Healthcare Cost Prediction Report

A Data Science Project

Author: Srilekha J   
Date: August 2025

# Executive Summary

This project aims to predict healthcare insurance costs using patient data. We analyzed the dataset through exploratory data analysis (EDA), conducted statistical tests, and applied machine learning models. Key findings show that smoking status, age, and BMI are the most important factors influencing healthcare charges. Random Forest outperformed Linear Regression in predicting charges.

# 1. Introduction

Healthcare cost prediction is important for insurance companies and hospitals to design fair pricing policies. The dataset contains demographic and health-related features such as age, sex, BMI, children, smoking habits, and region.

# 2. Dataset Overview

The dataset (insurance.csv) contains patient information with the following columns:  
- Age: Age of the individual  
- Sex: Gender (male/female)  
- BMI: Body Mass Index  
- Children: Number of children covered  
- Smoker: Smoking status (yes/no)  
- Region: Residential area  
- Charges: Medical insurance cost (target variable)

# 3. Exploratory Data Analysis (EDA)

We performed univariate and bivariate analysis to understand data distribution and relationships:

- calculating mean, variance, standard deviation, IQR, percentile, skewness and kurtosis, correlation.  
- Univariate: Histograms, boxplots, and countplots showed the distribution of age, BMI, charges, and categorical features.  
- Bivariate: Scatterplots, boxplots, and correlation heatmaps revealed strong relationships between smoking and charges, as well as age/BMI and charges.

# 4. Statistical Tests

- T-test: Charges for smokers are significantly higher than non-smokers (p < 0.05).  
- ANOVA: Regional differences in charges are statistically significant.  
- Correlation: Age and BMI are positively correlated with charges.

# 5. Modeling & Results

To evaluate the performance of various regression models, we applied a range of machine learning algorithms using **5-Fold Cross-Validation**. The following models were tested:

* Linear Regression
* Support Vector Machine (SVM)
* Random Forest Regressor
* Decision Tree Regressor
* Ridge Regression
* Lasso Regression
* XGBoost
* Gradient Boosting Regressor
* AdaBoost Regressor
* LightGBM Regressor

**Evaluation Metrics Used**

* **R² Score**: Indicates how well the model explains the variance in the target variable. Higher is better.
* **Mean Squared Error (MSE)**: Measures the average squared difference between predicted and actual values. Lower is better.
* **Mean Absolute Error (MAE)**: Measures the average absolute difference between predicted and actual values. Lower is better.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **R2Score** | **MeanSquaredError** | **MeanAbsoluteError** |
| **LinearRegression** | 0.742059 | 3.863930e+07 | 4375.440921 |
| **SVM** | -0.143712 | 1.713268e+08 | 8666.761701 |
| **RF** | 0.831677 | 2.521467e+07 | 2722.161888 |
| **DT** | 0.695701 | 4.558370e+07 | 3329.612259 |
| **Ridge** | 0.742371 | 3.859254e+07 | 4385.502600 |
| **Lasso** | 0.742062 | 3.863887e+07 | 4375.671456 |
| **Xg\_boost** | 0.802186 | 2.963226e+07 | 3193.907959 |
| **gradient\_boost** | 0.846539 | 2.298829e+07 | 2533.643695 |
| **ada\_boost** | 0.823975 | 2.636835e+07 | 3632.581991 |
| **LightGBM** | 0.849127 | 2.260055e+07 | 2502.636984 |

# 6. Insights & Observations

* Linear Regression, Ridge, and Lasso performed similarly, capturing linear patterns but failing to fully model non-linear relationships in the data.
* Support Vector Machine (SVM) performed poorly with a negative R² score, indicating that it failed to model the data effectively.
* Random Forest, Gradient Boosting, XGBoost, AdaBoost, and LightGBM showed significantly better results by capturing complex, non-linear relationships.
* Gradient Boosting and LightGBM outperformed all other models in terms of R² score and error metrics.
* LightGBM achieved the best performance overall, with the highest R² score (0.8491) and lowest MAE and MSE, indicating its robustness and accuracy.

# 7. Conclusion

Among all the tested models:

* **LightGBM Regressor** emerged as the **top performer**, making it the most suitable choice for this regression task.
* **Gradient Boosting** and **Random Forest** also demonstrated strong performance and can be considered viable alternatives.
* **SVM** was not appropriate for this problem, likely due to its sensitivity to feature scaling and inability to handle complex data without proper kernel tuning