Estimating Medicare Costs for AMI

Team 02 (Aditya Dube and Richard Budden)

4/15/2023

#-#-# Predicting Medicare for AMI #-#-#

```
  ## Step 0: Basic Setup - Install Packages / Load Libraries

  ## Step 1: Collecting Data
             part a) Import Datasets
             part b) Joining the Datasets

  ## Step 2: Explore / Prepare Data
             part a) Remove, Code, and/or Impute Data
             part b) Variable Selection
             part c) Group by State/Summarize Data

  ## Step 3: Visualization of Data
             part a) Histogram
             part b) United States Heat Map

  ## Step 4: Creating Training and Test Datasets

  ## Step 5: Build and Evaluate Linear Regression Model

  ## Step 6: Build and Evaluate CART Model

  ## Step 7: Build and Evaluate Artifical Neural Network  Model (Feeforward ANN)
```

```
## Step 0: Basic Setup - Install Packages / Load Libraries
```

```r
#install.packages("tree")

library(dplyr)
library(tidyverse)
library(dplyr)
library(ggplot2)
library(maps)
library(caret)
library(rpart)
library(rpart.plot)
library(rattle)
library(neuralnet)
#conflict_prefer("prune", "rpart")
```

```
## Step 1: Collecting Data       part a) Import Datasets
```

```r
# Load the three datasets
hospital_data <- read.csv("Hospital General Information.csv")
medicare_data <-
↪  read.csv("Medicare_Inpatient_Hospital_by_Provider_and_Service_2018_data.csv")
census_data <- read.csv("Census data.csv")
```

Initial look at the Structure of the Datasets

```r
##  Structure of the dataset
str(hospital_data)
```

```
## 'data.frame':    4793 obs. of  13 variables:
##  $ Provider_ID                              : int  10001 10005 10006 10007 10008 10011 10012 10016 1
##  $ Hospital.Name                            : chr  "SOUTHEAST ALABAMA MEDICAL CENTER" "MARSHALL MEDI
##  $ Address                                  : chr  "1108 ROSS CLARK CIRCLE" "2505 U S HIGHWAY 431 NO
##  $ City                                     : chr  "DOTHAN" "BOAZ" "FLORENCE" "OPP" ...
##  $ State                                    : chr  "AL" "AL" "AL" "AL" ...
##  $ ZIP_Code                                 : int  36301 35957 35631 36467 36049 35235 35968 35007 3
##  $ County.Name                              : chr  "HOUSTON" "MARSHALL" "LAUDERDALE" "COVINGTON" ...
##  $ Phone.Number                             : num  3.35e+09 2.57e+09 2.57e+09 3.34e+09 3.34e+09 ...
##  $ Hospital.Type                            : chr  "Acute Care Hospitals" "Acute Care Hospitals" "Ac
##  $ Hospital.Ownership                       : chr  "Government - Hospital District or Authority" "Go
##  $ Emergency.Services                       : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Meets.criteria.for.meaningful.use.of.EHRs: chr  "Y" "Y" "Y" "Y" ...
##  $ Hospital.overall.rating                  : chr  "3" "2" "2" "2" ...
```

```r
str(medicare_data)
```

```
## 'data.frame':    193003 obs. of  15 variables:
##  $ Provider_ID             : int  10001 10001 10001 10001 10001 10001 10001 10001 10001 10001 ...
##  $ provider_name           : chr  "Southeast Alabama Medical Center" "Southeast Alabama Medical Cen
##  $ street_address          : chr  "1108 Ross Clark Circle" "1108 Ross Clark Circle" "1108 Ross Clar
##  $ Rndrng_Prvdr_City       : chr  "Dothan" "Dothan" "Dothan" "Dothan" ...
##  $ Rndrng_Prvdr_State_Abrvtn: chr  "AL" "AL" "AL" "AL" ...
##  $ Rndrng_Prvdr_State_FIPS : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ ZIP_Code                : int  36301 36301 36301 36301 36301 36301 36301 36301 36301 36301 ...
##  $ Rndrng_Prvdr_RUCA       : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Rndrng_Prvdr_RUCA_Desc   : chr  "Metropolitan area core: primary flow within an urbanized area of
##  $ DRG_Cd                   : int  3 23 25 38 39 57 64 65 66 69 ...
##  $ DRG                      : chr  "\"ECMO OR TRACH W MV >96 HRS OR PDX EXC FACE, MOUTH & NECK W MAJ
##  $ Total_discharges         : int  13 33 26 11 64 30 115 107 17 53 ...
##  $ Ave_covered_charges      : num  368434 148677 118718 74449 46628 ...
##  $ Ave_total_payment        : num  81541 29062 22442 9546 6468 ...
##  $ Ave_medical_payment      : num  80435 27997 19592 7562 5073 ...
```

```r
str(census_data)
```

```
## 'data.frame':    33120 obs. of  2 variables:
##  $ ZIP_Code: chr  "8600000US00601" "8600000US00602" "8600000US00603" "8600000US00606" ...
##  $ NAME    : chr  "ZCTA5 00601" "ZCTA5 00602" "ZCTA5 00603" "ZCTA5 00606" ...
```

## Step 1: Collecting Data   part b) Joining the Datasets

Used the dplyr package in R to join multiple datasets based on a common variable. We Joined the medicare and hospital data by Provider ID and by Zip code

```r
final_data <- inner_join(medicare_data, hospital_data, by = c("Provider_ID","ZIP_Code"))
```

## Step 2: Explore / Prepare Data   part a) Remove, Code, and/or Impute Data

---

Data cleaning: Before analyzing the data, we need to clean it by removing missing values, fixing formatting issues, and dealing with outliers

```r
# Remove rows with missing values
final_data <- final_data[complete.cases(final_data), ]
head(final_data)
```

```
##   Provider_ID                    provider_name         street_address
## 1       10001 Southeast Alabama Medical Center 1108 Ross Clark Circle
## 2       10001 Southeast Alabama Medical Center 1108 Ross Clark Circle
## 3       10001 Southeast Alabama Medical Center 1108 Ross Clark Circle
## 4       10001 Southeast Alabama Medical Center 1108 Ross Clark Circle
## 5       10001 Southeast Alabama Medical Center 1108 Ross Clark Circle
## 6       10001 Southeast Alabama Medical Center 1108 Ross Clark Circle
##   Rndrng_Prvdr_City Rndrng_Prvdr_State_Abrvtn Rndrng_Prvdr_State_FIPS ZIP_Code
## 1            Dothan                        AL                       1    36301
## 2            Dothan                        AL                       1    36301
## 3            Dothan                        AL                       1    36301
## 4            Dothan                        AL                        1    36301
## 5            Dothan                        AL                        1    36301
## 6            Dothan                        AL                        1    36301
##   Rndrng_Prvdr_RUCA
## 1                 1
## 2                 1
## 3                 1
## 4                 1
## 5                 1
## 6                 1
##                                                                 Rndrng_Prvdr_RUCA_Desc
## 1 Metropolitan area core: primary flow within an urbanized area of 50,000 and greater
## 2 Metropolitan area core: primary flow within an urbanized area of 50,000 and greater
## 3 Metropolitan area core: primary flow within an urbanized area of 50,000 and greater
## 4 Metropolitan area core: primary flow within an urbanized area of 50,000 and greater
## 5 Metropolitan area core: primary flow within an urbanized area of 50,000 and greater
## 6 Metropolitan area core: primary flow within an urbanized area of 50,000 and greater
##   DRG_Cd                                                                  DRG
## 1      3 "ECMO OR TRACH W MV >96 HRS OR PDX EXC FACE, MOUTH & NECK W MAJ O.R.
## 2     23  CRANIOTOMY W MAJOR DEVICE IMPLANT OR ACUTE COMPLEX CNS PDX W MCC OR
## 3     25              CRANIOTOMY & ENDOVASCULAR INTRACRANIAL PROCEDURES W MCC
## 4     38                                     EXTRACRANIAL PROCEDURES W CC
## 5     39                               EXTRACRANIAL PROCEDURES W/O CC/MCC
## 6     57                    DEGENERATIVE NERVOUS SYSTEM DISORDERS W/O MCC
##   Total_discharges Ave_covered_charges Ave_total_payment Ave_medical_payment
## 1               13           368434.00         81540.923           80434.923
## 2               33           148677.12         29061.515           27996.576
## 3               26           118718.35         22441.769           19591.808
## 4               11            74449.18          9546.000            7561.818
## 5               64            46627.78          6468.297            5073.297
## 6               30            27139.97          6204.733            5178.900
##                      Hospital.Name                Address   City State
## 1 SOUTHEAST ALABAMA MEDICAL CENTER 1108 ROSS CLARK CIRCLE DOTHAN    AL
## 2 SOUTHEAST ALABAMA MEDICAL CENTER 1108 ROSS CLARK CIRCLE DOTHAN    AL
## 3 SOUTHEAST ALABAMA MEDICAL CENTER 1108 ROSS CLARK CIRCLE DOTHAN    AL
## 4 SOUTHEAST ALABAMA MEDICAL CENTER 1108 ROSS CLARK CIRCLE DOTHAN    AL
## 5 SOUTHEAST ALABAMA MEDICAL CENTER 1108 ROSS CLARK CIRCLE DOTHAN    AL
## 6 SOUTHEAST ALABAMA MEDICAL CENTER 1108 ROSS CLARK CIRCLE DOTHAN    AL
##   County.Name Phone.Number        Hospital.Type
## 1     HOUSTON   3347938701 Acute Care Hospitals
## 2     HOUSTON   3347938701 Acute Care Hospitals
## 3     HOUSTON   3347938701 Acute Care Hospitals
## 4     HOUSTON   3347938701 Acute Care Hospitals
## 5     HOUSTON   3347938701 Acute Care Hospitals
## 6     HOUSTON   3347938701 Acute Care Hospitals
##                            Hospital.Ownership Emergency.Services
## 1 Government - Hospital District or Authority                Yes
## 2 Government - Hospital District or Authority                Yes
## 3 Government - Hospital District or Authority                Yes
## 4 Government - Hospital District or Authority                Yes
## 5 Government - Hospital District or Authority                Yes
## 6 Government - Hospital District or Authority                Yes
##   Meets.criteria.for.meaningful.use.of.EHRs Hospital.overall.rating
## 1                                         Y                       3
## 2                                              Y                        3
## 3                                              Y                        3
## 4                                              Y                        3
## 5                                              Y                        3
## 6                                              Y                        3
```

```
      ## Step 2: Explore / Prepare Data     part b) Variable Selection
```

Data wrangling and variable selection: We use the tidyverse package in R to manipulate and select variables from our dataset.

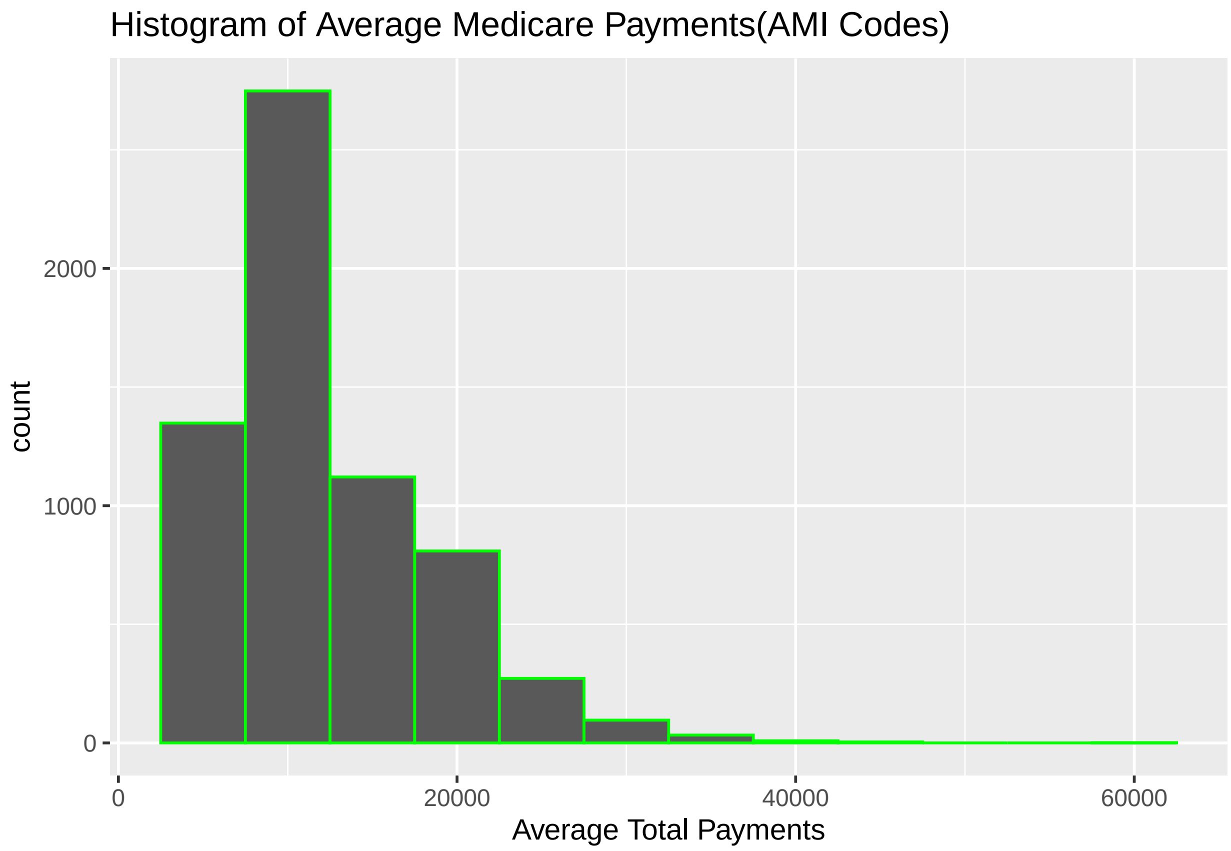
```
# Select relevant variables
final_data_AMI <- filter(final_data,DRG_Cd %in%
↪  c("246","247","248","249","250","251","280","281"))
#sum(final_data_AMI[, 'Total_discharges']) #total discharge sum = 258186


selected_data <- final_data_AMI %>%
  select("Provider_ID", "ZIP_Code","Ave_medical_payment", "Total_discharges",
  ↪  "Ave_covered_charges", "Ave_total_payment","Hospital.Type","Hospital.Ownership",
         "Hospital.overall.rating","State", "DRG_Cd")
attach (selected_data)

# Concerting character variable to string variables
selected_data[sapply(selected_data, is.character)] <-
↪  lapply(selected_data[sapply(selected_data, is.character)],
                                       as.factor)
```

```
    ## Step 2: Explore / Prepare Data     part c) Group by State/Summarize Data
```

```
# Calculate the average costs per state
Avg_Mdcr_Pymt_Amt  <- selected_data %>%
  group_by(State) %>%
  summarise(Avg_Mdcr_Pymt  = Ave_medical_payment)

Avg_Mdcr_Pymt_Amt <- Avg_Mdcr_Pymt_Amt[1:15537, ]
```

```
      ## Step 3: Visualization of Data     part a) Histogram
```

```r
# Create a histogram of average costs per state
ggplot(selected_data , aes(x = Ave_medical_payment)) +
  geom_histogram(binwidth = 5000, col = 'green') +
  labs(title = "Histogram of Average Medicare Payments(AMI Codes)", x = "Average Total
  ↪  Payments")
```



```
## Step 3: Visualization of Data     part b) United States Heat Map
```

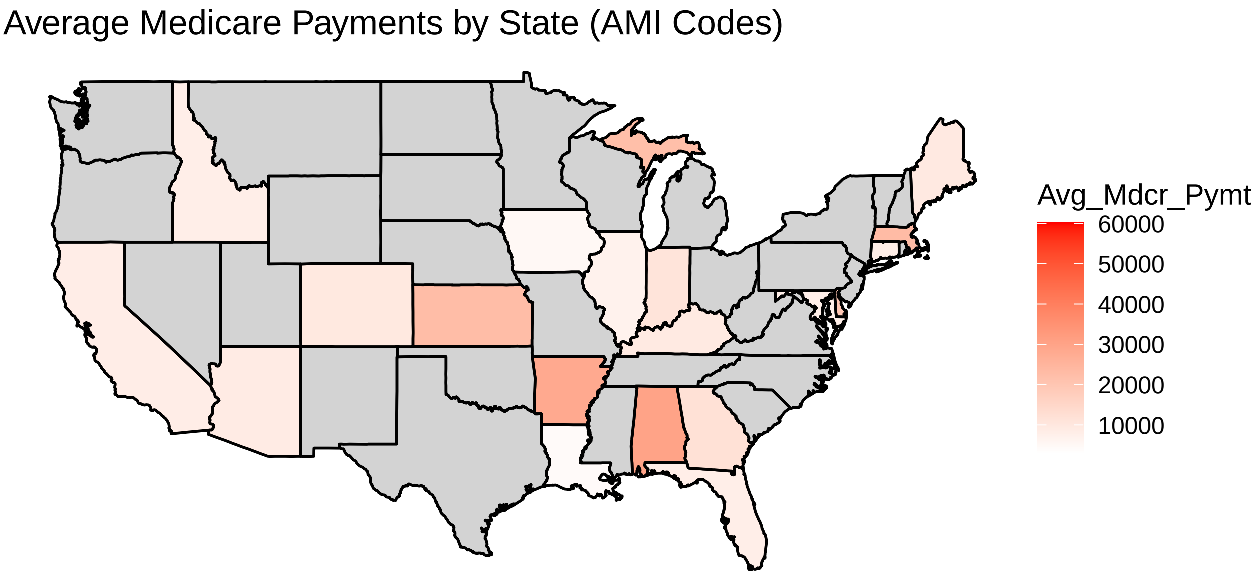
```r
# Create a US density map of average costs per state
us_map <- map_data("state")


usmap <- cbind(us_map,Avg_Mdcr_Pymt_Amt)

head(usmap)
```

```
##        long      lat group order  region subregion State Avg_Mdcr_Pymt
## 1 -87.46201 30.38968     1     1 alabama      <NA>    AK     30212.477
## 2 -87.48493 30.37249     1     2 alabama      <NA>    AK     16897.901
## 3 -87.52503 30.37249     1     3 alabama      <NA>    AK     13936.729
## 4 -87.53076 30.33239     1     4 alabama      <NA>    AK      8349.179
## 5 -87.57087 30.32665     1     5 alabama      <NA>    AK     20494.111
## 6 -87.58806 30.32665     1     6 alabama      <NA>    AK      9607.818
```

```
tail(usmap)
```

```
##            long      lat group order  region subregion State Avg_Mdcr_Pymt
## 15594 -106.3295 41.00659    63 15594 wyoming      <NA>  <NA>            NA
## 15595 -106.8566 41.01232    63 15595 wyoming      <NA>  <NA>            NA
## 15596 -107.3093 41.01805    63 15596 wyoming      <NA>  <NA>            NA
## 15597 -107.9223 41.01805    63 15597 wyoming      <NA>  <NA>            NA
## 15598 -109.0568 40.98940    63 15598 wyoming      <NA>  <NA>            NA
## 15599 -109.0511 40.99513    63 15599 wyoming      <NA>  <NA>            NA
```

```
ggplot(usmap, aes(x = long, y = lat, group = group, fill = Avg_Mdcr_Pymt)) +
  geom_polygon(color ="black") +
  scale_fill_gradient(low = "white", high = "red", na.value = "lightgrey") +
  theme_void() + coord_map() +
  labs(title = "Average Medicare Payments by State (AMI Codes)")
```

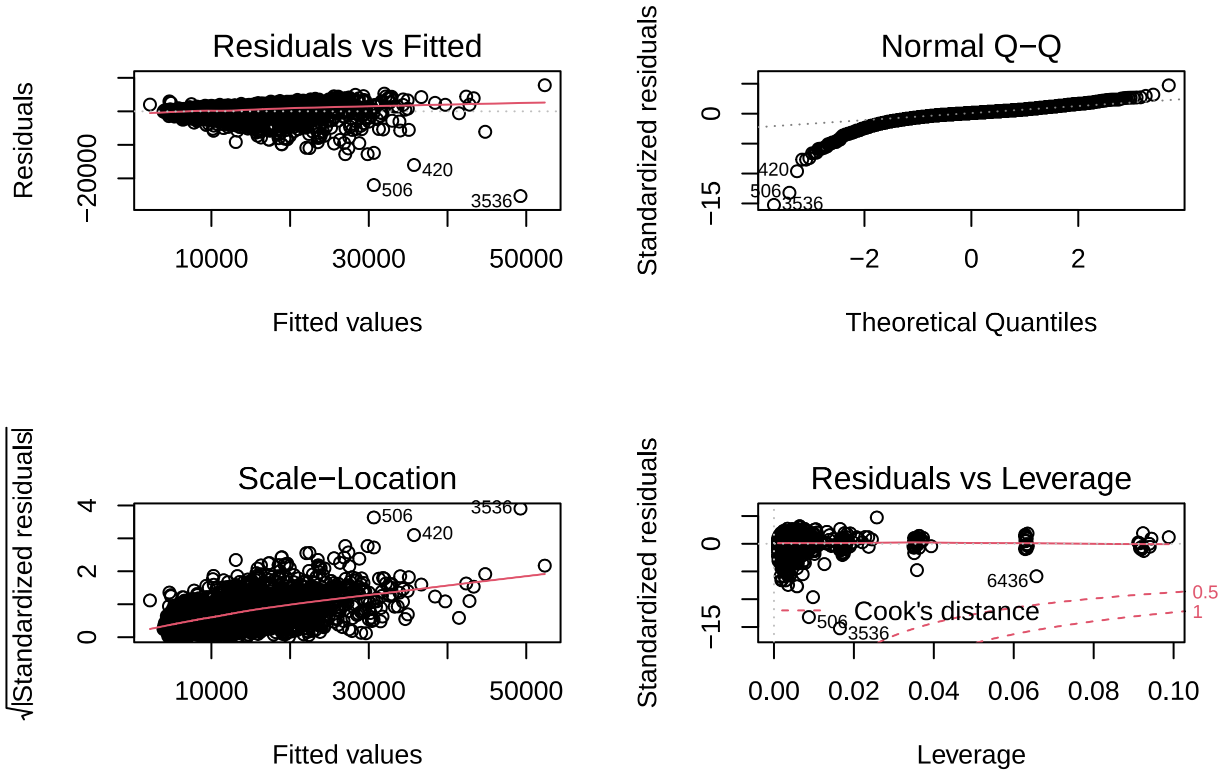


```
## Step 4: Creating Training and Test Datasets
```

```r
# Split the data into training and testing sets
set.seed(123)
library(caret)
trainIndex <- createDataPartition(selected_data$Ave_medical_payment, p = 0.7, list =
↪  FALSE)
train <- selected_data[trainIndex, ]
test <- selected_data[-trainIndex, ]
```

```
## Step 5: Build and Evaluate Linear Regression Model
```

```r
# Build the linear regression model
model <- lm(Ave_medical_payment ~ Total_discharges + Ave_covered_charges +
↪  Ave_total_payment+ Hospital.overall.rating+Hospital.Ownership, data = train)

# Predict on the testing set
predictions <- predict(model, newdata = test)


# Check for the model assumptions
par(mfrow = c(2, 2))
plot(model)
```



Evaluate the performance of the Linear Model

```
#model 1

RMSE(predictions, test$Ave_medical_payment)
```

```
## [1] 1518.246
```

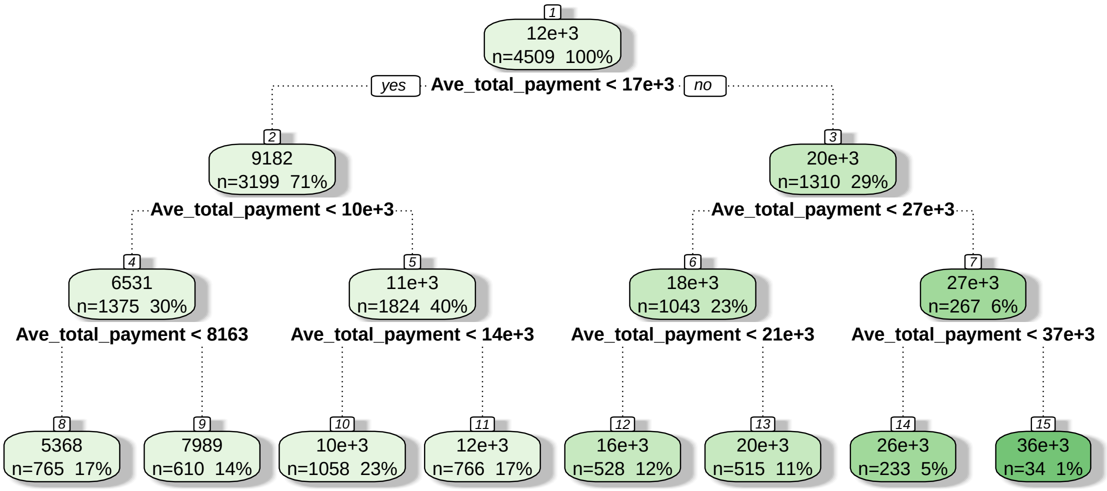
```
R2(predictions, test$Ave_medical_payment)
```

```
## [1] 0.9354307
```

```
## Step 6 Build and Evaluate CART Model (Regression Tree)
```

```
# Build the regression tree model
model2 <- rpart(Ave_medical_payment ~ Total_discharges + Ave_covered_charges +
↪  Ave_total_payment+ Hospital.overall.rating+Hospital.Ownership, data = train, method =
↪  "anova")
```

```r
#Plot
#install.packages("RGtk2")
# Plot the tree
fancyRpartPlot(model2)
```



Rattle 2023−Apr−21 21:37:21 richardbudden

```r
# Predict on the testing set
predictions <- predict(model2, newdata = test)

#Model 2
RMSE(predictions, test$Ave_medical_payment)
```

```
## [1] 1791.18
```

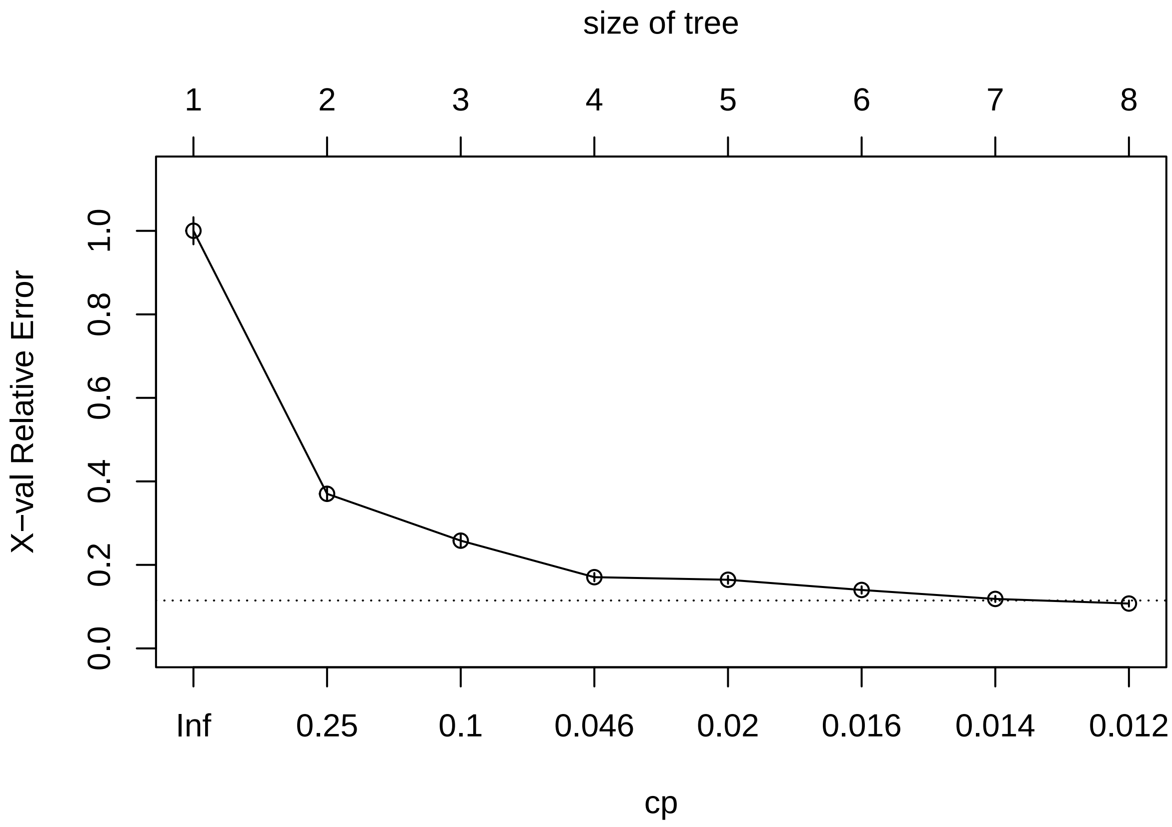
```r
R2(predictions, test$Ave_medical_payment)
```

```
## [1] 0.9095409
```

```r
printcp(model2)
```

```
##
## Regression tree:
## rpart(formula = Ave_medical_payment ~ Total_discharges + Ave_covered_charges +
##     Ave_total_payment + Hospital.overall.rating + Hospital.Ownership,
##     data = train, method = "anova")
##
## Variables actually used in tree construction:
## [1] Ave_total_payment
##
## Root node error: 1.6574e+11/4509 = 36756963
##
## n= 4509
##
##          CP nsplit rel error  xerror      xstd
## 1 0.631480      0  1.000000 1.00022 0.0324754
## 2 0.102234      1  0.368520 0.37021 0.0166698
## 3 0.101413      2  0.266286 0.25801 0.0149650
## 4 0.020514      3  0.164873 0.17056 0.0099140
## 5 0.018622      4  0.144358 0.16435 0.0095595
## 6 0.014060      5  0.125737 0.13989 0.0087609
## 7 0.013671      6  0.111677 0.11847 0.0073105
## 8 0.010000      7  0.098005 0.10742 0.0072899
```

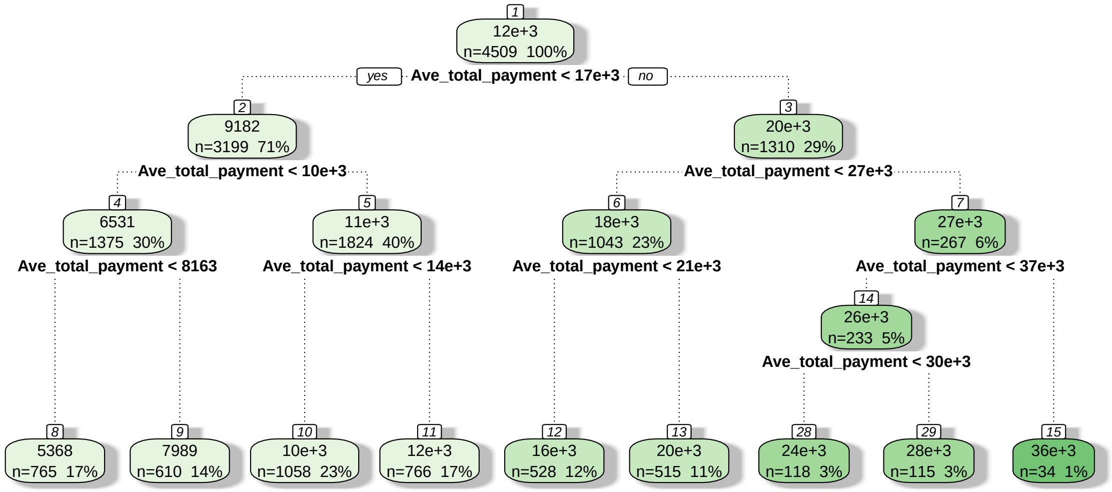
```
plotcp(model2)
```



```
pfit = prune(model2, cp=model2$ptable[which.min(model2$cptable[,"xerror"]), "CP"])
pfit
```

```
## n= 4509
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 4509 165737100000 12264.600
##    2) Ave_total_payment< 17330.01 3199  25871200000  9181.564
##      4) Ave_total_payment< 10316.22 1375   3215865000  6530.859
##        8) Ave_total_payment< 8163.443 765    420696500  5368.380 *
##        9) Ave_total_payment>=8163.443 610    464907400  7988.722 *
##      5) Ave_total_payment>=10316.22 1824   5711375000 11179.770
##       10) Ave_total_payment< 13688.9 1058   1404813000 10231.410 *
##       11) Ave_total_payment>=13688.9 766   2040764000 12489.630 *
##    3) Ave_total_payment>=17330.01 1310  35206200000 19793.320
##      6) Ave_total_payment< 26678.09 1043   9058654000 17981.000
##       12) Ave_total_payment< 20768.9 528   2383637000 16282.110 *
##       13) Ave_total_payment>=20768.9 515   3588702000 19722.770 *
##      7) Ave_total_payment>=26678.09 267   9339633000 26872.900
##       14) Ave_total_payment< 37147.43 233   4182570000 25509.750 *
##       15) Ave_total_payment>=37147.43 34   1757059000 36214.530 *
```

```
# Build the regression tree model
control_setting <- rpart.control(minsplit = 2, cp = .005, xval = 10)

model2 <- rpart(Ave_medical_payment ~ Total_discharges + Ave_covered_charges +
↪  Ave_total_payment+ Hospital.overall.rating+Hospital.Ownership, data = train, method =
↪  "anova", control = control_setting)

# Plot
#install.packages("RGtk2")
fancyRpartPlot(model2)
```



Rattle 2023−Apr−21 21:37:21 richardbudden

```
# Predict on the testing set
predictions <- predict(model2, newdata = test)

RMSE(predictions, test$Ave_medical_payment)
```

```
## [1] 1729.211
```

```
R2(predictions, test$Ave_medical_payment)
```

```
## [1] 0.9156873
```

Evaluate the performance of the CART model

```
#Model 2
RMSE(predictions, test$Ave_medical_payment)
```

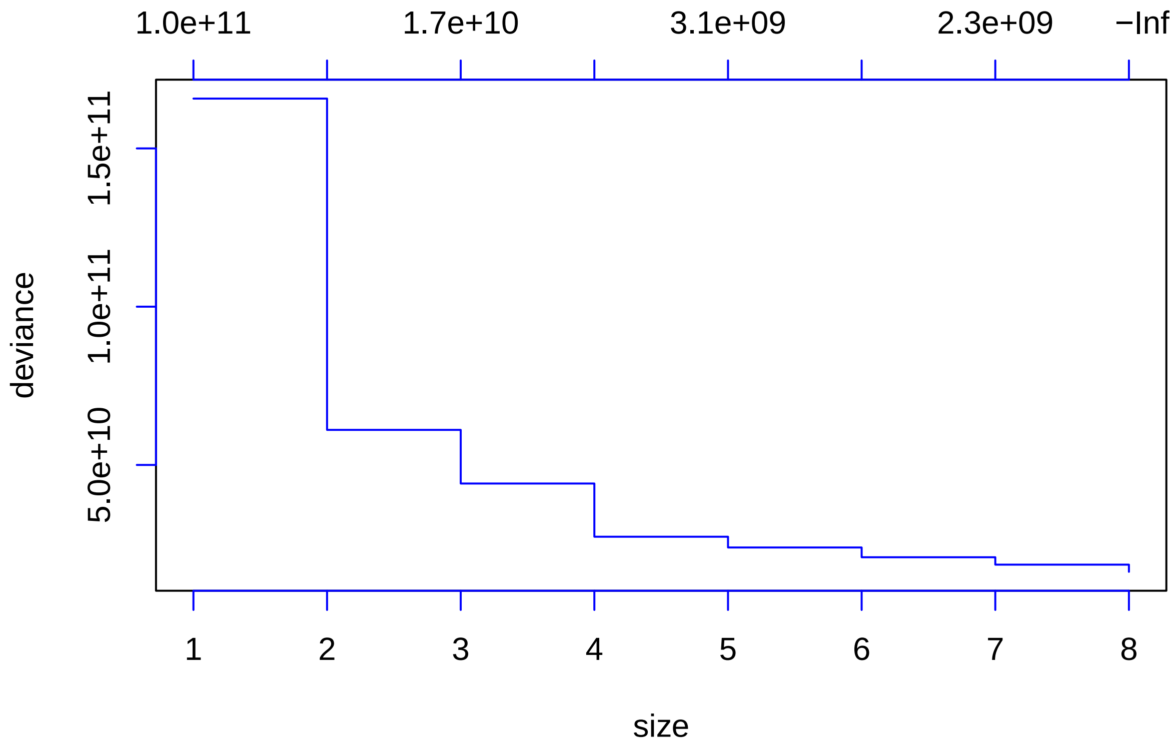
```
## [1] 1729.211
```

```
R2(predictions, test$Ave_medical_payment)
```

```
## [1] 0.9156873
```

Data pruning for Regression Tree

```
# Prune the tree, display pruned tree
library(tree)

tree.model <- tree(Ave_medical_payment ~ Total_discharges + Ave_covered_charges +
↪  Ave_total_payment+ Hospital.overall.rating+Hospital.Ownership, data = train)
prune.data <- prune.tree(tree.model)
plot(prune.data, col = "blue")
```



```
## Step 7 : Build and Evaluate Artifical Neural Network  Model (Feedforward ANN)
```

Feedforward ANN

```
## Scaling
scale_data <- final_data_AMI %>%
  select("Ave_medical_payment", "Total_discharges", "Ave_covered_charges",
  ↪  "Ave_total_payment")

set.seed(123)
```

```
trainIndex2 <- createDataPartition(scale_data$Ave_medical_payment, p = 0.7, list = FALSE)
train2 <- scale_data[trainIndex2, ]
test2 <- scale_data[-trainIndex2, ]
```

Min-Max Normalization and Scaling the input variable

```
#transform your factor to numeric.
#tranform it to a factor and the to numeric
selected_data$Hospital.Ownership <- as.numeric(as.factor(Hospital.Ownership))
selected_data$Hospital.overall.rating <- as.numeric(as.factor(Hospital.overall.rating))
```

```
selected_data$Ave_medical_payment <- (selected_data$Ave_medical_payment -
↪  min(selected_data$Ave_medical_payment)) / (max(selected_data$Ave_medical_payment) -
↪  min(selected_data$Ave_medical_payment))

selected_data$Total_discharges <- (selected_data$Total_discharges -
↪  min(selected_data$Total_discharges)) / (max(selected_data$Total_discharges) -
↪  min(selected_data$Total_discharges))

selected_data$Ave_covered_charges <- (selected_data$Ave_covered_charges -
↪  min(selected_data$Ave_covered_charges)) / (max(selected_data$Ave_covered_charges) -
↪  min(selected_data$Ave_covered_charges))

 selected_data$Ave_total_payment <- (selected_data$Ave_total_payment -
↪  min(selected_data$Ave_total_payment)) / (max(selected_data$Ave_total_payment) -
↪  min(selected_data$Ave_total_payment))

selected_data$Hospital.overall.rating <- (selected_data$Hospital.overall.rating -
↪  min(selected_data$Hospital.overall.rating)) /
↪  (max(selected_data$Hospital.overall.rating) -
↪  min(selected_data$Hospital.overall.rating))

selected_data$Hospital.Ownership <- (selected_data$Hospital.Ownership -
↪  min(selected_data$Hospital.Ownership)) / (max(selected_data$Hospital.Ownership) -
↪  min(selected_data$Hospital.Ownership))


set.seed(123)
inp <- sample(2, nrow(selected_data), replace = TRUE, prob = c(0.7, 0.3))
training_data <- selected_data[inp==1, ]
test_data <- selected_data[inp==2, ]
```

```
#from RBloggers "Selecting the number of neurons in the hidden layer of a neural network"
#A Variation of this rule suggests to choose a number of hidden neurons between one and
↪  the number of Inputs minus the number of outputs

#Upon our Model with 5 Inputs and 1 Output we set Hidden Levels at 4 (5-1=4)


set.seed(333)
model3 <- neuralnet(Ave_medical_payment ~ Total_discharges + Ave_covered_charges +
↪  Ave_total_payment+Hospital.Ownership+ Hospital.overall.rating,
```

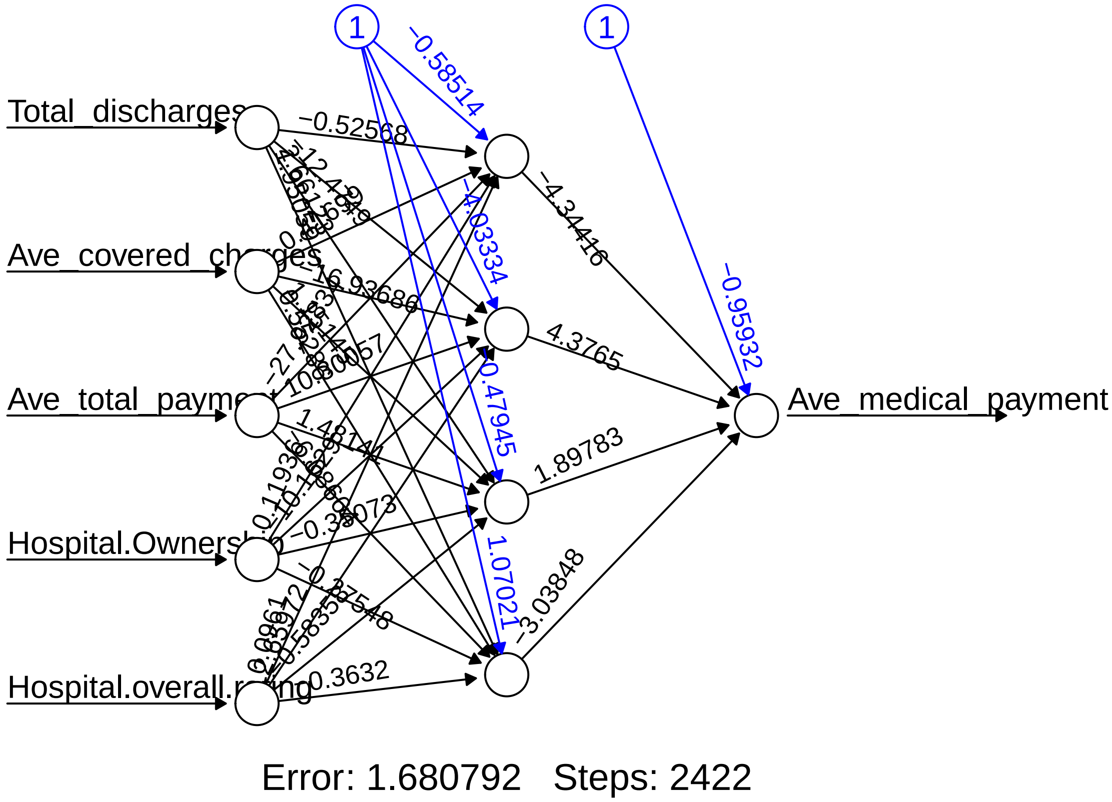
```r
                data = training_data,
                hidden = 4,
                linear.output = FALSE)

# Predict on the testing set
predictions <- predict(model3, test_data)

# Evaluate the performance of the model
RMSE(predictions, test_data$Ave_medical_payment)
```

```
## [1] 0.03022313
```

```r
R2(predictions, test_data$Ave_medical_payment)
```

```
##              [,1]
## [1,] 0.9235071
```

```r
# plot neural network
plot(model3, rep = "best")
```



1
1
Total_discharges
Ave_covered_charges
Ave_total_payment
Hospital.Ownership
Hospital.overall.rating
Ave_medical_payment
−0.58514
−4.03334
−0.47945
1.07021
−0.95932
−0.52568
−16.93686
1.48141
−0.36073
0.3632
−4.34416
4.3765
1.89783
−3.03848
Error: 1.680792   Steps: 2422