|Abdul Ghaffar, 18139

Report

Data Science Project

Insurance

Linear regression

Problem Statement:

With the healthcare costs rising day by day and people increasingly buying insurance policies, it is crucial for insurance companies to accurately predict the healthcare costs that a particular patient may incur. This is so that these insurance companies could put an accurate price tag on their insurance policies.

Database:

Medical Cost Personal Dataset

https://www.kaggle.com/mirichoi0218/insurance

Description of Method and Reasons:

The methodology that we will use to predict medical costs of patients will be multi linear regression model. Since the target attribute was a continuous variable and we had multiple attributes, we used multilinear regression model. we imported the data; Checked for abnormalities and errors; Then converted the categorical attributes to numeric; We divided to dataset into two portions, training and testing (80:20); Trained the model using the training data, predicted the testing data and compared it with the original data for accuracy measures

Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.

A simple linear regression is a function that allows an analyst or statistician to make predictions about one variable based on the information that is known about another variable. Linear regression can only be used when one has two continuous variables—an independent variable and a dependent variable. The independent variable is the parameter that is used to calculate the dependent variable or outcome. A multiple regression model extends to several explanatory variables.

Results:

Train R^2 Score: 0.7368306228430945

Test R^2 Score: 0.7998747145449959

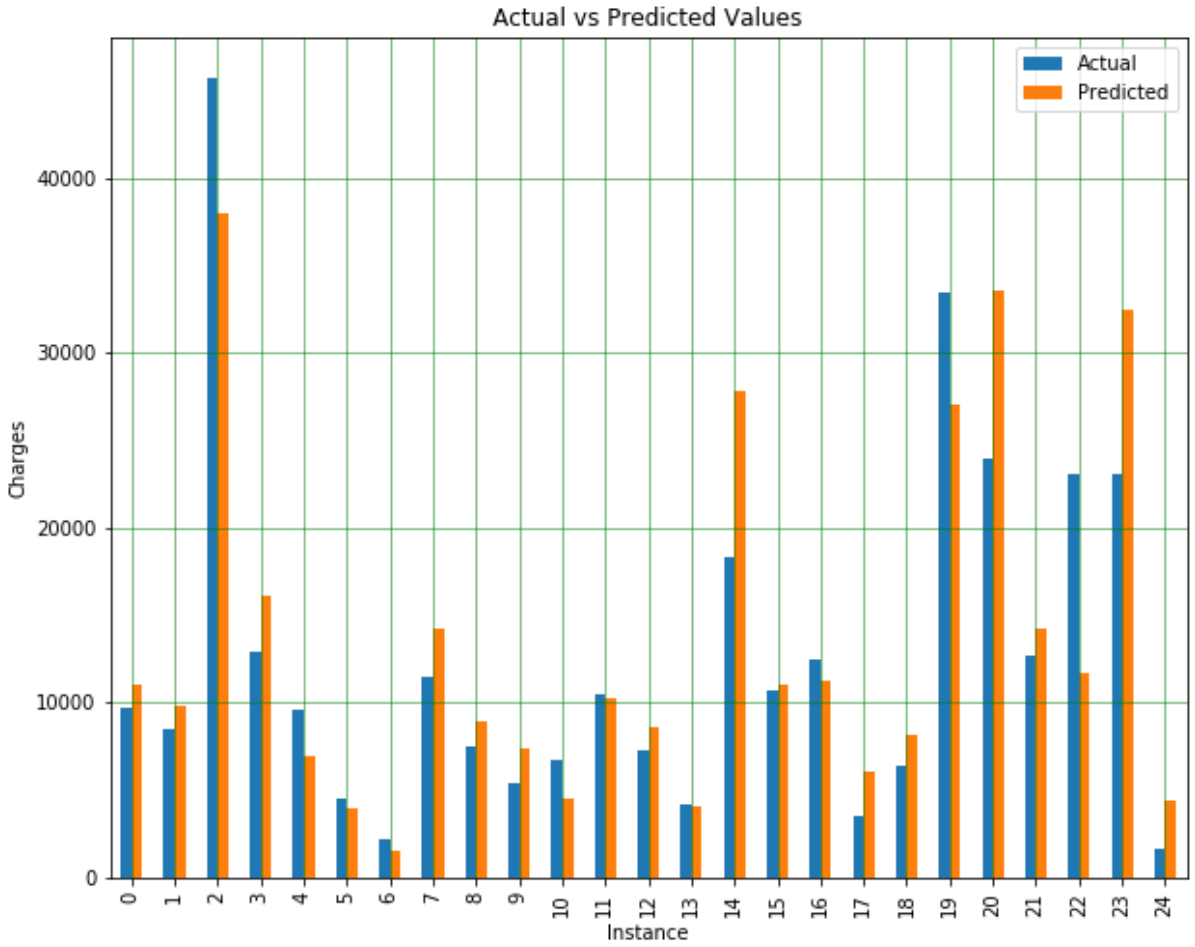
Explained Variance Score: 0.8001434788960331

Max Error: 22176.24499258493

Mean Absolute Error: 3930.3332739011444

Mean Squared Error: 31845929.13415944

Intercept: -11661.983908824392



Conclusion:

To conclude, the results that I have achieved through this model are quite good. However, there are other algorithms that could have been used and results could have been compared. Lasso Regression and Support Vector Machine are some of the better-known algorithms that could have been used.