Graduate Rotational Internship Program : June 2021

The Sparks Foundation
Data Science & Business Analytics Tasks - 1

Predicting Student's Score using Supervised ML

Owner: Reema Lad

Wednesday, June 16, 2021

Data Science & Business Analytics Tasks

Prediction using Supervised ML

(Level - Beginner)

- Predict the percentage of an student based on the no. of study hours.
- This is a simple linear regression task as it involves just 2 variables.
- You can use R, Python, SAS Enterprise Miner or any other tool
- Data can be found at http://bit.ly/w-data
- What will be predicted score if a student studies for 9.25 hrs/ day?
- Sample Solution : https://bit.ly/2HxiGGJ
- Task submission:
  1. Host the code on GitHub Repository (public). Record the code and output in a video. Post the video on YouTube
  2. Share links of code (GitHub) and video (YouTube) as a post on **YOUR LinkedIn profile**, not TSF Network.
  3. Submit the LinkedIn link in Task Submission Form when shared.

Project Case

Case Study Background

We need to predict the percentage of an student based on the number of study hours.

Problem Statement

What will be predicted score if a student studies for 9.25 hrs/ day?

Data Dictionary

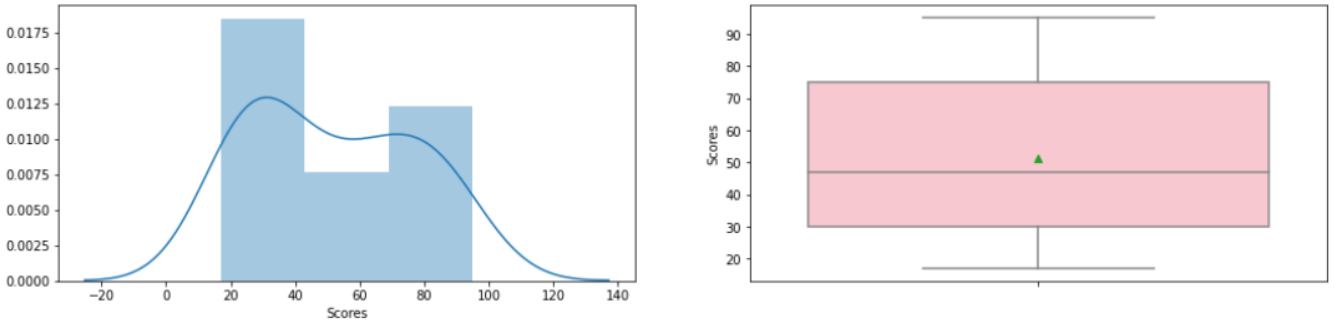
Variable	Definition
Hours	Hours Studied by student
Scores	Scores scored by student

Process

- Import Libraries
- Load Data
- Reading Raw Data
- Visualization, UniVariate - BiVariate Analysis, EDA
- Model Building
- Predictions

Understanding Data & EDA, Insights on Variables

Univariate Analysis of Scores (Target Variable)



Data Insights

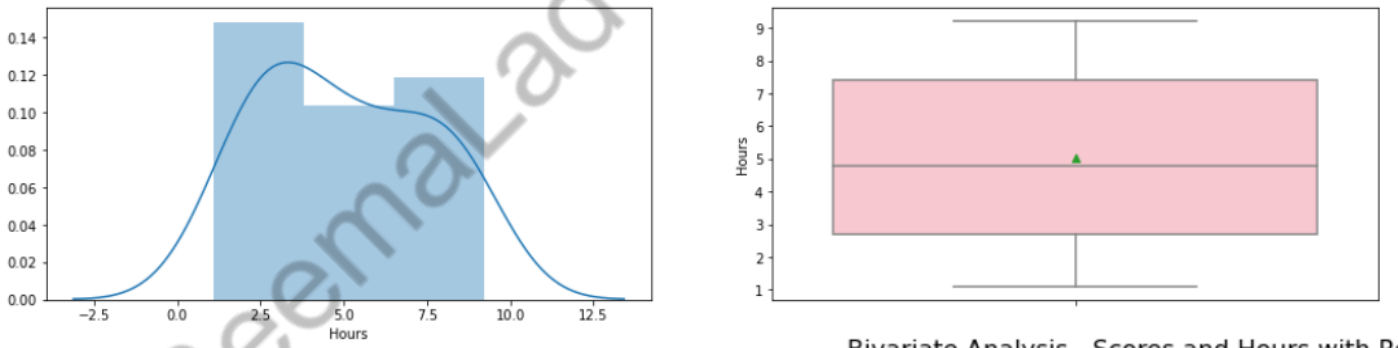
* 25 rows with two columns

* No missing data

* Target Variable : Score

* Predictor Variable : Hours

Univariate Analysis of Hours Studied (Predictor Variable)



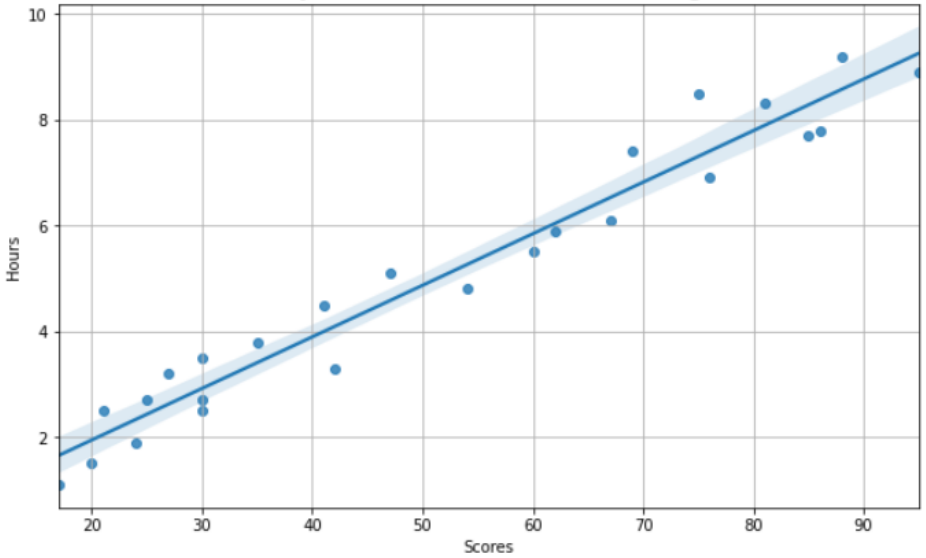
Variables Insights

* Both Target and Predictor Variables are almost normally distributed with no outliers

* Hours of studies have a positive corelation with Scores and are highly co-related

* As a linear relation is reflected, and only single predictor variable is available we will attempt building model using OLS (Ordinary Least Squared method) instead of standard LinearRegression

Bivariate Analysis - Scores and Hours with Regression Line



Insights - Train Test Split

* As data is very less, spliting Train Test into 50-50 to guage performance by pushing more data in Test set

Model - OLS (Ordinary Least Squares regression) (Using Stats Model)

OLS Regression Results

Dep. Variable:	Scores	**R-squared:**	0.936
Model:	OLS	**Adj. R-squared:**	0.929
Method:	Least Squares	**F-statistic:**	145.9
Date:	Fri, 18 Jun 2021	**Prob (F-statistic):**	2.75e-07
Time:	10:43:00	**Log-Likelihood:**	-36.515
No. Observations:	12	**AIC:**	77.03
Df Residuals:	10	**BIC:**	78.00
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.7773	5.080	0.547	0.597	-8.543	14.097
Hours	9.7853	0.810	12.078	0.000	7.980	11.591

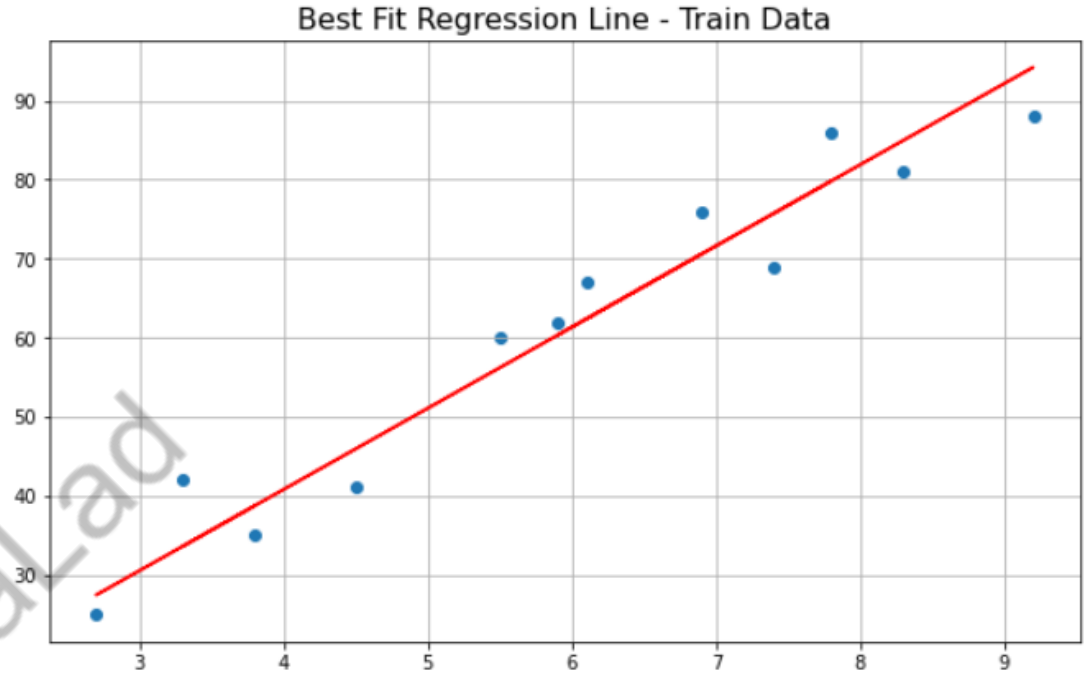
Omnibus:	6.908	**Durbin-Watson:**	1.469
Prob(Omni			
Kurt			

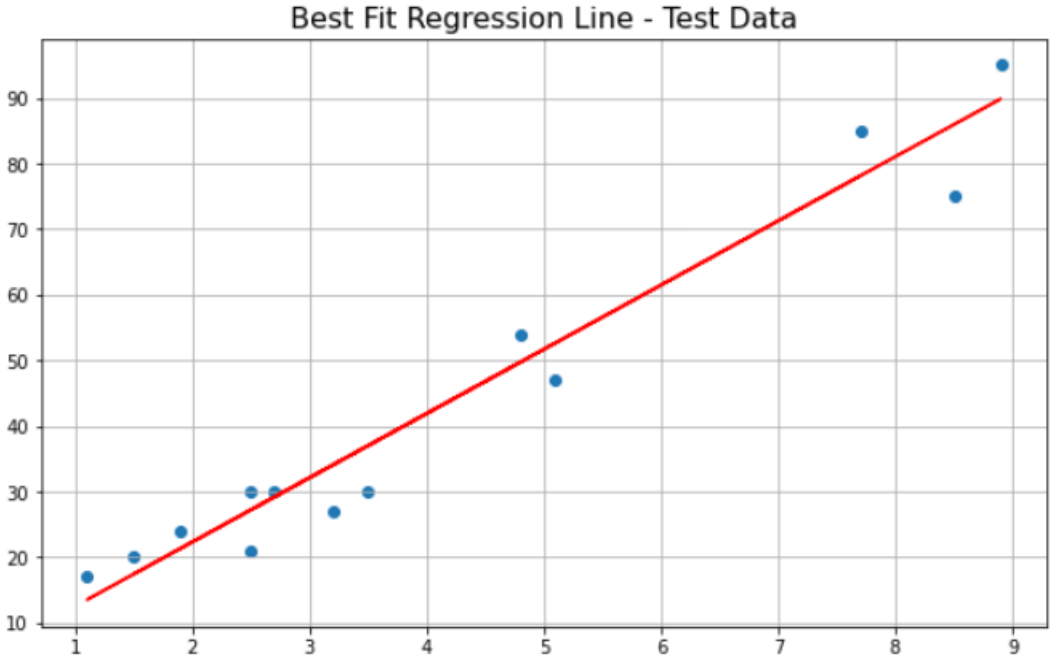
Model Insights

* RSquare Value on Train Data is 93.6%

* RSquared Value on Test Data is 95.06%

* The model is stable and can generalize on unseen test set





Prediction to User Input Predictor

Input Hours of studies per day to predict Score :

Input Hours of studies per day to predict Score : 9.25

```
Input Hours of studies per day to predict Score : 9.25
You have selected :  9.25  hours of studies per day
Predicted Score for:  9.25  hours of study per day, is:  [93.29160468] %
```

Problem Statement

What will be predicted score if a student studies for 9.25 hrs/ day?

Prediction

Predicted score is 93.29% if a student studies for 9.25 hrs/ day

Thank You