Medical Insurance Cost Prediction

Sabarinath U S, Ashly Mathew

Abstract: ***This is a medical insurance cost prediction model that uses a linear regression algorithm to predict the medical insurance charges of a person based on the given data. To predict things that have never been so easy. In this project used to predict values that wonder how Insurance amount is normally charged. This is a medical insurance cost prediction model that uses a linear regression algorithm to predict the medical insurance charges of a person based on the given data. This project on predicting medical insurance costs can serve various purposes and address several needs that are Accurate Pricing Insurance companies need accurate predictions of medical insurance costs to set appropriate premiums for policyholders. Predictive models can analyse historical data and various factors such as age, gender, pre-existing conditions, lifestyle habits, and geographic location to estimate future healthcare expenses accurately. This Prediction model achieves three regression methods accuracy that the linear regression gets an accuracy of 74.45 %, whereas Ridge regression and Support Vector Regression gets 82.59% word-level state-of-the-art accuracy. The Medical Insurance Cost Prediction project, proposes a comprehensive approach to predict the medical cost, aiming to develop a robust and accurate system capable of predicting the accurate cost for a particular individual. Leveraging linear regression, our proposed system builds upon the successes of existing models like different types of regressions like linear regression, Ridge regression and Support Vector regression. We will put the Regression algorithm into practice and evaluate how it performs in comparison to the other three algorithms. By comparing the performance of these three methodologies, this project aims to identify the most effective approach for medical insurance cost prediction. Through rigorous evaluation and validation processes, the selected model will provide valuable insights for insurance companies, policymakers, and individuals seeking to optimize healthcare resource allocation and financial planning strategies.***

Keywords: ***Regression, Linear regression, Ridge Regression, Support vector Regression.***

I. INTRODUCTION

The Medical Insurance Cost Prediction project, proposes a comprehensive approach to predict the medical cost, aiming to develop a robust and accurate system capable of predicting the accurate cost for a particular individual. Leveraging linear regression, our proposed system builds upon the successes of existing models like different types of regressions like linear regression, Ridge regression and [9][10]. We will put the Regression algorithm into practice and evaluate how it performs in comparison to the other three regression algorithms.

The project will start with data pre-processing and collecting. The next phase will be to do exploratory data analysis in order to learn more about the data and pinpoint significant characteristics of the dataset. To choose the optimum strategy, several machine learning methods will be tested, including Linear regression, Ridge regression and Support Vector regression will take place in the last stage. The model will be evaluated based on its accuracy and performance, and the best model will be selected for future deployment. In the realm of healthcare economics, accurately predicting medical insurance costs is paramount for insurance providers, policy makers, and individuals alike. Leveraging advanced machine learning techniques, this project aims to develop robust models for predicting medical insurance costs using three distinct methodologies like linear regression, ridge regression, and support vector regression (SVR).

Linear regression serves as the cornerstone of predictive modelling, offering a straightforward approach to understanding the relationship between independent variables (such as age, BMI, smoking status, and geographical region) and the dependent variable of medical insurance costs. By fitting a linear equation to the data, linear regression provides insights into how each predictor influences insurance costs and facilitates interpretable results. Building upon the principles of linear regression, ridge regression introduces regularization to mitigate multi co-linearity and over fitting issues. By adding a penalty term to the regression equation, ridge regression shrinks the coefficients, thereby reducing the model's sensitivity to outliers and noise in the data. This regularization technique enhances the model's generalization performance and stability, particularly in scenarios where the number of predictors exceeds the number of observations.

In contrast, support vector regression (SVR) offers a nonlinear approach to modelling insurance costs by mapping the input features into a high-dimensional feature space. By identifying the optimal hyper plane that maximizes the margin between data points and the regression line, SVR accommodates nonlinear relationships between predictors and insurance costs. By comparing the performance of these three methodologies, this project aims to identify the most effective approach for medical insurance cost prediction. Through rigorous evaluation and validation processes, the selected model will provide valuable insights for insurance companies, policymakers, and individuals seeking to optimize healthcare resource allocation and financial planning strategies.

Manuscript received on 04 April 2024 | Revised Manuscript received on 29 May 2024 | Manuscript Accepted on 15 June 2024 | Manuscript published on 30 June 2024.

**Correspondence Author(s)*

Sabarinath U S*, Department of Computer Science, St. Albert's College, Kochi (Kerala), India. E-mail: sabarinath348@gmail.com, ORCID ID: 0009-0008-2187-9356

Ashly Mathew, Department of Computer Science, St. Albert's College, Kochi (Kerala), India. E-mail: ashly.worklife25@gmail.com

II. LITERATURE REVIEW

[1] [6] [7] Orji, Ugochukwu & Ukwandu, Elochukwu. (2023).Here the Predictive modelling in healthcare continues to be an active actuarial research topic as more insurance companies aim to maximize the potential of Machine Learning (ML) approaches to increase their productivity and efficiency. In this paper, the authors deployed three regression-based ensemble ML models that combine variations of decision trees through Extreme Gradient Boosting (XGBoost), Gradient-boosting Machine (GBM were deployed to discover and explain the key determinant factors that influence medical insurance premium prices in the dataset. The dataset used comprised 986 records and is publicly available in the KAGGLE repository. The models were evaluated using four performance evaluation metrics, including R-squared (R2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). The results show that all models produced impressive outcomes; however, the XGBoost model achieved a better overall performance although it also expanded more computational resources, while the RF model recorded a lesser prediction error and consumed far fewer computing resources than the XGBoost model. It is the aim of the authors that the contributions of this study will help policymakers, insurers, and potential medical insurance buyers in their decision-making process for selecting the right policies that meet their specific needs.

[2] [8] F. Amato, G. Cozzolino, A. Mazzeo and S. Romano, a Semantic System for Diagnoses Suggestion and Clinical Record Management. The increasing average life expectancy is increasing the number of healthy elderly people and consequently the incidence of chronic diseases with a dramatic increase of healthcare costs. The main outcome applied is to the existing healthcare systems can increase their efficiency, improve quality of life and reduce costs providing alternatives to the traditional disease management and allowing disease prevention. Several research efforts are devoted to provide innovative and not-intrusive e-Health systems. In this work we present a methodology for diagnosis decision supporting based on big data analysis of available medical data. This methodology is aimed to assist medical personnel in the patient's diagnosis making. [3] C. Hao, J. Wang, W. Xu and Y. Xiao, "Prediction-Based Portfolio Selection Model Using Support Vector Machines," In this paper, the rate of the returns is predicted using AR-MRNN and SVM and then the prediction-based portfolio selection model using SVM and the prediction-based portfolio selection model using AR-MRNN are proposed. Compared with the performance of the prediction of the AR-MRNN predictor and the SVM predictor, we found that the accuracy of the SVM is superior to the AR-MRNN. Compared with the performance of the prediction-based portfolio selection model using SVM and using AR-MRNN with the mean-variance portfolio selection model, we found that the former is superior to the latter. Meanwhile, we also proved that the more accuracy of the prediction achieved, the higher the rate of the returns.

[4] Panay, Belisario & Baloian, Nelson & Pino, José & Peñafiel, Sergio & Sanson, Horacio & Bersano-Méndez, Nicolás. (2019). Predicting Health Care Costs Using Evidence Regression. Proceedings. People's health care cost prediction is nowadays a valuable tool to improve accountability in health care. In this work, we study if an interpretable method can reach the performance of black-box methods for the problem of predicting health care costs. We present an interpretable regression method based on the Dempster-Shafer theory, using the Evidence Regression model and a discount function based on the contribution of each dimension. Optimal parameters are learned using gradient descent. We used health to test our method, which includes medical checkups, exam results, and billing information from 2016 to 2017. We compared our model to an Artificial Neural Network and Gradient Boosting method. Our results showed that our transparent model outperforms the Artificial Neural Network and Gradient Boosting with an R 2 of 0 .44.

[5] Taloba AI, Abd El-Aziz RM, Alshanbari HM, El-Bagoury AH. Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning. J Health c Eng. 2022. Medical costs are one of the most common recurring expenses in a person's life. Based on different research studies, BMI, ageing, smoking, and other factors are all related to greater personal medical care costs. The estimates of the expenditures of health care related to obesity are needed to help create cost-effective obesity prevention strategies. Obesity prevention at a young age is a top concern in global health, clinical practice, and public health. To avoid these restrictions, genetic variants are employed as instrumental variables in this research. Using statistics from public huge datasets, the impact of body mass index (BMI) on overall healthcare expenses is predicted. A multiview learning architecture can be used to leverage BMI information in records, including diagnostic texts, diagnostic IDs, and patient traits. In this system model, linear regression analysis, naive Bayes classifier, and random forest algorithms were compared using a business analytic method that applied statistical and machine-learning approaches.

III. METHODS

Data collection is a critical aspect of the Medical Insurance Cost Prediction project as it involves obtaining ad diverse dataset for continuous data. The dataset used for this project was medical insurance cost prediction Type Dataset which consisted of a csv file which had over 7 columns namely age, sex, bmi, children, smoker, region, and charges since it is a supervised learning. The file consists of over 1338 entries or rows. One of the significant challenges in data collection is obtaining a representative sample of the target population The data collection process should also consider factors such as age, sex, bmi, children, smoker and region, as these can affect cost prediction and personality type. Pre-processing is a critical step in the Medical Insurance Cost Prediction project, as it involves cleaning and preparing the dataset for machine learning model training.

This step is essential to ensure data quality, improve model accuracy, and reduce the risk of over fitting. The first step in pre-processing is data cleaning, which involves removing irrelevant or noisy data that could affect model performance. In the case of this project, this could involve removing duplicates, incomplete or irrelevant communication samples, or samples from individuals with invalid or inconsistent Medical cost .The next is data normalization, which involves transforming the data to a standardized format. This step is essential for data comparison and reduces the risk of bias caused by differences in data formatting. When it comes to Medical Insurance Cost Prediction project, data normalization could involve standardizing communication samples to a particular language, format, or length. After data normalization, feature extraction is performed to identify relevant features that can be used to predict an individual's MIC personality type. In the case of the MICP this could involve extracting linguistic features, such as Age, Sex, Body mass index, is the person a smoker or not, the region where he/she lives etc. The Feature Engineering step is particularly important in cases where the extracted features may not be sufficient to predict the target variable accurately. Here, the feature engineering could involve creating new features such as the use of comparing the cost or is it higher or lower for the person with a certain income etc. Once the features have been identified, the next step is feature selection, which involves identifying the most important features for model training. Here the main step is that, we are providing categorical features as the input data to the machine that too identify from and predict the output, so here comes the main step that is converting the categorical features to numerical values. In this project we have split the dataset into 70-30, 70% being the training set and 30% being the testing set. There are several evaluation metrics that can be used to evaluate the performance model. Most used evaluation metrics is Accuracy score.

IV. RESULTS AND DISCUSSION

A. Results

In this project to develop a medical insurance cost prediction system using machine learning techniques, has been a significant endeavour with far-reaching implications for accessibility also throughout the development process, several key aspects have emerged that highlight the project's importance and potential impact. One of the primary motivations behind this project was to improve the accuracy and to implement the system with different regression. Through meticulous data collection, annotation, and model training procedures, we achieved an impressive accuracy of 74.45% for linear regression and 82.59% on our dataset with the Ridge & SVR regressions. This high level of accuracy underscores the effectiveness and reliability of the developed system in predicting the medical insurance cost of any individual. Comparative analyses against baseline methods further validated the superiority of our approach in terms of accuracy and efficiency. Moreover, qualitative analysis of model predictions and errors provided valuable insights into the system's strengths and weaknesses, guiding future improvements and optimizations.

Table 1. Accuracy Comparison for Each Model

Model Name	Accuracy (%)
Linear Regression	74.45
Ridge Regression	82.59
Support Vector Machine	82.59

B. Discussion

1. Implications: Implementing accurate medical insurance cost prediction models can revolutionize healthcare planning, insurance pricing, and resource allocation. By leveraging advanced regression techniques like linear regression, ridge regression, and support vector regression, stakeholders can optimize financial strategies, enhance risk management, and ensure equitable access to healthcare services.
2. Real-World Applications: The practical applicability of the Cost prediction system extends beyond its technical achievements, with potential applications in assistive technology, individual satisfaction, security, and beyond this also gets personalized premium estimation for policyholders, optimizing healthcare budget allocations for governments and insurance companies, and identifying high-risk patient populations for proactive intervention and cost containment strategies, ultimately improving healthcare affordability, accessibility, and efficiency.
3. Limitations: The potential inaccuracies due to simplified modelling assumptions, such as linear relationships between predictors and insurance costs, inability to capture all relevant factors influencing medical expenses, and reliance on historical data which may not fully represent future trends or unforeseen changes in healthcare dynamics.. Additionally, the system may encounter difficulties in interpreting non-standard or inaccurate information, null values, or languages not adequately represented in the training data.
4. Future Directions: Continued research and development efforts are essential to further enhance the accuracy, robustness, and usability of the prediction system. Future directions may include exploring advanced deep learning architectures, multimodal fusion techniques, and domain adaptation strategies to improve performance across diverse contexts. It could involve incorporating more advanced machine learning techniques like ensemble methods or deep learning to capture nonlinear relationships and interactions among predictors, integrating real-time data streams for dynamic cost predictions, and leveraging big data analytics for more granular insights into healthcare cost drivers and trends.

V. CONCLUSION

The project to develop a medical insurance cost prediction system using machine learning techniques has been a significant endeavour with far-reaching implications for accessibility and communication.

Throughout the development process, several key aspects have emerged that highlight the project's importance and potential impact. One of the primary motivations behind this project was to improve the accuracy and to implement the system with different regression. By leveraging different regression algorithms and techniques, we have been able to develop a system that can accurately predict the medical insurance cost of an individual with the provided insights. Additionally, the system holds promise in business settings and also for analytical purposes, where it can be used to predict the medical insurance cost or assist individuals with the provided details. It can also serve as a valuable tool for healthcare interpreters, providing them with a means to quickly and accurately calculate the particular cost for a particular individual. Moreover there are several avenues for further improvement and expansion of the system. This includes enhancing the accuracy, adding more details about the individual so that the accurate cost can be fetched and integrating the system into various devices and platforms for widespread use. Overall, the medical insurance cost prediction system has the potential to have a profound impact on society by promoting inclusivity, gaining trusty claims from the individuals and advancing technology for healthcare and social good. It represents a significant step forward in leveraging technology to address the needs of individuals with that every individual can compute their own medical cost that can pave the way for the individual to keep aware and also to make decisions regarding the healthcare.

DECLARATION STATEMENT

Funding	No, I did not receive
Conflicts of Interest	No conflicts of interest to the best of our knowledge.
Ethical Approval and Consent to Participate	No, the article does not require ethical approval and consent to participate with evidence.
Availability of Data and Material	Yes, It is relevant. The dataset used for this project was a CSV format file which includes data continuous values in tabular format is an openly available corpus containing numerous attributes thousands of data.
Authors Contributions	Each author has made an independent contribution to the article. The individual contributions of each author are presented below for clarity and transparency. Sabarinath U S is the main contributor and Ashly Mathew is the project guide.

REFERENCES

1. Orji, Ugochukwu & Ukwandu, Elochukwu. (2023). Machine Learning For An Explainable Cost Prediction of Medical Insurance. Machine Learning with Applications. 15. 100516. 10.1016/j.mlwa.2023.100516. https://doi.org/10.1016/j.mlwa.2023.100516
2. Amato, Flora & Cozzolino, Giovanni & Mazzeo, Antonino & Romano, Sara. (2016). A Semantic System for Diagnoses Suggestion and Clinical Record Management. 133-138. 10.1109/WAINA.2016.135. Matthews, T. F. Cootes, J. A. Bangham, S. Cox and R. Harvey, "Extraction of visual features for lipreading," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, pp. 198-213, Feb. 2002, doi: 10.1109/34.982900. https://doi.org/10.1109/34.982900
3. Hao, Cuiyan & Wang, Jiaqian & Xu, Wei & Xiao, Yuan. (2014). Prediction-Based Portfolio Selection Model Using Support Vector Machines. Proceedings - 2013 6th International Conference on Business Intelligence and Financial Engineering, BIFE 2013. 567-571. 10.1109/BIFE.2013.118. https://doi.org/10.1109/BIFE.2013.118
4. Panay, Belisario, Nelson Baloian, José A. Pino, Sergio Peñafiel, Horacio Sanson, and Nicolas Bersano. 2019. "Predicting Health Care Costs Using Evidence Regression" *Proceedings* 31, no. 1: 74. https://doi.org/10.3390/proceedings2019031074
5. Taloba AI, Abd El-Aziz RM, Alshanbari HM, El-Bagoury AH. Estimation and Prediction of Hospitalization and Medical Care Costs Using Regression in Machine Learning. J Healthc Eng. 2022 Mar 2;2022:7969220. doi: 10.1155/2022/7969220. PMID: 35281545; PMCID: PMC8906954. https://doi.org/10.1155/2022/7969220
6. Singh*, A., & Ramkumar, K. R. (2019). Validation of Machine Learning Models for Health Insurance Risks Assessment. In International Journal of Engineering and Advanced Technology (Vol. 9, Issue 1, pp. 4247–4256). https://doi.org/10.35940/ijeat.a1670.109119
7. hanafy, M., & Mahmoud, O. M. A. (2021). Predict Health Insurance Cost by using Machine Learning and DNN Regression Models. In International Journal of Innovative Technology and Exploring Engineering (Vol. 10, Issue 3, pp. 137–143). https://doi.org/10.35940/ijitee.c8364.0110321
8. Jani*, K. K., Srivastava, S., & Srivastava, R. (2019). Computer-Aided Diagnosis for Capsule Endoscopy: From Inception to Future. In International Journal of Recent Technology and Engineering (IJRTE) (Vol. 8, Issue 4, pp. 12261–12273).. https://doi.org/10.35940/ijrted8094.118419
9. Sistla, S. (2022). Predicting Diabetes u sing SVM Implemented by Machine Learning. In International Journal of Soft Computing and Engineering (Vol. 12, Issue 2, pp. 16–18).. https://doi.org/10.35940/ijsce.b3557.0512222
10. Muthukrishnan, Dr. R., & Prakash, N. U. (2023). Validate Model Endorsed for Support Vector Machine Alignment with Kernel Function and Depth Concept to Get Superlative Accurateness. In International Journal of Basic Sciences and Applied Computing (Vol. 9, Issue 7, pp. 1–5). https://doi.org/10.35940/ijbsac.g0486.039723

AUTHORS PROFILE

Sabarinath U S, currently pursuing Master of Science in Computer Science from the prestigious St. Albert's College (Autonomous), Ernakulam. Prior to this he had completed his Bachelor of Science degree in Mathematics & Computer Science from The Cochin College,Cochin, Ernakulam. His area of interests includes prominent fields like IoT, Computing, Networking, Designing. He is given attention to details as well as he is able to think outside the box, he loves to solve problems and has been keenly observing the latest technology. When he is not studying or working on new projects, he enjoys to play music instruments, explores the nature. He is an active member of the Computer Science community and coordinates in various events conducted.

Ms. Ashly Mathew is an experienced Assistant Professor with a blend of industry and academic expertise. She currently works at St. Albert's College (Autonomous) in Ernakulam, where she contributes to the academic community through teaching, research, and mentorship. She completed her undergraduate studies at BVM Holy Cross College, Pala, before pursuing Master of Computer Applications (MCA) from Santhigiri College of Computer Science, Vazhithala. Prior to entering academia, she acquired 2.3 years of valuable experience in the tech industry, which has given her practical insights to bring into the classroom. With 6.2 years of teaching experience, she has established herself as an educator.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of the Lattice Science Publication (LSP)/ journal and/ or the editor(s). The Lattice Science Publication (LSP)/ journal and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.