**Course Three Task3 Report**

**by**

**Shifeng (Steve) Li**

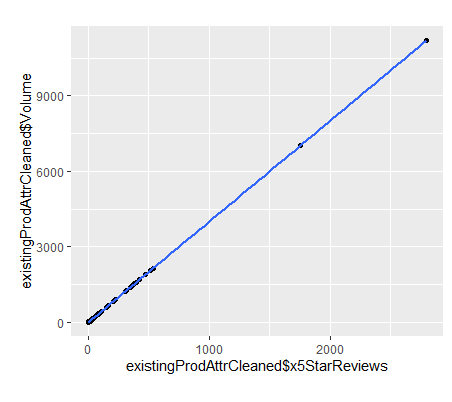
**02/26/2022**

In this project, the Blackwell Electronics historical sales data and new sales data sets are analyzed using R caret data analysis techniques to predict sales of difference product types, and access the impact the reviews have on sales.

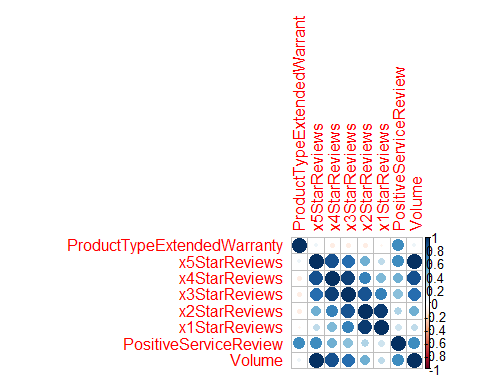
In the study, multiple regression algorithms are used to analyze through the historical data set , then based on the quantified prediction accuracy values comparison, the best algorithm is selected to do the prediction for the new data set. In the following, will go through the analysis steps.

1. **Initial Data Profiling**

As the standard data analysis procedure, went through the initial data set analysis to understand the data profile and structure, validated the data and visualized the data correlations.

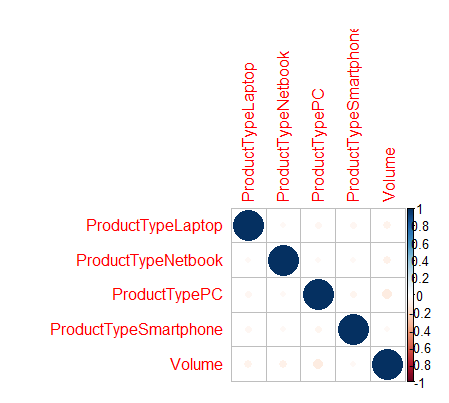


The linear correlation scattered plot shows the 5-star review has the most significant correlation to the sales volume.



The correlation heat map of reviews vs sales volume also shows the 5-star review has the most significant correlation to the sales volume.

The product types of PC, Netbook, Laptop, and Smartphone, however, do not have significant correlation to the sales volume. Smartphone has minor correlation to the sales volume.



1. **Build and train the different models**

In this part of process, using the R caret tools to build and train the different regression models using the three algorithms. First of all, by applying the generic regression prediction to the historical data set train and test, find that the PC, Laptop, Netbook, and Smartphone product types do not have good prediction accuracy level for forecasting the sales volume. However, the customer reviews do have significant impact to the sales volume. The following are the prediction accuracy level values of the 5star, 4star, and 3star reviews, came out from the generic regression model lm.

5-starCustomerReviews correlation with sales volume accuracy level

RMSE Rsquared MAE

0 1 0

4-starCustomerReviews correlation with sales volume accuracy level

RMSE Rsquared MAE

1346.9228104 0.6419054 806.1363444

3-starCustomerReviews correlation with sales volume accuracy level

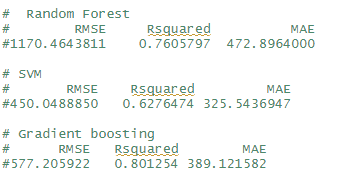
RMSE Rsquared MAE

2477.7053427 0.5602049 1128.691163

Then, built the three specific regression models using the three different algorithms: Support Vector Machine, Gradient Boosting, and Random Forest. For each algorithm, went through the process to build the model, complete the machine learning with training data, and do prediction on the test data, to obtain the quantified prediction accuracy statistical data set for each model. These quantified prediction accuracy data then be used to compare and select the best fit model for the prediction on the new data set.

1. **Model selection**

After built and trained the three models, now use the R caret resample summary to list the prediction accuracy values of the three different models from doing the predictions on the test data set.

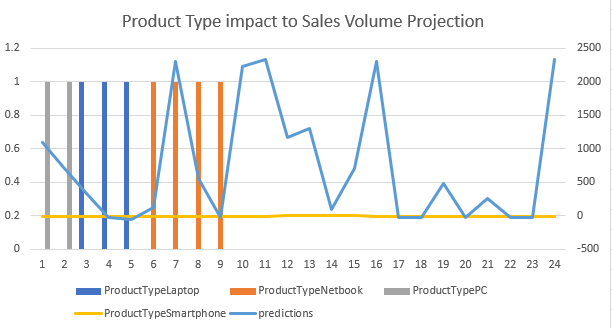


As we can see, the Gradient Boosting regression model has the better fit for the prediction.

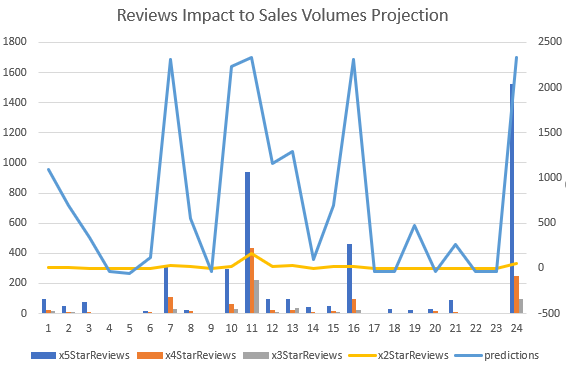
1. **Projection**

With the selected Gradient Boosting regression model, did the prediction on the new sales data set. And here are the top level of projections of product types impact to sales volume and customer reviews impact to the sales volume, as the following visualization plots.

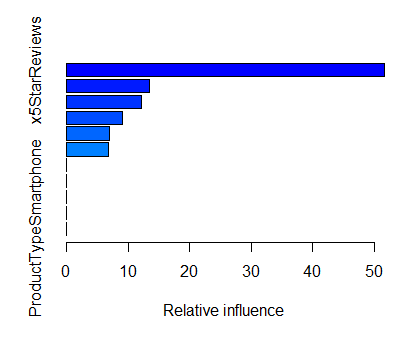
1. The product types have very limited impact to the sales volume projection.

****

1. Among the customer reviews, the 5-star review has the most significant impact to the sales volume projection.

****

1. The Gradient Boosting regression model’s Variable Importance Plot also shows that the 5-star customer review has the most significant impact to the sales volume, while the Smartphone product type only has minor impact.



So, from the analysis study, find that the customer reviews have good impact to the sales volume projection, while the four product types have very limited impact to the sales volume projection.

And, among the three tested algorithms, the Gradient Boosting regression algorithm based model has the best prediction accuracy level.

**Uploaded to: https://github.com/UTOct21DaPtSteve/Steve\_Course3**