Anly501 Decision Tree

Yangyi Li

```
# we set align to be the label
# we want to see if any other elements affect the align
marvel <- read.csv("marvel.csv")
names(marvel) <- tolower(names(marvel))

head(marvel)
```

```
##   page_id                                    name
## 1    1678             spider-man (peter parker)
## 2    7139       captain america (steven rogers)
## 3   64786 wolverine (james \\"logan\\" howlett)
## 4    1868   iron man (anthony \\"tony\\" stark)
## 5    2460                    thor (thor odinson)
## 6    2458            benjamin grimm (earth-616)
##                                   urlslug               id              align
## 1             \\/spider-man_(peter_parker)  secret identity    good characters
## 2       \\/captain_america_(steven_rogers)  public identity    good characters
## 3 \\/wolverine_(james_%22logan%22_howlett)  public identity neutral characters
## 4   \\/iron_man_(anthony_%22tony%22_stark)  public identity    good characters
## 5                   \\/thor_(thor_odinson) no dual identity    good characters
## 6            \\/benjamin_grimm_(earth-616)  public identity    good characters
##          eye       hair             sex gsm             alive appearances
## 1 hazel eyes brown hair male characters     living characters        4043
## 2  blue eyes white hair male characters     living characters        3360
## 3  blue eyes black hair male characters     living characters        3061
## 4  blue eyes black hair male characters     living characters        2961
## 5  blue eyes blond hair male characters     living characters        2258
## 6  blue eyes    no hair male characters     living characters        2255
##   first.appearance year
## 1           aug-62 1962
## 2           mar-41 1941
## 3           oct-74 1974
## 4           mar-63 1963
## 5           nov-50 1950
## 6           nov-61 1961
```

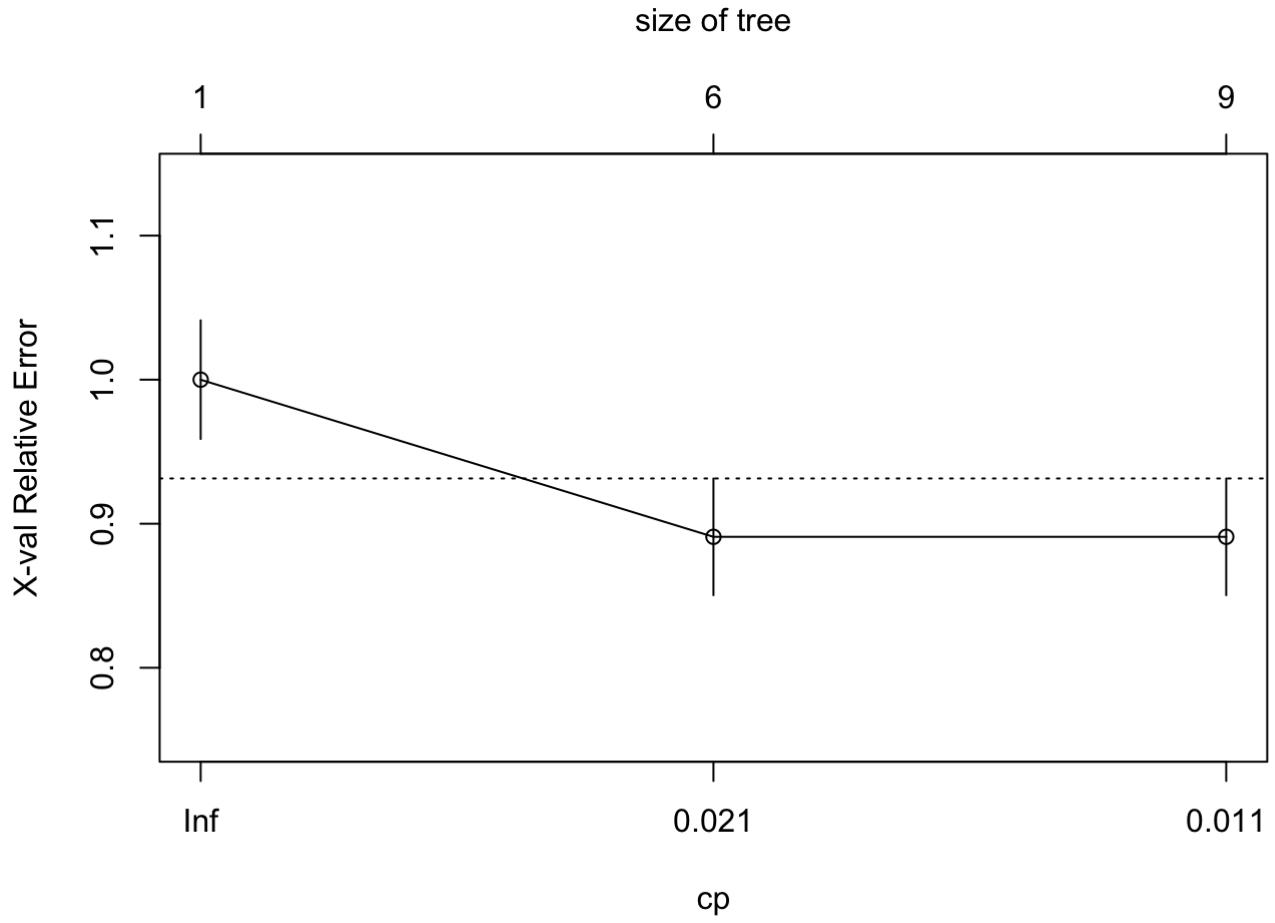
```
marvel <- marvel %>% select(id, align, sex, alive, appearances, year)
marvel <- marvel %>% select(align, everything())
marvel <- subset(marvel, align!="")
marvel <- subset(marvel, id!="")
marvel <- subset(marvel, sex!="")
marvel <- marvel %>% filter(sex == "male characters"|
           sex == "female characters")
marvel <- marvel %>% drop_na()
colnames(marvel)[1] <- "label"
marvel <- marvel[1:1000,]
head(marvel)
```

```
##                label               id             sex             alive
## 1    good characters  secret identity male characters living characters
## 2    good characters  public identity male characters living characters
## 3 neutral characters  public identity male characters living characters
## 4    good characters  public identity male characters living characters
## 5    good characters no dual identity male characters living characters
## 6    good characters  public identity male characters living characters
##   appearances year
## 1        4043 1962
## 2        3360 1941
## 3        3061 1974
## 4        2961 1963
## 5        2258 1950
## 6        2255 1961
```

```
DataSize=nrow(marvel)
TrainingSet_Size<-floor(DataSize*(3/4))
TestSet_Size <- DataSize - TrainingSet_Size

MyTrainSample <- sample(nrow(marvel),TrainingSet_Size,replace=FALSE)
MyTrainingSET <- marvel[MyTrainSample,]
MyTestSET <- marvel[-MyTrainSample,]
TestKnownLabels <- MyTestSET$label
```

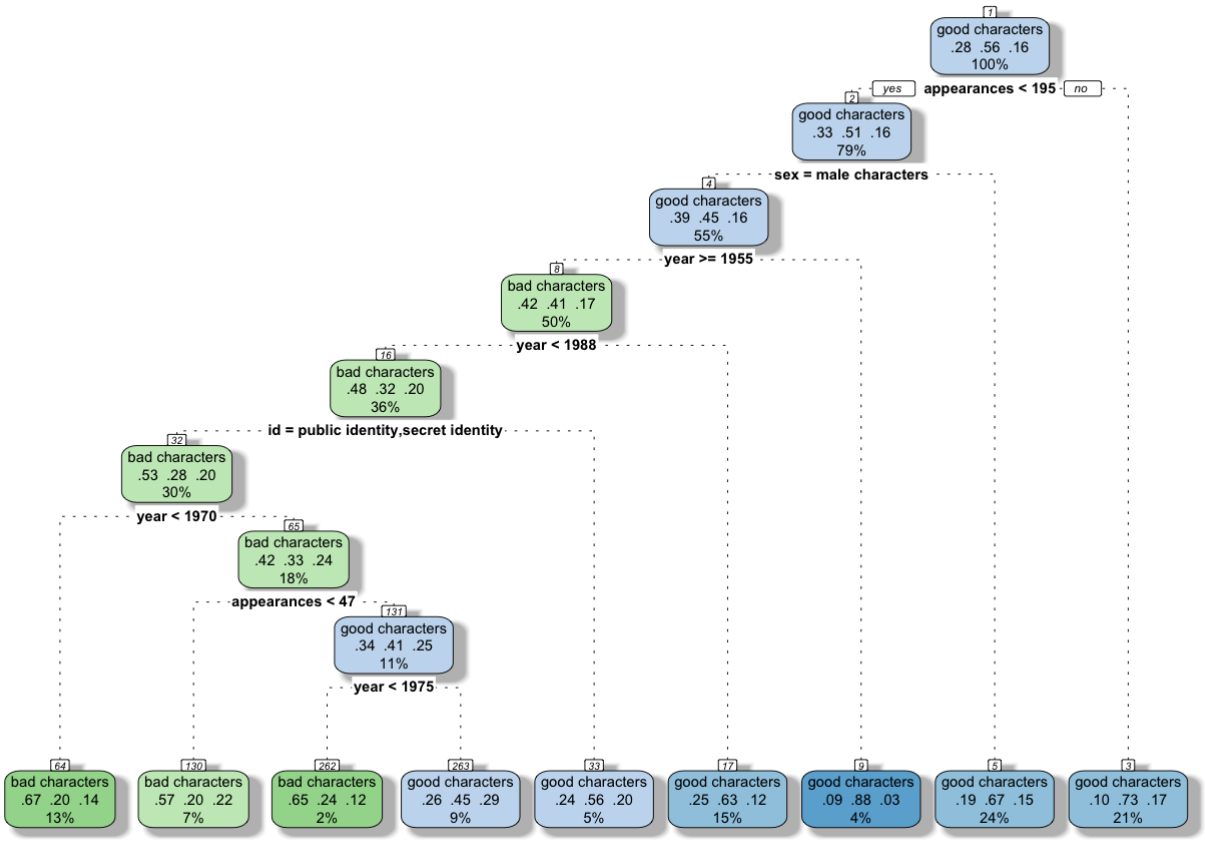
```
DT <- rpart(MyTrainingSET$label ~ ., data = MyTrainingSET, method="class")
plotcp(DT)
```




```
DT_Prediction= predict(DT, MyTestSET, type="class")
table(DT_Prediction,TestKnownLabels)
```

```
##                    TestKnownLabels
## DT_Prediction       bad characters good characters neutral characters
##   bad characters                24              14                   8
##   good characters               33             137                  34
##   neutral characters             0               0                   0
```

```
fancyRpartPlot(DT)
```



Rattle 2021-Nov-12 04:31:44 liyangyi

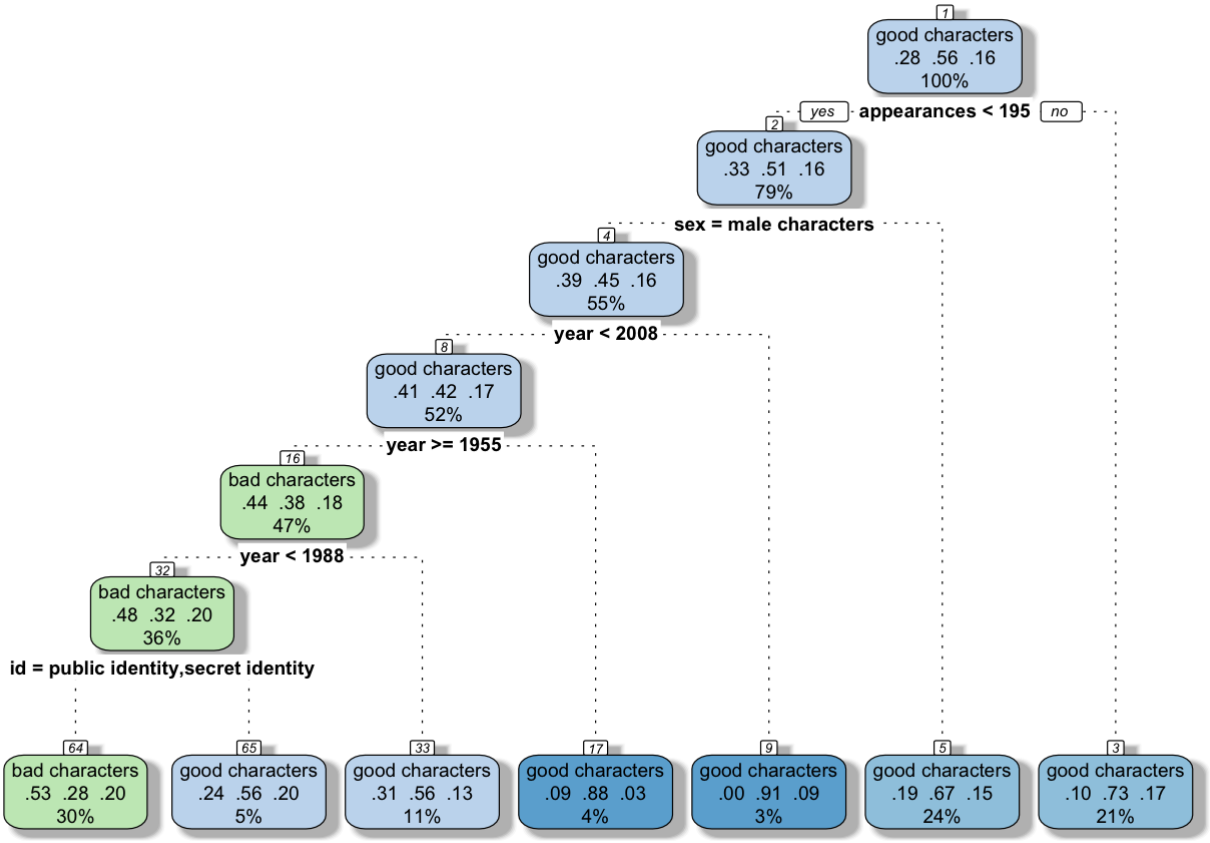
```
png(file="plot1.png", width=1600, height=1350)
rattle::fancyRpartPlot(DT,main="Decision Tree")
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
DT2 <- rpart(MyTrainingSET$label ~ ., data = MyTrainingSET,cp = .01, parms = list(split=
"information"), method="class")
DT_Prediction= predict(DT2, MyTestSET, type="class")
table(DT_Prediction,TestKnownLabels)
```

```
##                    TestKnownLabels
## DT_Prediction       bad characters good characters neutral characters
##   bad characters                28              20                15
##   good characters               29             131                27
##   neutral characters             0               0                 0
```

```
fancyRpartPlot(DT2)
```



Rattle 2021-Nov-12 04:31:48 liyangyi

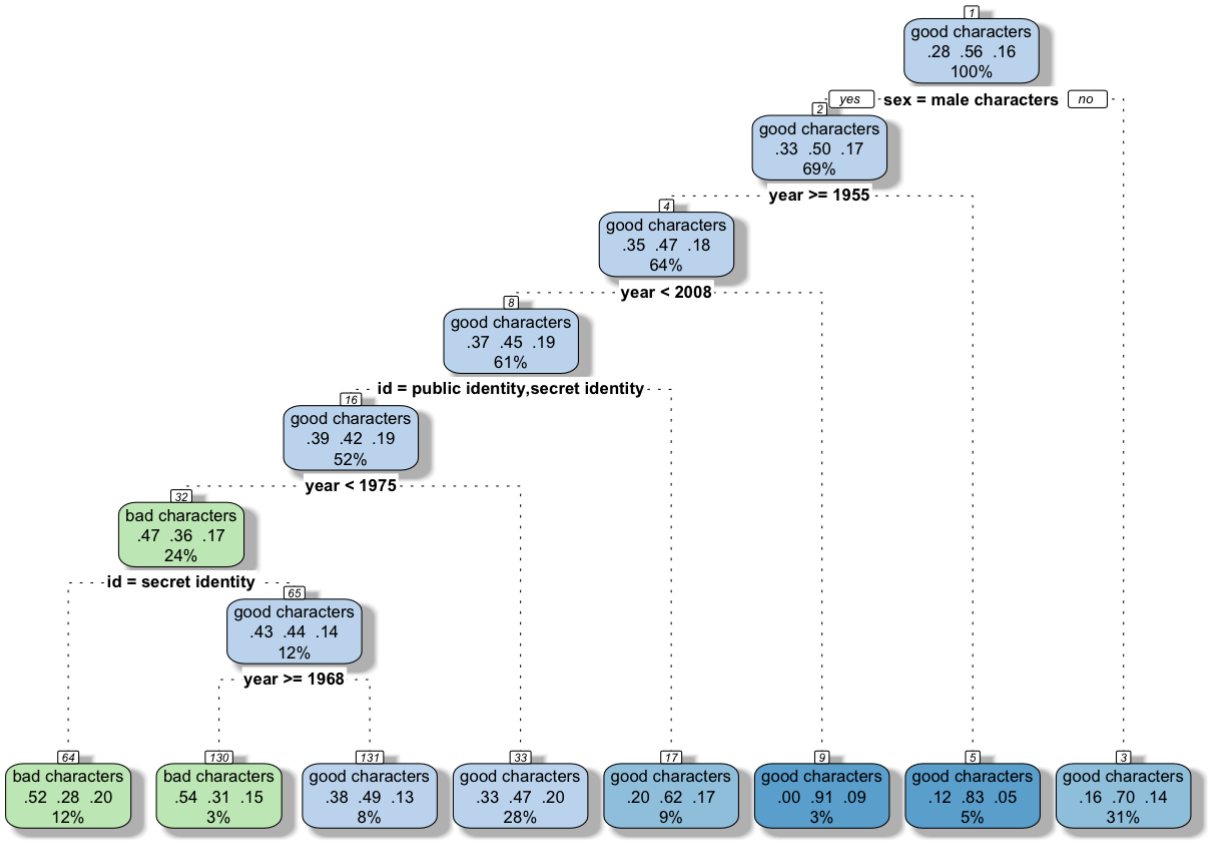
```
png(file="plot2.png", width=1600, height=1350)
rattle::fancyRpartPlot(DT2,main="Decision Tree")
dev.off()
```

```
## quartz_off_screen
##                 2
```

```
DT3 <- rpart(MyTrainingSET$label ~ year+sex+id, data = MyTrainingSET, method="class")
DT_Prediction= predict(DT3, MyTestSET, type="class")
table(DT_Prediction,TestKnownLabels)
```

```
##                    TestKnownLabels
## DT_Prediction       bad characters good characters neutral characters
##   bad characters                 9              11                  8
##   good characters               48             140                 34
##   neutral characters             0               0                  0
```

```
fancyRpartPlot(DT3)
```



Rattle 2021-Nov-12 04:31:52 liyangyi

```
png(file="plot3.png", width=1600, height=1350)
rattle::fancyRpartPlot(DT3,main="Decision Tree")
dev.off()
```

```
## quartz_off_screen
##                 2
```