|  |  |  |
| --- | --- | --- |
|  |  |  |
| Ramses Meza  May 2021  <https://github.com/lordtable/UTDSGC_C3T3> |  | Data Analytics Course 3.3: Predict Sale Volumes of certain product types |

Business Acumen

Blackwell Electronics needs to accurately predict the sale volumes of specific product types at its stores. The sales team has gathered some historic sales data that may be used to build predictive models based on several product-related and service-related attributes. Upon successfully building such model using Machine Learning techniques (using RStudio), then it must be applied to a new product data set in order to estimate Sale volumes for four product types of interest: PC, laptops, netbooks and smartphones. The successful calibration, testing and deployment of a predictive model can greatly impact all of Blackwell’s value chain.

**Data Management, Cleaning & Pre-processing**

The datasets were provided on a .csv file format and locally stored. Two files were provided: one containing the historical sales data (called *existingproductattributes2017*) and other file containing new data for which volumes need to be predicted (called *newproductattributes2017*). Both files are structured in a table format: each row represents an undisclosed transaction. Each column represents an attribute or feature, such as:

* **ProductType**: [categorical]
* **ProductNum**: Sequential number [numeric]
* **Price**: Transaction amount [numeric]
* **X5StarReviews**: # of customers top reviews of the product [numeric]. Decreasing star review attributes also exist with consistent naming.
* **PositiveServiceReview**: # of customers positive reviews of the sale experience [numeric]
* **NegativeServiceReview**: # of customers negative reviews of the sale experience [numeric]
* **Recommendproduct**: past customer qualifier for recommending the product [numeric]
* **BestSellersRank**: sales ranking of the product relative to a much wider (?) database [numeric]
* **ShippingWeight**: [numeric]
* **ProductDepth, ProductWidth** and **ProductHeight**: [numeric]
* **ProfitMargin**: [numeric]
* **Volume**: Sales volume. This is the target feature [numeric]

Each survey file was managed as a RStudio Dataframe. The only data cleaning/pre-processing needed consisted on the data type transformations to factor of the ***ProductType*** feature. Early screening of the data using the *summarytools* library indicated that no missing values (except ***BestSellersRank***), no duplicated instances were detected. The *existingproductattributes2017* dataset to be used for model training and testing consists of 80 observations/instances. Of these, 70% were retained as instances used for training all the predictive models to be later described, while 30% of the instances were kept for models’ testing and evaluation. The *newproductattributes2017* dataset has the same overall data quality description as *existingproductattributes2017*, but having an empty ***Volume*** feature. It contains 24 observations/instances. The ***Volume*** feature values of this file will be replaced by synthetic or model-generated values generated by a machine learning regression model, subject to the success on designing one based on the *existingproductattributes2017* dataset.

**Exploratory Data Analysis (EDA)**

EDA was performed on the *existingproductattributes2017* dataset using combinations of R libraries, such as *summarytools*, *explore* and *dlookr*. The first goal was to screen for data errors. Secondly, to get familiarity with each feature characteristics, such as ranges and normality checks. Then, attention was focused on features inter-dependence as depicted on Figure 1, aiming for feature selection and/or dimensionality reduction. It was found that top product and service reviews kind of features are highly mutually-correlated. Same applies to poorer review features. ***X5StarReviews*** is perfectly correlated with ***Volume***, which may lead to over-fitting if used. I have also decided to discard features related to the physical characteristics of the product (i.e. weight and height), as well as ***ProfitMargin***, since I consider them to have no causality on the target feature. This allowed for a significant feature reduction, as appreciated on the selected features shown on Figure 2.

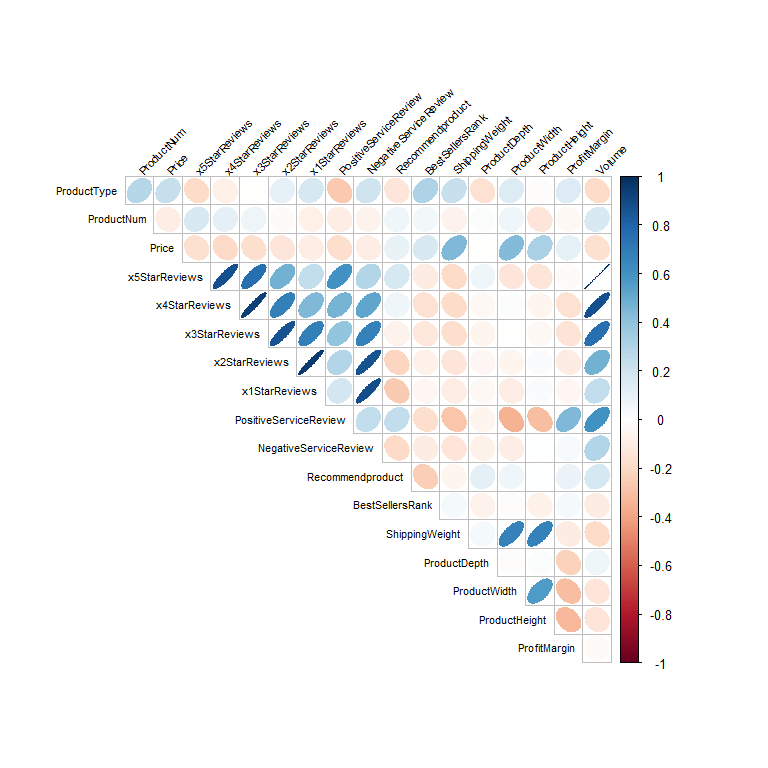


Figure 1: All Features correlation matrix from the *explore* library

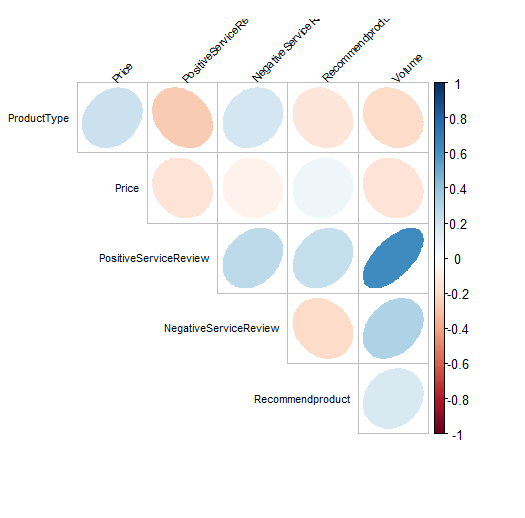


Figure 2: Selected (likely Volume-causal) Features correlation matrix from the *explore* library

**Models Tuning, Assessment & Final Model Selection**

Four (4) regression modelling algorithms were chosen to be built using the same selected features to predict the ***Volume*** variable: Linear Regression (LM), Supported-Vector Machines (SVM), Random Forest (RF) and Gradient Boosting (GB). I took advantage of the fact that most R models have a built-in capability for tuning parameters, that allows parameters selection based on a systematic testing and performance based on the root-mean squared error (RMSE) metric for each parameter set tested. The tuning was executed using the training data as follows:

* **Linear Regression (LR)**: The simplest regression algorithm, its parametrization has essentially null effects on the results. The adjusted R2=0.1764 and RMSE=1,330.
* **Supported-Vector Machines (SVM)**: Using a 10-fold cross-validation and an automatic grid parameter testing, which tested for the *epsilon* and *cost* best permutations aiming for the lowest RMSE for each combination. The best performing (lowest RMSE) tuned SVM model resulted from an *epsilon*=0.03 and *cost*=4. The corresponding R2= 0.8643 and RMSE=938.
* **Random Forest (RF)**: Using a 10-fold cross-validation and an automatic parameter testing, which tested for *mtry*. The best performing (lowest RMSE) tuned RF model resulted from an *mtry*=1. The corresponding R2= 0.556 and RMSE=807.
* **Gradient Boosting (GB)**: Using a 10-fold cross-validation and an automatic grid parameter testing, which tested for the *number of trees* and *interaction depth* best permutations aiming for the lowest RMSE for each combination. The best performing (lowest RMSE) tuned GB model resulted from a *ntree*=50 and *interaction depth*=2. The corresponding R2= 0.6285 and RMSE=2,165

I validated each tuned model by applying them to the test dataset, and gathered each model prediction metrics as shown on Table 1 for model performance comparison. The criteria to select the best performing model is based on such model having: 1.- the highest R2, 2.- the lowest RMSE*training*, and 3.- RMSE*test* that is lower than that for training. The latter criterion is needed in order to avoid picking an over-fitting model. At a first glance, SVM would be the logical choice. However, it was noticed that it yields negative predicted values, making this model unfit for final predictions. The same holds true for LM. GB is not suitable due to its very large RMSE, **so the best striking balance of performance metrics while still yielding reasonable predicted values is accomplished by the RF model, which I decided to use as the final predictive model**. Figure 3 is a scatter-plot of predicted vs actual ***Volume*** values based on the selected RF model. A perfect prediction would mean that all points would lie along the blue line. It can be seen that the final model makes predictions that are reasonably close to the actual values, although larger absolute departures are seen for larger volumes.

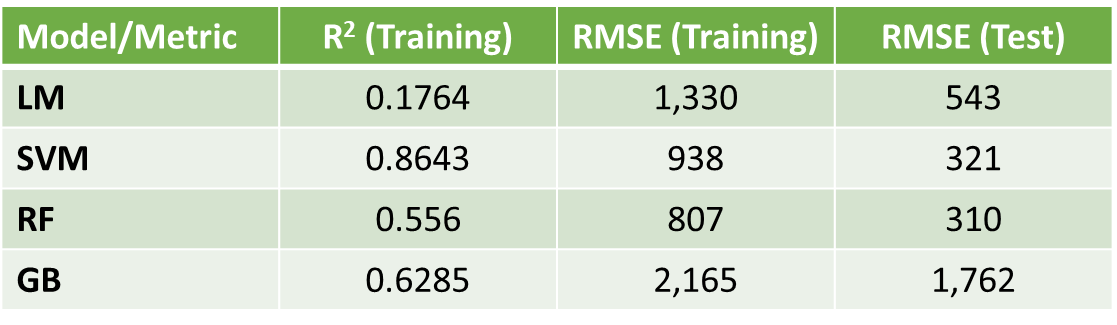


Table 1: Tuned-models metrics performance based on both training/calibration and validation/testing data

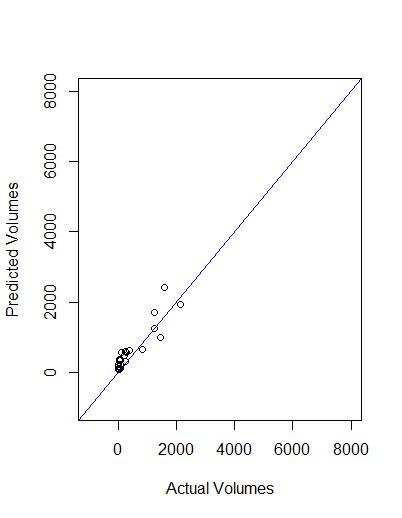


Figure 3: Predicted vs Actual Volume values, based on the tuned Random Forest regression model

For this model, the most important predictors or features are ***PositiveServiceReview*** and ***NegativeServiceReview***, as shown on Figure 4, somewhat in agreement with the insights drawn during the EDA. Features ***Price*** and ***ProductType*** play a secondary role while ***RecommendProduct***, play an insignificant role as predictor.

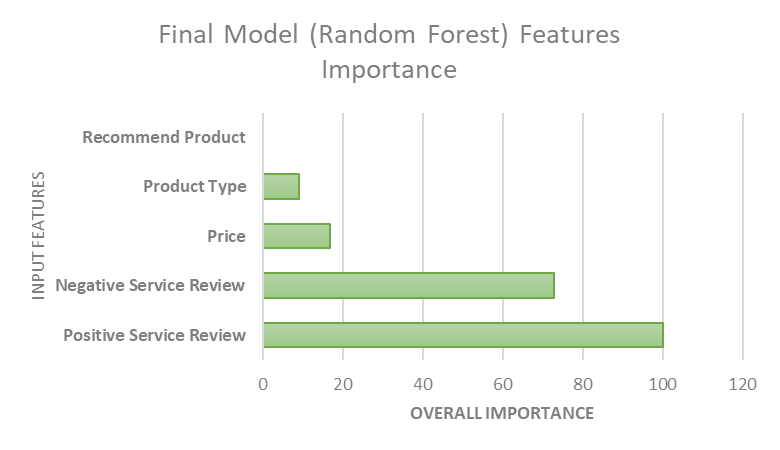


Figure 4: Variable Importance based on the final, tuned RF model.

**Prediction of sales volumes of selected products**

The final RF regression model was applied to a dataframe copy of the *newproductattributes2017* dataset (which already underwent the same editing and feature selection as the *existingproductattributes2017* dataset), effectively replacing the empty ***Volume*** nil values by predicted values generated by the calibrated and tuned RF model. Figure 5 shows the predicted volumes of the four (4) product types Blackwell is focusing on: Smartphones, netbooks, laptops and PCs. These expected sales volumes would impact the sales strategy, and imply adjustments on logistics (i.e., inventory, warehouse capabilities), marketing & advertisement and strategic supplier partnerships, either by taking advantage of the expected larger volumes in the case of smartphones and netbooks, or to make adjustments by focusing the effort on significantly improving the overall service experience (as inferred from the feature importance) to boost the volumes of laptops and PCs.

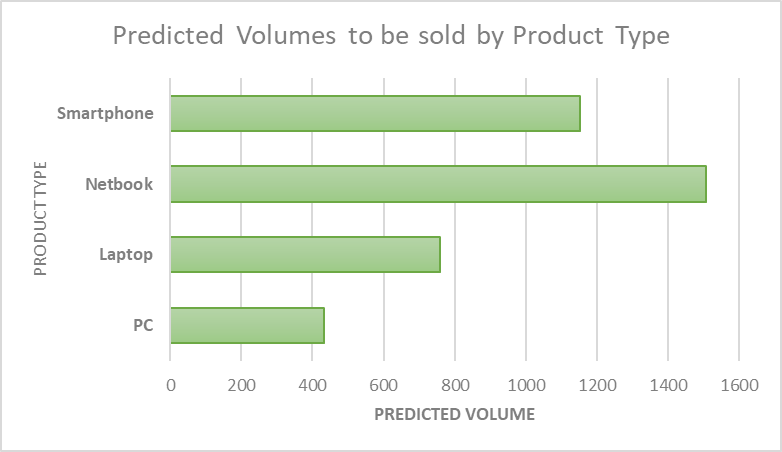


Figure 5: Predicted Volumes to be sold per selected product types.