Author: Piero Termini

Dec/2021

Insurance Price Predictor (Machine Learning)

About:

The following machine learning project has been done with the objective of creating a machine learning model to predict the yearly price of private health insurance in the United States based on the general data of the clients.

DATA:

The dataset used is a dataset obtained from a .csv file obtained from Kaggle - link:

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

PROCESS:

-EDA & Data Cleaning:

Initially the we started by loading the data into a jupyter notebook, we analyzed the variables of the dataset and after doing an analysis of the correlation between the variables and their importance.

As our intended goal is to determine who much should a client pay for their insurance based on their general characteristic, we selected the variable “charges” which would be the yearly cost of the policy as the target to analyze.

We proceeded to clean the data, dropping some variables which did not have any relevance for the objective of the model and creating a new .csv file to be used for the model training.

-Model Training:

We loaded the cleaned data and started working with it.

First, most of the categorical values were transformed into numeric values as the problem we face in this project is a regression problem (we want to get a price as the prediction result based on the other variables).

After that, we did a splitting of the data to train and test the models we were going to build.

After we had our train and tests set ready, we started creating and testing several regression models.

The models created were:

* LinnerRegression
* PolynomialRegression
* Lasso
* Ridge
* DecisionTreeRegressor
* RandomForestRegressor
* SVM
* KNN

After training and testing all of this model with the regular data, we proceded to use the method of StandarScaler to see if the performance of the models was improved and we retrained each one with the new scaled dataset.

With the scale testing done, it was observed that the best metrics results were obtained by the DecisionTreeRegressor, RandomForestRegressor and SVM models.

We performed a GridSearch to improve find the ideal parameters of each model.

The GridSearch results showed a clear winner with the RandomForest model.

We proceeded to test the model and the predictions and after checking that it worked, we went ahead and saved it for production.