Project Report : Predicting Medical Charges

1 Business Case: Defining the Problem and Objective

The core business challenge we address is the need for accurate financial forecasting within the insurance and healthcare sectors. Insurance companies require precise estimation of future liabilities to calculate premiums, manage risk portfolios, and ensure solvency. Individuals also benefit from transparent, data-driven cost predictions.

Our project objective is to develop a predictive model capable of estimating individual medical insurance charges based on readily available personal and health metrics. This is a regression problem, aiming to predict a continuous numerical value.

This objective is directly linked to the field of Data Science and Predictive Modeling. Our specialization lies in transforming complex, multi-variate data into actionable financial predictions using statistical learning techniques to manage uncertainty and support strategic decision-making. The necessity for predictive accuracy dictates the type of data we must collect: features that are known determinants of healthcare utilization and cost.

2 Dataset Description and Source

The dataset used is a publicly available resource, commonly used for educational and benchmarking purposes in healthcare economics modeling. It contains 1,338 records, where each entry represents an individual. The features collected are all known variables influencing health costs: age, sex, BMI (Body Mass Index), number of children, smoker status, and region of residence. The `charges` column serves as our target variable.

3 Data Exploration, Graphics, and Formalization of the Problem

Initial exploration revealed a clean dataset with no missing values, simplifying the early stages of the project. Analysis confirmed key relationships that would inform our model. For instance, the target variable `charges` exhibited a highly right-skewed distribution, meaning most costs are low, but a few individuals incur extremely high expenses. This skewness suggests that linear models might struggle, hinting at the need for transformation or non-linear models.

A crucial insight came from the correlation analysis. We observed a notably strong positive correlation between being a smoker and high charges, underscoring its predictive power. Age and BMI also showed moderate positive correlations with the target variable.

The problem is formally defined as a supervised regression task. Given:

$$X = (\text{age}, \text{sex}, \text{bmi}, \text{children}, \text{smoker}, \text{region}),$$

we aim to find a function f approximating medical charge y such that:

$$f(X) = y.$$

4 Preprocessing, Models, and Obstacle Mitigation

All categorical variables (sex, smoker, region) were encoded using OneHot Encoding. To mitigate scale imbalance, all features were scaled using Standard Scaler. The data was split into an 80% training set and a 20% test set.

The models selected for performance comparison were: Linear Regression (LR), Random Forest Regressor (RFR), Gradient Boosting Regressor (GBR), and KNN Regressor.

Data Skewness

The heavy right-skew in the `charges` variable risks violating the homoscedasticity assumption of linear regression, possibly leading to underestimation of high-cost patients.

5 Comparison of Model Results

Table 1 presents the performance metrics for the four models evaluated. Metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2.

Model	**MAE**	**RMSE**	R^2	**Interpretation**
Linear Regression	4181.19	5796.28	0.784	Baseline model. Captures global linear trends but cannot model nonlinear interactions such as the strong impact of smoking.
Random Forest Regressor	2525.65	4603.87	0.863	Strong improvement over linear regression. Learns nonlinear patterns and feature interactions, reducing prediction error significantly.
Gradient Boosting Regressor	**2404.90**	**4328.15**	**0.879**	Best-performing model overall. Sequential boosting progressively corrects residual errors, delivering excellent predictive accuracy.
KNN Regressor	3332.96	5426.54	0.810	Reasonable performance but inferior to ensemble methods. Sensitive to scaling and degraded by sparse high-dimensional data.

Table 1: Comparison of model performance metrics and interpretation.

6 Conclusion

This project explored the prediction of individual medical charges using multiple machine learning models. Our results show that:

- Linear Regression serves as a useful baseline but lacks the complexity to capture nonlinear relationships.
- Ensemble methods such as Random Forest and Gradient Boosting offer substantial improvements in accuracy.
- The Gradient Boosting Regressor is the best model, achieving the lowest error metrics and highest R^2.
- Smoking status, BMI, and age are the most influential predictors of medical expenditure.

Overall, the Gradient Boosting model provides the most reliable and robust predictions, effectively capturing the nonlinear patterns inherent in healthcare cost data.