Dhruvinsinh Rathod

Gandhinagar, Gujarat, India

Student at Pandit Deendayal petroleum university

Medical-Cost-Prediction-for-health-Insurance

ABSTRACT

Abstract In this thesis, I have analyse the personal health data to predict insurance amount for individuals. Different models of Machine learning has been trained as mentioned in System architecture. 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 R2 score of these models were compared. Xgboost is best suited in this case.

INTRODUCTION

The goal of this project is to 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. 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.

This 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

The primary source of data for this project was from Kaggle. The dataset is comprised of 1340 records with 6 attributes as input and 1 as output feature.

The dataset Content is as follows: Columns

age: age of primary beneficiary

sex: insurance contractor gender, female, male

bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

children: Number of children covered by health insurance / Number of dependents

smoker: Smoking

region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

charges: Individual medical costs billed by health insurance

	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520

MACHINE LEARNING

Machine learning can be defined as the process of teaching a computer system which allows it to make accurate predictions after the data is fed.

However, training has to be done first with the data associated. By filtering and various machine learning models accuracy can be improved.

Types of Machine Learning

1. Supervised Learning

   Supervised learning algorithms create a mathematical model according to a set of data that contains both the inputs and the desired outputs. Usually a random part of data is selected from the complete dataset known as training data, or in other words a set of training examples. Training data has one or more inputs and a desired output, called as a supervisory signal. What's happening in the mathematical model is each training dataset is represented by an array or vector, known as a feature vector. A matrix is used for the representation of training data. Supervised learning algorithms learn from a model containing function that can be used to predict the output from the new inputs through iterative optimization of an objective function. The algorithm correctly determines the output for inputs that were not a part of the training data with the help of an optimal function.

2. Unsupervised Learning

   In this learning, algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or clustering of data points. Test data that has not been labelled, classified or categorized helps the algorithm to learn from it. What actually happens is unsupervised learning algorithms identify commonalities in the data and react based on the presence or absence of such commonalities in each new piece of data. The main application of unsupervised learning is density estimation in statistics. Though unsupervised learning, encompasses other domains involving summarizing and explaining data features also.

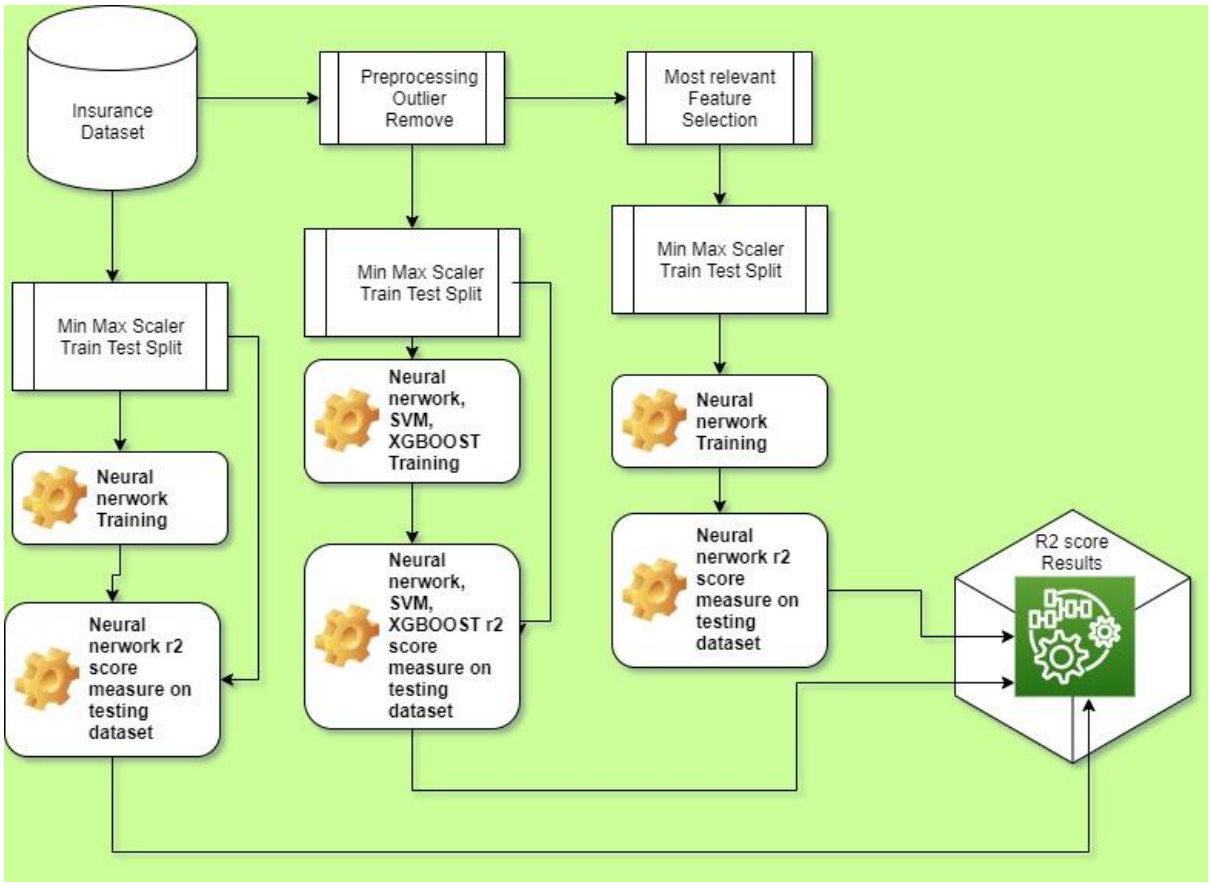
3. Reinforcement Learning

   Reinforcement learning is class of machine learning which is concerned with how software agents ought to make actions in an environment. These actions must be in a way so they maximize some notion of cumulative reward. Reinforcement learning is getting very common in nowadays, therefore this field is studied in many other disciplines, such as game theory, control theory, operations

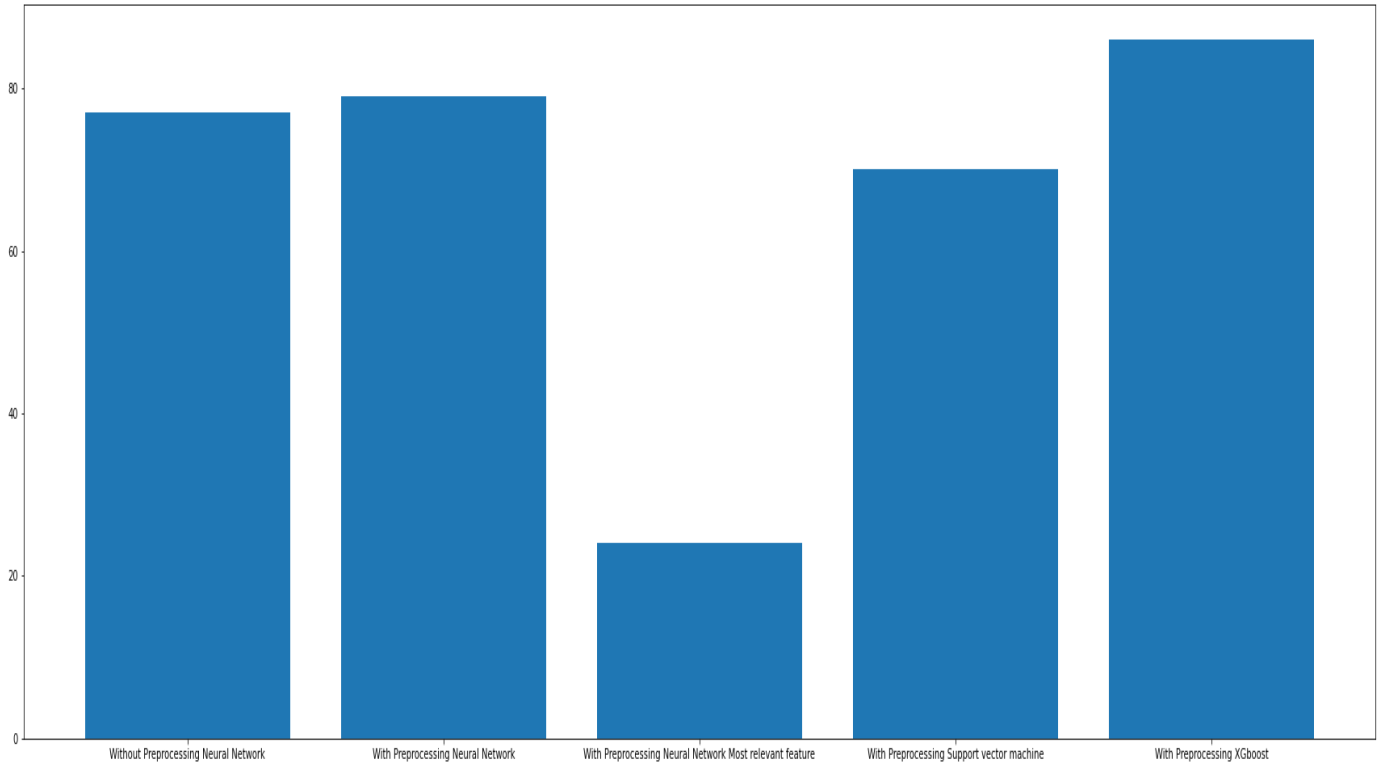
research, information theory, simulated-based optimization, multi-agent systems, swarm intelligence, statistics and genetic algorithms.

Here This project has treated as supervised Learning.

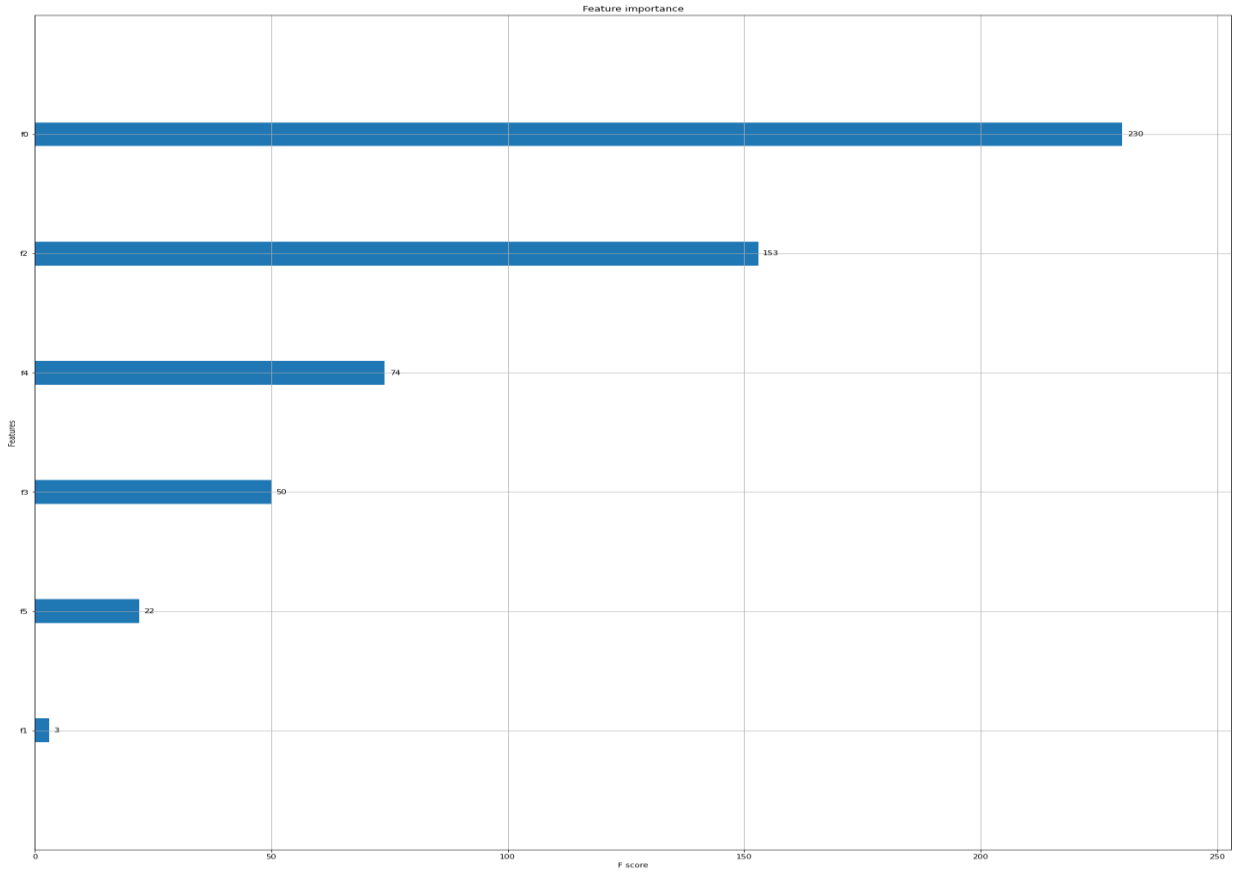
System Architecture



RESULTS



Feature Importance By XGBOOST



CONCLUSION

As part of this project total 5 models were trained.

Here is the r2 score for each

1)Without pre-processing neural network-0.77

2)With pre-processing neural network-0.79

3) With pre-processing neural network with most relevant feature-0.24

4) With pre-processing SVM-0.70

5) With pre-processing XGBoost-0.86

So here XGboost is providing Best and accurate prediction among all models.