Health Insurance Cost Prediction Using IBM Auto AI Service

Introduction

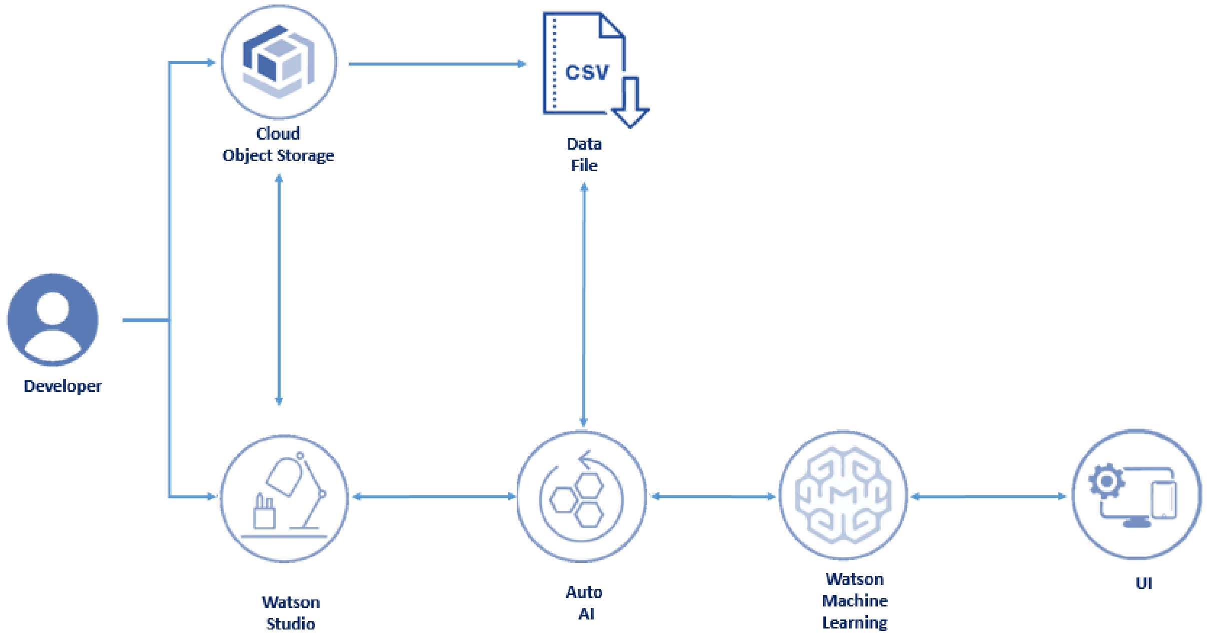
In this project, we study the effects of age, smoking, BMI, gender, and region to determine how much of a difference these factors can make on your insurance premium. By using our application, customers see the radical difference their lifestyle choices make on their insurance charges. By leveraging artificial intelligence (AI) and machine learning, we help customers understand just how much smoking increases their premium by predicting how much they will have to pay within seconds.

To build this project we will be using IBM AutoAI. You create a model from a data set that includes the age, gender, BMI, number of children, smoking preferences, region, and charges to predict the health insurance premium cost that an individual pays.

Services Used:

1. IBM Watson Studio
2. IBM Watson Machine Learning
3. Node-RED
4. IBM Cloud Object Storage

Architecture



Madhavi B Desai
Email ID : mbdesai@fetr.ac.in

Objectives of Project

1. Design of Health Insurance Premium Prediction model using IBM Auto AI Service and Insurance Premium Dataset

2. Comparative analysis of various Maching learning algorithms

3. Design of User Interface

Insurance Premium Dataset Description:

The insurance.csv dataset contains 1338 observations (rows) and 7 features (columns). The dataset contains 4 numerical features (age, bmi, children and expenses) and 3 nominal features (sex, smoker and region) that were converted into factors with numerical value designated for each level.

Link of Dataset

Insurance.csv file is obtained from the Machine Learning course website (Spring 2017) from Professor Eric Suess at http://www.sci.csueastbay.edu/~esuess/stat6620/#week-6.

Machine Learning Model Description

Random Forest Regressor

Random forest is a Supervised Learning algorithm which uses ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree.

Gradient Boosting Regressor

Gradient Boosting Regressors (GBR) are ensemble decision tree regressor models. At each step, a new tree is trained against the negative gradient of the loss function, which is analogous to (or identical to, in the case of least-squares error) the residual error. Gradient boosting involves three elements:

- A loss function to be optimized.
- A weak learner to make predictions.
- An additive model to add weak learners to minimize the loss function.

Experiment Results:

Table 1 displays the results of Cross Validation Score for health premium prediction

Table 1: Cross Validation Score Results for Health Premium Prediction

Sr No	Machine Learning Algorithm	RMSE	R^2	Explained Variance	MSE	MSLE	MAE	Median Absolute Error (MedAE)	Root Mean Squared Log Error (RMSLE)
1	**Random Forest Regressor**	**4,616.347**	**0.850**	**0.850**	**21,367,376.902**		**2,569.672**	**1,427.742**	
2	Gradient Boosting Regressor	5,363.872	0.798	0.800	28,780,955.212	0.273	3,065.557	1,161.706	0.522

User Interface Design

User interface design is developed using node-red services of IBM cloud. Node-red file is shown in Figure 1.

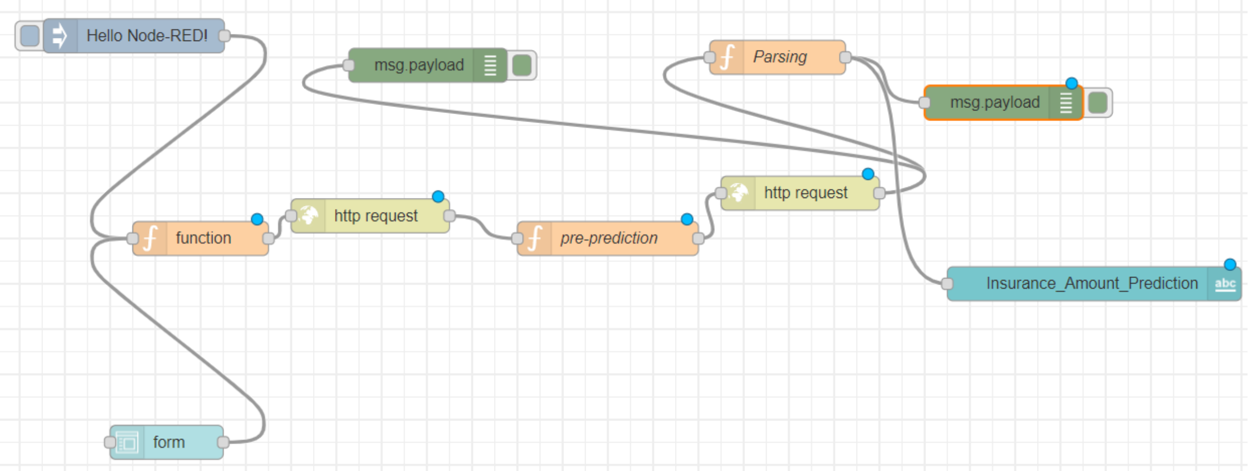
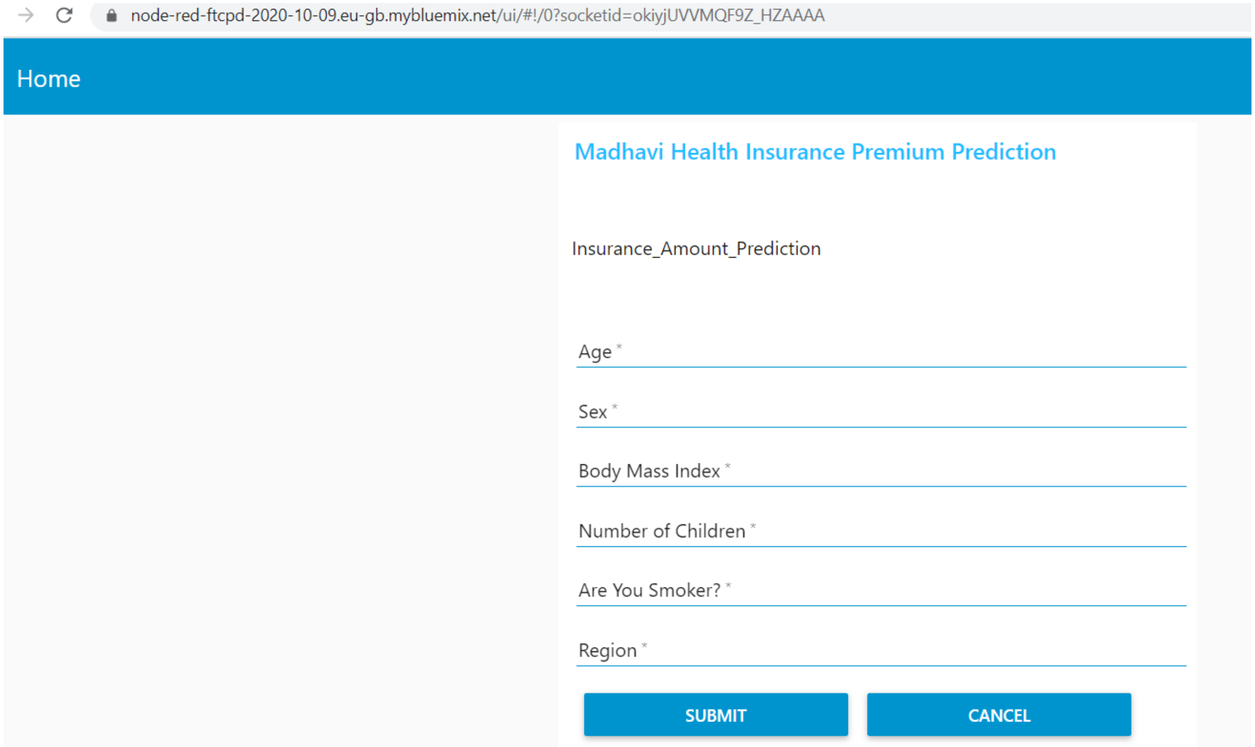


Figure 1: Node-Red flow for health premium prediction user interface

User interface design of health insurance premium prediction is shown below:



node-red-ftcpd-2020-10-09.eu-gb.mybluemix.net/ui/#!/0?socketid=okiyjUVVMQF9Z_HZAAAA

Home

Madhavi Health Insurance Premium Prediction

Insurance_Amount_Prediction

Age *

Sex *

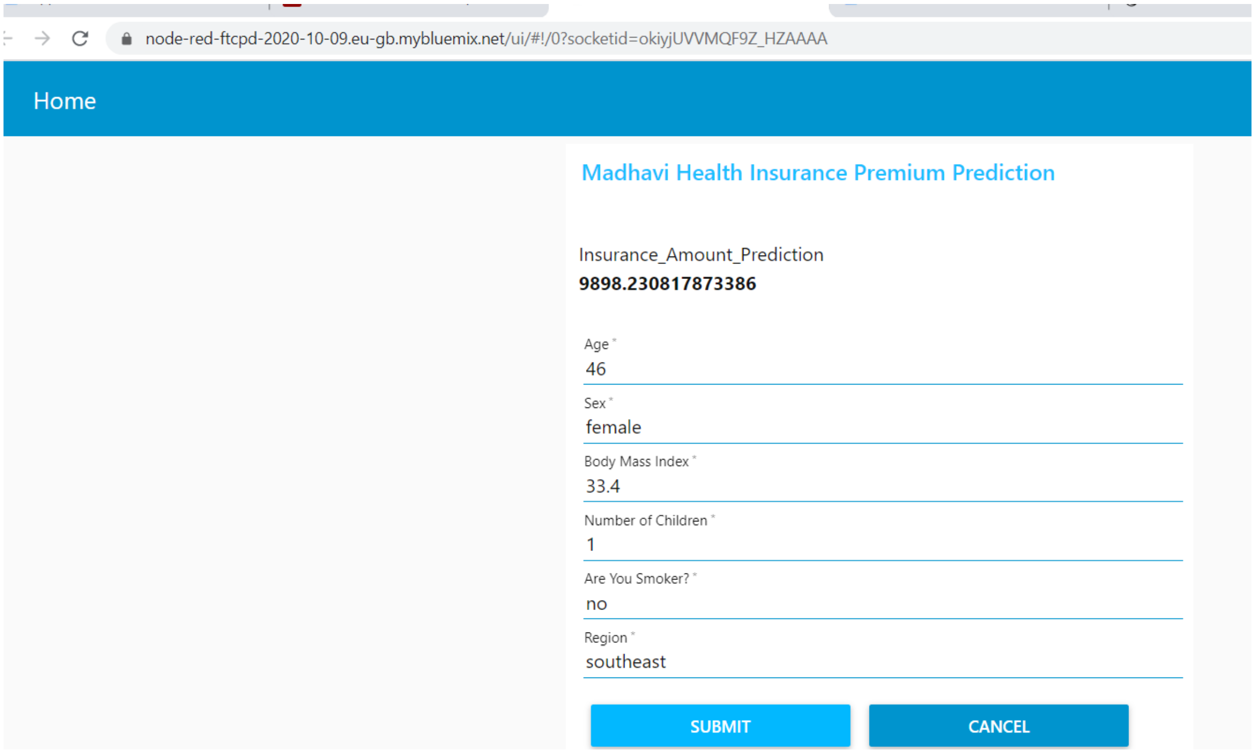
Body Mass Index *

Number of Children *

Are You Smoker? *

Region *

Figure 2 Health insurance premium prediction application



node-red-ftcpd-2020-10-09.eu-gb.mybluemix.net/ui/#!/0?socketid=okiyjUVVMQF9Z_HZAAAA

Home

Madhavi Health Insurance Premium Prediction

Insurance_Amount_Prediction
9898.230817873386

Age *
46

Sex *
female

Body Mass Index *
33.4

Number of Children *
1

Are You Smoker? *
no

Region *
southeast

Figure 3: Application output for premium prediction application

Conclusion:

From the results of table 1, we can observe that random forest regressor gives better result compare to gradient boosting regressor.

Author Details

Dr Madhavi Desai

Head and Associate Professor

R N G Patel Institute of Technology, Bardoli

Email Id : mbdesai@fetr.ac.in