1199.0	0.15	45.95	8264.9
Laptop	176	1999.0	0.23	15.94	7330.9
Netbook	178	400.0	0.08	56.60	1811.2
Netbook	180	329.0	0.09	1182.94	35026.8
Netbook	181	439.0	0.11	107.49	5190.7
Netbook	183	330.0	0.09	33.18	985.5
Tablet	186	629.0	0.10	1146.87	72138.1
Tablet	187	199.0	0.20	1666.74	66336.3
Smartphone	193	199.0	0.11	279.82	6125.3
Smartphone	194	49.0	0.12	610.47	3589.6
Smartphone	195	149.0	0.15	77.23	1726.0
Smartphone	196	300.0	0.11	153.47	5064.5
GameConsole	199	250.0	0.09	1134.69	25529.6
Display	201	140.0	0.05	22.35	156.4
Accessories	301	21.0	0.05	34.57	36.3
Accessories	302	8.5	0.10	54.78	46.6
Software	303	71.0	0.20	108.04	1533.9
Printer	304	200.0	0.90	80.99	14576.9
PrinterSupplies	305	21.0	0.30	15.33	96.5
ExtendedWarranty	306	100.0	0.40	7.39	295.6
GameConsole	307	425.0	0.18	1559.66	119313.8

4. 6 Technical Recommendations

Looking backwards on the task, we understood that more data within the different product categories is needed to make more accurate statements and predictions. However, as this is sometimes impossible such as in this example we need to bootstrap the dataset or the run loops to artificially increase the dataset.

For more information, check our code on Github:

Author username: ***mariafarres***

Repository: testingGit

Author username: ***FlorianJUnger***

Repository: Multiple-Regression-in-R