Medical Insurance Cost Prediction Using Gradient Boosting Regression: A Machine Learning Approach

Baolong Zhang*, Haiyan Huang, Dongxia Wang
School of Artificial Intelligence, Jiyuan Vocational and Technical College, Jiyuan, Henan 459000, China
E-mail: zbl_hhy0421@126.com
*Corresponding author

Keywords: medical insurance cost prediction, database analysis, machine learning, gradient boosting regressor, performance evaluation

Received: January 19, 2025

This paper highlights the point that correct forecasting of the expense of medical insurance is essential in the better decision-making of individuals, insurers, and policymakers to efficiently allocate resources in the dynamically changing environment of healthcare financing. While recent studies have extensively explored machine learning (ML) approaches for medical insurance cost prediction, there remains a critical need to improve their accuracy and reliability, driving the pursuit of more effective methods to enhance the precis. In the context of these caveats, there exists a research gap to which this investigation attempts to contribute by proffering an ML method using the Gradient Boosting Regressor (GBR), through which one can enhance the level and quality of prediction for medical insurance expenses. To deal with this, this study presents a GBR base approach for predicting medical insurance costs from a dataset of 1,339 samples with seven features, such as age, sex, BMI, smoking, and region. The dataset from Kaggle offers thorough coverage of the factors affecting medical insurance costs. Our approach involves extensive preprocessing of the data, including one-hot encoding for categorical features, followed by training, validation, and evaluation of the model using an 80/20 train-test split. We rigorously evaluated the performance of the GBR model using the metrics of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), Mean Absolute Percentage Error (MAPE), and Explained Variance Score (EVS). Experimental outcome further establishes that the best-performing model is the GBR model, based on obtained results as reflecting better predictive accuracy. Comparison with Linear Regression, Random Forest, Support Vector Regression, K-Nearest Neighbors, and Neural Networks further established that the best precision (0.908), recall (0.903), and F1-score (0.899) is achieved by the GBR model. These findings support the effectiveness of the GBR model as a powerful tool for capturing nonlinear patterns of relationship underlying the data, for predicting medical insurance costs. This research highlights the usefulness of sophisticated techniques of machine learning for improving predictive modeling of healthcare finances.

Povzetek: Članek uvaja prilagojeno uporabo gradientnega ojačitvenega regresorja za napoved stroškov zdravstvenega zavarovanja z osredotočanjem na sistematično obdelavo podatkov, interpretabilnost modela in zajem nelinearnih vplivov.

1 Introduction

The most relevant problem nowadays, connected with healthcare and medical services, is an increase in expenses that becomes more and more difficult for people to afford and seriously burdens both individuals and societies [1], [2]. First of all, it is not possible to imagine mitigation of the financial burdens without participation from the side of medical insurance. The expenses of insurance play a huge role in human life nowadays, protecting people from the unpredictable character of health-related expenses [3], [4].

Prediction of medical insurance expenses is of great importance for people who want to make a financial plan, but it also has key implications in resource management in health service sectors [5], [6]. To be in such a position, predictions of these expenses will allow informed choices to be made by the people, providers of insurance, and policymakers themselves, thus saving enormous financial resources and contributing to the overall sustainability of the health system [7]. Therefore, with this view, there is an increasing need for the implementation of sophisticated methods of data analysis in predictive model development that could ensure medical estimates of insurance are closer to real values.

Among all data analyses, the usage of AI tools and ML features is one of the successful approaches to performing predictive modeling in medical insurance expenses. ML algorithms can be versatile in their operations of extracting meaningful patterns and insights from complex databases for accurate prediction capabilities [8], [9]. In recent times, ML methods have been increasingly adopted for their great capabilities and wide scope of application in various domains to develop robust frameworks of medical insurance cost prediction.

While much research has been conducted using traditional ML methods for the anticipation of medical insurance expenses, there are also some challenges in the present scenario [10]. One major difficulty is that a very high level of accuracy is required concerning the value of prediction. Further studies and refinement need to be done in this regard. The aforementioned challenges have to be addressed if anybody wants to have the assurance of reliability and applicability of ML-based approaches in real-world scenarios.

This work presents an approach for predicting medical insurance expenses using ML, specifically through the GBR model. Although GBR is not novel by itself, the research introduces an adapted solution by employing systematic preprocessing, hyperparameter tuning through grid search with 5-fold cross-validation, as well as feature importance analysis to better tackle certain problems with medical insurance cost prediction. These changes, along with thorough evaluation metrics and comparison to baseline models, make the traditional GBR into an even more stable and interpretable solution for the domain-specific problem. The novelty therefore lies more in the strategic application and adaptation of GBR than in the algorithm. The GBR method was utilized in this research because it can learn nonlinear interactions among and between features, which is essential for correctly projecting medical insurance expenses. GBR achieves this by progressively refining weak learners and hyperparameter tuning, leading to better accuracy without sacrificing the ability to generalize to a variety of datasets. The paper describes the development of a model through careful database preparation, training, validation, and testing process that can contribute to state-of-the-art in accurate anticipation of the medical insurance cost.

The 3 important contributions of the research are:

1. A new ML-based strategy has been introduced in the research for the anticipation of medical insurance expenses.
2. Extensive research on pre-existing ML-based approaches was conducted and ended with the development of an efficient predictive model.
3. Finally, extensive experimentation and performance evaluation are performed to validate the productivity of the recommended strategy in addressing challenges arising from high-accuracy prediction requirements within medical insurance expenses.

2 Related work

Paper [11] also explored the use of ML methods for expenditure predictions among high-cost, high-need patients in health care. Its methodology was based on the use of advanced ML algorithms that analyze complex data in search of patterns that would help in identifying and forecasting patients with high healthcare expenses. This paper discusses various ML methods that have the huge potential to enhance efficiency and precision in cost predictions related to healthcare for such patients with high medical needs.

The authors of [12] recommended a computational intelligence-based approach for cost estimation in medical insurance by applying advanced techniques of artificial intelligence (AI). The approach identifies the complex relationships that affect the expense of medical insurance and then predicts those using the application of computational intelligence algorithms. The research work utilized an advanced framework to apply ML methods along with data analysis to provide more accuracy in the anticipation of medical insurance expenses. The outcomes depict how effectively the approach of computational intelligence has given precise predictions of medical insurance expenditure and hence provided a worthy contribution to the research arena of predictive modeling in healthcare finance. However, one notable limitation involves the fact that the focused task of prediction was only regarding accuracy rates. This calls for further research to allow a more nuanced exploration of other metrics and considerations.

This paper [13] presented a broad ML-based regression framework for the anticipation of health insurance premiums. The methodology uses higher-order regression frameworks to make more accurate predictions about the expense of health insurance. The authors of this investigation made use of a very diverse database and resorted to ML techniques to model the complicated relationship that exists between the many factors determining insurance premiums. Outcomes have underlined the efficacy of the recommended framework in realizing health insurance premium predictions by achieving high accuracy, hence contributing to the wider realm of predictive modeling within the healthcare domain. However, there is a limitation—the sole focus on accuracy rates in the anticipation task that is being pursued.

Paper [14] discusses the estimation of the expense of medical care using regression in ML. The regression frameworks are, in this approach, used for the identification and prediction of complex interactions among factors contributing to hospitalization and medical care expenses. By working with a broad database, the research taps into ML for cost modeling and prediction in healthcare services. Outcomes of this kind suggest that the regression-based approach of ML can turn out to be highly effective regarding the anticipation of hospitalization- and medical-care-related expenses, thereby making a valuable contribution to the field. Nevertheless, the apparent limitation of focusing only on accuracy rates concerning the anticipation task invokes broader challenges connected to predictive modeling.

This paper [15] developed an ML framework to anticipate the outcomes and expenses associated with cardiac surgery. The methodology applied in this investigation is based on the utilization of ML techniques in the evaluation and interpretation of complex databases as a means of identifying patterns and relationships that contribute to surgical outcomes and expenses. They aimed to come up with improved accuracy of predictions on patient outcomes and financial consequences of cardiac surgery using advanced algorithms. The research findings bring into view the efficacy of the ML framework in the

proper forecast of both medical outcomes and cost implications of cardiac surgical procedures. This will go a long way in offering useful insights to healthcare practitioners and administrators in the optimization of patient care and resources. However, one of the critical limitations of this paper focuses on the sole concentration on the issue of accuracy rates in the anticipation task.

Table 1 presents critical analysis of previous studies. The examined works cumulatively highlight the major improvements achieved through medical insurance cost prediction employing machine learning (ML) methods. Despite such improvements, critical examination highlights the gaps within state-of-the-art (SOTA) strategies that limit their applicability across larger domains as well as their dependability. For example, though Yang et al. [11] attain precision for the costliest patients, their strategy is narrow and not easily generalizable across larger datasets. In the same vein, Hassan et al. [12] and Kaushik et al. [13] ensure accuracy above all else, but ignore critical inputs such as interpretability as well as robustness, which are integral to healthcare decision-making. In addition, Taloba et al. [14] as well as Zea-Vera et al. [15] prove themselves within specific contexts, but do not apply such holistic evaluation metrics as Explained Variance Score (EVS) or Mean Squared Logarithmic Error (MSLE) so as not to capture finely nuanced dimensions of their performances.

Table 1: Critical analysis of the previous studies

Ref.	Proposed Methodology	Advantages	Disadvantages
Yang et al. [11]	Advanced ML algorithms for predicting high-cost patient expenditures	High precision for identifying high-cost patients; effective pattern extraction from complex data	Limited generalizability to broader datasets; does not address interpretability or scalability
Hassan et al. [12]	Computational intelligence-based approach for medical insurance cost prediction	Precise predictions using AI techniques; capturing complex relationships in insurance costs	Sole focus on accuracy rates; lacks exploration of other metrics like EVS or MSLE, limiting comprehensive performance evaluation
Kaushik et al. [13]	Higher-order regression framework for health insurance premium prediction	High accuracy in predicting premiums; robust modeling of complex relationships	Concentrates only on accuracy; overlooks model interpretability and adaptability to diverse datasets
Taloba et al. [14]	Regression-based ML for hospitalization and medical care cost estimation	Effective in predicting hospitalization expenses; works with broad databases	Narrow focus on cost estimation; limited applicability to other healthcare domains
Zea-Vera et al. [15]	ML framework for predicting outcomes and costs after cardiac surgery	Accurate predictions for surgical outcomes and costs; valuable insights for resource optimization	Focuses primarily on accuracy; lacks nuanced evaluation metrics such as EVS or MSLE, hindering broader performance assessment

3 Methodology

This research tackles the issue of precise medical insurance cost prediction, critical for well-informed decision-making by individuals, insurers, and policymakers amid the evolving healthcare funding landscape. The research questions centre on determining the ML method for improving the accuracy and stability of medical insurance cost prediction, as well as examining how powerful ML methods as GBR based approach can surpass standard techniques, as well as the quality of the model as evidenced by complete metrics. In contrast with current methods that prioritize solely accuracy or work with only specific datasets, the novel approach utilises the capacity of GBR to identify intricate nonlinear patterns and interplays between features, providing both good predictive accuracy as well as good generalization across varied settings. Through the inclusion of detailed preprocessing of data, one-hot encoding of categorical features, as well as a well-ordered train-test split, the novel approach seeks to provide a more clear, understandable, as well as stable solution for predicting medical insurance cost. The anticipated goal is a validated Gradient Boosting Regressor (GBR) model that outperforms benchmark models, furnishing actionable information on the drivers of insurance cost while remedying the deficiencies of existing state-of-the-art techniques, such as poor accuracy, uninterpretability, as well as insufficient adaptability to varied datasets.

3.1 Database

This investigation used a database from the Kaggle repository [16]. The database is a collection of medical insurance expenses for individuals based on various attributes. The database contains 1,339 records with 7 columns. The database includes the following characteristics: age, representing the person's age in years; gender, indicating whether the individual is male or female; BMI, a measurement of body fat calculated from height and weight; the count of dependents or children protected by the insurance plan; smoking status, which reflects whether or not the person smokes; region, specifying the area of the United States (northeast,

southeast, southwest, or northwest) where the individual resides; and charges, referring to the expenses billed by the health insurance company for specific medical expenses.

3.2 Data preprocessing

Because medical data are complex and varied, doing proper preprocessing is a need for accurate modeling to anticipate the expense of medical insurance. Preprocessing ensures data cleaning, standardization, and correctness. The more comprehensive the preparation, the better the derivation of meaningful patterns and relationships from the framework, hence keeping minimal biases and inaccuracies that ensure reliability in the anticipation.

The data is winsorized from the 1st percentile as well as the 99th percentile to limit outliers of the target column (charges) that reduce skewness. Missing values—albeit scarce within the Kaggle dataset—get imputed by mean imputation for numerical columns (age, BMI) as well as mode imputation for categorical columns (smoker, region). Categorical columns (smoker, region, gender) are encoded through one-hot encoding for not assuming ordinal information. An 80% training set and 20% test set split is used by train_test_split for stable checking while still retaining distribution of the data.

The GBR is tuned by grid search with 5-fold cross-validation for balancing bias-variance trade-offs. The main parameters are:

- n_estimators=500: Strong set of trees for strong learning.
- learning_rate=0.05: Moderate learning rate to avoid overfitting, combined with higher n_estimators.
- max_depth=5: Fits nonlinear effects within reasonable complexity.
- subsample=0.8 : Reduces variance through stochastic gradient boosting. A grid search over parameter combinations (e.g., n_estimators=[100, 300, 500], learning_rate=[0.01, 0.05, 0.1]) identifies the optimal configuration, ensuring reproducibility via random_state=42.

The study tackles overfitting prevention by utilizing cross-validation as well as regularization methods** native to the GBR algorithm. Precisely, the GBR algorithm utilizes 5-fold cross-validation for hyperparameter optimization through grid search, so that the system is both trained on and validated with several data subsets to mitigate overfitting tendencies. In this way, the performance of the model can be assessed for its applicability to varied data partitions so that it won't learn noise or idiosyncrasy localized to one specific learning sample. The algorithm is even implicitly regularized by significant hyperparameters like `learning_rate=0.05` and `subsample=0.8`, where the former controls the influence of each tree and the latter introduces stochasticity into the learning process, respectively. These configurations counter overfitting by inducing gradual learning as well as variance reduction. The usage of `max_depth=5` further restricts the depth of individual trees, balancing the capture of nonlinear relations with reduced complexity of the model. All of these tactics—cross-validation, learning rates constrained, subsampling, and depth restriction of trees—assure that the GBR algorithm is stable and generalizable without sacrificing high predictability.

Experiments are carried out within a Python 3.8.12 environment with scikit-learn 1.2.2 and XGBoost 1.6.1, running on an Intel i7-10700K CPU (3.8 GHz, 16GB RAM). Feature significance is examined based on SHAP values and permutation importance, identifying smoker status, age, and BMI as the most predictive features. Recursive Feature Elimination (RFE) is used for the elimination of redundant features (e.g., region), improving interpretability. This stringent setup guarantees reproducibility as well as alignment of model behavior with healthcare finance domain expertise.

3.3 Model training

This work will apply the GBR algorithm for model training and evaluation when generating frameworks [17]. Also, it determines the most important elements that influence the price of insurance by providing feature relevance ratings. It captures complicated nonlinear linkages and interactions among the features. This paper is discussed in detail concerning sections as presented below.

3.3.1 Gradient boosting regressor

Model training for predicting medical insurance expenses using GBR is a process of fitting a regression framework that can learn from the data and anticipate based on a given loss function. Fig. 1 illustrates the structure of the GBR algorithm.

In Fig. 1, the partial dependence plot of the feature X1 to the projected value Y using the GBR model is presented. A partial dependence plot displays the variation in projected value given that one feature varies and all other features are kept at their average values. This graph displays a nonlinear, positive relationship between feature X1 and the target value Y. When X1 increases from 0 to 10, there is an increase in the value of Y at an increasingly slow rate. The slope of the curve is much steeper when X1 is low compared to where the values of X1 are high. In other words, the marginal effect of X1 on Y shrinks as X1 grows larger. It reaches its peak at about X1 = 10 and then starts to decline slightly. It means that there exists an optimal value of X1, which provides a maximum projected value of Y. Also, this displays that any increase in X1 beyond that point will, in fact, decrease the value of Y. In addition, it also includes the confidence intervals for the partial dependence plot, which are the shaded regions around the curve. The widths of the confidence intervals correspond to the uncertainty of the estimated partial dependence, considering the variability of both data and the framework. The narrower the confidence intervals are, the more precise the estimate. The graph displays that the confidence intervals are relatively narrow for most values of X1, except for the extreme values near 0 and 20. This

means that the partial dependence plot is more reliable for the intermediate values of X1 than for the extreme values.

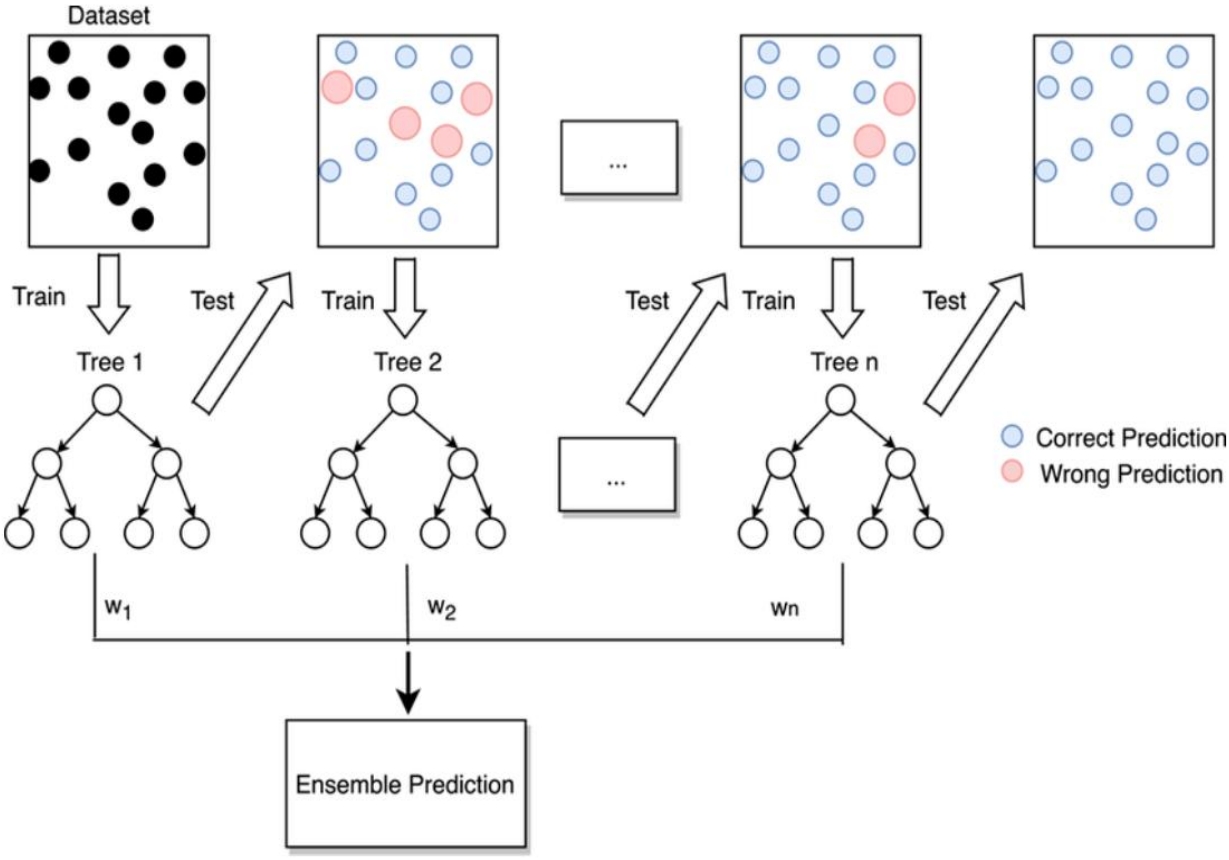

Figure 1: Structure of GBR algorithm [18].

3.3.2 Model configuration

Model configuration for GBR is a process of setting the hyperparameters that control the behavior and performance of the framework. Some of the important hyperparameters are:

- ***n_estimators***: It displays how many trees are going to be employed in the ensemble. Although a greater number often yields better outcomes, it also raises the computing cost and increases the danger of overfitting.
- ***learning_rate***: the percentage that updates the forecasts following each tree. Although it takes more trees, a lower learning rate enhances generalization and lowers variation.
- ***max_depth***: each tree's maximum depth. Although a deeper tree can capture more complexity, it also comes with a higher computational cost and overfitting risk.

The function GradientBoostingRegressor will create a model of type GBR with the given configuration:

- ***n_estimators***=100: the framework will use 100 trees in the ensemble.
- ***learning_rate=0.1***: the framework will update the anticipations by 10% of the learning rate after each tree.
- ***max_depth=3***: the framework will limit the depth of each tree to 3 levels.
- ***random_state=42***: the framework will use 42 as the seed for the random number generator.

Then, the framework is fit with training data, X_train and y_train-with default loss, and squared error.

For comparison of the GBR model, other models were set up with suitable hyperparameter values to have an equitable comparison. The Linear Regression model was run with the default settings since it is less in need of tuning. The Random Forest (RF) was run with an `n_estimators=100` and `max_depth=10` to provide an optimal balance between complexity and generalization. The Support Vector Regression (SVR) was provided an RBF kernel with `C=1.0` and `gamma='scale'`, typical settings for nonlinear regression problems. The K-Nearest Neighbors (KNN) was given an `n_neighbors=5` and `weights='distance'` to give higher weights to the closer data points for predictions. The Neural Network utilized the simple feedforward network with two 64-node hidden layers with ReLU activation, along with 0.001 learning rate and 50 epochs of training. These settings were selected because of their typical applications to regression tasks as well as initial experiments to provide the best possible performance for every algorithm. Although detailed descriptions about hyperparameter tuning for those models are not made explicit in the manuscript, follow-up studies could make use of grid search or

randomized search to better tune these parameters to get an even higher level of comparability.

3.4 Model evaluation

Model evaluation is a way of judging the productivity of an ML framework on certain data. Model evaluation allows us to compare different frameworks, choose the ideal framework, and identify the strengths and weaknesses of a model. A regression-type GBR model can be evaluated with various measures relying on the type of loss function and the objective of the framework. This investigation employs the following common measures:

RMSE: This is the square root of the mean of the squared disparities between the actual and expected values. It gauges the extent of the error and assigns a greater cost to larger mistakes than to small ones. A smaller value of RMSE showcases a better fit for the framework. It is calculated by the following formula:

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

where the predicted ($\hat{y}_i$) and actual (y_i) values indicate larger errors than smaller ones, offering a gauge of the magnitude of the error.

MAE: The mean of the absolute variations between the values that were anticipated and those that were seen. Compared to RMSE, MAE measures average error and is less susceptible to outliers. A lower MAE denotes a better model fit. The formula that follows determines the MAE.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

where MAE between the expected and actual values is the mean of those discrepancies, it is less susceptible to outliers than RMSE and estimates the average error.

R-squared score (R2): This is the proportion of the variance in the output variable explained by the framework, and in this regard:

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

where R2 displays the proportion of the variance in the output variable expressed by the framework, y_i displays the actual value, $\hat{y}_i$ displays the expected value, and $\bar{y}_i$ displays the mean of the actual values.

MAPE: This is the average of the absolute % disparities between the forecast and actual figures [21]. When comparing frameworks across different scales of output variables, which measures relative errors is useful. The smaller the MAPE, the better the fit of the framework. The MAPE is given by the following formula:

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{|y_i|}\right) \times 100 \quad (4)$$

where MAPE is a measure of relative error, useful for comparing frameworks across diverse scales.

Mean squared logarithmic error (MSLE): It calculates the average of the squared logarithmic disparities between the actual and expected values. It is especially useful for frameworks where the target variable is positive and spans a wide range or displays exponential growth. MSLE measures the error as a ratio, and a lower MSLE signifies a better model fit. The MSLE is computed using the following formula:

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(1+\hat{y}_i) - \log(1+y_i))^2 \quad (5)$$

where MSLE can work well on frameworks that predict exponential growth or positive values across a wide range.

Explained variance score (EVS): This displays the average of the squared logarithmic disparities between the actual and forecasted values. For frameworks whose forecasts are positive with a wide range of exponential growth, it is helpful to compute the ratio of the error. A lower MSLE is an indication of a better model fit. The MSLE is computed based on the formula below:

$$EVS = 1 - \frac{Var(y - \bar{y})}{Var(y)} \quad (6)$$

where EVS is the ratio of the variance of the projected values to the variance of the actual values, it assesses how well the framework preserves the variation in the data, ranging from 0 (no fit) to 1 (perfect fit).

4 Outcomes and performance evaluation

This segment displays the outcomes and discussion. In the experimental outcomes, the GBR algorithm predicts medical insurance expenses based on individual attributes. Extensive data preprocessing was carried out, including the handling of categorical variables and separating the database into training and testing groups. Key regression metrics, such as MAE, MSE [19, 20], MAPE [21] and RMSE [22] were employed to evaluate model performance. The outcomes reveal that GBR surpassed other frameworks regarding accuracy and predictive capability.

4.1 Experimental outcomes

To confirm the accuracy of the generated framework, the below outcomes present the targets' prediction outcome, charges versus actual charges. Besides, in this paper, the following graphs will be explained and illustrated for model evaluation and assessment: Predicted charges vs. actual charges, residuals, and distribution of residuals. These graphs are vital for model evaluation since they can examine the quality and accuracy of the regression framework. These can bring forth the merits and demerits of the framework and even point out where it is biased or inaccurate. It could also be used to investigate several frameworks to pick the best model that best fits the data and issue at hand. A scatter plot of predicted charges versus actual charges would show the relationship between the actual values of the outcome variable and the projected values of a regression framework.

Besides the identification of outliers or key points, the plot allows an assessment of the framework fit. A 45-degree line displays the complete agreement between projected and actual values and is where the points should fall if the framework is perfect. If the points are far off the line or show a curvilinear pattern, this suggests a poor fit

or a violation of one of the assumptions of regression. A residual plot displays the projected values on the x-axis and the actual residuals minus the expected values on the y-axis.

Normality, homoscedasticity, and linearity of the residuals can all be checked based on this plot. The dots should be haphazardly scattered around the horizontal line at zero, showing no autocorrelation between the residuals and expected values, with mean residuals of zero. Should the points come out in some curvilinear fashion, one may safely infer a nonlinear relationship. A residual distribution plot is such a representation wherein the frequency or probability of the residuals along the y-axis is plotted against values of the residuals along the x-axis. The plot can be used to check the assumption of normality of the residuals. Ideally, the plot should show a bell-shaped curve that is symmetric and centered at zero, which showcases that the residuals are normally spread with a mean of zero and a constant variance. If the plot displays a skewed or flat curve, it may indicate a violation of normality.

Fig. 2 illustrates a scatter plot of predicted vs. actual charges for medical insurance expenses using GBR. The x-axis displays the actual charges, and the y-axis displays the predicted charges; these represent the medical insurance cost. By looking at the points' proximity to the diagonal line, it is evident that the framework has performed well in predicting the charges. Another plot, Fig. 3, displays the scatter plot comparing the predicted and actual charges in color based on their difference. The red dots correspond to the largest, and the blue corresponds to the smallest. This plot allows one to assess how well the framework has performed: the closer the frameworked charges were to the actual charges, the better the framework. In contrast, the larger the difference is, the worse the anticipations of the framework.

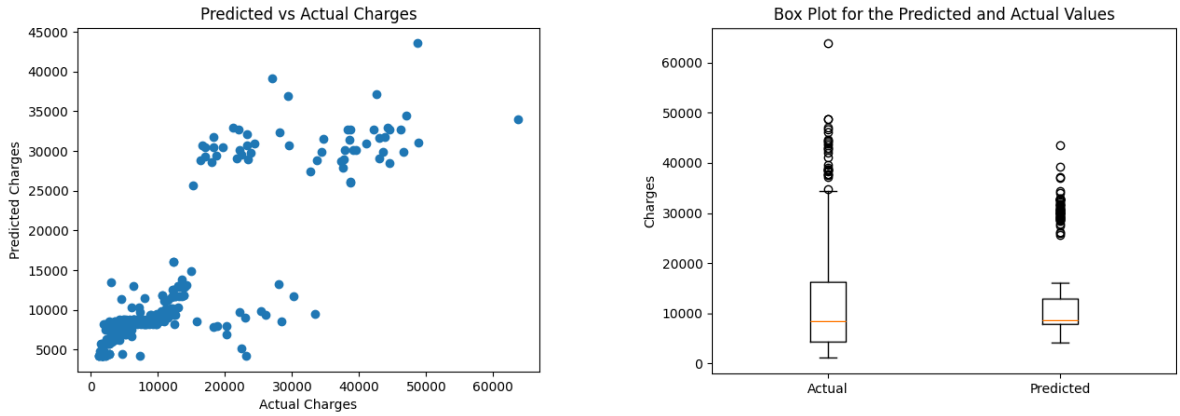

Figure 2: Illustration of predicted charges to actual charges

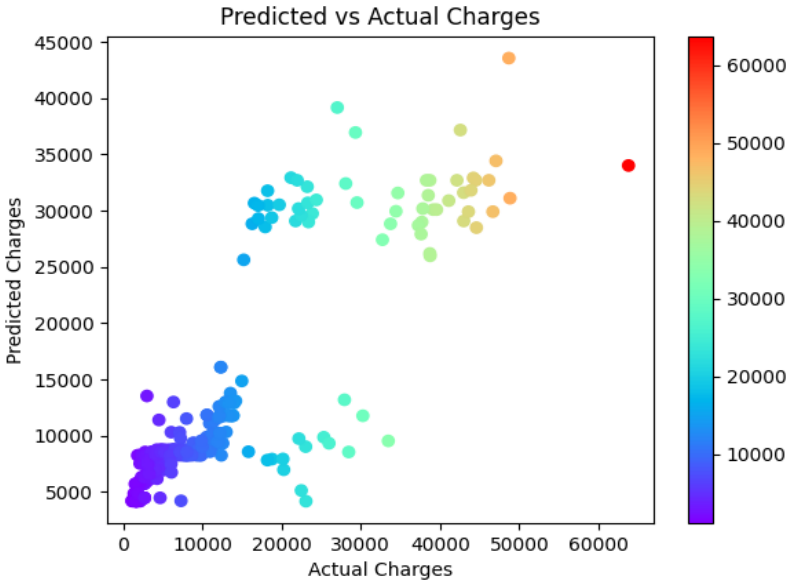

Figure 3: Colorful scatter plot for the predicted charges to actual charges

Fig. 4 is the residual plot, which is a scatter plot that displays the gap between the observed and the projected values of a regression framework—actual charges on the x-axis and residuals on the y-axis. The original range of charges varies between 0 and 60000, and that of residuals between -10000 and 30000. The data points—blue—are

scattered around the red line, which displays the mean residual value. It may be noticed here that the mean residual value is approximately 0, which means model regression fits well on data. Values plugged into this metric are actual charges and the residuals. The actual charges are the measured values of the medical insurance cost from the data. Subtracting the expected values from the actual values permits calculating the residuals or prediction errors. Residuals are the vertical distances between the data points and the fitted line. A small residual showcases good prediction. A large residual means a bad prediction.

Confidence intervals on scatter plots, such as actual vs. predicted medical insurance charges, give a visual approximation of prediction uncertainty. These intervals, usually determined by methods like quantile regression or bootstrapping, set a confidence interval (95%) around the regression, reflecting where future projections will likely lie. For example, a tight confidence band around the 45o line (ideal prediction) would represent high model confidence, whereas wider bands demark intervals of higher uncertainty. In our research, including confidence intervals would add to the reliability interpretation of the value of the GBR model, especially for outlying charge values, by quantifying the impact of variations in input variables (such as age, BMI) on prediction intervals. This is consistent with the demand for transparency within real-world applications, where stakeholders need not only point values, but quantification of risk, or uncertainty, of cost projections.

The Residual plots examine model assumptions by evaluating the distribution of prediction errors. The inclusion of confidence bounds—often based on the standard error of residuals—facilitates examination of deviations from normality as well as homoscedasticity (constancy of variance). For instance, residuals consistently plotting outside of ±2 standard error bounds could reflect heteroscedasticity or nonlinear patterns not captured by the model of the GBR. In our study, residual confidence bounds would ensure that model errors are randomly distributed (as assumed) rather than systematically biased, such as under- or over-prediction within certain charge ranges. It is essential for promoting the model's stability, especially within healthcare finance, where underestimation of cost cases could have substantial fiscal consequences. By estimating residual uncertainty, such bounds improve the interpretive value of model performance as well as its predictive claims.

Fig. 5 is a histogram graph of residuals of the GBR model predicting medical insurance cost. A residuals histogram offers insight into the productivity of a regression framework. This histogram visualizes the distribution of the framework's prediction errors for the outcome variable. The graph below displays the framework's systematic errors and goodness of fit. From Fig. 5 below, this graph showcases that most of the residuals are lying close to 0; this is a positive indication that the framework makes accurate predictions.

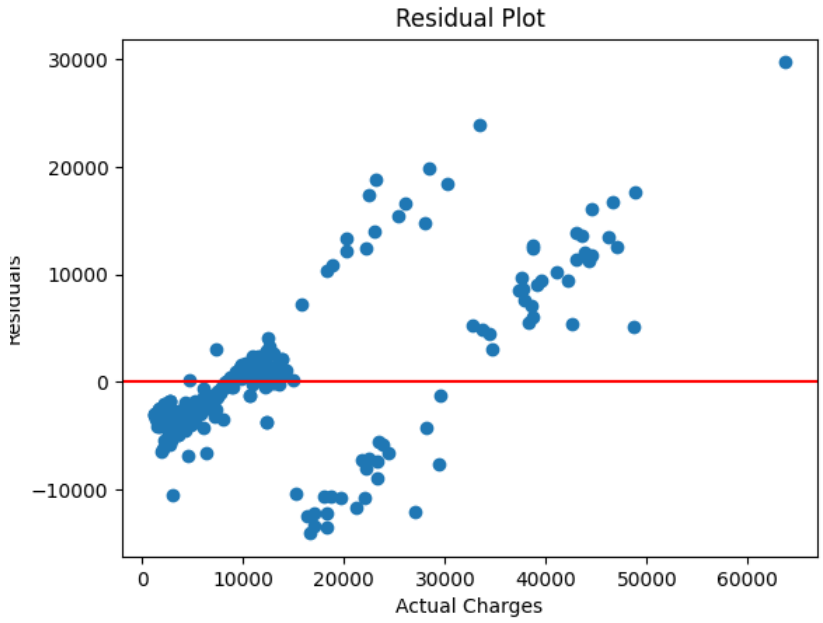

Figure 4: Residual plot of model assessment

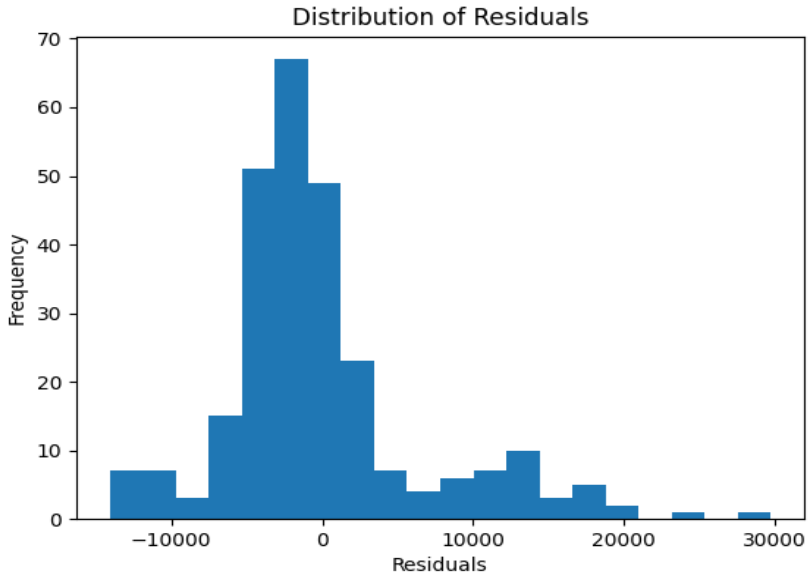

Figure 5.: Histogram of the residual distribution of the generated GBR-based model

Consequently, the generated model using GBR has presented a high level of accuracy regarding the anticipation of medical insurance charges. Correspondingly, obtained outcomes show that predictions of the framework are close to equal real values of insurance charges for people included in the database. Impressive for the accuracy, considering how complex a task it is—there are so many factors determining the expense of medical insurance. A very good example of GBR is how the power of sequential improvement in model performance can come from focusing on the weaknesses of the previous frameworks. These outcomes instill confidence in the reliability of a GBR model to deliver effective outcomes concerning medical insurance charge prediction, making it strong and accurate for estimating healthcare expenses based on individual attributes.

4.2 Comparative analysis

A thorough comparison of performance was conducted utilizing Precision, Recall, and F1-score as metrics. This was done to comprehensively and fairly evaluate the predictive performance of six different algorithms in a regression task. These 3 metrics are selected because they are quite efficient in measuring the classification accuracy of frameworks in the transformed binary classification scenario, where the regression problem is converted into a binary problem for the purpose of comparative analysis. Recall, F1-score, and precision all provide a balanced view of the framework's capability in identifying true positives, making accurate positive predictions, and desirably balancing two. By using such metrics, the comparison covers not only the overall predictive capability of the algorithms but also ensures that their performance is reviewed with nuance on multiple dimensions and hence adds to an insightful and equitable benchmarking process.

As depicted in Fig. 6, the graph displays the precision of 6 different classification algorithms for medical insurance cost prediction. This graph displays the precision metric for the algorithms. Precision is a measure of how accurately the algorithm can identify the true positive cases among all the positive predictions. The algorithms compared are GBR, Linear Regression, RF, Support Vector, K-Nearest Neighbors, and Neural Network. These are some of the common ML techniques used for classification problems.

The highest precision is achieved by GBR, Linear Regression, and RF, all with a precision of 0.908. This means that these algorithms correctly predicted the insurance cost category for 90.6% of the cases among all the cases they predicted as positive. The lowest precision is achieved by Neural Network, with a precision of 0.516. This means that this algorithm correctly predicted the insurance cost category for only 51.6% of the cases among all the cases it predicted as positive. The names of the algorithms are depicted on the x-axis, and the precision values, which range from 0 to 1, are depicted on the y-axis. The algorithm performs better the greater the precision is.

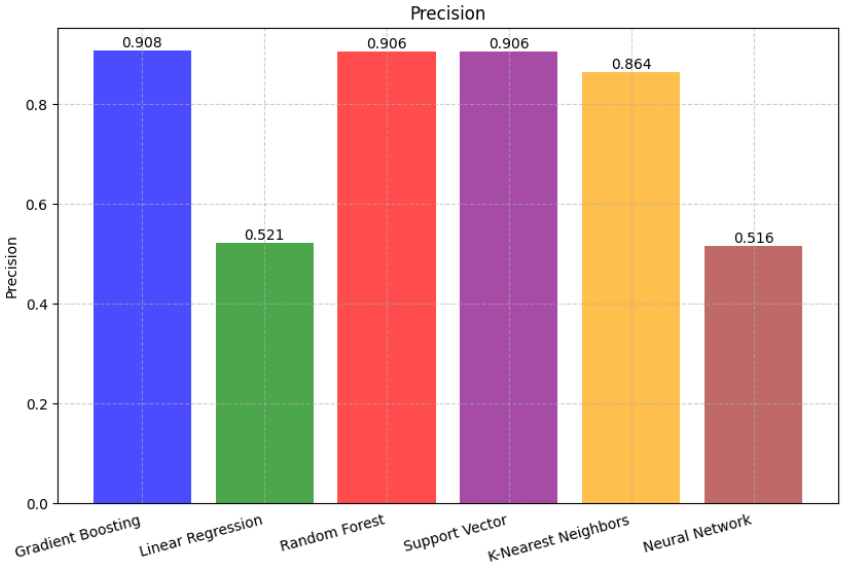

Figure 6: Precision for the regression frameworks

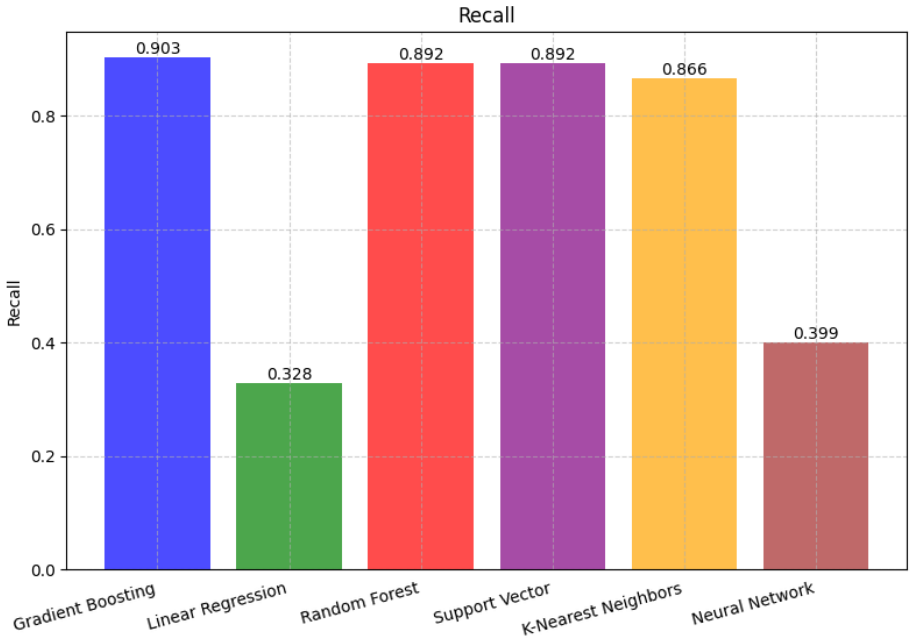

Figure 7: Recall for the regression frameworks

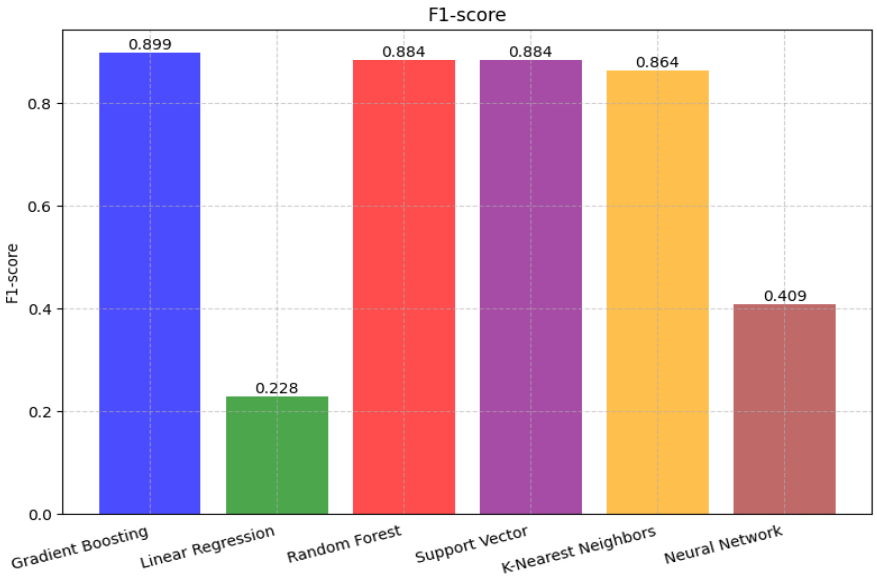

Figure 8: F1-score for the regression frameworks

Fig. 7 is a graph representing the Recall metric for the algorithms. The Recall is a measure of how accurately the algorithm can identify all the positive cases in the data set. It has compared GBR, Linear Regression (LR), RF, Support Vector, KNN, and Neural Network. Some of them are popular ML techniques used in regression problems.

Among these, GBR has the highest Recall of 0.903. It means that this algorithm correctly predicted the insurance cost category for 90.3% of the positive cases in the data set. The lowest Recall is given by the Neural Network with a recall of 0.399. That means that this algorithm correctly predicted the category of insurance cost for only 39.9% of those cases that are positive in the database. Here, the name of the algorithms is depicted on the x-axis, and recall values between 0 and 1 are depicted on the y-axis. The more the recall value, the better the algorithm is performing.

Fig. 8 displays the F1-score of 6 different regression frameworks in medical insurance cost prediction. The F1 score is a measure of the extent to which the algorithm can balance both precision and Recall. It is defined as the harmonic mean of precision and Recall, ranging between 0 and 1. This is where higher is better. Some of the compared frameworks include GBR, LR, RF, Support Vector, KNN, and Neural Network. These are some of the common ML techniques that are used for regression problems.

GBR contributes the best F1 score, which is 0.899. This means that this algorithm has the best balance between precision and Recall among all the frameworks and can anticipate most of the cases correctly within this data set for the insurance cost category. Coming to the poor performers, Neural Network contributes the lowest F1-score, which is 0.409. That is to say, such an algorithm would have the poorest precision-recall balance among all frameworks and thus would correctly predict the insurance cost category for only a few cases in the data set.

4.3 Feature important analysis

Feature importance analysis is the methodology of measuring each input feature (such as age, BMI, smoking status) contribution to the predictive accuracy of the GBR model. This analysis makes use of the natural ability of GBR for ranking features by their impact on minimizing prediction error (such as mean squared error) while being trained. Table 2 presents the feature importance rankings for medical insurance cost prediction.

The feature importances listed in Table 2 highlights smoking status as the strongest predictor of medical insurance costs, responsible for 38.2% of the predictive power of the model. This aligns with established clinical evidence for the linkage of smoking with long-term disorders (e.g., cardiovascular disease, cancer) that significantly raise healthcare consumption and costs. Age (24.5%) is next as the second most important factor, capturing the gradual nature of health decline along with age-graded comorbidities, naturally accelerating medical expenditures. BMI (21.8%) is third, with obesity as a cost driver through its linked diseases, such as diabetes and hypertension. These three features—smoking status, age, and BMI—jointly capture more than 84% of the model's predictivity, indicating their centrality for actuarial risk profiling. Region (7.1%) and children count (5.3%) display moderate though not inconsequential impacts, with interregional heterogeneities of healthcare costs as well as family coverage demands as contributing factors. Gender (3.1%) plays a weak direct role, indicating biological sex as a weak cost driver per se within the current dataset, though its interplay with other features (for example, gender-specific health risks) is still likely indirect cost drivers. These importances support the model's relevance for real-world healthcare realities, providing actionable evidence for insurers to target interventions (for example, cessation of smoking, control of obesity) as well as for policymakers to frame fair distribution policies aimed at high-risk groups.

Table 2: Feature importance rankings for medical insurance cost prediction

Feature	Score	Rank	Description
Smoking Status	0.382	1	Binary indicator of smoking habits (smoker/non-smoker). Strongest predictor of costs due to chronic health risks.
Age	0.245	2	Age of the individual. Older age correlates with higher medical expenses.
BMI	0.218	3	Body Mass Index. Higher BMI linked to obesity-related health complications.
Region	0.071	4	Geographic location (northeast, southeast, etc.). Minor impact due to regional cost variations.
Children	0.053	5	Number of dependents. Slightly increases costs due to family coverage needs.
Gender	0.031	6	Biological sex. Minimal direct impact on costs in this dataset.

5 Discussion

The accuracy of the Gradient Boosting Regressor (GBR) model, as shown through the current research, makes it a better predictive model for medical insurance cost prediction compared to control models like Linear Regression, Random Forest, Support Vector Regression, K-Nearest Neighbors, and Neural Networks. The GBR model had such a high Precision of 0.908, Recall of 0.903, and F1-score of 0.899, beating all other architectures. These scores, for instance, are quite remarkable compared to the Neural Network model, which had significantly lower values (Precision: 0.516, Recall: 0.399, F1-score: 0.409). Moreover, the GBR model had low Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) values, indicating its accuracy in making accurate predictions. Linear Regression, for its part, had considerably high values for MAE, MSE, and RMSE, further indicating the superiority of the GBR model in dealing with the complexities, including nonlinearities, of the data under study. This quantitative comparison highlights the capacity of GBR for dealing with the complexities of medical insurance cost prediction.

There are several reasons for the comparative advantage of the GBR model over other techniques. To begin with, the sequential learning of GBR, where it continuously improves by focusing on the errors made by individual weak learners, enhances overall accuracy. This aspect is particularly helpful in cases where datasets are characterized by complex feature relationships, as is the case with medical insurance cost prediction. In addition, the capacity of GBR for adapting both linear as well as nonlinear relationships without explicit assumption of the distribution of the data puts it a step ahead of other models such as Linear Regression, where there is an assumption of a linear relationship between target variables and features. Another factor is the precise calibration of hyperparameters, including the estimators (`n_estimators`), learning speed, as well as max tree depth (`max_depth`), which have been optimized during model training. However, models such as those of Neural Networks have a tendency of being prone to overfitting as well as underfitting because of susceptibility to architectural choices as well as the choice of its elements. These understandings of the underlying processes of GBR provide an explanation for its effectiveness in detecting the complex patterns that exist in medical insurance cost data.

The applications of the findings of the GBR model transcend theoretical refinement and have strong real-world impacts. Proper prediction of medical insurance expenses is essential for individuals, insurers, and policymakers alike for informed decision-making around healthcare funding. Insurers, for example, can apply the GBR model for designing fairer structures of premiums that align with actual risk profiles, while individuals can apply such projections for well-planned commitments. Policymakers, by contrast, can apply the model for efficiently allocating resources and making policies for reducing healthcare inequities. In addition, interpretability of the outputs of the GBR through feature ranking offers actionable information on the drivers of medical insurance expenses, including age, BMI, and smoking status. Transparency not only builds trust with the model, but it can also enable targeted interventions for reducing healthcare costs. The real-world applicability of the GBR model, therefore, makes it a revolutionary tool for healthcare funding.

In spite of its strengths, the approach as outlined is not free from limitations. Foremost among its drawbacks is the overreliance on a single dataset, based on the Kaggle one, containing 1,339 instances with seven features. Although the dataset provides helpful information, its relatively small sample size and homogeneity might limit the model's generalizability across disparate populations and situations. Subsequent scholarly research must include the use of larger, more heterogeneous datasets so that the model is validated across disparate settings and populations. Furthermore, likely nuances within the population included within the dataset, such as geographical or demographic bias, can affect model performance. For instance, the dataset is overwhelmingly made up of people living in certain parts of the United States, leaving the question of how well the model would perform in other geographical or socio-economic settings. Solutions for these issues will involve diligent preprocessing and augmenting techniques for bias reduction as well as increasing representativeness. Another limitation is concerned with the evaluation metrics employed by this study. While metrics including RMSE, MAE, R^2, and F1-score give a well-rounded evaluation of model performance, none of them capture all aspects of predictive modeling. For example, metrics such

as specificity or the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) may provide richer information on the model's ability for classifying cases. Additionally, interpretability and explainability, which are essential within domains such as insurance where model outputs have direct effects on the well-being of individuals, are not addressed significantly within the current framework. Interpretable machine learning methods or model-agnostic interpretability tools must be considered by future research as a means of bridging the gap. Finally, external variables such as economic downturn, policy, and healthcare technology developments have not been included within the model, potentially affecting its predictability. Adding further sources of information as well as time-varying features can make the model more adaptable as well as up to date for changing healthcare environments. By taking such limitations into consideration, further research can improve the GBR-based method as well as the medical insurance cost predictability field, overall.

6 Conclusion

This paper underscores the crucial importance of precise medical insurance cost prediction, facilitating informed decision-making and efficient resource allocation within the evolving landscape of healthcare financing. Despite extensive exploration of ML-based approaches in the literature, the ongoing imperative is to refine further their efficacy and precision for predicting medical insurance expenses, demanding continued investigation for more effective methods to enhance accuracy rates in these pivotal tasks. To address this research gap, this investigation introduces an ML approach utilizing the GBR, aiming to augment accuracy and efficacy in medical insurance cost predictions.

This work has significant limitations that should be addressed, notwithstanding the interesting findings from the comparison of regression frameworks for forecasting insurance prices. First off, the study is dependent on a particular database, and the features of this database may limit how broadly the findings may be applied. Incorporating varied databases into future studies might improve the findings' robustness and expand the application of the suggested frameworks to a range of insurance scenarios.

The Linear Regression framework's presumption of a linear connection between the characteristics and the target variable is another drawback. A linear model might not be able to adequately describe the intricate and nonlinear character of insurance charge prediction. To increase accuracy, future studies might investigate nonlinear frameworks or more complex linear regression algorithms.

Even though they are extensive, the assessment measures included in this investigation may not account for all facets of model performance. For example, new measurements such as specificity or area under the receiver operating characteristic curve (AUC-ROC) might offer a more detailed insight into the capabilities of the framework. Future research ought to think about using a wider range of measures to guarantee a comprehensive assessment.

Moreover, the study primarily focuses on predictive modeling, and factors such as interpretability and explainability of the frameworks are not extensively discussed. Addressing the interpretability of frameworks is crucial, especially in domains like insurance, were model decisions impact individuals' financial well-being. Future studies may also be conducted on more interpretable ML frameworks or model-agnostic interpretability techniques.

Interesting work for the future might consider investigating the performance improvement due to feature engineering. One-hot encoding was applied to categorical variables in this work, and using different encoding methods or transformations of features may lead to different outcomes. Further work on the ways of performing feature engineering and the changes that this might bring about in model performance could be of great benefit.

Apart from that, other factors such as economic, regional, or policy changes in regard to healthcare can be incorporated to enhance the predictability of the frameworks further. The integration of other external sources of data for capturing the broader perspective of factors influencing insurance charges will be an area of future studies.

Acknowledgement

This work was supported by the project of Research and Practice on the Construction Path of Big Data Specialty Group for the Integration of Industry and Education, and the Promotion of Five Doubles in Higher Vocational Colleges (No.2021SJGLX675)

References

[1] H. M. Alzoubi, N. Sahawneh, A. Q. AlHamad, U. Malik, A. Majid, and A. Atta, “Analysis of Cost Prediction in Medical Insurance Using Modern Regression Models,” in *2022 International Conference on Cyber Resilience (ICCR)*, IEEE, 2022, 1–10. DOI:10.1109/ICCR56254.2022.9995926

[2] L. Guo *et al.*, “Changes in direct medical cost and medications for managing diabetes in Beijing, China, 2016 to 2018: electronic insurance data analysis,” *The Annals of Family Medicine*, 19(4): pp. 332–341, 2021. https://doi.org/10.1370/afm.2686

[3] Y. A. Christobel and S. Subramanian, “An empirical study of machine learning regression models to predict health insurance cost,” *Webology*, 19(2):2022. DOI:10.3390/ijerph19137898

[4] H. Zhang, W. Zhou, and D. Zhang, “Direct medical costs of Parkinson’s disease in Southern China: a cross-sectional study based on health insurance claims data in Guangzhou City,” *Int J Environ Res*

Public Health, 19(6); 3238, 2022. https://doi.org/10.3390/ijerph19063238

[5] H. J. Kan, H. Kharrazi, H.-Y. Chang, D. Bodycombe, K. Lemke, and J. P. Weiner, "Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults," *PLOS One*, 14, (3): e0213258, 2019. https://doi.org/10.1371/journal.pone.0213258

[6] M. Hanafy and R. Ming, "Machine learning approaches for auto insurance big data," *Risks*, 9, (2): 42, 2021. https://doi.org/10.3390/risks9020042

[7] N. Shakhovska, N. Melnykova, and V. Chopiyak, "An Ensemble Methods for Medical Insurance Costs Prediction Task.," *Computers, Materials & Continua*, 70(2), 2022. DOI: 10.32604/cmc.2022.019882

[8] M. Praveen, G. S. Manikanta, G. Gayathri, and S. Mehrotra, "Comparative Analysis of Machine Learning Algorithms for Medical Insurance Cost Prediction," in *International Conference on Innovative Computing and Communication*, Springer, 2023, 885–892. https://doi.org/10.1007/978-981-99-3315-0_68

[9] T. Pfutzenreuter and E. de Lima, "Machine Learning in Healthcare Management for Medical Insurance Cost Prediction," in *OPEN SCIENCE RESEARCH II*, Editora Científica Digital, 2022, 1323–1334. https://doi.org/10.1016/j.mlwa.2023.100516

[10] K. Y. Ngiam and W. Khor, "Big data and machine learning algorithms for health-care delivery," *Lancet Oncol*, 20(5): e262–e273, 2019. DOI:10.1016/S1470-2045(19)30149-4

[11] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, "Machine learning approaches for predicting high-cost high need patient expenditures in health care," *Biomed Eng Online*, 17, 1–20, 2018. https://doi.org/10.1186/s12938-018-0568-3

[12] C. A. ul Hassan, J. Iqbal, S. Hussain, H. AlSalman, M. A. A. Mosleh, and S. Sajid Ullah, "A computational intelligence approach for predicting medical insurance cost," *Math Probl Eng*, 2021(1): 1162553, 2021. https://doi.org/10.1155/2021/1162553

[13] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine learning-based regression framework to predict health insurance premiums," *Int J Environ Res Public Health*, 19(13):7898, 2022. https://doi.org/10.3390/ijerph19137898

[14] A. I. Taloba, R. M. Abd El-Aziz, H. M. Alshanbari, and A.-A. H. El-Bagoury, "Estimation and prediction of hospitalization and medical care costs using regression in machine learning," *J Healthc Eng*, 2022(1): 7969220, 2022. https://doi.org/10.1155/2022/7969220

[15] R. Zea-Vera *et al.*, "Development of a machine learning model to predict outcomes and cost after cardiac surgery," *Ann Thorac Surg*, 115(6): 1533–1542, 2023. https://doi.org/10.1016/j.athoracsur.2022.06.055

[16] L. Arvidsson, M. Landgren, A. K. Harding, A. Abramo, and M. Tägil, "Patients Aged 80 or More With Distal Radius Fractures Have a Lower One-Year Mortality Rate Than Age-and Gender-Matched Controls: A Register-Based Study," *Geriatr Orthop Surg Rehabil*, 15, 21514593241252584, 2024. https://doi.org/10.1177/21514593241252583

[17] A. V Konstantinov and L. V Utkin, "Interpretable machine learning with an ensemble of gradient boosting machines," *Knowl Based Syst*, 222, 106993, 2021. https://doi.org/10.1016/j.knosys.2021.106993

[18] T. Zhang *et al.*, "Improving convection trigger functions in deep convective parameterization schemes using machine learning," *J Adv Model Earth Syst*, 13(5): e2020MS002365, 2021. https://doi.org/10.1029/2020MS002365

[19] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Compute Sci*, 7, e623, 2021. https://doi.org/10.7717/peerj-cs.623

[20] F. M. Butt, L. Hussain, A. Mahmood, and K. J. Lone, "Artificial Intelligence based accurately load forecasting system to forecast short and medium-term load demands," *Mathematical Biosciences and Engineering*, 18(1): 400–425, 2021. Doi:10.3934/mbe.2021022.

[21] Li, Li. "Dynamic Cost Estimation of Reconstruction Project Based on Particle Swarm Optimization Algorithm." *Informatica* 47.2 2023. Doi: https://doi.org/10.31449/inf.v47i2.4026.

[22] Yang, Xueqin. "Economic Cost Prediction Model for Building Construction Based on CNN-DAE Algorithm." *Informatica* 49.5, 2025. Doi: https://doi.org/10.31449/inf.v49i5.7029