



Health Insurance Premium Predictor

Bachelor of Technology in Computer Science

Report Submitted By :

Group: **Cs-47**

Name	Reg No
Rohit Kumar	20184009
Saurav Chaudhary	20184165
Vidushi Vigh	20184188
Rishu Kumar	20184010

Mentored By: **Dr. Dushyant Kumar Singh**
(Prof CSED Dept)

Project Description

Health Insurance companies have a tough task at determining premiums for their customers. While the health care law in the United States does have some rules for the companies to follow to determine premiums, it's really up to the companies on what factor's they want to hold more weightage to.

The objective of this case study is to predict the health insurance cost incurred by Individuals based on their age, gender, BMI, number of children, smoking habit and geo-location.



Approach

This is Supervised Regression Problem. I considered these as independent variables (columns) in the dataset:

1. Age
2. Sex
3. Body Mass Index (BMI)
4. Smoker (Yes/No)
5. Region (Where the Person Lives)

These independent variables are assumed to have statistical significance in determining dependent variable insurance premium charges of the customer.

Dataset Source: <https://www.kaggle.com/noordeen/insurance-premium-prediction>

Steps Involved :

1. Description of Buiness Problem
2. Importing Library & Dataset
3. Data Pre-Processing
4. Data Visualization
5. Training & Testing Set Split
6. Train & Test
7. Deploy

Technology Used

Model :

Using Multiple Linear Regression - a machine learning technique - I try to determine the most (statistically) significant factors (independent variables) that influence the premiums charged (dependent variable) by an insurance company. I predicted the costs based on the insurance data that I obtained from Kaggle.com.

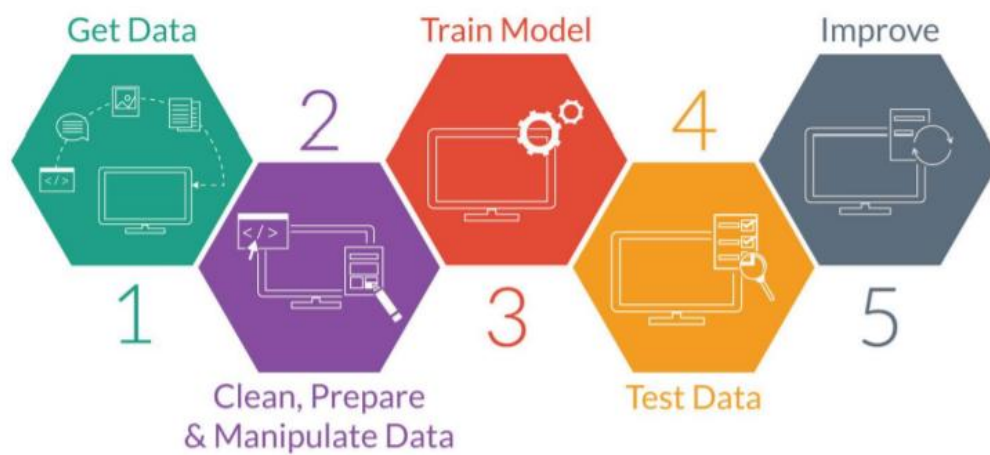
Multiple Linear Regression (Intuition)

- Multiple Linear Regression: examines relationship between more than two variables.
- Recall that Simple Linear regression is a statistical model that examines linear relationship between two variables only.
- Each independent variable has its own corresponding coefficient.

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n x_n$$

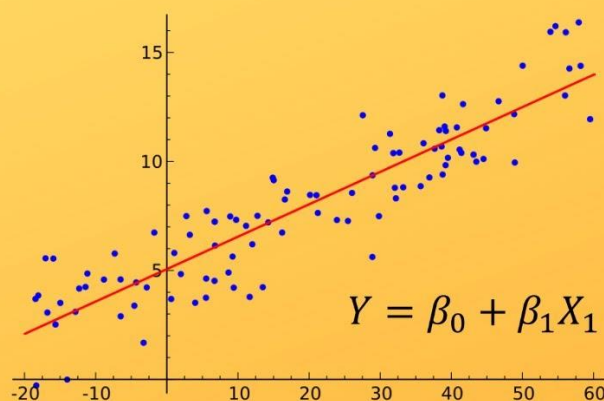
DEPENDANT VARIABLES
INSURANCE COST (\$)

INDEPENDENT VARIABLES
(AGE, SMOKING HABITS, REGION,..ETC)



The goal is to predict the premium charge which is a numeric outcome. So, regression model.

Multiple Linear Regression



number of predictors

Tools Used

Jupyter Notebook:

Jupyter notebooks can illustrate the analysis process step by step by arranging the stuff like code, images, text, output etc. It helps a data scientist to document the thought process while developing the analysis process. One can also capture the result as the part of the notebook.

Spyder:

Spyder is an acronym for Scientific Python development environment used in machine learning projects developed in Python language. It's also called an open source cross-platform IDE. Anaconda Spyder has some essential features like advanced editing, interactive testing, and debugging.

AWS SageMaker:

Amazon SageMaker is a fully managed service that provides every developer and data scientist with the ability to build, train, and deploy machine learning (ML).

Contribution

Rohit Kumar: -

I worked on calling an Amazon SageMaker model endpoint using Amazon API Gateway and AWS Lambda.

Then I along with Saurav Chaudhary implemented the machine learning model on the pre-processed data. I trained & tested for the given dataset and checked the accuracy.

Saurav Chaudhary: -

I worked on Data Pre-processing and alongside learn AWS Sage Maker to train test and deploy the model.

Then I along with Rohit Kumar implemented the machine learning model on the pre-processed data. I trained & tested for the given dataset and checked the accuracy.

Vidushi Vigh: -

I worked on the Data Visualization along with Rishu Kumar. I took a deep dive about why AWS for Machine Learning and came to know that AWS offers the broadest and deepest set of machine learning services and supporting cloud infrastructure, putting machine learning in the hands of every developer, data scientist and expert practitioner.

Rishu Kumar: -

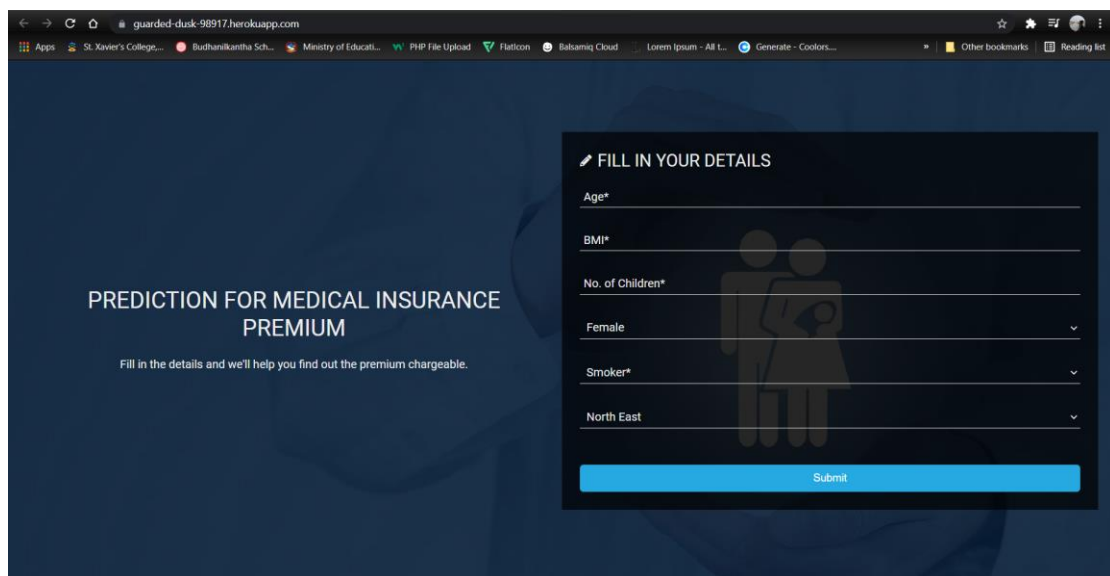
I worked on the Data Visualization along with Vidushi Vigh then I worked on understanding the Business Problem and Visualizing the Data. And I studied the various error calculation mechanism for assessing our model.

Prototype

Demo Video Link :

<https://drive.google.com/file/d/17Gmo7VG8Rfo3AJIH3vEpS8vvDImLczU1/view?usp=sharing>

ScreenShot :



The screenshot shows a web browser window with the URL `guarded-dusk-98917.herokuapp.com`. The page has a dark blue background with a faint silhouette of a person. The main heading is "PREDICTION FOR MEDICAL INSURANCE PREMIUM" in white. Below it, a subtext says "Fill in the details and we'll help you find out the premium chargeable." On the right side, there is a white form titled "FILL IN YOUR DETAILS" with a pencil icon. The form contains the following fields: "Age*" (text input), "BMI*" (text input), "No. of Children*" (text input), "Female" (dropdown menu), "Smoker*" (dropdown menu), and "North East" (dropdown menu). A blue "Submit" button is located at the bottom of the form.

Future Scope

1. Web App → Mobile App
2. Better UI for Customers
3. Improving Predictor Accuracy
4. Adding Extra Features