International Journal of Core Engineering & Management
Volume-7, Issue-11, 2024 ISSN No: 2348-9510

AN EFFECTIVE MACHINE LEARNING BASED REGRESSION TECHNIQUES FOR PREDICTION OF HEALTH INSURANCE COST

Rajesh Goyal
FS - Insurance
IBM - USA
Glen Mill, Pennsylvania, USA
Rajesh.nim@gmail.com

Abstract

Health insurance policies provide financial assistance to cover medical expenses and mitigate the financial impact of illnesses. Various factors contribute the cost of healthcare and health insurance. Predicting health insurance costs early can assist in determining the appropriate coverage amount and identifying potential benefits. The insurance business may benefit from ML's ability to increase policy efficiency. In healthcare, ML algorithms excel at forecasting high-cost medical expenses. The insurance industry may benefit from ML by making insurance program language more effective. This study uses a medical insurance cost dataset obtained from Kaggle, which has 986 records and 11 characteristics, to investigate the potential of ML-based regression algorithms for predicting health insurance premiums. The performance of three regression models—XGBoost, LR, and SVR—is evaluated using R2-score, RMAE, and MSE. After comparing it to the other regression techniques, XGBoost comes out on top. Results demonstrate that XGBoost outperforms both LR and SVR, achieving an R^2 of 86.47 and a significantly lower MAE of 14.42, indicating superior predictive capability. This study highlights the effectiveness of XGBoost in capturing the variance in health insurance costs based on customer attributes, paving the way for future research to explore more advanced techniques and broader datasets for enhanced prediction accuracy.

Keywords: Healthcare, insurance policy, medical cost prediction, machine learning models.

I. INTRODUCTION

An important problem in modern culture is the cost of healthcare. According to data compiled by the WHO, in 2016, healthcare costs throughout the world amounted to almost US$ 7.5 trillion, or around 10% of their GDP [1]. Healthcare cost forecasting by individuals has evolved into a powerful instrument for enhancing healthcare accountability. The healthcare industry generates vast amounts of patient-, disease-, and diagnosis-related data; yet, this data is underutilized and so fails to provide the value it should, especially when considering the financial burden on patients [2].

It is possible to protect one's financial stability from a wide range of risks with a health insurance coverage. There are various factors that influence both insurance and medical expenditures [3]. Many interested parties and health authorities rely on accurate individual healthcare cost estimates provided by prediction models. Accurate cost estimations are useful in helping healthcare delivery organizations and health insurers make long-term plans and priorities the distribution of scarce resources for care management. Additionally, by being aware of their expected future costs in

advance, patients may choose insurance plans with suitable rates and deductibles. The creation of insurance policies is influenced by these factors [4].

In order to construct models that can anticipate rates for new clients, it is necessary to use historical data on people's demographics, health factors, and insurance convergence in order to estimate medical insurance premiums using ML [5]. ML can improve the effectiveness of policy language in the insurance industry. ML algorithms are very effective in the healthcare industry for forecasting high-need, high-cost patient expenses [6][7].

The motivation for this study arises from the increasing complexity and variability of health insurance costs, which can significantly impact individuals and healthcare providers alike. As the healthcare landscape evolves, accurately predicting insurance costs becomes crucial for ensuring affordability and accessibility. Traditional methods often fall short in capturing the intricate relationships between customer attributes and health insurance expenses. By leveraging advanced machine learning-based regression techniques, this research aims to improve a predictive accuracy of health insurance costs, providing valuable insights for insurers and policymakers. The findings could lead to more informed decision-making, enabling stakeholders to better allocate resources and manage risks in an increasingly dynamic market.

1. Contribution of study

This research makes significant contributions to the field of predictive analytics, specifically focusing on health insurance cost forecasting using ML techniques. The main contributions of study are as follow:

- Utilize the medical insurance cost dataset for medical health insurance cost prediction.
- Perform pre-processing for cleaning, scaling, and selecting relevant features, ensuring higher accuracy and robustness in ML models.
- Apply machine learning models like SVR, XGBoost, and logistic regression.
- Evaluate model efficiency with error matrix like MAPE, RMSE, MSE, and R2-score.

2. Structure of paper

The paper's structure is arranged as follows: A summary of earlier studies on the cost of health insurance is given in Section II. The research process is described in Section III, along with the methodology, data pretreatment, and model selection that were used. In Section IV, the models' performance is analyzed and the experimental data is presented. Section V wraps up by summarizing the important discoveries and talking about their consequences.

II. LITERATURE REVIEW

The several machine learning techniques that may be used to predict healthcare costs are described in the recent papers that are included in this section. Some background studies are provide in below:

This study Garmdareh et al., (2023) suggests a novel model that utilizes regression-based ML techniques to forecast the Total Price of a patient's claim using past claims as a basis. The program then compares the anticipated amount to an actual number to determine the price difference. A claim's aberrant or fraudulent charges will be estimated using an absolute price difference criterion. It is decided to look at a set of 99,440 records from the RASA web site. Although DL has the lowest MAE during the training phase, decision trees have the lowest MAE during the testing

phase. That's why the decision tree is used to find anomalies. It can tell that about 17% of records aren't regular if they have at least a 30% variation. Professional human reviewers concur with almost half of the identified issues after reviewing the findings [8].

This study Nabrawi and Alanazi, (2023) create a health model that can instantly spot any fraud in Saudi Arabian health insurance claims. As accurately as possible, the model shows the biggest cause of scam. Three DL and ML methods were used on the labelled skewed dataset. Three healthcare companies in Saudi Arabia gave us the data set. Models such as LR, RF, and ANN were used. The dataset was balanced using the SMOT approach. To remove irrelevant characteristics, they used Boruta object feature selection. Metrics for validation included precision, accuracy, F1 score, specificity, recall, and AUC. The most important characteristics, according to RF classifiers, are policy type, education level, and age. These features achieved 98.21%accuracy, 98.08%precision, 100%recall, 99.03%F1 score, 80%specificity, and 90.00%AUC. Accuracy was 80.36 percent, precision was 97.62 percent, recall was 80.39 percent, F1 score88.17%, specificity was 80%, and AUC was 80.20%; all of these were produced via logistic regression. The results obtained via ANN were as follows: 94.64%accuracy, 98.00%precision, 96.08%recall, 97.03%F1 score, 80%specificity, and 88.04%AUC. Additional research on a bigger dataset is recommended, since all three models used in this predictive analytics study produced satisfactory accuracy and validation metrics [9].

This paper Panda et al., (2022) creates a ML algorithm-based real-time insurance cost prediction system called MLHIPS. This system will help market insurance businesses quickly and easily determine premium values, which will subsequently reduce health expenditure. To predict insurance premiums and evaluate the efficacy of the model, the suggested approach uses a variety of regression models, including Simple Linear, Ridge, Multiple Linear, Lasso, and Polynomial Regression. The Polynomial Regression model succeeded where the others failed, with an R-squared value of 0.80 and an RMSE of 5100.53 [10].

In Fursov et al., (2022) suggest DL architectures for handling insurance data, which includes detailed records of patients' visits and other personal information. The model's quality is enhanced by both the sequential and tabular components, which provide fresh insights into detecting health insurance fraud. The claims management process may be greatly enhanced by our method, as shown by empirical findings obtained using pertinent data from a health insurance company, which surpass advance models. A top competitor using modern models gets a ROC AUC value of 0.815, whereas we get 0.873. They further show that our designs are more resistant to corrupted data. Our methods are going to be useful for a lot of similar applications when insurers start to have access to more semi-structured event sequence data. This is especially true for variables with a lot of categories, like those in the ICD codes or other classification systems [11].

This research Nur Prasasti, Dhini and Laoh, (2020) create a prediction model that uses ML to identify motor insurance fraud. The research made use of actual data from an Indonesian motor insurance provider. The dataset exhibits a significant imbalance in the distribution of valid data versus data from policyholders who commit fraud. This study uses under sampling techniques and the SMOTE to address the unbalanced dataset issue. The supervised classifiers that have been recommended include RF, DT, and MLP. One way to measure how well a model works is by looking at its sensitivity, confusion matrix, and ROC curve. Among the classifiers evaluated, RF performed the best with a 98.5% accuracy rate [12].

The suggested work's Jyothsna et al., (2022) predict an individual's insurance expenses and to locate people, regardless of health issues, who have health insurance plans and medical records.

International Journal of Core Engineering & Management
Volume-7, Issue-11, 2024 ISSN No: 2348-9510

This research used a variety of regression models, including Multi-Linear, DT, RF, and GB Regression. The results showed that, with an accuracy of 87%, GB was the best strategy out of the bunch. It all comes down to training the Telegram-integrated chatbot to speak with the user and estimate the insurance premium using the optimal model[13]. The background study comparison between its dataset, models, performance and contribution is provided in Table 1.

Table 1: Background study comparison on medical health insurance cost prediction using ML and DL methods

Author	Dataset	Methodology	Performance	Limitation/contribution
Garmdareh et al. (2023) [8]	99,440 records from the RASA web portal	Regression-based models, Decision Tree for anomaly detection, Deep Learning for training	Decision Tree detected 17% of claims as abnormal with at least 30% deviation, deep learning had best training MAE	Decision Tree effective for anomaly detection. Only 50% of anomalies approved by expert assessors.
Nabrawi & Alanazi (2023) [9]	Health insurance data from 3 providers in Saudi Arabia	RF, LR, ANN; SMOTE for balancing; Boruta for feature selection	Random Forest: 98.21% accuracy, 100% recall, 90% AUC; ANN: 94.64% accuracy, 96.08% recall, 88.04% AUC	identified key features (policy type, education, age) for fraud detection. Further research on larger datasets is recommended.
Fursov et al. (2022) [11]	Health insurance data from an insurance company	Deep learning architectures combining sequential and tabular components	Achieved ROC AUC of 0.873, outperforming state-of-the-art models (0.815)	Improved fraud detection and robustness to data corruption. Focused primarily on sequential/tabular data integration.
Nur Prasasti et al. (2020) [12]	Real-world automobile insurance data (Indonesia)	Supervised classifiers: MLP, Decision Tree C4.5, RF; SMOTE, undersampling for balancing	Random Forest: 98.5% accuracy	Addressed imbalanced data in fraud detection. Lack of comparison with more recent models or deep learning techniques.
Panda et al., 2022[10]	Insurance cost data	Ridge, Lasso, Simple Linear, Multiple Linear, Polynomial Regression	Polynomial Regression: RMSE = 5100.53; R^2 = 0.80	The contribution is a real-time insurance cost prediction system with Polynomial Regression offering the best results. The limitation is the high RMSE value, indicating some inaccuracies in the predictions.
Jyothsna et al., 2022[13]	Health insurance and emergency department data	Multi-Linear, DT, RF, GB Telegram-integrated chatbot	Gradient Boosting: 87% accuracy	The study combines Gradient Boosting with a chatbot for user interaction, offering practical application in premium estimation. The limitation may be the accuracy ceiling at 87%.

1. Research gaps

The research gap identified from these studies lies in the limited exploration of advanced NLP) techniques and deep learning models for predicting insurance costs and claims. While traditional machine learning methods such as regression models and ensemble techniques (e.g., RF, GB) have demonstrated high accuracy in cost prediction, most studies focus on structured data, neglecting the potential of unstructured data like clinical notes or insurance documents. Additionally, there is a lack of comprehensive models that integrate demographic, medical, and behavioral factors simultaneously to improve predictive accuracy and reliability. Furthermore, although some studies have developed real-time systems for cost estimation, they often exhibit limitations in terms of interpretability and scalability, particularly for large and diverse datasets. This gap

International Journal of Core Engineering & Management
Volume-7, Issue-11, 2024 ISSN No: 2348-9510

underscores the need for more holistic, interpretable, and scalable models that can leverage both structured and unstructured data, while also addressing ethical concerns in healthcare insurance forecasting.

III. METHODS AND MATERIALS

In this research, machine learning-based regression techniques are applied to forecast health insurance costs employing a dataset sourced by Kaggle, consisting of 986 records and 11 features. After data collection, conduct pre-processing for data reliability. The data preprocessing steps include dropping irrelevant columns, handling missing and duplicate values, and applying standard scaling for normalization. Feature selection is carried out to retain the most relevant features, followed by splitting the dataset into training (70%) and testing (30%) sets. Then, regression models utilized in the study include SVR, LR, and XGBoost. These models are evaluated using performance metrics like R^2, MAE, RMSE, and MAPE, to assess their accuracy in forecasting continuous health insurance costs based on customer attributes. Figure 1 shows the whole process and a systematic flow diagram of the system that has been suggested.

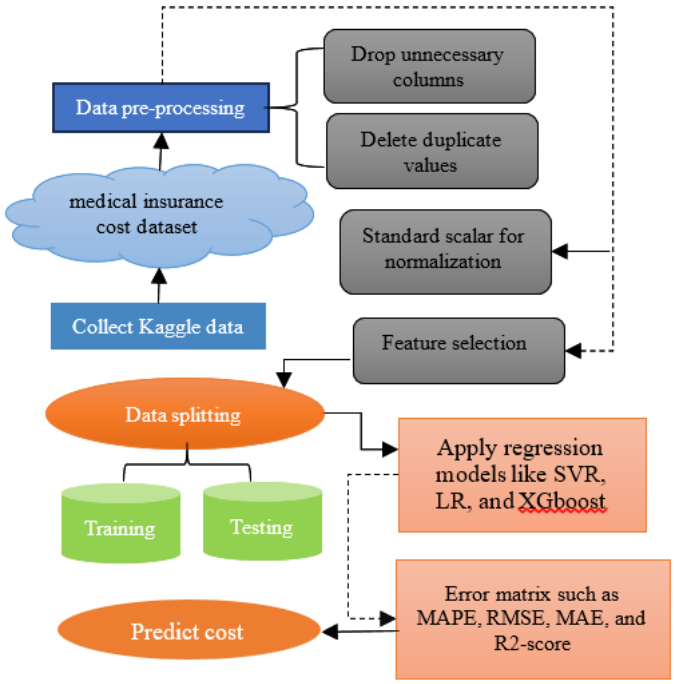

Figure 1: Flowchart for medical health insurance cost prediction

1. Data source

The dataset utilized for the analysis of medical insurance costs was obtained by a KAGGLE repository. A Medical Insurance Company Has Released Data For Almost 1000 Customers. Eleven characteristics or features and 986 records make up the dataset. As shown in figure 2's Pearson correlation heat map, it is critical to verify the associations between a few important characteristics in order to determine their correlation at this point in the study.

International Journal of Core Engineering & Management
Volume-7, Issue-11, 2024 ISSN No: 2348-9510

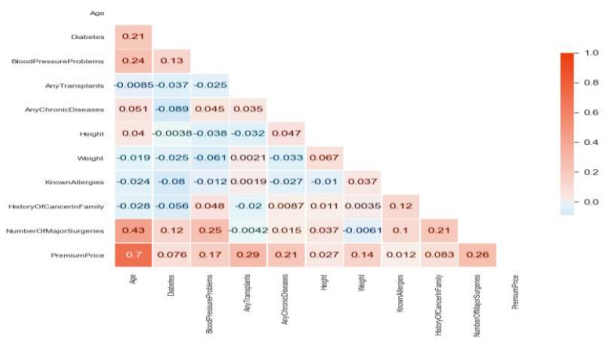

Figure 2: Correlation plot of features

The correlation plot in figure 2 visually represents the relationships between different features. Each matrix cell displays the correlation coefficient among the two variables; the colors indicate the strength and direction of the association. Lighter colors imply lesser or no association, whereas dark blue denotes significant negative correlation and dark red suggests high positive correlation.

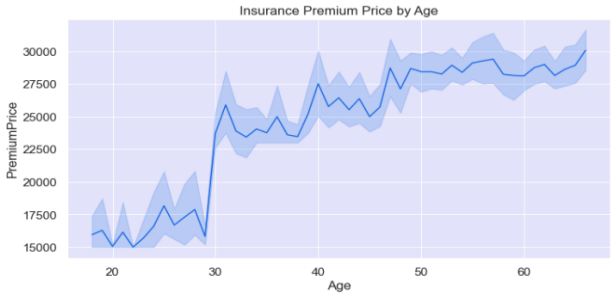

Figure 3: Insurance premium price by age

Figure 3 shows how insurance premiums change as people age. The y-axis represents the premium price, ranging from around 15,000 to 30,000, while the x-axis represents age, from 20 to 60 years. The line fluctuates, indicating that the premium price varies with age, and there's a shaded area around the line that might represent variability or confidence intervals.

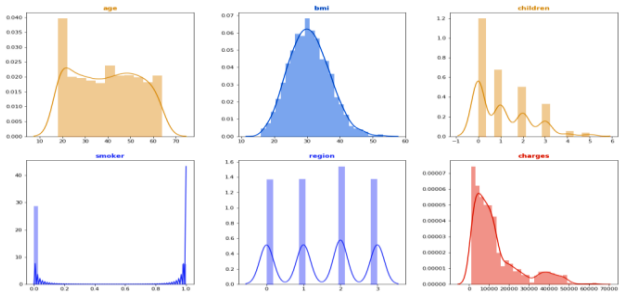

Figure 5: Histogram plot for Distribution of features

The figure 5 presents histograms showing the distribution of key features in a dataset, including age, BMI, number of children, smoker status, region, and charges (likely medical expenses). These visualizations reveal the frequency and spread of each feature, helping to identify central tendencies, variations, and potential patterns, which are valuable for statistical analysis and demographic insights.

International Journal of Core Engineering & Management
Volume-7, Issue-11, 2024 ISSN No: 2348-9510

2. Data preprocessing

Data preprocessing is a important step in data mining since it prepares databases for extraction by cleaning, organizing, and converting them. A number of procedures must be fulfilled before processing may begin. Several techniques include dataset reduction, cleaning, integration, and transformation. This project employed a number of processes to collect the data in the correct format. Below are the essential preprocessing steps:

- Drop unnecessary columns: Use the drop method to remove superfluous rows and columns from your datasets.
- Delete duplicate values: In this case, rows or columns with missing values are removed, ensuring that the remaining data is complete and ready for analysis.

3. Standard scalar for normalization

Another well-liked feature scaling method in ML is the standard scaler, often known as standardization. The method averages out all features with zero volatility. Most data will be around zero since this strategy does not alter the distribution of the data or restrict it to a certain period. This means that data outliers will persist after scaling. Standard scaling is defined in Equation 1.

$$x_{scaled} = \frac{x - \bar{x}}{\sigma} \ldots\ldots (1)$$

where:

- xscaled = scaled sample point
- x = sample point
- $\bar{x}$ = mean of the training samples
- σ = standard deviation of the training samples

4. Feature selection

Feature selection is a method for improving the accuracy of a dataset by deleting irrelevant or redundant characteristics using an evaluation index. Therefore, it is critical to separate the most pertinent and relevant elements from the data and eliminate any irrelevant or less significant information.

5. Data splitting

Next, the dataset was divided into a training set and a test set. There was a 70% allocation to training and a 30% allocation to testing of the total data.

6. Machine learning models

In the current experimentation, the regression models [14] used are as follows: [SVR, LR, and XGBoost]. These models are applied to predict continuous variables based on the provided dataset.

A. Support vector regressor

Classical ML techniques known as SVM analyze classification and regression problems by passing parameters into kernel functions such as linear, Gaussian, sigmoid, polynomial, etc [15].

International Journal of Core Engineering & Management
Volume-7, Issue-11, 2024 ISSN No: 2348-9510

B. XGBoost regressor

XGBoost is a decision-tree based ensemble learning method. Regression scenarios may make use of it by minimizing a loss function that evaluates the gap between the actual and expected goal values. Here is the mathematical model for XGBoost regression: (2):

$$y = f(x) \dots \dots (2)$$

where y is the predicted property price, x is a vector of input features (such as number of bedrooms, square footage, etc.), and f(x) is the XGBoost model that predicts y based on x. In order to calculate f(x), XGBoost constructs a network of decision trees that have been trained to minimize the loss function known as MSE. For the final forecast, the model averages the results from all the decision trees. The general form of the XGBoost regression model can be expressed as (3):

$$y = \sum(k = 1 \; to \; K) \; fk(x) \dots \dots \dots (3)$$

The forecast of the k-th decision tree is denoted by fk(x), while the total number of DT in the ensemble is represented by K. The prediction of each tree is a weighted sum of the leaf values of the tree, which are learned during training [16][17]. The XGBoost model's forecast for an input x is calculated by adding together the forecasts of all the ensemble decision trees.

C. Logistic regression (LR)

LR is a method for categorization based on the premise that the outcome is affected by one or more separate variables. To handle multi-class scenarios, one-vs-rest logistic regression (OVR) or MLR may be applied to LR, despite its primary use as a binary classifier [18].

7. Performance matrix

A key component of developing ML projects is model assessment, which facilitates the understanding of model performance and facilitates the explanation and presentation of model output. The objective then becomes to demonstrate how near the projected values are to the real values since it might be challenging to anticipate a regression model's precise value. Four performance assessment criteria were used in this study to assess the models: R2, MAE, RMSE, and MAPE.

A. R-Squared

The regression model's fit to the data is gauged by its R2 value. A stronger correlation between the model and the data is shown by higher R2 values. Within the interval of 0 to 1, R2 values are numerical. When the R2 number is 1, it implies that the model accurately predicts the response data, whereas an R2 value of 0 shows that the model does not explain any of the variability of the response data around its mean. The following is the formula (4) to get R2:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \dots \dots (4)$$

B. Mean Absolute Error (MAE)

MAE is a frequently utilized statistic to assess a prediction model's accuracy. It measures the typical magnitude of prediction mistakes independent of direction. Improved performance is suggested by a lower MAE number. To compute MAE, use formula (5):

International Journal of Core Engineering & Management
Volume-7, Issue-11, 2024 ISSN No: 2348-9510

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |(y_i - y_i^P)| \ldots . (5)$$

Where,
Y is an actual value,
Y is the forecasted value, and n is the number of observations.

C. RMSE (Root Mean Squared Error)

This statistic stands for the MSE. To measure the extent to which the model's predictions differ from the actual values, RMSE is used. Model performance is improved when the RMSE value is lower. The RMSE formula is (6):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - y_i^{\dot{P}})^2} \ldots \ldots (6)$$

D. Mean Absolute Percentage Error (MAPE)

According to MAPE, an average percentage discrepancy among predictions and their intended objectives in the dataset is the measure by which mistakes are calculated in percentage terms. Another way to look at MAPE is as a percentage of the MAE that was returned (7).

$$MAPE = \frac{1}{n}\sum_{i=1}^{n} / (\frac{(y_i - y_i^{\dot{P}})^2}{y_i}) /* 100 \ldots \ldots (7)$$

Together, these measures provide information about the model's predictive power and accuracy for the target variable.

IV. RESULT ANALYSIS AND DISCUSSION

Predicting the cost of health insurance was the aim of this research. Several regression approaches were employed in this investigation. It is advised that for this experiment, a computer with at least 16GB of RAM and an Intel processor generation of at least 9th technology or above be used. The following Table 2 provides the XGBoost model performance across the performance.

Table 2: XGBoost model performance on medical insurance cost dataset

Matrix	XGBoost
R2	86.47
MAE	1442.904
RMSE	2231.524
MAPE	5.906

International Journal of Core Engineering & Management
Volume-7, Issue-11, 2024 ISSN No: 2348-9510

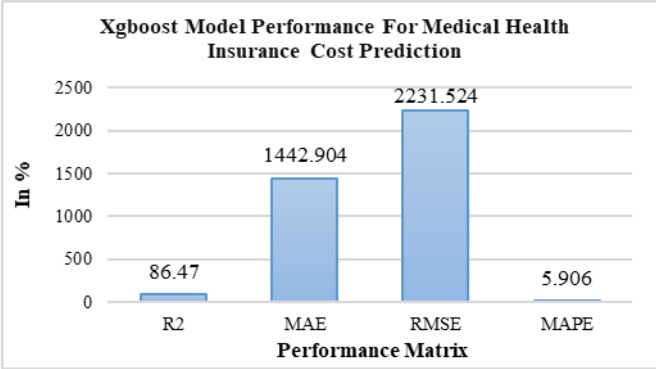

Figure 6: XGBoost model performance

The performance of the XGBoost model shows in above Table 2 and figure 6. The XGBoost model demonstrates strong performance with an R^2 of 86.47%, a MAE of 1,442.90, and a RMSE of 2,231.52, indicating accurate predictions with minimal error. The MAPE of 5.91% further confirms the model's low average prediction error, reflecting robust predictive capability.

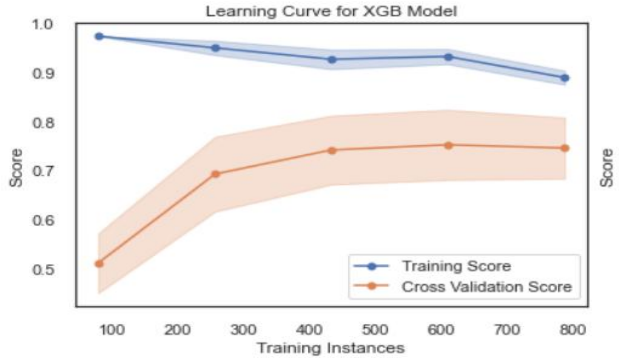

Figure 7: Learning curve for XGBoost

Figure 7 above depicts the XGBoost model's learning curve. The x-axis represents the number of training instances, ranging from 100 to 800, and the y-axis indicates the score, which could be accuracy or another evaluation metric. training score (blue) starts 1.0 high and slightly decreases, while the cross-validation score (orange) improves as more data is added, levelling off around 0.75.

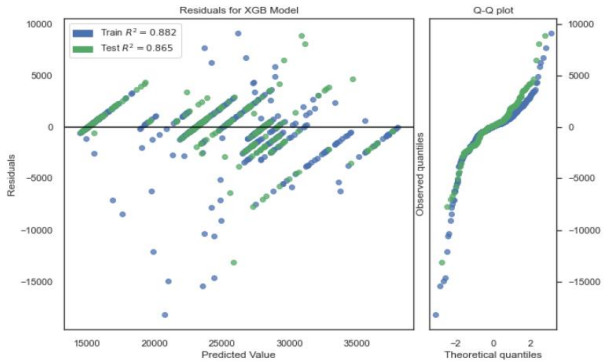

Figure 8: Residual plot for XGBoost model

International Journal of Core Engineering & Management
Volume-7, Issue-11, 2024 ISSN No: 2348-9510

The XGBoost model in Figure 8 shows strong performance, with a residual plot indicating scattered residuals for both training (R^2 = 0.882) and test data (R^2 = 0.865), suggesting unbiased predictions. The Q-Q plot reveals that the residuals are approximately normally distributed, indicating a well-fitted model on both datasets.

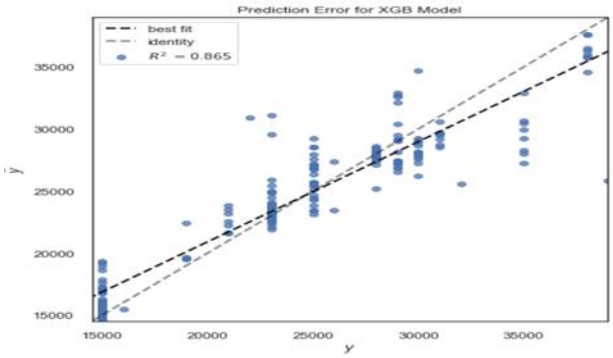

Figure 9: Prediction error plot for XGBoost

The prediction error plot for the XGBoost model in figure 9 effectively visualizes the alignment between predicted and actual values, featuring individual data points along with a dashed line for the best fit and a dotted line for perfect predictions. With an R^2 value of 0.865, the plot indicates that the model demonstrates strong predictive power, suggesting it captures the underlying patterns in the data well.

1. Comparative Analysis

The comparative analysis for predict the medical insurance cost using regressor model is present in this section. The models include SVR[19], LR [20], and XGBoost for comparison are implement on the medical insurance cost dataset. The following Table 3 shows the comparison of models on same dataset.

Table 3: Model Comparison for Predicting Medical Insurance Cost on Data

Models	R2	MAE
Logistic Regression	70.70	35.84
Support Vector Regressor	84.23	23.47
XGBoost regressor	86.47	14.42

Table 3 provides the comparison between model performance for medical health insurance cost prediction. The comparison of model performance reveals that XGBoost outperforms both LR and SVR in terms of R^2 and Mean Absolute Error (MAE). While LR shows an R^2 of 70.70 with an MAE of 35.84, and SVR achieves an R^2 of 84.23 with an MAE of 23.47, XGBoost leads with an R^2 of 86.47 and a significantly lower MAE of 14.42. This indicates that XGBoost not only explains more variance in the data but also produces predictions that are closer to the actual values, demonstrating its superior predictive capability.

V. CONCLUSION AND FUTURE SCOPE

The whole cost of medical bills resulting from an insured person's sickness is covered by health insurance. One significant way to manage health insurance expenditures is to keep costs under

control. In the modern world, managing the expense of health insurance has grown in importance. This study successfully demonstrates the application of ML-based regression techniques for predicting health insurance costs using a dataset sourced from Kaggle. The RMSE, MAE, R2, and MAPE metrics are the five model assessment metrics that we utilize to determine whether or not the model is successful. The analysis reveals that XGBoost significantly outperforms Linear Regression and Support Vector Regression in terms of R^2 and Mean Absolute Error, showcasing its effectiveness in capturing the complexities of the data. Results demonstrate that XGBoost outperforms both LR and SVR, achieving an R^2 of 86.47 and a significantly lower MAE of 14.42, indicating superior predictive capability. The findings underline the importance of employing advanced ML methods to enhance an accuracy of cost predictions, which can be invaluable for both insurers and policyholders. However, this research is not without its limitations. The dataset, while comprehensive, contains only 986 records, which may restrict the model's generalizability to larger and more diverse populations. The constraints should be addressed in future research by using bigger and more varied datasets that include a wider variety of socio-economic characteristics. If we want better predictions, we should look at more complex machine learning methods like DL and hybrid models.

REFERENCES

1. K. Xu et al., "Public spending on health: a closer look at global trends," 2018.
2. B. D. Sommers, "Health insurance coverage: What comes after the aca?," Health Aff., 2020, doi: 10.1377/hlthaff.2019.01416.
3. B. Milovic, "Prediction and decision making in Health Care using Data Mining," Int. J. Public Heal. Sci., 2012, doi: 10.11591/ijphs.v1i2.1380.
4. M. Kumar, R. Ghani, and Z. S. Mei, "Data mining to predict and prevent errors in health insurance claims processing," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2010. doi: 10.1145/1835804.1835816.
5. Prof. M. S. Patil, Kulkarni Sanika, and Khurpe Sanjana, "Medical Insurance Premium Prediction With Machine Learning," Int. J. Innov. Eng. Res. Technol., vol. 11, no. 5, pp. 5–11, 2024, doi: 10.26662/ijiert.v11i5.pp5-11.
6. S. R. Thota and S. Arora, "Neurosymbolic AI for Explainable Recommendations in Frontend UI Design - Bridging the Gap between Data-Driven and Rule-Based Approaches," no. May, pp. 766–775, 2024.
7. 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, 2018, doi: 10.1186/s12938-018-0568-3.
8. M. S. Garmdareh, B. S. Neysiani, M. Z. Nogorani, and M. Bahramizadegan, "A Machine Learning-based Approach for Medical Insurance Anomaly Detection by Predicting Indirect Outpatients' Claim Price," in 2023 9th International Conference on Web Research, ICWR 2023, 2023. doi: 10.1109/ICWR57742.2023.10139290.
9. E. Nabrawi and A. Alanazi, "Fraud Detection in Healthcare Insurance Claims Using Machine Learning," Risks, 2023, doi: 10.3390/risks11090160.
10. 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.
11. I. Fursov et al., "Sequence Embeddings Help Detect Insurance Fraud," IEEE Access, 2022, doi: 10.1109/ACCESS.2022.3149480.
12. I. M. Nur Prasasti, A. Dhini, and E. Laoh, "Automobile Insurance Fraud Detection using Supervised Classifiers," in 2020 International Workshop on Big Data and Information Security, IWBIS 2020, 2020. doi: 10.1109/IWBIS50925.2020.9255426.
13. C. Jyothsna, K. Srinivas, B. Bhargavi, A. E. Sravanth, A. T. Kumar, and J. N. V. R. S. Kumar, "Health Insurance Premium Prediction using XGboost Regressor," in Proceedings - International Conference on Applied Artificial Intelligence and Computing, ICAAIC 2022, 2022. doi: 10.1109/ICAAIC53929.2022.9793258.
14. H. Sinha, "ANALYZING MOVIE REVIEW SENTIMENTS ADVANCED MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING METHODS," Int. Res. J. Mod. Eng. Technol. Sci. (, vol. 06, no. 08, pp. 1326–1337, 2024.
15. S. Zou, C. Chu, N. Shen, and J. Ren, "Healthcare Cost Prediction Based on Hybrid Machine Learning Algorithms," Mathematics, 2023, doi: 10.3390/math11234778.
16. Wankar H, Dimble K, Dasgaonkar P, Chavan V, and Sayyad A, "Property Price Prediction Engine Using XGBoost Regression," Int. J. Creat. Res. Thoughts, vol. 11, no. 4, pp. 190–194, 2023.
17. J. Thomas, "Enhancing Supply Chain Resilience Through Cloud-Based SCM and Advanced Machine Learning: A Case Study of Logistics," J. Emerg. Technol. Innov. Res., vol. 8, no. 9, 2021.
18. J. Wu and Y. Asar, "On almost unbiased ridge logistic estimator for the logistic regression model," Hacettepe J. Math. Stat., 2016, doi: 10.15672/HJMS.20156911030.
19. M. hanafy and O. M. A. Mahmoud, "Predict Health Insurance Cost by using Machine Learning and DNN Regression Models," Int. J. Innov. Technol. Explor. Eng., 2021, doi: 10.35940/ijitee.c8364.0110321.
20. N. D. R. Dr. S. M. Iqbal, Sayali D. Ghatol, Prerana V. Jadhav, "Health Insurance Cost Prediction using Machine Learning Algorithms," Int. Res. J. Eng. Technol. e, vol. 11, no. 04, pp. 1381–1384, 2024, doi: 10.1109/ICECAA55415.2022.9936153.