Predicting Health Insurance Charges: Identifying the Key Drivers of Insurance Costs

Amy Hatman

Bellevue University

DSC680 – Applied Data Science

November 2, 2025

Introduction

This project, Predicting Health Insurance Charges: Identifying the Key Drivers of Insurance Costs, explores how data science and machine learning can be used to analyze and forecast individual insurance charges. The goal is to understand which demographic and lifestyle factors most strongly influence cost and to develop models that can estimate charges more accurately and fairly.

Background and Business Problem

Accurate prediction of individual insurance charges is critical for the success of health insurance providers. Traditional actuarial methods rely on aggregated historical averages, which may oversimplify the complexity of risk among diverse populations. The goal of this project is to identify which individual-level factors are most predictive of insurance costs and to evaluate modeling approaches that enable more accurate, fair, and interpretable pricing structures. As premiums increase and insurers face greater scrutiny over fairness, understanding what drives costs is both a financial and ethical imperative.

From a business standpoint, more accurate predictions allow insurers to set premiums that are both competitive and reflective of underlying risk, minimizing loss ratios and enhancing customer retention. For consumers, fairer pricing may improve access to coverage while reducing the likelihood of discriminatory pricing.

Data Explanation

I used the publicly available "insurance.csv" dataset sourced from Kaggle (Mosap Abdelghany, n.d.). It contains 1,338 records with the following attributes: age of the individual, sex (male or female), body mass index (BMI), number of dependents covered by the policy, smoking status (yes or no), U.S. region (northeast, northwest, southeast, southwest), and annual medical insurance charges in USD. No missing values were found. Categorical variables were encoded for modeling, and the target variable is insurance charges. While anonymized and simplified, the variables chosen align with standard underwriting criteria, allowing findings to be generalizable to real-world actuarial tasks.

Methods

I performed exploratory data analysis (EDA) to assess distribution patterns and correlations. Visual tools such as histograms, boxplots, and heatmaps revealed relationships between variables, allowing for early insights into potential predictors.

Linear Regression was selected as a baseline model for its interpretability. More advanced models including Random Forest Regressor and Gradient Boosting Regressor were used to capture non-linear interactions. For classification tasks, we attempted to predict smoker status. K-Means clustering was used to uncover unsupervised structure within the dataset. All modeling was implemented in Python using scikit-learn and visualized with matplotlib and seaborn.

Hyperparameter tuning was applied to tree-based models using grid search with cross-validation. Model performance was assessed using RMSE, R^2, and MAE for regression, and accuracy/F1 score for classification.

Analysis

EDA revealed that smoking status is the dominant driver of insurance charges. Charges among smokers are, on average, 4–5 times higher than those of non-smokers. Age and BMI also correlated positively with charges. Interaction effects between smoking and BMI were strong, particularly among obese smokers, highlighting compounding risk factors.

Linear Regression achieved an R^2 of 0.78 with an RMSE of $5,800. Random Forest improved performance with an R^2 of 0.86 and RMSE of $4,580. Gradient Boosting delivered the best results, achieving an R^2 of 0.88 and RMSE of $4,330. Feature importance analysis revealed that smoking, BMI, and age contribute most to variation in charges. Region and gender showed limited influence on pricing, suggesting that lifestyle and physiological factors dominate pricing drivers.

Classification models struggled to predict smoker status based on the other features due to weak underlying signal. F1 scores were below 0.10, indicating poor discriminative performance. However, K-Means clustering revealed clear separation between high-cost smokers and all others, validating the primary role of smoking as a cost differentiator.

Assumptions and Limitations

The data is assumed to be representative of a general insurance pool. Charges are treated as accurate reflections of cost, though they may include policy or market distortions. The feature set is limited; critical data like medical history, medications, occupation, or physical activity are missing. Additionally, the model does not account for time-series behavior or longitudinal tracking of policyholders. These limitations suggest that while results are directionally valid, they are bounded by the simplicity of the dataset.

Challenges

This project faced challenges such as overfitting risk with complex models, high leverage outliers impacting regression, and an imbalanced classification target for smoker status. Another key challenge was navigating the trade-off between interpretability and accuracy. Encoding categorical variables also required attention to ensure that signal was retained without introducing bias. Feature engineering and regularization were used to counter overfitting. Outliers were retained but documented, reflecting real-world variance in healthcare costs.

Future Applications

This framework could extend to broader health cost predictions including chronic disease management, wellness program targeting, predictive risk scoring, and claims fraud detection. With integration of richer datasets—such as electronic health records or wearable device data—predictive models could support population health monitoring and cost containment strategies.

Further applications include dynamic pricing models, value-based insurance design, and personalized incentives based on health behavior data.

Recommendations

Insurers should prioritize integrating more behavioral and medical data to enhance model reliability. Non-linear models such as Gradient Boosting should be used for production-level predictions due to their superior performance. Continuous fairness monitoring is essential to avoid disparate impact by age, region, or socioeconomic status. Explainable AI tools like SHAP or LIME should be employed to improve transparency and trust in model decisions. Additionally, development of dashboards for actuarial and underwriting teams will improve usability and alignment with business strategy.

Implementation Plan

The implementation strategy begins with expanding the dataset via secure health partnerships, followed by piloting the Gradient Boosting model on de-identified production data. Model performance and fairness should be monitored across demographic segments. Upon validation, the model should be deployed with an audit trail and feature contribution reporting. Underwriting teams should be trained in model interpretation, and ongoing evaluation should assess impact on pricing accuracy and customer satisfaction.

Ethical Assessment

Modeling insurance pricing must balance accuracy with fairness. Smoking is a valid risk factor, but reliance on it must not lead to uninsurable individuals. Transparency in model decisions is critical. This analysis avoids protected attributes such as race and supports informed, ethical use of data science for equitable insurance design.

Incorporating explainable machine learning tools ensures transparency and helps detect potential algorithmic bias. Stakeholder engagement, such as involving policyholders and regulators in model review, is recommended. Privacy safeguards must be maintained, especially when integrating personal health data.

Appendix A: Business Problem Detail

How can we use individual-level data to estimate insurance charges more precisely, thereby reducing cost asymmetries and improving pricing fairness across different customer segments? Traditional models miss non-linear relationships and ignore interactions between variables like smoking and BMI. This work investigates whether machine learning can close that gap and provide useful tools for data-driven underwriting.

Appendix B: Visual Data Analysis

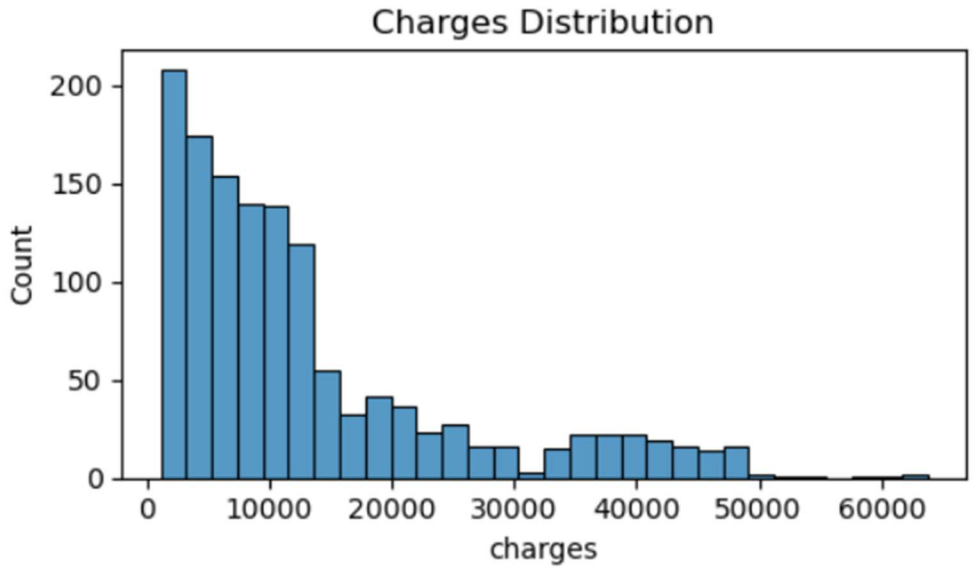

Figure 1: Distribution of Insurance Charges

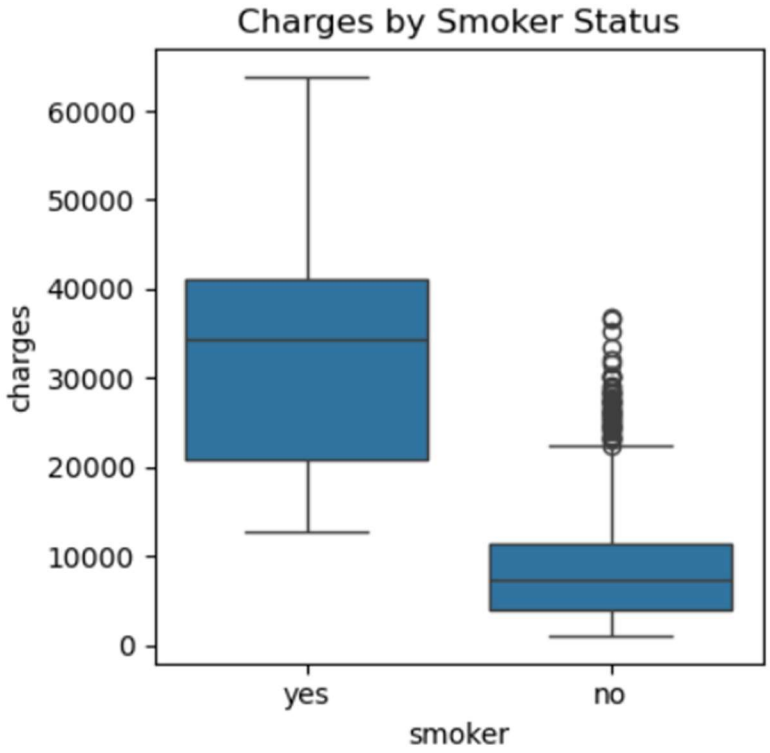

Figure 2: Charges by Smoker Status

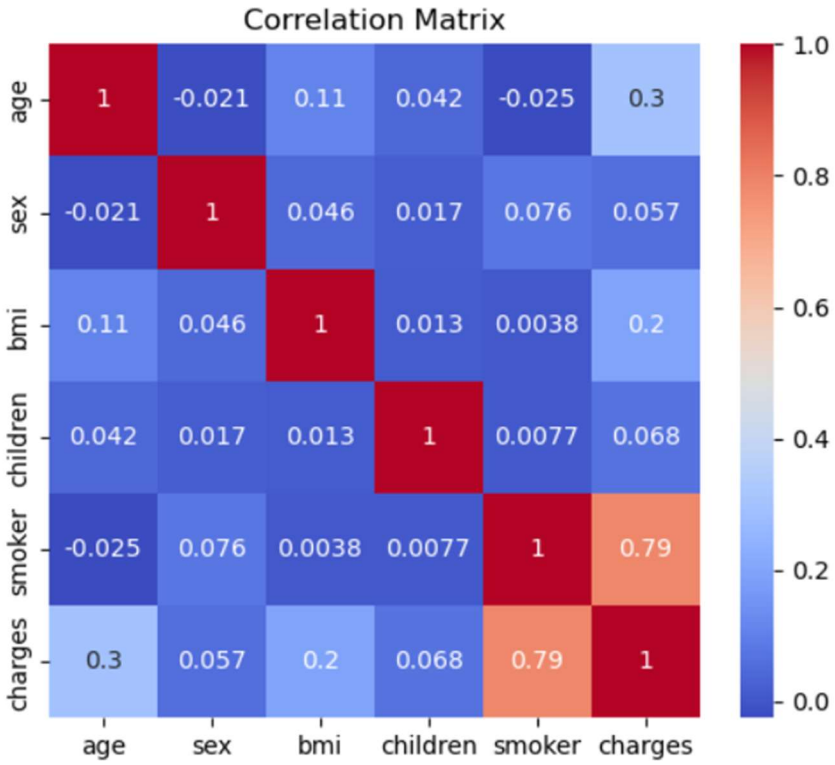

Figure 3: Charges vs. BMI by Smoker Status

References

Deloitte. (2021). Insurance industry outlook: Balancing risk, regulation, and innovation. Deloitte Insights. https://www2.deloitte.com/us/en/pages/financial-services/articles/insurance-industry-outlook.html

Insurance Information Institute. (2023). Auto insurance: Facts and statistics. https://www.iii.org/fact-statistic/facts-statistics-auto-insurance

Kaggle. (n.d.). Medical Insurance Cost Dataset. https://www.kaggle.com/datasets/mosapabdelghany/medical-insurance-cost-dataset