Predicting Insurance Charges

A Linear Regression Model

Emre Şentürk

123200133

13.11.2025

Project Overview

Using machine learning to predict medical insurance costs based on patient attributes.

The Dataset: insurance.csv

Key Features (Independent)

- Age
- Sex (male/female)
- BMI (Body Mass Index)
- Children
- Smoker (yes/no)
- Region

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

Target (Dependent)

Charges (Insurance Cost in $)

What is Linear Regression?

Concept

A statistical method that models the relationship between variables by fitting a linear equation to the data. It finds the "best-fit line" that minimizes the error between predicted and actual values.

Our Goal

Our goal is to find the weights (coefficients) for each feature (age, smoker, etc.) to build an equation that can accurately predict the final 'charges'.

Mathematical Formulation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \varepsilon$$

- y is the predicted **'charges'**.
- x_1, x_2... are the features (**age**, **bmi**, **smoker**).
- β_1, β_2... are the feature coefficients (weights) the model learns.
- β_0 is the intercept, or the baseline cost.

Training Process: Data Prep

Encoding

Converted categorical data (sex, smoker, region) into numerical values using One-Hot Encoding.

Splitting

Split the data into a training set (80%) to teach the model and a test set (20%) to evaluate it.

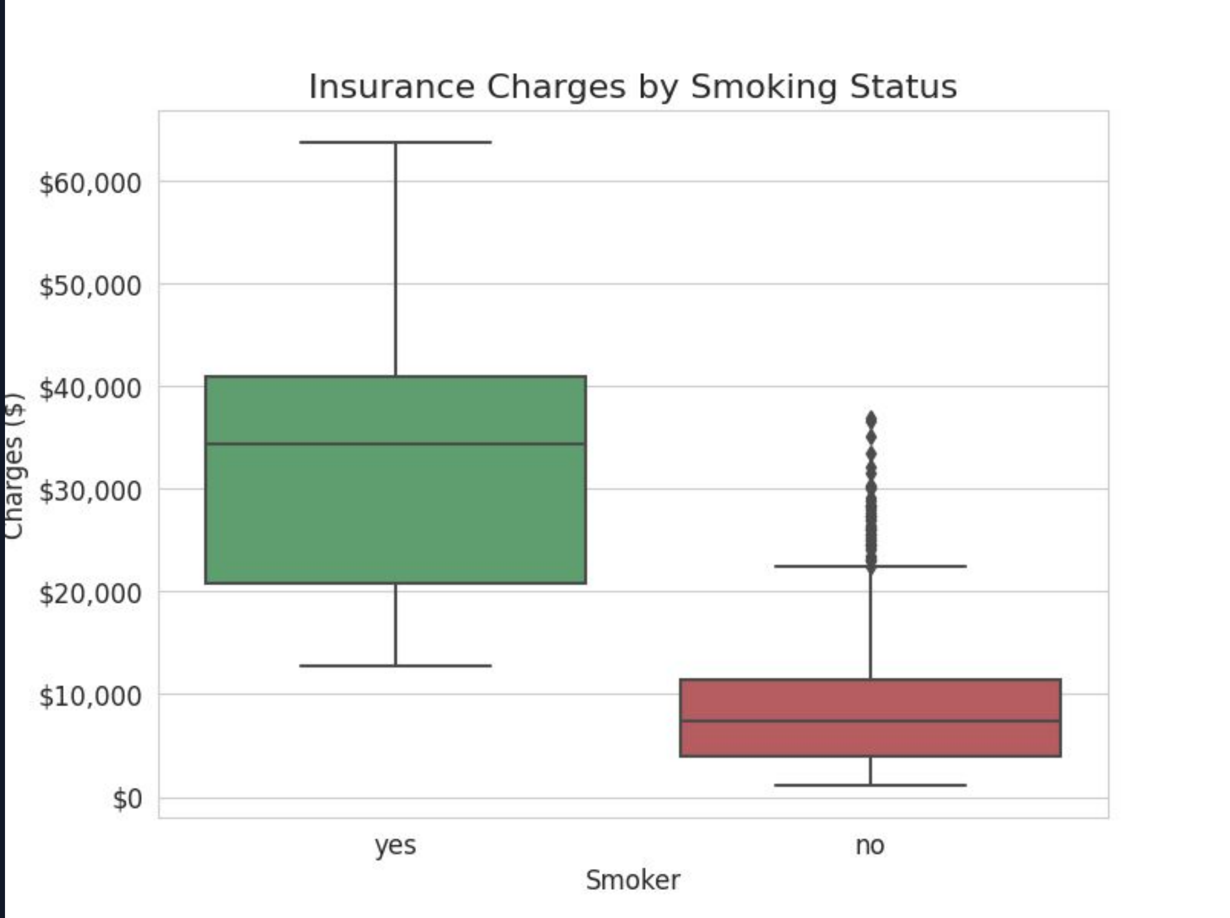
Scaling

Normalized numerical features (like age, BMI) using `StandardScaler` so they are on a common scale.

Inference Mechanism

How does it predict a new cost?

1. The new patient's data (e.g., age, bmi, smoker=yes) is encoded and scaled just like the training data.
2. The scaled features are plugged into the trained models equation.
3. The model multiplies each feature by its learned coefficient, adds the intercept, and outputs the final predicted charge.



Model Evaluation: R-Squared (R^2)

0.784

R-Squared Score

What it means

This means our model can explain approximately **78.4%** of the variance in the insurance charges.

This is a good score, indicating a strong relationship between our chosen features and the final cost.

Model Evaluation: MAE

$4181

Mean Absolute Error

What it means

On average, the model's predictions are off by about **$4,181** on the test set.

This is the average monetary error per prediction, which gives us a real-world sense of the model's accuracy.

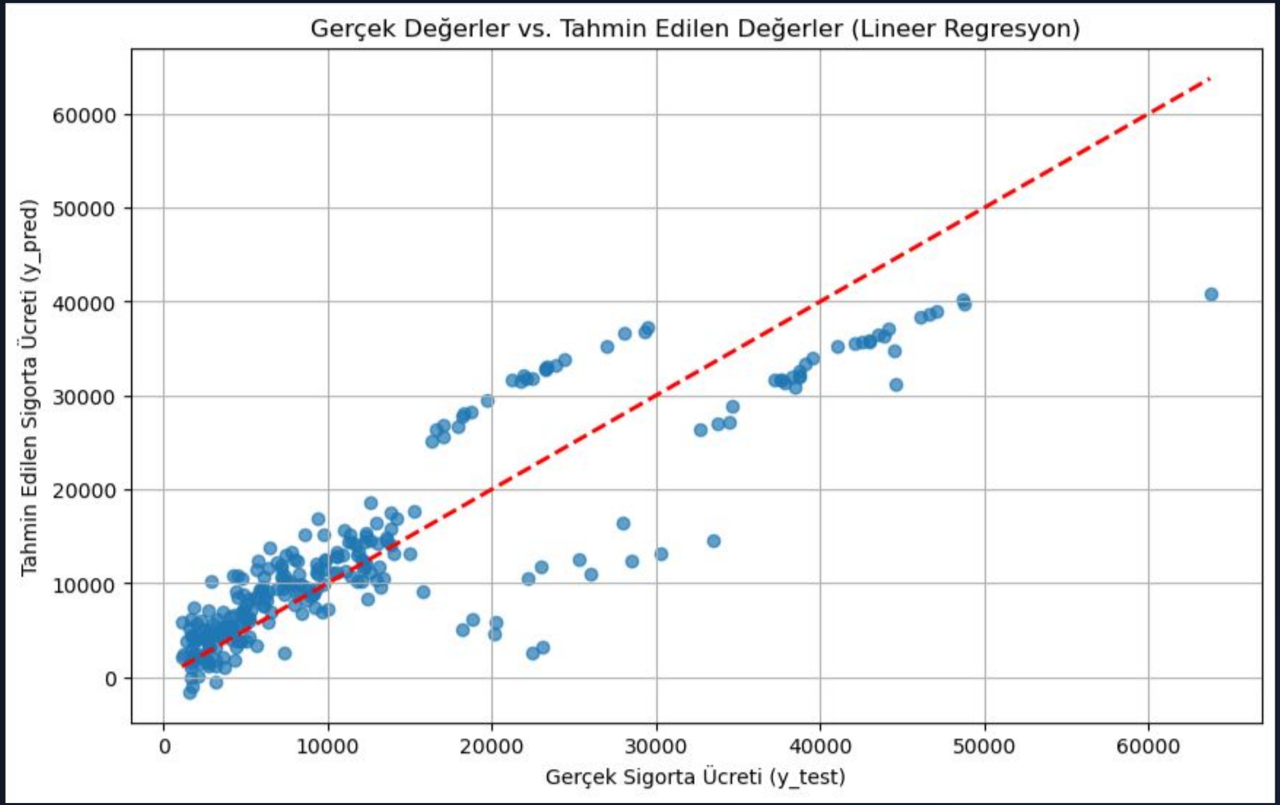
Illustration: Actual vs. Predicted

Visual Results

This scatter plot shows the model's performance on the test data.

- The X-axis is the **Actual** cost.
- The Y-axis is the **Predicted** cost.
- Points on the red line are perfect predictions.

Observation: The model follows the trend well, but struggles with very high-cost patients (the points at the top).



Conclusion

- The Linear Regression model achieved a respectable **R^2 of 0.784**.
- The model is effective at predicting low-to-mid-range costs but is less accurate for high-cost outliers.
- This confirms that features like age, BMI, and especially 'smoker' status are strong predictors of medical costs.