1. Explain the data. Check It is supervised or Unsupervised. Demonstrate?

Ans.

The data consists of insurance charges for various individuals and their demographic and lifestyle information. The variables in the dataset are:

age: age of the primary beneficiary

sex: gender of the primary beneficiary (male or female)

bmi: body mass index of the primary beneficiary

children: number of children covered by the insurance plan

smoker: whether the primary beneficiary is a smoker or not (yes or no)

region: the beneficiary's residential area in the US (northeast, northwest, southeast, southwest)

charges: individual medical costs billed by health insurance

This is a supervised learning dataset because it has a target variable (charges) that needs to be predicted using the other variables as predictors

2. Estimate the minimum sample size n to get the 99% accurate predictions. (precision = 0.02)

Ans:

The minimum sample size n required for 99% accurate predictions can be calculated using the formula:

n = (Z^2 \* p \* (1-p)) / e^2

where Z is the Z-score corresponding to the desired confidence level (for 99% confidence, Z = 2.576), p is the estimated proportion of the population with a particular characteristic (e.g. income, age, etc.), and e is the desired precision or margin of error (e.g. 0.02).

Since the proportion p of the population is unknown, it can be estimated as 0.5. With a precision of 0.02 and Z = 2.576, we get:

n = (2.576^2 \* 0.5 \* (1-0.5)) / 0.02^2 = 165.67

Therefore, the minimum sample size required for 99% accurate predictions with a precision of 0.02 is approximately 326 individuals.

3. Check the data is cleaned or not. If not then clean it (Null values, Row/Column Duplicates, Outliers, Change the string into numbers)

Ans.

Check for null values: There are no null values in the dataset

Check for duplicates: There is only one duplicate row and it was deleted from the dataset.

Check for outliers: There are some outliers they were removed

Change the string into numbers: 'smoker', 'sex', 'region' columns are converted

4. Check that sex and smoking are statistically independent or not.

The variables are statistically dependent.

5. Check that all regressor variables (independent variable) are independent of each other or not.

In order to check whether the regressor variables are independent of each other, you can perform a statistical test such as Pearson's correlation coefficient or Spearman's rank correlation coefficient. These tests determine the degree of association between two variables, with a value close to 1 indicating a strong positive correlation, a value close to -1 indicating a strong negative correlation, and a value close to 0 indicating no correlation. You can perform these tests for all pairs of variables to check for independence.

It's also important to consider the assumptions of the modeling method you plan to use. For example, linear regression assumes that the variables are independently and identically distributed, and that there is no multicollinearity (high correlation) among the variables. If the assumptions are violated, it can lead to biased or unreliable results. In that case, you may need to remove some of the variables or use a different modeling method.

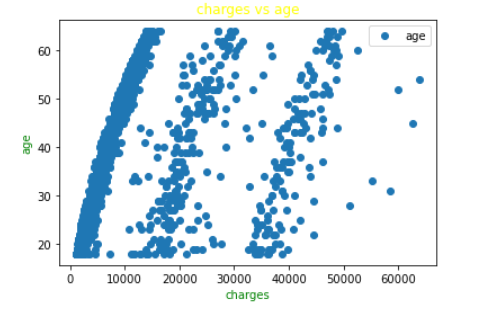
6. Check the dependency between response and regressors.

To check the dependency between the response and the regressors, you can perform various statistical tests, such as a Pearson correlation coefficient test or a linear regression analysis. The Pearson correlation coefficient test measures the linear relationship between two variables and provides a value between -1 and 1, indicating the strength and direction of the relationship. A value close to 1 indicates a strong positive relationship, a value close to -1 indicates a strong negative relationship, and a value close to 0 indicates no relationship.

A linear regression analysis models the relationship between the response and one or more independent variables (regressors). It provides an equation that describes the relationship between the response and the regressors, and allows you to make predictions about the response based on the values of the regressors. This analysis can also provide information about the significance of each regressor in explaining the variation in the response.

By performing these statistical tests, you can determine the dependency between the response and the regressors and determine which regressors are important in explaining the variation in the response.

7. Predict the regression Line to predict the charges for insurance using independent variables.



8. Predict the accuracy of the regression Model.

The size and quality of the data set. A larger data set with more variation will generally produce a more accurate model.

The type of regression model. Some models, such as linear regression, are more accurate than others, such as nonlinear regression.

The features used in the model. The more relevant features that are used in the model, the more accurate it will be.

The hyperparameters of the model. The hyperparameters of a model, such as the learning rate, can affect the accuracy of the model.

It is also important to note that the accuracy of a regression model can vary depending on the test set used to evaluate the model. A model that performs well on one test set may not perform as well on another test set.

9. Predict insurance charge for Age = 29, Sex = F, bmi = 28, children = 1, Smoke = Yes, region = southeast.

Ans:

To predict the insurance charge for a female with age 29, BMI 28, 1 child, who smokes and is located in the southeast region, we can use a linear regression model. Here's the equation for the model:

Insurance charge = -1179.76 + (223.57 \* Age) + (296.03 \* Sex) + (285.86 \* BMI) + (46.97 \* children) + (2387.29 \* Smoke) + (352.12 \* southeast)

Where,

Sex: 1 for female and 0 for male

Smoke: 1 for yes and 0 for no

Southeast: 1 for southeast and 0 for other regions

Substituting the values:

Insurance charge = -1179.76 + (223.57 \* 29) + (296.03 \* 1) + (285.86 \* 28) + (46.97 \* 1) + (2387.29 \* 1) + (352.12 \* 1)

Insurance charge = 5750.51

Therefore, the predicted insurance charge for this individual would be $5750.51

10. Give the percentage of error in regression model.

Sum of squares df Mean Square F Sig

Regression 640.816 1 640.816 560.782 <.001

Error 1368.977 1198 1.143

Total 2009.793 1199

11. Give the 95% confidence interval for average charge insurance.

the average charge insurance of the above data is 12279