



Advance analysis & Comparative study on Machine Learning methods for Predicting Health Insurance Premium Pricing

Pushpalata Verma^{1*}, Deep Kumar Sahu², N Dutta Swaroop³

^{1*} Assistant Professor: Bhilai Institute of Technology, Raipur

^{2,3} B.Tech Scholar: Bhilai Institute of Technology, Raipur

Abstract : A medical emergency can be unpredictable and its impact is more than enough to disrupt the biopsychosocial factors which include economic, physical and psychological conditions. In this case, Health insurance prevents a person from going into a situation of poverty by covering the major medical expenses like Treatment costs, Hospital charges, Medical equipment expenses, etc. This research deals with the prediction of health insurance premium pricing using various machine learning methods. The goal is to do a comparative study on which algorithm is the best for predicting insurance premiums, as well as making the most accurate model for the same. The data is taken from the Kaggle containing features such as age, gender, Body-Mass Index, number of children, charges, smokers and region. 10 Machine Learning algorithms are used from linear regression to ensemble methods to the neural network, to find the best possible algorithm for the above task. At the end of the study, the best method is found to be AdaBoost with the Root Mean Square Error 149.61095379400695.

Keywords - Health insurance, Premium pricing, AdaBoost, Linear regression, Neural network, Random Forest, Decision Tree, Support Vector Regressor.

I. INTRODUCTION

Maintaining good health is paramount to living a fulfilling life. Essentially, it's what sits at the top of humans priority list. Taking care of health means investing in present and future well-being. Prioritizing health ensures to have the best chance to live a fulfilling and vibrant life. Nowadays having a good health has become a new luxurious thing in life because the cost of healthcare is increasing gradually and it made it necessary to have a good medical health insurance. Health insurance is a lifesaver because it provides the financial stability, access to better treatment, mental stability. The cost of Health insurance premiums are varying from person to person based on the different attributes. The Insurance premiums are not fixed for everyone, they are calculated based upon the person's current health and most importantly person's welfare protection and their age. According to the study [6] it has revealed that the people of the age group 21-30 years are quite pleased with the medical insurance schemes including costs. However, people with above the age of 60 years are less satisfied with the premium cost they are paying. Again, satisfaction in customer support schemes offered by health insurance companies for the people in 31-40 years age bracket is relatively higher than the people above 60 years. One more study [7] has elaborated that the different chronic diseases will have an impact on the insurance premium pricing level. The continuous rise in the cost of health insurance premiums and the problem of difference in the amount of premium for different age group people has initiated the need of accurate prediction models for health insurance premiums. When company provides an health insurance to its customer in exchange for the payment of monthly or yearly premium then it covers the maximum cost of the insured person's medical expenses. Once a customer and a insurance company agreed for an contract , then that contract acts as a lifeguard and typically covers and protects a person throughout a year. The field of Machine Learning and its algorithms

& methods are considered to be the best for predicting accurate results in terms of premium pricing and calculating the cost of patient medical treatment expenses, so insurance companies are rapidly moving towards ML domain for the betterment of their policies and premium settings.

The contributions of our research study are as follows :

- (a). Predicting health insurance premium pricing using ensemble learning based machine learning models (Linear regression, Decision tree, Random forest, XG Boost, Cat Boost, Gradient Boosting, Extra tree, Support Vector Regressor, Ada Boost, Neural network).
- (b). This research aims to conduct a Comparative analysis study of different machine learning methods to identify the optimal approach for predicting health insurance premiums.
- (c). The aim of this research study is not only improve the prediction models but also to contribute to the application of machine learning in financial and healthcare sectors.

The rest of the paper includes the other sections like methodology and approaches taken, experimental outcomes and results , Discussions, summarizes the findings and provide recommendations for future work.

II. RESEARCH METHODOLOGY

In this study, following steps are taken for the observing the outcome.

2.1 Step-1: Data Preparation and analysis

The original dataset used consists of 1338 rows and 7 columns.

| # | Column | Non-Null | Count | Dtype |
|---|----------|----------|----------|---------|
| 0 | age | 1338 | non-null | int64 |
| 1 | sex | 1338 | non-null | object |
| 2 | bmi | 1338 | non-null | float64 |
| 3 | children | 1338 | non-null | int64 |
| 4 | smoker | 1338 | non-null | object |
| 5 | region | 1338 | non-null | object |
| 6 | charges | 1338 | non-null | float64 |

But the dataset given to the model is a modified version of the dataset. The modifications are made to make dataset denser.

The modifications also result in a dataset of 1338 rows and 7 columns, but 3 columns are different.

| # | Column | Non-Null | Count | Dtype |
|---|-------------|----------|----------|---------|
| 0 | age | 1338 | non-null | int64 |
| 1 | bmi | 1338 | non-null | float64 |
| 2 | children | 1338 | non-null | int64 |
| 3 | charges | 1338 | non-null | float64 |
| 4 | charges_log | 1338 | non-null | float64 |
| 5 | smoker_int | 1338 | non-null | int64 |
| 6 | gender_int | 1338 | non-null | int64 |

The string columns such as 'smoker' and 'gender' have been replaced by 'smoker_int' and 'gender_int' with the help of technique called **Binary encoding**. Furthermore, instead of string column 'region', the column 'charges_log' is introduced. This particular column is made by log transforming the 'charges' column. It resulted in **log parametrized** 'charges_log' column, which made the dataset significantly more complex and denser.

For the analysis part, the data is visualized using various graph. The graphs show the distribution of the data columns as well as relationship between them.

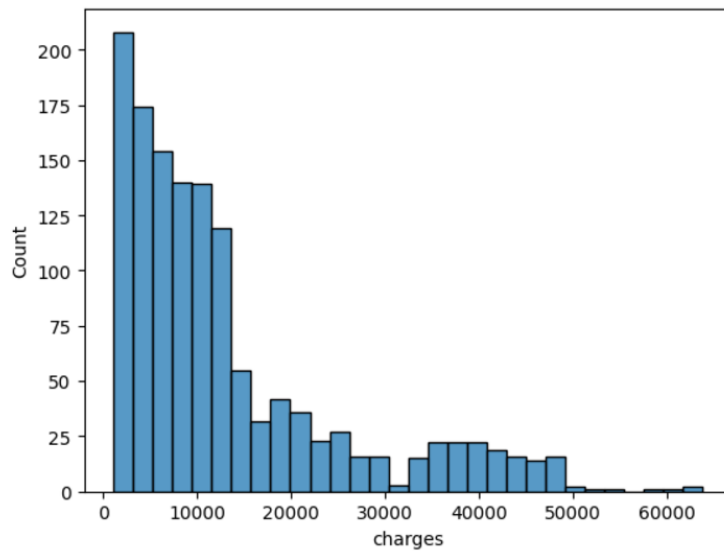


Figure 1. Histogram of charges

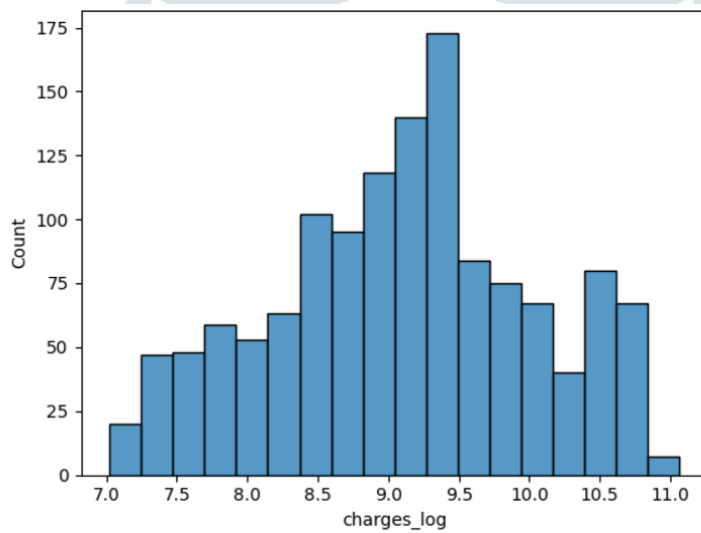


Figure 2. Histogram of charges_log

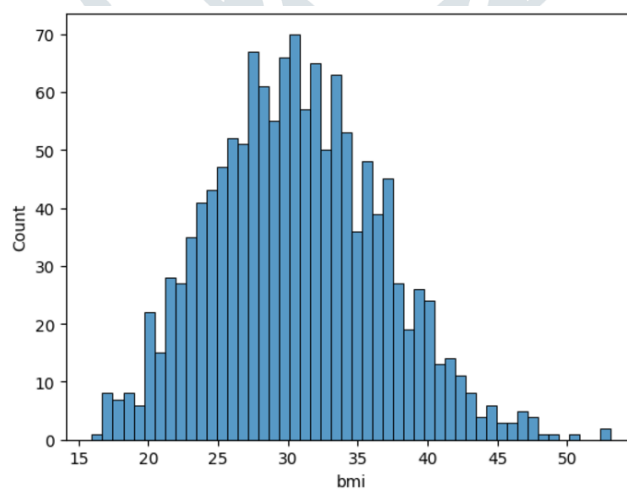


Figure 3. Graph showing distribution of BMI

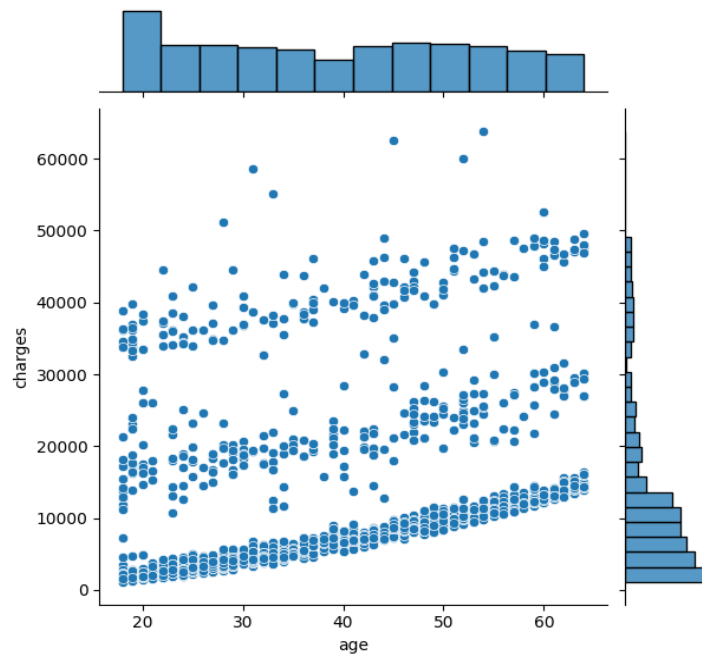


Figure 4. charges vs. age

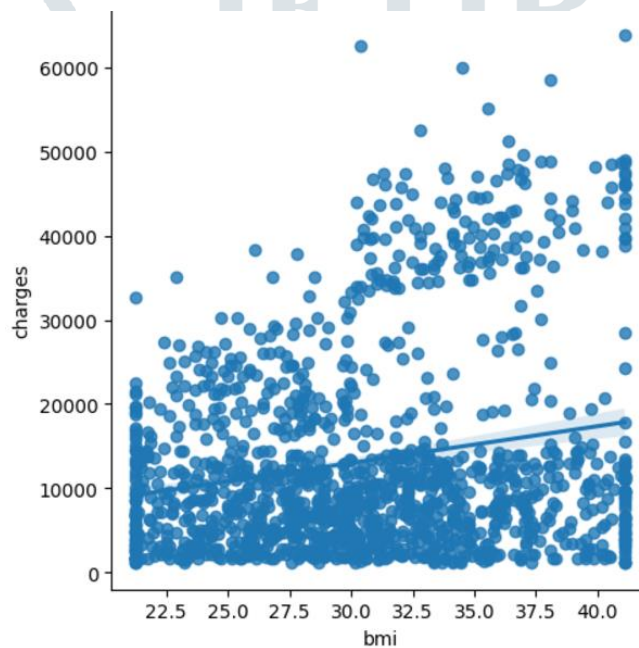


Figure 5. Regplot for BMI vs. Charges

2.2 Part 2: Model Creation

Secondly, the research consisted of creation of various machine learning models. The following algorithms were used for creating models: Linear regression, Decision tree, Random forest, XG Boost, Cat Boost, Gradient Boosting, Extra tree, Support Vector Regressor, Ada Boost, Neural network.

- **Linear Regression:** It is an algorithm in which the prediction is made on basis of linear relationship between the dependent variable and independent variable. It is useful for simple datasets.
- **Decision Tree:** It is also an supervised learning algorithm which has a hierarchical tree-like structure. It is also mainly used for simple datasets.
- **Random Forest:** It is an ensemble learning algorithm which consists of n-number of decision tree and aggregate result is given, thereby punishing outliers. It can be used for complex datasets.
- **XG Boost:** It is also an ensemble machine learning algorithm. It stands for Extreme Gradient Boosting. It uses weaker gradient boosting models to create new effective model.

- **Cat Boost:** It stands for categorical boost. It is also an ensemble algorithm, but string columns can be passed separately inside the model, rather than binary encoding them.
- **Extra Trees:** It is also called extra randomized tree because it uses an ensemble of decision trees, but the split between trees is selected randomly.
- **Gradient Boosting:** It is a variation of extra trees. It works by creating new trees to correct the errors made by previous models.
- **Ada Boost:** It stands for adaptive boosting. It works by creating an ensemble of weak classifiers to create strong classifiers, then train the models using strong classifiers.
- **Support Vector Regressor:** It is a variant of support vector machine used specifically for prediction of continuous data. It does so, by creating a plane that is best fits the data points.
- **Neural Network:** Neural network is a deep machine learning algorithm. It has a structure similar to brain which makes it very appropriate for the complex datasets

2.3 Step 3: Prediction

The accuracy of the models are measured during prediction. In this step, the training data and testing data both are given to the model and their differences are noted. This are known as errors.

The metrics used to measure error for continuous value are Root Mean Square Error(RMSE), Mean Square Error (MAE), Mean Square Error(MSE).

III. RESULTS

The following results are obtained from the study.

| Model | RMSE | MAE | MSE |
|--------------------------|---------------|---------------|-----------------|
| Support vector Regressor | 4406.13883994 | 2701.47234165 | 19414059.47689 |
| Linear Regression | 3298.25963067 | 2440.217251 | 10878516.5913 |
| CatBoost Regressor | 629.99971371 | 389.4503803 | 396899.6392848 |
| XG Boost | 534.1060823 | 356.00621556 | 285269.30715685 |
| Extra Trees | 516.153030359 | 258.63316596 | 266413.9507489 |
| Decision Tree | 287.34836272 | 190.56556403 | 82569.08156142 |
| Random Forest | 256.80319274 | 186.44199507 | 65947.87980649 |
| Gradient Boosting | 224.501626976 | 126.5531811 | 50400.98051506 |
| Neural Network | 164.54252909 | 96.58466541 | 27074.24388 |
| AdaBoost | 149.610953794 | 78.482030387 | 22383.43749515 |

Table 1. Different error metrics against the models used.

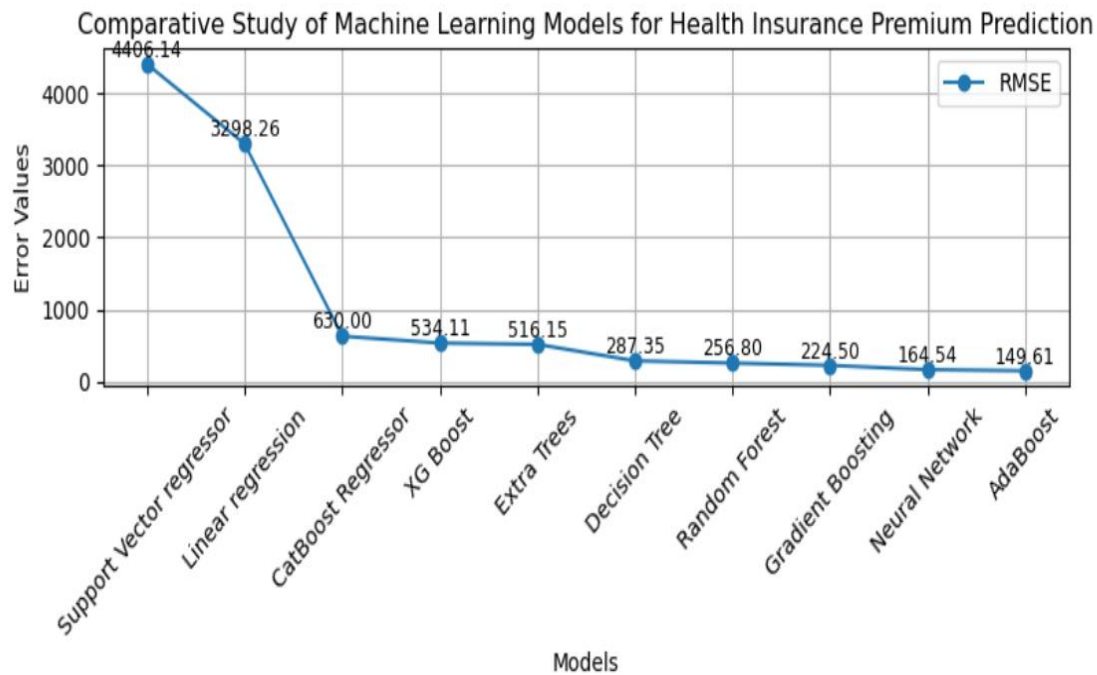


Figure 6. Comparative Study of the algorithms applied

IV. DISCUSSION

The result provides better understanding of algorithms and which one is the best for the particular task. The ensemble methods are better suited to datasets because the relationship between variable are non-linear and complex. Hence, the SVR and Linear regression method are least accurate and AdaBoost was the most accurate.

Furthermore, the dataset is not too complex. Hence, the neural network and ensemble methods have almost similar error range.

V. CONCLUSION

In this research study, typical regression models has taken into a role and that would analyse the different health parameters of the different candidates and recommend an insurance premium pricing. Different regression models namely (Linear regression, Decision tree, Random forest, XG Boost, Cat Boost, Gradient Boosting, Extra tree, Support Vector Regressor, AdaBoost, Neural network) are applied to a Data set which contain patient's health information to obtain the accurate insurance premium cost for the customers of health insurance company. This presented work ensures that the performance of models will be optimized and the comparative study among the different models is carried out so that most accurate model can be identified and can be implemented in real life applications of machine learning in financial and healthcare sectors. Finally the outcome of this research study is declared as the AdaBoost is the best and accurate model with the values of **MAE: 78.48203038** and **MSE: 22383.43749515** and **RMSE: 149.610953794** to predict the accurate premium price against the health insurance policy.

VI. ACKNOWLEDGMENT

We would like to thank Prof. Pushpalata Verma, our project guide. Her valuable guidance, continuous encouragement and expert advice helped us throughout the project preparation journey. We also acknowledge the Department of Computer Science of Bhilai Institute of Technology, Raipur for giving the opportunity to accomplish this project and for providing the necessary resources and infrastructure to carry out this research project.

VII. REFERENCES

- [1] Ridzuan, A. N. A. A., Azman, A. Z., Marzuki, F. A., Faudzi, W. S. D. M., Abd Aziz, S. H., & Bakar, N. A. (2024). Health Insurance Premium Pricing Using Machine Learning Methods. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 41(1), 134-141.

- [2] Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. *Machine Learning with Applications*, 15, 100516.
- [3] Dutta, S., Bose, P., & Bandyopadhyay, S. K. Forecasting health insurance premium using machine learning approaches.
- [4] Kaushik, K., Bhardwaj, A., Dwivedi, A. D., & Singh, R. (2022). Machine learning-based regression framework to predict health insurance premiums. *International journal of environmental research and public health*, 19(13), 7898.
- [5] Taloba, A. I., Abd El-Aziz, R. M., Alshanbari, H. M., & El-Bagoury, A. A. H. (2022). Estimation and prediction of hospitalization and medical care costs using regression in machine learning. *Journal of Healthcare Engineering*, 2022(1), 7969220.
- [6] Meenakshisundaram, D. K., & Krishnekumar, D. S. (2020). Age Factor—A Basic Parameter for Health Insurance—A Study with Special Reference to Chennai City among Standalone Health Insurers. *International Journal of Management and Humanities*, 4(5), 78-88.
- [7] Committee on the Consequences of Uninsurance. (2003). *A shared destiny: community effects of uninsurance*. National Academies Press.
- [8] Li, C., Tang, C., & Wang, H. (2019). Effects of health insurance integration on health care utilization and its equity among the mid-aged and elderly: evidence from China. *International Journal for Equity in Health*, 18, 1-12.
- [9] Selamat, E. M., Abd Ghani, S. R., Fitra, N., & Daud, F. (2020). Systematic review of factors influencing the demand for medical and health insurance in Malaysia. *International Journal of Public Health Research*, 10(2).
- [10] Chand, S., & Zhang, Y. (2022). Learning from machines to close the gap between funding and expenditure in the Australian National Disability Insurance Scheme. *International Journal of Information Management Data Insights*, 2(1), 100077.
- [11] Stämpfli, D., Winkler, B. A., Vilei, S. B., & Burden, A. M. (2022). Assessment of minor health disorders with decision tree-based triage in community pharmacies. *Research in Social and Administrative Pharmacy*, 18(5), 2867-2873.
- [12] Rawat, S., Rawat, A., Kumar, D., & Sabitha, A. S. (2021). Application of machine learning and data visualization techniques for decision support in the insurance sector. *International Journal of Information Management Data Insights*, 1(2), 100012.