Understanding Factors influencing Residential Property Prices through Predictive Modeling and EDA Insights'

Anthony Amadasun

Overview

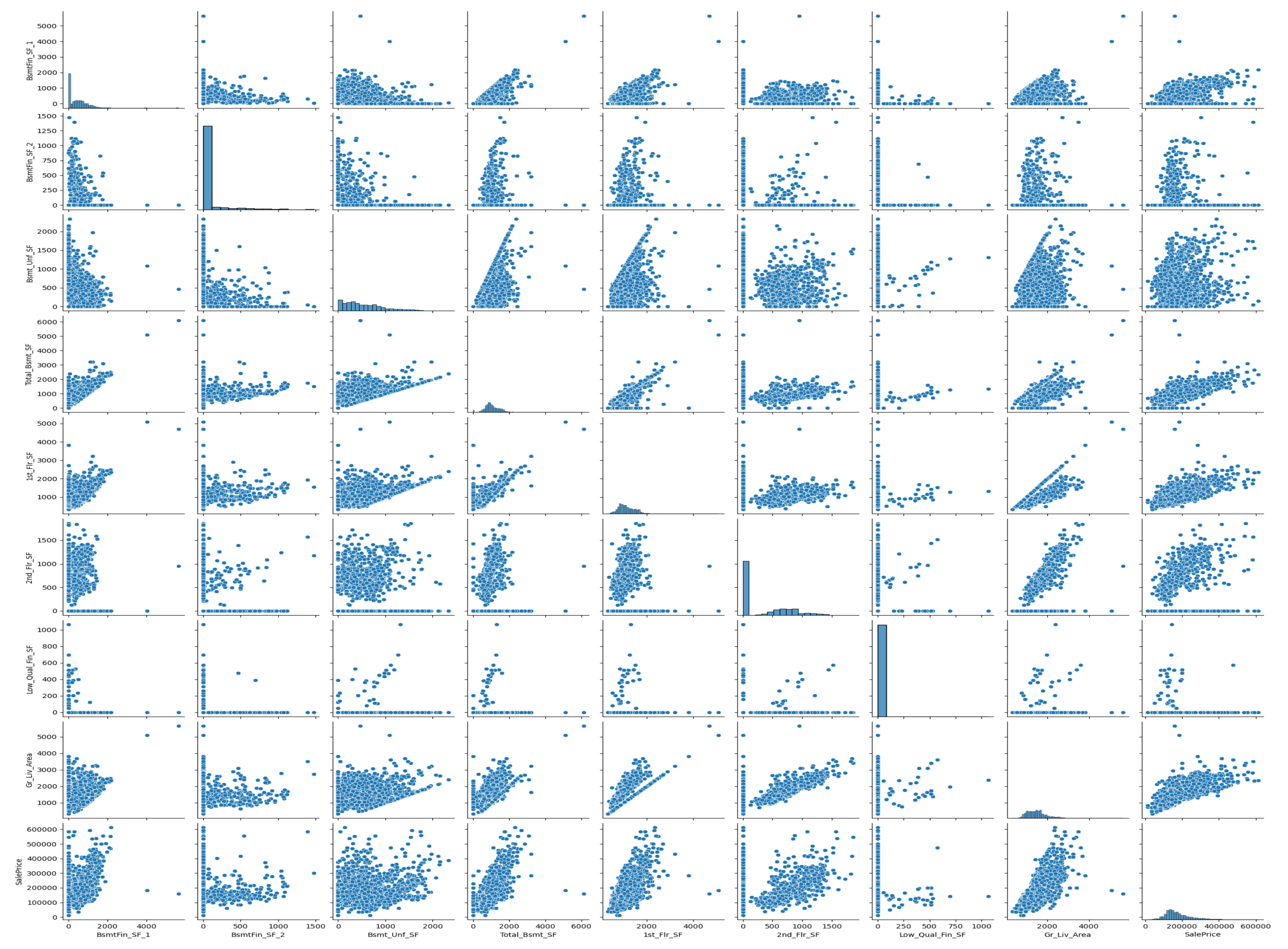
- This Project look at dataset containing information from the Ames Assessor's Office used in computing assessed values for individual residential properties sold in Ames, IA from 2006 to 2010.
- 82 features
- 2930 Observation
- Nearly 1500 Homes in Ames Iowa!!
- 23 nominal, 23 ordinal, 14 discrete, and 20 continuous variables

What are some of the objective?

- Predict sales price by minimizing the difference between predicted and actual values.
- performance on a separate test dataset should be comparable to the training dataset to show the model's generalizability
- EDA process: uncover meaningful visual and patterns that provide actionable insight into the data.

Data Cleaning

- Features dropped: BsmtFin_SF_1', 'BsmtFin_SF_2', 'Bsmt_Unf_SF', 'Total_Bsmt_SF','1st_Flr_SF', '2nd_Flr_SF’, Ms_Zoning, and 'Low_Qual_Fin_SF



Exploratory Data Analysis (EDA)

- **Pool QC, Misc Feature, Alley, Fence, Mas Vnr Type, Fireplace Qu, Garage Qual, Garage Finish, Garage Cond, Garage Type, Bsmt Exposure, BsmtFin Type 2, Bsmt Cond, and Bsmt Qual** have missing values because the corresponding features are not present for certain house. Rather than drop NA, replace this missing value with "None" category instead.

- **Lot Frontage** missing values indicate that information about the linear feet of street connected to the property is not available. Will need to fill missing values using imputation(the median or mean)

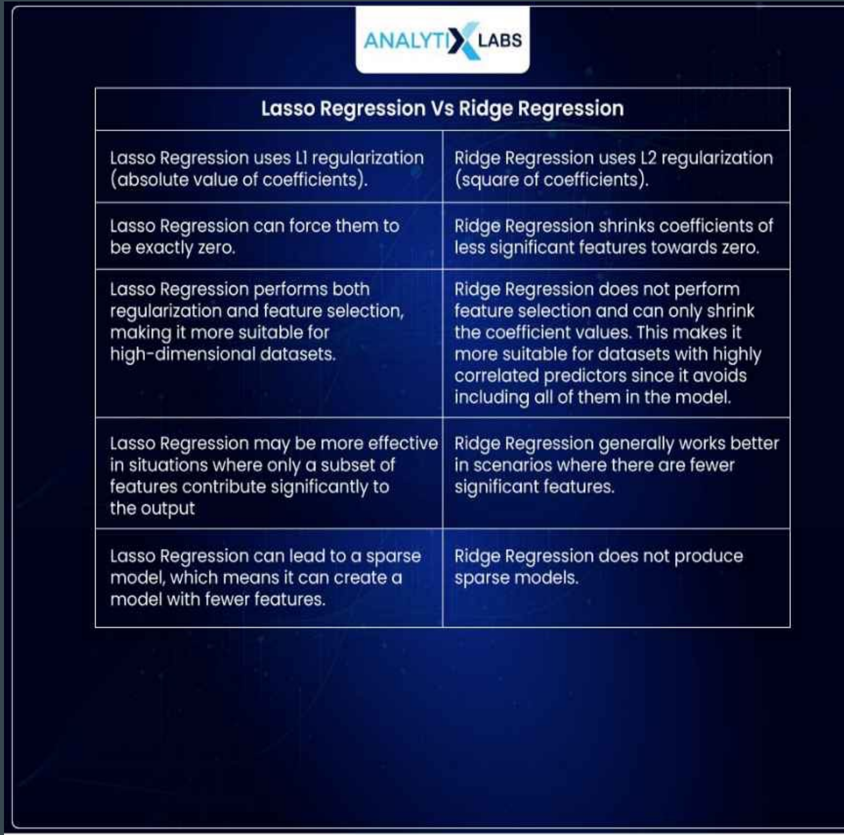
- **Mas Vnr Area, Garage Yr Blt, BsmtFin SF 1, BsmtFin SF 2, Bsmt Unf SF, Total Bsmt SF, Garage Cars Bsmt Full Bath, and Bsmt Half Bath**, missing value indicates the house doesnt have those features. Will need to fill missing value with 0

Model Building

- **combine any feature? Interactive term, One hot encoding**
- The neighborhood variables has 28 unique neighborhoods

Model building, evaluation, tuning, production model

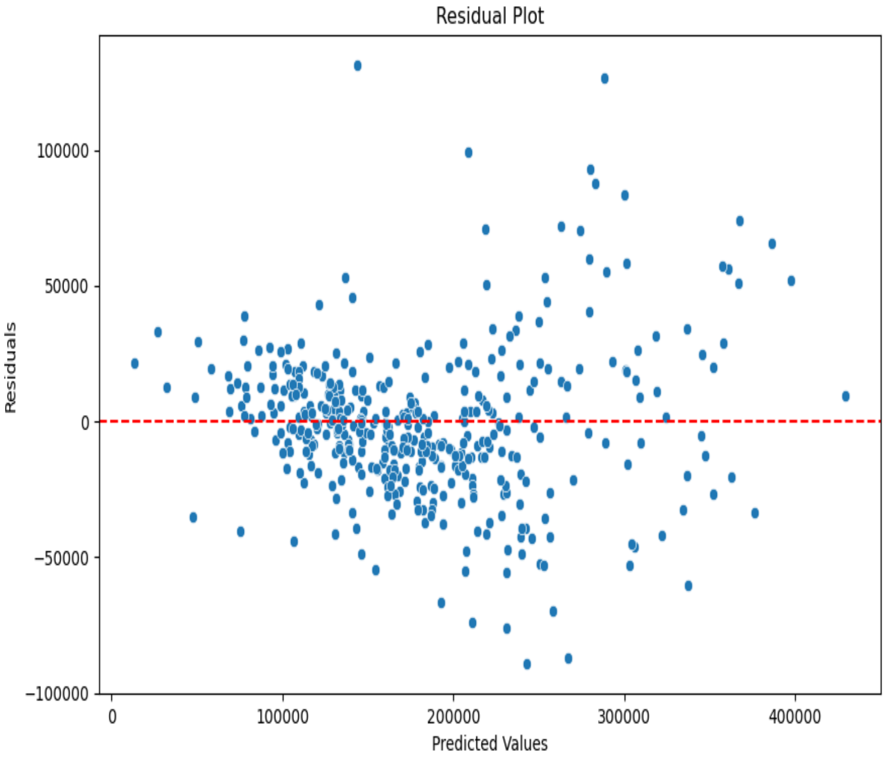
- Linear Regression: MAE: 20278.8879, R^2: 0.8648 R^2: 0.8648
- LASSO Regression: MAE: 20280.1919, R^2: 0.8649
- Ridge: MAE: 20277.9923, R^2: 0.8653
- Tuned LASSO Regression Results: MAE: 19902.55, R^2: 0.8687
- Tuned Ridge Regression Results: MAE: 19930.94, R^2 (Test): 0.8655
- Production Model = Lasso

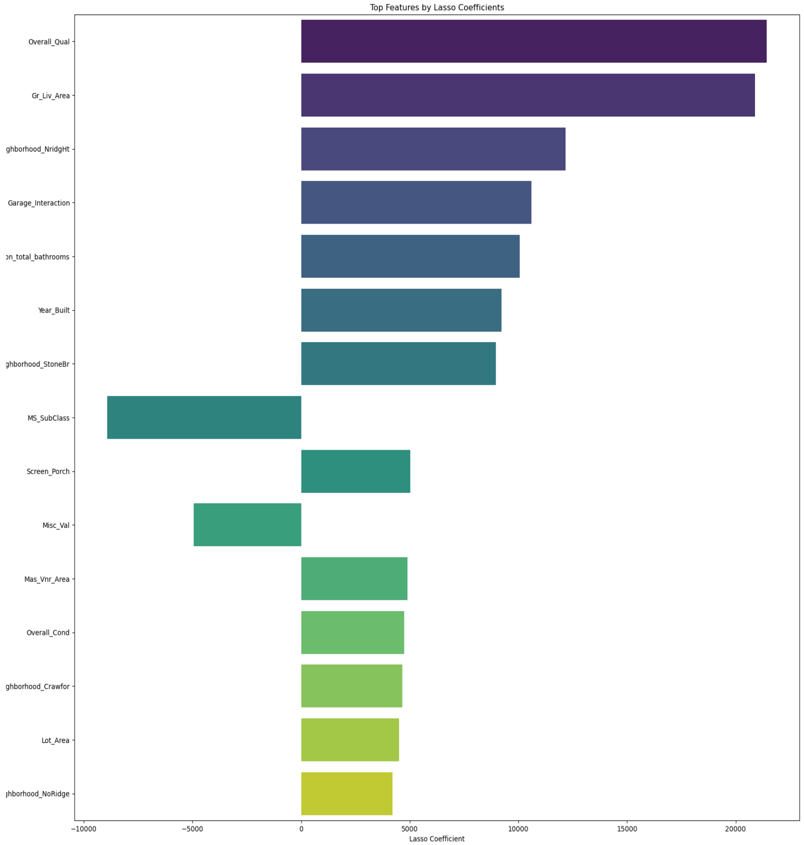


ANALYTIX LABS

Lasso Regression Vs Ridge Regression	
Lasso Regression uses L1 regularization (absolute value of coefficients).	Ridge Regression uses L2 regularization (square of coefficients).
Lasso Regression can force them to be exactly zero.	Ridge Regression shrinks coefficients of less significant features towards zero.
Lasso Regression performs both regularization and feature selection, making it more suitable for high-dimensional datasets.	Ridge Regression does not perform feature selection and can only shrink the coefficient values. This makes it more suitable for datasets with highly correlated predictors since it avoids including all of them in the model.
Lasso Regression may be more effective in situations where only a subset of features contribute significantly to the output	Ridge Regression generally works better in scenarios where there are fewer significant features.
Lasso Regression can lead to a sparse model, which means it can create a model with fewer features.	Ridge Regression does not produce sparse models.

Key Findings





Conclusion and actionable insight

- Buyers are willing to pay more for houses with better quality.
- More living space tends to increase the value of a property.
- Being in a certain neighborhood is associated with higher sale prices.
- Newer houses might be perceived as more valuable due to modern features and construction.
- Houses with more bathrooms or a specific combination may command higher prices.
- Thanks for your time!