BA305 Insurance Pricing Model

Utilizing Extreme Gradient Boosting to Develop Personalized Health Insurance Forecasts

By: Damian Grochowski, Ryan Whitehouse, Ole Nagorsen, Michael Lee, Reid Brock

1 Introduction

1.1 The Issue of Health Insurance

The focus of our project is to better understand how specific measurable health and lifestyle attributes, such as smoking, BMI, age, and location, affect a person's insurance prices. Insurance pricing is, fundamentally, a risk-assessment problem. There are many different customer attributes that determine their expected medical expenditures. Our analysis and project seek to determine which customer attributes are the strongest predictors of insurance costs.

1.2 Why Cost Transparency is Important

This problem is both economic and social; rising healthcare and insurance costs affect millions of consumers, with more than 100 million people in the United States being in debt due to rising medical costs.[1] In 2011, obesity increased the total yearly individual healthcare expenses by $1360, while smoking increased the total yearly individual healthcare expenses by $1046.[2] Insurance companies have a history of utilizing a risk classification technique. Risk classification is an efficient process that insurance companies use. This process classifies individuals with similar traits (smoking habits, BMI, etc.) into distinct risk categories.[3] Individuals with similar characteristics are combined and are charged similar premiums. This practice is more efficient because insurance companies have distinct premiums and costs they charge depending on which group a customer falls into. Specifically, insurance companies classify smoking risks into two or three main categories: non-smokers, smokers, and heavy-smokers. In fact, smokers are charged up to four times higher insurance premiums compared to non-smokers.[4]

Understanding which traits have a strong impact on higher insurance prices helps not only insurers make fairer and transparent pricing decisions, but also helps consumers better understand the financial impact of their behaviors. With the knowledge of what their insurance price is projected to be, consumers can gain a better understanding of whether they are overpaying for their health insurance. Additionally, the ability to gauge estimated costs could allow consumers to budget for them and potentially avoid going into medical debt.

1.3 The Techniques We Used

---

[1] "< the Sunday Story: The Unbearable Weight of Medical Debt." NPR, December 10, 2023. https://www.npr.org/transcripts/1198909604.

[2] An R. (2015). Health care expenses in relation to obesity and smoking among U.S. adults by gender, race/ethnicity, and age group: 1998-2011. *Public health*, *129*(1), 29–36. https://doi.org/10.1016/j.puhe.2014.11.003

[3] Ibid.

[4] Insurer Perspectives on Smoking Risks, accessed November 6, 2025, https://www.marshmclennan.com/assets/insights/publications/2019/oct/20191030_Insure_Perspectives_on_Smoking_Risks.pdf.

We first analyzed the distributions and relationships between all the variables in the dataset to specifically identify non-linear relationships, interactions between variables, which variables are strong predictors of insurance price, and which variables are weak predictors. Our team then tested multiple regression models, including linear regressions, decision trees, random forests, and new models such as Lasso/Ridge regression, Gradient boosting, and XGBoost, utilizing 5-fold cross-validation to ensure reliable performance estimates for all models. Our modeling strategy was motivated by our initial findings: many non-linear patterns, specific relationships between variables, and a very significant jump in insurance prices for smokers.

1.4 How Our Analysis Will Help Customers

This analysis provides insurance consumers with two core benefits: A high-accuracy model for predicting insurance prices, and a clear understanding of which variables have the greatest impact on insurance costs. As an economic benefit, this may allow customers to avoid being overcharged for their health insurance and budget for potential costs. As a health benefit, our analysis may uncover certain traits that have an outsized impact on insurance prices and indicate a behavior that is classified as "risky" by insurance firms. Factors that can be managed by customers and increase a customer's insurance costs dramatically may be ones that individuals can change to reduce their risky behavior.

2 Data Preprocessing

2.1 Source of the Data

Our dataset for this project is referred to as the *Medical Cost Insurance Dataset*.
Source Link: https://www.kaggle.com/datasets/kanzariachref/medical-insurance-cost-dataset

Although this dataset does not specifically specify the exact year this data was collected, the structure still reflects and resembles realistic U.S. health insurance data. This dataset is well-documented and primarily used for learning purposes, making it a suitable dataset for our modeling.

2.2 Population and Sampling

The original dataset contains 2,772 observations and 7 variables. Each row in the dataset represents an individual health insurance policy holder and the annual medical charges billed to their respective insurance. All seven variables are a combination of demographic and health-related information: the age (in years), the gender (male or female), the BMI (body mass index, which is measured as a weight in kilograms divided by the height in meters squared), the number of children covered by the insurance, smoking status (non-smoker or smoker), and the individual's region of residence (Northwest, Northeast, Southwest, Southeast). The target

variable in this dataset is the annual medical insurance price in USD. The dataset contains both categorical and numerical features. The gender variable is coded as 1 (male) and 2 (female). Smoking status is 0 (non-smoker) and 1 (smoker). Region is 1 (Northwest), 2 (Northeast), 3 (Southwest), and 4 (Southeast). The dataset provides a holistic overview of an individual's demographic information, health-risk habits/indicators, and their respective insurance expenses.

2.3 Preprocessing/Data Cleaning

The first step in our data cleaning process was to identify any missing or null values. Our team noticed that the Age and Smoking columns contained the character "?," implying missing or unknown values for those variables. Our team then summed up the total amount of null values in the Age variable, which was 4, and in the Smoker variable, which was 7, adding up to 11 null values in the dataset. We created a variable "row_with_null" that contains an index of all the rows with a "?". Our team then dropped all the rows with those missing values.

Both the Age and Smoker variables were stored as objects in the dataset, so our team converted Age into an integer and Smoker into a binary classification: 0 represents a non-smoker and 1 represents a smoker. After conducting a final consistency check of our dataset, checking for any missing columns or rows, we ended up with a total of 2761 entries with 7 variables. Variables such as sex and region were also string-based, so our team applied One-Hot encoding to region, creating a separate binary column for each region so that the model can correctly interpret the data without assuming a numerical order with the variables (1 and 2 for sex for running models).

2.4 Identifying Correlations

After preprocessing and cleaning the dataset, our team created a correlation matrix to identify key correlations between all the variables. We noticed immediately that there is a strong positive correlation of 0.790 between smoking and insurance price. This not only implies that if a person smokes, then there is an increase in their insurance price, but also suggests that there is a big discrepancy in insurance price between smokers and non-smokers. The Age variable had a moderate positive correlation of 0.297 with insurance price, suggesting that older individuals tend to have higher insurance charges. The BMI variable had a weaker, but still noticeable, positive correlation of 0.1999 with insurance price. The other variables, number of children, gender, and region, have smaller correlations with insurance price, suggesting these variables have less predictive power on their own.

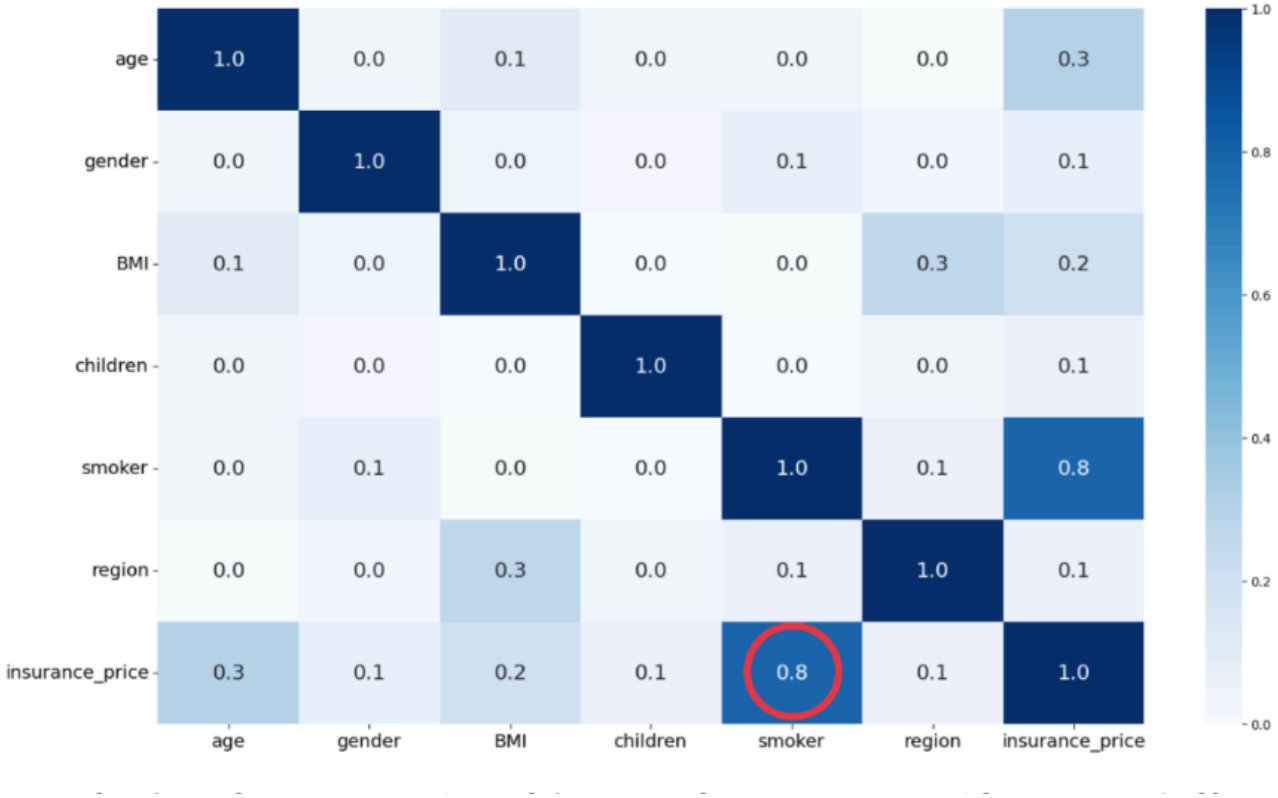

Figure 1: Correlation Matrix Between All 7 Variables

2.5 Addressing Outliers

Our team also noticed that there are several individuals with very high insurance charges: there are 14 individuals with insurance prices exceeding $50,000, and only 6 with insurance prices over $60,000. While at first this might be concerning, a deeper analysis of these individuals confirmed that most of these individuals were smokers, older, and had a high BMI. Therefore, our team determined that these values are not anomalies but rather extreme, plausible medical cases. In outside research, we found that even a "moderate, less extreme stay" in the ICU can cost over $4,300 a day, suggesting that extreme cases can skyrocket an individual's expenses well above $50,000.[5] Given this context, our team decided not to remove these extreme cases. Additionally, we also realized that if we were to remove these variables entirely, it would distort the true distribution of insurance price, specifically for high-cost individuals. We intentionally kept all individuals with high insurance prices.

3. Methodology

[5] "Critical Care Statistics," Society of Critical Care Medicine (SCCM), accessed December 1, 2025, https://www.sccm.org/communications/critical-care-statistics.

3.1 Ways We Looked at the Data

The main goal was to analyze how different variables affect an individual's insurance price. We first looked at each variable individually, examining how their age, gender, level of BMI, smoking status, and region they reside affect the insurance price. We used scatter plots to visualize the relationship between the predictor variable (insurance price) and other numerical variables, such as age and BMI. We utilized boxplots to visually describe the distribution of different variables. All of our visualizations gave us critical insights into how prevalent non-linearity is in our dataset. This guided our decision to utilize more flexible models.

3.2 Modeling Approach

We trained and tested 10 different models, the results of each we put into the table below.

Model	MAE	RMSE	R^2
Baseline	9132.437	12156.556	0
KNN Regressor	1272.606	4189.186	0.879
Linear Regression	4181.771	6063.585	0.75
Polynomial Regression	2870.173	4766.909	0.845
Ridge Regression	4182.268	6063.584	0.75
Lasso Regression	4182.846	6063.102	0.75
Decision Tree Regressor	**607.44**	2857.938	0.944
Random Forest Regressor	1235.035	2567.749	0.955
Gradient Boosting Regressor	1273.326	2635.217	0.952
Hist Gradient Boosting Regressor	1528.816	2784.487	0.947
XGBoost Regressor	643.087	**2234.407**	**0.966**

Table 1: Cross Model Comparison based on MAE, RMSE, and R^2

Our modeling strategy includes a combination of simple models and a few new, more complex models to fully and accurately measure the complexity of this dataset. To ensure a fair comparison, three metrics were used: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). Each of these metrics captures a different quality of our models. The MAE metric measures the average insurance price, the dollar difference between predictions and the actual insurance prices. A low MAE score is ideal because it signals that the model, on average, predicts closer to the true dollar amount. The RMSE metric penalizes larger errors, due to first squaring the error, significantly increasing the values of errors compared to MAE. Similarly, a low RMSE score is better. Finally, the R^2 metric measures how much variance in the insurance prices can be explained by a specific model. Ideally, a good model is one with a high R^2 value, values closer to 1, because it indicates the model is more accurate. When

choosing the best model, our team realized that relying on only one metric is not ideal. Different models were considered "better" based on different metrics, so we evaluated each model based on its MAE, RMSE, and R^2 values, giving us a more comprehensive and reliable overview.

3.3 Models and Why We Used Them

In total, we tested 10 different models. The very first model we created is the baseline model, comparing the mean insurance price with each person's real insurance. The baseline model served as the comparison benchmark for all subsequent models: all other models should beat the baseline model's RMSE and MAE scores of 12,156 and 9132, respectively. Once we had this baseline, we trained and tested the different regression models to compare them to it. To do this, we created a function that utilized k-fold cross-validation on any model of our choice. The function requires an input of the model we are using, the X and Y variables, and the number of splits. For the purpose of our data, we used 5 folds, which is the standard for datasets of our size. The reason we used cross-validation to evaluate our models is that it provides a more reliable and robust estimate of model performance compared to other methods, such as holdout.

For each of our regression models, we defined a target variable Y (insurance_price) and the features X (everything else). Our linear models utilized a pipeline to define our models, to account for any one-hot encoding that was necessary, and for ease of use. For our non-linear/tree models, we utilized GridSearchCV to find the most optimal parameters before testing them. We tested familiar models such as linear regression, decision tree, KNN regressor, and random forest. The first new model we tested is called a polynomial regression. The purpose is to capture curved, non-linear relationships, which our dataset highlights. This resulted in an improved RMSE and R^2 score, 4766.909 and 0.845, proving the non-linear relationships in our dataset. Both the Lasso and Ridge regressions reduce overfitting in the dataset and can handle collinearity. The results of both were very similar to our linear regression output, which is why we determined non-linear models to be more accurate. We also tested a Gradient Boosting Regression, Histogram Gradient Regression, and a KNN regression. Gradient Boosting builds a sequence of decision trees, where each new tree corrects the mistakes from the previous one, focusing on non-linear patterns. A Histogram Gradient Regression is much faster by grouping continuous features into bins/histograms.

The final model we tested is called the XGBoost regressor. XGBoost regressor starts with a simple prediction mean, then builds trees sequentially, corrects any errors from the previous tree, and utilizes capabilities such as pruning to prevent overfitting. The key advantages of using XGBoost over other models are that it is more accurate at handling non-linear relationships and is incredibly fast and accurate. Overall, the XGBoost Regressor gave us an RMSE score of 2234.407, an MAE score of 643.087, and an R^2 of 0.966, outperforming the previous models by a wide margin. The XGBoost regressor also determined the feature importances, showcasing

the smoker (0.885) variable having the highest impact on insurance prices, with BMI (0.040) and age (0.034) having a moderate effect.

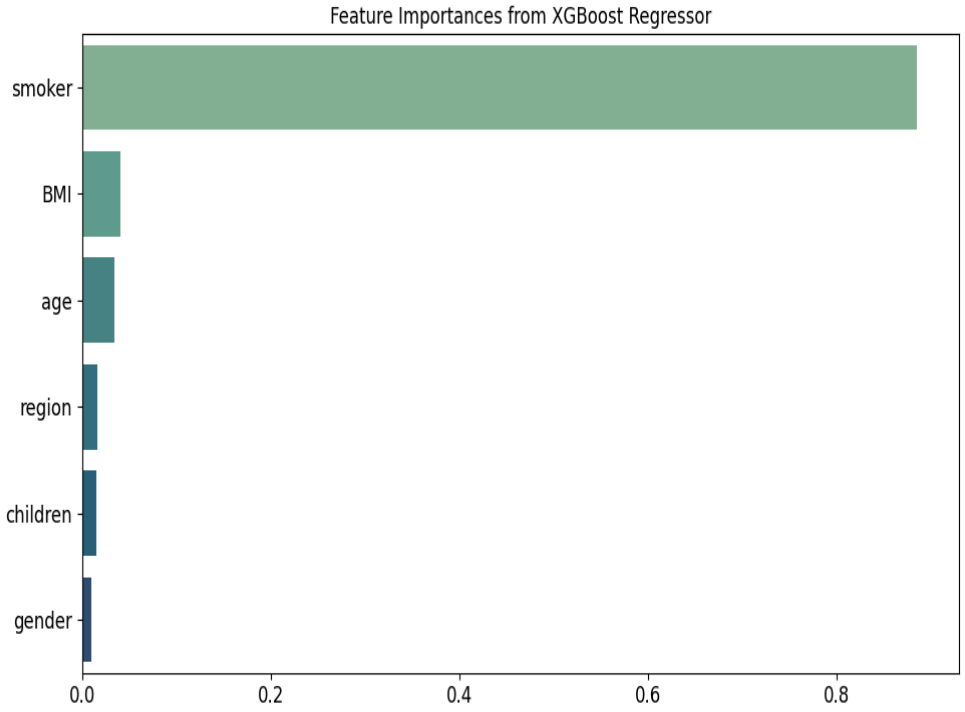

Figure 2: XGBoost Feature Importance

4 Findings

4.1 Overall Findings

Concluding the comparison of models, we found that XGBoost had the best overall performance when it came to predicting insurance costs. This is evident by the model having the largest R^2 value at 0.966 and the lowest RMSE at 2234.407. The nature of these values highlights how much more accurate XGBoost is when compared to other models. Under this model, we analyzed the most important features in determining insurance pricing. Our research discovered that when determining insurance costs, smoking is a strong factor in the final calculation. This is highlighted by XGBoost discovering that smoking determines 88.5% of an individual's insurance costs. Furthermore, under a subsample of entries (gathered from the table below), the average cost of insurance cost increase for smokers is $19,768.84. Between the findings of XGBoost and the significant average increase in costs, it is clear that smoking is the key determinant of insurance pricing.

Sample_ID	Mock_Name	Original_Smoker_Status	Predicted_Non_Smoker_Price	Predicted_Smoker_Price	Price_Increase_If_Smoker
1362	Ole	0	**$8,965.83**	**$19,895.90**	**$10,930.07**
2543	Reid	0	**$4,530.18**	**$17,303.95**	**$12,773.77**
2229	Michael	0	**$11,731.44**	**$23,315.32**	**$11,583.88**
2048	Ryan	0	**$3,945.26**	**$33,153.18**	**$29,207.92**
446	Damian	0	**$4,685.57**	**$39,034.12**	**$34,348.55**

Table 2: Random Sample Smoking Insurance_Price Increase

4.2 How this Helps Patients

Our model helps patients by giving them clear, data-driven insight into how their personal characteristics influence their insurance costs, making an otherwise unclear pricing system far more understandable and actionable. By showing that smoking is the strongest driver of premiums, followed by BMI and age, it allows patients to see which factors meaningfully affect what they pay and which factors, such as gender, number of children, or region, contribute very little. This transparency enables patients to estimate what their insurance should reasonably cost, identify when a quoted premium seems unusually high, and understand why insurers place them into specific risk categories. Because two of the major cost drivers, smoking and obesity, are within a patient's control, the model also provides strong feedback on which health decisions can significantly reduce long-term insurance costs and reflect meaningful differences in overall health risk. The model's ability to benchmark expected premiums helps patients detect possible overcharging, compare plans with a more informed baseline, and plan for future expenses based on all of their personal factors. Ultimately, it empowers patients to make more informed decisions about coverage, budgeting, and lifestyle choices that affect their long-term financial and medical outcomes.

5 Areas of Improvement and Failed Model Attempt

While our model is accurate to insurance pricing, it is specifically built around a few key factors. Expanding our research to other health factors would likely increase the accuracy of our predictions. Specific factors such as family health history, underlying health conditions, and diet could all be important factors that could make our model more accurate, yet this would require significant resources towards gathering a large amount of sufficient data.

Another area of improvement would be to obtain a larger sample of medium to high-risk health insurance subscribers for our dataset. Such an addition would allow us to effectively implement a two-part model. We first have a classification element that groups users into low, medium, and high cost, then train 3 distinct regression models based on each price group. Our first draft of the model used a random forest classification and a XGBoost regression creating a model able to achieve a MAE of 1115 and R^2 of 0.876 making it a strict downgrade to our regular XGBoost model, but we believe this discrepancy was due to the low number of users in

medium and high price range 397 and 339 respectively in reference to the low cost class having 2025 entries.

Due to the lack of data on non-low-risk subscribers, additional models for the hypothetical bucketing solution are impossible. However, by collecting a significant amount of data for medium and high risk subscribers, these models would be feasible. Creating a more accurate and tailored model that allows for a better prediction of insurance subscriber costs. Our ideal final step would be to create a user interface where individuals could input their own personal health information and receive a projected health care price.

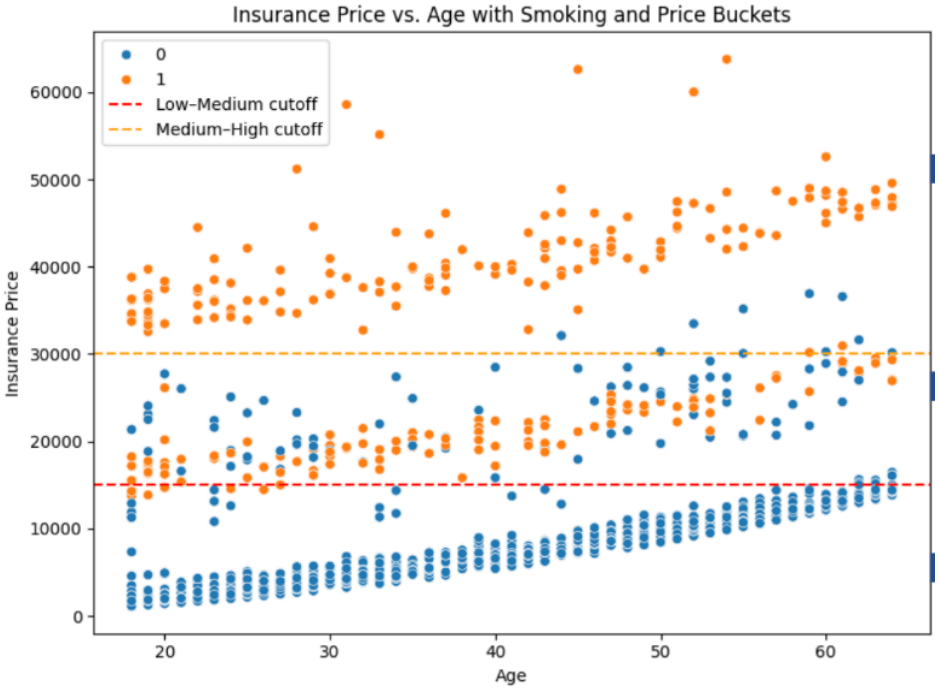

Figure 3: Breakdown of price profiles used within the two part model

6 Conclusion

6.1 Our Approach

We began the project following the step by step instructions left in class, visualizing our model, getting to know our data and searching for unique insights, and once we saw the age price diagram we had our big breakthrough. It was a multi-linear model and this breakthrough informed our future model construction. Focusing on tree based models instead of linear models. Then we began to expand our research outside of class content to find and create superior

models. Leading to the discovery of XGBoost and experimentation with a two part model which unfortunately fell apart due to lack of data points.

6.2 Final Thoughts

The XGBoost model returned the most accurate predictive results and assigned the largest importance towards insurance cost to the smoking factor. We believe that we were successful in building an accurate and interpretable insurance pricing model that predicts a patient's projected cost based on their individual factors. Our final XGBoost model not only predicts charges with high accuracy but also clearly shows that smoking is by far the strongest cost driver, with BMI and age playing secondary roles and variables like gender, region, and number of children contributing relatively little. With a larger set of data and more variables, this model could be an even more effective method for individuals to predict their insurance costs and make impactful decisions in regards to their insurance provider, budget, and health.

7 Appendix

7.1 Appendix A: Explanation of the features of the dataset

Feature	Description
Age	The age of the patient between with a range of 18 to 64.
BMI (Body Mass Index)	The Body Mass Index of patients (their weight divided by the square of their height) with a range of 15 to 53.
Gender	The gender of the patient (either male or female).
Number of Children	The patient's total number of dependents that are covered by their insurance.
Region	The patient's geographic region . These include the Northwest, Northeast, Southwest, and Southeast.
Smoker	Whether the patient self-identifies as a smoker or a non-smoker.
Insurance Price	The total annual medical charges billed by the patient's insurer in U.S. dollars.

Bibliography

Barlow P;McKee M;Reeves A;Galea G;Stuckler D; "Time-Discounting and Tobacco Smoking: A Systematic Review and Network Analysis." International journal of epidemiology. Accessed November 12, 2025. https://pubmed.ncbi.nlm.nih.gov/27818375/.

"Critical Care Statistics." Society of Critical Care Medicine (SCCM). Accessed December 1, 2025. https://www.sccm.org/communications/critical-care-statistics.

Insurer Perspectives on Smoking Risks. Accessed November 6, 2025. https://www.marshmclennan.com/assets/insights/publications/2019/oct/20191030_Insure_Perspectives_on_Smoking_Risks.pdf.

Liber AC;Drope JM;Graetz I;Waters TM;Kaplan CM; "Tobacco Surcharges on 2015 Health Insurance Plans Sold in Federally Facilitated Marketplaces: Variations by Age and Geography and Implications for Health Equity." American journal of public health. Accessed November 5, 2025. https://pubmed.ncbi.nlm.nih.gov/26447913/.

"Risk Pooling: How Health Insurance in the Individual Market Works." Actuary.org. Accessed November 5, 2025. https://actuary.org/risk-pooling-how-health-insurance-in-the-individual-market-works/.

"< the Sunday Story: The Unbearable Weight of Medical Debt." NPR, December 10, 2023. https://www.npr.org/transcripts/1198909604.

***AI was used in part for the generation of the code, in particular for the two-part model**