Brand Prediction

C3t2- Sugitha Devarajan

Overview

- Marketing conducted a survey of existing customer and one of the objectives of the survey was to find out which of two brands of computers our customers prefer. This information will help Blackwell Electronics decide with which manufacturer they should pursue a deeper strategic relationship
- In the survey conducted there are some incomplete questions to the brand variable and the request is to build a model in R to predict the incompletes and help marketing with their strategy.

Data

- There are 3 files given
  - CompleteResponse csv
  - Survey Keys
  - SurveyIncomplete csv
- CompleteResponse csv has 7 variables
  - Salary – excluding bonus
  - Age – Customer age
  - E-Level – Education level
  - 0-Less than highschool,1 high school,2-some college,3 – 4 yr. degrees, and 4 is masters
  - Primary car – 1 to 20 options
  - Zip code – 9 regions in the us
  - Credit – credit available to the customer
  - Brand – 0 Acer and 1 Sony

```
> summary(df_pb)
     salary           age            elevel           car             zipcode          credit           brand      
 Min.   : 20000   Min.   :20.00   Min.   :0.000   Min.   : 1.00   Min.   :0.000   Min.   :     0   Min.   :0.0000  
 1st Qu.: 52082   1st Qu.:35.00   1st Qu.:1.000   1st Qu.: 6.00   1st Qu.:2.000   1st Qu.:120807   1st Qu.:0.0000  
 Median : 84950   Median :50.00   Median :2.000   Median :11.00   Median :4.000   Median :250607   Median :1.0000  
 Mean   : 84871   Mean   :49.78   Mean   :1.983   Mean   :10.52   Mean   :4.041   Mean   :249176   Mean   :0.6217  
 3rd Qu.:117162   3rd Qu.:65.00   3rd Qu.:3.000   3rd Qu.:15.75   3rd Qu.:6.000   3rd Qu.:374640   3rd Qu.:1.0000  
 Max.   :150000   Max.   :80.00   Max.   :4.000   Max.   :20.00   Max.   :8.000   Max.   :500000   Max.   :1.0000  
```

```
> str(df_pb)
'data.frame':   9898 obs. of  7 variables:
 $ salary : num  119807 106880 78021 63690 50874 ...
 $ age    : int  45 63 23 51 20 56 24 62 29 41 ...
 $ elevel : int  0 1 0 3 3 3 4 3 4 1 ...
 $ car    : int  14 11 15 6 14 14 8 3 17 5 ...
 $ zipcode: int  4 6 2 5 4 3 5 0 0 4 ...
 $ credit : num  442038 45007 48795 40889 352951 ...
 $ brand  : int  0 1 0 1 0 1 1 1 0 1 ...
> names(df_pb)
[1] "salary"  "age"     "elevel"  "car"     "zipcode" "credit"  "brand"
> sum(is.na(df_pb)) #no na
[1] 0
```

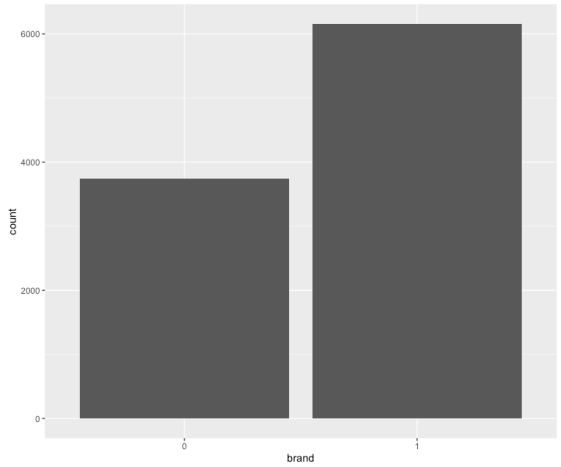
Data Preprocessing

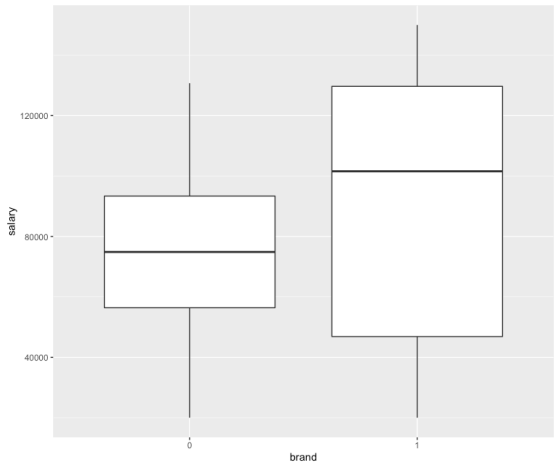
- There was no missing data in Complete Response csv
- Brand variable is converted to factor for modeling purpose

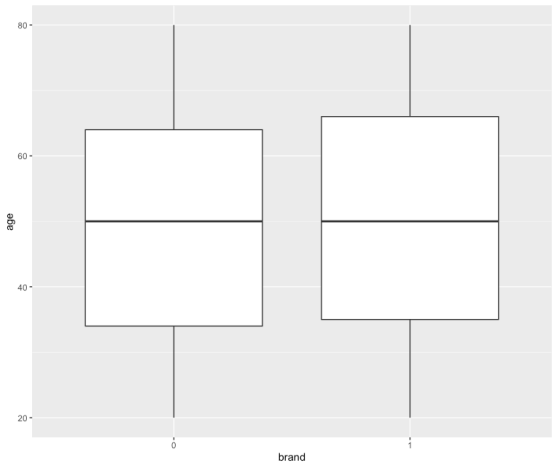
Correlation

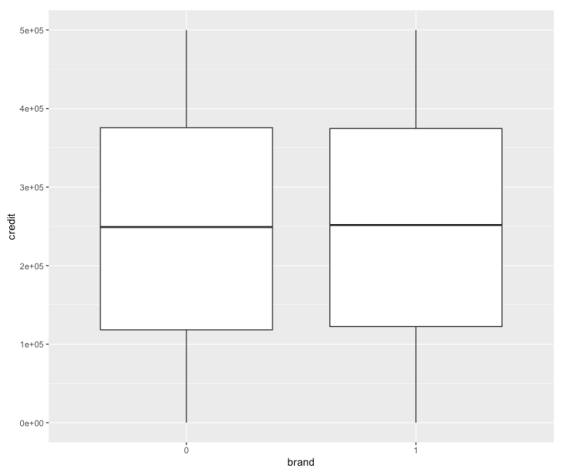
- There was not much correlation found between variables

```
> print(correlationMatrix)
               salary          age        elevel           car       zipcode       credit        brand
salary   1.000000000  0.007978566 -6.620234e-03 -6.090575e-03 -0.005471132 -0.025126808  0.206489883
age      0.007978566  1.000000000 -5.830340e-03  1.024607e-02  0.003681375 -0.004400692  0.013713286
elevel  -0.006620234 -0.005830340  1.000000e+00 -4.676852e-05  0.018095400  0.002720642 -0.004828912
car     -0.006090575  0.010246067 -4.676852e-05  1.000000e+00  0.001526528 -0.010329137  0.005923147
zipcode -0.005471132  0.003681375  1.809540e-02  1.526528e-03  1.000000000  0.004962011  0.004665088
credit  -0.025126808 -0.004400692  2.720642e-03 -1.032914e-02  0.004962011  1.000000000  0.005688438
brand    0.206489883  0.013713286 -4.828912e-03  5.923147e-03  0.004665088  0.005688438  1.000000000
```


count
6000
4000
2000
0
0
1
brand


salary
120000
80000
40000
0
1
brand


age
80
60
40
20
0
1
brand


credit
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
0
1
brand

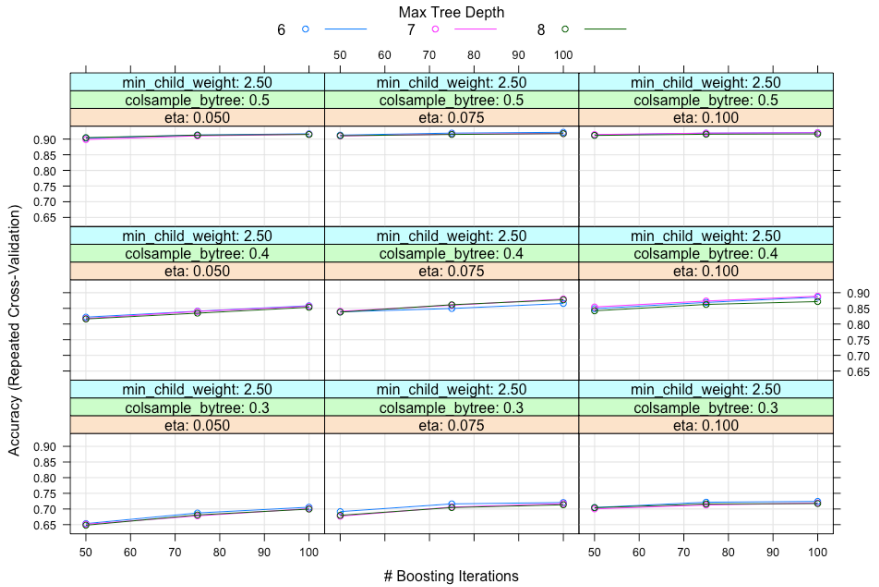
EDA

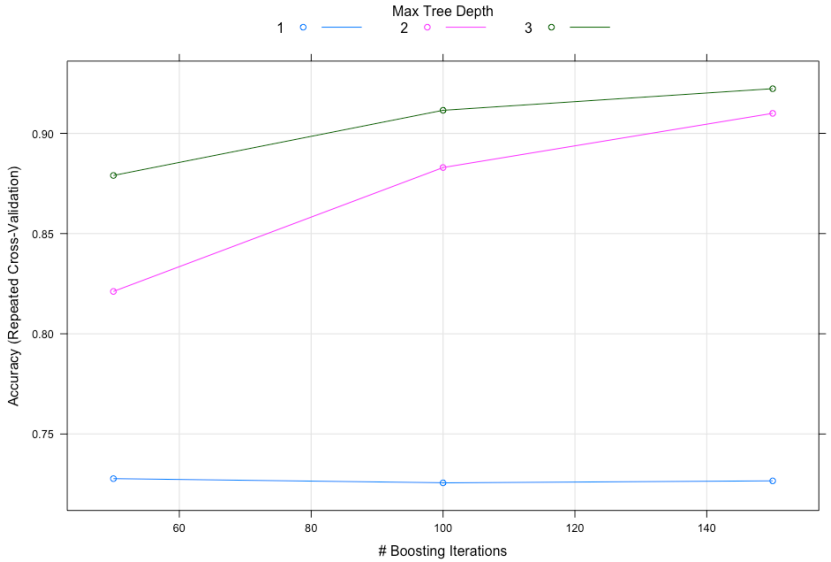
Modeling

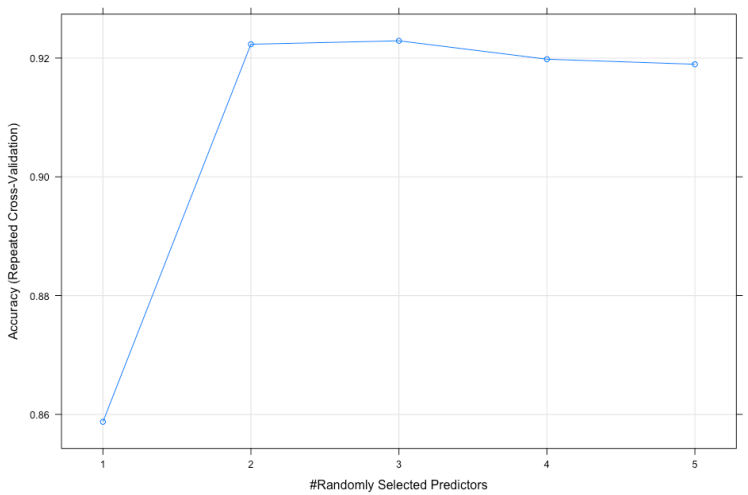
- Complete response csv data is split in to 75 and 25 for training and testing.
- Train control is set to 10-fold CV repeated 3 times for all my models
- Model selected are XgbTree, GBM, RF, and C5.0
- Xgbtree was trained with manual tune grid.
- GBM was trained with automatic tuning grid
- Random Forest was manually tuned 5 different mtry values.
- C5 was tunes with winnow set to false and trials c(1:100) and model equal to tree.

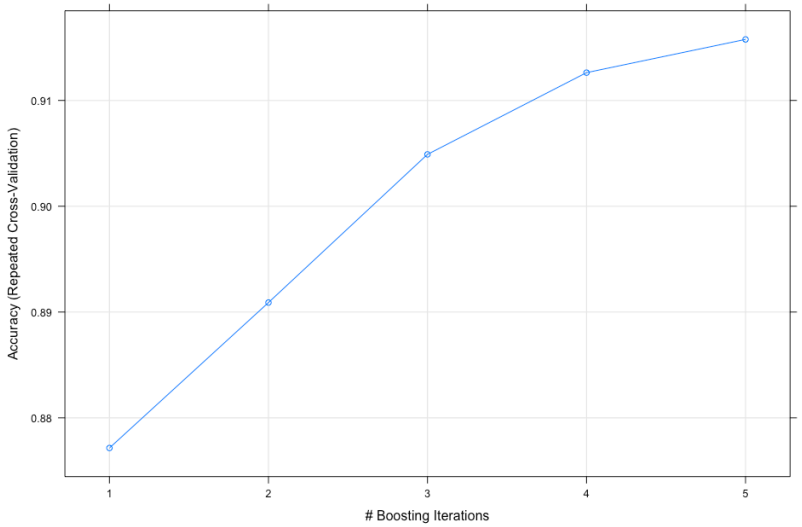
Fit Plot – All models

RF model seems to predict better than rest of the models









Variable Importance – All models

- All models say Salary and age are the most important variable

```
> gbmImp_xg
xgbTree variable importance

         Overall
salary  0.609680
age     0.351646
credit  0.021331
car     0.008357
zipcode 0.005207
elevel  0.003779
> gbmImp1
gbm variable importance

         Overall
salary  1727.807
age     1525.684
credit    26.320
car       10.290
zipcode    3.856
elevel     2.259
> rfImp
rf variable importance

        Overall
salary  1806.36
age     1158.32
credit   225.24
car      129.68
zipcode   98.77
elevel    71.03
> c5Imp
C5.0 variable importance

        Overall
salary   100.00
age       85.30
car       12.34
credit     8.74
zipcode    8.15
elevel     6.44
>
```

Model Selection - RF due to Accuracy

```
Call:
summary.resamples(object = resample_results)

Models: GBM, RF, C5.0
Number of resamples: 30

Accuracy
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
GBM  0.8936743 0.9152086 0.9244265 0.9222784 0.9299191 0.9407008    0
RF   0.9004038 0.9154313 0.9232323 0.9229055 0.9272972 0.9487871    0
C5.0 0.8869448 0.9088156 0.9131317 0.9157664 0.9229475 0.9460916    0

Kappa
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
GBM  0.7765805 0.8221654 0.8399537 0.8357387 0.8516180 0.8739858    0
RF   0.7902875 0.8202569 0.8377232 0.8361705 0.8460365 0.8908663    0
C5.0 0.7635802 0.8057545 0.8150285 0.8213275 0.8373849 0.8852823    0
```

```
> confusionMatrix(preds_rf, bp_testing$brand)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  834   83
         1  102 1455

               Accuracy : 0.9252
                 95% CI : (0.9141, 0.9353)
    No Information Rate : 0.6217
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8404

 Mcnemar's Test P-Value : 0.1857

            Sensitivity : 0.8910
            Specificity : 0.9460
         Pos Pred Value : 0.9095
         Neg Pred Value : 0.9345
             Prevalence : 0.3783
         Detection Rate : 0.3371
   Detection Prevalence : 0.3707
      Balanced Accuracy : 0.9185

       'Positive' Class : 0
```

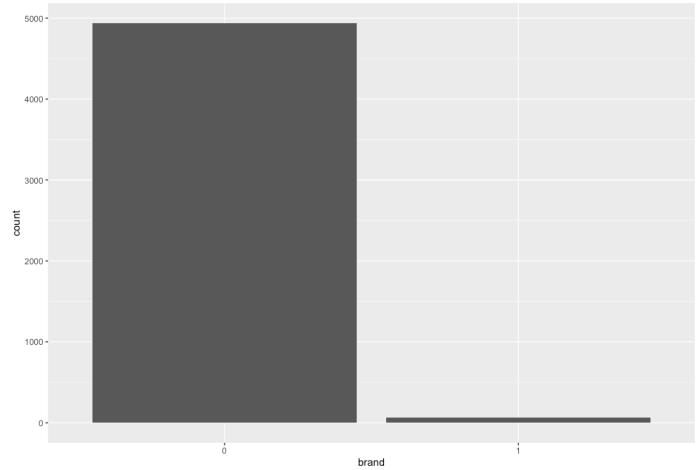
```
> confusionMatrix(preds_c5, bp_testing$brand)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  832   91
         1  104 1447

               Accuracy : 0.9212
                 95% CI : (0.9099, 0.9315)
    No Information Rate : 0.6217
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.832

 Mcnemar's Test P-Value : 0.3902

            Sensitivity : 0.8889
            Specificity : 0.9408
         Pos Pred Value : 0.9014
         Neg Pred Value : 0.9329
             Prevalence : 0.3783
         Detection Rate : 0.3363
   Detection Prevalence : 0.3731
      Balanced Accuracy : 0.9149

       'Positive' Class : 0
```

```
> confusionMatrix(preds_gbm, bp_testing$brand)
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0  846   81
         1   90 1457

               Accuracy : 0.9309
                 95% CI : (0.9202, 0.9406)
    No Information Rate : 0.6217
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8528

 Mcnemar's Test P-Value : 0.5407

            Sensitivity : 0.9038
            Specificity : 0.9473
         Pos Pred Value : 0.9126
         Neg Pred Value : 0.9418
             Prevalence : 0.3783
         Detection Rate : 0.3420
   Detection Prevalence : 0.3747
      Balanced Accuracy : 0.9256

       'Positive' Class : 0
```

Confusion Matrix of Test data prediction

Predicting the Survey Incomplete

```
> postResample(preds_rf, bp_testing$brand)
 Accuracy     Kappa
0.9252223 0.8403981
> postResample(preds_final_ic_rf, df_ic_bp$brand)
  Accuracy      Kappa
0.38300000 0.01182682
```



Since we are comparing corrupted brand against the complete data the accuracy dropped, its not the model ;)

Conclusion

- We found from the prediction that Sony is widely selected in the 15000 survey records.