House Prices

# House Prices: Advanced Prediction Techniques (Kaggle)

Ritesh Kumar

2024WI\_MS\_DSP\_422-DL\_SEC61: Practical Machine Learning

Module 3 Assignment

# House Prices: Advanced Prediction Techniques

# Donald Wedding and Narayana Darapaneni

# January 21, 2024

**Approach**

We start with a comprehensive Exploratory Data Analysis (EDA), commencing with the extraction of descriptive statistics for the 'SalePrice' column, serving as the target variable for prediction. This entails computing the count, mean, standard deviation, minimum, quartiles, and maximum values.

A graph with blue lines

Description automatically generatedA graph with a blue rectangular bar

Description automatically generated

The script meticulously identifies columns with missing data, presenting the count and names of these columns. We remove all features that exhibit over 90% missing values. For the remaining features, we infer that rows with missing values indicate the absence of those features in the respective properties. Consequently, we substitute all missing values with 0.A graph of a number of data

Description automatically generated

Post-handling missing data, categorical features are recognized based on unique values, with those containing eight or fewer unique values considered as such. Further refinement involves dropping columns with less than 1% of values for more than half of the categories, a strategic dimensionality reduction.

The script endeavors to engineer new features pertinent to prediction, summing up square footage and porch-related columns. A Chi-Square test evaluates categorical variable correlation, ensuring independence.

Relationship of SalePrice with the most important categorical and numerical features is explored.

A screenshot of a graph

Description automatically generated A screenshot of a computer

Description automatically generated

All numerical features are analyzed for outliers, and only two of 28 numerical features are found to be having more than 10% outliers.

Random Forest algorithm is used for identification of the most influential features, encompassing both numerical and categorical variables, that exert the strongest impact on the 'SalePrice.

A graph with blue and white bars

Description automatically generated

Two new variables – TotalSF and Porch are created by adding all the variables for the covered area and porch (in sf), respectively. And, the relationship of SalePrice with both is explored.

A graph showing a number of blue dots

Description automatically generatedA graph showing a number of blue dots

Description automatically generated

Highly correlated features, both categorical and numerical, are identified, and removed.

Feature list is refined. The script specifies features for removal, optimizing the dataset for model training.

Application of the 'OneHotEncoder' from sklearn.preprocessing converts categorical features to a one-hot numeric array, enhancing compatibility with classification and regression algorithms.

Ridge regression, Lasso regression, and ElasticNet models are trained, incorporating 5-fold cross-validation to gauge accuracy and prevent overfitting.

Subsequently, the script prepares test data ('test.csv') mirroring operations applied to the training data. Trained models predict test data outcomes, stored in CSV files.

These files were uploaded on Kaggle (<https://www.kaggle.com/riteshrk>) , and the scores are:

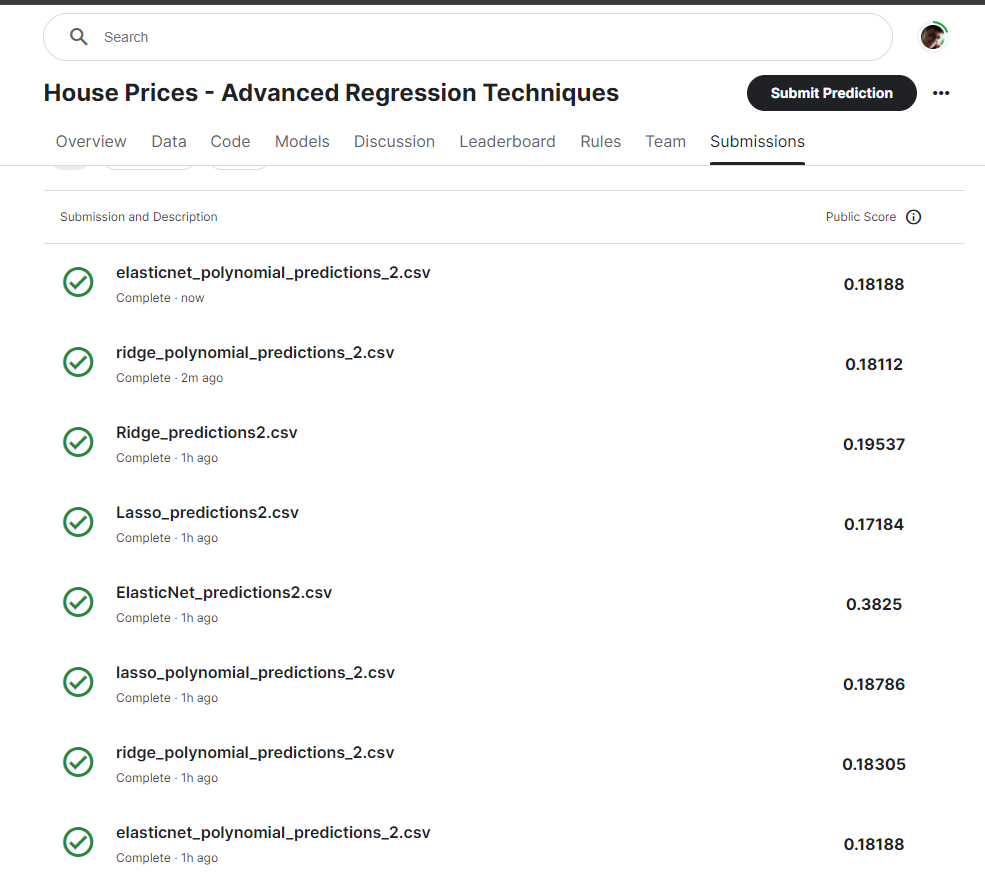
* Elastic Net: 0.3825
* Ridge: 0.19537
* **Lasso: 0.17184**

Next, we used to scikit-learn on Lasso to perform a grid-search for hyperparameter tuning (polynomial\_degree: 1 to 3, alpha [50, 100, 250]), and Cross-Validation = 5, and the best parameters were found to be alpha = 250 and polynomial\_degree = 1. Kaggle submission score is 0.18786.

After that, we used to scikit-learn on Ridge to perform a grid-search for hyperparameter tuning (polynomial\_degree: 1 to 3, alpha [10, 15, 20, 25, 30, 35, 40]), and Cross-Validation = 5, and the best parameters were found to be alpha = 25 and polynomial\_degree = 1. Kaggle submission score is 0.18112.

Finally, we used to scikit-learn on ElasticNet to perform a grid-search for hyperparameter tuning (polynomial\_degree: 1 to 3, alpha [0.01, 0.05, 0.1], l1\_ratio [0.4, 0.5, 0.6], and Cross-Validation = 5, and the best parameters were found to be alpha = 0.05, l1\_ratio = 0.5 and polynomial\_degree = 1. Kaggle submission score is 0.18188.

The submission results can be viewed at <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/submissions>.



The best score 0.17184 on Kaggle is returned by our Lasso model.

The jupyter notebook has been uploaded on Kaggle and can be viewed at <https://www.kaggle.com/riteshrk/lasso-ridge-elasticnet/edit>.

We have identified the 15 features with the highest coefficients from the optimal model i.e.

the Lasso:

A screenshot of a computer

Description automatically generated

**Coefficients:** The coefficients returned by a LASSO regression model represent the strength and type of the relationship between each feature and the target variable, SalePrice.

Breakdown of the coefficients:

1. LotArea: A larger lot area significantly increases property values.
2. KitchenAbvGr: A higher number of above-ground kitchens (like in a duplex or apartment building) seems to have a very high positive influence on property value.
3. BedroomAbvGr\_8: Properties with eight bedrooms above ground are valued higher, possibly indicating large family homes or commercial residential investments like B&Bs.
4. GarageYrBlt: Newer garages seem to add to property value, suggesting that modern construction or renovations are favorable.
5. FullBath\_2: Having two full bathrooms above grade is a significant positive for property value.
6. YearRemodAdd: Recently remodeled or added constructions boost the property's value.
7. LotFrontage: Having a larger street connected to the property marginally increases value.
8. Fireplaces\_1: The presence of at least one fireplace adds value, indicating a preference for this feature.
9. MiscVal: Miscellaneous values also play a role in increasing property value, suggesting unique features or amenities are important.
10. MasVnrType\_Stone: A stone masonry veneer is highly valued, indicating a preference for stone finishes.
11. Foundation\_CBlock: Cinder block foundations are valued, possibly for their durability or the property style they signify.
12. LandSlope\_Gtl: Properties with a gentle slope have a slight positive impact on value.
13. HalfBath\_0: No half baths surprisingly increase value, which may indicate a preference for more full baths or simpler layouts.
14. OverallQual: Overall quality has a strong positive impact on value, as expected.
15. FireplaceQu\_0: Having no fireplace or a low-quality fireplace reduces the value, indicating the desirability of higher-quality fireplaces.

**Management Recommendations:** Given these findings, several strategic recommendations can be made:

1. Focus on Quality and Modern Features: Prioritize high-quality construction and modern amenities like new garages and stone masonry.
2. Maximize Value-Adding Features: Include at least two full bathrooms and ensure there are fireplaces in the properties, as they have shown to increase value significantly.
3. Highlight Unique Features: Any unique or miscellaneous features that a property possesses should be emphasized in listings, as they can add substantial value.
4. Invest in Kitchen and Bedroom Optimizations: Properties with multiple kitchens or a larger number of bedrooms fetch higher prices and might appeal to investors in rental properties or large family homes.
5. Consider Lot Characteristics: Maintain or improve lot frontage and ensure the property has a gentle slope, as these aspects have positive influences on value.
6. Capitalize on Renovations: Market the value of recent remodels or additions, as they are shown to increase property values.

By leveraging these insights, property managers, real estate agents, and developers can make informed decisions to enhance property appeal and market value effectively.

**Code**