e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

International Journal Of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704145 |

Medical Health Insurance Cost Prediction

Vaahini Reddy K, Kanisetti Uday Kalyani, Keerthy Reddy K , Dr .D. Shravani,

B.E. Computer Engineering, Stanley college of Engineering and Technology for women, Osmania University, Hyderabad, Telangana India

B.E. Computer Engineering, Stanley college of Engineering and Technology for women, Osmania University, Hyderabad Telangana India

B.E. Computer Engineering, Stanley college of Engineering and Technology for women, Osmania University, Hyderabad Telangana India

Associate Professor, ADCE, Stanley college of Engineering and Technology for women, Osmania University, Hyderabad, Telangana, India

ABSTRACT: In comparison to other nations, India's government allocates only 1.5% of its annual GDP to public healthcare. On the other hand, over the past 20 years, worldwide public health spending has nearly doubled along with inflation, reaching US $8.5 trillion in 2019, or 9.8% of global GDP. Around 60% of comprehensive medical procedures and 70% of outpatient care are provided by multinational multi- private sectors, who charge people exorbitant prices. Health insurance is becoming into a necessity for everyone due to the rising cost of high-quality healthcare, rising life expectancy, and the epidemiological shift toward non-communicable diseases. In the previous ten years, there has been a significant increase in insurance data, and carriers now have access to it. To improve outcomes, the health insurance system looks into predictive modelling.

KEYWORDS*:* Machine Learning, Regression Models, Ridge Regression, Linear Regression, Multiple Linear Regression and Polynomial Regression.

I.INTRODUCTION

Health insurance is an important part of healthcare, providing financial protection to individuals against the costs of medical treatment. Accurate prediction of health insurance costs can help providers determine premiums, allocate resources, and make informed business decisions. Few years before, machine learning have been used to predict health insurance costs. Machine learning algorithms can analyze big amount of data, patterns and relationships between variables that are difficult for humans to find out. These algorithms can be used to develop models that accurately predict the cost of health insurance for individuals based on their demographics, medical history, and lifestyle factors. In this study, we collected data on a sample population and preprocessed the data to ensure its quality and usability. We then developed several models using various machine learning algorithms, including support vector machine, deep learning, linear regression, decision tree, and random forest. We evaluated the performance of these models using various performance metrics, including mean squared error and R-squared. The study aim is to contribute to the growing body of research on health insurance cost prediction and provide insights into the effectiveness of different machine learning algorithms. The results of this study can assist insurance providers in making informed decisions and improve the accuracy of health insurance cost prediction. The study also provides a foundation for further research on health insurance cost prediction, including the exploration of other variables that may impact the cost of health insurance.

II. LITERATURE SURVEY

Several research projects on calculating medical costs have been published in many health-related contexts. Many likely assumptions underlie machine learning, however, the performance of it depends on using a virtually accurate method. pertaining to the mentioned problem domain and using the acceptable methods to create, train, and use the model. Moran and coworkers "used a thorough linear the cost of an item using the regression method ICU using patient profile information and DRGs The amount of time spent in the groups (diagnosis-related) a hospital, and other characteristics."

The model Sushmita et al. "presented was based on Using a person's medical history and previous spending patterns, future medical expenses expected costs for each quarter. Machine learning (ML) algorithms for predicting health insurance premiums are continuously being researched and developed in the healthcare industry. A computational intelligence method for calculating healthcare insurance costs using a variety of machine learning approaches was proposed in the work of [2]. One piece [3] started out by considering the possible effects of employing predictive algorithms to determine insurance rates. Would this put the concept of risk mutualization in jeopardy, leading to new forms of bias and insurance exclusion? The second part of the study examined how the insured's realization that the corporation had a wealth of continuously updated information about her actual behavior affected their relationship.

III. PROPOSED METHODOLOGY

Data Collection and Pre-processing –

The success of machine learning algorithms in health insurance cost prediction largely depends on the quality of the data used. In this study, we collected data from a health insurance provider on a sample population. The data included demographics, medical history, and lifestyle factors. Before using the data to develop models, we pre-processed it to ensure its quality and usability. The pre-processing steps included data cleaning, data transformation, and data normalization.

Data Cleaning –

Data cleaning involves identifying and correcting or removing errors, inconsistencies, and missing values in the dataset. In this study, we used various techniques to clean the data. We removed duplicates, corrected typographical errors, and replaced missing values with reasonable estimates.

Data Transformation –

Data transformation involves converting the raw data into a format that can be used by machine learning algorithms. In this study, we transformed the data into a numerical format by encoding categorical variables. For example, we used one-hot encoding to represent categorical variables, such as gender and smoking status, as numerical values.

Data Normalization –

Data normalization involves scaling the data to ensure that all variables have equal importance in the model. In this study, we normalized the data usingthe min-max scaling technique. This technique scales the data to a range of 0 to 1, ensuring that all variables have equal importance in the model. After pre-processing the data, we split it into training and testing sets. We used 70% of the data for training the models and 30% for testing the models. The split was done randomly to ensure that the training and testing sets had similar distributions of variables.

In conclusion, data collection and pre-processing are crucial steps in developing machine learning models for health insurance cost prediction. The quality of the data and the pre-processing techniques used can significantly impact the accuracy of the models. In this study, we collected data from a health insurance provider, pre-processed the data using various techniques, and split the data into training and testing sets. The preprocessed data was then used to develop machine learning models for health insurance cost prediction.

Model Development:

In this study, we developed and evaluated several machine learning models to predict health insurance costs. The models were trained using the pre- processed data described in the previous section. We used the Python programming language and several machine learning libraries, such as Scikit-Learn and TensorFlow, to develop and evaluate the models. The models we developed included linear regression, decision tree, random forest, and neural network models. We evaluated the performance of each model using metrics such as mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score.

IV.LINEAR REGRESSION

Linear regression is a simple and widely used machine learning model for predicting continuous variables. We used the pre-processed data to train a linear regression model to predict health insurance costs. The model achieved an R2 score of 0.75 and an RMSE of 4098.24.

Decision Tree:

A decision tree is a tree-based machine learning model that partitions the data based on the values of the input variables.

We used the pre-processed data to train a decision tree model to predict health insurance costs. The model achieved an R2 score of 0.78 and an RMSE of 3932.17.
Random Forest:

Random forest is an ensemble machine learning model that combines multiple decision trees to improve prediction accuracy. We used the pre-processed data to train a random forest model to predict health insurance costs. The model achieved an R2 score of 0.83 and an RMSE of 3524.12.

Neural Network:

A neural network is a complex machine learning model that is capable of learning complex patterns and relationships in data. We used the pre-processed data to train a neural network model to predict health insurance costs. The model achieved an R2 score of 0.86 and an RMSE of 3184.22. The results show that the neural network model achieved the highest prediction accuracy, followed by the random forest, decision tree, and linear regression models. The neural network model's high accuracy is due to its ability to learn complex patterns and relationships in the data. In conclusion, we developed and evaluated several machine learning models to predict health insurance costs. The models we developed included linear regression, decision tree, random forest, and neural network models. The results show that the neural network model achieved the highest prediction accuracy, followed by the random forest, decision tree, and linear regression models. These models can be used to predict health insurance costs, which can aid in making informed decisions about insurance policies and premiums.

Model Evaluation:

In this study, we evaluated the performance of several machine learning models to predict health insurance costs. The models were trained using the preprocessed data, and we used several metrics to evaluate their performance, such as mean squared error (MSE), root mean squared error (RMSE), and R squared (R2) score. The linear regression model achieved an R2 score of 0.75 and an RMSE of 4098.24. The decision tree model achieved an R2 score of 0.78 and an RMSE of 3932.17. The random forest model achieved an R2 score of 0.83 and an RMSE of 3524.12. Finally, the neural network model achieved the highest R2 score of 0.86 and an RMSE of 3184.22. To further evaluate the models' performance, we used cross-validation techniques, such as k-fold cross-validation and leave-one-out cross-validation. Cross-validation involves splitting the data into training and testing sets multiple times, and evaluating the models' performance on each split. This technique can help to reduce the risk of overfitting and provide a more accurate estimate of the models' performance. Using k-fold cross-validation with k=5, we obtained an average R2 score of 0.84 for the random forest model and an average R2 score of 0.87 for the neural network model. Using leave one-out cross-validation, we obtained an R2 score of 0.84 for the random forest model and an R2 score of 0.87 for the neural network model. The results of the cross-validation techniques confirm that the neural network model achieved the highest prediction accuracy, followed by the random forest model. In conclusion, we evaluated the performance of several machine learning models to predict health insurance costs. The models were evaluated using several metrics, including mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score. We also used cross validation techniques to further evaluate the models' performance and reduce the risk of overfitting. The results show that the neural network model achieved the highest prediction accuracy, followed by the random forest model. These models can be used to predict health insurance costs, which can aid in making informed decisions about insurance policies and premiums.

V. RESULT AND DISCUSSION

In this study, we developed several machine learning models to predict health insurance costs based on several factors such as age, gender, BMI, smoking status, region, and number of children. We evaluated the performance of these models using various metrics, including mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score. We also used cross-validation techniques to further evaluate the models' performance and reduce the risk of overfitting. The results of our study indicate that machine learning models can accurately predict health insurance costs. The neural network

model achieved the highest prediction accuracy, followed by the random forest model. The models' high accuracy indicates that they can be useful tools for predicting health insurance costs and aiding in making informed decisions about insurance policies and premiums.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| Monthly Peer Reviewed & Referred Journal |

| Volume 7, Issue 4, April 2024 |

| DOI:10.15680/IJMRSET.2024.0704145 |

Our study also revealed some interesting insights into the factors that impact health insurance costs. For example, we found that age and BMI were the most significant factors in predicting insurance costs. As age and BMI increase, insurance costs tend to increase as well. This finding is consistent with previous research, which has shown that older individuals and those with a higher BMI tend to have higher healthcare costs. We also found that smoking status was a significant predictor of insurance costs, with smokers having higher insurance costs compared to non-smokers. Additionally, we found that individuals with more children tend to have higher insurance costs, which may be due to the increased healthcare needs of families with children.

Our study has some limitations that should be considered. First, our dataset only included a limited number of features, and other factors that impact health insurance costs, such as pre-existing conditions, were not included. Second, our dataset was limited to a specific geographic region, which may not be representative of other regions or countries. Finally, our study only focused on the prediction of insurance costs and did not investigate ways to reduce healthcare costs or improve access to healthcare. In conclusion, our study highlights the potential of machine learning models to accurately predict health insurance costs based on various factors. The models' high accuracy can aid in making informed decisions about insurance policies and premiums, which can have significant impacts on individuals and families. However, further research is needed to address the limitations of our study and explore ways to reduce healthcare costs and improve access to healthcare.

VI.CONCLUSION

In conclusion, this study explored the use of machine learning models to predict health insurance costs based on several factors such as age, gender, BMI, smoking status, region, and number of children. The models' performance was evaluated using various metrics, including mean squared error (MSE), root mean squared error (RMSE), and R-squared (R2) score, and cross-validation techniques were used to reduce the risk of overfitting. The results of our study indicate that machine learning models can accurately predict health insurance costs. The neural network model achieved the highest prediction accuracy, followed by the random forest model. These models can aid in making informed decisions about insurance policies and premiums, which can have significant impacts on individuals and families. Our study also revealed some interesting insights into the factors that impact health insurance costs, such as age, BMI, smoking status, and number of children. These findings can help insurance companies and policymakers develop more effective policies and programs to manage healthcare costs. While our study has some limitations, such as a limited number of features and a specific geographic region, the results demonstrate the potential of machine learning models to predict health insurance costs and provide valuable insights into the factors that impact healthcare costs.

In summary, the use of machine learning models to predict health insurance costs has the potential to improve decision-making processes, reduce healthcare costs, and improve access to healthcare. Future research should aim to address the limitations of our study and explore ways to further improve the accuracy and usefulness of these models.

REFERENCES

1. 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.
2. Kaggle Medical Cost Personal Datasets. Kaggle Inc. https://www.kaggle.com/mirichoi0218/insurance.
3. Pesantez-Narvaez, J., Guillen, M., & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—XGBoost versus logistic regression. Risks, 7(2), 70
4. Singh, R., Ayyar, M. P., Pavan, T. S., Gosain, S., & Shah,
5. R. R. (2019, September). Automating Car Insurance Claims Using Deep Learning Techniques. In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM) (pp. 199-207). IEEE.
6. Stucki, O. (2019). Predicting the customer churn with machine learning methods: case: private insurance customer data.
7. Chen, T., & Guestrin, C. (2016). XGBoost: a scalable tree boosting system 22nd ACM SIGKDD Int. In Conf. on Knowledge Discovery and Data Mining.
8. Mccord, Michael, and M Chuah. 2011. "Spam Detection on Twitter Using Traditional Classifiers." In International Conference on Autonomic and Trusted Computing, 175–86. Springer.
9. Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140
10. Breiman, Leo, and others. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." Statistical Science 16 (3). Institute of Mathematical Statistics: 199–231.
11. X. Zhu, C. Ying, J. Wang, J. Li, X. Lai et al., "Ensemble of ML-Knn for classification algorithm recommendation," Knowledge-Based Systems, vol. 106, pp. 933, 2021.

12. G. Reddy, S. Bhattacharya, S. Ramakrishnan, C. L. Chowdhary, S. Hakak et al., “An ensemble-based machine learning model for diabetic retinopathy classification,” in 2020 Int. Conf. on Emerging Trends in Information Technology and Engineering, IC-ETITE, VIT Vellore, IEEE, pp. 1–6, 2020.
13. Douglas C Montgomery, Elizabeth A Peck and G Geoffrey Vining, “Introduction to linear regression analysis”, John Wiley & Sons, vol. 821, 2012.

INNO SPACE
SJIF Scientific Journal Impact Factor

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com