TSF Task 1:Prediction using Supervised ML

Nirnoy Ghosh , Data Science & Business Analytics Intern, The Sparks foundation

2023-02-19

1. Installtion and calling the required library

```
#install.packages("readr")
#install.packages("dplyr")
library(readr)
library(dplyr)
```

2. Import the data and print the data.

```
data=read.csv("C:/Users/ASUS/OneDrive/Desktop/Nirnoy/Placement/TSF/TSFtask1.csv",header=TRUE)
data
```

```
##    Hours Scores
## 1    2.5     21
## 2    5.1     47
## 3    3.2     27
## 4    8.5     75
## 5    3.5     30
## 6    1.5     20
## 7    9.2     88
## 8    5.5     60
## 9    8.3     81
## 10   2.7     25
## 11   7.7     85
## 12   5.9     62
## 13   4.5     41
## 14   3.3     42
## 15   1.1     17
## 16   8.9     95
## 17   2.5     30
## 18   1.9     24
## 19   6.1     67
## 20   7.4     69
## 21   2.7     30
## 22   4.8     54
## 23   3.8     35
## 24   6.9     76
## 25   7.8     86
```

3. Checking the null value

```
is.null(data)
```

```
## [1] FALSE
```

From the output it is clear that there is no null value

4. Checking some statistical property.

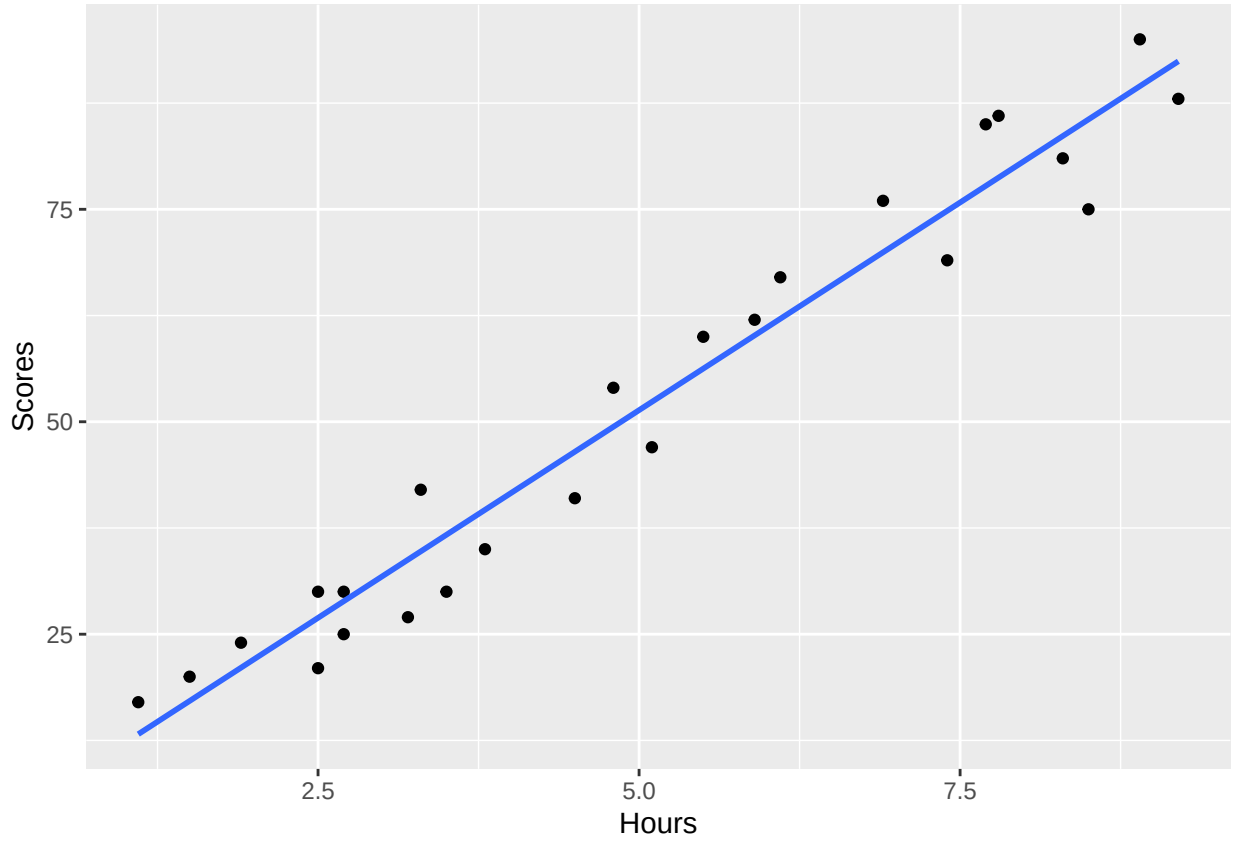
```
summary(data)
```

```
##      Hours           Scores
##  Min.   :1.100   Min.   :17.00
##  1st Qu.:2.700   1st Qu.:30.00
##  Median :4.800   Median :47.00
##  Mean   :5.012   Mean   :51.48
##  3rd Qu.:7.400   3rd Qu.:75.00
##  Max.   :9.200   Max.   :95.00
```

5. Plot the data to understnd what kind of relationship has between two variables.

```
#install.packages("ggplot2")
library(ggplot2)
ggplot(data = data, aes(x= Hours, y = Scores)) +
  geom_point() +
  geom_smooth(method= "lm",se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The above diagram shows us the relationship of our model

6. Now we will calculate the coefficients of our model by using the given data.

```
#Finding the coefficients of the data using linear model
data_coeff=lm(Scores~Hours, data= data)
data_coeff
```

```
##
## Call:
## lm(formula = Scores ~ Hours, data = data)
##
## Coefficients:
## (Intercept)        Hours
##       2.484        9.776
```

The model equation will be score= intercept + (hours * time)

```
summary(data_coeff)
```

```
##
```

```
## Call:
## lm(formula = Scores ~ Hours, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.578  -5.340   1.839   4.593   7.265
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.4837     2.5317   0.981    0.337
## Hours         9.7758     0.4529  21.583   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.603 on 23 degrees of freedom
## Multiple R-squared:  0.9529, Adjusted R-squared:  0.9509
## F-statistic: 465.8 on 1 and 23 DF,  p-value: < 2.2e-16
```

Here from the R squared valu we can say that 95.09% of scores can be explained by the study hours.

7. We will calculate the expected or predicted score of the given data.

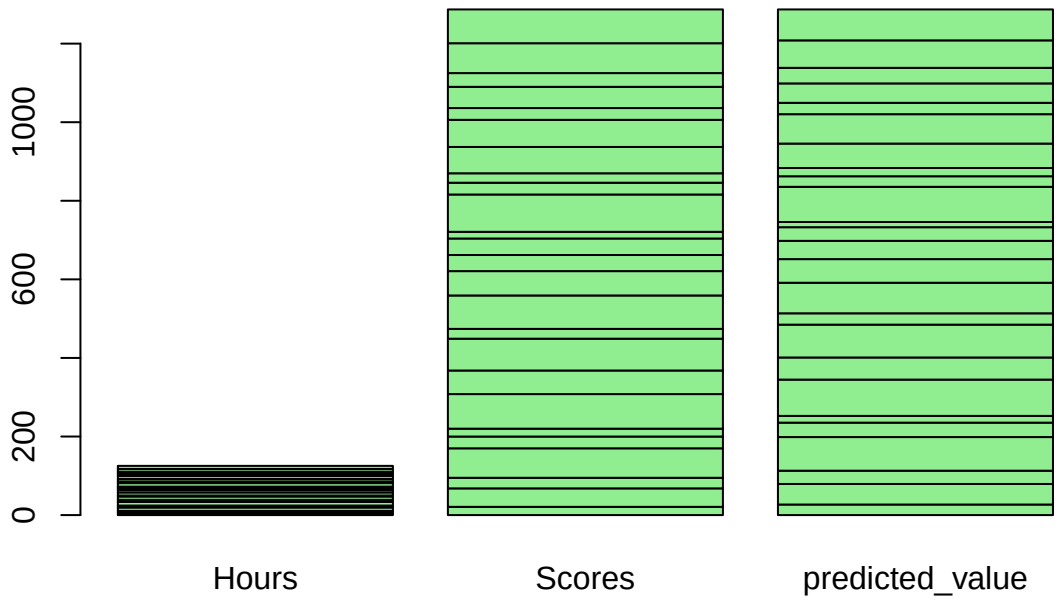
```
coef(data_coeff)
```

```
## (Intercept)       Hours
##    2.483673    9.775803
```

```
data = data %>%
  mutate(predicted_value=fitted(data_coeff))
head(data)
```

```
##   Hours Scores predicted_value
## 1   2.5     21        26.92318
## 2   5.1     47        52.34027
## 3   3.2     27        33.76624
## 4   8.5     75        85.57800
## 5   3.5     30        36.69899
## 6   1.5     20        17.14738
```

```
barplot(as.matrix(data),col="light green")
```



8. Calculating the final score.

```
score = coef(data_coeff)[[1]]+(coef(data_coeff)[[2]]*9.25)
score
```

```
## [1] 92.90985
```

If a student studies for 9.25 hrs/ day,predicted score will be 92.91

Thank You