Final Report
Capstone – Boston Housing Prices

Problem Statement

Prices of house always fluctuate and sometimes housing prices are hard to predict. With the housing market constantly changing and fluctuating it would be beneficial for buyers and real estate agents give reasonable prediction of housing prices.

By using the Boston Ames Housing Data set, I used regression models to predict housing prices with the categories given. This can help agents and buyers understand prices in the Boston area.

The data used and processed was a preliminary analysis before a more detailed analysis was done.

Data Wrangling

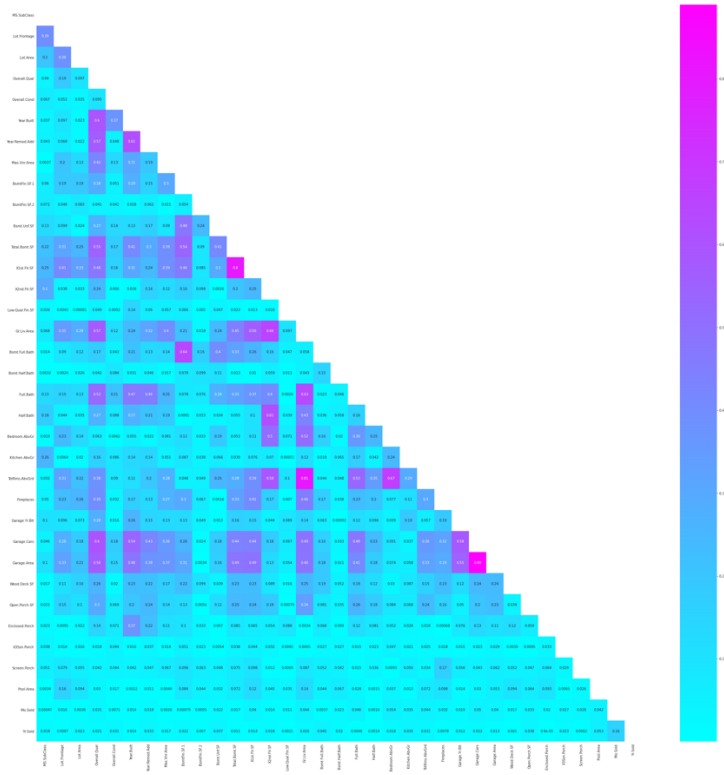
The raw dateset from the Boston dataset is 2,930 rows and 81 columns. The data set was usable and was not large. No reduction was necessary.

Exploratory Data Analysis

Since this was a preliminary analysis the exploration of data was used. First I took a look at the shape of the data. The data was skewed but the variance of the data was still within normal limits.

Missing values and null values were filled by determining the type of variable along with the number of missing values. Data that had Na or words that replaced 0 were filled with 0. Categorical data with null values were filled with the respective no value descriptor to that variable. Like Garage area's null values were filled with 0. Garage finish and garage condition was filled with NoG for no garage. Garage year built was filled with 0 which meant that there was never a garage built. If there ar over 90% missing values the variable was deleted for example Misc value, fence, pool quality, misc.features were all delted. Ally was included by adding a third type named noal which meant no alley.

The next I created a correlation matrix to compare the variables to Sale Price

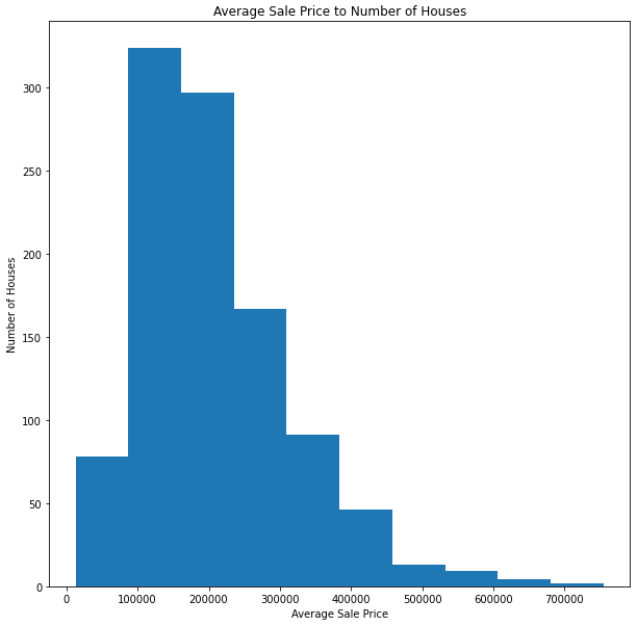


The Correlation matrix shows that there are 4 columns that are extremely not correlated less than .1. Once of which is month the house was sold. The month sold data had a normal distribution amongst the dataset but that only means that the most house sold was during the June months but with the preliminary findings it has

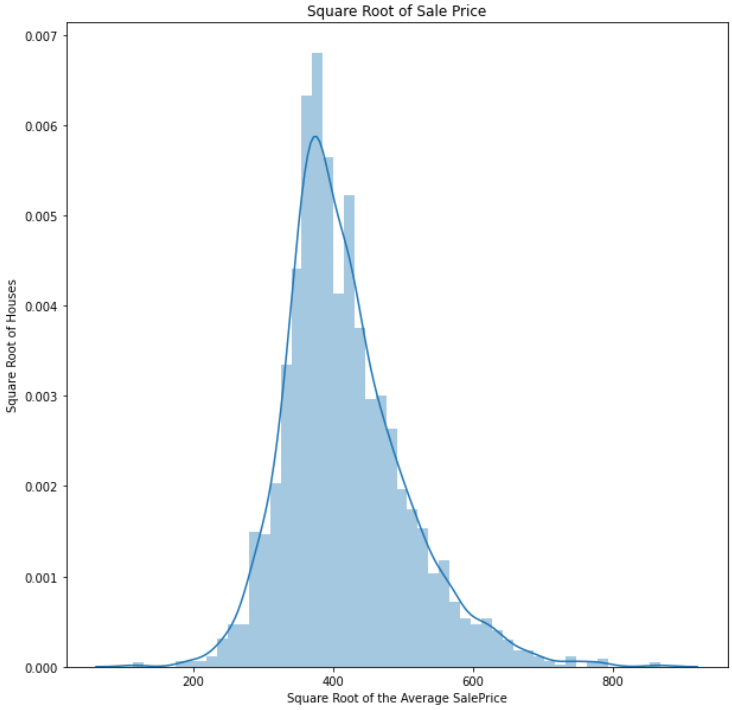
a .06 correlation to Sale Price. The other variables that had less than .1 correlation would be 3 season porch are, screen porch area, and pool area. Those variables were dropped.
The other variable that was dropped was Overall quality of house. It has a correlation of .96. This logically makes sense the higher the overall quality of the house the higher the sale price so this would be eliminated from our analysis.

Preprocessing

Sales Price was explored to determine the distribution of the data set. The distribution was a slightly skewed to the right. Where there are more data points in the lower end of the sale price.



In order to normalize the dataset, taking the square root of Sales Price created a better distribution.



Once Sale Price was adjusted. Data types in the data were adjusted from Object type to category type. This is necessary for when creating dummie variables for the categorical data points. Once the data types were correct, categorical variables were transformed into dummie variables. The data set was split 75% training set and 25% testing set.

Modeling

Seven different models were used to predict housing prices. Since the data was not linear the outcome for Linear Regression was bad. Other models used were Ridge, Gradient Boosting, Random Forest, Lasso, and ElasticNet Regressor. All presented with a R^2 score of .90 and above. Lasso had the best out come with a mean squared error (mse) of 595.89 and mean absolute error of 16.00. Both metrics of root mean square and mean absolute error is shown but since there were a lot of outliers in sales price and within our data point a better representation of the analysis would be using Mean Absolute error. Since the data was normalize and adjusted we would have to take the square of the data to get real life predictions. Which changes our results in having an error of about $354,025.00 with a standard deviation of $256.00.

	Explained Variance (r^2 score)	mean^2 error	root mean^2 error	mean absolute error
Linear Regression	-1.14E+21	8.36E+24	2.8921E+12	1.1149E+11
Ridge	0.917	605.997	24.617	16.968
Gradient Boosting Regressor	0.909	664.556	25.779	17.445
Random Forest Regressor	0.901	727.977	26.981	18.304
Lasso(alpha=95)	0.919	595.885	24.411	16.791
ElasticNet(ratio=.95)	0.917	605.997	24.617	16.968

Conclusion and Future Analysis

The data gives good overview of predicted prices in Boston. Minimizing error would be the next step in the process. More preprocessing is necessary. Possibly using gradient boosting to determine which features are best to predict Sale Price then using the top 10 features to predict the prices may have a better outcome. Another would be feature engineering where sale price is a range which lowers the variance which can decrease the mean square error as well as the mean absolute error. Also, by feature engineering Sale Price it could fix the skewness of the data so less data manipulation is needed to normalize the data.