Machine Learning Exercise

Chen Chen

For this exercise, I want to apply machine learning techniques on my dataset. I applied two types of machine learning algorithms and both are supervised algorithm. I first ran linear regression analysis on my dataset, because I want to explore if the predictor - total deaths/damages is a function of the earthquake variables - depth/magnitude/location(lat,lon)/intensity.

High residual and low R-squared value in the summary of this model suggests a bad fit.

```
summary(model1)
```

```
##
## Call:
## lm(formula = TOTAL_DEATHS ~ FOCAL_DEPTH + EQ_PRIMARY + INTENSITY +
##     LATITUDE + LONGITUDE, data = df_eq_gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12574  -5071  -2347    822 229070
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -29396.051   9095.856  -3.232  0.00139 **
## FOCAL_DEPTH     -4.789     38.116  -0.126  0.90011
## EQ_PRIMARY    2579.113   1381.193   1.867  0.06302 .
## INTENSITY     1872.967    786.637   2.381  0.01801 *
## LATITUDE        50.497     52.875   0.955  0.34049
## LONGITUDE       10.689     13.159   0.812  0.41737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17410 on 251 degrees of freedom
##   (1755 observations deleted due to missingness)
## Multiple R-squared:  0.06531,    Adjusted R-squared:  0.04669
## F-statistic: 3.508 on 5 and 251 DF,  p-value: 0.004401
```

I then included the economic status of a country represented by the GDP and the population in the regression analysis. Inspecting the summary of this model suggests a slight increase in performance of this model, but it was still not a good fit model.

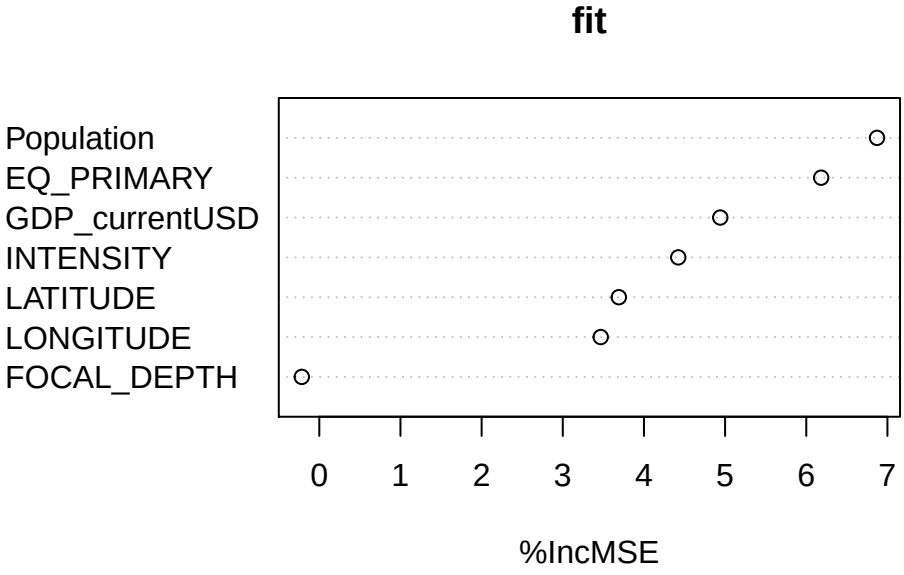
```
model2 <- lm(TOTAL_DEATHS ~ FOCAL_DEPTH + EQ_PRIMARY + INTENSITY + LATITUDE + LONGITUDE
             + GDP_currentUSD + Population,data = df_eq_gdp)
summary(model2)
```

```
##
## Call:
## lm(formula = TOTAL_DEATHS ~ FOCAL_DEPTH + EQ_PRIMARY + INTENSITY +
##     LATITUDE + LONGITUDE + GDP_currentUSD + Population, data = df_eq_gdp)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -22102  -5239  -1374   1977 219126
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.603e+04  1.157e+04  -3.114  0.00214 **
## FOCAL_DEPTH     2.414e+01  6.273e+01   0.385  0.70083
## EQ_PRIMARY      3.053e+03  1.727e+03   1.768  0.07867 .
## INTENSITY       2.047e+03  1.006e+03   2.035  0.04330 *
## LATITUDE        3.940e+01  6.644e+01   0.593  0.55390
## LONGITUDE      -9.284e+00  1.778e+01  -0.522  0.60217
## GDP_currentUSD  2.611e-10  1.753e-09   0.149  0.88177
## Population      1.418e-05  4.637e-06   3.058  0.00256 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19070 on 186 degrees of freedom
##   (1818 observations deleted due to missingness)
## Multiple R-squared:  0.1173, Adjusted R-squared:  0.08412
## F-statistic: 3.532 on 7 and 186 DF,  p-value: 0.001382
```

The simple linear regression analysis did not seem to work well for this dataset. And it is time to try something different! My mentor suggested giving Random Forest algorithm a try. Random Forest performs regression analysis by constructing a multitude of decision trees at training time and outputting the mean result of the individual trees.

```
set.seed(130)
fit <- randomForest(TOTAL_DEATHS ~ FOCAL_DEPTH + EQ_PRIMARY + INTENSITY +
                    LATITUDE + LONGITUDE + GDP_currentUSD + Population,
                    data = df_eq_gdp,
                    importance = TRUE,
                    na.action = na.omit,
                    ntree = 2000)
varImpPlot(fit,type=1)
```



```
fit
```

```
##
## Call:
##  randomForest(formula = TOTAL_DEATHS ~ FOCAL_DEPTH + EQ_PRIMARY +      INTENSITY + LATITUDE + LONGITU
##                Type of random forest: regression
##                      Number of trees: 2000
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 408750305
##                     % Var explained: -3.45
```

The variable importance plot suggests that population and earthquake magnitude are relatively more important than the other variables. But the large mean of squared residuals and the negative value for the % of variance explained by this model indicates that random forest algorithm was not adequate to define a model that relates the total deaths/damages to the earthquake variables.