GALGOTIAS UNIVERSITY
(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

HEALTH INSURANCE AMOUNT PREDICTION

Submitted by

ROHAN NEGI

1613101585

16SCSE101127

in partial fulfillment for the award of the degree of

Bachelor of Technology

IN

Computer Science and Engineering

SCHOOL OF COMPUTER SCIENCE

Under the Supervision of

Dr. Kuldeep Singh Kaswan

Associate Professor

APRIL / MAY 2020

TABLE OF CONTENTS

CHAPTER 1

Abstract

Health insurance market is a crucial market and everyone needs some level of health care. Health insurance is one of the most significant investment an individual makes every year. This study is an effort to find mathematical models to predict future amount and verify results using regression models.

In this project, health data has been used to analyze and predict insurance amount for individuals. Three regression models naming Multiple Linear Regression, Decision tree Regression, Gradient Boosting Decision tree Regression have been used to compare and contrast the performance of these algorithms.

To test and verify the model, Health data was used as inputs for training the models and the predicted amount was compared with the actual data to compare the accuracies of the models. It was found that multiple linear regression and gradient boosting algorithms performed better than the linear regression and decision tree. It was found that gradient boosting is the winner, although its performance is comparable to multiple regression, but it takes much less computational time to achieve the same performance metrics.

CHAPTER 2

Introduction

1.1 Overall description

Medical costs that occur due to illness, accidents or any other health reasons are considerably expensive, by having health insurance, an individual is not liable for paying the entire medical costs of the procedure. Health insurance is one of the most significant investment an individual makes every year. This study is an effort to find mathematical models to predict future amount and verify results.

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables.

We use various Regression model which takes the input from the data set and predict the amount. Later it compares the amount with original data set and check for accuracy.

1.2 Purpose

The purpose of this analysis is to build a prediction model to predict future health insurance amount as well as comparing various regression models in order to select the most accurate one. The most accurate regression model is then opted for the prediction of insurance amount.

This application helps people to estimate their medical costs that may occur due to illness, accidents or any other health reasons so that they can plan and make necessary financial adjustments earlier in their life.

Since medical costs are incredibly expensive, this application hopes to use statistical knowledge in order to figure out an estimate amount, leading to a more healthier and peaceful living conditions.

1.3 **Motivation and scope**

Health insurance is a necessity nowadays, and almost every individual is linked with a government or private health insurance company. Factors determining the amount of insurance vary from company to company. Also people in rural areas are unaware of the fact that government of India provide free health insurance to those below poverty line. It is very complex method and some rural people either buy some private health insurance or do not invest money in health insurance at all. Apart from this people can be fooled easily about the amount of the insurance and may unnecessarily buy some expensive health insurance

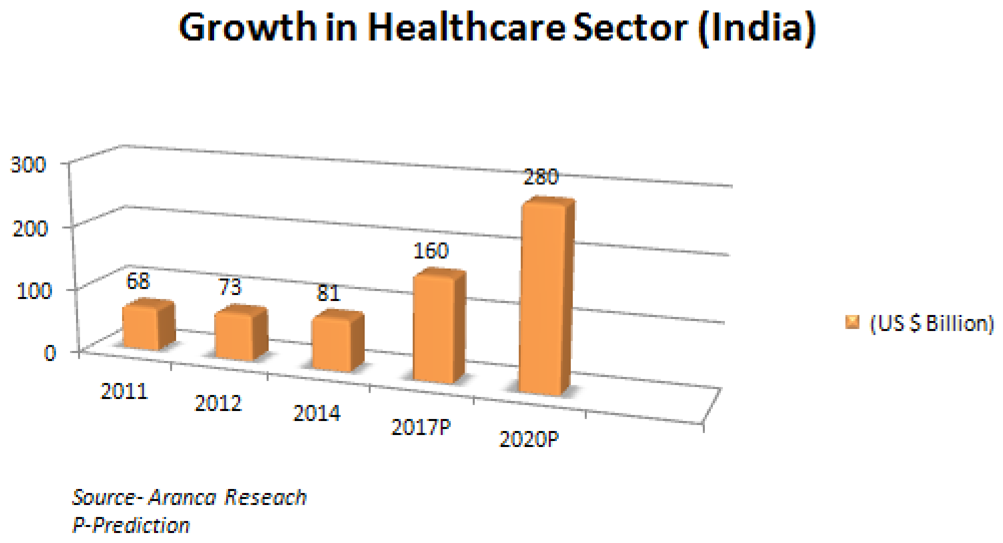
- The Scope is to create a project which allows a person to get an idea about the necessary amount required according to their own health status.
- Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project.
- The factors which affect the premium amount and the degree by which they affect the premium amount.

CHAPTER 3

Literature Survey

According to National Family Health Survey-3, the private medical sector remains the primary source of health care for 70% of households in urban areas and 63% of households in rural areas. The study conducted by IMS Institute for Healthcare Informatics in 2013, across 12 states in over 14,000 households indicated a steady increase in the usage of private healthcare facilities over the last 25 years for both Out Patient and In Patient services, across rural and urban areas. In terms of healthcare quality in the private sector, a 2012 study by Sanjay Basu et al., published in PLOS Medicine, indicated that health care providers in the private sector were more likely to spend a longer duration with their patients and conduct physical exams as a part of the visit compared to those working in public healthcare.

The figure shows the growth of the healthcare industry in India over the past years. The growth is exponentially increasing and according to the data, this industry will continue to grow.



Source- Aranca Reseach
P-Prediction

Growth in Health Care Sectors(India) through the years.

Health insurance market is a crucial market and everyone needs some level of health care. Health insurance is one of the most significant investment an individual makes every year. This project is an effort to find mathematical models to predict future amount.

CHAPTER 4

Proposed System

In this project, three regression models are evaluated for individual health insurance data. The health insurance data was used to develop the three regression models, and the predicted premiums from these models were compared with actual premiums to compare the accuracies of these models.

1.Linear Regression

Basic regression methods can be used to explore more complex relationships and more socially significant questions. Simple linear regression can be used to model one dependent variable and one independent variable. Linear regression with multiple variables is also known as multiple linear regression. Multiple linear regression models the relationship between dependent and independent variables

2.Decision Tree Regression

Decision tree modeling is done by constructing a decision tree for the entire input data. In the Decision tree input variables are represented in the branches and output values are represented in the leaves. Decision tree is one of the predictive modeling approaches used in statistics, data mining and machine learning.

Decision tree algorithm is a top-down approach that begins with a root node (or feature) and then selects a feature at each step that gives the best split of the data set, as measured by the information gain of this split.

3.Gradient Boosting Regression

Gradient Tree Boosting or Gradient Boosted Regression Trees (GBRT) is a generalization of boosting to arbitrary differentiable loss functions. GBRT is an accurate and effective off-the-shelf procedure that can be used for both regression and classification problems. Gradient Tree Boosting models are used in a variety of areas including Web search ranking and ecology.

Gradient boosting regressor takes the weak decision tree and builds stronger decision tree by incrementing the wrongly identified data by a previous decision tree. This approach allows gradient boosting decision tree to focus less on easy to predict decision tree and more on difficult cases that result in high computation time to train the data.

Sample Of Data Set

A sample from the actual dataset containing first 5 records. There are 6 attributes in the dataset. They are age, gender, bmi, children, smoker and charges.

	age	gender	bmi	children	smoker	charges
0	**19**	**female**	**27.900**	**0**	**yes**	**16884.92400**
1	**18**	**male**	**33.770**	**1**	**no**	**1725.55230**
2	**28**	**male**	**33.000**	**3**	**no**	**4449.46200**
3	**33**	**male**	**22.705**	**0**	**no**	**21984.47061**
4	**32**	**male**	**28.880**	**0**	**no**	**3866.85520**

Factors Affecting Amount Prediction:

In a dataset not every attribute has an impact on the prediction. Whereas some attributes even decline the accuracy, so it becomes necessary to remove these attributes from the features of the code. Removing such attributes not only help in improving accuracy but also the overall performance and speed.

- **Pre-existing medical conditions:** The policyholder or applicant will need to provide your own health records to ensure there aren’t any pre-existing medical conditions. But if, you do have any pre -existing conditions, then the company can choose to allow it in their policies or can even decide not to cover it, and if the insurance company cannot cover it under the health insurance then the policyholder will need to bear the costs. Thereby increasing and affecting the premium.
- **Family medical history:** If the policyholder’s family have certain medical ailments their premiums may be higher than others. No one can do anything with their genes. If the policyholder’s family has a medical history of illnesses such as heart diseases, cancer or any other, that puts you at a risk

and it increases the individual’s rate of premium.

- **Body Mass Index (BMI):** People with high BMI have a significantly higher rate of premium than people with normal BMI. The reason again being this can lead to various ailments such as heart problems, joint problems, diabetes, to name a few. People with higher BMI may even need specialized treatment, for normal procedures like pregnancy. Thereby making even simple process a little tedious and affects the premium rates.
- **Injurious substances:** Most insurance companies increase their rates of premium for their insurance plans and at times even refuse to insure people who have the habit of smoking, chewing tobacco or snuff. Since they are most prone to getting life threatening diseases like cancer. Thereby affecting the rates of premium.
- **Gender:** Many policies have a difference in premium rates for men and women, the 3 reasons for this experts say are - Women are more likely to visit doctors, take prescriptions, and be subject to chronic diseases.
- **Age:** Most young individuals have premiums at much lower rates since they have fewer identified and unidentified diseases than older individuals. Young policyholders are less likely to have health problems and are more likely not to visit a doctor.
- **Choice of profession:** Policyholders working in environments with hazardous substances, radiation, chemicals, and jobs with high risk of injuries like constructions have to end up paying higher premiums as per insurance companies since they’re prone to risk of cardiovascular diseases.
- **Marital status:** It’s still unclear if married people live longer and healthier lives, but the insurance premiums generally lower in rates. The men generally reap better benefits with this status change.
- **No insurance yet:** If you’re not previously insured, the insurance companies generally charge a higher rate of premium. The insurers believe that previously uninsured individuals would make frequent trips to doctors and hospitals to start reaping benefits of the health insurance policy.

Software/Hardware requirements:

Software Requirements:

- Operating System : Windows 7 and higher versions, macOS and Linux
- Programming Language : Python
- Libraries : Numpy, Pandas, Matplotlib

Hardware Requirements

- RAM : 4 GB and more
- Processor : Intel® Core™ i3 processor and above
- Hard Disk : 2 GB and more

CHAPTER 5

Design and Implementation

1. Gathering Raw Data

Raw data set of hundreds of people is gathered from the website kaggle. Using their free API, the stock prices could be downloaded in JSON and CSV format. So for the data set of this project the the data set was downloaded from kaggle in CSV format. These data sets includes but are not limited to :

- **Age :** Primary beneficiary's age
- **Sex:** Gender of the primary beneficiary, male or female.
- **BMI**: Body mass index provides an understanding of body, weights that are high or low relative to the height, using the ratio of height to weight, ideally it needs to be in between the range of 18.5 to 24.9.
- **Children**: Number of dependents/ Number of children covered by the health insurance.
- **Smoker**: If the person is a smoker or not.
- **Charges**: Medical costs of an individual billed by Health Insurance.

2. Data Cleaning

Most of the time, there will be discrepancies in the captured data such as incorrect data formats, missing data, errors while capturing the data. This is important for a data science project because the results' accuracy depends a lot on the data used. The presence of missing, incomplete, or corrupted data leads to wrong results while performing any functions such as count, average, mean etc. These inconsistencies must be removed before doing any analysis on data.

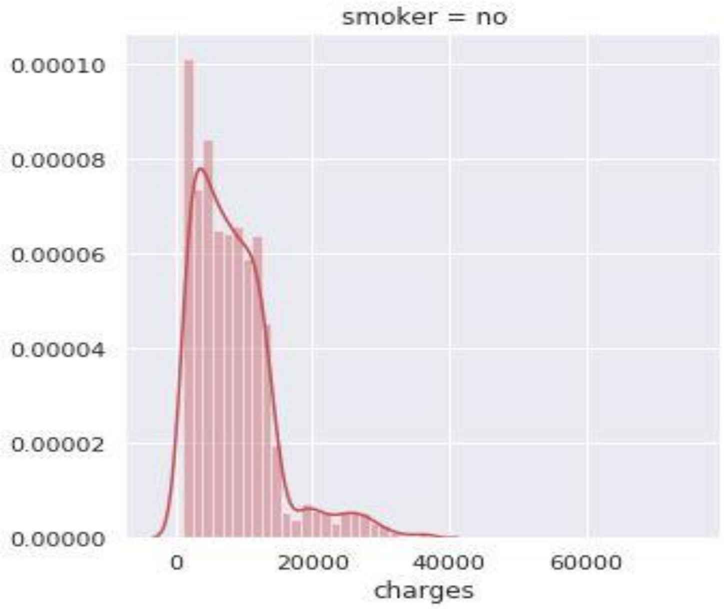
3. Features Selection

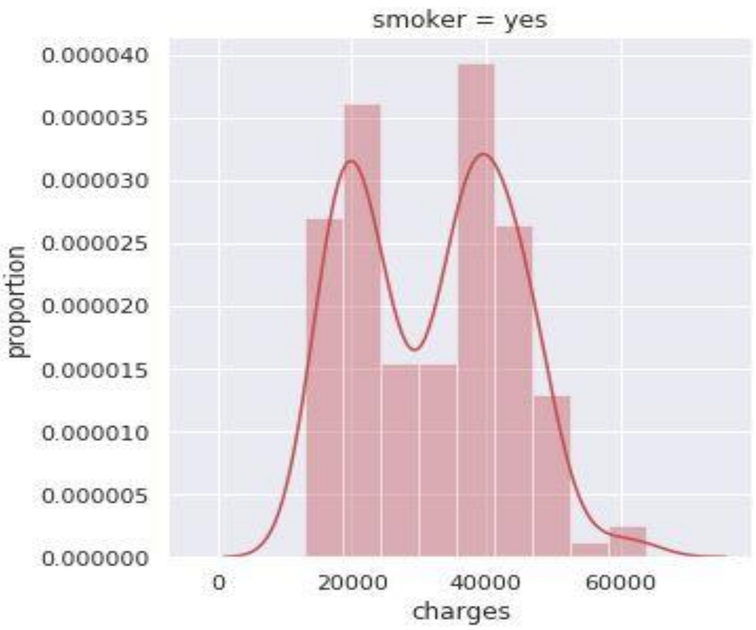
In this step, the features which contains the most information and contribute the least noise to the data are selected and the rest of the features are removed from the data.

In this system only the close value of the data is kept and the rest of the features are dropped from the data set. For instance:

- **Body Mass Index (BMI)**: People having high BMI have a significantly higher rate of premium than people having normal BMI. Since high BMI leads to various ailments such as heart problems, joint problems, diabetes, to name a few.
- **Marital status**: It’s still doubtful if married people live longer and healthier lives, so it is considered as an ineffective attribute for prediction.

This data shows that smokers pay more for healthcare, making smoking an important attribute for prediction.





4. Prediction

We have gathered the data, cleaned it, selected the important features all that is left is using the three prediction models to predict the charges and evaluate the accuracy using the actual

value from the data set. Accuracy defines the degree of correctness of the predicted value of the insurance amount. The model predicted the accuracy of model by using different algorithms, different features and different train test split size. The size of the data used for training of data has a huge impact on the accuracy of data. The larger the train size, the better is the accuracy. The model predicts the premium amount using multiple algorithms and shows the effect if each attribute on the predicted value.

CHAPTER 6

Result

Graphs plotted shows the predicted and the actual value with respect to the particular independent variable.

Results were predicted and a rough idea about required amount was known.

A) USING LINEAR REGRESSION:

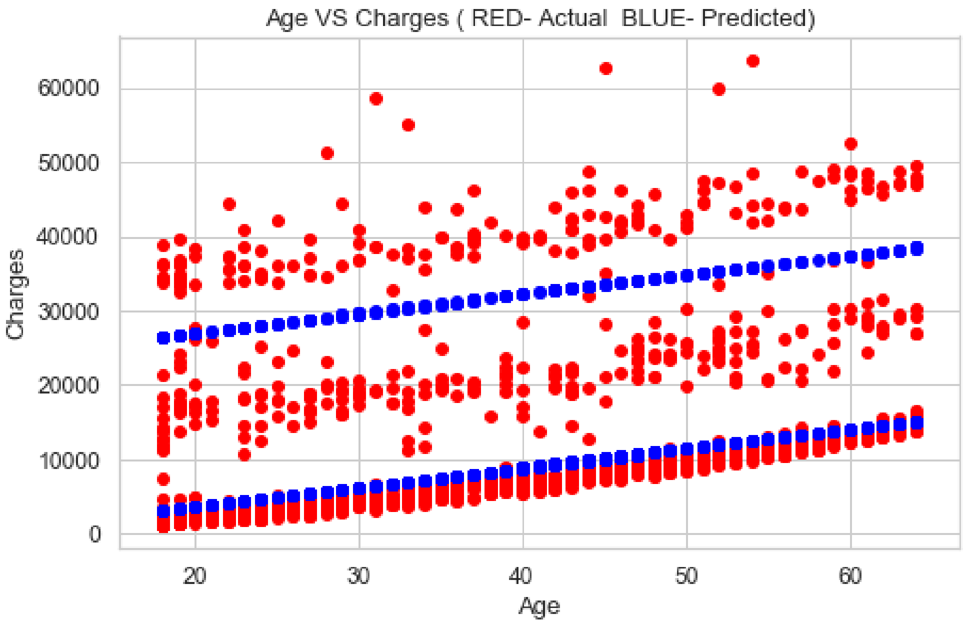


Fig 1 (Age vs Charges)

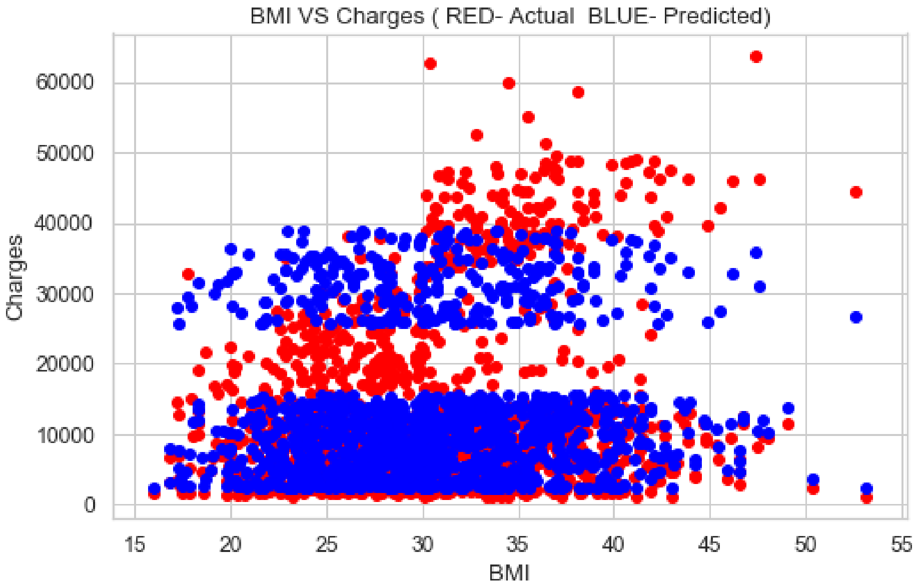


Fig 2 (BMI vs Charges)

B) USING DECISION TREE REGRESSION:

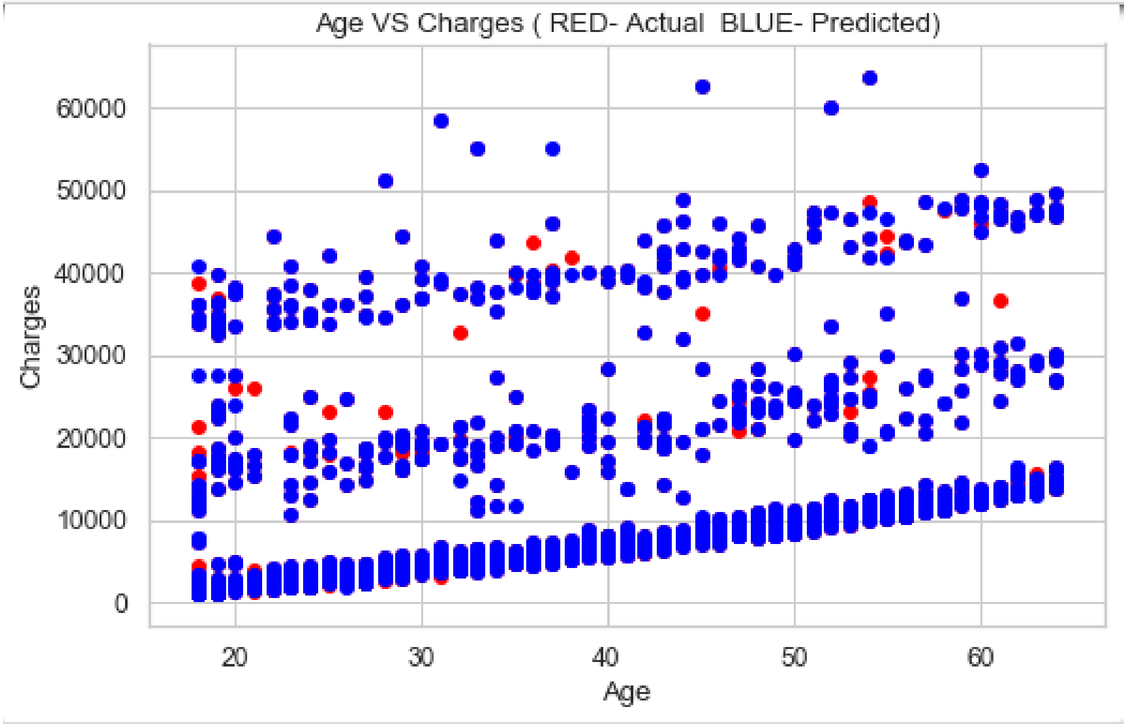


Fig 3 (Age vs Charges)

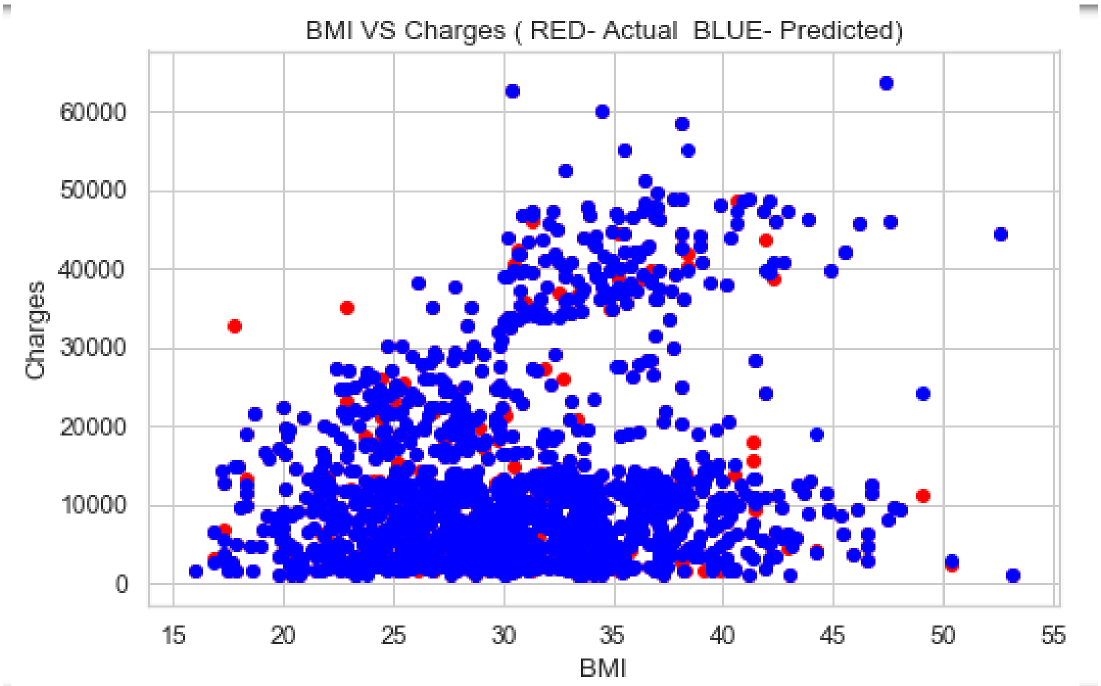


Fig 4 (BMI vs Charges)

C) **USING GRADIENT BOOSTING REGRESSION:**

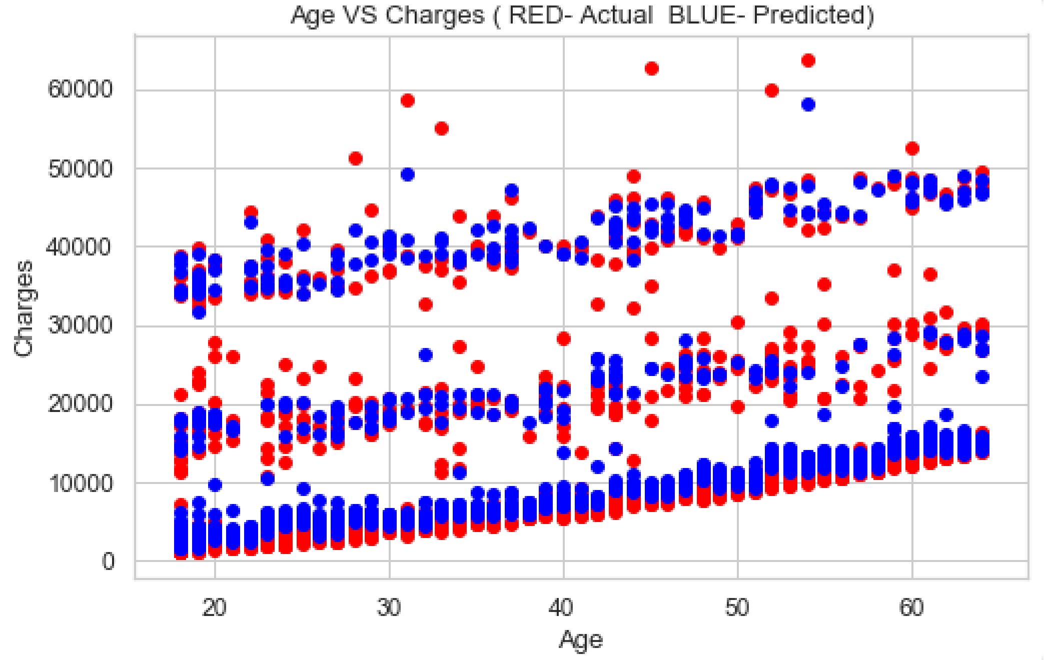


Fig 5 (Age vs Charges)

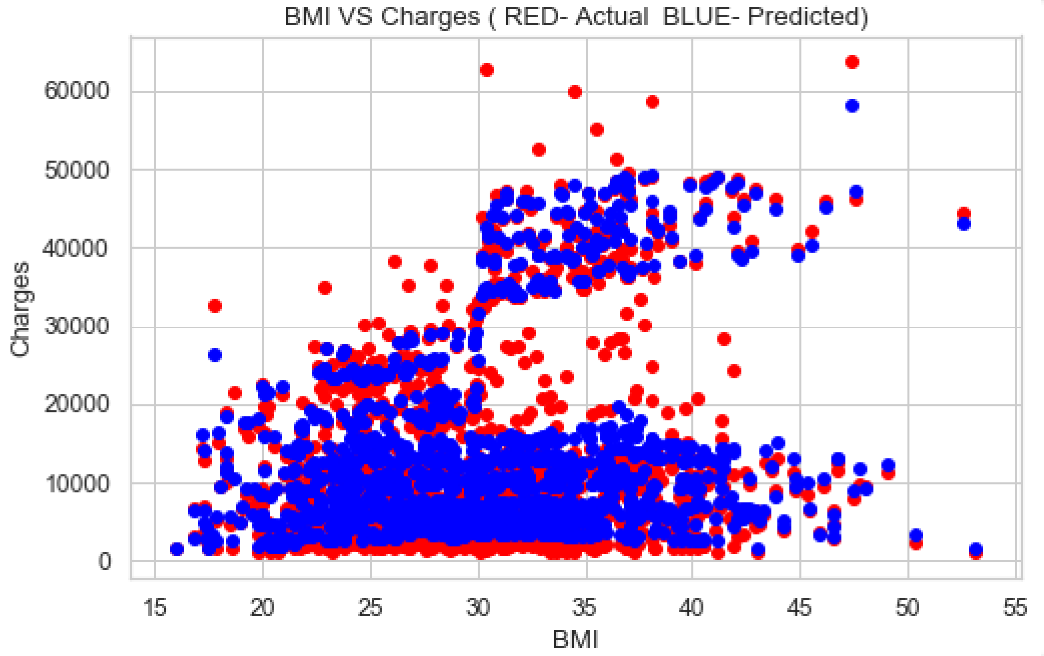


Fig 6 (BMI vs Charges)

The table below depicts the accuracy of the predicted value to the actual value using each algorithm taking different attributes as input for the prediction. The table will help us in understanding the overall effect of each attribute in calculating the premium amount.

The accuracy values (in percentage) taking different attributes as input are:

	Linear Regression	Decision Tree Regression	Gradient Boosting Regressor
Age	8-13	2-13	2-20
Gender	0	0	0
Smoker	50-64	57-69	57-70
BMI	0-4	0-1	0
Age + Gender	0-15	0-1	0-15
Age + Smoker	9-12	2-4	6-17
Age + BMI	0-9	0	0-11
Gender + Smoker	59-65	56-69	59-67
Gender + BMI	2-10	0	0-3
Smoker + BMI	61-80	58-74	74-81
Age + Gender + Smoker	67-75	57-64	67-75
Age + Gender + BMI	2-17	0	0-14
Gender + Smoke + BMI	59-70	54-77	70-86
Age + Smoke + BMI	67-82	70-79	80-93
Age + Gender + Smoke + BMI	69-82	67-80	84-94

CHAPTER 7

Conclusion

Graphs plotted shows the predicted and the actual value with respect to the particular independent variable. Similarly various other variables like age, smoking etc were predicted and evaluated. The predicted charges that were calculated using the regression techniques were compared with actual charges taken from data set to compare the accuracy of these techniques.

Various factors were used and their effect on predicted amount was examined. It was observed that a person's age and smoking status affects the prediction most in every algorithm applied. Attributes which had no effect on the prediction were removed from the features.

Gradient Boosting Regression technique which is built upon decision tree comes out to be the best performing technique compared to others when taking all effective attributes into account.

CHAPTER 8

References

[2] https://www.moneycrashers.com/factors-health-insurance-premium-costs/

[3] https://www.bankbazaar.com/health-insurance/top-10-factors-affecting-health-insurance-premium.html

[4] https://en.wikipedia.org/wiki/Healthcare_in_India

[5] https://www.kaggle.com/mirichoi0218/insurance

[6] https://economictimes.indiatimes.com/wealth/insure/what-you-need-to-know-before-buying-health-insurance/articleshow/47983447.cms?from=mdr

[7] https://www.policygenius.com/health-insurance/learn/health-insurance-basics-and-guide/

[8] https://www.investopedia.com/terms/r/regression.asp

[8] https://en.wikipedia.org/wiki/Regression_analysis