**Assignment- Random Forest**

1. Company Data: Problem Statement

A cloth manufacturing company is interested to know about the segment or attributes causes high sale.

Approach - A Random Forest can be built with target variable Sales (we will first convert it in categorical variable) & all other variable will be independent in the analysis.

Solution:

#Company Data -Cloth manufacturing Company

Company\_Data

library(randomForest)

model<-randomForest(Company\_Data$Sales~.,data=Company\_Data,ntree=1000)

#view the forest results

print(model)

#importance of the variable-Lower Gini

print(importance(model))

#prediction

pred<-predict(model,Company\_Data)

table(pred,Company\_Data$Sales)

Iteration in R:

library(randomForest)

> model<-randomForest(Company\_Data$Sales~.,data=Company\_Data,ntree=1000)

> #view the forest results

> print(model)

Call:

randomForest(formula = Company\_Data$Sales ~ ., data = Company\_Data, ntree = 1000)

Type of random forest: regression

Number of trees: 1000

No. of variables tried at each split: 3

Mean of squared residuals: 2.658447

% Var explained: 66.58

> #importance of the variable-Lower Gini

> print(importance(model))

IncNodePurity

CompPrice 289.28591

Income 229.23684

Advertising 280.39418

Population 182.92934

Price 757.78550

ShelveLoc 751.89443

Age 335.62981

Education 128.32773

Urban 25.49452

US 39.53052

pred<-predict(model,Company\_Data[,-1])

> table(pred,Company\_Data$Sales)

pred 0 0.16 0.37 0.53 0.91 1.42 1.82 2.05 2.07 2.23 2.34 2.52 2.66

2.16652355263158 1 0 0 0 0 0 0 0 0 0 0 0 0

2.32393011904763 0 1 0 0 0 0 0 0 0 0 0 0 0

.

.

.

. pred 13.39 13.44 13.55 13.91 14.37 14.9 15.63 16.27

2.16652355263158 0 0 0 0 0 0 0 0

2.32393011904763 0 0 0 0 0 0 0 0

[ reached getOption("max.print") -- omitted 398 rows ]

Iteration 2: Taking US as a variable

Company\_Data

library(randomForest)

model<-randomForest(Company\_Data$US~.,data=Company\_Data,ntree=1000)

#view the forest results

print(model)

#importance of the variable-Lower Gini

print(importance(model))

#prediction

pred<-predict(model,Company\_Data)

table(pred,Company\_Data$US)

Solution in R:

|  |
| --- |
| Call:  randomForest(formula = Company\_Data$US ~ ., data = Company\_Data, ntree = 1000)  Type of random forest: classification  Number of trees: 1000  No. of variables tried at each split: 3  OOB estimate of error rate: 9.25%  Confusion matrix:  No Yes class.error  No 124 18 0.12676056  Yes 19 239 0.07364341  > #importance of the variable-Lower Gini  > print(importance(model))  MeanDecreaseGini  Sales 14.885041  CompPrice 9.175715  Income 9.565801  Advertising 106.456139  Population 12.692884  Price 9.899213  ShelveLoc 2.273467  Age 11.066954  Education 5.543590  Urban 1.342522  > #prediction  > pred<-predict(model,Company\_Data)  > table(pred,Company\_Data$US)    pred No Yes  No 142 0  Yes 0 258 |
|  |
| |  | | --- | |  | |

Iteration 3: Using Urban as a variable;

#Company Data -Cloth manufacturing Company

Company\_Data

library(randomForest)

model<-randomForest(Company\_Data$Urban~.,data=Company\_Data,ntree=1000)

#view the forest results

print(model)

#importance of the variable-Lower Gini

print(importance(model))

#prediction

pred<-predict(model,Company\_Data)

table(pred,Company\_Data$Urban)

Solution in R:

Call:

randomForest(formula = Company\_Data$Urban ~ ., data = Company\_Data, ntree = 1000)

Type of random forest: classification

Number of trees: 1000

No. of variables tried at each split: 3

OOB estimate of error rate: 33%

Confusion matrix:

No Yes class.error

No 3 115 0.97457627

Yes 17 265 0.06028369

> #importance of the variable-Lower Gini

> print(importance(model))

MeanDecreaseGini

Sales 21.859341

CompPrice 20.550313

Income 25.413399

Advertising 12.208970

Population 24.907257

Price 19.913603

ShelveLoc 5.009606

Age 21.463756

Education 12.740419

US 1.819601

> #prediction

> pred<-predict(model,Company\_Data)

> table(pred,Company\_Data$Urban)

pred No Yes

No 118 0

Yes 0 282

1. Fraud Check: Problem Statement

Use Random Forest to prepare a model on fraud data treating those who have taxable\_income <= 30000 as "Risky" and others are "Good"

#Fraud Check new

Fraud\_check

library(randomForest)

model<-randomForest(Fraud\_check$Urban~.,data=Fraud\_check,ntree=1000)

#view the forest results

print(model)

#importance of the variable-Lower Gini

print(importance(model))

#prediction

#taxable income<=30000 as "Risky" and others are "Good"

type<-ifelse(Fraud\_check$Taxable.Income<=30000,"Risky","Good")

model<-data.frame(df,type)

plot(model$type)

barplot(table(model$type))

table(model$type)

**Iteration In R**

> library(randomForest)

> model<-randomForest(Fraud\_check$Urban~.,data=Fraud\_check,ntree=1000)

> #view the forest results

> print(model)

Call:

randomForest(formula = Fraud\_check$Urban ~ ., data = Fraud\_check, ntree = 1000)

Type of random forest: classification

Number of trees: 1000

No. of variables tried at each split: 2

OOB estimate of error rate: 43.5%

Confusion matrix:

NO YES class.error

NO 162 136 0.4563758

YES 125 177 0.4139073

> #importance of the variable-Lower Gini

> print(importance(model))

MeanDecreaseGini

Undergrad 10.90827

Marital.Status 17.60345

Taxable.Income 97.49096

City.Population 92.06635

Work.Experience 66.45454

> #prediction

> #taxable income<=30000 as "Risky" and others are "Good"

> type<-ifelse(Fraud\_check$Taxable.Income<=30000,"Risky","Good")

> model<-data.frame(df,type)

> plot(model$type)

> barplot(table(model$type))

> table(model$type)

Good Risky

476 124

Bar Plot:

