

Healthcare Cost Patterns and Prediction: Investigating Personal Datasets using Data Analytics

Md Aminul Islam^{1,*[0000-0002-2535-6519]}, Anindya Nag^{2[0000-0001-6518-8233]},

Pretam Chandra³, SM Firoz Ahmed Fahim⁴,

Md Mozammel Hoque⁵

¹ School of Engineering, Computing, and Mathematics, Oxford Brookes University, Oxford, UK

² Computer Science & Engineering Discipline, Khulna University, Khulna -9208, Bangladesh

³Department of Computer Science & Engineering, Adamas University, Kolkata -700126, India

⁴Department of Computer Science & Engineering, American International University Bangladesh. Dhaka, Bangladesh

⁵Gannon University, United States

talukder.rana.13@gmail.com, anindyanag@ieee.org,
pretam591@gmail.com, firozfahimm@gmail.com,
mozammel2030@gmail.com

Abstract. The present study introduces a health insurance prediction system that leverages machine learning methodologies. In contemporary times, there has been a notable increase in endeavors focused on tackling this matter since the significance of health insurance as a research topic has markedly escalated following the pandemic. The dataset employed in this research comprises 1338 observations 7 columns and corresponds to individual medical expenditures in the United States, available at the Kaggle platform. The dataset encompasses a variety of variables utilized in the prediction of insurance prices, including age, gender, BMI, smoking status, and number of children. The researchers used machine learning models, including neural networks, XAI, and auto modeling, to determine the correlation between pricing and the attributes. The training process involved partitioning the dataset into an 80-20 ratio for training and evaluation. Consequently, the system achieved an accuracy rate of 97% by Gradient Boosting, but we corrected it to 92% by Gradient Boosting Regressor by encoding and hyper-tuning. Also, among predictive machine learning models, Random Forest had the best accuracy i.e., of 83.44%.

Keywords: Machine Learning, Insurance, Medical Cost, Prediction.

1 Introduction

In today's world, many of us require medical coverage. Annual premiums differ depending on the type of medical care. Medical expenditures are difficult to estimate due to the wide range of health issues. Some conditions are more common among

certain demographics. Smokers are more likely to develop lung cancer, whereas obese people are more likely to develop cardiac problems [1]. Insurers invest significant time and money in creating algorithms that accurately forecast medical expenditures. As data scientists, we may be given real-world patient data in the form of (insurance.csv) with seven columns: numerical variables (age, BMI, children, and medical cost) and categorical variables (sex, smoker, region). Utilizing customer information is critical for many businesses [2]. Regarding an insurance company, client traits such as those stated below can be crucial when making decisions. As a result, the capacity to analyze and extract value from such data can be invaluable.

This essential contribution of this work is outlined as follows:

- To develop a prediction model for estimating medical costs using supervised learning approaches.
- Finding relevant variables and evaluating their significance in forecasting medical expenses, which are made possible through data exploration, correlation analysis, and regression modeling.
- To delve thoroughly into the data to unearth some useful information.
- Thus, finding analytical results will provide essential insights into the factors influencing medical expenses and the effectiveness of the constructed models.

The paper is structured into five distinct sections. Section 2 provides a comprehensive literature review, while Section 3 outlines the research methods and materials employed. Section 4 presents the analysis and results derived from the study, and finally, Section 5 concludes the paper.

2 Literature Survey

The following investigation expands upon comparing various machine learning models for Bank Fraud detection as presented by multiple researchers.

The utilization of data analytics and machine learning techniques in predicting healthcare costs has garnered significant attention in recent research. This literature survey delves into key studies that explore healthcare cost patterns and prediction utilizing personal datasets.

Abdelmoula et al., (2021) [1] presented a machine learning-based prediction tool for hospitalization costs. The study leveraged algorithms to forecast hospitalization costs, offering insights into cost estimation accuracy. Baro et al. (2022) [2] focused on predicting hospitalization using health insurance data. Their research utilized data-driven methodologies to anticipate hospitalization events, contributing to enhanced risk assessment. Subroto et al., (2022) [3] employed tree-based algorithms to predict informal workers' willingness to pay for national health insurance after tele-collection. Their approach provided valuable insights into determining individuals' preferences regarding health insurance participation. T et al. (2023) [4] conducted medical insurance cost analysis and prediction using machine learning. Their work facilitated a better understanding and prediction of medical insurance costs through data analytics. Vijayalakshmi et al., (2023) [5] implemented a medical insurance price prediction

system using regression algorithms, enhancing the accuracy of cost estimation. Dong et al., (2021) [6] enhanced interpretability using decision tree regression with an insurance dataset, contributing to a clearer understanding of the underlying cost patterns. Bora et al., (2022) [7] explored the interpretation of machine learning models using XAI in health insurance datasets, promoting transparency and insights into predictive models. Bhatia et al., (2022) [8] investigated health insurance cost prediction using machine learning techniques, striving for accurate estimation of insurance costs. D et al., (2022) [9] utilized machine learning algorithms to predict health insurance costs, aiding in cost estimation for healthcare services. Jyothisna et al., (2022) [10] employed XGboost regressor to predict health insurance premiums, contributing to improved premium estimation accuracy. Sailaja et al., (2021) [11] proposed a hybrid regression model for medical insurance cost prediction and recommendation, offering a comprehensive approach to cost estimation. Mary Chittilappilly et al., (2023) [12] conducted a comparative analysis of optimizing medical insurance prediction using genetic algorithms and other machine learning methods, striving for improved prediction accuracy. Wedanage et al., (2021) [13] focused on forecasting healthcare costs in Australia using health insurance claims data, contributing to informed cost predictions. Panda et al., (2022) [14] explored health insurance cost prediction using regression models, enhancing the accuracy of cost estimation. Aggarwal et al., (2022) [15] predicted health insurance amounts using supervised learning techniques, aiding in precise amount estimation. Alzoubi et al. (2022) [16] analyzed cost prediction in medical insurance using modern regression models, contributing to an improved understanding of cost patterns. B et al. (2022) [17] explored end-to-end encryption and prediction of medical insurance costs, combining data privacy with accurate prediction techniques. Xiaoqun et al., (2022) [18] proposed an improved K-means clustering model based on a support vector machine for health insurance cost prediction, enhancing the accuracy of prediction techniques. Harale et al., (2022) [19] empirically analyzed predictive models for insurance claim classification, aiding in classification and prediction of healthcare claims. Kikuchi et al., (2022) [20] investigated disease prediction from officially anonymized medical and healthcare big data, contributing to early disease identification. Dutta et al., (2021) [21] developed a data mining-based approach for modeling health insurance claims, offering insights into efficient claim processing and prediction.

In summary, the presented studies collectively emphasize the importance of data analytics and machine learning in understanding healthcare cost patterns and predicting costs accurately. These works contribute to informed decision-making and improved healthcare cost management through data-driven insights. This literature survey highlights the diverse range of approaches and methodologies employed in healthcare cost prediction and analysis, contributing to the understanding of cost patterns and the enhancement of healthcare management strategies.

Table 1. Summary of related work

Ref.	Data set	Best Model	Accuracy
[6]	Insurance Dataset	Decision Tree Regression	81%
[10]	Health Insurance Data	XGBoost Regressor	87%
[12]	Medical Insurance Data	TPOT generated: Gradient Boosting Regressor with Lasso Lars CV	87.4%
[16]	Medical Insurance Data	Gradient Boosting Regressor	87.8%
[21]	Health Insurance Claims Data	Random Forest Regressor	86.2%

3 Methodology and Data Analytics

The authors can divide this paper into four main stages: Dataset Description, Data Pre-processing, EDA, Model preparation, and training. Figure 1 depicts the flow of data in this entire process.

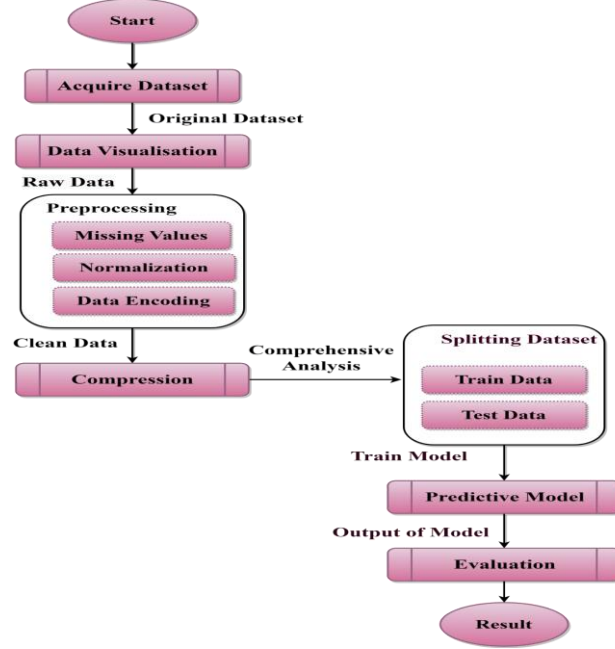


Fig. 1. Proposed Model

3.1 Dataset Description

The dataset under investigation contains information regarding medical costs incurred by individuals meeting specific criteria. This comprises 1338 observations alongside 7 columns, corresponding to individual medical expenditures in the United States as available on the Kaggle platform. It forms a part of a broader statistical learning project focusing on insurance analysis. The dataset attributes are as follows:

Age: The age of the primary beneficiary for the insurance coverage. Gender: The gender of the insurance holder is categorized as female or male. Body Mass Index (BMI): A numerical measure of an individual's body weight relative to height. A healthy BMI range is typically between 18.5 and 24.9 kg/m². Children: The number of dependents or children covered by health insurance. Smoker: Indicates whether the individual is a smoker or not. Region: The geographical region within the United States where the insurance recipient resides, categorized into northeast, southeast, southwest, and northwest. Medical Cost: The monetary value of medical expenses billed by health insurance companies for everyone [21].

3.2 Data Pre-processing

The given dataset is labeled for all variables, and we must predict medical costs based on each variable. So, the input and output are labeled, which tends to be supervised learning [22]. Now, Classification and regression are connected to prediction. Classification involves identifying values or objects that pertain to a particular category.

ry. In contrast, the regression method predicts a response value based on a series of outcomes. Classification is preferred over regression when the model's output must identify the category to which each data point in a dataset belongs [23]. In the given case, the output label is medical cost, which would be a continuous value (neither discrete nor grouped), which means it must be a regression analysis, which is multi-linear regression as it has one dependent variable and more than 2 independent variables including numerical and categorical [24].

3.3 Exploratory Data Analysis

The EDA phase of the research process involves an in-depth exploration of the dataset to uncover patterns, relationships, and insights that can inform subsequent model preparation and analysis [25]. This stage is crucial as it helps understand the dataset's structure and characteristics. The Medical Insurance dataset, in our case, has some noticeable trends in the data. Let's uncover those and try to gain a deeper understanding using Python Plotting libraries and various kinds of plots:

Considering features of BMI and age, this line plot offers a clear and concise visual representation of how average BMI [26] changes across different age groups, potentially shedding light on age-related trends in health and lifestyle factors among the individuals represented in the dataset.



Fig. 2. Average BMI per age

Violin plots are utilized when observing the distribution of numerical data and are particularly useful when comparing the distributions of multiple groupings. Peaks, valleys, and tails of each group's density curve can be contrasted to identify areas of similarity and difference. Frequently, additional elements, such as box plot quartiles, are added to a violin plot to provide additional means of group comparison [27].

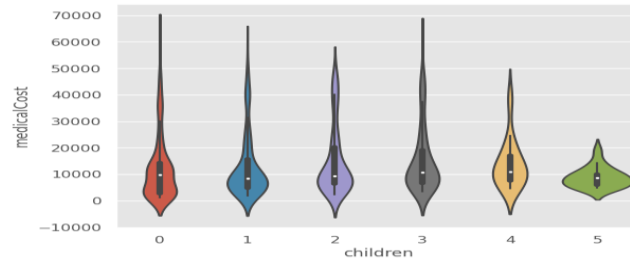


Fig. 3. Violin plot showing “Distribution of Medical Costs by Number of Children”

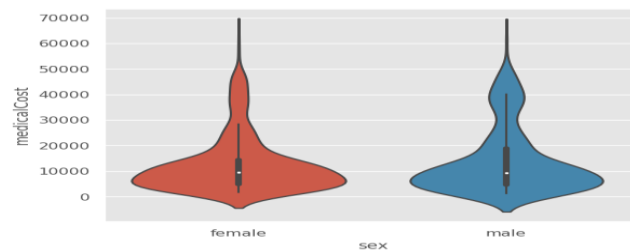


Fig. 4. Distribution of Medical Costs by Gender

The violin plot in figure 3 tells that the distribution of medical cost is more when the number of children is 4. The violin plots in figure 4 tells that the distribution of medical cost is almost the same and doesn't change with Gender.

The horizontal bar plot effectively communicates the average medical cost variations across different numbers of children, enabling quick and clear comparisons between categories. It provides valuable insights into the relationship between family size and medical expenses.

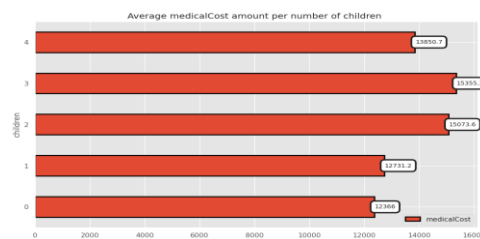


Fig. 5. Average medicalCost amount per number of children

From the plot, it's very clear that when the number of children's is 3, the average medicalCost is the most.

The joint plot is a compelling visualization to explore and understand the potential relationship between age and medical costs, providing valuable insights for further analysis and model development.

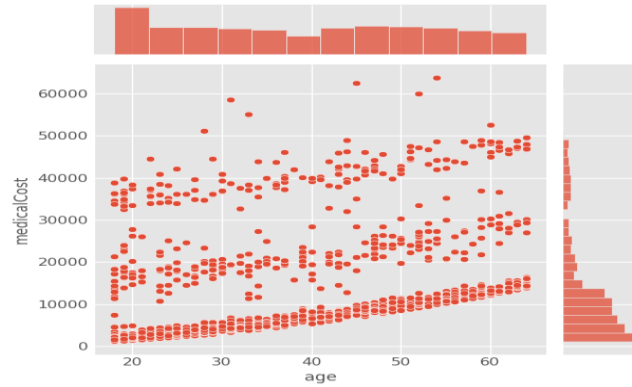


Fig. 6. Relationship Between Age and Medical Costs

The pie chart effectively communicates the distribution of total medical costs among different regions within the dataset, providing an overview of the regional contributions to the overall medical cost amount.

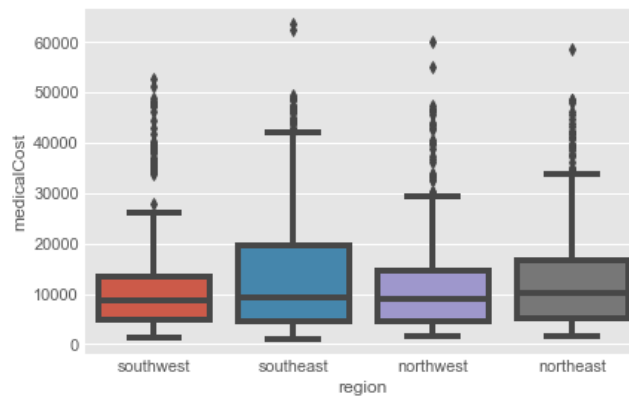


Fig. 7. Distribution of Total Medical Costs by Region

In general, men are spending more than women. The authors anticipated a massive gap, but there is only a small one which is clear from Figure 8.

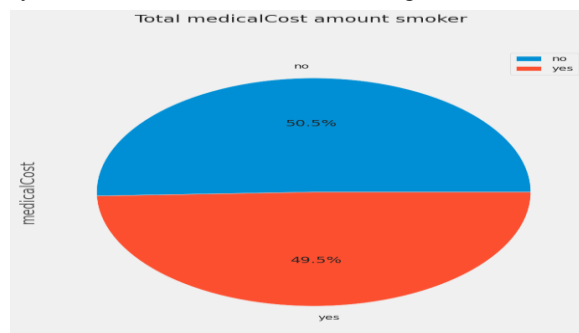


Fig. 8. Distribution of Total Medical Costs by Smoking Habits

The box plot effectively communicates the distribution of medical costs for different smoking habits, providing insights into potential differences in healthcare expenses between individuals with varying smoking habits.

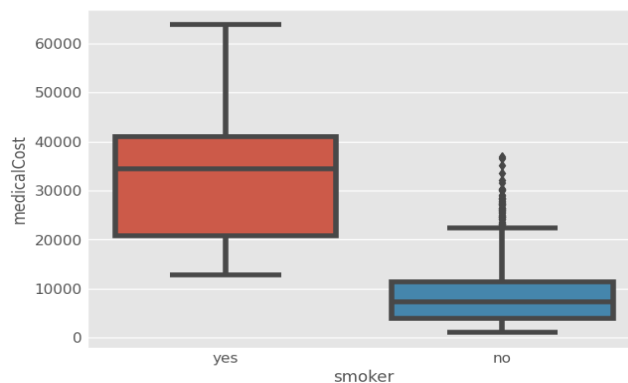


Fig. 9. Distribution of Medical Costs by Smoking Habits (Box Plot)

From figure 9, it's very clear that the person who is a "Smoker" has a significantly higher range of medical costs as the longer the box, the more dispersed is the data. The smaller, the less dispersed the data.

Now, let's see a 3D scatter plot where the "age" values will be along the x-axis, the "bmi" values will be along the y-axis, and the "medicalCost" values will be along the z-axis. The associated "sex" value will decide the colour of each point.

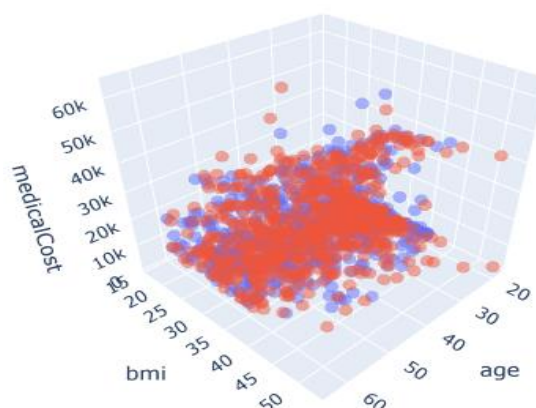


Fig. 10. 3D Scatter Plot of Age, BMI, and Medical Costs to determine Sex using plot color of red as male and blue as female.

This visualization is helpful to understand how age, BMI, and medical costs change between females. It may uncover patterns or correlations between these variables that can be useful for additional research or modelling.

3.4 Key Insights

Most of the participants are in their 20s. Men pay a more significant average rate for insurance than women do. In the study, there are more non-smokers than smokers. The Southeast has the highest insurance rates, followed by the Northeast, the Northwest, and the Southwest. A participant's insurance premium is higher than a member's insurance premium if they do not have any children. Only a slender correlation (corr 0.067) exists between the factors and the number of children.

4 Analysis and Result Discussion

The complex relationships between demographic, lifestyle, and regional factors and personal healthcare expenses are the focus of our investigation. We discovered essential trends through various exploratory data analysis techniques that shed light on the distribution and volatility of medical costs across multiple populations. A thorough analysis of these patterns offers a detailed comprehension of the variables influencing healthcare costs in the investigated group. Our visualizations highlight the value of investigating the data from many perspectives. For instance, the scatter plot matrix offered insights into pairwise correlations between variables, and the 3D scatter plots made possible by color-coded gender differentiation, allowing for the simultaneous assessment of age, BMI, and medical costs. The distribution of medical expenses across several categories is also illuminated by box plots and violin plots, highlighting potential deviations and outliers that demand additional research. Further, metrics like mean absolute error, mean squared error, and R-squared were used to quantitatively assess the model's performance.

4.1.1 Data Correlation

The resulting heatmap in Figure 11 provides a visual representation of the correlations between numerical variables in the dataset. Each cell in the heatmap is colored according to the strength and direction of the correlation.

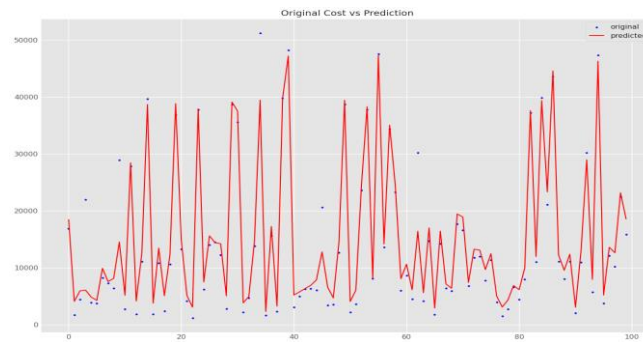


Fig. 11. Numerical Data Correlation

As we previously mentioned, the most significant link between numerical values is seen between medical cost and age (0.30). Also, the smoking status of a person has the most impact on the price. We have it, and that is the most essential element. The heatmap makes clear the three most crucial factors influencing medical costs: Smoking, age, and BMI. By taking categorical variables into account using one-hot encoding, we may create a somewhat different heatmap.

4.1.2 Gradient Boosting Regressor

GBR model is adequate with an accuracy of 89%. The plot in Figure 12 provide insights into how well the model predicts insurance costs.

**Fig. 12.** GBR Model accuracy plot

4.1.3 Random Forest Regressor

The authors used a metric named R-squared score. This measures the proportion of the variance in the dependent variable (y_{test}) that is predictable from the independent variables (X_{test}). It indicates how well the model's predictions match the actual target values. The R-squared score ranges between 0 and 1, where a higher value indicates a better fit of the model to the data. In our case, the score was 0.83.

The authors performed training and evaluation of different regression models and then compiled the results into a Data Frame for comparison. Table 2 describes the accuracy values.

Table 2. Accuracy values for 5 best Predictors

Sl. No.	Model	Accuracy
1	Linear Regression	0.9235671
2	Random Forest	0.9235671
3	Gradient Boosting	0.9703313
4	SVM	-0.053583

Gradient boosting regression has the highest accuracy, which is 92%. SVM shows negative results because of categorical values. These results are unacceptable, as we used a label encoder for absolute values.

4.1.4 Linear Regression using Conversion Method as Ordinal Encoding

Each unique category value is allocated an integer value in ordinal encoding. For instance, regions are denoted by 0, 1, 2, and 3. This is referred to as ordinal or integer encoding, and it is readily reversible. Often, integer values commencing at zero are used. For certain variables, ordinal encoding may suffice. There is a natural ordering between the integer values, and machine learning algorithms may be able to comprehend and exploit this ordering. Ordinal variables have a raw encoding. It imposes an ordinal relationship on categorical variables where no such relationship exists. This can pose issues, so a one-hot encoding may be an alternative.

4.1.5 Multiple Linear Regressor

Figure 13 provides a comprehensive overview of multiple linear regression, including data preprocessing, model fitting, prediction, and visualization of the results. The scatter plot with the regression line helps you assess the model's performance in predicting insurance charges.

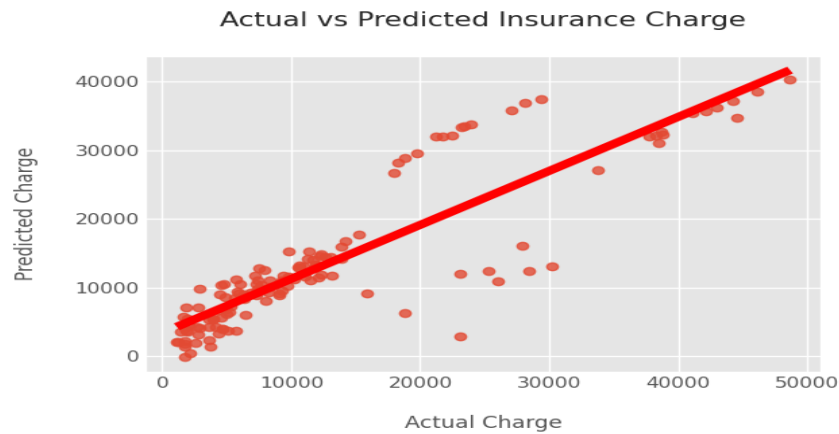


Fig. 13. Multiple Linear Regression

4.1.6 LightGBM (LGB) Model

The R-squared score for LGB model as observed was 0.85. Figure 15 provides insights into how well the LGB model's predictions match the true values. The scatter plot helps to visually assess the quality of the predictions, and the calculated metrics provide quantifiable performance indicators.

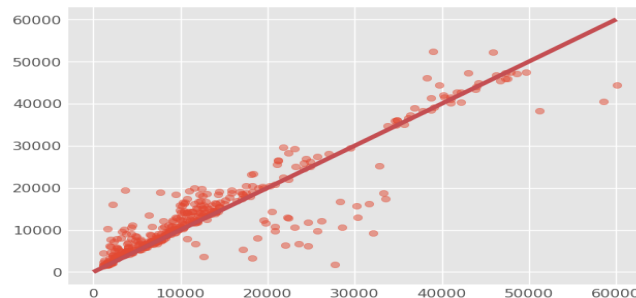


Fig. 14. LGB model's predictions

So, in the scatter plot created by `plt.scatter`, the x-axis represents the values in `y_test`, and the y-axis represents the values in `y_pred1`. Points falling close to this line indicate a strong positive correlation between `y_test` and `y_pred1`.

4.1.7 RNN

Figure 15 plots the training and validation losses plotted on the y-axis and the number of epochs plotted on the x-axis. The learning curve reveals how well the learning of the model is and whether it is overfitting or underfitting. The scatter plot in Figure 16 allows a visual assessment of how well the model's predictions align with the true values. Ideally, the points should fall close to a diagonal line, indicating a strong correlation between the true and predicted values.

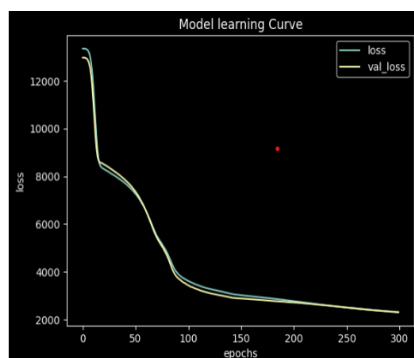


Fig. 15. RNN Model Learning Curve

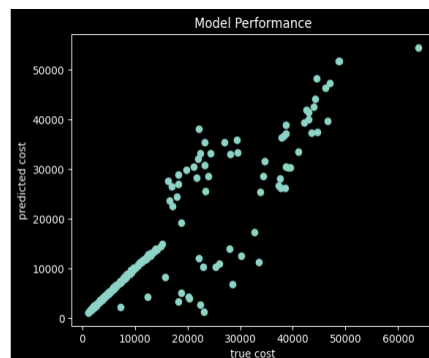


Fig. 16. RNN Model Performance

4.1.8 Predictions and Best Model

This plot in Figure 17 allows the visual comparison of the original and predicted insurance costs along the x-axis. It helps understand how well the model's predictions align with the actual values across the dataset.

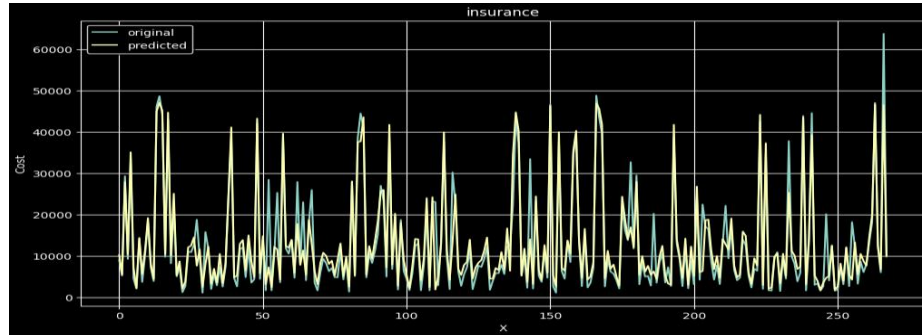


Fig. 17. Comparison of the original and predicted insurance costs

Finally, the authors used a method named “GridSearchCV” to search for the best hyperparameters for multiple regression algorithms and the method returned a data frame with the best scores and parameters for each algorithm. Figure 18 summarizes the accuracy values for the best 5 predictors.

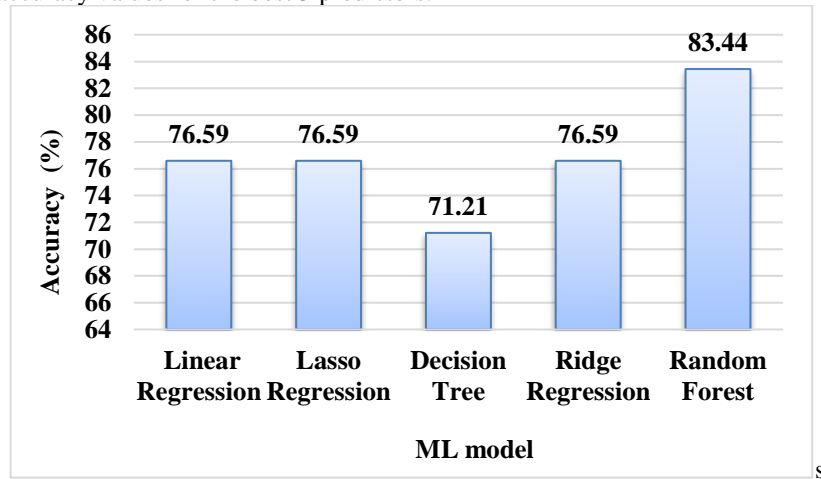


Fig. 18. Accuracy values for 5 best Predictors

The bar plot in Figure 18 makes it evident that the Random Forest model, with an accuracy of 83.44%, provides the most excellent match for estimating the costs related to medical insurance. Later, the GBR model with 92% accuracy was developed.

5 Conclusion

A predictive model was developed using a dataset consisting of seven variables to estimate the expenses associated with medical insurance. The features encompassed all the relevant attributes necessary for forecasting insurance costs. In the implementation, various regression methods were employed in Python. We analyzed using Linear Regression, Decision Tree Regression, Lazy Predict, Interpret ML, Lasso Regression, Ridge Regression, Random Forest Regression, ElasticNet Regression, KNN,

Support Vector Regression, K Nearest Neighbour Regression, and ANN Regression. The evaluation of model performance involved the utilization of seven metrics: MSE, RMSE, accuracy, MAE, MAPE, R-squared (R²), Adjusted R-Squared (Adj. R²), and Explained Variance Score (EVS). Initially, we achieved accuracy at 97%, but we identified it as faulty for the label encoder. Later, we had the GBR model with 92% accuracy. The model can contribute to societal welfare for faster policymaking and decision-making in medical insurance cost prediction. Future researchers can build real-life projects in action because of this study.

References

1. Abdelmoula, B., Torjmen, M., & Abdelmoula, N. B. (2021). Machine learning based prediction tool of hospitalization cost. 2021 22nd International Arab Conference on Information Technology (ACIT). <https://doi.org/10.1109/acit53391.2021.9677110>
2. Baro, E. F., Oliveira, L. S., & de Souza Britto Junior, A. (2022). Predicting hospitalization from health insurance data. 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC). <https://doi.org/10.1109/smc53654.2022.9945601>
3. Mansur, R., & Subroto, A. (2022). Using tree-based algorithm to predict informal workers' willingness to pay national health insurance after Tele-Collection. 2022 10th International Conference on Information and Communication Technology (ICoICT). <https://doi.org/10.1109/icoict55009.2022.9914901>
4. T, Thejeshwar., T, S. Harsha., V, V. Krishna., & R, Kaladevi. (2023). Medical Insurance Cost Analysis and prediction using machine learning. 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA). <https://doi.org/10.1109/icidca56705.2023.10100057>
5. Vijayalakshmi, V., Selvakumar, A., & Panimalar, K. (2023). Implementation of medical insurance price prediction system using regression algorithms. 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT). <https://doi.org/10.1109/icssit55814.2023.10060926>
6. Dong, S., & Fei, D. (2021). Improve the interpretability by decision tree regression: Exemplified by an insurance dataset. 2021 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI). <https://doi.org/10.1109/icceai52939.2021.00065>
7. Bora, A., Sah, R., Singh, A., Sharma, D., & Ranjan, R. K. (2022). Interpretation of machine learning models using XAI - A Study on Health Insurance Dataset. 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO). <https://doi.org/10.1109/icrito56286.2022.9964649>
8. Bhatia, K., Gill, S. S., Kamboj, N., Kumar, M., & Bhatia, R. K. (2022). Health insurance cost prediction using machine learning. 2022 3rd International Conference for Emerging Technology (INCET). <https://doi.org/10.1109/incet54531.2022.9824201>
9. D, R., K, M. S., & J, D. (2022). Health insurance cost prediction using machine learning algorithms. 2022 International Conference on Edge Computing and Applications (ICECAA). <https://doi.org/10.1109/icecaa55415.2022.9936153>
10. Jyothsna, C., Srinivas, K., Bhargavi, B., Sravanth, A. E., Kumar, A. T., & Kumar, J. N. V. R. S. (2022). Health Insurance Premium prediction using xgboost regressor. 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC). <https://doi.org/10.1109/icaaic53929.2022.9793258>

11. Sailaja, N. V., Karakavalasa, M., Katkam, M., M, D., M, S., & Vasundhara, D. N. (2021). Hybrid regression model for medical insurance cost prediction and recommendation. 2021 IEEE International Conference on Intelligent Systems, Smart and Green Technologies (ICISSGT). <https://doi.org/10.1109/icissgt52025.2021.00029>
12. Mary Chittilappilly, R., Suresh, S., & Shanmugam, S. (2023). A comparative analysis of optimizing medical insurance prediction using genetic algorithm and other machine learning algorithms. 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI). <https://doi.org/10.1109/accai58221.2023.10199979>
13. Wedanage, K. W., Wickramasuriya, R., Win, K. T., & Perez, P. (2021). Forecasting healthcare cost in Australia using health insurance claims data. 2021 17th International Computer Engineering Conference (ICENCO). <https://doi.org/10.1109/icenco49852.2021.9698885>
14. Panda, S., Purkayastha, B., Das, D., Chakraborty, M., & Biswas, S. K. (2022). Health insurance cost prediction using regression models. 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON). <https://doi.org/10.1109/com-it-con54601.2022.9850653>
15. Aggarwal, S., & Anmol. (2022). Health Insurance Amount Prediction using supervised learning. 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS). <https://doi.org/10.1109/ictacs56270.2022.9988256>
16. Alzoubi, H. M., Sahawneh, N., AlHamad, A. Q., Malik, U., Majid, A., & Atta, A. (2022, October). Analysis Of Cost Prediction In Medical Insurance Using Modern Regression Models. In 2022 International Conference on Cyber Resilience (ICCR) (pp. 1-10). IEEE.
17. B, J., Ghosh, A., & Kumar R, J. A. (2022). End-to-end encryption and prediction of Medical Insurance Cost. 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI). <https://doi.org/10.1109/icoei53556.2022.9777238>
18. Xiaoqun, L., & Run, L. (2022). An improved K-means clustering model based on support vector machine for health insurance cost prediction. 2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI). <https://doi.org/10.1109/icetci55101.2022.9832085>
19. Harale, A., Dubey, Y., Gupta, V., Motghare, A., Chakole, M., & Pathade, A. (2022). Empirical analysis of predictive models for insurance claim classification. 2022 International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS). <https://doi.org/10.1109/icetems56252.2022.10093335>
20. Kikuchi, H., Ito, S., Ikegami, K., & Shindo, S. (2022). Diseases prediction from officially anonymized medical and healthcare big data. 2022 IEEE International Conference on Big Data (Big Data). <https://doi.org/10.1109/bigdata55660.2022.10075547>
21. Dutta, K., Chandra, S., Gourisaria, M. K., & Harshvardhan, G. M. (2021, April). A data mining based target regression-oriented approach to modelling of health insurance claims. In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1168-1175). IEEE.
22. JRIese, F.M. and Keller, S., 2020. Supervised, semi-supervised, and unsupervised learning for hyperspectral regression. In *Hyperspectral Image Analysis: Advances in Machine Learning and Signal Processing* (pp. 187-232). Cham: Springer International Publishing.
23. Rukhsar, L., Bangyal, W.H., Nisar, K. and Nisar, S., 2022. Prediction of insurance fraud detection using machine learning algorithms. *Mehran University Research Journal of Engineering & Technology*, 41(1), pp.33-40.
24. Gogtay, N.J., Deshpande, S.P. and Thatte, U.M., 2017. Principles of regression analysis. *J. Assoc. Physicians India*, 65(48), pp.48-52.

25. Islam, M. A., Nag, Anindya, et. al., (2023). Utilization of Encoding, Early Stopping, Hyper Parameter Tuning, and Machine Learning Models for Bank Fraud Detection. 10.36227/techrxiv.24319243.
26. Islam, M. A., et. al., (2023). Socioeconomic Factors Influencing Breastfeeding Duration in Bangladesh: An Analysis of the Bangladesh Demographic and Health Survey. BMJ Open. 10.1101/2023.10.15.23297049.
27. Chartio. (2022). Violin Plot: The Complete Guide. [Online]. Available: <https://chartio.com/learn/charts/violin-plot-complete-guide/>