HEALTH INSURANCE CHARGES PREDICTION WITH REGRESSION

IBM Supervised Machine Learning: Regression - Final Assignment

By:

Azeed Soliat Omolara

TABLE OF CONTENT

Instructions:

In this Assignment, you will demonstrate the data regression skills you have learned by completing this course. You are expected to leverage a wide variety of tools, but also this report should focus on present findings, insights, and next steps. You may include some visuals from your code output, but this report is intended as a summary of your findings, not as a code review.

The grading will center around 5 main points:

- Does the report include a section describing the data?
- Does the report include a paragraph detailing the main objective(s) of this analysis?
- Does the report include a section with variations of linear regression models and specifies which one is the model that best suits the main objective(s) of this analysis.
- Does the report include a clear and well-presented section with key findings related to the main objective(s) of the analysis?
- Does the report highlight possible flaws in the model and a plan of action to revisit this analysis with additional data or different predictive modeling techniques?

About the Data:

This dataset was gotten from kaggle, the dataset is has various features that can be used to predict health insurance of individuals. Below are the summary of the columns.

Columns:

- **Age**: age of primary beneficiary
- **Sex**: insurance contractor gender, female, male
- **Bmi**: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- **Children**: Number of children covered by health insurance / Number of dependents
- **Smoker**: Smoking
- **Region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **Charges**: Individual medical costs billed by health insurance

Objectives:

The main objective of this analysis is to predict personal health insurance charges of individuals based on their features. other objectives of this analysis include the following;

- To explore relationships of the various features.
- To know the key determinant features for insurance charges
- To know the correlation between insurance charges and other features.

Data Cleaning and Exploration:

- The dataset comprises of 1338 rows and 7 columns.
- The dataset has 1338 non_null values, which implies that the datasets has no missing value since it's equal to the number of rows of the dataset. Also, the dataset has 2 integer features, 3 object features, and 2 float features.
- The dataset has 676 males and 662 females.
- The dataset has 1064 non-smokers and 274 smokers.
- The dataset has 364 southeast entries, 325 southwest entries, 325 northwest entries, and 324 northeast entries
- The dataset has 2 duplicates entries, which was dropped and shorten the length of the dataset to 1336.
- All the object features in the dataset was convert to numerical features.

Exploratory Data Analysis and Engineering:

- The statistical summary of the dataset is checked to show the standard deviation, count, mean, min, max, and interquartile range of the dataset.
- The correlation of the various features of the dataset to check for the distribution of their relationships, multicollearity among independents features are being explored to know better about the dataset.
- Seaborn **Pairplot** function is being used to show the visualization of the correlations between all the features of the dataset.
- Seaborn Regplot is being used to show the kind of relationships between features like Age, BMI, children, and the target feature **Charges.**
- The dataset is being divided into the independent and target features. The independent features includes all the features in the dataset excluding **Charges** while the target feature only includes **Charges.**
- All the independent features were scaled and transformed to a small scale using the **StandardScaler** function on Sklearn.
- All the data were later divided into training and testing for the machine learning prediction using the **train_test_split** function on Sklearn.

Linear Models:

To predict the health insurance charges with features like age, gender, bmi, and among others. The following models were used:

- Linear Regression model as the baseline model
- Ridge Regression for L2 Regularization
- Lasso Regression for L1 Regularization
- Polynomial Features as an added features to the model.

Summary and Insights:

From the three regression models, the following insights can be drawn:

- The Linear regression has the R_score for the training set is 75% while the testing score is 74% without any polynomial or regularization features added. The visualization of the actual values vs. the predicted values shows that the predicted values is slightly more than the actual values and the model is not well fitted.
- The Ridge regression without polynomial feature training set R_score is 74%, while the testing score is also 74% compare to the linear regression score this is slightly worse. The visualization of the actual values vs. the predicted values shows that the model is not well fitted. The Ridge regression with the polynomial feature training set R_score is 84% and the testing set is 80% compare to the linear regression and the ridge regression without polynomial features, the score is better. The visualization of the actual values vs. predicted values show that the predicted values fitted the actual values better than the linear regression and ridge regression without polynomial features.
- The Lasso regression without polynomial feature training set R_score is 74%, while the testing score is also 74% compare to the linear regression score this is slightly worse, however it is same with the Ridge regression without polynomial features. The visualization of the actual values vs. the predicted values shows that the model is not well fitted. The Lasso regression with the polynomial feature training set R_score is 84% and the testing set is 80% compare to the linear regression and the ridge regression without polynomial features, the score is better. However, compare to the ridge regression with polynomial features the score is slightly worse. The visualization of the actual values vs. predicted values show that the predicted values fitted the actual values better than the linear regression, ridge regression without polynomial features, and lasso regression without polynomial features, but slightly worse compare to the ridge regression with polynomial features.

Key Findings:

- The dataset do well with polynomial features compare to using only Linear regression and both regularizations alone.
- In terms of accuracy and explainablity of the model, the Ridge Regression with polynomial features is the best for the main objectives of the analysis

Next Steps:

- The dataset need more features for it to be a more better model, so there is need to add more relevant features to the dataset in the future.
- Also, there is need to add more dataset in the future, because the more the data the better the model.

Thanks for Reading!!!