Prediction of insurance costs

Hüseyin Durmaz

3/24/23

Task

The task assigned to us is the prediction of insurance costs.

Cost is a continues variable, so we need to establish a regression model.

Description of the dataset

```
# install.packages("caret")
#install.packages("tidyverse")
library(tidyverse)
library(caret)
# Importing the data set
insurance <- read.csv("insurance.csv")

# Take a loot the data set
str(insurance)
```

```
'data.frame':   1338 obs. of  7 variables:
 $ age     : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex     : chr  "female" "male" "male" "male" ...
 $ bmi     : num  27.9 33.8 33 22.7 28.9 ...
 $ children: int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker  : chr  "yes" "no" "no" "no" ...
 $ region  : chr  "southwest" "southeast" "southeast" "northwest" ...
 $ charges : num  16885 1726 4449 21984 3867 ...
```

```r
# transform categorical to factor

insurance$sex = as.factor(insurance$sex)
insurance$smoker = as.factor(insurance$smoker)
insurance$region = as.factor(insurance$region)
```

It consists 1338 observation and 7 variables. It means that there are 1338 rows and 7 columns.

Age, bmi, children and charges are *numerical* variables. Sex, smoker and region are *categorical* variable

Splitting and training a lineer model

Before training a regression analysis, we split the data we have into two parts as test and train.

```r
set.seed(31)
index <- sample(1 : nrow(insurance), round(nrow(insurance) * 0.80))
train <- insurance[index, ]
test <- insurance[-index, ]
```

To make regression on the train dataset using the lm function.

```r
lm <- lm(charges ~ ., data = train)
summary(lm)
```

```
Call:
lm(formula = charges ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-11287.2  -2896.1   -936.1   1577.5  24902.9

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -12220.12    1103.53 -11.074  < 2e-16 ***
age               258.87      13.29  19.472  < 2e-16 ***
sexmale          -121.09     368.03  -0.329 0.742208
```

```
bmi                351.20      31.61  11.110  < 2e-16 ***
children           550.38     151.45   3.634 0.000292 ***
smokeryes        23934.09     453.16  52.816  < 2e-16 ***
regionnorthwest   -746.28     528.17  -1.413 0.157964
regionsoutheast  -1308.87     528.83  -2.475 0.013478 *
regionsouthwest  -1572.41     532.47  -2.953 0.003216 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5986 on 1061 degrees of freedom
Multiple R-squared:  0.7578,    Adjusted R-squared:  0.7559
F-statistic: 414.9 on 8 and 1061 DF,  p-value: < 2.2e-16
```

r square is 0.75 thats a good fit. We see that age, bmi, and smoker have a positive relationship We dont see a relationship between sex and region.

Measuring Performance

```r
predictedcharges = predict(lm, test)

head(predictedcharges)
```

```
        3          4         23         29         39         44
6838.9117  3429.1336  2985.3765  -477.5459 34082.0784  7966.8128
```

```r
error = test$charges - predictedcharges

rmse = sqrt(mean(error^2))

rmse
```

```
[1] 6378.75
```

Charges is a continous variable, so RMSE is the best metric for this data set RMSE value is positive and pretty high, it may be underfitting. It looks like there is a underfitting(step X1) because rmse value and meaning value's difference is not enough.

```r
newobs <- data.frame(
    age <- 31,
    sex <- "female",
    bmi <- 18,
    children <- 1,
    smoker <- "yes",
    region <- "northwest"
)

  model <- train(
  charges ~ .,
  data <- insurance,
  method <- "lm"
)
tahmin <- predict(model, newdata <- newobs)

tahmin
```

```
       1
26100.56
```

This is the estimated premium in the dataset I created.