Jacob Collins

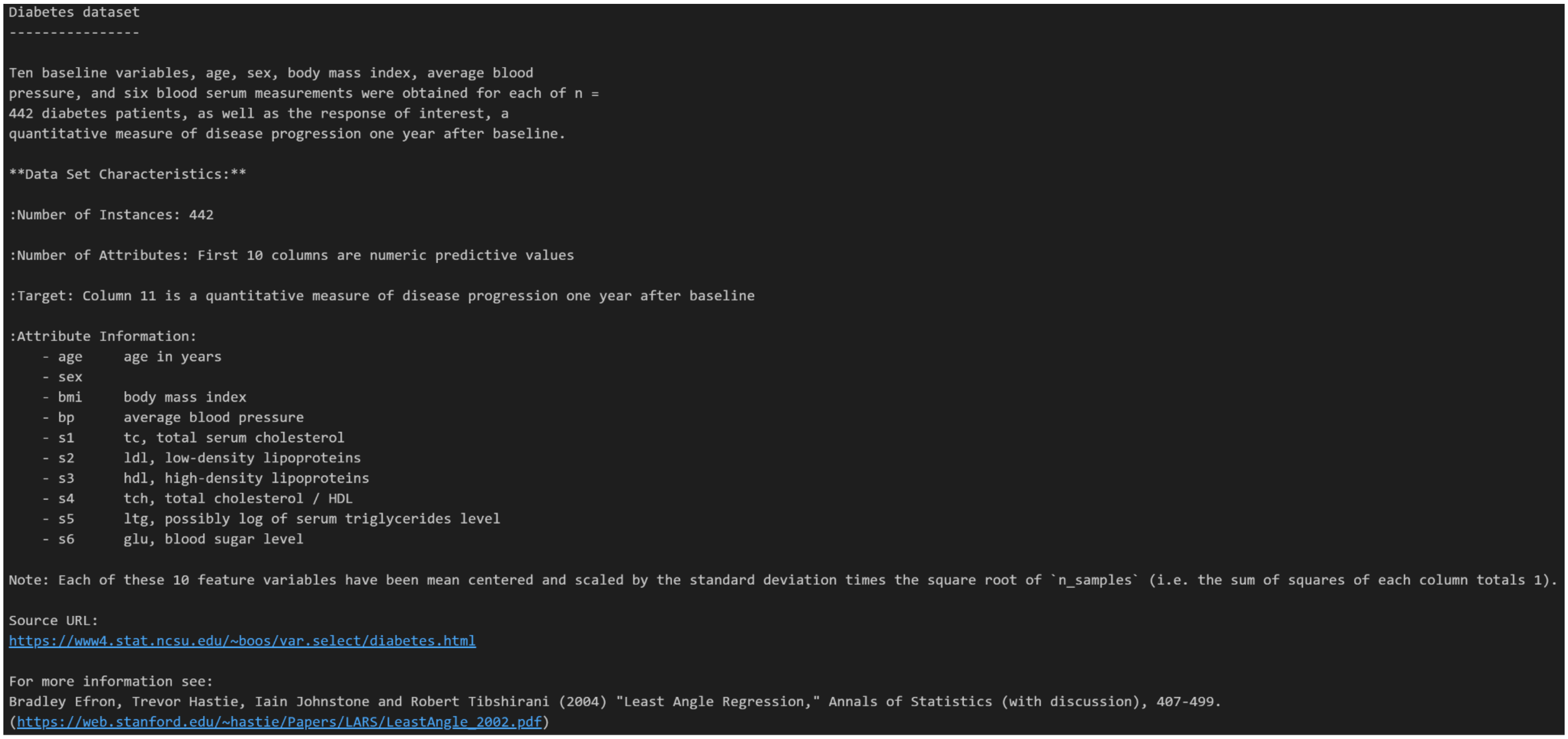
Dr. Shen

CSCI 485

February 6, 2025

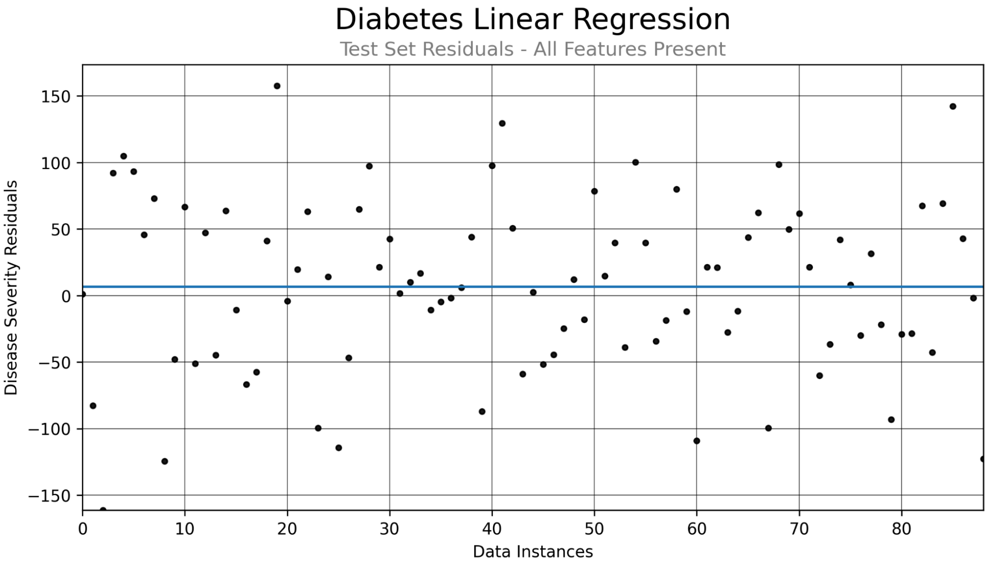
CSCI 485 Assignment#1: Recursive Feature Elimination with Linear Regression

Task 1: Dataset Exploration



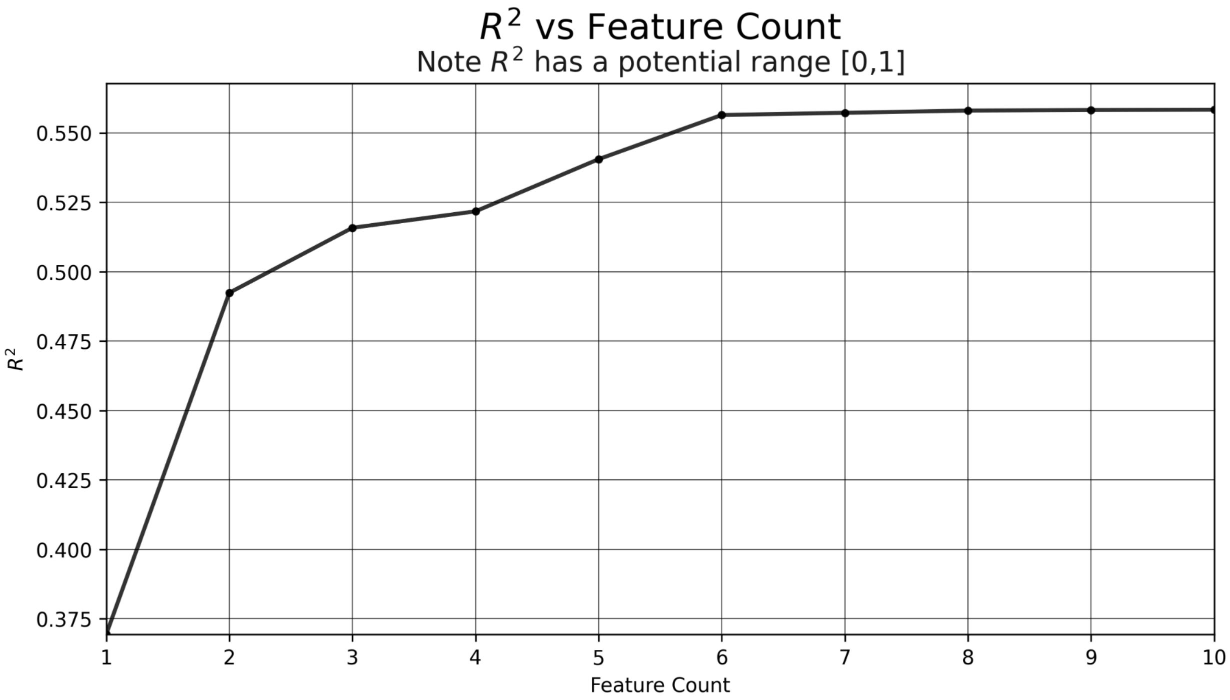
```
Diabetes dataset
----------------

Ten baseline variables, age, sex, body mass index, average blood
pressure, and six blood serum measurements were obtained for each of n =
442 diabetes patients, as well as the response of interest, a
quantitative measure of disease progression one year after baseline.

**Data Set Characteristics:**

:Number of Instances: 442

:Number of Attributes: First 10 columns are numeric predictive values

:Target: Column 11 is a quantitative measure of disease progression one year after baseline

:Attribute Information:
    - age     age in years
    - sex
    - bmi     body mass index
    - bp      average blood pressure
    - s1      tc, total serum cholesterol
    - s2      ldl, low-density lipoproteins
    - s3      hdl, high-density lipoproteins
    - s4      tch, total cholesterol / HDL
    - s5      ltg, possibly log of serum triglycerides level
    - s6      glu, blood sugar level

Note: Each of these 10 feature variables have been mean centered and scaled by the standard deviation times the square root of `n_samples` (i.e. the sum of squares of each column totals 1).

Source URL:
https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html

For more information see:
Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani (2004) "Least Angle Regression," Annals of Statistics (with discussion), 407-499.
(https://web.stanford.edu/~hastie/Papers/LARS/LeastAngle_2002.pdf)
```

Task 2: Linear Regression Model



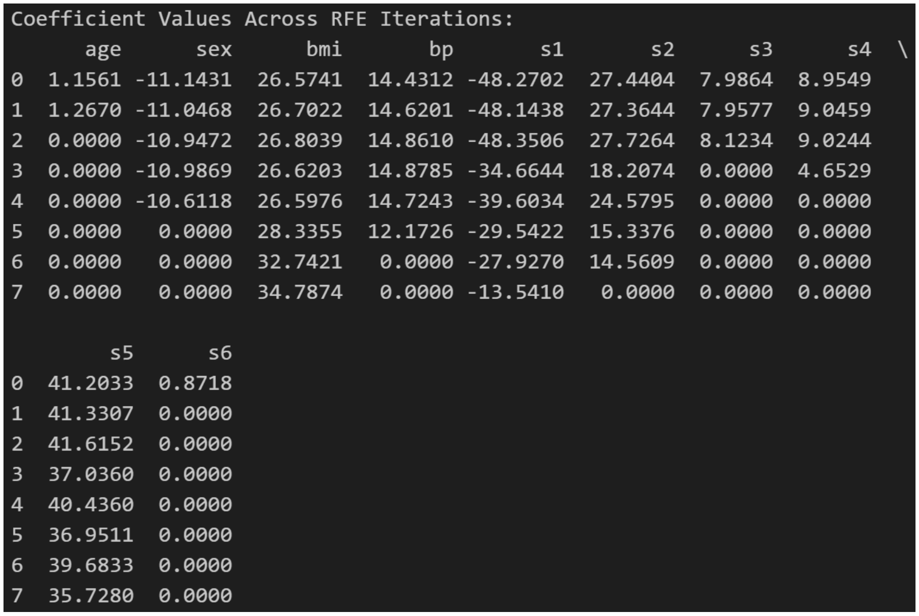
The initial model had an R^2 of 0.3512.

Task 3: Implement Recursive Feature Elimination (RFE)



Optimal number of features with threshold of 0.01 is 3: bmi, s1, s5.

Task 4: Analyze Feature Importance



```
Coefficient Values Across RFE Iterations:
      age      sex      bmi       bp       s1       s2      s3      s4  \
0  1.1561 -11.1431  26.5741  14.4312 -48.2702  27.4404  7.9864  8.9549   
1  1.2670 -11.0468  26.7022  14.6201 -48.1438  27.3644  7.9577  9.0459   
2  0.0000 -10.9472  26.8039  14.8610 -48.3506  27.7264  8.1234  9.0244   
3  0.0000 -10.9869  26.6203  14.8785 -34.6644  18.2074  0.0000  4.6529   
4  0.0000 -10.6118  26.5976  14.7243 -39.6034  24.5795  0.0000  0.0000   
5  0.0000   0.0000  28.3355  12.1726 -29.5422  15.3376  0.0000  0.0000   
6  0.0000   0.0000  32.7421   0.0000 -27.9270  14.5609  0.0000  0.0000   
7  0.0000   0.0000  34.7874   0.0000 -13.5410   0.0000  0.0000  0.0000   

        s5      s6  
0  41.2033  0.8718  
1  41.3307  0.0000  
2  41.6152  0.0000  
3  37.0360  0.0000  
4  40.4360  0.0000  
5  36.9511  0.0000  
6  39.6833  0.0000  
7  35.7280  0.0000  
```

Selected features matched the original ranking. See notebook for details.

Task 5: Reflection

What did you learn about feature selection using RFE?

Feature selection using RFE is a neat process, and is a very straightforward way to narrow down our feature selection to focus on the factors that matter most.

RFE helps to provide a clearer direction for further study, and also helps reducing the dimensional complexity of the dataset while removing the least useful features.

How does RFE compare to other feature selection methods like LASSO in terms of methodology and results?

RFE exists outside of a model, and is more like a tool to get information about how the model could be simplified while maximizing R^2.

LASSO, on the other hand, is built into a model, and explicitly changes the coefficients of features based on their importance.

In LASSO, the feature removal is like a byproduct of its broader usefulness, whereas RFE exists only to remove features.

What insights can you draw about the dataset from the selected features?

Our results make intuitive sense:

- **BMI** is a weight indicator, which we would expect to be highly correlated with diabetes.
- **s1** represents cholesterol levels, which is also intuitively known to be correlated with diabetes.
- **s5** is our final pick, and what's interesting is that in the dataset description, s5 is defined as "possibly log of serum triglycerides level", which seems to imply that the meaning of s5 is potentially unknown. Regardless, it is the only one of our selected features with a negative coefficient, implying that higher serum triglyceride levels may be correlated with lower diabetes severity.