Regression Modeling

Predicting Sale Prices of Houses
in Ames, Iowa
Between 2006 and 2010 inclusive

Overview

1. Problem Statement
2. Exploratory Data Analysis
3. Baseline Regression Modeling
4. Regularized Regression Modeling
5. Conclusion and Recommendations

Problem Statement

To predict the sale price of a house in Ames, Iowa sold between 2006 and 2010 inclusive, and to identify the most influential factor on the sale price of a house in Ames within the same timeframe

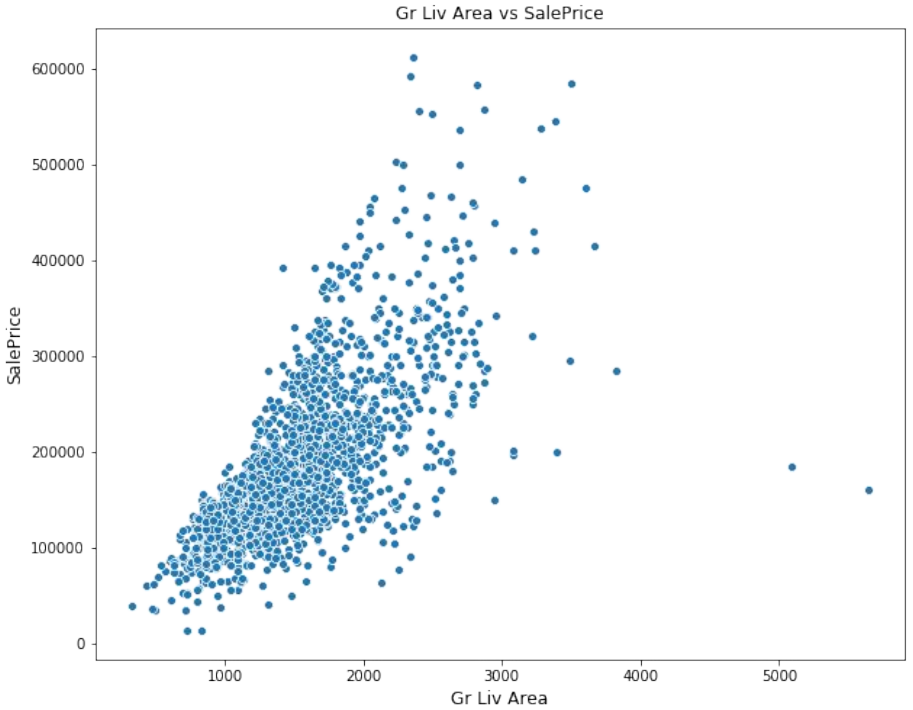
Exploratory Data Analysis

- Data Types

Discrete	Continuous	Nominal	Ordinal
'Year Built', 'Year Remod/Add', 'Bsmt Full Bath', 'Bsmt Half Bath', 'Full Bath', 'Half Bath', 'Bedroom AbvGr', 'Kitchen AbvGr', 'TotRms AbvGrd', 'Fireplaces', 'Garage Yr Blt', 'Garage Cars', 'Mo Sold', 'Yr Sold'	'Lot Frontage', 'Lot Area', 'Mas Vnr Area', 'BsmtFin SF 1', 'BsmtFin SF 2', 'Bsmt Unf SF', 'Total Bsmt SF', '1st Flr SF', '2nd Flr SF', 'Low Qual Fin SF', 'Gr Liv Area', 'Garage Area', 'Wood Deck SF', 'Open Porch SF', 'Enclosed Porch', '3Ssn Porch', 'Screen Porch', 'Pool Area', 'Misc Val'	'PID', 'MS SubClass', 'MS Zoning', 'Street', 'Alley', 'Land Contour', 'Lot Config', 'Neighborhood', 'Condition 1', 'Condition 2', 'Bldg Type', 'House Style', 'Roof Style', 'Roof Matl', 'Exterior 1st', 'Exterior 2nd', 'Mas Vnr Type', 'Foundation', 'Heating', 'Central Air', 'Garage Type', 'Misc Feature', 'Sale Type'	'Lot Shape', 'Utilities', 'Land Slope', 'Overall Qual', 'Overall Cond', 'Exter Qual', Exter Cond', 'Bsmt Qual', 'Bsmt Cond', 'Bsmt Exposure', 'BsmtFin Type 1', 'BsmtFin Type 2', 'Heating QC', 'Electrical', 'Kitchen Qual', 'Functional', 'Fireplace Qu', 'Garage Finish', 'Garage Qual', 'Garage Cond', 'Paved Drive', 'Pool QC', 'Fence'

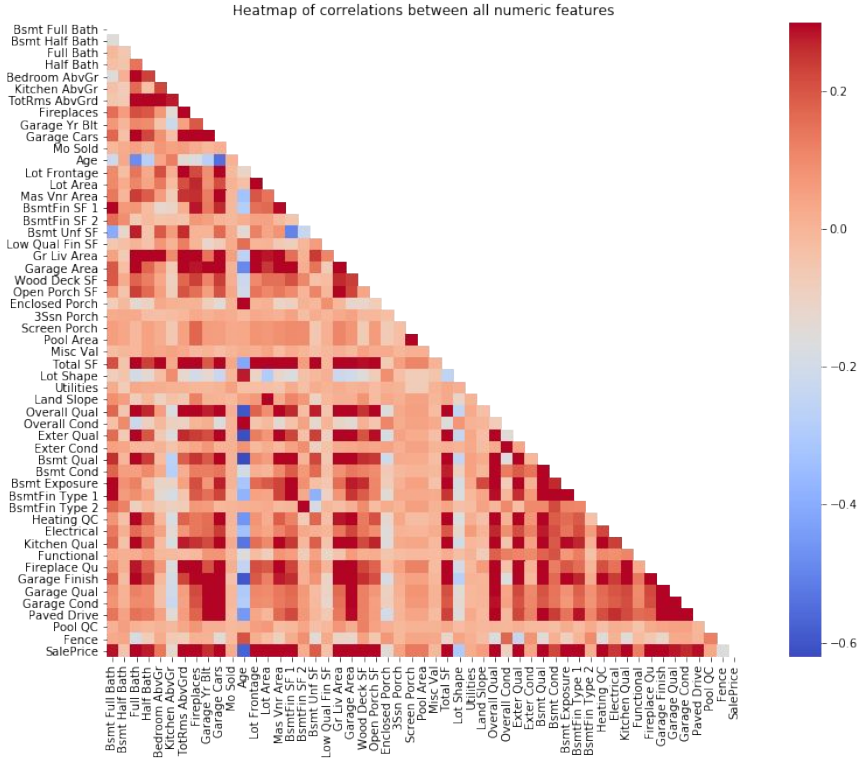
Exploratory Data Analysis

- Scatterplot of SalePrice vs GrLivArea
  - Observed two outliers with GrLivArea > 4000
  - Removed outliers



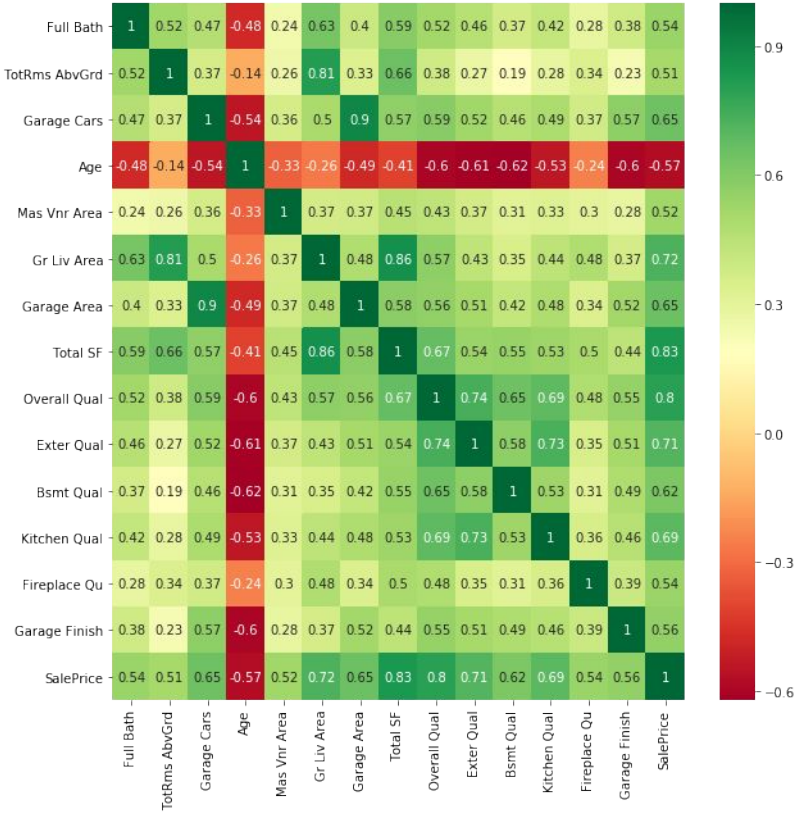
Exploratory Data Analysis

- Heatmap was plotted to investigate multicollinearity
- There are features in the dataset that describe a common attribute of the house such as Garage Cars, Garage Yr Blt.
- These features tend to exhibit high correlation among themselves



Exploratory Data Analysis

- Age is highly correlated with all the other variables except Fireplace Qu, TotRms AbvGrd



Baseline Regression Modeling

- All features were added into the model
- RMSE value: 20715.89
- Adjusted R2: 0.924

Regularized Regression Model

- A subset of features were first selected
  - Numeric features
    - Selected those with highest correlation with target variable, SalePrice
    - Dropped those with high multicollinearity e.g. Age
  - Nominal features
    - Using the ANOVA test, those with p-value < 0.05 were selected
- Regularized regression models built:
  - RMSE
    - Ridge: 27464.229
    - Lasso: 27394.428
    - ElasticNet: 27394.474
- Lasso was selected as it had the lowest RMSE

Regularized Regression Model

- The selected lasso regression model had 45 features:
- The features with the highest coefficients are:
  - Total SF, Total SF^2
  - Suggests that the size of a house influences the sale price of the same house far greater than other features

Conclusion and Recommendation

- The lasso regression model with a selected number of numeric and categorical features had the lowest RMSE
- Take note that this model may not necessarily be reliable when predicting sale price houses in other cities where the climate is not the same
- Also, the model may be not be reliable when predicting sale prices of houses in Ames beyond the time period of 2006 and 2010