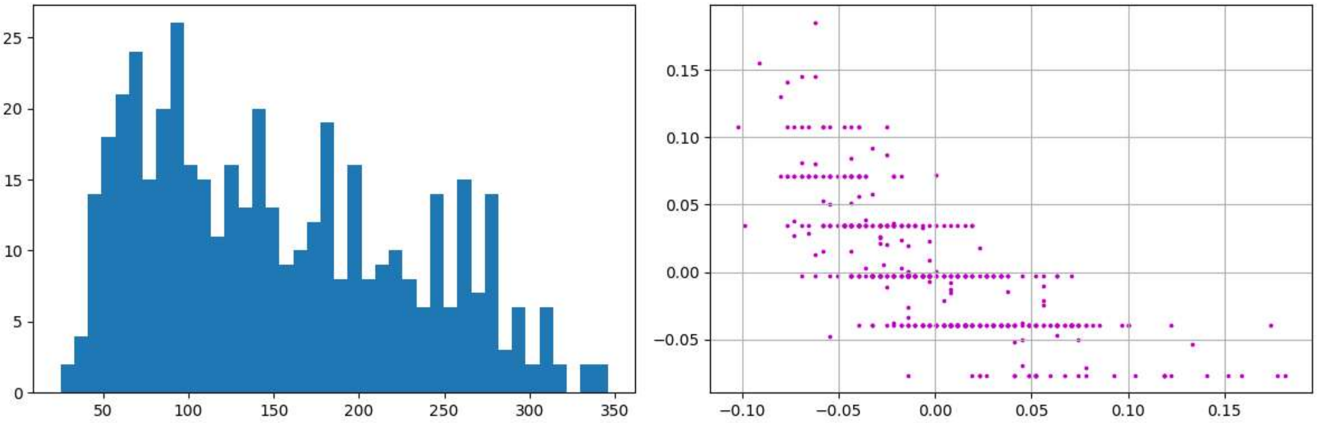
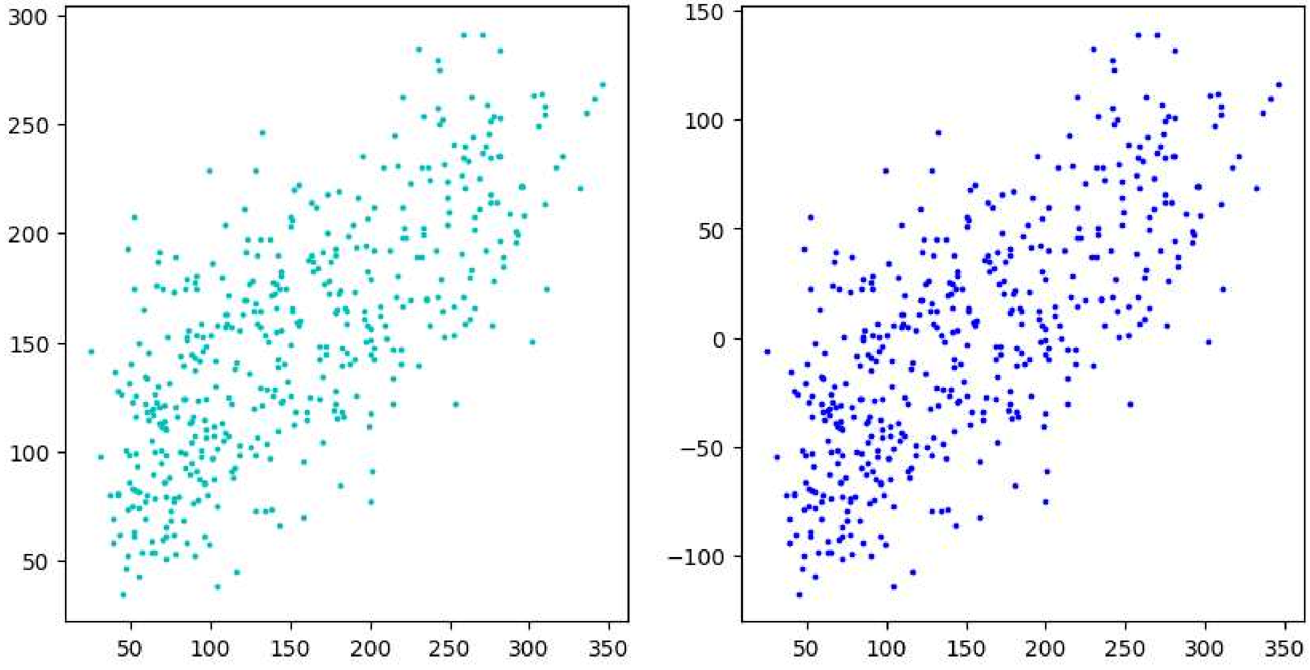
CO544 – Machine Learning
E/19/432 – Wickramaarachchi U.I.
Lab 04

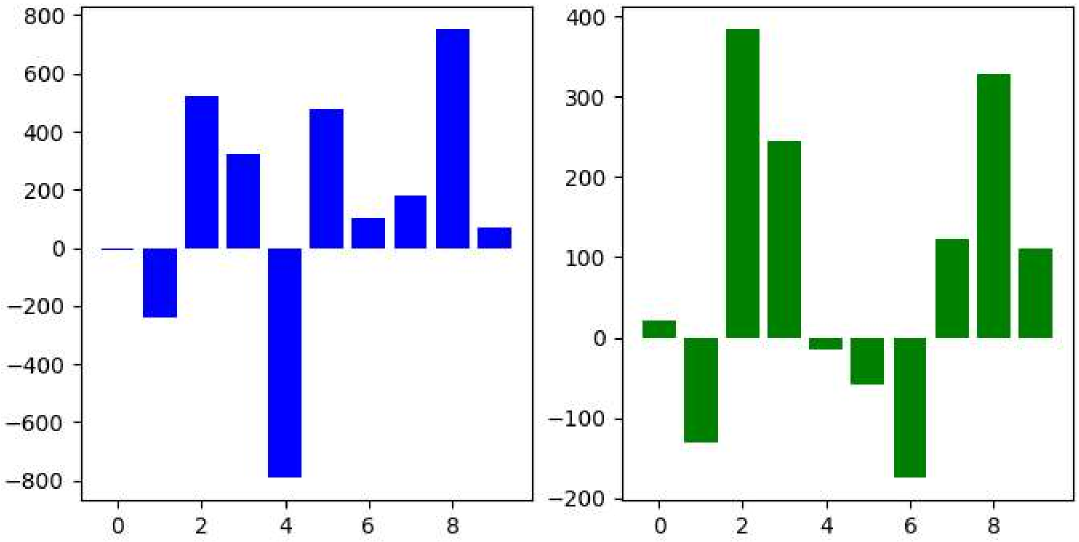
1.

Number of samples: 442
Number of features: 10
Shape: (442,)

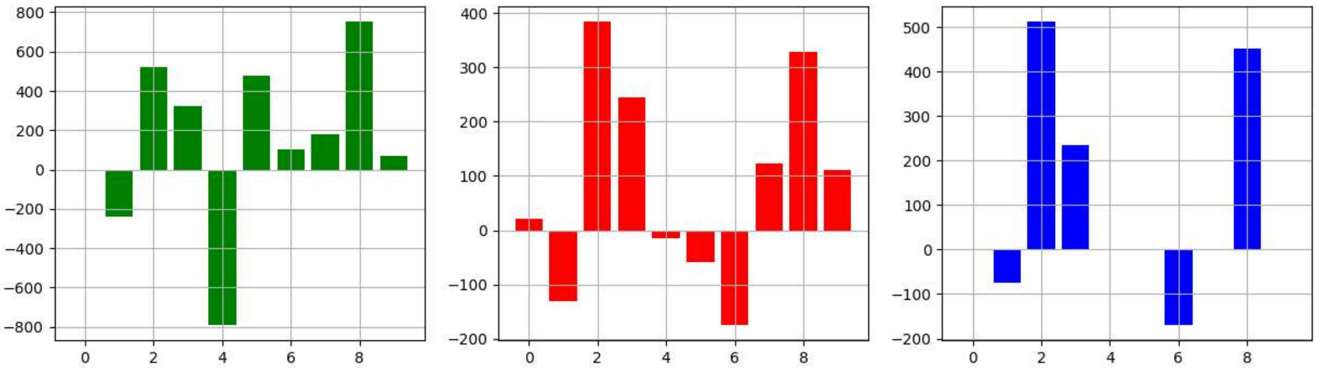




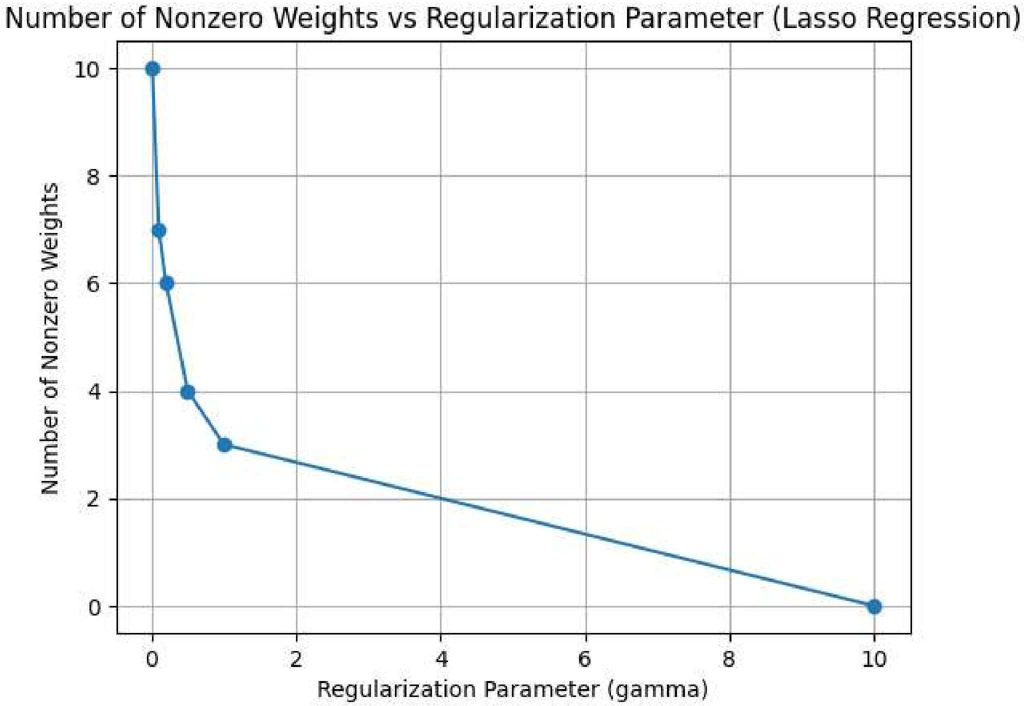
2.


800
600
400
200
0
−200
−400
−600
−800
0
2
4
6
8
400
300
200
100
0
−100
−200
0
2
4
6
8

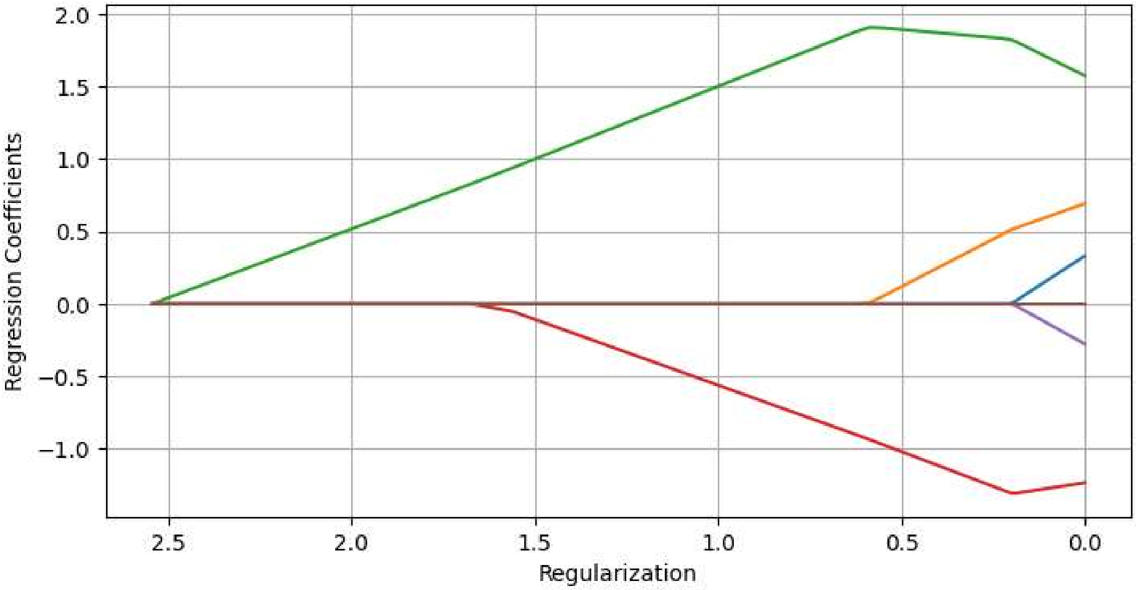
3. a.


800
600
400
200
0
−200
−400
−600
−800
0
2
4
6
8
400
300
200
100
0
−100
−200
0
2
4
6
8
500
400
300
200
100
0
−100
−200
0
2
4
6
8

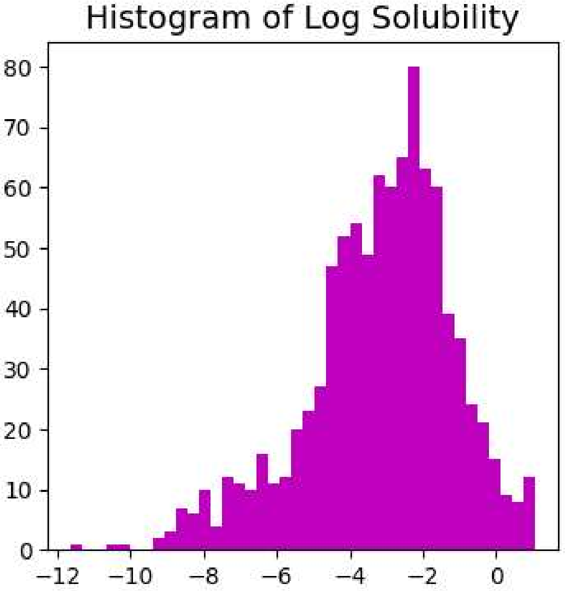
The number of non-zero weights decrease as the regularization parameter(γ) increases.

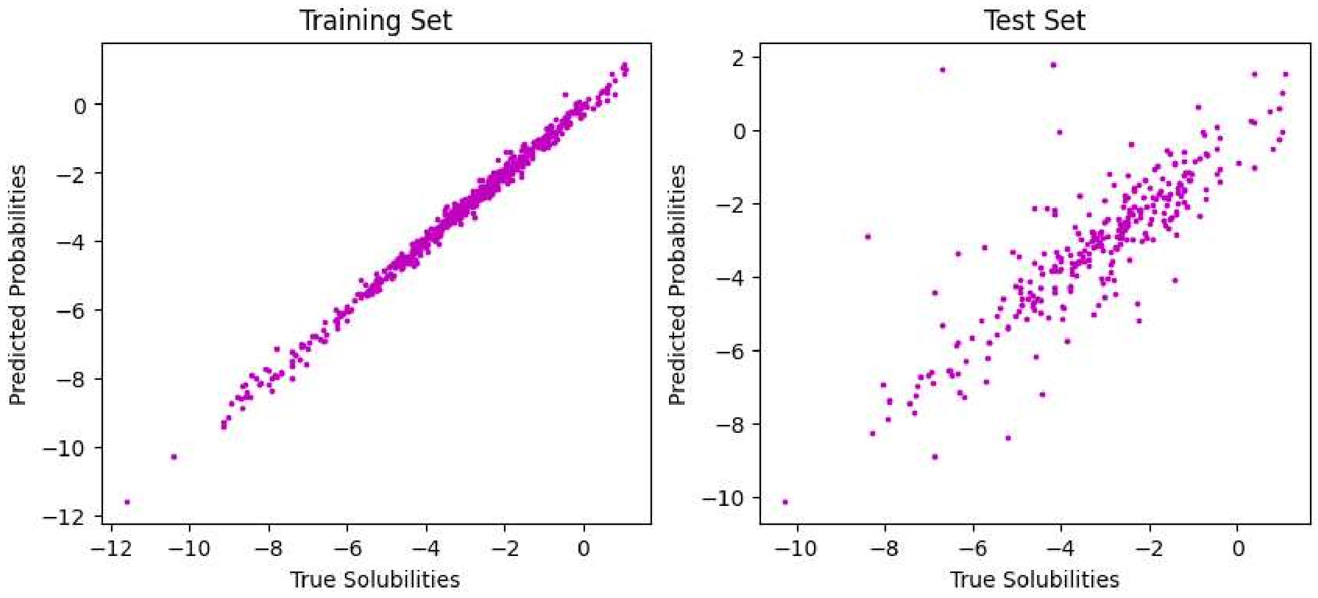


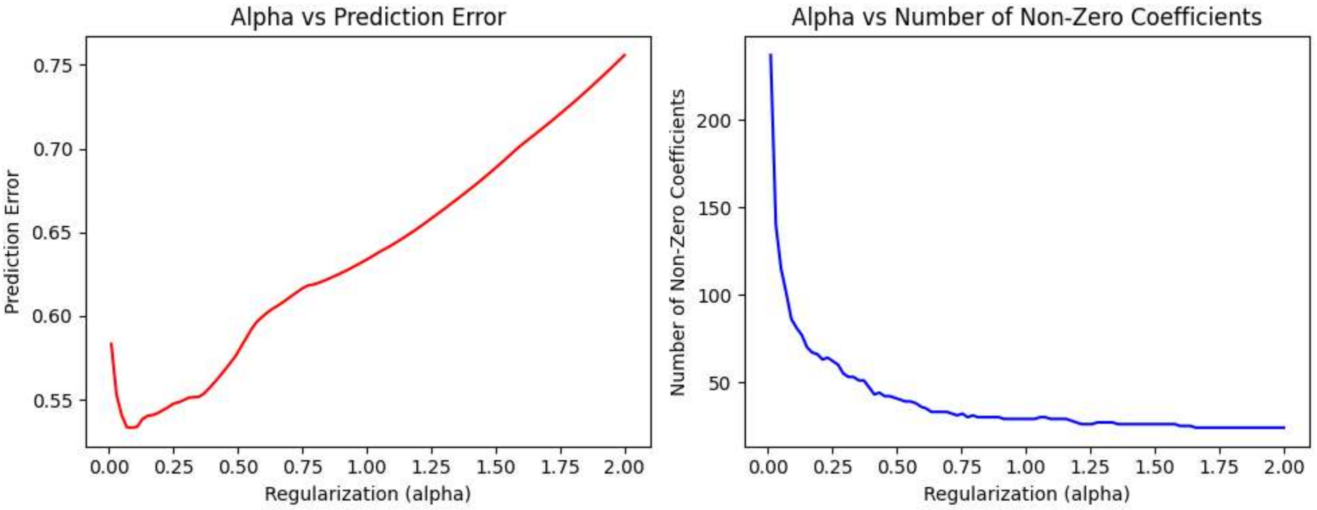
Yes, features with nonzero weights are considered more meaningful or important. This is because Lasso regression performs both variable selection and regularization in order to increase the prediction accuracy and interpretability of the statistical model it produces.


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
Regression Coefficients
2.5
2.0
1.5
1.0
0.5
0.0
Regularization

4.


Histogram of Log Solubility
80
70
60
50
40
30
20
10
0
−12
−10
−8
−6
−4
−2
0





Graphs of how the prediction error(on the test data) and the corresponding number of non zero coefficients change with increasing regularization.

Selecting top 10 features

Top 10 features: ['P_VSA_v_3', 'P_VSA_p_3', 'nCb-', 'MLOGP', 'MLOGP2', 'ALOGP', 'ALOGP2', 'BLTF96', 'BLTD48', 'BLTA96']

SelectKBest is a scikit-learn method that picks the top k features using a chosen scoring function. Here, we use f_regression, which measures the correlation between each feature and the target (solubility). Features with higher F-values are deemed more important for prediction

Prediction accuracy with these selected features when compared to using all the features and a quadratic regularizer.

MSE with all features: 2.7545856098040398
MSE with top 10 features only: 0.8978906081406246

As we can see the MSE value is lower when using only the top 10 features, so using those 10 feature provide a more accurate result.

Comparing the results to Part 3

When comparing with Part 3, there is an increase in the number of non-zero coefficients. And the regularization parameter have changed when compared to Part 3 for both Lasso and Tikhonov regularizations