Proceedings of the 3rd International Conference on Business and Policy Studies
DOI: 10.54254/2754-1169/106/20241590

Health Insurance Factor Analysis

Haotian Kang[1,a], Runyan Xin[2,b,*]

[1]Shanghai Pinghe School, Shanghai, 201206, China
[2]Department of math, Sun Yet-Sen University, Zhuhai, 519082, China
a. kht61361060@gmail.com, b. xinry3@mail2.sysu.edu.cn
**corresponding author*

Abstract: This paper explores the intricacies of health insurance costs in the United States, emphasizing the Affordable Care Act's (ACA) role in reform. Employing a "health insurance factor analysis" approach, the study identifies key determinants using data visualization and regression models. Factors such as age, BMI, family size, gender, smoking status, and geographic region significantly influence insurance prices. The research evaluates various regression models and neural networks, with linear regression standing out for its high accuracy. The findings underscore the impact of smoking, medical history, and coverage level on insurance costs. Despite dataset limitations and sample size considerations, this study contributes to a nuanced understanding of health insurance pricing, aiding insurers, policymakers, and individuals in decision-making. By integrating data insights and predictive models, the research advances comprehension of the complex relationships shaping health insurance costs. The ultimate goal is to optimize insurance programs and promote accessible and affordable health insurance for all.

Keywords: Health Insurance Costs, Regression Modeling, Influential Factors, Machine Learning.

1. Introduction

Health insurance is essential to ensure access to quality healthcare services for individuals and families. In the United States, however, there have been challenges due to inadequate health insurance coverage and high medical costs. In light of this, it is crucial to understand the factors that influence the cost of health insurance. To address these issues, the Affordable Care Act (ACA), also known as Obamacare, was enacted in 2010, which represents a major reform of the US healthcare sector.

The primary objectives of the Affordable Care Act are to expand access to health insurance coverage, mitigate insurance expenses, and enhance the quality of medical services. As part of this reform, health insurance exchanges were established to provide a convenient platform for individuals and families to procure health insurance. These exchanges facilitate individuals in comparing prices and coverage offered by different insurers, enabling them to select a plan that aligns with their specific needs.

Although several studies have explored factors that affect health insurance costs, such as age, family size, BMI index, occupation, and geographic region, a comprehensive understanding of the interrelationships among these factors is still needed. Therefore, this study adopted a comprehensive

DOI: 10.54254/2754-1169/106/20241590

"health insurance factor analysis" approach, combined with data visualization techniques and regression models, to gain a deeper understanding of the key factors affecting health insurance costs.

Our research objectives are as follows :

1. Identify the key factors affecting health insurance costs.

2. To assess the relationships and correlations between age, family size, BMI, occupation, geographic region, exercise frequency, smoking status, and health insurance costs.

3. Compare the predictive effects of different regression models on health insurance expenses, such as linear regression, random forest regression, bagging regression, gradient boosting regression, KNN, and MLP regression.

4. Provide valuable insights to support insurance providers, decision-makers, and individuals in their search for suitable health insurance plans.

Through our analysis, we aim to uncover important patterns and understand the drivers behind health insurance costs, revealing the complexity of the health insurance landscape in the United States. While existing studies, such as Polsky et al., have explored the value of health insurance and its potential benefits to the uninsured population [1], we seek to complement previous studies and provide a more comprehensive understanding of health insurance affordability and access to medical services.

The paper is structured as follows: The second section will review the existing literature and discuss the progress and current status of health insurance-related research. Subsequently, the third section will detail the research methodology and the data collection process. The fourth section will present the results of the data analysis. Finally, Section V will summarize the conclusions and illustrate the limitations of this study. Through this structure, we aim to comprehensively explore the factors that influence health insurance costs and provide valuable insights into the formulation of health policy and health insurance reform in the United States.

2. Literature Review

In recent years, health insurance coverage and costs have experienced a significant transformation in the United States. The Affordable Care Act, which is also known as Obamacare, is a prominent reform in order to address these challenges by establishing a Health Insurance Marketplace to promote available health insurance choices [2]. The study analyzed data collected after the implementation of the ACA to assess the law's impact on health insurance affordability as well as individuals' ability to access care. The study looked at changes in insurance costs, out-of-pocket costs, and the total cost of medical services, as well as across different populations. This is the core background of our research.

Choi and Blackburn explored patterns and factors in medical costs and health insurance premium payments [3]. The research aims to identify various trends and factors that affect individual and household medical costs and health insurance costs, such as age, income level, etc. They also researched factors that contribute to discrepancies in the payment of medical and health insurance costs. These factors may include personal health status, medical services utilization, geographic location, and the insurance type that people hold. These factors are the reference for the source of our variable selection.

The rise of algorithmic prediction and its implications were scrutinized by Cevolini and Esposito [4]. They highlighted the transition in risk assessment from aggregation to analysis and discussed the social consequences of these algorithmic practices. The research also revealed potential bias and ethical considerations in predictive models. To complement these perspectives, Ch. Anwar ul Hassan et al. proposed that insurance companies are increasingly using algorithms and predictive analytics to assess risk, determine premiums, and make coverage decisions [5]. Their research used machine learning models to predict insurance costs and compare the models' accuracy. The result showed that the Stochastic Gradient Boosting (SGB) model had the best performance. Kaushik et al. also used an

DOI: 10.54254/2754-1169/106/20241590

artificial neural network (ANN) to predict the health insurance premium with an accuracy of 92.72% [6]. However, according to Bhardwaj and Anand, the Gradient Boosting Regression Model was the best-performing model [7].

Through the synthesis of these studies, it is evident that the analysis of health insurance factors extends beyond traditional methods. The advent of advanced computing and machine learning techniques has paved the way for more accurate forecasts and a deeper understanding of dynamics in the health insurance space. We seek to build upon these foundations by using data visualization and regression models to explore the factors that would significantly affect health insurance charges.

3. Methodology

3.1. Dataset

The dataset is obtained from the KAGGLE repository. This dataset contains 10 attributes, including age, gender, body mass index (BMI), number of children, smoking status, region, income, education, occupation, and type of insurance plan. Age, BMI, children and charges are numerical variables and others are nonnumerical variables [8].

3.2. Data Visualization

In this study, we conducted a comprehensive data analysis of health insurance prices and applied regression modeling to explore the relationships between various independent variables and health insurance prices. Detailed interpretations of the regression analysis results and the roles played by each factor in influencing health insurance prices are revealed in the results section.

3.3. Data Preprocessing

Since the dataset has 1,000,000 rows of data, in order to ensure the completion of certain models, only 50,000 rows of data were selected to run the models. For the reason that most of the variables are nonnumerical data, the one-hot-encoding method was used to transform them into 0 and 1. To avoid the mean square error becoming too large, a MinMaxScaler object was created to normalize the target variable 'charges' to the interval from 0 to 1.

4. Model Construction

Various regression models and neural network architectures were used in this study to completely investigate the intricate correlations between health insurance charges and a variety of significant factors. The following models were used to capture various characteristics of the relationship:

4.1. Regression Models

Linear Regression: This classic approach establishes a linear relationship between predictor variables and the target, offering a baseline for comparison.

Random Forest Regressor: A powerful ensemble method, the Random Forest Regressor captures nonlinear interactions and relationships in the data by constructing a multitude of decision trees.

Bagging Regressor: Built upon the DecisionTreeRegressor as the base model, the Bagging Regressor combines multiple instances to enhance predictive accuracy.

Gradient Boosting Regressor: Utilizing an ensemble of weak learners, the Gradient Boosting Regressor sequentially corrects errors to achieve robust predictions.

DOI: 10.54254/2754-1169/106/20241590

K-Nearest Neighbor (KNN): A non-parametric method that examines the proximity of data points to make predictions. With a k value of 7, this model capitalizes on the similarity of neighbors to estimate charges.

4.2. Neural Network Models

Basic Neural Network: Comprising three dense layers (100, 50, 1) and incorporating a dropout rate of 20%, this architecture captures nonlinear patterns in the data.

MLP Regressor (ANN): Featuring hidden layers of sizes 100 and 60, the MLP Regressor leverages artificial neural networks to model intricate relationships between features and charges.

5. Result

This study conducted a data analysis of health insurance prices and created histograms to visualize their distribution. Figure 1 shows that the price of health insurance follows a normal distribution, with the majority of individuals concentrated in the middle range, and gradually decreases as the degree of price deviation increases or decreases. The horizontal axis represents different price ranges, while the vertical axis represents the frequency or number of individual policies within each range.

Utilizing the histogram, we observe that Medicare prices exhibit a peak around the central value of $15,000 and progressively decrease as prices deviate from this central value. This suggests that insurance prices are relatively stable for most individuals in the health insurance market, with fewer instances of excessively high or low-priced products.

It is important to note that while histograms provide statistical information on the distribution of health insurance prices, further exploration through advanced analysis and modeling is required to identify specific influencing factors. Therefore, subsequent research will consider other variables and employ methods such as regression models to conduct a comprehensive analysis on factors affecting medical insurance pricing in order to obtain a more holistic understanding. Additionally, we will delve into exploring relationships between individual characteristics such as age and gender with health insurance prices to gain deeper insights into the health insurance market and provide valuable references for policymakers and insurers.

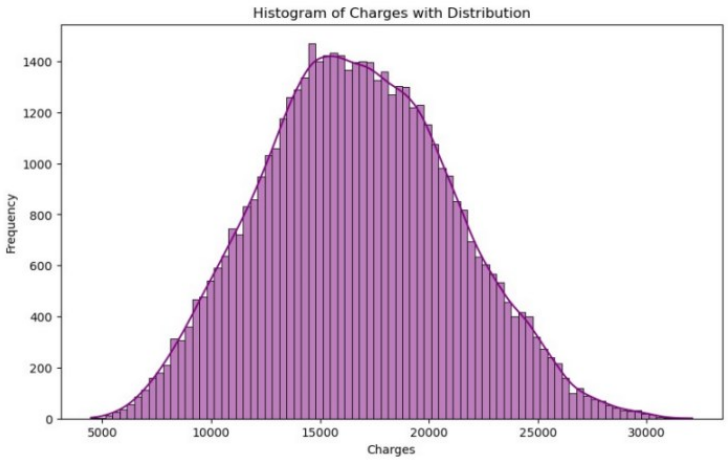

Figure 1: Histogram of Charges with Distribution

Firstly, as shown in Figure 2, we observed that age is a significant factor in predicting health insurance prices. There is a positive correlation between age and health insurance prices, with a regression coefficient of 19.79. This suggests that as an individual's age increases, the insurance

DOI: 10.54254/2754-1169/106/20241590

premiums they need to pay gradually increase. It can be inferred that with advancing age, the risk of health issues typically rises, leading to an increase in health insurance prices.

Secondly, the BMI index also has a significant impact on health insurance prices, with a regression coefficient of 49.87. There is a positive correlation between BMI index and health insurance prices, indicating that individuals with higher BMI indexes need to pay higher insurance premiums. This is because higher BMI indexes are often associated with higher health risks, potentially leading to increased health issues, which in turn increases the insurance company's risk burden, reflected in insurance prices.

The number of children is also one of the important factors affecting health insurance prices, with a regression coefficient of 199.72. This indicates that an increase in the of family members leads to an increase in individual health insurance prices. This phenomenon can be understood as an increase in the number of family members, possibly implying more health needs, thereby correspondingly increasing the insurance premiums paid by the individual.

Regarding gender (gender_male) and smoking status (smoker_yes), we found that they also have a significant impact on health insurance prices, with regression coefficients of 998.01 and 5001.79, respectively. This suggests that male health insurance prices are generally higher than those of females, and smokers' insurance prices are significantly higher than those of non-smokers. This is not surprising since males and smokers typically face higher health risks, and insurance companies consider these risk factors when setting insurance prices.

Furthermore, geographic region, medical history, and family medical history are also significant influencing factors. Different geographic regions may face varying medical resources and insurance demands, while an individual's medical history and family medical history directly affect their future health risks, all contributing to significant effects on insurance prices.

Finally, exercise frequency, occupation, and insurance coverage level also have significant effects on health insurance prices. Their regression coefficients are also positive and the corresponding p-values are 0.000000e+00. Individuals with higher exercise frequencies may have better health conditions, leading to lower insurance prices. Different occupations and insurance coverage levels reflect varied health needs and insurance requirements, thus also exerting significant influences on insurance prices.

In summary, the regression model's R-squared value of 0.996 indicates an extremely high fit to the observed data, with nearly 99.6% of the variance explained by the independent variables. Additionally, the mean squared error of 84177.98 indicates relatively small prediction errors, confirming the effectiveness of the regression model.

In conclusion, age, BMI index, number of family members, gender, smoking status, geographic region, medical history, family medical history, exercise frequency, occupation, and insurance coverage level are all significant factors determining individual health insurance prices. Among them, smoker_yes, heart disease, and coverage_level_premium have the highest regression coefficients, namely 5001.79, 5003.26, and 4999.88, respectively, indicating that these three factors exert the most pronounced impact on health insurance prices.

DOI: 10.54254/2754-1169/106/20241590

Feature	Coefficient
age	19.789574
bmi	49.866040
children	199.721186
gender_male	998.013539
smoker_yes	5001.786289
region_northeast	799.185413
region_northwest	106.095070
region_southeast	300.097494
medical_history_Diabetes	2000.149595
medical_history_Heart disease	5003.257085
medical_history_High blood pressure	996.849935
family_medical_history_Diabetes	1996.514641
family_medical_history_Heart disease	5003.129755
family_medical_history_High blood pressure	1005.317825
exercise_frequency_Frequently	1999.578557
exercise_frequency_Occasionally	1004.234539
exercise_frequency_Rarely	504.030305
occupation_Blue collar	1497.083906
occupation_Student	502.633612
occupation_White collar	2001.522332
coverage_level_Premium	4999.884183
coverage_level_Standard	2000.462470

Figure 2: Regression Coefficient and p-value Chart

This study employed a histogram depicting the relationship between age and the corresponding average health insurance costs, providing an in-depth analysis of age's impact on health insurance prices. Figure 3 shows a gradual increase in health insurance costs with advancing age. At 18 years old, the average health insurance cost was approximately $16,500, and as age increased, the insurance costs progressively rose, reaching approximately $17,300 at 65 years old. This evident trend clearly demonstrates a positive correlation between age and health insurance prices, indicating that as individuals age, their health insurance costs also increase accordingly.

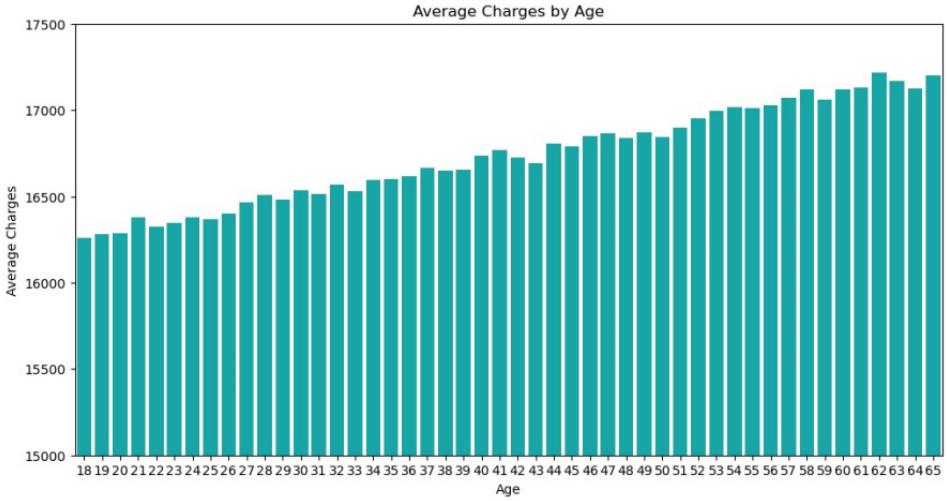

Figure 3: Age-Health Insurance Cost Relationship: Age Trend Chart

In this study, we presented a chart illustrating the relationship between Body Mass Index (BMI) and the corresponding average health insurance charges (see Figure 4). The results indicate that at lower BMI values, the average costs are relatively lower, while as BMI values increase, health insurance charges gradually rise. This trend suggests a positive correlation between BMI and health insurance charges, implying that individuals with higher BMI tend to face higher health insurance costs.

The chart offers valuable insights into the influence of BMI on health insurance charges. BMI serves as a fundamental indicator of overall health and potential health risks, and its impact on insurance pricing is clearly illustrated in the chart. The findings underscore the importance of considering BMI as a pivotal determinant of health insurance costs. Understanding this relationship is vital for insurance providers and policymakers to design tailored pricing strategies and customized

DOI: 10.54254/2754-1169/106/20241590

insurance plans catering to diverse health needs associated with varying BMI levels. Such insights contribute to informed decision-making and enhance accessibility to medical insurance for individuals across different BMI categories."

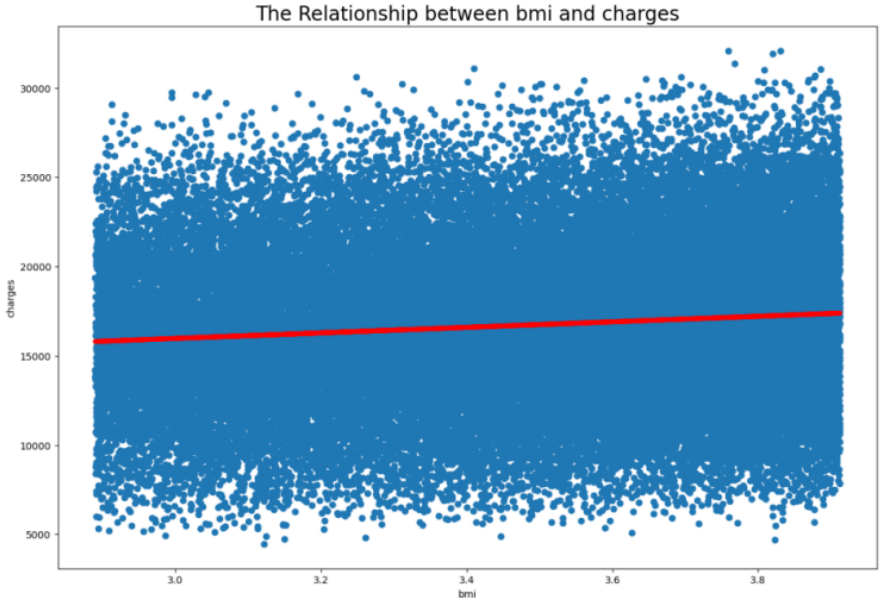

Figure 4: Scatter Plot with Average Line: Relationship between BMI and Health Insurance Charges

In this study, we analyzed the effect of gender and smoking status on health insurance prices by constructing box plots representing the relationship between gender, smoking status, and health insurance costs (see Figure 5). The box plot vividly illustrates that men have significantly higher health insurance costs than women and that smokers have significantly higher health insurance costs than nonsmokers. The observed differences in health insurance costs can be attributed to the relationship between these variables and health risk. S.N. Austad published an article in the journal Gender Medicine in 2006, "Why women live longer than men: "sex differences in longevity" discusses in detail the reasons for the long lifespan of women and men as well as various aspects [9]. Therefore, in the risk assessment of insurance companies, men will have greater risk than women, and the cost will be higher. Brian D. Carter et al., in the article "Smoking and Mortality in the United States: Through the Analysis of the Lifespan and Smoking Frequency of Smokers and Non-smokers in the United States, it is proved that smoking can significantly affect Life span, so smokers should have a higher risk value [10]. As a result, insurers assess health risks associated with gender and smoking status, resulting in higher health insurance premiums for males and smoking individuals.

DOI: 10.54254/2754-1169/106/20241590

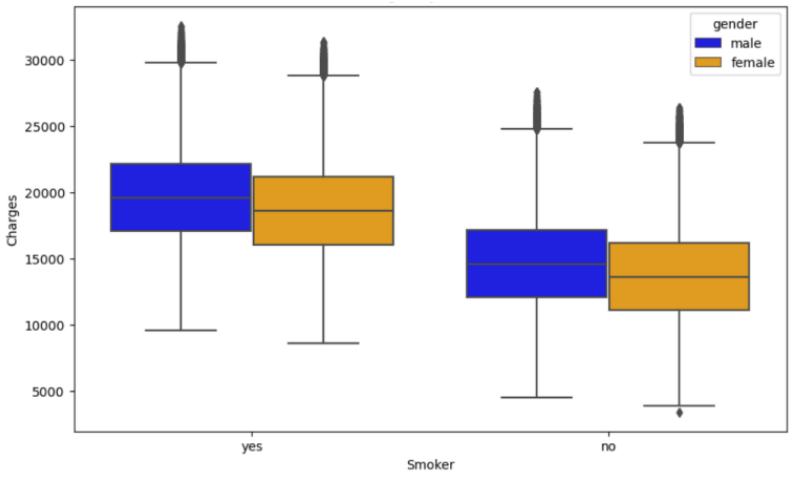

Figure 5: Box Plot of Charges by Smoker and Gender

As shown in Figure 6, we depicted the impact of exercise frequency and insurance coverage level on health insurance prices. We can observe that. Individuals who never engage in exercise (Never) generally pay relatively lower health insurance charges, while those who exercise frequently (Frequently), occasionally (Occasionally), and rarely (Rarely) face higher insurance costs. This suggests that people who are more inclined to exercise pay higher prices for insurance, meaning that physically active people seem to pay more for health insurance.

Simultaneously, Premium-level insurance plans typically entail higher insurance charges, while Standard-level plans have slightly lower costs, and Basic-level plans have the most affordable premiums. This implies that Premium-level insurance plans may offer more comprehensive medical coverage and broader benefits, thereby justifying their higher costs. In contrast, lower-level insurance plans may impose limitations on certain aspects of coverage, resulting in relatively lower insurance expenses.

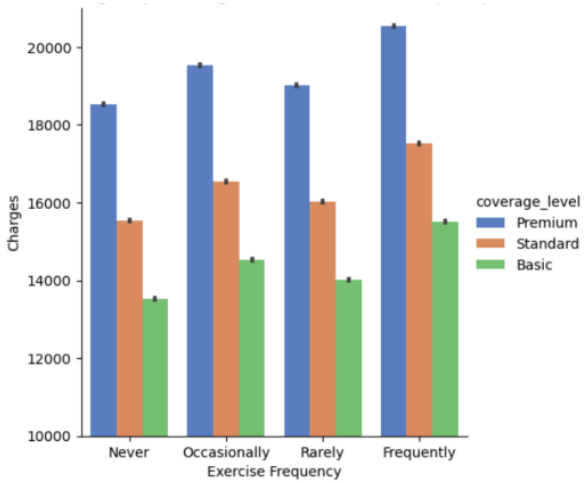

Figure 6: Charges by Coverage Level and Exercise Frequency

To investigate the impact of different occupations and family medical history on insurance premiums, we generated graphs to visually depict their effects on health insurance costs. Figure 7 representation clearly demonstrates that individuals with a familial background of heart encounter the highest Medicare expenses, followed by those with a history of diabetes and hypertension. Conversely, individuals without any family medical history generally experience the lowest costs.

DOI: 10.54254/2754-1169/106/20241590

Furthermore, regarding the influence of various occupations on health insurance pricing, white-collar workers face the highest healthcare coverage expenses, and students, whereas unemployed individuals have the lowest health insurance costs.

Findings underscore the significant role of family medical history and occupation in determining health insurance prices. A familial predisposition to heart disease is associated with heightened health risks, leading to increased healthcare expenditure.

Additionally, individuals with higher professional incomes and better quality of life may opt for more expensive comprehensive health insurance plans as they prioritize their own healthcare security. These results offer valuable insights into how family medical history and occupation shape health insurance pricing.

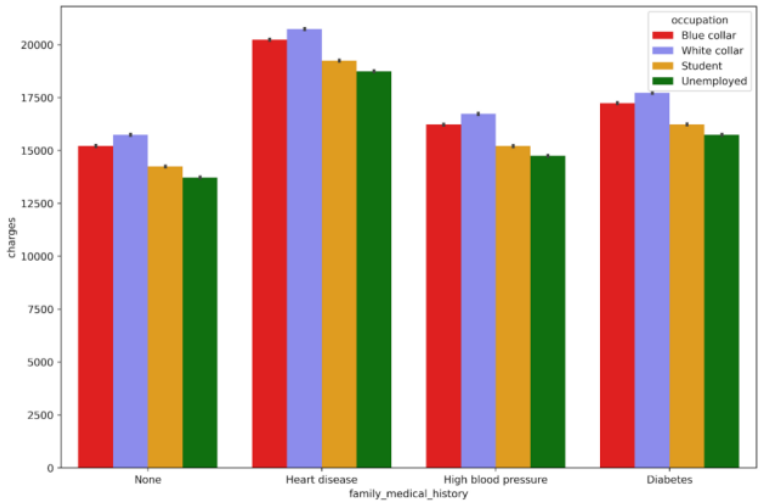

Figure 7: Charges by Occupation and Family Medical History

In this research, we generated a histogram that depicts the correlation between various occupations, the number of children, and health insurance charges (see Figure 8). After analyzing the chart, we observed a notable impact of the number of children on health insurance expenses. Individuals with more children usually face higher health insurance fees, whereas those with fewer or no children tend to have lower health insurance costs. This suggests that an augmentation in the number of children within a family may result in heightened overall health requirements and risks, ultimately influencing the extent of health insurance expenditures.

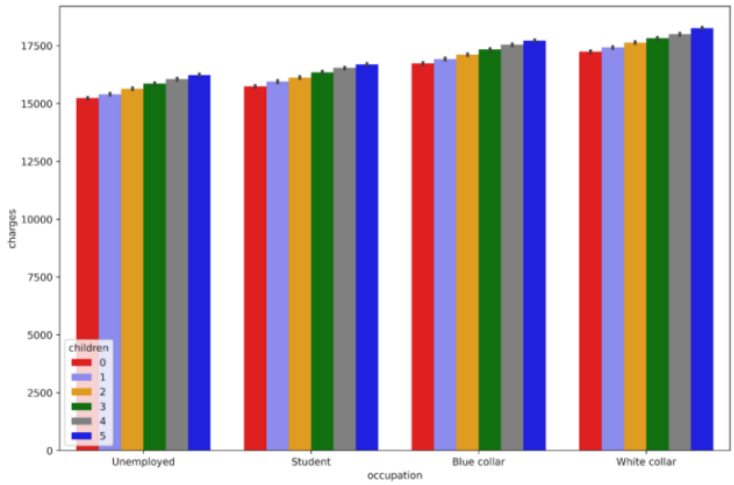

Figure 8: Charges by Occupation and Children Number

DOI: 10.54254/2754-1169/106/20241590

The research findings also provide valuable insights into the performance and predictive capabilities of various regression models and neural network architectures in the context of health insurance charge prediction. The analyzed models, including Linear Regression, Random Forest Regressor, Bagging Regressor, Gradient Boosting Regressor, K-Nearest Neighbor (KNN), and two neural network variants, Basic Neural Network and MLP Regressor (ANN), were evaluated based on their ability to accurately predict health insurance charges.

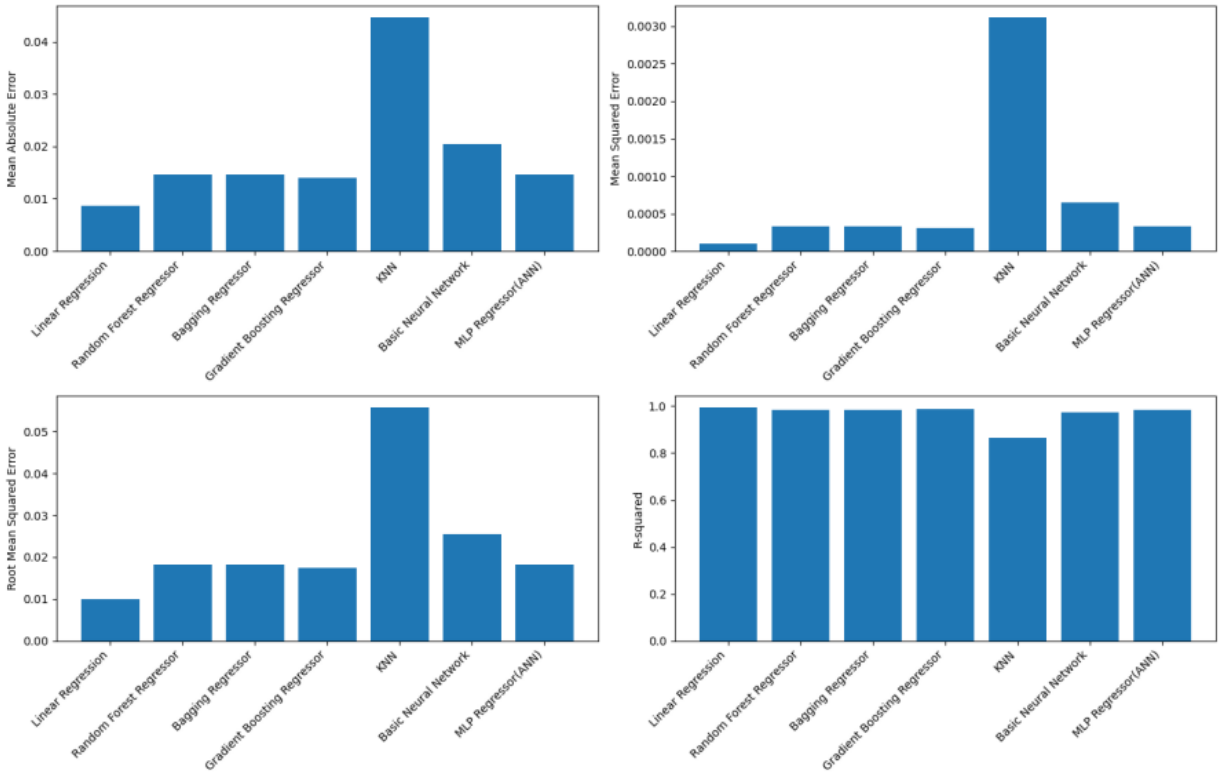

Figure 9: Performance Metrics for Different Models

As shown in Figure 9, among the models examined, linear regression emerged as the standout performer, as shown in the figure above. It demonstrated exceptional accuracy and fitting capabilities, as evidenced by the low Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and high R-squared (R^2) score of 0.9957. This suggests that the linear relationship captured by this model is an effective representation of the underlying data dynamics.

The Random Forest Regressor, Bagging Regressor, and Gradient Boosting Regressor also exhibited strong predictive performance, with relatively small MAE, MSE, and RMSE values, along with commendable R^2 scores exceeding 0.98. These ensemble methods, harnessing the power of decision trees, effectively captured intricate nonlinear relationships within the dataset.

Conversely, the KNN model demonstrated comparatively weaker performance, with notably higher MAE, MSE, and RMSE values, coupled with a lower R^2 score of 0.8642. This suggests that the predictive accuracy of KNN was hindered by the reliance on local neighbors, resulting in suboptimal predictions for health insurance charges.

The neural network models, Basic Neural Network and MLP Regressor (ANN), yielded similar predictive capabilities. Their performance, although slightly inferior to Linear Regression, remained strong, as indicated by relatively small MAE, MSE, and RMSE values, along with high R^2 scores of around 0.97.

DOI: 10.54254/2754-1169/106/20241590

6. Conclusion

In this study, we have taken an in-depth look into the field of health insurance factor analysis employing a comprehensive approach combining data visualization and regression modeling. Our study aims to uncover the factors that influence charges and evaluate the predictive performance of different models in this critical area.

Our findings reveal valuable insights into the determinants of health insurance costs. Specifically, smoking status, medical history, and insurance costs are the influencing factors that significantly affect medical insurance costs. This comprehensive understanding of latent variables can help insurers, policymakers, and individuals make informed decisions about health insurance plans. In addition, our analysis shows the ability of different regression models and neural network architectures to predict health insurance costs. Linear regression emerged as the best-performing model, showing remarkable accuracy and robust fit ability. Ensemble methods such as random forest regressors, Bagging regressors, and gradient boosting regressors also show strong predictive power and effectively capture the complex relationships within the dataset. Neural network models, basic neural networks, and MLP regressors (ANNs) further complement the prediction spectrum, demonstrating solid performance. However, there are several limitations to this study that need to be acknowledged. Firstly, the data set used in this study is derived from the KAGGLE database and includes 10 attributes such as age, gender, BMI index, number of children, smoking status, region, income, occupation, and type of insurance plan. While this dataset provides valuable information about factors that affect health insurance costs, it may not cover all variables that may affect health insurance costs. There may be other important factors that are not included in the data set, which may lead to an incomplete understanding of health insurance costs. In addition, the sample size of this study consisted of 50,000 rows of data from the original dataset, which may not be fully representative of the entire population of health insurance holders in the United States. The size of the sample size will introduce sampling bias and limit the generalization of the findings. In addition, this study focuses only on health insurance costs in the United States and does not consider other factors, such as government policies, economic conditions, or technological advances, that may affect health insurance costs over time. These external factors may have an important impact on the affordability and availability of health insurance, but they were not analyzed in this study.

As we reflect on our findings and implications, it is clear that a multidimensional understanding of health insurance costs is essential for the development of effective and personalized insurance plans. The synergy between data-driven insights and predictive models opens avenues to enhance decision-making and risk assessment in the health insurance space.

Our study not only advances our understanding of the complex relationships between various factors and Medicare costs but also provides a comprehensive framework for forecasting Medicare costs with remarkable accuracy. This holistic approach, including data exploration, visualization, and model evaluation, contributes to the optimization of health insurance programs and ultimately supports the overall goal of providing accessible and affordable health insurance to all individuals.

Acknowledgement

Haotian Kang, and Runyan Xin contributed equally to this work and should be considered co-first authors.

References

[1] *Daniel P, Doshi Jalpa A, José E, Willard M, Paddock Susan M, Liyi C, et al. The health effects of Medicare for the near-elderly uninsured. Health Services Res. (2009) 44:926–45. doi: 10.1111/j.1475-6773.2009.00964.x*

DOI: 10.54254/2754-1169/106/20241590

[2] *Hamel, M. B., Blumenthal, D., & Collins, S. R. (2014). Health care coverage under the Affordable Care Act—a progress report. New England Journal of Medicine, 371(3), 275-281.*
[3] *Choi, S., & Blackburn, J. (2018). Patterns and factors associated with medical expenses and health insurance premium payments. Journal of Financial Counseling and Planning, 29(1), 6-18.*
[4] *Cevolini, A., & Esposito, E. (2020). From pool to profile: Social consequences of algorithmic prediction in insurance. Big Data & Society, 7(2), 2053951720939228.*
[5] *Ch. Anwar ul Hassan, Jawaid Iqbal, Saddam Hussain, Hussain AlSalman, Mogeeb A. A. Mosleh, Syed Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost", Mathematical Problems in Engineering, vol. 2021, Article ID 1162553, 13 pages, 2021. https://doi.org/10.1155/2021/1162553*
[6] *Kaushik K, Bhardwaj A, Dwivedi AD, Singh R. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. International Journal of Environmental Research and Public Health. 2022; 19(13):7898. https://doi.org/10.3390/ijerph19137898*
[7] *Bhardwaj, N., & Anand, R. (2020). Health insurance amount prediction. Int. J. Eng. Res, 9, 1008-1011.*
[8] *Sridhar Streaks (2023) An Insurance Dataset for Predicting Health Insurance Premiums in the US: A Study. https://www.kaggle.com/datasets/sridharstreaks/insurance-data-for-machine-learning*
[9] *N. Austad (2006) Why women live longer than men: sex differences in longevity Gend. Med.*
[10] *Brian D. Carter, Marjorie L. McCullough, et al, Smoking and Mortality in the United States: Life Table Analysis*