Predicting U.S. Healthcare Expenditures Using Machine Learning

Dolon Talukder
Professor: Dr. Christelle Scharff
Seidenberg Graduate Capstone
Pace University, Seidenberg School of CSIS

Abstract

Healthcare costs in the United States have continued to rise, placing financial strain on patients and institutions. This project leverages the 2022 MEPS HC-243 dataset to identify drivers of healthcare expenditure and predict total annual costs (TOTEXP22) using machine learning models. Through exploratory data analysis and predictive modeling, key cost indicators such as insurance type, age, and source of payment are extracted to inform future forecasting and financial planning tools.

Research Question

What factors are most influential in determining total annual healthcare spending in the U.S.? Can predictive models accurately estimate individual healthcare costs based on demographic, socioeconomic, and payer-related features?

Related Work

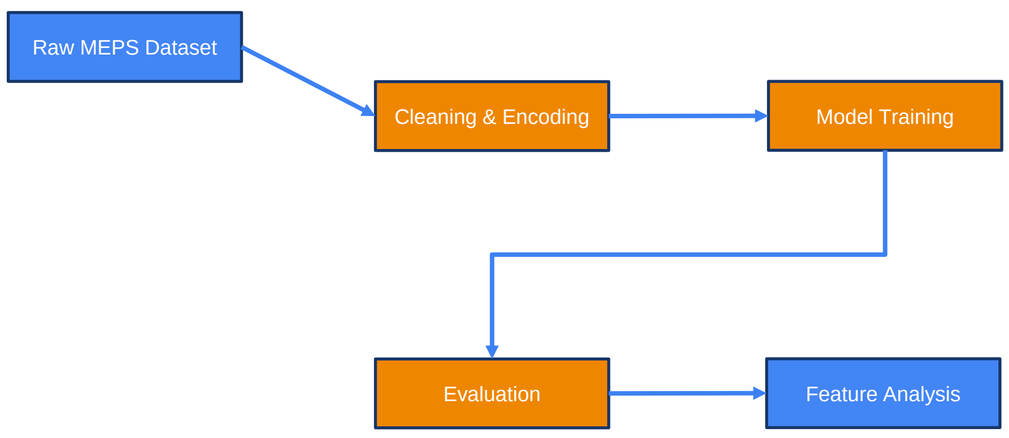
Existing research using MEPS data has demonstrated strong correlations between healthcare expenditure and factors such as insurance coverage, income level, and chronic health conditions. Prior studies have used statistical modeling and basic regression approaches. This project extends that work by incorporating multiple machine learning models and conducting comparative analysis to improve accuracy and interpretability.

Dataset

The MEPS HC-243 (2022) Full-Year Consolidated File contains data on 22,431 individuals and over 1,400 features, including demographics, health status, expenditure, and insurance type. After cleaning, one-hot encoding, and feature selection, the analysis focused on a refined subset of predictors for modeling TOTEXP22 (total healthcare expenditure).

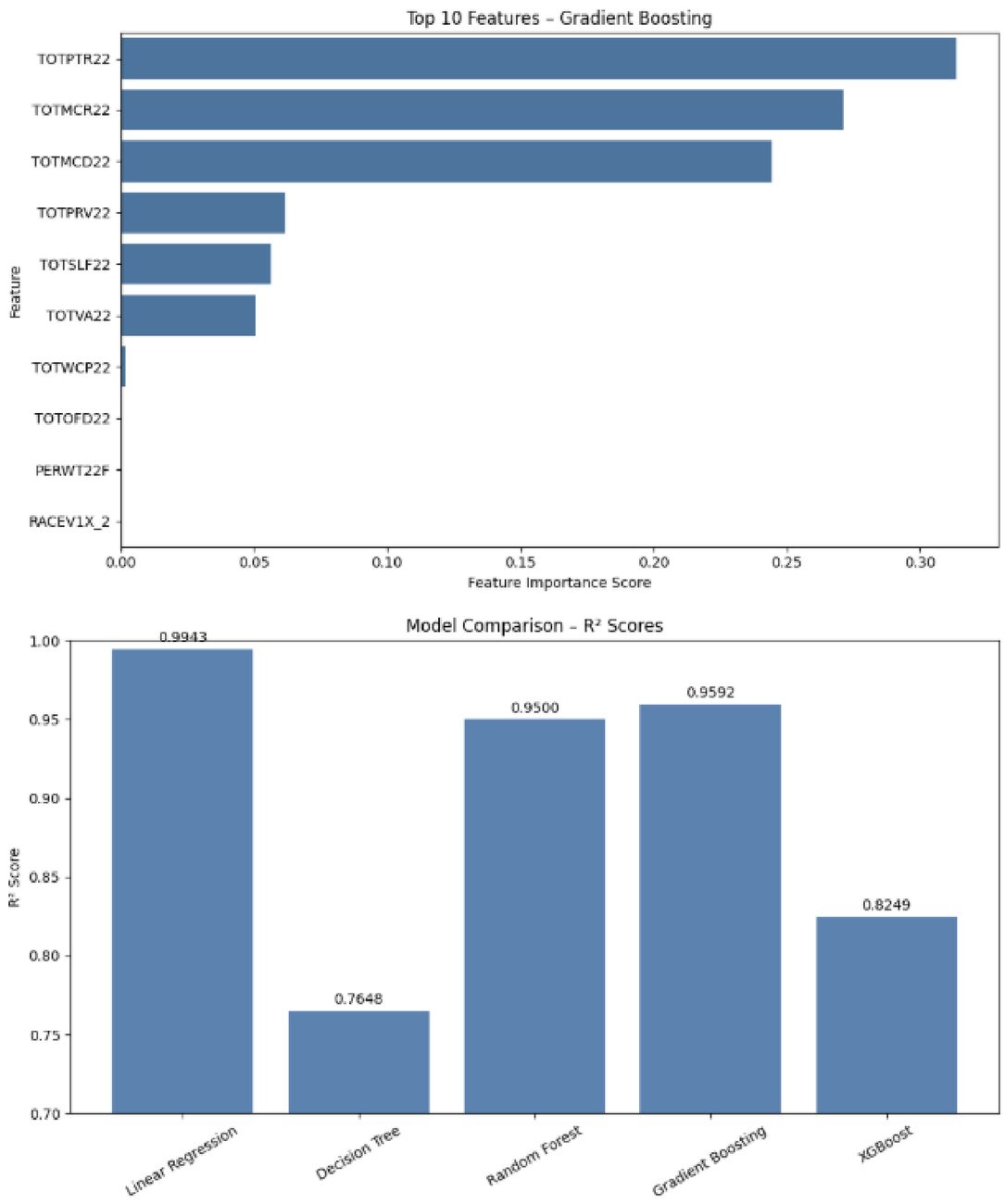
Methodology

The workflow began with data cleaning and exploratory analysis, addressing missing values, skewed distributions, and outliers. Categorical variables were transformed using one-hot encoding. Five models were implemented and evaluated: Linear Regression, Decision Tree, Random Forest, Gradient Boosting, and XGBoost. Model performance was assessed using MAE, RMSE, and R^2. Feature importance analysis was conducted using the Gradient Boosting model to interpret the strongest cost drivers.



Results

Among all models tested, XGBoost produced the most accurate predictions, with the lowest error metrics across MAE, RMSE, and R^2. Features related to insurance type, patient age, number of office visits, and source of payment ranked highest in predictive power. Exploratory analysis also revealed significant right-skew in expenditure distribution and disparities across geographic regions.



Conclusion & Future Work

The modeling pipeline successfully predicted healthcare expenditure with reasonable accuracy and interpretability. The project highlights the potential for machine learning to support healthcare cost forecasting and policy design. Future work may include the integration of temporal variables, more granular health condition codes, and deeper tuning of model hyperparameters to further improve predictive performance.

References

- 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, 1362–1371. https://pmc.ncbi.nlm.nih.gov/articles/PMC5977561/
- Nomura, Y., Ishii, Y., Chiba, Y., Suzuki, S., Suzuki, A., Suzuki, S., Morita, K., Tanabe, J., Yamakawa, K., Ishiwata, Y., Ishikawa, M., Sogabe, K., Kakuta, E., Okada, A., Otsuka, R., & Hanada, N. (2021). Does last year's cost predict the present cost? An application of machine learning for the Japanese area-basis public health insurance database. International Journal of Environmental Research and Public Health, 18(2), 565. https://www.mdpi.com/1660-4601/18/2/565
- Chen, Z., Siltala-Li, L., Lassila, M., Malo, P., Vilkkumaa, E., & Saaresranta, T. (2023). Predicting visit cost of obstructive sleep apnea using electronic healthcare records with transformer. IEEE Access, 11, 50313–50325. https://ieeexplore.ieee.org/document/10128115
- Sepetis, A., Zaza, P. N., Rizos, F., & Bagos, P. G. (2022). Identifying and predicting healthcare waste management costs for an optimal sustainable management system: Evidence from the Greek public sector. International Journal of Environmental Research and Public Health, 19(16), 9821. https://www.mdpi.com/1660-4601/19/16/9821
- Sohn, J. H., Chen, Y., Lituiev, D., Yang, J., Ordovas, K., Hadley, D., Vu, T. H., Franc, B. L., & Seo, Y. (2022). Prediction of future healthcare expenses of patients from chest radiographs using deep learning: A pilot study. Scientific Reports, 12, Article 8679. https://www.nature.com/articles/s41598-022-12551-4
- Zou, S., Chu, C., Shen, N., & Ren, J. (2023). Healthcare cost prediction based on hybrid machine learning algorithms. Mathematics, 11(23), 4778. https://www.mdpi.com/2227-7390/11/23/4778