Grenze International Journal of Engineering and Technology, January Issue

Improved ML Model and XAI Methods for Medical Insurance Premium

Laxminarayanamma K[1], Chakradhar B[2], Govardhan R[3] and Bhagya Chandra Raj V[4]
[1-4]Institute of Aeronautical Engineering, Hyderabad, India
Email: k.laxminarayanamma@iare.ac.in, bchakradhar2000, govardhan.reddiboina, bhagyachandraraj}@gmail.com

***Abstract*—This paper explores the integration of a machine learning model within a web-based application for the real-time prediction of healthcare insurance premiums based on individual health data.**
Utilizing a RandomForest regression model initialized with a set of hyperparameters. To optimize the model's performance, a GridSearchCV is employed. The model claims an R2 score of 0.901, the study aims to enhance the accuracy and efficiency of insurance premium estimations, offering a scalable solution that can significantly improve decision-making processes in the insurance sector.
The model was trained and evaluated on a dataset comprising health attributes from 986 individuals, achieving robust performance metrics.
Subsequently, the model was embedded into a Flask web application, providing a user-friendly interface for instant premium estimations.
The deployment on a server facilitated the efficient handling of multiple user requests with minimal latency.
This work also implements the Methods in Explainable AI (XAi) designed to clarify how algorithms make decisions, ensuring that patients, healthcare managers, and insurers can all understand and trust these processes.
This clarity is vital because it helps reveal and address any biases that may be concealed within the inherently opaque operations of these algorithms.

***Index Terms*—RandomForest, Explainable AI, and Flask.**

I. INTRODUCTION

Medical insurance is an essential element of the healthcare sector, enabling individuals to receive needed medical care without shouldering the entire expense.
Calculating medical insurance premiums involves a detailed process that considers various factors including age, gender, pre-existing health conditions, and lifestyle choices.
The objective of this project is to provide individuals with an estimation of the necessary premium based on their personal health profile.
This estimation will enable them to compare and consider different insurance plans and benefits offered by various companies, armed with knowledge about the baseline cost relevant to their situation. [1] “A number of measures, including R-squared (R2), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and mean absolute percent error (MAPE)”, will additionally be used in this research to evaluate the precision and efficacy of its models.

Grenze ID: 01.GIJET.11.1.324

II. Literature Review

A. Machine Learning for an Explainable Cost Prediction of Medical Insurance

In their paper, the authors implemented three regression-based ensemble machine learning models—"Extreme Gradient Boosting (XGBoost), Gradient-boosting Machine (GBM), and Random Forest (RF)—to predict medical insurance costs" [1]. They employed Individual Conditional Expectation (ICE) plots in addition to SHapley Additive exPlanations (SHAP) techniques, which are forms of Explainable Artificial Intelligence (XAI), to determine and elucidate the primary drivers of medical insurance prices in their dataset. Performance measures are used to assess the models.

The results indicated that while all models performed well, the XGBoost model achieved superior overall performance despite higher computational demands. Conversely, the RF model demonstrated lower prediction error, with an R^2 value of 84.046, and required significantly less computational resources compared to XGBoost (Ugochukwu Orji et al., 2024).

B. Prediction of Medical Premium Price

In this study, the authors implemented ten machine learning models: Bagging (IO variables), Single Regression Tree, Random Forest, Boosting, Backward Selection, Lasso Regression, Forward Selection, Ridge Regression, Principal Component Regression (PCR) with 9 directions, and PCR with 5 directions. Among these models, Bagging explained the highest percentage of premium price variability at 79.94%, while PCR with 5 components explained the lowest at 43.56%. The authors suggest that further analysis could involve tuning the parameters of the Bagging model, such as adjusting the number of trees (Elena Denner et al., 2021).

C. Medical Insurance Premium Prediction Using Regression Models

In this work, "we assessed the accuracy of four different machine learning models—linear regression, ridge regression, support vector machine (SVM), and random forest regression—in forecasting health insurance prices" [1]. These models were trained and tested on a dataset comprising demographic data and medical insurance claims. To improve model performance, preprocessing and feature engineering were used to the data. We evaluated and compared the models' efficacy using mean squared error (MSE) and R-squared [3].(Kamma Lakshmi Narayana et al., 2023).

D. A Survey on Explainable Artificial Intelligence (XAI).

As we enter the era of the fourth industrial revolution, the swift and extensive integration of artificial intelligence (AI) into our everyday lives is driving a substantial transformation towards a more algorithm-driven society. Despite these remarkable advancements, a significant challenge to the broad adoption of AI-based systems remains their opacity [7]. These systems, often referred to as black boxes, offer potent predictive capabilities but lack transparency and direct explainability [7]. This has reignited the discussion of whether explainable AI (XAI) is necessary [7].

The topic of XAI research has great potential to improve the transparency as well as trustworthiness of AI-powered systems. It is acknowledged that this is necessary for AI to advance steadily and unhindered [7]. Through a comprehensive review of the literature, we examine existing approaches, discuss emerging trends, and outline major research trajectories in the field of XAI. (AMINA ADADI et al., 2018)

III. Existing Work

Numerous studies have focused on predicting medical insurance premiums, often achieving prediction accuracies ranging between 70% and 80%. Despite these advances, many of these solutions suffer from limited usability and lack the user-friendly interfaces necessary for practical application by customers. As a result, customers often remain unaware of the specific factors and processes involved in determining their insurance premiums. Furthermore, inefficiencies in the calculation methods can lead to inaccuracies, with many claims being processed incorrectly. This highlights the need for more transparent, efficient, and user-centric solutions in the field of medical insurance premium prediction

IV. Proposed Work

In this study, we propose the integration of a machine learning model within a web application to enable real-time prediction of healthcare insurance premiums using individual health data. the RandomForest regression model, optimized with GridSearchCV, is chosen for its suitability for real-time prediction tasks. The system is trained on health data from 986 individuals and embedded into a Flask web application, providing a user-friendly interface for instant premium estimations. Furthermore, Explainable AI (XAI) techniques are incorporated to ensure the transparency of the model. This comprehensive approach aims to enhance decision-making accuracy and efficiency within the insurance sector.

V. METHODOLOGY

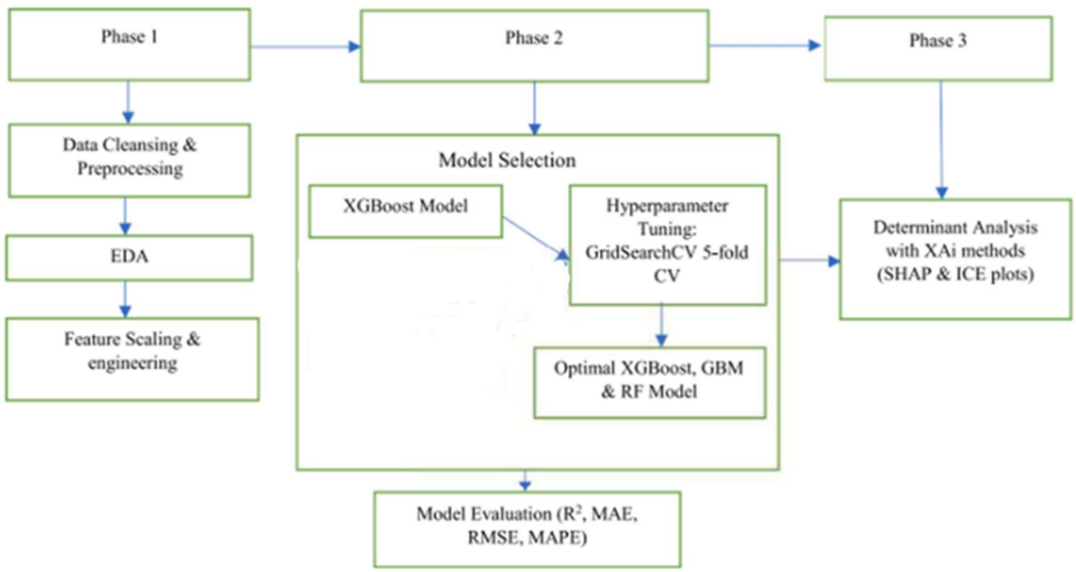

Figure 1. Methodology of insurance premium prediction system

A. Setup for the experiment

The experiments were conducted using Python libraries on VS code on an AMD Ryzen 5 5600H with Radeon RX 5500M Graphics, and 24GB RAM.

B. Source of data and preparation

"The 986 records and 11 attributes/features that make up the medical insurance cost dataset used in this study were taken from Kaggle's repository" [1]. Thorough data pretreatment was done before analysis began, and no duplicates were found despite checks for missing and duplicate variables. Furthermore, the Standard Scalar approach was used to normalize the data [1].

The original publication date of the Kaggle.com dataset, "Medical Insurance Premium Prediction," was August 4, 2021[2]. The simulated annual medical coverage expenses for 986 consumers are included in this dataset. The premium amounts are given in Indian Rupees or around 0.013 US dollars per INR. Data from the customer comprises ten variables relating to health: Height (cm), Weight (kg), Family History of Cancer (0 or 1), Diabetes Status (0 or 1), Blood Pressure Issues (0 or 1), Any Major Organ Transplants (0 or 1), Any Chronic Diseases (0 or 1), Age (years), and Number of Major Surgeries (0, 1, 2, or 3) [2].

C. Analyzing exploratory data

A data analysis methodology called exploratory data analysis (EDA), which commonly makes use of visual tools, is used to highlight the salient elements of a dataset [1].

D. Machine learning model

Following preprocessing, all features were retained for the experiment. "After that, the dataset was split into training and test sets, with 20% of the data going toward testing and the remaining 80% going toward training" [2]. The training dataset was used to create regression models that predicted medical insurance costs thanks to this partitioning, and the test dataset was used to assess the models' performance [1].

Ensemble machine learning methods, such as Random Forest, are renowned for their ability to mitigate the instability associated with individual models by averaging predictions from multiple trees. This characteristic makes ensemble models particularly well-suited for healthcare predictive modeling, offering robustness and reliability in predicting medical insurance costs.

The code defines a grid of hyperparameters for the `RandomForestRegressor` model. “This grid includes multiple configurations for n_estimators (number of trees in the forest), max_depth (maximum depth of each tree), min_samples_leaf (minimum number of samples required at a leaf node), and min_samples_split (minimum number of samples necessary to split a node)” [8].

To find the optimal hyperparameters, `GridSearchCV` is employed. This approach uses a 5-fold cross-validation process to assess the model's performance for every combination after doing an exhaustive search across the given parameter grid. To maximize the R-squared metric—a measure of the percentage of variation in the dependent variable that can be predicted from the independent variables—the {scoring} parameter is set to {r2}.

E. Model evaluation approach

R-squared:

“Calculates the percentage of the dependent variable's variation that can be predicted based on the independent variables” [1].

Mean Absolute Error:

Calculates the average error size over a series of forecasts without taking direction into account [1].

Root Mean Squared Error (RMSE):

“Calculates the average squared difference between the expected and actual values, or the square root of that difference” [1].

F. Explainable AI (XAi)

While machine learning (ML) models are generally known for their reliability and efficiency, their complexity often renders them as black-box models. “This complexity poses challenges in quantifying the contribution of each feature to the model's predictions “[9]. In machine learning, the trade-off between interpretability and accuracy is still a major problem, particularly when it comes to the explainability of models (Gunning, 2019). The Explainable Artificial Intelligence (XAI) initiative was started by the Defense Advanced Research Projects Agency (DARPA) in order to overcome this problem [1].

Shapley additive explanations (SHAP)

SHAP is a technique that measures each feature's contribution to the model's output in order to provide light on the mechanisms underlying certain occurrences. Game theory and its tenets, such as giving attributes that don't affect the model's prediction zero attribution, are the foundation of the SHAP concept [1].

Individual conditional expectation (ICE) plots

Partial Dependence Plot (PDP), which depicts the average partial connection between a collection of predictors and the expected response, is the source of Individual Conditional Expectation (ICE) plots [1]. ICE charts visually break down the conventional PDPs, displaying a feature's partial dependency for each instance independently. The PDP is represented as the average of all ICE curves [1].

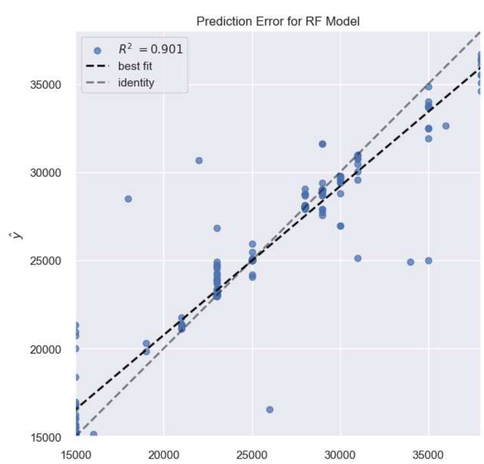

Figure 2. Prediction Error for RF model

VI. RESULT

Hyperparameters are vital in determining the development and performance of machine learning models. Properly tuned models demonstrate superior performance. Grid Search Cross-Validation ('GridSearchCV') was used to determine which parameters were optimal for each machine learning technique. A well-known technique for examining every potential combination of hyperparameters and ensuring the model employs the best values to get a more precise generalization performance estimate is GridSearchCV. This method involves initially setting a range of values for each hyperparameter. Through iterative cycles, the best-performing combination of hyperparameters is identified and applied to the learning process.

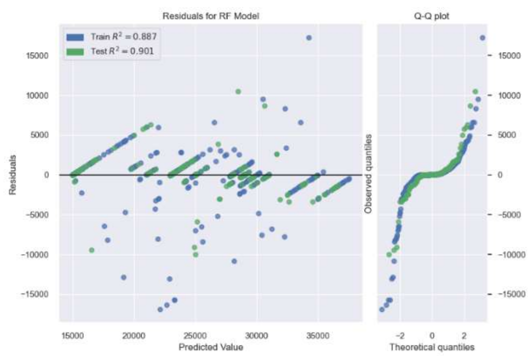

Figure 3. Residuals Plot to Evaluate RF Performance

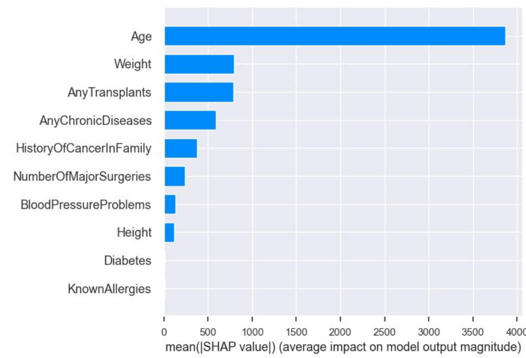

Figure 4. SHAP values of health parameters

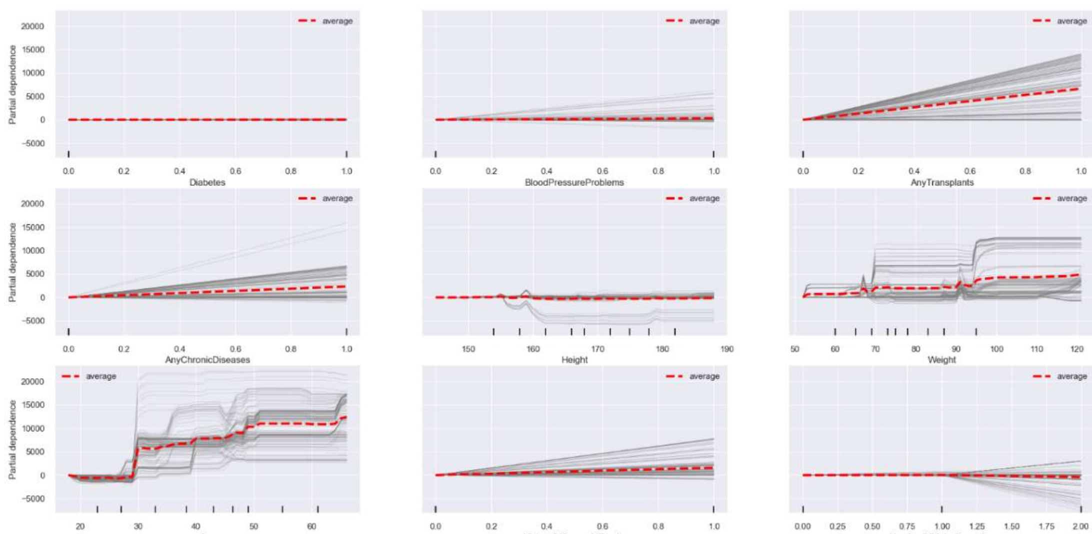

Figure 5. Individual Conditional Expectation (ICE) and Partial Dependence (PD) plot for XAi

TABLE I. RESULTS

Metric	Value
Test R2 score	0.9010125
MAE Score	985.790952
RMSE Score	2054.53205
MAPE Score (%)	4.2616013
Prediction Time (Sec)	0.0427443
Memory Used (MB)	405.87109
CV Score (R2)	0.80704356
Best Parameters	{"max_depth": 20, "min_samples_leaf": 1, "min_samples_split":10, "n_estimators": 500}

VII. CONCLUSION

In this work, we thoroughly assessed an ML model for estimating the cost of insurance premiums. To optimize the model's performance, Grid Search Cross-Validation (GridSearchCV) was utilized for rigorous hyperparameter tweaking during the model's development process. A maximum depth of 20, a minimum of 1 sample per leaf, a minimum of 10 samples needed to divide an internal node and 500 estimators were found to be the optimal hyperparameters.

With a Test R2 value of 0.901, our results showed a high degree of prediction accuracy and the ability of the model to explain 90.1% of the variation in the test data. The accuracy of the model's predictions was demonstrated by the MAE and RMSE, which were 985.79 and 2054.53, respectively. Furthermore, 4.26% for the MAPE highlights the correctness and dependability of the model.

The computational efficiency was also noteworthy, with the grid search process taking approximately 394 seconds and the prediction time being a mere 0.043 seconds. Cross-validation, which yielded an R2 score of 0.807, took around 4.92 seconds, showcasing the model's robustness and generalizability. The memory usage was measured at 405.87 MB, indicating a reasonable resource consumption given the complexity of the model.

Overall, our findings highlight the significance of hyperparameter tuning in enhancing model performance and ensuring accurate predictions. Transparency and interpretability were enhanced with the addition of SHAP values and ICE charts. These findings clear the path for more effective and trustworthy prediction models by demonstrating the potential of sophisticated ML techniques in the insurance sector.

REFERENCES

[1] Ugochukwu Orji, Elochukwu Ukwandu, "Machine learning for an explainable cost prediction of medical insurance" 2024. (references)

[2] Elena Denner, "Prediction of Medical Premium Price" 2021

[3] Kamma Lakshmi Narayana, Yogesh, Putta Kowshik, "Medical Insurance Premium Prediction Using Regression Models" unpublished.

[4] P Sudheer Benarji, Kollipara Praveen Kumar, Manishetty Sushanth, Meghana Gogi, Donuru Neeraja "Minimum Health Insurance Premium prediction using health parameters," 09 Sep 2023.

[5] Preet Jayendrakumar Modi, Vraj Jatin Naik, "Insurance Management with Premium Prediction," (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429 Volume 9 Issue XII Dec 2021

[6] Sahana L B, Anjali B, Sarojini Naik, P M Shreenidhi, "Prediction Of Health Insurance Premium Costs With IBM Auto Ai Service", 2 IJCRT Volume 10, Issue 7 July 2022

[7] Amina adadi and mohammed berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)" 2021

[8] Dario Passos, Puneet Mishra," A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks" 2022

[9] Jerome H, Friedman, "Greedy Function Approximation: A Gradient Boosting Machine" 2001