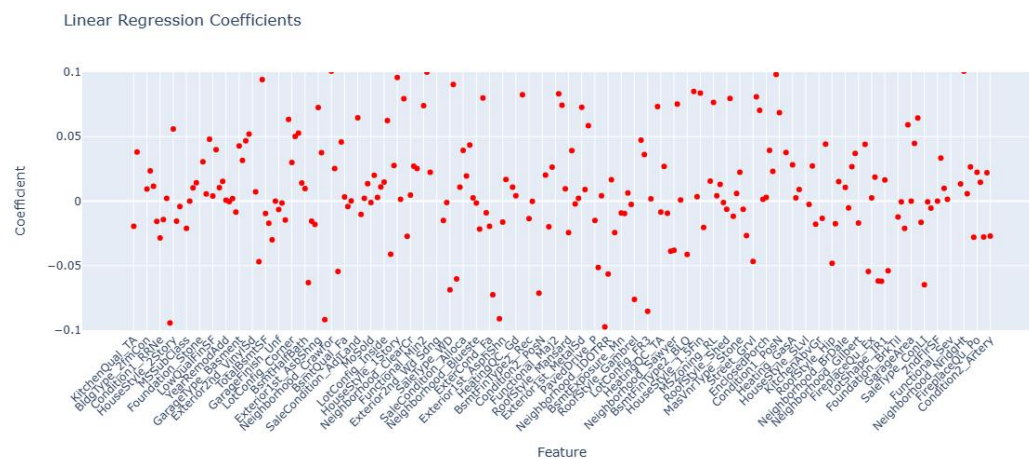


1. Abstracts

The updated project aims to regularize the model and compares the performances among OLS, Lasso and Elastic Net. Then it selects the optimal model based on the MSE (Mean Square Error) and the risk of overfitting.

2. OLS Regression

Firstly, using the dataset handled by preprocessing method mentioned in last assignment, it is easy to generalize the OLS result and coefficients distribution.



Plot 1: OLS Coefficients Distribution

3. Data Scaling

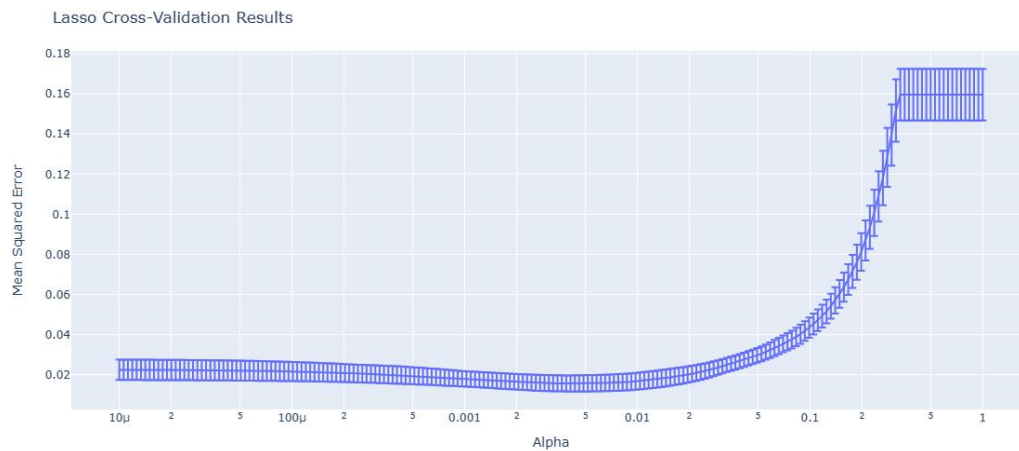
Why choose to standardize data? Because in regularization, the normalized data can help control the penalty term, making the feature selection more stable and accurate.

4. Applying Lasso

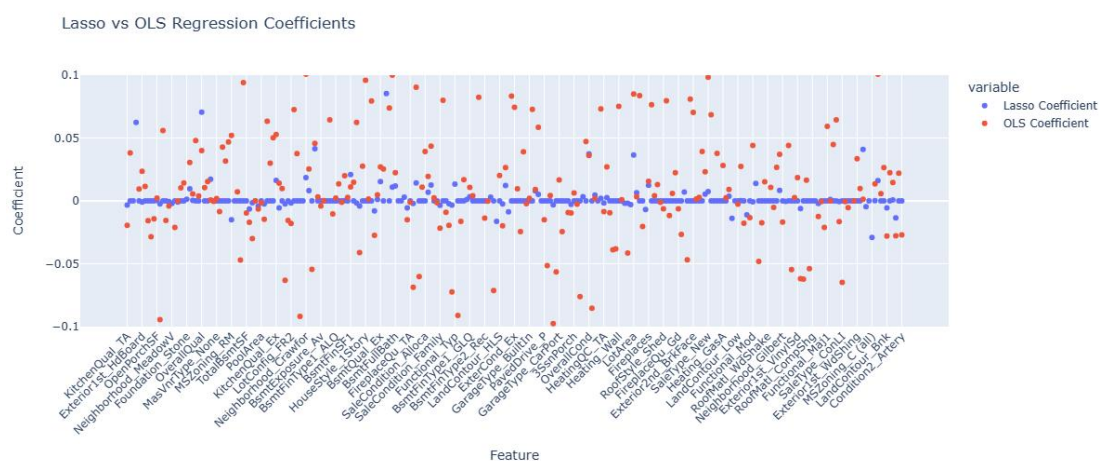
After scaling, we perform a 5-fold cross-validation to select the alpha value with the best performance as the optimal hyperparameter based on the model's mean square error on each alpha value.

To visualize the effect of Lasso, we make the graph about the coefficients between

these models. It is clearly to see that the Lasso model penalizes large coefficients to small and even to zero and make the model more interpretable.



Plot 2: Lasso Cross-Validation Results



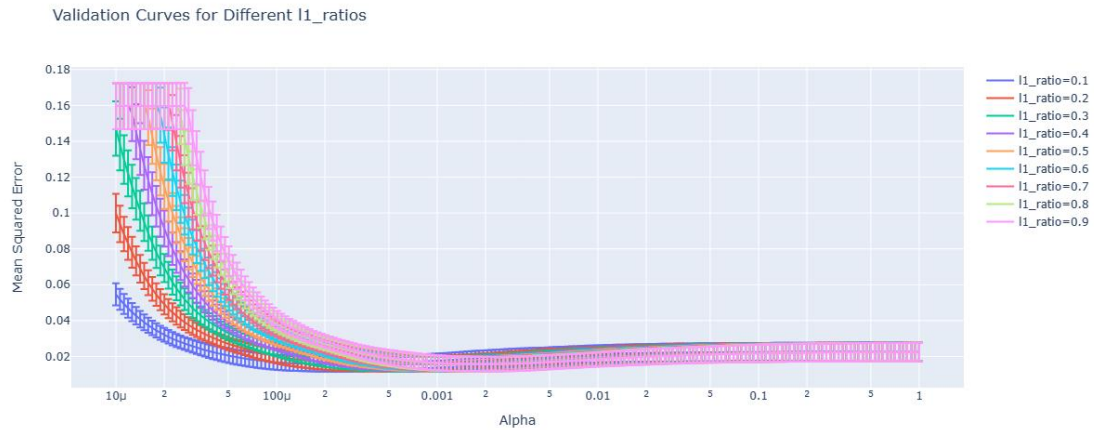
Plot 3: OLS and Lasso Coefficients Distribution

5. Applying Elastic Net

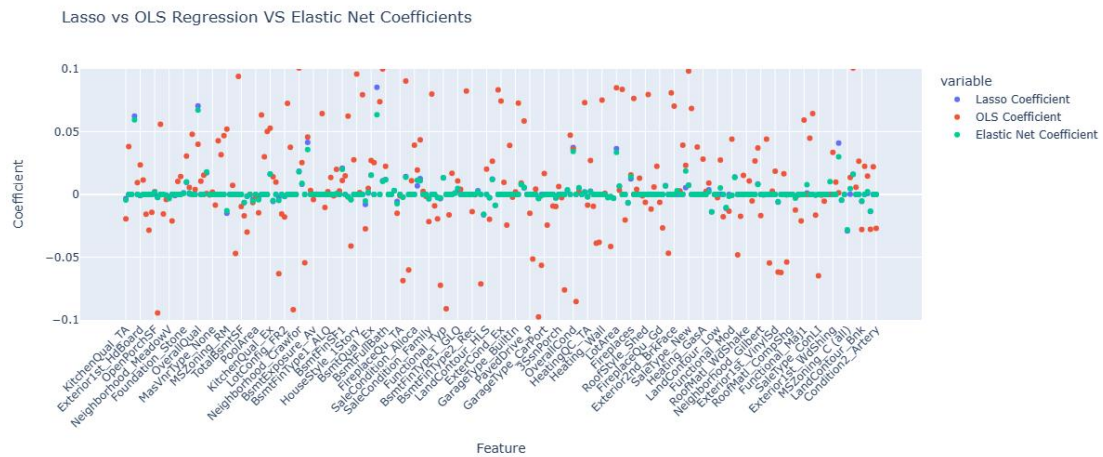
Furthermore, to deal with the problem about multicollinearity, we also employ the Elastic Net model and select the optimal regularization parameter as α and the optimal L1 regularization weight as λ_1 ratio.

After the parameter selection in Elastic Net, we put all three models in a graph together to watch their coefficients. We can see that the coefficients distribution in the Lasso model tends to be sparse, with a large portion being 0, while the Elastic Net

model maintains a balance between feature selection and coefficient stability.



Plot 4: Elastic Net Cross-Validation Results



Plot 5: OLS, Lasso and Elastic Net Coefficients Distribution

6. Model Selection and Prediction

Finally, although OLS has a lower MSE, in order to prevent overfitting and improve prediction performance, we plan to select parameters processed by the Lasso model for a new round of housing price prediction. According to Kaggle's results, it is evident that the new model has better prediction accuracy.

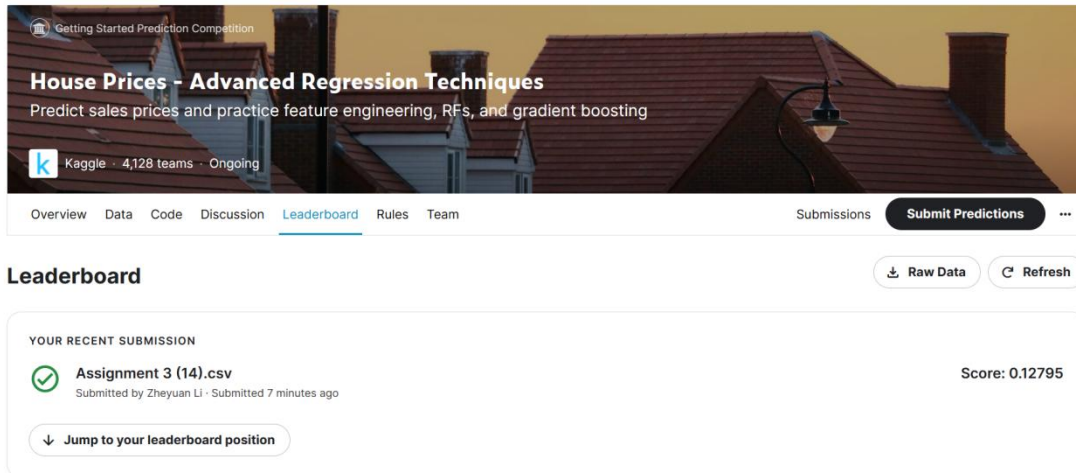
OLS	Lasso	Elastic Net
0.009589	0.011743	0.011811

Table 1: Mean Square Error in OLS, Lasso and Elastic Net

Python Code Link:

https://colab.research.google.com/drive/1S3J6Ufu047ufZLHQCK-46cqmzw1XmuQD?usp=drive_link

Kaggle Score:



The image shows the Kaggle competition page for "House Prices - Advanced Regression Techniques". The header includes the competition title, a description "Predict sales prices and practice feature engineering, RFs, and gradient boosting", and the Kaggle logo with "4,128 teams · Ongoing". The navigation bar has links for Overview, Data, Code, Discussion, Leaderboard (active), Rules, and Team. On the right, there are links for Submissions and a Submit Predictions button. Below the navigation bar, the "Leaderboard" section is displayed, featuring a "Raw Data" button and a "Refresh" button. The "YOUR RECENT SUBMISSION" section shows a green checkmark icon, the filename "Assignment 3 (14).csv", the submitter "Submitted by Zheyuan Li", the time "Submitted 7 minutes ago", and the score "Score: 0.12795". A button labeled "Jump to your leaderboard position" is also present.

Getting Started Prediction Competition

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Kaggle · 4,128 teams · Ongoing

Overview Data Code Discussion **Leaderboard** Rules Team Submissions **Submit Predictions** ...

Leaderboard

Raw Data Refresh

YOUR RECENT SUBMISSION

✓ **Assignment 3 (14).csv** Score: 0.12795
Submitted by Zheyuan Li · Submitted 7 minutes ago

↓ Jump to your leaderboard position