CSCI 485 Assignment #1

Recursive Feature Elimination with Linear Regression

Github link: https://github.com/Angelmartinez-20/CSCI485_Spring25_Angel_Martinez

---

**Dataset Exploration**

The Diabetes Dataset from the sklearn library had the following features:

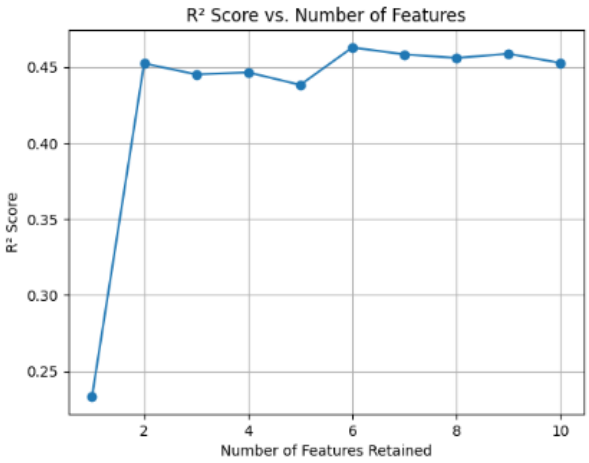
1. age - Age of the patient
2. sex- Gender of the patient
3. bmi - Body Mass Index
4. bp - Average blood pressure
5. s1 - Total serum cholesterol (TC)
6. s2 - Low-density lipoproteins (LDL)
7. s3 - High-density lipoproteins (HDL)
8. s4 - Total cholesterol / HDL (TCH)
9. s5 - Log of serum triglycerides level (LTG)
10. s6 - Blood sugar level (GLU)

The target variable represents a quantitative measure of disease progression one year after baseline.

**Recurrent Feature Implementation**

After Initially applying linear regression across all features, The current model's performance suggests that 45.26% of the variance in the target variable can be explained using all the features. This means that the model still leaves approximately 54.74% of the variation unexplained.

Using a convergence threshold of 0.01, the optimal number of features was 2.



**Feature Importance Analysis**

When comparing the initial and final features after applying RFE, the coefficients show slight changes for each feature, with age being dropped at the end. The top three features suggest that Total Serum Cholesterol (s1), Log of Serum Triglycerides Level (s5), and Body Mass Index (BMI) have the most impact on disease progression one year after baseline, compared to the other features. Then the table below represents the features values at each iteration.

```
Initial Features:
 s1     44.448856
s5      35.161195
bmi     25.607121
s2      24.640954
bp      16.828872
s4      13.138784
sex     11.511809
s3       7.676978
s6       2.351364
age      1.753758
dtype: float64
```

```
Final Features:
 s1     44.680849
s5      35.555151
bmi     25.624624
s2      25.151517
bp      17.143839
s4      12.903924
sex     11.258951
s3       7.882757
s6       2.577454
age           NaN
```

| | 1 features | 2 features | 3 features | 4 features | 5 features | 6 features | 7 features | 8 features | 9 features | 10 features |
|---|---|---|---|---|---|---|---|---|---|---|
| bmi | 47.141122 | 34.561598 | 34.824858 | 32.642633 | 28.225480 | 26.309836 | 26.052658 | 25.999687 | 25.624624 | 25.607121 |
| s5 | NaN | 26.852197 | 32.487841 | 37.404489 | 34.800355 | 38.357441 | 31.749939 | 36.389705 | 35.555151 | 35.161195 |
| s1 | NaN | NaN | -10.895940 | -28.295763 | -31.282091 | -40.632693 | -31.524621 | -45.228296 | -44.680849 | -44.448856 |
| s2 | NaN | NaN | NaN | 17.263257 | 19.483143 | 28.114579 | 16.330902 | 25.759819 | 25.151517 | 24.640954 |
| bp | NaN | NaN | NaN | NaN | 14.841764 | 16.948653 | 17.538042 | 17.607527 | 17.143839 | 16.828872 |
| sex | NaN | NaN | NaN | NaN | NaN | -10.241663 | -11.197798 | -11.121222 | -11.258951 | -11.511809 |
| s4 | NaN | NaN | NaN | NaN | NaN | NaN | 8.835330 | 13.254423 | 12.903924 | 13.138784 |
| s3 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | 8.091813 | 7.882757 | 7.676978 |
| s6 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | 2.577454 | 2.351364 |
| age | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | 1.753758 |

**Reflection**

What did I learn about RFE

- Recursive Feature Elimination is a neat algorithm that iteratively removes the least important features and evaluates the model's performance along the way. This can lead to discovering the most important features for a target variable and also reduce computational cost in further analysis by utilizing only the optimal number of features compared to the entire dataset. This is definitely something I will remember and utilize when handling datasets and their feature importance.

How does it compare to LASSO

- Unlike using an iterative approach to remove features one at a time, LASSO uses a regularization technique that applies an L1 penalty to the regression coefficients. It doesn't require multiple iterations, and the number of features selected depends on the strength of the penalty parameter. Therefore, LASSO is more ideal for larger datasets with many features, as it is more direct and less computationally expensive. However, RFE does offer more control since you go through each iteration of removing a feature one at a time.

Insights about the dataset and selected features

- Based on the optimal number of features selected by the model, Total Serum Cholesterol (s1) and Log of Serum Triglycerides Level (s5) have the highest impact on disease progression one year after baseline, compared to the other features. However, I would like to point out that when graphing the $R^2$ values against the number of features retained, I found it odd that sometimes the $R^2$ value would drop when the number of features increased. My current understanding of the subject is that more features mean more data to explain the variability. Therefore, $R^2$ should continue to rise as more features are included. I would like to review this report with someone else to validate my results.