MEDICAL INSURANCE PREMIUM PREDICTION

Advanced Analysis

GROUP 05

Kavindi Chamathka	- s16367
Nethmi Sansala	- s16252
Kavindu Weerasekara	- s16076

Abstract

Accurately predicting medical insurance costs is crucial for both insurers and policyholders, as several factors influence premium pricing. This study utilizes machine learning techniques to develop a predictive model for healthcare insurance costs using a dataset from Kaggle. Various regression-based models, including Multiple Linear Regression, XGBoost, Random Forest, LASSO, and Gradient Boosting were evaluated to forecast insurance premiums based on key attributes such as age, weight, diabetes status, and height. The performance of these models was assessed using R², MSE, and RMSE metrics. The findings offer valuable insights for insurance companies seeking to optimize premium calculations and for policymakers aiming to manage healthcare costs, demonstrating the potential of ML in improving transparency and efficiency in the healthcare insurance industry.

Table of Contents

List of Figures

List of Tables

Introduction

Healthcare systems, particularly in developing countries, face significant challenges due to the heavy reliance on out-of-pocket payments, creating barriers to universal health coverage and contributing to inefficiency and inequity. Health insurance, a mechanism designed to mitigate the financial risk associated with medical expenses, plays a critical role in improving access to healthcare services. However, high insurance premiums often prevent individuals from obtaining coverage, resulting in delayed access to medical care and higher mortality rates. Accurate prediction of healthcare costs is essential for both insurance companies and policyholders, as it allows insurers to set fair premiums and helps individuals choose appropriate coverage. This study leverages machine learning algorithms to predict health insurance premiums based on demographic factors and health conditions such as age, weight, and diabetes status. By comparing the performance of various models, including Multiple Linear Regression, XGBoost, Random Forest, LASSO, and Gradient Boosting, the research aims to offer valuable insights into improving premium estimation and cost management in the healthcare sector. Given the rising medical costs and the increasing demand for affordable coverage, this study contributes to developing more efficient and accessible healthcare insurance models.

Description of the question

The primary objective of this study is to identify the key factors affecting medical insurance premiums and understand how these premiums are calculated. Through descriptive analysis, the structure of the dataset is explored, trends are identified, and the relationships between various factors are investigated. The questions to be addressed are as follows:

- What are the main factors affecting medical insurance premiums?
- How are the medical insurance premiums calculated based on these factors?

Subsequently, predictive models are built to quantify the influence of these factors on insurance premiums. These models provide deeper insights into the calculation process, offering a more accurate understanding of the determinants of medical insurance costs.

Description of the data set

Variable Name	Description
Age	Age of customer
Diabetes	Whether The Person Has Abnormal blood sugar Levels
BloodPressureProblems	Whether the person has abnormal blood pressure levels
AnyTransplants	Any major organ transplants
AnyChronicDiseases	Whether customer suffers from chronic ailments like asthma, etc.
Height	Height of customer
Weight	Weight of customer
KnownAllergies	Whether the customer has any known allergies
HistoryOfCancerInFamily	Whether any blood relative of the customer has had any form of
NumberOfMajorSurgeries	The number of major surgeries that the person has had
PremiumPrice	Yearly premium price

Table 1: Dataset Description

The '**Medical Insurance Premium Prediction**' dataset obtained from Kaggle which contains 986 observations of customers with 11 variables was used for the analysis.

Important Results of the Descriptive Analysis

To understand the structure of the dataset and identify key factors influencing medical insurance premiums, both visual and statistical analyses were performed.

For categorical variables, boxplots and hypothesis testing were conducted to examine their impact on premium price. The results indicate that diabetes status, blood pressure problems, any transplants, any chronic diseases, history of cancer in the family, affect the premium price. Number of major surgeries showed a little impact while known allergies status does not show a noticeable impact.

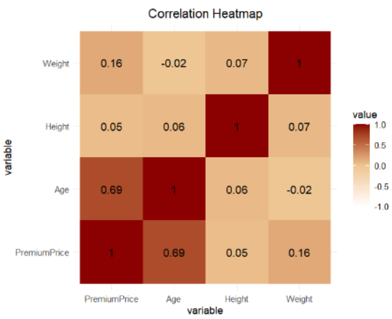

Figure 1: Correlation Heatmap

Among numerical variables, age, height, and weight, were analyzed. Age exhibits a strong influence on premium price, while weight has a slight effect. In contrast, height has only a minimal impact on premium price. Additionally, no multicollinearity was detected among the predictor variables, ensuring that independent variables do not exhibit high correlation.

Furthermore, Factor Analysis for Mixed Data (FAMD) was performed to explore potential clusters within the dataset and assess how different variables collectively influence premium price. The results provide insights into relationships among categorical and numerical predictors, enhancing the understanding of premium variations. No distinct clusters were observed.

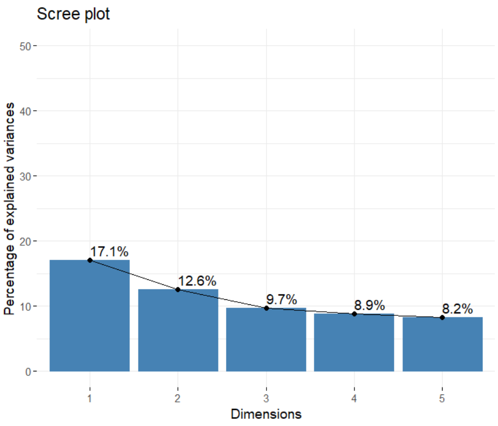

Figure 3: Scree Plot

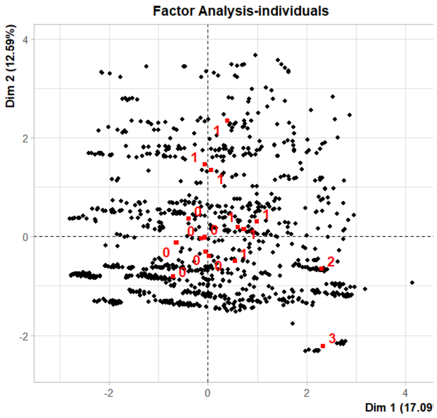

Figure 2: Factor Analysis-Individuals

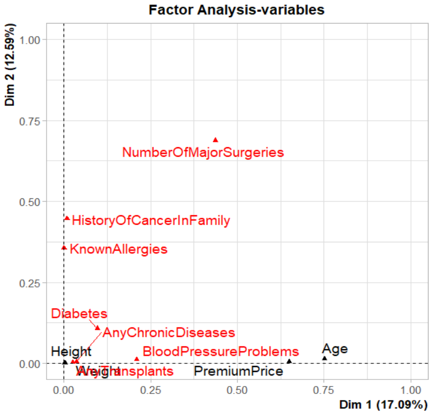

Figure 4: Factor Analysis-Variables

Important Results of Advanced Analysis

The dataset was split into two subsets: 80% for training and 20% for testing. The models were trained using the training set and evaluated on the test set to assess their performance. Evaluation metrics such as R-squared, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) were used to compare the models' effectiveness in predicting the premium prices.

1. Multiple Linear Regression (MLR)

Multiple Linear Regression models the relationship between a dependent variable and multiple independent variables. In this study, the dependent variable is medical insurance charges, and the independent variables include age, diabetes status, weight and etc. Here, best subset selection method was used to determine the optimal number of predictors, which resulted in a model with eight variables providing the best fit. The performance of the model was evaluated as follows using training and test sets to assess its fit and generalization to new data.

Data set	R2	MSE	RMSE
Training	**0.6440822**	**14105915**	**3755.784**
Testing	**0.6576598**	**12482729**	**3533.091**

Table 2: R2, MSE & RMSE values for MLR

2. Decision Trees

A decision tree is a supervised machine learning algorithm that can be used for both classification and regression tasks. In this study we used Recursive Feature Elimination method to select the best set of predictor variables. Then by using selected set of variables were used for pruning the tree for optimal cp and the performance for the optimal model was evaluated for both training and test sets using the following metrics.

Data set	R2	MSE	RMSE
Training	**0.7928**	**8236640.55**	**2869.9548**
Testing	**0.7837**	**9360597.43**	**2816.7669**

Table 3: R2, MSE & RMSE values for Decision Trees

3. Random Forest

Random Forest approach is a supervised learning algorithm that builds multiple decision trees which are known as forest and connect them together for more accurate and stable predictions. Like decision trees we performed Recursive Feature Elimination to get the best set of predictor variables, and it selected the variables other than the "diabetes" variable. Then we used hyperparameter tuning with cross validation to further enhance our model performance and evaluated the model performance using the following metrics.

Data set	R2	MSE	RMSE
Training	**0.9215322**	**3101036.12**	**1760.9759**
Testing	**0.7589486**	**9792083.828**	**3129.2305488**

Table 4: R2, MSE & RMSE values for Random Forest

4. XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful machine algorithm based on gradient boosting. We first attempted to use hyperparameter tuning with cross validation but values that were optimized for the training set were not suitable for test set. We observed that our model was overfitted. Hence, we decided to use manual tuning of hyperparameter by observing both training and test MSE values. Then by observing the feature importance ranking we removed least important variable and again fitted a new XGBoost model. The model performed slightly better than the model with all the predictor variables. However, considering that we are tuning this model manually, we settled on the model with the all-predictor variables as the best possible XGBoost model and evaluated its model performance using following metrics.

Data set	R2	MSE	RMSE
Training	**0.8958**	**4128353**	**2031.834885**
Testing	**0.8089**	**6986249**	**2643.1514205**

Table 5: R2, MSE & RMSE values for XGBoost

5. Lasso Regression

The Lasso Regression is a regression method based on Least Absolute Shrinkage and Selection Operator. We aimed to omit some variables which were identified as less important in the descriptive phase and in that case, Lasso was helpful because it shrink the coefficients of the variables Height and Known Allergies towards zero. The performance of the model was evaluated using training and test sets to assess its fit and generalization to new data as follows.

Data set	R2	MSE	RMSE
Training	**0.6433636**	**14134393**	**3759.574**
Testing	**0.6533592**	**12639540**	**3555.213**

Table 6: R2, MSE & RMSE values for Lasso Regression

6. Ridge Regression

Ridge regression was used to prevent overfitting and improve performance by making ridge suitable when all predicators are expected to have some contribution for the response rather omitting them entirely. This made the coefficients minimal and the coefficient of Height much smaller. Its performance was evaluated and obtained the following results for both training and testing data.

Data set	R2	MSE	RMSE
Training	**0.6409161**	**14231395**	**3772.452**
Testing	**0.6502397**	**12753285**	**3571.174**

Table 7: R2, MSE & RMSE values for Ridge Regression

Issues encountered and proposed solutions

1. Hyperparameter Tuning leading to decrease in model performance:
   - Tuning hyperparameters sometimes resulted in a decrease in evaluation metrics, occasionally causing overfitting.
   - **Solution:** Decided to manually set the hyperparameters to avoid overfitting and improve the stability of the model.
2. Skewed Response Variable:
   - The response variable exhibited right skewness, which can impact the performance of regression models. Although a log transformation was applied to the response variable, it didn't lead to a noticeable improvement in the model's performance.
   - **Solution:** Considered tree-based models, such as decision trees, random forests, and gradient boosting models, which do not rely on the normality assumption and may better capture nonlinear relationships in the data.

Discussion and Conclusions

The results obtained by Advanced Analysis can be summarized as follows:

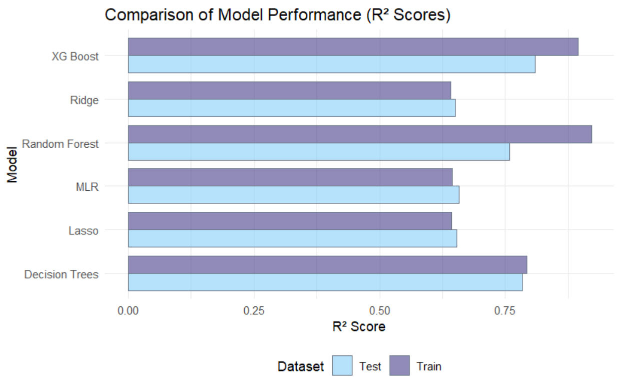

Figure 6: Comparison of model performance (R- Squared Value)

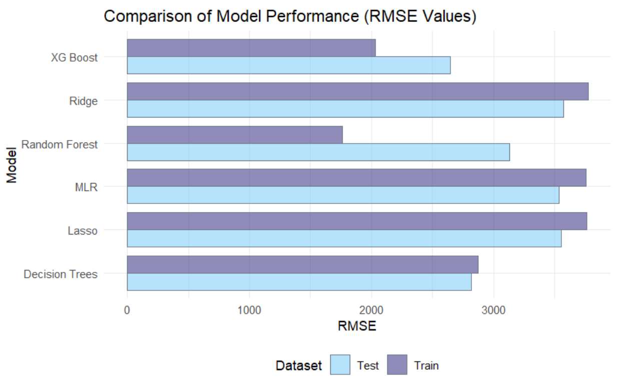

Figure 5: Comparison of model performance (RMSE Values)

The evaluation of model performance based on R² scores and RMSE values provides key insights into the predictive accuracy and generalization ability of the six machine learning models. The R² score analysis indicates that XG Boost achieved high values for both training and testing datasets, demonstrating strong predictive power. Additionally, the RMSE plot reveals that XG Boost has the lowest test error, confirming its ability to minimize prediction mistakes effectively. While there is a slight performance gap between training and testing, it remains the most reliable model among all the evaluated approaches.

Random Forest also showed a high R² score for training, but its significant drop in testing performance indicates overfitting, which is further confirmed by its relatively high RMSE on the test set. Decision Trees displayed moderate performance with a smaller gap between training and testing scores, suggesting better generalization compared to Random Forest. However, its accuracy remains lower than XG Boost.

Linear models such as Multiple Linear Regression (MLR), Lasso, and Ridge Regression had lower R² scores, indicating they struggle to capture complex patterns in the data. However, their smaller differences between training and testing results suggest they are less prone to overfitting. Ridge Regression, in particular, showed a balanced performance with moderate RMSE values, demonstrating the effectiveness of regularization in controlling variance. While these models may not provide the highest accuracy, their simplicity and interpretability make them valuable in scenarios where model transparency is essential.

Overall, XG Boost emerges as the best-performing model for predicting medical insurance premiums due to its high predictive accuracy and lower error rates. Despite other models showing certain advantages in generalization or interpretability, XG Boost provides the best balance between accuracy and reliability. Therefore, XG Boost is selected as the final model for this task, ensuring the most effective premium predictions based on the available data.

Appendix

1. Link for the dataset: https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction
2. Click here for R codes

References

1. https://www.geeksforgeeks.org/random-forest-approach-for-regression-in-r-programming/
2. https://www.geeksforgeeks.org/how-to-use-xgboost-algorithm-for-regression-in-r/
3. https://forecastegy.com/posts/does-random-forest-need-feature-scaling-or-normalization/
4. https://www.geeksforgeeks.org/decision-tree-in-r-programming/
5. https://www.geeksforgeeks.org/lasso-regression-in-r-programming/
6. https://www.irjet.net/archives/V11/i4/IRJET-V11I4171.pdf
7. https://wjarr.com/sites/default/files/WJARR-2023-1355.pdf
8. https://www.ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/1944/725
9. https://www.researchgate.net/publication/348559741_Predict_Health_Insurance_Cost_by_using_Machine_Learning_and_DNN_Regression_Models
10. https://www.irjmets.com/uploadedfiles/paper/issue_2_february_2024/49200/final/fin_irjmets1707496442.pdf