Predicting medical charges using machine learning

Nigel K. Gondo

Brief Description

The purpose of this project is to conduct a regression analysis using machine learning, to predict medical cost of individuals based on certain criteria. These criteria are, if the individual is a smoker or not, has children, what their body mass index is, and where they are from. The train split train split will be utilised for the purpose of training our model and see how well it can fair in predicting.

Setting up the directory to work in

```
setwd('C://Users//Nigel Gondo//Documents//Portfolio Projects//ML')
```

Importing csv file

```
df_insurance <- read.csv('insurance.csv')
head(df_insurance)

##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

checking the structure of the dataset

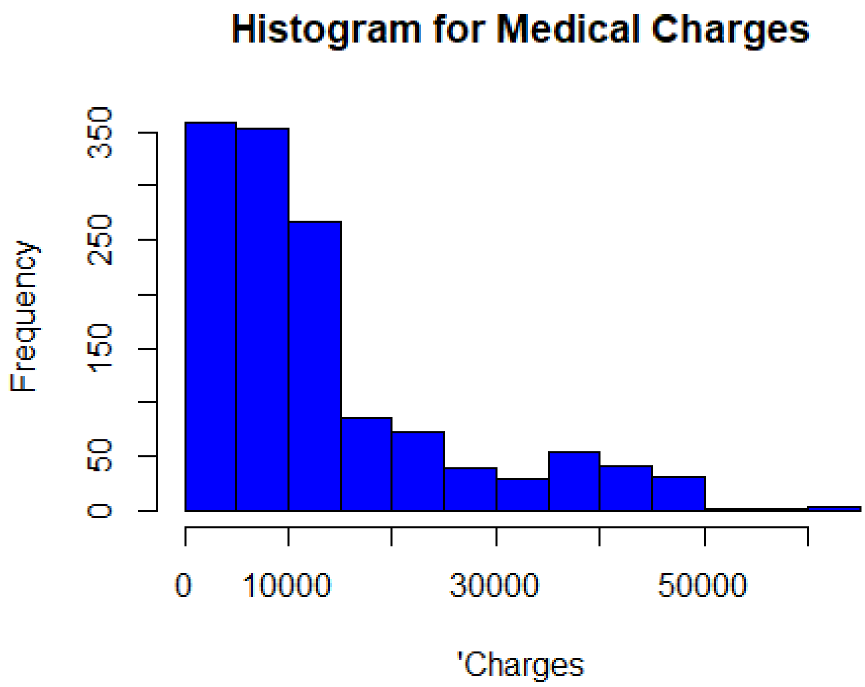
```
str(df_insurance)

## '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 ...
```

Checking for null values (no null values if it sums up to zero)

```
sum(is.null(df_insurance))

## [1] 0
```

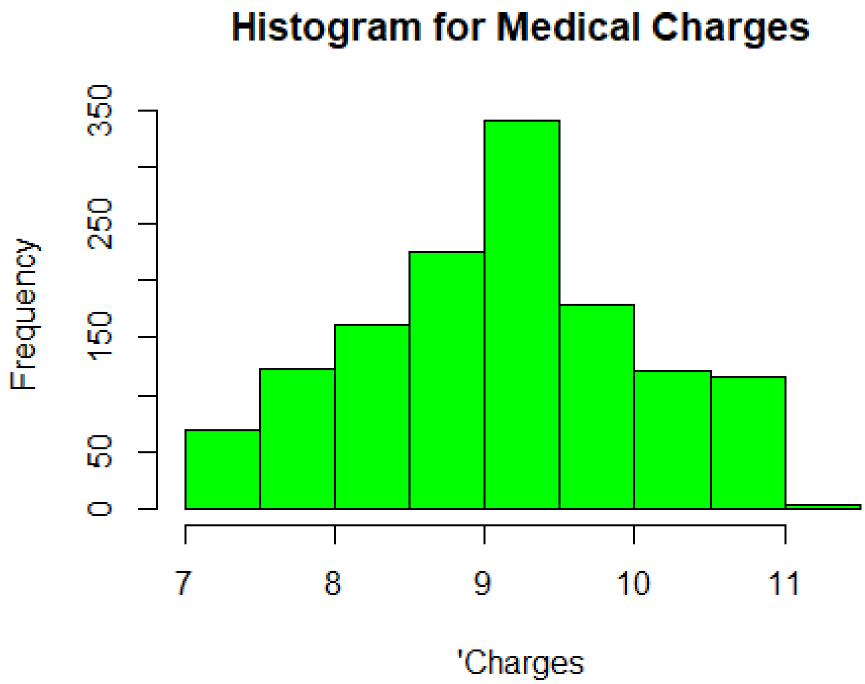
Summary statistics and EDA

```
summary(df_insurance)

##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21                      Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                      Max.   :53.13   Max.   :5.000
##     smoker             region             charges
##  Length:1338        Length:1338        Min.   : 1122
##  Class :character   Class :character   1st Qu.: 4740
##  Mode  :character   Mode  :character   Median : 9382
##                                        Mean   :13270
##                                        3rd Qu.:16640
##                                        Max.   :63770

hist(df_insurance$charges,
     main = 'Histogram for Medical Charges',
     xlab = "'Charges",
     border = 'black',
     col = 'blue',
     )
```



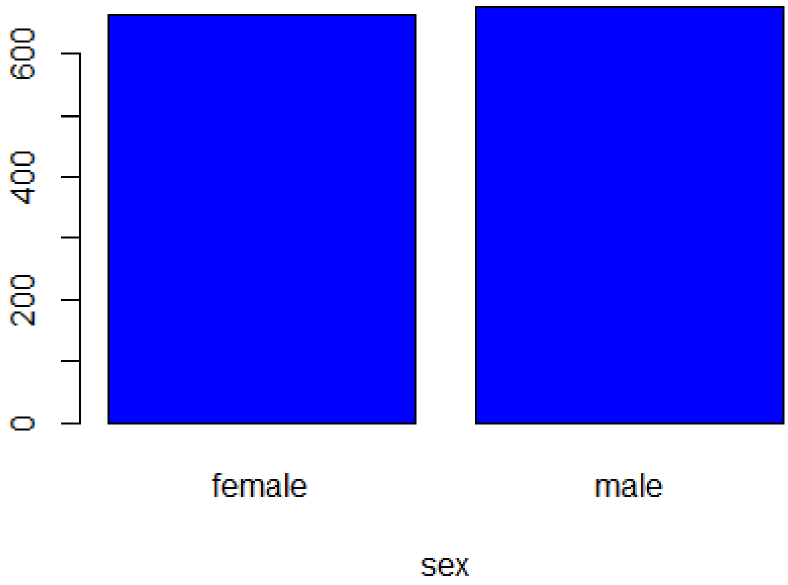
creating a normally distributed column of charges

```
charges_norm_dist <- log(df_insurance$charges)

hist(charges_norm_dist,
     main = 'Histogram for Medical Charges',
     xlab = "'Charges",
     border = 'black',
     col = 'green',
     )
```



Bar chart showing gender count

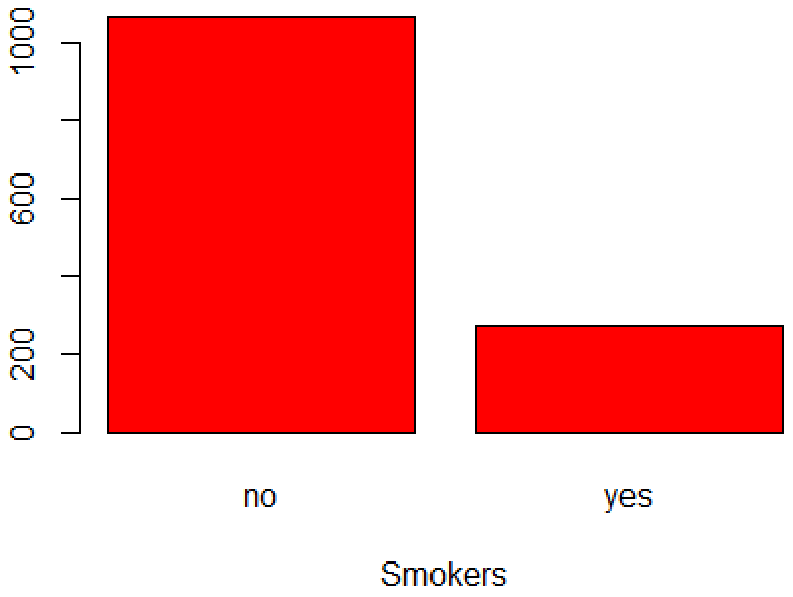
```
count_sex <- table(df_insurance$sex)
barplot(count_sex,
        names.arg=rownames(count_sex),
        xlab = 'sex',
        border ='black',
        col = 'blue')
```



The number of male and females is about even

Visualising smokers and non-smokers

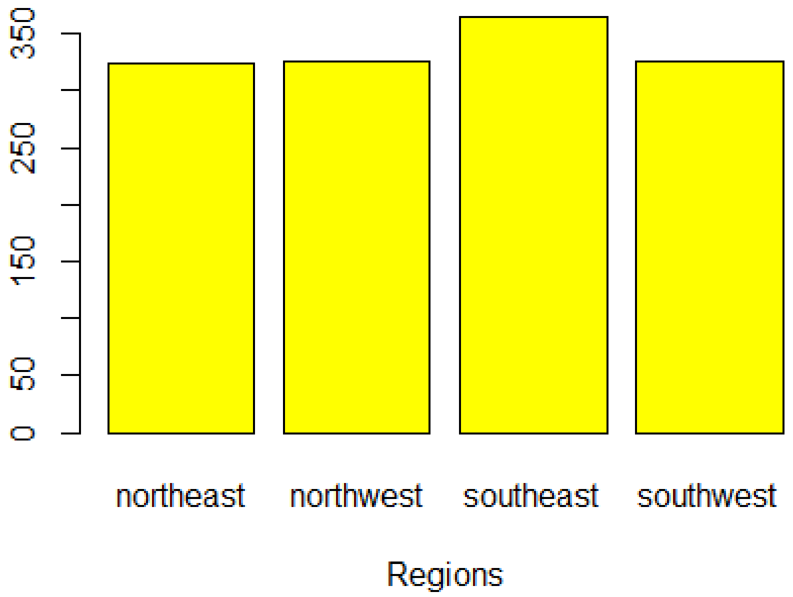
```
count_smokers <- table(df_insurance$smoker)
barplot(count_smokers,
        names.arg=rownames(count_smokers),
        xlab = 'Smokers',
        border ='black',
        col = 'red')
```



Non-smokers out number the smokers by more than 50%

Visualising the different regions the clients are from

```
count_region <- table(df_insurance$region)
barplot(count_region,
        names.arg=rownames(count_region),
        xlab = 'Regions',
        border ='black',
        col = 'yellow')
```



Creating dummy variables

```
sex_male <- ifelse(df_insurance$sex == 'male', 1, 0)
sex_female <- ifelse(df_insurance$sex == 'female', 1, 0)
smoker_yes <- ifelse(df_insurance$smoker == 'yes', 1, 0)
smoker_no <- ifelse(df_insurance$smoker == 'no', 1, 0)
region_northeast <- ifelse(df_insurance$region == 'northeast', 1, 0)
region_northwest <- ifelse(df_insurance$region == 'northwest', 1, 0)
region_southeast <- ifelse(df_insurance$region == 'southeast', 1, 0)
region_southwest <- ifelse(df_insurance$region == 'southwest', 1, 0)
```

Creating data frame for regression

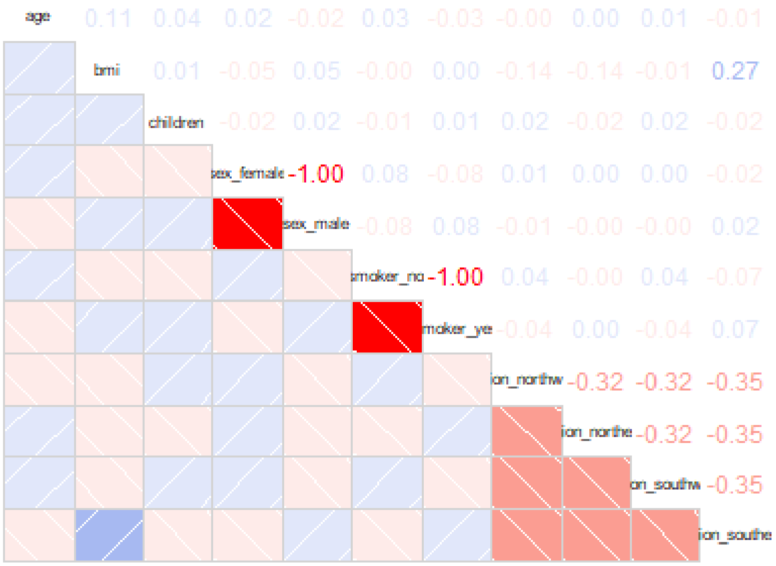
```
df_insurance2 <- data.frame( age = df_insurance$age,
                             bmi = df_insurance$bmi,
                             children = df_insurance$children,
                             #charges = df_insurance$charges,
                             charges_norm_dist = charges_norm_dist,
                             sex_female = sex_female,
                             sex_male = sex_male,
                             smoker_no = smoker_no,
                             smoker_yes = smoker_yes,
                             region_northwest = region_northwest,
                             region_northeast = region_northeast,
                             region_southwest = region_southwest,
                             region_southeast = region_southeast)
```

```
head(df_insurance2)
```

```
##   age    bmi children charges_norm_dist sex_female sex_male smoker_no
## 1  19 27.900        0          9.734176          1        0         0
## 2  18 33.770        1          7.453302          0        1         1
## 3  28 33.000        3          8.400538          0        1         1
## 4  33 22.705        0          9.998092          0        1         1
## 5  32 28.880        0          8.260197          0        1         1
## 6  31 25.740        0          8.231275          1        0         1
##   smoker_yes region_northwest region_northeast region_southwest
## 1          1                0                0                1
## 2          0                0                0                0
## 3          0                0                0                0
## 4          0                1                0                0
## 5          0                1                0                0
## 6          0                0                0                0
##   region_southeast
## 1                0
## 2                1
## 3                1
## 4                0
## 5                0
## 6                1
```

Correlation map

```
library(corrgram)
df_insurance_indep_values <- df_insurance2[-c(4)]
corrgram(df_insurance_indep_values, lower.panel=panel.shade,
upper.panel=panel.cor)
```



Splitting the model in training and test data set

```
library(caTools)
set.seed(42)
sliptting_data <- sample.split(df_insurance2$charges_norm_dist, SplitRatio =
0.75)
train <- subset(df_insurance2, sliptting_data = 'TRUE')
test <- subset(df_insurance2, sliptting_data = 'FALSE')
```

modeling the data

```
#train$charges_norm_dist <- exp(train$charges_norm_dist)

model <- lm(charges_norm_dist ~., data = train)

summary(model)

##
## Call:
## lm(formula = charges_norm_dist ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07186 -0.19835 -0.04917  0.06598  2.16636
##
## Coefficients: (3 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       8.3522677  0.0818065 102.098  < 2e-16 ***
## age               0.0345816  0.0008721  39.655  < 2e-16 ***
## bmi               0.0133748  0.0020960   6.381 2.42e-10 ***
## children          0.1018568  0.0100995  10.085  < 2e-16 ***
## sex_female        0.0754164  0.0244012   3.091  0.00204 ** 
## sex_male                 NA         NA      NA       NA    
## smoker_no        -1.5543228  0.0302795 -51.333  < 2e-16 ***
## smoker_yes               NA         NA      NA       NA    
## region_northwest  0.0934092  0.0351023   2.661  0.00788 ** 
## region_northeast  0.1571967  0.0350828   4.481 8.08e-06 ***
## region_southwest  0.0282445  0.0344925   0.819  0.41301    
## region_southeast         NA         NA      NA       NA    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.4443 on 1329 degrees of freedom
## Multiple R-squared:  0.7679, Adjusted R-squared:  0.7666 
## F-statistic: 549.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

The R^2 of the testing set is reasonable as the measures of variability for the target variable is 76.7% . However the model can be considered to be under fitted.

predicting the model

```
pred <- predict(model, test)

## Warning in predict.lm(model, test): prediction from a rank-deficient fit
may be
## misleading

modelEval <- cbind(test$charges_norm_dist, pred)
colnames(modelEval) <- c('Actual', 'Predicted')
modelEval <- as.data.frame(modelEval)
head(modelEval)

##     Actual Predicted
## 1 9.734176  9.486137
## 2 7.453302  7.973939
## 3 8.400538  8.513171
## 4 9.998092  8.336224
## 5 8.260197  8.384231
## 6 8.231275  8.289660

mse <- mean((modelEval$Actual - modelEval$Predicted)^2)
mse

## [1] 0.1960605
```

The mean squared error for both testing set is relatively low which which is about 19.6% respectively