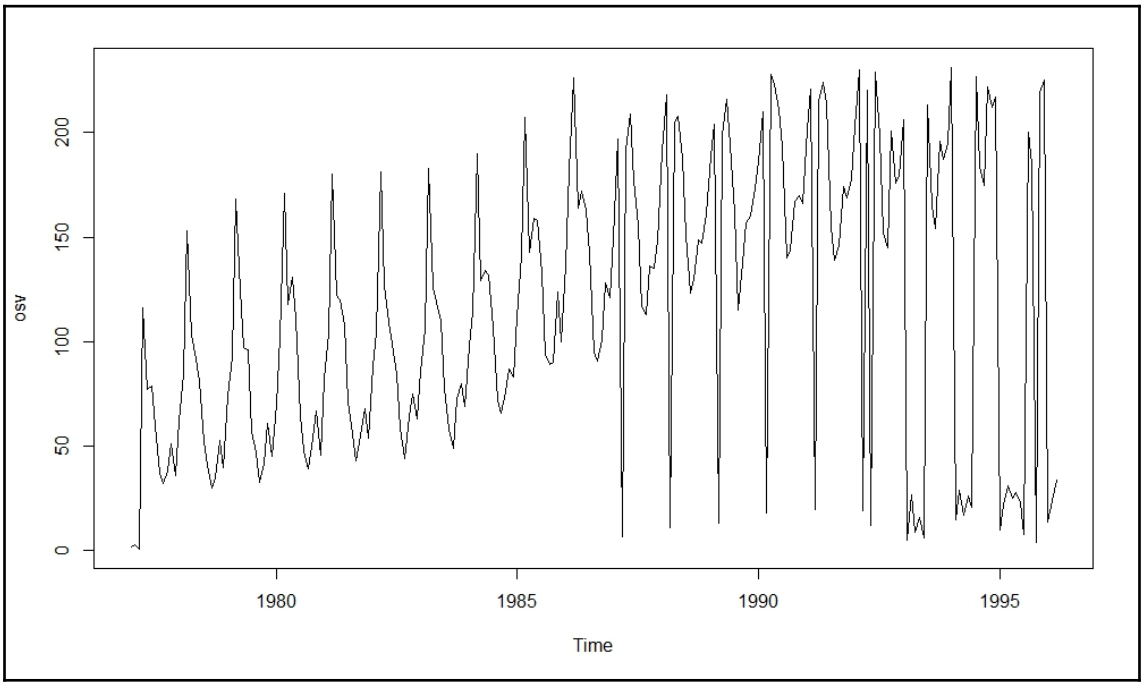
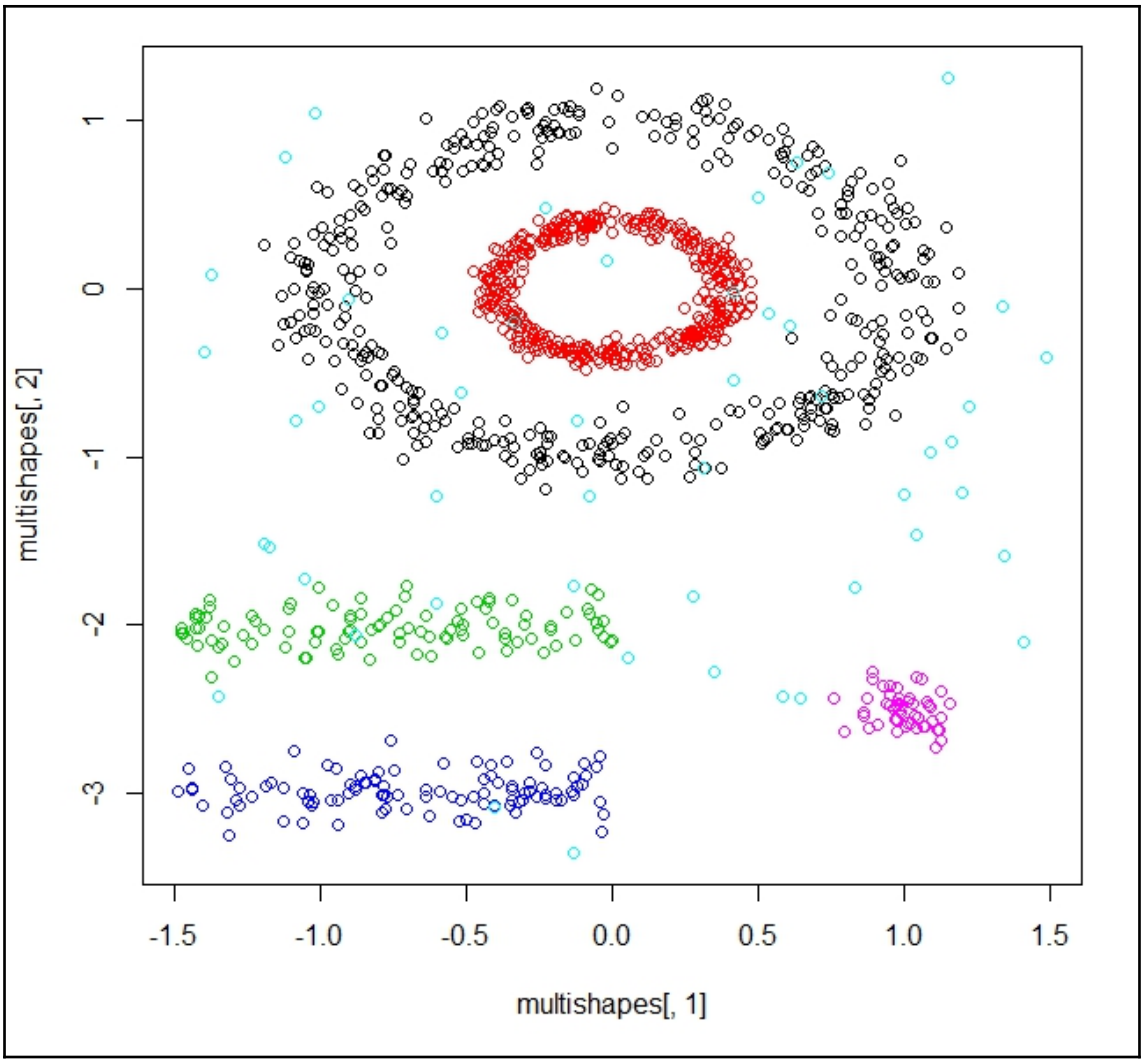
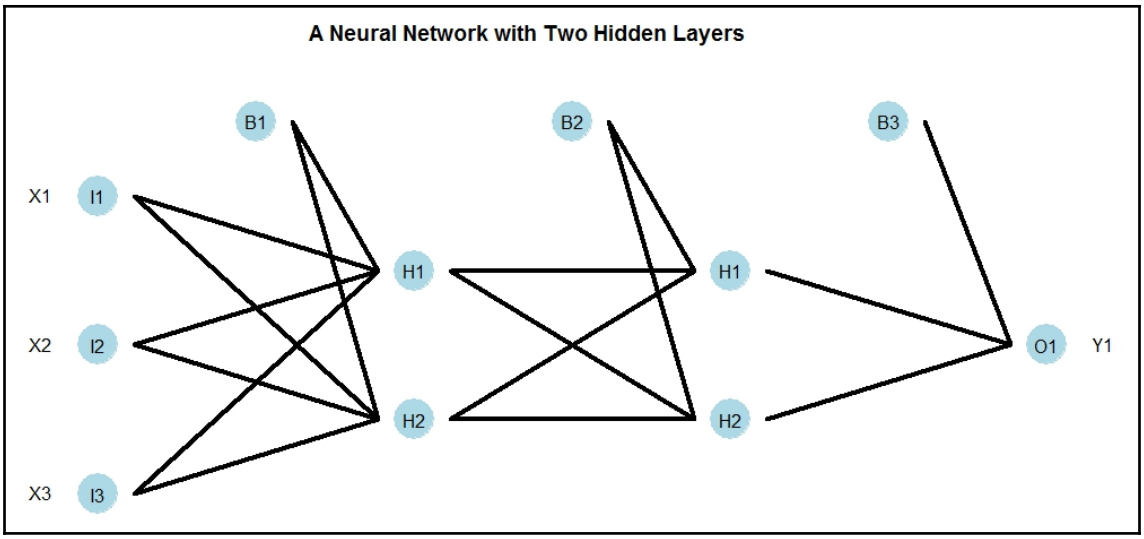
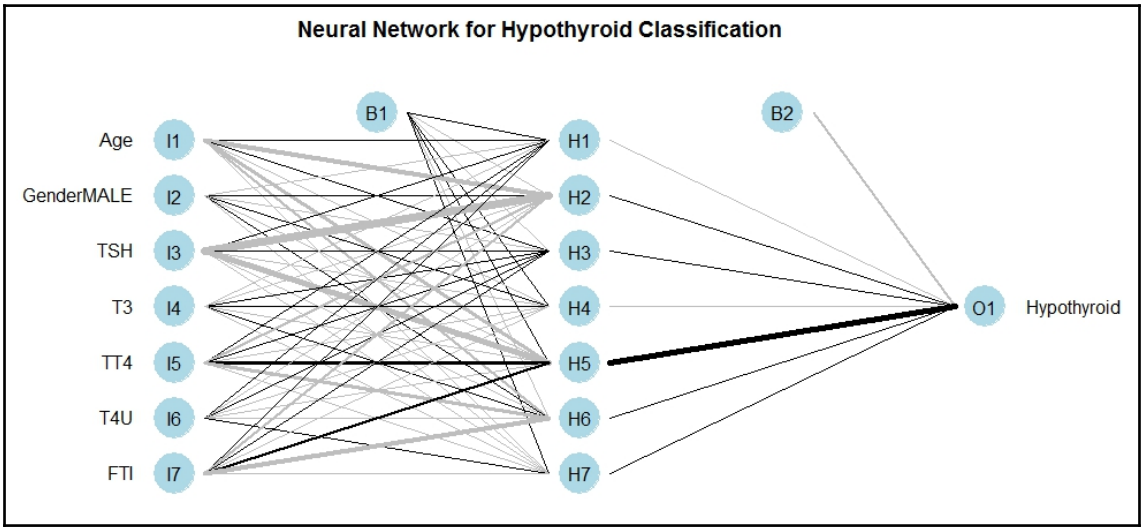
1
Graphics

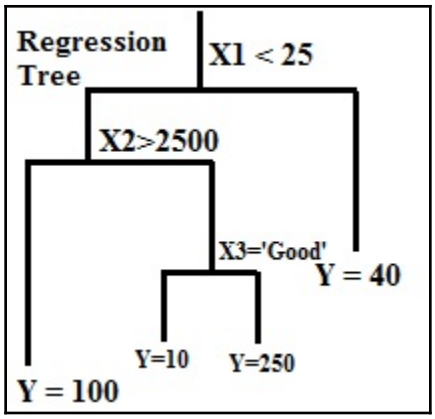
Chapter 1: Introduction to Ensemble Techniques

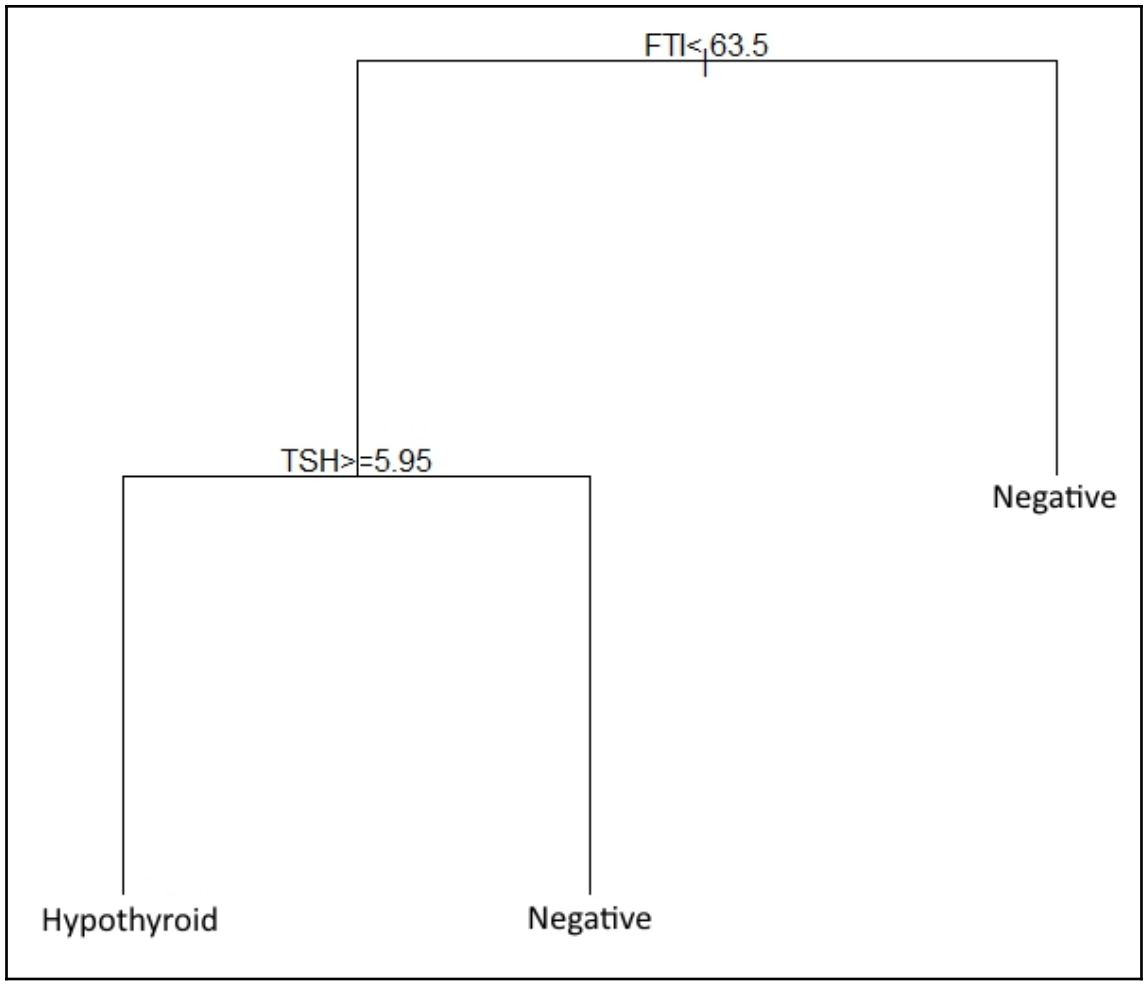



multishapes[, 2]
1
0
-1
-2
-3
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
multishapes[, 1]


A Neural Network with Two Hidden Layers
B1
B2
B3
X1
I1
X2
I2
X3
I3
H1
H2
H1
H2
O1
Y1


Neural Network for Hypothyroid Classification
B1
B2
Age
I1
GenderMALE
I2
TSH
I3
T3
I4
TT4
I5
T4U
I6
FTI
I7
H1
H2
H3
H4
H5
H6
H7
O1
Hypothyroid

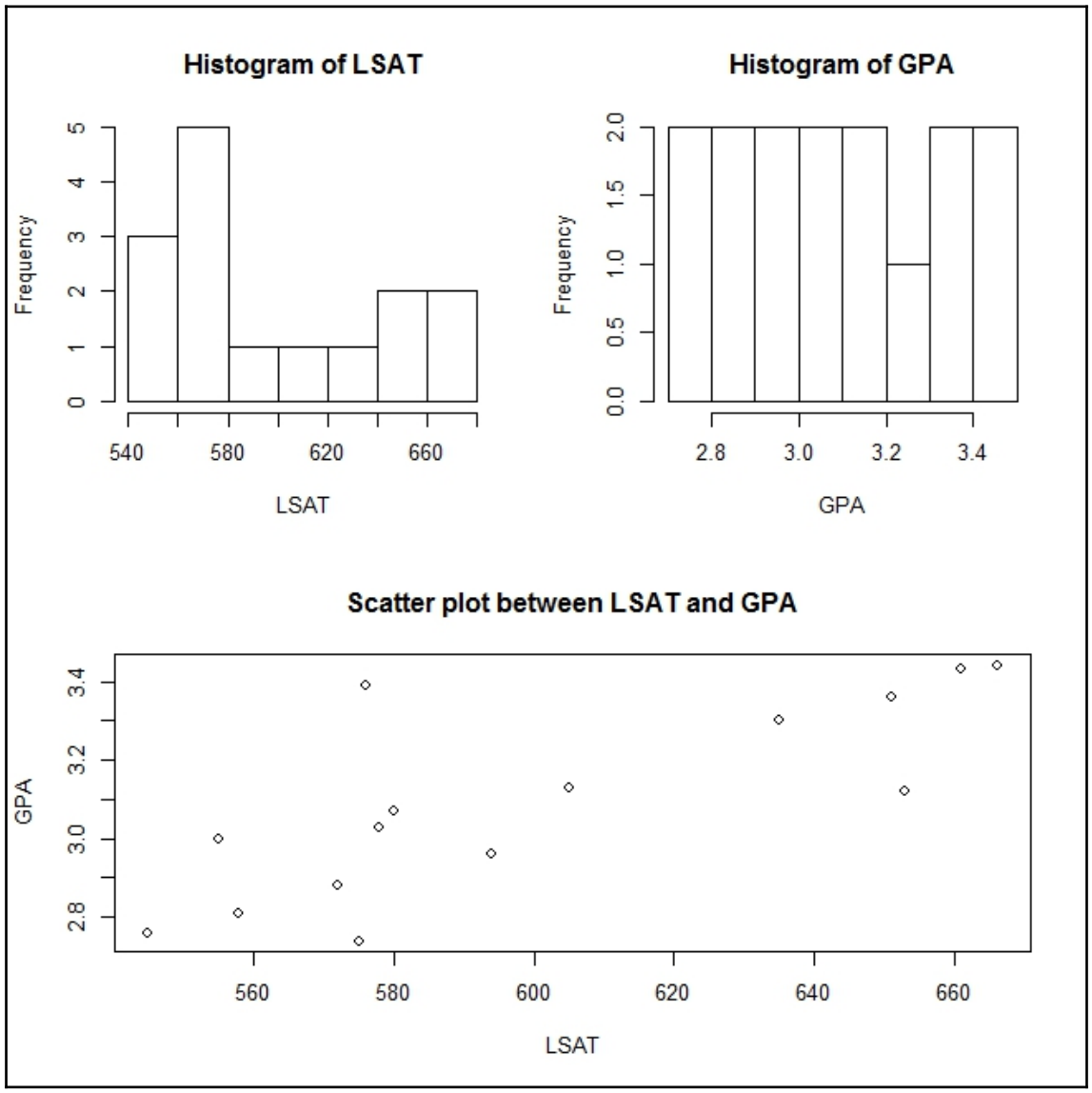

Regression Tree
X1 < 25
X2>2500
X3='Good'
Y = 40
Y=10
Y=250
Y = 100

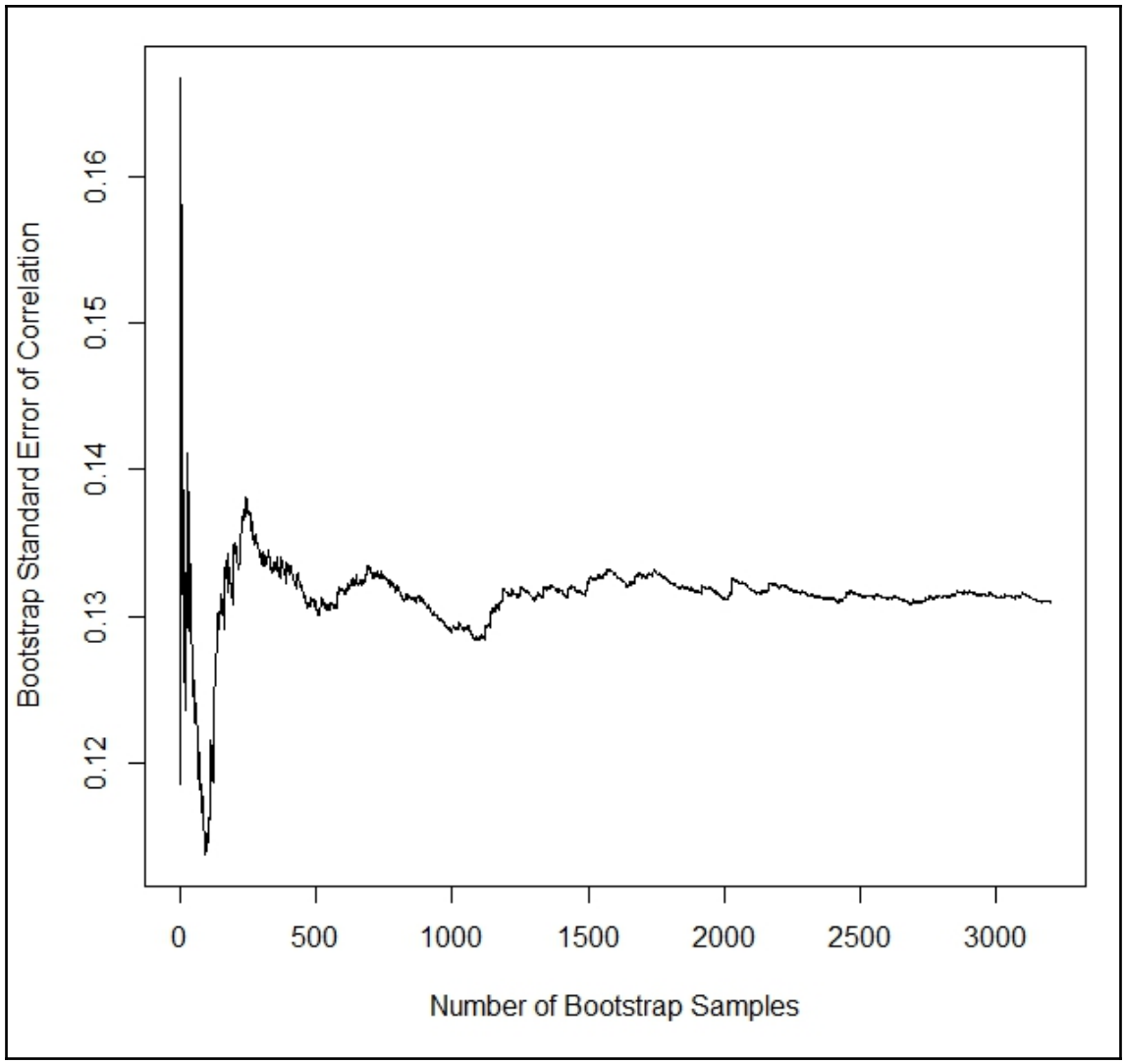


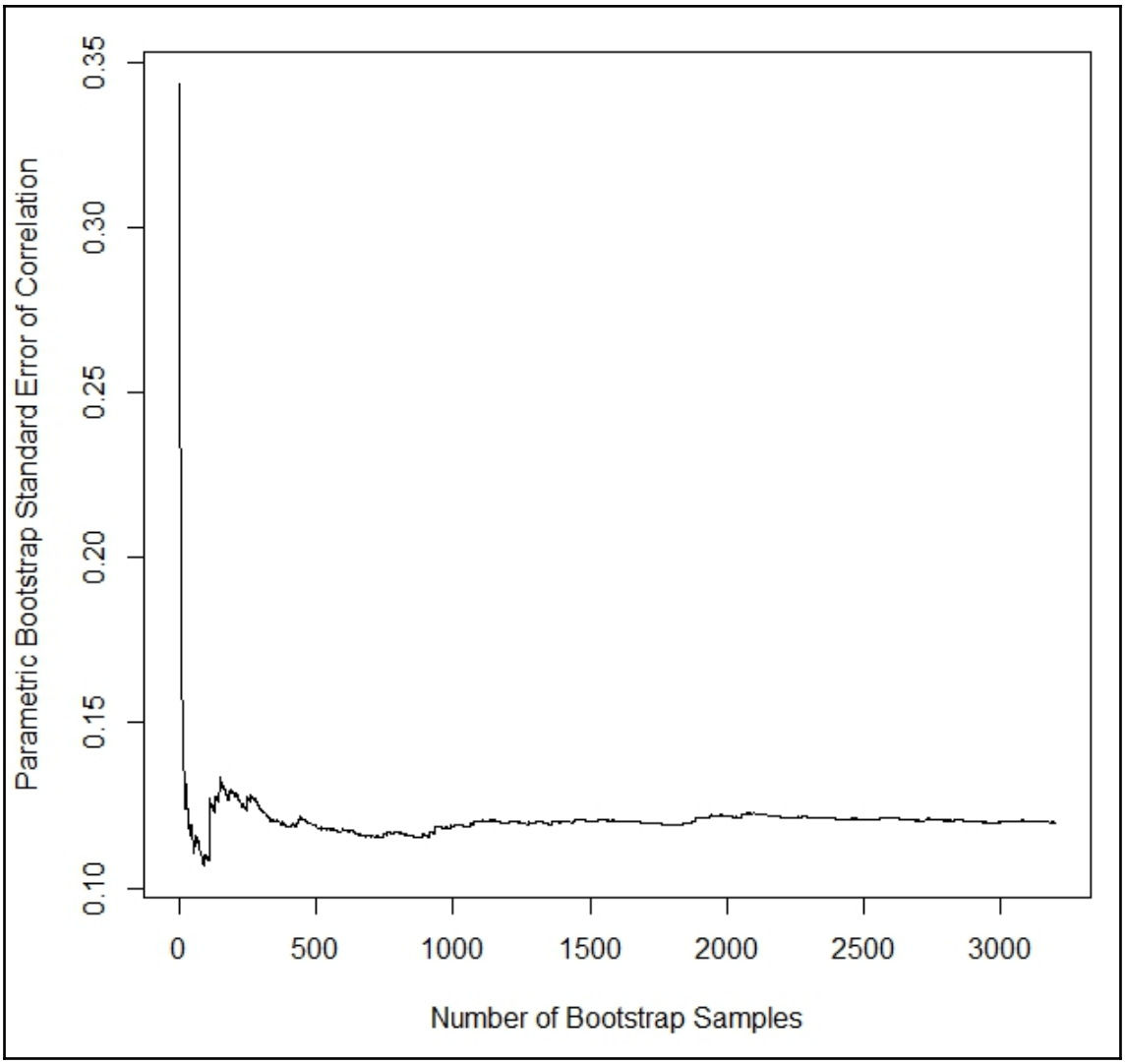
Dataset/ Model	Hypothyroid	Waveform	German	Iris	Pima Indian Diabetes
Neural Network	98.27%	88.40%	72.52%	100.00%	67.32%
Logistic Regression	97.33%	88.73%	75.72%	100.00%	75.10%
Naïve Bayes	97.33%	86.01%	80.83%	100.00%	78.21%
Decision Tree	98.74%	84.35%	70.61%	100.00%	75.88%
SVM	98.43%	91.71%	75.40%	100.00%	76.65%

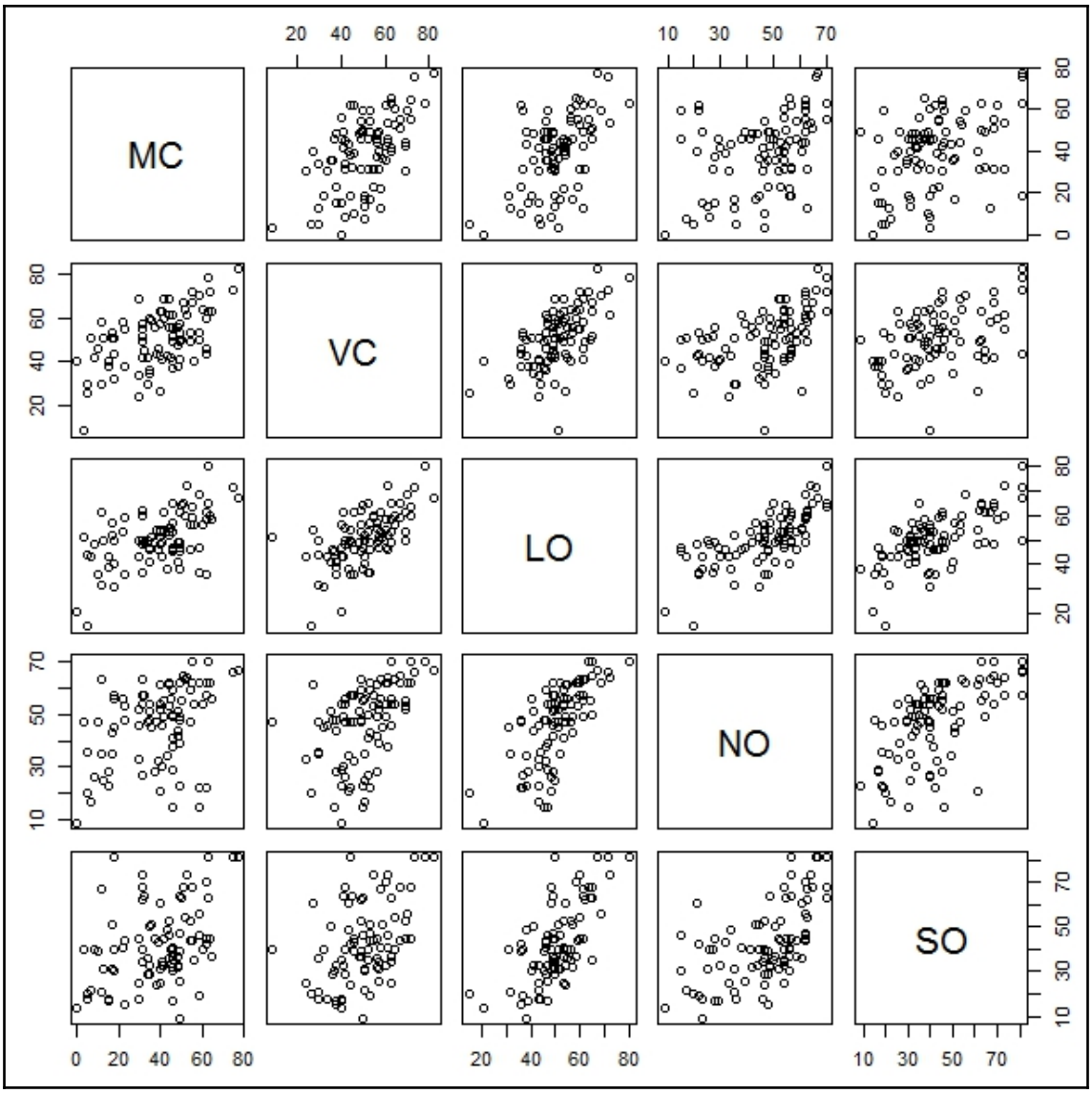
```
1 package is needed for this model and is not installed. (frbs). Would you like to
try to install it now?
1: yes
2: no

selection:
```

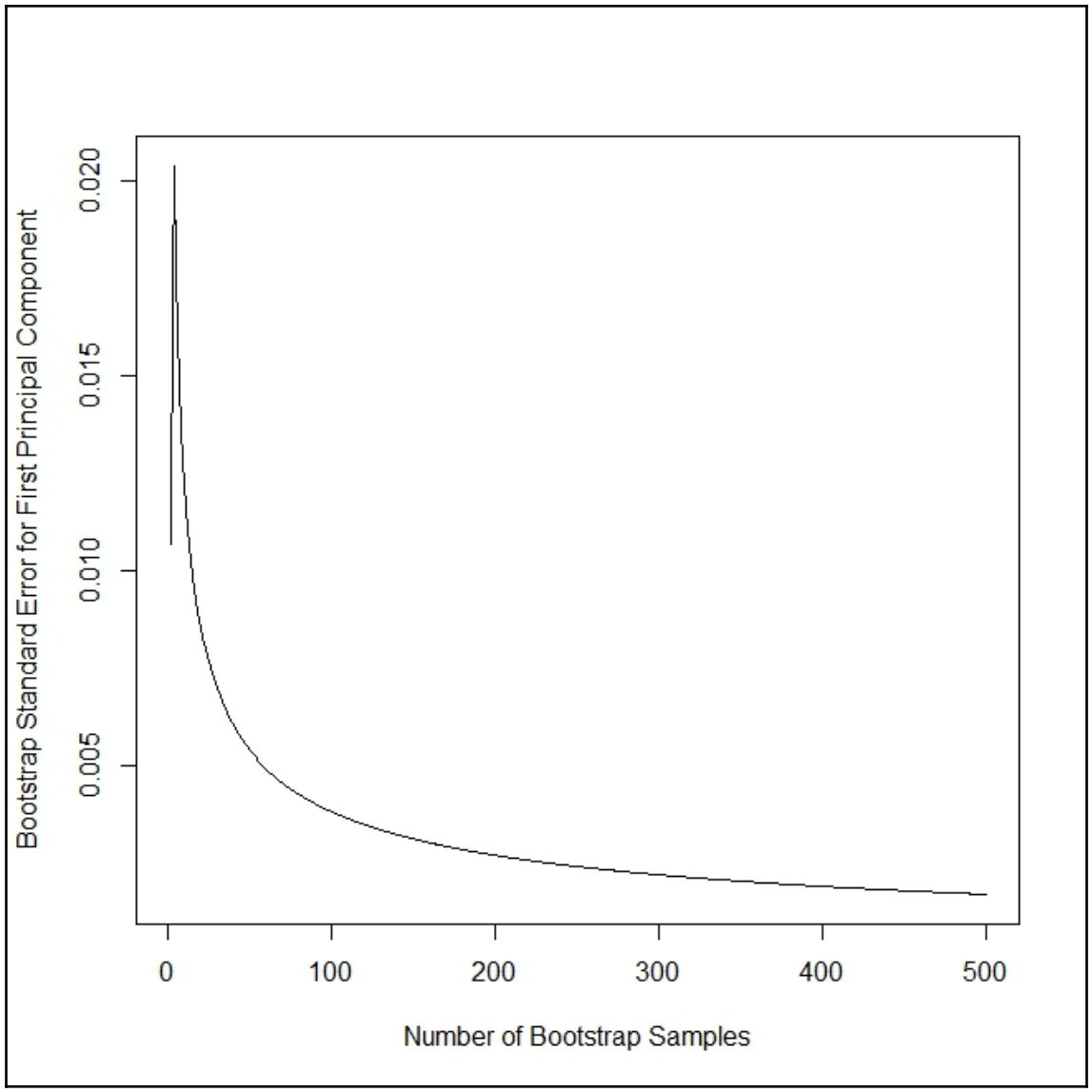
Chapter 2: Bootstrapping



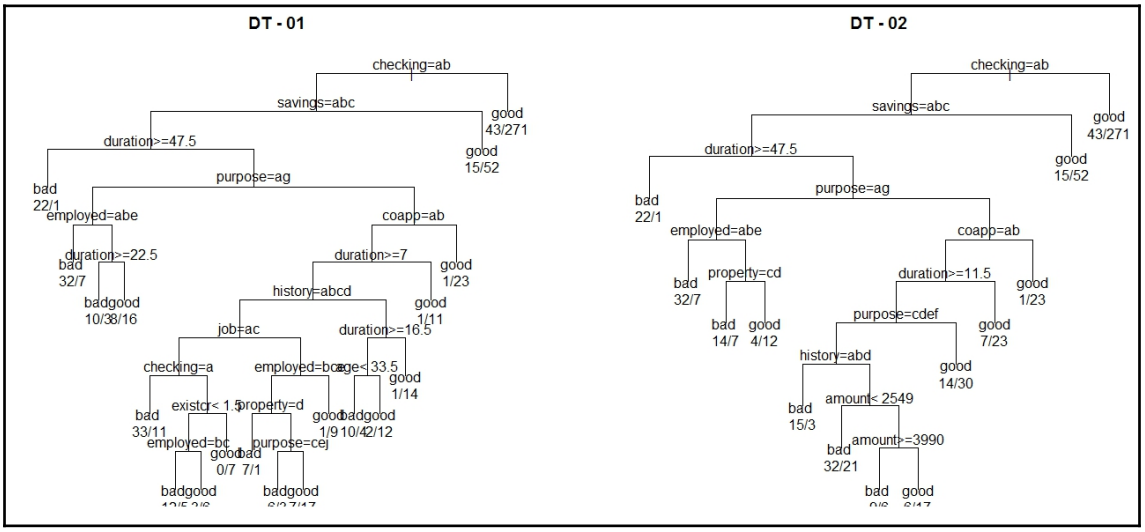

Bootstrap Standard Error of Correlation
0.12
0.13
0.14
0.15
0.16
0
500
1000
1500
2000
2500
3000
Number of Bootstrap Samples

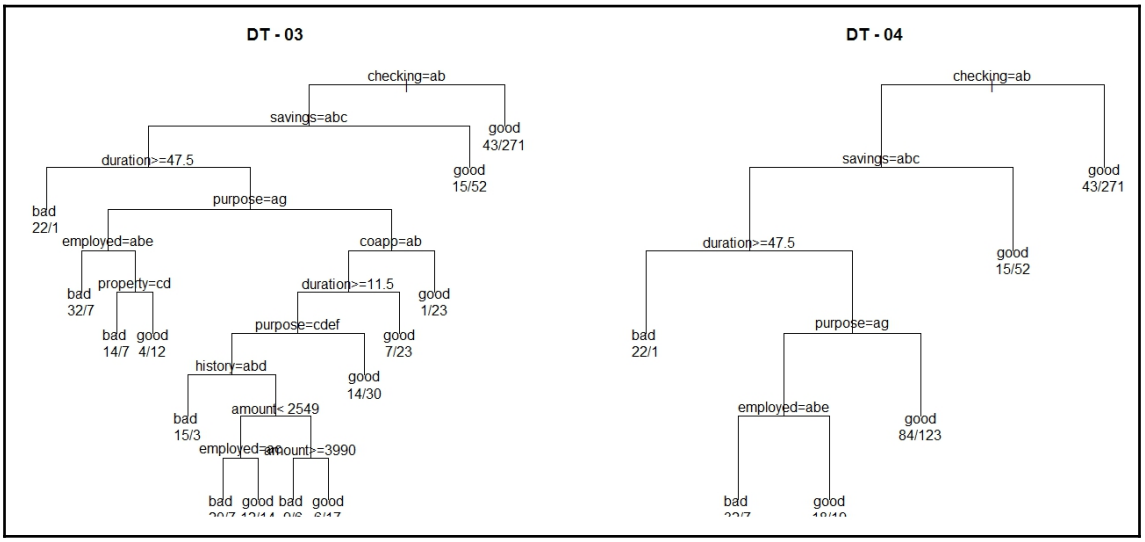

Parametric Bootstrap Standard Error of Correlation
0.10
0.15
0.20
0.25
0.30
0.35
0
500
1000
1500
2000
2500
3000
Number of Bootstrap Samples

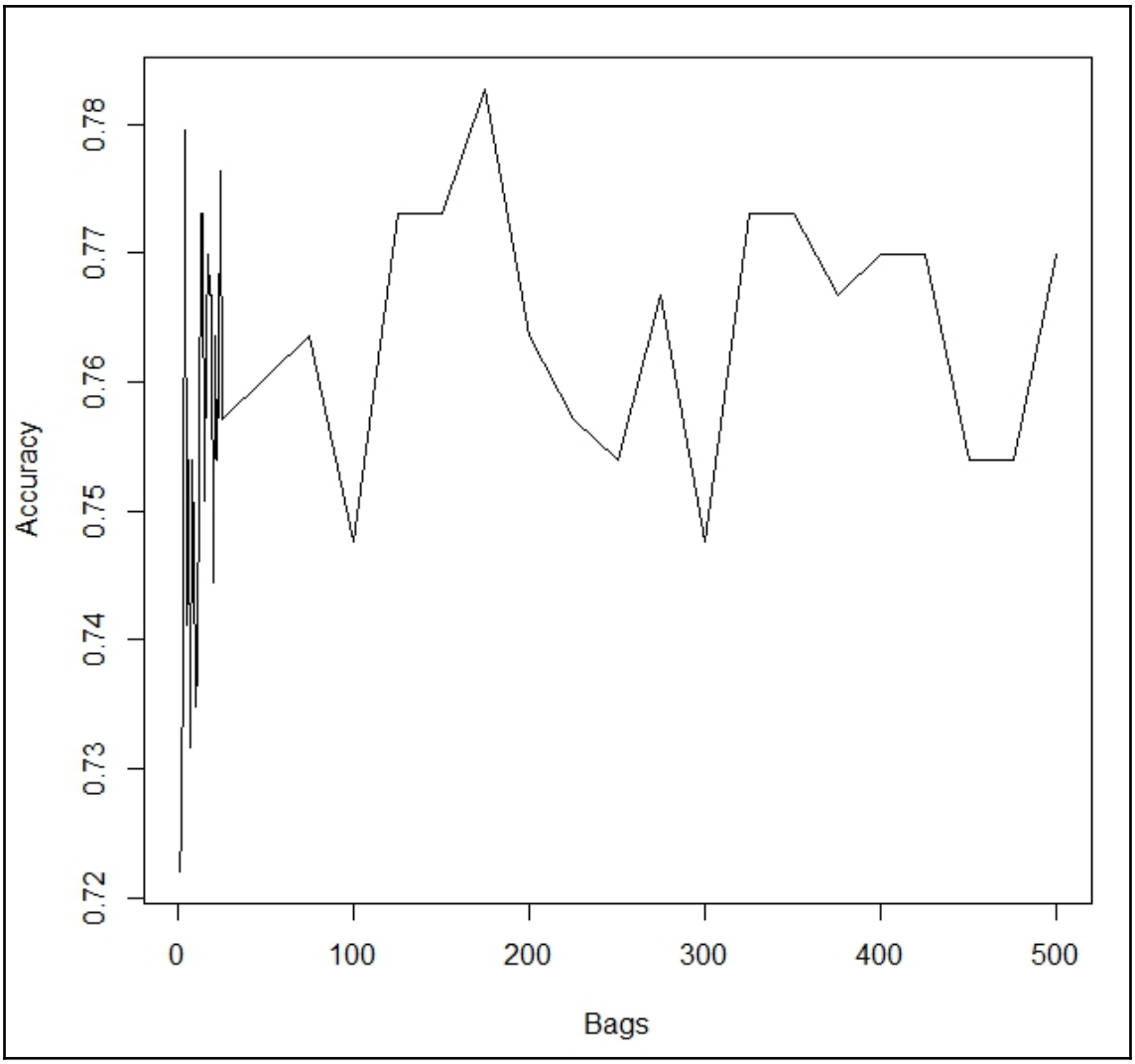

MC
VC
LO
NO
SO
20 40 60 80
10 30 50 70
0 20 40 60 80
20 40 60 80
10 30 50 70
10 30 50 70
20 40 60 80
10 30 50 70
0 20 40 60 80
20 40 60 80
10 30 50 70

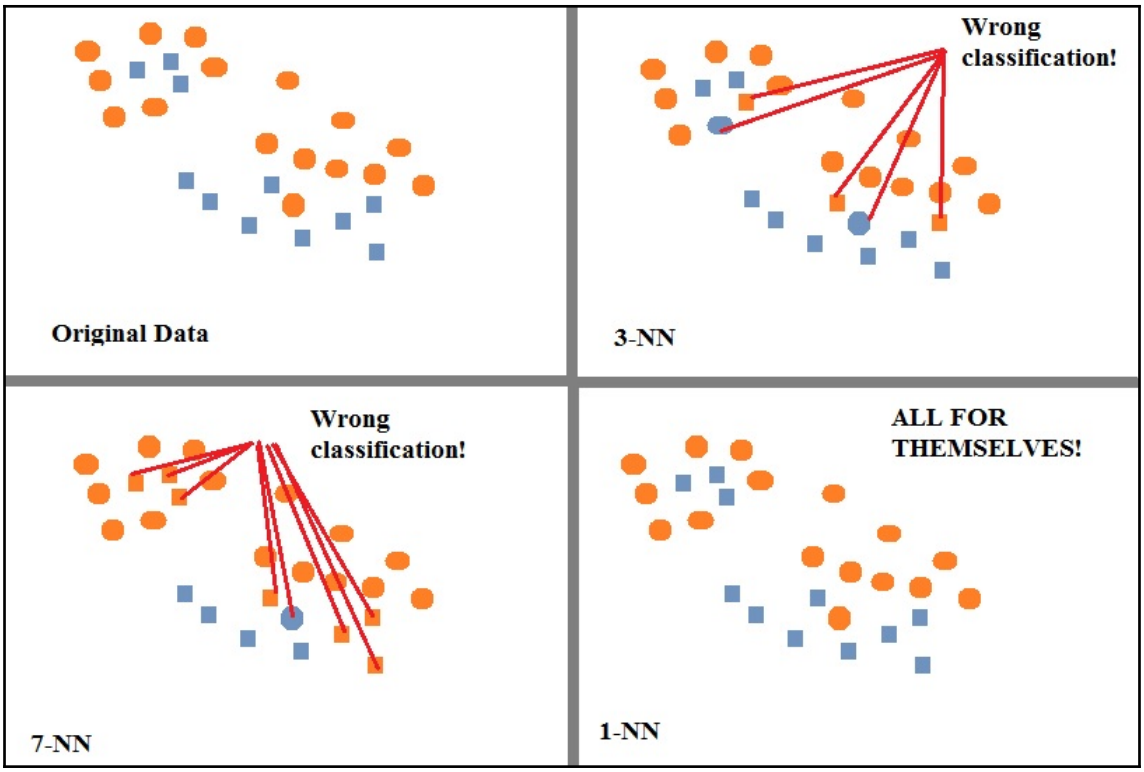

Bootstrap Standard Error for First Principal Component
0.005
0.010
0.015
0.020
0
100
200
300
400
500
Number of Bootstrap Samples

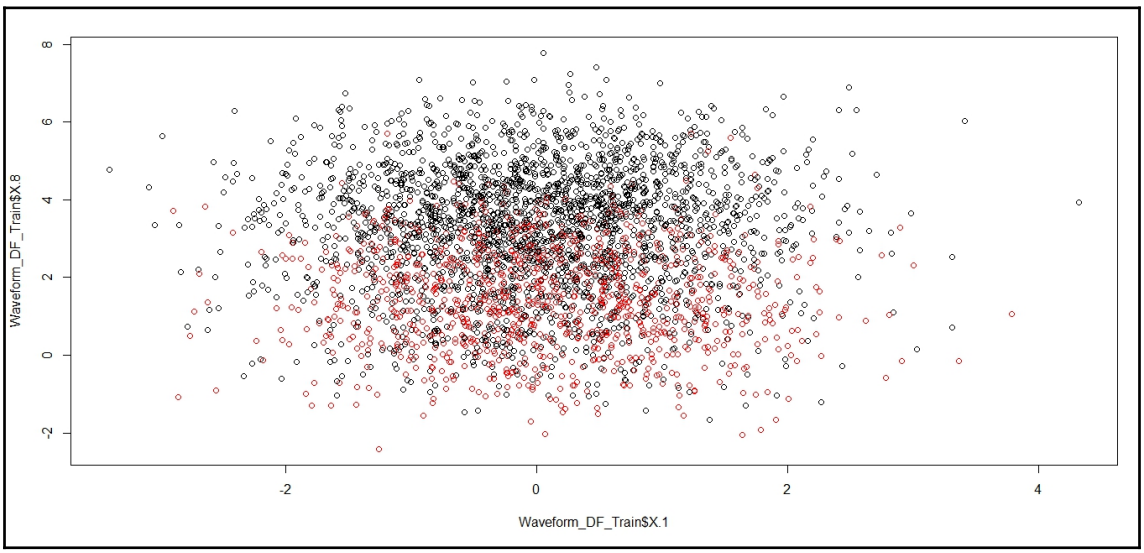
Chapter 3: Bagging

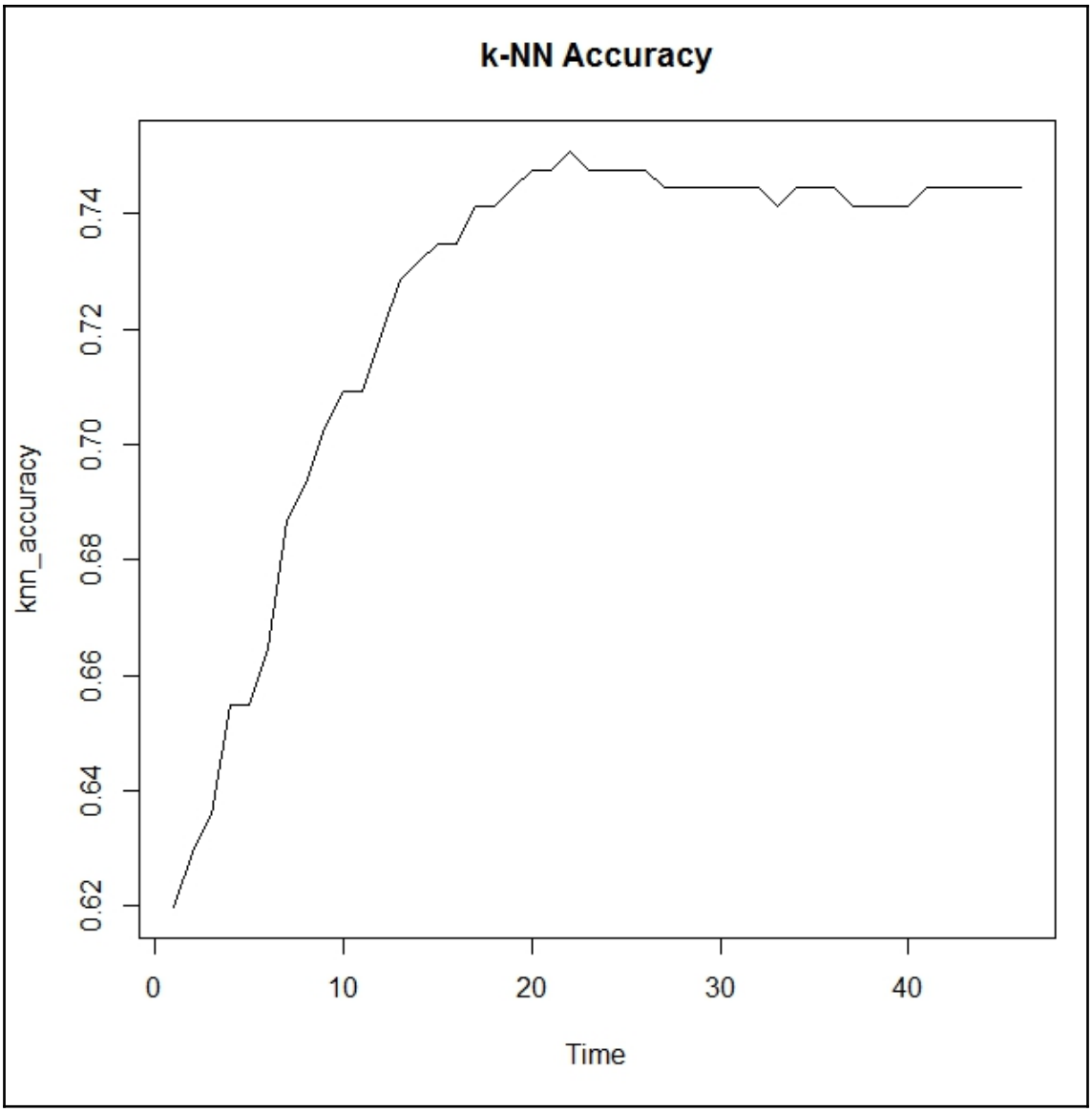




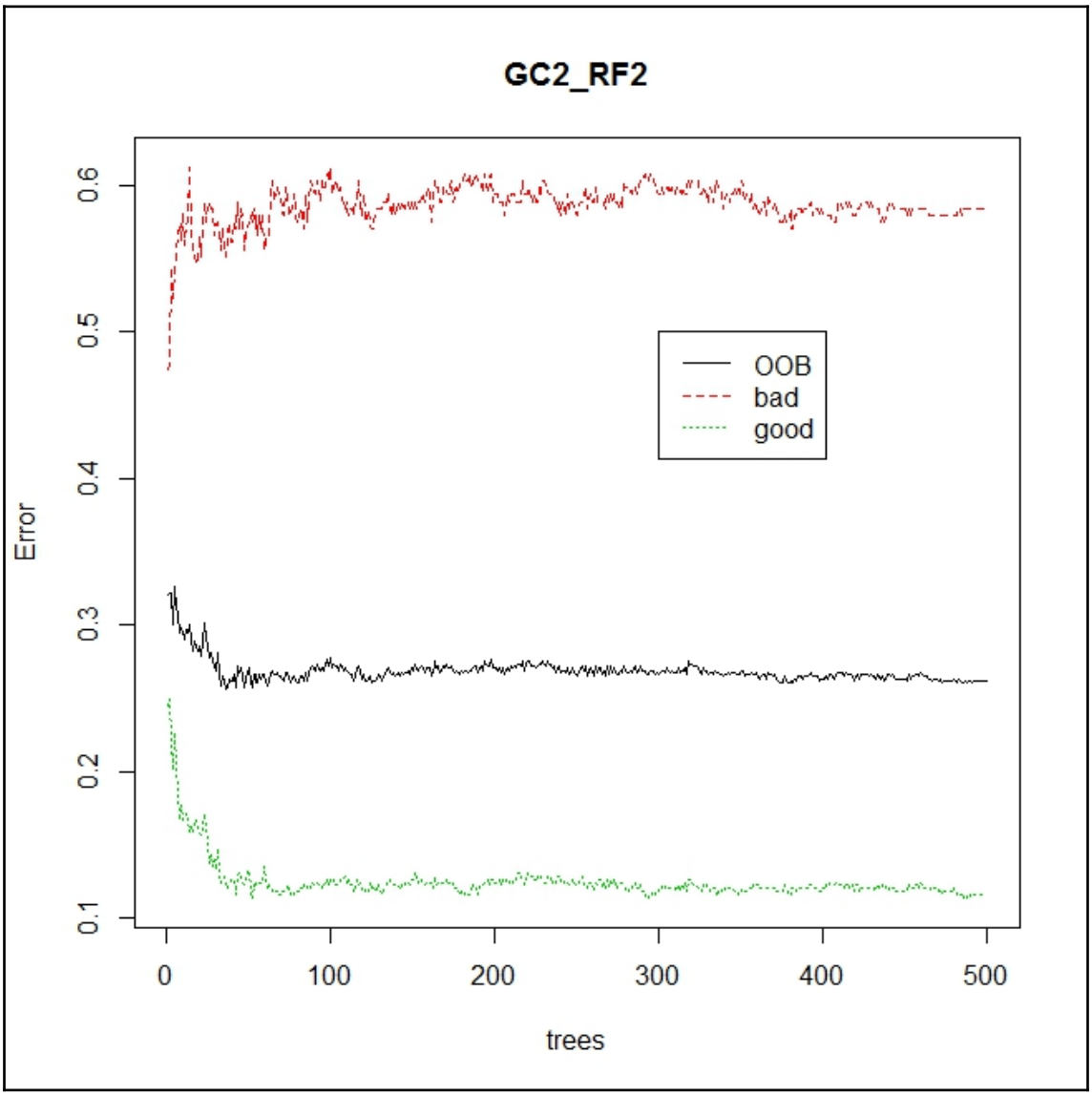

Accuracy
0.72
0.73
0.74
0.75
0.76
0.77
0.78
0
100
200
300
400
500
Bags


Original Data
Wrong classification!
3-NN
Wrong classification!
7-NN
ALL FOR THEMSELVES!
1-NN

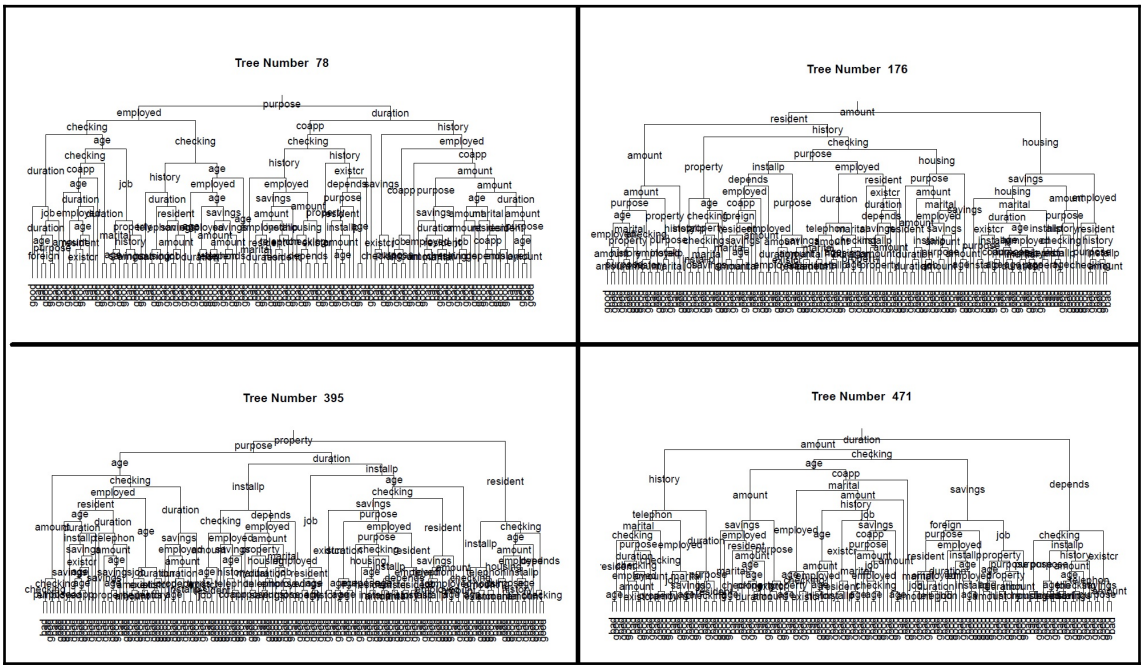

Waveform_DF_Train$X.8
Waveform_DF_Train$X.1
-2
0
2
4
6
8

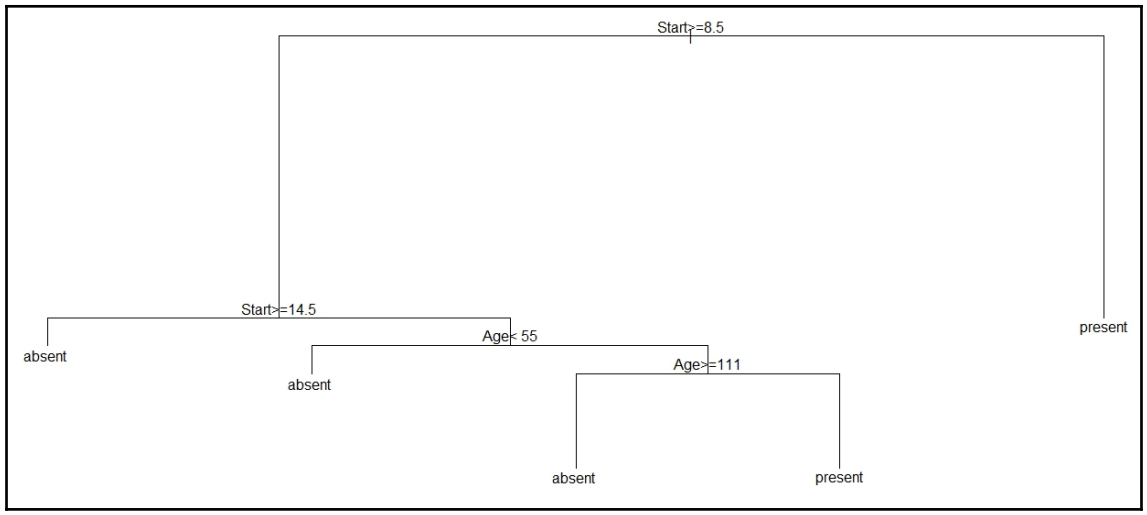

k-NN Accuracy
knn_accuracy
0.62
0.64
0.66
0.68
0.70
0.72
0.74
0
10
20
30
40
Time

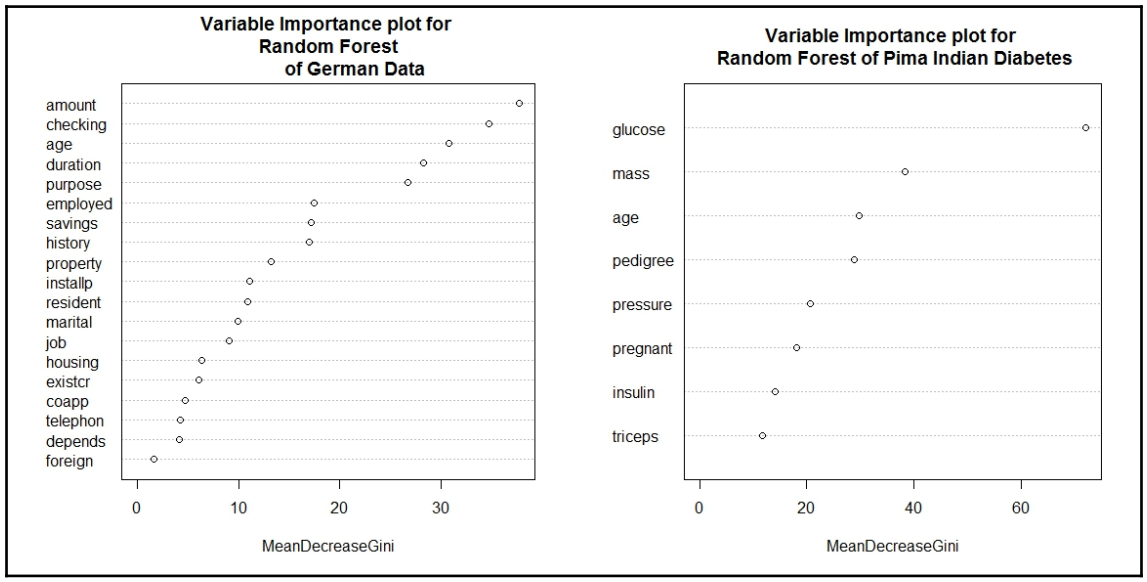
Chapter 4: Random Forests

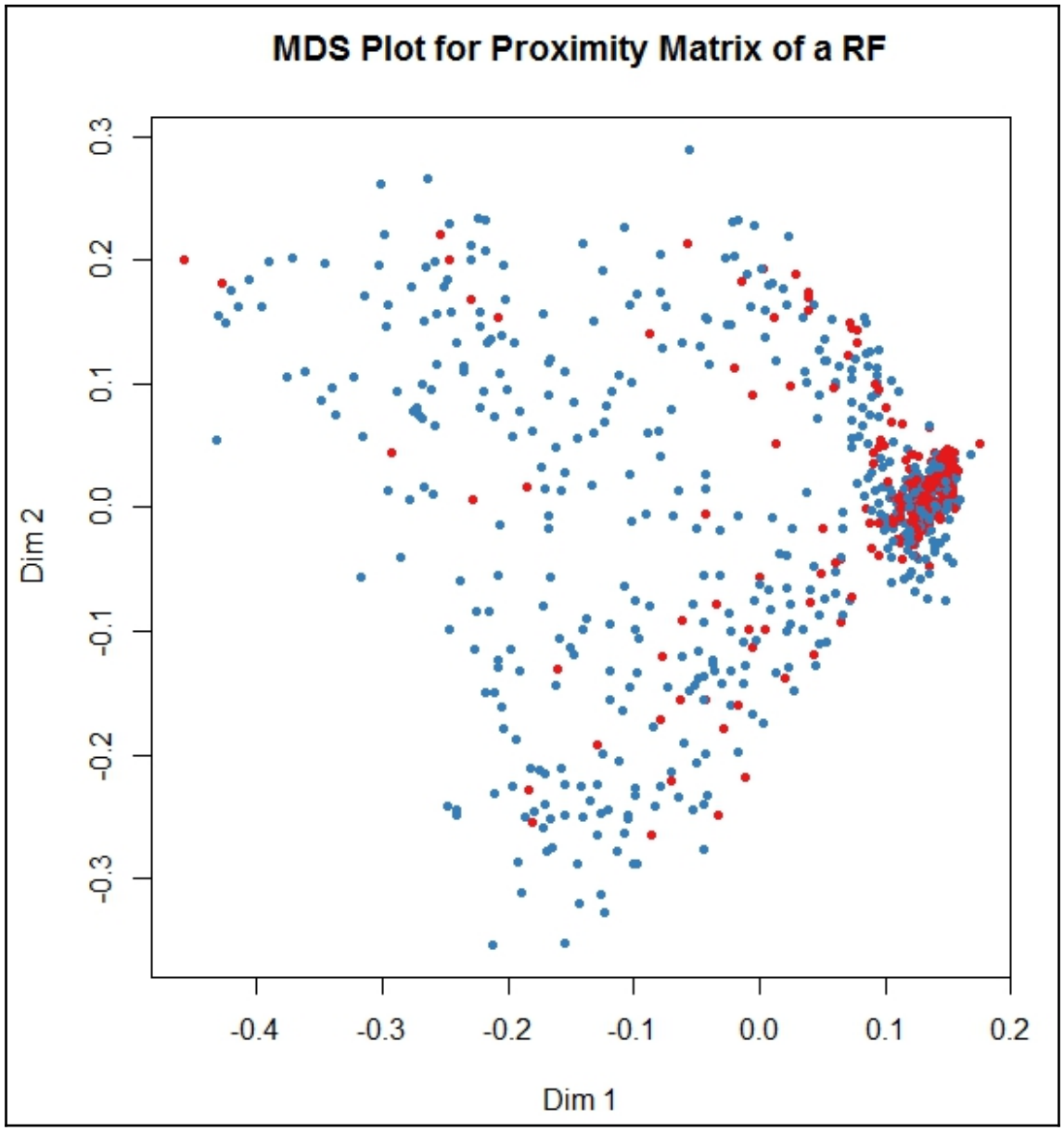


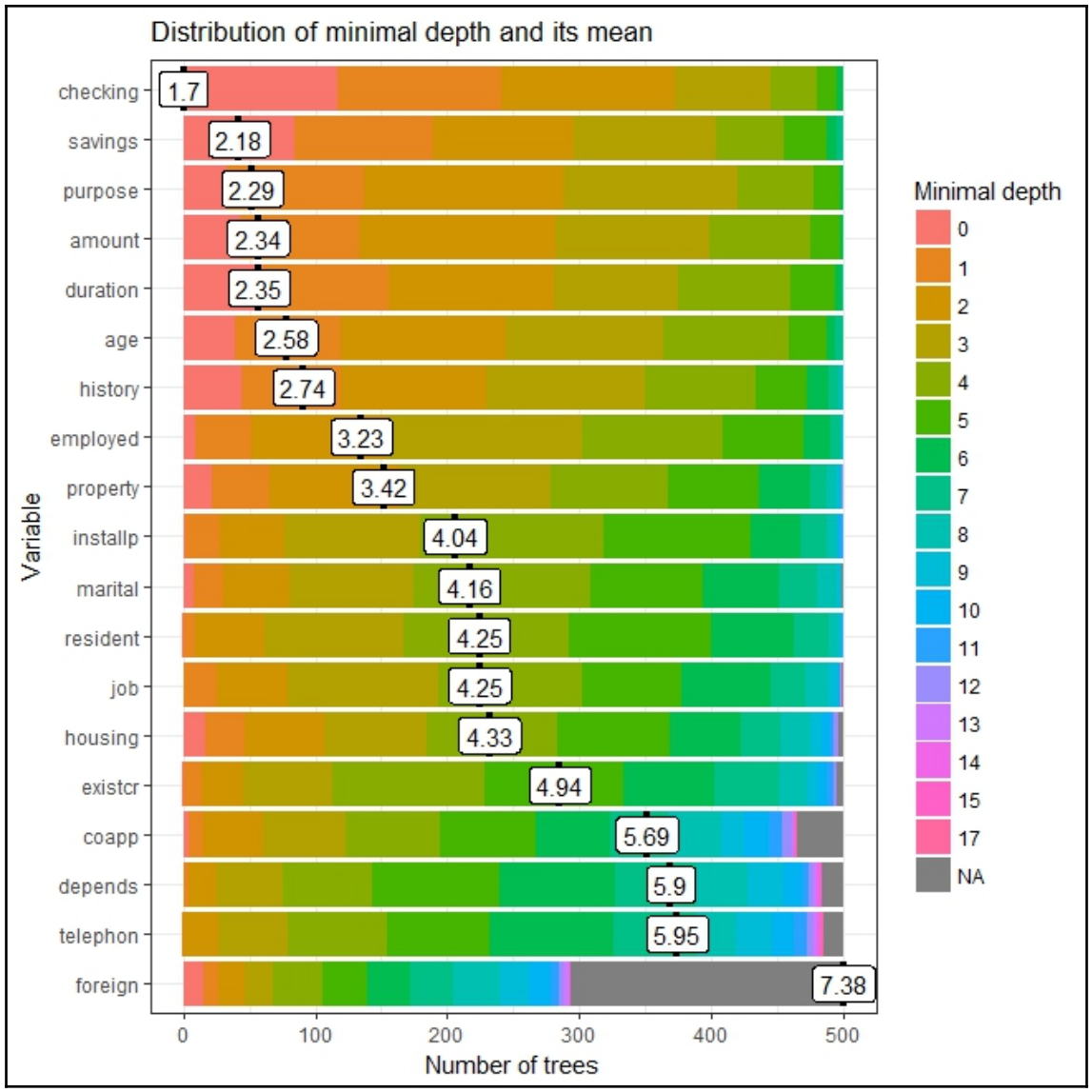
```
plot_RF <- function(RF){
  n <- RF$ntree
  for(i in 1:n){
    tt <- getTree(RF,i,labelVar = TRUE)
    dt <- to.dendrogram(tt)
    plot(dt,center=TRUE,edgePar=list(t.cex=1,p.col=NA,p.lty=0),
         yaxt='n',horiz=TRUE)
    print(i)
  }
}
```



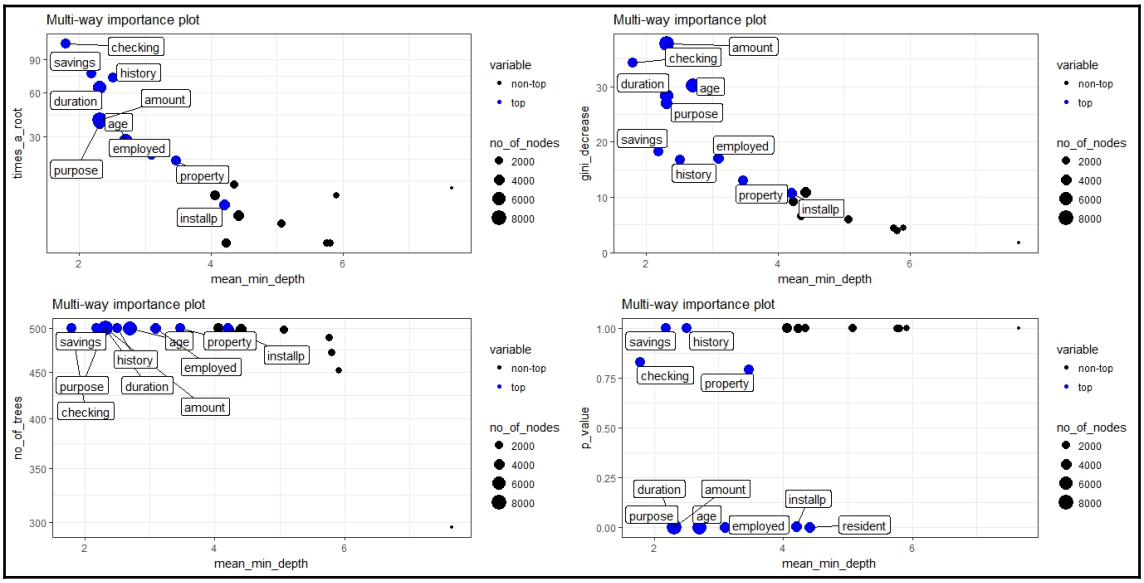

Start>=8.5
Start>=14.5
Age< 55
Age>=111
absent
absent
absent
present
present

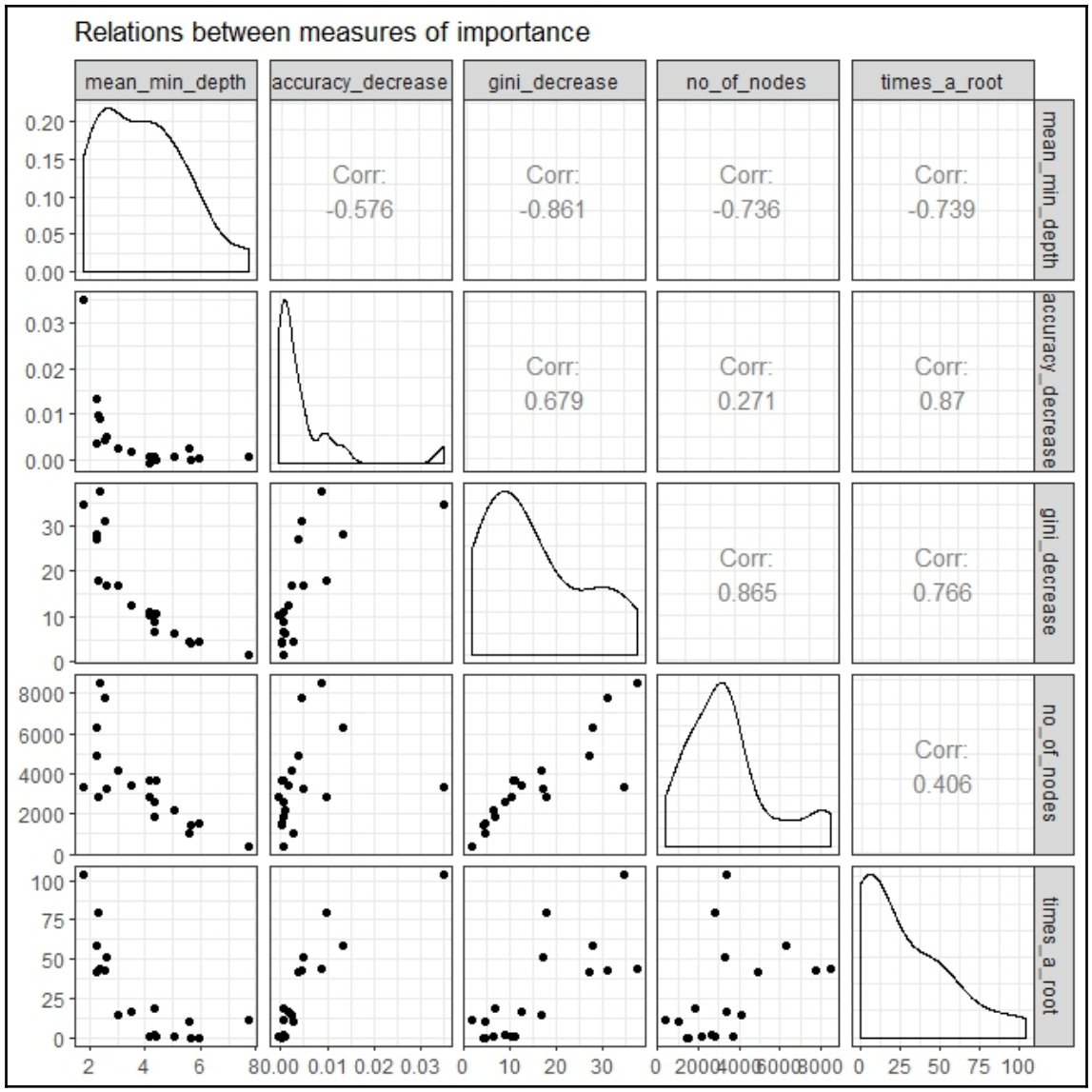

Variable Importance plot for
Random Forest
of German Data
amount
checking
age
duration
purpose
employed
savings
history
property
installp
resident
marital
job
housing
existcr
coapp
telephon
depends
foreign
0
10
20
30
MeanDecreaseGini
Variable Importance plot for
Random Forest of Pima Indian Diabetes
glucose
mass
age
pedigree
pressure
pregnant
insulin
triceps
0
20
40
60
MeanDecreaseGini

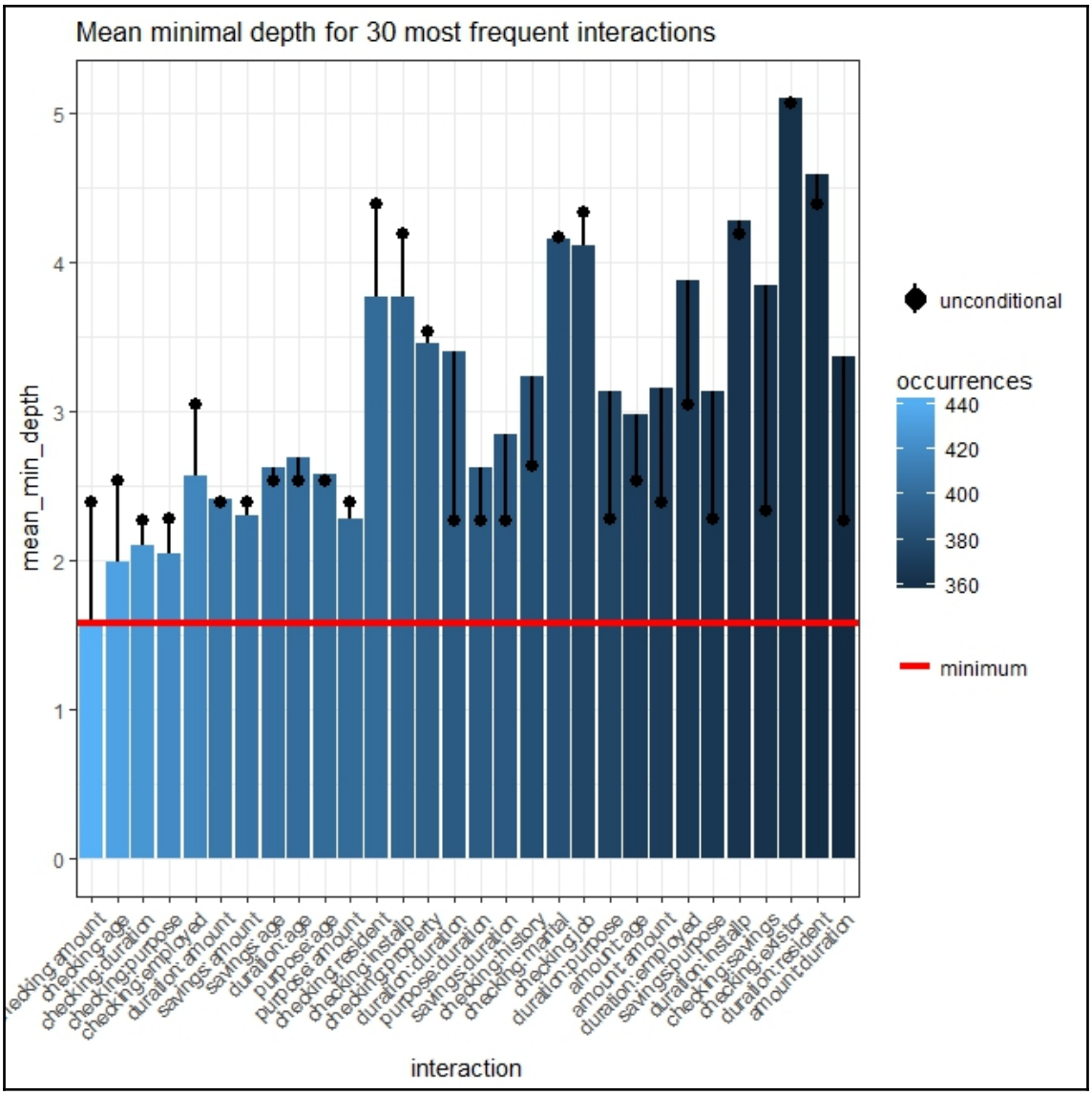

MDS Plot for Proximity Matrix of a RF
Dim 2
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
Dim 1

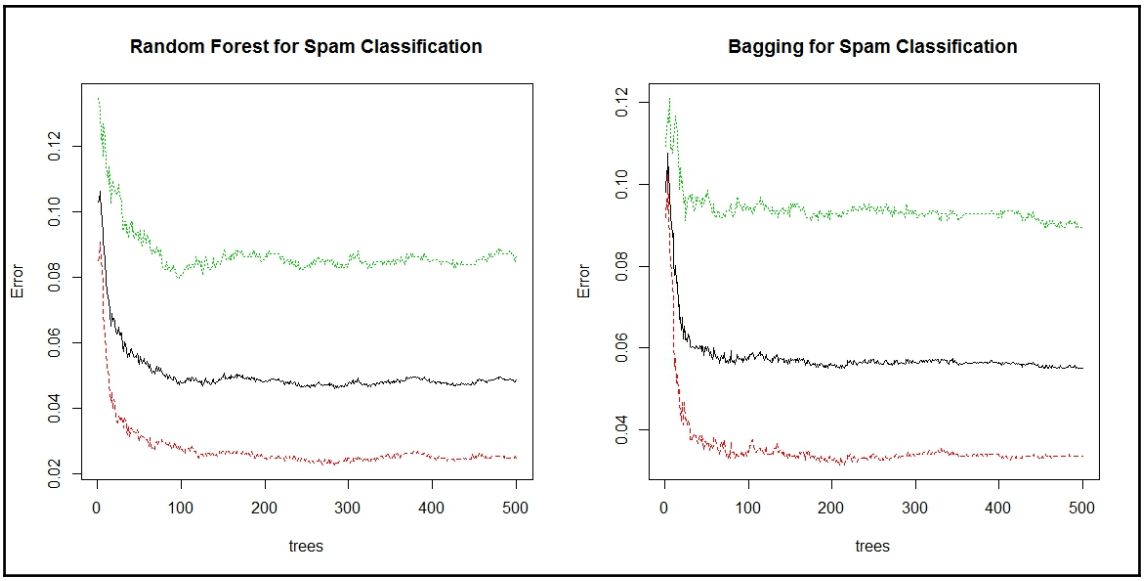

Distribution of minimal depth and its mean
checking
savings
purpose
amount
duration
age
history
employed
property
installp
marital
resident
job
housing
existcr
coapp
depends
telephon
foreign
1.7
2.18
2.29
2.34
2.35
2.58
2.74
3.23
3.42
4.04
4.16
4.25
4.25
4.33
4.94
5.69
5.9
5.95
7.38
Variable
0
100
200
300
400
500
Number of trees
Minimal depth
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
17
NA

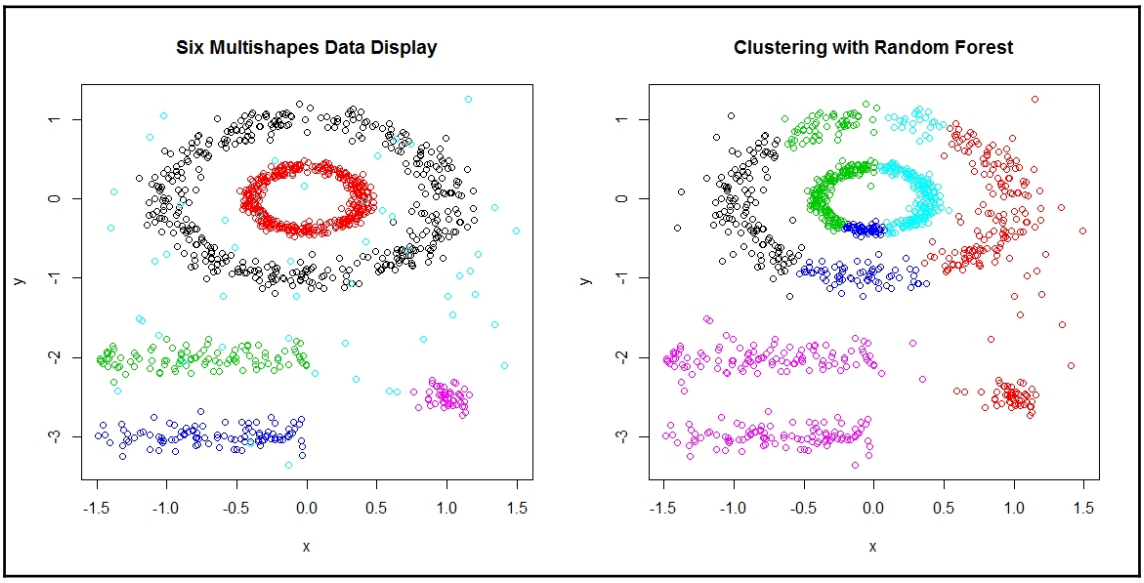
```
> GC2_RF4_VIM
   variable mean_min_depth no_of_nodes accuracy_decrease gini_decrease no_of_trees times_a_root  p_value
1       age           2.5        7750         0.00455         31.2         500          43  0.0e+00
2    amount           2.4        8512         0.00890         37.7         500          44  0.0e+00
3  checking           1.8        3347         0.03490         34.6         500         104  9.8e-01
4     coapp           5.6        1009         0.00272          4.5         456          10  1.0e+00
5   depends           5.6        1461         0.00022          4.3         486           0  1.0e+00
6  duration           2.3        6319         0.01343         28.1         500          59  0.0e+00
7  employed           3.0        4147         0.00254         16.8         500          14  2.3e-31
8   existcr           5.1        2186         0.00084          6.2         499           1  1.0e+00
9   foreign           7.8         403         0.00067          1.6         300          11  1.0e+00
10  history           2.6        3290         0.00499         17.0         500          51  1.0e+00
11  housing           4.3        1845         0.00071          6.6         496          19  1.0e+00
12 installp           4.2        3699         0.00065         11.1         500           1  2.4e-05
13      job           4.3        2612         0.00061          9.0         500           2  1.0e+00
14  marital           4.2        2847        -0.00059         10.3         499           1  1.0e+00
15 property           3.5        3401         0.00180         12.6         500          17  8.6e-01
16  purpose           2.3        4930         0.00372         27.3         500          42 4.3e-129
17 resident           4.4        3684         0.00024         10.7         500           1  6.9e-05
18  savings           2.3        2820         0.00995         17.9         500          80  1.0e+00
19 telephon           5.9        1539         0.00034          4.6         487           0  1.0e+00
```



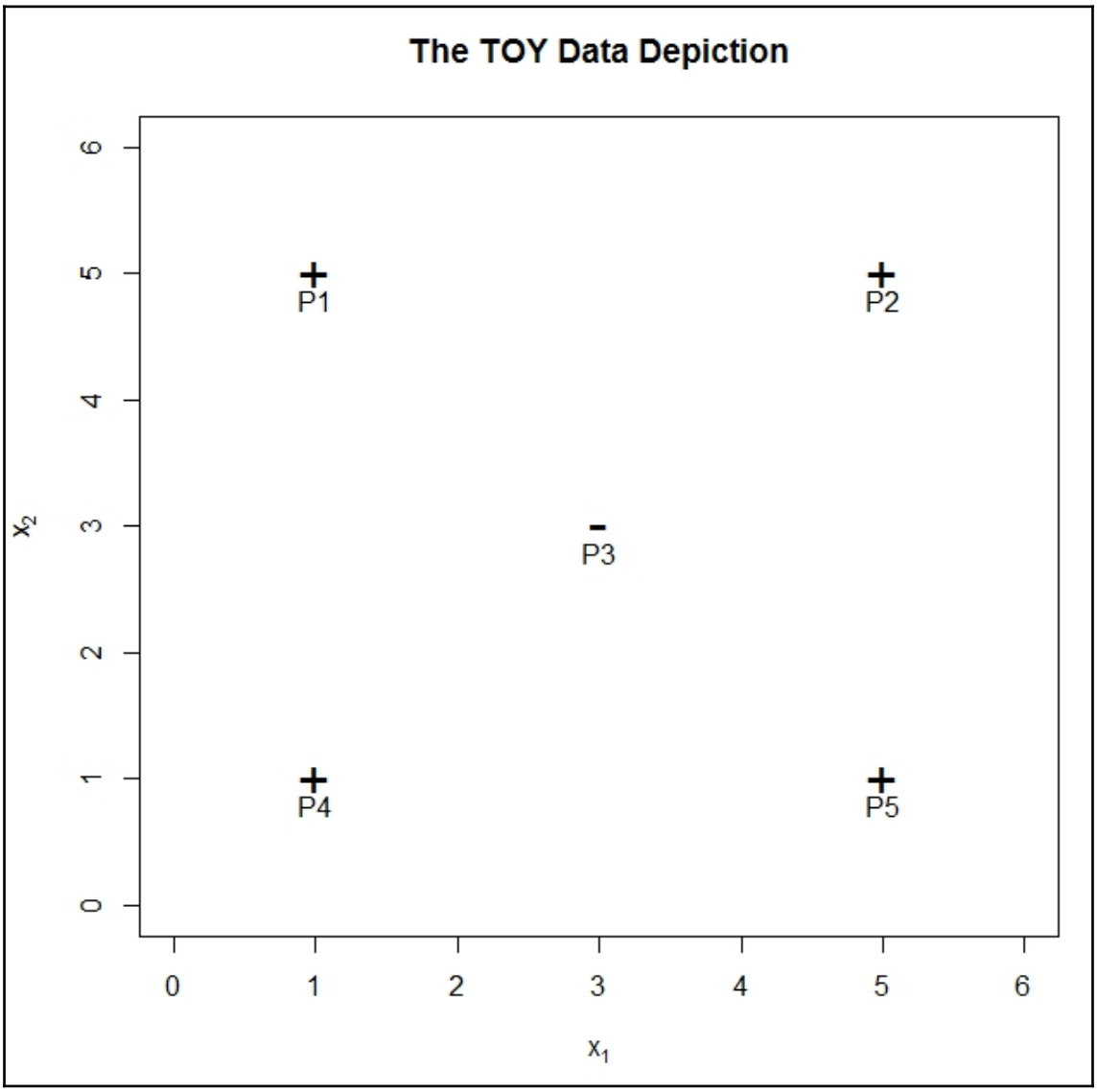

Relations between measures of importance
mean_min_depth
accuracy_decrease
gini_decrease
no_of_nodes
times_a_root
Corr: -0.576
Corr: -0.861
Corr: -0.736
Corr: -0.739
Corr: 0.679
Corr: 0.271
Corr: 0.87
Corr: 0.865
Corr: 0.766
Corr: 0.406

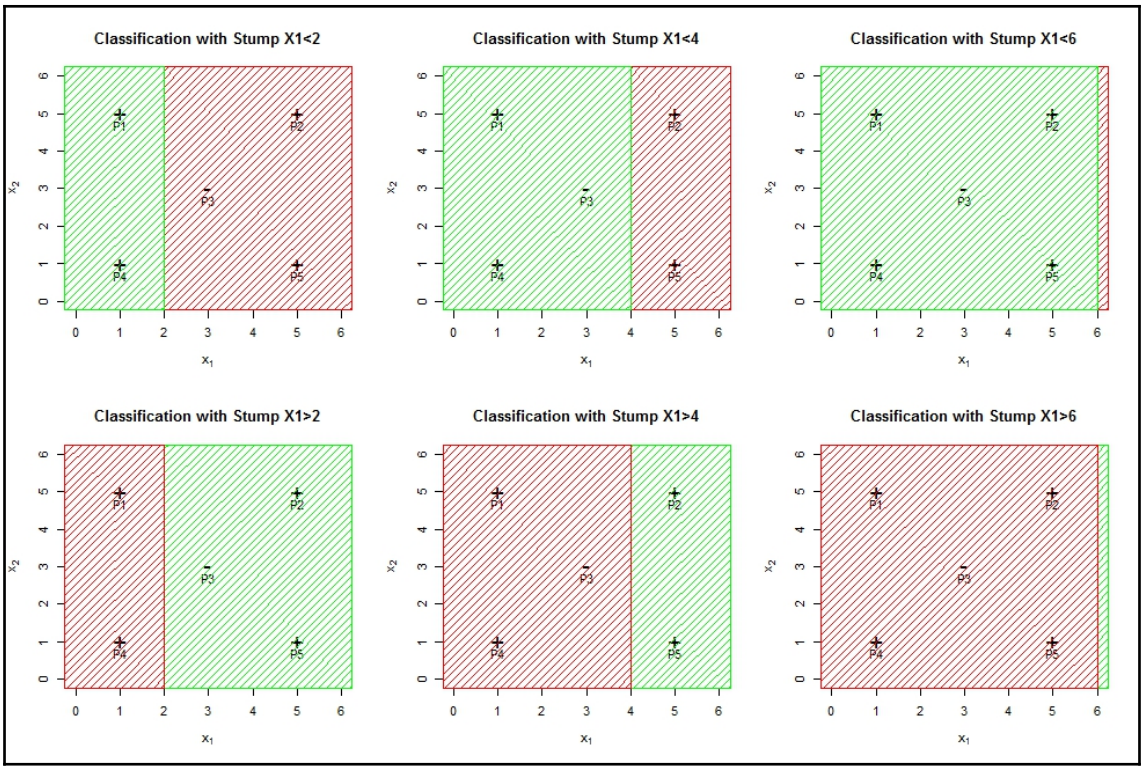

Mean minimal depth for 30 most frequent interactions
mean_min_depth
0
1
2
3
4
5
interaction
checking:amount
checking:age
checking:duration
checking:purpose
checking:employed
duration:amount
savings:amount
savings:age
duration:age
purpose:age
purpose:amount
checking:age
checking:resident
checking:installp
checking:property
duration:duration
purpose:duration
savings:duration
checking:history
checking:marital
checking:job
duration:purpose
amount:age
amount:amount
duration:employed
savings:purpose
duration:installp
checking:savings
checking:existcr
duration:resident
amount:duration
unconditional
occurrences
440
420
400
380
360
minimum

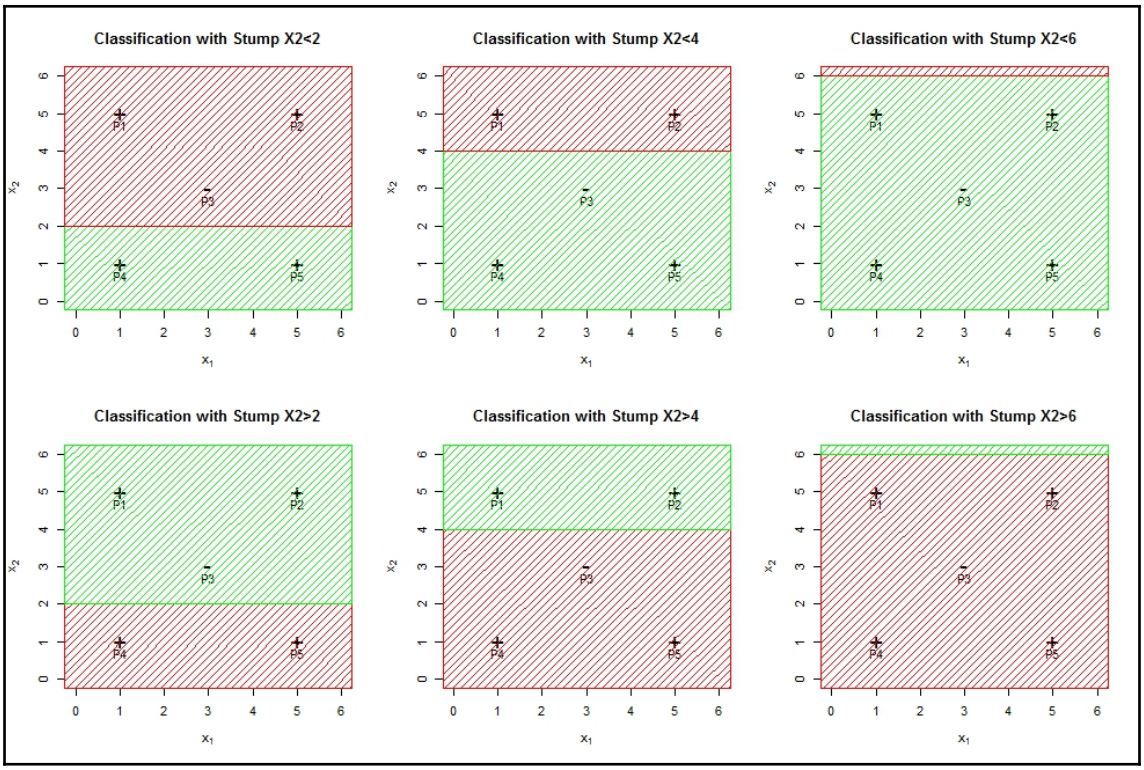

Random Forest for Spam Classification
Error
0.02
0.04
0.06
0.08
0.10
0.12
0
100
200
300
400
500
trees
Bagging for Spam Classification
Error
0.04
0.06
0.08
0.10
0.12
0
100
200
300
400
500
trees

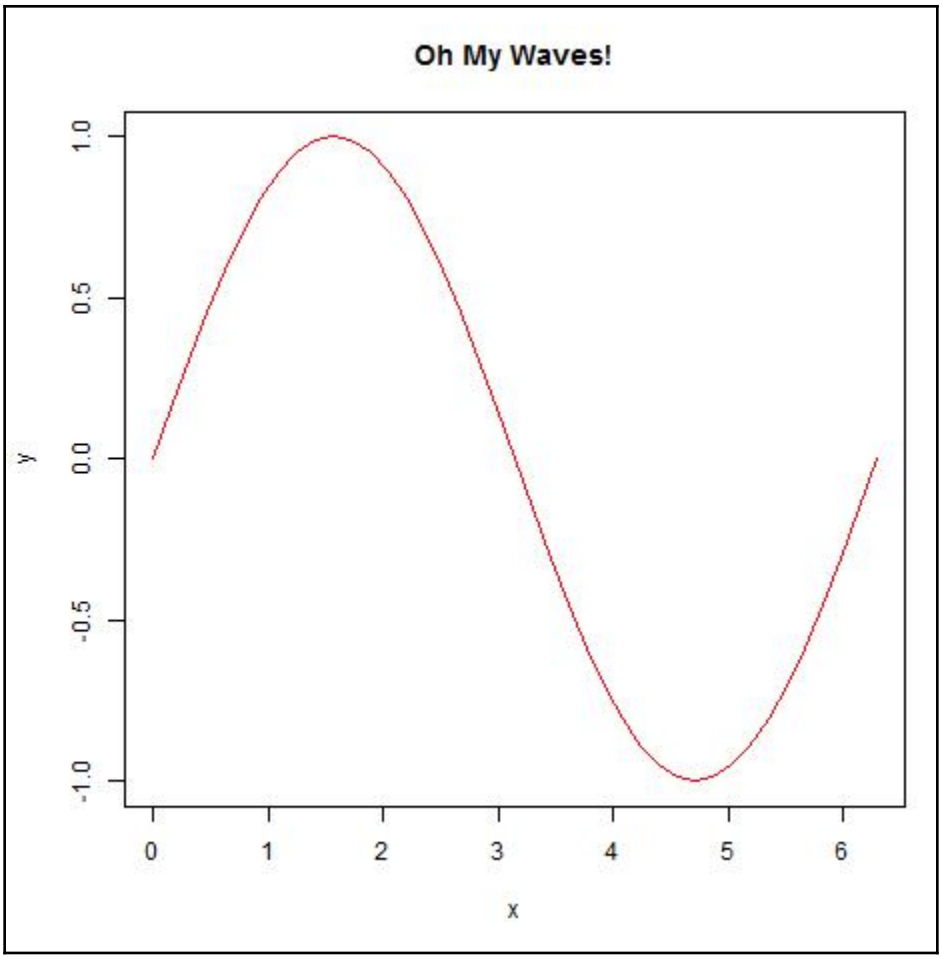

Six Multishapes Data Display
y
-3
-2
-1
0
1
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
x
Clustering with Random Forest
y
-3
-2
-1
0
1
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
x

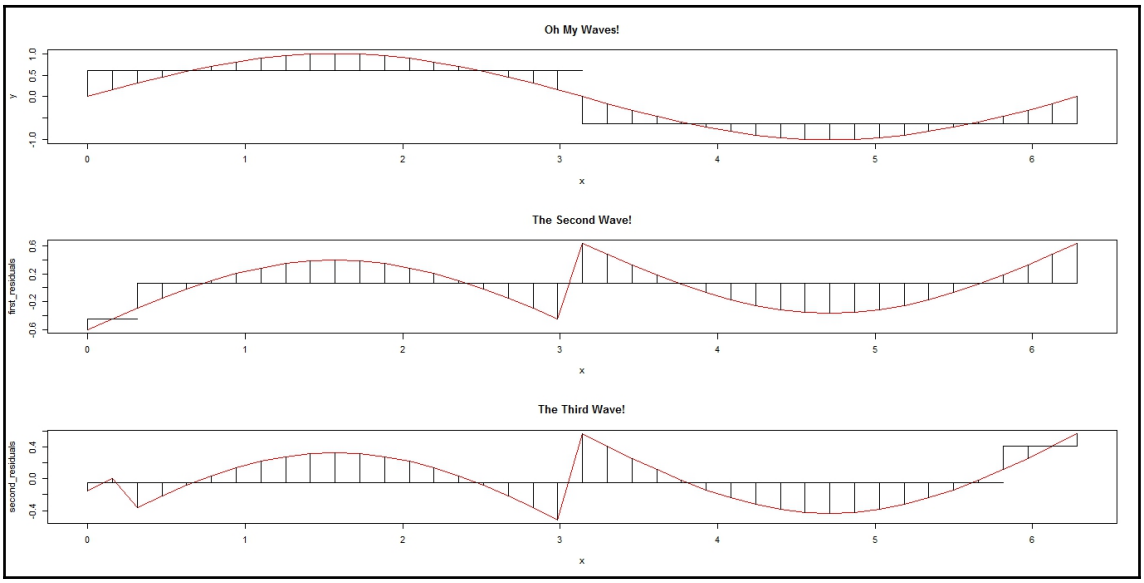
Chapter 5: The Bare Bones Boosting Algorithms

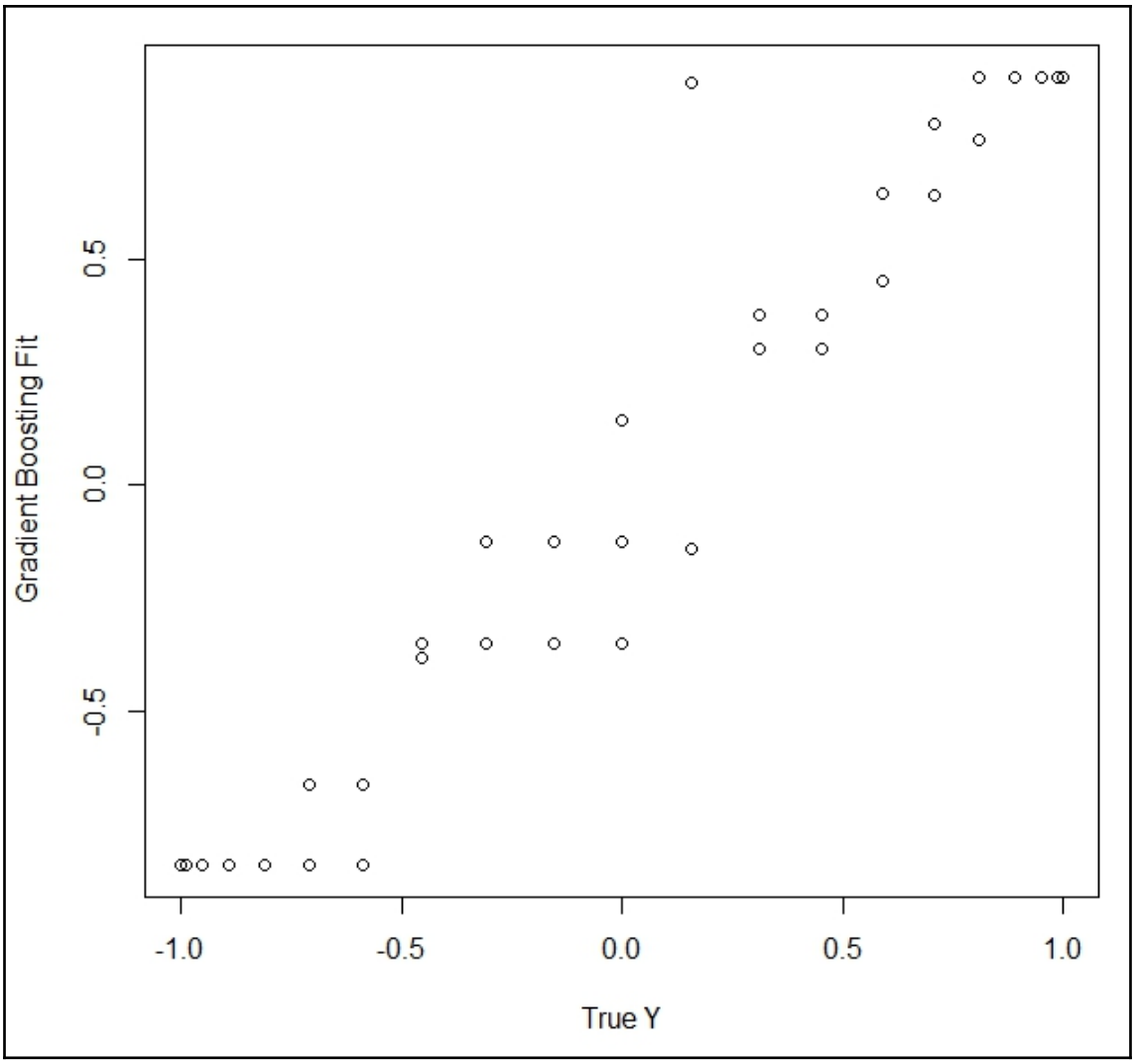


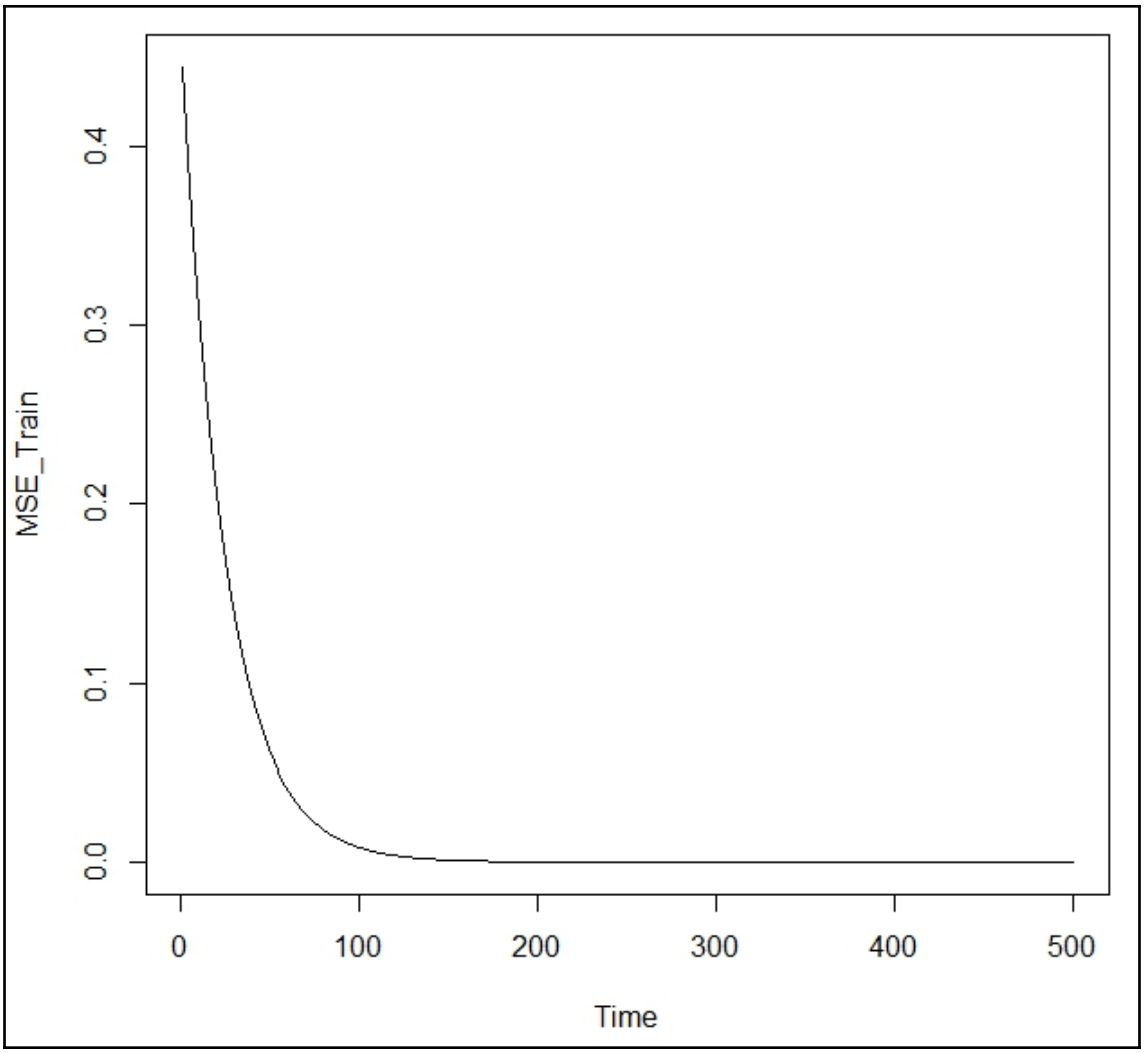

Classification with Stump X1<2
Classification with Stump X1<4
Classification with Stump X1<6
Classification with Stump X1>2
Classification with Stump X1>4
Classification with Stump X1>6
P1
P2
P3
P4
P5
x1
x2


Classification with Stump X2<2
Classification with Stump X2<4
Classification with Stump X2<6
Classification with Stump X2>2
Classification with Stump X2>4
Classification with Stump X2>6
P1
P2
P3
P4
P5
x1
x2

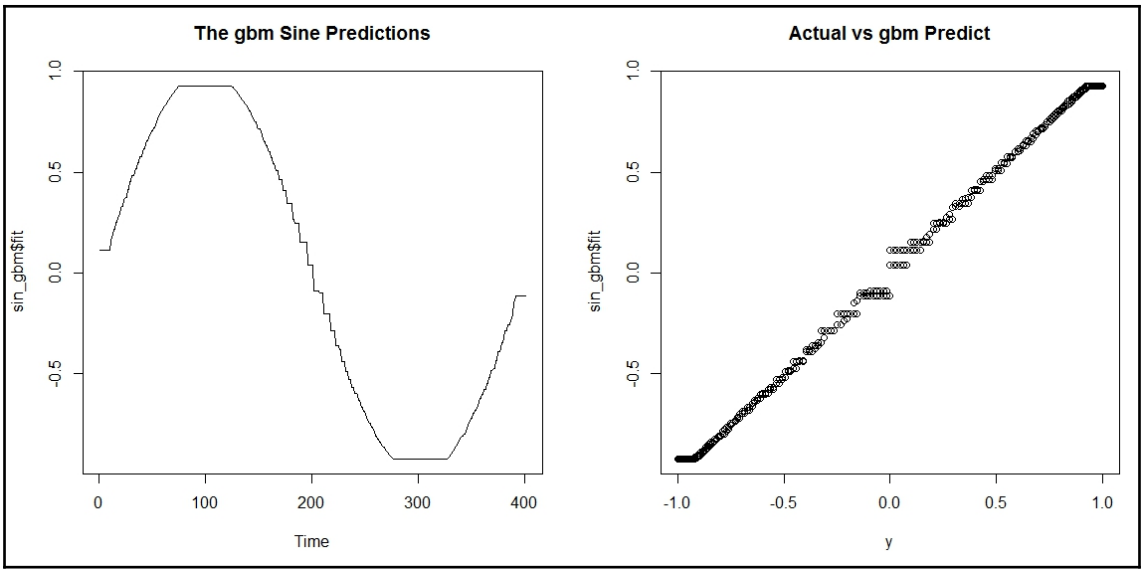

Oh My Waves!
y
1.0
0.5
0.0
-0.5
-1.0
0
1
2
3
4
5
6
x


Oh My Waves!
y
x
The Second Wave!
first_residuals
x
The Third Wave!
second_residuals
x


Gradient Boosting Fit
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
True Y

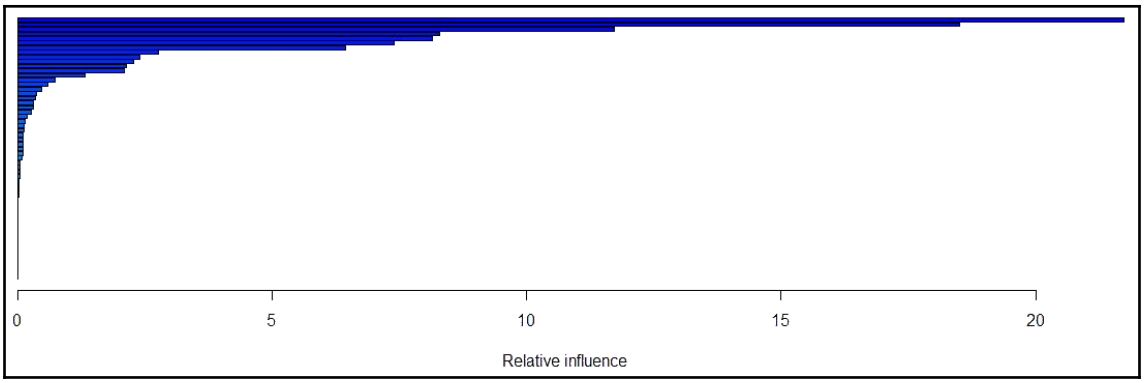


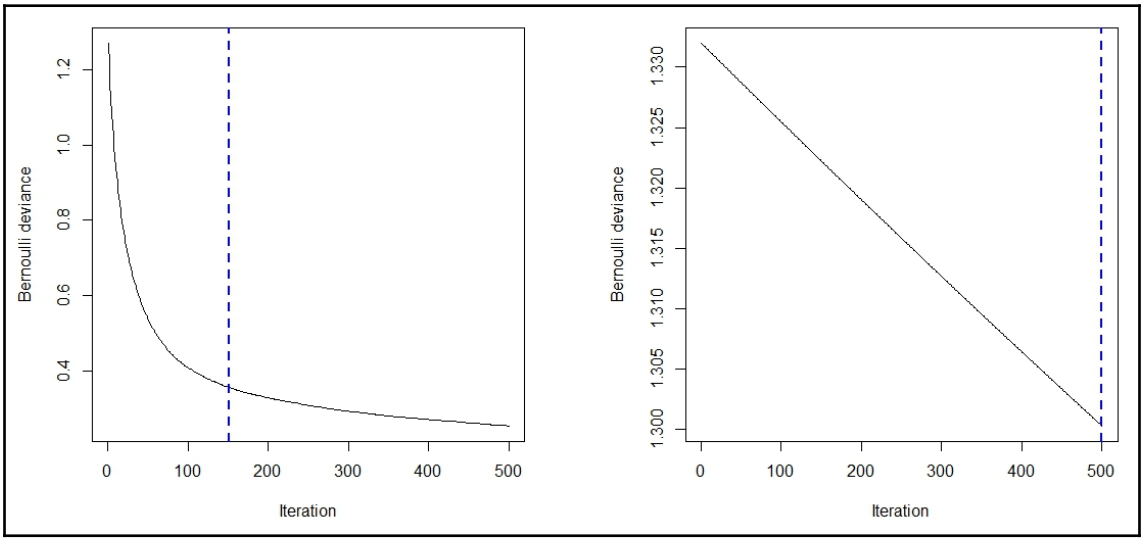

```
> boosting
function (formula, data, boos = TRUE, mfinal = 100, coeflearn = "Breiman",
    control, ...)
> gbm
function (formula = formula(data), distribution = "bernoulli",
    data = list(), weights, var.monotone = NULL, n.trees = 100,
    interaction.depth = 1, n.minobsinnode = 10, shrinkage = 0.001,
    bag.fraction = 0.5, train.fraction = 1, cv.folds = 0, keep.data = TRUE,
    verbose = "CV", class.stratify.cv = NULL, n.cores = NULL)
```

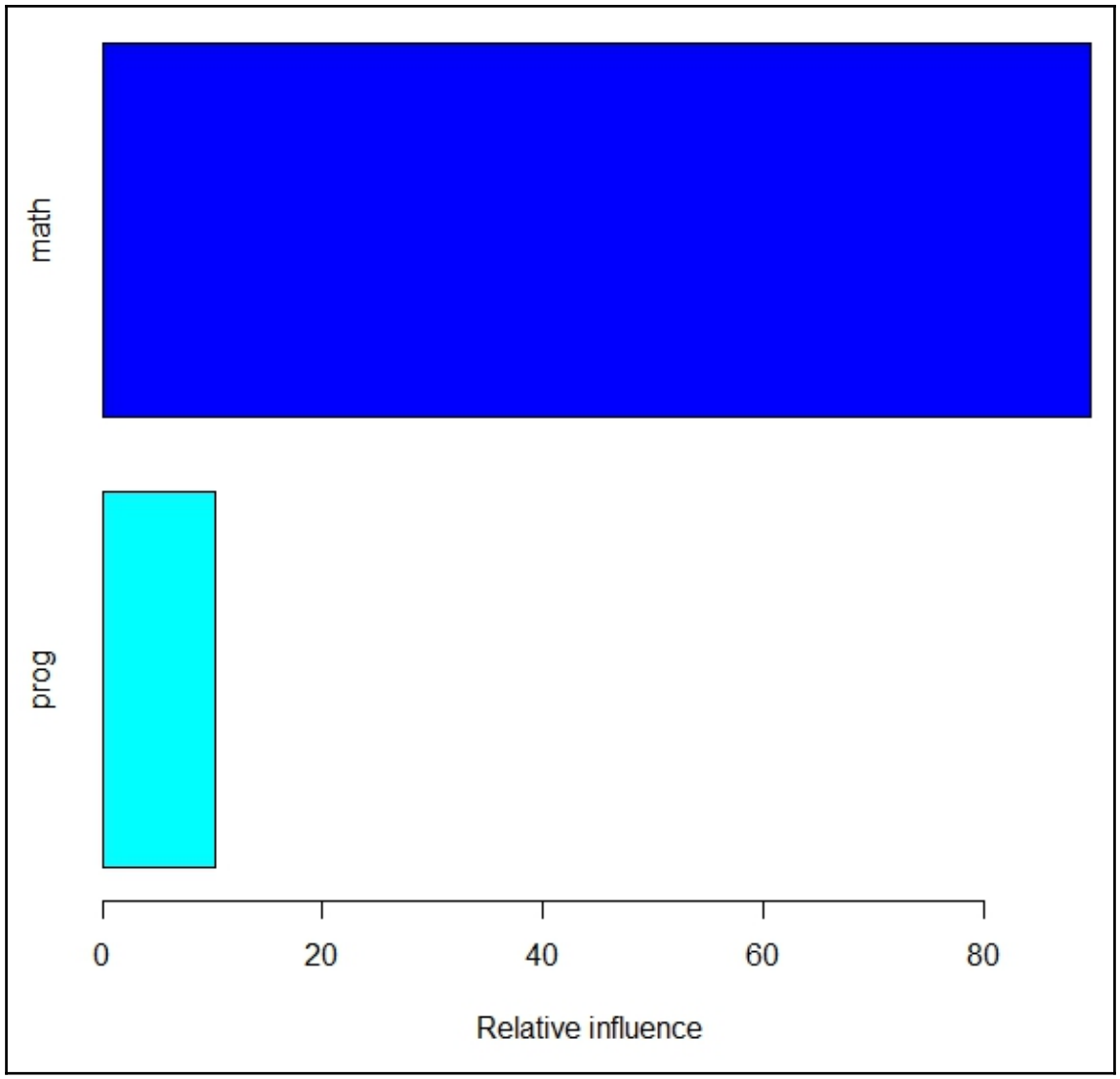

The gbm Sine Predictions
sin_gbm$fit
1.0
0.5
0.0
-0.5
0
100
200
300
400
Time
Actual vs gbm Predict
sin_gbm$fit
1.0
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
1.0
y

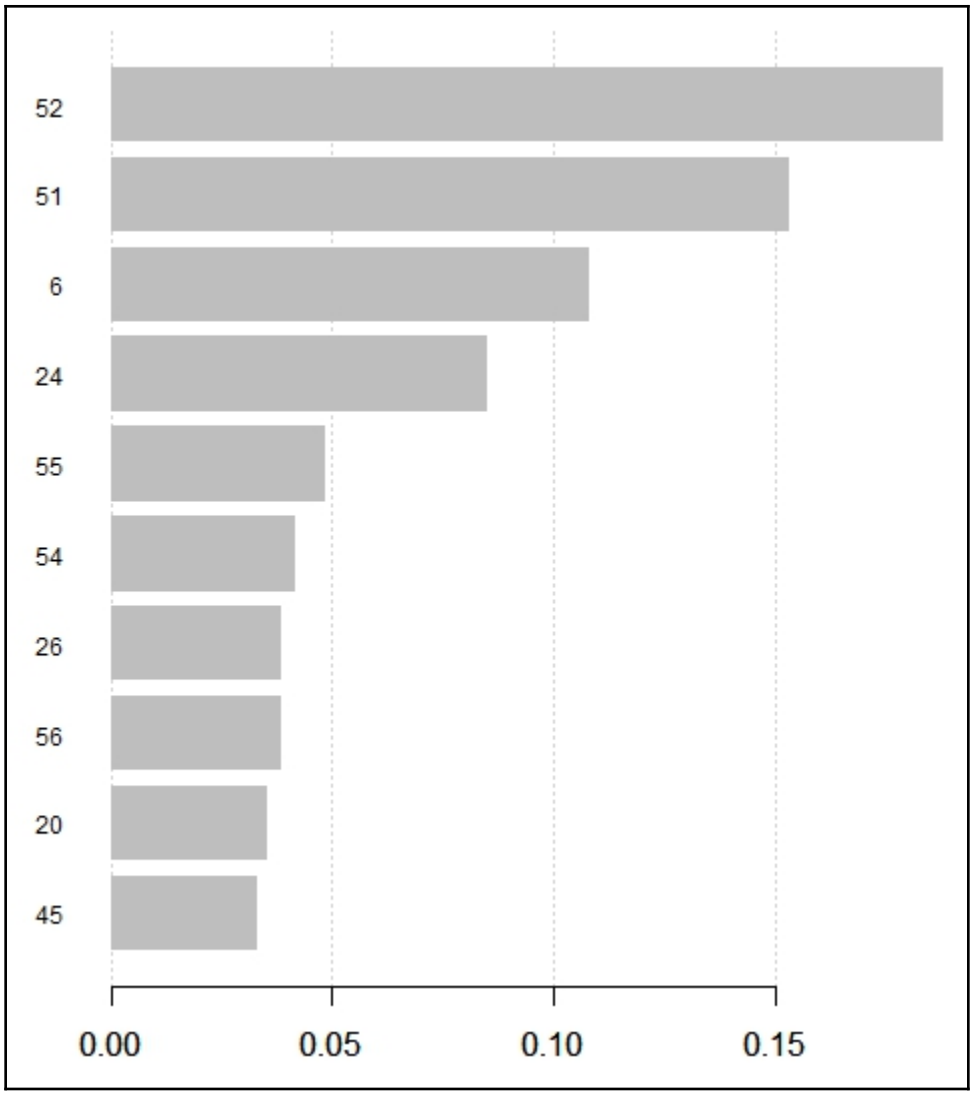
Chapter 6: Boosting Refinements

	mb1	mb2	mb3	mb4	mb5
1	-0.2375	0.30382	0.341887	0.3332	0.400
2	-0.2375	0.30382	0.341887	0.3332	0.368
3	-0.7479	-0.27177	0.072900	0.2614	0.317
4	-0.0146	0.52949	0.681289	0.5718	0.624
5	-0.6307	-0.17281	0.105200	0.2788	0.323
6	-0.6307	-0.15168	0.030059	0.2297	0.323
7	-0.1922	0.09987	0.161135	0.4676	0.368
8	-0.6069	-0.19883	0.081071	0.2535	0.326
9	-0.2375	0.00700	0.179740	0.2616	0.338
10	-0.2009	0.30655	0.333049	0.2877	0.354
11	-0.5702	0.02514	0.169440	0.3162	0.325
12	-0.2375	0.11115	0.295751	0.3105	0.351
13	-0.6307	-0.15168	0.009839	0.2310	0.339
155	-0.3786	-0.11153	0.066041	0.2508	0.329
156	-0.1922	-0.36783	0.009775	0.2550	0.323
157	-0.3181	0.18975	0.206236	0.2970	0.347
158	-0.1922	0.18975	0.184170	0.3522	0.343
159	-0.3181	-0.03722	0.235859	0.2496	0.356
160	-0.7479	0.20146	0.169046	0.3433	0.336
161	-0.3786	0.20146	0.307440	0.3752	0.353
162	-0.3181	-0.32505	0.130531	0.2373	0.340
163	-0.2009	0.67787	0.514666	0.5280	0.424
164	-0.3181	0.12438	0.130471	0.2513	0.357
165	-0.3786	0.26382	0.060728	0.2829	0.326

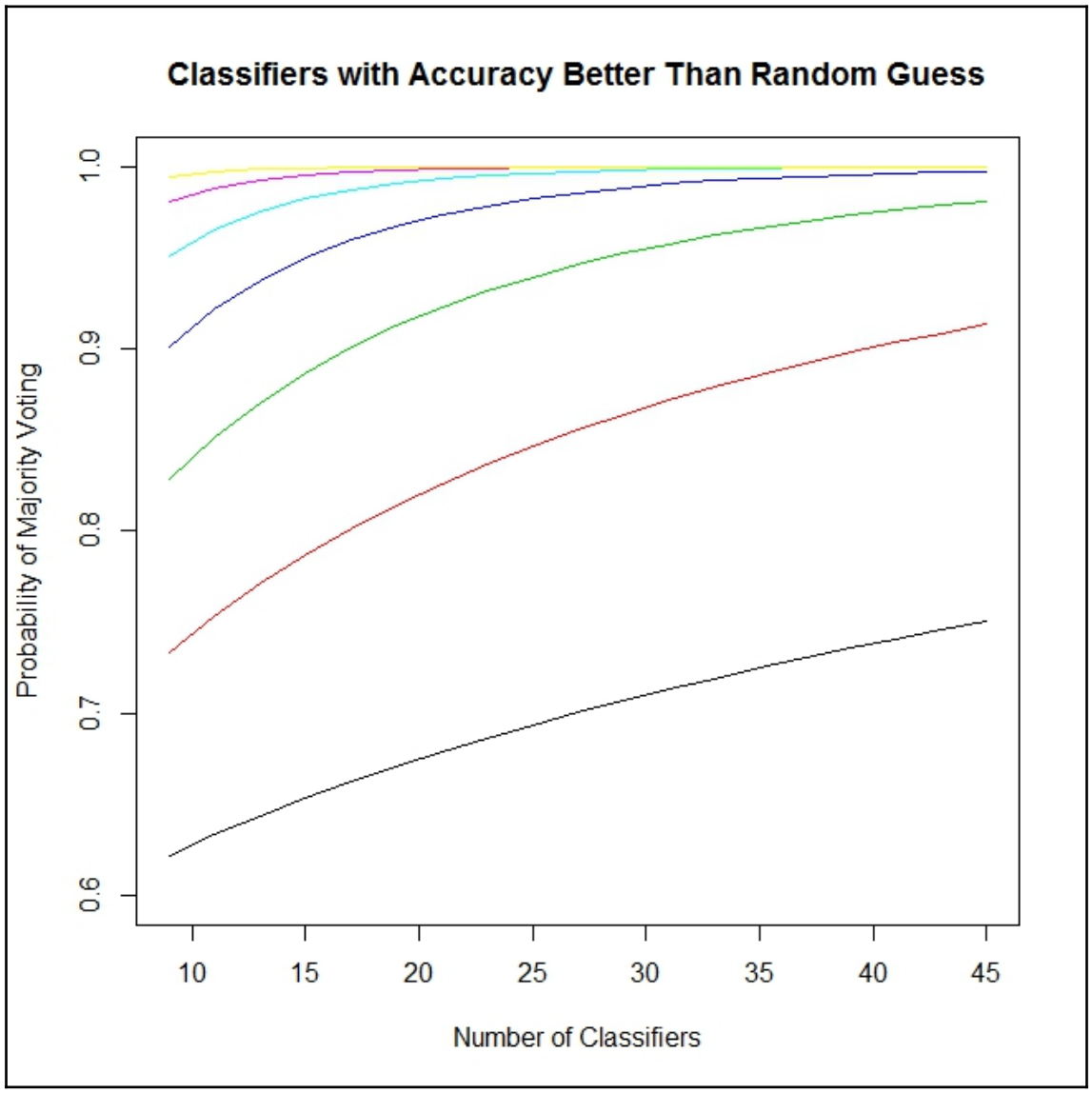

0
5
10
15
20
Relative influence

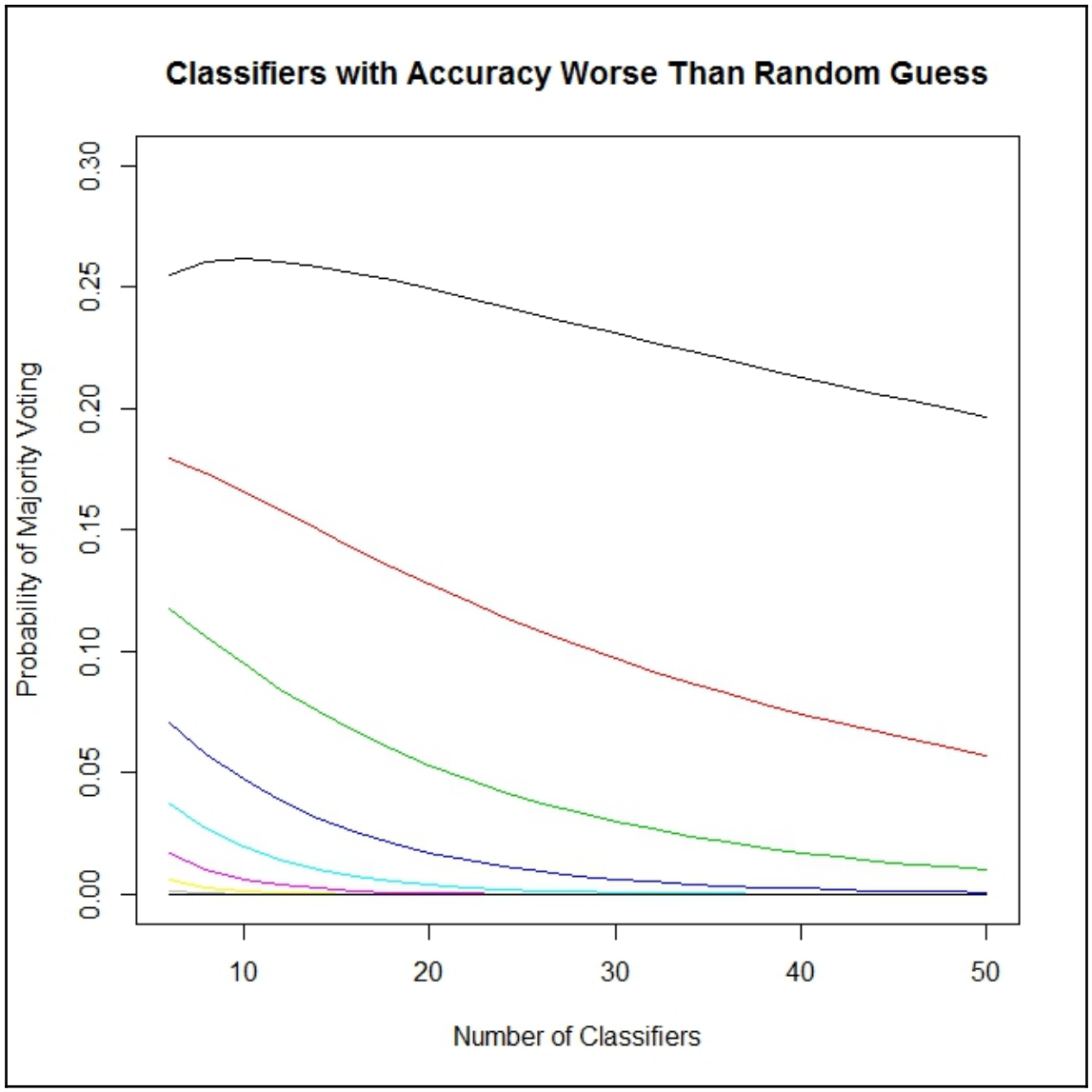

Bernoulli deviance
0.4
0.6
0.8
1.0
1.2
0
100
200
300
400
500
Iteration
Bernoulli deviance
1.300
1.305
1.310
1.315
1.320
1.325
1.330
0
100
200
300
400
500
Iteration

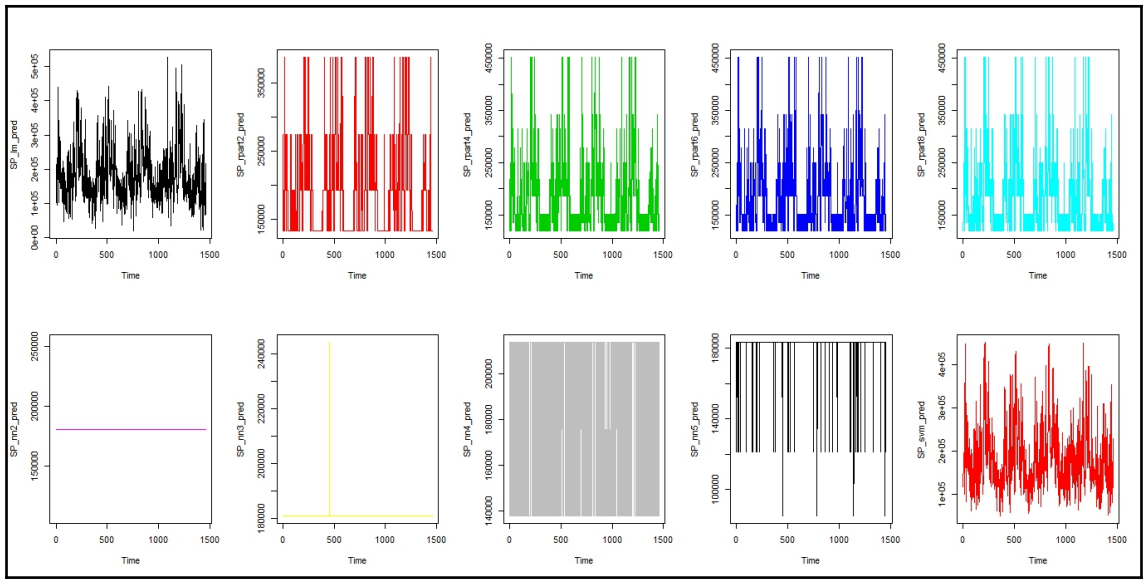

math
prog
0
20
40
60
80
Relative influence

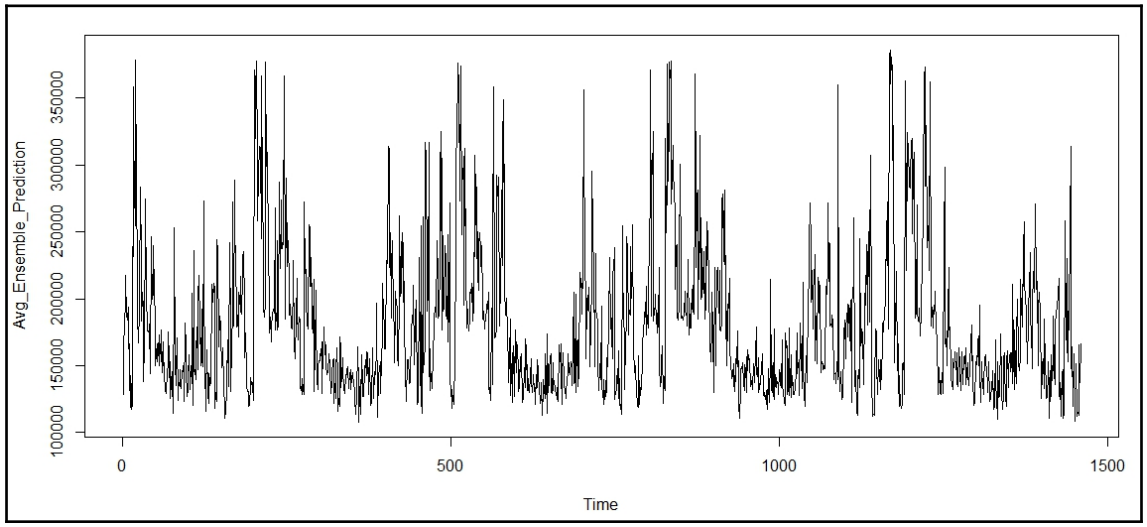

52
51
6
24
55
54
26
56
20
45
0.00
0.05
0.10
0.15

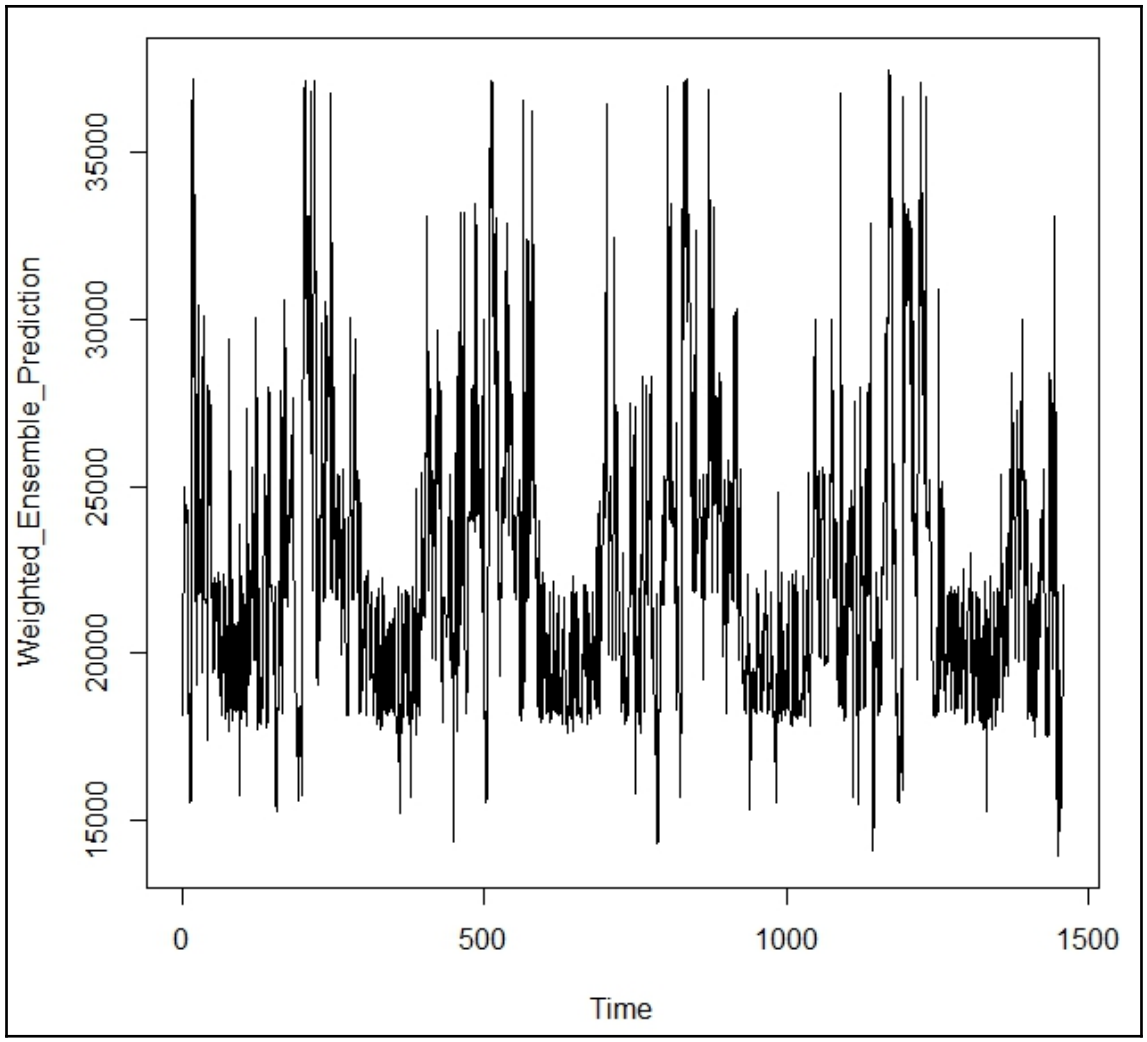
Chapter 7: The General Ensemble Technique



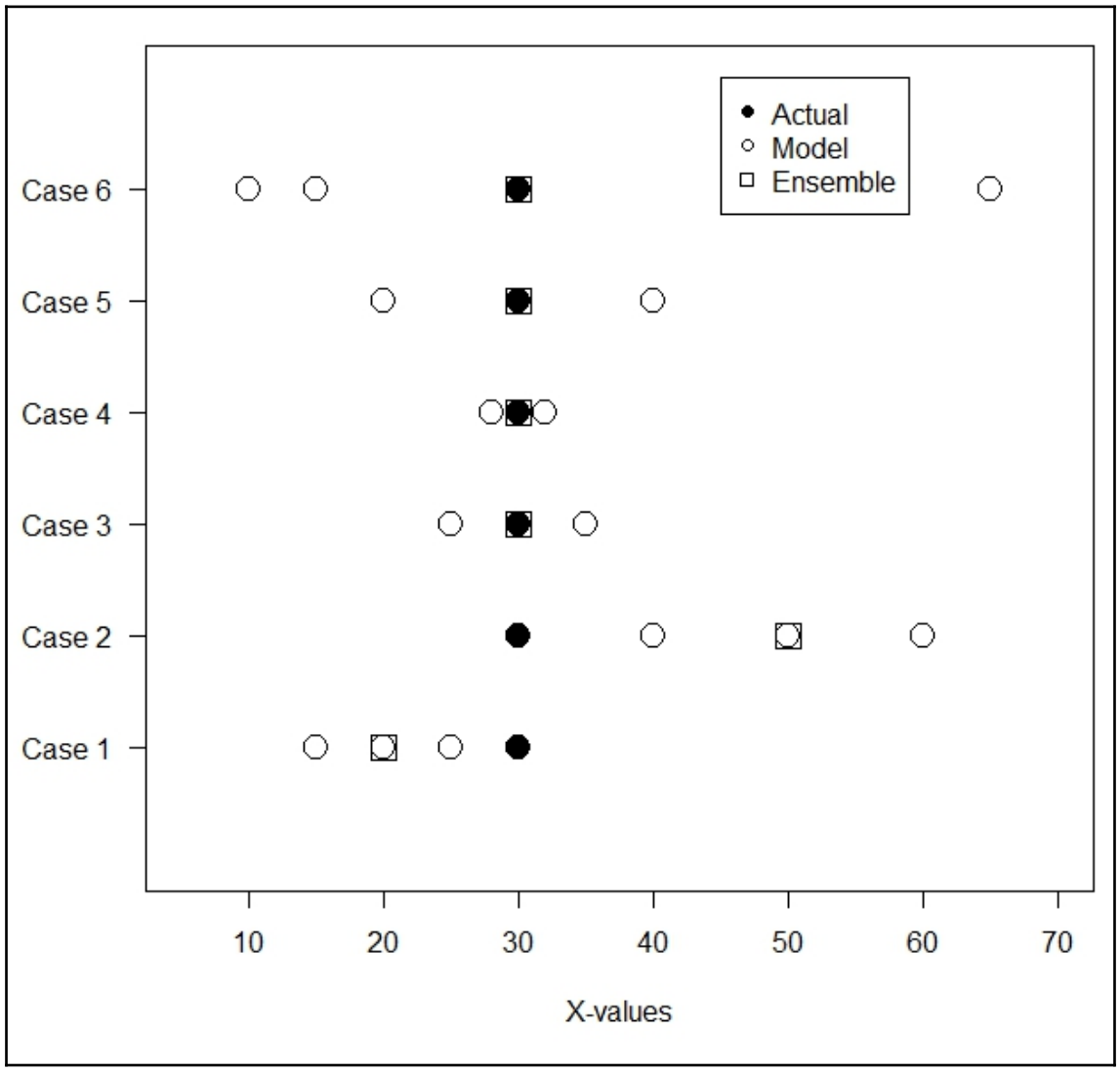

Classifiers with Accuracy Worse Than Random Guess
Probability of Majority Voting
0.00
0.05
0.10
0.15
0.20
0.25
0.30
10
20
30
40
50
Number of Classifiers

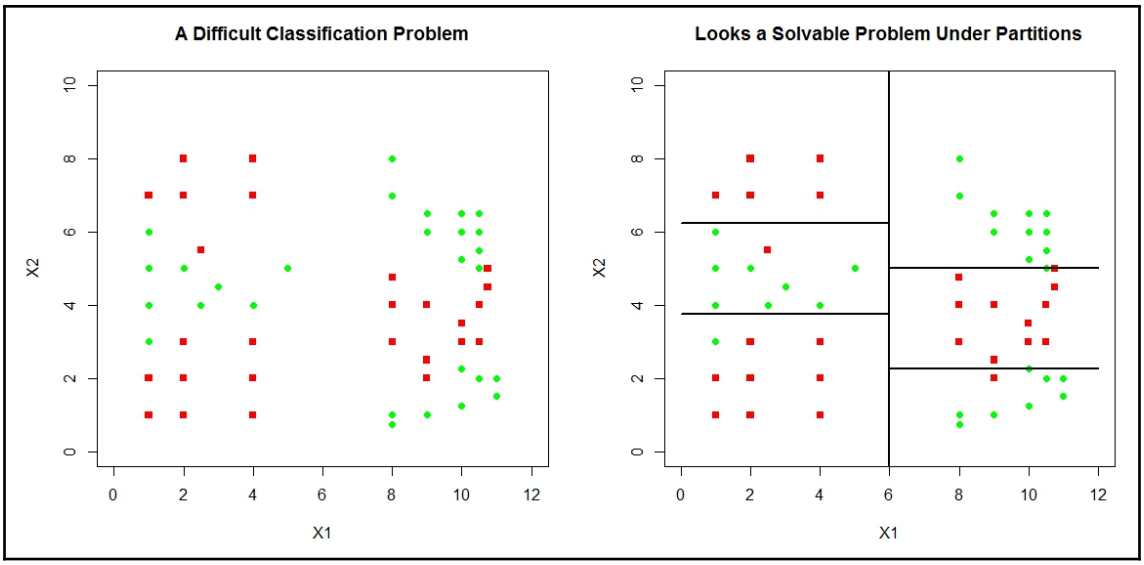

SP_lm_pred
SP_rpart2_pred
SP_rpart4_pred
SP_rpart6_pred
SP_rpart8_pred
SP_nn2_pred
SP_nn3_pred
SP_nn4_pred
SP_nn5_pred
SP_svm_pred
Time


Avg_Ensemble_Prediction
Time

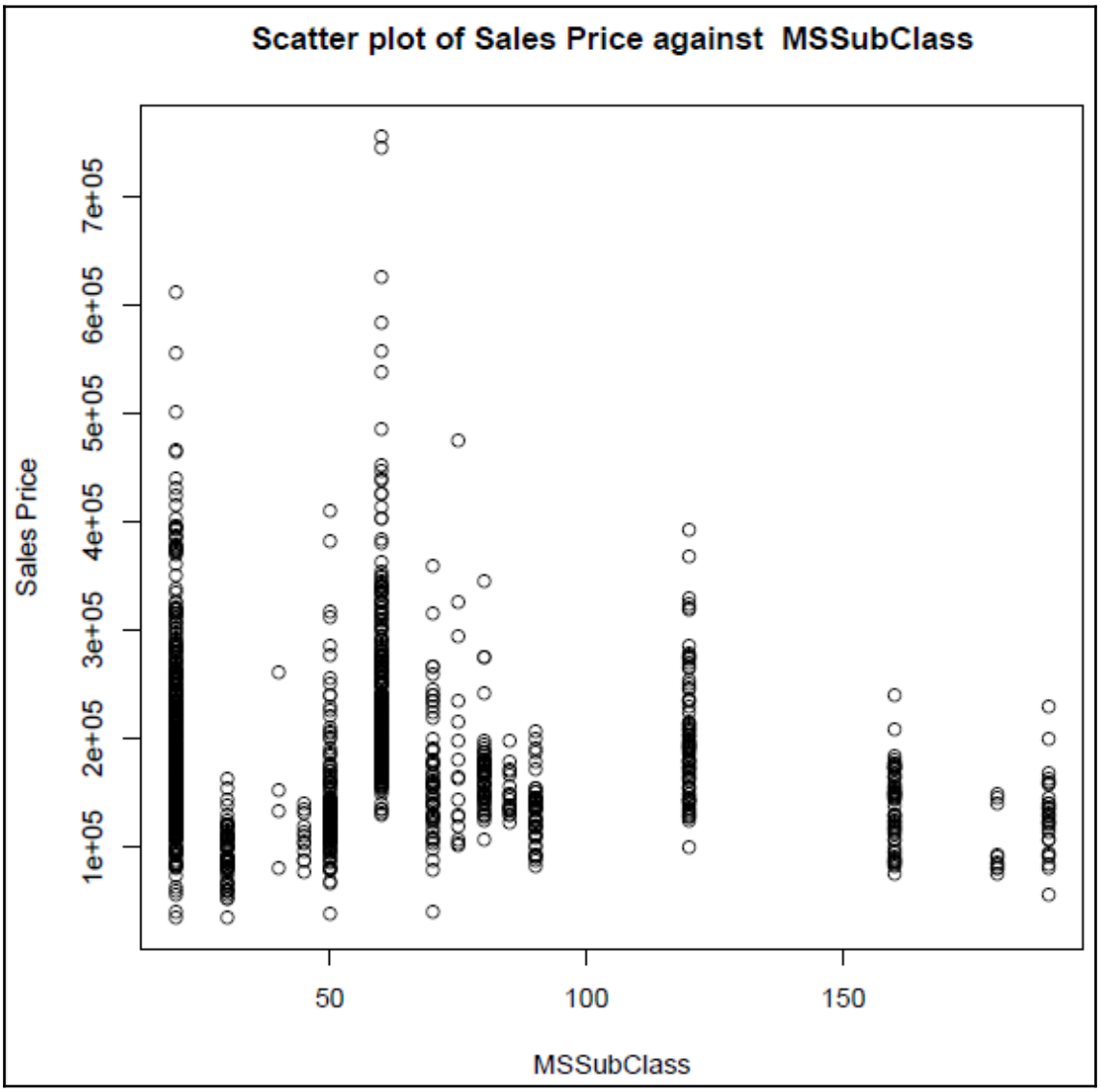

Weighted_Ensemble_Prediction
15000
20000
25000
30000
35000
0
500
1000
1500
Time

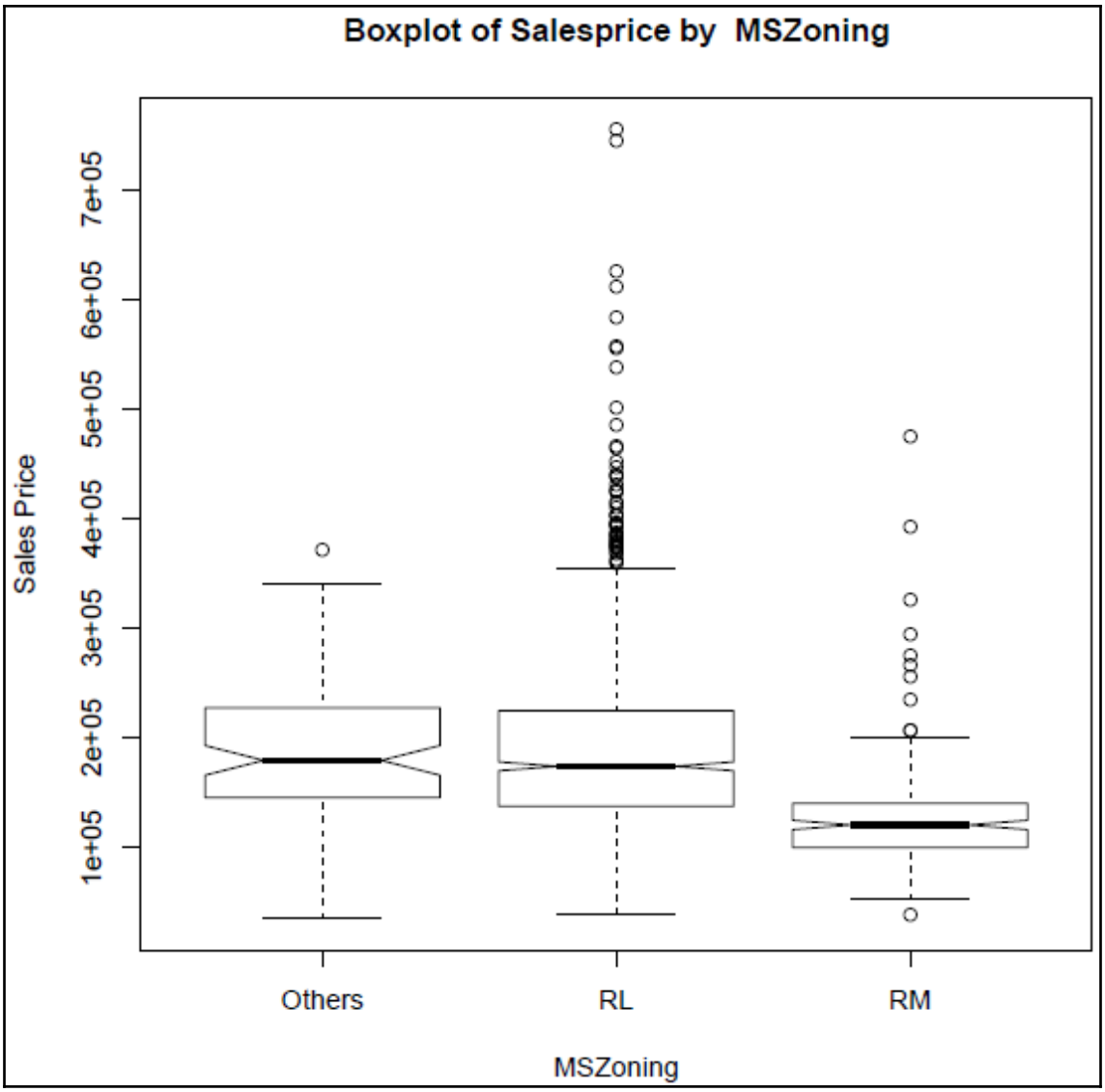
Chapter 8: Ensemble Diagnostics

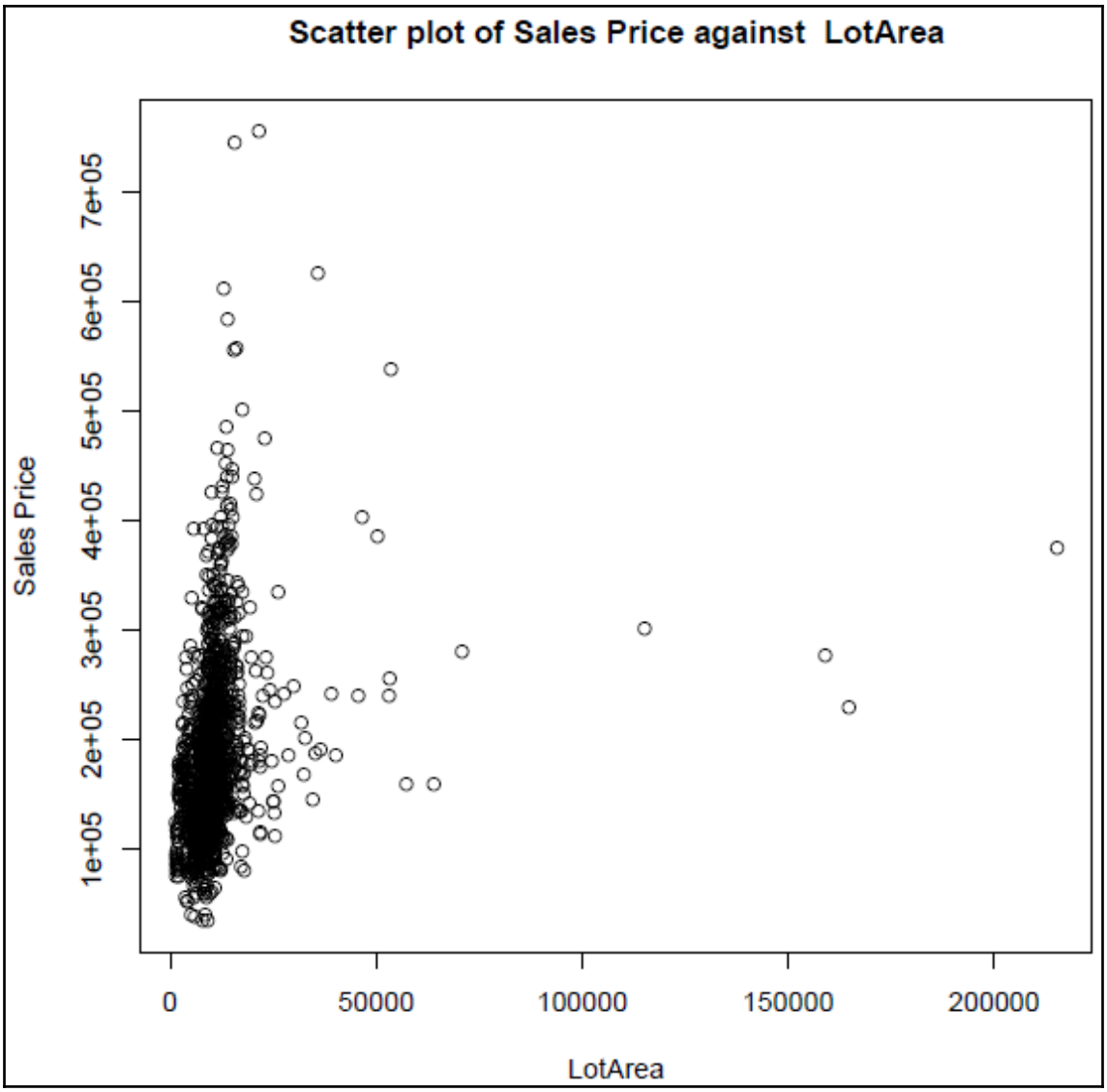



A Difficult Classification Problem
X2
X1
0
2
4
6
8
10
12
Looks a Solvable Problem Under Partitions
X2
X1

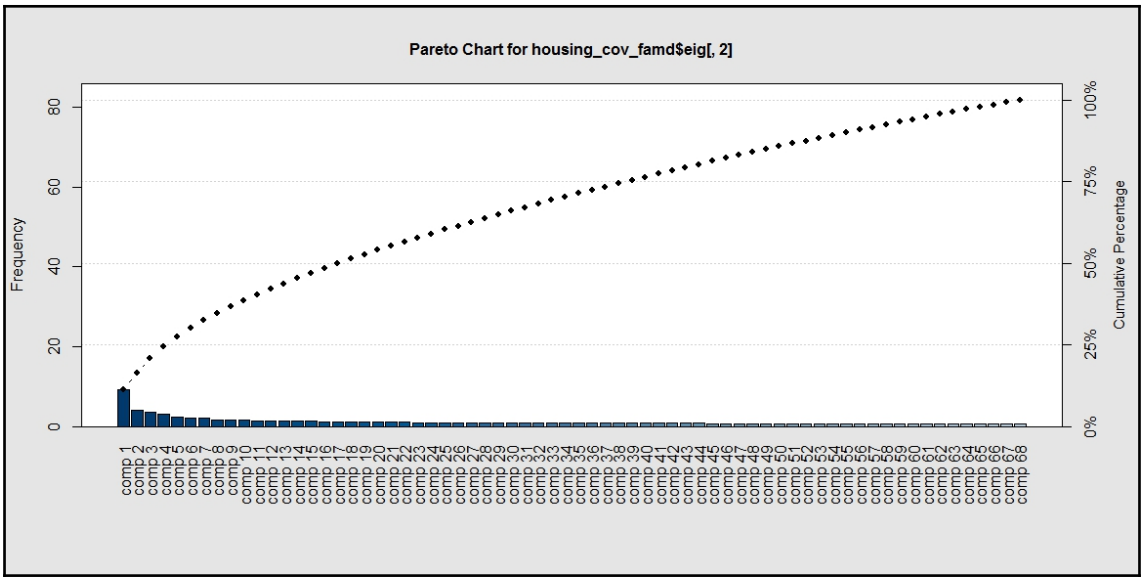
Chapter 9: Ensembling Regression Models

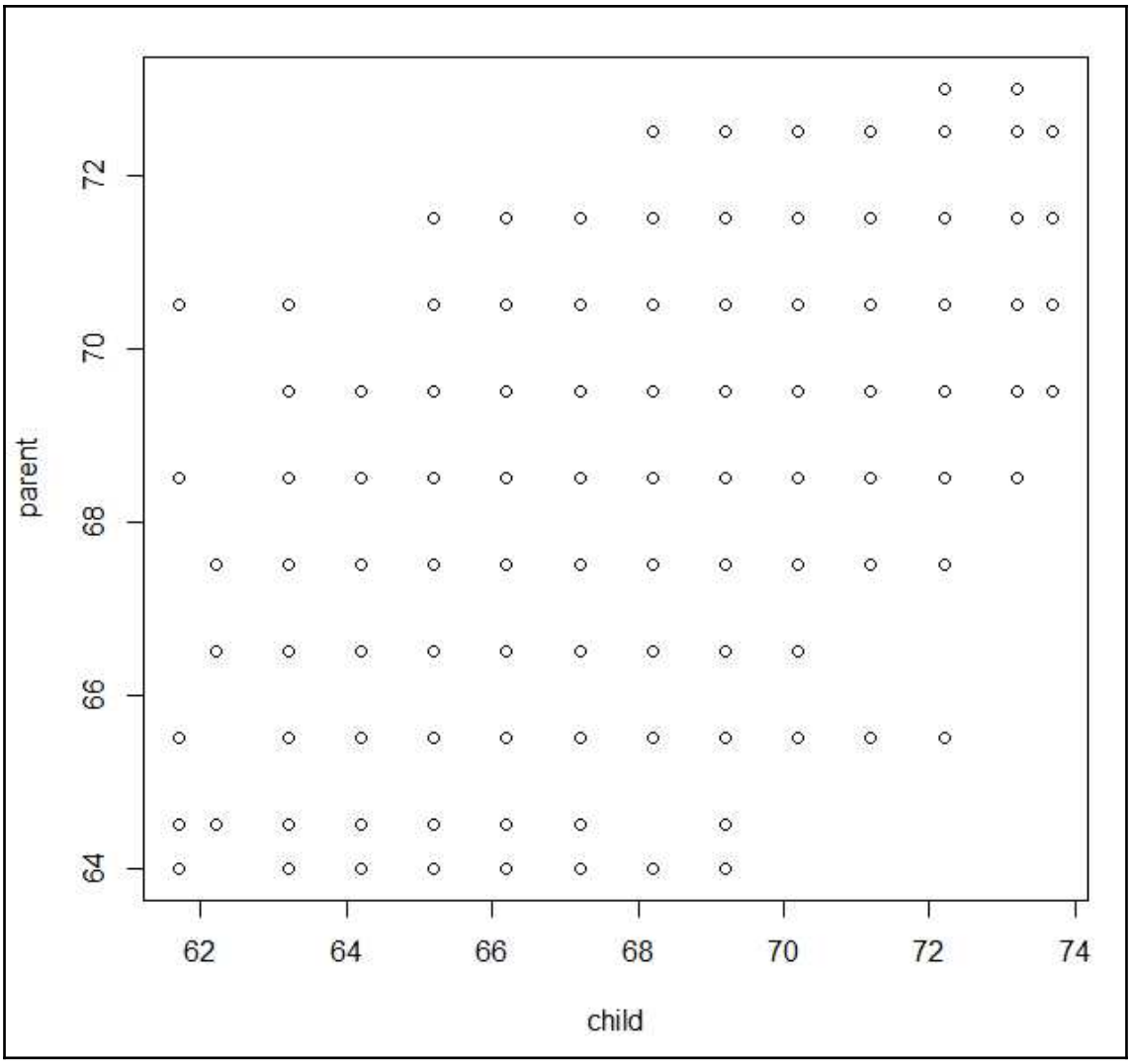


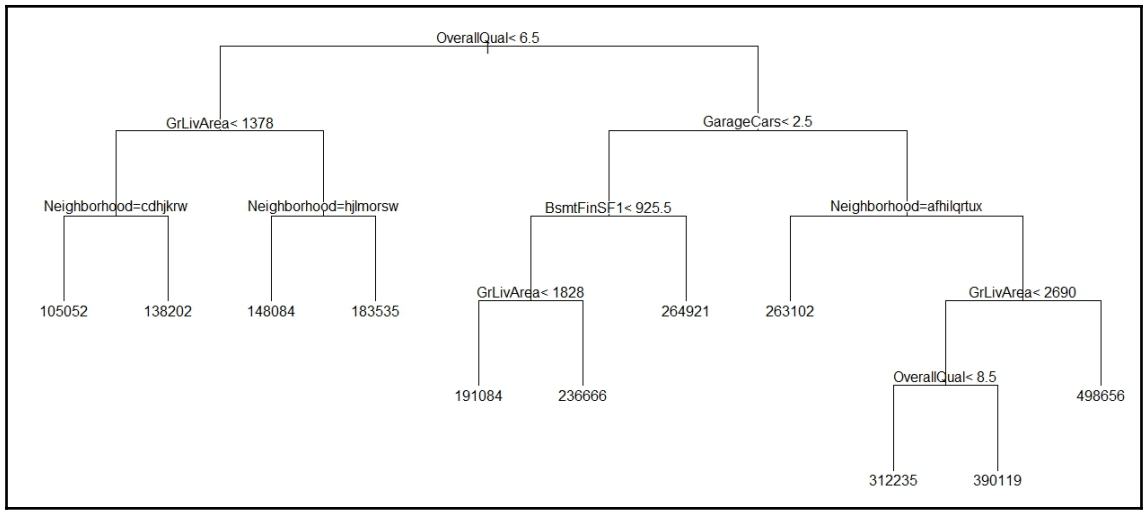

Boxplot of Salesprice by MSZoning
Sales Price
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
Others
RL
RM
MSZoning

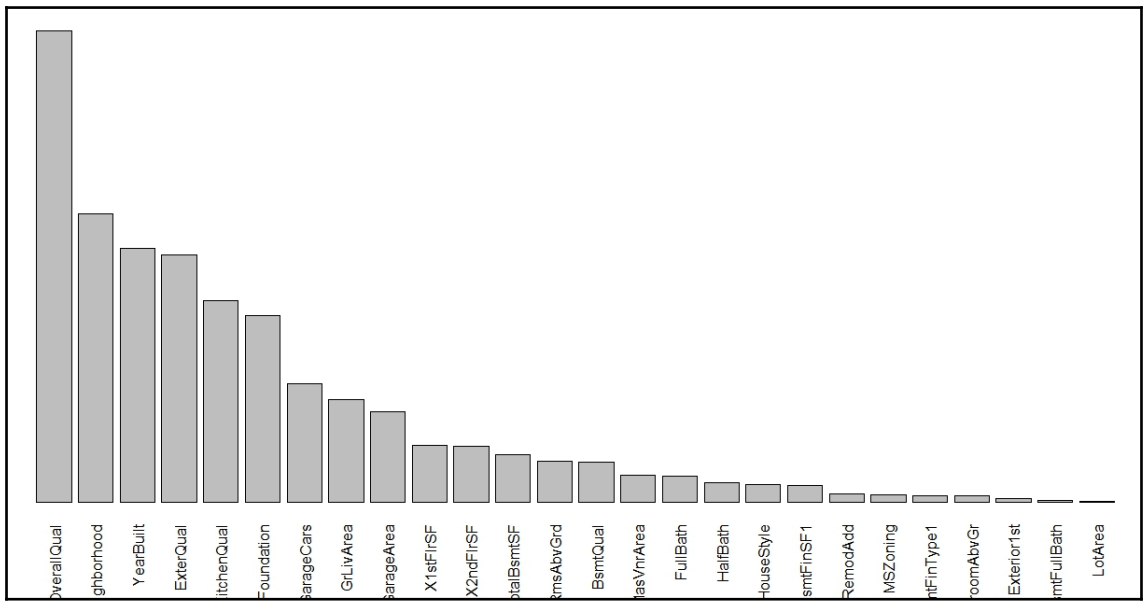

Scatter plot of Sales Price against LotArea
Sales Price
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
0
50000
100000
150000
200000
LotArea

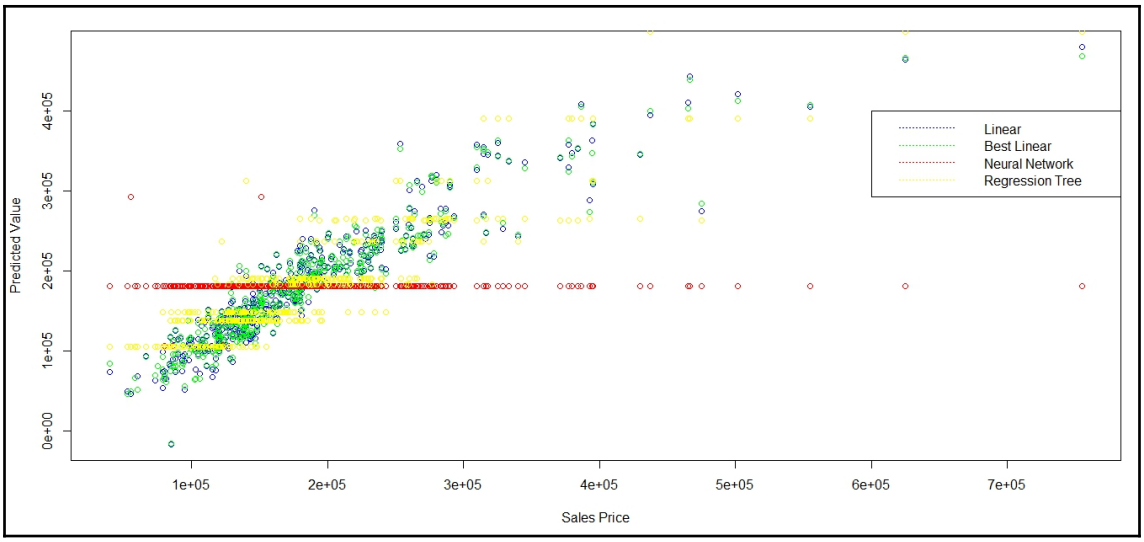
Scatter plot of Sales Price against TotalBsmtSF
Sales Price
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
0
1000
2000
3000
4000
5000
6000
TotalBsmtSF

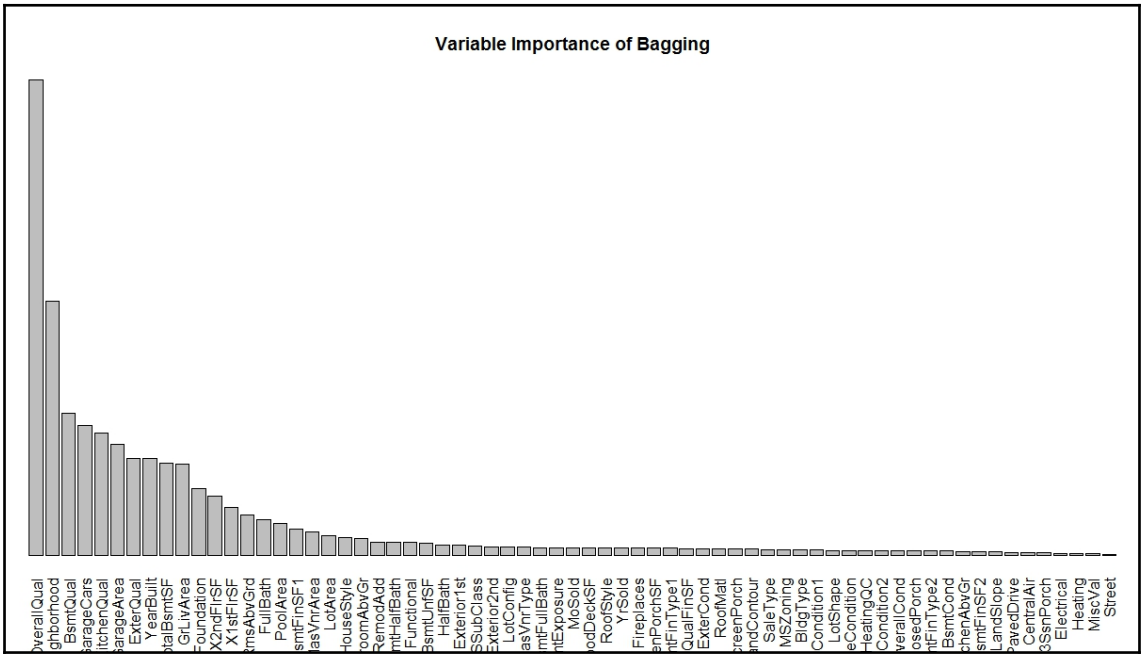

Pareto Chart for housing_cov_famd$eig[, 2]
Frequency
Cumulative Percentage
0
20
40
60
80
0%
25%
50%
75%
100%
comp 1
comp 2
comp 3
comp 4
comp 5
comp 6
comp 7
comp 8
comp 9
comp 10
comp 11
comp 12
comp 13
comp 14
comp 15
comp 16
comp 17
comp 18
comp 19
comp 20
comp 21
comp 22
comp 23
comp 24
comp 25
comp 26
comp 27
comp 28
comp 29
comp 30
comp 31
comp 32
comp 33
comp 34
comp 35
comp 36
comp 37
comp 38
comp 39
comp 40
comp 41
comp 42
comp 43
comp 44
comp 45
comp 46
comp 47
comp 48
comp 49
comp 50
comp 51
comp 52
comp 53
comp 54
comp 55
comp 56
comp 57
comp 58
comp 59
comp 60
comp 61
comp 62
comp 63
comp 64
comp 65
comp 66
comp 67
comp 68

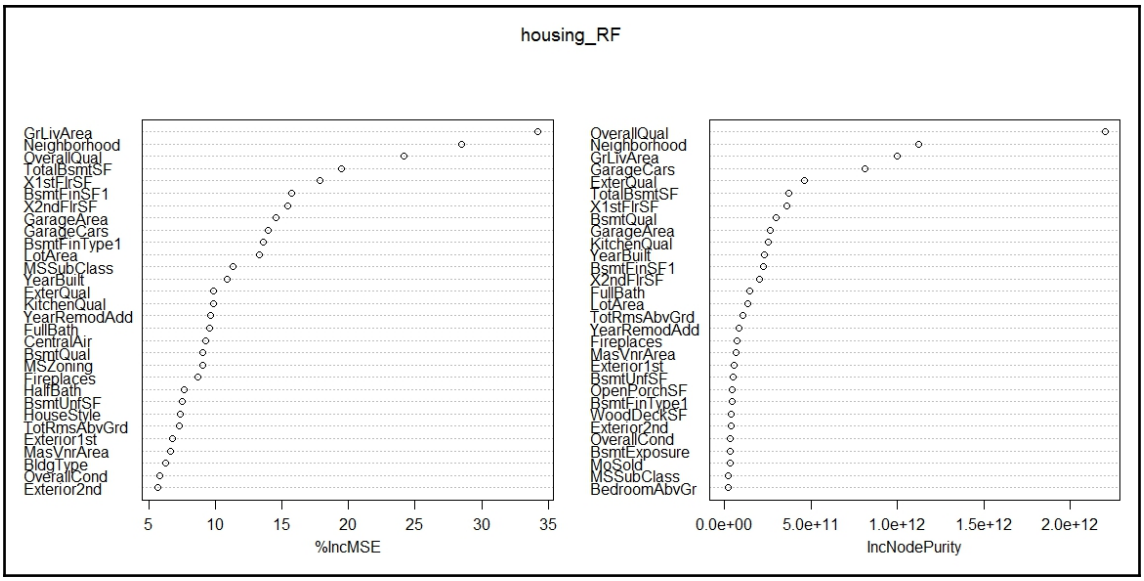

parent
child
64
66
68
70
72
62
64
66
68
70
72
74

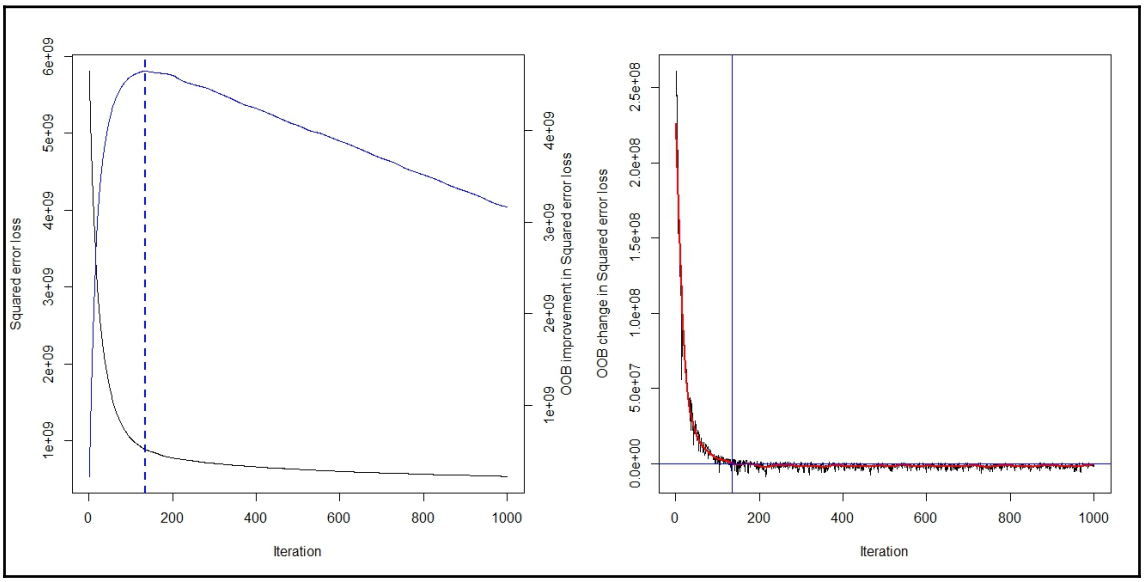

OverallQual< 6.5
GrLivArea< 1378
GarageCars< 2.5
Neighborhood=cdhjkrw
Neighborhood=hjlmorsw
BsmtFinSF1< 925.5
Neighborhood=afhilqrtux
105052
138202
148084
183535
GrLivArea< 1828
264921
263102
GrLivArea< 2690
191084
236666
OverallQual< 8.5
498656
312235
390119

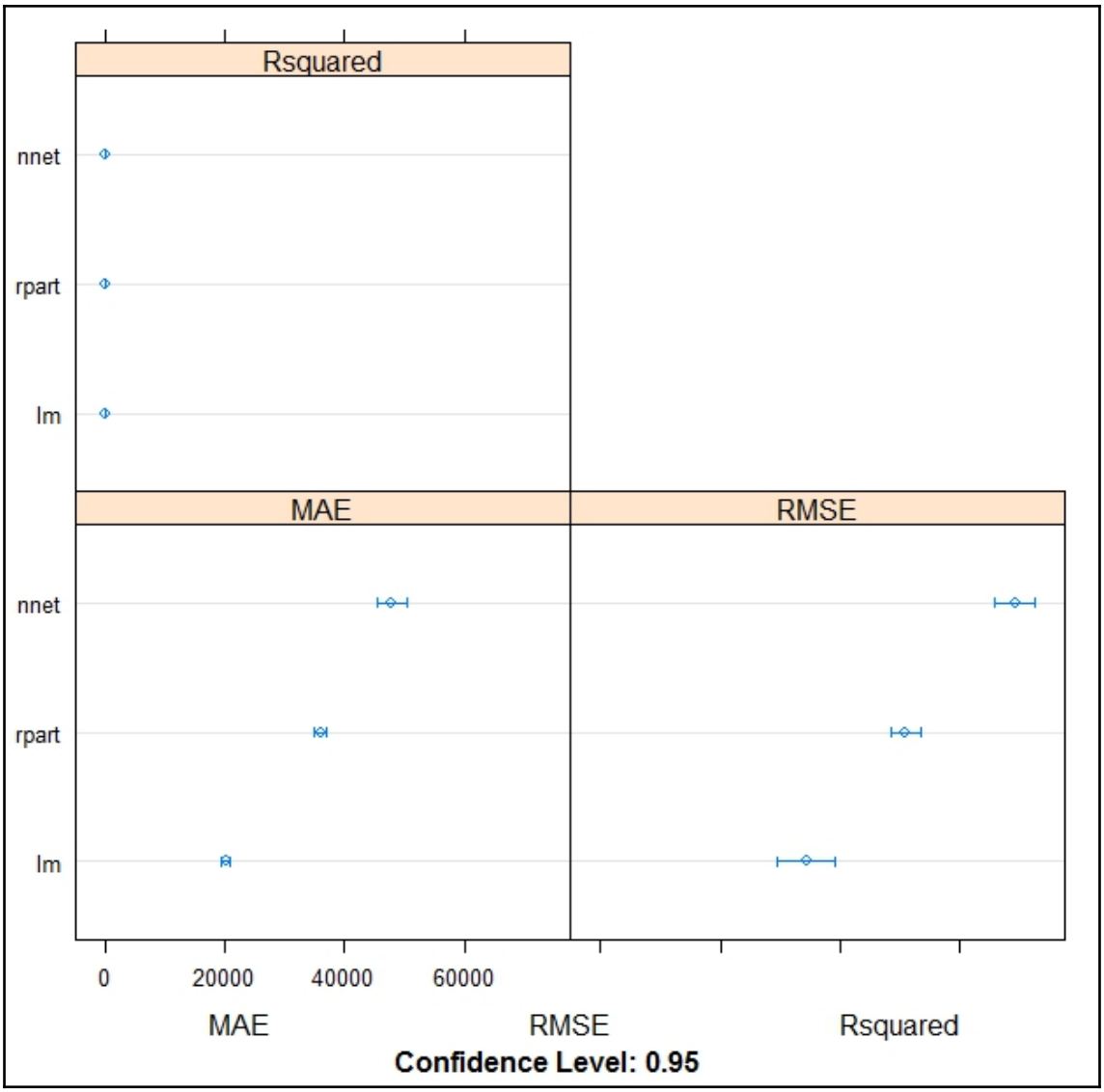

OverallQual
ghborhood
YearBuilt
ExterQual
itchenQual
Foundation
arageCars
GrLivArea
arageArea
X1stFlrSF
X2ndFlrSF
talBsmtSF
RmsAbvGrd
BsmtQual
lasVnrArea
FullBath
HalfBath
HouseStyle
smtFinSF1
RemodAdd
MSZoning
ntFinType1
roomAbvGr
Exterior1st
rmtFullBath
LotArea


Predicted Value
0e+00
1e+05
2e+05
3e+05
4e+05
1e+05
2e+05
3e+05
4e+05
5e+05
6e+05
7e+05
Sales Price
Linear
Best Linear
Neural Network
Regression Tree

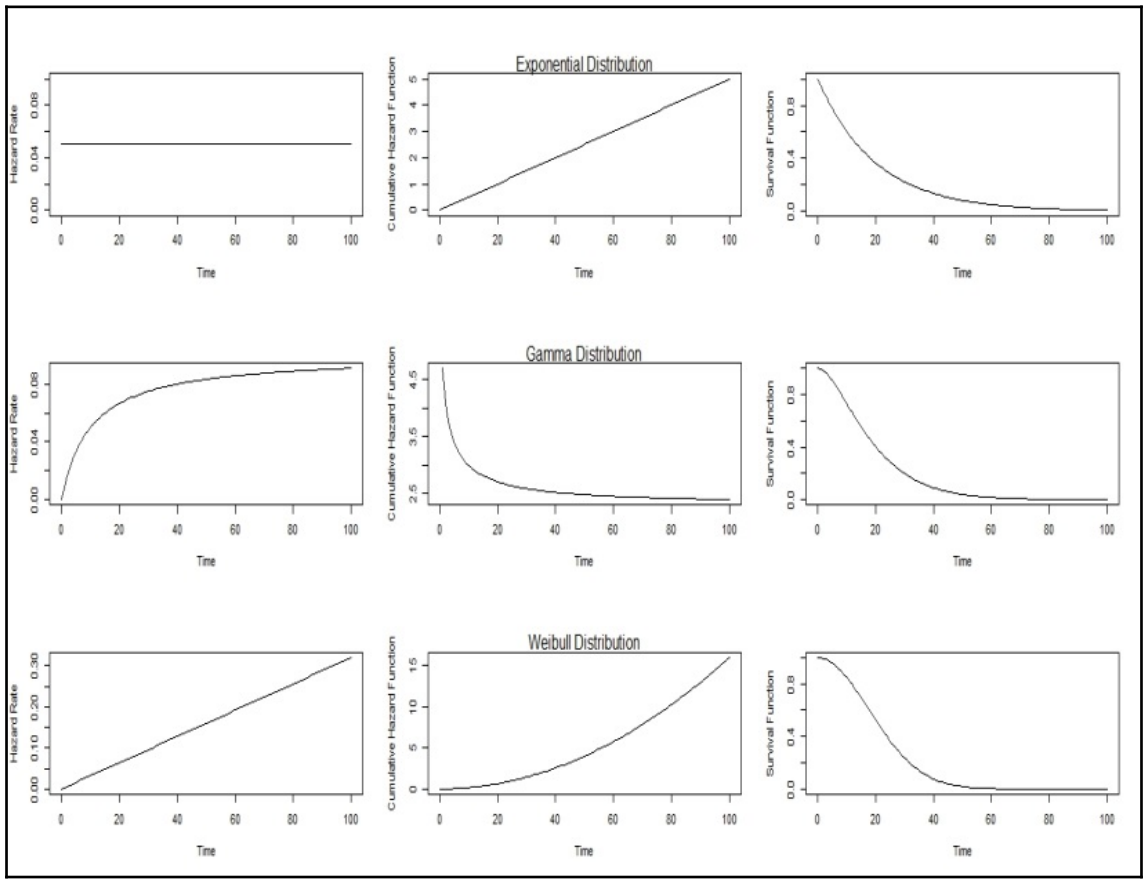

Variable Importance of Bagging
OverallQual
ghborhood
BsmtQual
arageCars
itchenQual
arageArea
ExterQual
YearBuilt
otalBsmtSF
GrLivArea
Foundation
X2ndFlrSF
X1stFlrSF
tmsAbvGrd
FullBath
PoolArea
smtFinSF1
asVnrArea
LotArea
HouseStyle
oomAbvGr
RemodAdd
mtHalfBath
Functional
BsmtUnfSF
HalfBath
Exterior1st
SSubClass
Exterior2nd
LotConfig
asVnrType
mtFullBath
mtExposure
MoSold
odDeckSF
RoofStyle
YrSold
Fireplaces
enPorchSF
mtFinType1
QualFinSF
ExterCond
RoofMatl
creenPorch
andContour
SaleType
MSZoning
BldgType
Condition1
LotShape
eCondition
HeatingQC
Condition2
verallCond
osedPorch
ltFinType2
BsmtCond
chenAbvGr
smtFinSF2
LandSlope
avedDrive
CentralAir
3SsnPorch
Electrical
Heating
MiscVal
Street

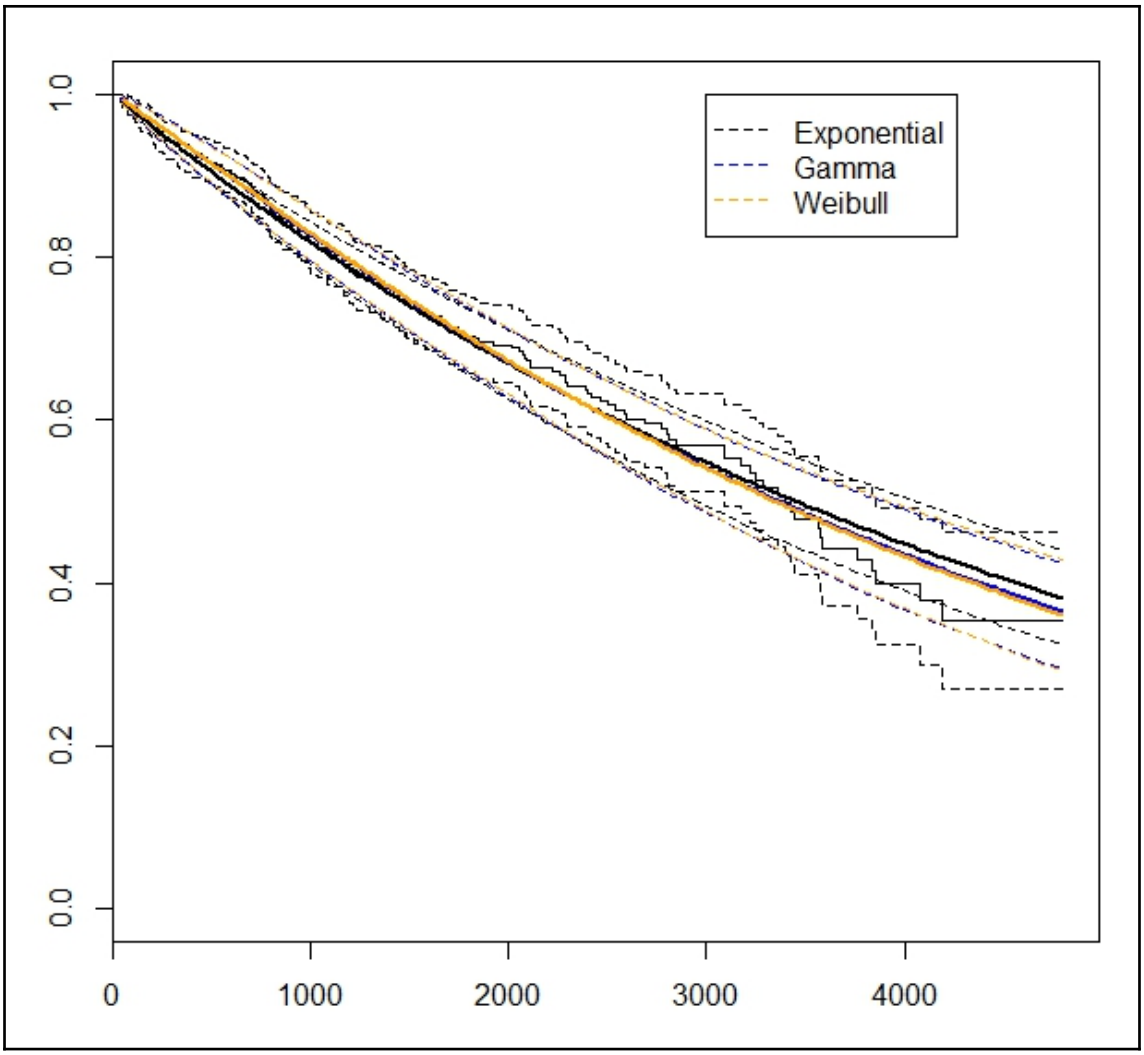

housing_RF
GrLivArea
Neighborhood
OverallQual
TotalBsmtSF
X1stFlrSF
BsmtFinSF1
X2ndFlrSF
GarageArea
GarageCars
BsmtFinType1
LotArea
MSSubClass
YearBuilt
ExterQual
KitchenQual
YearRemodAdd
FullBath
CentralAir
BsmtQual
MSZoning
Fireplaces
HalfBath
BsmtUnfSF
HouseStyle
TotRmsAbvGrd
Exterior1st
MasVnrArea
BldgType
OverallCond
Exterior2nd
5
10
15
20
25
30
35
%IncMSE
OverallQual
Neighborhood
GrLivArea
GarageCars
ExterQual
TotalBsmtSF
X1stFlrSF
BsmtQual
GarageArea
KitchenQual
YearBuilt
BsmtFinSF1
X2ndFlrSF
FullBath
LotArea
TotRmsAbvGrd
YearRemodAdd
Fireplaces
MasVnrArea
Exterior1st
BsmtUnfSF
OpenPorchSF
BsmtFinType1
WoodDeckSF
Exterior2nd
OverallCond
BsmtExposure
MoSold
MSSubClass
BedroomAbvGr
0.0e+00
5.0e+11
1.0e+12
1.5e+12
2.0e+12
IncNodePurity

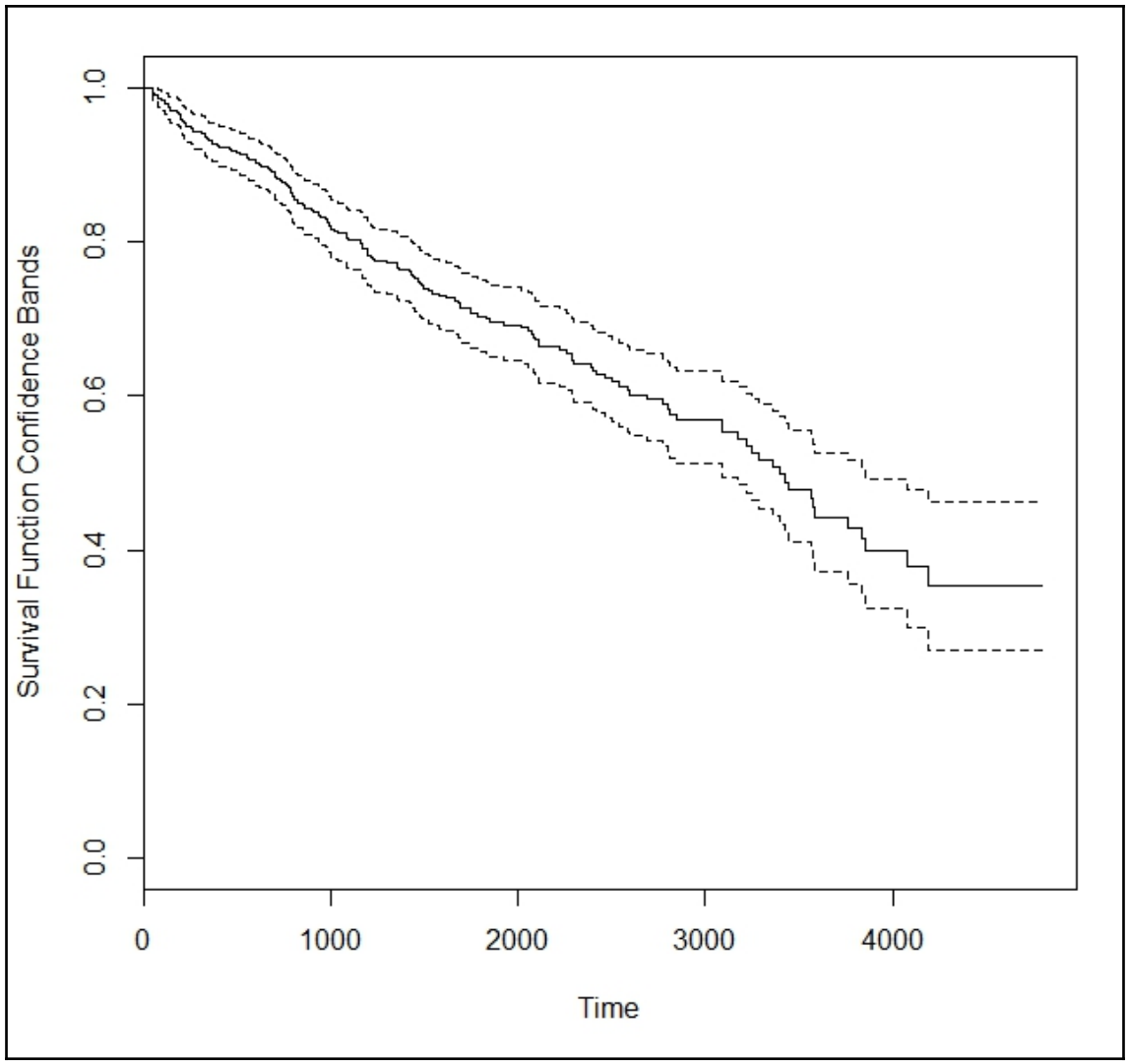

Squared error loss
1e+09
2e+09
3e+09
4e+09
5e+09
6e+09
OOB improvement in Squared error loss
1e+09
2e+09
3e+09
4e+09
0
200
400
600
800
1000
Iteration
OOB change in Squared error loss
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
0
200
400
600
800
1000
Iteration

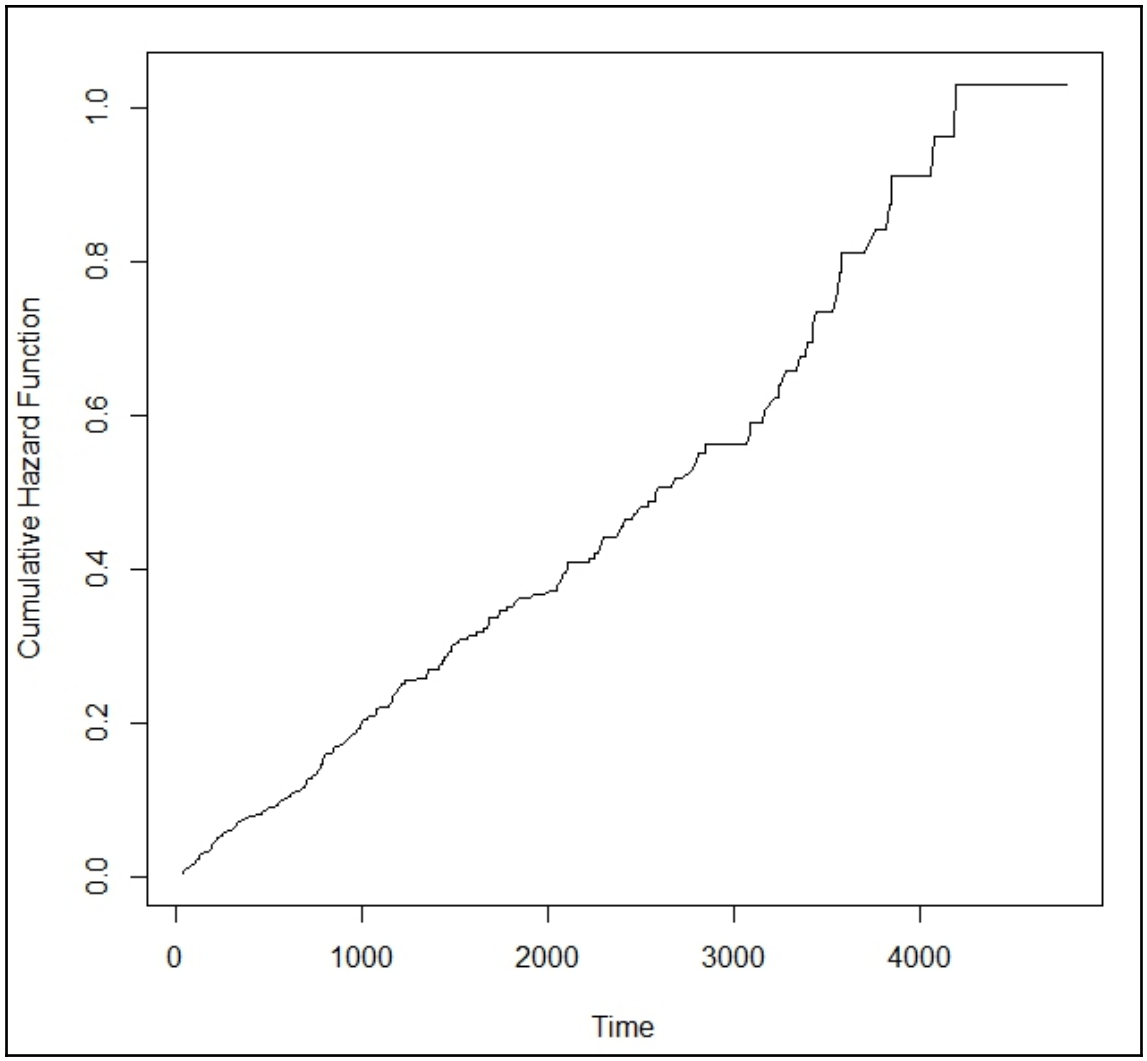

Rsquared
nnet
rpart
lm
MAE
RMSE
nnet
rpart
lm
0
20000
40000
60000
MAE
RMSE
Rsquared
Confidence Level: 0.95

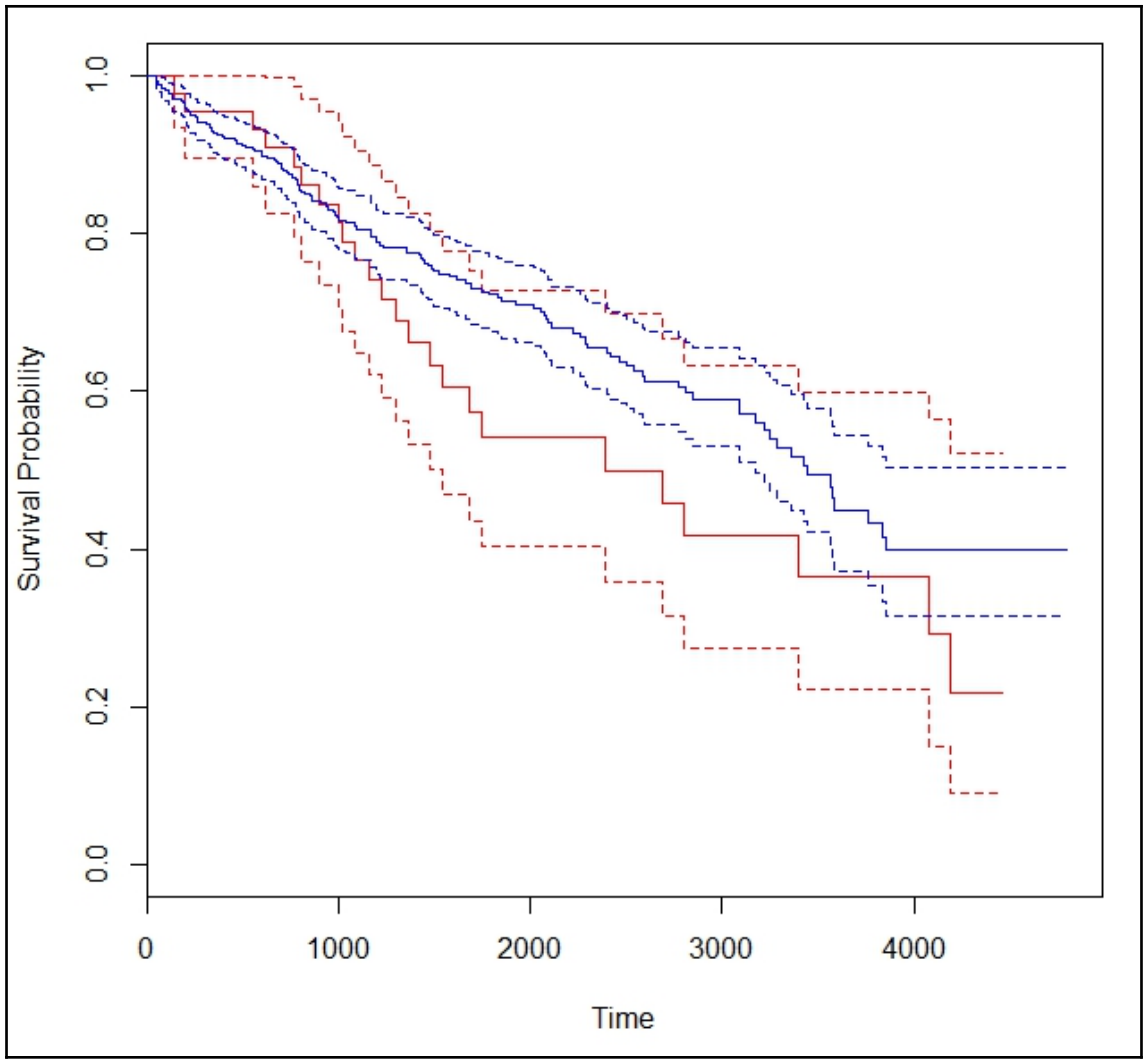
Chapter 10: Ensembling Survival Models

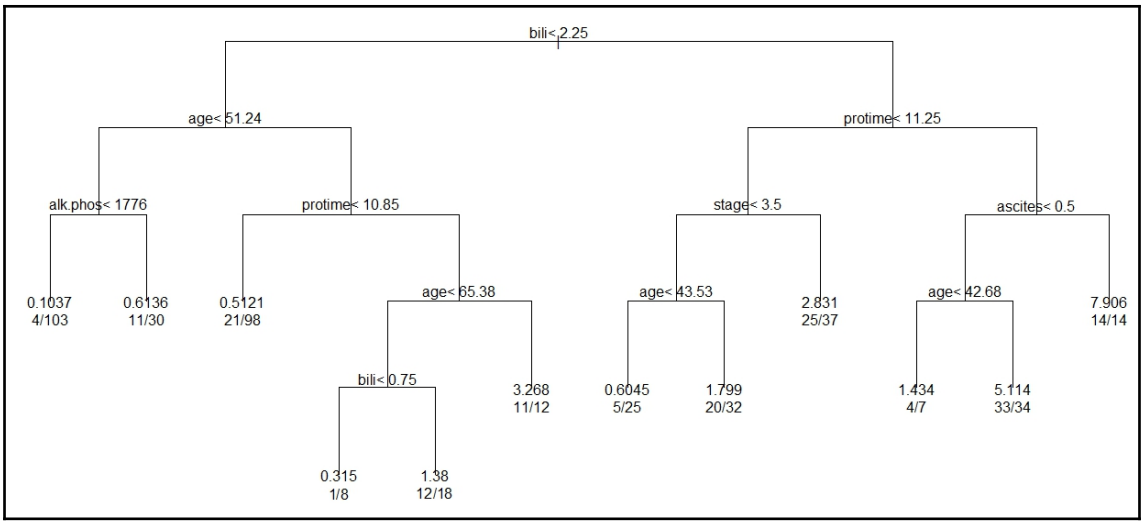


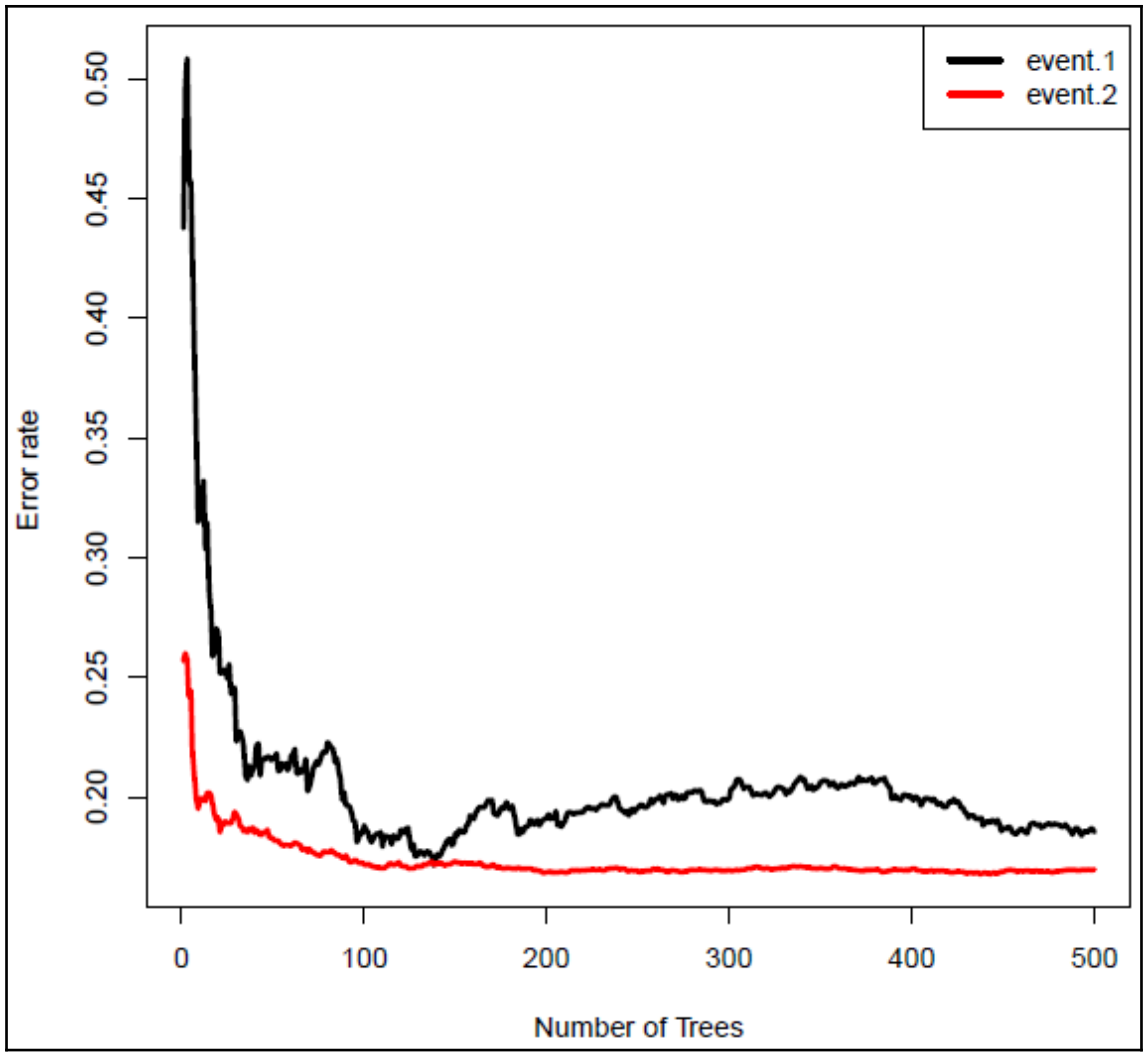

Exponential
Gamma
Weibull
1.0
0.8
0.6
0.4
0.2
0.0
0
1000
2000
3000
4000


Survival Function Confidence Bands
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
Time


Cumulative Hazard Function
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
Time


Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
1000
2000
3000
4000
Time


bili< 2.25
age< 51.24
protime< 11.25
alk.phos< 1776
protime< 10.85
stage< 3.5
ascites< 0.5
0.1037
4/103
0.6136
11/30
0.5121
21/98
age< 65.38
age< 43.53
2.831
25/37
age< 42.68
7.906
14/14
bili< 0.75
3.268
11/12
0.6045
5/25
1.799
20/32
1.434
4/7
5.114
33/34
0.315
1/8
1.38
12/18


event.1
event.2
Error rate
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0
100
200
300
400
500
Number of Trees