1. Abstracts

This new project makes improvements from various perspectives. Nonparametric methods, variables transformation, and ensemble prediction will be covered.

2. LOWESS Modeling

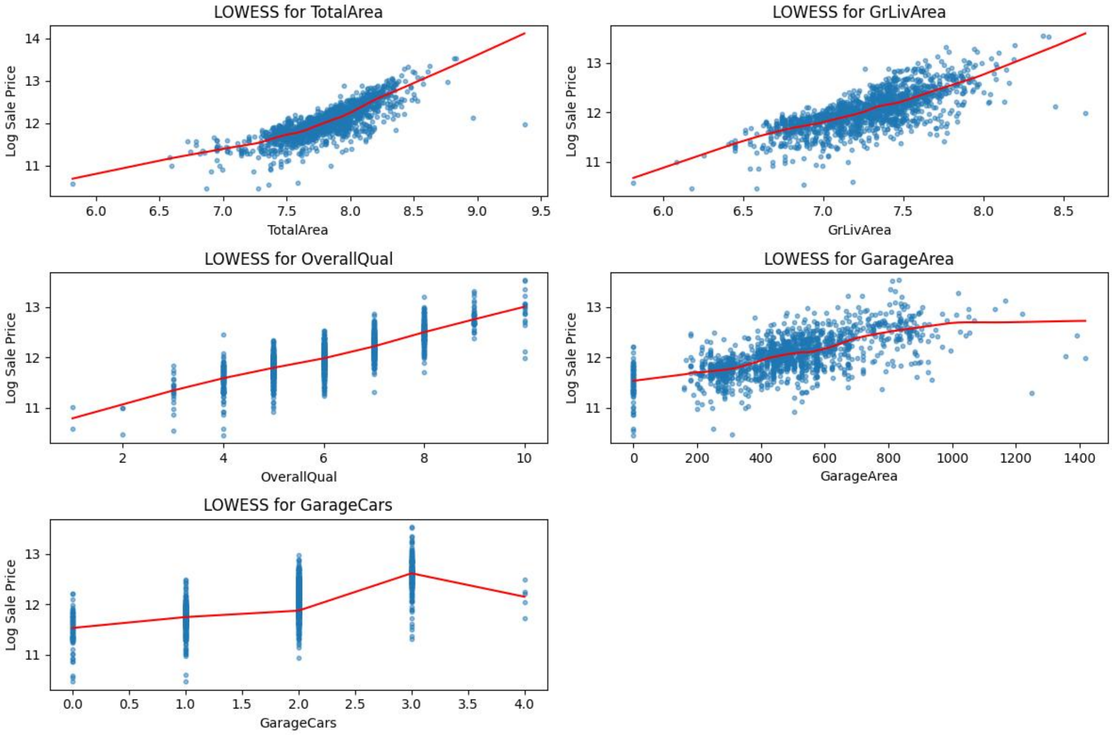
Some numerical features may not exhibit a linear relationship with sales prices, so the LOWESS method is introduced to smooth these features. Recognizing that excessive use of nonparametric models can lead to overfitting and time-consuming computations, the project selects the top 5 features with the highest correlation in LOWESS.

Feature	**Correlation**
OverallQual	0.8172
TotalArea	0.8071
GrLivArea	0.7303
GarageCars	0.6806
GarageArea	0.6509
1stFlrSF	0.609
FullBath	0.5948
YearBuilt	0.5866
YearRemodAdd	0.5656
TotRmsAbvGrd	0.5344

Table 1: 10 Features with the Highest Correlation with Sales Price

In order to strike a balance between the original and smoothed features, the features that undergo LOWESS processing will be added to the existing features, rather than replacing them directly.

In the plot of LOWESS, it can be seen that it captures local features in the data through weighted least square, reducing the impact of outliers.



Plot 1: LOWESS Results

3. Variables Transformation and Scaling

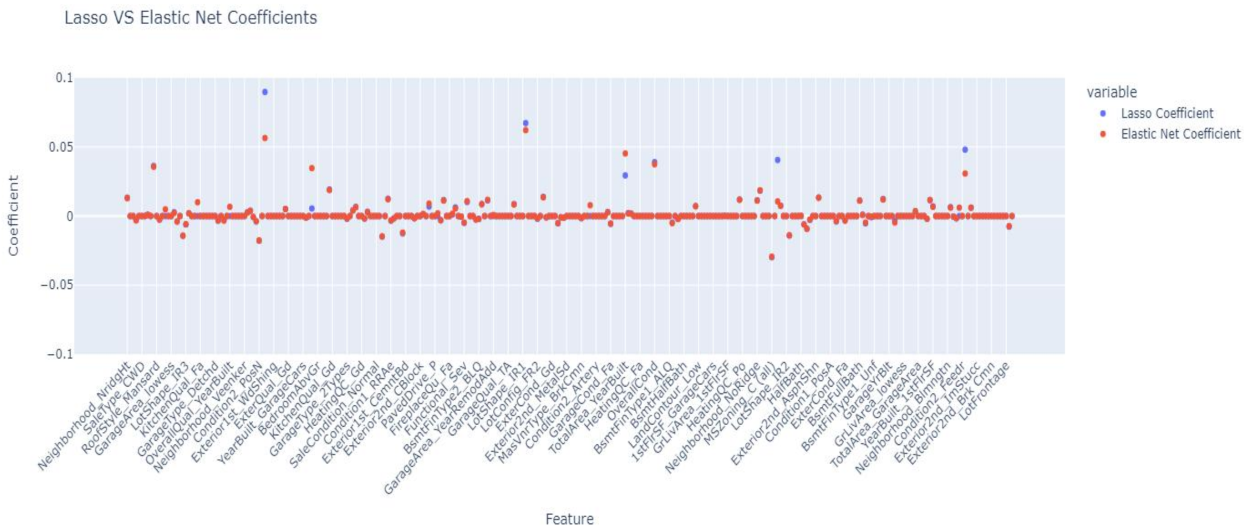
Interactions between different features can reveal deeper relationships and enhance prediction accuracy. Therefore, it is beneficial to transform certain features by creating their interaction terms.

Creating interactions for all variables would result in over 30,000 combinations. To improve efficiency, the project selects 10 features from Table 1, generates two-way interaction terms for them as new features, and incorporates these terms into the dataset. Prior to regularization, the dataset is also standardized to control the penalty terms.

4. Regularization

Lasso and Elastic Net are applied in regularization. Simply using Lasso for regularization may result in multicollinearity. Therefore, the project introduced

Elastic Net that utilizes L2 regularization to handle small coefficients. From the graph, it can be seen that the participation of Elastic Net makes regularization more flexible, rather than directly punishing coefficient to 0.



Plot 2: Lasso and Elastic Net Coefficients Distribution

5. Ensemble Prediction

Generally, model selection is based on MSE, but considering the advantages of Elastic Net, the project combines the results of Lasso and Elastic Net, taking the average of their prediction results, and establishes an ensemble model for prediction. Finally, the Kaggle score indicates that it has made progress compared to the previous assignment.

Lasso	Ensemble Model	Elastic Net
0.01155	0.011604	0.011664

Table 2: Mean Square Error in Lasso, Ensemble Model and Elastic Net

Python Code Link:

https://colab.research.google.com/drive/1qPUbnCaHkERT0xvMuQU_Vznt17tj7J5_?usp=drive_link

Kaggle Score: