DA ASSIGNMENT1 PART(III)

NAME: SNEHA SAVARKAR Subject code: OEIT6 UID: 2019120055

Correlation:

Variables within a dataset can be related for lots of reasons.

For example:

- One variable could cause or depend on the values of another variable.
- One variable could be lightly associated with another variable.
- Two variables could depend on a third unknown variable.

It can be useful in data analysis and modeling to better understand the relationships between variables. The statistical relationship between two variables is referred to as their correlation.

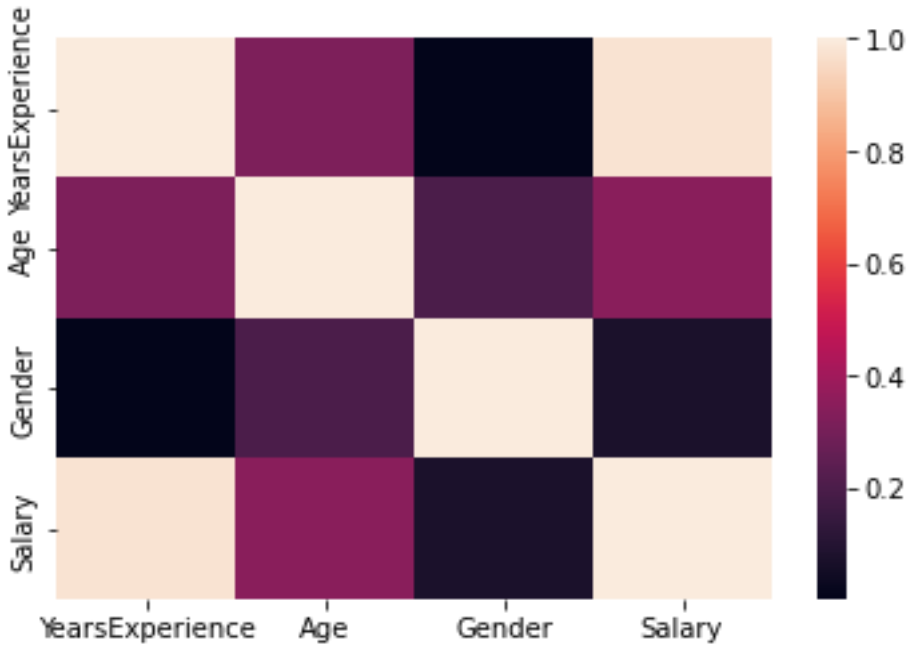
A correlation could be positive, meaning both variables move in the same direction, or negative, meaning that when one variable's value increases, the other variables' values decrease. Correlation can also be neural or zero, meaning that the variables are unrelated.

- Positive Correlation: both variables change in the same direction.
- Neutral Correlation: No relationship in the change of the variables.
- Negative Correlation: variables change in opposite directions.
- The performance of some algorithms can deteriorate if two or more variables are tightly related, called multicollinearity. An example is linear regression, where one of the offending correlated variables should be removed in order to improve the skill of the model.

One' may also be interested in the correlation between input variables with the output variable in order provide insight into which variables may or may not be relevant as input for developing a model.

The structure of the relationship may be known, e.g. it may be linear, or we may have no idea whether a relationship exists between two variables or what structure it may take. Depending what is known about the relationship and the distribution of the variables, different correlation scores can be calculated.

Output:



As we can see from the above heatmap that the salary and the years of experience have positive correlation between them. Where as the other factor such as age, gender are positively correlated.

Perhaps if we do, PCA on this dataset, we might get better accuracy.

PCA

{1: 0.9999999076659841, 2: 0.99999999921403, 3: 0.9999999997476197}

As the dataset is quite small we can see that almost the entire data is captured in al the variables this gives us an idea that our data is ready to try modelling.

Linear Regression:

Linear regression is a predictive modeling technique for predicting a numeric response variable based on one or more explanatory variables. The term "regression" in predictive modeling generally refers to any modeling task that involves predicting a real number (as opposed classification, which involves predicting a category or class.). The term "linear" in the name linear regression refers to the fact that the method models data with linear combination of the explanatory variables. A linear combination is an expression where one or more variables are scaled by a constant factor and added together. In the case of linear regression with a single explanatory variable, the linear combination used in linear regression can be expressed as:

response = intercept + constant $*$ explanatory

The right side if the equation defines a line with a certain y-intercept and slope times the explanatory variable. In other words, linear regression in its most basic form fits a straight line to the response variable. The model is designed to fit a line that minimizes the squared differences (also called errors or residuals.).

Training and test Dataset:

So in order to work on the and to test if the model we have created is working for all the conditions/dataset will be divinding the data into two parts. As the training dataset will be much bigger than the test dataset. Will be dividing it into 2/3 and 1/3 rspectively. Our test data set will be 33.33% of our entire dataset. Entirely will be dividing the dataset into 4 parts training set, test set, depending variable of training set and depending variable of test set.

What are exactly the coefficients b0 and b1 in the Simple Linear Regression equation:

'Salary = b0 + b1 × Experience'

b0 is the salary you get with no experience and b1 is the increase in salary per year.

Why do we take the squared differences and simply not the absolute differences?

Because the squared differences makes it easier to derive a regression line. Indeed, to find that line we need to compute the first derivative of the loss error function, and it is much harder to compute the derivative of absolute values than squared values

Training the Simple Linear Regression model on the Training set

With the help of sklearn will train our training set to perform linear regression.

Output:

From the above graph we can see that the original data which is a red scatter plot and the data that we have predicted with the help of linear regression is nearly the the same. It is the best fit for the dataset which we have selected.

As we perform the regression model on training set we also need to see what happens to our test set with same condition. The above graph represents the Linear regression model of the test set and if we compare the two , training set and the test set results it is nearly equal.

Making a single prediction

What will be the salary of an employee with 12 years of experience

With the help of our model will try to predict it:

[138967.5015615]

Therefore, our model predicts that the salary of an employee with 12 years of experience is $ 138967,5.

Final Linear Regression Equation with the values of coefficients:

[9345.94244312]
26816.192244031183

Therefore, the equation of our simple linear regression model is:

Salary= 9345.94 × YearsExperience + 26816.19
With the help of this equation we can say that if the experience is zero that is the starting salary of a fresher will be 26816.19.

Conclusion:

From this we can conclude that we have successfully performed linear regression on a dataset.