Volume 1 No 6, 2025
Journal of Global Research in Multidisciplinary Studies (JGRMS)
Research Paper
Available online at https://saanvipublications.com/journals/index.php/jgrms/index

Machine Learning-Based Cost Prediction of Health Insurance Claims Using Medical Dataset

Dr. Dinesh Yadav
Associate Professor
CSE Department
St. Andrews Institute of Technology & Management, Gurugram, Haryana, India
dinesh.yadav@saitm.ac.in

***Abstract*—It allows companies to fix the prices of policies, evaluate risks and divide resources among different areas in health care. When the costs of health insurance claims are difficult to anticipate, healthcare companies are less prepared for financial planning and managing potential risks. Since more medical data is now accessible, machine learning (ML) approaches can help us model the complex reasons behind cost changes. This paper recommends using Light Gradient Boosting Machine (LightGBM) to predict the amount of money paid for health insurance claims from a medical insurance dataset. Just before dividing it into training and testing sets, the dataset with numbers and categories is prepared by encoding and handling any missing values. Using the new model results in significant improvements over regular regression models, showing an R-squared (R^2) value of 86.81%, MAE of 2381.57 and an RMSE of 4450.43. Experimental tests and visual aid prove that the model can catch non-linear patterns and boost the accuracy of predictions. This study points out that using LightGBM can make health insurance cost predictions more accurate.**

***Keywords*—Health Insurance, Cost Prediction, Machine Learning, Medical Dataset, Regression Analysis.**

I. INTRODUCTION

The role of public health is very important in society and it impacts people at local, national and global levels. Human lives and public health are constantly under threat from natural calamities, pandemics, shortages in medical resources, and other crises that amplify the vulnerability of populations [1][2]. Individuals may encounter unforeseen and unavoidable health-related events at any stage of life, highlighting the necessity for robust healthcare systems and financial planning mechanisms such as insurance [3].

The healthcare expenditures accounting for nearly 30% of the GDP, managing these costs effectively is becoming increasingly critical [4]. The growing demand for medical services, particularly as the baby boomer generation approaches Medicare eligibility, necessitates innovative solutions to predict and manage health insurance costs [5]. According to recent trends, the projected increase it is estimated that medical costs will increase by 7.0% each year in 2024, more than previous rates and requiring new steps to manage costs [6][7].

A larger part of the literature in health insurance is now devoted to developing and refining claim modeling for efficient setting of premiums [8][9]. Still, it is difficult for traditional actuarial techniques to take into account the complicated, nonlinear links between all the variables like demographics, lifestyle, location and medical records. Because of these limits, firms may use advanced tools to improve their forecasting results [10].

In the past few years, ML and AI have become important tools for insurers, allowing them to discover new patterns in large datasets [11][12]. Using these techniques brings greater precision in predicting costs, which enhances risk management and the choice of prices. Unlike mechanical engineering, in health insurance analytics, extra knowledge is required late in the development journey because of the need for big data and uncertain changes in health risks [13][14].

This work uses a set of medical features in a dataset such as age, Body Mass Index (BMI), smoking status, marital status, family history and location, to estimate health insurance expenses. Varying ML algorithms are reviewed to see which has the best predictive power [15][16][17]. Also, a web application using Streamlet is created, so users can check insurance costs in real time and benefit from the model [18]. How actuarial science differs from AI-powered modeling is developing new ways to price medical insurance. By adjusting insurance plans using new scientific and technological tools, everyone can enjoy fairer and less expensive healthcare.

A. Significance and contribution

It significantly helps health insurance cost prediction by using ML to improve both precision and efficiency in estimating medical claim costs. It is very important for insurance companies to predict health insurance claims accurately, as this helps them manage risks, set the right premiums and distribute their resources wisely. Using the LightGBM regressor among other ML approaches, this study solves the difficulty of modelling complex and non-linear trends found in healthcare cost data. Having a real-world dataset and processing it thoroughly before using it strengthens the predictive model and helps reduce worries about finances and improves operations in healthcare insurance.

- Utilization of a real-world medical insurance dataset comprising 1,338 records with both categorical and numerical attributes relevant to healthcare costs.
- Comprehensive data cleaning involves getting the dataset ready for modelling, missing values must be handled, duplicates must be eliminated, and categorical variables must be encoded.
- Development and evaluation of a LightGBM regression model, alongside comparative analysis with

Ridge Regression and Multiple Linear Regression models.
- Check and confirm model accuracy and performance using R^2, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

B. Justification and Novelty

The reason for this research is that health insurance claim costs are growing more diverse and complex, and traditional linear models have difficulties with them. Using Light GB Machine (LightGBM), which is a strong and efficient gradient boosting framework, the suggested approach handles challenges in regression by modelling complex relationships and working with lots of variables accurately. The novelty of this work lies in the tailored application of LightGBM to a real-world medical insurance dataset and comprehensive comparative analysis against established regression models. This integration demonstrates enhanced predictive performance, providing a robust solution for healthcare cost forecasting that has been underexplored in existing literature.

C. Structure of the Paper

This is how the study is organized: Section II provides pertinent research on estimating health insurance claim costs, and Section III explains the steps and resources used. Section IV displays the results of the suggested system's experiments. Section V wraps up the inquiry and presents a summary of its conclusions.

II. Literature Review

This section discusses several review articles on ML-Based Cost Prediction of Health Insurance Claims Using Medical Datasets. Table 1 summarizes, methods used, dataset details, key findings, and identified limitations or suggestions for future research.

E et al. (2025) community, to access healthcare services such as insurance policies, LIC, ICICI, HDFC ERGO, Star Healthcare, which offer benefits for claiming an amount for their medical expenses. The dataset encompasses three categories such as generalized data, hospitalized data and claim data. Each record in the dataset represents an individual's health insurance charges along with corresponding demographic and health-related characteristics. This research use ML methods to classify the claim amounts that have been adopted by individuals based on the hospitalized expenditure and insurance charges paid by the individuals. In the health sector, ML techniques are used in the prediction of insurance claims by building predictive models from historical data on claims, patient demographics, treatment types, and results [19].

Sharma and Jeya (2024) addresses the increased utility of health insurance estimations after the COVID-19 outbreak they are in a context where many efforts are trying to address this important issue, their study takes a dataset from Kaggle, with 1338 entries that impressively capture the nuances of American medical spending. In order to forecast health insurance costs, their research looks at a number of variables, including age, sex, BMI, smoking, and the number of children. It is distinct because it uses linear regression to examine the robust correlation between these variables and insurance premiums. They carefully partitioned their data set, with 70% for training, and achieved an impressive 81.3% accuracy [20].

Vijayalakshmi, Selvakumar and Panimalar (2023) the 24-feature dataset, which included all pertinent characteristics required for insurance cost prediction, was employed. Regression methods, including Linear Regression, Decision Tree Regression, Lasso Regression, Ridge Regression, Random Forest Regression, ElasticNet Regression, Support Vector Regression, K Nearest Neighbor Regression, and Neural Network Regression in R programming, were used in the implementation. R-squared (R^2), Explained Variance Score (EVS), MSE, RMSE, MAE, Mean Absolute Percentage Error (MAPE), and Adjusted R^2 (Adj. R^2) were the seven metrics used to assess the model's performance. A R-squared score of 0.9533 indicated that Random Forest Regression fared better. With this prediction system, the medical profession will have less manual labor and more accurate predictions [21].

Paikaray et al. (2023) deals with the prediction About making a health insurance claim for example, if a person makes a claim on their health insurance depending on several criteria. When human brains are used for prediction, organizations encounter a variety of flaws that often lead to more administrative labor for reworking and, ultimately, very little accuracy. Because ML algorithms can manage large volumes of data, they may be the solution to the aforementioned issues. SVM, DT, Artificial Neural Networks (ANN), and LRM are used to analyze the dataset in this study Machine Learning-Based Regression Model to Predict Health Insurance Claim [22].

Ghosh (2023) study use a balanced method and modify the threshold value to provide a balanced trade-off between accuracy and recall metrics, allowing the ML algorithms to produce objective output. Following dataset balance and threshold value modification, the final result produced an 89% accuracy rate for the LR model, with precision and recall metrics receiving scores of 40% and 39%, respectively. When applied to an unbalanced dataset, the same model produced comparable accuracy; however, the precision and recall values were 53% and 13%, respectively. Similar outcomes were seen with RF, which received a very low recall value (4%), although having the greatest accuracy value (67%). It follows that applying balancing procedures to unbalanced datasets and modifying the threshold value results in a far more impartial and balanced metric analysis [23].

Bora et al. (2022) the two ML algorithms in the proposed work Random Forest and Multiple Linear Regression are followed by an XAI description of the expected outcomes. Using model-specific techniques based on Microsoft's Interpret ML package, they first provide a brief overview. LIME, SHAP, and model-agnostic methodologies are then used to further explain the estimated cost of insurance premiums. The study's importance lies in improving user experience and fostering trust between the user and the computer. Since the domain experts may examine the characteristics that have the greatest impact on the result and provide their knowledge, these methods can aid in verifying the accuracy of the prediction models [24].

Table 1: Summary of background study for the Cost prediction of Health insurance claims

Author	Methods	Dataset	Key Findings	Limitations & Future Work
E et al. (2025)	ML grouping based on claim amount, hospitalized expenditure, and insurance charges	Dataset with 3 categories: generalized, hospitalized, and claim data	ML used to classify claim amounts based on demographic and health-related features	Detailed evaluation metrics not reported; potential for expanding real-time implementation
Sharma & Jeya (2024)	Linear Regression	Kaggle dataset with 1338 entries (age, sex, BMI, smoking, children)	Achieved 81.3% accuracy using 70/30 train-test split; identified strong factor-cost relationships	Focused only on Linear Regression; could explore ensemble methods or deep learning for improved accuracy
Vijayalakshmi, Selvakumar & Panimalar (2023)	ElasticNet, RF, Lasso, Ridge, Decision Tree, Linear, SVR, KNN, and NN (in R)	Dataset with 24 features for insurance cost	Random Forest Regression performed best with $R^2 = 0.9533$; reduced manual prediction effort	Requires real-world deployment validation; no mention of data imbalance handling
Paikaray et al. (2023)	ANN, Decision Tree, SVM, Logistic Regression	Health insurance claim prediction dataset	ML models can significantly reduce manual errors in predicting claims	Further work could focus on improving accuracy and exploring deep learning methods
Ghosh (2023)	Logistic Regression, Random Forest; threshold adjustment & balancing	Imbalanced dataset with insurance features	Balancing techniques improved precision-recall tradeoff; achieved 89% accuracy with balanced data	Need for real-world testing; poor recall in imbalanced settings; expand to other ML models
Bora et al. (2022)	Multiple Linear Regression, Random Forest, XAI (LIME, SHAP, Microsoft Interpret ML)	Insurance premium dataset	Enhanced model interpretability and trust using XAI; helped explain prediction outcomes	Future work could include more robust model generalization; computational cost of interpretability methods

III. Methodology

The methodology for ML-Based Cost Prediction of Health Insurance Claims Using a Medical Dataset involves a structured workflow as illustrated in Figure 1. Initially, medical insurance cost data is collected and subjected to preprocessing steps, including handling of missing values and removal of duplicates to ensure data quality. Label encoding is used to convert categorical variables into numerical representations so that ML algorithms may use the dataset. Following processing, the dataset is split 80:20 into training and testing groups. To anticipate the expenses of health insurance claims, a regression-based predictive modelling technique is used using the (LightGBM or LGBM). Standard regression measures like R^2, RMSE, and MSE are used to assess the model's performance. To evaluate the model's accuracy and general efficacy in cost prediction, the last step is to analyze the forecast outcomes.

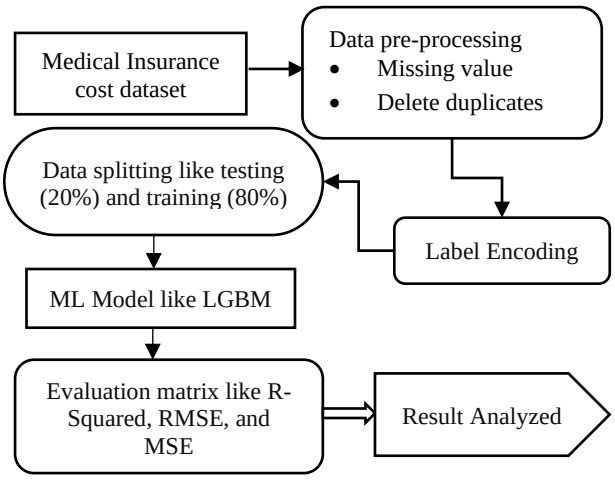

Fig. 1. Flowchart for Cost Prediction of Health Insurance Claims

The following steps of proposed methodology are briefly discussed below:

A. Data Collection

The dataset of medical insurance expenses used to create the ML model for cost prediction was obtained from Kaggle and comprises 1,338 records with seven attributes. Among these attributes, three contain categorical data (while the other four are numerical (age, BMI, number of children, and insurance prices). Examples of these include gender, smoking status, and area. The dataset is divided into two subsets: 80% for training and 20% for testing to guarantee efficient training and assessment. Regression models that can estimate health insurance costs are constructed using the training data, and their performance and generalizability are assessed using the testing data. While maintaining a distinct subset for objective performance evaluation, their data partitioning technique guarantees that the model discovers underlying patterns from a sizable percentage of the data.

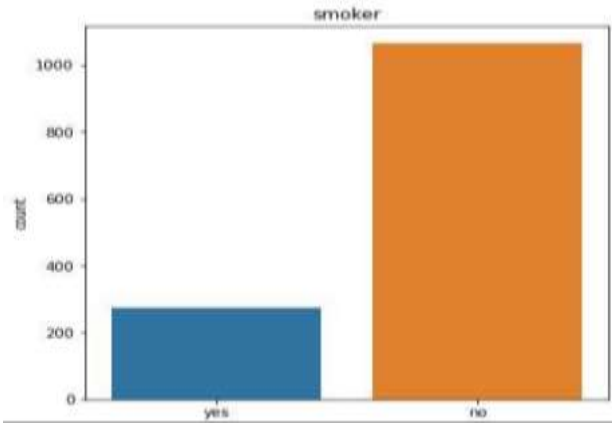

Fig. 2. Checking Smoker vs Non-Smoker

Figure 2 is a bar plot showing the distribution of people in the medical insurance dataset according to whether or not they smoke. It demonstrates that most of the people are not smokers, with over 1000 entries, while smokers are significantly fewer, around 250. This imbalance may influence insurance cost predictions and model training.

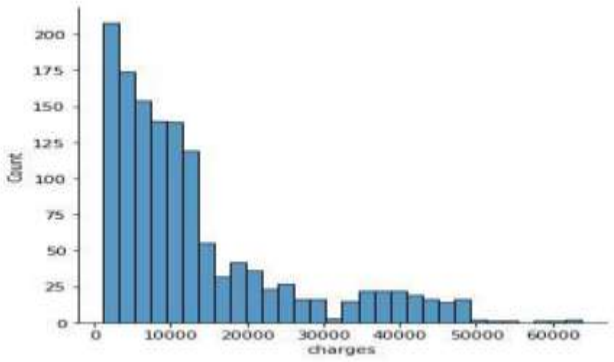

Fig. 3. Distribution of Charge Value

Figure 3 presents the distribution of charges. The graph clearly shows a right-skewed distribution, with a high concentration of charges occurring at lower values, predominantly below 15,000. As the charge values increase, the frequency of occurrences significantly decreases, indicating that a smaller number of instances are associated with higher charges. There are some minor peaks observed in the higher charge ranges, particularly around 35,000 to 45,000, suggesting a secondary, less frequent cluster of charges.

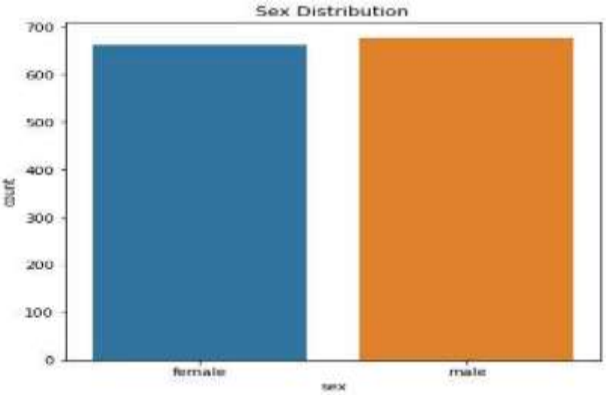

Fig. 4. Sex Distribution Female vs Male

Figure 4: Bar plot illustrating the distribution of individuals by sex in the medical insurance dataset. The dataset is nearly balanced, with a slightly higher number of males than females, around 675 each. This balanced distribution ensures that gender bias is minimal in model training for predicting insurance costs.

B. Data Preprocessing

Three columns are numerical, and three are categorical. The category values cannot be accommodated by a ML model, as computers are unable to comprehend this text value. Consequently, it shall assign numerical values to those categories' attribute labels. It changes "female" to 1 and "male" to 0 in the "sex" field. It also changes the other two columns to have numerical values.

- **Missing Values:** Dealing with missing values is the first stage in data cleansing. Absent data were detected using the isnull (), sum () function. Records containing null values were removed using the dropna () method to maintain data integrity.
- **Delete Duplicates:** Make use of the Remove Duplicates method to guarantee that duplicate data is removed from the dataset forever.

C. Data Splitting

Dataset splitting is a crucial process in ML modelling that supports the model at every stage, from training to evaluation. The dataset is divided into two distinct subsets: the training set and the testing set. During the experimental phase, 80% of the data is utilized for training and 20% for testing.

D. Regression with LightGBM Regressor (LGBM)

The GBDT technique is implemented by the LightGBM framework, which facilitates effective parallel training, quicker training speeds, reduced memory use, improved accuracy, and distributed assistance for swiftly processing large amounts of data [25][26][27]. Instead of level-wise splitting of decision trees during growth, it follows a strategy where leaves are created in a leaf-wise fashion and a limit is put on tree depth. The tree keeps finding the leaf that provides the most split gain and continues to split it. In this way, leaf-wise is better at reducing errors and achieving more accurate results with the same number of split points versus level-wise [28][29]. It is shown in Equation (1):

$$\hat{y}_i = \sum_{m=1}^{M} f_{m(x)_i}, fm \in F \quad (1)$$

E. Evaluation Metrics

Evaluating models is important during ML project development since it lets you judge their performance and present their results more clearly. But it is not easy to predict the exact value of a regression model which is why it try to explain how close the predictions and actual observations are. It evaluated the models by applying 3 metrics: R^2, MAE and RMSE.

1) R Square

R^2 is the squared form of the Correlation Coefficient (R) and is used to assess how well a model fit. It also denotes the proportion of the predicted price that is explained by the features (see Equation (2)).

$$R^2 = 1 - \frac{\sum_i^n (y_i - \hat{y}_i)^2}{\sum_i^n (y_i - \bar{y})^2} \quad (2)$$

2) Mean Absolute Error

MAE is calculated to accurately reflect prediction errors (see Equation (3)).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}i)^2 \quad (3)$$

3) Root Mean Squared Error

RMSE calculates the deviation (standard deviation) of the residuals which shows the quality of the relationship in the model. (see Equation (4)) [30].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_i)^2} \quad (4)$$

Where, y_i is the actual value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the total number of cases (n) and the average of the values should be used to calculate the standard deviation.

IV. Result Analysis And Discussion

This section shows the findings from ML experiments carried out for predicting cost in health insurance claims using the Medical Dataset. ML models were trained on a system with a 3.3 GHz Intel dual-core i6 processor, 1 TB of RAM and Windows 10 as its operating system. Evaluation of the model's results was done using key regression stats as shows in Table 2: The R^2as well as the MAE, and the RMSE. The

R^2 of 0.8681, MAE of 2381.5670 and RMSE of 4450.4333 for the LightGBM Regression model mean it performed well in estimating health insurance claim costs.

Table 2: ML and DL models on the Medical Dataset for Cost Prediction of Health Insurance Claims

Performance Measures	LightGBM Regression
R squared	86.81
MAE	2381.5670
RMSE	4450.4333

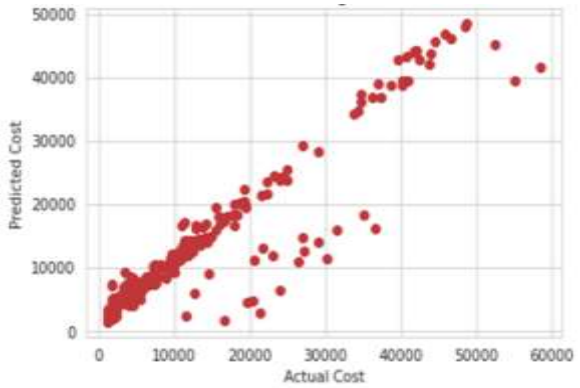

Fig. 5. Predicted Cost using LightGBM Regression

Figure 5 illustrates the relationship between using the LightGBM regression model, the actual and anticipated costs of health insurance. Every red dot in the chart is a prediction and it shows the actual insurance cost against the predicted cost with the x- and y-axes. Since the points are close to the diagonal line, there is a clear positive correlation which means the model is successful at catching the key features in the data. However, a few deviations from the diagonal highlight instances where the prediction error is relatively higher. Overall, the plot visually confirms the effectiveness of the LightGBM model in accurately forecasting medical insurance expenses.

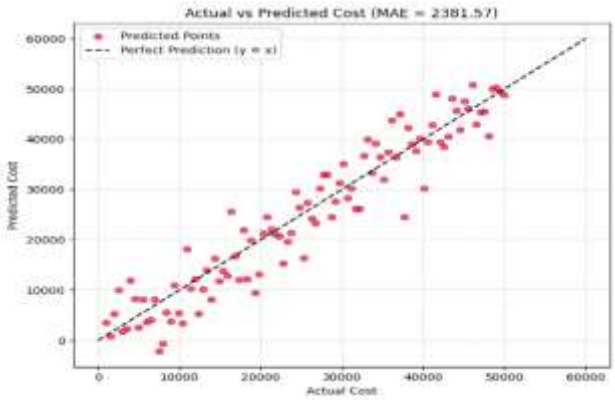

Fig. 6. MAE for LightGBM Regressor

Figure 6 illustrates the regression performance of the LightGBM model in predicting healthcare costs based on the medical dataset. The scatter plot presents the relationship between actual and predicted cost values, with each point representing an individual prediction. The ideal prediction line (y = x), when anticipated expenses precisely match actual values, is shown by the dashed diagonal line. Data being close to the line reveals that the model makes accurate predictions. The MAE of 2381.57, seen in the figure title, shows the typical difference between predicted and actual medical insurance costs, confirming that the model is precise.

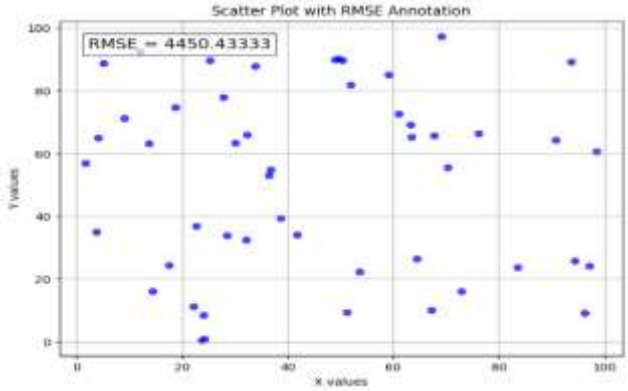

Fig. 7. Scatter Plot with RMSE Annotation

Figure 7 shows how the actual and predicted values relate, and the RMSE is written on the graph as 4450.4333. It allows us to see the average size of prediction errors which helps us understand how close the model is to estimating the target variable correctly. When RMSE is low, the model can better forecast the data and the present RMSE level allows you to measure how well the model matches the data's underlying patterns.

A. *Comparative Analysis and Discussion*

The lightweight model LightGBM Regressor, along with DT and RF, showed their performance on the Medical Insurance cost dataset to help predict medical costs and distribute resources more efficiently. Accuracy and how well the model worked under different conditions were evaluated using R-Square, MAE and RMSE. Table 3 shows a comparison of three regression models. In comparison with other models, the LightGBM Regressor is best because it has an R-Square of 86.81 and an RMSE of 4450.4333. Ridge Regression manages an R-Square of 78.38 and an RMSE of 4652.06, which is not as good as LightGBM. Compared to SVD and Ridge Regression, Multiple Linear Regression turns out to be the worst, with an R-Square of 75 and the biggest RMSE of 7523.98, meaning there are much bigger differences between predicted and actual prices.

Table 3: ML models comparison of ML Models for Health Insurance Cost Prediction

Performance Measures	LightGBM Regressor	Ridge[31]	Multiple Linear Regression[32]
R- Square	86.81	78.38	75
RMSE	4450.43333	4652.06	7523.98

Its good performance and an R^2 score of 86.81% prove that the proposed LightGBM-based regression model accurately estimates health insurance costs. It exceeds Ridge Regression and Multiple Linear Regression in their ability to find accurate solutions and to keep errors to a minimum. Working well with challenging and detailed data, LightGBM can find hidden non-linear patterns in the medical data. Because of its gradient boosting model, it can cope well with new information which makes it valuable for predicting healthcare costs in real-life scenarios.

V. Conclusion And Future Work

Insurance companies are under a lot of pressure to predict how much they will pay for claims as healthcare prices go up. Using predictive models in health insurance is vital for making

resource use more efficient and cutting financial risks. Because medical data is more accessible, ML techniques are now used for accurate cost prediction. It looks at the ways ML methods are used to forecast healthcare claim amounts. A practical and effective ML solution is described to predict health insurance claim costs using real medical records. LightGBM achieved R^2 result of 86.81%. Such results suggest that the model can find complex relationships and work well with new data. Because the model performs well, it may be useful in helping insurance organizations manage costs and plan resources more effectively. If the number of expected claims is known, it helps stakeholders handle finances wisely, avoid risks and improve the way health insurance works.

The researchers will focus on broadening the data to record more patient experiences, which could make the model suitable for different cases. Adding information about medical history, how the patient is treated, and hospital records can give a wider picture of what drives healthcare costs. Ensuring that real-time insurance platforms include explainable AI parts will increase trust and transparency in decision-making by the system.

REFERENCES

[1] L. N. Srinivasagopalan, "Predicting health insurance premiums using machine learning : A novel regression- based model for enhanced accuracy and personalization," vol. 19, no. 01, pp. 1580–1592, 2023.

[2] A. Balasubramanian, "Intelligent Health Monitoring: Leveraging Machine Learning and Wearables for Chronic Disease Management and Prevention," *Int. J. Innov. Res. Eng. Multidiscip. Phys. Sci.*, vol. 7, no. 6, pp. 1–13, 2019.

[3] S. R. P. Madugula and N. Malali, "Adversarial Robustness of AI-Driven Claims Management Systems," *Int. J. Adv. Res. Sci. Commun. Technol.*, pp. 237–246, Mar. 2025, doi: 10.48175/IJARSCT-24430.

[4] B. Chaudhari and S. C. G. Verma, "Synergizing Generative AI and Machine Learning for Financial Credit Risk Forecasting and Code Auditing," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 2, 2025.

[5] S. Singamsetty, "Neurofusion: Advancing Alzheimer's Diagnosis with Deep Learning and Multimodal Feature Integration," *Int. J. Educ. Appl. Sci. Res.*, vol. 8, no. 1, pp. 23–32, 2021.

[6] D. Ramya, S. K. Manigandan, and J. Deepa, "Health Insurance Cost Prediction using Machine Learning Algorithms," in *International Conference on Edge Computing and Applications, ICECAA 2022 - Proceedings*, 2022. doi: 10.1109/ICECAA55415.2022.9936153.

[7] P. Chatterjee, "AI-Powered Payment Gateways : Accelerating Transactions and Fortifying Security in Real-Time Financial Systems," *Int. J. Sci. Res. Sci. Technol.*, 2023.

[8] Anmol, S. Aggarwal, and A. Jahan Badhon, "Medical Insurance Cost Prediction Using Machine Learning Algorithms," *Lect. Notes Networks Syst.*, vol. 467, no. October, pp. 271–281, 2023, doi: 10.1007/978-981-19-2538-2_27.

[9] G. Mantha, "Transforming the Insurance Industry with Salesforce: Enhancing Customer Engagement and Operational Efficiency," *North Am. J. Eng. Res.*, vol. 5, no. 3, 2024.

[10] P. Chatterjee, "Cloud-Native Architecture for High-Performance Payment System," vol. 10, no. 4, pp. 345–358, 2023.

[11] Md Mohtaseem Billa, "Medical Insurance Price Prediction Using Machine Learning," *J. Electr. Syst.*, vol. 20, no. 7s, pp. 2270–2279, 2024, doi: 10.52783/jes.3962.

[12] R. P. Mahajan, "Transfer Learning for MRI image reconstruction: Enhancing model performance with pretrained networks," *Int. J. Sci. Res. Arch.*, vol. 15, no. 1, pp. 298–309, Apr. 2025, doi: 10.30574/ijsra.2025.15.1.0939.

[13] C. Hennebold, K. Klöpfer, P. Lettenbauer, and M. Huber, "Machine Learning based Cost Prediction for Product Development in Mechanical Engineering," *Procedia CIRP*, vol. 107, no. March, pp. 264–269, 2022, doi: 10.1016/j.procir.2022.04.043.

[14] A. Polleri, R. Kumar, M. M. Bron, G. Chen, S. Agrawal, and R. S. Buchheim, "Identifying a Classification Hierarchy Using a Trained Machine Learning Pipeline," U.S. Patent Application No. 17/303,918, 2022

[15] W. Mimra, J. Nemitz, and C. Waibel, "Voluntary pooling of genetic risk: A health insurance experiment," *J. Econ. Behav. Organ.*, vol. 180, pp. 864–882, Dec. 2020, doi: 10.1016/j.jebo.2019.04.001.

[16] S. Wawge, "Evaluating Machine Learning and Deep Learning Models for Housing Price Prediction," *IJARSCT*, vol. 5, no. 11, pp. 367–377, 2025.

[17] R. Q. Majumder, "Machine Learning for Predictive Analytics: Trends and Future Directions," *Int. J. Innov. Sci. Res. Technol.*, vol. 10, no. 4, pp. 3557–3564, 2025.

[18] N. Malali, "Using Machine Learning to Optimize Life Insurance Claim Triage Processes Via Anomaly Detection in Databricks: Prioritizing High-Risk Claims for Human Review," *Int. J. Eng. Technol. Res. Manag.*, vol. 6, no. 6, 2022, doi: 10.5281/zenodo.15176507.

[19] P. E, K. M. N, S. S. S, D. D, J. S, and A. P. V, "Prediction of Insurance Claims for Health Sector using Machine Learning Techniques," in *2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT)*, IEEE, Feb. 2025, pp. 1312–1318. doi: 10.1109/IDCIOT64235.2025.10914851.

[20] A. Sharma and R. Jeya, "Prediction of Insurance Cost through ML Structured Algorithm," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, IEEE, Feb. 2024, pp. 495–500. doi: 10.1109/IC2PCT60090.2024.10486304.

[21] V. Vijayalakshmi, A. Selvakumar, and K. Panimalar, "Implementation of Medical Insurance Price Prediction System using Regression Algorithms," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, Jan. 2023, pp. 1529–1534. doi: 10.1109/ICSSIT55814.2023.10060926.

[22] B. K. Paikaray, T. Samantara, D. Rout, and S. Mishra, "Machine Learning-Based Regression Model to Predict Health Insurance Claim," in *Proceedings of 2023 IEEE 2nd International Conference on Industrial Electronics: Developments and Applications, ICIDeA 2023*, 2023. doi: 10.1109/ICIDeA59866.2023.10295226.

[23] S. Ghosh, "Health claim propensity model using Machine Learning," in *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*, IEEE, Dec. 2023, pp. 504–509. doi: 10.1109/DeSE60595.2023.10469173.

[24] A. Bora, R. Sah, A. Singh, D. Sharma, and R. K. Ranjan, "Interpretation of machine learning models using XAI - A study on health insurance dataset," in *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)*, IEEE, Oct. 2022, pp. 1–6. doi: 10.1109/ICRITO56286.2022.9964649.

[25] S. Singamsetty, "Lightweight Reg Net-Driven Deep Learning Framework for Enhanced Classification of Neurodegenerative Diseases from MRI Images," *Int. J. Adv. Eng. Sci. Res.*, vol. 10, no. 1, pp. 28–37, 2023.

[26] S. J. Wawge, "A Survey on the Identification of Credit Card Fraud Using Machine Learning with Precision , Performance , and Challenges," vol. 10, no. 4, 2025.

[27] K. K. Nimavat and R. Kumar, "Updating Machine Learning Training Data Using Graphical Inputs," 17178360, 2022

[28] H. Du, L. Lv, A. Guo, and H. Wang, "AutoEncoder and LightGBM for Credit Card Fraud Detection Problems," *Symmetry (Basel).*, 2023, doi: 10.3390/sym15040870.

[29] S. R. Sagili, S. Chidambaranathan, N. Nallametti, H. M. Bodele, L. Raja, and P. G. Gayathri, "NeuroPCA: Enhancing Alzheimer's disorder Disease Detection through Optimized Feature Reduction and Machine Learning," in *2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, IEEE, Jul. 2024, pp. 1–9. doi:

10.1109/ICEEICT61591.2024.10718628.

[30] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," *Mach. Learn. with Appl.*, 2024, doi: 10.1016/j.mlwa.2023.100516.

[31] G. K. Patra, C. Kuraku, S. Konkimalla, V. N. Boddapati, and M. Sarisa, "An Analysis and Prediction of Health Insurance Costs Using Machine Learning-Based Regressor Techniques," *J. Comput. Eng. Technol.*, vol. 12, no. 3, pp. 102–113, 2021.

[32] S. Panda, B. Purkayastha, D. Das, M. Chakraborty, and S. K. Biswas, "Health Insurance Cost Prediction Using Regression Models," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, COM-IT-CON 2022*, 2022. doi: 10.1109/COM-IT-CON54601.2022.9850653.

