CO544 – Lab 4

E/19/166

Jayathunga W.W.K.

1. The dataset, which included details on 442 diabetic patients, was examined. Ten characteristics are specific to each patient's situation. The target variable for each patient's stage of diabetes progression was the focus of the investigation. A histogram is used in the first investigation to visualize this goal variable and display the number of patients in each disease stage. This aids in our comprehension of the prevalence of various progressions and the location of the central tendency. Furthermore, in order to perhaps uncover links between two particular attributes (numbers six and seven) from the dataset, a scatter plot examines their association.

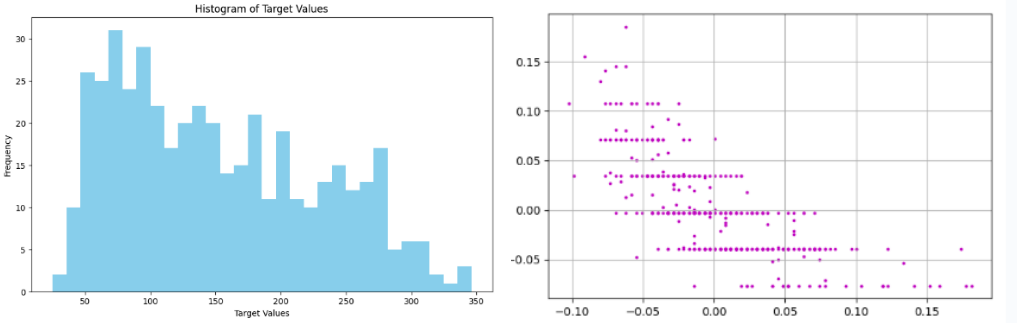

Figure 1: Histogram of the Targets and Pair-Wise Scatters

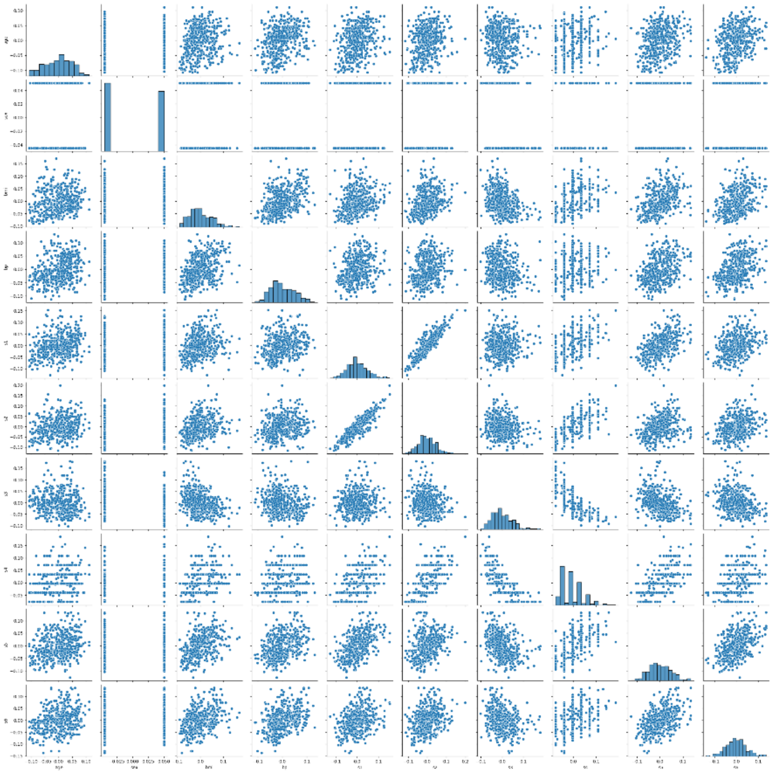

Figure 2:All Pair-Wise Scatters

Next, two machine learning methods for forecasting the course of diabetes were evaluated. Although they employ distinct strategies, both techniques are based on linear regression models. The first makes use of the linear regression function included in the scikit-learn toolkit. The pseudo-inverse approach is used in the second. It's interesting to note that when the actual illness stages are plotted against the phases that each technique predicts, the results look quite similar. This implies that, given the information at hand, both approaches predict diabetes progression fairly well for this specific dataset.

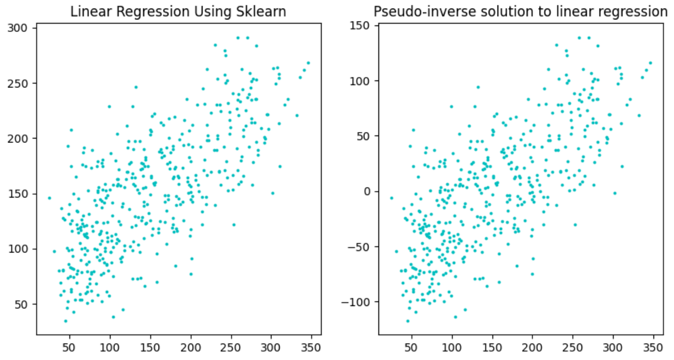

Figure 3: Comparison of the Pseudo-Inverse Solution to Sklearn Output

2. Given a linear model

$$y=Xw+\epsilon$$

where y is an n×1 dependent variable,

X is an n×p matrix of independent variables, w is a p×1 vector of parameters to be estimated, ϵ is an n×1 error term. The objective of Tikhonov regularization is to minimize the Residual Sum of Squares (RSS) plus a penalty term:

$$w_{min}\{||y-Xw||^{2}{}_{2}+\gamma||w||^{2}{}_{2}\}$$

where γ is the regularization parameter. The first term is just the RSS, and the second term is the penalty term.Taking the derivative of this expression with respect to w and setting it to zero gives the normal equations for Tikhonov regularization:

$$(X'X+\gamma I)w=X'y$$

Solving for w gives the Tikhonov regularization coefficient estimates:

$$w\hat{}\,ridge=(X'X+\gamma I)-1X'y$$

where I is the identity matrix.Overfitting, or the excessive dependence of linear regression models on training data, can result in subpar performance on unknown data. We can employ a method known as L2 regularization, or Tikhonov regularization, to stop this.

Large weight models are penalized in L2 regularization. This is accomplished by including a term in the error computation that penalizes the sum of squares for each of the model's weights. The strength of the regularization is determined by a parameter called gamma (γ), which governs this penalty term. The weights shrink more near zero the higher the gamma.

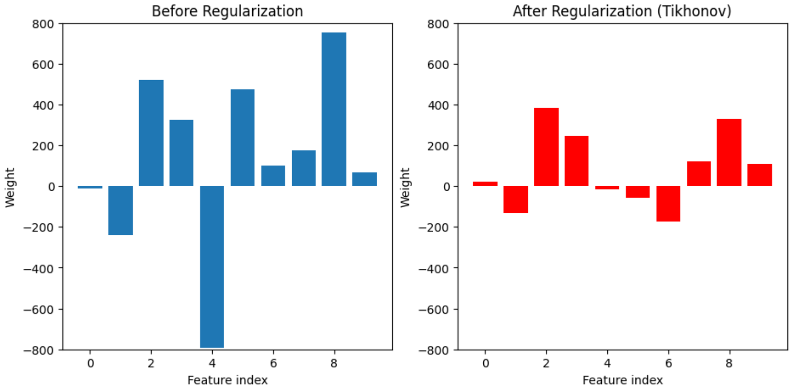

Figure 4:Comparison of Weights Without Regularization and With Regularization

The weights from two models—one without regularization (blue bars) and one with L2 regularization (red bars)—are plotted against each other in Figure 3. Weight values in the non-regularized model are substantially higher. These high weights may be a factor in overfitting. On the other hand, because the penalty term pushes the weights closer to zero, the regularized model has substantially smaller weights. This encourages the use of a more basic model that performs better on unknown data and is less prone to overfit. Essentially, L2 regularization aids in finding a compromise between training data fitting and preserving a more straightforward model that performs well in novel scenarios.

3. Regular regression models are complex and challenging to interpret since they might incorporate a large number of features. This is addressed with L1 regularization, commonly referred to as Lasso, which builds sparse models. Some weights are driven to absolutely zero by Lasso, in contrast to L2, which decreases all weights. In doing so, the model's unnecessary features are effectively eliminated, simplifying and improving its understandability.

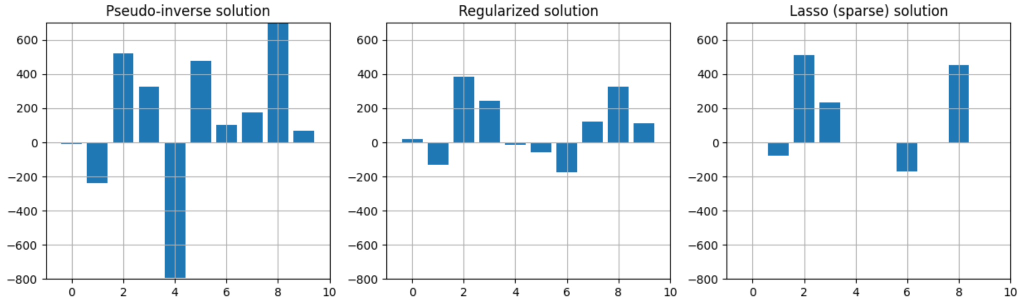

Figure 5:Comparison of Weights

This idea is shown visually in Figure 4. Weights from a non-regularized model (several features used) are displayed in the blue bars. Because of shrinkage, the red bars (L2) display decreased weights. Last but not least, a sparse model with many weights set to zero is displayed by the green bars (Lasso). By concentrating on what really matters, this sparsity can even enhance performance on unseen data by assisting in the identification of the most crucial traits for prediction.

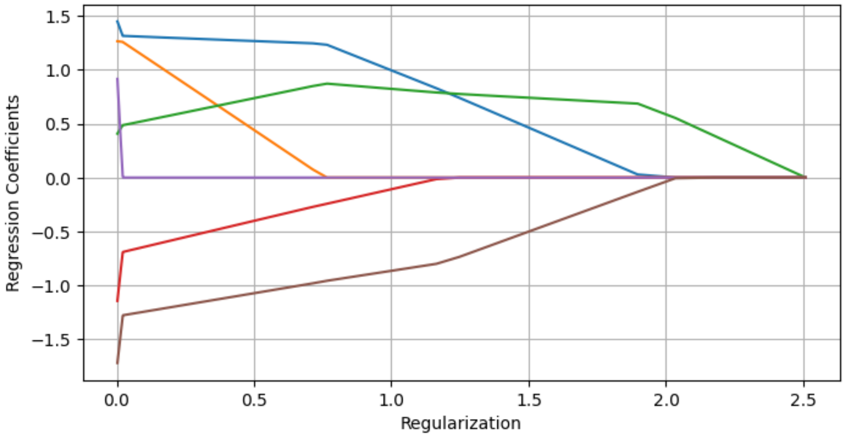

Figure 6:Regularization Path

This plot reveals how Lasso performs feature selection. It shows the coefficients of a six-variable model changing as the "alpha" value (regularization strength) increases. Each line represents a coefficient, and as alpha goes up, some coefficients shrink to zero. This highlights Lasso's core principle: pushing unimportant features out of the model by driving their coefficients to zero, resulting in a sparse and interpretable model.

4.

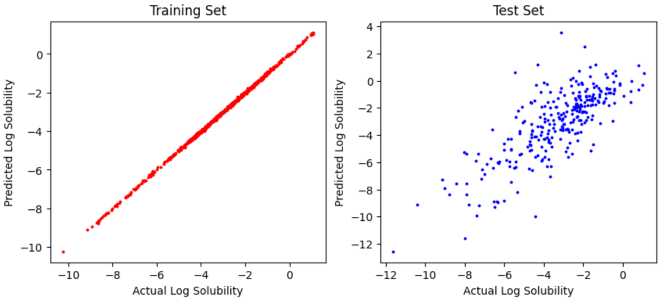

Figure 7:Predicted Solubility vs True Solubility for training and test set

Training data made up 70% of the total, and testing data made up 30%. The model performs well on both sets, as seen in Figure 6. As anticipated, it fits the training data more precisely. The model's fit for the test data varies a little more, which is expected given that the data was not utilized for training.

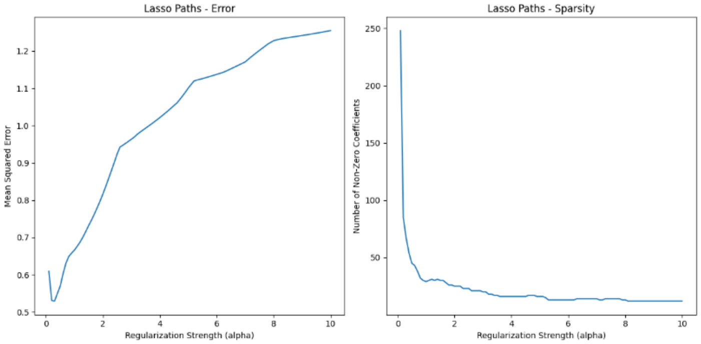

Figure 8:Variation of Prediction Error (on the test data) and The Corresponding Number of Non-zero Coefficients

The model's capacity for learning and generalization is strongly impacted by the selection of the regularization coefficient (alpha). Even after considerable training, the model struggles to converge when alpha is very tiny (poor regularization). This implies that the model is overfitting the training set and is unable to come up with a good answer for unknown data. This is visually shown in Figure 7, where the error on the test set starts to rise as alpha grows (regularization intensifies). This suggests that the model is become too conservative and might be overlooking significant data patterns.In other words, it's critical to find the ideal alpha balance. Overfitting can result from too little regularization, while excessive regularization might oversimplify the model and prevent it from capturing the complexity of the data.

Top ten features: TIC2, Vx, P_VSA_p_2, MLOGP2, CSI, P_VSA_LogP_8, P_VSA_v_3, H_D/Dt, Wi_Dz(v), P_VSA_MR_1

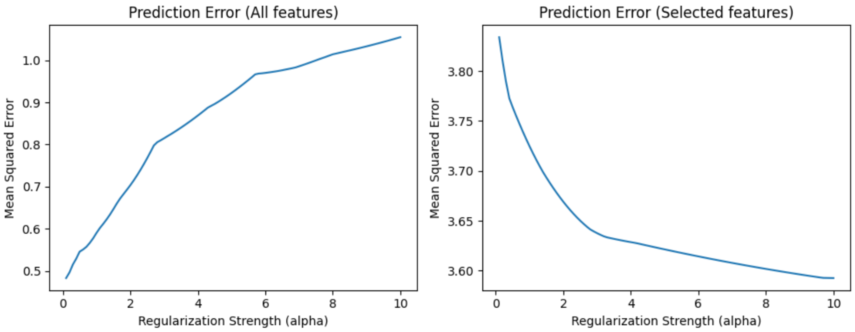

Figure 9:The Variation of Prediction Error with All Featurs and Selected Features

Reducing the number of carefully selected features in a model can enhance its regularization performance. This is due to the fact that regularization penalizes intricate models with lots of features. On the other hand, accuracy with heavy regularization may suffer if all features are used. By concentrating on the most crucial characteristics, feature selection assists in preventing this. Regularization and feature selection work together to provide a more broadly applicable model. This indicates that the model avoids overfitting to the training set, resulting in good performance on unknown data.