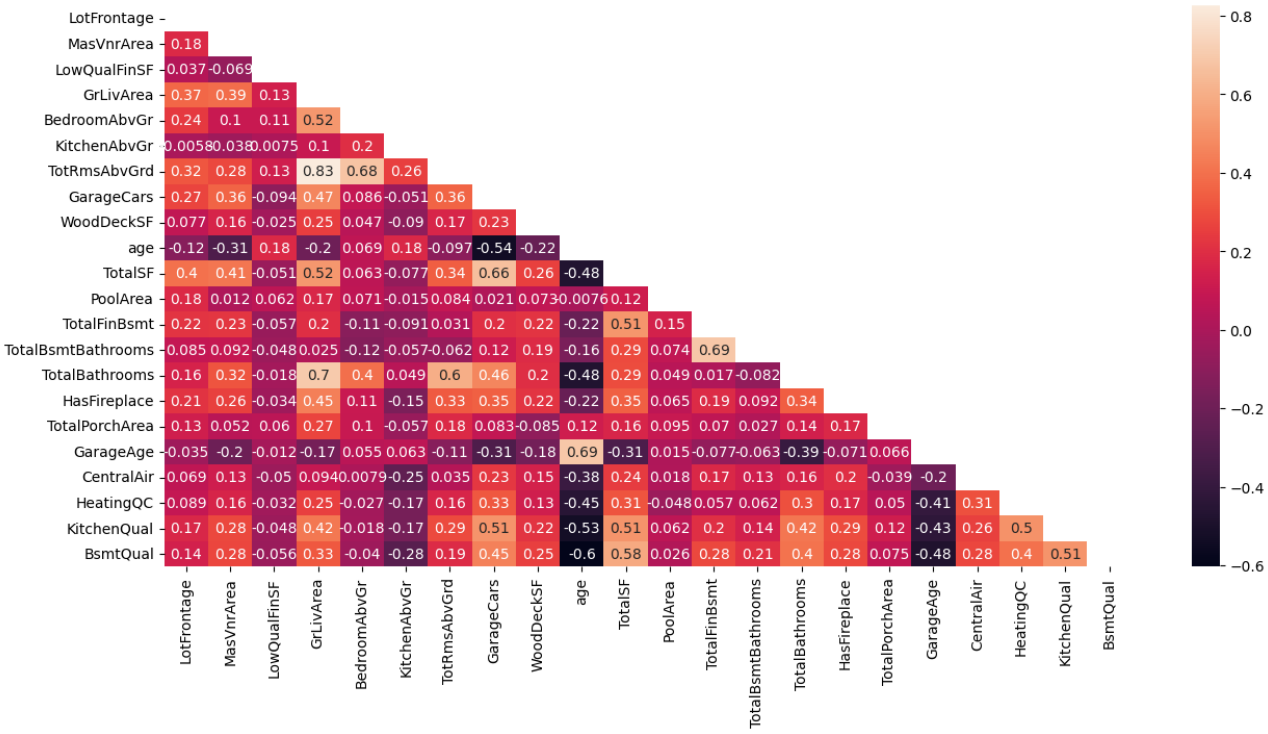
Numerical dataset

General Information on dataset:

the dataset used is House Prices - Advanced Regression Techniques from Kaggle. This dataset provides information related to residential properties, aiming to predict the sale price of houses. The key attributes include details about the building class, zoning classification, lot features (frontage, area, shape), property access and configuration, neighborhood, overall quality and condition, construction details (year built, remodeled), roof characteristics, exterior features, basement details, heating and air conditioning systems, electrical setup, room dimensions, garage attributes, and additional features like porch, pool, and fence. The target variable is "SalePrice," representing the property's sale price. Analyzing this dataset can offer insights into factors influencing housing prices and facilitate the development of predictive models. The dataset contains 80 features with 1460 training examples.

The correlation between Features.



Implementation details:

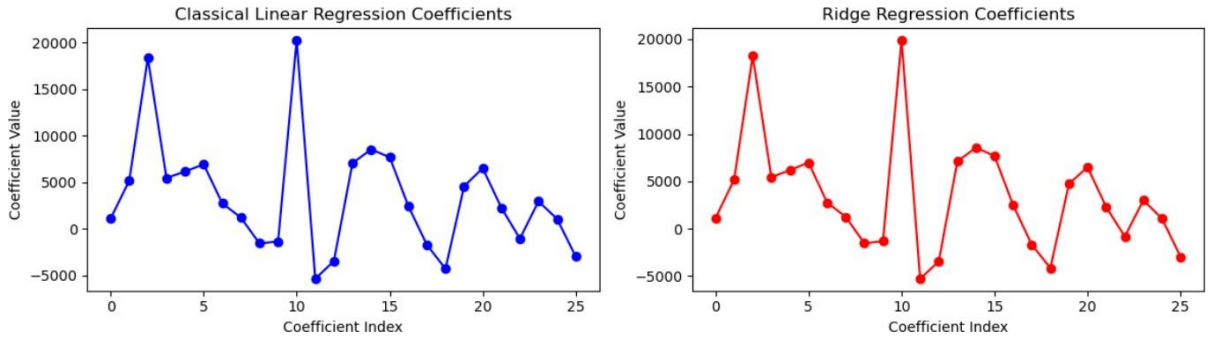
After using feature selection and engineering, only 27 features have been used with all training examples(1460). Those names are

['LotFrontage', 'LotArea', 'OverallQual', 'OverallCond', 'MasVnrArea',
'ExterQual', 'BsmtQual', 'HeatingQC', 'CentralAir', 'LowQualFinSF',
'GrLivArea', 'BedroomAbvGr', 'KitchenAbvGr', 'KitchenQual',
'TotRmsAbvGrd', 'GarageCars', 'WoodDeckSF', 'PoolArea', 'SalePrice',
'age', 'TotalSF', 'TotalFinBsmt', 'TotalBsmtBathrooms',
'TotalBathrooms', 'HasFireplace', 'TotalPorchArea', 'GarageAge']
all the columns with shape (1460,1).

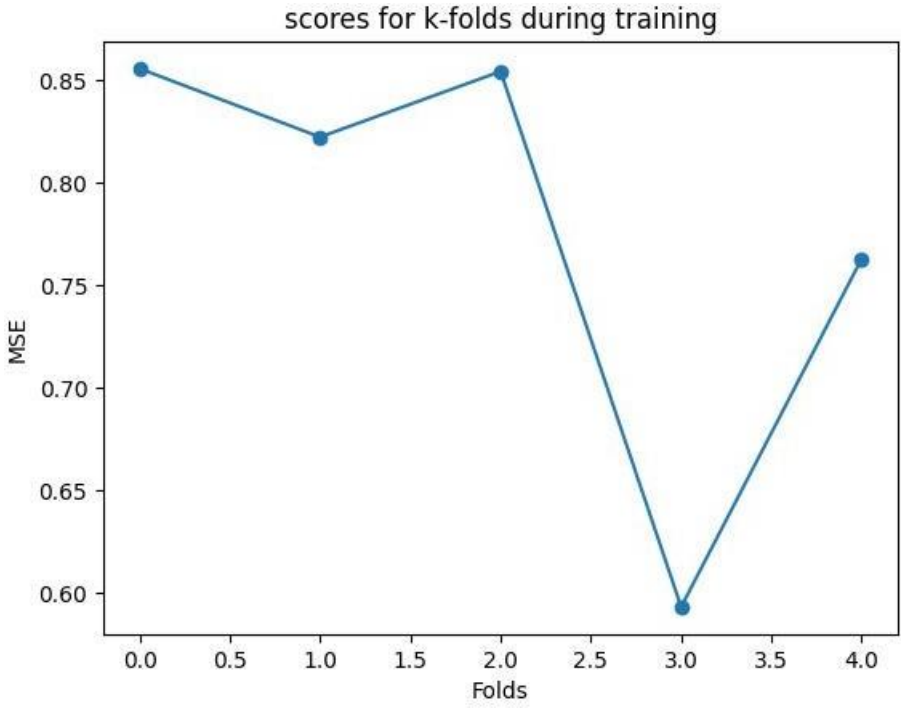
the dataset was divided into two parts(test_set and training_set) with a ratio of 80% for the training set and 20% for the testing set, then cross-validation was used in th e regularized linear regression(Ridge) model with 5 splits, and the splits were chos en randomly such that four splits were used for the training and one for testing in e ach iteration. In the regularized linear regression(Ridge) model, the grid was used t o select the best hyperparameter alpha from the values [0.01,0.1,1,2,5]. In the k-ne arest neighbors regressor, for loop was initialized in different k values to calculate which one would minimize the mean square error. The value of k=3 was the best v alue with an error of 0.83.

Results Details

for linear regression, the cv score mean was 77.696223%, and the accuracy was 87. 6867%. the ridge had a close accuracy with 87.6794%.graph of the coefficients of the two models:



The CV scores is shown below:



1. **High Mean Test Score (0.8):** A mean test score of 0.8 is relatively high and suggests that, on average, the model is able to make accurate predictions on the held-out validation sets during cross-validation.
2. **Low Variability (Standard Deviation ≈ 0.1):** The standard deviation of around 0.1 indicates that the individual test scores are clustered around the mean of 0.8, and there is limited variability or uncertainty in the model's performance across different folds.
3. **The model was stable with different alpha**

for the knn model, the accuracy was 85.943 %. different k values were examined and the value 3 was the best as shown below :