Medical Cost Prediction: A Regression Analysis Journey

Project Overview

This project is an end-to-end machine learning application that predicts individual medical costs billed by health insurance. The goal was not only to build an accurate predictive model but also to understand the key factors that drive insurance charges. This analysis follows an iterative modeling process, starting with a simple baseline and progressively building a more complex and accurate model.

Methodology & Findings

The project progressed through three key modeling stages, each revealing deeper insights into the data.

Stage 1: Baseline Model (Simple Linear Regression)

A baseline model was established using a single feature, Body Mass Index (bmi), to predict medical charges.

- **Result:** The model yielded a very low **R-squared of ~4%**.
- **Conclusion:** This confirmed that bmi alone is insufficient for making accurate predictions and that a more comprehensive set of features was necessary.

Stage 2: Enhanced Model (Multiple Linear Regression)

The model was improved by including all relevant features: age, bmi, children, sex, smoker, and region. Categorical features were numerically encoded using One-Hot Encoding.

- **Result:** The model's performance improved dramatically, achieving an **R-squared of ~78%**.
- **Analysis of Coefficients:** By inspecting the model's coefficients, it was determined that being a **smoker was the single most significant factor**, increasing predicted costs by over $23,000. age and bmi were also strong positive predictors.
- **Model Limitation:** An "Actual vs. Predicted" plot revealed that the model's predictions became less accurate for higher-cost individuals, suggesting a non-linear relationship in the data.

Stage 3: Advanced Model (Polynomial Regression)

To address the limitations of the linear model, **Polynomial Features (degree 2)** were engineered. This technique creates interaction terms and squared features (e.g., age^2, age * bmi), allowing the linear model to fit more complex, non-linear patterns.

- **Result:** This advanced model achieved the highest performance, with an **R-squared of ~85%**.
- **Conclusion:** The significant increase in accuracy confirms that medical costs have a non-linear relationship with patient attributes. The Polynomial Regression model successfully captured these complex patterns, making it the best-performing model for this task.

Model Performance Summary

Model	R-squared Score	Key Takeaway
Simple Linear Regression	~4%	A single feature is not enough.
Multiple Linear Regression	~78%	Multiple features provide good predictive power.
Polynomial Regression	**~85%**	**Capturing non-linearity is key to high accuracy.**

This iterative process demonstrates a complete analytical workflow: starting simple, identifying model weaknesses through diagnostics, and applying advanced techniques to build a robust and highly accurate predictive model ready for deployment.