# A Generalized Linear Model to Predict Health Insurance Cost

## Introduction

This project details the development of a predictive model for health insurance charges using a publicly available dataset. As an aspiring actuary, I've focused on creating a model that is not only accurate but also interpretable, allowing for a clear understanding of the factors that drive costs. Using a Generalized Linear Model (GLM), this analysis identifies key risk factors such as smoking, age, and BMI, and provides a framework for predicting individual healthcare expenses. The project demonstrates proficiency in data exploration, statistical modeling, and communicating technical findings, which are core skills for risk analysis in the insurance industry.

## Data

The insurance.csv dataset, which contains 1,338 records of health insurance data, was used for this analysis. The dataset was obtained from Kaggle, where it was published by Mosap Abdel-Ghany. While the specific origin of the data is not stated on the Kaggle page, the file is widely used in data science tutorials and projects for its clear structure and relevant features.

The dataset contains seven features:

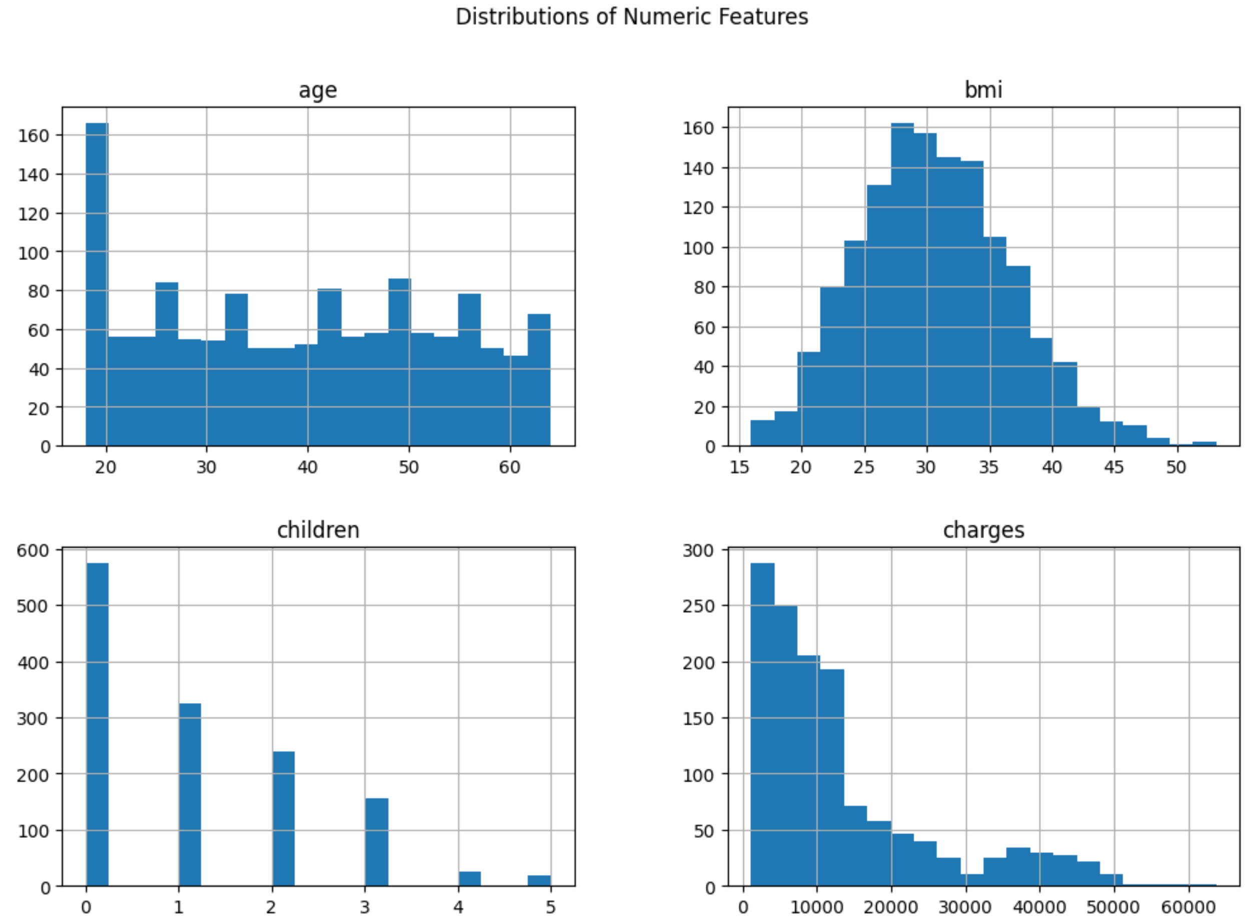
| Feature | Description | Type |
|---|---|---|
| age | Age of primary beneficiary | integer |
| sex | Gender of beneficiary | string (female/male) |
| bmi | Body Mass Index | decimal |
| children | Number of children covered by health insurance | integer |
| smoker | Smoking status of the beneficiary | string (yes/no) |
| region | Residential region in the US | String (northeast, northwest, southeast, southwest) |
| charges | Medical insurance cost billed to the beneficiary | decimal |

The dataset is complete, with no missing values in any feature.

## Exploratory Data Analysis

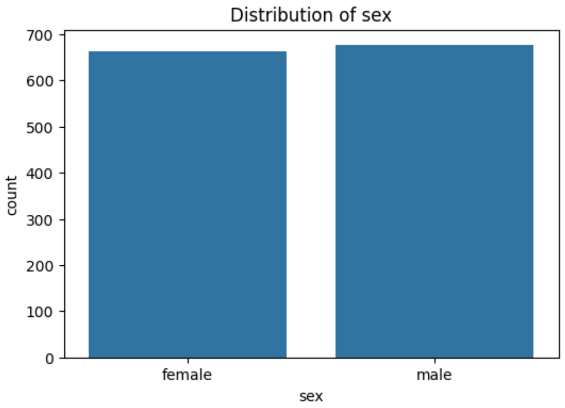
*Univariate Analysis*

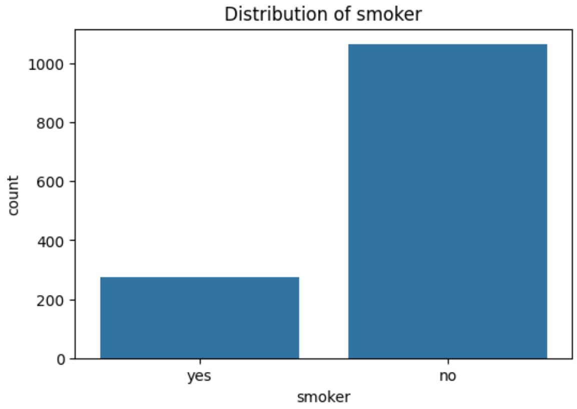
Histograms and boxplots for age, bmi, children, and charges were generated. The charges variable showed a highly skewed distribution, with a long tail to the right, indicating that a small number of individuals incur very high medical costs.

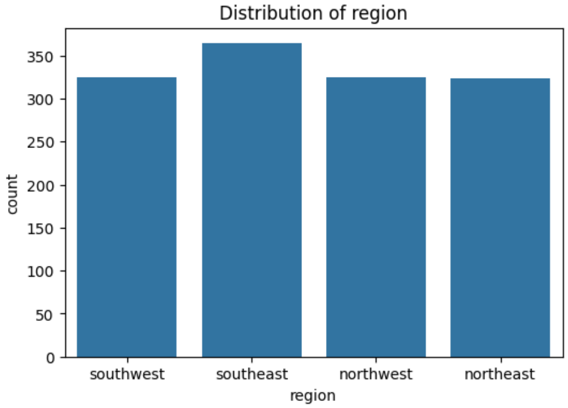


The boxplots confirmed this skewness and highlighted several high-value outliers. In contrast, age and bmi had more symmetric distributions, though bmi had some notable outliers at the high end.

Count plots for sex, smoker, and region showed a nearly even split between males and females, and a large majority of individuals who are non-smokers (~80%). The data was fairly evenly distributed across the four U.S. regions.



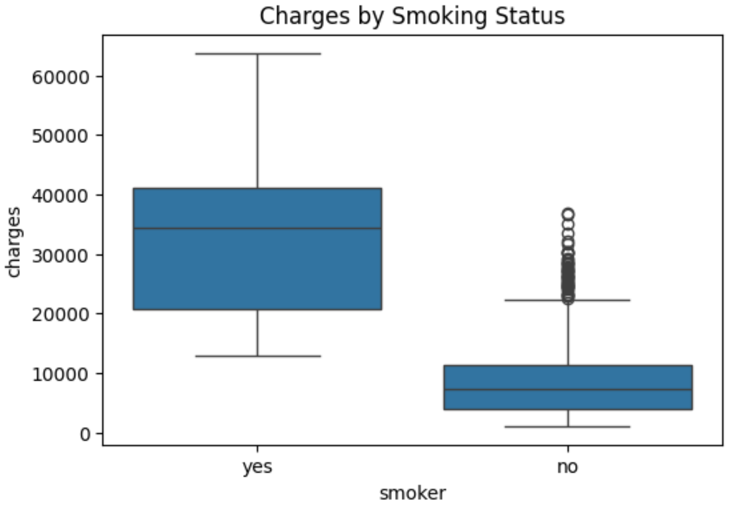




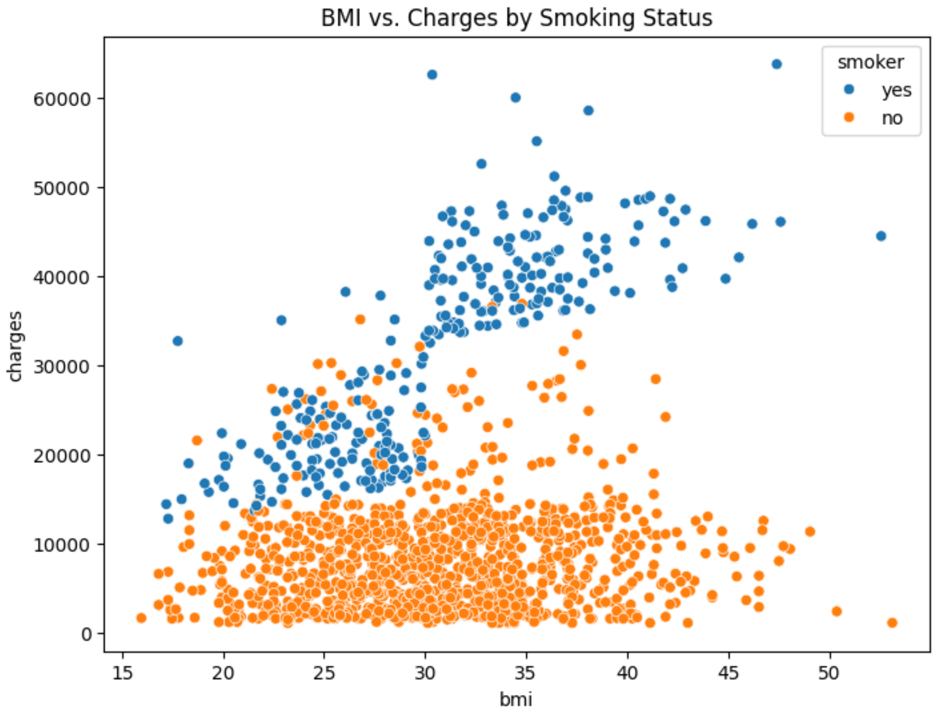
*Bivariate Analysis*

This analysis focused on how each feature relates to the target variable, charges.

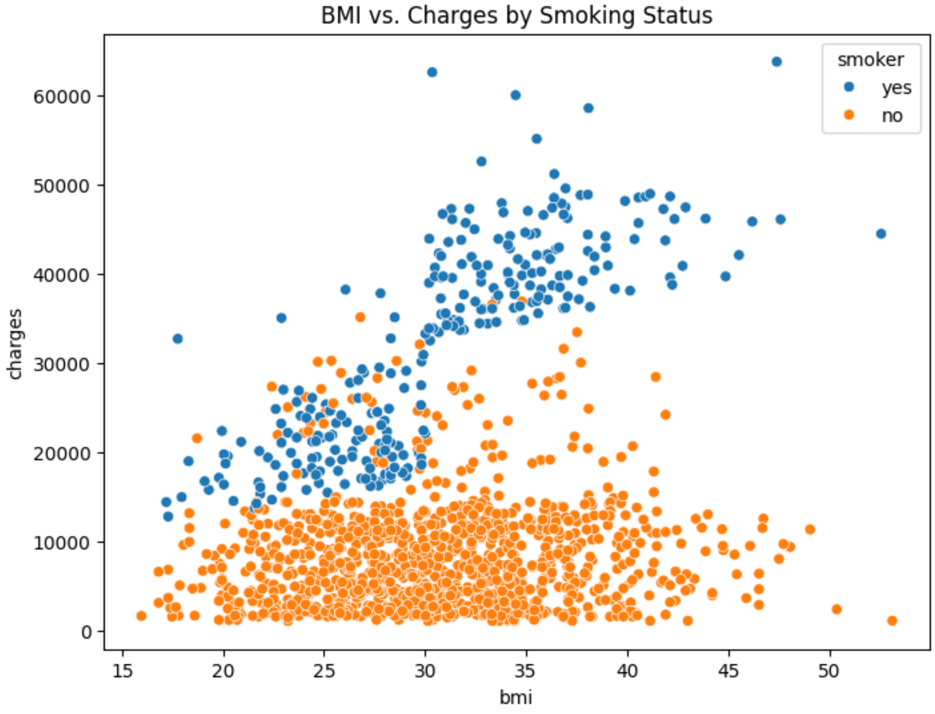
*Smoker vs. Charges:* The most striking relationship was between smoking status and charges. The boxplot and average charges calculation showed that smokers' costs are significantly higher than non-smokers' costs, a difference of over $23,000 on average.



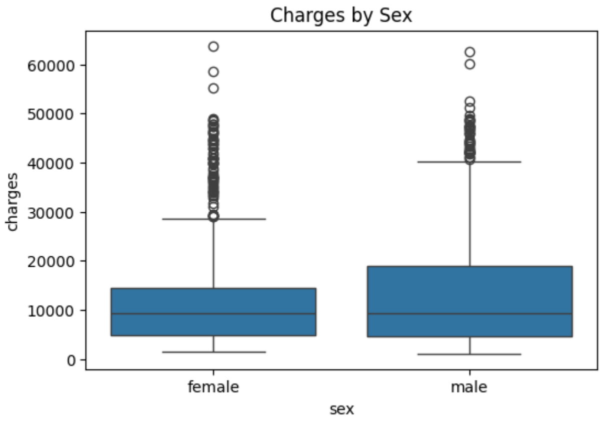
*Age vs. Charges:* A scatter plot of age vs. charges showed a clear positive correlation, which is more pronounced for smokers, indicating that older smokers tend to have the highest charges.

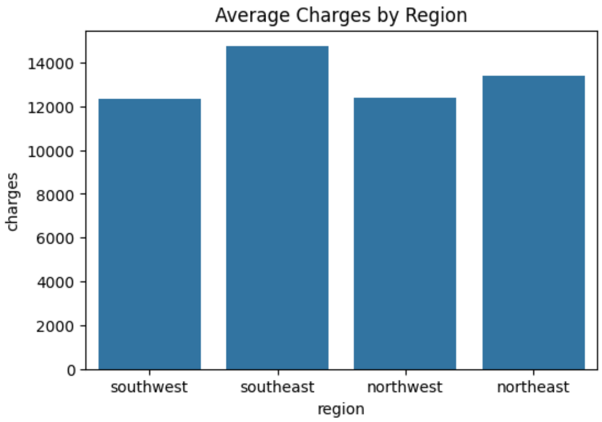


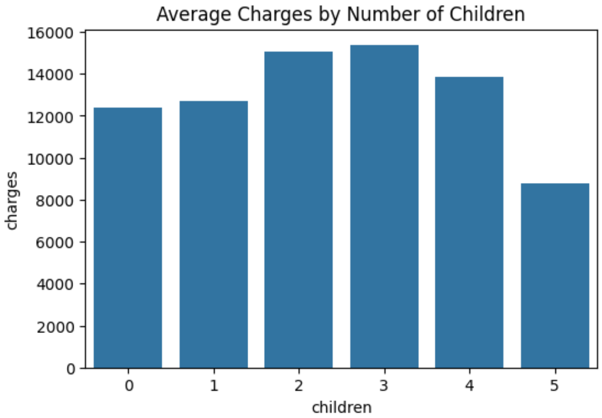
*BMI vs. Charges:* Similarly, bmi showed a positive correlation with charges. This relationship is also heavily influenced by smoking status, with a stronger upward trend observed among smokers.



*Other Features:* region and children also showed a positive relationship with charges, though their impact was less pronounced compared to smoking and age. The sex variable, however, showed no significant difference in average charges between males and females.

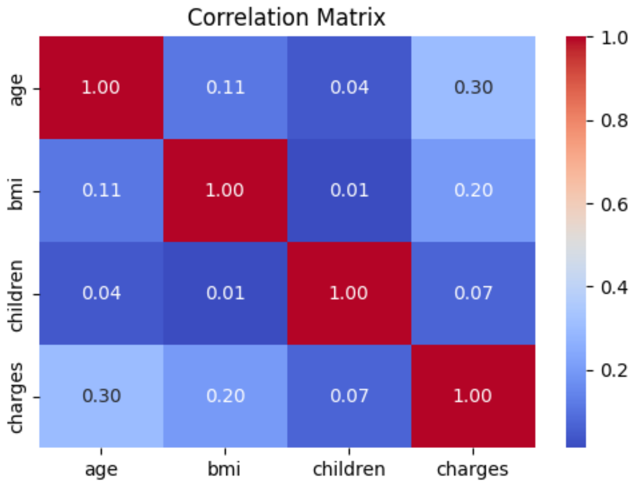






## *Correlation Analysis*

A heatmap was used to visualize the pairwise correlations between the numeric variables (age, bmi, children, charges). The heatmap confirmed the findings from the bivariate analysis: age and charges have a moderate positive correlation, while bmi and charges have a weaker positive correlation. The correlation between children and charges was almost negligible. The heatmap provided a quick, comprehensive overview of the linear relationships within the numeric data.



*Pareto Principle and Cost Skewness*

A quick analysis was performed to see if the Pareto Principle (or 80/20 rule**)** applies to the distribution of healthcare costs. By calculating the cumulative sum of charges, it was found that the highest-cost patients drive a disproportionate amount of the total spend. Specifically, the top 5% of patients accounted for 17.58% of all costs.

## Generalized Linear Model (GLM) Approach

A Generalized Linear Model (GLM) was selected for this project for its suitability in handling the specific characteristics of the insurance data. While a simple linear regression model is a common starting point, it assumes that the target variable is normally distributed. As the EDA revealed, our charges variable is highly skewed and strictly positive, which violates this core assumption.

A GLM provides a flexible framework that extends the linear model to accommodate non-normal distributions. It does this by using a link function and a family of distributions that are better suited to the data. This approach is standard practice in actuarial science and is often used for pricing and reserving models because it produces results that are both robust and easily interpretable.

To properly model the insurance charges, the following specifications were chosen for the GLM:

- **Distribution Family**: The Gamma family was chosen because it is ideal for modeling variables that are continuous, strictly positive, and right-skewed, such as insurance costs.
- **Link Function**: The log link function was used to transform the relationship between the predictors and the target variable. This ensures that the predicted charges will always be positive ($\exp(\beta_0 + ...)$ will always be > 0), and it allows the model to interpret the coefficients as multiplicative effects on the charges, which is intuitive for this type of data (e.g., a 10% increase in charges).

The model was fit using the statsmodels library in Python with the formula:

charges ~ age + bmi + children + sex + smoker + region

Categorical variables (sex, smoker, region) were automatically encoded by the library to be included in the model, using a reference category for comparison (e.g., sex[T.male] compares males to the reference category of females).

**Model Results and Interpretation**

The GLM was highly effective in identifying the key drivers of health insurance charges. The model's strong performance is evidenced by a Pseudo R-squared of 0.6833, indicating that it explains a substantial portion of the variance in charges. A key advantage of using a log link function is that the coefficients can be interpreted as multiplicative effects on the charges. This means we can directly understand how each variable proportionally impacts costs.

The most significant finding is the profound impact of smoking. The coefficient for smoker[T.yes] is 1.5003, which translates to smokers' charges being approximately 4.48 times higher ($e^{1.5003}$) than non-smokers', holding all other factors constant. This finding confirms smoking as the single most important predictor of insurance costs in the dataset.

Other factors also have a significant, albeit smaller, effect. For each additional year of age, charges are predicted to increase by about 2.9% ($e^{0.0286}$), and each one-unit increase in BMI is associated with a rise of about 1.4% ($e^{0.0141}$). Similarly, each additional child is predicted to increase charges by around 8.4% ($e^{0.0842}$). While regional differences were also observed, with the Southeast and Southwest having significantly lower costs than the Northeast, the model found that gender has no statistically significant effect on charges after accounting for the other variables. These clear, interpretable results are a major benefit of using a GLM for this type of analysis.

**Predictive Scenarios**

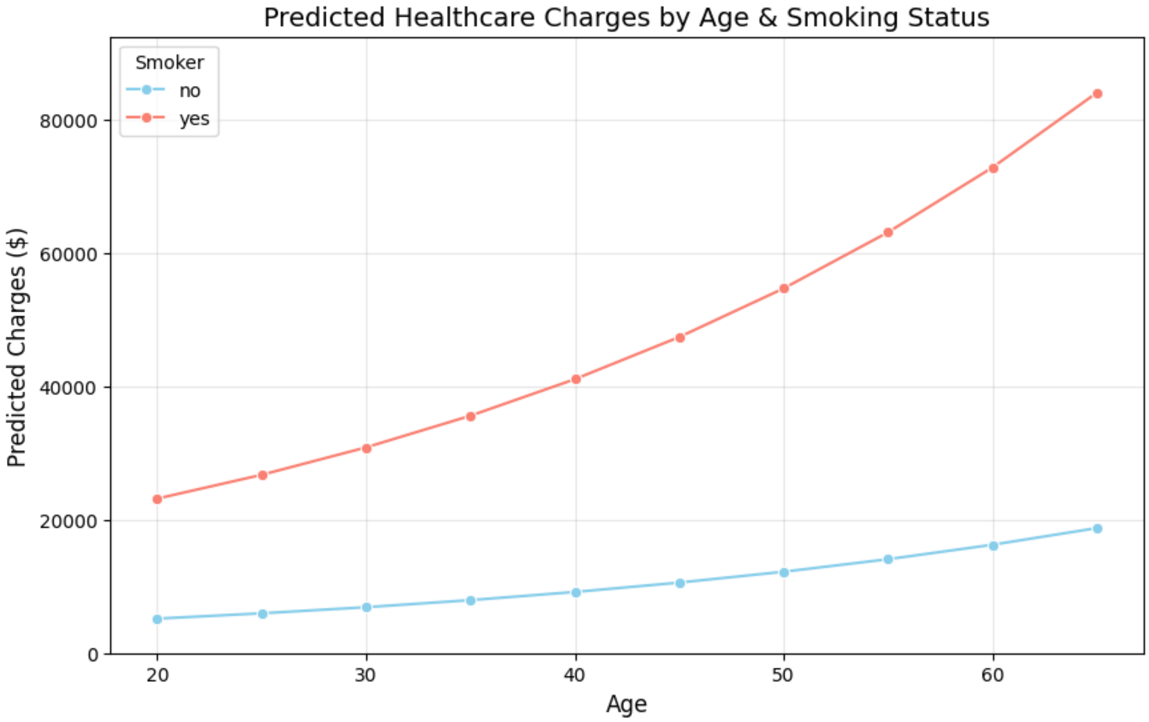
To illustrate the model's utility, four distinct profiles were created, ranging from a low-risk to a high-risk individual. The model was used to predict the insurance charges for each profile, as shown in the table below.

- *Low-Risk Profile:* A 25-year-old non-smoking female with a healthy BMI and no children is predicted to have very low charges, around $4,505.
- *Moderate-Risk Profile:* As risk factors accumulate, the predicted charges rise. A 45-year-old non-smoker with a higher BMI and two children is predicted to have charges of over $10,584.
- *High-Risk Profile:* The impact of smoking is most evident in the high-risk profiles. A 60-year-old male smoker with a high BMI and three children is predicted to have charges approaching $70,000, highlighting the dramatic cost increase associated with smoking, age, and BMI.

These scenarios demonstrate the model's ability to provide a granular, data-driven cost estimate for an individual based on their specific characteristics, which is critical for insurance pricing and risk assessment.

## Visualizations

A line plot of predicted charges across different ages for smokers and non-smokers clearly visualizes the model's key finding. It shows a steady, moderate increase in charges with age for non-smokers, while charges for smokers rise exponentially and are significantly higher at every age, especially in later years.



## Conclusion

This project successfully developed an interpretable GLM to model and predict health insurance charges. The analysis confirmed that smoking is the most significant determinant of costs, far outweighing the effects of age, BMI, and other demographic factors. The model's coefficients provide clear, actionable insights into how each variable contributes to an individual's predicted costs, a critical capability for actuarial pricing and risk management. While this GLM serves as a robust baseline, future work could explore more advanced machine learning techniques to further enhance predictive accuracy, such as incorporating interaction terms or using models like Gradient Boosting. Ultimately, this project serves as a strong foundation for understanding the complex drivers of healthcare costs and applying data-driven techniques to solve real-world problems.