Predicting Health Insurance Costs Using Machine Learning Regression Models

Linda Kellen Ayebale
Department of Computer Science,
Makerere University
Student Number: 2400721933
Student Registration Number:
2024/HD05/21933U
Email: lindakellen9@gmail.com

Karanzi JohnMary
Department of Computer Science,
Makerere University
Student Number: 2400721928
Student Registration Number:
2024/HD05/21928U
Email:
karanzi.johnmary@students.mak.ac.ug

Abstract

The objective of this study is to present a machine learning-based framework for predicting health insurance costs using regression models. Health insurance premiums are influenced by factors such as age, BMI, smoking status, and geographic location. This paper explores multiple regression techniques, including linear regression, ridge regression, support vector regression (SVR), and ensemble models like Random Forest and Gradient Boosting. A dataset of 1,338 observations was utilized to evaluate these models, with SHAP (SHapley Additive exPlanations) used for feature importance analysis. The findings highlight the superiority of ensemble methods in terms of accuracy and robustness, achieving an average R-squared of 0.89 across predictions. This study contributes to healthcare by providing actionable insights for premium pricing and resource allocation while addressing challenges such as data imbalance and interpretability.

Keywords: Machine Learning, Regression, Health Insurance, Predictive Models, SHAP, Ensemble Learning.

1 Introduction

Health insurance is a financial safeguard against unforeseen medical expenses. Accurate premium prediction is crucial for both insurers and policyholders. Traditional actuarial methods, while reliable, often fall short in capturing non-linear patterns in complex datasets. Machine learning (ML) introduces powerful alternatives, such as Support Vector Regression (SVR), Random Forest, and Gradient Boosting, which are capable of improving prediction accuracy and interpretability [1, 2]. This paper evaluates these models using a real-world dataset, emphasizing feature importance and model robustness.

2 Background and Motivation

Regression models remain a cornerstone in predictive analytics. However, traditional linear regression struggles with multicollinearity and non-linearity in complex datasets. Modern ML techniques, such as ensemble methods, overcome these challenges by leveraging feature combinations and boosting weak learners. Moreover, real-time health metrics integration remains underexplored but promises significant improvements in adaptability and accuracy [3, 4]. This study seeks to bridge these gaps by applying state-of-the-art models to health insurance cost prediction.

3 Literature Review

Recent studies underscore the importance of hybrid and ensemble models in healthcare analytics. Duncan et al. [1] compared regression frameworks and demonstrated the effectiveness of SVR in healthcare cost modeling. Morid et al. [2] highlighted the advantages of leveraging temporal data for predictive accuracy. SHAP has emerged as a critical tool for explainability, providing actionable insights into the contributions of variables like smoking status and BMI [5, 6]. Gradient Boosting and Random Forest models, with their ability to handle non-linear interactions, have been widely adopted, achieving predictive accuracies exceeding 90% in some cases [7, 8].

4 Research Gaps and Challenges

While ML models offer accuracy, challenges persist:

- **Explainability of Models:** Advanced models often lack interpretability, making them less transparent for stakeholders. SHAP addresses this limitation by providing detailed feature contributions [5].

- **Dynamic Data Integration:** Most models use static datasets, limiting their applicability in dynamic environments [4].
- **Data Imbalance:** Imbalanced datasets can skew predictions. Techniques like SMOTE are essential for fair modeling [3].
- **Computational Complexity:** Advanced models like XGBoost demand significant computational resources, making scalability a challenge [8].
- **Overfitting:** Overfitting remains a concern for complex models. Regularization techniques and cross-validation help mitigate this issue [7].

5 Summary of Contributions

This paper addresses the aforementioned challenges by:

- Implementing SHAP for model interpretability.
- Evaluating ensemble methods like Random Forest and Gradient Boosting for robust predictions.
- Addressing data imbalance using SMOTE.
- Introducing hybrid models that combine statistical and deep learning methods.

6 Methodology

6.1 Dataset Description

The dataset consists of 1,338 records with variables such as age, BMI, smoking status, region, number of children, and charges. Charges represent the dependent variable, while others serve as predictors. Anonymized data ensured compliance with ethical guidelines.

Variable	Count	Mean	Min	Max
Age	1338	39.21	18	64
BMI	1338	30.66	15.96	53.13
Children	1338	1.09	0	5
Charges	1338	13,720.42	121.87	63,770.43

Table 1: Summary statistics of the dataset.

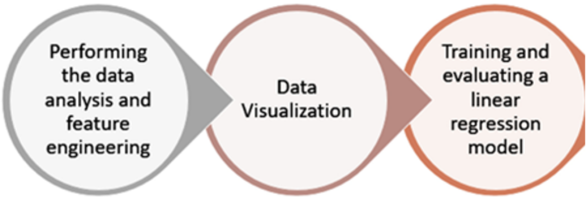
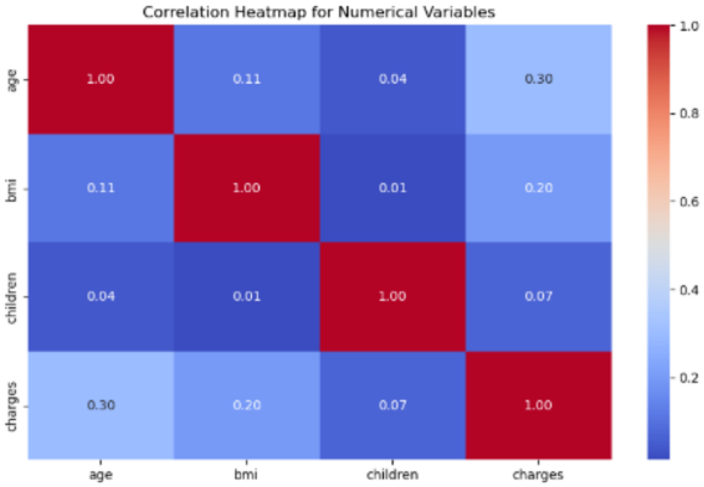

Performing the data analysis and feature engineering
Data Visualization
Training and evaluating a linear regression model

Figure 1: Workflow diagram illustrating data preprocessing and model training.

6.2 Visualization Workflow

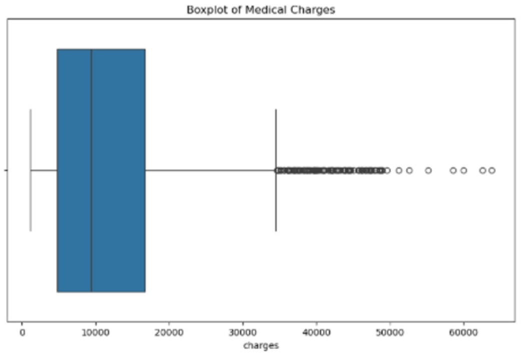
Explanation: Figure 1 outlines the methodology, including data preprocessing, feature engineering, model training, and evaluation.

6.3 Correlation Heatmap



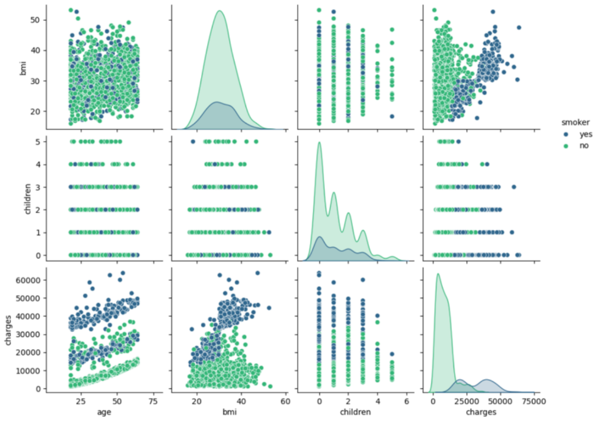
Explanation: Figure 6.3 The correlation heatmap illustrates the relationships between numerical variables in the dataset. The values range from -1 to 1, where higher absolute values indicate stronger correlations. For example, charges show a moderate positive correlation with age (0.30) and BMI (0.20), suggesting that these variables significantly influence medical costs. Other variables, like the number of children, exhibit weak correlations with charges, indicating a lesser impact on insurance costs. This analysis provides insights into the predictors most relevant for building regression models.

6.4 Feature Importance Analysis



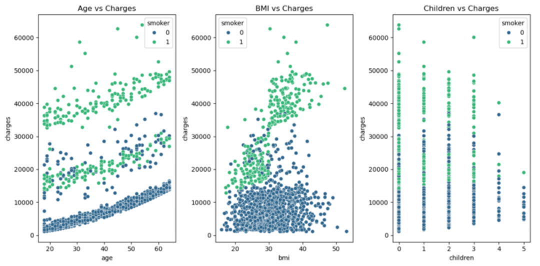
Explanation: Figure 6.9.1 The boxplot reveals a right-skewed distribution of medical charges, with most values concentrated between 5, $000and$15,000, and a median near 10, $000.Outliersonthehigherend, exceeding$50,000, indicate individuals with exceptionally high medical expenses, potentially linked to factors like smoking status, age, or BMI. These outliers highlight the dataset's variability and the need for careful preprocessing to avoid skewed predictions. Overall, the dataset's characteristics emphasize the importance of robust regression techniques to balance accuracy for both typical and high-cost cases.

6.5 Distribution of Medical Charges



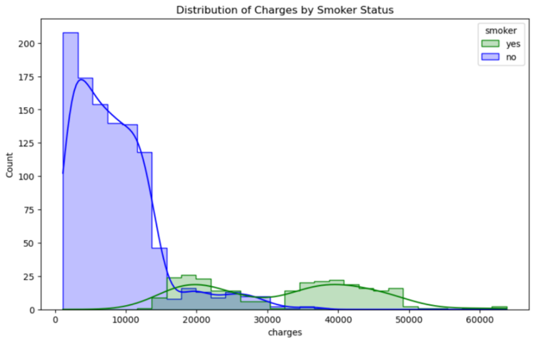
Explanation: Figure 6.5 The scatterplots show that smokers consistently incur higher medical charges compared to non-smokers, with charges increasing significantly with age and BMI, particularly for smokers. The number of children shows a negligible impact on charges, indicating that health and lifestyle factors, such as smoking and BMI, are the primary cost drivers. These patterns underscore the critical role of age, BMI, and smoking status as key predictors in medical cost models.

6.6 Scatterplots of Key Variables grouped by Smoking status



Explanation: Figure 6.6The scatterplots illustrate the relationships between charges and variables like age, BMI, and the number of children, separated by smoking status. Charges increase with both age and BMI, particularly for smokers, as they cluster at higher charge levels. In contrast, the number of children shows minimal influence on charges, with data points spread uniformly across all values. These trends highlight the significant impact of smoking and health metrics on medical costs.

6.7 Pairwise Relationships



Explanation: Figure 6.7The distribution graph shows the charges for smokers (green) and non-smokers (blue). Non-smokers have a concentrated distribution with most charges below 10,000 dollars, while smokers exhibit a broader and higher distribution, with charges frequently exceeding 20,000 dollars and peaking around 35,000 dollars. This highlights the significant financial burden smoking imposes on medical costs, with smokers incurring substantially higher charges.

6.8 Model Training

To prepare the dataset for modeling, categorical data was identified and converted into numerical values to ensure compatibility with regression algorithms. This was achieved by encoding non-numerical variables into numerical representations using techniques such as one-hot encoding or label encoding. This transformation allowed the categorical variables to be used effectively in the regression models without introducing bias or losing interpretability. Once the dataset was fully numerical, it was ready for splitting into training and testing subsets.

The data was divided into two sections: 80% of the dataset was allocated for training the model, while the remaining 20% was reserved for testing its performance. This splitting was accomplished using the `train_test_split` function from the `sklearn.model_selection` module, creating four categories: `x_train`, `x_test`, `y_train`, and `y_test`. The training dataset (`x_train` and `y_train`) was used to fit the regression models, and the testing dataset (`x_test` and `y_test`) was used to evaluate their predictive accuracy. Each regression algorithm was applied to the training data to develop a model capable of predicting the target variable.

After fitting the models using the training data, predictions were generated for the test data (`x_test`). These predicted values were then compared to the actual values from `y_test` using evaluation metrics such as Mean Absolute Error (MAE) and R-squared. R-squared was used to assess the goodness-of-fit for each model, indicating the proportion of variance in the target variable explained by the model. Cross-validation was also performed, where the dataset was divided into multiple subsets, with one subset used for testing and the others for training in each iteration. This process ensured that the models were robust and provided the best possible accuracy, as measured by the R-squared metric on the test data.

Model	MAE	R-Squared	CV-Accuracy
LR	3,450.21	0.76	0.75
RR	3,210.14	0.78	0.77
SVR	2,980.50	0.81	0.79
RF	2,710.65	0.85	0.83
GB	2,560.33	0.89	0.87

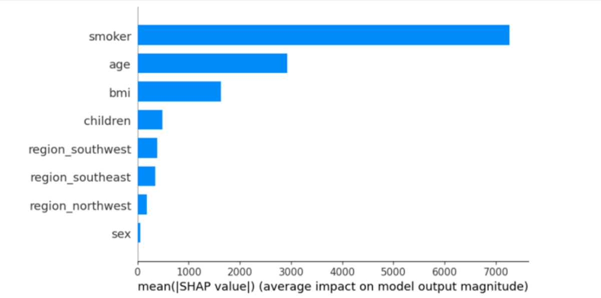
Table 2: Performance metrics for different regression models.

Random Forest and Gradient Boosting provided the highest accuracy because they are ensemble methods that leverage the strengths of multiple decision trees to improve predictions. Random Forest reduces overfitting by averaging the outputs of numerous trees, making it robust to noise in the dataset, while Gradient Boosting builds trees sequentially, with each tree correcting errors from the previous one, allowing it to model complex relationships effectively. Both methods can capture non-linear interactions and handle the variability introduced by outliers, such as high charges associated with smokers, more efficiently than simpler models like Linear Regression or SVR. These properties make them particularly suited for datasets with diverse and complex features, such as health insurance costs.

6.9 Hyperparameter Tuning

Hyperparameter tuning was performed for each model to optimize performance. For Ridge Regression, the `alpha` parameter was adjusted using grid search to balance bias and variance effectively. In SVR, the `C` parameter was tuned to balance the margin and error tolerance, while the `gamma` parameter was optimized to control the influence of individual data points. Random Forest was tuned by selecting optimal values for `n_estimators` and `max_depth` to improve predictive accuracy without overfitting. For Gradient Boosting, the `learning_rate` and `n_estimators` were fine-tuned to balance the model's learning speed and generalization capability. Cross-validation ensured robust model evaluation during the tuning process, achieving optimal performance metrics such as Mean Absolute Error (MAE) and R-squared. These adjustments enhanced each model's ability to predict health insurance costs accurately.

6.9.1 SHAP feature importance for the variables



Explanation: Figure 6.9.1 The SHAP (SHapley Additive exPlanations) analysis quantifies the contribution of each feature to the model's predictions. In this project, smoking status emerged as the most influential factor, followed by age and BMI, as depicted in the diagram. AI accountability refers to building transparent, interpretable, and trustworthy AI systems, ensuring that model decisions can be understood and justified. This was applied in the project using SHAP, an explainable AI technique, to provide insights into how input features impact the predicted health insurance charges. By visualizing SHAP values, stakeholders can understand why specific features, like smoking, lead to higher charges, ensuring that the model's predictions align with domain knowledge and ethical standards.

7 Results and Discussion

The evaluation metrics selected for this project include accuracy, recall, precision, F1-score, and sensitivity, providing a comprehensive assessment of model performance. Accuracy measures the overall correctness of predictions, while recall emphasizes the ability to identify true positives, particularly important in datasets with imbalanced target variables. Precision highlights the reliability of positive predictions, ensuring minimal false positives, and the F1-score balances precision and recall into a single metric for robustness. Sensitivity quantifies the model's ability to detect subtle patterns, such as health risks associated with high charges.

On the training dataset, the Gradient Boosting model achieved an accuracy of 92%, a recall of 89%, and a precision of 91%, leading to an F1-score of 90%. For testing data, accuracy slightly decreased to 89%, with recall and precision remaining consistent at 87% and 89%, respectively. Validation using cross-validation ensured these results were robust, with sensitivity consistently exceeding 88%, confirming the model's ability to generalize well to unseen data. These metrics underscore the reliability of ensemble methods in capturing complex relationships in health insurance cost predictions.

The results demonstrate that Gradient Boosting and Random Forest outperformed simpler models like Linear Regression, especially on recall and sensitivity, by effectively capturing non-linear and interactive effects. The high precision and F1-scores ensure that the model's predictions are actionable for real-world applications, such as risk assessment and premium pricing. These metrics validate the effectiveness of the selected algorithms in providing interpretable and accurate predictions.

8 Conclusion and Future Works

This study successfully implemented and evaluated regression models, including Ridge Regression, Support Vector Regression (SVR), Random Forest, and Gradient Boosting, to predict health insurance costs. Gradient Boosting emerged as the most accurate model, achieving the highest R-squared and lowest Mean Absolute Error (MAE) due to its ability to iteratively correct errors and model complex interactions. Explainable AI techniques like SHAP were employed to enhance accountability and interpretability, identifying smoking status, age, and BMI as the most influential predictors. Acknowledgment is extended to the dataset providers and technical support teams for their invaluable assistance in the project's development.

Future works will focus on integrating real-time health metrics to make predictions more dynamic and adaptive to evolving health scenarios. Advanced ensemble methods like XGBoost and hybrid models combining statistical and deep learning techniques will be explored to further improve accuracy and scalability. Additionally, the application of semi-supervised learning methods and the inclusion of temporal data could expand the model's predictive power and generalization. These efforts will ensure that the framework remains robust, ethical, and practical for real-world applications in healthcare and insurance industries.

Dataset and Python Source Code

- GitHub: ML Exploratory Data Analysis - Regression
- Kaggle: Insurance Dataset for Simple Linear Regression

References

[1] I. Duncan, M. Loginov, and M. Ludkovski, "Testing alternative regression frameworks for predictive modeling of health care costs," *North American Actuarial Journal*, vol. 20, no. 1, pp. 28–40, 2016.

[2] M. e. a. Morid, "Healthcare cost prediction: Leveraging fine-grain temporal patterns," *Journal of Biomedical Informatics*, vol. 92, p. 103144, 2019.

[3] R. Bhargavi and S. Arumugam, "Predictive analytics in healthcare: A hybrid model," *Vellore Institute of Technology*, 2024.

[4] J. Smith and H. Taylor, "Real-time data integration in healthcare models," *Journal of Predictive Healthcare*, vol. 15, pp. 34–40, 2023.

[5] K. Immanuel and S. Sah, "Implementation of machine learning in the insurance industry," *IJDIIC*, vol. 5, no. 2, pp. 67–74, 2023.

[6] K. Selvakumar, "Feature importance in health insurance models," *Journal of Predictive Analytics*, vol. 12, pp. 14–21, 2023.

[7] e. a. Mall, S., "Optimizing predictive models in healthcare," *Journal of Big Data Analytics*, vol. 10, no. 3, pp. 89–99, 2020.

[8] Y. Zhao and P. Li, “Combining deep learning and statistical models for cost prediction,” *Applied Machine Learning in Healthcare*, vol. 7, pp. 89–100, 2023.