**Course Three Task2 Report**

**by**

**Shifeng (Steve) Li**

**02/13/2022**

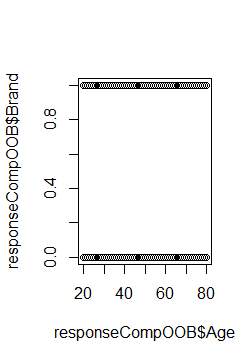
In this project, the Blackwell Electronics customer survey data has been analyzed using R caret data analysis techniques to go through the data science process steps to gain the good accuracy to identify co-related patterns and to develop prediction model for Blackwell. In the following, will describe the different phases of the process.

1. **Initial Data Profiling**

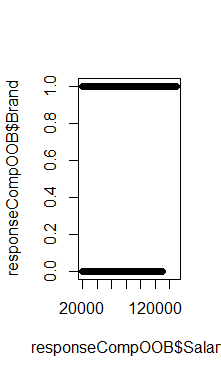
As the standard data analysis procedure, went through the initial data set analysis to understand the data profile and structure, validated the data and visualized the data correlations.

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

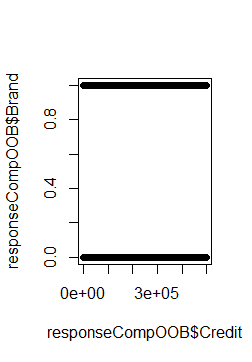
See the following three individual plots



Age – Brand plot

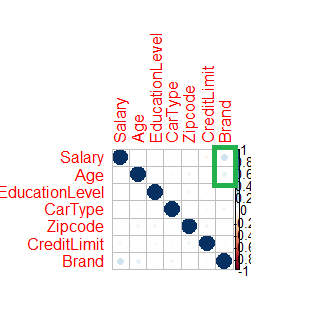


Age – Brand plot

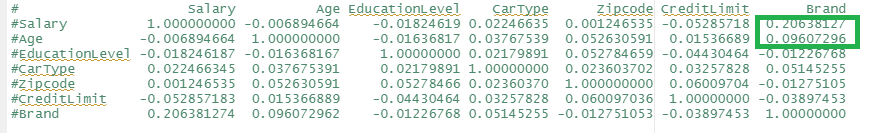


Credit – Brand plot

The correlation matrix plot also provides the supporting information that Brand selection is more related to the Salary and Age.



Now, 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**

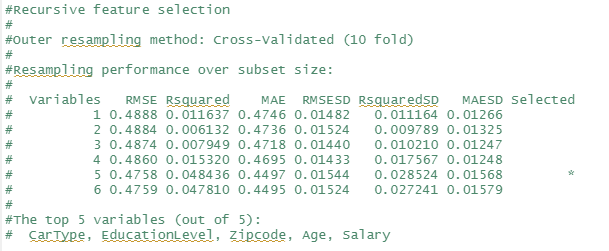
It is time to use the R caret RFE control to run through different models, and let the runtime computational process to walk through various of feature combination in the Blackwell data set. Then identify the top related features that will be used for prediction.

In this project, processed through three algorithms to find the top variables/features that can be potentially used for predict Brand.

1. Linear Model

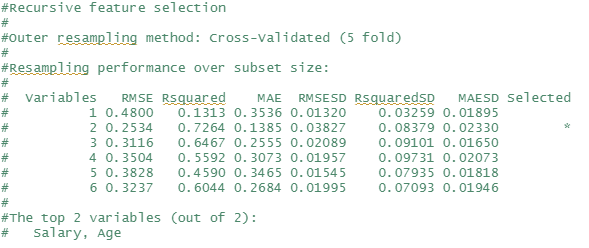
The first tried algorithm is linear model, the R RFE control did the recursive feature selection for the Blackwell data set using the linear Model, and produced the summary as the following. Not ideal, as most of the features in the data set are in the radar.

So, let’s see how other algorithm goes.



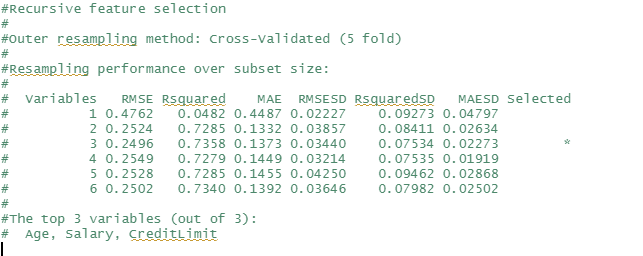
1. Random Forest

The next tried algorithm is Random Forest, the R RFE control did the recursive feature selection for the Blackwell data set using the Random Forest Classifier, and produced the summary as the following. Now we can see the identified top variables scope is more focused to Salary and Age.



1. Bagged Trees

The next tried algorithm is Bagged Trees, the R RFE control did the recursive feature selection for the Blackwell data set using the Bagged Trees algorithm, and produced the summary as the following. Now we can see the identified top variables are Salary and Age, and CreditLimit.



During running through the R RFE control computational process to do feature selection, need to turn on/register the cluster to utilize the multiple CPU cores on the computer to provide the needed computing power.

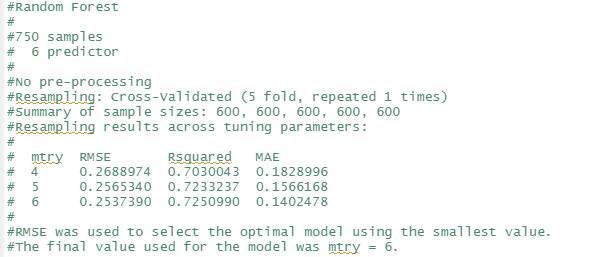
Now, it is time to choose the proper algorithm for building the model to train the data and then do the predication.

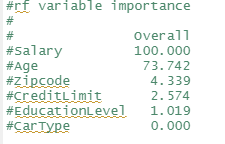
1. **Build and train the different models with Blackwell data set.**

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

1. Random Forest

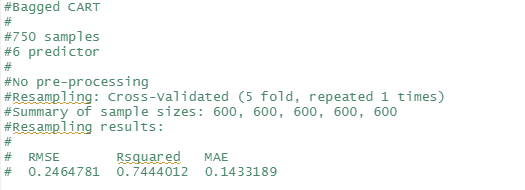
Again, the Random Forest model, when running through the training data, shows the Salary and Age features are important for predicting Brand.

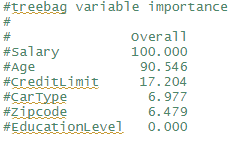




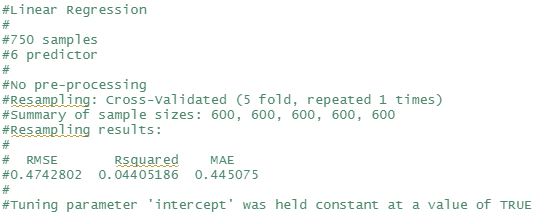
1. Bagged Tree

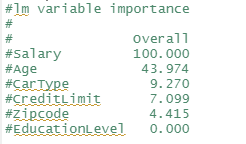
Similarly, the Bagged Tree model, when running through the training data, shows the Salary and Age features are important for predicting Brand.





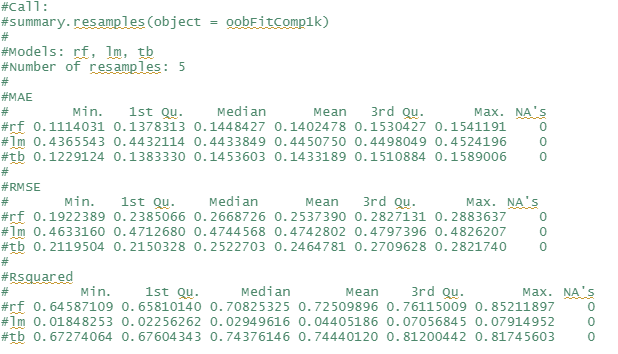
1. Linear Model





1. **Model selection**

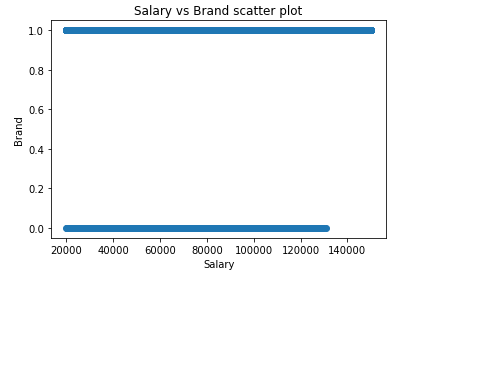
After built and trained the three models, now use the R caret resample summary to list the benchmarking attributes of the three models.



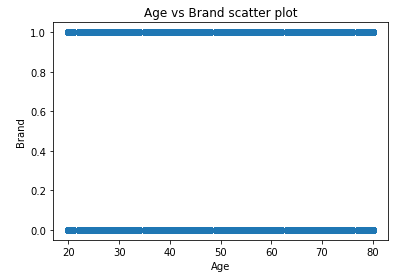
As we can see, the Random Forest and Bagged Tree algorithms based models have the better fit for forecasting the Brand selection, based on the Salary and Age features.

1. **Projection**

The highest salary group customers tend to buy Sony Brand.

****

The age factor, though identified as one of the top variables, does not contribute significant difference to the Brank choice.

****

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