Name: Alvee Jawad Chowdhury

Answer 1

Summary of the 4 models

Method	Best Tune	Elapsed	Resampled Accuracy	Test Accuracy
Logistic Model with CV	NA	0.30	0.6616	0.6628
Single Decision Tree using a 5-fold CV	0.0173333	0.25	0.7504	0.6615
random forest model using a OOB	1.0000000	2.28	0.8404	0.8428
random forest model using a 5-fold	1.0000000	6.76	0.8396	0.8425

Answer 2

Summary of the 3 models

Method	Resampled RMSE	Test RMSE
Boosted Tree Model	307.8612	277.3268
Random Forest Model	292.0605	258.5086
Bagged Tree Model	300.9585	262.7169

Answer 3

```
## Support Vector Machines with Linear Kernel
##
## 536 samples
##  17 predictor
##   2 classes: 'CH', 'MM'
##
## Pre-processing: centered (17), scaled (17)
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 429, 430, 429, 428, 428
## Resampling results across tuning parameters:
##
##   C         Accuracy   Kappa
##    0.03125  0.8619559  0.7038711
##    0.06250  0.8582349  0.6964787
##    0.12500  0.8601040  0.7008693
##    0.25000  0.8675637  0.7173054
##    0.50000  0.8675637  0.7172340
##    1.00000  0.8638081  0.7088879
##    2.00000  0.8638254  0.7091919
##    4.00000  0.8582349  0.6968904
##    8.00000  0.8619386  0.7046755
##   16.00000  0.8619559  0.7049208
##   32.00000  0.8564003  0.6929784
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was C = 0.25.
```

```
##  Accuracy     Kappa
## 0.7902622 0.5566757
```

```
## Loading required package: e1071
```

```
##
## Call:
## svm(formula = Purchase ~ ., data = oj_trn, method = "polynomial",
##     degree = 2, trControl = trainControl(method = "cv", number = 5),
##     preProcess = c("center", "scale"))
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  radial
##        cost:  1
##
## Number of Support Vectors:  249
##
##  ( 127 122 )
##
##
## Number of Classes:  2
##
## Levels:
##  CH MM
```

```
##  Accuracy     Kappa
## 0.8089888 0.5919755
```

```
## Support Vector Machines with Radial Basis Function Kernel
##
## 536 samples
##  17 predictor
##   2 classes: 'CH', 'MM'
##
## Pre-processing: centered (17), scaled (17)
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 429, 428, 430, 428, 429
## Resampling results across tuning parameters:
##
##   C     sigma  Accuracy   Kappa
##   0.25  0.125  0.8264210  0.6162188
##   0.25  0.250  0.7984525  0.5463166
##   0.25  0.500  0.7760389  0.4845926
##   0.25  1.000  0.7462878  0.4021429
##   0.25  2.000  0.7108241  0.3027014
##   0.50  0.125  0.8301243  0.6299935
##   0.50  0.250  0.8058599  0.5766822
##   0.50  0.500  0.7908537  0.5416448
##   0.50  1.000  0.7777866  0.5061942
##   0.50  2.000  0.7629548  0.4578369
##   1.00  0.125  0.8264037  0.6255117
##   1.00  0.250  0.8002348  0.5697382
##   1.00  0.500  0.7964612  0.5607280
##   1.00  1.000  0.7889320  0.5432544
##   1.00  2.000  0.7721262  0.5018542
##   2.00  0.125  0.8207609  0.6155857
##   2.00  0.250  0.8095626  0.5922777
##   2.00  0.500  0.7946270  0.5601761
##   2.00  1.000  0.7833595  0.5357018
##   2.00  2.000  0.7665014  0.4926210
##   4.00  0.125  0.8245515  0.6238199
##   4.00  0.250  0.8151704  0.6063285
##   4.00  0.500  0.7908710  0.5526628
##   4.00  1.000  0.7721615  0.5109841
##   4.00  2.000  0.7758652  0.5135017
##   8.00  0.125  0.8300897  0.6364280
##   8.00  0.250  0.8133189  0.6022398
##   8.00  0.500  0.7853158  0.5410643
##   8.00  1.000  0.7815425  0.5315776
##   8.00  2.000  0.7721442  0.5087854
##
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were sigma = 0.125 and C = 0.5.
```

```
##  Accuracy     Kappa
## 0.7940075 0.5536610
## Random Forest
##
## 536 samples
##  17 predictor
##   2 classes: 'CH', 'MM'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 429, 428, 429, 428, 430
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    2    0.8189087  0.6142391
##    9    0.8432963  0.6700407
##   17    0.8432264  0.6717601
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 9.
```

```
## [1] 0.7846442
```

Summary of the 4 models

Model	Kernel	Resampled Accuracy	Test Accuracy
SVM	linear	0.8675637	0.7902622
SVM	polynomial	0.8768657	0.8089888
SVM	radial	0.8301243	0.7940075
random forest	N/A	0.8432963	0.7846442

The SVM with polynomial Kernel is the best model since it has the highest test accuracy.

Answer 4

a. time taken for rf oob method

```
## elapsed
##    2.28
```

time taken for rf 5-fold cv

```
## elapsed
##    6.76
```

```
##  elapsed
## 2.964912
```

The OOB method is faster. The time taken to tune for OOB is about three times that of the 5-fold CV method. We expect this to be around five times. b) They choose the same model.

c) Logistic Model performed the worst since a non-linear decision boundary is needed. Single Tree performed better than the logistic model but not the better than the random forest models. It will have non-linear decision boundaries unlike the logistic model but it will be boxed areas since it uses binary splits and not spiral. Random Forest: From part (c) we know they fit the same model. Random forest is the best model out of all we tested. Instead of a single tree, it will use a lot of trees to give a combined result which better match the spiral data.

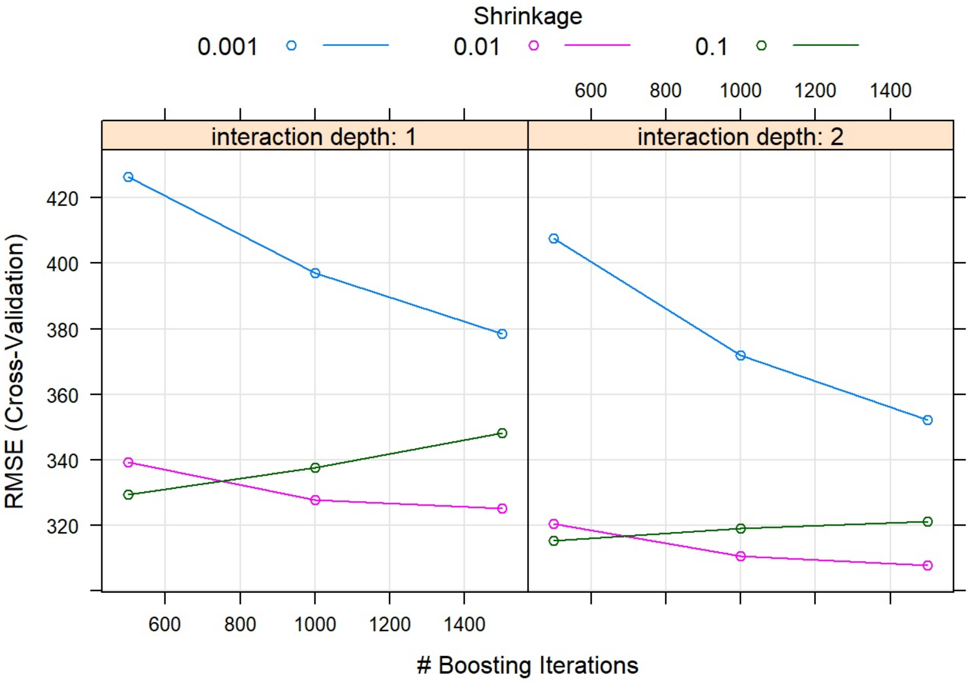
Answer: 5

a.

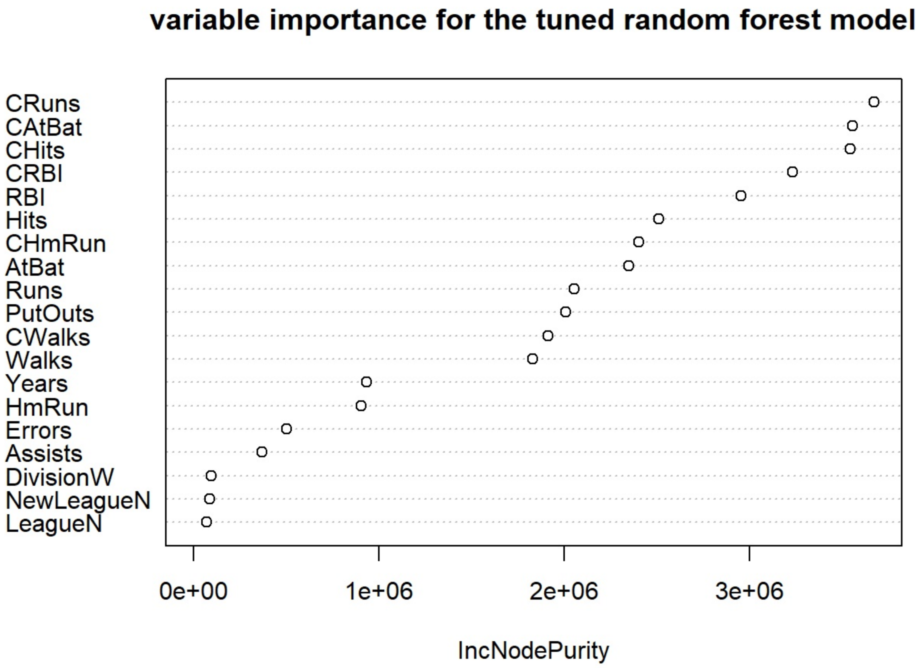
▸
5

1 row | 1-1 of 2 columns

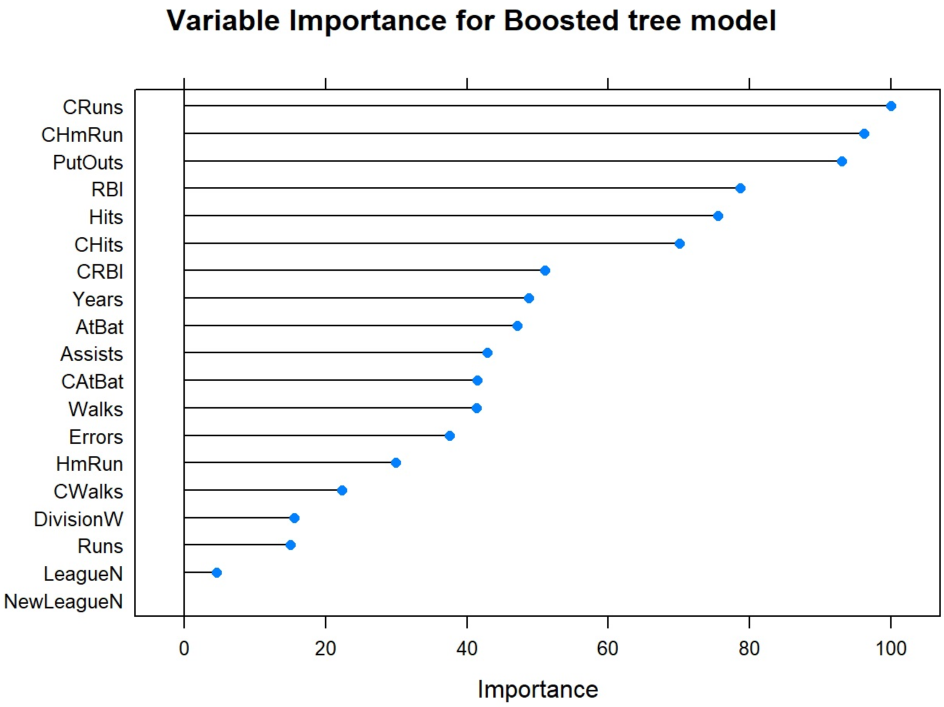
b. The plot that shows the tuning results for the tuning of the boosted tree model:



c. The plot of the variable importance for the tuned random forest model



d. The plot of the variable importance for the tuned boosted tree model



e. According to the random forest model, the three most important predictors are:

```
## [1] "CRuns"  "CAtBat" "CHits"
```

f. According to the boosted tree model, the three most important predictors are:

```
## [1] "CRuns"   "CHmRun"  "PutOuts"
```