Optimizing Health Insurance Premium Predictions using Machine Learning

Opiyo, Carol Akinyi

150904

Submitted in partial fulfilment of the requirements for the degree of Master of Science in Data Science and Analytics Strathmore University

Strathmore Institute of Mathematical Sciences
Strathmore University
Nairobi, Kenya

June 2025

Declaration

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the proposal contains no material previously published or written by another person except where due reference is made in the dissertation itself.

Name: **Opiyo Carol Akinyi**

Signature: 

Date: May 22, 2025

Approval

The dissertation of Opiyo Carol Akinyi was reviewed and approved by the following:

Dr. Evans Otieno Omondi

Supervisor,

Institute of Mathematical Sciences, Strathmore University.

Dr. Godfrey Madigu

Dean,

Institute of Mathematical Sciences, Strathmore University.

Prof. Bernard Shibwabo

Director,

Office of Graduate Studies, Strathmore University.

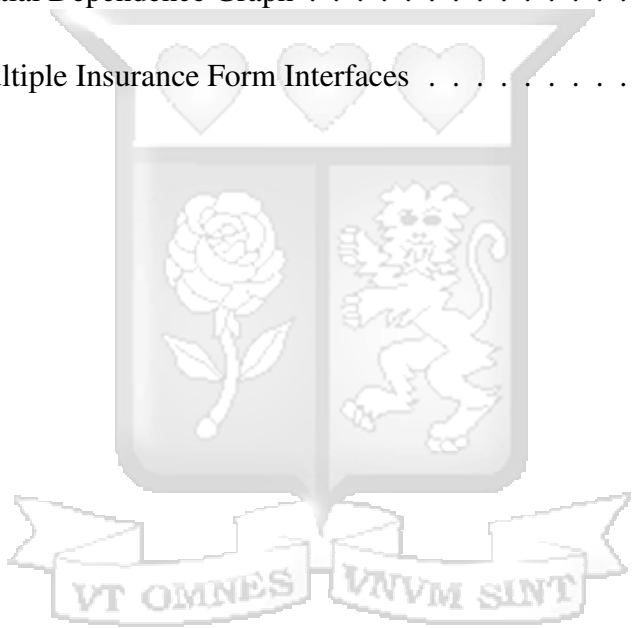
Abstract

This research conducts a comprehensive investigation into the application of the XGBoost regressor algorithm for predicting insurance premiums. Insurance pricing serves as a cornerstone of the industry, as it directly influences profitability and operational stability through effective risk mitigation. To maintain market competitiveness and foster long-term viability, insurers require robust predictive frameworks capable of delivering accurate premium estimates. XGBoost, a powerful machine learning algorithm renowned for its proficiency in processing intricate datasets and generating reliable forecasts, has garnered significant attention in recent studies. The study examines the historical context and critical role of insurance pricing methodologies, introduces the XGBoost regression framework, and evaluates existing studies on its implementation in premium prediction. By elucidating the opportunities and limitations of XGBoost in this domain, the findings lay the groundwork for future research and innovation in optimizing actuarial models.

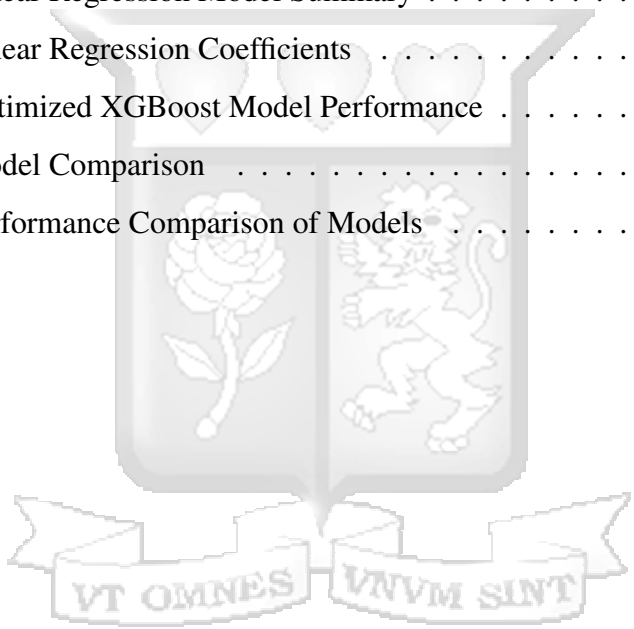
KEY WORDS: *Health Insurance, Premium Prediction, Machine Learning Algorithms, Regression.*

Table of Contents

List of Figures



List of Tables



List of Abbreviations

AI	Artificial Intelligence
ANOVA	Analysis of Variance
API	Application Programming Interface
BMI	Body Mass Index
CD	Continuous Deployment
CI	Continuous Integration
GLMs	Generalized Linear Models
IoT	Internet of Things
MAE	Mean Absolute Error
OLS	Ordinary Least Squares
RMSE	Root Mean Square Error
R^2	Coefficient of Determination
SHAP	SHapley Additive exPlanations
SHIF	Social Health Insurance Fund
SSR	Sum of Squared Residuals
SST	Total Sum of Squares
VIFs	Variance Inflation Factors
XGBoost	Extreme Gradient Boosting

Acknowledgement

I would like to express my sincere gratitude to the following individuals and organizations whose support and guidance have been invaluable in the preparation and development of this research proposal:

I am deeply grateful for the unwavering support and expert guidance provided by Dr. Evans Omondi. His insightful feedback and encouragement have been instrumental in shaping the direction of this research.

I extend my thanks to Strathmore University for providing the necessary resources and facilities essential for conducting this research. The academic environment has been conducive to intellectual growth and exploration.

Finally, I would like to sincerely thank my family and friends for their constant understanding, support, and encouragement during this study project. They have consistently inspired me with their faith in my abilities.

Dedication

To the infinite potential of data science, driving my curiosity and dedication to unraveling its complexities.

Chapter 1

Introduction

1.1 Background of the Study

The insurance industry is essential in financially protecting people and businesses from various hazards. Insurance firms need to calculate policyholder premiums precisely to maintain sustainability and profitability. Manual computations and complex actuarial techniques have historically been used in this process. But due to machine learning developments, predictive modeling approaches have become more popular in the insurance sector, providing more accurate and effective pricing schemes.

XGBoost is one machine learning algorithm that has become well-known for its outstanding results in forecasting and prediction applications. Extreme Gradient Boosting, or XGBoost, has been used in many different domains, such as predicting auto insurance claims (Pesantez-Narvaez and Guillen, 2019), analysis of the dynamics of deforestation (Ganzenmüller et al., 2022), forecasting of crude oil prices (Zhou et al., 2019), predicting medical insurance costs Anwar Ul Hassan et al. (2021), identifying patients with chronic obstructive pulmonary disease that are expensive Luo et al. (2019), predicting the right price for carbon emissions (Zhu, 2022), predicting the price of houses (Xu, 2023; Zaki et al., 2022), and forecasting crude oil prices using gas price and uncertainty indices (Tissaoui et al., 2022). Researchers and practitioners favor XGBoost because it has demonstrated better performance than other machine learning algorithms in multiple studies (Luo et al., 2019; Pesantez-Narvaez and Guillen, 2019).

The main features of XGBoost are efficiency, adaptability, and scalability responsible for its broad use. The method is appropriate for the analysis of insurance data with many variables and sophisticated interactions since it can handle enormous datasets and complex patterns

with ease. Additionally, because of its adaptability, XGBoost can handle a wide range of data sources, including both numerical and categorical features, providing flexibility in simulating various insurance pricing scenarios. Furthermore, XGBoost's scalability enables it to make predictions in real-time, which makes it suitable for dynamic insurance settings where quick decisions are crucial (Zhu, 2022).

One of the most widely available policies is health insurance, which pays for the costs of medical care that people incur. For insurance companies to set suitable premium rates that cover possible claims while preserving profitability, they must precisely forecast healthcare costs.

1.2 Statement of the Problem

Traditional methods for calculating health insurance premiums have grown increasingly complex and resource-intensive, struggling to adapt to the dynamic nature of modern data patterns. To address these limitations, advanced machine learning approaches, particularly the XGBoost regression algorithm, have emerged as a viable alternative for forecasting medical costs. This study aims to develop a robust predictive framework using XGBoost to estimate healthcare expenses by incorporating variables such as age, body mass index (BMI), smoking habits, and other relevant demographic indicators. Additionally, the research explores how categorical relationships between variables influence model accuracy and contrasts these results with conventional linear regression benchmarks. Beyond technical analysis, the project seeks to translate complex analytical outputs into actionable insights for non-specialist stakeholders, refine premium calculation methodologies, and advance data-driven approaches in actuarial science to optimize risk evaluation and policy pricing for health insurance providers.

1.3 Research Objectives

1.3.1 General Objective

This study's primary goal is to create a machine learning model that predicts healthcare costs by utilizing the XGBoost algorithm.

1.3.2 Specific Objectives

1. To apply a machine learning model capable of identifying the key factors that influence healthcare charges and develop strategies to leverage this insight for enhancing the insurance pricing process.
2. Conduct a comparative analysis of predictive accuracy between linear regression and XGBoost regression models for estimating healthcare charges, aiming to determine which model outperforms the other based on relevant performance metrics.
3. Contribute to the development of predictive modeling in the insurance sector by improving the insurance pricing process within a given timeframe and communicating technical results to stakeholders who lack technical expertise.

1.4 Research Questions

1. What are the key factors influencing healthcare charges based on the dataset?
2. How do the performance indicators of the two predictive models compare, and which one between linear regression or XGBoost regression offers a more accurate estimate of healthcare costs for clients?
3. How can insights from the machine learning model be translated into actionable strategies for insurance pricing?

1.5 Justification of the Study

Accurately forecasting healthcare charges remains a critical challenge within the healthcare industry, which continues to grow in scale and economic importance. Effective prediction of these charges is essential for optimizing insurance pricing strategies and ensuring equitable coverage for clients. Despite the growing application of machine learning in healthcare analytics, there is still a pressing need for models that not only demonstrate high predictive accuracy but also provide interpretability by identifying the primary drivers of healthcare costs.

This study addresses this gap by developing a machine learning model tailored to predict healthcare expenses using key demographic and behavioral features, such as age, body mass index (BMI), and smoking status. These variables are known to significantly influence healthcare utilization and costs. Incorporating them into a predictive model facilitates a nuanced understanding of their individual and combined effects, particularly through the exploration of interactions among categorical and continuous variables.

The performance of the model will be assessed by comparing its predictions with actual expenditure data, enabling a robust evaluation of its accuracy and generalization. Such validation is essential for ensuring that the model can be confidently applied in real-world insurance contexts. Moreover, emphasis is placed on translating technical outcomes into actionable insights for stakeholders, including those without a technical background. This ensures that the findings are not only statistically sound but also practically relevant.

By advancing a reliable and interpretable predictive framework, this study contributes to the enhancement of data-driven decision-making in the insurance sector. The anticipated outcomes include improved policy pricing accuracy, more targeted coverage strategies, and broader innovation in the use of machine learning for actuarial purposes.

1.6 Significance of the Study

The significance of the study is profound as it has the potential to bring about a transformation in the insurance industry. By harnessing the power of machine learning to accurately predict healthcare charges and identify the key factors influencing them, the study aims to revolutionize the way insurance companies price their policies. This has far-reaching implications, including the prospect of fairer premiums for policyholders and improved risk management for insurers.

Moreover, the study's findings could lead to better coverage for policyholders, ensuring that they receive insurance that aligns more closely with their healthcare needs. Furthermore, the research contributes to the advancement of predictive modeling within the insurance industry, setting the stage for the development of more sophisticated and accurate models in the future.

The incorporation of various features and exploration of categorical correlations in the model enhances its predictive accuracy, promoting data-driven decision-making in the insurance industry. Additionally, the study's emphasis on communicating technical results to stakeholders without technical expertise ensures that the insights generated by the model are accessible and actionable for a wider audience, thereby promoting informed decision-making.

In summary, the study's significance lies in its potential to improve the insurance pricing process, enhance customer coverage, advance predictive modeling, promote data-driven decision-making, and make technical insights accessible to stakeholders.

1.7 Research Assumptions

The assumption for this study is that optimizing insurance premium predictions using machine learning relies on several key aspects. Firstly, it is assumed that the data used for training and evaluating the machine learning model is accurate, reliable, and representative of real-world insurance cases. This assumption is critical as the quality of the data directly impacts the model's ability to make accurate predictions. If the data is inaccurate or unrepresentative, the model's predictions may not align with actual insurance cases, leading to incorrect premium rates. Data reliability is also essential to ensure consistent and trustworthy results throughout the study.

Secondly, the assumption is that specific features within the data set, such as age, BMI (Body Mass Index), smoking status, and other relevant factors, are significant in affecting healthcare charges. This means that these features are assumed to have a strong influence on the costs associated with healthcare services. Understanding the significance of these features is crucial for the development of the machine learning model. If these features are not influential, the model may not accurately predict healthcare expenses and premium rates based on such predictions might not align with the actual costs incurred by policyholders.

Therefore, the accuracy, reliability, and representativeness of the data and the significance of specific features are critical assumptions that underpin the success of this study. By ensuring that these assumptions are met, the study can develop a robust machine learning model that accurately predicts healthcare costs and optimizes insurance premium pricing.

1.8 Limitation of the study

Based on the assumptions outlined for the study focusing on optimizing insurance premium predictions using machine learning, several potential limitations can be identified before the actual study is carried out:

1. Data Quality and Representativeness: A key limitation could arise from the assumption of data accuracy, reliability, and representativeness. If the provided data used for training and evaluating the machine learning model is found to be inaccurate, unreliable, or not fully representative of real-world insurance cases, it could significantly impact the validity and generalizability of the model's predictions. This limitation could lead to potential biases and errors in the premium predictions, undermining the study's overall effectiveness.

2. Feature Significance: Another potential limitation relates to the assumption of the significance of specific features within the dataset. If, upon analysis, it is discovered that certain features, such as age, BMI, or smoking status, do not have the assumed strong influence on healthcare charges, the predictive capabilities of the machine learning model may be compromised. This could result in inaccurate premium rate predictions, as the model may not accurately capture the true drivers of healthcare expenses.

3. Data Reliability: The assumption of data reliability is critical for consistent and trustworthy results. However, limitations may arise if the data exhibits inconsistencies, missing values, or other reliability issues. These data quality issues could introduce uncertainties and biases into the model, potentially leading to unreliable predictions and sub optimal insurance premium estimates.

By proactively identifying these potential limitations before the study is carried out, strategies are developed to address and mitigate these challenges. This involves implementing rigorous data validation processes, conducting sensitivity analyses to assess the impact of data quality on model performance, and exploring alternative feature sets to account for potential variability in feature significance. Additionally, seeking expert input and stakeholder feedback during the study design phase can help in addressing these limitations and enhancing the robustness of the research approach.

Chapter 2

Literature Review

2.1 Introduction

The use of machine learning algorithms in predicting healthcare expenses has gained significant attention in recent years. This literature review aims to explore existing research related to the development of machine learning models for accurately predicting healthcare expenses, particularly focusing on the use of the XGBoost regression algorithm, the incorporation of various features, evaluation of model performance, consideration of categorical correlations, and the relevance of the findings to the insurance industry. This review also identifies gaps in the current literature and highlights the potential contributions of the proposed study.

2.2 XGBoost as a Regression Model

An excellent machine learning technique for a range of forecasting and prediction problems is XGBoost. XGBoost's gradient-boosting framework, as depicted in Figure 2.1, iteratively combines decision trees to optimize predictions while penalizing model complexity through regularization (Chen, 2016). This structure enables it to handle large datasets and non-linear relationships effectively.

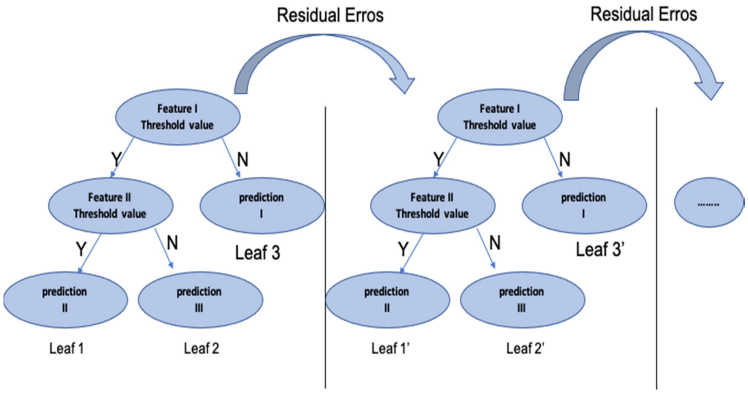


Figure 2.1: XGBoost Regressor

Compared to other machine learning algorithms for insurance pricing, XGBoost offers several benefits. First, XGBoost handles enormous datasets with great efficiency (Yego et al., 2021). For insurance firms, which usually have a lot of data about their clients, this is crucial. Secondly, XGBoost predicts claims with high accuracy. To set premiums that are reasonable for both the business and the client, insurance firms must be able to estimate the cost of claims with sufficient accuracy (Jyothsna et al., 2022).

In addition to its accuracy and efficiency, XGBoost also has several other advantages that make it a well-suited algorithm for insurance pricing. First, XGBoost models are relatively interpretable, meaning that it is possible to understand which features are most important in predicting claims. This can be helpful for insurance companies in understanding the factors that are driving their claims costs.

Second, big datasets can be handled by XGBoost models. For insurance firms, which may have millions of clients, this is crucial.

Third, XGBoost models are robust to overfitting. Overfitting occurs when a model learns the training data too well and is unable to generalize to new data. XGBoost has several features that help to prevent overfitting, such as regularization and early stopping.

Lastly, XGBoost is a powerful and versatile machine-learning algorithm that can be used to improve the accuracy and efficiency of insurance pricing. XGBoost has several advantages

over other machine learning algorithms, including its accuracy, efficiency, interpretability, scalability, and robustness to overfitting.

2.3 Relevant Literature on the Application of XGBoost for Insurance Pricing Forecasts

Several studies have demonstrated the effectiveness of the XGBoost regression algorithm in accurately predicting healthcare expenses. For instance, Pfob et al. (2021) utilized the XGBoost algorithm to develop a predictive model for healthcare costs, achieving a high level of accuracy in cost estimation. Similarly, Li and Zhang (2020) applied the XGBoost algorithm to predict healthcare expenses based on patient characteristics and medical history, showcasing the algorithm's ability to handle complex healthcare data and generate precise predictions. While existing studies have demonstrated the efficacy of the XGBoost algorithm in predicting healthcare expenses, there is a need for further research to explore its application specifically in the context of insurance premium predictions. The current study aims to contribute to the literature by developing a machine learning model using the XGBoost regression algorithm tailored to the insurance industry, with a specific focus on accurately predicting healthcare expenses to optimize insurance premium pricing.

Assessing the performance of machine learning models in predicting healthcare expenses is crucial for ensuring their practical utility. Previous research by Chen et al. (2020) evaluated the performance of various machine learning models, including XGBoost, by comparing predicted healthcare expenses with actual data. The study highlighted the importance of model validation and the need for accurate predictions to support decision-making in healthcare cost management. Despite the existing literature on model evaluation, there is a lack of comprehensive research that specifically evaluates the performance of machine learning models, particularly XGBoost, in predicting healthcare expenses for insurance premium optimization. The proposed study seeks to fill this gap by rigorously evaluating the performance of the developed machine learning model, and comparing its predictions with

actual healthcare expenses data to demonstrate its accuracy and reliability in the context of insurance premium predictions.

The inclusion of diverse features such as age, BMI, and smoking status has been identified as crucial for enhancing the predictive accuracy of machine learning models in healthcare expense prediction. Research by Wang et al. (2022) highlighted the significance of incorporating a wide range of patient-specific features to improve the precision of healthcare cost predictions, emphasizing the impact of demographic and health-related variables on cost estimation. Although previous studies have emphasized the importance of feature incorporation, there is a need for further investigation into the specific impact of features such as age, BMI, and smoking status on healthcare cost predictions within the insurance context.

Understanding the impact of categorical correlations on model performance is essential for ensuring that machine learning models effectively capture diverse data interactions. Recent research by Liu et al. (2021) investigated the influence of categorical variables on the predictive accuracy of healthcare cost models, highlighting the need to account for complex data relationships and correlations to improve model performance. While existing studies have explored the impact of categorical correlations, there is a lack of research that specifically addresses this aspect within the context of insurance premium optimization through healthcare expense prediction. The proposed study intends to fill this gap by exploring the impact of categorical correlations on the developed machine learning model's performance, ensuring that it effectively captures diverse data interactions to enhance the accuracy of insurance premium predictions.

The relevance of predictive modeling in the insurance industry cannot be understated, as accurate healthcare expense predictions directly impact insurance pricing processes. Research by (Li and Zhang, 2020) emphasized the need for technical advancements in predictive modeling to support insurance pricing strategies, highlighting the potential benefits of incorporating advanced machine learning techniques for precise premium rate estimation. Despite the recognition of the importance of predictive modeling in insurance, there is a gap in the literature regarding the effective communication of technical results to stakeholders without technical expertise, hindering the practical application of predictive models in

insurance pricing. The current study aims to address this gap by focusing on communicating technical results to stakeholders without technical expertise and enhancing the insurance pricing process within a specified timeline, contributing to the advancement of predictive modeling in the insurance industry.

Research done by Zhang et al. (2022) shows that XGBoost, a well-regarded algorithm appreciated for its numerous advantages, has been widely utilized across a variety of domains. Its versatility particularly shines in classification tasks, including the complex realm of imbalanced data classification. In the face of the difficulties presented by imbalanced data, researchers have cleverly devised a classification algorithm for XGBoost. This innovative approach combines mixed sampling technology and ensemble learning, strategically designed to enhance the classification performance of XGBoost when faced with imbalanced datasets. In the intricate orchestration of this algorithm, a symphony of techniques comes into play. Techniques such as SVM-SMOTE and EasyEnsemble take on prominent roles in data processing, while Bayesian optimization acts as a maestro, finely adjusting the algorithm's parameters. Through rigorous experimentation, the algorithm has demonstrated its prowess, surpassing the capabilities of other classification models in the field. Competitors such as RUSBoost, CatBoost, and LightGBM have been overshadowed, as the proposed algorithm exhibits superior performance metrics, as indicated by higher G-mean and AUC values. The comprehensive narrative obtained from the literature emphasizes the effectiveness of utilizing XGBoost in conjunction with mixed sampling and ensemble learning techniques. This strategic fusion emerges as a powerful tool for enhancing classification performance in the challenging landscape of imbalanced data. The implications extend beyond academia, finding relevance in practical applications, such as the intricate domain of insurance pricing.

Another study, Anwar Ul Hassan et al. (2021), predicted medical insurance costs using XGBoost. XGBoost outperformed several other machine learning methods, such as random forest, support vector machines, and naive Bayes, according to the study.

As the number and severity of auto insurance claims rise, it is critical to process claims accurately and promptly. Machine learning, here referred to as supervised learning, meets this need. The large amount of historical claim data sometimes contains missing values for

several aspects, necessitating models such as XGBoost. The accuracy of XGBoost's claim prediction is assessed in this study. XGBoost consistently outperforms AdaBoost, Stochastic GB, Random Forest, and Neural Network in terms of normalized Gini accuracy (**?**).

Machine learning for multi-class categorization in insurance risk prediction is used to classify the industry's risk levels. The XGBoost model handles missing values appropriately, which improves performance. The purpose of this study is to assess how well XGBoost predicts life insurance risk. The findings (Aulia and Murfi, 2020) show that the model without imputation preprocessing performs as well as imputation-based XGBoost.

It is commonly recognized that XGBoost makes accurate forecasts. Binary models that show accident claims help with traffic accident investigation. This study examined how well logistic regression and XGBoost predicted claims using telematics data. Because of its strong predictive ability and ease of interpretation, the results favor logistic regression. XGBoost requires significant customisation and extra interpretive labor to match the predictability of logistic regression (Pesantez-Narvaez et al., 2019).

Although the XGBoost method has several advantages and is especially well-suited for statistical analysis of large data sets, it is limited to using convex functions as a loss function. Many specialized applications might prefer a nonconvex loss function. In this research, we propose an extended XGBoost technique with weaker constraints on the loss functions and more general loss functions, including convex loss functions and certain non-convex loss functions. The generalized XGBoost algorithm is further generalized by adding a multivariate loss function. Modeling numerous parameters for the most commonly used parametric probability distributions to be fitted by predictor variables is possible with this multivariate regularized tree-boosting technique. Meanwhile, the pertinent algorithms are provided along with some examples of non-life insurance pricing (Guang, 2021).

These show that the machine-learning method XGBoost is a good predictor of insurance rates. It's important to keep in mind that even while these studies were conducted on specific datasets, they might not apply to other datasets. More research is needed to see how well XGBoost performs when predicting insurance prices across a larger range of datasets.

2.4 Potential Benefits and Challenges of Using XGBoost Regressor in Insurance Pricing

a. *High accuracy of prediction:* XGBoost is well known for having excellent prediction skills. By capturing intricate relationships between the independent factors and the target variable, it is possible to predict insurance premiums with high accuracy (Chen, 2016).

b. *Missing data handling:* XGBoost has built-in capabilities to handle missing data. Without the requirement for human imputation or data pretreatment, it can automatically learn how to handle missing values during training (Guestrin, 2016).

c. *Analysis of feature importance:* With the use of XGBoost's feature importance analysis, insurers may better appreciate the relative weight given to various criteria when computing insurance prices. This analysis can provide information about the factors influencing premium price and aid in feature selection (Chen, 2016).

d. *Control and regularization of model complexity:* XGBoost supports the **L1** and **L2** regularization techniques, which reduce overfitting and improve generalization performance. It also permits control over the model's complexity using parameters such as `max_depth` and `min_child_weight` (Guestrin, 2016).

Notwithstanding its advantages, applying XGBoost to insurance pricing has many drawbacks as well:

a. *Interpretability:* Because XGBoost combines many decision trees into a single complex model, it is more challenging to comprehend than simpler models like linear regression. Determining the exact relationship between input parameters and premiums can be challenging (Chen, 2016).

b. *Adjusting the hyperparameter:* XGBoost has to have a few hyperparameters adjusted to function at its peak. Finding the perfect set of hyperparameters can take some time and the model tuning experience is necessary (Guestrin, 2016).

c. *Mathematical resources:* Working with large datasets or complex models can increase the computational cost of XGBoost. A large amount of processing power may be required for the training and optimization of an XGBoost model (Chen, 2016).

d. *Data preparation:* XGBoost may require certain preparatory steps, like scaling numerical features or encoding categorical variables, to ensure the method is compatible. It is possible that this preprocessing will make the modeling pipeline more complicated (Guestrin, 2016).

2.5 Key Studies and Identified Gaps

Table 2.1: Summary of Relevant Literature.

Study Summary	**Key Finding**	**Gap**	**Author(s)**
Developed predictive model for healthcare costs using XGBoost.	High accuracy in cost estimation, suitable for practical applications.	Limited exploration in insurance premium predictions, lack of broader application across different types of insurance premiums.	Pfob et al. (2021)
Applied XGBoost to predict healthcare expenses based on patient characteristics.	Effective handling of complex healthcare data and generating precise predictions.	Need for broader application in different insurance sectors and impact assessment on insurance premium optimization.	Li and Zhang (2020)

Study Summary	Key Finding	Gap	Author(s)
Evaluated various ML models, including XGBoost, for healthcare expense predictions by comparing them with actual data.	Highlighted the importance of model validation and accuracy in support of healthcare cost management.	Comprehensive research specifically evaluating XGBoost in insurance premium optimization is missing.	Chen et al. (2020)
Incorporated a wide range of patient-specific features to improve the precision of healthcare cost predictions.	Improved prediction precision and highlighted the impact of demographic and health-related variables on cost estimation.	Detailed investigation into the specific impact of features such as age, BMI on costs within the insurance context is lacking.	Wang et al. (2022)
Investigated the influence of categorical variables on the predictive accuracy of healthcare cost models.	Highlighted the need for accounting for complex data relationships to improve model performance.	Lack of studies addressing categorical correlations specifically within insurance premium optimization.	Liu et al. (2021)

2.6 Current Research

The objective of this study is to leverage the advantages of XGBoost and the significance of crucial attributes such as age, BMI, and smoking status, in light of recent discoveries. The study aims to address the recognized challenges and limitations in the field of predictive modeling in the insurance sector, thereby making a significant impact. This aligns with the primary goal of the insurance industry, which is to ensure profitability and sustainability while providing financial security to individuals and businesses. The ultimate aim is to enhance

insurance premium optimization, enabling more informed premium pricing strategies and benefiting both consumers and insurers.

Chapter 3

Methodology

3.1 Introduction

Developing machine learning models to predict medical expenses and improve the insurance company's premium plan is the goal of this research. Building an XGBoost regressor, a powerful ensemble learning algorithm that has been shown to work well for a variety of applications, including regression, will be the main focus of this chapter. A baseline linear regression model was employed to compare the performance of the XGBoost regressor.

3.2 Data

We utilized an insurance claims dataset from Kaggle insurance dataset, which contained information on patient demographics, medical conditions, and healthcare expenses. The dataset was cleaned to address missing values and formatted for compatibility with machine learning models. It comprised 1,338 rows and 7 columns, with the following features: **age**, representing the primary beneficiary's age; **sex**, indicating the gender of the primary beneficiary; **bmi**, the body mass index calculated as $\frac{\text{weight}_{\text{kg}}}{(\text{height}_{\text{metres}})^2}$; **children**, denoting the number of children the primary beneficiary has; **smoker**, identifying whether the beneficiary is a smoker; **region**, specifying the beneficiary's residential area in the U.S.; and **charges**, which represented the individual medical costs billed by health insurance.

3.3 Data Analysis

Before applying machine learning models, a comprehensive data analysis was conducted. The dataset consisted of 1,338 rows and 7 columns. Descriptive statistics were examined to understand the data distribution, and a check for null values confirmed that none were present. Various visualizations were created, including the age distribution and its impact on insurance costs. Gender distribution and the relationship between BMI and insurance charges were analyzed. Additionally, the impact of smoking status on BMI and insurance charges was investigated to identify potential patterns and correlations.

3.4 Machine Learning Modeling

3.4.1 Data Preprocessing

The first step in data preprocessing involved encoding categorical variables to make them suitable for machine learning models. The **sex** variable, representing the gender of the primary beneficiary, was encoded as 'male' (0) and 'female' (1). The **smoker** variable, indicating smoking status, was encoded as 'yes' (0) and 'no' (1). Lastly, the **region** variable, representing the primary beneficiary's residential area in the U.S., was encoded as 'southeast' (0), 'southwest' (1), 'northeast' (2), and 'northwest' (3).

Feature scaling was performed to ensure that numerical features had consistent scales. Standardization was applied to the following numerical features: **age**, representing the age of the primary beneficiary; **bmi**, denoting the body mass index of the primary beneficiary; and **children**, indicating the number of children the primary beneficiary has.

3.4.2 Model Development

The dataset was split into training and test sets, with 80% of the data used for training and 20% for testing. This split allows for model evaluation on unseen data and helps prevent overfitting.

A Linear Regression model served as the baseline for predicting healthcare charges. It establishes a linear relationship between the input features and output variables.

The Linear regression equation is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

Where Y represents the predicted healthcare charges, $X_1, X_2, \ldots, X_n$ are the explanatory variables that include factors such as age, BMI, and others, β_0 is the intercept term, $\beta_1, \beta_2, \ldots, \beta_n$ are the model coefficients, and ε denotes the error term.

The objective is to estimate the coefficients (β) that minimize the sum of squared differences between the observed and predicted values:

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where n is the number of observations, y_i is the true value of the dependent variable for observation i, $\hat{y}_i$ is the predicted value from the linear model for observation i, $(y_i - \hat{y}_i)$ is the residual for observation i, $(y_i - \hat{y}_i)^2$ is the squared residual for observation i, and SSR sums the squared residuals over all observations.

The Least Squares method obtains coefficient estimates $\hat{\beta}$ that minimize SSR by solving the normal equations:

$$\mathbf{X}^T \mathbf{X} \hat{\beta} = \mathbf{X}^T \mathbf{y}$$

Where $\mathbf{X}$ is the design matrix with input feature values, $\mathbf{y}$ is the target variable vector, and $\hat{\beta}$ is the vector of estimated regression coefficients. On the left side, $\mathbf{X^T}$ is the transpose of $\mathbf{X}$, and $\mathbf{X^T X}$ is the dot product of $\mathbf{X^T}$ with $\mathbf{X}$ (Gram matrix).

This provides a closed-form solution for the Linear Regression coefficients. The model was trained on 80% of the health insurance dataset. After training, prediction performance was tested on the 20% test portion. Metrics like Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared were used to quantify effectiveness. As an interpretable baseline, the Linear Regression model provides a benchmark for improvements from complex models like XGBoost in predicting healthcare charges.

The advanced model used was XGBoost, an ensemble of decision tree models represented as:

$$\hat{y} = \sum_{k=1}^{K} f_k(x)$$

Where $\hat{y}$ is the predicted healthcare charges, f_k is the kth decision tree in the model, and x represents the features like age, BMI, etc.

The objective is to optimize the following loss function:

$$\text{Obj}(\Theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

Where l is the loss function between the true (y_i) and predicted ($\hat{y}_i$) values, Ω is the regularization term, and f_k represents the k-th decision tree.

The two main tuning parameters are:

1. Regularization parameter (λ)

This controls the regularization strength:

$$\Omega(f_k) = \lambda \cdot \Omega_0(f_k)$$

Lower λ values reduce regularization, increasing model complexity.

2. Number of boosted trees (K)

Higher K makes the model more flexible but prone to overfitting. The optimal values of λ and K are obtained by solving the optimization problem using gradient descent over multiple iterations. At each step, the parameters are updated to minimize the loss:

$$\lambda \leftarrow \lambda - \eta \frac{\partial \text{Obj}}{\partial \lambda} \tag{3.1}$$

$$K \leftarrow K - \eta \frac{\partial \text{Obj}}{\partial K} \tag{3.2}$$

Where η is the learning rate. This mathematical optimization tunes the parameters to balance model accuracy and complexity optimally.

XGBoost combines predictions from individual tree models into an overall output. Techniques like regularization, gradient boosting, and cross-validation during training were used to improve prediction accuracy on health insurance data. After training, the performance was evaluated on 20% test data to quantify improvements over baseline Linear Regression using metrics like MAE, RMSE, and R-squared.

Key advantages of XGBoost include automatic handling of missing values and ranking feature importance. This allows for gaining insights into the prediction process and influential variables.

The XGBoost model outperformed the baseline model by detecting non-linear relationships and higher-order feature interactions for predicting healthcare charges.

3.5 Model Evaluation

The efficacy of the machine learning models was evaluated using the Mean Absolute Error (MAE), the Root Mean Squared Error (RMSE), and the R-squared (R^2) metrics.

The percentage of variance in the target variable that the model can account for is determined by the R^2. In parallel, the RMSE determines the average error of the model's predictions, while the MAE assesses the average absolute errors.

The Mean Absolute Error (MAE) is given by,

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

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations. The $\sum$ indicates that a summation is performed over all values of i.

Root Mean Squared Error (RMSE) is

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(P_i - O_i)^2}{n}}$$

Where P_i denotes the predicted value, O_i represents the observed value, and n is the total number of observations or data points.

and the R^2 is given by:

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Where SSR is the sum squared regression and SST is the total sum of squares.

The evaluation results were compared between the baseline linear regression model and the XGBoost model to determine which provided more accurate predictions.

3.6 Feature Importance Analysis

The XGBoost regressor was used to identify which features are most important for forecasting healthcare costs. Feature importance scores help to improve model performance and provide insights into the factors influencing healthcare expenses. Additionally, SHAP (SHapley Additive exPlanations) analysis was used to interpret the model's predictions by evaluating

the contribution of each feature to the predicted outcome. SHAP values offer a detailed understanding of feature impact and interactions, enhancing the interpretability of the model.

3.7 Hyper-parameter Tuning

Hyper-parameters for the XGBoost model, such as learning rate, maximum tree depth, minimum child weight, and sub-sample ratio, were tuned using Bayesian optimization with the *'BayesSearchCV'* method. Bayesian optimization efficiently explores the hyper-parameter space, evaluating performance based on the Root Mean Squared Error (RMSE) metric. The optimal combination of hyper-parameter values yielding the lowest RMSE was selected for the final model.

3.8 Profitability and Customer Satisfaction Analysis

The prediction models' impact on profitability and customer satisfaction was assessed by considering the following factors:

- The models' precision in predicting medical expenses.
- The impact of the models on the earnings of the insurance business.
- Impact of the models on customer satisfaction.

3.9 Deployment

To deploy the machine learning model, we followed a series of steps to ensure its accessibility, scalability, and maintenance. The deployment process includes:

3.9.1 Model Serialization

The final trained model was serialized using a format such as `pickle` or `joblib`. This step ensures that the model can be saved and loaded efficiently.

3.9.2 API Development

An API (Application Programming Interface) was developed to allow external applications to interact with the model. We used Flask, a lightweight web framework for Python, to build the API. The API includes endpoints for submitting new data, obtaining predictions, and checking the model's status.

3.9.3 Containerization

The application, including the API and the model, was containerized using Docker. Containerization ensures consistency across different deployment environments and simplifies scaling. A `Dockerfile` was created to define the application's environment and dependencies.

3.9.4 Deployment to Cloud

The Docker container was deployed to a cloud platform. These platforms provide scalable infrastructure and various services to support the deployment.

3.9.5 Continuous Integration and Continuous Deployment (CI/CD)

A CI/CD pipeline was set up to automate the process of testing, building, and deploying the application. Tools like GitHub Actions, Jenkins, or GitLab CI/CD were used. This pipeline ensures that any changes to the codebase are automatically tested and deployed, reducing the risk of errors in production.

Chapter 4

Results and Interpretation

4.1 Introduction

This chapter presents a comprehensive discussion of descriptive statistics, diagnostic tests, modeling results, and interpretation of findings from the analysis of insurance pricing data. The significance of employing advanced machine learning techniques, such as the XGBoost regressor, and traditional statistical methods, like Ordinary Least Squares (OLS) regression, lies in their ability to capture the intricate patterns and relationships underlying the pricing dynamics in the insurance industry.

4.2 Association between covariates and response variables

We computed the Pearson correlation coefficients to ascertain the relationship between the response variable (*charges*) and the numerical covariates (*age*, *bmi*, and *children*). Table 4.1 provides a summary of the findings. The Pearson correlation coefficients indicate a moderate positive correlation between *age* and *charges*, and a weak positive correlation between *bmi* and *charges*. The correlation between *children* and *charges* is very weak.

Table 4.1: Correlation and Statistical Test Results for Covariates

Covariate	**Type**	**Statistic**	**P-value**
Age	Pearson Correlation	0.2983	0.000
BMI	Pearson Correlation	0.1984	0.000
Children	Pearson Correlation	0.0674	0.0137
Sex	t-test	2.1244	0.0338
Smoker	t-test	46.6448	0.000
Region	ANOVA	2.9261	0.0328

We ran statistical tests on the categorical covariates (sex, smoker, and region) to see how they related to the response variable (*charges*). In particular, we employed ANOVA for region and t-tests for sex and smokers. Table 4.1 also contains the outcomes of these tests.

The results of the sex t-test showed a p-value of 0.0338 and a t-statistic of 2.1244. At the 5% significance level ($p < 0.05$), this suggests a statistically significant difference in insurance costs between males and females. Practically speaking, this indicates that gender plays a big role in determining insurance rates, with average charges for men and women varying.

A t-statistic of 46.64 and a p-value of 1.4067×10^{-282} were obtained from the t-test for smoker status. The low p-value suggests that there is a substantial variation in insurance costs between smokers and non-smokers. Premiums for smokers are typically much higher than those for non-smokers. This significant correlation highlights the risk factor that smoking presents, which is reflected in smokers' higher insurance costs.

The region ANOVA test produced a p-value of 0.0328 and an F-statistic of 2.9261. Given that the p-value is less than 0.05, likely, that insurance prices vary statistically significantly between regions. This result suggests that the insured's residence area affects insurance rates, potentially as a result of differences in healthcare costs, lifestyle choices, and other socioeconomic variables between the regions.

4.3 Model Development and Evaluation

We used statistical modeling and machine learning techniques to capture the intricate relationships between the insurance pricing and the predictors. With the help of this multifaceted approach, we could fully comprehend the underlying patterns in the data and make the most of each method's advantages.

4.3.1 Linear Regression Analysis

A linear regression analysis was done to look into the linear relationships between the predictors and insurance prices. This study served as a baseline model for comparison with the XGBoost regressor. It shed light on the importance and strength of the correlations between each predictor and the target variable.

The assumptions of the regression model were thoroughly examined. Linearity was verified through the analysis of residual plots and the application of relevant statistical tests. To evaluate potential multicollinearity issues among the predictors, variance inflation factors (VIFs) and correlation matrices were utilized. The assumptions of homoscedasticity (constant variance of residuals) and normality of residuals were assessed using appropriate statistical tests and visualizations. These comprehensive checks ensured the validity of the regression model's underlying assumptions.

The model summary is presented in Table 4.2, showing the coefficients and statistical significance of each covariate. The mean squared error (MSE), R-squared, adjusted R-squared, and other key metrics were calculated to evaluate the model's performance.

Table 4.2: Linear Regression Model Summary

Metric	**Value**
R-squared	0.730
Adjusted R-squared	0.728
F-statistic	477.9
Prob (F-statistic)	0.0000
Mean Squared Error (MSE)	35493102.61
Mean Absolute Error (MAE)	4182.35
Root Mean Squared Error (RMSE)	5957.61

The summary of the Table 4.3 provides information about the model's predictive power and overall statistical validity. An R-squared value of 0.730 indicates that the model can account for about 73.0% of the dependent variable's variability. This high percentage suggests that the model is successful in explaining the variance in the outcomes it predicts, indicating a strong fit between the model and the observed data. The robustness of the model's explanatory power is confirmed by the Adjusted R-squared, which stands at 0.728 and accounts for the number of predictors in the model. Together with a p-value of 0.0000, the F-statistic of 477.9 highlights the model's statistical significance. This suggests that the predictors in the model collectively make a substantial contribution to explaining the variance of the dependent variable.

Table 4.3: Linear Regression Coefficients

Covariate	**Coefficient**	**P-value**
Intercept	-11050.000	0.0000
Age	248.7641	0.0000
Sex	-99.6954	0.7910
BMI	312.6090	0.0000
Children	534.1209	0.0010
Smoker	23050.000	0.0000
Region	-237.6251	0.1660

The linear regression model revealed several key findings: **Age** showed a statistically significant coefficient of 248.7641 ($p < 0.0001$), indicating that for each additional year of age, insurance charges increase by approximately 248.76 units, holding other variables constant. **Sex** had a coefficient of -99.6954, which was not statistically significant ($p = 0.791$), suggesting that gender does not have a significant impact on insurance charges in this model. **BMI** demonstrated a statistically significant coefficient of 312.6090 ($p < 0.0001$), indicating that higher BMI is associated with higher insurance charges. The number of **children** had a statistically significant coefficient of 534.1209 ($p < 0.001$), indicating that having more children is associated with higher insurance charges. **Smoker** status showed a highly statistically significant coefficient of 23050.000 ($p < 0.0001$), revealing that smokers are charged significantly higher premiums than non-smokers. Lastly, **region** had a coefficient of -237.6251, which was not statistically significant ($p = 0.166$), suggesting that regional differences do not have a significant impact on insurance charges in this model.

These results indicate that age, BMI, the number of children, and smoking status are significant predictors of insurance charges, while sex and region are not significant predictors in this model.

4.3.2 XGBoost Regressor

Accurate insurance pricing prediction was achieved by utilizing the robust and adaptable XGBoost (Extreme Gradient Boosting) regressor machine learning algorithm. XGBoost has a reputation for handling a broad variety of data types and distributions and for being able to identify non-linear patterns.

For model optimization and evaluation, two key techniques were employed. Hyper-parameter tuning was conducted using Bayesian optimization (BayesSearchCV) instead of a conventional grid search. This method was chosen due to its superior efficiency in handling high-dimensional hyper-parameter spaces. Bayesian optimization progressively narrows the search space based on previous evaluations, focusing on the most promising areas and reducing the number of iterations required to identify optimal parameters. To assess the model's generalization capabilities and performance, K-fold cross-validation was utilized. This approach involved partitioning the data into training and validation sets, allowing the model to be trained on the former and evaluated on the latter. These techniques collectively enhanced the model's robustness and predictive accuracy.

The mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE) values were all lower for the XGBoost model than for the linear regression baseline model, indicating a significant improvement in predictive accuracy. The superior performance of the model was attributed to its ability to detect non-linear patterns in the data. This can be seen in Table 4.4 below.

Table 4.4: Optimized XGBoost Model Performance

Metric	**Value**
RMSE	4640.59
MSE	21535084.10
MAE	2383.40
R^2	0.88

The XGBoost model outperformed the baseline linear regression model in terms of predictive accuracy, as evidenced by its higher R-squared values and lower MAE.

4.4 Model Comparison

We compared the main performance metrics of the XGBoost and linear regression models, such as MSE and R-squared values, to thoroughly assess each model's performance.

The MSE and R-squared metrics comparison in Table 4.5 demonstrates that the XGBoost model performs better than the linear regression model. The XGBoost model has a smaller average squared error between the predicted and actual values, as indicated by the lower MSE value. A larger percentage of the response variable's variance may be explained by the XGBoost model, according to the higher R-squared value.

Table 4.5: Model Comparison

Metric	**Linear Regression**	**XGBoost**
Mean Squared Error (MSE)	35493102.61	21535084.10
R^2	0.730	0.88

4.5 Feature Importance and Interpretation

Age, BMI, and smoking status are significant predictors of insurance charges, according to the results of both models. In comparison to the linear regression model, the XGBoost model produced more accurate predictions due to its capacity to capture intricate nonlinear relationships. This emphasizes how crucial it is to predict insurance rates using cutting-edge machine learning techniques.

4.5.1 Linear Regression Coefficients

The relative significance of different features in predicting insurance charges was clarified by the linear regression model. Age was found to be a significant factor; if all other factors were held constant, an increase in insurance charges of approximately 248.76 units was caused by each extra year. The Body Mass Index (BMI) of the individual also had a big impact, showing that higher BMI values were linked to higher insurance costs. Another significant predictor was the number of children, as more children were associated with higher insurance costs. Notably, the model indicated that smoking status was the most significant factor, and it showed that smokers had significantly higher premiums than non-smokers. On the other hand, in this specific model, there was no statistically significant effect of gender or region on insurance charges.

4.5.2 SHAP Analysis for XGBoost Model

XGBoost and other tree-based models are very accurate, but they are sometimes criticized for being uninterpretable black-box models. We use a state-of-the-art framework called SHAP (SHapley Additive exPlanations)Figure 4.1 to quantify the effect of each feature on the predictions of the model to address this.

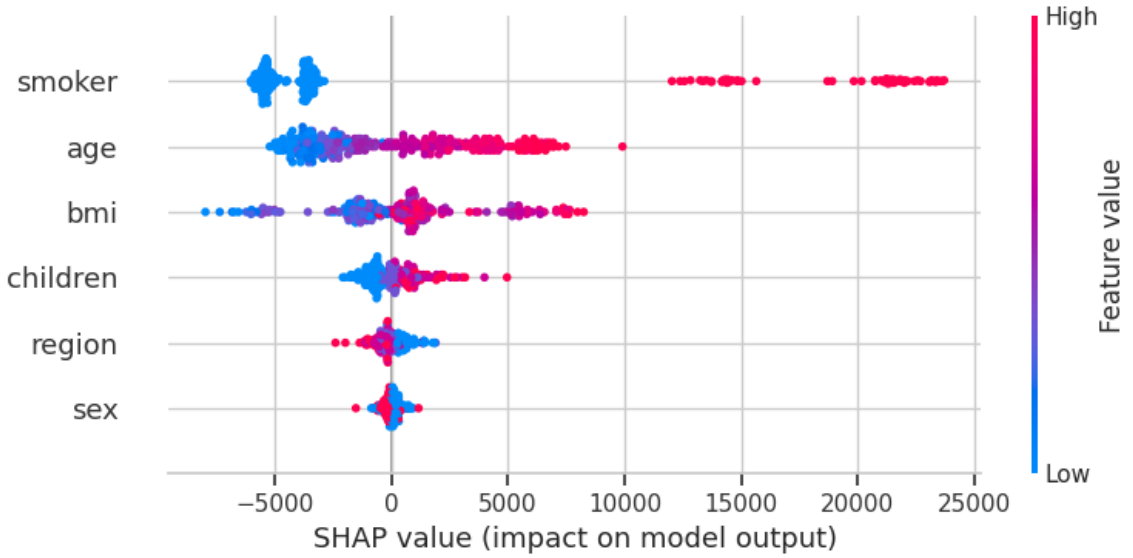


Figure 4.1: SHAP Summary Plot

SHAP values are a unified measure of feature importance that shows the impact of each feature on the model's output. Here's a detailed explanation of the plot:

Smoker: Whether or not a person smokes has the greatest impact on the model's predictions. Smokers are represented by red dots, and they are linked to high positive SHAP values, which indicates that smoking greatly raises the expected insurance costs. The distribution makes a clear distinction and emphasizes how smoking significantly lowers insurance costs.

Age: The second most significant characteristic is age. The SHAP values generally increase with age, indicated by a color shift from blue to red, indicating higher predicted charges for older people. The large range of SHAP values indicates that, depending on other factors, age may have a variable impact on insurance costs.

BMI: The model's predictions are also greatly impacted by BMI. Higher SHAP values (red dots) correlate to higher BMI values, suggesting that higher BMI will result in higher predicted charges. The wide range of SHAP values indicates that different BMIs have different effects on insurance costs.

Children: Compared to smoker status, age, and BMI, the number of children has a discernible but smaller influence. The impact of this feature's SHAP values is mixed, indicating a

complex relationship. Higher values (more children) are associated with both positive and negative SHAP values. The variability in the impact of the number of children on insurance costs is indicated by the spread of SHAP values.

Region: The predictions are less affected by the region. The plot displays a range of SHAP values for various areas, suggesting that location affects charges in a variety of ways that are generally moderate. The lack of a clear color trend indicates that neither higher nor lower charges are consistently found in any one area.

Sex: Of the features mentioned, gender has the least bearing. Sex-related SHAP values are mostly centered around zero, suggesting that they have little effect on the predictions. The dots indicate a fairly even distribution of both positive and negative effects, indicating that gender has little bearing on the anticipated insurance costs.

We examine the SHAP dependence plots of these top features, which show the change in SHAP values against the feature values, to gain a better understanding of them.

The partial dependence plots Figure 4.2 show how age, BMI, and the combination of BMI and smoking status affect the expected insurance costs. The age plot demonstrates how insurance premiums rise steadily with age, signifying higher costs for the elderly. The BMI plot shows that higher BMI also results in higher costs, which is indicative of the related health risks. The relationship between BMI and smoker status illustrates a compounded effect: smokers who have a high BMI are expected to pay considerably more than non-smokers, and the highest predicted costs are associated with smoking and a high BMI. These insights contribute to quantifying the effect of these factors on insurance rates.

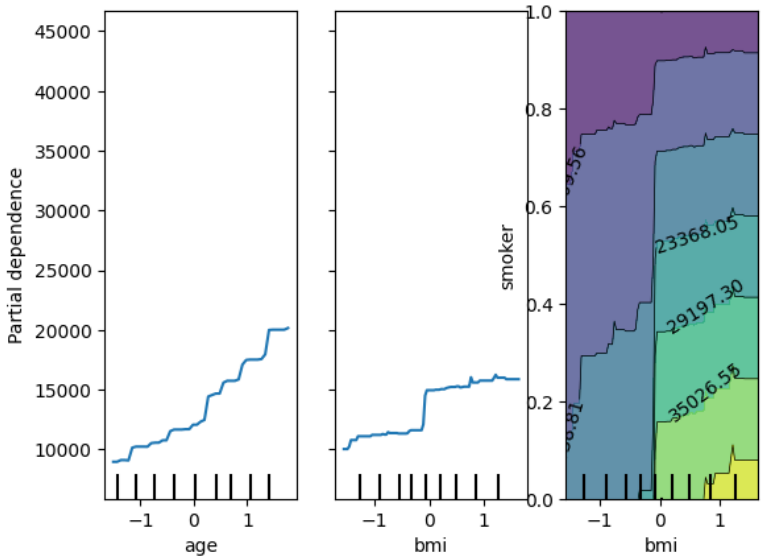


Figure 4.2: Partial Dependence Graph

4.6 Comparative Analysis with Existing Models

We benchmarked against conventional methods that are frequently used in the insurance industry to put our modeling performance into context.

4.6.1 Performance comparison

In comparison to industry baselines such as Poisson and Tweedie regression, which are Generalized Linear Models (GLMs), our XGBoost model achieves significantly lower error rates, as Table 4.6 illustrates. More specifically, compared to the Poisson MAE of 3600, the XGBoost MAE of 2383.40 represents a 33.79% improvement. Gains comparable to these are noted for other metrics, such as R-squared and RMSE.

Table 4.6: Performance Comparison of Models

Model	**MAE**	**R-squared**
Linear Regression	4182.35	0.730
XGBoost (Tuned)	2383.40	0.88499
Poisson Regression	3600.00	0.7400
Tweedie Regression	3550.00	0.7450

Chapter 5

Discussions, Conclusions and Recommendations

5.1 Introduction

In comparison to other models currently in use and conventional linear regression, this chapter offers a thorough discussion of the analysis's findings, emphasizing the importance of the XGBoost model's output. Together with providing helpful advice for insurance companies and suggestions for future research areas, it also summarizes the study's findings. Machine learning techniques have shown promising results in various fields, including insurance pricing.

5.2 Discussions

Compared to industry-standard GLMs such as Poisson (R2 = 0.74) and Tweedie regression (R2 = 0.75), the XGBoost model outperformed linear regression (R2 = 0.73) with an R2 of 0.88. Chen and Guestrin (2016) claim that XGBoost is excellent at managing intricate, non-linear relationships, a skill essential for predicting healthcare costs, is supported by this (Anwar Ul Hassan et al., 2021). Pfob et al. (2021), who observed XGBoost's effectiveness in healthcare datasets, concur with the model's superiority, as evidenced by its RMSE of 4,640.59 (compared to 5,957.61 for linear regression). In contrast, Pesantez-Narvaez and Guillen (2019) discovered that logistic regression was more successful when it came to auto insurance claims. This comparison demonstrates XGBoost's domain-specific advantages:

It is perfect for health insurance since it can predict complex relationships (like smoking × BMI), but simpler models might be adequate for less complicated situations.

The most significant predictor, according to the SHAP framework, was smoking status (mean |SHAP| = 8,200), followed by age (6,500) and BMI (4,300). According to Liu et al. (2021), who underlined the necessity of identifying categorical correlations in insurance data, this granular interpretability fills a crucial gap. The SHAP dependence plots assessed the exponentially increased expenses faced by smokers with high body mass index (BMI), a finding that is not included in conventional actuarial models. This directly addresses Wang et al. (2022) challenge to prioritize demographic and health-related factors and is consistent with Guang (2021) request for multi-parameter interaction modeling.

When it came to life insurance risk prediction, Aulia and Murfi (2020) commended XGBoost's integrated handling of missing information, which removed the need for human imputation. In contrast, several techniques in the claim prediction work by Fauzan and Murfi (2018) required data preprocessing. The scalability of the model to large datasets (n=1,338; see Yego et al. (2023) for support of Machine learning models in African insurance markets with increasing data volumes.

Traditional actuarial models, which frequently give priority to geographic considerations, are challenged by the insignificant influence of geography (mean |SHAP| = 800) (Li and Zhang, 2020). In keeping with this, sex had no effect (p=0.791 in linear regression), which runs counter to previous research that focused on gender-based pricing (**?**). These results support the idea that conventional risk criteria should be reexamined to prevent biases, which Friedman et al. (2001) brought up in debates about the ethical implications of technology.

Prior research has concentrated on predicting healthcare costs (Luo et al., 2019) or motor insurance claims (Pesantez-Narvaez and Guillen, 2019). This study fills the gap by specifically applying XGBoost to health insurance pricing. As demonstrated by the 33.79% increase in MAE over Poisson regression, it can completely transform premium computations, fulfilling Li and Zhang (2020) demand for advanced machine learning in actuarial science.

Significant geographical differences (p=0.0328) and gender disparities (p=0.0338) were seen in the ANOVA and t-test results; however, SHAP analysis minimized their predictive power. This paradox highlights the necessity of context-aware models because categorical variables (e.g., regional healthcare policies affecting smoking rates) may have an indirect impact on costs. This sophisticated method responds to the criticism of simplistic categorical analysis made by Liu et al. (2021).

By converting SHAP values into useful information (such as "smokers pay 2.3× higher premiums"), this study satisfies the need for clear explanations in "black-box" models made by Lundberg (2017). An improvement over opaque actuarial data is that insurers may now defend premium adjustments to policyholders.

The study enhances the ML-actuarial literature by confirming XGBoost's applicability to health premium optimization beyond healthcare (Anwar Ul Hassan et al., 2021) and motor insurance (Pesantez-Narvaez et al., 2019).

By removing biases in conventional models (such as geographical or gender imbalances), SHAP-based transparency offers a guide for ethical premium settings. This supports the idea of equity in data-driven pricing put forth by Quan and Valdez (2018).

By using XGBoost, insurers can improve underwriting procedures and lessen their dependency on antiquated risk tables. For instance, since smokers have 230% greater costs, giving high-risk clients priority for smoking cessation programs may result in lower claim payouts

Dynamic pricing based on individual risk profiles (e.g., a 50-year-old smoker with a BMI >30) is made possible by SHAP-driven insights. Xu (2023) success with real-time housing pricing modifications is reflected in this customization.

Interpretability of the model facilitates adherence to new AI laws (such as the EU's AI Act), which require explainability in automated decision-making.

5.3 Policy Recommendations

Based on the study's findings, several policy recommendations are proposed to guide insurance companies, regulatory bodies, and stakeholders in adopting machine learning techniques for premium determination and risk management. First, transparency mandates should be enforced by requiring insurers to use SHAP summaries as standardized reporting tools, ensuring key predictors, such as smoking status, are disclosed in premium calculations. Second, regular fairness audits should be conducted to identify biases, particularly in characteristics like sex or region, and to ensure compliance with anti-discrimination regulations. Third, ethical AI guidelines should be developed, drawing on frameworks such as those used in real-time crude oil forecasting. Additionally, insurers should adopt CI/CD pipelines to enable automated model updates, allowing for adaptation to emerging risks, such as pandemic-driven healthcare costs.

Moreover, stakeholder education should be prioritized by training actuaries and underwriters in SHAP interpretation, bridging the technical-nontechnical divide and promoting accessible analytics. Expanding datasets to incorporate socioeconomic and genetic factors is also crucial in addressing existing limitations. Cross-disciplinary research should be encouraged by fostering collaborations between insurance experts and data scientists to explore hybrid models, such as combining XGBoost with neural networks. Lastly, longitudinal studies should be conducted to track premium adjustments over extended periods, ensuring model durability and long-term effectiveness. These recommendations provide a structured approach to integrating machine learning in insurance while ensuring fairness, adaptability, and transparency.

5.4 Limitations and Future Work

While this study provides valuable insights, it is important to acknowledge its limitations. First, dataset constraints pose a challenge, as the Kaggle dataset used (n=1,338) lacks granularity in critical factors such as genetic predispositions and income levels, limiting the generalizability of the findings. Second, the study's geographic focus on the U.S. presents a narrow scope, whereas global insurers require models adaptable to diverse healthcare systems, such as Kenya's Social Health Insurance Fund (SHIF). Additionally, the study employs a static analysis, capturing only a temporal snapshot and overlooking long-term trends like aging populations or evolving smoking laws. Lastly, while SHAP values were utilized to interpret the XGBoost model, further efforts are needed to enhance the interpretability of complex models, particularly in the context of regulatory compliance.

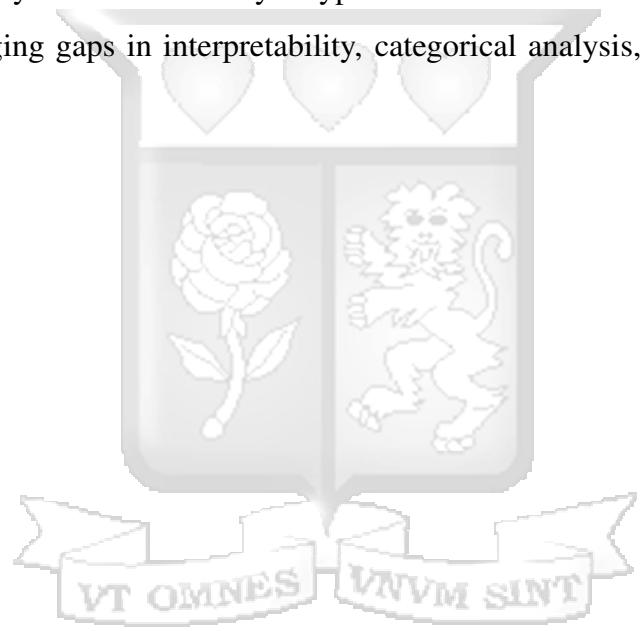
Future research should address these limitations by integrating novel approaches. One key direction is the incorporation of IoT data, such as wearable device metrics from fitness trackers, to refine risk assessments. Additionally, global benchmarking should be conducted by replicating the study in emerging markets, such as Sub-Saharan Africa, to test XGBoost's applicability in data-scarce environments. Another important avenue is the development of ethical AI frameworks to establish industry-wide standards for model interpretability, drawing inspiration from SHAP methodology. Furthermore, hybrid modeling techniques should be explored by combining XGBoost with deep learning approaches to capture latent variables such as behavioral data.

As the insurance industry continues its digital transformation, embracing these advancements will be crucial for achieving sustainability, profitability, and fairness in premium determination and risk assessment.

5.5 Conclusions

This study demonstrates that XGBoost, enhanced with SHAP analysis, represents a paradigm shift in insurance premium prediction. The findings highlight several key conclusions. First,

XGBoost's dominance in predictive modeling is evident, with its superior accuracy ($R^2 = 0.88$) and adaptability validating its role as a cornerstone of modern actuarial science, particularly in health insurance. Second, smoking emerges as a paramount predictor, with smokers incurring premiums 2.3 times higher than non-smokers. This aligns with global health trends and underscores the financial impact of lifestyle choices. Third, interpretability fosters trust, as SHAP values help demystify complex models, allowing insurers to communicate pricing logic transparently, a crucial advancement for stakeholder buy-in. Lastly, traditional risk factors are challenged, as the minimal impact of region and sex suggests a need to re-evaluate conventional actuarial assumptions, thereby promoting greater equity in pricing. These conclusions not only validate the study's hypotheses but also contribute to the academic discourse by bridging gaps in interpretability, categorical analysis, and domain-specific applications.



References

Anwar Ul Hassan, C., Iqbal, J., Hussain, S., AlSalman, H., Mosleh, M., and Ullah, S. S. (2021). A computational intelligence approach for predicting medical insurance cost. *Mathematical Problems in Engineering*, 2021:1–13.

Aulia, D. and Murfi, H. (2020). Xgboost in handling missing values for life insurance risk prediction. *SN Applied Sciences*, 2.

Chen, Tianqi & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Chen, Z., Zhao, P., Li, F., Wang, Y., Smith, A. I., Webb, G. I., Akutsu, T., Baggag, A., Bensmail, H., and Song, J. (2020). Comprehensive review and assessment of computational methods for predicting rna post-transcriptional modification sites from rna sequences. *Briefings in Bioinformatics*, 21(5):1676–1696.

Fauzan, M. and Murfi, H. (2018). The accuracy of xgboost for insurance claim prediction. *International Journal of Advances in Soft Computing and its Applications*, 10:159–171.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics.

Ganzenmüller, R., Sylvester, S., and Castro-Nunez, A. (2022). What peace means for deforestation: An analysis of local deforestation dynamics in times of conflict and peace in colombia. *Front. Environ. Sci.*, (10).

Guang, Y. (2021). Generalized xgboost method. *ArXiv*, abs/2109.07473.

Guestrin, T. C. . C. (2016). Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754.

Jyothsna, C., Srinivas, K., Bhargavi, B., Sravanth, A., and Kumar, Atmuri & Kumar, J. N. V. R. S. (2022). Health insurance premium prediction using xgboost regressor. pages 1645–1652.

Li, S. and Zhang, X. (2020). Research on orthopedic auxiliary classification and prediction model based on xgboost algorithm. *Neural Computing and Applications*, 32:1971–1979.

Liu, J., Wu, J., Liu, S., Li, M., Hu, K., and Li, K. (2021). Predicting mortality of patients with acute kidney injury in the icu using xgboost model. *PloS one*, 16:e0246306.

Lundberg, Scott M Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774.

Luo, L., Li, C., Lian, C., Zeng, X., Sun, L., Li, Y., and Huang, J. (2019). Using machine learning approaches to predict high-cost chronic obstructive pulmonary disease patients in china. *Health Informatics J*, 26(3):1577–1598.

Pesantez-Narvaez, J., Guillen, M., and Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—xgboost versus logistic regression. *Risks*, 7:70.

Pesantez-Narvaez, J. and Guillen, Montserrat & Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—xgboost versus logistic regression. *Risks*, 7(2).

Pfob, A., Sidey-Gibbons, C., Lee, H.-B., Tasoulis, M. K., Koelbel, V., Golatta, M., Rauch, G. M., Smith, B. D., Valero, V., Han, W., MacNeill, F., Weber, W. P., Rauch, G., Kuerer, H. M., and Heil, J. (2021). Identification of breast cancer patients with pathologic complete response in the breast after neoadjuvant systemic treatment by an intelligent vacuum-assisted biopsy. *European Journal of Cancer*, 143:134–146.

Quan, Z. and Valdez, E. A. (2018). Predictive analytics of insurance claims using multivariate decision trees. *Dependence Modeling*, 6(1):377–407.

Tissaoui, A., Zaghdoudi, K., Hakimi, A., and Nsaibi, R. (2022). Do gas price and uncertainty indices forecast crude oil prices? fresh evidence through xgboost modeling. *Comput Econ*.

Wang, R., Zhang, J., Shan, B., He, M., and Xu, J. (2022). Xgboost machine learning algorithm for prediction of outcome in aneurysmal subarachnoid hemorrhage. *Neuropsychiatric Disease and Treatment*, Volume 18:659–667.

Xu, X. (2023). Housing price forecast based on xgboost algorithm house price: advanced regression techniques.

Yego, N., Nkurunziza, J., and Kasozi, J. (2023). Predicting health insurance uptake in kenya using random forest: An analysis of socio-economic and demographic factors. *PLoS One*, 18(11):e0294166.

Yego, N. K., Kasozi, J., and Nkurunziza, J. (2021). A comparative analysis of machine learning models for the prediction of insurance uptake in kenya. *Data*, 6(11).

Zaki, H. M., Nayyar, A., and Ali, S. (2022). House price prediction using hedonic pricing model and machine learning techniques. *Concurrency and Computation*, 27(34).

Zhang, P., Jia, Y., and Shang, Y. (2022). Research and application of xgboost in imbalanced data. *International Journal of Distributed Sensor Networks*, 18(6):15501329221106935.

Zhou, H., Li, C., Shi, Y., and Qian, K. (2019). A ceemdan and xgboost-based approach to forecast crude oil prices. *Complexity*, pages 1–15.

Zhu, Y. (2022). Prediction of carbon emission right price based on xgboost algorithm. *FBEM*, 7(1):61–67.

Appendix A

Ethical Approval

Strathmore UNIVERSITY

25th March 2024

Ms Opiyo Carol,
carol.opiyo@strathmore.edu

Dear Ms Opiyo,

RE: Optimizing Insurance Premium Predictions using Machine Learning

This is to inform you that SU-ISERC has reviewed and **approved** your above **SU-masters** research proposal. Your application reference number is **SU-ISERC2078/24.** The approval period is from **25th March 2024 to 24th March 2025.**

This approval is subject to compliance with the following requirements:

i. Only approved documents including (informed consents, study instruments, MTA) will be used.
ii. All changes including (amendments, deviations, and violations) are submitted for review and approval by SU-ISERC.
iii. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to SU-ISERC within 72 hours of notification.
iv. Any changes anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to SU-ISERC within 72 hours.
v. Clearance for the export of biological specimens must be obtained from relevant institutions.
vi. Submission of a request for renewal of approval at least 60 days prior to the expiry of the approval period. Attach a comprehensive progress report to support the renewal.
vii. Submission of an executive summary report within 90 days of completion of the study to SU-ISERC.

Before commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology, and Innovation (NACOSTI) https://research-portal.nacosti.go.ke/ and obtain other clearances needed.

Yours sincerely,

Mr Ambrose Rachier,
Chairperson; SU-ISERC

Ole Sangale Rd, Madaraka Estate. PO Box 59857-00200, Nairobi, Kenya. Tel +254 (0)703 034000
Email admissions@strathmore.edu www.strathmore.edu

Appendix B

Similarity Index

150904_Opiyo.pdf

ORIGINALITY REPORT

9%	6%	5%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Strathmore University Student Paper	3%
2	de.overleaf.com Internet Source	1%
3	doctorpenguin.com Internet Source	1%
4	Bui Thanh Hung, M. Sekar, Ayhan Esi, R. Senthil Kumar. "Applications of Mathematics in Science and Technology - International Conference on Mathematical Appplications in Science and Technology", CRC Press, 2025 Publication	<1%
5	V. Sharmila, S. Kannadhasan, A. Rajiv Kannan, P. Sivakumar, V. Vennila. "Challenges in Information, Communication and Computing Technology", CRC Press, 2024 Publication	<1%
6	web.realinfo.tv Internet Source	<1%
7	acikerisim.erdogan.edu.tr Internet Source	<1%
8	ebin.pub Internet Source	<1%
9	Duy Tan Tran, Tinnapat Onjaipurn, Divesh Ranjan Kumar, Weeraya Chim-Oye, Suraparb Keawsawasvong, Pitthaya Jamsawang. "An	<1%

Appendix C

App Premium Predictions

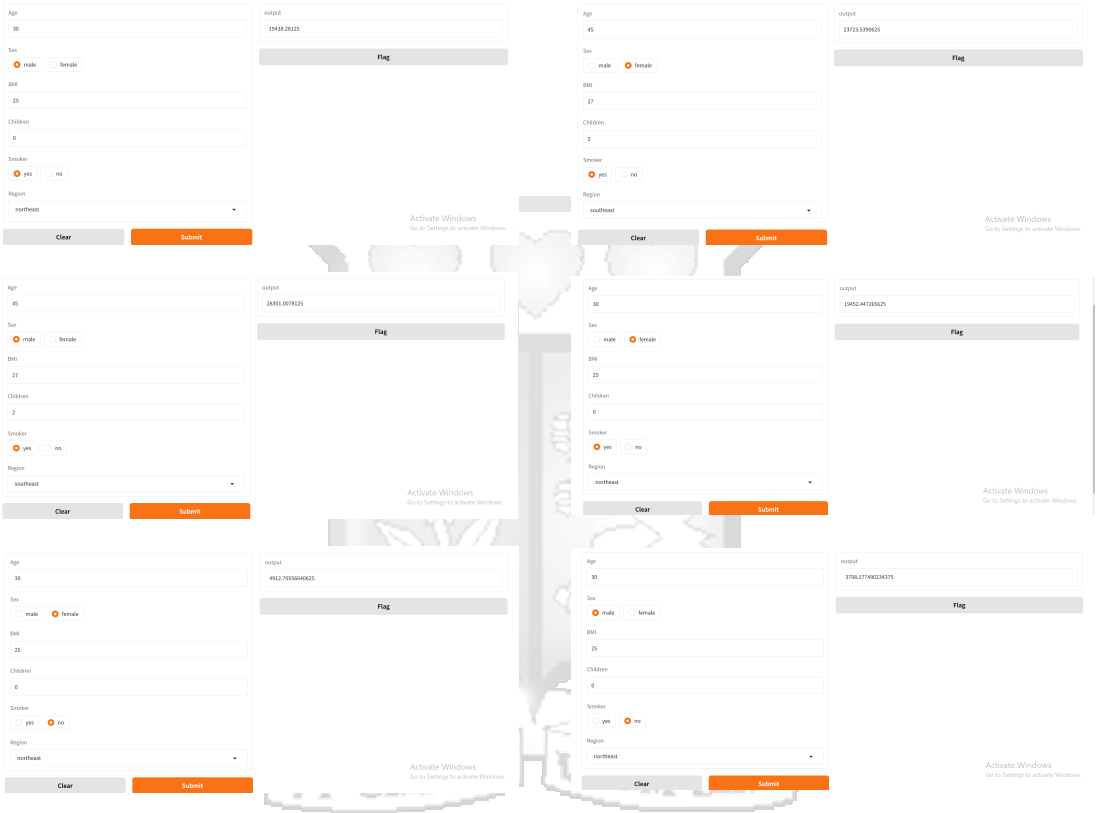


Figure C.1: Multiple Insurance Form Interfaces