# Task 3: Multiple Regression in R

This informal report contains a brief overview of the procedures and the results obtained in this third task. In this task I began by pre-processing the data and developing an exploratory data analysis to comprehend the data.

The objective of this task is to train multiple regression models on a dataset and apply the said model on a different dataset to generate predictions. Briefly, we are intended to create a model capable of forecasting the weekly sales volume in a small selection of products. As can be seen below, this nominal variable is slightly unbalanced:

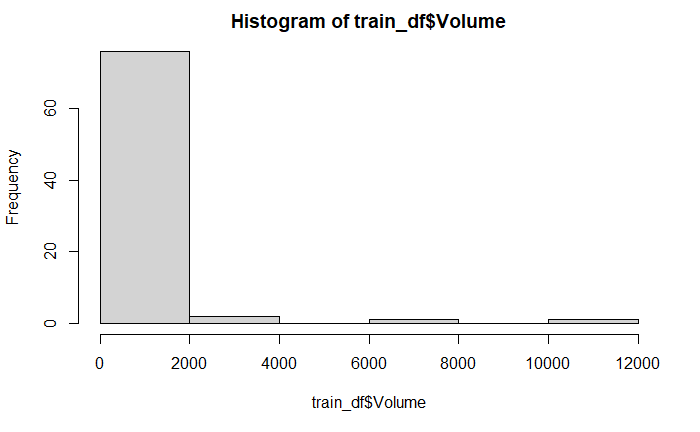


Figure 1 - Target count distribution

Both datasets (train and test) contain the same information, and the behavior of each column is similar in both the datasets. The only exception is that the target variable, Volume, has been filled by default with ‘0’ in the testing dataset. The remaining variables are valid.

In order to forecast the target, a dataset with 17 other variables has been provided. Briefly, these variables contain the information of the product type, price and profit margins, customer reviews and dimensions.

In this exercise I was asked to develop 4 models: Linear Regression; Support Vector Machine; Random Forest & Gradient Boosting. The data processing and procedure for training each of these models was identical. Before training each model I simply divided the pre-processed dataset in train and test sets (the latter contained 25% of the data). Now, I will briefly discuss each model individually below.

## Model 1: Linear Regression:

In this first model, I didn’t use any cross validation as this was just an exploratory model. The performance of this model could not capture the complexity of the target. The main metrics can be seen in the following table:

|  |  |  |
| --- | --- | --- |
| **RMSE** | **R2** | **MAE** |
| 688 | 0.33 | 447 |

It’s worth adding that I chose the x5StarReviews as the correlation exploration showed that review based variables had a high correlation with the target and therefore could be more informative. Linear models, generalized linear models, and nonlinear models are examples of parametric regression models because we know the function that describes the relationship between the response and explanatory variables. This dataset is small and the data does not seem to follow the distributional requirements of parametric methods, (as can be seen in the R markdown script) the data is highly skewed and many features are categorical, not numerical. Additionally, the dependent variable does not seem to respect a linear relation with any of the independent variables.

## Model 2: Support Vector Machine:

In this second model, I utilized 10 fold cross-validation with an automatically tuning grid of length size 10. In order to choose the optimization metric I took into consideration that both RMSE and R2 quantify how well a regression model fits a dataset and that the RMSE tells us how well a regression model can predict the value of the response variable in absolute terms while R2 tells us how well a model can predict the value of the response variable in percentage terms. Therefore, I used the RMSE metric as the point of this regression is to forecast the absolute Volume as accurately as possible (also the results were better utilizing this metric).

After training I ran the model on the test dataset and obtained the following metric results for the model:

|  |  |  |
| --- | --- | --- |
| **RMSE** | **R2** | **MAE** |
| 156 | 0.93 | 127 |

The value of the Root Mean Squared Error is 198, taking into account that the average Volume is 700, this error is not terrible, but it is not great as well. The same goes for a Mean Absolute Error of 152. The R squared is not far from ‘1’ which means that this models is an alright fit for this task.

For this model the overall relative importance of the used features is:

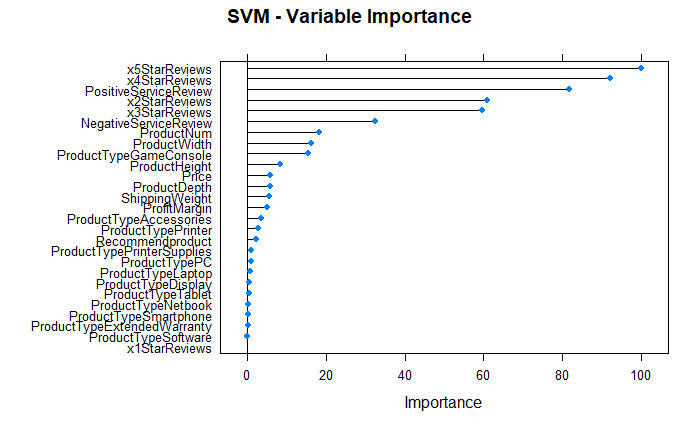


Figure - SVM variable importance plot

## As was predicted in the correlation section, the Review variables are the most relevant towards forecasting this target.

## Model 3: Random Forest:

Once again, I utilized 10 fold cross-validation with an automatically tuning grid of length size 15. In order to choose the optimization metric I took into consideration that both RMSE and R2 quantify how well a regression model fits a dataset and that the RMSE tells us how well a regression model can predict the value of the response variable in absolute terms while R2 tells us how well a model can predict the value of the response variable in percentage terms. Therefore, I used the RMSE metric as the point of this regression is to forecast the absolute Volume as accurately as possible (also the results were better utilizing this metric).

After training I ran the model on the test dataset and obtained the following metric results for the model:

|  |  |  |
| --- | --- | --- |
| **RMSE** | **R2** | **MAE** |
| 53.35 | 0.99 | 30.24 |

The value of the Root Mean Squared Error is 198, taking into account that the average Volume is 700, this error is not terrible, but it is not great as well. The same goes for a Mean Absolute Error of 152. The R squared is not far from ‘1’ which means that this models is an alright fit for this task.

For this model the overall relative importance of the used features is:

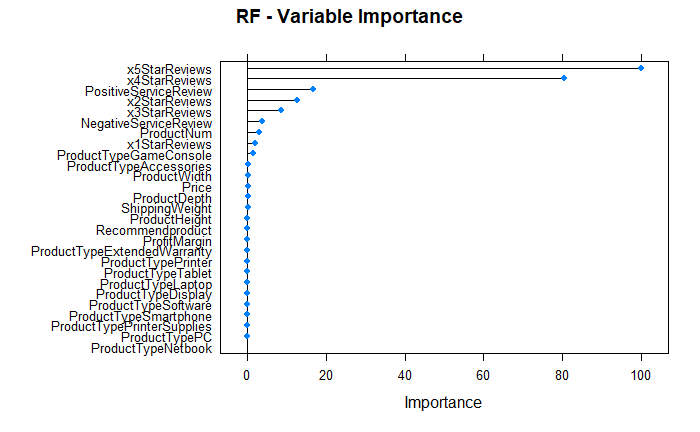


Figure - RF variable importance plot

## As was predicted in the correlation section, the Review variables are the most relevant towards forecasting this target. In this case, the relative importance of the third most relevant onwards has dropped significantly. This model is capable of providing a better result with less information.

## Model 4: Gradient Boosting:

In this last model, I utilized 10 fold cross-validation with an automatically tuning grid of length size 26. Once again I used the RMSE as the optimization metric.

After training I ran the model on the test dataset and obtained the following metric results for the model:

|  |  |  |
| --- | --- | --- |
| **RMSE** | **R2** | **MAE** |
| 360.26 | 0.89 | 324.21 |

The value of the Root Mean Squared Error is 198, taking into account that the average Volume is 700, this error is not terrible, but it is not great as well. The same goes for a Mean Absolute Error of 152. The R squared is not far from ‘1’ which means that this models is an alright fit for this task.

For this model the overall relative importance of the used features is:

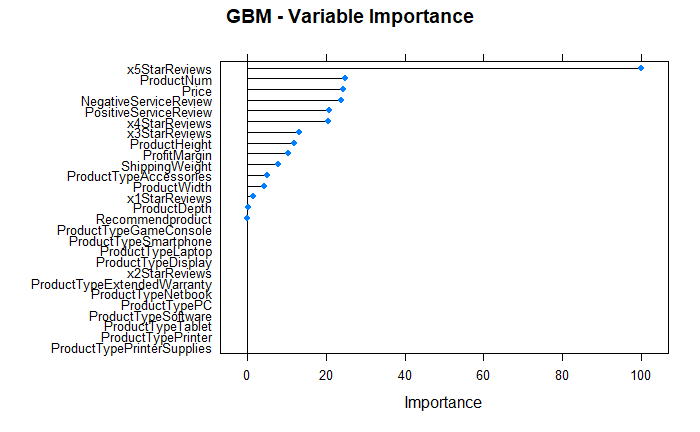


Figure - GB variable importance plot

## As was predicted in the correlation section, the Review variables are the most relevant towards forecasting this target.

## Model Comparison

First of all, I must add that the models above utilized many irrelevant features. If this was a real situation another iteration of training should be added to this process in order to remove the useless features from the independent variables. This could provide a slight improvement in the performance of each model.

Having created the 4 models, it is now time to pick the one with better performance in order to create the predictions in the previously untouched dataset. Below are the plots of each performance metric individually:

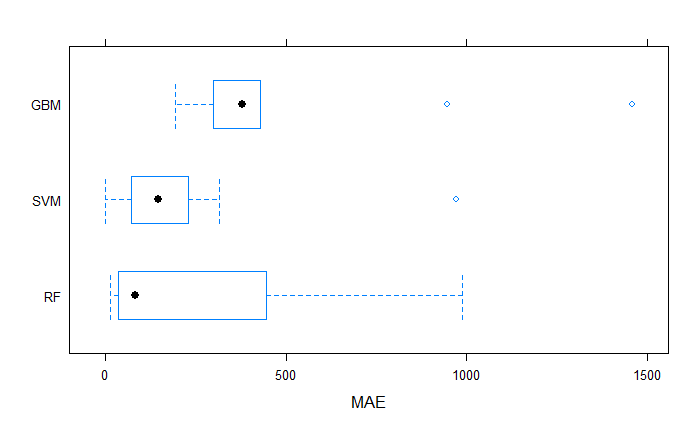
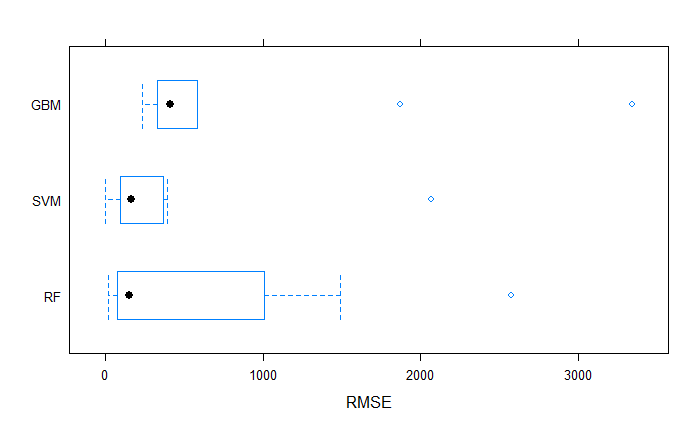


Figure - Mean Absolute Error for the 3 best models Figure 6 – Root mean squared error for the 3 best models

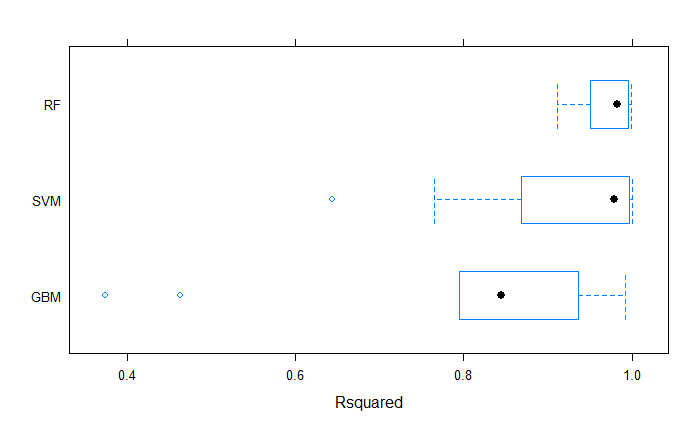


Figure 7 – R2 for the 3 best models

Looking firstly at the MAE plot, the RF has the smallest median out of the 3, but the third quartile and maximum values is clearly larger than the remaining. The RMSE plot tells an identical story as the former. Finally, the R2 is the only graph where the RF is beyond any doubt the model: the median is higher than the other two and the minimum and first quartile likewise.

In conclusion, I believe that the RF is arguably the best performing model and therefore I will recreate this model with less independent variables and perform the final predictions using this model.

## Random Forest retraining

In this second version of Random Forest, I started with the 5 top features obtained previously, when training the RF and then compare the obtained results to what was generated previously. I will then either increase or decrease the number of features and verify how the quality of the model varies. Note, I only used the top 5 because as observed in the Variable Importance plot, the importance of the remaining ones is almost null. Once again I used the RMSE as the optimization metric.

After training I ran the model on the test dataset and obtained the following metric results for the model:

|  |  |  |
| --- | --- | --- |
| **RMSE** | **R2** | **MAE** |
| 34.65 | 0.99 | 23.1 |

For this model the overall relative importance of the used features is:

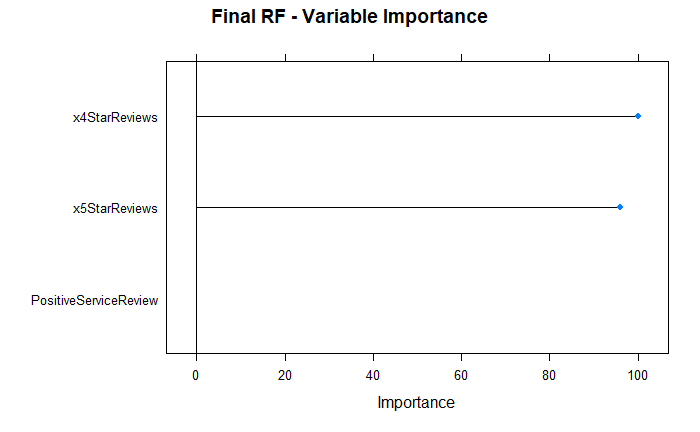


Figure 8 - GB variable importance plot

I verified that as I reduce the number of features decreases, the quality of the model increases. It seems, this model is capturing the noise and overfitting likewise to the multilinear regression model. Therefore I will conclude that in order to develop a usable model more data would be required. This dataset seems to be complex and too small to take any conclusion.

Below I have represented an histogram of the Volume in the original train dataset and of the final predictions generated with the RF from this second iteration in order to verify if the initial distribution is respected:

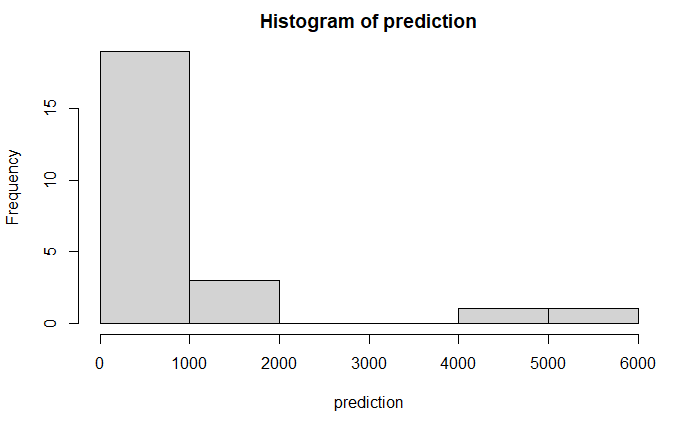
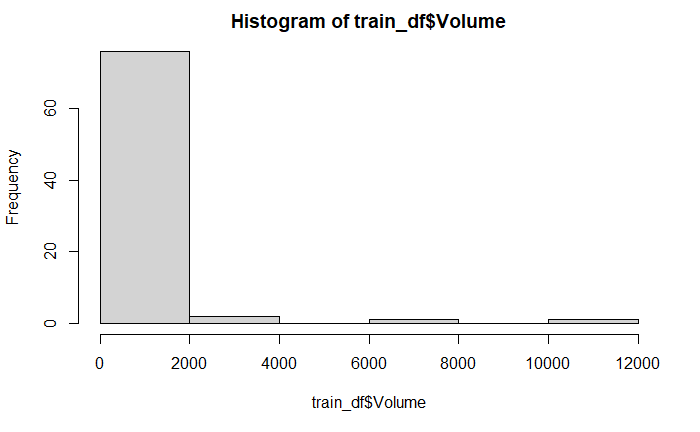


Figure 9 – Histogram of the Volume in the training dataset (left) and the Volume of the predictions (right)

It can be seen that this proportion is pretty much the same for both datasets, meaning that the test and train datasets contain clients with similar behaviors.

# Conclusion

I was asked to predict the sales of four different product types: PC, Laptops, Netbooks and Smartphones. This dataset is extremely short and complex therefore all conclusions should be taken lightly, a more in depth dataset is required to confidently conclude anything.

Anyhow, I have created 4 machine learning models to forecast the volume of sales in a set of different product types. After testing several models with reasonable performance I picked the best out of the 4 and came to the following conclusions regarding the Volume of sales:

|  |  |  |  |
| --- | --- | --- | --- |
| **Type of Product** | **Product ID** | **Predicted Future Sale Volume** | **Total Volume sale p/category** |
| PC | 171 | 437.505333333333 | 620.888 |
| 172 | 183.383066666667 |
| Laptops | 173 | 265.3524 | 303.513 |
| 175 | 31.1997333333333 |
| 176 | 6.96119999999999 |
| Netbooks | 178 | 83.2684 | 1409.842 |
| 180 | 1164.21146666667 |
| 181 | 149.070133333333 |
| 183 | 13.2924 |
| Smartphones | 193 | 444.7368 | 1650.514 |
| 194 | 597.300533333334 |
| 195 | 256.0768 |
| 196 | 352.399866666667 |

Table - Predicted Volumes for the new products

Additionally, as has been suggested, the variables related to reviews are extremely important towards the future volume of sales. As can be seen below, the most high reviews a product has, the higher the sales volume was (figure 9) and will be in the future (figure 10).

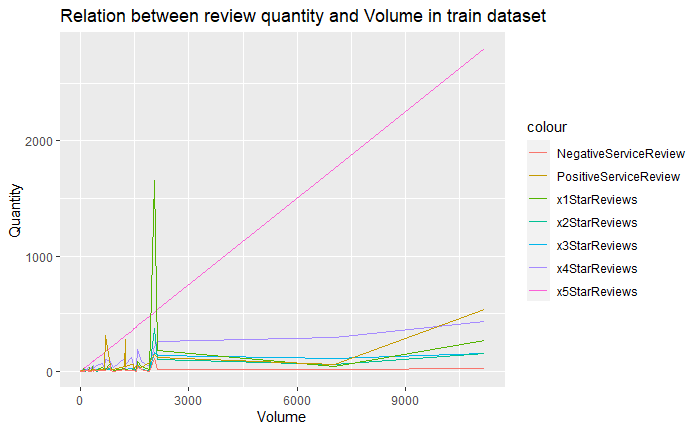


Figure 10

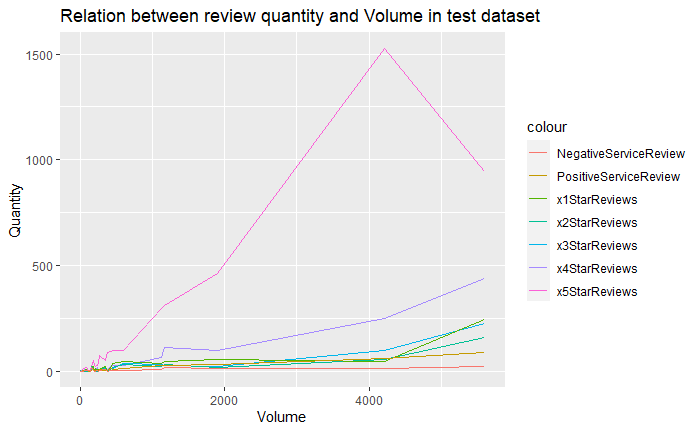


Figure 11