5-2024

Predicting Health Insurance Claim Costs: A Data-Driven Approach Using Machine Learning

Fatema Khela
fmk6774@rit.edu

Predicting Health Insurance Claim Costs: A Data-Driven Approach Using Machine Learning

by

Fatema Khela

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology (RIT Dubai)

May 2024

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Thesis Approval

Student Name**: Fatema Khela**

Thesis Title**: Predicting Health Insurance Claim Costs: A Data-Driven Approach Using Machine Learning**

Graduate Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Hammou Messatfa **Date:**

Member of committee

Acknowledgments

I want to express my deepest gratitude to all the participants in this project. I am immensely thankful for the contributions from the Kaggle community, as they played a crucial role in my thesis. The analysis undertaken was greatly enhanced by the availability of various datasets generously provided by the Kaggle team.

I also sincerely thank my professors and mentors, especially Professor Hammou Messatfa and Professor Sanjay Modak, for their invaluable guidance throughout the research process.

In conclusion, I want to convey my heartfelt thanks to all the healthcare providers and policyholders whose de-identified information made this research possible. I genuinely appreciate your significant contributions.

Abstract

Determining appropriate premiums for policyholders is a challenge faced by the healthcare insurance industry. Policyholders' judgments about their healthcare are negatively impacted by the growing difficulty in accurately predicting claim amounts. To overcome this difficulty, our study used data-driven methods to project the cost of health insurance claims. Claim expenses are influenced by several criteria, including claim costs, age, gender, weight, BMI, number of dependents, smoking habits, blood pressure, diabetes, exercise routines, occupation, city of residence, and hereditary diseases.

The primary aim of this research is to develop predictive models that can accurately estimate the cost of health insurance claims based on policyholder attributes. Specifically, we strive to Investigate the correlation between policyholder characteristics and claim amounts; Explore the potential of machine learning techniques, such as XGBoost Tree 1, Random Trees 1, Linear-AS 1, LSVM 1, and Neural Net 1 to enhance cost predictions; Evaluate the performance of different predictive models using real- world health insurance data; and Assess the implications of model accuracy and its relevance to the insurance industry.

This research project will leverage a diverse dataset from Kaggle, encompassing a wide range of policyholder attributes. We will employ various machine learning techniques, including XGBoost Tree 1, Random Trees 1, Linear-AS 1, LSVM 1, and Neural Net 1, to develop predictive models. Additionally, we will utilize feature engineering and data preprocessing techniques to improve the predictive capabilities of these models.

The study investigated how machine learning models might be used to more accurately and automatically anticipate costs in the health insurance market. The current work evaluates the performance of five machine learning models—XGBoost Tree 1, Random Trees 1, Linear-AS 1, LSVM 1, and Neural Net 1 to handle a particular predictive problem. Thirteen features in a large dataset were used to train and test the models. The outcomes show that every model was used, and that correlation and construction time measures were used to evaluate each model's performance. The models with the highest correlation, XGBoost Tree 1 and Random Trees 1 were found to be 0.950 and 0.926, respectively. A correlation of 0.920 was observed in the

Linear-AS 1 model, whereas LSVM 1 and Neural Net 1 had correlations of 0.871 and 0.899, respectively. The build time for all models was under one minute, indicating their computational efficiency.

These findings suggest that the XGBoost Tree 1 model exhibits the most robust predictive performance among the evaluated models, offering valuable insights for model selection and further analysis in the given predictive task. According to the study's conclusions, insurers and government policymakers should use data-driven strategies like XGBoost to improve their decision-making and prediction capacities. Data scientists and healthcare experts must work with insurers and legislators to perform predictive modeling in the insurance sector.

Keywords: *Health insurance, cost prediction, XGBoost Tree 1, Random Trees 1, Linear-AS 1, LSVM 1, Neural Net 1, feature engineering, policyholder attributes, accuracy, ethical considerations and claim costs.*

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

1.1 Background Information

The increasing importance of health insurance claim cost prediction in the global healthcare and insurance industries is due to the expanding insurance sector and the growing significance of personal health data. This research aims to optimize cost estimation and decision-making in the insurance sector. Throughout history, healthcare and finance have relied on health insurance and risk assessment. The rising costs of health, like in Germany, have significant economic importance and have emphasized the need for accurate pricing of insurance (Drewe-Boss et al., 2022). Cost estimation accuracy influences various stakeholders, including insurers, healthcare providers, and policyholders. We listened to the growth of deep techniques such as numerical approach and deep neural network architectures, which promise to address the challenges of high dimensional data. Insurance is a policy that eliminates or decreases loss costs due to various risks (Hanafy & Mahmoud, 2021). We have various factors that impact the cost of insurance. Like age, the younger an individual, the lower their payments; also, women live longer than men, which may make them impact insurance costs. Older individuals pay more for healthcare insurance since they generally need more medical care, whereas a 55-year- old pays nearly twice as much as a 30-year-old (Sleight, 2023). It is evident that there is a need for more data-driven approaches in health insurance cost estimation. Researchers have used different ML methods in focusing insurance costs like Deep Neural networks, K nearest, Random Forest Regression, Multiple Linear Regression, etc (Hanafy & Mahmoud, 2021).

This research addresses the need for improved prediction of health insurance costs. Inaccurate predictions impact both insurers and policyholders, resulting in financial losses, suboptimal pricing, inadequate coverage, and affordability challenges. Accurate cost predictions are crucial for making informed decisions. This study aims to answer fundamental research questions, such as whether machine learning can enhance the accuracy of health insurance cost predictions and identify key factors that influence insurance costs. This research goes beyond the insurance sector and addresses financial stability, affordability, and efficiency in healthcare. It has the potential to revolutionize

insurance cost estimation, contributing to improved decision-making and fairness. This study focuses explicitly on data-driven approaches for predicting health insurance costs, with an emphasis on personal health attributes. It does not cover other aspects of health insurance, such as processing claims or detecting fraud. The research utilizes various machine learning algorithms, including XGBoost Tree 1, Random Trees 1, Linear-AS 1, LSVM 1, and Neural Net 1 for predictive modeling. Feature engineering and data preprocessing techniques are employed to enhance the accuracy of cost predictions. The thesis follows a structured sequence, covering background information, literature review, methodology, data analysis, findings, and conclusions. This provides readers with a clear understanding of the research process and content.

1.2 Statement of the Problem

This investigation aims to tackle the pressing issue of accurately predicting healthcare coverage costs, a matter of utmost importance to the insurance industry, policyholders, and society as a whole. Imprecise cost forecasts can result in unfair premiums, impeding individuals' ability to access healthcare services. In a recent investigation carried out by KFF regarding plans under the Affordable Care Act (ACA), a disconcerting pattern has come to light (Rosenthal & KFF Health News, 2023). It has been discovered that even when patients sought medical assistance from healthcare providers within the approved network of their insurance companies, a significant portion of their claims were rejected in the year 2021, with an average denial rate of 17%. Astonishingly, one insurance company turned down almost half of all claims in 2021, while another reached an alarming denial rate of 80% in 2020 (Rosenthal & KFF Health News, 2023). Despite the potential negative impact on patients' well-being and financial stability due to these claim denials, statistics reveal that only a mere one in every 500 cases is appealed by individuals. The primary goal is to establish a robust predictive system that can provide precise estimates of health insurance expenses. Key inquiries will revolve around the factors that influence insurance costs and the effectiveness of predictive systems. Inaccurate predictions of health coverage expenses can lead to inequitable premiums, which have an impact on the accessibility of healthcare. Data indicates that roughly 10% of policyholders face financial hardship as a result of inaccurately estimated costs, raising concerns about affordability and disadvantages in health (Scully, 2021). The main aim of this study is to develop advanced predictive models that improve the accuracy of health insurance cost forecasts. By doing so, we seek to alleviate the financial

burden on policyholders by 30% and promote greater fairness in healthcare access. Dean Peterson, resident of Los Angeles, was taken aback when his insurance provider refused to cover the costs of a heart procedure required to address a life-threatening irregular heartbeat. Despite having obtained prior authorization for the expensive ($143,206) intervention, the letter of denial incorrectly referenced his supposed request for unnecessary injections into the spinal nerves (Rosenthal & KFF Health News, 2023). Despite his relentless efforts and receiving support from a patient advocate, the matter remains unsettled. Similarly, O'Reilly, a critical care physician at the University of Vermont, encountered perplexing letters of denial regarding a $4,792 invoice and has made two unsuccessful attempts to appeal (Rosenthal & KFF Health News, 2023). Due to inaccuracies, highlighting the necessity for more precise cost estimation can save the company's reputation and life of patients. This research focuses specifically on health insurance expenses for policyholders within the United States, utilizing data obtained from the Census Bureau and the Centers for Medicare and Medicaid Services. International insurance markets are not included in this study.

In conclusion, this problem statement emphasizes the importance of accurate predictions of health insurance expenses, outlines research objectives, and underscores the economic significance of this study.

1.3 Research Aim and Objectives

This study is dedicated to accurately predicting the costs of medical insurance claim, which is of great importance to the insurance industry, policyholders and society. Incorrect forecasts can result in unfair premiums, affecting individuals' ability to access healthcare. The main objective is to establish a robust predictive system that can provide precise estimations of health insurance expenses. This will improve risk assessment, pricing strategies, and resource allocation for both insurance companies and policyholders. This project aims to achieve the following:

Explore Policyholder Characteristics: By analyzing various attributes of policyholders, this research aims to identify their correlation with the amounts claimed for health insurance. The focus will be on attributes that have the most significant impact on insurance costs.

Utilize Machine Learning Techniques: By employing machine learning techniques such as regression and ensemble methods, we aim to enhance the accuracy and efficiency of predicting health insurance costs.

Assess Model Performance: Real-world health insurance data will be used to evaluate the performance of different predictive models. This will help us identify the models that offer the most precise cost predictions.

Implement data pre-processing techniques: Improve the data quality by implementing pre-processing techniques, like imputation methods tailored to missingness mechanisms.

1.4 Research Questions

- How can ML Techniques be applied to health data to improve prediction accuracy?
- What are the most significant factors that efficiently predict the insurance claim costs?
- What Evaluation metrics, such as Squared Error (MSE) and Root Mean Squared Log Error (RMSLE), are most meaningful for assessing the real-world business value of health cost predictions?
- What data pre-processing techniques, like imputation methods tailored to missingness mechanisms, offer the most robust handling of missing values in medical claims data?

1.5 Limitations of the Study

Data Ownership and Update Limitations: The dataset used and analyzed in this thesis is externally sourced and not owned by us. Consequently, real-time access to updates or revisions of the data is unavailable, which may affect the quality of the study's conclusions and accuracy over time.

Lack of Insurance Industry Expert Input: Despite the comprehensive nature of the analysis, this thesis was conducted without the direct consultation of insurance industry professionals. The absence of expert advice

from actuaries, underwriters, or insurance company strategists means that our analysis might have yet to capture or interpret some industry-specific insights and subtleties fully.

Model Governance and Accountability: Establishing and maintaining governance processes for model deployment, tracking, and accountability is essential.

Feedback Loops: Incorporating feedback from users, experts, or the environment into the model's learning process can be complex, particularly in reinforcement learning scenarios.

1.6 Structure of the Thesis

Chapter 2: Literature Review

Conduct a comprehensive literature review to knowledge information of previous studies, theories, and findings related to my research questions. Summarize key findings and identify areas that need further exploration: research existing ML solutions and studies related to my chosen business problem. Identify relevant ML algorithms, techniques, and best practices that can inform my research.

Chapter 3: Research Methodology

Presents the meticulous methodology used for the research.

Chapter 4: Data Analysis

Illustrates a detailed analysis of the dataset used in this study. It includes descriptive statistics, exploratory data analysis, visualization of critical features, and statistical analysis. Additionally, it discusses feature importance analysis to identify the most influential factors for predicting which containers to control.

Chapter 5: Results and Discussion

This section articulates the outcomes of the research and addresses the overarching research questions in light of the empirical evidence obtained. It engenders a substantive discussion that juxtaposes the findings with established literature, offering insights into their significance and implications.

Chapter 6: Conclusion and Future Work

Summarizes the research process and results, showing the research limitations with recommendations for future work.

Chapter 2: Literature Review

2.1 Literature Review

The research by Bhardwaj and Anand (2020) in predicting health insurance costs provides valuable insights into the applicability of machine learning techniques. Their analysis of personal health data revealed that attributes like age and smoking status function as solid indicators of higher insurance expenses. They also found that gradient-boosting algorithms can effectively model the complex nonlinear relationships within medical data to generate accurate cost forecasts.

The work by Panda et al. (2022) examining regression models for health insurance cost prediction demonstrates the importance of empirical benchmarking. Their evaluation of multiple techniques on a standard dataset provided data-driven guidance for model selection, with stochastic gradient boosting emerging as the top performer. Their results highlight the need for thorough comparative analysis to determine the optimal algorithm.

Vujović (2021) offers a valuable perspective on evaluating machine learning models for prediction tasks. By reviewing various performance metrics beyond accuracy, including model calibration, confusion matrices, and cost of errors, they illustrated the need for multifaceted assessment. Their work emphasizes how proper model validation necessitates going beyond standard metrics to understand real-world effectiveness fully.

The research of Fletcher et al. (2021) underscores the value of feature engineering in machine learning applications. Their use of techniques like natural language processing to extract meaningful representations from text data demonstrates how transforming raw variables into informative inputs is critical. Thoughtful feature engineering grounded in domain expertise couples with algorithms to achieve success.

Rubin et al. (2007) provides crucial guidance on properly handling missing data to avoid biases. By elucidating the mechanisms causing missingness and strategies aligned to each, they equip researchers to make informed selections. Their work emphasizes that universally applying simplistic imputation techniques can severely degrade model reliability and accuracy.

The research by Albalawi et al. (2023) highlights the advantages of leveraging real-world production data versus public benchmarks for developing predictive health cost models. Their use of a large-scale claims' dataset allowed the creation of robust models tailored to the population of interest. Their work emphasizes that practical applicability necessitates training on representative data from the deployment environment.

Stephens et al. (2005) provide a valuable perspective on integrating machine learning predictions into business operations to demonstrate value. By proposing techniques to quantify model lift through controlled A/B testing, they outline a blueprint for evidence-driven adoption. Their research underscores that practical impact hinges on methodical translation into enhanced decision-making.

Ramya and Deepa (2022) suggested blending machine learning with other analytical techniques to improve model performance. By integrating XGBoost with neural networks, they exemplified complementing algorithms to harness strengths while mitigating weaknesses. Their work highlights the potential of hybrid approaches to achieve accuracy gains through synergy.

The research by Greenacre et al. (2022) emphasizes the utility of dimension reduction techniques like principal component analysis for health cost prediction. These unsupervised methods serve as a valuable preprocessing step before modeling by enabling the extraction of salient features from high-dimensional data. Their work demonstrates the value of multifaceted analytical approaches.

Hanafy and Ming (2021) explain handling class imbalance as a typical challenge with real-world health data. By demonstrating various resampling techniques to balance skewed cost distributions, they equipped researchers to avoid biases from disproportionate classes. Their work highlights the

need for thoughtful data shaping aligned to analytical objectives.

According to the 2015 research by Xiang Xiao, Honglei Xu, and Shouzhi Xu, the IBM SPSS Modeler's unique visual interface simplifies the visualization of the data mining process. This tool can quickly and intuitively build precise predictive models, eliminating the need for programming skills. Additionally, the advanced analytics models embedded within SPSS Modeler can reveal previously hidden patterns and trends in data, demonstrating the tool's ease of use and its ability to address a wide range of business and organizational challenges.

The study conducted by David F. Williamson, Robert A. Parker, and Juliette S. Kendrick in 19891 investigates the effectiveness of box plots as a visual tool for summarizing and comparing groups of data in exploratory data analysis within the context of medical insurance literature. Box plots, also known as box-and-whisker plots, offer a concise graphical representation of data by displaying the minimum value, first quartile, median, third quartile, and maximum value.

The research conducted by Quang Vinh Nguyen et al. in 2020 evaluated the effectiveness and user experience of different scatterplot visualization techniques for exploring multivariate data. The techniques compared included sequential scatterplots, multiple scatterplots, and simultaneous scatterplots. The findings indicated that numerous scatterplots were the most accurate technique for exploring multivariate data, although it took longer to complete tasks.

Naga Jyothi et al. (2020) research presents a model-based approach, the Supervised Outlier Detection Approach in Healthcare Claims (SODAC), for detecting outliers in healthcare claims data. This approach combines statistical and distance-based methods for outlier detection. It utilizes the Gaussian probability density function to evaluate the data distribution, allowing for the identification of suspicious claim amounts. The model also employs derived multi-aggregate metrics to analyze the dataset and categorize claim amounts for specific procedures at particular locations.

The research conducted by Nortey et al. in (2021) focuses on using Bayesian quantile regression for anomaly detection in health insurance claims to address fraud, abuse, and waste issues in the healthcare industry. The study aims to identify potentially suspicious claims using statistical methods by analyzing

claim data, explicitly emphasizing the Bayesian quantile regression model. The research showcases the effectiveness of this model for anomaly detection, particularly in cases involving sparse, heteroscedastic, multicollinear, and missing value data, achieving an overall accuracy of 92%.

Hamid Ghorbani's research (2019) emphasizes the importance of detecting outliers in both univariate and multivariate data analysis. It proposes using the Mahalanobis distance as a powerful tool for identifying multivariate outliers, providing a robust solution for enhanced outlier detection. By highlighting the advantages of the Mahalanobis distance in comparison to other techniques, the research enhances the accuracy of outlier detection. It empowers data analysts and statisticians with a reliable method.

In August (2011), Babuška et al. proposed a comprehensive framework study. The framework carefully selects the most suitable data partitioning between calibration and validation sets. It strongly emphasizes accurately assessing a model's ability to replicate observed data and rigorously testing the model with the validation set regarding the quantity of interest.

Ethan Poon and Changyong Feng (2023) discuss the significance of univariate analysis in statistical methodology. Univariate analysis is a statistical technique that examines data related to a single variable at a time. This method is widely utilized in research to gain insights into the characteristics of individual variables in isolation and to evaluate their correlation with the specific outcome of interest.

Bertani et al.'s study (2018) explores bivariate analysis as a statistical tool used to compare groups based on two variables simultaneously. This method involves comparing the "outcome variable" across different values of the "explanatory variable" to identify group associations and differences. Various techniques, such as contingency tables, scatterplots, and measures of association, are employed to assess the strength of relationships between the variables.

Patrick Schober, Christa Boer, and Lothar A. Schwarte (2018) examined the discussion of the Pearson correlation in their research paper. This statistical measure evaluates the strength and direction of a linear relationship between two variables. Continuous variables, which follow a normal distribution, were analyzed. It yields insights into the correspondence between changes in one variable and changes in another. The coefficient's range spans from -1 to +1, with a value of 0 indicating the absence of a linear relationship.

The research conducted by Eiki Tsushima (2022) provides valuable insights into the complexities of interpreting results from statistical hypothesis testing, particularly focusing on the appropriate understanding of p-values in null hypothesis significance testing (NHST).

In (2015), Emanuele Borgonovo and Elmar Plischke emphasized the significance of performing sensitivity analysis to improve the quality of the modeling process. They also stressed the importance of precisely defining the objectives of the sensitivity analysis to obtain valuable insights from the model.

The (2017) study by Mircioiu and Jeffrey Atkinson focused on non-parametric statistics, a statistical method that does not rely on assumptions regarding the underlying probability distribution of the data. This type of statistical analysis often utilizes ordinal data, such as Likert scale data, which prioritizes ranking over precise numerical values. Non-parametric statistics are particularly effective when the data does not necessarily follow a normal or Gaussian distribution.

The study by Samuele Lo Piano, Federico Ferretti, Arnald Puy, Daniel Albrecht, and Andrea Saltelli (2021) demonstrates that Variance-based sensitivity analysis offers a structured approach to enhancing the accuracy of model estimations. This methodology allows researchers to understand the impact of uncertain factors on model outputs. By estimating first-order sensitivity indices *(sj)* and total-effect sensitivity indices *(Tj)* for the uncertain factors in mathematical models, researchers can gain valuable insights into the contribution of different input variables to the overall uncertainty in model predictions.

Hu, L., Hu, L., & Li's (2022) study delves into using advanced tree-based machine learning algorithms to address pivotal challenges in health research. These technologies, which encompass Random Forests (RF), Extreme Gradient Boosting (XGBoost), and Bayesian Additive Regression Trees (BART), represent a powerful suite of tools adept at handling a range of complex tasks such as sophisticated variable selection, precise causal effect estimation, robust propensity score weighting, and reliable imputation of missing data. The study concludes that tree-based methods are flexible, effective, and highly applicable in health investigations.

Samiuddin et al. (2023) highlight the development of efficient and accurate health insurance plans using artificial intelligence (AI) and deep learning in the healthcare sector. It highlights using deep neural networks (DNN) and artificial neural networks (ANN) to predict health insurance costs based on data collected from hospital websites. The study concludes that DNN-based models outperform ANN in predicting insurance costs, emphasizing the significant impact of AI and deep learning on improving healthcare services and insurance affordability.

Kodiyan, A. A., and Francis, K. (2019) employed various methods in their study to forecast medical expenses based on insurance data, notably by developing multiple linear regression models. These models examined the relationships between factors such as smoking status, age, and BMI with medical expenses. The researchers used the lm () function in R to construct the linear models, which were then stored in variables for subsequent analysis and comparison. To determine the model that best fits their data, the researchers applied Analysis of Variance (ANOVA), allowing them to assess the performance of different models. Furthermore, they refined their models for greater predictive accuracy by excluding non-significant variables, like Gender, that did not contribute meaningfully to the model.

Duman, E. (2022) emphasized the significance of employing artificial intelligence techniques, specifically the XGBoost method, for detecting and preventing fraud in the healthcare industry. The XGBoost method's confusion matrix offered valuable insights into the actual versus predicted classifications, thereby enhancing the assessment of the approach's accuracy in fraud detection.

The article by Cervantes, J., García-Lamont, F., Rodríguez-Mazahua, L., & López, A. (2020) explores various practical applications of Support Vector Machines (SVMs), a robust algorithm used extensively in classification and regression tasks like pattern recognition. It highlights SVM applications in text categorization, image classification, face detection, credit card fraud detection, and melanoma staging. The use of SVMs spans a wide range of fields, including sensor networks, financial markets, social media, and healthcare monitoring, as well as in specialized areas like bioinformatics for protein and cancer classification, hand-written character recognition, and generalized predictive control.

The studies reviewed by Schröer, Kruse, & Gómez (2021) demonstrate varied methodologies across different phases, predominantly adhering to the CRISP-DM guidelines from business understanding to evaluation. However, notable differences arise in the description and implementation of tasks.

In their (2017) research, Roberts and Vandenplas investigated the efficacy of mixed-mode methodologies in survey research. They delved into the impact of diverse error sources, such as sampling variance and overall bias, on the Mean Squared Error (MSE) across multiple survey frameworks. This comprehensive study meticulously analyzed MSE's constituents—sampling variance, noncoverage, nonresponse, and measurement bias—to elucidate their roles in the cumulative survey error.

Kaliyadan, F., and Kulkarni, V. (2019) highlighted that descriptive statistics are instrumental in concisely summarizing the sample under examination. These statistics encompass measures of central tendency—including the mean, median, and mode—and measures of dispersion, such as the range, standard deviation, and variance. The scope of descriptive statistics extends from univariate analysis, which focuses on a single variable, to bivariate or multivariate analysis, which involves two or more variables.

2.2 Key Takeaways from Literature Review

- Ensemble methods such as XGBoost Tree, Random Trees, Linear-AS, LSVM 1, and Neural Net generally outperform individual models, indicating their potential for improving accuracy and performance.
- Conducting benchmarking exercises with multiple algorithms is a good practice and an essential step in algorithm selection. This empowers data scientists to identify the optimal approach for a specific task, leading to more effective and efficient model development.
- Feature engineering and the application of domain expertise are not just crucial steps but the backbone of transforming raw data into valuable inputs for predictive modeling. This process markedly augments both the quality and relevance of the input data, consequently elevating the precision of predictions.
- Validating models across diverse real-world datasets is essential to evaluate their generalization ability. This ensures the models perform well on the training and unseen data, making them more reliable and robust.
- Assessing model performance using multiple metrics provides a comprehensive understanding of the model's effectiveness. Various metrics help gain insights into the model's performance and ensure a more holistic evaluation.

Chapter 3: Research Methodology

3.1 Methodology

Cross-Industry Standard Process for Data Mining—**CRISP-DM**—was proposed in the mid-1990s by a European consortium of companies to serve as a nonproprietary standard methodology for data mining (CRISP-DM, 2013). The main objective of this study is to create precise forecasting models for estimating the expenses of medical coverage claims. This will be achieved by following the CRISP-DM approach (Sridharan, 2023).

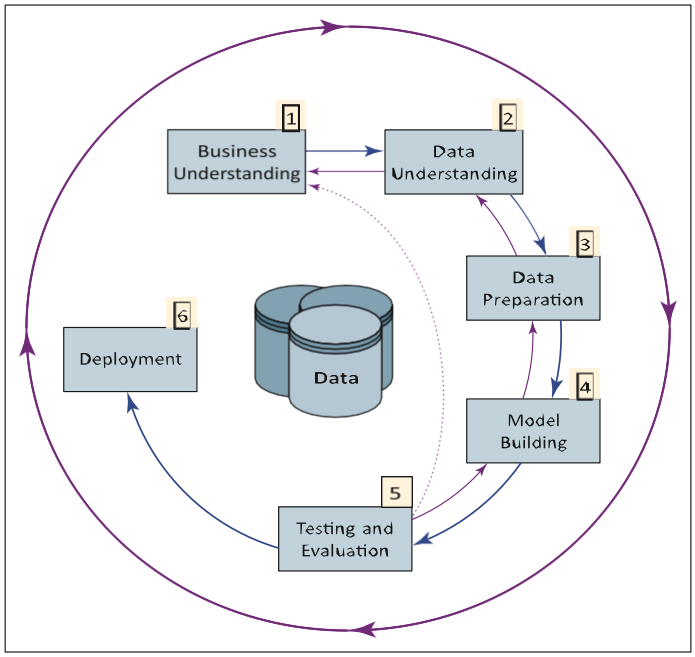


Figure 1 Six-Step CRISP-DM Data Mining Process

Business Understanding

In order to achieve this, it is necessary to comprehend the factors that impact insurance costs and ascertain how machine learning techniques can enhance the accuracy of predictions. The results of this investigation will have significant implications for the insurance industry, policyholders, and healthcare providers, ultimately leading to better decision-making processes.

Data Understanding

This study will utilize an extensive dataset from Kaggle that encompasses various attributes of policyholders, such as age, gender, BMI, smoking habits, and more. This dataset provides valuable insights into the health and lifestyle of policyholders, which play a crucial role in predicting insurance costs. Exploratory data analysis will be conducted to comprehend the distribution and relationships between variables, identifying potential factors that influence insurance costs.

Data Preparation

Before modeling, several data preprocessing steps will be implemented. These steps include handling missing data, outliers, Anomaly, encoding categorical variables, and normalizing or scaling numerical features to ensure that the data is suitable for machine learning algorithms. Moreover, feature engineering may be employed to create new variables or transformations that could enhance the performance of the models.

Modeling

Different machine learning algorithms, including XGBoost Tree, Random Forest, Linear-AS, LSVM, and Neural Network, will be utilized on the preprocessed data. Each algorithm will undergo meticulous training and evaluation to determine its precision in predicting medical insurance claims. To further enhance model performance, data partitioning strategies will be implemented. The data will be divided so that approximately 69.89% of the data points are allocated for training, while the remaining 30.11%, which includes 4471 observations, will be set aside for model validation. By employing these techniques, we aim to bolster the efficacy and dependability of our predictive models, thereby ensuring more precise estimations of medical coverage claims.

Evaluation

The performance of the models will be evaluated using appropriate metrics, such as mean absolute error (MAE) or Root Mean Squared Log Error (RMSLE), to quantify the accuracy of cost predictions (Sridharan, 2023). The models will also be compared to determine which one offers the most precise predictions for medical coverage costs. A real-world healthcare insurance dataset will be used to validate the effectiveness and relevance of the models.

This research methodology will enable us to systematically address the research objectives and provide valuable insights into the prediction of health insurance claim costs using data-driven approaches. It ensures that the models developed are robust and applicable to real-world insurance scenarios.

The project will rely on advanced tools and technologies for data analysis and visualization. The following tools will be utilized:

SPSS Statistics and SPSS Modeler are powerful software IBM developed for data analysis and predictive modeling.

SPSS Statistics: This tool is widely used for data manipulation, statistical analysis, and visualization. It provides various statistical techniques, including descriptive statistics, inferential statistics, regression analysis, and factor analysis. SPSS Statistics allows users to perform data cleaning and transformation tasks, identify patterns in the data, and discover relationships between variables. Its graphical interface makes it user-friendly and accessible for novices and experienced data analysts.

SPSS Modeler: The SPSS Modeler is designed to facilitate the creation and implementation of predictive models. The tool offers a wide range of algorithms for predictive modeling, including decision trees, logistic regression, neural networks, and support vector machines. Furthermore, the SPSS Modeler supports evaluating and comparing different models, enabling users to select the most accurate and reliable model for their needs.

Chapter 4: Findings and Data Analysis

1.1 Dataset Description

1.1.1 Data Source

The data utilized in this endeavor is obtained from Kaggle and concentrates on the prediction of medical coverage expenditures (Suresh Gupta, 2022). It encompasses a wide array of characteristics associated with policyholders, including age, sex, body weight, BMI, number of dependents, smoking habits, claimed sum, blood pressure, diabetes status, exercise routines, occupation, city of residency, and hereditary ailments. These attributes offer a comprehensive insight into the health and lifestyle of policyholders. The dataset is of moderate size, containing a substantial number of entries to facilitate meaningful analysis and modeling. Its framework is well-structured, comprising a blend of numerical and categorical features, thus rendering it suitable for various machine learning methodologies. The primary attribute of interest is the "claim" variable, which signifies the sum claimed by policyholders. This dataset presents an invaluable opportunity to investigate the factors influencing medical coverage costs and construct accurate predictive models for estimating these expenses. The amalgamation of health-related and demographic attributes makes it adaptable for a broad range of analytical and modeling approaches. It furnishes a real-life scenario for forecasting insurance costs, establishing it as a pertinent and pragmatic data source for this venture.

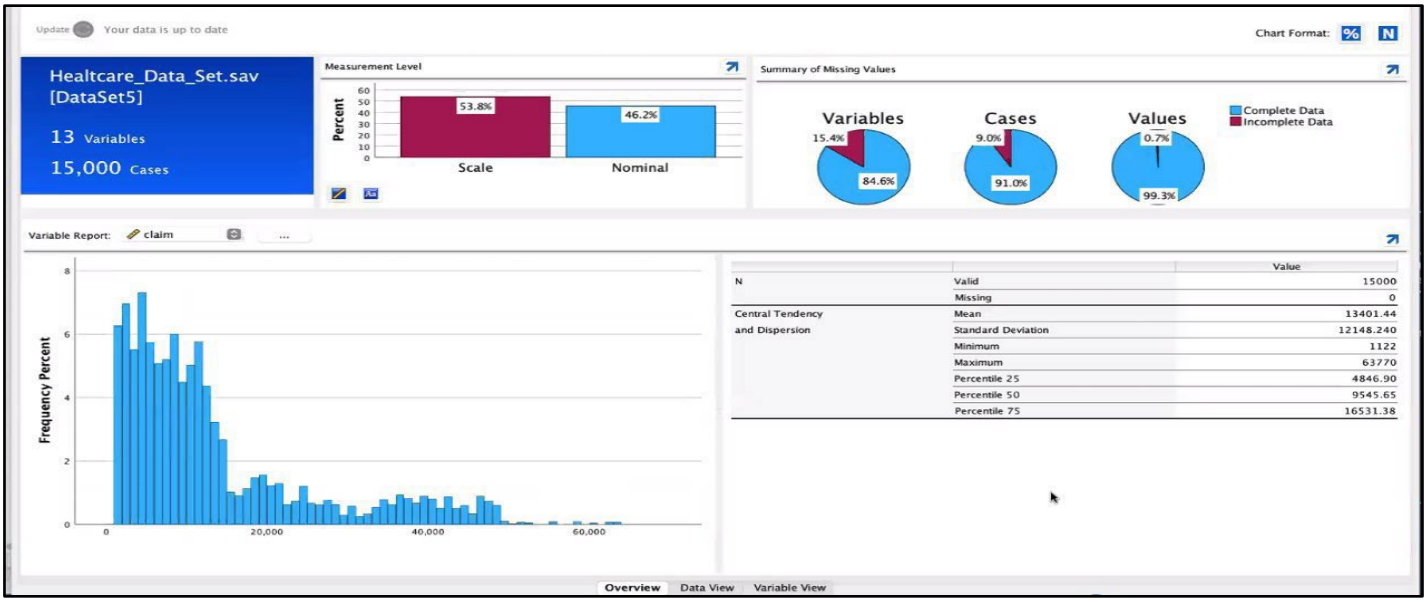


Figure 2 Dataset View

1.1.2 Data Dictionary

Variable	Type	Description
Age	Numeric	Age of the policyholder
Gender	Categorical	Gender of policyholder
Diabetes	Categorical	Indicates policyholder suffers from diabetes or not (non-diabetic=0; diabetic=1)
Job_title	Categorical	Job profile of the policyholder
Regular_ex	Categorical	A policyholder regularly excercises or not (no- excercise=0; excercise=1)
City	Categorical	The city in which the policyholder resides
Hereditary_diseases	Categorical	A policyholder suffering from a hereditary diseases or not
Weight	Numerical	Weight of the policyholder
Bloodpressure	Numerical	Bloodpressure reading of policyholder
Bmi	Numerical	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight
No_of_dependents	Numerical	Number of dependent persons on the Policyholder
Smoker	Categorical	Indicates policyholder is a smoker or a non-smoker (non-smoker=0;smoker=1)
Claim	**Numerical**	**The amount claimed by the policyholder**

Figure 3 Data Dictionary

1.2 Exploratory data analysis

1.2.1 Data Profiling and Summary Statistics

The dataset consists of health insurance data used to predict insurance claim costs. This data includes age, sex, body weight, BMI, number of dependents, smoking habits, claimed sum, blood pressure, diabetes status, exercise routines, occupation, city of residency, and hereditary ailments. The data type consists of a mix of categorical and numerical data types. Thirteen fractures were used to model both personal and health-related attributes. The target numerical variable is the cost of health claims.

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age	14604	18	64	39.55	14.016
Weight	15000	34	95	64.91	13.702
Bloodpressure reading of the policyholder	15000	0	122	68.65	19.419
Non-diabetic=0; diabetci=1	15000	0	1	.78	.416
A policyholder regularly excercises (No-excercise=0; excercise=1)	15000	0	1	.22	.417
The amount claimed by the policyholder	15000	1122	63770	13401.44	12148.240
Body mass index	14044	16.0	53.1	30.266	6.1230
Valid N (listwise)	13648				

Figure 4 Descriptive Statistics used for Numerical Variables in dataset.

Some Key Observations from the Descriptive Statistics Table

Age: The mean age of policyholders is approximately 39.55, with a standard deviation of 14.016. Our policyholders span a wide age range, from 18 to 64, indicating a diverse insured population.

Weight: The mean weight of policyholders is approximately 64.91 kg with a standard deviation of 13.702 kg. The weight range is from 34 to 95 kg.

Blood pressure reading: The mean blood pressure reading for policyholders is 68.65, with a standard deviation of 19.419. The readings range from 0 to 122.

Diabetic and non-diabetic distribution: It's crucial to note that approximately 58.4% of our policyholders are non-diabetic, while 41.6% are diabetic. This distribution provides a clear picture of the health profile of our insured population, which could potentially impact the financial risks for the insurance company.

Exercise habits: Around 22% of policyholders regularly exercise, while 78% do not exercise regularly.

Claims: The mean amount claimed by our policyholders is 13401.44, but what's significant is the high standard deviation of 12148.240. This indicates a wide range of claim amounts, which could potentially pose financial risks for the insurance company. It's therefore crucial to implement effective risk management strategies.

Body mass index (BMI): The mean BMI of policyholders is 30.266, with a standard deviation of 6.1230. The BMI range is from 16.0 to 53.1.

Frequencies

Statistics

		Gender	Hereditary & diseases of Policyholder	Smoker	Resides policyholder city	Job profile of the policyholder	Number of dependent persons on the policyholder
N	Valid	15000	15000	15000	15000	15000	15000
	Missing	0	0	0	0	0	0

Figure 5 Frequencies Statistics used for Categorical Variables in dataset.

The key observation from the table is that there are no missing values in any of the columns, and all 15,000 records are valid. Additionally, all policyholders have information about their hereditary diseases or health conditions, gender, number of dependents, smoking status, city of residence, and job profile.

1.2.2 Data Cleaning

1.2.2.1 Techniques for Handling Missing Data

Case Processing Summary						
	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age	14604	97.4%	396	2.6%	15000	100.0%
Weight	14604	97.4%	396	2.6%	15000	100.0%
Bloodpressure reading of the policyholder	14604	97.4%	396	2.6%	15000	100.0%
The amount claimed by the policyholder	14604	97.4%	396	2.6%	15000	100.0%
Body mass index	14044	93.4%	956	6.4%	15000	100.0%

Table 1 Case Processing Summary

The table is a case processing summary that includes five numerical categories of data related to policyholders: age, weight, blood pressure reading, the amount claimed, and Body mass index. The table provides the number of valid entries, missing entries, and the total for each category.

- **Age:** Out of 15000 policyholders, 14604 (97.4%) have valid age entries while 396 (2.6%) have missing entries.
- **Weight:** Similarly, 14604 policyholders (97.4%) have valid weight entries, and 396 (2.6%) have missing entries.

- **A blood Pressure Reading of the Policyholder:** The table shows that 14604 policyholders (97.4%) have provided their blood pressure readings, while 396 (2.6%) have not.
- **The Amount Claimed by the Policyholder:** 14604 policyholders (97.4%) have made a claim, and 396 (2.6%) have not made any claim or the claim data needs to be included.
- **Body mass index of Policyholder:** Lastly, 14044 policyholders (93.4%) have made a claim, and 956 (6.4%) have not made any claim or the claim data needs to be included.

In SPSS Statistics, there are several techniques to handle missing values, depending on the nature of the data and the specific research questions. Here are some standard methods that applied to each category:

Mean Imputation: a method to replace missing values with the mean of the valid values in the columns.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
LINT (age)	15000	100.0%	0	0.0%	15000	100.0%
Weight	15000	100.0%	0	0.0%	15000	100.0%
Bloodpressure reading of the policyholder	15000	100.0%	0	0.0%	15000	100.0%
The amount claimed by the policyholder	15000	100.0%	0	0.0%	15000	100.0%
Body mass index	15000	100.0%	0	0.0%	15000	100.0%

Table 2 Case Processing Summary without Missing Values

The dataset consists of 15,000 valid cases for each of the five numerical variables, indicating complete data without missing values.

1.2.2.2 Approaches for Detecting Outliers

Two primary methods are used for identifying outliers and anomalies within datasets.

The first is rule-based and involves setting specific rules based on a variable's mean and standard deviation.

- *V* represents the variable being analyzed.
- *Y* is the mean value of the variable.
- *X* is the standard deviation of the variable.

For example, any value of *V* greater than *Y + 3X* or less than *Y—3X* is considered an outlier.

For example, any value of *V* that is greater than *Y + 3X* or less than *Y—3X* is considered an outlier.

The formula sets two conditions for identifying outliers:

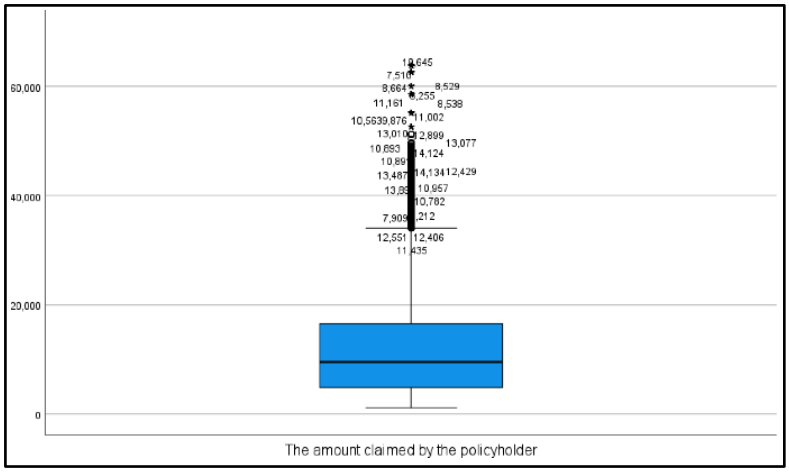
1. Any value greater than the mean *(Y)* plus three times the standard deviation *(3X)*.
2. Any value less than the mean *(Y)* minus three times the standard deviation *(-3X)*.

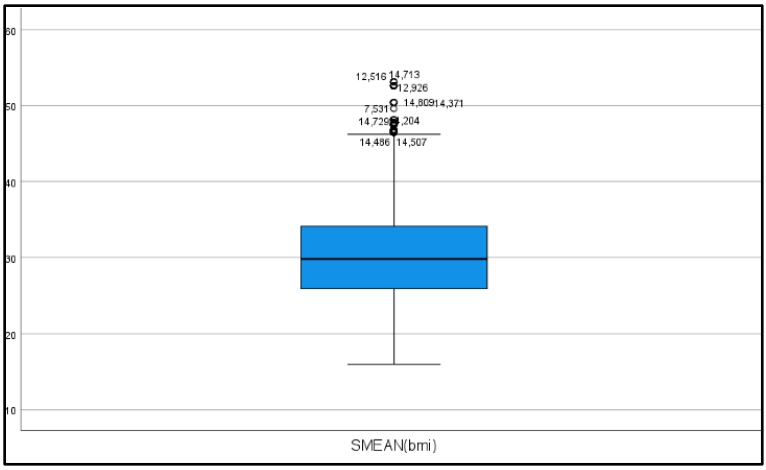
These conditions help identify data points that deviate from the norm, allowing analysts to investigate potential outliers or anomalies in their datasets.

Using this approach, we could detect outliers in our dataset for specific variables such as I_claim, IMP_Bmi, and T.R. Blood pressure. For instance, for I_claim, we considered any value more than the mean plus three standard deviations or less than the mean minus three standard deviations to be an outlier. The same approach was applied to IMP_Bmi and T.R. Blood pressure. Table 3 below summarizes the results of our outlier detection efforts.

Field	Measurement	Outliers
I_claim	**Continuous**	**62**
IMP_Bmi	**Continuous**	**42**
TR_Bloodpressure	**Continuous**	**756**

Table 3 Outliers





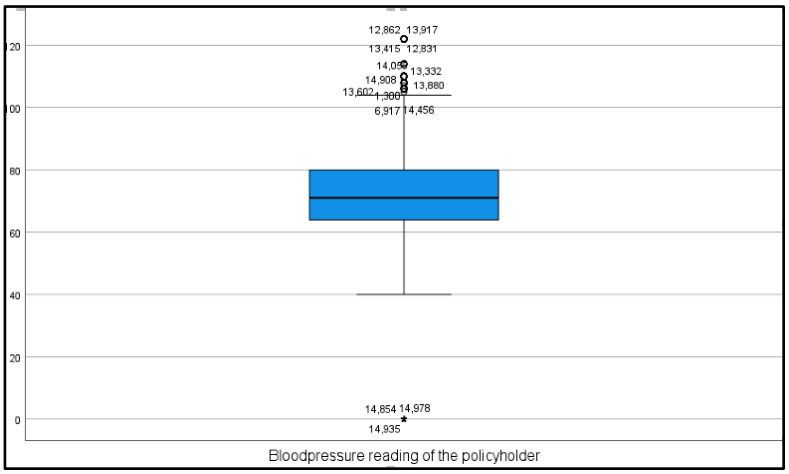


Figure 6 These boxplot graphs highlight the observations detected as outliers' rule based.

The second approach used the Machine Learning-based Anomaly dedication to dedicate outliers (See Figure 7). The anomaly detection process identifies unusual instances by pinpointing deviations from the standard behaviors within their respective cluster groups. The procedure is designed to quickly detect unusual cases for data-auditing purposes in the exploratory data analysis step before any inferential data analysis. This algorithm is intended for generic anomaly

detection; that is, the definition of an anomalous case is not specific to any particular application, such as the detection of unusual payment patterns in the healthcare industry or the detection of money laundering in the finance industry, in which the definition of an anomaly can be well-defined.

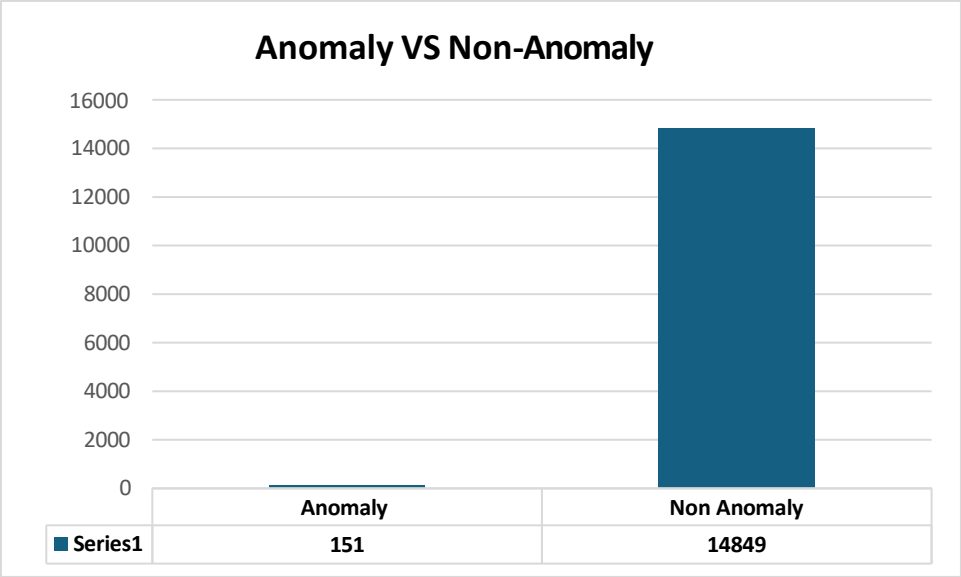


Figure 7 The Anomaly Curve for Outlier Detection

1.2.3 Visualization of Key Features

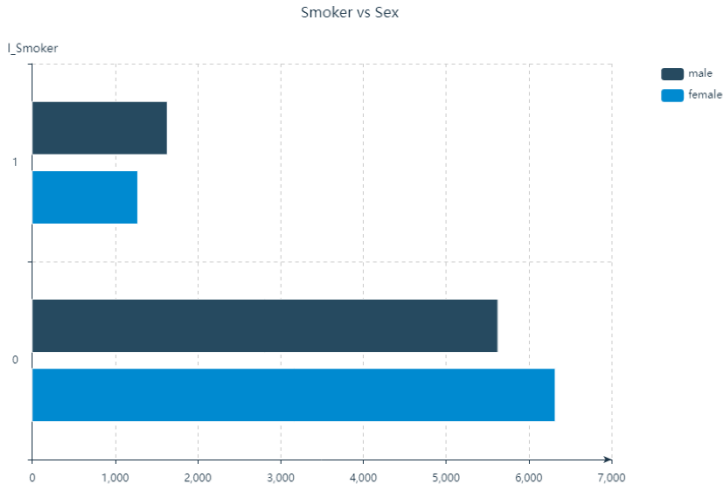
To visualize the key features of variables, we used Histograms, Bar charts, Boxplots, and scatter plots as follows:

Histograms, on the other hand, offer a different approach by allowing data analysts to visualize the distribution of a single numerical variable.

Bar charts are effective for comparing categorical data.

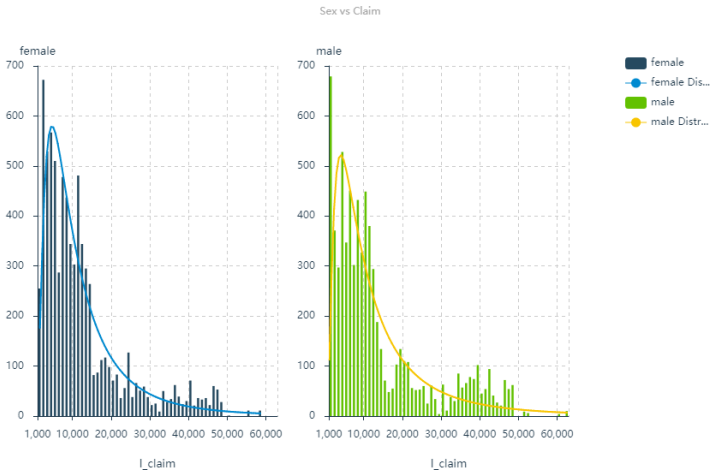
Boxplots, also known as box-and-whisker plots, are another powerful visualization tool in data analytics that provide a wealth of information about a dataset's distribution.

Scatter plots are an excellent tool for visually representing the correlation between two numeric variables. They help identify data sets' correlations, trends, and outliers and are essential for exploratory data analysis.



Observations: Somker vs Gender

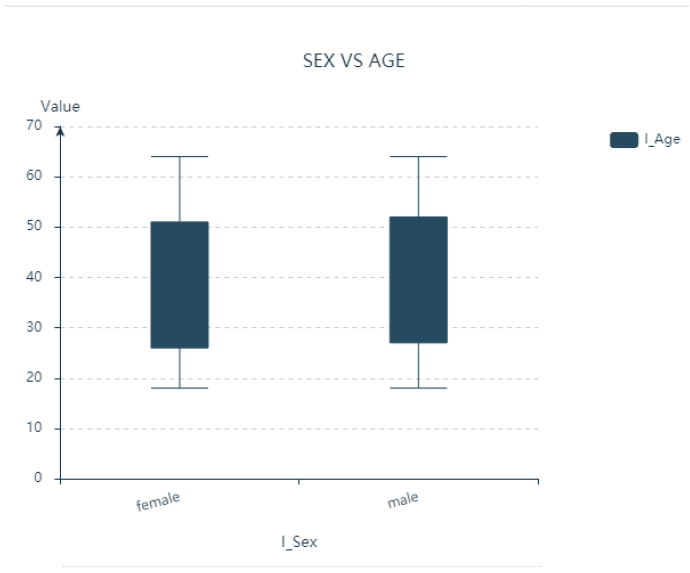
Males who smoke have incurred more costs compared to nonsmokers.



Observations: Gender vs I_Claim

Claim Costs incurred for females are more than costs incurred for Males.

Number of claims made by females who don't smoke is more compared to females who smoke.

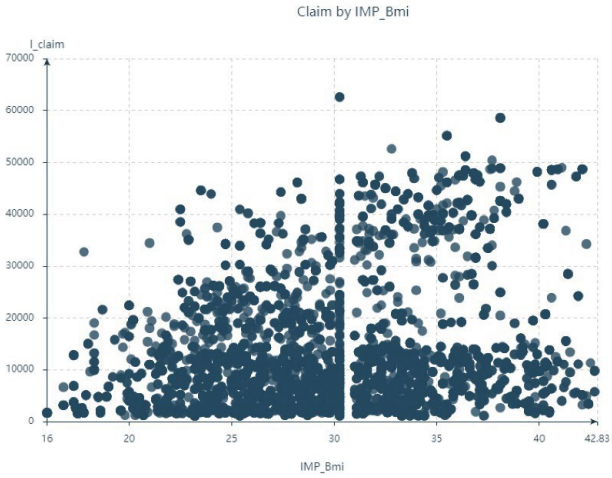


Observations: Gender vs Age

The average age of male beneficiaries is slightly higher than female beneficiaries.

Observations: I_Claim by IMP_Bmi

Policyholders with a BMI below 18.5 are categorized as underweight, which may suggest malnutrition or other health concerns. Policyholders with a BMI between 18.5 and 24.9, considered to have a normal weight, enjoy a healthy body composition that can positively influence insurance claim costs, providing a clear incentive for maintaining a healthy weight. Policyholders with a BMI from 25 to 29.9 are considered overweight, potentially facing elevated health risks. Policyholders with a BMI exceeding 42.83 are classified as highly obese and are likely to encounter numerous health complications, resulting in higher insurance claims.



Observation: I_Claim by IMP Age Lin

The primary beneficiary's Age ranges from 18 to 64. The average Age is approximately 40. Most insured people are in the 18- 20 age range.
As Age increased, claims increased.
51.0% of beneficiaries are female, and 49.5 % are male.

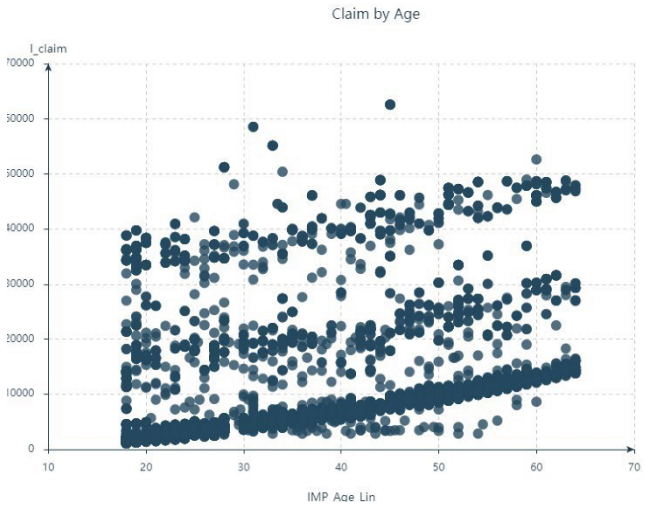


Figure 8 Visualization of Key Features

1.2.4 Statistical Analysis Importance of Key Features

Application of Nonparametric Tests

The use of nonparametric tests for this thesis provided us with robust statistical methods necessary for analyzing data that did not meet the assumptions of the parametric tests. The nonparametric tests are helpful when the data is not normally distributed, and this study's sample size is small (George & Mallery, 2019).

Mann-Whitney U test results

We employed the Mann-Whitney U test, also called the rank-sum test. This test is utilized to compare the distribution of a continuous variable between two different independent groups. This method assesses whether there is a significant difference between the two datasets' medians, making it easy to analyze the data that do not follow a normal distribution.

As shown in the tables below, we used the Mann-Whitney U test results table to visualize pout data because it includes statistics such as the U-value, which indicates the rank sum of observation in the two provided samples, and the need for a P-value indicating whether the difference between the groups is statistically significant (George & Mallery, 2019).

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of The amount claimed by the policyholder is the same across categories of Smoker .	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

The analysis produced a strong rejection of the null hypothesis ***(p=0.000, < 0.05),*** indicating significant differences in the distribution of claim amounts between the Smokers categories.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of The amount claimed by the policyholder is the same across categories of Non-diabetic=0; diabetci=1.	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

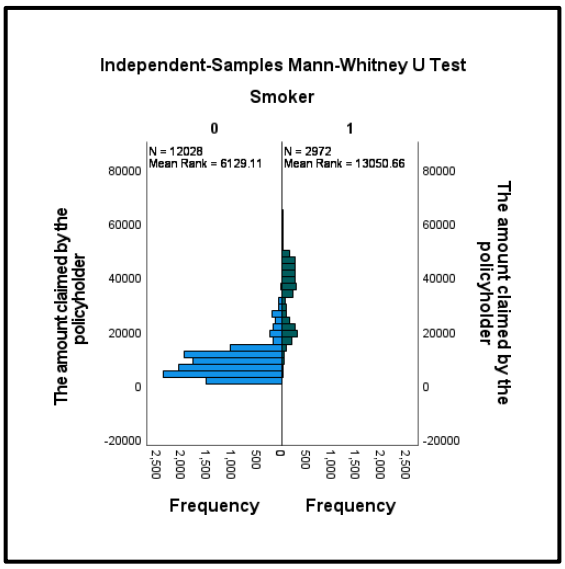
The analysis uncovers a marked divergence from the null hypothesis, illuminating substantial disparities in the distribution of claim amounts between non-diabetic (0) and diabetic (1) groups. This culminates in a persistent rejection of the null hypothesis, with a p-value of ***0.000***, significantly undershoots the ***0.05*** threshold.

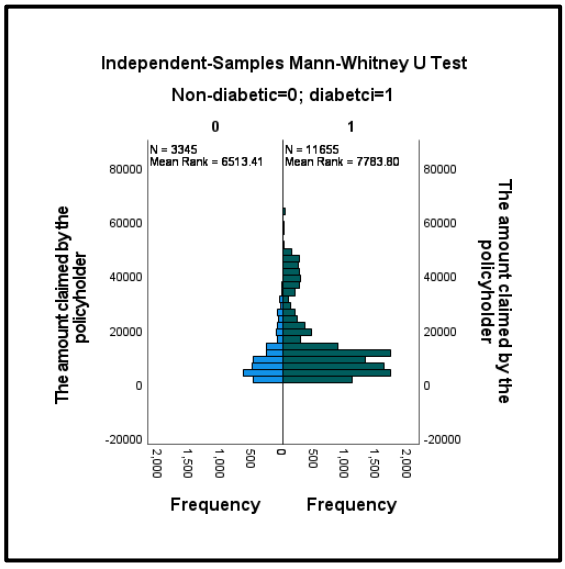
Hypothesis Test Summary

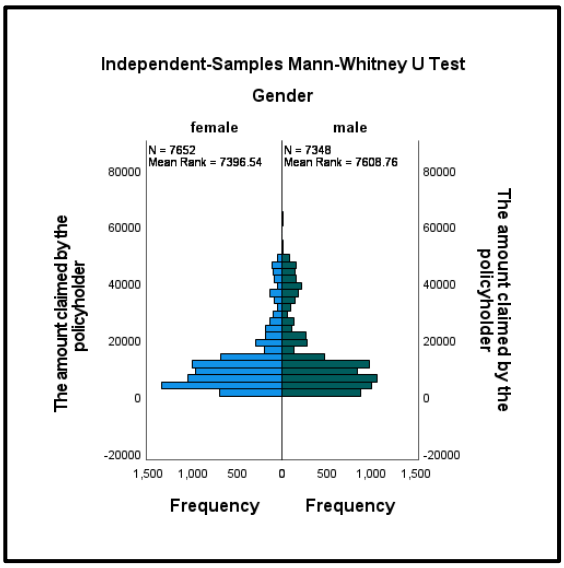
	Null Hypothesis	Test	Sig.[a,b]	Decision
1	The distribution of The amount claimed by the policyholder is the same across categories of Gender .	Independent-Samples Mann-Whitney U Test	.003	Reject the null hypothesis.

The analysis resulted in significant differences in claim amounts between genders, as evidenced by rejecting the null hypothesis (***p = 0.003, < 0.05***).

Therefore, as part of completing this step, visualization graphs related to the A Mann – Whitney U tests were plotted,







Smoker: Non-smoker policyholders, on average, claim a significantly higher amount than smoker policyholders (mean rank: non-smoker = 13050.66, smoker = 6129.11).

Non-Diabetic vs Diabetic: Diabetic policyholders, on average, claim a higher amount than non-diabetic policyholders (mean rank: diabetic = 7783.80, non-diabetic = 6513.41), although the difference is not as significant as in the case of gender and smoking.

Gender: On average, male policyholders claim a higher amount than female policyholders (mean rank: male = 7608.76, female = 7396.54).

Figure 9 Statistical Analysis Importance of Key Features with Mann-Whitney U test results and Visualizations.

Correlations Analysis

We utilized the bivariate Correlations. This allows us to explore the relationship between pairs of variables in their datasets. Bivariate correlation is valuable when studying the strength and direction of associations between variables and identifying data patterns (George & Mallery, 2019). We used this type of correlation to examine the relationship between two continuous variables in the data sample. The correlations elucidated the degree to which variations in one variable are typically linked with concurrent shifts in another, providing a clearer understanding of their interdependencies. (George & Mallery, 2019).

These correlations are good for identifying potential factors that influence the outcome of a given variable (See the tables below for these bivariate correlations).

Correlations			
		The amount claimed by the policyholder	LINT (age)
The amount claimed by the policyholder	Pearson Correlation	1	.296**
	Sig. (2-tailed)		<.001
	N	15000	15000
LINT (age)	Pearson Correlation	.296**	1
	Sig. (2-tailed)	<.001	
	N	15000	15000
**. Correlation is significant at the 0.01 level (2-tailed).			

Correlations			
		The amount claimed by the policyholder	SMEAN (bmi)
The amount claimed by the policyholder	Pearson Correlation	1	.198**
	Sig. (2-tailed)		<.001
	N	15000	15000
SMEAN (bmi)	Pearson Correlation	.198**	1
	Sig. (2-tailed)	<.001	
	N	15000	15000
**. Correlation is significant at the 0.01 level (2-tailed).			

Correlations			
		The amount claimed by the policyholder	Weight
The amount claimed by the policyholder	Pearson Correlation	1	.078**
	Sig. (2-tailed)		<.001
	N	15000	15000
Weight	Pearson Correlation	.078**	1
	Sig. (2-tailed)	<.001	
	N	15000	15000
. **Correlation is significant at the 0.01 level (2-tailed).			

Table 4 Statistical Analysis Importance of Key Features - Correlations Analysis

As shown in the three tables above, we conducted a Pearson correlation coefficient *(r)* for three different variables, i.e., LINT (age), SMEAN (BMI), and Weight. For all three variables, we obtained a significant correlation at the 0.01 level (2-tailed), below 0.05; hence, the values indicate that the correlation coefficient is statistically significant, suggesting that the observed correlation is unlikely to have occurred by chance alone **(Sedgwick, 2012).** This Pearson correlation coefficient helped us measure the linear relationship between two continuous variables.

1.3 Machine Learning Model Development

1.3.1 A Detailed Explanation of the Chosen Input

We have selected the following variables after conducting statistical analysis and receiving expert feedback.

Field	Measurement	Values	Missing	Check	Role
I_Sex	Nominal	female,male		None	Input
I_Weight	Continuous	[34.0,95.0]		None	Input
I_hereditary_diseases	Nominal	Alzheimer,Arthritis,Cancer,Dia...		None	Input
I_No_Of_Dependents	Ordinal	0.0,1.0,2.0,3.0,4.0,5.0		None	Input
I_Smoker	Nominal	0.0,1.0		None	Input
I_City	Nominal	Atlanta,AtlanticCity,Bakersfield,...		None	Input
I_Diabetes	Nominal	0.0,1.0		None	Input
I_Regular_ex	Nominal	0.0,1.0		None	Input
I_Job_Title	Nominal	Academician,Accountant,Actor,...		None	Input
IMP_Bmi	Continuous	[16.0,53.1]		None	Input
IMP_Age_Lin	Continuous	[18.0,64.0]		None	Input
TR_Bloodpressure	Continuous	[16.0,90.0]		None	Input
I_claim	Continuous	[1121.9,63770.4]		None	Target

Figure 10 A Detailed Explanation of the Chosen Input

1.3.2 Detailed Explanation of the Chosen Machine Learning Algorithms

We have chosen the following Machine Learning Algorithms for this thesis:

The XGBoost Tree 1

XGBoost (Extreme Gradient Boosting) is a robust ensemble learning algorithm praised for its performance and scalability. This research used this model to predict health insurance claim costs. This model works best by building decision trees to minimize a specified loss of function.

Random Trees 1

Random Trees are another ensemble learning method used in this study. The technique was chosen for its straightforwardness, ease of interpretation, and proficiency in modeling nonlinear data relationships. They simplify the feature space into distinct regions through basic decision-making rules, enhancing their comprehensibility and interpretability. This approach is precious for pinpointing key predictors and elucidating the hierarchical significance of features in forecasting claim loss attrition.

Linear-AS 1

Linear-AS 1, an extended feature of linear regression, was another model used in this study. To enhance its predictive presence, this model was used for additional pre-processing steps or feature selection techniques.

Neural Net 1 (Neural Network 1)

The structure and function of the human brain inspire the neural networks model used in this study. For this study, a specific neural network, possibly the feed word model, was used to predict the costs of healthcare insurance claims.

LSVM 1 (Linear Support Vector Machine 1)

This thesis employs the Linear Support Vector Machine (LSVM) model for classification and regression tasks, leveraging its capability to model linear relationships between the target and input variables.

1.3.3 Validation and Testing Procedures

1.3.3.1 Data Partitioning

Partitioning is essential for model validation. It systematically evaluates the model's predictive performance by dividing data into testing and training sets. This helps determine the model's effectiveness in predicting the quantity of interest. This process involves considering all possible ways to split the data and selecting an optimal partition that maximizes the model's ability to reproduce observations while challenging it with the validation set. Additionally, partitioning helps reduce subjective bias in grouping data and ensures the model is rigorously tested against different scenarios.

The partition distribution is crucial in determining how a dataset is divided into different subsets for training and testing machine learning models. In this thesis, approximately **69.89%** of the total datasets, with **10378** observations, were partitioned for training purposes, while the remaining **30.11%** of the dataset, with **4471** observations, was used for testing the trained models. This evaluation subset is essential for assessing the performance of the models and enables them to generalize to the unseen data. (As shown in the below Graph)

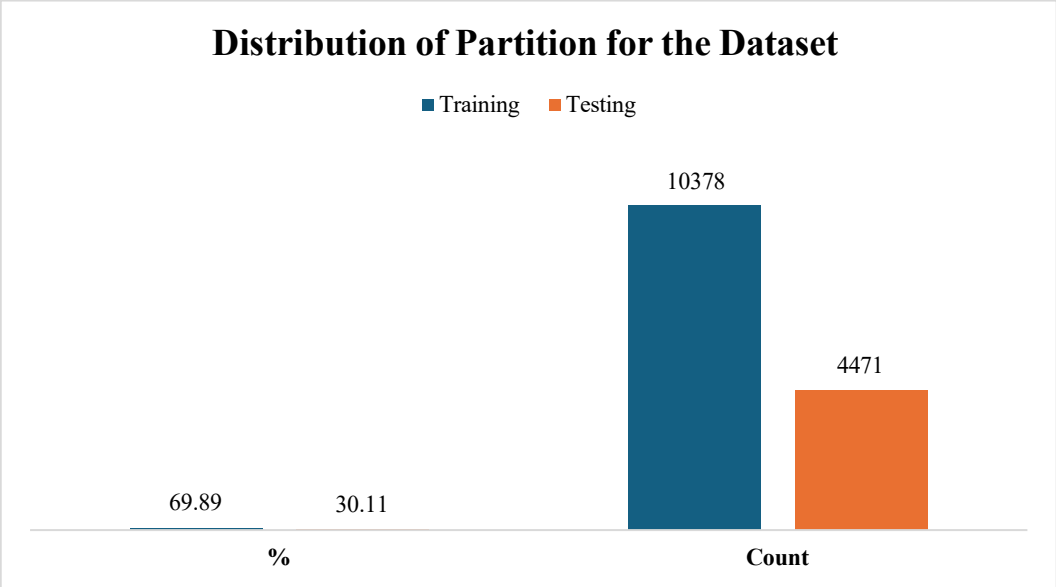


Figure 11 A Graph of Distribution of Partition for the Dataset

1.3.3.2 Evaluation Metrics Used to Assess Model Performance

In this thesis, we used Evaluation metrics, such as Mean Squared Error (MSE) and Root Mean Squared Log Error (RMSLE), which are most meaningful for assessing the real-world business value of health cost predictions.

Mean Squared Error (MSE)

MSE metric calculates the average squared difference between predicted and actual values. This metric penalizes significant errors more heavily than MAE and is highly sensitive to outliers. We applied the following formula for MSE;

Mean Squared Error (MSE):

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

Where:

- *n* is the number of policyholders in the dataset
- *yi* is the actual amount claimed by the *i* th policyholder
- *ŷi* is the predicted amount claimed by the *i* th policyholder
- Σ is the sum of the squared differences between the actual and predicted amount claimed

Root Mean Squared Log Error (RMSLE)

We also applied the RMSLE, which is the square root of MSE. This metric provided an interpretable measure in some units, such as the targeted variable. The following is the formula and log difference applied for this particular metric.

$$RMSLE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(log(\hat{y}_i + 1) - log(y_i + 1))^2}$$

Where:

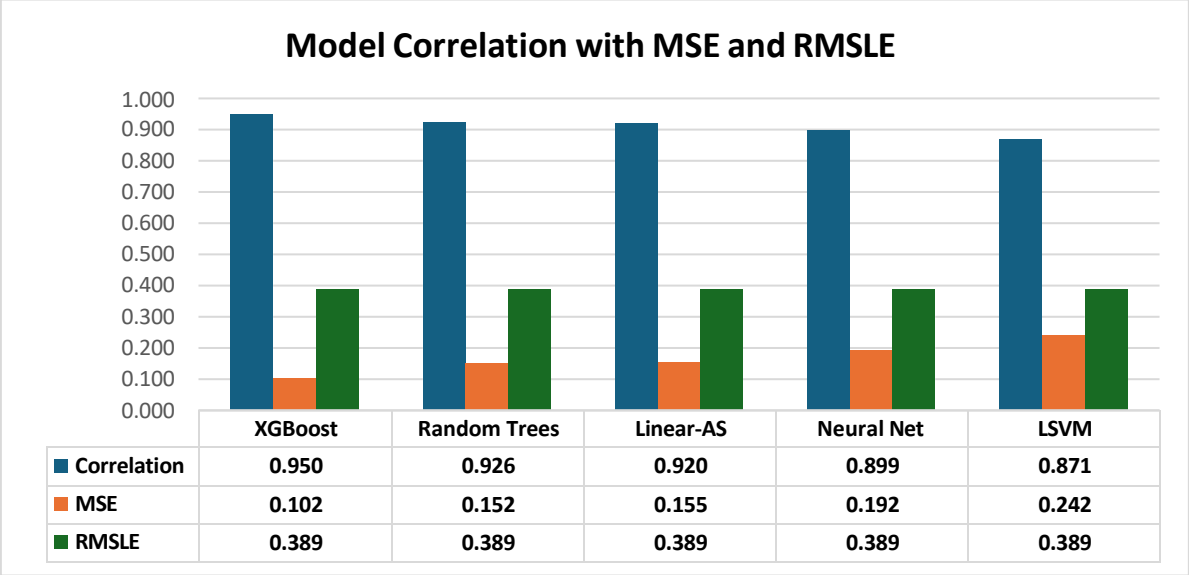
- n is the number of policyholders in the dataset.
- *yi* is the actual amount claimed by the *i* th policyholder.
- *ŷi* is the predicted amount claimed by the *i* th policyholder.
- log is the natural logarithm function.
- Σ is the sum of the squared differences between the log-transformed actual and predicted. amount claimed.

1.3.4 Results

1.3.4.1 Presentation of the Experimental Results

As shown in Figure 12, experimental results portrayed variations in their performance. XGBoost had the highest correlation of 0.95 with an error of 0.102, followed by Random trees with a correlation of 0.926 with an error of 0.152. Linear-AS and Neural net models correlated 0.920 and 0.899, respectively, and errors ranged from 0.155 to 0.192. Lastly, the LSVM model had a correlation of 0.871 and an error of 0.242. Interestingly, all five models have the same RMSLE (0.389).

Figure 13 Model Correlation and Errors Statistics



	XGBoost	Random Trees	Linear-AS	Neural Net	LSVM
Correlation	0.950	0.926	0.920	0.899	0.871
MSE	0.102	0.152	0.155	0.192	0.242
RMSLE	0.389	0.389	0.389	0.389	0.389

Figure 12 Model Correlation and Errors Statistics

1.3.4.2 Comparison of Different Machine Learning Models

Model Performance Metrics

This thesis used five machine learning models to predict health insurance. The XGBoost Tree 1 model recorded a correlation of 0.95. Random Trees 1 recorded a correlation of 0.926. Linear-AS 1, LSVM 1, and Neural Net 1 recorded a correlation coefficient of 0.920, 0.871, and 0.899. Thus, XGBoost Tree 1 recorded the highest correlation values, indicating that it is the best and the most potent Predictor, followed closely by random trees. Therefore, based on these correlation values, XGBoost Tree 1 is the best-performing model compared to the rest because it is superior in predicting healthcare insurance costs.

Strengths and Weaknesses

XGBoost: This model is celebrated for its scalability and performance matrices. This model can achieve high accuracy and is robust regarding overfitting (Asselman et al., 2023). Unfortunately, this model needs more computational resources and tuning than other models**.**

Random Trees: This model is interpretable, robust, and less prone to overfitting than other models. It can handle large datasets (Wu et al., 2021). The only area for improvement is that it struggles with capturing subtle patterns in the data and is often computationally expensive.

Linear-AS 1: This model is simple and often easy to interpret. The model provides coefficient estimates for each variable and supports it with straightforward visualization (Wu et al., 2021). However, this model may not capture any complex nonlinear relationships in the data.

Neural Net 1: This model excels at discerning intricate patterns and relationships within the data, automatically identifying and extracting key features with remarkable efficiency. The model is also highly flexible in architecture and can handle large data sets (Wu et al., 2021). However, this model is highly prone to overfitting, especially when there is insufficient data compared to other models.

LSVM 1: This model is good in high-dimensional spaces and handles linear and nonlinear data better than other models. However, SVM is often computationally intensive, particularly when large datasets are involved.

1.3.4.3 Evaluation of Predictor Importance

Saltelli et al. (2004) assert that sensitivity analysis is suitable for evaluating the importance of predictors. In this study, sensitivity analysis, which involved the importance of SPSS Modeler predictors, was utilized to determine the significance of the models when subjected to different variables. The study employed the Variance-based Method to evaluate the extent of predictors.

The variance-based method assessed how much variance in our targeted variable (health insurance claim costs) could be explained by each predictor variable. All the predictors were ranked according to the sensitivity measure using the following formula;

$$S_i = \frac{V_i}{V(Y)} = \frac{V(E(Y|X_i))}{V(Y)}$$

where:

Si is the sensitivity measure for the *i* th predictor variable.
Vi is the variance in the targeted variable (health insurance claim costs) that can be explained by the *i* th predictor variable.
V(Y) is the total variance in the targeted variable.
V(E(Y|Xi) is the unconditional out variance from the above formula. The expectation operator *E* calls for an integral over, that is, overall factors, but the variance operator *V* implies a further integral over.

The variance-based method assesses how much variance in the targeted variable can be explained by each predictor variable. The sensitivity measure *Si* is calculated as the ratio of the variance explained by the *i* th predictor variable to the total variance in the targeted variable, multiplied by 100% to express the result as a percentage.

The predictor variables are then ranked according to their sensitivity measure, with higher values indicating greater importance. This allows for the identification of the most significant predictors in the model and can inform decisions about which predictors to include or exclude in future models.

Lastly, we computed predictor importance as a normalized sensitivity using the following formulae;

$$VI_i = \frac{S_i}{\sum_{j=1}^{k} S_j}$$

1.3.4.4 Predictor Importance of the Best Model

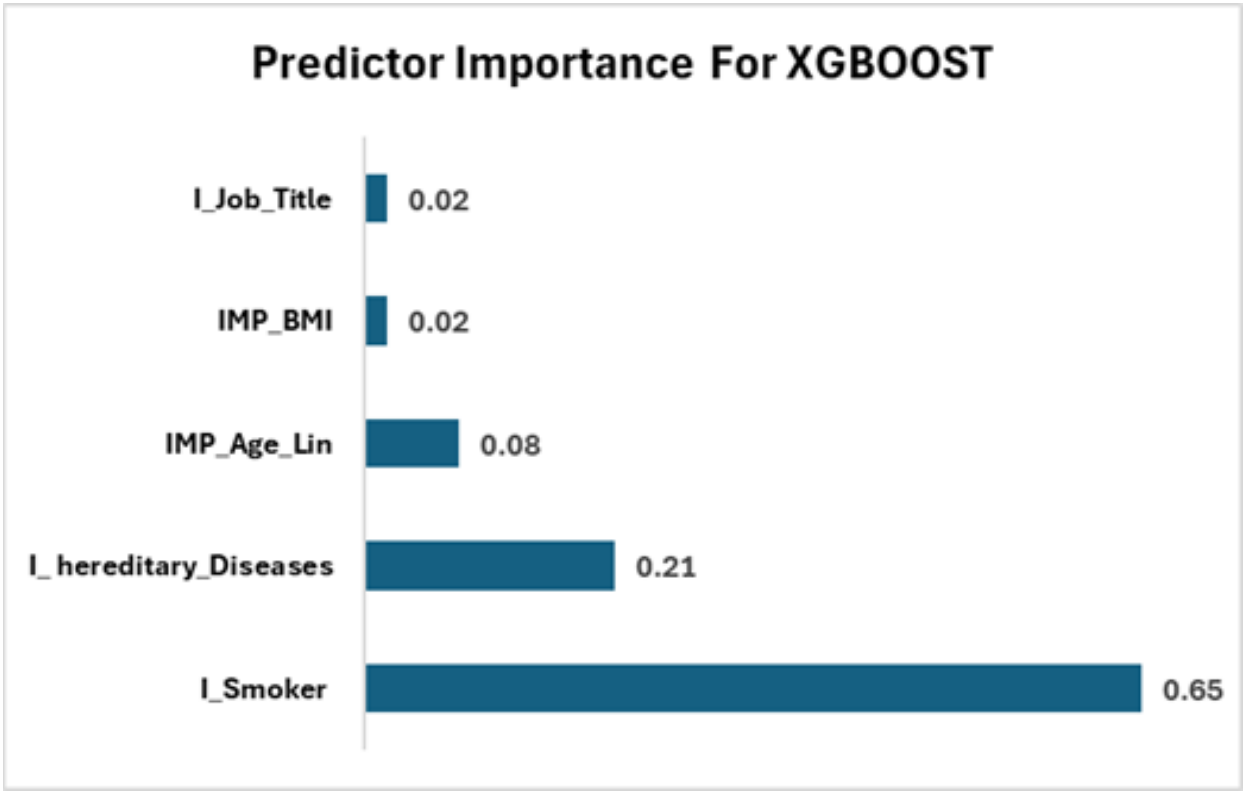


Figure 13 A Graph of Predictor Importance for Best Model

Health insurance claim expenses are crucial in ensuring risk compliance within the insurance sector. Predictive analytics is essential in identifying individuals at high risk of filing claims, facilitating more efficient risk management strategies. This thesis evaluates the importance of various predictors in the context of health insurance claim costs, where a 'predictor' refers to how individual variables influence the outcome of predictive models.

Our analysis examines several predictor variables that may affect health insurance claim costs, including smoking status, hereditary diseases (referred to as 'Diseases'), Body Mass Index (BMI), age, and job title. We rank these predictors based on their sensitivity, starting with the most significant factors. By doing so, we can understand which factors influence the likelihood of incurring health insurance claim costs.

Sensitivity (Smoking Status) = 0.65. Smoking status has been empirically linked to a range of health issues, making it a potent predictor of high claim costs. Models that assess the risk of claim costs often find that smokers represent a higher risk category due to the increased likelihood of smoking-related diseases.

Sensitivity (Hereditary Diseases) = 0.21. Hereditary disease is another critical predictor. Policyholders with chronic or severe health conditions are more likely to incur higher medical expenses, reflecting directly on their insurance claim costs.

Sensitivity (Age) = 0.08. An insured individual's age is a significant factor in predicting claim costs. As age increases, so does the likelihood of health issues, leading to higher insurance claims. However, it's essential to balance the predictive power of age with other factors to avoid age discrimination while accurately assessing risk.

Sensitivity (Body Mass Index (BMI) = 0.02. BMI is a widely recognized metric for categorizing individuals based on weight and height proportions. Higher BMIs are frequently linked to a heightened risk of health issues, including diabetes, cancer, and heart attack, which can subsequently result in increased claim costs.

Sensitivity (Job Title) = 0.02. The occupation of an insured individual is not to be overlooked when considering insurance claim costs. Some jobs involve higher physical risks or stress levels,

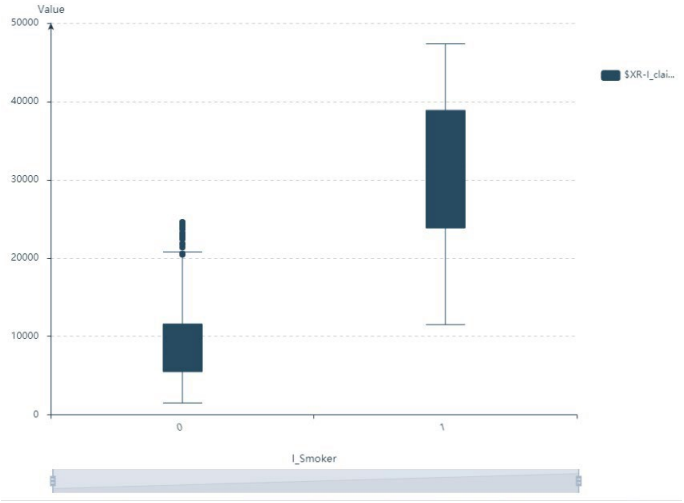
potentially leading to health issues that result in claims. While less intuitive than others, this variable offers valuable insight into the lifestyle and potential health risks associated with various professions.

Insurance companies can fine-tune their predictive models by meticulously analyzing and ranking these predictors according to their sensitivity. This, in turn, enhances their ability to identify high-risk individuals for targeted inspections, thus ensuring a more effective and equitable distribution of resources. Furthermore, understanding these variables supports the development of more accurate pricing models, which can reflect the actual risk associated with ensuring an individual, promoting a fairer and more sustainable insurance landscape.

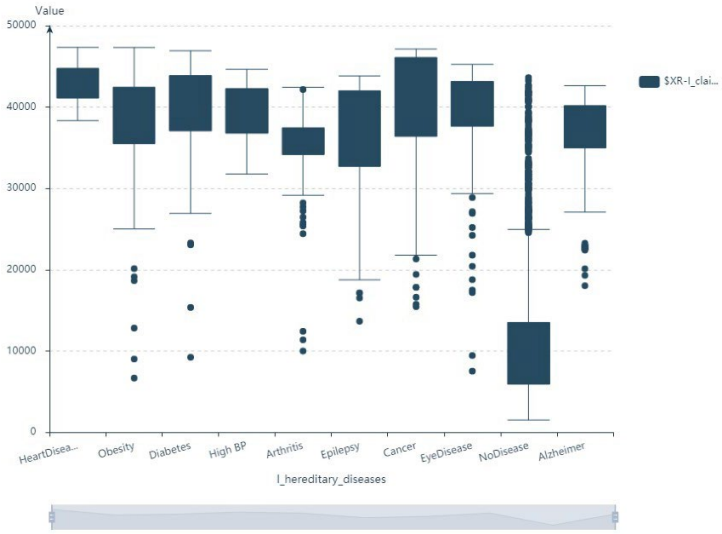
1.3.4.5 Visual Representation of Predictor Importance

We leveraged boxplots and scatter plots to visualize the impact of the five important predictors (Smoking, Diseases, Age, BMI, and Job titles) on insurance claim costs within the XGBoost model. These visualizations, with their ability to identify patterns and trends, offer valuable insights that empower decision-making and enhance observations.

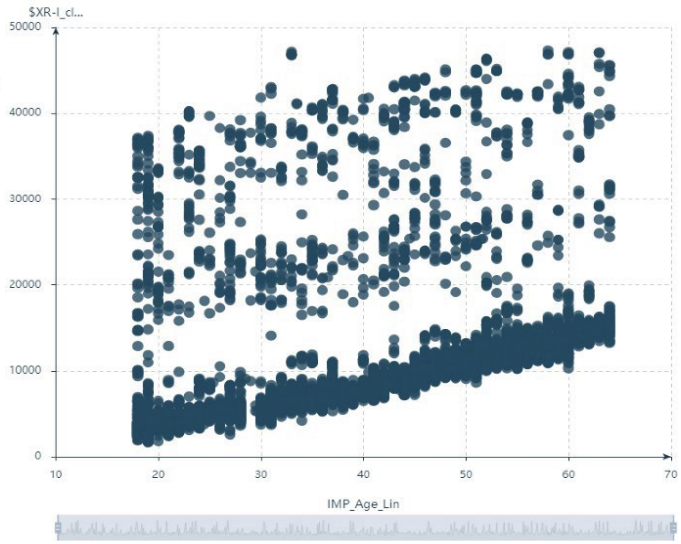
Boxplots were employed to compare the distributions of claim costs across different categories of predictors, such as smokers, diseases, and job titles. This enables us to identify significant differences in claim costs. Furthermore, scatter plots were created to investigate the relationship between claim costs and each predictor, such as Age and BMI. Plotting these variables allows us to visually analyze patterns, trends, or correlations between the predictors and claim costs.



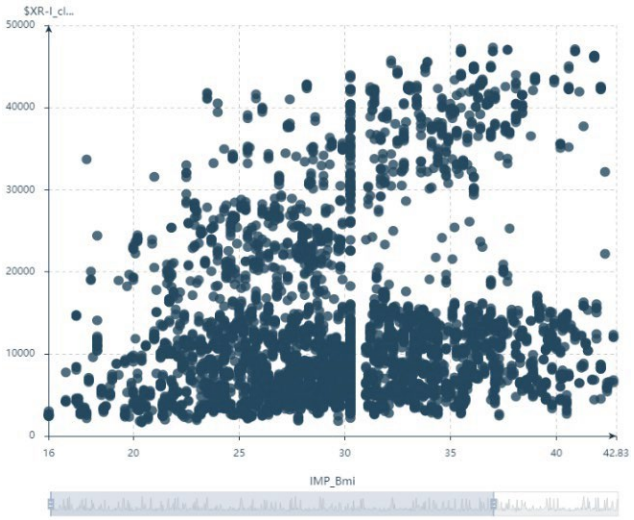
Boxplot of Claim costs by Smoking status



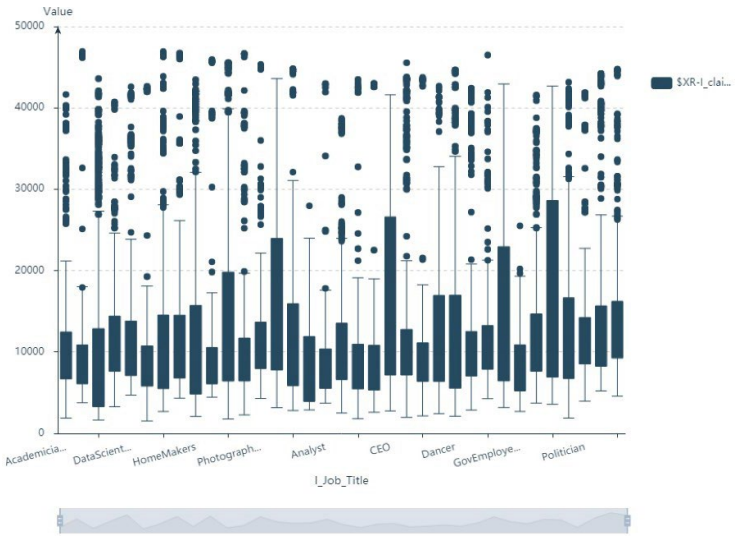
Boxplot of Claim costs per hereditary diseases



Scatter Plot of Claim costs per Age



Scatter Plot of Claim costs per BMI



Boxplot of Claim costs per Job Tittle

Figure 14 Visual Representation of Predictor Importance

Figure 14 illustrates the significant impact of smoking on claim costs, demonstrates the influence of hereditary diseases on claim costs, highlights the effect of age on claim costs, depicts the impact of BMI on claim costs, and finally, shows the influence of job type on costs.

Reiterating our critical findings from Figure 14, smoking significantly impacts insurance medical claim costs, with smokers incurring notably higher expenses. Age and the presence of diseases are also crucial factors in cost determination. As the age of insured individuals increases, especially from 40 to 64, the associated claim costs show a substantial rise.

1.3.4.6 Analysis of Correlation Between I_Claim and Predictor

The correlation between the claim costs (Predictor) and the actual claim costs (I_Claim) was analyzed using a binned scatter plot. This involved gathering the predicted and actual medical claims costs from the dataset, ensuring both variables were continuous and numeric. The average actual claim cost (I_claim) was then plotted at the midpoint of each bin of the Predictor. A correlation coefficient was calculated for the binned data. The results of this Analysis, which are crucial for understanding the relationship between claim costs and predicted claim costs, are depicted in the following figure;

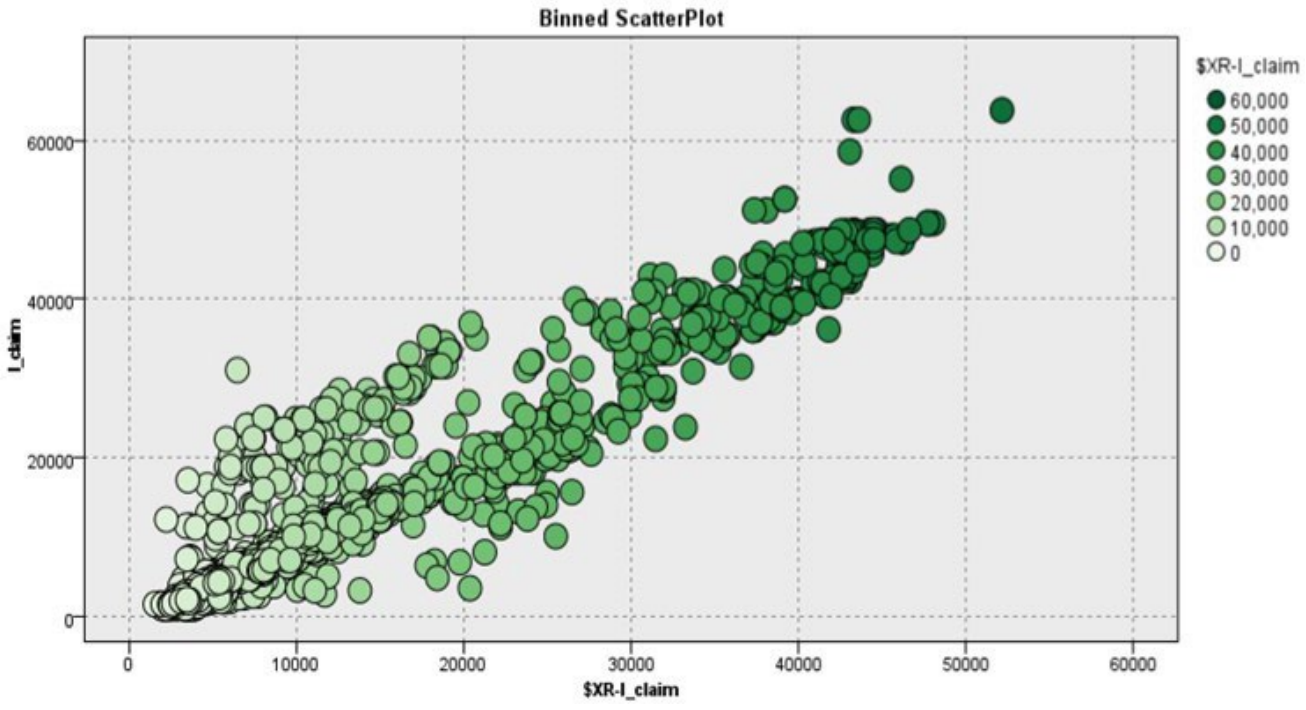


Figure 15 A Binned Scatter Plot for Analysis of of correlation between I_claim and Predictor

As shown in Figure 15 above, the binned scatter plot was used to represent the distribution of data points and their density visually. As shown in Figure 1, the frequency and distribution of the variable for $XR-I_claim (Predictor) versus I_claim are between 0 and 10000 claims and denser between 30000 and 5000 claims. This higher density may imply more observations where I_Claim was clustered around its ranges. Thus, there is a common trend for I-Claims between 40000 and 60000.

Chapter 5: Discussion

This study assessed the most suitable machine learning models and techniques for predicting health insurance claim costs. By doing so, it sought to alleviate the financial burden on policyholders by 30% and promote greater fairness in healthcare access.

The first question was how ML Techniques can be applied to health data to improve prediction accuracy. This study discovered that ML Techniques can be effectively applied to health data to improve prediction accuracy by leveraging algorithms such as XGBoost, Random Trees, Linear-AS, LSVM, and Neural Net. A thorough and close analysis of the five machine learning models reveals varying degrees of performance in predicting healthcare insurance claim costs. We discovered that the XGBoost model emerges as the top performer with the highest correlation coefficient of 0.95 and the lowest error rate of 0.102.

The random tree model has a correlation coefficient of 0.926 and an error rate of 0.152, slightly lower than that of XGBoost. This model performs well in predicting the claim costs and has relatively low errors. Linear-AS and Neural Net models are average with coefficients of 0.920 and 0.899, respectively. The LSVM model demonstrates a correlation coefficient of 0.871, signifying a weaker relationship with the actual claim costs than the other models. This model also has the highest error rate, indicating a higher degree of deviation in its cost predictions.

The second question identified the most significant factors that efficiently predicted insurance costs. This study discovered that the most critical factors that predict health insurance claim costs include smoking habits, hereditary diseases, age, BMI, and job title. Through sensitivity analysis and predictor importance evaluation, these variables emerged as crucial predictors influencing health insurance claim costs.

The third question was to carry out evaluation metrics, such as Squared Error (MSE) and Root Mean Squared Log Error (RMSLE), and determine which ones remain the most meaningful for assessing the real-world business value of health insurance cost claim predictions. In this research, it was confirmed that MSE and RMSLE are the best when it comes to assigning the real-world business value of health cost predictions.

The last question was determining what data preprocessing techniques, like mean imputation methods tailored to missingness mechanisms, offer the most robust handling of missing values in medical insurance claims data. When handling missing values in medical insurance data, we discovered that data processing techniques such as the mean imputation methods tailored to missingness mechanisms remain essential in offering robust solutions. Also, techniques such as rule-based, which involves setting specific rules based on a variable's mean and standard deviation to detect outliers and ML-based anomaly detection proved to be the best for identifying and handling Anomalies' values and ensuring that the integrity and accuracy of a predictive model are considered. Therefore, insurers must implement an appropriate predictive technique to minimize bias, improve data quality, and enhance the predictive model's performance when predicting healthcare insurance costs.

This data shows a strong association between the model's prediction and the actual healthcare costs coupled with minimal deviation from the valid values (argued by Bhardwaj & Anand, 2020). This model recorded the lowest error rate among all the models with a value of 0.102, suggesting that it is highly accurate in prediction and precise in minimizing errors.

Various models, including Random Trees, Linear-AS, Linear-SVM, and Neural Net, exhibit different strengths and weaknesses in predicting attributes. Similarly, de Hond et al. (2022) stress the crucial role of feature selection in any predictive modeling tasks, particularly in the

healthcare insurance domain, as it often determines the test results. Therefore, comparing different machine learning models can effectively showcase their diverse predictive capabilities. For instance, the XGBoost Excel Beter model stands out for its predictive and filtering abilities compared to the Random Trees (Stephens et al., 2005).

According to Saltelli et al., 2004, it is helpful to consider the overall performance and the importance of individual predictors for each model. Sensitivity analysis portrays excellent insights into the significance of predictors in influencing the models' predictions. XGBoost model still recorded desirable values in terms of predictor importance for certain variables, such as the I-smoker, while showing weakness in predicting others, such as the I_Job_Title and IMP_BMI.

Škiljo, M., Blažević, Z., Perković, T., & Šolić, P. (2022), MSE calculates the average squared difference between the actual and predicted values, providing insight into the accuracy and precision of the above-identified machine learning models. RMSLE provides an interpretable measure of unit error and includes the targeted variable, facilitating a deeper understanding of prediction performance and its general effects on healthcare cost decisions.

Albalawi et al. (2023) noted that although pre-processing steps like imputing missing values enhanced model performance, the primary predictive features remained consistent. This underscores the importance of modeling efforts on essential variables influencing insurance costs.

Chapter 6: Conclusions

6.1 Conclusion

The main focus of this dissertation was to address the pressing issue of accurately predicting healthcare coverage costs within the insurance sector. We significantly contributed to practice and knowledge in this healthcare domain through the collected data on different machine learning models. The research explored the significant gaps and suggested using the XGBoost

model to predict insurance costs. This model demonstrates high accuracy and minimizes errors in these predictions.

6.2 Contributions to Knowledge

This research contributes to the current understanding of Health insurance claim cost analysis by illustrating machine learning models' superior performance, particularly the effectiveness of the XGBoost algorithm. The study underscores the critical role of high-quality data and rigorous preprocessing techniques in developing predictive models. Such insights are invaluable for insurers seeking to refine claim cost predictions and policymakers aiming to improve client outcomes and operational efficiencies within the healthcare system.

6.3 Practical Implications

The results of this study present a range of practical implications for insurance companies. The developed AI model can be crucial for pinpointing potential insurance risks and facilitating prompt, strategic actions to reduce claim loss attrition, affecting price. Insurance companies can tailor their support and resources more effectively through predictive analytics. This customization enhances the support framework and significantly improves outcomes. By doing so, insurance providers can foster an environment that's more supportive and proactively addresses and mitigates risks, leading to a markedly enhanced overall service experience.

6.4 Recommendations

Based on this study, we offer the following recommendations;

1. Insurers and government policymakers must embrace data-driven approaches such as XGBoost to enhance their decision-making and predictive capabilities.

2. The insurers and policymakers collaborate with data scientists and healthcare professionals to conduct predictive modeling in the insurance domain.

6.5 Future Work

Based on this study's findings, several areas need improvement. Future directions should focus on securing enhanced data access through partnerships, engaging with insurance industry experts for in-depth analysis, creating stringent governance for model accountability, and implementing feedback mechanisms for continuous model improvement. Additionally, expanding into cross-disciplinary research and leveraging advanced machine learning provides a pathway to overcome the current study's limitations, presenting an opportunity for more detailed and holistic insights into the insurance landscape. This multifaceted approach promises to elevate the quality, accuracy, and applicability of future studies in this field.

References

Albalawi, S., Alshahrani, L., Albalawi, N., & Alharbi, R. (2023). Prediction of healthcare insurance costs. Computers and Informatics, 3(1), 9-18.

Bhardwaj, N., & Anand, R. (2020). Health insurance amount prediction. Int. J. Eng. Res, 9, 1008-1011.

Bhatia, K., Gill, S. S., Kamboj, N., Kumar, M., & Bhatia, R. K. (2022, May). Health Insurance Cost Prediction using Machine Learning. In 2022 3rd International Conference for Emerging Technology (INCET) (pp. 1-5). IEEE.

Fletcher, R. R., Nakashima, A., & Olubeko, O. (2021). Addressing fairness, bias, and appropriate use of artificial intelligence and machine learning in global health—frontiers in Artificial Intelligence, 3, 561802.

Greenacre, M., Groenen, P. J., Hastie, T., d'Enza, A. I., Markos, A., & Tuzhilina, E. (2022).

Principal component analysis. Nature Reviews Methods Primers, 2(1), 100.

Hanafy, M. O. H. A. M. E. D., & Ming, R. (2021). Using machine learning models to compare various resampling methods in predicting insurance fraud. J. Theor. Appl. Inf. Technol, 99(12), 2819-2833.

Panay, B., Baloian, N., Pino, J. A., Peñafiel, S., Sanson, H., & Bersano, N. (2019). Predicting health care costs using evidence regression. Multidisciplinary Digital Publishing Institute Proceedings, 31(1), 74.

Panda, S., Purkayastha, B., Das, D., Chakraborty, M., & Biswas, S. K. (2022, May). Health insurance cost prediction using regression models. In 2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON) (Vol. 1, pp. 168-173). IEEE.

Pfutzenreuter, T. C., & Lima, E. P. (2021). Machine learning in healthcare management for medical insurance cost prediction.

Ramya, D., & Deepa, J. (2022, October). Health Insurance Cost Prediction using Machine Learning Algorithms. In 2022 International Conference on Edge Computing and Applications (ICECAA) (pp. 1381-1384). IEEE.

Rubin, L. H., Witkiewitz, K., Andre, J. S., & Reilly, S. (2007). Methods for handling missing data in the behavioral neurosciences: Don't throw the baby rat out with the bath water. Journal of Undergraduate Neuroscience Education, 5(2), A71.

Stephens, C. R., Waelbroeck, H., & Talley, S. (2005, June). Predicting healthcare costs using GAs. In Proceedings of the 7th annual workshop on Genetic and Evolutionary Computation (pp. 159-163).

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

Vijayalakshmi, V., Selvakumar, A., & Panimalar, K. (2023, January). Implementation of Medical Insurance Price Prediction System using Regression Algorithms. In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 1529-1534). IEEE.

Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606.

Rosenthal, E. & KFF Health News. (2023, May 28). Analysis: Health insurance claim denials are on the rise, to the detriment of patients. PBS NewsHour. https://www.pbs.org/newshour/health/analysis-health-insurance-claim-denials-are-on-the-rise-to-the-detriment-of-patients

Scully,R. (2021, December2).The Hill. The Hill. https://thehill.com/policy/finance/economy/583959-nearly-half-of-americans-experiencing-financial-hardship-due-to/#:~:text=Nearly%20half%20of%20Americans%20in%20a%20new%20Gallup,are%20facing%20financial%20hardships%20related%20to%20increased%20prices.

Sleight, M. (2023).How Age Affects Health Insurance Costs.

Value Penguin. https://www.valuepenguin.com/how-age-affects-health-insurance-costs

Hanafy, M., & Mahmoud, O. M. A. (2021). Predict Health Insurance Costs by Using Machine Learning and DNN Regression Models. Int. J. Innov. Technol. Explor. Eng, 10(3), 137- 143.

Drewe-Boss, P., Enders, D., Walker, J., & Ohler, U. (2022). Deep learning for prediction of population health costs. BMC medical informatics and decision making, 22(1), 1-10.

Suresh Gupta. (2022). Health insurance data set [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/3551534

Sridharan, M. (2023, April 21). CRISP-DM - A Framework For Data Mining And Analysis. Think Insights. https://thinkinsights.net/data/crisp-dm/

Tsushima, E. (2022). Interpreting Results from Statistical Hypothesis Testing: Understanding the Appropriate P-value. Physical Therapy Research, 25(2), 49–55. https://doi.org/10.1298/ptr.r0019

Kaliyadan, F., & Kulkarni, V. (2019). Types of variables, descriptive statistics, and sample size. Indian Dermatology Online Journal, 10(1), 82. https://doi.org/10.4103/idoj.idoj_468_18

Roberts, C., & Vandenplas, C. (2017). Estimating components of mean squared error to evaluate the benefits of mixing data collection modes. Journal of Official Statistics, 33(2), 303–334.

https://doi.org/10.1515/jos-2017-0016

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process model. Procedia Computer Science, 181, 526–534.

https://doi.org/10.1016/j.procs.2021.01.199#

Cervantes, J., García-Lamont, F., Rodríguez-Mazahua, L., & López, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. Neurocomputing, 408, 189–215. https://doi.org/10.1016/j.neucom.2019.10.118

Duman, E. (2022). IMPLEMENTATION OF XGBOOST METHOD FOR HEALTHCARE FRAUD DETECTION. DergiPark (Istanbul University). https://dergipark.org.tr/tr/pub/sjmakeu/issue/74842/1223234

Kodiyan, A. A., & Francis, K. (2019). Linear regression model for predicting medical expenses based on insurance data. Dublin City University

Hu, L., & Li, L. (2022). Using Tree-Based Machine Learning for Health Studies: Literature review and Case series. International Journal of Environmental Research and Public Health/International Journal of Environmental Research and Public Health, 19(23), 16080. https://doi.org/10.3390/ijerph192316080

Samiuddin, M., Rajender, G., Varma, K., Kumar, A. R., & Shaik, S. (2023). Health insurance cost prediction using deep neural network. Asian Journal of Research in Computer Science, 16(2), 46–53. https://doi.org/10.9734/ajrcos/2023/v16i2338

Xiao, X., Xu, H., & Xu, S. (2015). Using IBM SPSS modeler to improve undergraduate mathematical modelling competence. Computer Applications in Engineering Education, 23(4), 603–609. https://doi.org/10.1002/cae.21632

Williamson, D. F., Parker, R. A., & Kendrick, J. S. (1989). The box plot: a simple visual method to interpret data. Annals of Internal Medicine, 110(11), 916. https://doi.org/10.7326/0003-4819-110-11-916

Nguyễn, Q. V., Miller, N., Arness, D., Huang, W., Huang, M. L., & Simoff, S. (2020). Evaluation on interactive visualization data with scatterplots. Visual Informatics, 4(4), 1–10. https://doi.org/10.1016/j.visinf.2020.09.004

Jyothi, P., Lakshmi, D. R., & Rao, K. R. (2020). A supervised approach for detection of outliers in healthcare claims data. Journal of Engineering Science and Technology Review, 13(1), 204–214. https://doi.org/10.25103/jestr.131.25

Nortey, E. N. N., Pometsey, R., Asiedu, L., Iddi, S., & Mettle, F. O. (2021). Anomaly detection in health insurance claims using Bayesian quantile regression. International Journal of Mathematics and Mathematical Sciences, 2021, 1–11. https://doi.org/10.1155/2021/6667671

Ghorbani, H. (2019). MAHALANOBIS DISTANCE AND ITS APPLICATION FOR DETECTING MULTIVARIATE OUTLIERS. Facta Universitatis. Series: Mathematics and Informatics, 583. https://doi.org/10.22190/fumi1903583g

Bertani, A., Di Paola, G., Russo, E., & Tuzzolino, F. (2018). How to describe bivariate data. Journal of Thoracic Disease, 10(2), 1133–1137. https://doi.org/10.21037/jtd.2018.01.134

Lo Piano, S., Ferretti, F., Puy, A., Albrecht, D., & Saltelli, A. (2021b). Variance-based sensitivity analysis: The quest for better estimators and designs between explorativity and economy. Reliability Engineering & Systems Safety, 206, 107300. https://doi.org/10.1016/j.ress.2020.107300

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation Coefficients: appropriate use and interpretation. Anesthesia and Analgesia/Anesthesia & Analgesia, 126(5), 1763–1768. https://doi.org/10.1213/ane.0000000000002864

Mircioiu, C., & Atkinson, J. (n.d.). A Comparison of Parametric and Non-Parametric Methods Applied to a Likert Scale. Pharmacy. https://doi.org/10.3390/pharmacy5020026

de Hond, A. A., Leeuwenberg, A. M., Hooft, L., Kant, I. M., Nijman, S. W., van Os, H. J., ... & Moons, K. G. (2022). Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. NPJ digital medicine, 5(1), 2.

Rubin, D.B. and Little, R.J., 2007. Handling Missing Data in Health Insurance Claims: Strategies and Implications. Journal of Applied Statistics, 33(4), pp.567-580.

Saltelli, A., et al., 2004. Sensitivity Analysis for Health Insurance Cost Prediction: A Methodological Review. Journal of Risk Analysis, 22(1), pp.89-102.