Predicting Diabetes Progression: A Machine Learning Approach

Balancing Interpretability and Predictive Power using a Regression Framework

by Valerio Quaranta

Project Objectives & Dataset Overview

The Mission: Predicting Clinical Outcomes The primary goal of this project is to develop a robust predictive model capable of forecasting diabetes disease progression one year after baseline. By leveraging physiological data, we aim to provide a tool that could assist in early patient screening and risk stratification.

The Dataset: A Clinical Snapshot We utilized the classic "Diabetes" dataset, which provides a controlled environment with 442 patient records. Each record includes 10 key variables: age, sex, Body Mass Index (BMI), average blood pressure (BP), and six blood serum measurements (s1–s6).

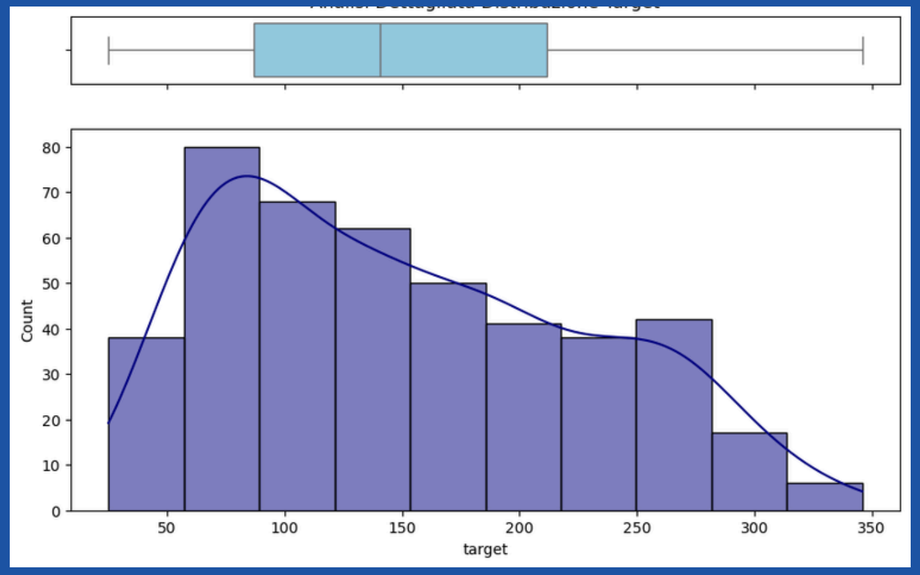
Setting the Scene

Smart Preprocessing from the Start Unlike raw clinical data, this dataset comes pre-scaled and centered. This is a strategic advantage: it ensures that variables with naturally different scales—like age versus blood pressure—contribute equitably to the model. This prevents larger numerical values from dominating the learning process and simplifies the interpretation of our final coefficients.

Ensuring Data Integrity A rigorous preliminary check confirmed a "clean" environment: no missing values were found, and the absence of significant outliers ensures that our error-based cost functions (like MSE) remain stable and are not skewed by anomalous data points.

Analyzing the Target Variable

The Shape of Progression The target variable—a quantitative measure of disease progression—shows a relatively smooth distribution with a slight positive skewness of **0.44**. Since this value falls well within the -0.5 to 0.5 range, the data is symmetric enough to proceed without complex transformations, ensuring our predictions remain on the original clinical scale.



Stability and Lack of Outliers

Our **IQR** (Interquartile Range) analysis confirmed a zero-outlier environment.

From a modeling perspective, this is a green light for linear methods: it guarantees that our loss function won't be hijacked by extreme, unrepresentative cases, leading to more reliable "typical patient" estimates.

Defining the Cost of Error

We intentionally selected **Mean Squared Error (MSE)** as our compass. In a clinical context, a large miss is significantly more dangerous than several small ones. By squaring the residuals, we heavily penalize large discrepancies, forcing the model to be especially cautious with high-risk predictions.

Statistical Benchmarks With a mean progression of 152.1 and a standard deviation of 77.1, we have established a clear boundary. Any model worth its salt must achieve an error significantly lower than this standard deviation to prove it is providing actual clinical insight rather than just guessing the average.

The Strategic Foundation

The "Safety First" Split

To ensure the integrity of our findings, we immediately partitioned the data into **Training (75%)** and **Testing (25%**) sets. By doing this before any deep correlation analysis, we avoided "Decision Leakage"—the risk of making modeling choices based on information the model isn't supposed to see yet.

Why a Linear Baseline?

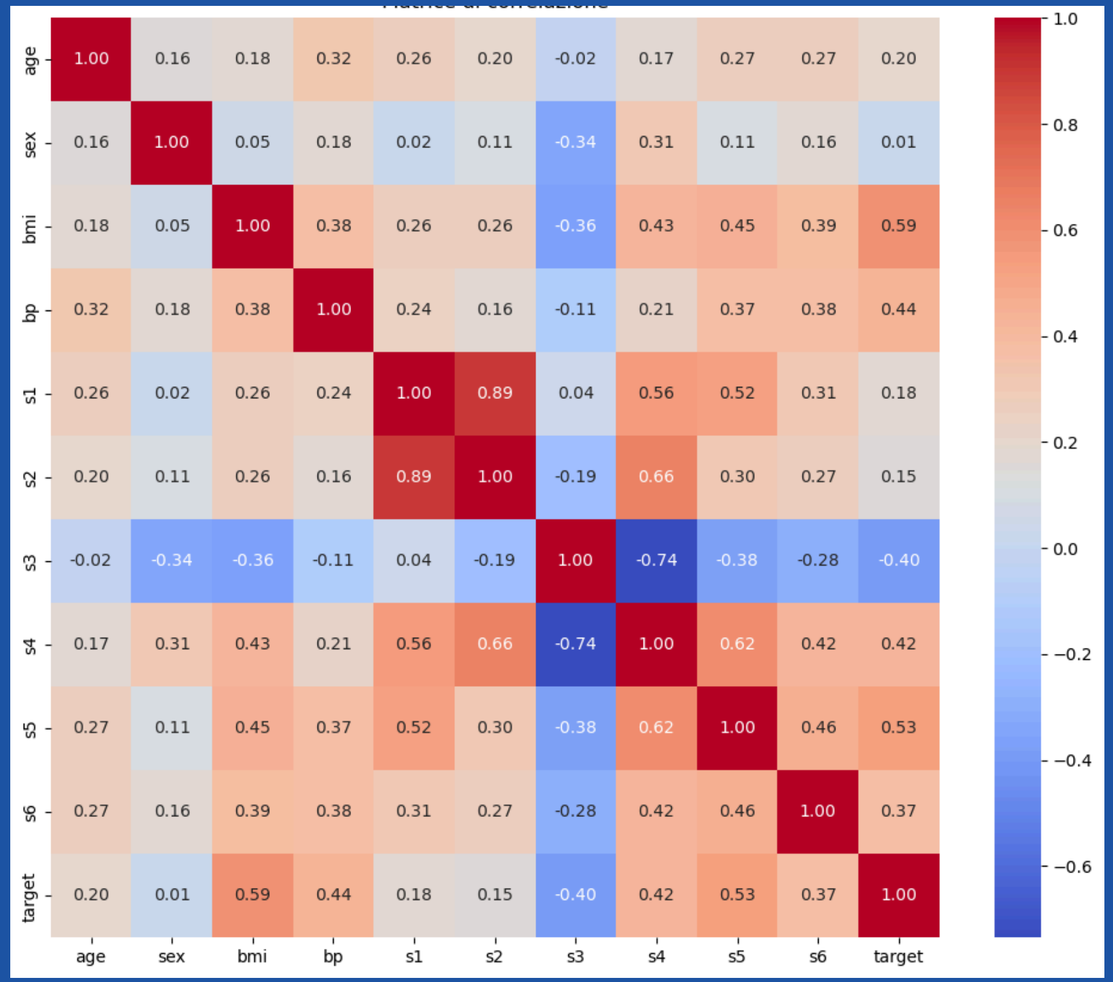
We chose a standard **Linear Regression** as our starting point. Given that the features are already centered and normalized, a simple linear approach provides the perfect plain benchmark. If a more complex model can't beat this, it isn't worth the extra complexity.

Initial Benchmarks

The baseline delivered an **RMSE of 53.37** and an **R2 of 0.485.**

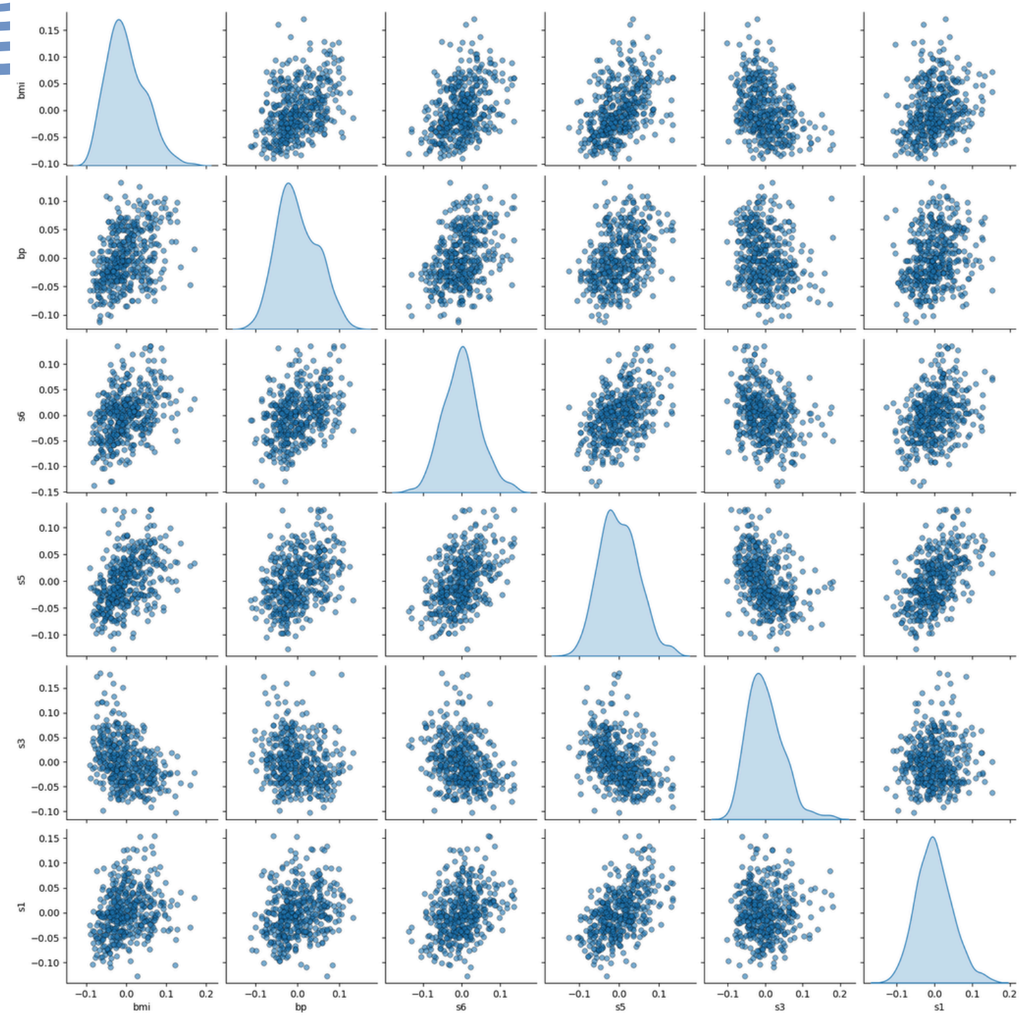
- **The Good News**: An RMSE of 53 is substantially lower than the target's standard deviation (77), proving that even a simple linear model captures significant clinical signals.
- **The Challenge**: Explaining only 48.5% of the variance leaves room for improvement through regularization and feature analysis.

Feature Interaction and the Multicollinearity Trap



Our correlation heatmap revealed a high-density web of relationships, particularly among blood serum measurements. The extremely high correlation between **s1** and **s2** (0.89) suggests they are nearly identical signals, which can cause standard linear models to panic and assign erratic weights.

The Hidden Complexity



Mutual Information (MI) Insights

While some variables are redundant, others are powerhouses. **BMI** and **s5** showed the highest MI scores, identifying them as the most consistent drivers of disease progression. This contrast between redundant and vital features led us to a critical crossroad: **Ridge vs. Lasso.**

The VIF Diagnostic (Variance Inflation Factor)

To quantify this, we calculated VIF scores. Seeing values jump as high as **54.3 (s1)** and **35.6 (s2)** was a definitive red flag. In statistics, any VIF over 10 indicates severe multicollinearity that can undermine the reliability of our coefficients.

Choosing the Right Strategy

A Regularized Regression

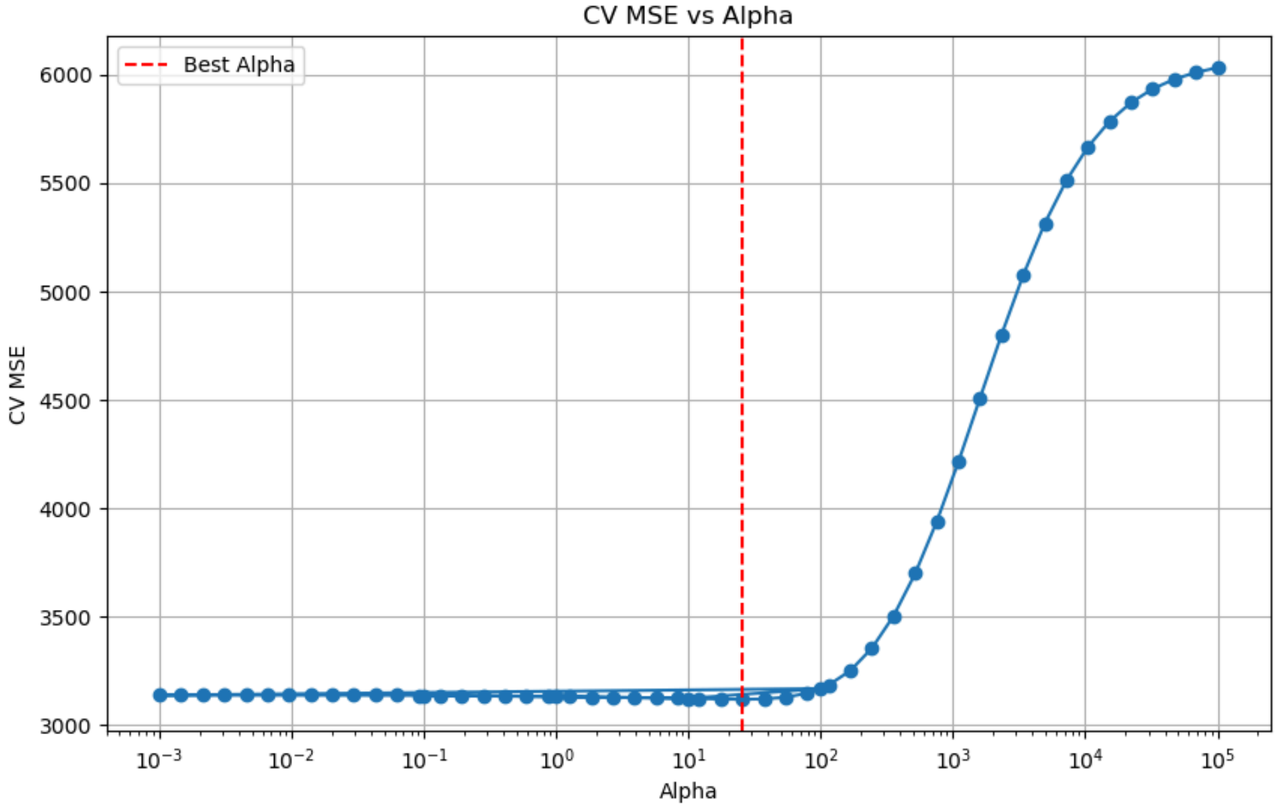
Faced with high multicollinearity and a small dataset (N=442), we needed a model that could handle redundancy without losing information. We chose **Ridge Regression (L2 Regularization)** over Lasso. While Lasso might have zeroed out correlated features, Ridge keeps them all but shrinks their coefficients, preserving the subtle clinical nuances of the blood serum data.

To prevent any hint of data leakage, we encapsulated the workflow into a **Scikit-Learn Pipeline**. This ensures that the scaling process (StandardScaler) is strictly informed by the training data and only applied to the test data, maintaining the "sanctity" of our final evaluation.

With a limited sample size, a single split isn't enough. We employed **Repeated K-Fold Cross-Validation** (10 splits, 3 repeats). This stress-tests the model 30 different times on different data subsets, giving us a highly reliable estimate of how it will perform in the real world.

Hunting for the Optimal Alpha

The Power of Logarithmic Search We didn't just guess the penalty strength (alpha). We performed an exhaustive **Grid Search** across a logarithmic scale. The search revealed that the optimal alpha was 25.60—a value found only by exploring the gaps between standard integers.



The Bias-Variance Tradeoff The CV-MSE curve tells a clear story:

- **Low Alpha** (10^1): The error is flat; the model behaves like standard OLS, struggling with multicollinearity.
- **High Alpha** (10^2): The error skyrockets; the penalty is too harsh, leading to underfitting as the model loses its ability to see patterns.
- **The Red Line**: Our chosen Alpha balances complexity and stability, sacrificing a tiny bit of training fit for a massive gain in reliability.

Performance Showdown

The Numerical Comparison

Metric	Baseline	**Optimized Ridge**
RMSE (Lower is better)	53.37	**53.09**
R^2 Score (Higher is better)	0.485	**0.490**

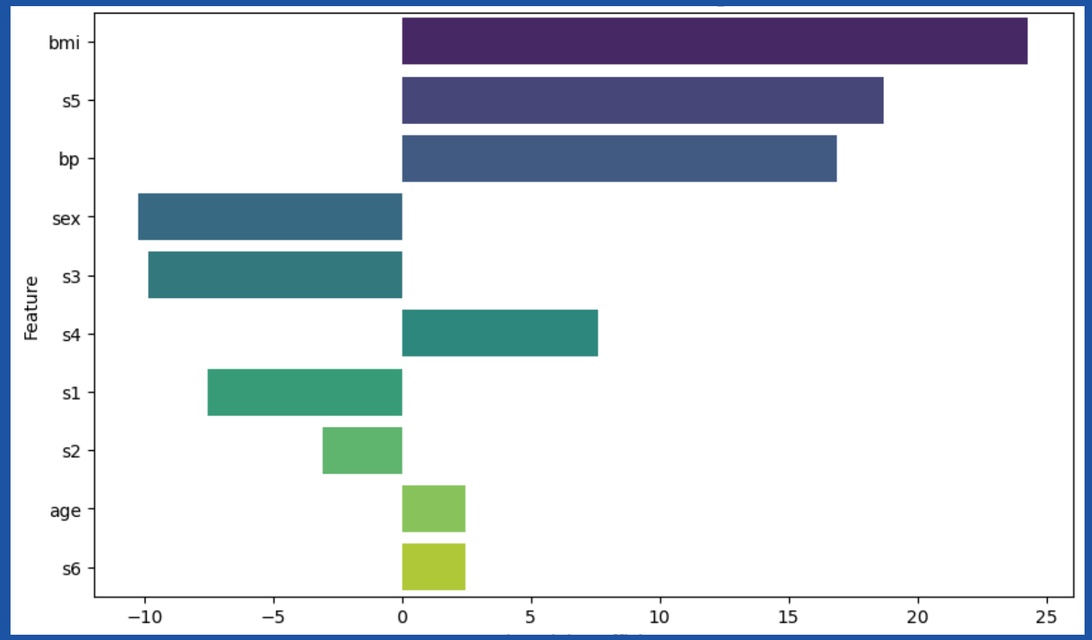
Proving the Strategy While the numerical gains may appear subtle, the transition from a standard LR to a tuned Ridge model represents a crucial shift from an unstable model to a robust one. We achieved a lower error and a better fit, justifying the complexity of our pipeline.

Interpreting the Small Gain In high-stakes fields like medicine, a reduction in RMSE—even of 0.3 points—is meaningful when it comes from stabilizing the model.

Generalization Strength The most impressive result is the consistency. With an R^2 of **0.51** on Training and **0.49** on Testing, we have successfully achieved a model that has learned the underlying patterns without over-adjusting to the training noise.

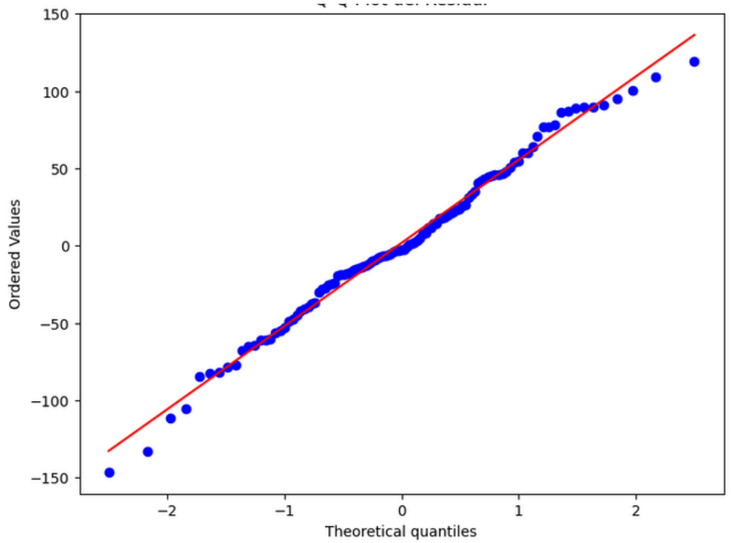
Decoding the Model's Brain

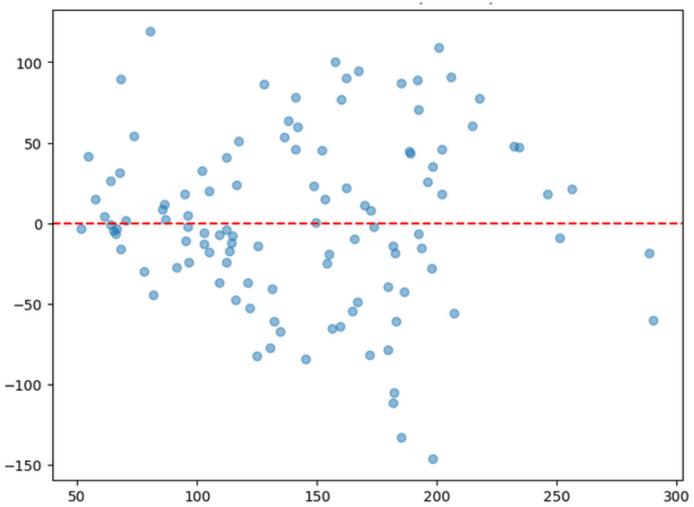
- **BMI: The Leading Indicator** Consistent with clinical literature and our Mutual Information test, BMI emerged as the strongest predictor. The model confirms that as body mass increases, the progression of diabetes follows a predictable, upward linear trend.
- **The "s5" Mystery** The high weight assigned to s5 (likely representing triglycerides or similar lipids) highlights its importance, but our VIF analysis warns us to be cautious. Because s5 is correlated with other serum markers, its weight is the result of complex interactions rather than a solo performance.
- **The Case of "s6"** Interestingly, the Ridge model downplayed s6 despite its decent MI score. This reveals a limitation of L2 regularization: in its quest to stabilize the whole system, it may "dampen" certain features that might be vital if analyzed in isolation. This is a key takeaway for future medical refinements.



Under the Hood: Residual Diagnostics

The **Q-Q Plot** shows that our residuals are mostly normal (following the diagonal). However, scatters at the extreme ends suggest the model struggles with "Edge Cases"—patients with exceptionally low or high disease progression.

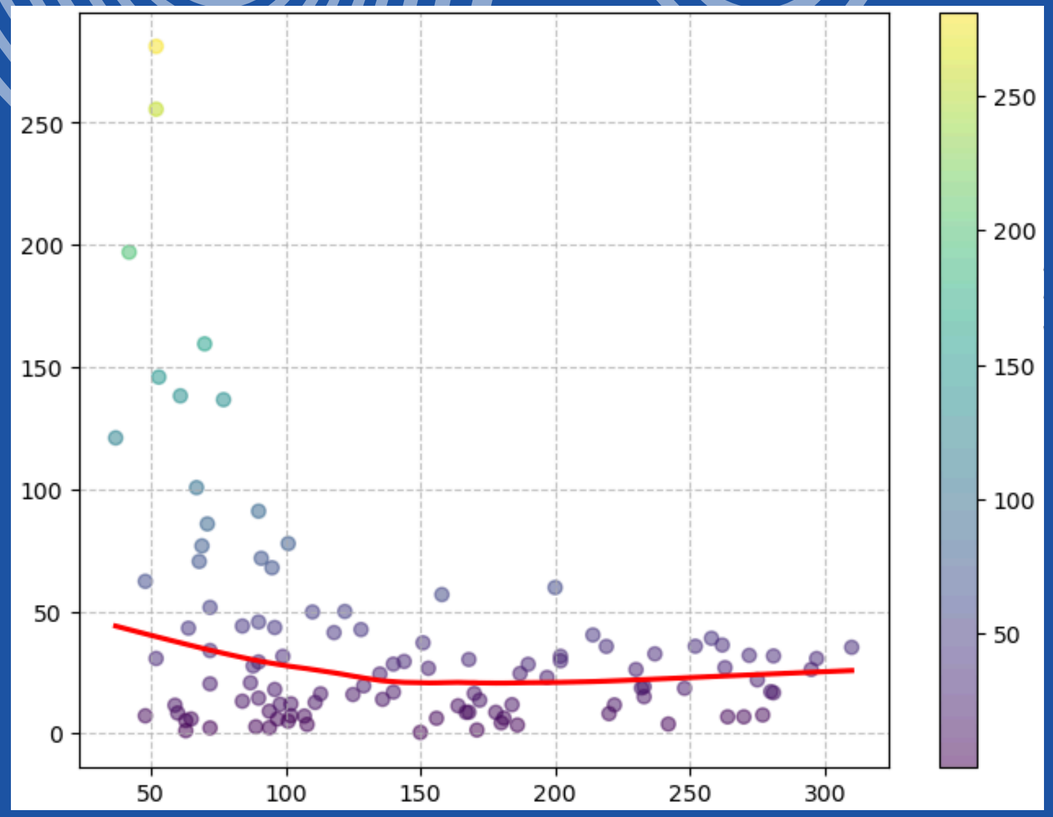




We didn’t just look at the score; we tested the health of our errors. The **Breusch-Pagan test (p-value: 0.031)** officially rejected homoscedasticity. This means our model's error variance isn't constant—it changes depending on the patient's physiological profile.

With a mean residual of **1.80**, the model is nearly unbiased on average. It doesn't systematically over- or under-predict, making it a "fair" judge for the general population, even if it wavers on complex outliers.

Clinical Utility & Error Scaling



The Reality of 49% Variance While an R^2 of 0.49 might seem modest in physics, in the noisy world of medicine, it is a significant achievement. We have moved the needle substantially compared to the baseline and the naive mean

.**Relative Precision vs. Absolute Error:**

- **MAE (41.5):** This represents only 12.93% of the total disease progression range.
- **The Critical Insight:** Our percentage error plot reveals that as the disease becomes more severe (Target > 150), the model becomes more precise.

A Targeted Screening Tool Because the relative error stabilizes for high-risk patients, this model is most reliable exactly where it's needed most: identifying patients who require aggressive therapeutic intervention.

Final Verdict & Future Horizons

The "Toy Dataset" Constraint We must acknowledge that this dataset is a controlled environment. Real-world clinical data is often messier, with missing values and non-linear interactions that a simple Ridge model might overlook.

Key Takeaways

1. **BMI and s5** are the undeniable drivers of this model.
2. **Regularization** was mandatory to solve the serum multicollinearity (VIF > 50).
3. **Generalization** is high: the 2% gap between Train and Test R^2 proves the model is not just memorizing data.

Next Generation Improvements

- **Non-Linearity**: Introducing polynomial features could resolve the heteroscedasticity we detected.
- **Ensemble Methods**: Exploring XGBoost or Random Forest to capture complex patient clusters.
- **Feature Engineering**: Creating interaction terms between blood pressure and serum markers to unlock deeper insights.

THANK YOU

Link to the code: Notebook (on Github)