Income Prediction with Decision Trees

Business Understanding

Objective is to use information about existing customers of a financial services company to develop a model that predicts whether a customer has an income of $50,000 or more. The motivation for this problem is to identify potential high-income customers from a prospective customer database that was recently purchase

This problems is going to be solved using a classification tree

Data Understanding

library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.2 ✓ purrr 0.3.4  
## ✓ tibble 3.0.3 ✓ dplyr 1.0.2  
## ✓ tidyr 1.1.2 ✓ stringr 1.4.0  
## ✓ readr 1.3.1 ✓ forcats 0.5.0

## ── Conflicts ───────────────────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

income <- read\_csv("Student/Data/income.csv", col\_types = "nffnfffffnff")  
glimpse(income)

## Rows: 32,560  
## Columns: 12  
## $ age <dbl> 50, 38, 53, 28, 37, 49, 52, 31, 42, 37, 30, 23, 32…  
## $ workClassification <fct> Self-emp-not-inc, Private, Private, Private, Priva…  
## $ educationLevel <fct> Bachelors, HS-grad, 11th, Bachelors, Masters, 9th,…  
## $ educationYears <dbl> 13, 9, 7, 13, 14, 5, 9, 14, 13, 10, 13, 13, 12, 11…  
## $ maritalStatus <fct> Married-civ-spouse, Divorced, Married-civ-spouse, …  
## $ occupation <fct> Exec-managerial, Handlers-cleaners, Handlers-clean…  
## $ relationship <fct> Husband, Not-in-family, Husband, Wife, Wife, Not-i…  
## $ race <fct> White, White, Black, Black, White, Black, White, W…  
## $ gender <fct> Male, Male, Male, Female, Female, Female, Male, Fe…  
## $ workHours <dbl> 13, 40, 40, 40, 40, 16, 45, 50, 40, 80, 40, 30, 50…  
## $ nativeCountry <fct> United-States, United-States, United-States, Cuba,…  
## $ income <fct> <=50K, <=50K, <=50K, <=50K, <=50K, <=50K, >50K, >5…

Exploring the data

summary(income)

## age workClassification educationLevel educationYears   
## Min. :17.00 Private :22696 HS-grad :10501 Min. : 1.00   
## 1st Qu.:28.00 Self-emp-not-inc: 2541 Some-college: 7291 1st Qu.: 9.00   
## Median :37.00 Local-gov : 2093 Bachelors : 5354 Median :10.00   
## Mean :38.58 ? : 1836 Masters : 1723 Mean :10.08   
## 3rd Qu.:48.00 State-gov : 1297 Assoc-voc : 1382 3rd Qu.:12.00   
## Max. :90.00 Self-emp-inc : 1116 11th : 1175 Max. :16.00   
## (Other) : 981 (Other) : 5134   
## maritalStatus occupation relationship   
## Married-civ-spouse :14976 Prof-specialty :4140 Husband :13193   
## Divorced : 4443 Craft-repair :4099 Not-in-family : 8304   
## Married-spouse-absent: 418 Exec-managerial:4066 Wife : 1568   
## Never-married :10682 Adm-clerical :3769 Own-child : 5068   
## Separated : 1025 Sales :3650 Unmarried : 3446   
## Married-AF-spouse : 23 Other-service :3295 Other-relative: 981   
## Widowed : 993 (Other) :9541   
## race gender workHours   
## White :27815 Male :21789 Min. : 1.00   
## Black : 3124 Female:10771 1st Qu.:40.00   
## Asian-Pac-Islander: 1039 Median :40.00   
## Amer-Indian-Eskimo: 311 Mean :40.44   
## Other : 271 3rd Qu.:45.00   
## Max. :99.00   
##   
## nativeCountry income   
## United-States:29169 <=50K:24719   
## Mexico : 643 >50K : 7841   
## ? : 583   
## Philippines : 198   
## Germany : 137   
## Canada : 121   
## (Other) : 1709

Data Preparation

Splitting test and train data

set.seed(1234)  
sample\_set <- sample(nrow(income), round(nrow(income) \* .75), replace = FALSE)  
  
income\_train <- income[sample\_set, ]  
income\_test <- income[-sample\_set, ]

Review data set balance

income %>%  
 select(income) %>%  
 table(exclude = NULL) %>%  
 prop.table() %>%  
 round(4) \* 100

## .  
## <=50K >50K   
## 75.92 24.08

Review test data set balance

income\_test %>%  
 select(income) %>%  
 table(exclude = NULL) %>%  
 prop.table() %>%  
 round(4) \* 100

## .  
## <=50K >50K   
## 76.89 23.11

Review train data set balance

income\_train %>%  
 select(income) %>%  
 table(exclude = NULL) %>%  
 prop.table() %>%  
 round(4) \* 100

## .  
## <=50K >50K   
## 75.59 24.41

Balancing the Training Data Set

library(DMwR)

## Loading required package: lattice

## Loading required package: grid

## Registered S3 method overwritten by 'quantmod':  
## method from  
## as.zoo.data.frame zoo

set.seed(1234)  
  
income\_train <- SMOTE(income ~ ., data.frame(income\_train), perc.over = 100, perc.under = 200)  
  
income\_train %>%  
 select(income) %>%  
 table(exclude = NULL) %>%  
 prop.table() %>%  
 round(4) \* 100

## .  
## <=50K >50K   
## 50 50

Modeling

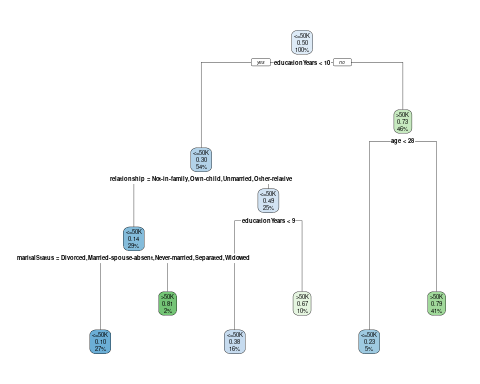
Training the model

library(rpart)  
  
income\_mod <-   
 rpart (  
 income ~ .,   
 method = "class",   
 data = income\_train  
 )

Model Evaluation

Plot the Decision Tree

library(rpart.plot)  
rpart.plot(income\_mod)



Confusion Matrix

income\_pred <- predict(income\_mod, income\_test, type = "class")  
income\_pred\_table <- table(income\_test$income, income\_pred)  
income\_pred\_table

## income\_pred  
## <=50K >50K  
## <=50K 4765 1494  
## >50K 546 1335

Prediction Accuracy

sum(diag(income\_pred\_table)) / nrow(income\_test)

## [1] 0.7493857