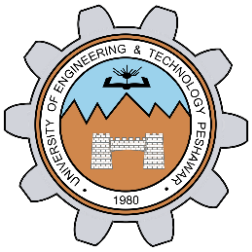
ASSIGNMENT #3



Fall 2021

Data Analytics

Submitted by: **Shah Raza**

Registration No.: **18PWCSE1658**

Class Section: **B**

"On my honor, as student of University of Engineering and Technology, I have neither given nor received unauthorized assistance on this academic work."

Student Signature: _____

Submitted to:

Engr. Naina Said

February 26, 2022

Department of Computer Systems Engineering

University of Engineering and Technology, Peshawar

Task 1:
Take any dataset from Kaggle and perform linear regression on it in python. Take snippets of the code. Write down the results of regression analysis and present your interpretation of the results.

Dataset:

Link: https://www.kaggle.com/karthickveerakumar/salary-data-simple-linear-regression

Task: Predict salary based on experience using simple linear regression.

Code:

```
# Importing the libraries
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd

# Importing the dataset
dataset = pd.read_csv('Salary_Data.csv')
X = dataset.iloc[:, :-1].values
y = dataset.iloc[:, -1].values

# Splitting the dataset into the Training set and Test set
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 1/3, random_state
= 0)

# Training the Simple Linear Regression model on the Training set
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(X_train, y_train)

# Predicting the Test set results
y_pred = regressor.predict(X_test)

# Visualising the Training set results
plt.scatter(X_train, y_train, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Salary vs Experience (Training set)')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()

# Visualising the Test set results
plt.scatter(X_test, y_test, color = 'red')
plt.plot(X_train, regressor.predict(X_train), color = 'blue')
plt.title('Salary vs Experience (Test set)')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()
```

Output:

Training Set:

Simple Linear Regression works in such a way that it finds a line (resulting from the intercept and slope) that has the minimum distance from all the points in the training set. This method is known as the Ordinary Least Square method.

Testing Set:

As we can see, the line obtained with the help of training set when applied to the testing set gives pretty accurate results.

Task 2:
Can the intercept in linear regression be negative? Is the intercept always meaningful? Please research and present your findings.

Answer:
Yes, the intercept in linear regression can be negative. This simply means that the expected value of the dependent variable will be less than 0 when all the independent variables are set to 0. The intercept is meaningful but not always. Typically, it is the overall relationships between the variables that will be of the most importance in a linear regression model, not the value of the constant. Constants in a simple regression equation do not always make practical sense. They may be positive or negative but impractical. For example, an illustration shows a correlation between height and weight with an equation of $Y' = a + bX = -159.31 + 4.62X$. As expected, the slope, b, is positive. The Y-intercept, a, however is negative and it is of no practical predictive value. It states that someone who has zero height weighs minus 159.31 pounds. Which makes no sense.

Task 3:
We have seen coefficient of determination r-squared in our last lecture. There is another commonly used quantity called adjusted r-squared. What is the different between the two? Why do we need adjusted r-squared when we have r-squared?

Answer:
Every time you add an independent variable to a model, the R-squared increases, even if the independent variable is insignificant. It never declines. Whereas Adjusted R-squared increases only when independent variable is significant and affects dependent variable. Adjusted R squared value is more reliable and accurate in determining the efficiency of the model. Adjusted R squared includes the number of independent variables in its formula whereas the R squared does not.