House Price Prediction

Can we predict the price of a house based on multiple parameters and features by examining the prices of other houses?

Introduction:

Homebuyers tend to have certain requirements when looking for a house to purchase. These requirements could be attributes such as number of bedrooms, square footage, location, age of the house, and most importantly, the price of the house. With data describing residential homes in Ames, Iowa, the goal is to create a model that confidently predicts the price of each home.

Data Wrangling:

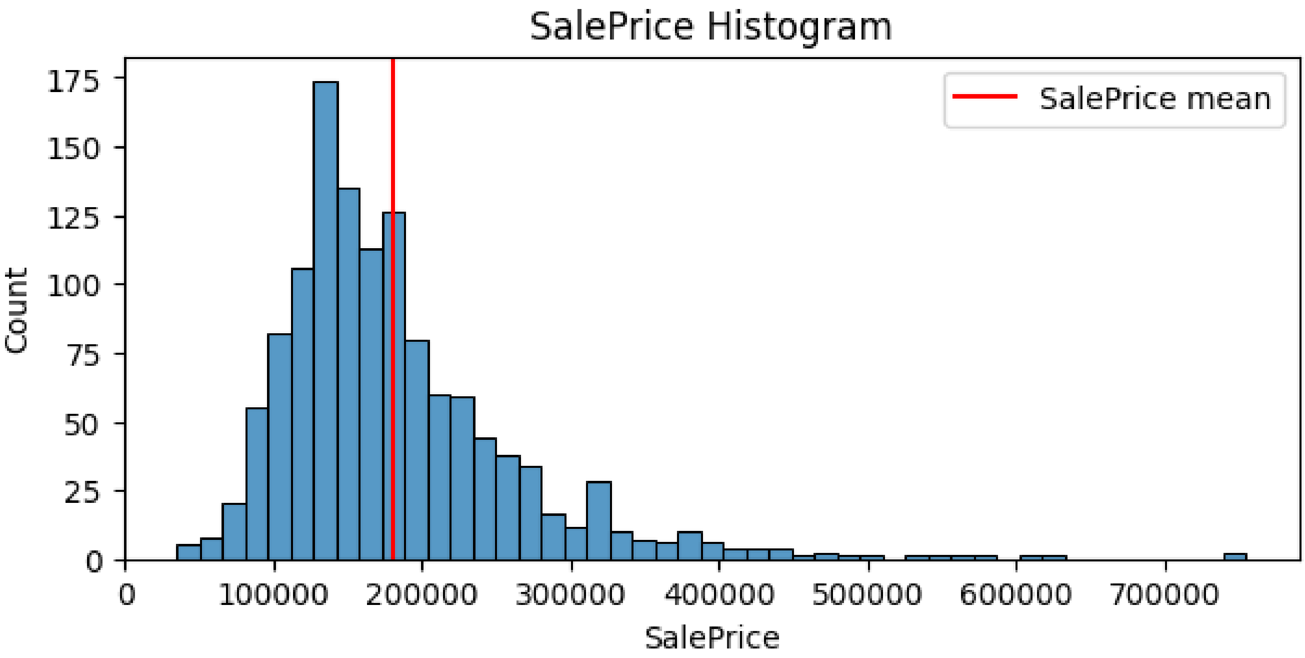
The data provided by Kaggle contained three files, train.csv, test.csv and a .txt file containing description of the data. The data from the train.csv file will be the only data used since a train/test split will be created from this data to train the model. The train.csv file contains 1259 entries with 81 columns about multiple factors describing a house.

The majority of cleaning required handling missing data. There were four columns, PoolQC, MiscFeature, Alley, and Fence that had missing data on over 80% of the entries. These columns were dropped. FireplaceQu also had a large percentage (47.6%) of missing data and will also be dropped since they will not provide any useful insights.

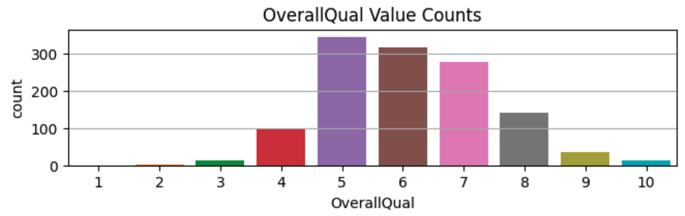
There were also a couple of columns that had heavily imbalanced classes and therefore would not provide any useful insights. These columns were also dropped. There were also a couple of features that very closely resembled each other and were therefore redundant, such as GarageCars and GarageArea. GarageCars was dropped. After all this processing, there were 59 remaining useful columns.

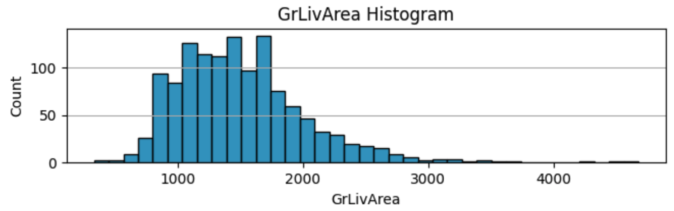
Exploratory Data Analysis:

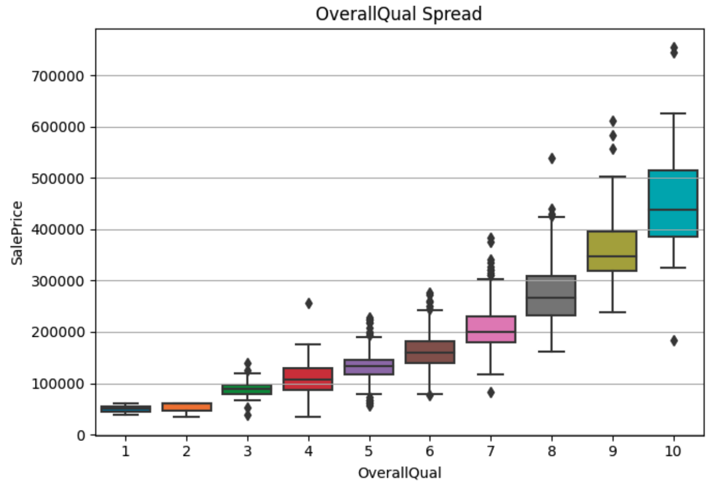
There were a few notable discoveries found while exploring each feature. Firstly, the spread of the target is right skewed with an average of 181,114 and a standard deviation of 80,588.

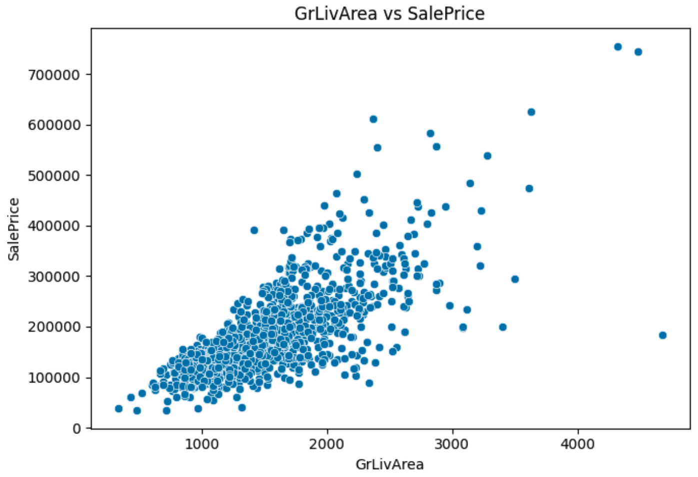


Each feature was visualized in a manner that depended on the type of data. Categorical and discrete features were visualized by count plots and box plots while continuous features were visualized by histograms and scatter plots.


OverallQual Value Counts

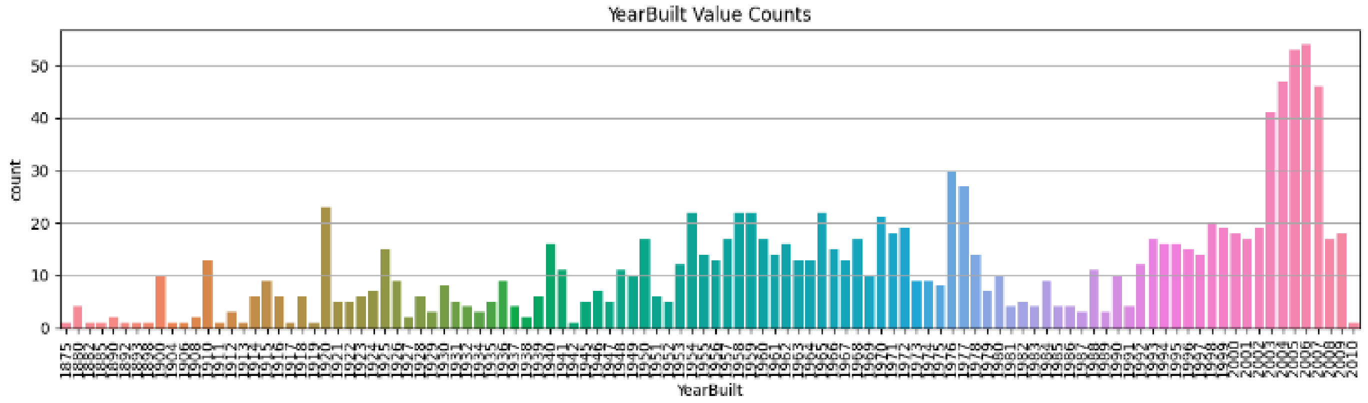

GrLivArea Histogram

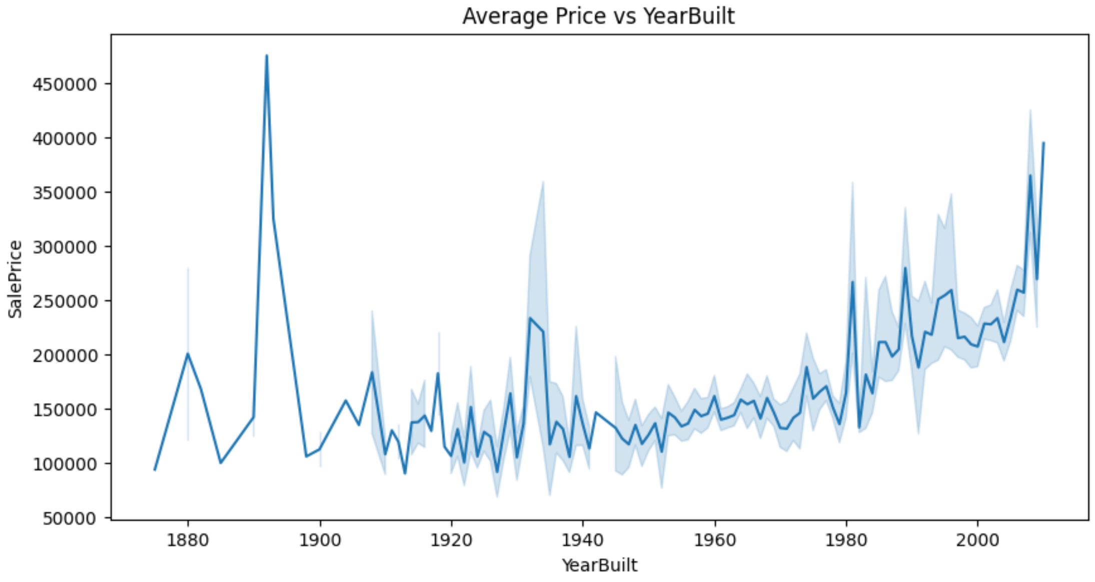

OverallQual Spread


GrLivArea vs SalePrice

There are a few features that showed a strong relationship with the sale price, including overall quality (OverallQual) and above ground living area(GrLivArea).

There is also a notable increase in average price with YearBuild. There is a lot more variation in average sale price for houses built before 1910 since there was significantly fewer samples from those years.


YearBuilt Value Counts


Average Price vs YearBuilt

Preprocessing and Training:

Since there are multiple categorical features, they had to be one-hot encoded in order to properly use them when training the model. All features with object types were one-hot encoded. This increased the total number of features to 221. A 70/30 train/test split was then conducted on the dataset.

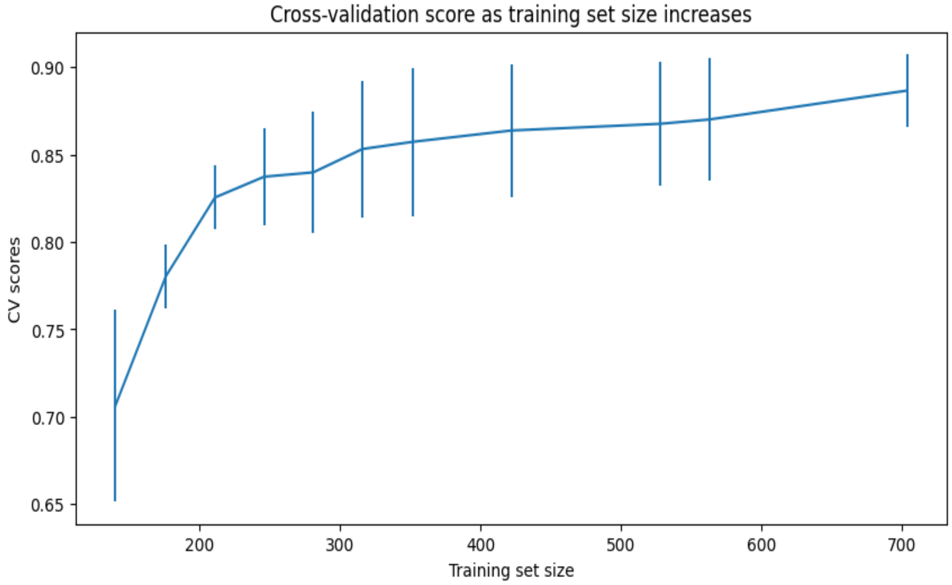
There were a handful of features that still had some missing values. These values were imputed with the median of the column since some of these columns were discrete and not continuous.

The data was then scaled around 0 using a standard scaler algorithm.

Modeling and Evaluation:

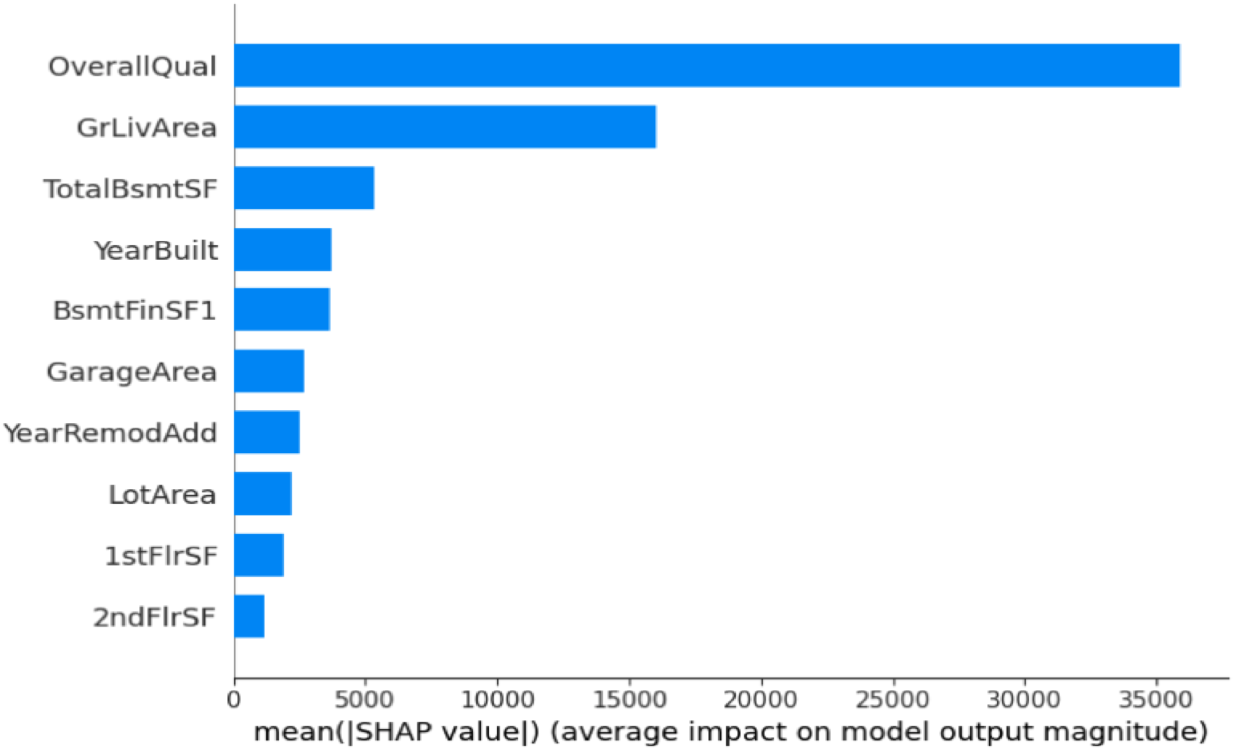
The grid search method with five-fold cross-validation and roc_auc scoring function was used on a random forest regressor model to find the optimal hyperparameters for the model. The best model was chosen with a max_depth of 15 and 1000 n_estimators.

The r-squared score on the train data was 0.986 while the r-squared score on the test data was 0.80. A five-fold cross validation function was run which resulted in an average score of 0.886 and standard deviation of 0.020.

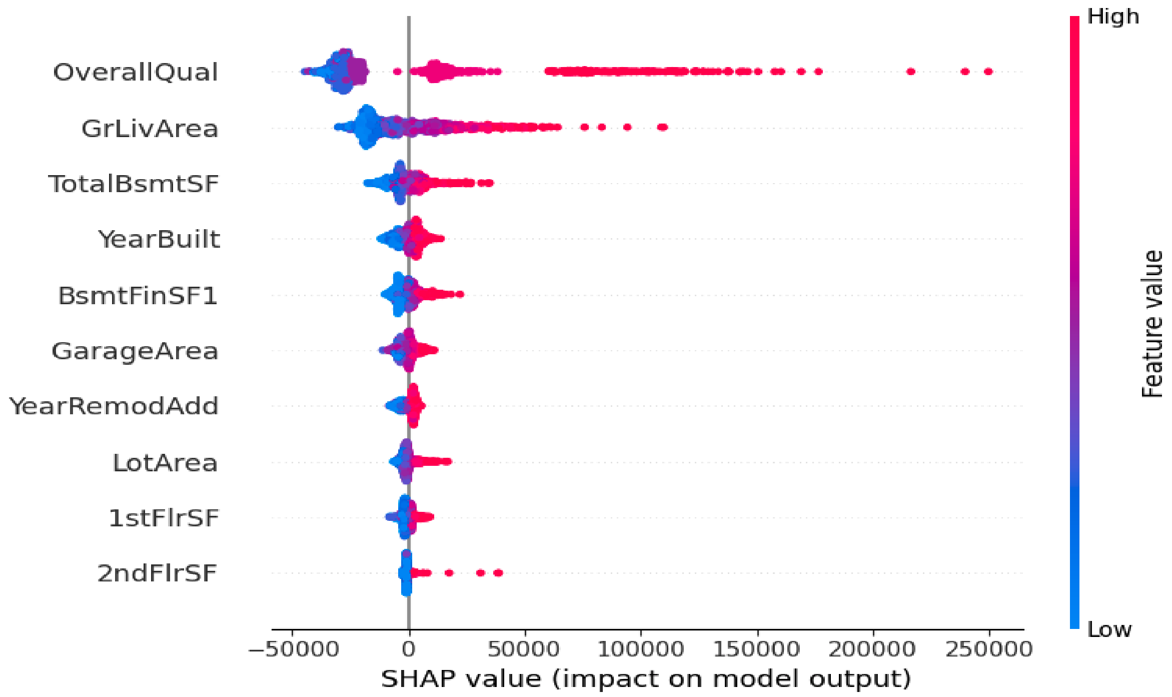


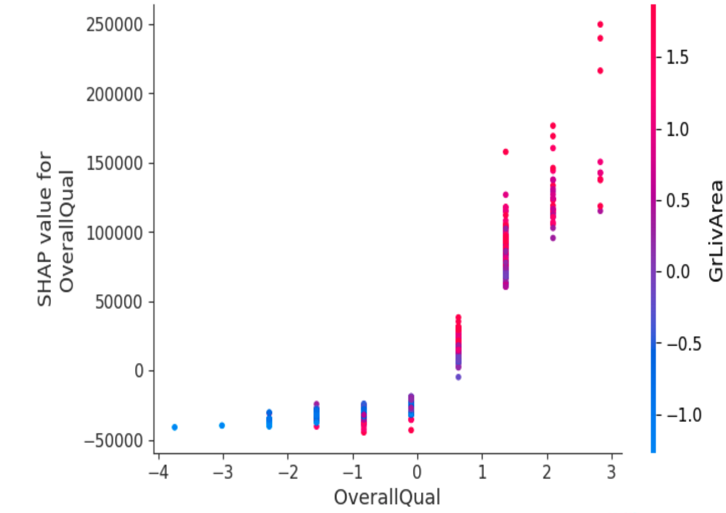
The mean absolute error was computed using five-fold cross validation. The average mean absolute error was 17,302.68 with a standard deviation of 746.15.

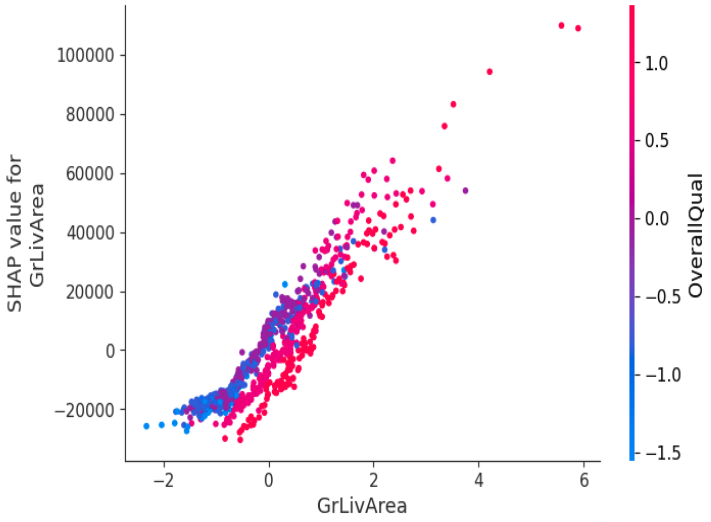
A learning curve was used to examine if acquiring more data would be beneficial. The graph indicates that plenty of data was used to train the model. The most influential features in accurately predicting the sale price of a house were the overall quality (OveralQual) of the house and the above ground living area (GrLivArea). The dependency

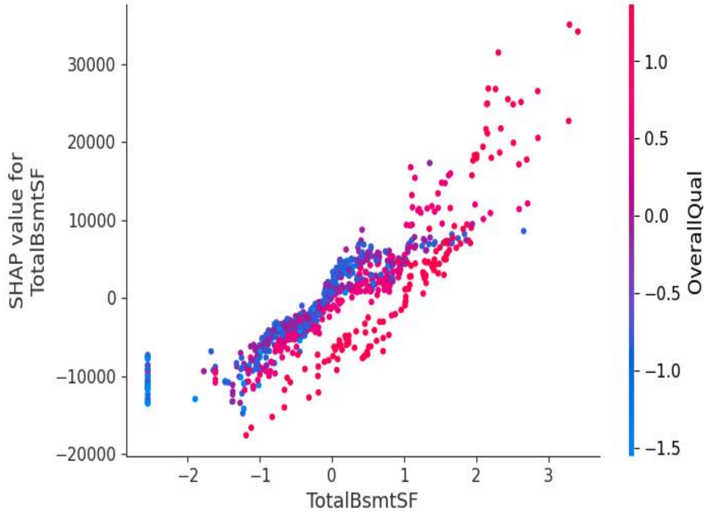


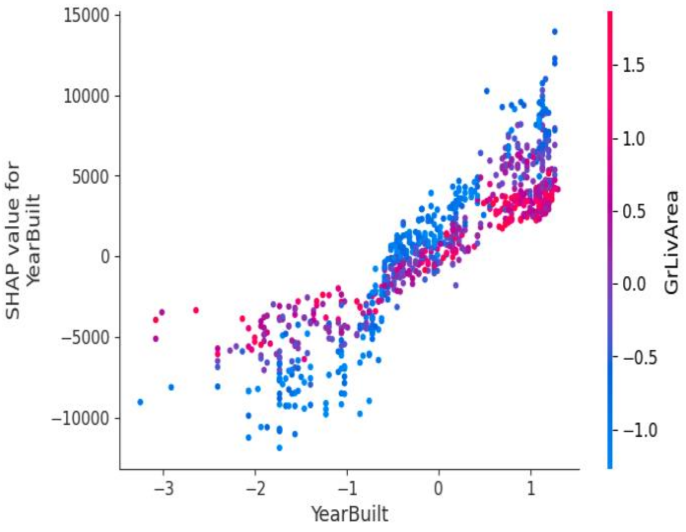
plots of the 10 most important features indicate strong relationships between the top ten most important features and the sale price.











Conclusion:

With an r-squared score of 0.80 on the test data, an average mean absolute error of 17,302.68 with standard deviation of 746.15, the random forest regressor model is accurate in predicting the house prices of houses with overall quality and above ground living area being features most strongly correlated with the house sale price. Total square feet of the basement and the year in which the house was built were also strong indicators of the sale price.