Insurance Charges Predictor

End-to-End Machine Learning Project (Regression)

Author: Phuc Vo

Date: January 10, 2026

This report documents the full ML workflow: EDA, data validation, preprocessing and feature engineering, model training and evaluation, and deployment as a FastAPI service with a Streamlit UI.

Table of Contents

1. Introduction

The goal of this project is to predict medical insurance charges (charges) from basic demographic and lifestyle information. The final deliverable includes a trained model, a serving API, and a small web UI.

2. Dataset

The dataset contains 1338 records and 7 columns. Target variable: charges. Features: age, sex, bmi, children, smoker, region.

Table 1. Descriptive statistics (numeric columns).

stat	age	bmi	children	charges
mean	39.21	30.66	1.09	13270.42
std	14.05	6.10	1.21	12110.01
min	18.00	15.96	0.00	1121.87
25%	27.00	26.30	0.00	4740.29
50%	39.00	30.40	1.00	9382.03
75%	51.00	34.69	2.00	16639.91
max	64.00	53.13	5.00	63770.43

3. Exploratory Data Analysis (EDA)

EDA focuses on distributions and relationships with the target. The dataset used here contains no missing values. Figures below highlight the target distribution and the strong association between smoker status and charges.

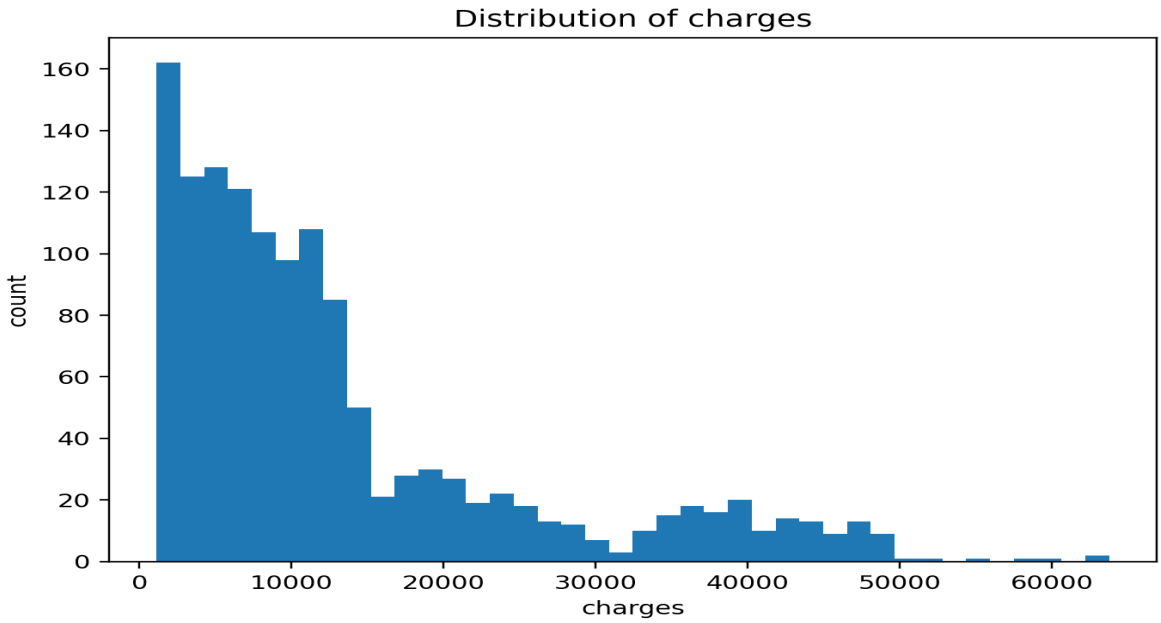

Figure 1. Distribution of charges (right-skewed).

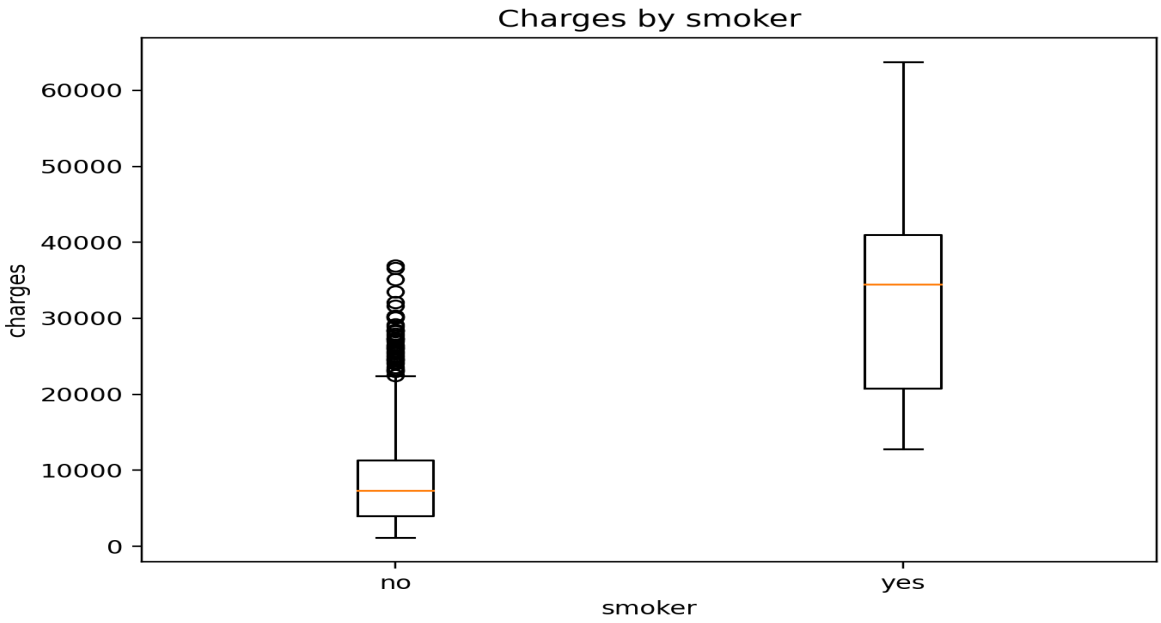

Figure 2. Charges by smoker status.

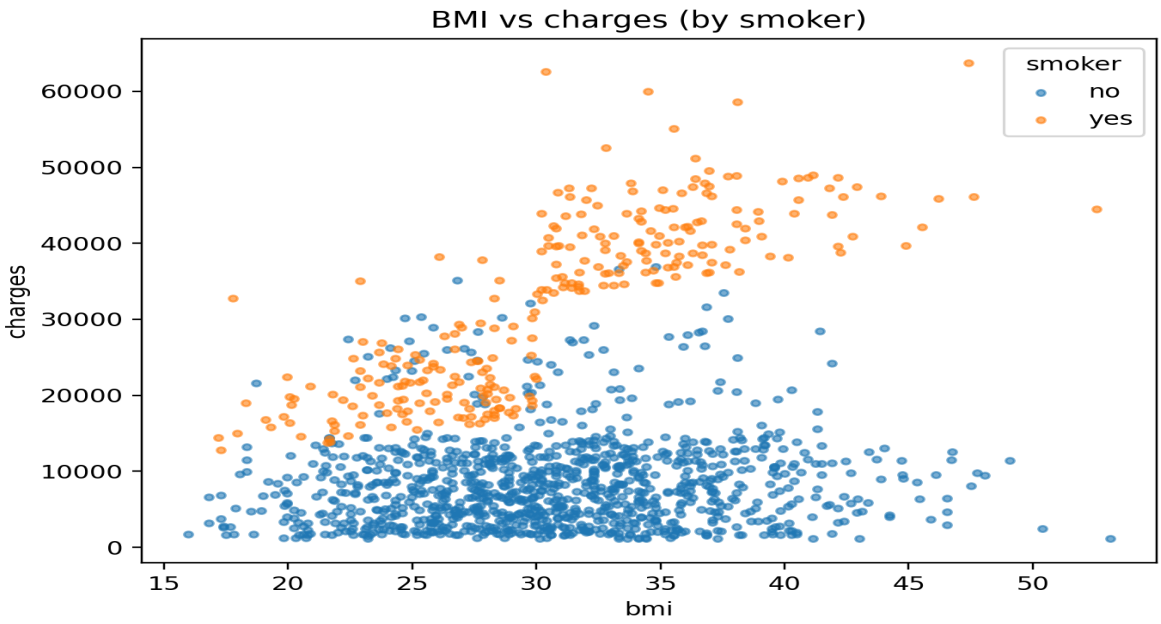

Figure 3. BMI vs charges (by smoker).

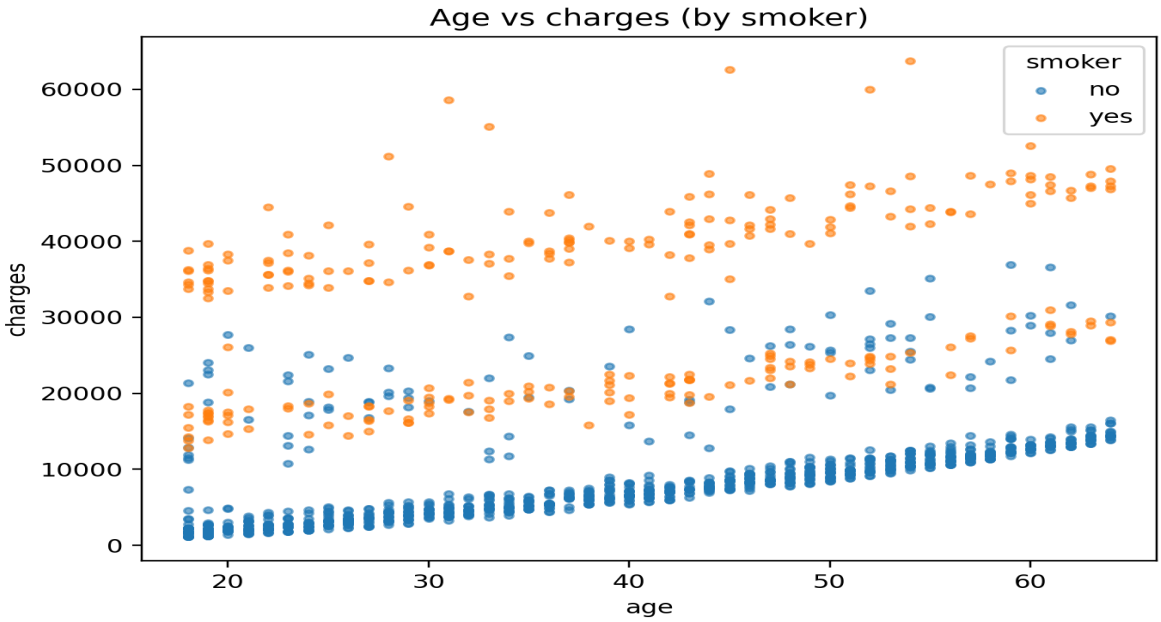

Figure 4. Age vs charges (by smoker).

4. Data Validation

A schema validation layer is applied in both training and serving to ensure consistent data quality. It checks required columns, categorical domains, and numeric ranges before making predictions.

Field	Type	Constraints / Allowed values
age	int	0-120
sex	str	female | male
bmi	float	10-70
children	int	0-20
smoker	str	no | yes
region	str	northeast | northwest | southeast | southwest
charges (target)	float	> 0 (training data)

5. Preprocessing and Feature Engineering

Categorical variables are one-hot encoded. Numeric variables are imputed (median) and standardized. Feature engineering adds BMI category, age group, smoker indicator, and interaction terms (BMI x smoker, age x smoker). All transformations are included in a single sklearn Pipeline that is saved and reused during inference.

6. Model Training

Data is split into train and test sets using an 80/20 split with random_state=42. Models trained: Linear Regression (required), Ridge Regression, Random Forest Regressor. The best model is selected based on R2 on the test set and saved for deployment.

7. Evaluation and Results

Evaluation uses MAE, MSE, and R2 on the test set. Selected model: random_forest.

Model	MAE	MSE	R2
random_forest	2430.35	1.98e+07	0.8728
ridge	2767.23	2.06e+07	0.8671
linear_regression	2762.71	2.07e+07	0.8667

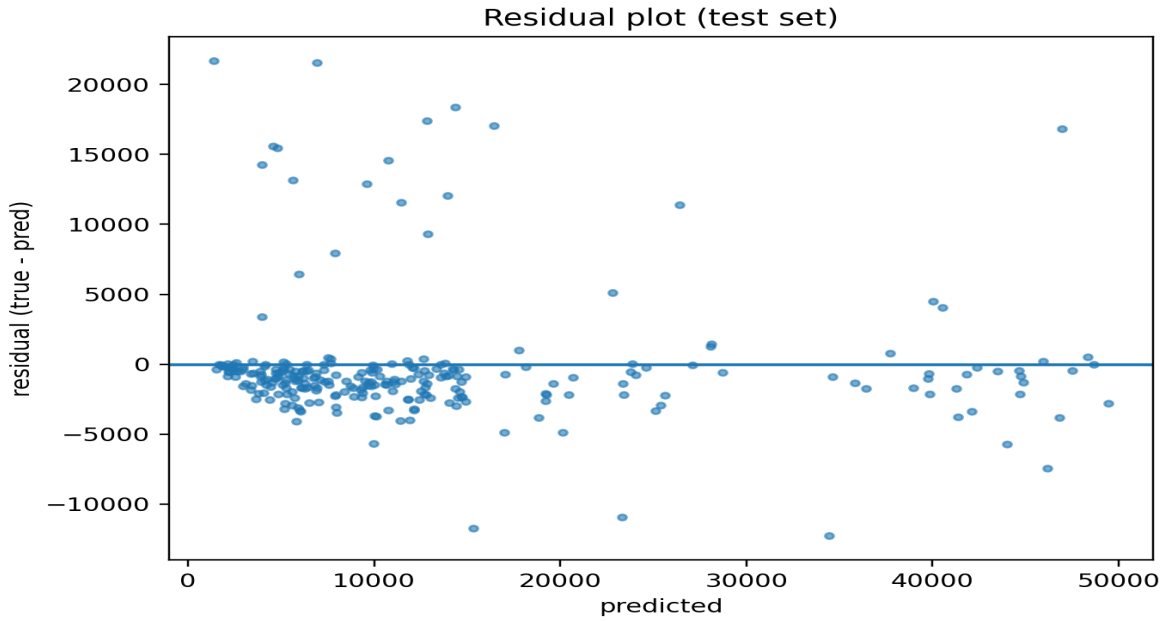

Figure 5. Residual plot (test set).

7.1 Feature importance (Random Forest)

Feature importances are computed from the trained Random Forest on the transformed feature space. The BMI x smoker interaction is the dominant driver, consistent with EDA observations.

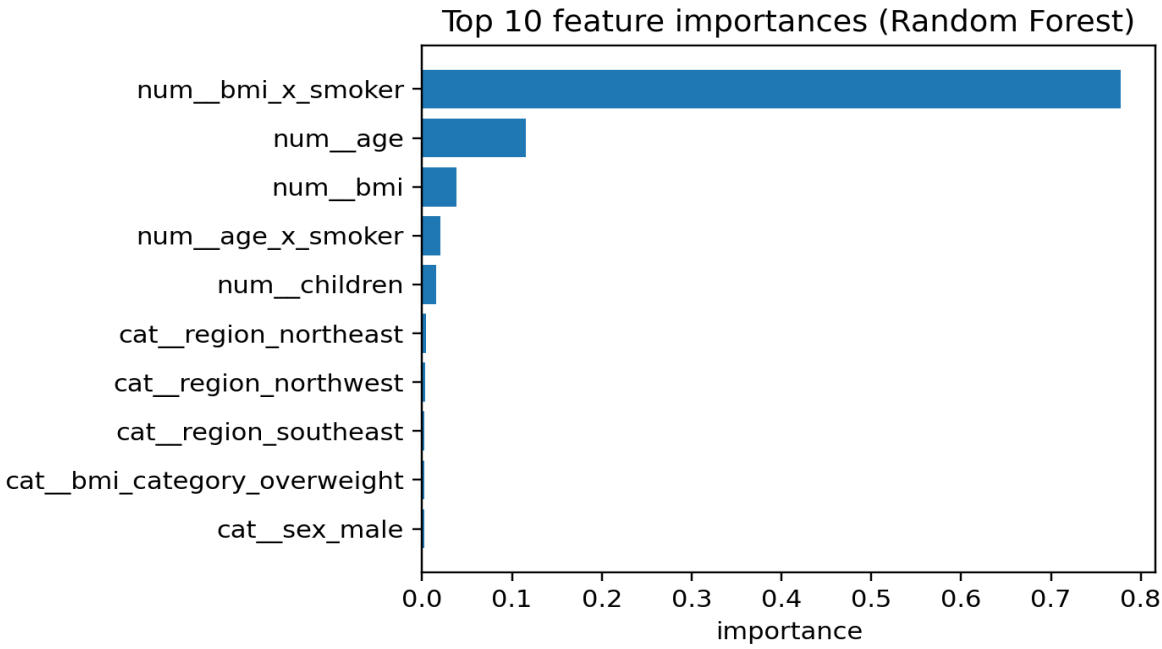

Figure 6. Top 10 feature importances.

Feature (after preprocessing)	Importance
num__bmi_x_smoker	0.7776
num__age	0.1154
num__bmi	0.0381
num__age_x_smoker	0.0202
num__children	0.0153
cat__region_northeast	0.0044
cat__region_northwest	0.0038
cat__region_southeast	0.0028
cat__bmi_category_overweight	0.0027
cat__sex_male	0.0026

8. Deployment

The best pipeline is deployed with FastAPI. The main endpoint POST /predict accepts JSON input and returns a JSON response with the predicted charges. A Streamlit app provides a simple form-based interface that calls the API.

Endpoints
- GET /health
- POST /predict
- POST /predict_batch

Example request
{"age":30,"sex":"male","bmi":28.0,"children":0,"smoker":"no","region":"southeast"}

9. Limitations

The dataset is relatively small and may not represent real-world underwriting or pricing rules. The model captures correlations, not causality, and performance is estimated from a single train-test split.

10. Future Work

Future improvements include cross-validation and tuning, trying boosting models, adding prediction intervals, and deploying with monitoring/logging for model drift.

11. Conclusion

This project demonstrates a complete end-to-end Machine Learning workflow for predicting medical insurance charges, from exploratory data analysis and data validation to model training, evaluation, and deployment. Among the evaluated models, Random Forest achieved the best performance on the held-out test set, indicating it captures non-linear patterns between the input features and charges. The final solution is deployed through a FastAPI endpoint and a lightweight Streamlit interface, enabling users to interact with the model in a simple and reproducible way.

Key takeaways

• A reproducible pipeline (validation + preprocessing + model) reduces serving/training mismatch.
• Random Forest performed best on this dataset under the current split and feature set.
• The API + web UI completes the product loop from user input to prediction output.