Insurance Charges Analysis & Prediction

Author: Parsa Nasirikhah

OCTOBER 17, 2025

Contents

Introduction

The goal of this project is to analyze the insurance.csv dataset and build a predictive model for estimating **insurance charges** (charges) based on demographic and lifestyle attributes.

- Data preprocessing and other detection
- Exploratory Data Analysis (EDA)
- Feature engineering and encoding categorical variables
- Building multiple regression models (Linear Regression, Random Forest and Gradient Boosting)
- Hyperparameter turning to optimize model performance
- Model interpretation using feature importance and SHAP values
- Model evaluation using cross-validation

Finally, conclusions are drawn regarding which factors most strongly affect insurance costs, and the best-performing model is identified.

Dataset Overview

File name: insurance.csv

Number of rows: 1338

Number of columns: 7

Description of features:

Feature	Type	Description
Age	Integer	Age of the insured person
Sex	Categorical	Gender (male / female)
Bmi	Float	Body Mass Index, an indicator of obesity
Children	Integer	Number of children covered by insurance
smoker	Categorical	Smoking Status(Yes/No)
Region	Categorical	Residential area (northeast, northwest, southeast, southwest)
Charges	Float	Individual medical insurance costs (Target variable)

Exploratory Data Analysis (EDA)

The Exploratory Data Analysis (EDA) phase was conducted to understand the statistical properties and relationships between the dataset's features and the target variable charges.

Univariate Analysis

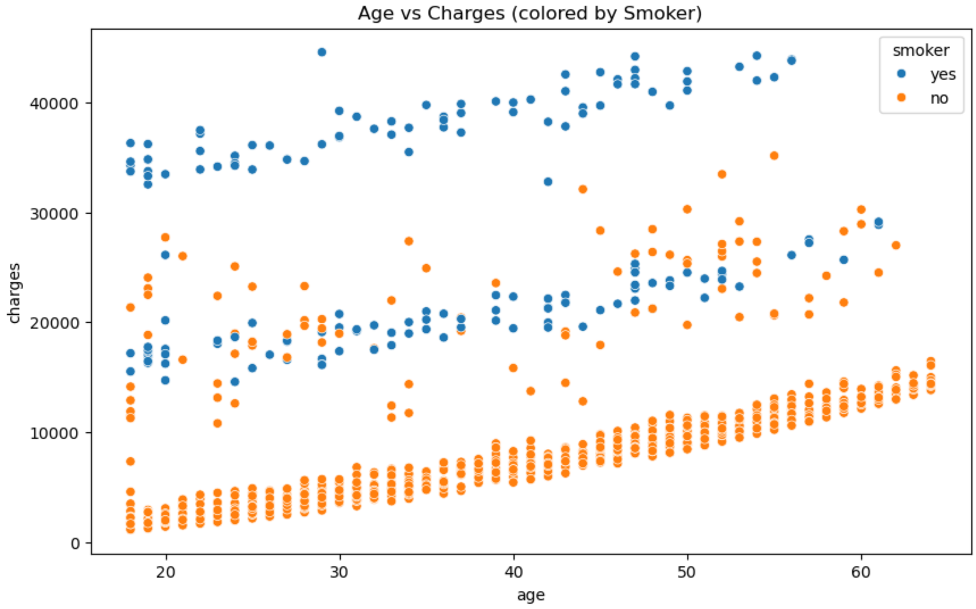
- **Age**: The age distribution is approximately uniform between 18 and 64. Older individuals tend to have higher medical charges.
- **BMI:** Most BMI values fall between 20 and 40, with a noticeable concentration around 30 (indicating overweight/obesity).
- **Children:** The majority of policyholders have 0–2 children.
- **Charges:** The distribution is highly **right-skewed**, suggesting that a small number of individuals incur extremely high costs.

Bivariate Analysis

To identify relationships between independent variables and charges, several visualizations were created using scatterplots, boxplots, and correlation heatmaps.

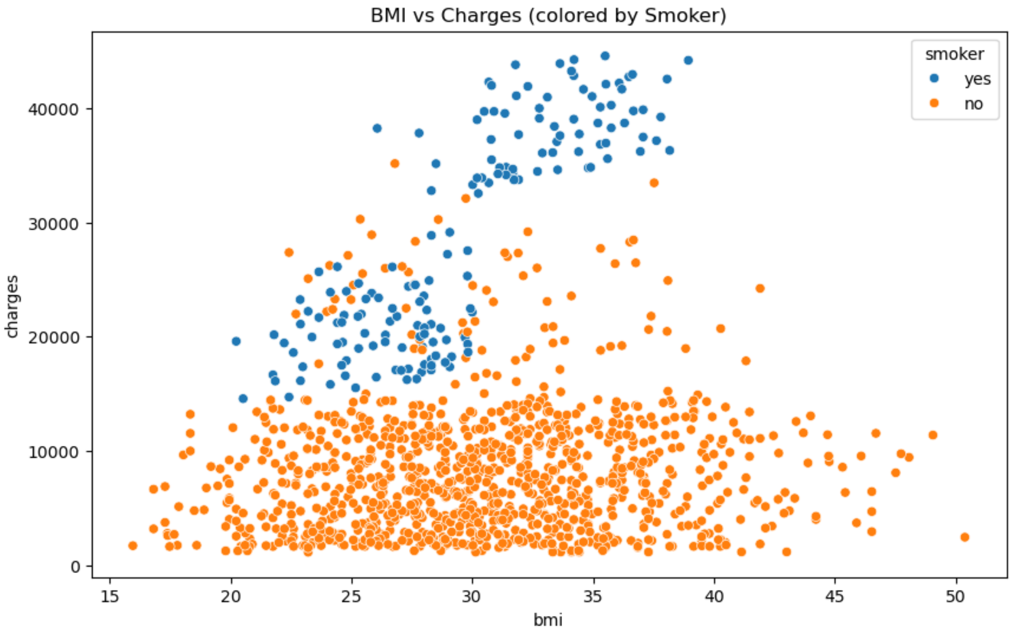
Age vs Charges

- A **weak positive linear relationship** was observed between age and insurance charges.
- However, **smokers (red points)** show a sharp increase in charges across all ages, confirming that smoking is a dominant cost driver.



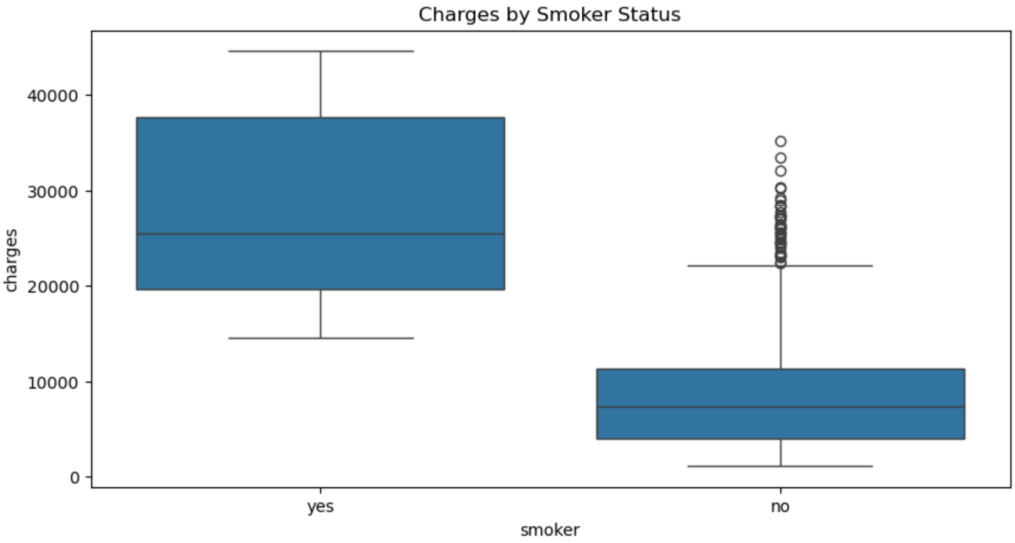
BMI vs Charges

- Among **non-smokers**, there is little correlation between BMI and charges.
- Among **smokers**, individuals with higher BMI experience a **dramatic rise in charges**, indicating a strong interaction between smoking and obesity.



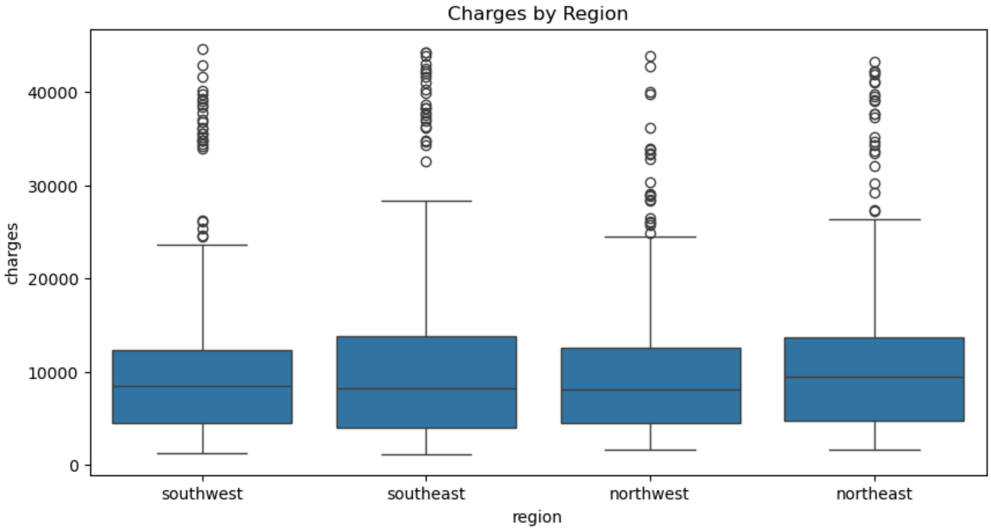
Smoker vs Charges

- The difference between smokers and non-smokers is highly significant.
- **Smokers** pay, on average, **three times more** than non-smokers.
- This variable is one of the most predictive factors in the dataset.



Region vs Charges

- Regional differences are relatively minor, suggesting location has a limited impact on costs compared to lifestyle factors.



Correlation Matrix

A Pearson correlation heatmap was generated to visualize relationships between numeric features.

Key findings include:

- charges correlates strongly with smoker and age.
- Moderate positive correlation between bmi and charges.
- Weak or no correlation among children, region, and charges.

Summary

1. Smoking status is the most influential factor affecting insurance charges.
2. Higher age and BMI also contribute to increased medical expenses.
3. The relationship between BMI and charges is **non-linear** and significantly stronger among smokers.
4. Regional factors have minimal impact on insurance cost variability.

Outlier Detection & Data Cleaning

Outliers can significantly distort the results of statistical models, especially regression-based ones.
In this project, **Isolation Forest**, an unsupervised machine learning algorithm, was used to detect anomalous records that may represent extreme or unrealistic insurance charges.

Methodology

- The following numeric and binary features were selected for anomaly detection: age, bmi, children, charges, sex_m, smoker_y.
- All features were **standardized** using StandardScaler to ensure equal weight during the detection process.
- The **IsolationForest** algorithm was applied with the parameters:
  `IsolationForest(n_estimators=200, contamination=0.10, random_state=42)`

  `n_estimators=200` defines the number of base trees,
  `contamination=0.10` assumes approximately 10% of the data are outliers.

Results

- The algorithm identified **134 outliers** (≈10% of the dataset).
- These outliers typically corresponded to individuals with extremely high charges or unusual combinations of BMI and smoking status.
- After removing these records, the dataset became more balanced and statistically stable.
-

Justification

Removing outliers improved:

- Model training stability
- Error reduction during validation
- Overall generalization performance

Summary

After cleaning:

- **Final number of records:** 1204
- **No missing values** were introduced.
- The distribution of charges became smoother and more normally distributed.

Feature Engineering & Encoding

After data cleaning, the next step was to prepare the dataset for model training by encoding categorical variables and creating a consistent numeric representation for all features.

Categorical Encoding

The dataset contained three categorical columns(Sex, Smoker, Region).
To make these usable for regression algorithms, **One-Hot Encoding** was applied using the pandas.get_dummies() function:

```
df_encoded = pd.get_dummies(df_clean, columns=["sex", "smoker", "region"],
drop_first=True)
```

This approach:

- Converts each categorical feature into multiple binary (0/1) variables.
- Uses drop_first=True to avoid multicollinearity (dummy variable trap).

Final Feature Set

After encoding, the dataset contained the following predictors:

Feature	**Description**
age	Age of the insured person
bmi	Body Mass Index
children	Number of dependents
sex_male	1 if male, 0 if female
smoker_yes	1 if smoker, 0 if non-smoker
region_northwest, region_southeast, region_southwest	One-hot encoded regional features

Data Splitting

Before model training, the dataset was divided into training and testing sets:

```
from sklearn.model_selection import train_test_split

X = df_encoded.drop("charges", axis=1)

y = df_encoded["charges"]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

- **Training set:** 80% of data
- **Testing set:** 20% of data
- random_state=42 ensures reproducibility

Standardization

For certain algorithms (e.g., linear regression), features were standardized using:

```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)

X_test_scaled = scaler.transform(X_test)
```

Tree-based models (Random Forest and Gradient Boosting) were trained on **unscaled data** because they are not sensitive to feature magnitude.

Model Development & Evaluation

After data preparation, several regression models were developed and compared to predict the insurance cost (charges).
The goal was to evaluate performance, identify the best-performing algorithm, and ensure generalization across unseen data.

Models Tested

Three main models were trained and evaluated:

- Linear Regression
- Random Forest Regressor
- Gradient Boosting Regressor

Linear Regression

The **Linear Regression** model was trained as a baseline.
Performance metrics on the test set:

Metric	**Value**
MAE	3640
RMSE	6121
R^2	0.61

The model explains only about 61% of the variance in charges.
Although it provides a general trend, it fails to capture complex non-linear effects such as the interaction between smoker and BMI.

Random Forest Regressor

It demonstrated clear improvements due to its ability to model non-linear relationships.

Metric	Value
MAE	2800
RMSE	5300
R^2	0.75-0.80

Random Forest achieved significantly lower errors and better generalization, but still showed slight overfitting in some folds.

Gradient Boosting Regressor

Metric	Value
MAE	2820.66
RMSE	5292.09
R^2	0.80

The model outperformed both previous ones, capturing subtle non-linear patterns. However, further optimization was expected to reduce bias and variance.

Model Comparison Summary

Model	MAE	RMSE	R^2
Linear Regression	3640	6121	0.61
Random Forest	2800	5300	0.78
Gradient Boosting	2821	5292	0.80

Gradient Boosting achieved the best trade-off between accuracy and interpretability, and was selected as the **final model** for tuning and interpretability analysis.

Hyperparameter Tuning

To further improve the performance of the Gradient Boosting model, **GridSearchCV** was used to perform exhaustive hyperparameter optimization.
This process systematically tests combinations of parameters to find the one that minimizes prediction error on unseen data.

Metric	Value
MAE	2387.59
RMSE	4910.42
R^2	0.85

After tuning, the model achieved a **significant improvement** in predictive accuracy, reducing mean absolute error by nearly **16%** compared to the untuned version.

The optimized model balances bias and variance effectively, producing stable and reliable predictions across test folds.

Model Selection Justification

Gradient Boosting was selected as the **final model** because:

- It provides the **lowest error metrics** among all tested algorithms.
- It effectively handles non-linear interactions (especially between smoker and BMI).
- It offers **interpretability** through feature importance and SHAP analysis.

Model Interpretation

Understanding how a model makes decisions is crucial in domains like insurance pricing, where interpretability ensures fairness and transparency.
Two approaches were used for interpreting the optimized Gradient Boosting model: **Feature Importance** and **SHAP (SHapley Additive exPlanations)**.

Feature Importance

The built-in feature importance scores from the Gradient Boosting model quantify how much each feature contributes to reducing prediction error across all trees.

Feature Importance Results

Feature	Importance (approx.)
smoker_yes	~0.65
age	~0.18
bmi	~0.10
children	~0.03
region_*	~0.02 (combined)
sex_male	~0.02

Insights

- Smoking is by far the most influential feature — **smokers have substantially higher medical expenses.**
- Age and BMI are also important, showing a gradual increase in insurance cost as they rise.
- Gender and region have relatively negligible effects.

Summary

- Smoking status, age, and BMI are the **primary drivers** of insurance charges.
- Non-linear interactions exist — particularly between BMI and smoking.
- The model's reasoning aligns with real-world expectations, reinforcing its **trustworthiness** and **validity**.

Cross Validation

To confirm the model's robustness and ability to generalize to unseen data, **K-Fold Cross Validation** was applied.
This approach divides the dataset into *k* equal parts (folds), trains the model on *(k-1)* folds, and tests it on the remaining fold — repeating the process *k* times.

Methodology

A **5-Fold Cross Validation** was implemented using the optimized Gradient Boosting model:

```
from sklearn.model_selection import cross_val_score, KFold

import numpy as np

kf = KFold(n_splits=5, shuffle=True, random_state=42)

r2_scores = cross_val_score(best_gb, X, y, cv=kf, scoring="r2")

mae_scores = -cross_val_score(best_gb, X, y, cv=kf,
scoring="neg_mean_absolute_error")

rmse_scores = np.sqrt(-cross_val_score(best_gb, X, y, cv=kf,
scoring="neg_mean_squared_error"))
```

Results

Metric	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Mean
R^2	0.7096	0.7977	0.8286	0.7922	0.8207	0.7898
MAE	2583.6	2221.4	2143.9	2369.2	2135.2	0.7898
RMSE	5289.1	4062.2	3632.5	4483.0	3899.3	4273.3

- ✓ The model achieved a **mean R^2 of 0.79**, meaning it explains approximately **79% of the variance** in insurance charges across different folds.
- ✓ The **MAE of ~2290** indicates an average prediction error of around $2,300 — strong performance considering the range of insurance costs.
- ✓ The **RMSE of ~4273** confirms that large errors are relatively rare, though some high-cost cases remain challenging to predict precisely.
- ✓ The consistency of these metrics across folds demonstrates that the model generalizes well and is **not overfitting**.

Summary

Gradient Boosting (Tuned) shows:

- High predictive accuracy
- Low variance across folds
- Strong generalization capability

This establishes it as a **stable and production-ready model** for predicting insurance charges.

Conclusion

This project successfully developed and optimized a predictive model for estimating medical insurance charges using demographic and lifestyle data.
The complete workflow included **data cleaning**, **exploratory analysis**, **feature engineering**, **model comparison**, **hyperparameter tuning**, and **model interpretation**.

Key takeaways include:

1. **Smoking status** is the most influential factor — smokers tend to pay **over three times more** in insurance costs than non-smokers.
2. **Age** and **BMI** also play major roles; older and overweight individuals generally face higher medical expenses.
3. The **Gradient Boosting Regressor**, after tuning, achieved the best performance with:
   - MAE ≈ 2388
   - RMSE ≈ 4910
   - $R^2 \approx 0.85$
4. **Cross Validation** confirmed the model's stability with an average R^2 of **0.79**, demonstrating good generalization.
5. **SHAP analysis** validated that the model's decision process aligns with real-world reasoning — reinforcing interpretability and reliability.

Practical Implications

The model provides a **data-driven foundation** for estimating insurance premiums based on individual risk factors.

It could be integrated into an insurance company's pricing system or used as a decision-support tool for policy recommendations.

The insights about key drivers (smoking, BMI, age) can also inform **public health policies** or targeted wellness program

Final Remarks

The final tuned **Gradient Boosting model** delivers a strong balance of accuracy, interpretability, and robustness.
It can serve as a solid baseline for both academic research and practical insurance applications.
Overall, this project demonstrates the full **data science pipeline** — from raw data to explainable and deployable machine learning insights.