Healthcare Cost Analytics & Modeling

Independent Data Analytics Project, November 2025

Executive Summary

This project aims to identify the main drivers of individual healthcare costs and to predict medical expenses based on basic patient information. The dataset, derived from Kaggle's *Medical Cost Personal Dataset*, contains approximately 1,300 records with seven key variables: age, sex, BMI, children, smoker status, region, and medical charges.
The analysis followed a structured consulting approach: (1) data audit and cleaning, (2) exploratory data analysis (EDA), (3) baseline and advanced model comparison, and (4) interpretation and business recommendations. The ultimate goal was to balance model accuracy with interpretability, ensuring actionable insights for insurers and healthcare planners.

Data Preparation and Methods

A thorough data audit was performed, removing one duplicate, imputing missing values using medians and modes, and winsorizing outliers in medical charges at the 1–99% percentile. New engineered features included categorical bins for BMI and age, and interaction terms capturing the joint effect of smoking with age and BMI.
The modeling process started with a baseline Linear Regression model (R^2 = 0.82, MAE ≈ €4,043). A polynomial and interaction expansion was then tested but led to overfitting (R^2 = 0.81, MAE ≈ €4,270). Finally, a Ridge regression model incorporating selected domain-based interactions (age × smoker, BMI × smoker) achieved a significantly improved performance (R^2 = 0.89, MAE ≈ €2,708).

Model	R^2	MAE (€)	Comment
Linear Regression (Base)	0.82	4,043	Strong baseline, high interpretability
Polynomial + Interactions	0.81	4,270	Overfitting, limited generalization
Ridge + Smoker Interactions	0.89	2,708	Best trade-off between accuracy and robustness

Key Insights

Smoking emerged as the most influential driver of medical costs. On average, smokers incurred expenses over €23,000 higher than non-smokers. Age and BMI were also positively correlated with costs, and their impact was magnified among smokers. Regional differences existed but were secondary compared to individual risk factors.
The regularized Ridge model successfully balanced predictive accuracy and interpretability. It demonstrated that complexity adds value only when guided by domain knowledge, in this case, the inclusion of business-relevant interactions. The 10-fold cross-validation confirmed the model's robustness (MAE: €2,908 ± 148; R^2: 0.838 ± 0.032).

Business Recommendations

- Develop targeted prevention and smoking cessation programs to reduce long-term healthcare expenditures.
- Introduce risk-based pricing strategies for insurance premiums incorporating smoking and BMI factors.
- Apply explainable machine learning approaches (e.g., regularized regression) for transparent risk assessment.
- Continue monitoring predictive stability across time and regions to ensure fairness and compliance.

Guillermo Encinas