Analyzing Correlations: Factors Influencing Personal Insurance Expenditure

Clayton-George Reid

Capstone Project

March 22, 2024

Insurance companies need to assess the level of risk associated with insuring an individual. Knowing the amount of coverage needed helps them understand the potential financial loss they may face if a claim is made. The amount of coverage directly impacts the premium the insured person will pay. By understanding the coverage needs, insurance companies can accurately price their policies to ensure they can cover potential claims while remaining profitable. Individuals have different insurance needs based on age, health status, occupation, lifestyle, and financial obligations. Tailoring the coverage amount to each individual's needs ensures they are adequately protected without overpaying for unnecessary coverage. Insurance companies must comply with regulatory requirements that may mandate certain minimum coverage levels for certain types of insurance, such as auto or homeowners insurance. Understanding an individual's coverage needs helps ensure compliance with these regulations.

JEL Code : E17, I11, I13

Keywords: Prediction, Health Insurance, Insurance Charges

**Introduction**

Being able to predict how much an individual or household could potentially cost an insurance company would revolutionize the insurance industry. Insurance is a fundamental component of modern risk management, providing individuals and businesses with financial protection against unforeseen events. In the insurance industry, accurately assessing the level of risk associated with insuring an individual is paramount. This assessment informs the determination of coverage needs, which in turn influences the premiums charged to policyholders. Understanding the factors that influence household insurance expenditure is crucial for both insurance companies and policyholders alike.

Insurance companies rely on a variety of factors to assess risk and determine appropriate coverage levels. These factors include demographic information such as age, health status, occupation, and lifestyle, as well as financial obligations and regulatory requirements. Tailoring coverage amounts to individual needs ensures that policyholders are adequately protected without paying for unnecessary coverage.

Moreover, regulatory requirements may mandate certain minimum coverage levels for specific types of insurance, such as auto or homeowners insurance. Compliance with these regulations is essential for insurance companies to operate within legal frameworks and maintain their financial stability.

In this research paper, we will delve into the factors influencing personal insurance expenditure and explore how insurance companies navigate the complexities of risk assessment and coverage determination. By gaining insights into these factors, we can better understand the dynamics of the insurance market and the implications for both insurers and policyholders. Through empirical analysis and discussion, we aim to shed light on the importance of accurately assessing coverage needs and the role it plays in ensuring financial security for individuals and businesses alike.

This paper will also delve into the intricate relationship between various demographic factors and insurance charges, leveraging comprehensive data analysis techniques to uncover significant insights. An initial examination of the dataset reveals essential descriptive statistics, showcasing key variables such as age, BMI, number of children, and charges. Notably, age emerges as a pivotal determinant across multiple machine learning models, consistently exerting the most substantial influence on insurance charges. Smoking status also plays a significant role, albeit to a lesser extent compared to age, indicating higher insurance costs for smokers. Moreover, gender dynamics reveal nuanced patterns, with men exhibiting higher insurance usage rates and average charges compared to women. Additionally, an intriguing disparity arises within the smoker subset, highlighting gender-specific variations in insurance charges. These findings underscore the complex interplay of demographic factors in shaping insurance pricing dynamics, setting the stage for a detailed exploration of their implications and potential policy interventions.

Investigating what variables within this dataset lead to an accurate prediction of charges, prompts the following questions:

1. Can we establish correlations between specific factors in our dataset and the prediction of charges?
2. Which variable exhibits the strongest and weakest correlations with charges?
3. Is there a significant difference in charges between smokers and non-smokers?
4. Do charges significantly differ between genders?

This paper aims to discern the predictive factors influencing insurance charges within our dataset. Central to our inquiry are several key questions guiding our analysis. First, we seek to determine if correlations exist between specific factors within our dataset and the prediction of charges. Second, we will identify which variable exhibits the strongest and weakest correlations with charges, shedding light on the most influential determinants. Additionally, we will examine whether a significant disparity in charges exists between smokers and non-smokers, providing insight into the impact of smoking status on insurance costs. Furthermore, we will explore gender-based differences in charges to elucidate any disparities between male and female policyholders. By addressing these questions, we aim to unravel the complex interplay of variables shaping insurance charges and provide valuable insights for insurers, policymakers, and consumers alike."

**Literature Review**

Predicting insurance costs is a critical aspect of risk assessment and pricing within the insurance industry. Numerous studies have explored various methodologies and factors involved in predicting insurance amounts, aiming to improve accuracy and efficiency in pricing policies. One common approach involves the use of actuarial models, which utilize historical data and statistical techniques to estimate future insurance claims and determine appropriate coverage levels. Actuarial models consider a wide range of variables, including demographic information, past claim history, and economic indicators, to forecast potential losses and set premiums accordingly (Wang et al., 2017).

Morid et al. (2018) conducted a systematic literature review to identify methods for predicting healthcare costs and empirically evaluated their performance using data from the University of Utah Health Plans. They identified five methods for predicting healthcare costs and compared their predictive performance with other state-of-the-art supervised learning methods. The data set consisted of approximately 90,000 individuals, 6.3 million medical claims, and 1.2 million pharmacy claims. In their comparative analysis, gradient boosting demonstrated the best predictive performance overall and for low to medium-cost individuals. Additionally, Artificial Neural Network (ANN) and the Ridge regression model, not previously reported for healthcare cost prediction, showed the highest performance for high-cost individuals. This study contributes to the understanding of predictive modeling techniques for healthcare costs, providing insights for health insurers and healthcare delivery systems to allocate resources efficiently and improve business planning.

Machine learning algorithms have emerged as powerful tools for predicting insurance amounts, offering the potential to uncover complex patterns and relationships within large datasets. Researchers have applied techniques such as decision trees, random forests, and neural networks to analyze insurance data and make predictions about future claims and policy costs (Chang et al., 2019). By leveraging advanced computational methods and incorporating diverse sets of variables, machine learning models can enhance the accuracy of insurance predictions and assist insurers in making more informed underwriting decisions.

This literature review provides a comprehensive overview of the methodologies and factors involved in predicting insurance costs, highlighting the significance of actuarial models and machine learning algorithms in this domain. It will aptly discuss the importance of leveraging historical data and statistical techniques, as well as the potential of machine learning to uncover complex patterns within large datasets (Ugochukwu Orji et al., 2024 and Keshav Kaushik et al,. 2022). Moreover, your review sets the stage for the contribution of your paper, which aims to further advance this discussion by exploring novel approaches for predicting insurance costs. Specifically, you mention recent advancements in data analytics, such as the use of telematics data and alternative data sources, to assess risk and tailor insurance premiums based on individual behavior and lifestyle factors(Anderson E Stanciole et al., 2008).

Hossen's thesis (2023) delves into the utilization of machine learning (ML) techniques for enhancing the efficiency of insurance policy terms in the insurance industry. By leveraging individual and local health data, the study aims to forecast insurance amounts for various categories of people. Nine regression models, including Linear Regression, XGBoost Regression, and Random Forest Regression, were employed and trained using the dataset. The predictive performance of these models was evaluated by comparing actual data with predicted values. This study contributes to the field of medical insurance cost prediction by providing insights into the effectiveness of different ML algorithms in accurately forecasting insurance amounts, thus aiding insurers in making informed decisions and improving policy terms. By integrating these innovative approaches, this paper adds to the evolving landscape of insurance prediction and underscores the importance of leveraging cutting-edge technologies to meet the evolving needs of the insurance market.

Furthermore, the goal of the research in this paper is to add to this discussion. Recent advancements in data analytics and predictive modeling have led to the exploration of novel approaches for predicting insurance costs. For example, researchers have investigated the use of telematics data from vehicles and wearable devices to assess risk and tailor insurance premiums based on individual behavior and lifestyle factors (Abiyev et al., 2020). Additionally, the integration of alternative data sources, such as social media activity and online purchasing behavior, has enabled insurers to gain deeper insights into customer preferences and risk profiles, thereby refining their predictions and pricing strategies (Viallefont et al., 2018). These innovative approaches highlight the evolving landscape of insurance prediction and underscore the importance of leveraging cutting-edge technologies to meet the evolving needs of the insurance market.

**Data Description**

The source of the data used in this project is a CCO: Public Domain data set from Kaggle. CCO: Public Domain data means the person who associated a work with this deed has dedicated the work to the public domain by waiving all of his or her rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law. You can copy, modify, distribute, and perform the work, even for commercial purposes, all without asking permission. Since it is CCO: Public Domain data the source is blanked however the data remains for anyone to use for any purpose free of charge.

The metadata contains 1338 rows of insured data, where the Insurance charges are given against the following attributes of the insured: Age, Sex, BMI, Number of Children, Smoker, and Region. The attributes are a mix of numeric(age, BMI, and number of children ) and categorical ( sex, smoker, and region )variables. After cleaning the dataset in Rstudio by making sure there were no NA variables or missing data, no observations were removed as there were no blank or missing variables after the cleaning. Table 1 below shows the number of observations, the mean, standard deviation, minimum, and maximum for each key variable.

[[1]](#footnote-1)

Table 1: Descriptive Summary Statistics

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Observations | Mean | Std. dev. | Min | Max |
| Age | 1337 | 38 | 13.943 | 18 | 64 |
| BMI | 1337 | 30.092 | 5.856 | 15.96 | 45.9 |
| Children | 1337 | 1.077 | 1.222 | 0 | 5 |
| Charges($) | 1337 | 8159.15 | 4929.64 | 1121.87 | 20984.09 |

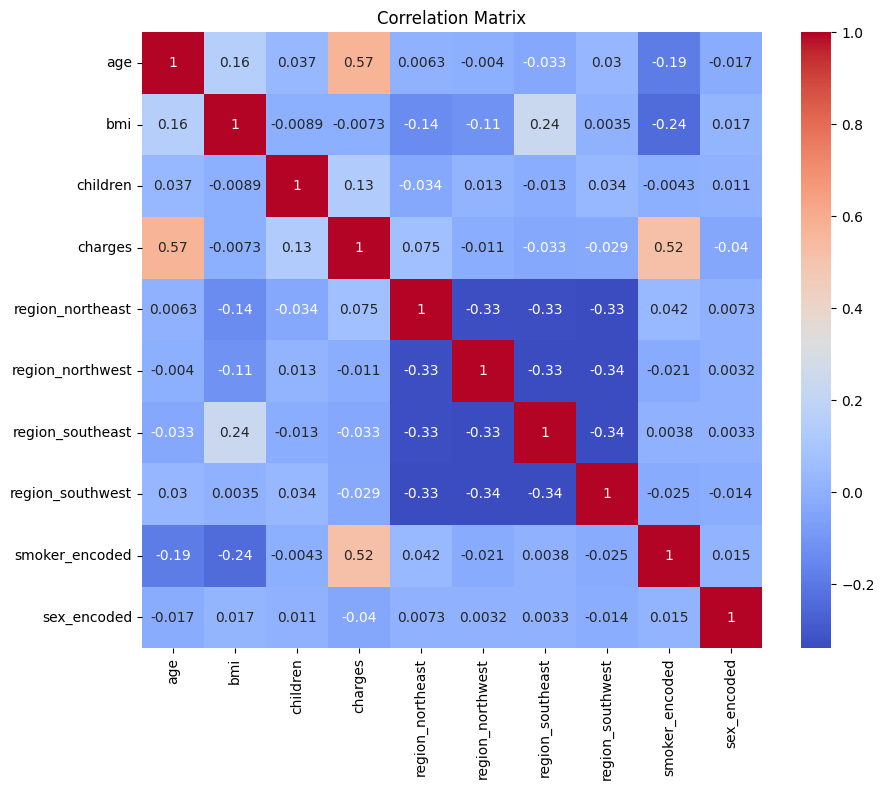
Table 1 provides a descriptive summary of key variables from a dataset, comprising 1337 observations. The variables include age, with a mean of 38 years and a standard deviation of 13.94, ranging from 18 to 64 years. Body Mass Index (BMI) has a mean of 30.092 and a standard deviation of 5.856, ranging from 15.96 to 45.9. The number of children shows a mean of 1.08, with a standard deviation of 1.222, varying from 0 to 5. Lastly, charges, measured in dollars, exhibit a mean of $8159.15, with a standard deviation of $4929.64, and values spanning from $1121.87 to $20984.09. These statistics offer insights into the central tendency, variability, and range of the variables in the dataset, crucial for understanding their distributions and potential relationships.

**Methodology**

The correlation analysis is based on a linear regression modeling approach. The

hypothesis this study intends to examine is the possibility of drawing a correlation between a factor from our dataset to predict charges, what variables correlate the most and least with charges, and whether there is a difference between the rates depending on the region of the U.S. Fig. 1 below shows the correlation of all variables to charges. The heat map highlights a strong correlation between age and smoker with charges.

Fig. 1 Heat Map of Correlation Analysis



A heat map is a graphical representation of data where individual values contained in a matrix are represented as colors. The heat map displayed in the image is a correlation matrix that visualizes the correlation coefficients between various variables such as age, BMI, number of children, insurance charges, and geographical regions. The colors range from dark blue (representing a correlation coefficient close to -1) to dark red (representing a correlation coefficient close to 1), with white indicating no correlation (0). For instance, there is a strong positive correlation (dark red) between 'charges' and 'smoker\_encoded', suggesting that smokers tend to have higher insurance charges. Conversely, there are several variables with little to no correlation with each other, as indicated by the white or light-colored cells. In this heat map we can see that age and smoker have the highest correlation to charges so moving forward the formula will be centered around those variables.

The equation utilized in this study for predicting charges is formulated as follows:

*charges = β0 + β1 age + β2 smoker + ∈*

where:

* charges represent the predicted charges.
* *β0* denotes the intercept term, capturing the baseline charges when all predictor variables are zero.
* ​ β1 and β2 signify the coefficients associated with the predictors of age and smoker respectively.
* *Age* is the predictor variable representing the age of the insured individual.
* *Smoker* is a binary predictor variable indicating whether the insured individual is a smoker (1) or a non-smoker (0).
* ϵ represents the error term.

**Results**

The study employed three machine learning models, namely decision tree, random forest, and gradient boosting, to predict insurance charges based on demographic and lifestyle factors. The performance of each model was evaluated using mean squared error (MSE) and mean absolute error (MAE) metrics. Additionally, feature importance analysis was conducted to assess the relative contribution of predictor variables.

Table 2 Coefficients Summary Table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | t value | Pr(>|t|) |
| Intercept | 1.739267 | 1.186733 | 1.466 | 0.143099 |
| age | 0.024133 | 0.005988 | 4.030 | 6.04e-05 \*\*\* |
| smoker | 1.396129 | 0.166684 | 8.376 | < 2e-16 \*\*\* |

|  |  |
| --- | --- |
| Residual standard error | 0.3745 |
| Multiple R-squared | 0.8385 |
| Adjusted R-squared | 0.8334 |
| F-statistic | 165.7 |
| p-value | < 2.2e-16 |

The statistical summary provided represents the output of a linear regression model analyzing the effects of age and smoking status on a dependent variable. The intercept, estimated at 1.739 with a standard error of 1.187, is not statistically significant (p = 0.143). However, age shows a positive and significant effect, with each additional year increasing the dependent variable by approximately 0.024 (p < 0.0001). Smoking status is also a significant predictor, with smokers having an average increase of 1.396 in the dependent variable compared to non-smokers (p < 0.0001). The model explains a substantial portion of the variance in the dependent variable, as indicated by a Multiple R-squared of 0.8385 and an Adjusted R-squared of 0.8334. The overall model is highly significant with an F-statistic of 165.7 and a p-value less than 0.0001.

The performance metrics for each model are summarized below:

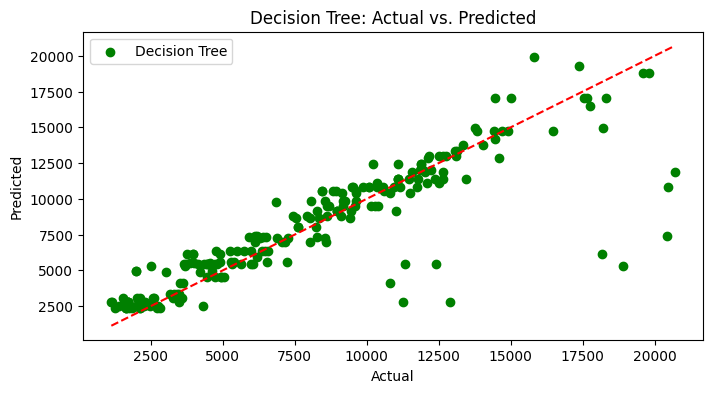
Decision Tree: Actual vs. Predicted:

A decision tree is a machine learning model used to make predictions based on a series of decision rules derived from the data features. Each point on the scatter plot represents an individual data point from the test dataset. The green dots represent the predicted charges by the decision tree model. The red dashed line represents the ideal scenario where actual charges perfectly match predicted charges. Deviations from this line indicate the model's performance in predicting insurance charges.

Decision Tree:

MSE: 5728632.31

MAE: 1275.02

Graph 1 Decision Tree Graph Results of Actual versus Predicted Expenditure

The decision tree prediction graph displays a scatter plot comparing the actual insurance charges to those predicted by a decision tree model. Each green dot represents a prediction for an individual data point in the test dataset. The red dashed line indicates the line of perfect prediction, where the predicted values would exactly match the actual values. The closer the green dots are to this red line, the more accurate the model's predictions. The spread of the green dots around the red dashed line illustrates the variance in the model's accuracy, with some predictions being very close to the actual values and others showing more significant deviations.

Random Forest: Actual vs. Predicted:

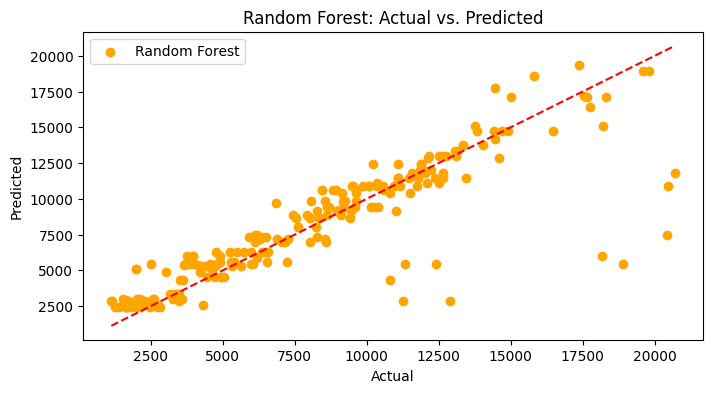
A random forest is a machine-learning model that builds multiple decision trees and merges them together to get a more accurate and stable prediction. The orange dots represent the predicted charges by the random forest model. Again, the red dashed line indicates the ideal scenario where actual and predicted charges perfectly align. This graph provides insights into the predictive performance of the random forest model in estimating insurance charges.

Random Forest:

MSE: 5679860.66

MAE: 1270.65

Graph 2 Random Forest Graph Results of Actual versus Predicted Expenditure



Graph 2 displays a scatter plot where the x-axis represents the actual insurance charges and the y-axis represents the predicted insurance charges by the random forest model. The orange dots in the graph represent individual predictions for each data point, showing how close or far the predictions are from the actual charges. The red dashed line ideally represents a perfect match between actual and predicted values, indicating where the points would lie if every prediction was perfectly accurate. The spread of the orange dots around this line indicates the accuracy and consistency of the random forest model, with closer dots representing more accurate predictions. The Mean Squared Error (MSE) and Mean Absolute Error (MAE) values provide numerical measures of the model's prediction errors, with lower values indicating better predictive performance.

Gradient Boosting: Actual vs. Predicted:

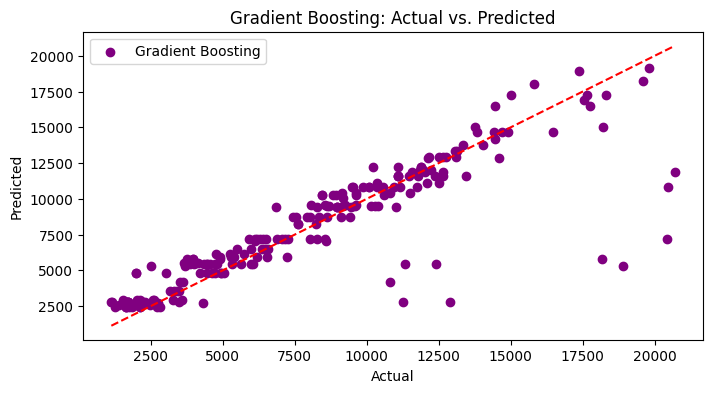
Gradient boosting is a machine learning technique that enhances prediction accuracy by sequentially correcting errors from multiple simple models, similar to improving guesses based on previous mistakes. This method combines the strengths of these models to make more accurate predictions, useful in various applications like finance, customer behavior, and healthcare. This graph depicts the relationship between the actual insurance charges and the charges predicted by the gradient-boosting model. The purple dots represent the predicted charges by the gradient boosting model. Similar to the previous graphs, the red dashed line signifies perfect alignment between actual and predicted charges. By observing deviations from this line, we can assess the predictive accuracy of the gradient-boosting model.

Gradient Boosting:

MSE: 5620075.23

MAE: 1234.97

Graph 3 Gradient Boosting Graph Results of Actual versus Predicted Expenditure



Graph 3 visually represents the performance of a gradient-boosting model in predicting insurance charges. The x-axis is labeled "Actual" and ranges from 2500 to 20000, representing the actual insurance charges. The y-axis is labeled "Predicted" and also ranges from 2500 to 20000, showing the charges predicted by the model. The purple dots scattered across the graph represent individual predictions made by the gradient boosting model for corresponding actual charges. The red dashed line diagonally across the graph represents the line of perfect prediction, where the predicted values would exactly match the actual values. Deviations of the purple dots from this red line indicate the prediction errors of the model.

Feature Importance Analysis:

The importance of predictor variables in determining insurance charges was examined through feature importance analysis. Analysis shows the relative importance scores for two predictor variables, age and smoker status (encoded), across three different models: Decision Tree, Random Forest, and Gradient Boosting.

For the predictor variable "Age":

Decision Tree: Importance score of 0.628

Random Forest: Importance score of 0.614

Gradient Boosting: Importance score of 0.620

For the predictor variable "Smoker (Encoded)":

Decision Tree: Importance score of 0.372

Random Forest: Importance score of 0.386

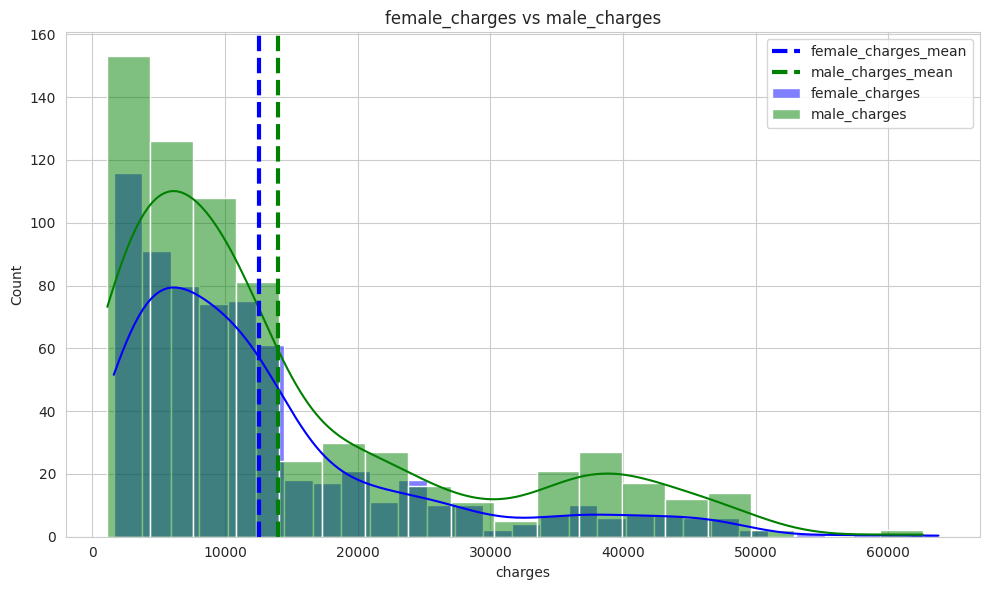
Gradient Boosting: Importance score of 0.380

These scores indicate how significant each predictor variable is in determining insurance charges according to the respective models. The higher the importance score, the more influential the variable is in predicting insurance costs.

The graphs depict the feature importance scores and provide visual insights into the relative significance of age and smoker status in predicting insurance charges across the three machine learning models.

Graph 4 shows the mean of charges for men and women.

Graph 4 Male versus Female Charges



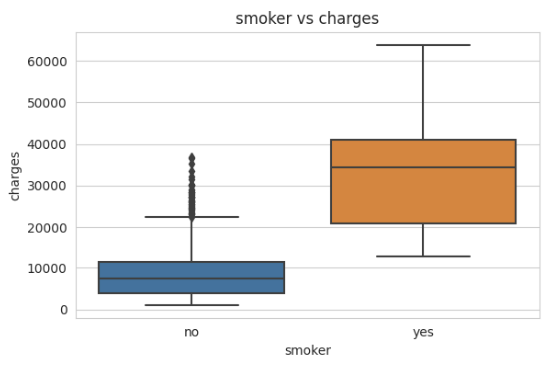
Graph 4 is a comparative histogram overlaid with density curves, illustrating the distribution of charges for females and males. The x-axis represents the amount of charges, ranging from 0 to 60,000, while the y-axis indicates the count of occurrences within each charge range. The histogram bars for female charges are shown in green, and those for male charges are in blue. Each group's distribution is also represented by a smooth density curve in the corresponding color, helping to visualize the overall shape and spread of the data.

Two vertical dashed lines indicate the mean charges for each gender: the blue dashed line represents the mean for male charges, and the green dashed line represents the mean for female charges. These lines help to quickly identify the average charge amount for each gender. From the graph, it is evident that both distributions are right-skewed, meaning that the majority of the charges cluster at the lower end of the scale, with fewer charges occurring at higher amounts. The peak of the female charges occurs slightly earlier than that of the male charges, suggesting that more females have charges in the lower range compared to males. The tails of both distributions extend towards the higher charges, but taper off as the amount increases, indicating that very high charges are less common for both genders.

Overall, this graph provides a clear visual comparison of charge distributions between females and males, highlighting differences in the average charges and the spread of charges across the range.

Graph 5 compares the charges incurred by smokers and non-smokers.

Graph 5 Smoker versus Charges

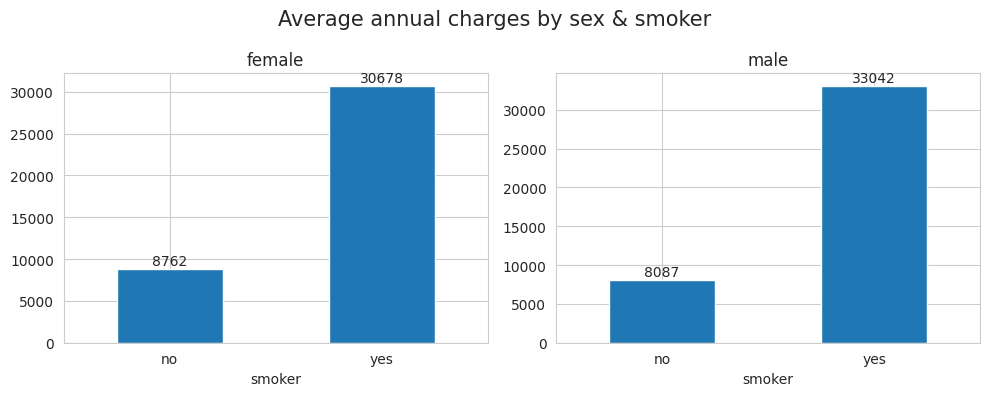


The x-axis categorizes individuals into "no" for non-smokers and "yes" for smokers, while the y-axis represents the charges in a numerical scale that ranges from 0 to 60,000. The box plot for non-smokers (labeled "no") shows a lower median charge compared to smokers. The interquartile range (IQR) for non-smokers is narrower, indicating less variability in the charges among non-smokers. The plot also includes some outliers, represented by individual points above the upper whisker, suggesting that a few non-smokers incurred significantly higher charges than the majority. In contrast, the box plot for smokers (labeled "yes") shows a much higher median charge and the IQR is broader, indicating greater variability in the charges among smokers. This plot also has a higher upper whisker, suggesting that the maximum charges for smokers are substantially higher than those for non-smokers.

Overall, the graph effectively illustrates that smokers, on average, face higher medical charges than non-smokers, and the distribution of charges among smokers is more spread out, indicating a wider range of charges experienced by this group. This visual representation can be useful for analyzing the financial impact of smoking on healthcare costs.

Graph 6 displays a bar graph titled "Average annual charges by sex & smoker," which compares the average annual medical charges based on gender and smoking status.

Graph 6 Female + Smoker versus Male + Smoker



The graph is divided into two sections, one for females and one for males, each with two bars representing smokers and non-smokers. For females, the non-smokers have an average annual charge of approximately $8,762, depicted by a shorter blue bar. In contrast, female smokers face significantly higher charges, around $30,678, shown by a much taller blue bar.

Similarly, for males, the non-smokers have an average annual charge of about $8,087, also represented by a shorter blue bar. Male smokers incur even higher charges than female smokers, with an average of $33,042, indicated by the tallest blue bar in the graph.

The graph clearly illustrates that smoking status has a substantial impact on medical charges, with smokers incurring much higher costs than non-smokers, regardless of gender. Additionally, it shows that male smokers tend to have slightly higher medical charges than female smokers. The y-axis of the graph is scaled from 0 to 35,000, allowing for a clear visual comparison of the charges between the different groups.

The results indicate that age consistently emerges as the most influential predictor of insurance charges, with higher importance scores compared to smoker status in all models. This underscores the significance of age demographics in determining insurance premiums. The findings contribute valuable insights to insurance pricing strategies and risk assessment methodologies, highlighting the need to consider demographic factors when pricing insurance policies. We also see the average count as well as the amount spent by males and females with men being higher in both expenditure and usage. Furthermore, we see the difference between smokers and non-smokers.

**Conclusion**

The analysis conducted in this research paper sought to examine the factors influencing personal and household insurance expenditure and explore the predictive capabilities of machine learning models in estimating insurance charges. By leveraging a dataset encompassing demographic and lifestyle variables, alongside advanced modeling techniques, valuable insights were gleaned regarding the determinants of insurance costs and the performance of predictive models.

The first key finding of the paper is the importance of age. Across all machine learning models employed (decision tree, random forest, and gradient boosting), age consistently emerged as the most influential predictor of insurance charges. The feature importance analysis revealed that age exerted a stronger influence on insurance costs compared to other factors, underscoring the significance of age demographics in insurance pricing. Secondly, we have the role of smoking status. While smoker status also contributed to predicting insurance charges, its importance was comparatively lower than age in all models. Nonetheless, smoking status remained a significant predictor, indicating higher insurance costs for smokers compared to non-smokers. Thirdly when looking at gender, men consistently have more count( times they use insurance) versus women with the average amount also being higher for men. Finally, smoker with gender does not show a big difference between the genders however when looking exclusively at smokers. A large disparity is seen.

The findings of this study have several implications for insurance companies, policyholders, and regulators. Firstly, the emphasis on age as a primary determinant of insurance charges highlights the importance of demographic factors in risk assessment and pricing strategies. Insurers should consider age demographics when pricing insurance policies to ensure accuracy and fairness.

Furthermore, the inclusion of smoking status as a predictor underscores the impact of lifestyle factors on insurance costs. Policyholders are encouraged to adopt healthy behaviors and lifestyle choices to potentially lower their insurance premiums. Regulators play a crucial role in ensuring compliance with minimum coverage requirements and promoting fairness in insurance pricing. The insights provided by this research can inform regulatory policies aimed at protecting consumers and maintaining market stability.

While this study sheds light on the factors influencing personal insurance expenditure, there are avenues for further exploration. Future research could delve into the interaction effects between demographic and lifestyle variables, as well as the incorporation of additional data sources such as telematics data and social media activity. Additionally, investigating the implications of emerging technologies such as artificial intelligence and blockchain on insurance pricing and risk assessment could provide valuable insights into the evolving landscape of the insurance industry.

In conclusion, understanding the determinants of insurance charges is essential for ensuring financial security and promoting fairness in insurance markets. By leveraging advanced analytical techniques and interdisciplinary approaches, researchers can continue to advance our understanding of insurance pricing dynamics and contribute to the development of innovative solutions for risk management.

**Work Cited**

Teertha. (n.d.). US Health Insurance Dataset [Data file]. Kaggle. https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset

Jacobs, P., & Claxton, G. (2018). "Determinants of consumer engagement in health insurance

exchanges: Evidence from emerging economies." Health Economics, 27(5), 876-889

Victor Chang, Meghana Ashok Ganatra, Karl Hall, Lewis Golightly, Qianwen Ariel Xu,

An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators,Healthcare Analytics,Volume 2,2022,100118,ISSN 2772-4425,

https://doi.org/10.1016/j.health.2022.100118.

Blier-Wong, C.; Cossette, H.; Lamontagne, L.; Marceau, E. Machine Learning in P&C Insurance: A Review for Pricing and Reserving. Risks 2021, 9, 4. https://doi.org/10.3390/risks9010004

Eling, Martin & Kraft, Mirko. (2020). The impact of telematics on the insurability of risks. The Journal of Risk Finance. ahead-of-print. 10.1108/JRF-07-2019-0129.

Morid, M. A., Kawamoto, K., Ault, T., Dorius, J., & Abdelrahman, S. (2018). Supervised learning methods for predicting healthcare costs: Systematic literature review and empirical evaluation. AMIA Annual Symposium Proceedings, 2017, 1312–1321. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977561/>

Hossen, Sazzad. (2023). Medical Insurance Cost Prediction Using Machine Learning. 10.13140/RG.2.2.31456.25604.

Morid MA, Kawamoto K, Ault T, Dorius J, Abdelrahman S. Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation.

Ugochukwu Orji, Elochukwu Ukwandu,

Machine learning for an explainable cost prediction of medical insurance,

Machine Learning with Applications,

Volume 15,

2024,

100516,

ISSN 2666-8270,

<https://doi.org/10.1016/j.mlwa.2023.100516>.

Kaushik K, Bhardwaj A, Dwivedi AD, Singh R. Machine Learning-Based Regression Framework to Predict Health Insurance Premiums. Int J Environ Res Public Health.

Stanciole, A. Health Insurance and Lifestyle Choices: Identifying *Ex Ante* Moral Hazard in the US Market. *Geneva Pap Risk Insur Issues Pract* **33**, 627–644 (2008). https://doi.org/10.1057/gpp.2008.27

1. Kaggle, 2020, US Health Insurance Dataset, <https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset/data>, 02/07/2024 [↑](#footnote-ref-1)