Data Analysis Report: Regression Modeling on the Diabetes Dataset

1 Introduction

This report documents the comprehensive process of developing and evaluating several regression models to predict the progression of diabetes in patients. The primary objective is to build an accurate and interpretable model by systematically exploring feature preprocessing, applying regularization techniques, and engineering interaction features.

The methodology is grounded in the use of the `scikit-learn` Diabetes dataset. The analysis begins by establishing a performance baseline using standard Linear Regression and then compares its efficacy against regularized models, namely Ridge and Lasso regression. This comparison is conducted across two distinct phases: an initial phase using only the 10 original predictive features, followed by a second phase that incorporates second-degree interaction features.

By documenting each step—from data inspection to final model diagnostics—this report provides a clear narrative of the analytical journey, culminating in the selection of an optimized model that balances performance with simplicity.

2 Data Loading and Preprocessing Analysis

2.1 Dataset Characteristics

The analysis utilizes the Diabetes dataset provided by `scikit-learn`, which contains data from 442 patients.

- **Number of Instances:** 442
- **Number of Attributes:** 10 numeric predictive features
- **Target Variable:** Quantitative measure of disease progression one year after baseline

The 10 predictive features include baseline physiological and blood serum measurements (Table 1).

2.2 Analysis of Pre-Applied Normalization

The Diabetes dataset, as loaded from `scikit-learn`, has already been normalized so that each feature has a mean of 0 and a squared length of 1:

Table 1: Description of Predictive Features

age	Age (years)
sex	Sex
bmi	Body mass index (kg/m^2)
bp	Average blood pressure
s1	TC – T-Cell count (cholesterol-related)
s2	LDL – Low-Density Lipoproteins (bad cholesterol)
s3	HDL – High-Density Lipoproteins (good cholesterol)
s4	TCH – Total Cholesterol / HDL*
s5	LTG – Log of serum triglycerides level
s6	GLU – Blood sugar level (glucose)

$$x_{\text{norm}} = \frac{x_{ij} - \text{mean}(x_j)}{\sqrt{\sum (x_{kj} - \text{mean}(x_j))^2}}$$

This differs from the `StandardScaler` normalization, which uses the standard deviation in the denominator. Comparing both methods confirms that the `scikit-learn` version indeed matches the formula above.

Table 2: Statistical Properties of Different Scaling Methods

Data Version	**Mean**	**Std. Deviation**
Raw	Varies by feature	Varies by feature
Diabetes-Normalized	≈ 0	≈ 0.0476
Standard-Scaled	≈ 0	≈ 1.0

3 Exploratory Data Analysis (EDA)

EDA was conducted to identify relationships between features and detect multicollinearity.

3.1 Pair Plot Analysis

The pair plot revealed clear positive linear relationships between the target variable and `bmi` and `bp`. Kernel Density Estimates along the diagonal confirmed that the normalized data are centered near zero.

3.2 Correlation Heatmap

The correlation heatmap revealed that:

- **Top correlated features:** `bmi` (0.586), `s5` (0.566), `bp` (0.441)
- **Most negative correlation:** `s3` (-0.395)
- **High collinearity:** `s1` and `s2` ($r = 0.897$)

This justified the use of regularization (Ridge, Lasso) to manage correlated predictors.

3.3 Box Plot and Histogram Analysis

Box plots highlighted typical ranges and outliers, while histograms quantified skewness. Example: feature `s4` showed a positive skewness of 0.89.

4 Modeling Phase 1: Original Features

Three models—Linear, Ridge, and Lasso—were trained on the 10 original normalized features.

4.1 Model Performance Summary

Table 3: Model Performance on Original Features

Model	Test R^2	Key Notes
Linear Regression	0.453	Baseline model
Ridge Regression (CV)	0.457	Best $\alpha = 10.0$
Lasso Regression (CV)	0.467	Best $\alpha = 1.0$; retained 9 features

A fixed-penalty Lasso ($\alpha = 5$) reduced active features to 5 and achieved $R^2 = 0.465$.

4.2 Selected Features (Lasso, $\alpha = 5$)

1. bmi: 25.85
2. s5: 18.66
3. bp: 11.92
4. sex: -2.51
5. s3: -8.26

Table 4: Model Performance with Interaction Features

Model	**Test R^2**	**Key Findings**
Linear Regression	0.478	Slight improvement, mild overfitting
Ridge Regression (CV)	0.500	Best $\alpha = 100.0$
Lasso Regression (CV)	0.514	Best $\alpha = 1.0$; retained 32 features

5 Modeling Phase 2: Interaction Features

PolynomialFeatures (degree=2, interactions only) expanded the feature set to 55.

5.1 Selected Features (Lasso, $\alpha = 5$)

Nine non-zero coefficients were retained:

1. bmi: 25.69
2. s5: 18.94
3. bp: 11.58
4. bmi * bp: 2.47
5. age * sex: 2.07
6. bmi * s6: 1.14
7. age * bp: 0.50
8. sex: -2.52
9. s3: -7.95

6 Final Model Selection and Diagnostics

6.1 Performance

The final Linear Regression model using the 9 Lasso-selected features achieved:

$$R^2_{\text{test}} = 0.535$$

This represents a 17% improvement over the baseline Linear Regression (0.453) and validates that regularization-driven feature selection enhances both interpretability and accuracy.

6.2 Diagnostic Plots

Predicted vs. Actual: Points closely follow the diagonal line, indicating a strong linear relationship. **Residuals vs. Predicted:** Residuals are centered around zero with mild heteroscedasticity. **Q–Q Plot:** Residuals approximately follow a normal distribution. **Histogram:** Symmetric and bell-shaped residual distribution.

7 Conclusion and Reflection

This analysis demonstrates how performance improves systematically through feature engineering and regularization. The experiment highlighted the bias-variance tradeoff:

- **Reducing bias:** Polynomial interactions captured nonlinear effects.
- **Controlling variance:** Regularization prevented overfitting.

Interpretability: Despite model expansions, `bmi`, `s5`, and `bp` consistently emerged as key predictors.

The final model balances simplicity and performance, achieving $R^2 = 0.535$ with only nine interpretable features. This underscores that in applied machine learning, *meaningful simplicity often outperforms unnecessary complexity.*