

*Remote Summer Internship Program 2020 Machine Learning, Career Basic Program
Smartinternz, SmartBridge*

Health Insurance Cost Prediction Using Watson Auto AI

**Internship Report
by
Bhavnesht Mittal**

03/08/2020-03/09/2020

Contents

1. Program Information	1
2. Internship Description - Problem Statement	2
3. Design of Project	3
4. Machine Learning Model	3
4.1 Algorithm Formulated.....	3
4.2 Setting up the enviroment.....	3
4.3 Data Acquisition.....	4
4.4 Applying Auto AI experiment.....	4
4.5 Random Forest Regression.....	5
4.5.1 Introduction.....	5
4.5.2 Working.....	6
4.6 Node Red Flow.....	6
4.7 Form of ML model.....	7
5. Overview of Internhip Experience	8

1. Program Information

Summer Internship Program by Smartbridge is an annual initiative taken to teach and prepare students across the globe for industry experience. They believe that experiential learning and development in a professional like environment can only bridge the gap between students and industries opening ways for both of them to achieve better results. This initiative enables students to better their resume to embark upon a successful industrial journey.

They provide various roles for internship according to the possible aptitude of the students such as Artificial Intelligence, Machine Learning, Internet of Things etc. A project is assigned to students individually with access to the platform wherein the students code and deploy their model just like in industries. This not only introduces them to environments like IBM cloud but also helps them in understanding the industrial environment better.

Classes are organised according to the technologies to be taught and doubt sessions are taken up by the mentors to help students complete their projects too.

2. Internship Description - Problem Statement

Health Insurance Cost Prediction Using Watson Auto AI

My role at the internship was of a Machine Learning Engineer and the project given to me was to generate a Regression based Machine Learning model.

Health Insurance companies have a tough task at determining premiums for their customer. While the health care law in any country does have some rules for companies to follow to determine premiums, it's really up to the companies on what factor/s they want to hold more weightage. Companies should know the most important factors and how much statistical importance do they hold.

This project involved building of a machine learning model to predict Health Insurance Cost for a customer given various features like age of the customer, sex of the customer, whether the customer is a smoker or not, how many children do the customer has, Body Mass Index (BMI), and the region of the customer.

This model works on the dataset from kaggle to evaluate insurance cost of different people depending on the previous records.

Model is built using Watson Auto AI.

The IBM Watson Studio services are used to auto AI the experiment. A Node - RED flow is built to integrate the ML services or Auto AI.

3. Design of project

Methodology used

Work flow for this project involves acquiring the data for the project from kaggle and Auto AI experiment is used to develop the model which will further predict the insurance cost of the customer.

4. Machine Learning Model

4.1 Algorithm formulated to solve the given problem statement

Algorithm Steps:

Step 1: Import the dataset.

Step 2: Apply Auto AI experiment on the dataset.

Step 3: This Auto AI experiment applied will automatically generate number of models and will help to choose the optimum model with the help of comparison between all the generated model's RMSE values.

Step 4: The model with lowest RMSE value is selected.

Step 5: Then the model is needed to be deployed.

Step 6: After Deployment, the model can be tested on IBM Watson studio itself.

4.2 Setting Up the Environment

An IBM cloud account was set up to access various services to create and deploy the model. The following services have been used in the project:

1. Watson Studio - This is where the Auto AI experiment is applied on the dataset acquired from kaggle to get the optimum model for prediction .

2. Node Red - Node Red is the front end application that uses interconnecting nodes to interact with machine learning services of the cloud and the model to show predictions when inputs are given.

4.3 Data Acquisition

4.4 Applying Auto AI experiment on dataset

Model will be generated and optimum model will be selected.

Algorithm used but Auto AI experiment will be displayed on the page.

The algorithm used for this dataset which gives optimum model is Random forest Regressor.

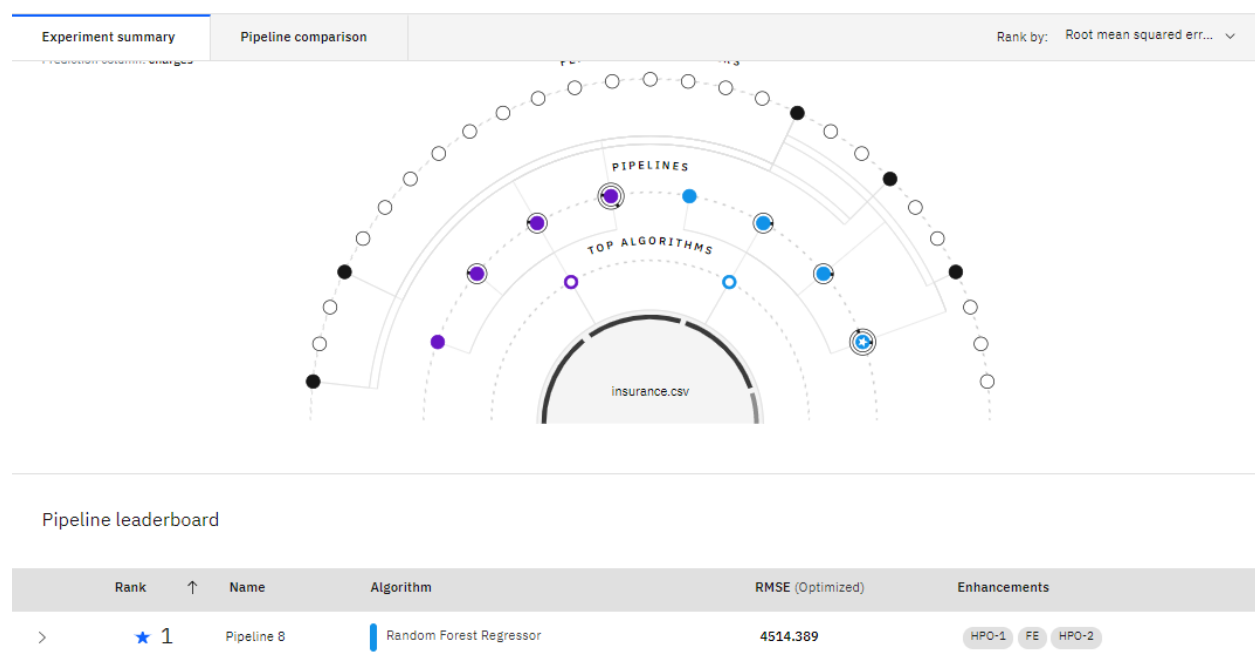


fig. 1

4.5 Random Forest Regression

4.5.1 Introduction

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate 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. Random decision forests correct for decision trees' habit of overfitting to their training set.

This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample.

In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

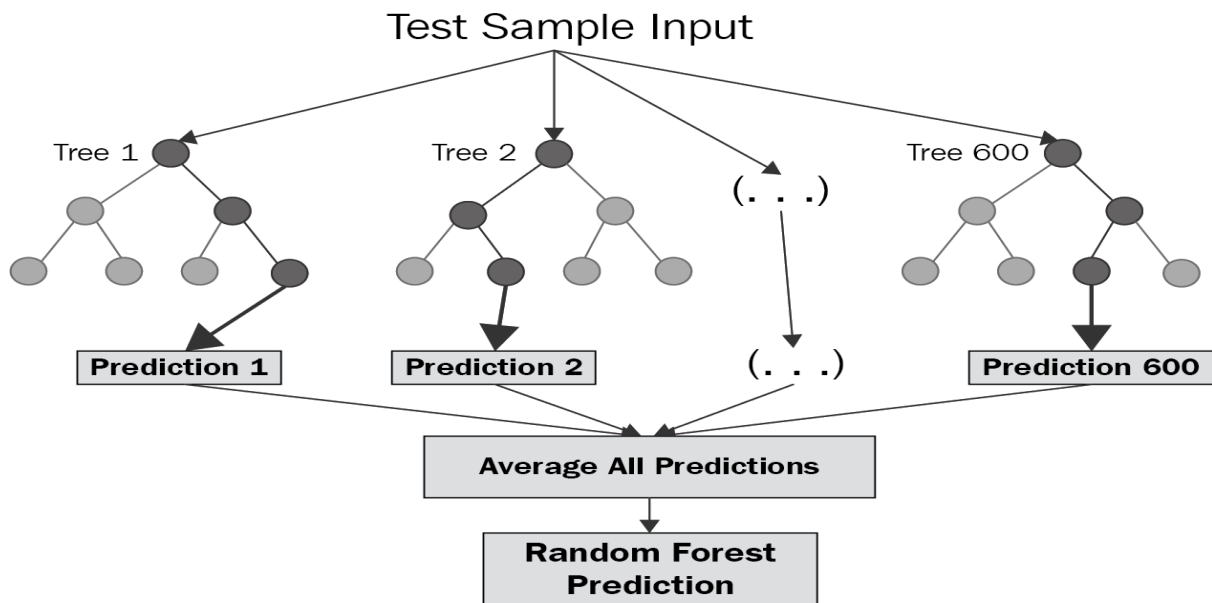


fig. 2

4.5.2 Working of Random Forest Regressor

It works in four steps:

- 1) Select random samples from a given dataset.
- 2) Construct a decision tree for each sample and get a prediction result from each decision tree.
- 3) Perform a vote for each predicted result.
- 4) Select the prediction result with the most votes as the final prediction.

4.6 Node Red Flow

A Node RED starter application was created to implement the front end of the project. In the starter application, nodes are dragged and dropped to create a flow to integrate the application with the machine learning model.

The following form appears after deployment of the app wherein the user can input values and life expectancy prediction is displayed according to the inputs.

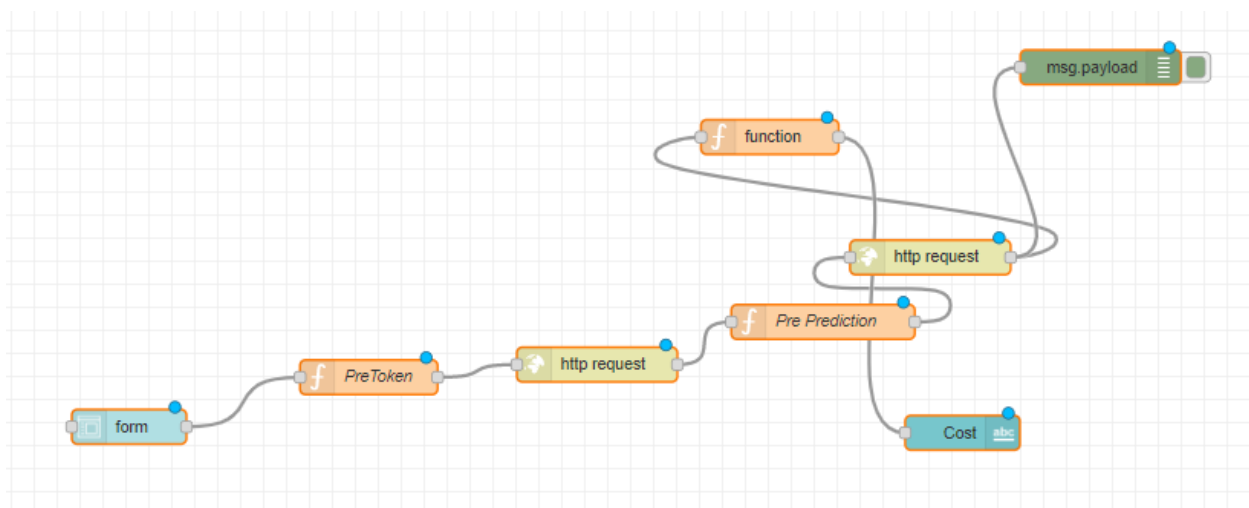
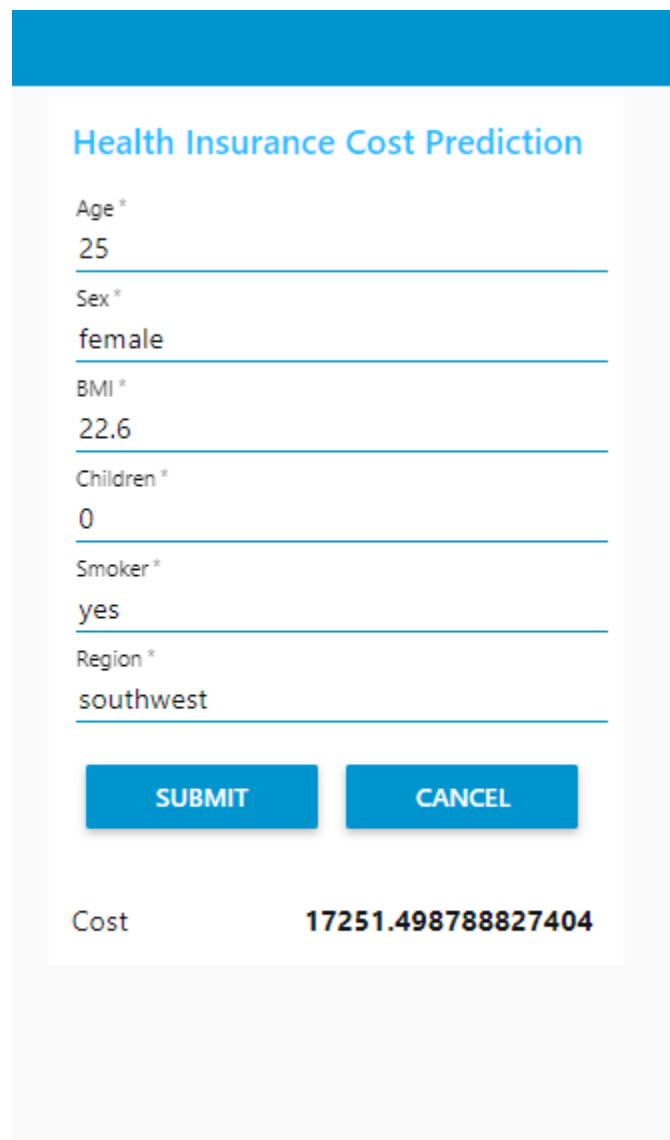


fig. 3

4.7 Form of ML model



A screenshot of a web form titled "Health Insurance Cost Prediction". The form contains six input fields with labels and asterisks indicating required fields: "Age *", "Sex *", "BMI *", "Children *", "Smoker *", and "Region *". The values entered are "25", "female", "22.6", "0", "yes", and "southwest" respectively. Below the inputs are two blue buttons labeled "SUBMIT" and "CANCEL". At the bottom, the predicted "Cost" is displayed as "17251.498788827404".

Field	Value
Age *	25
Sex *	female
BMI *	22.6
Children *	0
Smoker *	yes
Region *	southwest
Cost	17251.498788827404

fig. 4

5. Overview of Internship Experience

During my internship experience with Smart Internz, I was able to develop my Machine Learning skills. I particularly found the IBM cloud experience new and useful in improving my industrial skills. Although I found the Node RED service quite challenging, I found it to be valuable in developing my front end integration skills.