









Analysis and Prediction of Health Insurance Cost Using Machine Learning Approaches

Dileep Kumar Kadali ^{1,*}, M. Lakshmi Narayana ², V.S.N. Murthy ³, Srinivasa Rao Dangeti ⁴, Yugandhar Bokka ⁵, Samatham Chandra Sekhara Rao ⁶

^{1,3,4,5,6} Shri Vishnu Engineering College for Women, Bhimavaram, A.P, India
²S.R.K.R Engineering College, Bhimavaram, A.P, India

*dileepkumarkadali@gmail.com

Abstract. The intensifying cost of healthcare needs tools for up-to-date insurance ranges. Machine Learning approaches for predicting individual healthcare insurance costs are analyzed with the help of patient records, and a personalized cost estimation model empowers individuals, particularly in rural areas, to navigate complex insurance options. Unlike existing solutions, our model does not predict specific company costs but provides a personalized cost range. To overcome to proposed this paper focuses on affordability and informed decision-making and addresses challenges like limited health literacy and lack of awareness of government-provided schemes. The machine learning algorithms are Gradient Boosting and Random Forest to achieve high accuracy enabling all individuals, especially those in underserved communities, to make informed healthcare investment decisions.

Keywords: Machine Learning, Estimation Model, Random Forest, Gradient Boosting etc.

1 Introduction

Navigating the complex and often opaque world of healthcare insurance can be a daunting task, particularly for individuals facing affordability challenges or limited access to information [11]. expand more the ever-rising cost of healthcare further compounds this issue, creating a significant barrier to accessing quality care for many [16]. This study presents a novel Machine Learning (ML)-based approach for predicting personalized healthcare insurance costs, aiming to empower individuals, especially those in underserved communities, to make informed decisions about their insurance selection [14]. The challenges facing individuals within the healthcare insurance system are multifaceted [15]. The escalating cost of healthcare and intricate insurance structures often create substantial financial burdens, particularly for those in rural areas and underserved communities where limited resources and information compound the difficulty in navigating options [8]. Lack of financial literacy further exacerbates the challenge, leaving

© The Author(s) 2024

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,

https://doi.org/10.2991/978-94-6463-471-6_55

individuals unsure about navigating complex financial products and making informed choices. Furthermore, existing methods for predicting healthcare insurance costs often focus on specific companies or averages, leaving individuals unsure about their financial responsibility [10]. Exclamation This lack of personalized cost estimation can lead to confusion, misinformed decisions, and potentially unnecessary expenses. Individuals might opt for overly expensive plans due to a lack of transparency or settle for inadequate coverage due to cost concerns [12]. Adding to these challenges, many individuals, particularly in rural areas, remain unaware of government-provided health insurance schemes. expand more This lack of awareness further restricts access to essential healthcare services for financially vulnerable populations, exacerbating existing health disparities [13].

2 Literature Review

This writing review starts with an outline of medical coverage zeroing in on well-being data, looking for models, and lastly dynamic hypotheses pertinent to healthcare coverage education research. It then talks about generally speaking well-being education research, well-being proficiency estimations, and medical coverage education research. Moreover, it investigates decisions in health care coverage navigation and existing examination around here. Finally, the review examines involving semi-organized interviews in medical coverage proficiency research. A Far-reaching Investigation of Medical Services Bigdata The board, Examination and Logical Programming. This paper gives a far-reaching report on medical services large information, stressing the utilization of enormous information examination for demonstrating, surmising, grouping, expectation, bunching, relapse, and other nonexclusive methodologies in medical care [1]. Anticipating Engine Protection Cases Utilizing Telematics Information XGBoost versus Calculated Relapse. This study analyzes calculated relapse and XGBoost approaches for anticipating mishap claims utilizing telematics information, high-lighting strategic relapse's interpretability and prescient limit [3]. Computerizing Vehicle Protection Cases Utilizing Profound Learning Method. This paper examines vehicle protection claims utilizing profound learning methods, zeroing in on handling pictures of harmed vehicles to gauge fixed costs [5]. Foreseeing Client Beat with AI Strategies. This proposal investigates client stir expectations utilizing AI, zeroing in on the use of beat forecast ideas in a Finnish insurance agency [4]. Different Ascription for Missing Information in Epidemiological and Clinical Exploration: Potential and Entanglements. This article surveys the utilization of various attributions for dealing with missing information in epidemiological and clinical exploration, talking about its advantages, entanglements, and ongoing use in clinical diaries [2]. A controlled ML model for recognizing the verification of mosquitoes from the backscattered optical sign was made in the survey [7]. The survey showed that the optical sensor joined with coordinated ML can be a sensible elective means for noticing the mosquito people. The farsighted oversaw ML approach for the appraisal of the bet recurrent at the outset periods of oral tongue squamous cell carcinoma has been made in the work [6]. The result of the audit showed the limit of directed ML to expect locoregional reshapes. Regulated ML

calculations which incorporate help vector machines, straight discriminant examination, and K-closest neighbour calculations were utilized to distinguish dementia in the work. The consequence of the review showed that the calculations are fit for anticipating dementia [9].

3 Existing Methodology

The current system utilizes linear regression analysis for predicting healthcare insurance costs. Linear regression models the association flanked by a dependent variable (y) and one or more descriptive variables (x). While there have been advancements in this field, certain limitations persist. Simple Linear Regression, for example, uses only one illustrative variable. However, the method has several disadvantages:

- Limited to analyzing only two columns of the dataset.
- Focuses on open and closed values.
- The accuracy of predictions is often unsatisfactory.

These limitations highlight the need for more advanced techniques, such as those proposed in our framework, to progress the accuracy and strength of health insurance cost predictions.

4 Proposed Model

Based on our study addresses these challenges by presenting a data-driven approach that leverages the power of machine learning to:

Develop a Personalized Cost Estimation Model: We utilize diverse patient data and ML algorithms like Random Forest and Gradient Boosting to predict personalized healthcare insurance cost ranges for individuals. This empowers them to make informed choices based on their unique health profile and financial constraints, fostering affordability and financial well-being.

Focus on Underserved Communities: We recognize the specific needs of individuals in rural areas and underserved communities who often face limited access to healthcare resources and information. Our model aims to bridge this gap by providing clear and accessible cost estimations, promoting health equity and informed consent.

Increase Awareness of Government Schemes: By incorporating data on government-provided health insurance schemes into our model, we aim to increase awareness and encourage individuals to explore these options, further expanding access to affordable healthcare coverage.

Empowering Individuals: By providing personalized cost estimates, our model empowers individuals to make informed decisions about their healthcare insurance, promoting financial well-being and informed consent. This allows them to choose plans that align with their financial reality and healthcare needs.

Promoting Health Equity: Our focus on underserved communities and rural areas contributes to addressing socioeconomic disparities in healthcare access. By enabling

informed choices and facilitating access to affordable coverage, we can help narrow the healthcare gap and promote health equity for all.

Expanding Access to Care: Increased awareness of government-provided health insurance schemes can enable more individuals to access essential healthcare services. This can have a significant impact on the health outcomes of vulnerable populations and reduce the burden on healthcare systems.

Improving Healthcare Efficiency: By offering data-driven insights into individual cost profiles, our model can potentially aid healthcare providers and insurers in developing more efficient and targeted insurance options. This can lead to more sustainable healthcare systems and improved health outcomes.

This research holds great promise for improving healthcare access and affordability, particularly for underserved communities. This paper details the development and evaluation of our ML-based personalized healthcare insurance cost prediction model, outlining its methodology, results, and potential future directions. We hope this contribution paves the way for a more equitable and accessible healthcare system. Healthcare insurance prediction system using the Gradient Boosting algorithm. Gradient Boosting is a supervised learning algorithm that builds an ensemble of decision trees sequentially. Gradient Boosting builds trees one at a time, each tree correcting errors made by the previous one. This iterative process allows Gradient Boosting to continuously improve the model's accuracy. Gradient Boosting is highly accurate and often outperforms other machine learning algorithms. Its container handles an assortment of data types, counting numerical and definite data. Gradient Boosting is robust against overfitting, thanks to its regularization parameters. It automatically handles missing data, reducing the need for data preprocessing. It provides feature importance scores, allowing users to understand the most influential factors in the model's predictions. By using Gradient Boosting for healthcare insurance prediction, this system aims to improve prediction accuracy and provide valuable insights into the factors influencing insurance costs.

5 Methodology Working Architecture

The working architecture of a machine learning model involves several key stages shown in Fig 1. Firstly, data is collected from various sources and preprocessed to make it suitable for analysis. Succeeding, related features are selected or mined from the data. A suitable machine learning model is then chosen and trained using the prepared data. The model is evaluated for its performance using a separate testing dataset. Hyperparameters of the model are tuned to improve its performance. Finally, the trained model is deployed to predict new data in a real-world setting.

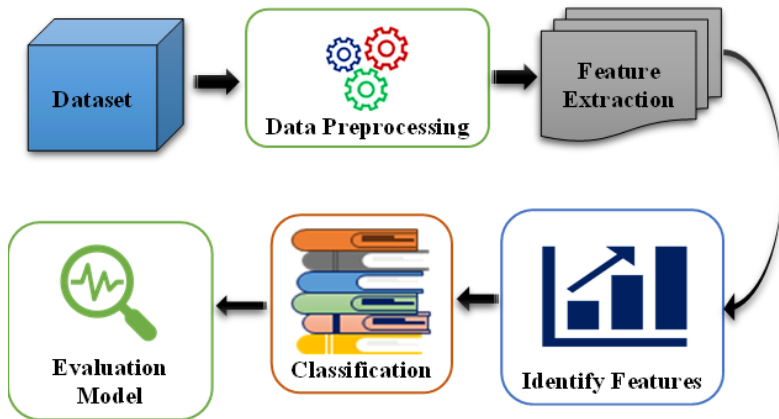


Fig. 1. Methodology Model Evaluation Process

6 RESULTS AND DISCUSSION

Our framework encompasses various techniques and methods to build a health insurance cost prediction model:

Data Collection: Relevant data is gathered from bases such as medical records, surveys, and insurance claims, ensuring it is complete and accurate we used a dataset from the Kaggle website.

Data Preprocessing: The collected data is cleaned and transformed to make it suitable for analysis, addressing missing values, duplicates, and outliers, and converting categorical variables to numerical data.

Data Partitioning: The preprocessed data is divided into train and test datasets, with a representative ratio of train data and test data are 70:30 or 80:20 respectively.

Feature Selection: The most relevant features for the model are selected based on their correlation with the target variable (health insurance cost), avoiding highly correlated features to prevent overfitting.

Model Training: Machine learning models like Linear Regression, Support Vector Machine (SVM), Random Forest trained and Gradient Boosting models are applied to the training dataset using selected features. The learning models show their prediction values shows in Fig 2.

Model Evaluation: The trained model is evaluated on the testing dataset using metrics such as mean squared error, mean absolute error, and R-squared to assess its performance and generalization.

Hyperparameter Tuning: The Gradient Boosting model's performance is further improved by tuning hyperparameters such as the number of trees, maximum depth, and minimum samples shown in Fig 3.

Model Deployment: Once optimized, the model is deployed to predict health insurance costs for new patients, taking inputs such as age, sex, BMI, and smoking habits to provide predicted costs shown in Fig 4.

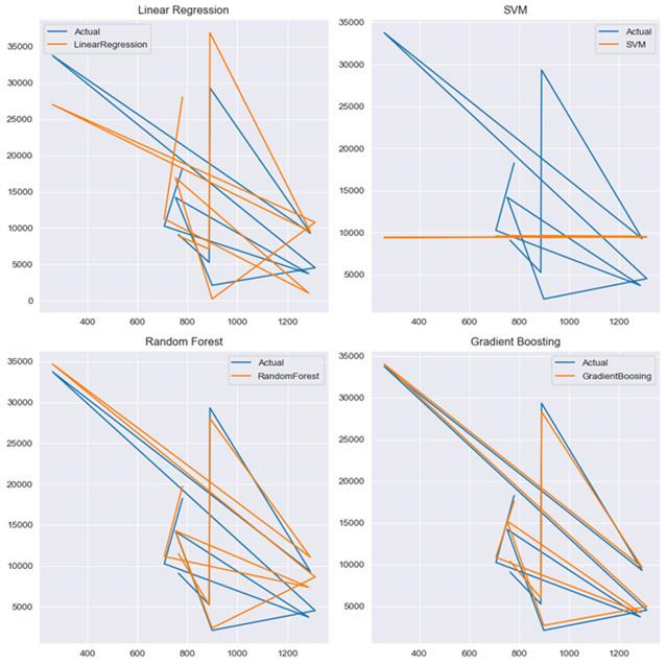


Fig. 2. Predict Model based on Evaluation Process

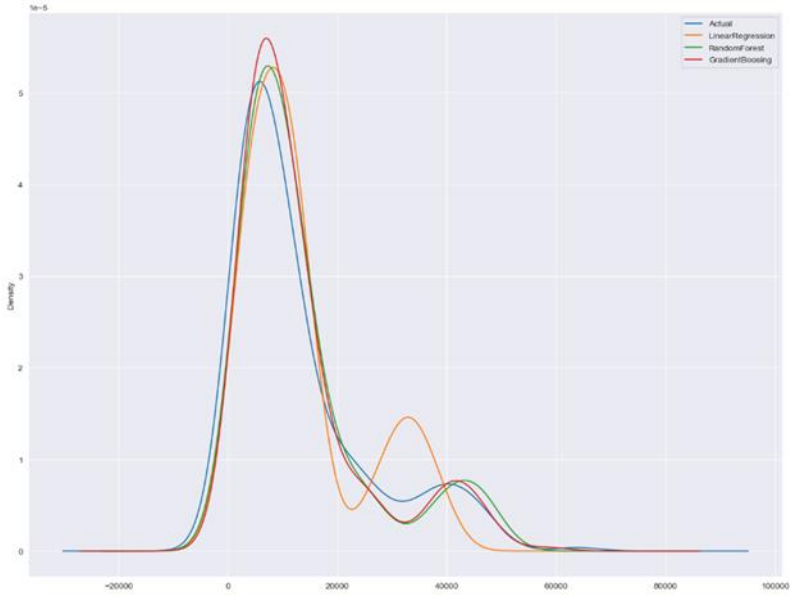


Fig. 3 Comparing the Model Evaluation Process

Health Insurance Cost Prediction

Enter your age: 35

Gender: Male Female

Enter your BMI Value: 32.50

Enter number of children: 3

Smoker: Yes No

Region: North East North West South East South West

Your Insurance cost is 7040.31

Fig.4 User Interactive Model

7 CONCLUSION

In this paper, machine learning models present a compelling solution for tasks traditionally performed on the user interface module. Machine learning can efficiently analyze large volumes of data, simplifying and streamlining health insurance operations. This technological advancement is poised to save time and money for both policyholders and insurers. It allows insurance experts to concentrate on enhancing the policyholder experience. Machine learning's impact extends to patients, hospitals, physicians, and insurance providers, enabling it to execute tasks currently performed by humans but at a significantly faster pace and lower cost. The model was estimated by key recital metrics such as MSE, MAE, RMSE, R-square, and adjusted R-square, achieving an accuracy of 92.72%. Additionally, correlation information forms a matrix designed to visualize the relationship between them many factors and insurance charges. In further extension is possible to evaluate metrics for all kinds of models in machine learning and then based on time complexity choose the reliable model on that.

References

1. Shakhovska, N., Melnykova, N., & Chopiyak, V. "An Ensemble Methods for Medical Insurance Costs Prediction Task". *Computers, Materials & Continua*, 70(2).
2. Kayri, M., Kayri, I., & Gencoglu, M. T. "The performance comparison of Multiple Linear Regression, Random Forest and Artificial Neural Network by using photovoltaic and atmospheric data". In *2017 14th International Conference on Engineering of Modern Electric Systems (EMES)* (pp. 1-4). IEEE.
3. D. K. Kadali, R. Mohan, N. Padhy, S. C. Satapathy, N. Salimath, and R. D. Sah, "Machine learning approach for corona virus disease extrapolation: A case study," *International*

- Journal of Knowledge-based and Intelligent Engineering Systems, vol. 26, no. 3, pp. 219–227, Dec. 2022, doi: 10.3233/kes-220015.
4. Kowshalya, G., & Nandhini, M. (2018, April). Predicting fraudulent claims in automobile insurance. In 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 1338-1343). IEEE.
 5. Avanija, J., G. Sunitha, and K. Reddy Madhavi. "Semantic Similarity based Web Document Clustering Using Hybrid Swarm Intelligence and FuzzyC-Means." *Helix* 7, no. 5 (2017): 2007-2012.
 6. Gupta, S., & Tripathi, P. (2016, February). An emerging trend of big data analytics with health insurance in India. In 2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH) (pp. 64-69). IEEE.
 7. D. K. Kadali, D. Raju, and P. V. R. Raju, "Cluster query optimization technique using Blockchain," in *Cognitive science and technology*, 2023, pp. 631–638. doi: 10.1007/978-981-99-2742-5_65.
 8. P. Maddula, P. Srikanth, P. K. Sree, P. B. V. R. Rao and P. T. S. Murty, "COVID-19 prediction with Chest X-Ray images using CNN," 2023 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE), Bengaluru, India, 2023, pp. 568-572, doi: 10.1109/IITCEE57236.2023.10090951.
 9. D. K. Kadali, R. Mohan, and M. C. Naik, "Enhancing crime cluster reliability using neutrosophic logic and a Three-Stage model," *Journal of Engineering Science and Technology Review*, vol. 16, no. 4, pp. 35–40, Jan. 2023, doi: 10.25103/jestr.164.05.
 10. D. K. Kadali, M. C. Naik, and K. N. Remani, "Estimation of data parameters using cluster optimization," in *Lecture notes on data engineering and communications technologies*, 2022, pp. 331–342. doi: 10.1007/978-981-19-2600-6_23
 11. A. S. Mallesh, N. Pamarthi, P. T. S. Murty, P. K. Sree, T. Daniya and B. Maram, "Smart System for Early Detection of Agricultural Plant Diseases in the Vegetation Period," 2023 1st DMIHER International Conference on Artificial Intelligence in Education and Industry 4.0 (IDICAIEI), Wardha, India, 2023, pp. 1-6, doi: 10.1109/IDICAIEI58380.2023.10406672.
 12. D. K. Kadali, "Cluster optimization for similarity process using De-Duplication," Sep. 01, 2016. <https://ijsrd.com/Article.php?manuscript=IJSRDV4I60433>
 13. D. K. Kadali, J. Mohan, and Y. Vamsidhar, "Similarity based Query Optimization on Map Reduce using Euler Angle Oriented Approach," <https://www.ijser.org/>, Jan. 2012,
 14. K. N. Remani, V. S. Naresh, S. Reddi, and D. K. Kadali, "Crime data optimization using neutrosophic logic-based game theory," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 15, Mar. 2022, doi: 10.1002/cpe.6973.
 15. D. K. Kadali and R. Mohan, "Shortest route analysis for High-Level Slotting using Peer-to-Peer," in *Apple Academic Press eBooks*, 2022, pp. 113–122. doi: 10.1201/9781003048367-10.
 16. D. K. Kadali, R.N.V.J. Mohan, "Risk minimization Process on crime cluster data," Apr. 30, 2023. <https://www.tijer.org/viewpaperforall?paper=TIJER2304443>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

