Recursive Feature Elimination with Linear Regression
MATH/CSCI 485 – Assignment #2
Ashish Dixit

Introduction

This report investigates Recursive Feature Elimination (RFE) using linear regression on the Diabetes dataset from scikit-learn. The objective is to analyze feature importance, evaluate model performance as features are removed, and determine the optimal subset of predictors for disease progression.

Dataset Description

Dataset Overview

- 442 samples
- 10 standardized features
- Target: quantitative measure of diabetes progression after one year
- Features include demographic (age, sex) and biological measurements (bmi, bp, s1–s6)

All features are standardized (mean ≈ 0, equal variance), allowing coefficient magnitudes to be compared directly.

```
=== Dataset Shapes ===
X shape: (442, 10)
y shape: (442,)

=== Feature Names ===
['age', 'sex', 'bmi', 'bp', 's1', 's2', 's3', 's4', 's5', 's6']

=== X Summary Statistics ===
                age           sex           bmi            bp            s1  \
count  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02
mean  -2.511817e-19  1.230790e-17 -2.245564e-16 -4.797570e-17 -1.381499e-17
std    4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02
min   -1.072256e-01 -4.464164e-02 -9.027530e-02 -1.123988e-01 -1.267807e-01
25%   -3.729927e-02 -4.464164e-02 -3.422907e-02 -3.665608e-02 -3.424784e-02
50%    5.383060e-03 -4.464164e-02 -7.283766e-03 -5.670422e-03 -4.320866e-03
75%    3.807591e-02  5.068012e-02  3.124802e-02  3.564379e-02  2.835801e-02
max    1.107267e-01  5.068012e-02  1.705552e-01  1.320436e-01  1.539137e-01

                s2            s3            s4            s5            s6
count  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02  4.420000e+02
mean   3.918434e-17 -5.777179e-18 -9.042540e-18  9.268604e-17  1.130318e-17
std    4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02  4.761905e-02
min   -1.156131e-01 -1.023071e-01 -7.639450e-02 -1.260971e-01 -1.377672e-01
25%   -3.035840e-02 -3.511716e-02 -3.949338e-02 -3.324559e-02 -3.317903e-02
50%   -3.819065e-03 -6.584468e-03 -2.592262e-03 -1.947171e-03 -1.077698e-03
75%    2.984439e-02  2.931150e-02  3.430886e-02  3.243232e-02  2.791705e-02
max    1.987880e-01  1.811791e-01  1.852344e-01  1.335973e-01  1.356118e-01

=== Target Summary Statistics ===
count    442.000000
mean     152.133484
std       77.093005
min       25.000000
25%       87.000000
50%      140.500000
75%      211.500000
max      346.000000
Name: target, dtype: float64

=== Train/Test Shapes ===
X_train: (353, 10) X_test: (89, 10)
y_train: (353,) y_test: (89,)
```

Baseline Linear Regression

Baseline Performance

Test R^2:

$$R^2 = 0.4526$$

Interpretation:

The model explains approximately 45% of the variance in disease progression, indicating moderate predictive performance.

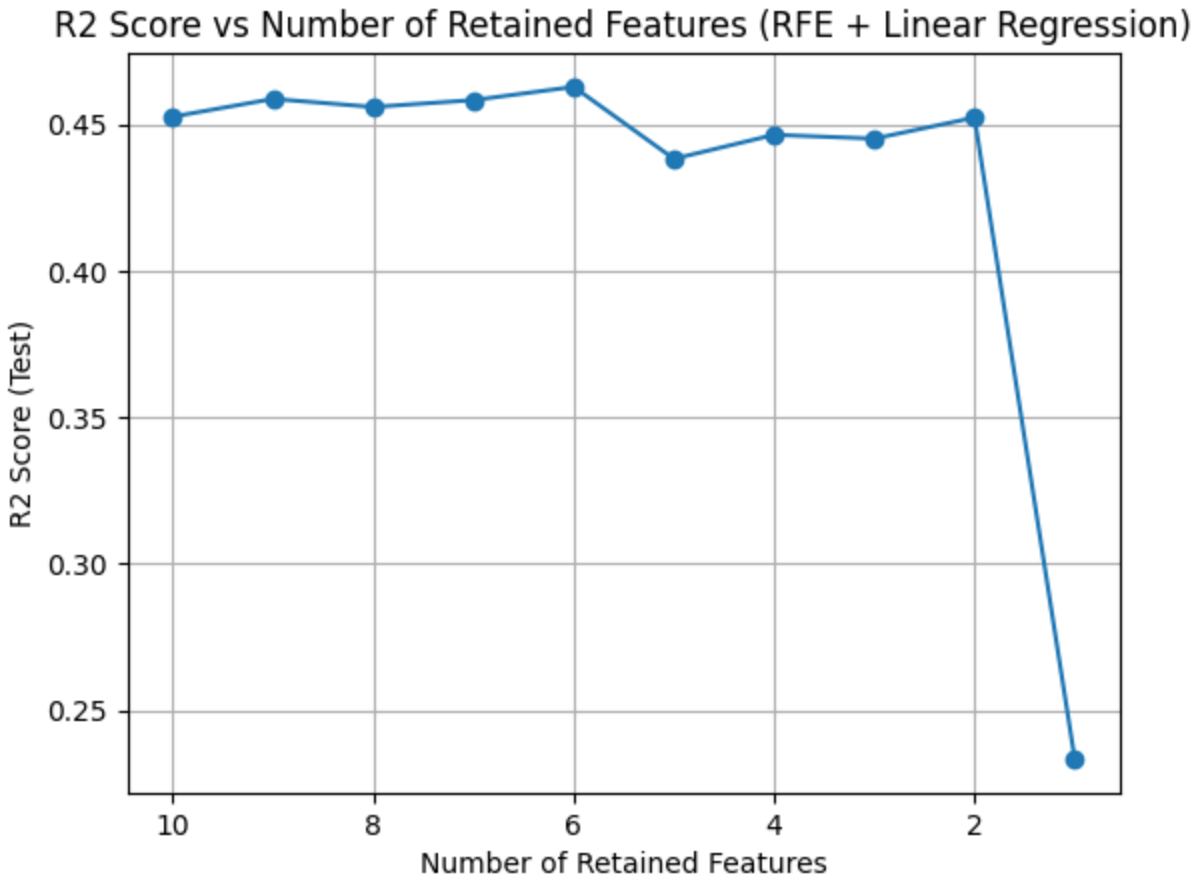
Table 1 – Baseline Feature Importance

```
=== Baseline Linear Regression ===
Baseline R2: 0.4526027629719196

=== Baseline Coefficients ===
s5     736.198859
bmi    542.428759
s2     518.062277
bp     347.703844
s4     275.317902
s3     163.419983
s6      48.670657
age     37.904021
sex   -241.964362
s1    -931.488846
dtype: float64

=== Baseline Feature Ranking (abs coef) ===
s1     931.488846
s5     736.198859
bmi    542.428759
s2     518.062277
bp     347.703844
s4     275.317902
sex    241.964362
s3     163.419983
s6      48.670657
age     37.904021
dtype: float64
```

The most influential predictors are s1, s5, and bmi. The large magnitude of s1 suggests a strong inverse relationship with disease progression.

Recursive Feature Elimination

R^2 vs Number of Features



R^2 values across feature counts:

```
=== RFE Results (R2 by number of features) ===
10 features -> R2 = 0.4526
 9 features -> R2 = 0.4587
 8 features -> R2 = 0.4559
 7 features -> R2 = 0.4583
 6 features -> R2 = 0.4628
 5 features -> R2 = 0.4382
 4 features -> R2 = 0.4464
 3 features -> R2 = 0.4451
 2 features -> R2 = 0.4523
 1 features -> R2 = 0.2334
```

Optimal Number of Features

Using a threshold of 0.01 for significant R^2 change, the first major drop occurs when reducing from 6 to 5 features (drop ≈ 0.0246).

Therefore:

Optimal number of features = 6

Performance slightly improves from 10 features (0.4526) to 6 features (0.4628), showing that removing weaker predictors reduces noise and improves generalization.

Feature Importance at Optimal Model

```
=== Features Selected at Optimal K ===
['sex', 'bmi', 'bp', 's1', 's2', 's5']

=== RFE Ranking (1 = selected, higher = eliminated earlier) ===
sex    1
bmi    1
bp     1
s1     1
s2     1
s5     1
s4     2
s3     3
s6     4
age    5
dtype: int64
```

Table 2 – Final Model Coefficients

```
=== Final Model at Optimal K ===
Final R2: 0.46277670793202996

=== Final Coefficients (sorted by abs value) ===
s1    -851.515734
s5     803.121285
s2     591.093315
bmi    557.314167
bp     350.178667
sex   -215.267423
dtype: float64

=== Top 3 Features (by abs coefficient, optimal model) ===
['s1', 's5', 's2']
```

Three Most Important Features

1. s1

2. s5
3. s2

Interpretation:

Blood serum measurements (s1, s2, s5) dominate predictive power. BMI also plays a significant role. Age was eliminated early, suggesting limited predictive value in this dataset.

Comparison: Baseline vs Final Model

Baseline Top Features:

s1, s5, bmi, s2, bp

Final Selected Features:

sex, bmi, bp, s1, s2, s5

The final model largely preserves the highest-ranked baseline features, indicating consistency between coefficient magnitude and recursive elimination. RFE removed weaker contributors such as age, s3, s4, and s6.

Reflection

What Was Learned About RFE

RFE iteratively removes the least important feature based on model coefficients. The results show that reducing dimensionality does not necessarily reduce predictive performance. In this case, eliminating four features slightly improved R^2.

RFE vs LASSO

RFE is a wrapper method requiring repeated model fitting. LASSO is an embedded method that applies L1 regularization to shrink coefficients toward zero.

Key differences:

- RFE eliminates features explicitly.
- LASSO performs continuous shrinkage.
- RFE is computationally more expensive.
- LASSO solves a single optimization problem.

Dataset Insights

- Biological markers are stronger predictors than demographic features.
- Age contributes minimal predictive value.
- Multiple blood measurements contain overlapping predictive information.