Drzewa klasyfikacyjne

Kamil Kukiełka

2023-05-22

Przygotowanie danych

Na początek zaimportujmy naszą bazę danych oraz potrzebne bibilioteki

```
library(readxl)
dane <- read_excel("data.xlsx")
library(MASS)
library(maptree)
library(rpart)
library(rpart.plot)
library(party)
library(class)
```

Tworzymy ramkę danych a następnie usuwamy braki danych

```
raw_data <- data.frame(dane)
data <- na.omit(data.frame(raw_data))
head(data)
```

```
##     Age Income Marital.Status Education.Level Approved.Loan
## 1 Young   High         Single            High            No
## 2 Young Medium         Single            High           Yes
## 3 Young Medium         Single          Medium           Yes
## 4 Young Medium        Married          Medium            No
## 5   Old   High        Married             Low            No
## 6   Old Medium        Married             Low           Yes
```

Tworzenie zbioru uczącego oraz testowego

Aby utworzyć nasze zbiry uczący bierzemy w sposób losowy rekordy z naszej ramki danych i przydzielamy je do naszych zbiorów

```
indexes <- sample(1:nrow(dane),nrow(dane)/2,replace = FALSE)

ZU = dane[indexes, ]
ZT = dane[-indexes, ]

head(ZU)
```

```
## # A tibble: 6 x 5
##   Age    Income `Marital Status` `Education Level` `Approved Loan`
##   <chr>  <chr>  <chr>            <chr>             <chr>
## 1 Old    High   Married          High              No
## 2 Young  High   Single           Medium            Yes
## 3 Young  High   Married          Medium            No
## 4 Young  High   Married          High              No
## 5 Medium Medium Married          Medium            Yes
## 6 Medium High   Married          Low               Yes
```

```
head(ZT)
```

```
## # A tibble: 6 x 5
##   Age    Income `Marital Status` `Education Level` `Approved Loan`
##   <chr>  <chr>  <chr>            <chr>             <chr>
## 1 Young  High   Single           High              No
## 2 Old    High   Married          Low               No
## 3 Medium Low    Married          High              Yes
## 4 Old    Medium Single           Low               No
## 5 Medium High   Married          Low               Yes
## 6 Medium Medium Married          Low               Yes
```

Tworzymy model drzewa

Tworzymy go za pomocą funkcji rpart, wykorzystyjąc zbiór uczący. Jak że nasze zmienne są jakościowe wybieramy metodę "class"

```
drzewo <- rpart(ZU$`Approved Loan`~.,ZU,method = "class")
```

Dla naszego drzewa możemy sprawdzić jego parametry oraz liczbę gałęzi

```
drzewo$parms
```

```
## $prior
##   1   2
## 0.5 0.5
##
## $loss
##      [,1] [,2]
## [1,]    0    1
## [2,]    1    0
##
## $split
## [1] 1
```

```
drzewo$numresp
```

```
## [1] 4
```

Jako że wygenerowane potrafią być bardzo rozległe możemy spróbować je przyciąć. W naszym przypadku liczba gałęzi jest dość mała, jednak i tak możemy sprawdzić czy przycięcie go nie sprawi że będzie lepsze

```
model.opt<-which.min(drzewo$cptable[,4])
cp.opt<-drzewo$cptable[model.opt,1]
drzewo2<-prune(drzewo,cp=1)
```

Weryfikacja naszych modeli

Teraz kiedy mamy już nasze modele możemy sprawdzić ich dopasowanie do zbioru uczącego dla naszego pierwotnego drzewa

```
Pred1<-predict(drzewo,ZU,type = "class")

print("tabela dobroci klasyfikacji")
table(predykacja=Pred1,prawdziwe=ZU$`Approved Loan`)

print("obliczanie błędu predykcji")
blad1<-mean(Pred1 != ZU$`Approved Loan`)
blad1
```

```
## [1] "tabela dobroci klasyfikacji"
##           prawdziwe
## predykacja No Yes
##        No  11   4
##        Yes 14  21
## [1] "obliczanie błędu predykcji"
## [1] 0.36
```

oraz dla przyciętego drzewa

```
Pred2<-predict(drzewo2,ZU,type = "class")

print("tabela dobroci klasyfikacji")
table(predykacja=Pred2,prawdziwe=ZU$`Approved Loan`)

print("obliczanie błędu predykcji")
blad2<-mean(Pred2 != ZU$`Approved Loan`)
blad2
```

```
## [1] "tabela dobroci klasyfikacji"
##           prawdziwe
## predykacja No Yes
##        No  25  25
##        Yes  0   0
## [1] "obliczanie błędu predykcji"
## [1] 0.5
```

Jak możemy zauważyć nasze przycięte drzewo daje wiekszy błąd predykcji, jednak nie pozbywajmy się go jeszcze, gdyż może mieć znacznie mniejszy błąd predykcji na zbiorze testowym ## Testowanie na zbiorze testowym

```
TPred1<-predict(drzewo,ZT,type = "class")
bladT1<-mean(TPred1 != ZT$`Approved Loan`)
print("Błąd predykcji drzewa")
bladT1

TPred2<-predict(drzewo2,ZT,type = "class")
bladT2<-mean(TPred2 != ZT$`Approved Loan`)
print("Błąd predykcji drzewa Przyciętego")
bladT2
```

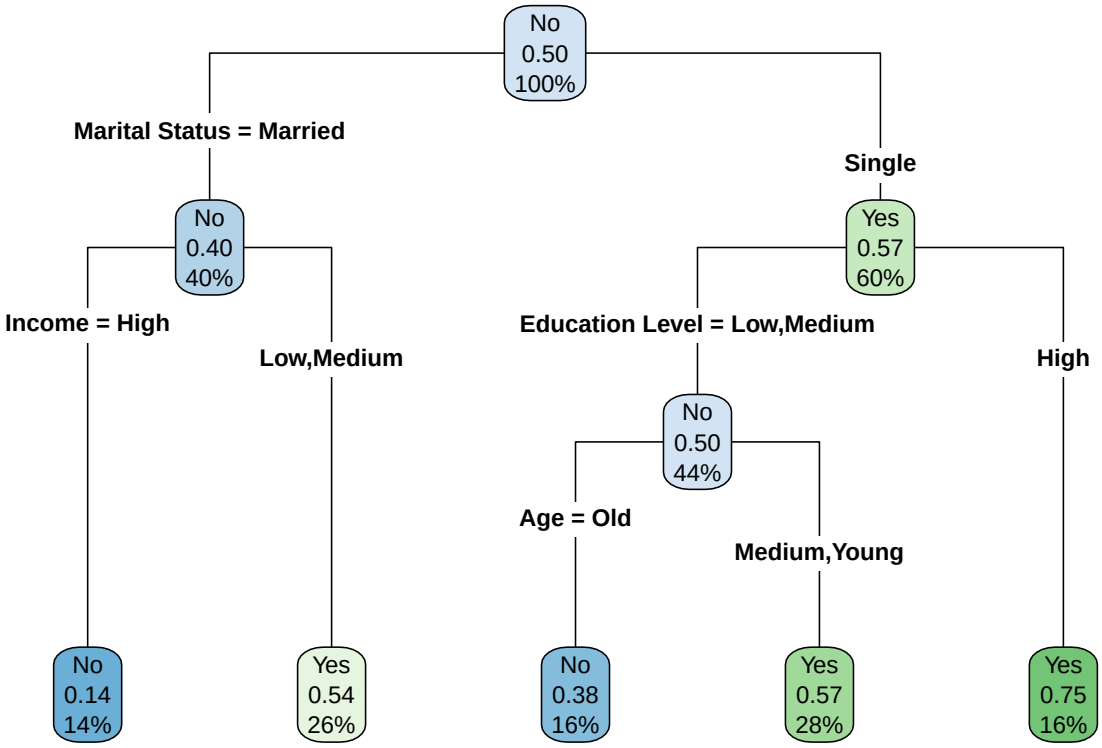
```
## [1] "Błąd predykcji drzewa"
## [1] 0.48
## [1] "Błąd predykcji drzewa Przyciętego"
## [1] 0.4
```

Teraz możemy ocenić który z naszych modeli mamy wybrać. Nalezy jednak pamiętać, że podział na zbior uczący i testowy jest losowy więc po ponownym odpaleniu kodu możemy uzyskać inny rezultat

Rysowanie drzewa

Kiedy zdecydujemy, który model jest lepszy możemy go narsyować, aby bardziej zwizualizować sobie jak działa dany model

```
rpart.plot(drzewo,type = 4,extra="auto")
```



```
rpart.plot(drzewo2,type = 4,extra="auto")
```