Insurance Cost Prediction

Prepared By:
Shubham Chavan
Shreya Naik
Jayesh Raikwar
Vasundhara Chilakala

(Group No: 29)

Under Guidance of:

Pranav Jaipurkar
M.E.
Knowledge Solutions India

College Name:

N.B. Navale Sinhgad College Of Engineering, Solapur.

Bheemanna Khandre Institute of technology, Bhalki

Global Nature Care Sangathan Group Of Institutions

Rajeev Gandhi Memorial college of Engineering and Technology

Table of Contents

***Abstract*:**

In this, we analyze the personal health data to predict insurance amount for individuals. Four regression models naming Multiple Linear Regression (MLR), Random Forest Regression (RFR), Multiple Linear Regression with Principal Component Analysis (MLR with PCA) and Random Forest Regression with Principal Component Analysis (RFR with PCA) have been used to compare and contrast the performance of these algorithms. Dataset was used for training the models and that training helped to come up with some predictions. Then the predicted amount was compared with the actual data to test and verify the model. Later the accuracies of these models were compared.

Introduction

The goal of this project is to allow 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. This can help a person in focusing more on the health aspect of an insurance rather than the futile part.

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 the 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.

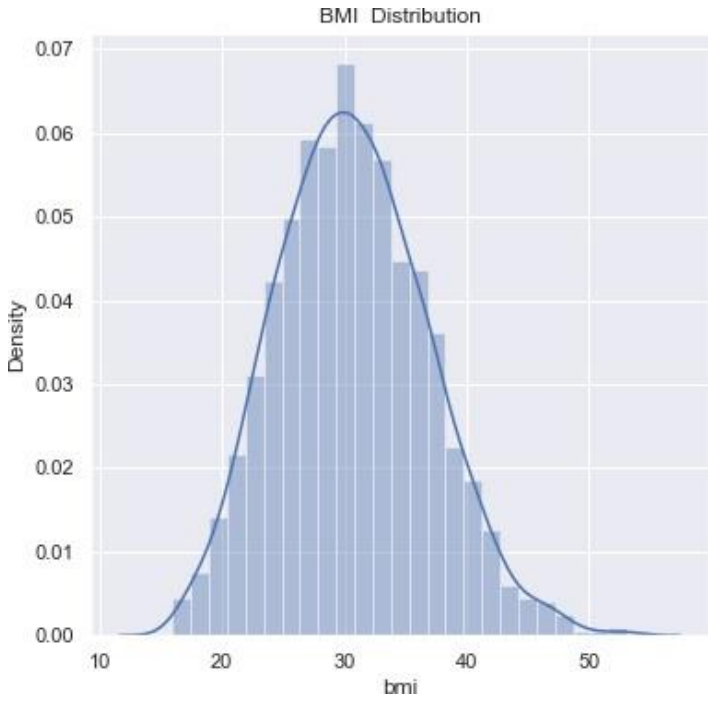
Our project does not give the exact amount required for any health insurance company but gives enough idea about the amount associated with an individual for his/her own health insurance. Prediction is premature and does not comply with any particular company so it must not be only criteria in selection of a health insurance. Early health insurance amount prediction can help in better contemplation of the amount needed. Where a person can ensure that the amount, he/she is going to opt is justified. Also, it can provide an idea about gaining extra benefits from the health insurance.

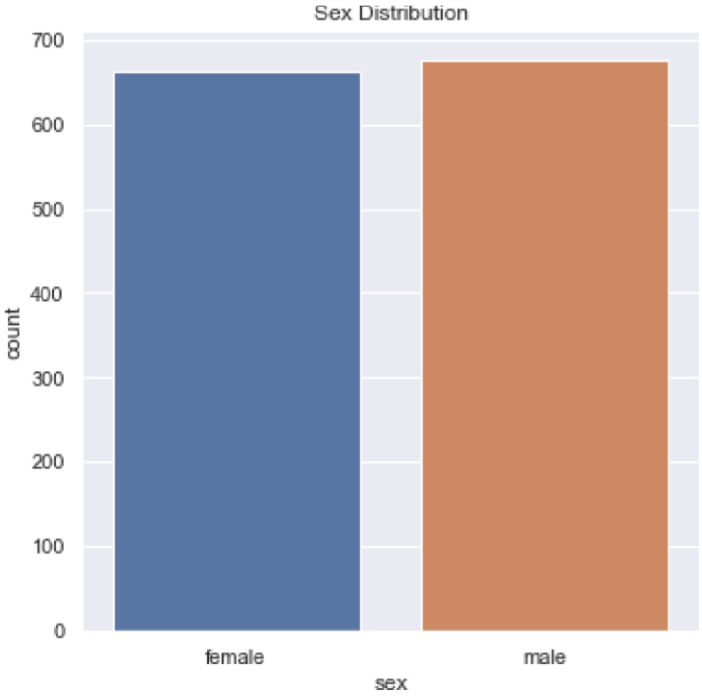
Dataset Used

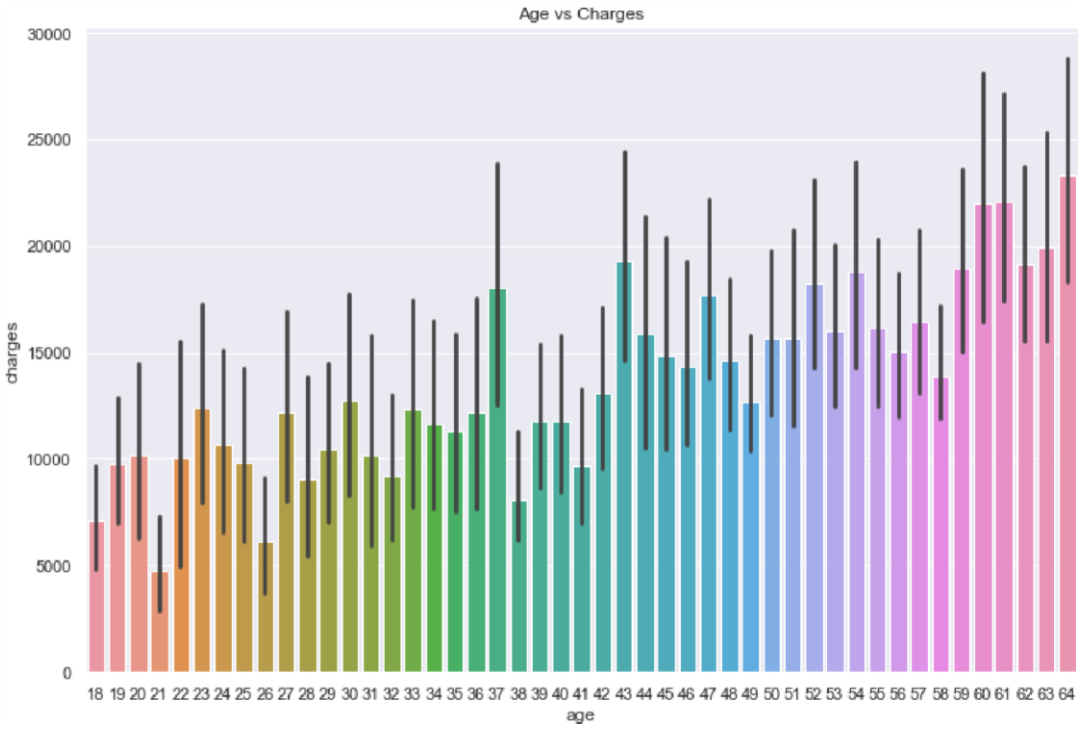
To create the claim cost model predictor, we obtained the data set through the csv files provided by teachers. The data set includes seven attributes, the data set is separated into two-part the first part called training data, and the second called test data; training data makes up about 80 percent of the total data used, and the rest for test data The training data set is applied to build a model as a predictor of medical insurance cost year and the test set will use to evaluate the regression model. The following table shows the Description of the Dataset.

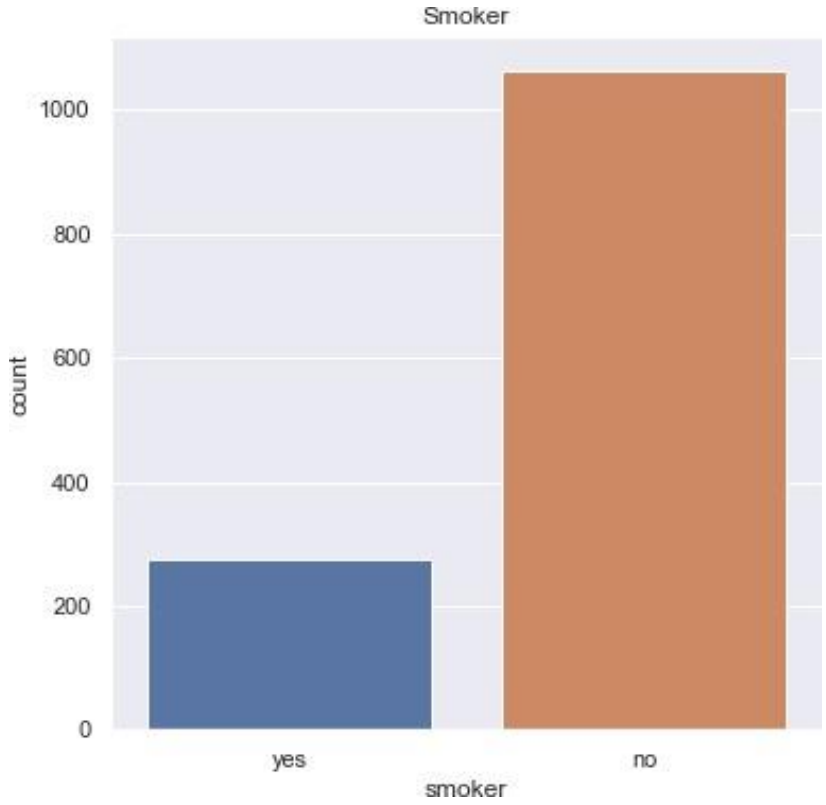
Name	Description
Age	Age of Client
BMI	Body mass index
Number of Kids	Number of children of client
Gender	Male/Female
Smoker	Whether the client is a smoker or not
Region	Where the client lives Southwest, Southeast, Northwest or Northeast
Charges (target variable)	Medical Cost of the client to pay

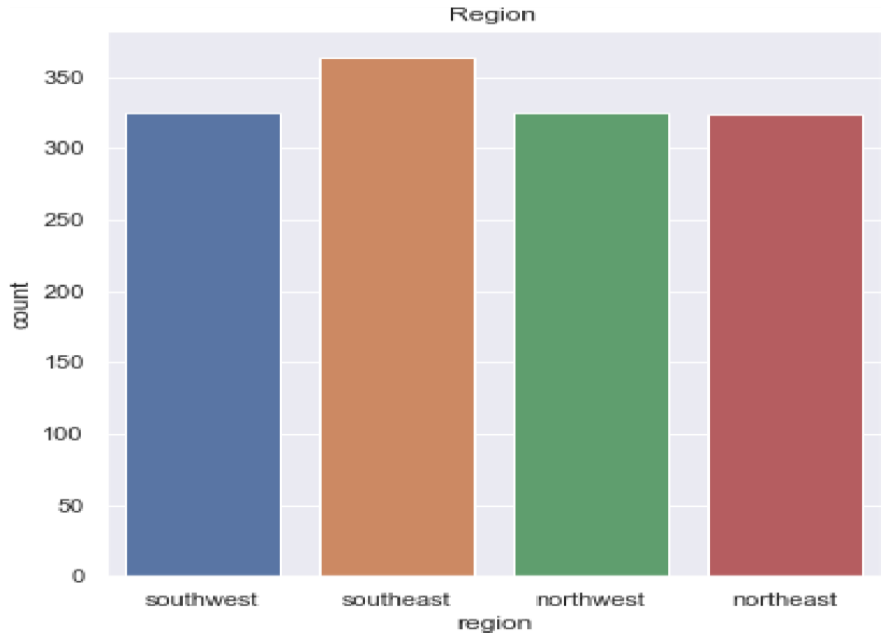
Various Graphs of Data Comparison

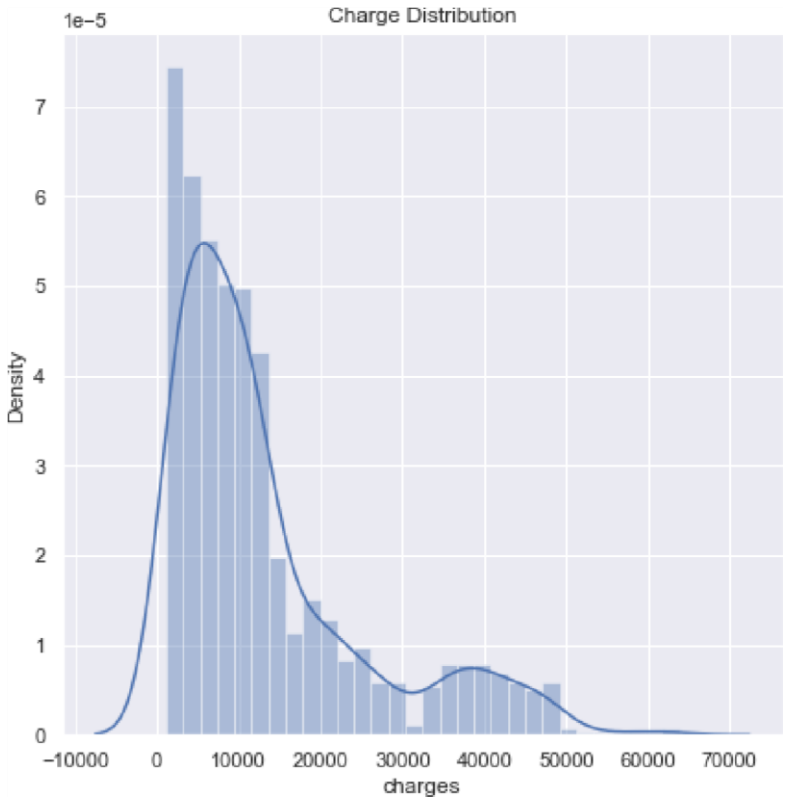


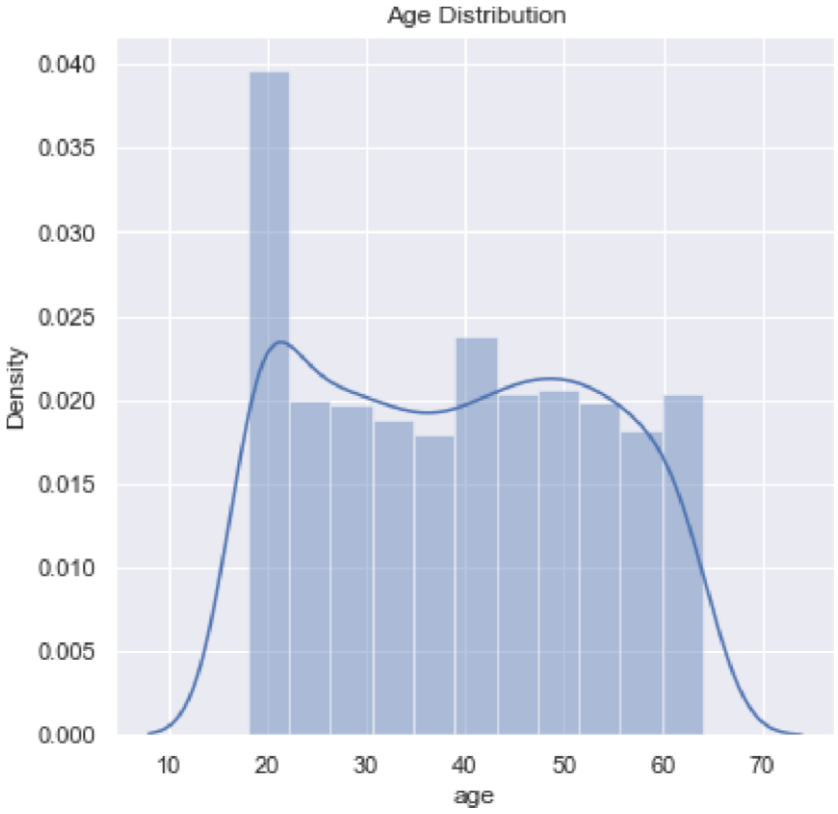


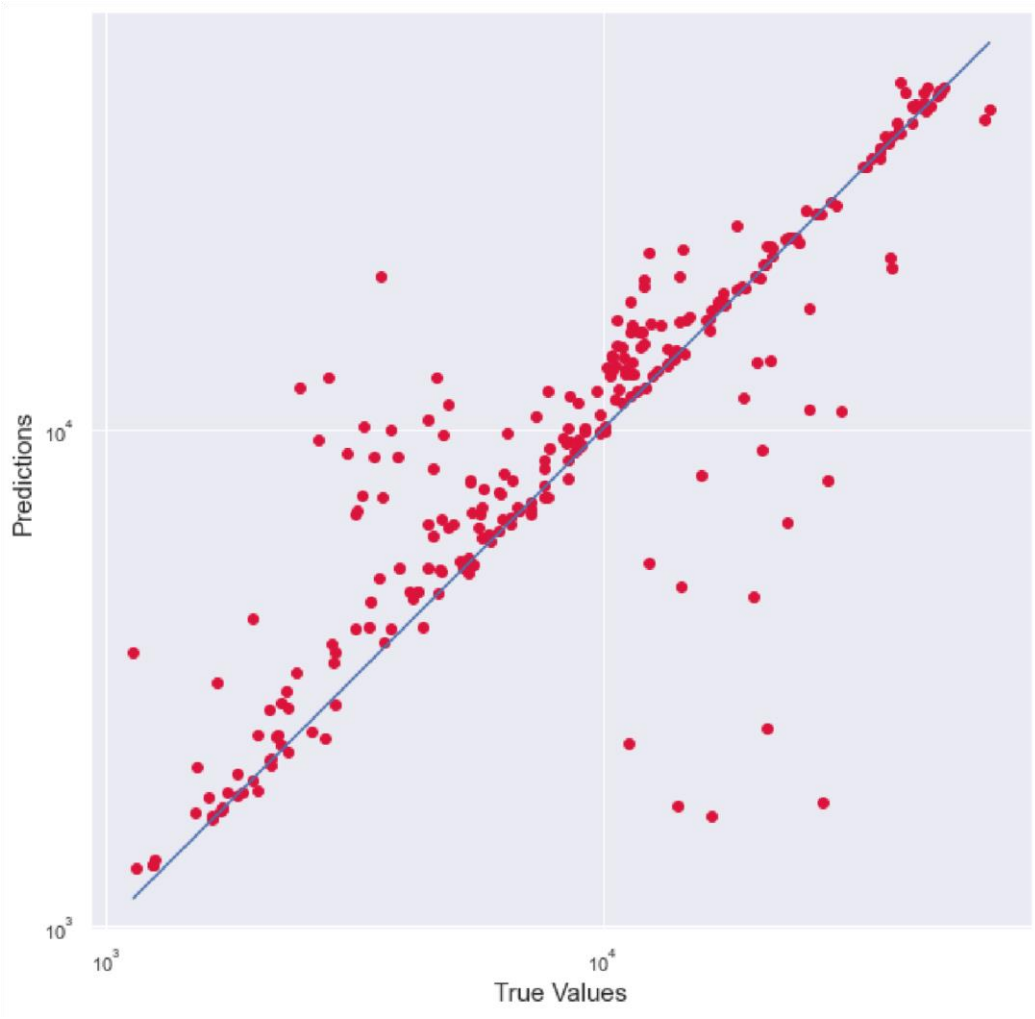

Age vs Charges
charges
30000
25000
20000
15000
10000
5000
0
18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64
age


Smoker
count
1000
800
600
400
200
0
yes
no
smoker


Region
350
300
250
200
150
100
50
0
count
southwest
southeast
northwest
northeast
region


Charge Distribution
1e−5
7
6
5
4
3
2
1
0
Density
−10000
0
10000
20000
30000
40000
50000
60000
70000
charges


Age Distribution
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
Density
10
20
30
40
50
60
70
age


Predictions
10^4
10^3
10^3
10^4
True Values

Software Libraries Used in Python:

- Pandas
- Numpy
- Matplotlib
- Seaborn
- Sklearn

Algorithm:

- **Multiple Linear Regression:**
  - Multiple linear regression can be defined as extended simple linear regression. It comes under usage when we want to predict a single output depending upon multiple input or we can say that the predicted value of a variable is based upon the value of two or more different variables. The predicted variable or the variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable) and the variables being used in predict of the value of the dependent variable are called the independent variables (or sometimes, the predict, explanatory or regressor variables).

- **Random Forest Regression:**
  - Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships. Disadvantages, however, include the following: there is no interpretability, overfitting may easily occur, we must choose the number of trees to include in the model.

- **Principal Component Analysis:**
  - Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. It is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables. PCA is the most widely used tool in exploratory data analysis and in machine learning for predictive models. Moreover, PCA is an unsupervised statistical technique used to examine the interrelations among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit.

Designing and Implementation

1. Data Preparation & Cleaning

a. The data included various attributes such as age, gender, body mass index, smoker and the charges attribute which will work as the label for the project.

b. The data was in structured format and was stores in a csv file format. The data was imported using pandas library. 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.

2. Training

a. Once training data is in a suitable form to feed to the model, the training and testing phase of the model can proceed.

3. Prediction

a. The model was used to predict the insurance amount which would be spent on their health. The model used the relation between the features and the label to predict the amount.

b. 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.

CONCLUSION

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

Amount prediction focuses on persons own health rather than other companies insurance terms and conditions. The models can be applied to the data collected in coming years to predict the amount. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance amount.