A Medical Price Prediction System using Boosting Algorithms through Machine Learning Techniques

Renu Manoj
M.Tech. Scholar, Department of CSE
SISTec-R, Bhopal (M.P), India

Prof. Ajeet Shrivastava
Professor, Department of CSE
SISTec-R, Bhopal (M.P)

prof. Rohit Bansal
Asst. Professor, Department of CSE
SISTec-R, Bhopal (M.P)

---***---

ABSTRACT

The health care insurance cost plays a vital role in developing medical facilities. To provide better medical facilities, it is very essential to forecast the cost of medical insurance which is one of the possibilities to enhance medical facilities. The paper deals with predicting the cost of the health insurance which has to be paid by the patient. Here various data mining regression algorithms such as decision tree, random forest, polynomial regression and linear regression are implemented to achieve the best prediction analysis. A comparison has been done between the actual and predicted expenses of the prediction premium and eventually, a graph has been plotted on this basis which will enlighten us to choose the best-suited regression algorithm for the insurance policy prediction. After the execution of these regression algorithms for prediction, correctness has been measured by the Coefficient of determination (r2_score), Root Mean Squared Error (RMSE) and Mean Squared Error (MSE) of each algorithm to check for the best-suited algorithm. Random Forest Regression was the best algorithm with an r2_score of 0.862533 which can be used in its best possible way for the correct prediction of the health insurance cost.

Keywords: Keywords: Natural Language Processing, Tokens, Features, Training & Testing Data, Model or classifier, Naïve Bayes, SVM, Bernoulli.

I. INTRODUCTION

Health insurance market is a crucial market because one-third of GDP [1] is spent on health insurance in the United States, 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 premiums and verify results using regression models. 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. According to the Office of Health and Human Services (HHS) [2], the total health service budget for the fiscal year 2015 is around 1100 billion dollars.

There are several health care systems around the world. For example, single payer system followed by Canada where premiums are paid by taxes, government health care system followed by the United Kingdom where healthcare is the responsibility of central government.

There are no existing tools to the best of our knowledge that can predict future premiums based on historic data. Therefore, there is a need to conduct research to find the premiums across the United States.

1.1 Health Service Area

Here Researches explained the major domain or service area where Health Service will be utilized. They found many areas out of them important area is given below:

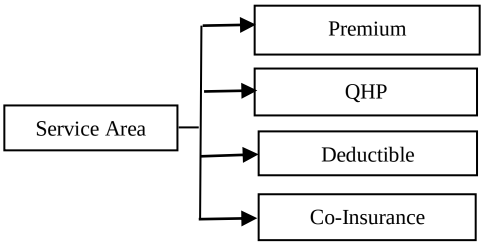

Figure 1: Service Area

Premium: Generally, every member from family some amount has to paid for the health insurance every month. This amount are knowns as premium, other costs are also paid for health care, including a deductible, co-payments, and coinsurance.

Qualified Health Plan (QHP): QHP is defined as a government certified insurance plan that provides essential benefits such as emergency services, maternity care etc. In India we can say that is similar to Ayushman Bharat Yojana.

Deductible: The minimum amount has to pay for covered health care services before any insurance plan starts to pay.

1.2 Data Mining Tools

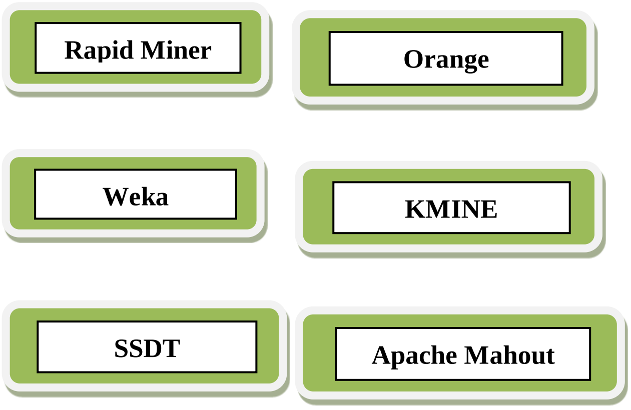

Figure 2: Data Mining Tools

In Figure 2 researchers explained the different data mining tools available for processing the given data.

Generally, Weka is Java based tools which is very efficient for Data processing and finding the behaviour and patterns from the taken data. Apache Mahout is very efficient for Big Data.

1.3 Objective of Work

Sentiment The Motto of This researcher is that In Today's hectic life everybody has to take concert to their Health status because this will define that how much amount will be required in near future. So, any Insurance company comes to some conclusion after behaviour of Health of anyone. We all know that in previous some years through out the globe from poor people to rich people everybody is taking health plan because they know if they come under illness their economic condition may suffer them and their family. So, for their safety they also try to book any good plan that is less in premium cast and will give better cover in near future. This is very fast and growing industries.

II. RELATED WORK

Here In this section, research efforts from the finding the information and machine learning techniques are explained. Many papers have explained the issue of claim prediction. Authors suggested, "Predicting motor insurance claims using telematics data" in 2019. This research Explained for better performance of logistic regression and XGBoost techniques to forecast the presence of accident claims by a small number and results showed. LR is very Effective in some cases. this system takes pictures of the damaged car as inputs and produces relevant details, such as costs of repair to decide on the amount of insurance claim and locations of damage. Thus, the predicted car insurance claim was not taken into account in the present analysis but was focussed on calculating repair costs [4].

Health care sector is one of the biggest sectors in terms of the global economy. According to the World Bank, in 21 century, health care expenditures accounted for 10% approx. of the world's total gross domestic product (GDP). Additionally, per capita health expenditures have been increasing day by day from last 10-20 years. In the many countries, the Centres for Medicare & Medicaid Services (CMS) reported that health care accounted for 17.5% of the national GDP [5].

III. PROBLEM IDENTIFICATION

Many research's is doing their work in this area. Authors have learned several things from this study. We come to conclusion that Insurance Sector is very growing in terms of number of customers growing as well as every year's number of new Insurance company comes into existence. These industries have huge contribution into GDP. So here huge concern and analysis needed. Besides looking at various approaches and models, we also focus on important aspects in the Machine Learning. This will enable you to use these methodologies in the future.

IV. ALGORITHMS

Step 01: Store Data from Kaggle Repository

Step 02: Import Prior Libraries

Step03: Now Import our Required Dataset

Step04: Apply Feature Extraction

a) Bivariate Analysis

b) Multivariate Analysis

Step 05: Visualize Data for better understanding

a) Descriptive Features

b) Distribution of Different Features

Step06: Applying Machine Learning Algorithms

Step07: Apply Different Model

a) Linear Regression

b) Random Forest

c) Gradient Boosting

Step08: Repeat Step07 for many times with different Algorithms

Step09: Finally Compare Results with performance parameters like RMSE Score & R2 Score.

Step10: Stop

IV. Flow Diagram

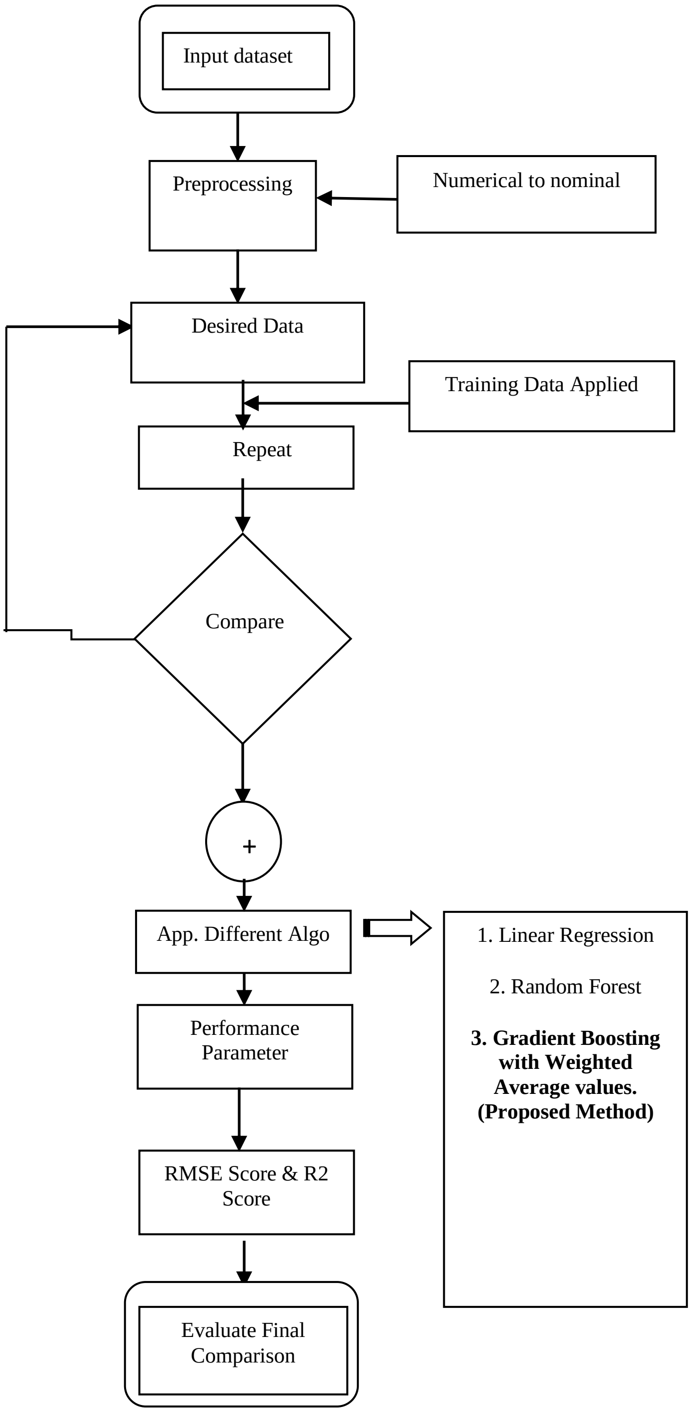

Figure 3: Flow Diagram

In In figure 3 At first step, we need to fetched Data from any external source or we can collect Data from Local Market but for better Analysis we are Fetching our Data from Kaggle. That is very reliable Data Source through Word Wide. In Next Step we need to Fetched Different Libraries for processing our Data. At very next Step that is Third Step we need to process our Data for next step Here we have many processing Mechanism. We are using Numerical to Nominal Data Conversion or also using Uni-Variant and Multi-Variant Data Processing. At Next Step i.e., Fourth Step we need to repeat it for different Data split. Then after we will reach at step 5 where we will apply Different Machine Learning Algorithms and Finally, we will apply our own Proposed Methods i.e., Gradient Boosting with Weighted Average values. Here we will adding average values of previous implemented mechanism. At Final Step i.e., Sith step we will have to find Performance Measures i.e., RMSE Score & R2 Score will give clear views of proposed method and Existing one. At Final Step we will compare these given Results. We can say that Our Proposed Methods gives better Result.

V. PROCESS OF IMPLEMENTATION

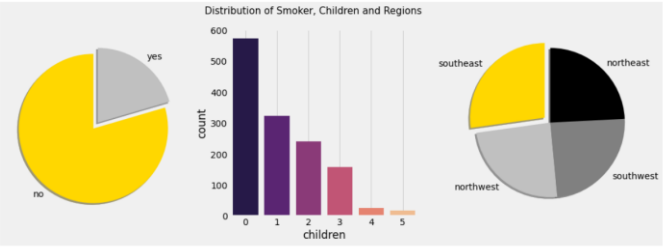

Figure 4: Distribution of smoker, children and region (Pia & Bar Graph)

In the Figure 4 reserchers Explained How Distribution gives clear Results :

1) We can say that we have an equal number of people of all ages

2.) Where maximum people have bmi around 30

3.)Finally Expenses are seem to be right skewed.(Learn Skewness from probablity and statistical)

Note: Because Expenses are skewed in nature so that can be either transform this column using log transformation or square root transformation. And this can be converted into normal Distribution.

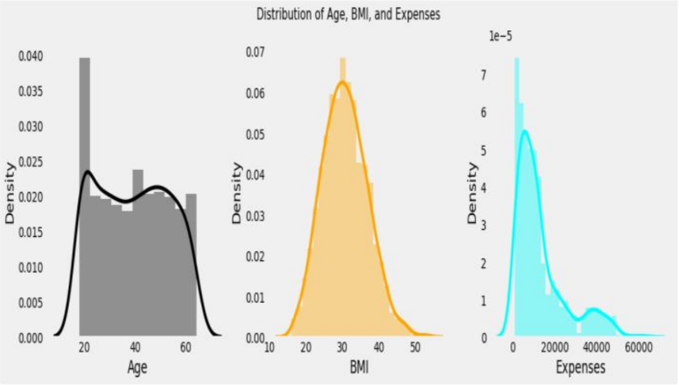

Figure 5: Distribution of smoker, children and region (Area Graph)

In figure 5 its clear picture that how Distribution of Smokers, Children and region gets effected.

Output Results

In the table 1 we are trying to show the Accuracy Results of Different Algorithms which we implemented.

Table: 1

Algorithms	Train Size (in %)	Test Size (in %)	Accuracy
Linear Regression	70	30	0.79
Random Forest	70	30	0.87
Gradient Boosting with Weighted Average values (Proposed Methods)	70	30	0.89

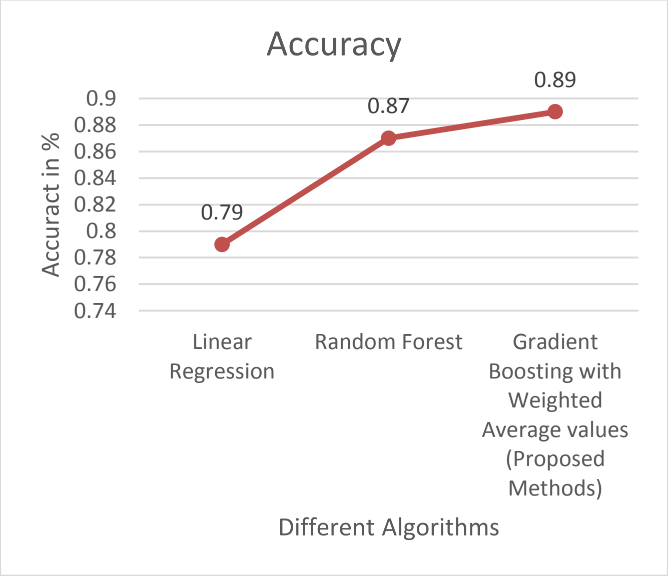

Figure 6: Different Algorithms Comparison Line Graph

In figure 6 By the analysis of above graph we can say that we find that Liner Regression ,Random Forest, Gradient Boosting with Weighted Average values (Proposed Method)gives Different Results respectively. Here We are going to proposed some tuning mechanism so that Results get improved. Here we are Implemented an weighted averaging model We tooked 50% weight to gradient boosting ,30% weight to random forest and 20% weight to linear regression. Due to to the above Data selection we are getting better Results.

VI. CONCLUSION

Here researches conclude that when we Implemented Number of Machine Learning Algorithms for finding best results in terms of performance. We Finds the major Features of given Data set are Smoking Behaviour, Age Group, Region Where They Lived and many more. Most Important Factors to Predict the Medical Expenses of a Patient is Smoking Behaviour, Age, Gender, Number of Children, the Region also have a good impact on determining the Medical Expenses. Finally, we Implemented Different Machine Learning Algorithms Like Linear Regression, Random Forest & Finally our Proposed Method gives results respectively i.e., 79%, 87 % and finally 89%. When we looked into our proposed methods then we can claim that with existing system our proposed methods give better results in terms of Accuracy. Apart from Accuracy we explained various Graphs from which we come for better understanding about our Implemented Models.

VII. FUTURE SCOPE

In this investigation the future works focus on applying some other techniques to improving the performances of these methods for up to maximum extent. Another concept that can be implemented Deep learning in place of machine learning technology. The reason behind this is best and efficient techniques using nowadays. Deep learning is also introduced nowadays which is becoming more popular for classification purpose. So, we can also implement deep learning in future work also.

REFERENCES

[1] Sonali Vyas, Rajeev Ranjan, Navdeep Singh, Arohan Mathur , "Review of Predictive Analysis Techniques for Analysis Diabetes Risk" , 978-1-5386-9346-9 ©2019 IEEE.

[2] Gaurav Tripath, Rakesh Kumar , "Early Prediction of Diabetes Mellitus Using

Machine Learning " ,2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Noida, India. June 4-5, 2020.

[3] Messan Komi, J un Li ,"Application of Data Mining Methods in Diabetes Prediction",2017 2nd International Conference on Image, Vision and Computing.

[4] Bakshi Rohit Prasad,Sonali Agarwal ,"Modeling Risk Prediction of Diabetes – A Preventive Measure"

[5] J. N. Myhre, I. K. Launonen, S. Wei and F. Godtliebsen, "Controlling blood glucose levels in patients with type 1 diabetes using fitted qiterations and functional features," 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP), Aalborg, pp. 1-6, 2018.

[7] Zhang, Y., Lin, Z., Kang, Y., Ning, R., & Meng, Y. A FeedForward Neural Network Model For The Accurate Prediction Of Diabetes Mellitus.

[8] Kadhm, M. S., Ghindawi, I. W., &Mhawi, D. E. (2018). An Accurate Diabetes Prediction System Based on K-means Clustering and Proposed Classification Approach. International Journal of Applied Engineering Research, 13(6), 4038-4041.

[9] Sundaram, N. M. (2018). An Improved Elman Neural Network Classifier for classification of Medical Data for Diagnosis of Diabetes. International Journal of Engineering Science, 16317.

[10] Deeraj Shetty,Kishor Rit, Sohail Shaikh, Nikita Patil,"Diabetes Disease Prediction Using Data Mining",2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)

[11] Muhammad Azeem Sarwar,Nasir Kamal,Wajeeha Hamid,Munam Ali Shah,"Prediction of Diabetes Using Machine Learning Algorithms in Healthcare",Proceedings of the 24th International Conference on Automation & Computing, Newcastle University,Newcastle upon Tyne, UK, 6-7 September 2018

[12]Md. Faisal Faruque,Asaduzzaman,Iqbal H. Sarker,"Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus",2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, 2019

[13] C. C. a. A. Semanskee, "Analysis of UnitedHealth Group's Premiums and Participation in ACA Marketplaces," 2016.

[14] "How the Affordable Care Act Has Improved America's Ability to Buy Health Insurance on Their Own," The Commonwealth Fund, 2017.

[15] I. D. Henry G. Dove, and Arthur Robb, "A prediction model for targeting low-cost,highrisk members of managed care organizations," The American Journal of Managed Care, 2003.