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

The scaling transformation used in `StandardScaler` standardizes features to have a mean of 0 and a standard deviation of 1.

2.3 Scaling the Training Data

The scaling transformation applied to the training data is expressed mathematically as:

$$\mathbf{X}_{\text{train,scaled}} = \frac{\mathbf{X}_{\text{train}} - \boldsymbol{\mu}_{\text{train}}}{\boldsymbol{\sigma}_{\text{train}}}$$

Where:

- $\mathbf{X}_{\text{train}}$: is the original feature value from the training data.
- $\boldsymbol{\mu}_{\text{train}}$: is the mean of the feature values computed from the training data.
- $\boldsymbol{\sigma}_{\text{train}}$: is the standard deviation of the feature values computed from the training data.

2.4 Scaling the Test Data

For the test data, the scaling transformation is applied using the parameters learned from the training data:

$$\mathbf{X}_{\text{test, scaled}} = \frac{\mathbf{X}_{\text{test}} - \boldsymbol{\mu}_{\text{train}}}{\boldsymbol{\sigma}_{\text{train}}}$$

Where:

- $\mathbf{X}_{\text{test}}$: is the original feature value from the test data.
- $\boldsymbol{\mu}_{\text{train}}$ and $\boldsymbol{\sigma}_{\text{train}}$: are the mean and standard deviation computed from the training data, respectively.

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, $\alpha = 5$)

While the cross-validated Lasso model achieved the highest predictive performance, its primary purpose in that configuration is to maximize accuracy. To enhance model interpretability, a second Lasso model was fitted with a higher, fixed penalty ($alpha = 5$). The goal of this step was not to improve the R^2 score but to force more feature coefficients to exactly zero, thereby identifying a more parsimonious set of influential predictors. This model yielded an R^2 score of 0.465 and reduced the number of active features from 10 to 5. The five remaining features and their corresponding coefficients are listed below, sorted from most positive to most negative impact:

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

4.3 Reflection First Analysis

This initial modeling phase confirmed that the Lasso regression model with cross-validation gave the highest R^2 score. Unlike standard Linear Regression, Ridge and Lasso introduce

a penalty term (regularization) to shrink coefficient values, which helps prevent overfitting, especially when predictors are correlated. Ridge shrinks all coefficients towards zero, whereas Lasso can shrink them to exactly zero, effectively performing feature selection. For Lasso, a higher alpha value increases the penalty strength, resulting in a simpler model with fewer active features. With a solid baseline established, the analysis proceeds to the next phase to investigate whether performance can be further improved by introducing interaction features.

5 Modeling Phase 2: Interaction Features

Building on the insights from the first modeling phase, this section details the strategic decision to introduce interaction features. This phase investigates whether modeling the combined effects and non-linear relationships between predictors can reduce model bias and lead to a meaningful improvement in predictive accuracy. PolynomialFeatures (degree=2, interactions only) expanded the feature set to 55.

5.1 Feature Engineering

To capture potential interactions, the feature set was expanded using `scikit-learn`'s `PolynomialFeatures` class, as part of the second set of experiments. Second-degree interaction features were generated for all pairs of the original 10 predictors by setting the argument `interaction_only=True`. This process resulted in a new, expanded feature space containing **55 features**. This new set includes the original 10 predictors plus 45 unique interaction terms (e.g., age × sex, bmi × bp). This expanded feature set, including the original and interaction terms, was subsequently standardized using `StandardScaler` before being used for model training.

5.2 Model Performance Comparison with Interaction Features

The same three regression models were trained using the new 55-feature dataset. The performance of each model on the test set is summarized in the table below.

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

5.3 Feature Selection with Lasso (Expanded Set)

As in the previous phase, a Lasso model with a fixed, higher penalty (alpha = 5) was fitted to the polynomial feature set. This was done to distill the most impactful features from the expanded set and improve interpretability. This approach yielded an R^2 score of 0.490 and successfully reduced the feature space from 55 to just 9 non-zero coefficients. The nine remaining features and their coefficients are listed below, sorted by impact:

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

It is crucial to note that the top three predictors—bmi, s5, and bp—retained their primary importance and similar coefficient magnitudes across both modeling phases. The inclusion of interaction terms like bmi * bp and age * sex refined the model but did not fundamentally alter the core story, suggesting these interactions represent second-order effects.

6 Final Model Selection and Diagnostic Evaluation

After exploring models with both original and interaction features, the analysis culminated in the construction of a final, optimized model. This model was built using only the most impactful features identified through the Lasso regularization process in the previous phase. This section validates the performance of this final model and evaluates its adherence to key statistical assumptions.

6.1 Final Model Specification and Performance

The final step involved fitting a simple Linear Regression model using only the top 9 features selected by the fixed-penalty ($\alpha = 5$) Lasso model from the polynomial feature set. This approach combines the interpretability of a simple linear model with the predictive power gained from feature engineering and rigorous feature selection. This final, parsimonious model achieved a test R^2 score of **0.535**.

This final R^2 of **0.535** is the culmination of our structured approach. It represents a **17**% relative improvement over the baseline Linear Regression model (0.453) and a **4**% improvement over the best-performing model from Phase 2 (0.514). The final model's performance validates our hypothesis that a parsimonious model, built from a carefully curated set of engineered features, would outperform both the naive baseline and the overly complex, fully-featured polynomial model. This demonstrates that the optimal strategy was not just adding complexity, but using regularization as a tool for targeted feature selection from an enriched feature space.

6.2 Diagnostic Plots

Beyond a single performance metric like R^2, visual diagnostics are essential for validating a model's underlying assumptions and ensuring its reliability.

6.3 Predicted vs. Actual Plot

This plot compares the model's predicted values against the actual values from the test set. The points generally follow the 45-degree diagonal line, which, combined with the R^2 score of 0.535, indicates a reasonably good fit and a moderate positive correlation between predictions and reality.

6.3.1 Predicted vs. Actual Plot

This plot compares the model's predicted values against the actual values from the test set. The points generally follow the 45-degree diagonal line, which, combined with the R^2 score of 0.535, indicates a reasonably good fit and a moderate positive correlation between predictions and reality.

6.3.2 Residuals vs. Predicted Plot

This plot shows the model's errors (residuals) against its predicted values. The residuals are randomly scattered around the horizontal zero line, with no discernible pattern or curve. This suggests that the model's assumptions of linearity and constant variance

(homoscedasticity) largely hold. While there is a slight hint of a wider spread at higher predicted values (mild heteroscedasticity), there are no strong non-linear patterns that the model failed to capture.

6.3.3 Normality of Residuals (Q-Q Plot and Histogram)

The Q-Q plot and the histogram of residuals were used to check the assumption that the model's errors are normally distributed.

- **Q-Q Plot**: Most data points lie close to the diagonal reference line, implying the residuals follow an approximately normal distribution.
- **Histogram**: The histogram shows a symmetric, bell-shaped distribution of residuals centered at zero.

Together, these plots provide strong support for the assumption that the model's errors are unbiased and normally distributed, reinforcing the validity of the final model.

With the final model selected and validated, the report concludes with an overall reflection on the findings.

7 Conclusion and Final Reflection

This comprehensive analysis demonstrates a systematic approach to regression modeling, highlighting how performance can be iteratively improved through feature engineering and regularization. The investigation consistently showed that regularized models like Ridge and Lasso improved generalization and delivered better performance on unseen data compared to the baseline linear model.

7.1 Bias-Variance Trade-Off

The two-phase experimental design provided a clear illustration of the bias-variance trade-off.

- **Reducing Bias**: Expanding the feature space with second-degree interaction terms allowed the models to capture more complex, non-linear relationships. This reduced model bias, as evidenced by the increase in R^2 scores across all model types in Phase 2.
- **Controlling Variance**: Regularization was crucial for managing the increased complexity. By penalizing large coefficients, both Ridge and Lasso prevented the

models from overfitting to the training data, thereby controlling variance and improving performance on the test set. The final model, which used Lasso for feature selection before fitting a simple linear regression, struck an effective balance between these two competing forces.

7.2 Interpretability and Complexity

A key finding from this analysis is that despite the performance gains from interaction terms, the core set of influential predictors remained remarkably consistent across all experiments. Features such as `bmi`, `s5`, and `bp` consistently emerged as the most significant drivers of disease progression.

This observation highlights an important principle in practical machine learning: the ultimate goal is often meaningful simplicity. For this particular dataset, the relationships between the predictors and the target appear to be predominantly linear and additive. While the inclusion of polynomial interactions provided a marginal performance boost, the added complexity may not be justified when the primary insights remain unchanged. The most robust and interpretable story is told by a handful of key features, a finding that **regularization helped to clarify and confirm**.

In conclusion, the analysis successfully produced a well-performing, validated regression model and reinforced the value of a structured, iterative approach to model building.