**Course Three Task2 Report**

**by**

**Shifeng (Steve) Li**

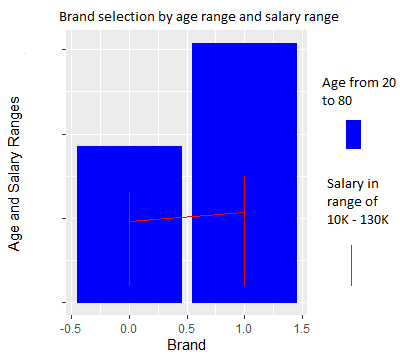
**02/20/2022**

In this project, the Blackwell Electronics customer survey data has been analyzed using R caret data analysis techniques to go through the data science process steps to try classifiers on the data features to find ones with the good accuracy, then to develop prediction model for Blackwell to predict what type of customer would choose the type of product brand. In the following, will describe the different phases of the process.

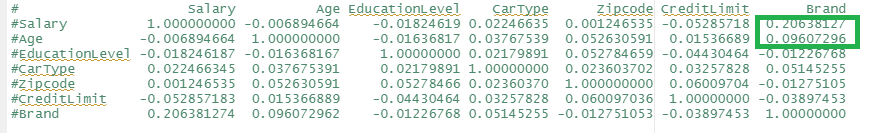
1. **Initial Data Profiling**

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 initial visualizations show the customers’ Age, Salary features have some noticeable relation to the Brand.



And, the correlation matrix quantified data also provide some supporting information that Salary and Age are have more correlation to Brand.



1. **Apply computational techniques to identify / select the key features for prediction**

It is time to use the R caret RFE control to run through different models, and let the runtime computational process to walk through various of feature combination in the Blackwell data set. Then identify the top related features that will be used for prediction.

The Random Forest and Bagged Trees classifiers have been used to identify the top variables/features that can be potentially used for predict Brand. Both classifiers identified the top variables are Salary and Age.

1. **Build and train the different models with Blackwell data set.**

In this part of process, using the R caret tools to build and train the different models that use Random Tree and Bagged Tree algorithms. The following are the quantified measurements of the three models training results.

1. Random Forest

Again, the Random Forest model, when running through the training data, shows the Salary and Age features are important for predicting Brand.

mtry RMSE Rsquared MAE

3 0.2387419 0.7594380 0.1318598

4 0.2367611 0.7612949 0.1150816

RMSE was used to select the optimal model using the smallest value.

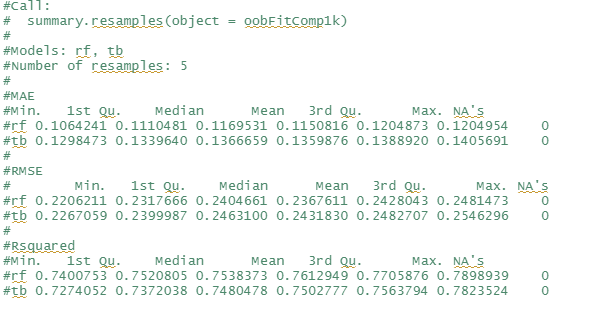
The final value used for the model was mtry = 4.

1. Bagged Tree

Similarly, the Bagged Tree model, when running through the training data, shows the Salary and Age features are important for predicting Brand.

1. **Model selection**

After built and trained the three models, now use the R caret resample summary to list the benchmarking attributes of the two models.



As we can see, the Random Forest classification model has the better fit for forecasting the Brand selection, based on the Salary and Age features.

1. **Projection**

When applying the trained Random Forest classification model and the trained TreeBag classification model to the new data set, the incompleted survey data, the predication

From Random Tree Classification Model

RMSE Rsquared MAE

0.745041897 0.005677289 0.612150467

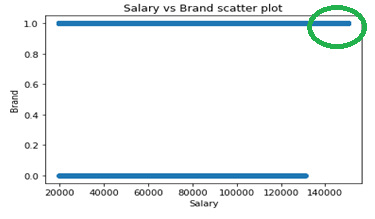
From TreeBag Classification Model

RMSE Rsquared MAE

0.73170065 0.00520051 0.61133302

So, the Random Forest classification model does the better prediction for the new data set.

And, the following visualization tells that the highest salary group customers tend to buy Sony Brand.



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