Predicting Medical Costs Using Regression Analysis

A Data-Driven Approach to Healthcare Insights

LINDA KELLEN AYEBALE (2400721933)
KARANZI JOHNMARY (2400721928)

November 21, 2024

Problem Statement

Objective: To predict individual medical costs billed by health insurance, using demographic and lifestyle factors such as age, BMI, smoking status, and region.

Why is this important?

- Helps insurers estimate risk and costs accurately.
- Aids in policy pricing for different demographics.
- Provides actionable insights for reducing high medical expenses.

Research Questions

1. How do demographic factors like age and region influence medical costs?
2. What is the impact of smoking on medical costs compared to other features?
3. Does having more children significantly increase medical costs?
4. How do BMI levels affect medical costs across different regions?
5. Can interactions between features (e.g., smoking and BMI) reveal hidden patterns in cost prediction?

Dataset Overview and Cleaning

Dataset Description:

- Number of Records: 1338
- Features: Age, BMI, Number of Children, Smoking Status, Region, Charges.
- Target Variable: `charges` (medical costs billed).

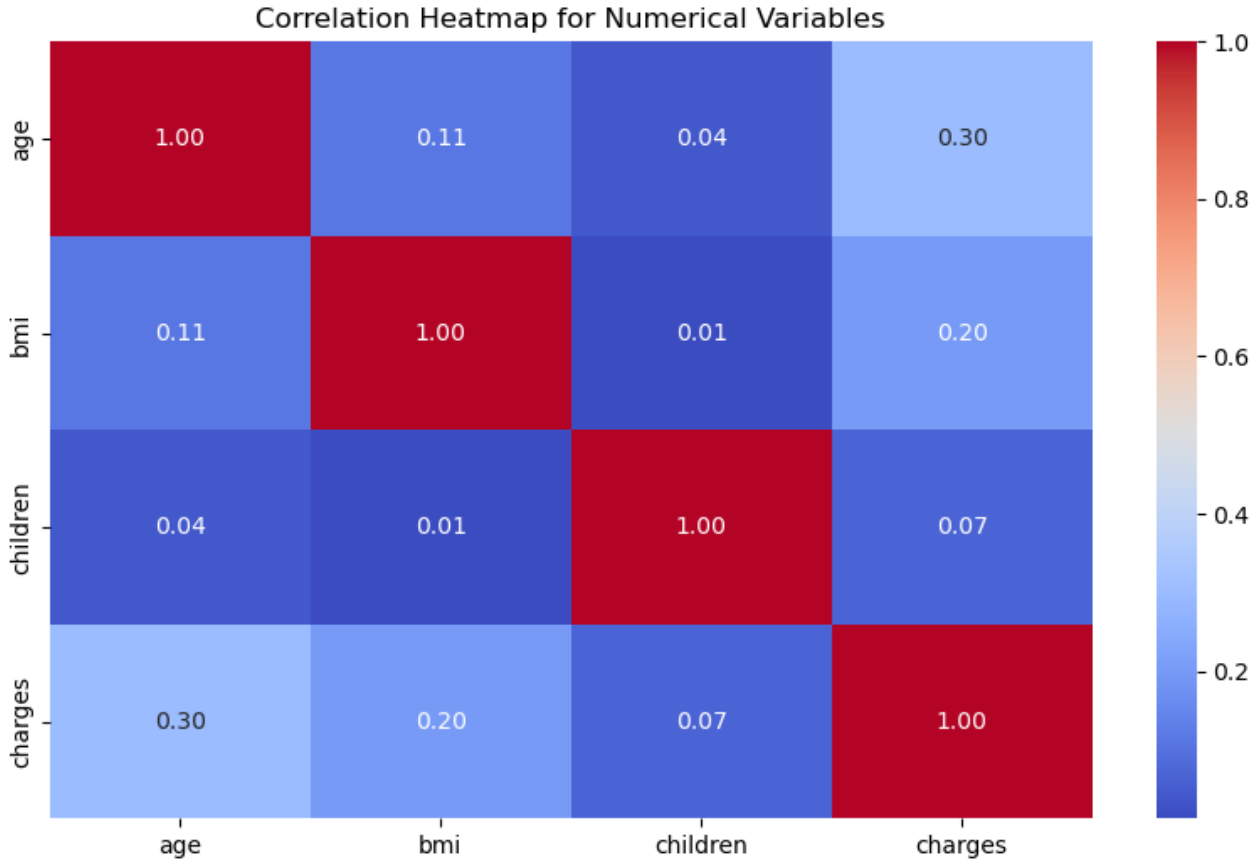
Data Cleaning Steps:

- Checked for and confirmed no missing values.
- Categorical variables (`sex`, `smoker`, `region`) were encoded using label encoder and map encoder.

Correlation Matrix

Correlation Insights:

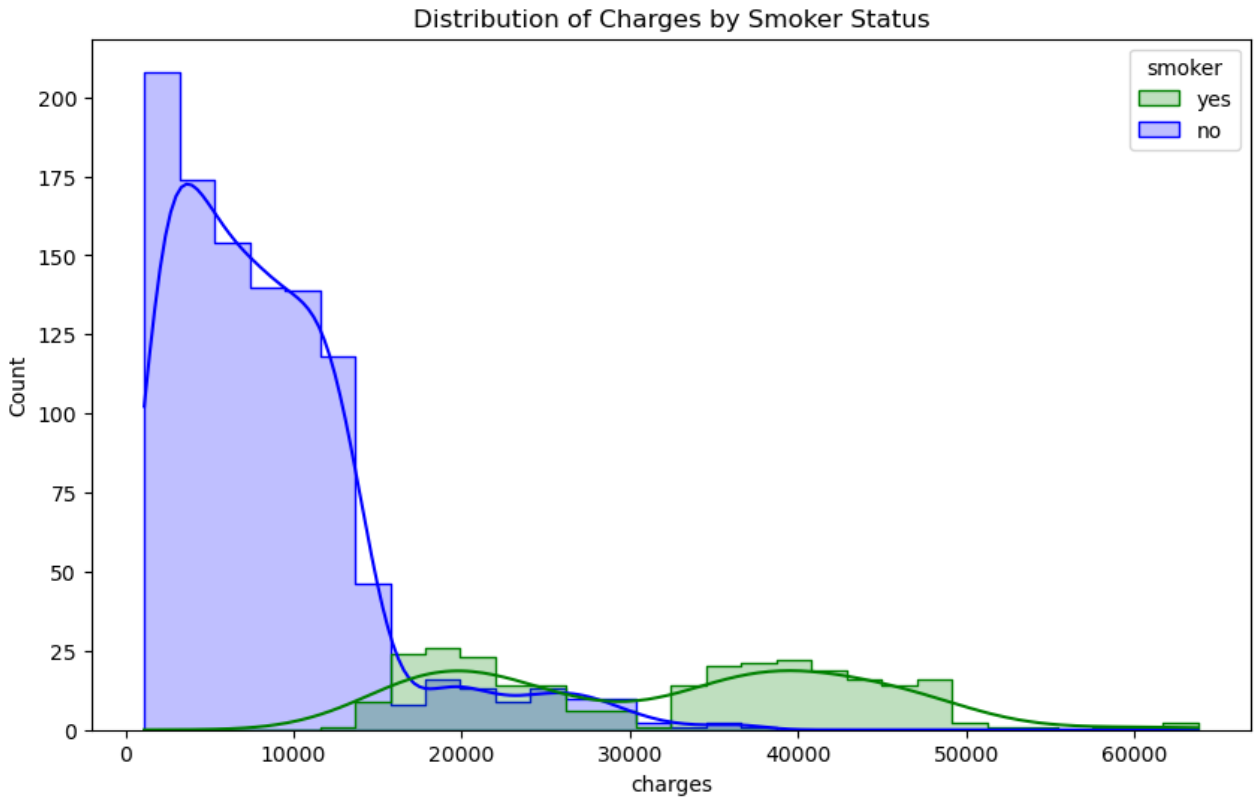
- Age is the most correlated feature with charges, suggesting its strong influence.
- Number of children and bmi have weaker correlations.



Smoking: The Most Influential Factor

Key Insight: Smoking (`smoker`) has the strongest impact on medical costs:

- Smokers incur significantly higher medical costs than non-smokers.



Data Splitting and Training

Data Splitting:

- Dataset split into Training (70%), Validation (10%), and Testing (20%).
- Random state was set for reproducibility.

Training Process:

- Linear Regression:
  - Used as a baseline model.
  - Assumes a linear relationship between features and target.
- Random Forest:
  - Hyperparameters tuned using grid search (e.g., number of trees, max depth).
  - Captures non-linear relationships and interactions between features.

Model Results: Regression vs Random Forest

Regression Metrics:

Dataset	MAE	MSE	RMSE	R^2 Score
Training	4251.53	3.77e+07	6144.20	0.742
Validation	3868.68	3.06e+07	5532.54	0.776
Testing	4295.34	3.54e+07	5946.35	0.766

Table: Linear Regression Results

Random Forest Metrics:

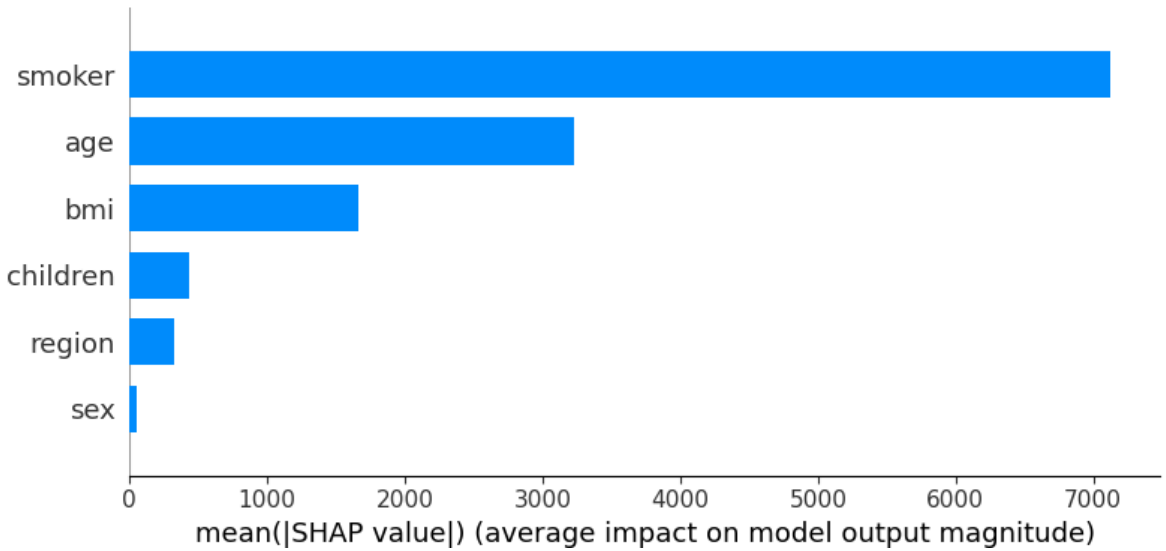
Dataset	MAE	MSE	RMSE	R^2 Score
Training	1039.27	3.38e+06	1838.65	0.977
Validation	2655.03	2.33e+07	4828.26	0.830
Testing	2548.64	2.09e+07	4567.29	0.862

Table: Random Forest Results

Explainability with SHAP

SHAP Analysis Highlights:

- Smoking (`smoker`) is the most influential factor, significantly driving up costs.
- Age and BMI also contribute heavily to predictions.



Conclusion

Key Findings:

- Smoking, BMI, and age are the most influential factors in predicting medical costs.
- Random Forest is the best-performing model for this dataset.
- SHAP analysis provides clear insights into feature importance.

Future Work:

- Incorporate additional features for better predictions (e.g., exercise habits, diet).
- Explore other advanced models like Gradient Boosting for comparison.

Thank you for your attention!