Expand.grid? did not try it

# Classification: Predict which brand of products customers prefer

* if customer responses to some survey questions (e.g. income, age, etc.) enable us to predict the answer to the brand preference question.
* If we can do this with confidence, I would like you to make those predictions and provide the sales team with a complete view of what brand our customers prefer.
* run and optimize at least two different decision tree classification methods in R - C5.0 and RandomForest - and compare which one works better for this data set.
* the file labelled **CompleteResponses.csv** is the data set you will use to train your model and build your predictive model.
  + It includes 10,000 fully-answered surveys and the key to the survey can be found in ***survey\_key.csv***.
* The file labelled **SurveyIncomplete.csv** will be your main test set (the data you will apply your optimized model to predict the brand preference).
  + You'll be applying your trained and tested model to this data to prepare the model for production.

***Customer Brand Preferences Report.*** A report in a Zip file that includes:

* A brief summary in Word or Excel of your methods and results that includes:
  + The classifiers you tried.
  + The classifier you selected to make the predictions, including a rationale for selecting the method you did and the level of confidence in the predictions.
  + The predicted answers to the brand preference question for the instances of survey results that are missing that answer.
  + A chart that displays the customer preference for each brand based on the combination of the actual answers and the predicted answers to the brand preference survey question.
* The results of each classifier you ran exported from R

## Understanding the dataset

Salary

yearly salary without bonus

### Age

### Education level

|  |  |
| --- | --- |
| Value | Description |
| 0 | Less than High School Degree |
| 1 | High School Degree |
| 2 | Some College |
| 3 | 4-Year College Degree |
| 4 | Master's, Doctoral or Professional Degree |

Car

|  |  |
| --- | --- |
| Value | Description |
| 1 | BMW |
| 2 | Buick |
| 3 | Cadillac |
| 4 | Chevrolet |
| 5 | Chrysler |
| 6 | Dodge |
| 7 | Ford |
| 8 | Honda |
| 9 | Hyundai |
| 10 | Jeep |
| 11 | Kia |
| 12 | Lincoln |
| 13 | Mazda |
| 14 | Mercedes Benz |
| 15 | Mitsubishi |
| 16 | Nissan |
| 17 | Ram |
| 18 | Subaru |
| 19 | Toyota |
| 20 | None of the above |

### Zipcode

|  |  |
| --- | --- |
| Value | Region |
| 0 | New England |
| 1 | Mid-Atlantic |
| 2 | East North Central |
| 3 | West North Central |
| 4 | South Atlantic |
| 5 | East South Central |
| 6 | West South Central |
| 7 | Mountain |
| 8 | Pacific |

Divide it into 4 regions as we did in the previous task.

### Credit

The amount of credit that is available to the respondent

### Brand

Choice between Acer = 0 or Sony = 1 which they prefer.

### Notes on exploring the data and survey key

The respondents are only able to choose between Acer and Sony, but maybe they prefer to buy a different brand altogether, such as Apple.

The questionnaire assumes all respondents drive a car, there is no option of “I have no car”, only “none of the above”, which could mean that the respondent has no car or owns a different brand.

The yearly salary is filled in, but it is not clear if people enter their gross of net salary. Some people focus on net salary, so it is possible that this is not completely clear.

How are the customers samples and why did they ask about their car brand.

Attributes are all transformed to numerical values: salary, age, education level, car, zipcode, credit, brand. Of which “brand” will be our dependent variable.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Salary** | **Age** | **Level of education** | **Car** | **Zip Code** | **Credit** | **Brand** |
| Min | $20.000 | 20 | 0 | 1 | 0 | 0 | 0 |
| 1st Qu.: | $52.082 | 35 | 1 | 6 | 2 | $120.807 | 0 |
| Median : | $84.950 | 50 | 2 | 11 | 4 | $250.607 | 1 |
| Mean | $84.871 | 50 | 2 | 11 | 4 | $249.176 | 1 |
| 3rd Qu.: | $117.162 | 65 | 3 | 16 | 6 | $374.640 | 1 |
| Max | $150.000 | 80 | 4 | 20 | 8 | $500.000 | 1 |

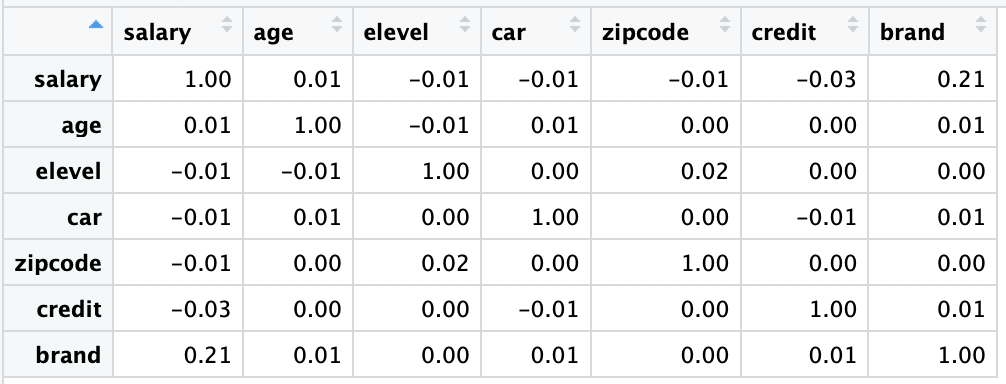
Figure Descriptive statistics Complete Survey

First, the descriptive statistics of both datasets are explored and a correlation matrix is made to see if there are correlations between the attributes. Salary and brand have a correlation of 0.21, all other attributes have very low correlations.

Second, boxplots are made of all the attributes to see if there are important outliers that must be dealt with. There are no outliers and the boxplots seem quite evenly distributed. Several scatterplots are made to see the relation between the attributes and brand. But no clear correlation is discernable, except a small one between salary and brand. Respondents with high salaries tend to prefer Sony a bit more often.

With ggplot it is possible to add a third feature into the graph. If salary, age and brand preference are combined an interesting image appears.

It could be possible that there are regional differences that are invisible now but are hiding in this dataset. Also, in the previous datasets the regions were divided into 4 regions, here there are 8 different zipcodes. For future research it is advised to do some feature selection and engineering to be able to make comparisons between the datasets



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Salary** | **Age** | **Level of education** | **Car** | **Zip Code** | **Credit** | **Brand** |
| Min | $20.000 | 20 | 0 | 1 | 0 | 0 | Na |
| 1st Qu.: | $52.242 | 35 | 1 | 6 | 2 | $121.878 | Na |
| Median : | $85.969 | 50 | 2 | 11 | 4 | $250.870 | Na |
| Mean | $85.560 | 50 | 2 | 11 | 4 | $249.510 | Na |
| 3rd Qu.: | $118.380 | 65 | 3 | 16 | 6 | $375425 | Na |
| Max | $150.000 | 80 | 4 | 20 | 8 | $500.000 | Na |

Figure 2 Descriptive statistics Incomplete Survey

* What parameters to tune?
* What search method to use to locate good algorithm parameters?
* What test options to use to limit overfitting the training data?

Direct from thehelp page for the *randomForest()* function in R:

* **mtry**: Number of variables randomly sampled as candidates at each split.
* **ntree**: Number of trees to grow.

## C5.0, 10 fold cross validation, tunelength =2

C5.0

7424 samples

6 predictor

2 classes: '0', '1'

Pre-processing: scaled (9), centered (9)

Resampling: Cross-Validated (10 fold, repeated 1 times)

Summary of sample sizes: 6682, 6681, 6682, 6681, 6682, 6681, ...

Resampling results across tuning parameters:

winnow trials Accuracy Kappa

FALSE 95 0.9224142 0.8355817

TRUE 91 0.9217431 0.8343297

Tuning parameter 'model' was held constant at a value of tree

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were trials = 95, model = tree and winnow = FALSE.

### Calculated Performance Across Resamples

Accuracy Kappa

0.9163298 0.8229055

Overall Importance

salary 100.00

age 100.00

elevel.Q 70.89

elevel^4 70.89

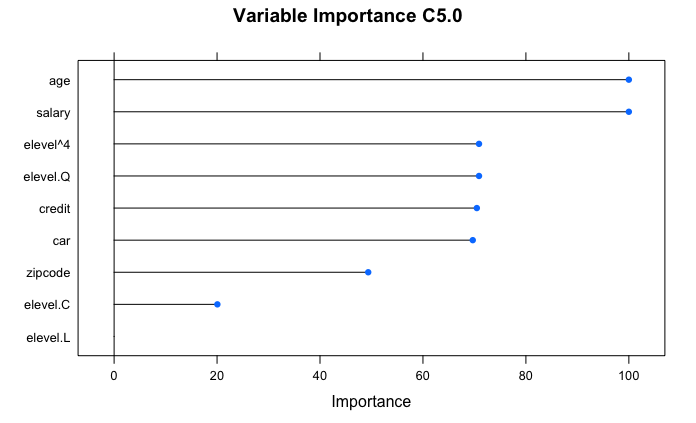
credit 70.47

car 69.68

zipcode 49.37

elevel.C 20.06

elevel.L 0.00



## Random Forest Model, 10-fold Cross Validation, parameterGrid

Random Forest

7424 samples

6 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 1 times)

Summary of sample sizes: 6682, 6681, 6682, 6681, 6682, 6681, ...

Resampling results across tuning parameters:

mtry Accuracy Kappa

1 0.6219019 0.0004421814

2 0.8930546 0.7714677326

3 0.9229558 0.8366806264

4 0.9236293 0.8380590496

5 0.9229558 0.8366412487

6 0.9217429 0.8340187215

7 0.9208002 0.8320109496

Accuracy was used to select the optimal model using the largest value.

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

### Calculated Performance Across Resamples

Accuracy Kappa

0.9183508 0.8270031

variable importance

Overall

salary 100.0000

age 61.5619

credit 12.1583

car 6.2617

zipcode 4.2845

#filter on zipcode=0

CompleteResponses[CompleteResponses$zipcode==1, ]

CompleteReponsesZip1 <- CompleteResponses[CompleteResponses$zipcode==1, ]

summary(CompleteReponsesZip1)

> summary(CompleteReponsesZip0)

salary age elevel car zipcode credit brand

Min. : 20000 Min. :20.00 Min. :0.000 Min. : 1.00 Min. :0 Min. : 0 Min. :0.000

1st Qu.: 52357 1st Qu.:34.00 1st Qu.:1.000 1st Qu.: 6.00 1st Qu.:0 1st Qu.:114635 1st Qu.:0.000

Median : 85804 Median :50.00 Median :2.000 Median :10.00 Median :0 Median :243432 Median :1.000

Mean : 85399 Mean :49.97 Mean :1.957 Mean :10.48 Mean :0 Mean :243964 Mean :0.623

3rd Qu.:118046 3rd Qu.:65.00 3rd Qu.:3.000 3rd Qu.:15.00 3rd Qu.:0 3rd Qu.:374107 3rd Qu.:1.000

Max. :150000 Max. :80.00 Max. :4.000 Max. :20.00 Max. :0 Max. :500000 Max. :1.000

> summary(CompleteReponsesZip1)

salary age elevel car zipcode credit brand

Min. : 20000 Min. :20.00 Min. :0.000 Min. : 1.00 Min. :1 Min. : 0 Min. :0.0000

1st Qu.: 55283 1st Qu.:35.00 1st Qu.:1.000 1st Qu.: 6.00 1st Qu.:1 1st Qu.:119323 1st Qu.:0.0000

Median : 86445 Median :50.00 Median :2.000 Median :10.00 Median :1 Median :249557 Median :1.0000

Mean : 86285 Mean :50.41 Mean :1.978 Mean :10.45 Mean :1 Mean :247578 Mean :0.6201

3rd Qu.:117642 3rd Qu.:66.00 3rd Qu.:3.000 3rd Qu.:15.00 3rd Qu.:1 3rd Qu.:374845 3rd Qu.:1.0000

Max. :150000 Max. :80.00 Max. :4.000 Max. :20.00 Max. :1 Max. :500000 Max. :1.0000

> summary(CompleteReponsesZip2)

salary age elevel car zipcode credit brand

Min. : 20000 Min. :20.00 Min. :0.00 Min. : 1.00 Min. :2 Min. : 0 Min. :0.0000

1st Qu.: 53672 1st Qu.:33.00 1st Qu.:1.00 1st Qu.: 5.00 1st Qu.:2 1st Qu.:133547 1st Qu.:0.0000

Median : 84739 Median :48.00 Median :2.00 Median :11.00 Median :2 Median :261533 Median :1.0000

Mean : 85294 Mean :48.53 Mean :1.94 Mean :10.63 Mean :2 Mean :256161 Mean :0.6142

3rd Qu.:118332 3rd Qu.:64.00 3rd Qu.:3.00 3rd Qu.:16.00 3rd Qu.:2 3rd Qu.:379552 3rd Qu.:1.0000

Max. :150000 Max. :80.00 Max. :4.00 Max. :20.00 Max. :2 Max. :500000 Max. :1.0000

> summary(CompleteReponsesZip3)

salary age elevel car zipcode credit brand

Min. : 20000 Min. :20.0 Min. :0.00 Min. : 1.00 Min. :3 Min. : 0 Min. :0.0000

1st Qu.: 48323 1st Qu.:33.0 1st Qu.:1.00 1st Qu.: 5.00 1st Qu.:3 1st Qu.:125060 1st Qu.:0.0000

Median : 80584 Median :48.0 Median :2.00 Median :10.00 Median :3 Median :255117 Median :1.0000

Mean : 82405 Mean :48.8 Mean :1.95 Mean :10.24 Mean :3 Mean :252140 Mean :0.6269

3rd Qu.:114664 3rd Qu.:64.0 3rd Qu.:3.00 3rd Qu.:15.00 3rd Qu.:3 3rd Qu.:378463 3rd Qu.:1.0000

Max. :150000 Max. :80.0 Max. :4.00 Max. :20.00 Max. :3 Max. :500000 Max. :1.0000

> summary(CompleteReponsesZip4)

salary age elevel car zipcode credit brand

Min. : 20000 Min. :20.00 Min. :0.000 Min. : 1.00 Min. :4 Min. : 0 Min. :0.0000

1st Qu.: 49800 1st Qu.:34.50 1st Qu.:1.000 1st Qu.: 6.00 1st Qu.:4 1st Qu.:117230 1st Qu.:0.0000

Median : 85543 Median :51.00 Median :2.000 Median :11.00 Median :4 Median :235624 Median :1.0000

Mean : 84770 Mean :50.03 Mean :1.943 Mean :10.83 Mean :4 Mean :241164 Mean :0.6293

3rd Qu.:119195 3rd Qu.:66.00 3rd Qu.:3.000 3rd Qu.:16.00 3rd Qu.:4 3rd Qu.:360544 3rd Qu.:1.0000

Max. :150000 Max. :80.00 Max. :4.000 Max. :20.00 Max. :4 Max. :500000 Max. :1.0000

> summary(CompleteReponsesZip5)

salary age elevel car zipcode credit brand

Min. : 20000 Min. :20.00 Min. :0.000 Min. : 1.00 Min. :5 Min. : 0 Min. :0.000

1st Qu.: 52351 1st Qu.:36.00 1st Qu.:1.000 1st Qu.: 6.00 1st Qu.:5 1st Qu.:115753 1st Qu.:0.000

Median : 84611 Median :52.00 Median :2.000 Median :11.00 Median :5 Median :252041 Median :1.000

Mean : 84877 Mean :51.03 Mean :2.014 Mean :10.61 Mean :5 Mean :250603 Mean :0.611

3rd Qu.:116570 3rd Qu.:67.00 3rd Qu.:3.000 3rd Qu.:16.00 3rd Qu.:5 3rd Qu.:382011 3rd Qu.:1.000

Max. :150000 Max. :80.00 Max. :4.000 Max. :20.00 Max. :5 Max. :500000 Max. :1.000

salary age elevel car zipcode credit brand

Min. : 20000 Min. :20.00 Min. :0.000 Min. : 1.0 Min. :6 Min. : 0 Min. :0.0000

1st Qu.: 53279 1st Qu.:35.00 1st Qu.:1.000 1st Qu.: 5.0 1st Qu.:6 1st Qu.:122068 1st Qu.:0.0000

Median : 84414 Median :50.00 Median :2.000 Median :10.0 Median :6 Median :259919 Median :1.0000

Mean : 84547 Mean :49.78 Mean :2.048 Mean :10.4 Mean :6 Mean :253202 Mean :0.6182

3rd Qu.:115982 3rd Qu.:65.00 3rd Qu.:3.000 3rd Qu.:15.0 3rd Qu.:6 3rd Qu.:381172 3rd Qu.:1.0000

Max. :150000 Max. :80.00 Max. :4.000 Max. :20.0 Max. :6 Max. :500000 Max. :1.0000

> summary(CompleteReponsesZip7)

salary age elevel car zipcode credit brand

Min. : 20000 Min. :20.00 Min. :0.000 Min. : 1.00 Min. :7 Min. : 0 Min. :0.0000

1st Qu.: 54339 1st Qu.:35.00 1st Qu.:1.000 1st Qu.: 6.00 1st Qu.:7 1st Qu.:121358 1st Qu.:0.0000

Median : 88239 Median :50.00 Median :2.000 Median :11.00 Median :7 Median :246176 Median :1.0000

Mean : 86645 Mean :50.09 Mean :1.975 Mean :10.63 Mean :7 Mean :246213 Mean :0.6002

3rd Qu.:118101 3rd Qu.:65.00 3rd Qu.:3.000 3rd Qu.:15.00 3rd Qu.:7 3rd Qu.:369412 3rd Qu.:1.0000

Max. :150000 Max. :80.00 Max. :4.000 Max. :20.00 Max. :7 Max. :500000 Max. :1.0000

salary age elevel car zipcode credit brand

Min. : 20000 Min. :20.00 Min. :0.000 Min. : 1.00 Min. :8 Min. : 0 Min. :0.000

1st Qu.: 50754 1st Qu.:35.00 1st Qu.:1.000 1st Qu.: 5.00 1st Qu.:8 1st Qu.:124041 1st Qu.:0.000

Median : 83153 Median :49.00 Median :2.000 Median :10.00 Median :8 Median :254582 Median :1.000

Mean : 83713 Mean :49.39 Mean :2.035 Mean :10.43 Mean :8 Mean :250987 Mean :0.652

3rd Qu.:115209 3rd Qu.:65.00 3rd Qu.:3.000 3rd Qu.:15.50 3rd Qu.:8 3rd Qu.:377312 3rd Qu.:1.000

Max. :150000 Max. :80.00 Max. :4.000 Max. :20.00 Max. :8 Max. :500000 Max. :1.000

Traincontrol = cross validation

It is part of the model, so we make sure it is not overfitting.

What is a ROC curve?

## RESULTS C5.0:

#run the model

> C50Fit

C5.0

7424 samples

6 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Bootstrapped (25 reps)

Summary of sample sizes: 7424, 7424, 7424, 7424, 7424, 7424, ...

Resampling results across tuning parameters:

model winnow trials Accuracy Kappa

rules FALSE 1 0.8789394 0.7484852

rules FALSE 10 0.9149515 0.8191052

rules FALSE 20 0.9167639 0.8229599

rules FALSE 30 0.9173201 0.8240941

rules FALSE 40 0.9173347 0.8242099

rules FALSE 50 0.9176291 0.8248104

rules FALSE 60 0.9182139 0.8260059

rules FALSE 70 0.9182161 0.8260085

rules FALSE 80 0.9186116 0.8268582

rules FALSE 90 0.9188473 0.8273511

rules TRUE 1 0.8790327 0.7487916

rules TRUE 10 0.9156522 0.8205755

rules TRUE 20 0.9170749 0.8235902

rules TRUE 30 0.9175599 0.8246452

rules TRUE 40 0.9173087 0.8241395

rules TRUE 50 0.9178956 0.8254373

rules TRUE 60 0.9181157 0.8258395

rules TRUE 70 0.9183346 0.8263493

rules TRUE 80 0.9185708 0.8267834

rules TRUE 90 0.9186726 0.8269776

tree FALSE 1 0.8649274 0.7172456

tree FALSE 10 0.9159281 0.8215445

tree FALSE 20 0.9163239 0.8225764

tree FALSE 30 0.9173063 0.8247322

tree FALSE 40 0.9172767 0.8246488

tree FALSE 50 0.9177166 0.8255545

tree FALSE 60 0.9179944 0.8261202

tree FALSE 70 0.9177628 0.8256871

tree FALSE 80 0.9176023 0.8253805

tree FALSE 90 0.9172035 0.8245077

tree TRUE 1 0.8648886 0.7172204

tree TRUE 10 0.9166580 0.8231089

tree TRUE 20 0.9162966 0.8225399

tree TRUE 30 0.9173541 0.8248702

tree TRUE 40 0.9171011 0.8243344

tree TRUE 50 0.9172471 0.8246562

tree TRUE 60 0.9176283 0.8254278

tree TRUE 70 0.9181108 0.8264708

tree TRUE 80 0.9178060 0.8258223

tree TRUE 90 0.9176445 0.8254693

Accuracy was used to select the optimal model using the largest value.

The final values used for the model were trials = 90, model = rules and winnow = FALSE.

## TEST:

Accuracy Kappa

0.9191593 0.8285015

<http://topepo.github.io/caret/random-hyperparameter-search.html>

The default method for optimizing tuning parameters in train is to use a [grid search](http://topepo.github.io/caret/model-training-and-tuning.html#grids). This approach is usually effective but, in cases when there are many tuning parameters, it can be inefficient. An alternative is to use a combination of [grid search and racing](http://topepo.github.io/caret/adaptive-resampling.html). Another is to use a [random selection of tuning parameter combinations](http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf) to cover the parameter space to a lesser extent.

For this reason, it may be inefficient to use random search for the following model codes: ada, AdaBag, AdaBoost.M1, bagEarth, blackboost, blasso, BstLm, bstSm, bstTree, C5.0, C5.0Cost, cubist, earth, enet, foba, gamboost, gbm, glmboost, glmnet, kernelpls, lars, lars2, lasso, lda2, leapBackward, leapForward, leapSeq, LogitBoost, pam, partDSA, pcr, PenalizedLDA, pls, relaxo, rfRules, rotationForest, rotationForestCp, rpart, rpart2, rpartCost, simpls, spikeslab, superpc, widekernelpls, xgbDART, xgbTree.

VarImp()

##### Calculation Of Variable Importance For Regression And Classification Models

only 20 most important variables shown (out of 34)

Overall

salary 100.00

age 100.00

car5 32.22

zipcode4 28.02

car17 24.77

car16 19.80

car20 19.64

elevel.L 18.31

zipcode3 17.61

car14 15.80

elevel.Q 14.08

elevel.C 13.20

car15 11.52

zipcode7 11.52

credit 11.52

car4 9.93

zipcode1 8.96

car8 6.91

zipcode6 5.86

car3 0.92

Blockage: Why does is say elevel.L, elevel.Q and elevel.C?

## RandomForest, default on training data

Call:

randomForest(formula = brand ~ ., data = training)

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 2

OOB estimate of error rate: 8.03%

Confusion matrix:

0 1 class.error

0 2496 312 0.11111111

1 284 4332 0.06152513

## > print(rf2) ntrees = 100

Call:

randomForest(formula = brand ~ ., data = training, ntree = 100, mtry = 4, importance = TRUE, proximity = TRUE)

Type of random forest: classification

Number of trees: 100

No. of variables tried at each split: 4

OOB estimate of error rate: 8.11%

Confusion matrix:

0 1 class.error

0 2470 338 0.12037037

1 264 4352 0.05719237

## print(rf2) ntrees=500

Call:

randomForest(formula = brand ~ ., data = training, mtry = 4, importance = TRUE, proximity = TRUE)

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 4

OOB estimate of error rate: 7.83%

Confusion matrix:

0 1 class.error

0 2484 324 0.11538462

1 257 4359 0.05567591

>print(rf3

Call:

randomForest(formula = brand ~ ., data = training, ntree = 200, mtry = 5, importance = TRUE, proximity = TRUE)

Type of random forest: classification

Number of trees: 200

No. of variables tried at each split: 5

OOB estimate of error rate: 8.31%

Confusion matrix:

0 1 class.error

0 2447 361 0.12856125

1 256 4360 0.05545927  
  
> print(rf3) ntrees =500

Call:

randomForest(formula = brand ~ ., data = training, mtry = 5, importance = TRUE, proximity = TRUE)

Type of random forest: classification

Number of trees: 500

No. of variables tried at each split: 5

OOB estimate of error rate: 8.22%

Confusion matrix:

0 1 class.error

0 2453 355 0.12642450

1 255 4361 0.05524263

#default ntrees = 500

rfmtry4 <- randomForest(brand~.,

data = training,

mtry=4,

importance=TRUE,

proximity=TRUE)

print(rfmtry4)

plot(rfmtry4)

## > rfFit

Random Forest

991 samples

6 predictor

2 classes: '0', '1'

No pre-processing

Resampling: Cross-Validated (10 fold, repeated 3 times)

Summary of sample sizes: 891, 891, 893, 892, 892, 893, ...

Resampling results across tuning parameters:

mtry Accuracy Kappa

1 0.6350409 0.05416854

2 0.8452848 0.66393277

3 0.8960958 0.77916315

4 0.9149251 0.81994306

5 0.9189896 0.82811507

6 0.9163025 0.82183289

7 0.9163061 0.82222287

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was mtry = 5.

> #Calculate Performance Across Resamples

> postResample(pred.da, testing$brand)

Accuracy Kappa

0.9178174 0.8249072

> #see variable importance

> varImp(rfFit)

rf variable importance

Overall

salary 100.0000

age 64.9711

credit 14.8576

car 8.8374

zipcode 5.2325

elevel.L 0.9772

elevel.C 0.6818

elevel^4 0.3305

elevel.Q 0.0000