// [1]

Medical Insurance Cost Prediction Using Machine Learning

D.Sri Manjunadh -221FA04721
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
SECTION:3G-CSE
BATCH-14

Sampath -221FA04464
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
SECTION:3G-CSE
BATCH-14

Rohitha-221FA04476
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
SECTION:3G-CSE
BATCH-14

Chiranjeevi-221FA04484
Department of Computer Science And Engineering
Vignan's Foundation for Science, Technology and Research
SECTION:3G-CSE
BATCH-14

***Abstract*—Insurance is a policy that reduces or eliminates the expenses associated with decreasing returns brought on by various risks. The price of insurance is influenced by a number of factors. These elements have an impact on how insurance plans are developed. The efficiency of insurance policy terms in the insurance industry can be enhanced using machine learning (ML). In this work, we use individual and local health data to forecast insurance amounts for various categories of people. To compare the effectiveness of these algorithms, nine regression models—Linear Regression, XGBoost Regression, Lasso Regression, Random Forest Regression, Ridge Regression, Decision Tree Regression, KNN Model, Support Vector Regression, and Gradient Boosting Regression—were utilized. The models were trained using the dataset, and some predictions were then made using the training data. The model was then put to the test and confirmed by contrasting the actual data with what was predicted to be abundant. These models' accuracy was compared subsequently. The optimal method to the XGBoost MAE 2381.567, MSE 19806356.6067, RMSE value 4450.4433, and R squared value of 0.8681 is provided in this report. Gradient Boosting and Random Forest, with R squared values of 0.8679 and 0.8382, respectively, are two further top models.**

***Index Terms*—Healthcare; Insurance; Regression, Machine Learning, Prediction, Data analysis.**

I. Introduction

A sector that is quickly growing globally is digital health. The number of digital health businesses has doubled globally during the last five years [1]. Health insurance faces two significant obstacles in industrialized nations: growing health care costs and an increase in the number of people without coverage. A broad-based political movement to address these issues is emerging as a result of this power. Governments in the area have pledged hundreds of millions of dollars to advance the digital health industry. Individual health insurance plays a crucial role in the healthcare system, particularly for people with rare diseases [2], for whom medical and preventative insurance can help cut down on treatment expenses. The world in which we live is a dangerous and unknowable place. houses, companies, buildings.In the last two decades, universal health care has been almost doubled, surpassing US $ 8.5 trillion in 2019, or 9.8% of global GDP [18]. People's happiness and health are fundamental to their existence. In the last two decades, universal health care has been almost doubled, surpassing US $ 8.5 trillion in 2019, or 9.8% of global GDP [17]. These products employ money to make up for the risks; as a result, the costs of some risks are reduced or even eliminated. [3]. A crucial component of the medical industry is medical insurance. Furthermore, novel ranking techniques with machine learning algorithms are applied to classify cost prediction in health insurance [16]. The factors of training time and accuracy are looked at. The bulk of machine learning algorithms only require a brief time of training. Because [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9265373/#B1-ijerph-19-0789819] numerous factors influence the insurance premium of a health insurance policy, the premium amount varies from person to person.Deep learning models can also find hidden patterns, but their usage in real time is constrained by the training period [4].

Identify applicable funding agency here. If none, delete this.

II. Literature Review

This section demonstrates the research being done on information exploration and machine learning methods. Several articles have addressed the topic of claim prediction. Providing decision aids that increase accuracy without sacrificing simplicity could strengthen the policy approach that relies on individuals making the best health insurance cost-based decision methods [15].. heir work not only showcases the potential of mobile technology to transform healthcare delivery but also underscores the importance of user-centric design, data privacy, and equitable access in shaping the future of mHealth applications [11]. The results and vibrations indicated that logistic regression is a superior model to XGBoost

for the reasons of its interpretability and predictability [7]. Without taking into account predicted cost and claim scope, the research listed above identify claims problems. [20] use data mining techniques, explicitly clustering algorithms and classification trees, and insurance claim data of nearly 500,000 members throughout a three-year period. Based on the data gathered from the medical expenses from the first two years, a justified third-year health- care cost projection is made This section demonstrates the research being done on information exploration and machine learning methods. Several articles have addressed the topic of claim prediction. By addressing key challenges and leveraging emerging technologies, these startups have the potential to drive significant advancements in healthcare quality, accessibility, and affordability, ultimately transforming the future of healthcare [8].. In a relatively small number of cases, this study compared the effectiveness of logistic regression and XGBoost strategies in predicting the occurrence of accident states. The results and vibrations indicated that logistic regression is a superior model to XGBoost for the reasons of its interpretability and predictability [7]. Without taking into account predicted cost and claim scope, the research listed above identify claims problems.With few exceptions, most regression studies to date have not addressed this problem [14] . hThis section demonstrates the research being done on information exploration and machine learning methods. The longitudinal approach, combined with a focus on policy implications, enhances our understanding of the complex dynamics surrounding access to orphan drugs and underscores the importance of evidence-based policymaking in this critical area of healthcare [9].Gupta and Tripathi advocate for the development of robust data governance frameworks and collaborative partnerships between insurers, healthcare providers, and technology vendors to overcome these challenges and unlock the full potential of big data analytics [10]. The results and vibrations indicated that logistic regression is a superior model to XGBoost for the reasons of its interpretability and predictability [7]. Without taking into account predicted cost and claim scope, the research listed above identify claims problems. *Providing decision aids that increase accuracy without sacrificing simplicity could strengthen the policy approach that relies on individuals making the best health insurance cost-based decision methods [12].*

III. Methodology

A. Dataset Description We obtained the data set from the Kaggle website [5] in order to calculate the cost of this model prediction. The data set is split into two categories: training data and test data, and it has seven attributes as listed in table I. The majority of the data used is for testing, with just around 20% being used for training. The training data set is used to create a model thatforecasts medical insurance costs by year, and the test data set is used to assess the regression model. The table below contains the dataset description.

IV. Data Collection and Preprocessing

There were 1338 rows and 7 columns in our data set. The charges variable, which has a float value, is our aim. Maximum number of individuals in our dataset range in age from 18 to 22.5, and the majority of them are male. Few have more than three children, and the majority of them have a BMI between 29.26 and 31.16. In this dataset, four main regions are taken into account: northeast, northwest, southeast, and southwest. The largest concentration of smokers is in the southeast, where 1064 out of 1338 people smoke. Here are some data visualizations. Only numerical values are presented. Standard deviations and average values for categorical variables are absent. .[13] From these variables, each one of these attributes has some contribution to estimating the cost of the insurance, which is our dependent variable.. The median number is higher than the average in the ”charges” column. It implies that the price of health insurance is unfairly skewed. Once we make those things visible, we will clearly grasp this. We therefore begin by displaying the charge column’s distribution

Three columns are numerical and three are categorical. Our machine learning model cannot suit the category values because computers cannot understand this text value. Therefore, we will give those categories qualities numerical labels. We change ”female” to 1 and ”male” to 0 in the ”sex” field. We also change the other two columns to have numerical values.

V. Data Model Specification

D. Model Specification The goal of the study is to forecast insurance costs based on a variety of factors, including age, sex, the number of children, location, BMI, and whether or not a person smokes. All of these characteristics aid in our ability to calculate the price of health insurance. Several regression models are used in this study to calculate the cost of health insurance. There are two portions to the data. Model testing is done in the other portion, whereas model training is done in the first. Data is used for training 80% of the time and testing 20%. We compute the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared value (RE), and Mean Squared Error (MSE) for each model to see how accurate it is in predicting costs. We compare them after generating those numbers for each model since it shows us the accurate result.

VI. Model Specification

D. Model Specification The goal of the study is to forecast insurance costs based on a variety of factors, including age, sex, the number of children, location, BMI, and whether or not a person smokes. All of these characteristics aid in our ability to calculate the price of health insurance. Several regression models are used in this study to calculate the cost of health insurance. There are two portions to the data. Model testing is done in the other portion, whereas model training is done in the first. Data is used for training 80% of the time and testing 20%. We compute the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared value (RE), and Mean Squared Error (MSE) for each model to see how

accurate it is in predicting costs. We compare them after generating those numbers for each model since it shows us the accurate result.

VII.

VIII. REFERENCES

[1] ”Digital Health 150: The Digital Health Startups Transforming the Future of Healthcare — CB Insights Research”, CB Insights Research, 2022. [Online]. Available: https://www.cbinsights.com/research/report/digitalhealth-startups-redefining-healthcare. [Accessed: 10- Sep- 2022]

[2] J. H. Lee, “Pricing and reimbursement pathways of new ophan drugs in South Korea: A longitudinal comparison. in healthcare,” Multidisciplinary Digital Publishing Institute, vol. 9, no. 3, pp. 296, 2021.

[3] 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

[4] N. Shakhovska, S. Fedushko, I. Shvorob and Y. Syerov, “Development of mobile system for medical recommendations,” Procedia Computer Science, vol. 155, pp. 43–50, 2019

[5] Medical Cost Personal Datasets: https://www.kaggle.com/datasets/mirichoi0218/insuran ce

[6] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, ”Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression, ” Risks, vol. 7, no. 2, p. 70, Jun. 2019, doi: 10.3390/risks7020070.

[7] M. hanafy and O. Mahmoud, ”Predict Health Insurance Cost by using Machine Learning and DNN Regression Models”, International Journal of Innovative Technology and Exploring Engineering, vol. 10, no. 3, pp. 137-143, 2021. Doi: 10.35940/ijitee.c8364.0110321

[8] ”Digital Health 150: The Digital Health Startups Transforming the Future of Healthcare — CB Insights Research”, CB Insights Research, 2022. [Online]. Available: https://www.cbinsights.com/research/report/digital-health-startups-redefining-healthcare. [Accessed: 10- Sep- 2022].

[9] J. H. Lee, “Pricing and reimbursement pathways of new ophan drugs in South Korea: A longitudinal comparison. in healthcare,” Multidisciplinary Digital Publishing Institute, vol. 9, no. 3, pp. 296, 2021.

[10] 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.

[11] N. Shakhovska, S. Fedushko, I. Shvorob and Y. Syerov, “Development of mobile system for medical recommendations,” Procedia Computer Science, vol. 155, pp. 43–50, 2019.

[12] M. hanafy and O. Mahmoud, ”Predict Health Insurance Cost by using Machine Learning and DNN Regression Models”, International Journal of Innovative Technology and Exploring Engineering, vol. 10, no. 3, pp. 137-143, 2021. Doi: 10.35940/ijitee.c8364.0110321

[13] Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Prediction vehicles insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 704. Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah, R. R. (2019, September). Automating Vehicle Car Insurance Claims Using Deep Learning Techniques

[14] C. A. Powers, C. M. Meyer, M. C. Roebuck and B. Vaziri, ”Predictive modeling of total healthcare costs using pharmacy claims data: A comparison of alternative econometric cost modeling techniques”, *Med. Care*, vol. 43, pp. 1065-1072, 2005.

[15]MC Politi, E Shacham, AR Barker, N George, N Mir, S Philpott et al., ”A Comparison Between Subjective and ObjectiveMethods of Predicting Health Care Expenses to Support Consumers’”.

[16] G. Satya Mounika Kalyani, Rama Parvathy L, “A Novel Ranking Approach to Improved Health Insurance Cost Prediction by Comparing Linear Regression to Random Forest”, Journal of Survey in Fisheries Sciences, 2023, 10(1S) 2030-2039.

[17]“Global Expenditure on Health”, WHO annual report 2021, [Online].Available:https://www.who.int/newsroom/events/detail/2021/12/15/default-calendar/global-spending-onhealth-2021

[18] “Health Insurance of India’s missing middle”, Niti Ayog India, Oct 2021, [Online]. Available: https://www.niti.gov.in

[19]Health Insurance Premium Prediction with Machine Learning.[(accessed on 9 May 2022)]. Available online: https://thecleverprogrammer.com/2021/10/26/health-insurance-premium-prediction-with-machine-learning/

[20] Bertsimas, M. V. Bjarnad´ottir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala, and G. Wang, “Algorithmic prediction of health-care costs,” Operations Research, vol. 56, no. 6, pp. 1382–1392, 2008.