Predicting Medical Insurance Costs: A Linear Regression Case Study

From Raw Data to an Actionable Prediction Model

Algorithm: Linear Regression
Tools: Python, Pandas, Scikit-Learn, Seaborn, Matplotlib

Setting insurance premiums is a high-stakes balancing act.

The Manual Approach

Traditional estimation is slow, inconsistent, and struggles to capture complex risk factors. The process is **time-consuming** and prone to **human error.**

Time-consuming, prone to error

inconsistent, and struggles to capture complex risk factors

Complexity and confusion

The Automated Solution

To build an automated model that **predicts** individual medical costs **accurately** and **instantly** using patient data.

Build an automated model

predicts individual medical costs accurately

instantly using patient data

Our investigation began with data from 1,338 beneficiaries.

Age
(Numeric: 18-64 years)

Sex
(Categorical: Male/Female)

BMI
(Numeric: Body Mass Index)

Children
(Numeric: Dependents)

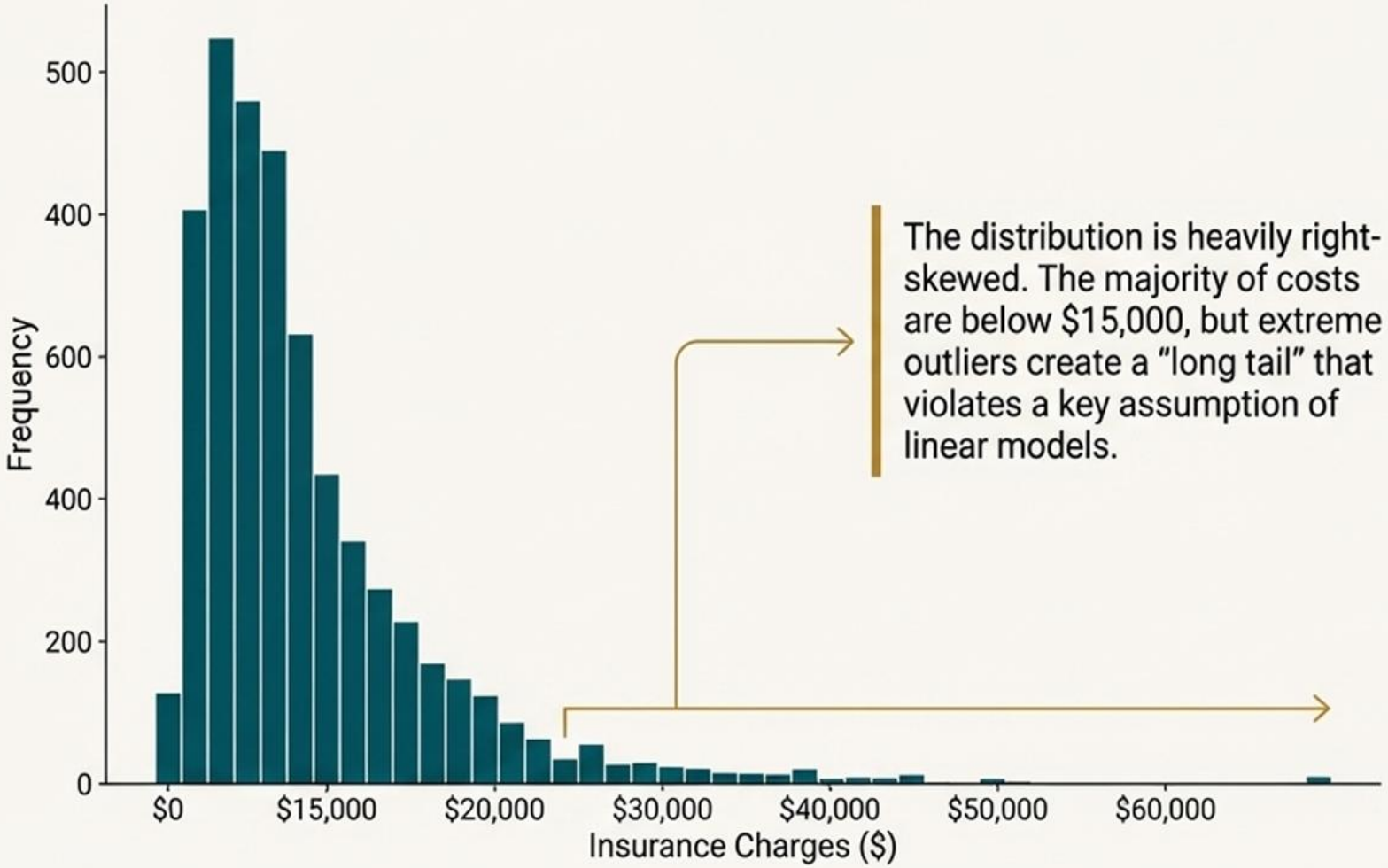
Smoker
(Categorical: Yes/No)

Region
(Categorical: SW, SE, NW, NE)

Charges
(Numeric: Target Variable)

The Value to Predict.

A few high-cost cases were distorting the entire picture.



The distribution is heavily right-skewed. The majority of costs are below $15,000, but extreme outliers create a "long tail" that violates a key assumption of linear models.

Decision: Apply a Log Transformation.

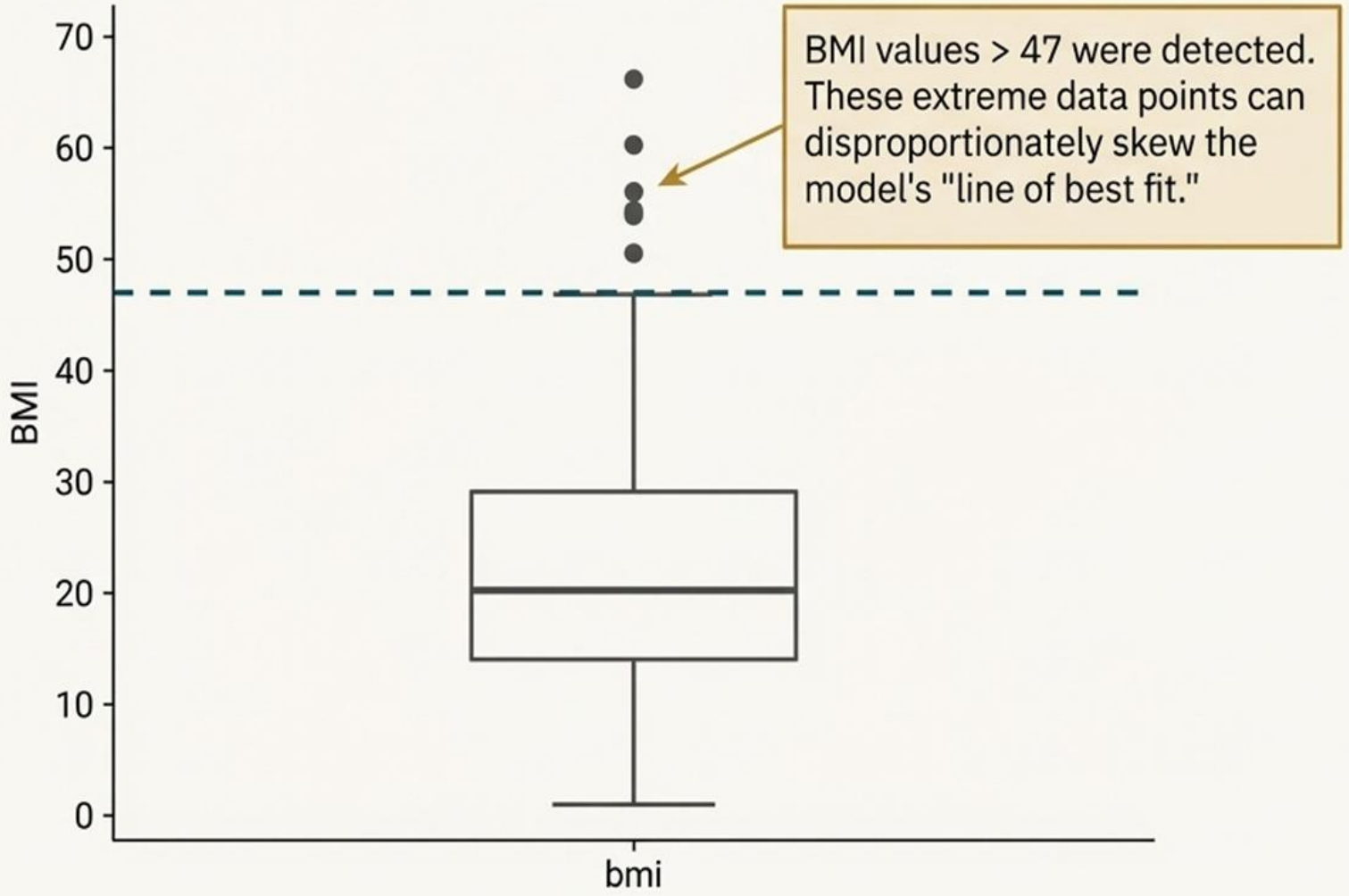
Why?: To normalize the distribution. This compresses the high-cost outliers and prevents the model from being disproportionately influenced by rare, extreme claims, improving its performance on more common cases.

It also helps satisfy the "Homoscedasticity" assumption required by Linear Regression.

The data clearly showed one behavior drives costs more than any other.

Correlation analysis confirmed that the `smoker` feature has the strongest relationship with with insurance `charges`, followed by `age` and `bmi`.

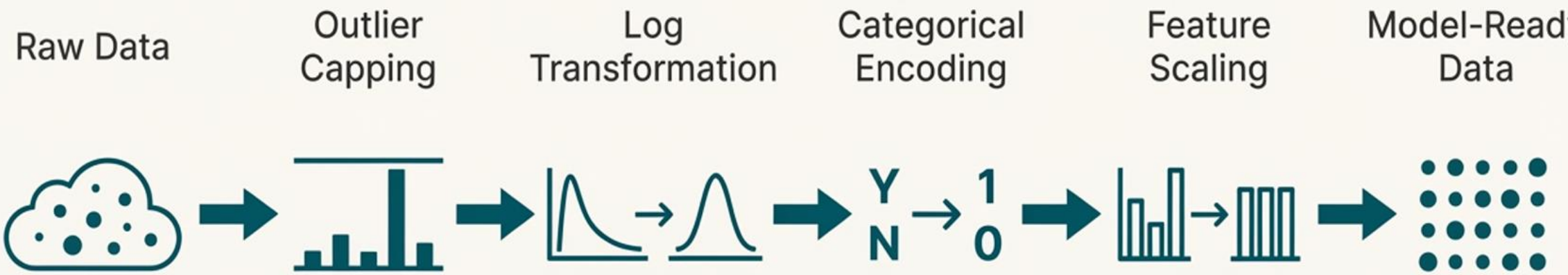
We identified and corrected for unrealistic BMI values.



Decision: Cap outliers at the 99th percentile (approx. 47). This technique is also known as **Winsorization.**

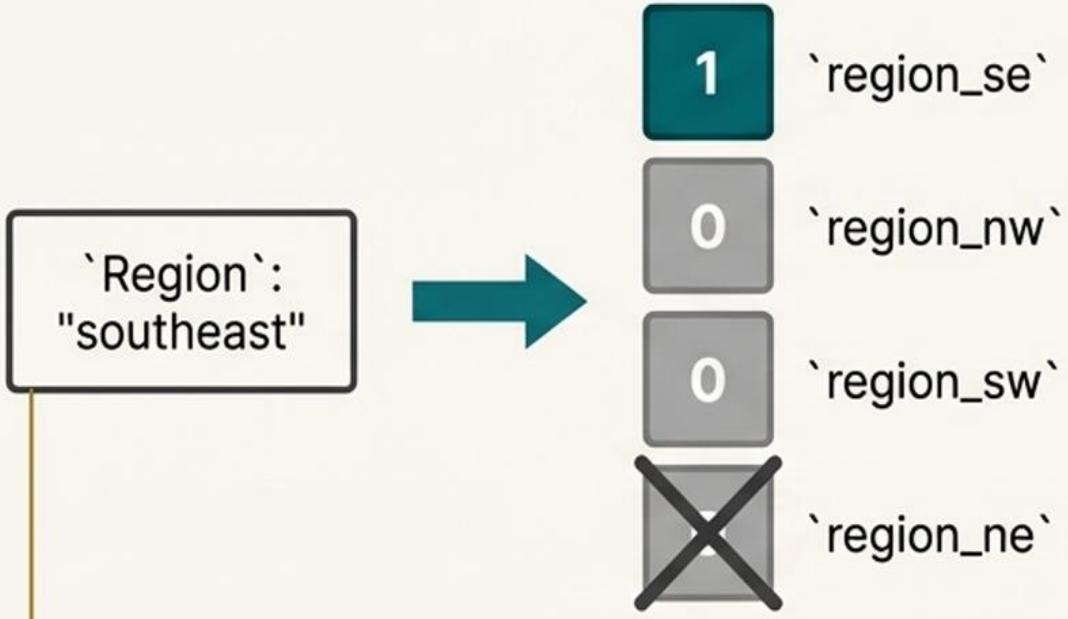
Why?: To prevent extreme, and possibly erroneous, data points from degrading the model's predictive accuracy for the general population.

We systematically translated raw data into a language the model understands.



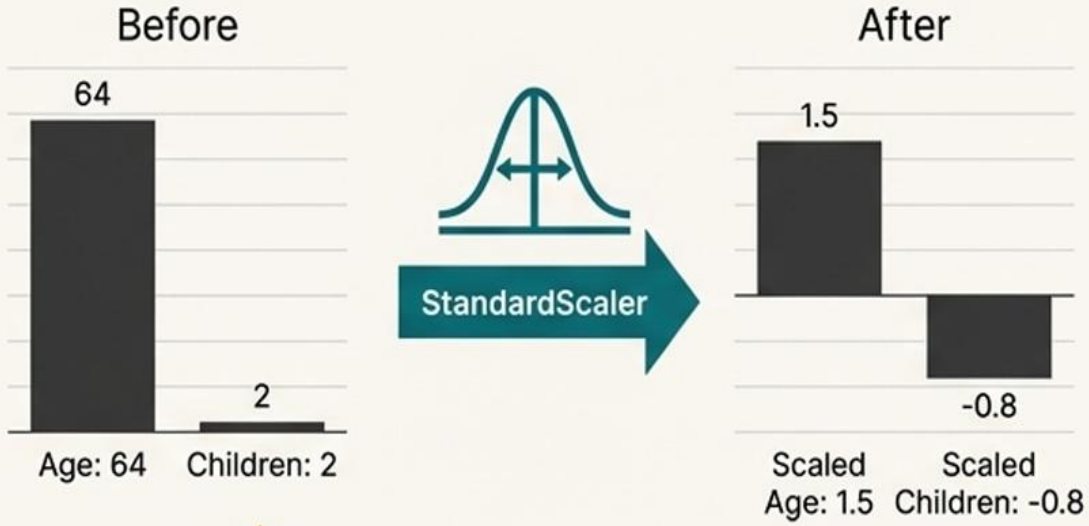
Every feature was meticulously calibrated for the model.

Translating Categories into Numbers



We used One-Hot Encoding for `Region` and applied `drop_first=True` to avoid the 'Dummy Variable Trap' (Multicollinearity), ensuring model stability.

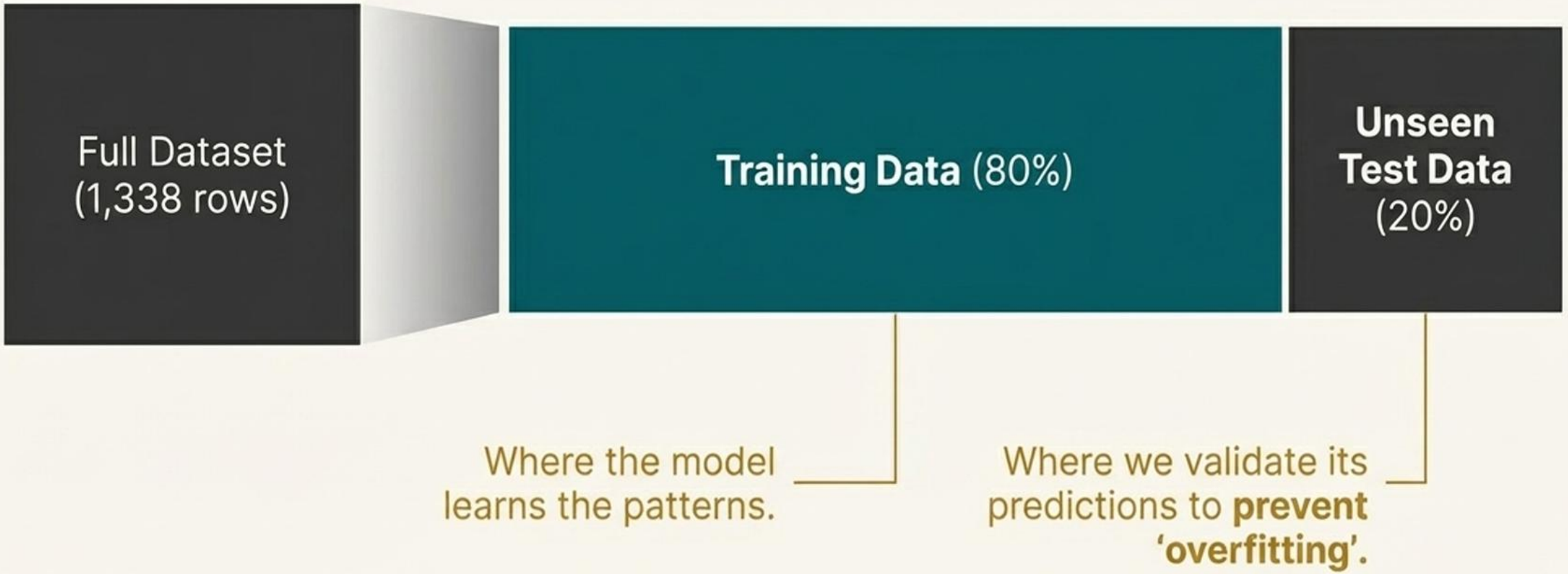
Standardizing Numerical Scales



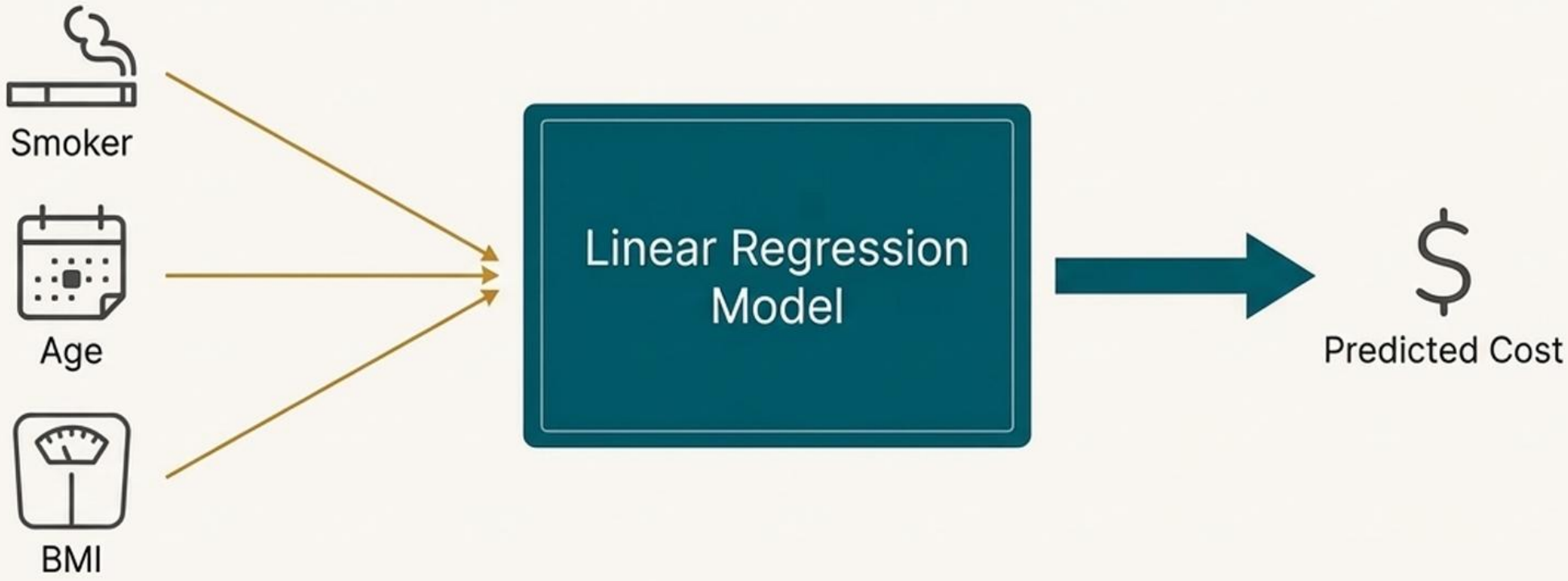
Why Scale?

To prevent the model from incorrectly assuming features with larger numeric ranges (like `age`) are inherently more important than features with smaller ranges (like `children`). Scaling ensures all features are judged on the same footing.

We held back 20% of the data to rigorously test the model's real-world performance.



Our model learns a simple formula to connect patient traits to cost.



The model assigns a “weight” (coefficient) to each feature, quantifying its impact on the final cost. Afte ɑimpule-positify tracter in the final cost. After training, the model assigned the highest positive coefficient to the `Smoker` feature, confirming it ꙅ it as the most significant factor.

$$\text{Predicted Cost} \approx (w_1 * \text{Smoker}) + (w_2 * \text{Age}) + (w_3 * \text{BMI}) + \ldots + \text{intercept}$$

The model can explain 78% of the variation in medical costs.

R^2 Score

0.78

Our model successfully explains 78% of the variance in medical costs. This is considered a strong result for a linear model in this domain.

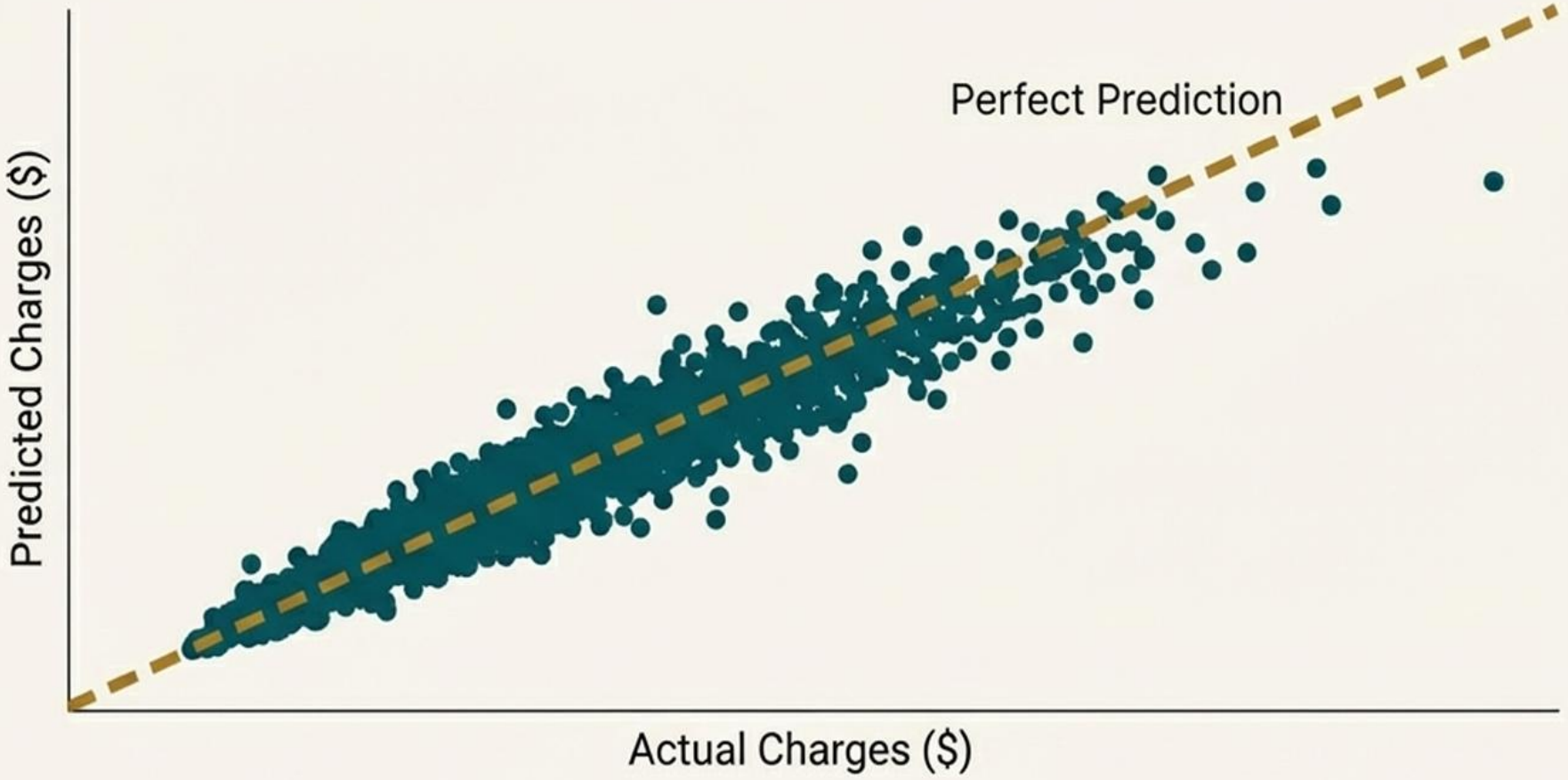
Mean Absolute Error (MAE)

~$4,200

On average, our model's predictions are within $4,200 of the final hospital bill.

Important Note: Predictions were made on the 20% unseen test data. The original predictions were in log form and converted back to real dollars using `np.expm1()` before calculating the error.

Our model's predictions closely track the actual insurance charges.



The tight clustering of points around the line demonstrates the model's strong predictive power on data it has never seen before.

The model provides instant cost estimates for any given profile.

We successfully built a robust prediction pipeline from start to finish.

1. **Key Insight:** Smoking, age, and BMI are the dominant drivers of medical costs. A model that ignores them will be inaccurate.

2. **Critical Process:** Thorough data exploration and preprocessing (handling skewness, outliers, and scaling) are non-negotiable for building an accurate model.

3. **Strong Result:** Linear Regression provides a powerful, interpretable, and reliable baseline for cost prediction, achieving a 78% R^2 score on this dataset.

The next step is to capture more complex, non-linear relationships.

Pushing Accuracy Beyond 80%

We recommend exploring advanced, non-linear models like **Gradient Boosting (XGBoost)** or **Random Forest**.

Why This Works

These models can automatically capture complex interactions that linear models cannot, such as the combined effect where a high BMI is significantly more costly *only if* the individual is also a smoker.

Project Repository: [Medical-Insurance-Cost-Prediction-Using-LR]

Thank You