Khangai Enkhbat

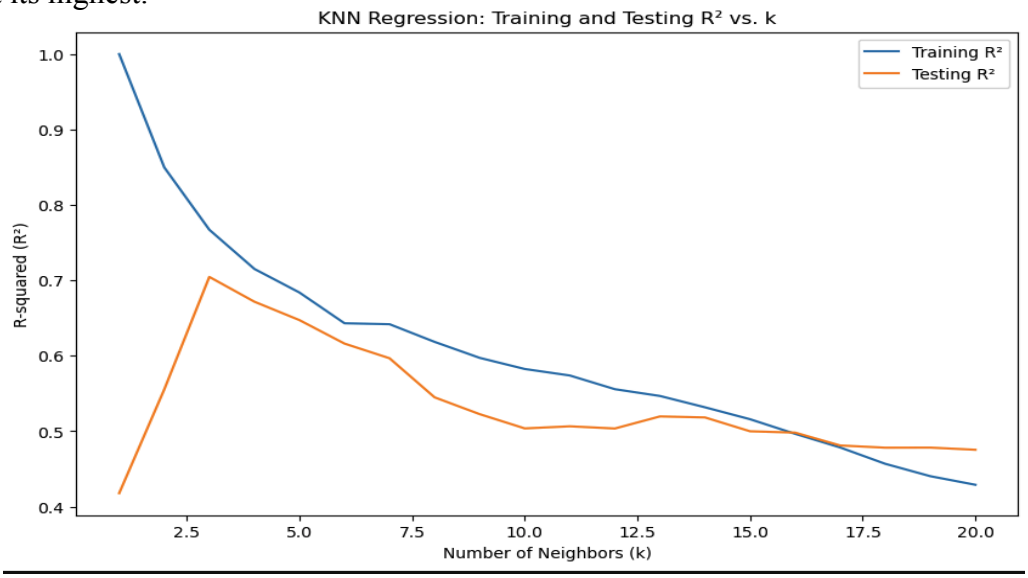
02 Feb 2025

Analysis

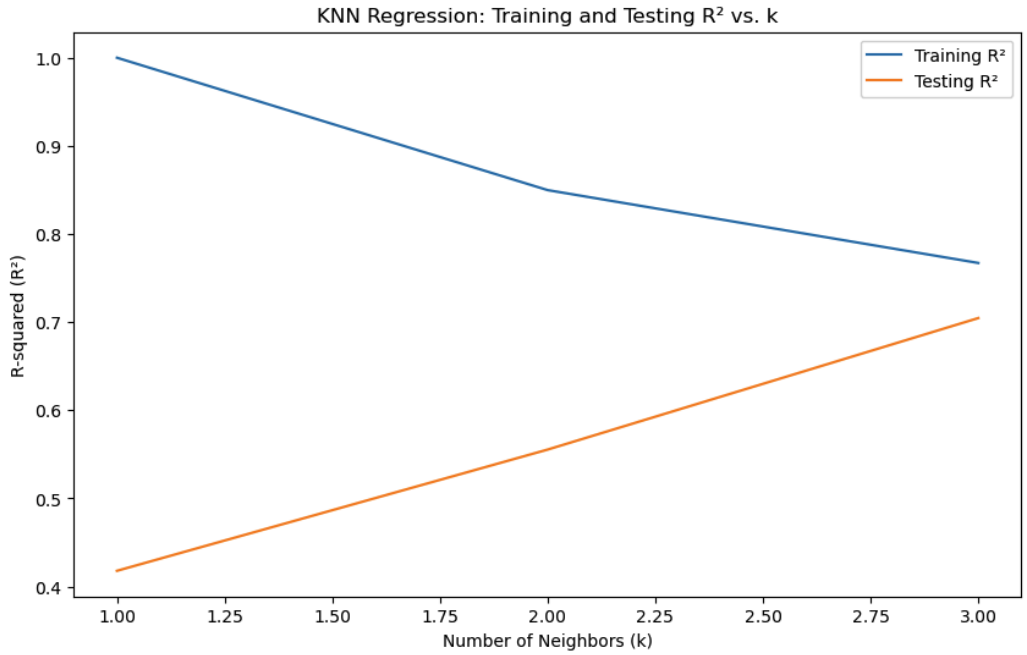
KNN REGRESSION

In this test, I created and tested the KNN Regression model, with the base K value of 20(which I changed later on). The dataset is “Boston House-Price Data” and I have plotted the result using “matplotlib”.

So to analyze the results both the training and testing data’s R^2 is lower than I thought with Training R^2 at 0.4404 and Testing at 0.47 when our K value is 20. As the k value increases, the model starts averaging more neighbors, reducing variance and leading to lower R^2 on training data. So to find the fitting K value I looked at the graph and found a value where Test R^2 is at its highest.



When K is 3 or 4 the Testing R^2 is maximized to its highest so that will be the optimal k value. When K starts to increase the model starts to be more dull and it starts underfitting. Now



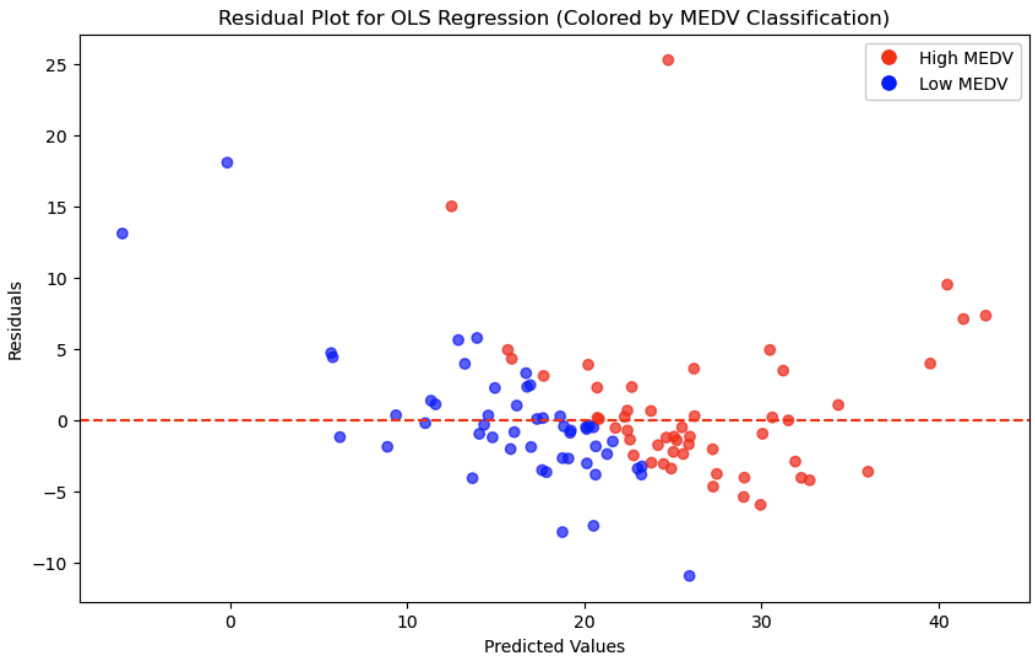
let's see the graph when we choose the optimal K value.

Such inferred tendencies of overfitting as k is shown on a very small range. At k=1, there is a 1.0 training R^2 value, meaning that the model is indeed memorizing the training data perfectly, a condition indicative of extreme overfitting or overtraining. However, the testing R^2 becomes a lot lower at 0.4179, indicating that it does not generalize well to unseen data. Both training R^2 values drop as k progresses from 2 to 3; 0.8498 down to 0.7672 proves that the model is starting to generalize much better but still bears tendencies of overfitting. The model has improved from 0.5555 to 0.7046 in testing R^2, which shows that it is learning meaningful patterns rather than complete memorization from training sets. Therefore, at k=1, the model is deeply overfitting, exhibited by the high training R^2 but low test R^2. While at k=3, this model was a much better generalization reflected in the increased testing R^2. If k goes into larger

numbers, the testing R^2 could decline because of underfitting and too much smoothing. The optimal k is around 3 to 5, where the test R^2 maximizes both bias and variance performance.

OLS REGRESSION

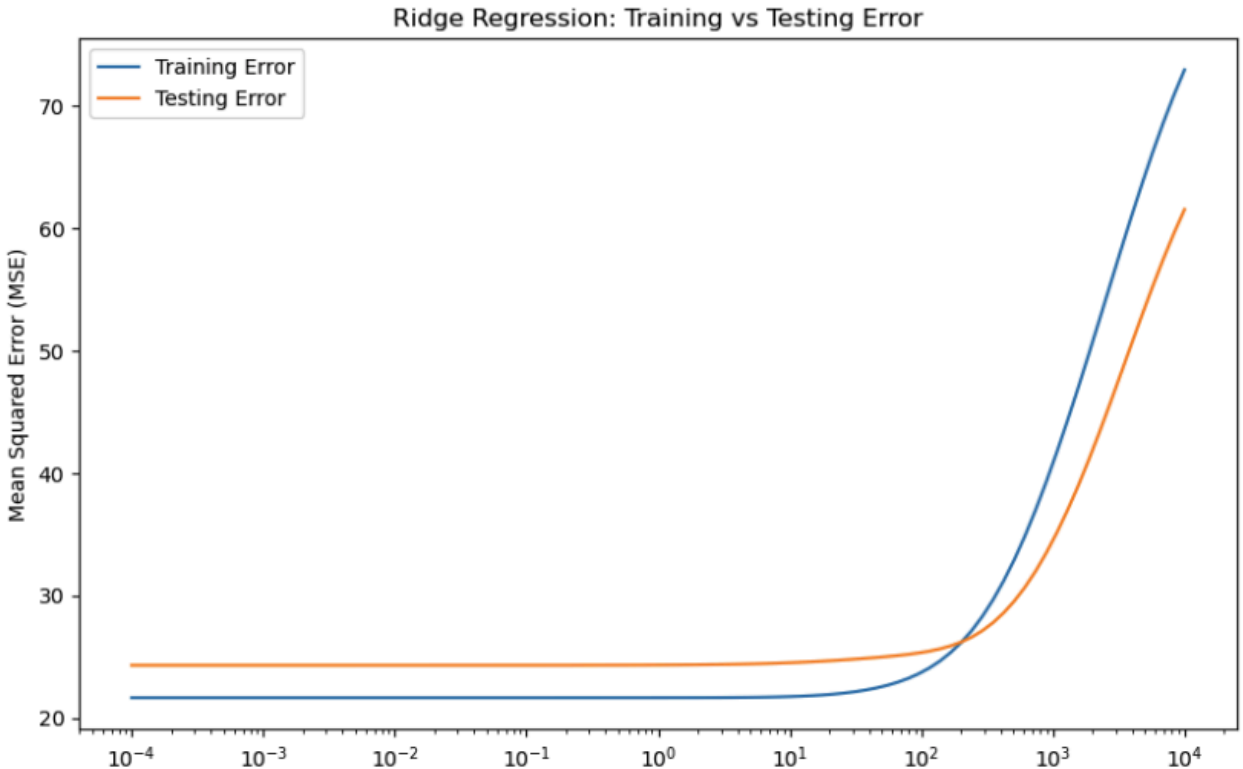
For the analysis of the result for this OLS model performed Trainin R^2 of 0.75, Testing R^2 of 0.67 and the MSE was 21.64(Training) and 24.29(Testing). Training R^2 (0.75) means that 75% of the variance in the training data is explained by the model, this model is a good fit. On the other hand, R^2(0.67) for Testing the score drops from 0.75 to 0.67 indicating some overfitting which the model performance goes a little worse on unsees data. To unpack the results for Mean Squared Error(MSE), Training MSE was 21.64 and Testing MSE was 24.29. The increase in MSE from training to testing further confirms that this model is overfitting to the data.



The R^2 values indicate that the model explains a reasonable amount of variance in the data however the drop in the R^2 and increase in MSE from training to testing suggested that the model is overfitting. The plot shows that residual plots may reveal patterns, indicating that the linear model is not capturing all the complexity of the data. So in conclusion the OLS Regression model is performing reasonably well on the dataset, explaining 75% of the variance in the training data ($R^2 = 0.75$) and 67% in the testing data ($R^2 = 0.67$). However, the drop in R^2 and the increase in Mean Squared Error (MSE) from training (21.64) to testing (24.29) indicate overfitting, as the model performs slightly worse on unseen data.

RIDGE REGRESSION

Training error (Blue) As "α" increases the training error increases, this is because "α" makes the model less flexible and less able to fit the training data perfectly. The same goes for Testing error(Red) however even as "α" increased testing error it was much lower than the training error(Blue) which indicates that regularization is helping reduce overfitting. For the optimal "α" it can seen as 0.1 which is 10^-1.



I have also made a plotting in R^2, in the plot, as “α” increases the training R^2 score decreases because it makes the model becomes less flexible. For testing R^2 score as α increases, the testing R^2 score improves, indicating that regularization is helping the model generalize better.

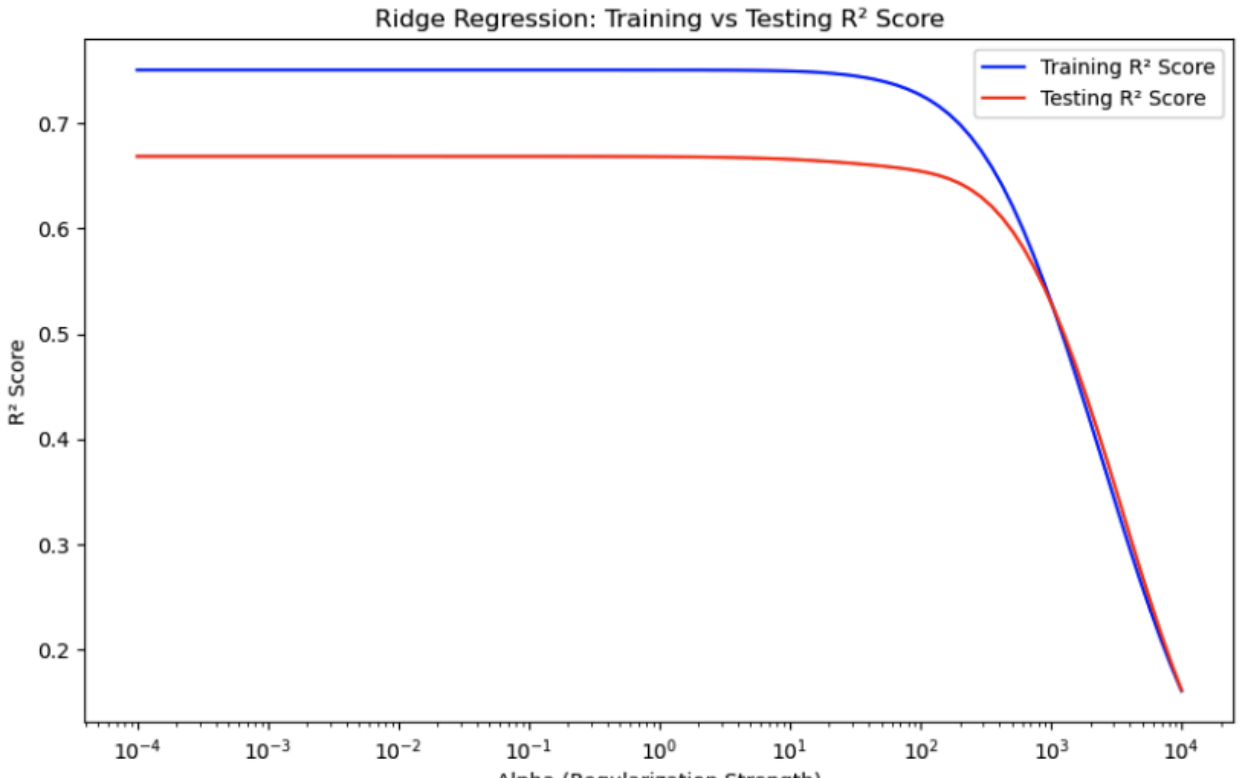
After a certain point, the testing R^2 score starts to decrease, indicating underfitting. Similar to MSE the optimal “α” is the one that maximizes the testing R^2 is around 0.1. Since both MSE and R^2 have the same “α” this value minimizes the MSE and maximizes the R^2 score. Starting from the 0.0001 value till 10 or 10^1 it looks flat but 0.1(10^-1) is where both of them are maximized and minimized.

Overfitting (Small α):

- High training error.
- High testing error.
- Low training R^2 score.
- Low testing R^2 score.

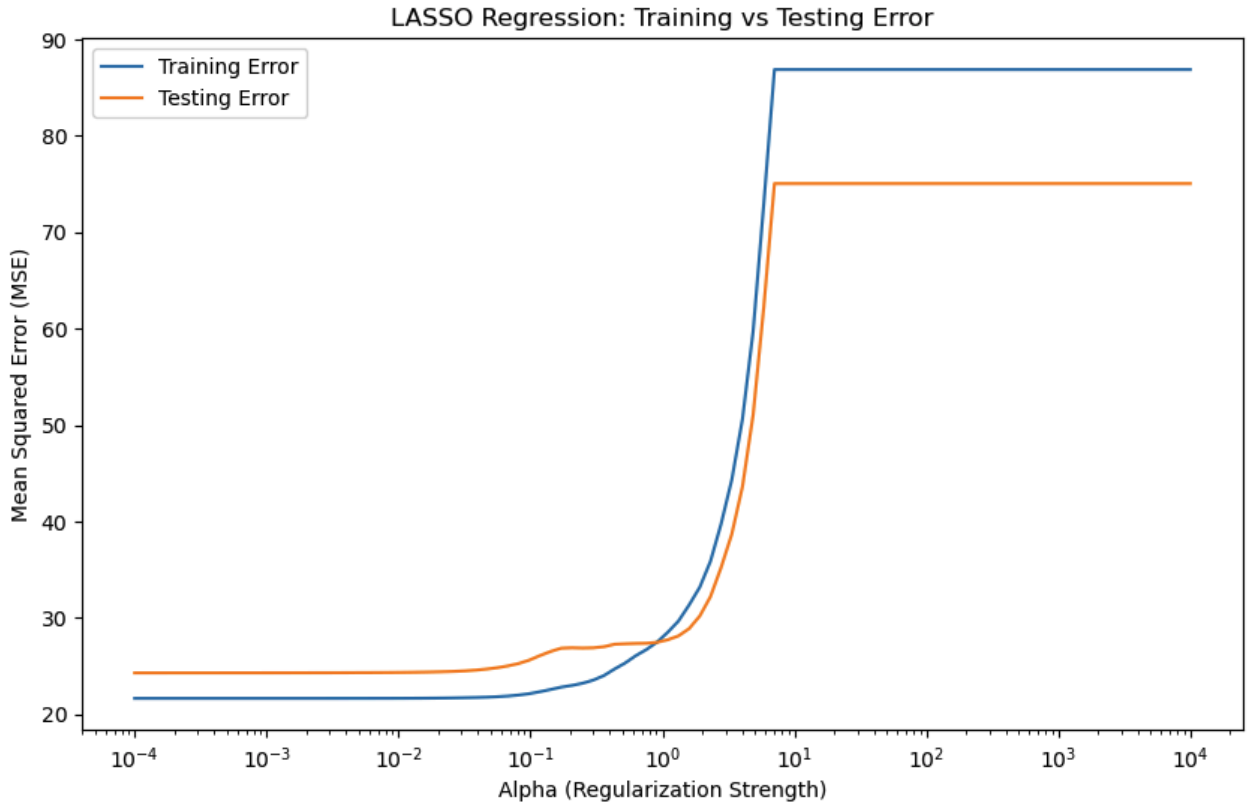
Underfitting (Large α):

- Low training error.
- High testing error.
- High training R^2 score.
- Low testing R^2 score.



LASSO REGRESSION

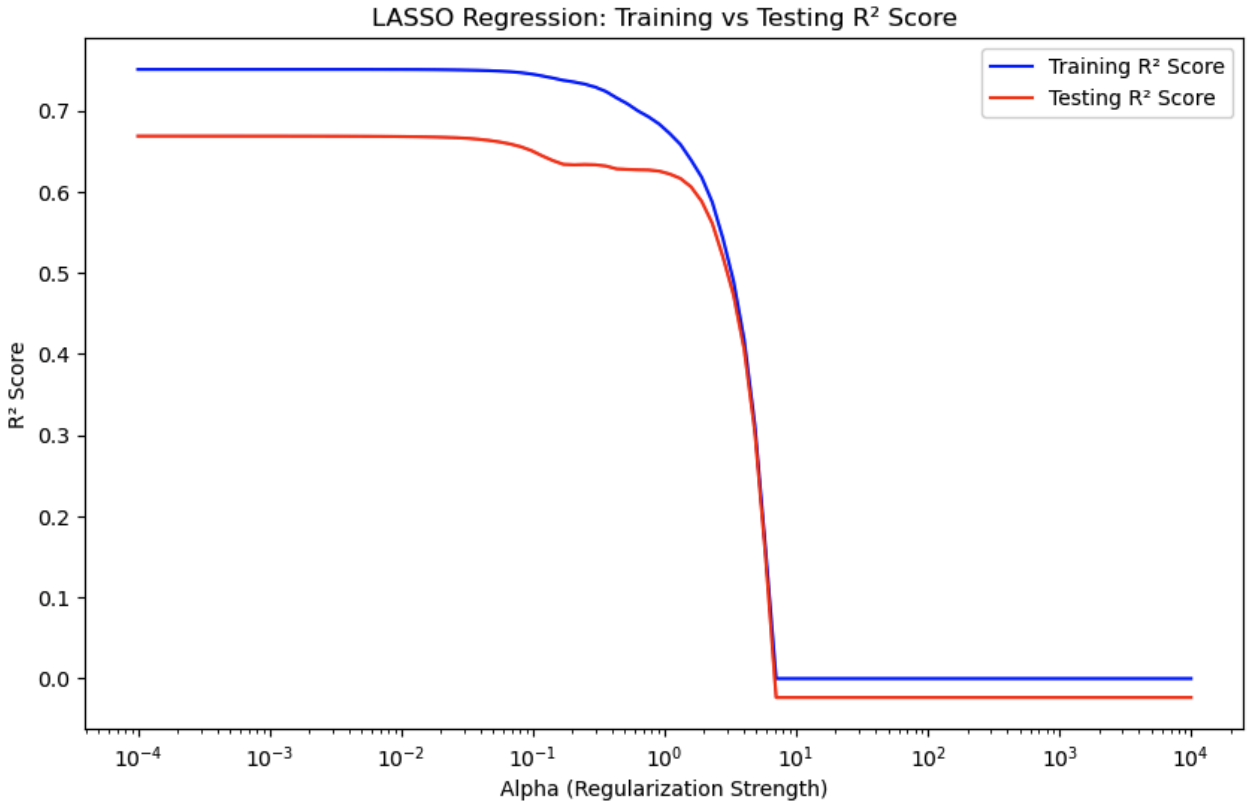
Training error(Blue) increases as “α” increases due to “α” penalizing the model's coefficients more making it less responsive. On the other hand, as α increases the Testing result(Red) decreases compared to Training, and at a certain point Testing error increases but does not exceed the Training error.



For R^2 when “α” increases the score decreases for Training R^2 and Testing R^2 the score improves indicating the model generalizes better.

For both plots, the optimal “α” is around 0.01 (10^-2) where the testing error is minimized and the testing R^2 score is maximized. At α = 0.01, the model achieves a good

balance between bias and variance,m where it is flexible enough to capture the underlying patterns in the data and regularized enough to avoid overfitting (low variance).



Overfitting (Small α):

- Low training error.
- High testing error.
- High training R^2 score.
- Low testing R^2 score.

Underfitting (Large α):

- High training error.
- High testing error.
- Low training R^2 score.
- Low testing R^2 score.

The LASSO Regression model with $\alpha = 0.01$ is neither underfitting nor overfitting. It achieves a good balance between bias and variance, performing well on both training and testing data.

KNN is the best-performing model for this dataset due to its ability to capture non-linear relationships, resulting in the lowest testing error and highest R^2 score. However, it is slower and less interpretable compared to linear models. LASSO is a strong alternative, particularly if you need a linear model with feature selection, as it performs well and offers better interpretability. Ridge serves as a good baseline model but is outperformed by both KNN, OLS, and LASSO in terms of predictive accuracy and generalization.