Sales prediction report

Multiple Regression analysis

Introduction

Blackwell Electronics is expanding its product portfolio, and the sales teams is struggling to select with product to launch. This report covers the data analytics activities of multiple regression analysis exploring the data to predict the profitability of new products. Its objective is to help Blackwell to choose the best new products based in the best profits.

Results

General Insights

In order to predict the profitability of the new products, we used the existing data from 80 products that we currently have on sales. The database consists of the product type, its price, as well as its customer review (1 to 5 stars review, service review, customer recommendation, and best seller rank). As well as the product physical characteristics (weight, width, depth, height), and lastly the profit of each product and the volume of sales.

Before doing the predictions, we did some preprocessing of the dataset. We dummified the product types in order to transform this categorical feature to binary. After that we took off the best sellers rank because it contains missing data and we weren't able to replace it. We also removed the profit margin since this isn't an important feature for the model. We also removed 2 outliers found in the volume feature, since they were too high, and could affect the prediction.

We did a correlation matrix to check correlations among the features. This showed that 5 stars had a perfect correlation with volume and 4 and 2 stars had a high correlation with 3 and 1 stars respectively. So we took these attributes off to address the collinearity. The figure 1 shows us the final correlation matrix that we found. This matrix shows us that negatives reviews has correlation with the product type Software while positives reviews has correlation with Extended Warranty. We also can see that 2 stars reviews has correlation with Software and 4 stars reviews has correlation with Game Console. This makes sense since its natural that lower stars should be at the same product as the negative review.

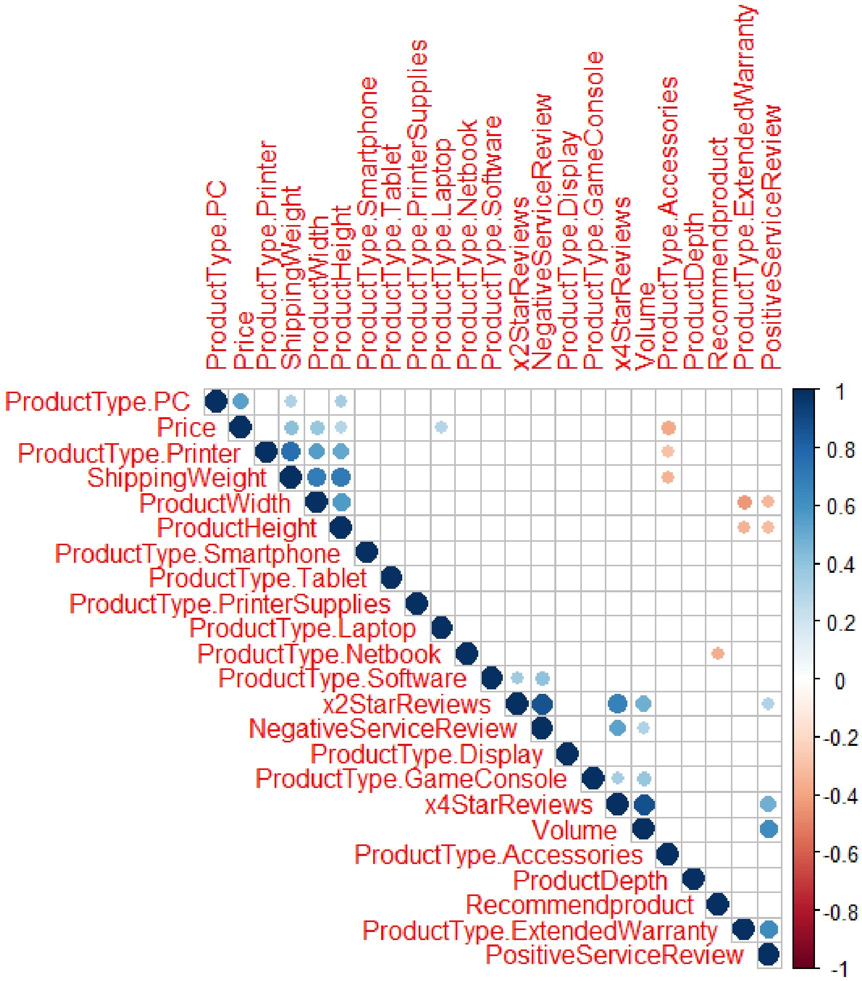

Figure 1: Correlation matrix of all the features of the dataset. Notice that 1 is a perfect positive while -1 is a perfect negative correlation. Only significant ($p < 0.01$) correlations are shown.

Predicting profitability

Once all the preprocessing of the data was done, we explored several models in order to see which one was the best to predict the volume of sales. We tested linear model, k-NN, SVM, and random forest (rf) models. We decided to select the random forest model, using a value of *mtry* equal to 19, as this model showed one off the best results regarding the fit of the model (R^2=0.916), and one of the lowest variance on the errors (175.481) (full results in the Apendix).

Based on these results, we were able to predict the volume of the new products, and using the profit margin and the price of the products, the profit was calculated as shown on Table 1 and Figure 2. The most profitable product type is PC, followed by products of NetBook and Laptop categories. Keep in mind that some products are predicted to have a very low volume, so for these products, we advise against launching these products, since most likely will not sell much.

Table 1: Ranking of the potential new products, ordered from highest to lowest profitability. The top products for each category are highlighted.

ProductType	*ProductNum*	*Price*	*ProfitMargin*	*Volume.prediction*	*Profit*
PC	171	699	0.25	466.98	81604.1
PC	172	860	0.2	200.93	34560.2
Laptop	173	1199	0.1	202.97	24336.0
Laptop	175	1199	0.15	34.25	6160.7
Laptop	176	1999	0.23	17.31	7957.2
Netbook	178	399.99	0.08	60.36	1931.5
Netbook	180	329	0.09	1251.14	37046.3
Netbook	181	439	0.11	119.87	5788.6
Netbook	183	330	0.09	30.95	919.2
Smartphone	193	199	0.11	444.86	9738.1
Smartphone	194	49	0.12	581.90	3421.6
Smartphone	195	149	0.15	76.96	1720.0
Smartphone	196	300	0.11	102.33	3376.7

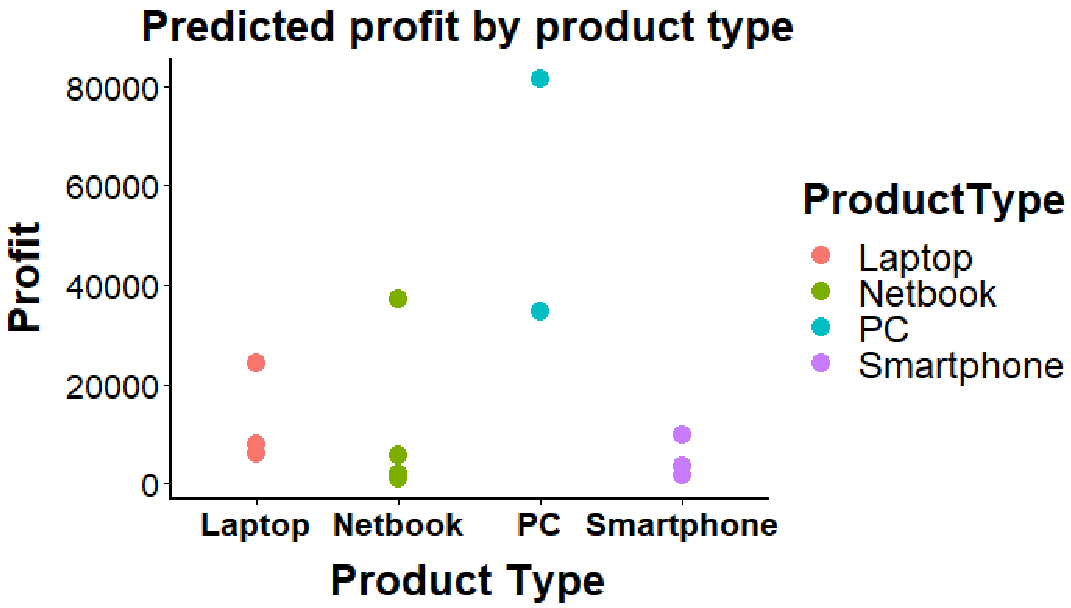

Figure 2: Predicted profit of different product types.

Relationship of Volume with Customer and Service reviews

We explored the relationship of the total sales volumes and the reviews given by our customers. The data was normalized before the analysis. We confirmed the positive correlation observed in the correlation matrix in which 5 star reviews are highly correlated with the volume. For The other types of customer reviews (4 to 1 stars) the correlation was very low (Figure 3). When taking into account the 4 product types of interest, we observe different patterns (Figure 4). For Laptops, the volume is not affected by the customer reviews, and for the Netbook and Smartphone is very important to have 5 star reviews, since is highly correlated with volume.

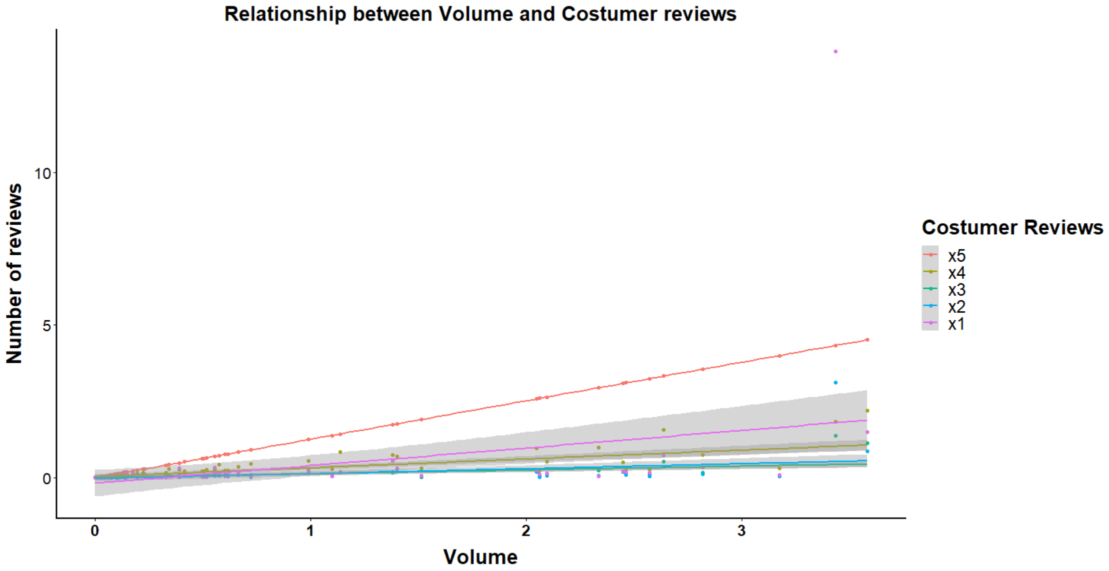

Figure 3: Relationship between Volume and the number of customer reviews.

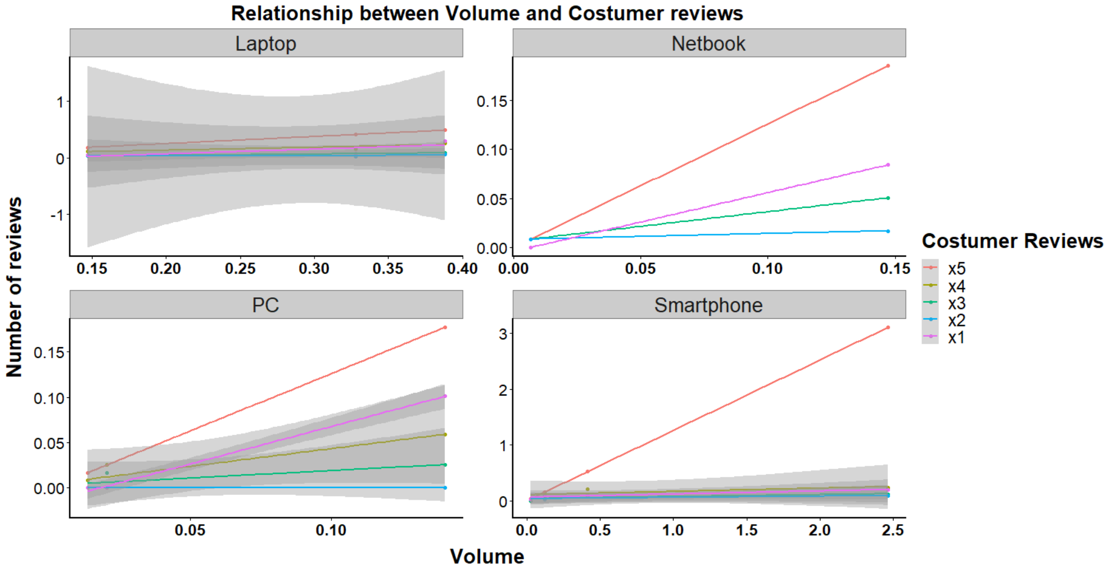

Figure 4: Relationship between Volume and the number of customer reviews, based on 4 product types: Laptop, Netbook, PC, and Smartphone.

The relationship between volume and the service review was similar. We observe a positive correlation with the positive service reviews, but no correlation on the negative service reviews (Figure 5). When taking into account the 4 product types, we observe again that for Laptops service reviews are not affecting the sale volume. Nevertheless, there is a positive relation on the other 3 product types between volume and positive service review. Please take into consideration that for the product types chosen, the data is very limited, with only 4 PC's, 4 Smartphones, 3 Laptops, and 2 Netbooks.

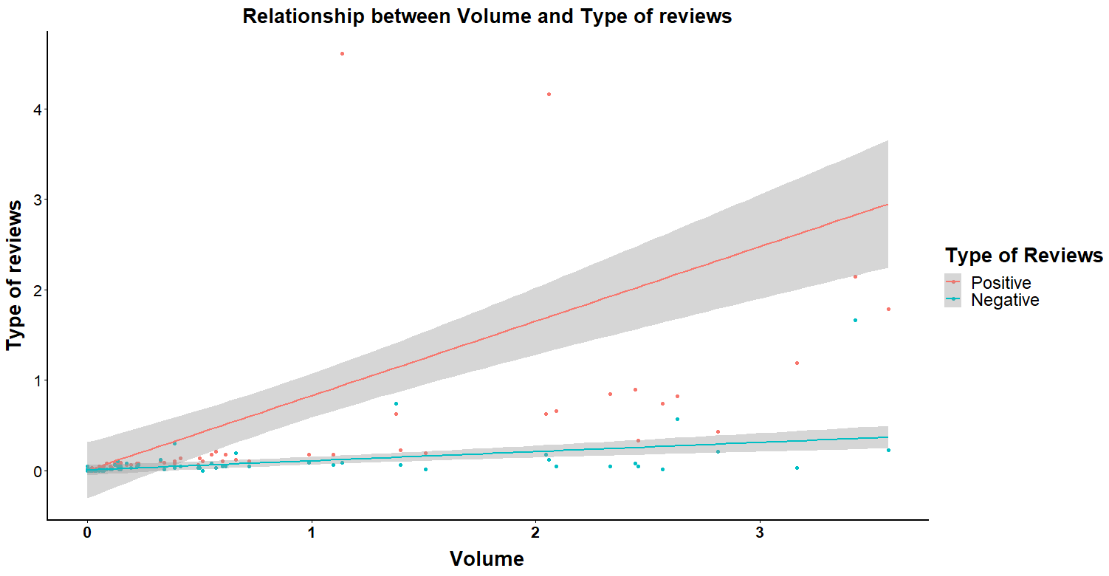

Figure 5: Relationship between Volume and the Service reviews (positive/negative).

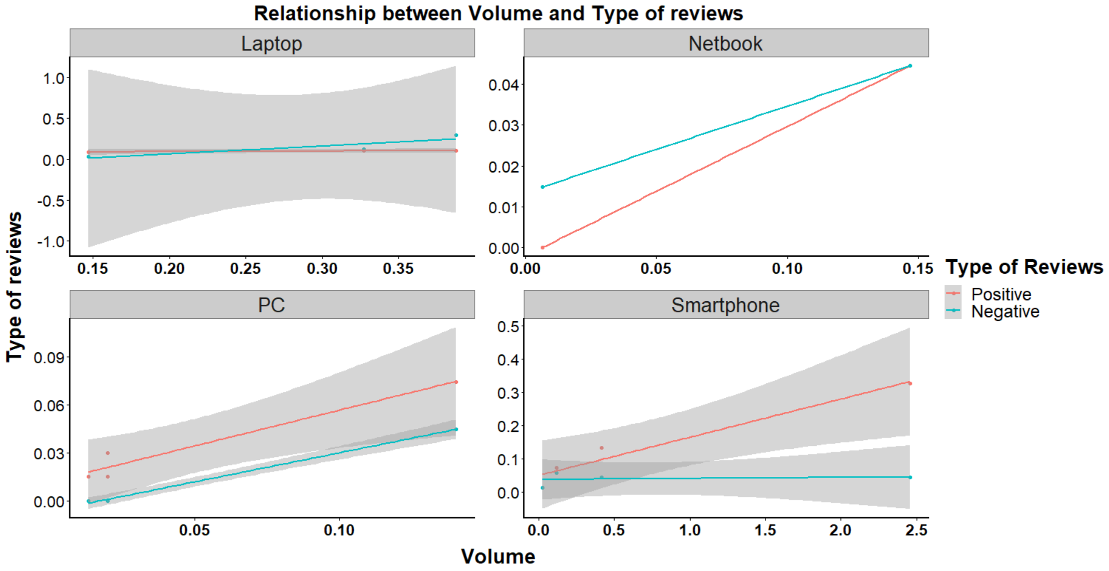

Figure 6: Relationship between Volume and Service reviews (positive/negative), based on 4 product types: Laptop, Netbook, PC, and Smartphone.

Conclusions and recommendations

Based on our existing database, we investigated the data and performed a multiple regression analysis to predict the likely volume of 4 product types that the sales team were considering to launch. Using the estimated volume, we calculated the profit, and suggest the most profitable products to consider. Therefore, we recommend the sales team to focus on the products with the following numbers: 171, 173, 178 and 193. This recommendation is because these products should have the best monthly profits to the company considering different products types.

Please keep in mind that although the model is well supported, the data used is very small and not diverse enough. This is important since better data allows a better model to different types of analysis, so if we had a more diverse data we could give a more general model. Therefore, we suggest repeating the model using a bigger and more diverse data to improve the model.

Appendix

Output of the model used for the prediction of the Volume.

```
rfFit <- train(Volume ~ .,
            data = TrainSet,
            method = "rf",
            trControl = ctrl,
            preProcess = c("center", "scale"),
            tuneLength = 20)
```

```
Random Forest

60 samples
22 predictors

Pre-processing: centered (22), scaled (22)
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 53, 55, 53, 54, 53, 54, ...
Resampling results across tuning parameters:

 mtry  RMSE      Rsquared   MAE
  2   308.3306  0.8350608  229.5990
  3   275.3392  0.8468316  197.0894
  4   252.9536  0.8475172  176.4806
  5   235.8546  0.8613665  161.4304
  6   225.5338  0.8680058  153.6551
  7   222.2856  0.8689901  150.1415
  8   208.5100  0.8743791  140.7181
  9   202.5014  0.8825676  131.8152
 10    194.7322  0.8863698  125.9307
 11    189.5591  0.8944122  121.4033
 12    185.9986  0.8937302  118.4062
 13    179.7797  0.9040080  110.9021
 14    183.0698  0.8971953  111.5983
 15    181.2935  0.9038915  111.3765
 16    180.5331  0.9081523  108.2717
 17    178.3136  0.9100543  106.2013
 18    176.7471  0.9111652  105.7669
 19    175.4810  0.9155711  103.7359
 20    179.9703  0.9138694  106.6394
 22    179.0026  0.9162149  105.0747

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 19.
```