Housing Price Predictions - Ames, IA

Johannes Potgieter, Jose Cuan, Gerald Bourdeau

Introduction

In modern times prospective buyers of properties are able to view a large selection of property stock online with good quality pictures, virtual tours and complete descriptions. However, this has not eliminated the need for investors or home buyers to physically view properties before putting in an offer to purchase.

The aim of our analysis is to provide these stakeholders with model to predict homstments.e sale prices with high accuracy using various machine learning algorithms. The prediction of the sales price is enriched by the highlighting of key features that influence home sale prices most directly. Investors, speculators and home buyers can therefore make a more informed decision as to the types of homes and features they should focus on to maximise either short-term profits or lock in long-term value on their investments.

Background Information:

Our analysis is focused on a housing data set from the city of Ames in Iowa.

Ames is a city in Story County, Iowa, United States, located approximately 30 miles (48 km) north of Des Moines in central Iowa. It is best known as the home of Iowa State University (ISU), with leading agriculture, design, engineering, and veterinary medicine colleges. A United States Department of Energy national laboratory, Ames Laboratory, is located on the ISU campus.

In 2020, Ames had a population of 66,427, making it the state's ninth largest city in the state. Iowa State University was home to 33,391 students as of fall 2019, which make up approximately one half of the city's population.

Data Understanding and EDA

- The data set provided represents as sample of approximately 2600 houses in the city with about 81 features. The features are broken down into 20 Continuous, 15 Discrete, 23 Categorical Nominal and 23 Categorical Ordinal features.

- The target variable of our analysis is the Sales Price.

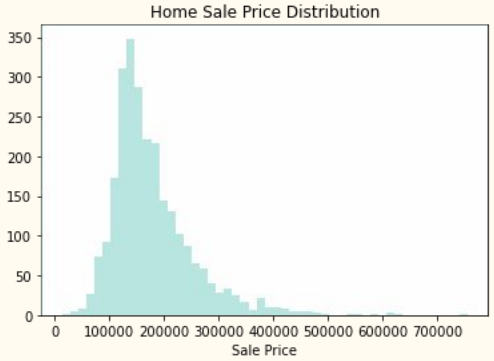
- We explore the various techniques available in ML to discern the best approach and decide on a model which provides the best prediction on house price

- Below follows an Exploratory Data and Visualization Analysis to gain insight into the data with regards to useful features, limitations, outliers and missing values. As data scientists the understanding of data and mutation to enhance the data set is key to an accurate and valuable analysis which will be useful for our stakeholders.

Exploratory Data Analysis

- Visualisation to understand the data set.
- Identification and rectification of potential important data with missing values.
- Identification and handling of outliers
- Feature Engineering and Feature Selection

Normalization of Target Data



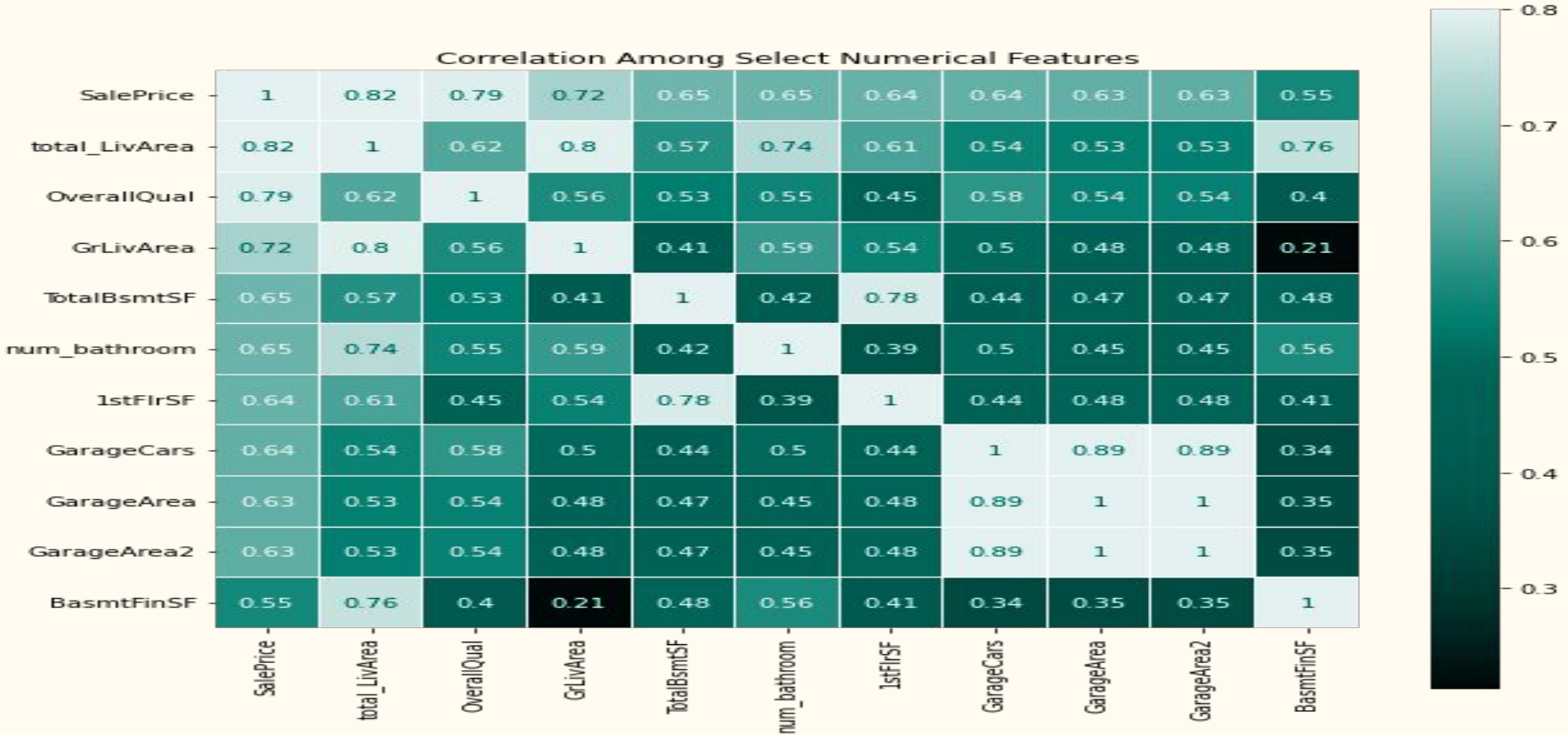
Distribution of Sale Price

Log Transform of Sale Price

Log Transform w/o Outliers

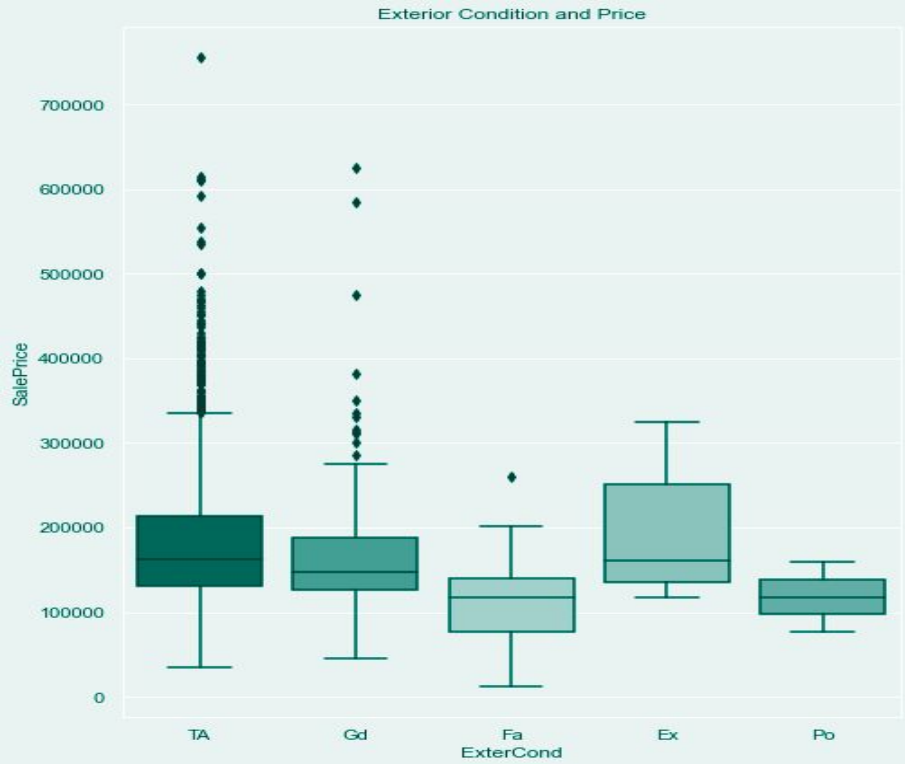
- After Logarithmic transform of target data, outliers were removed.

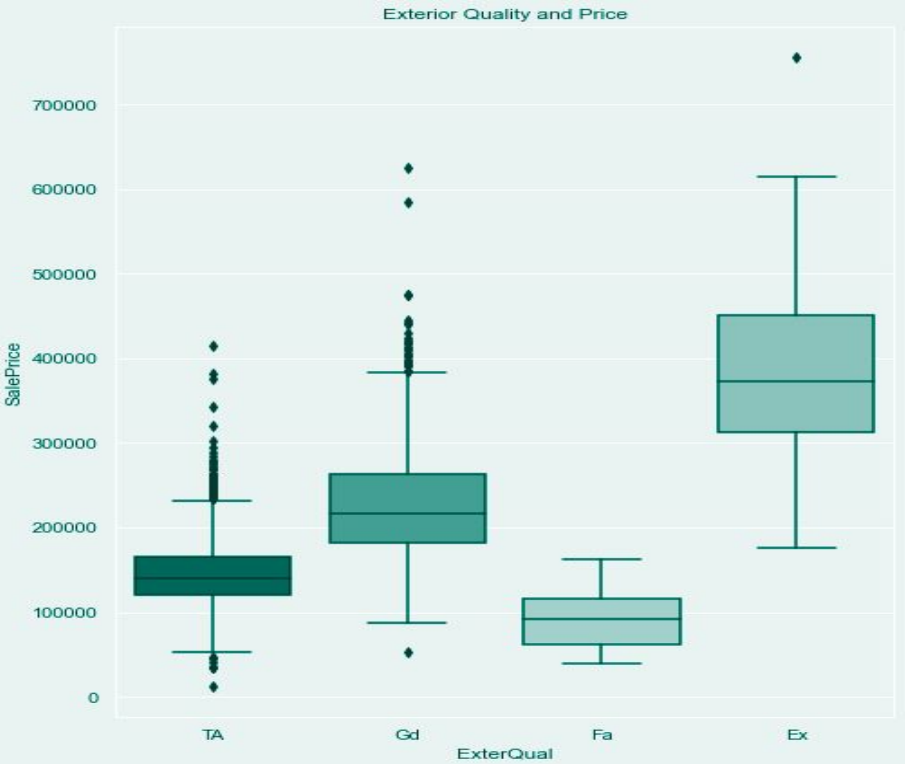
Correlation Matrix:



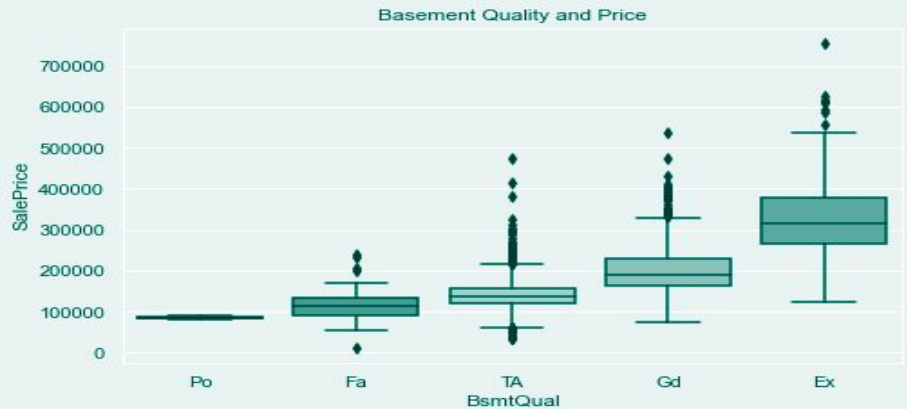
Total Living Area vs Sales Price. Outliers are evident!

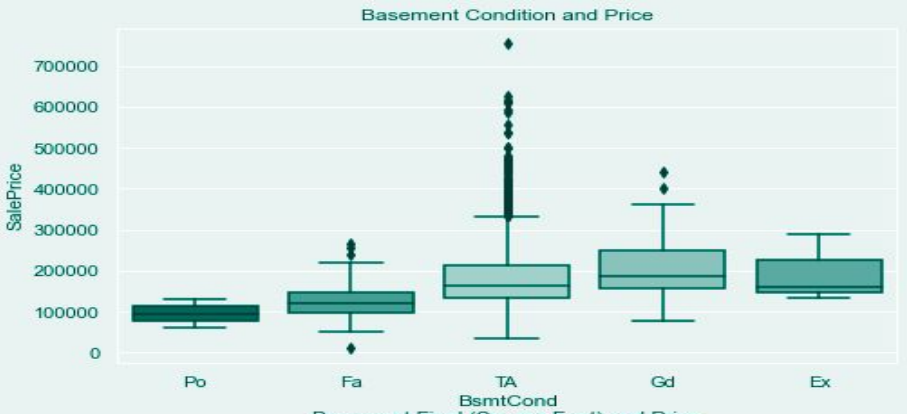
Exterior condition/quality vs Sales Price

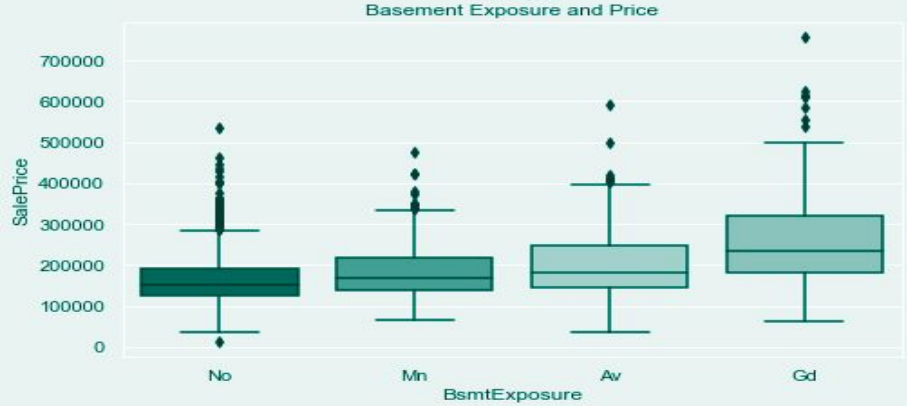


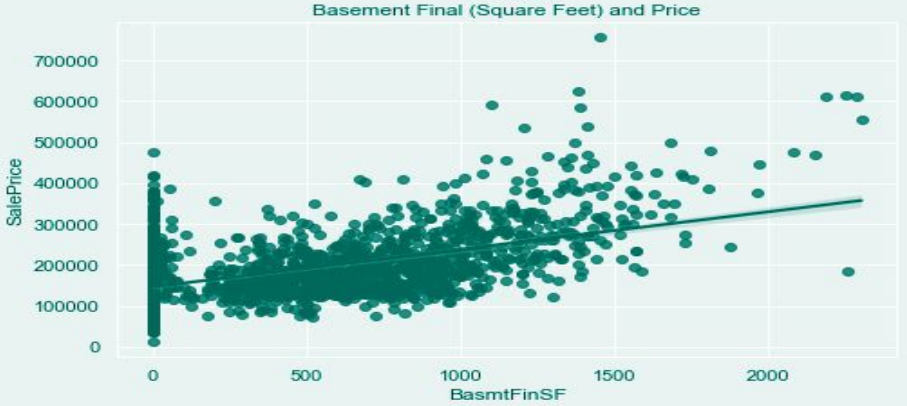


Basement Features vs Sales Price

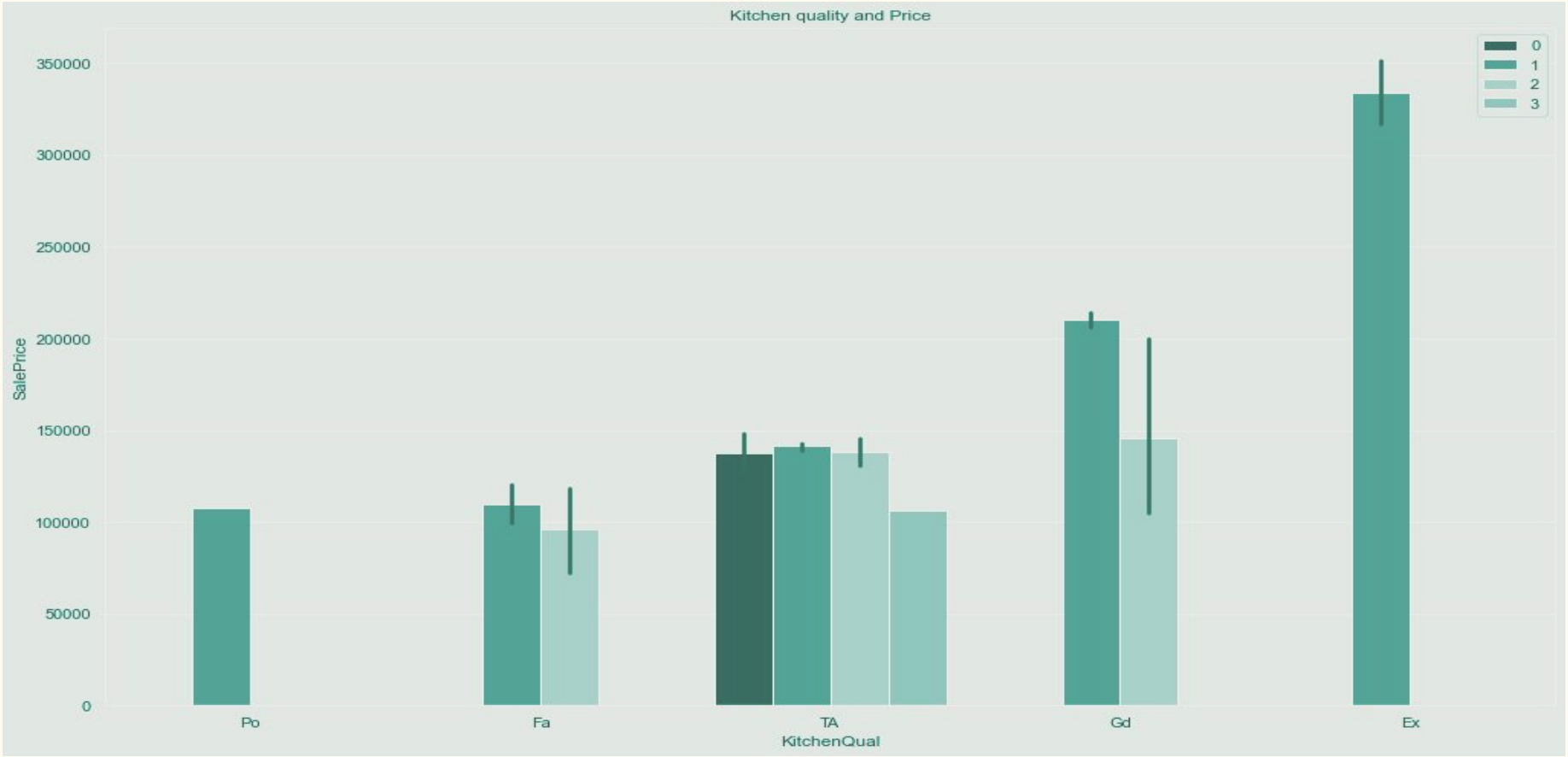






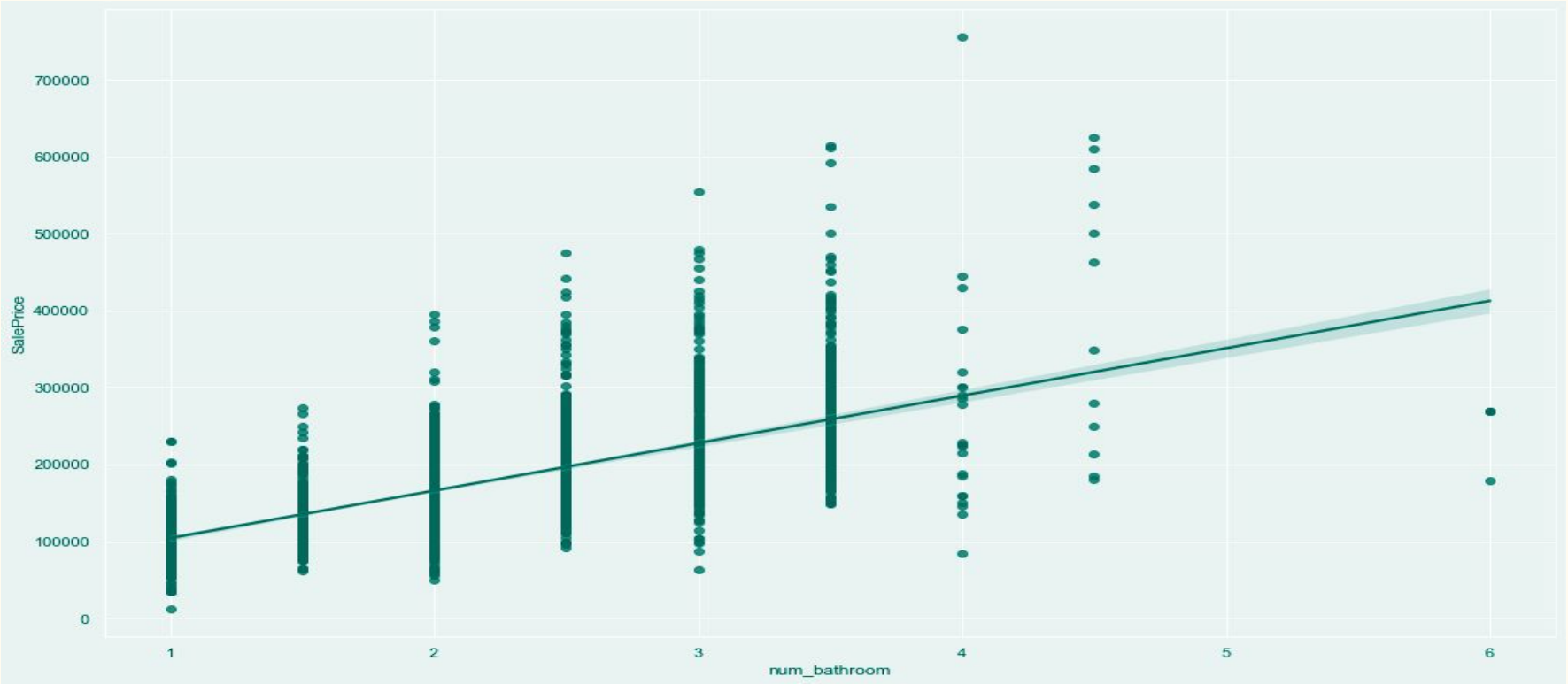


Kitchen Quality vs Sales Price

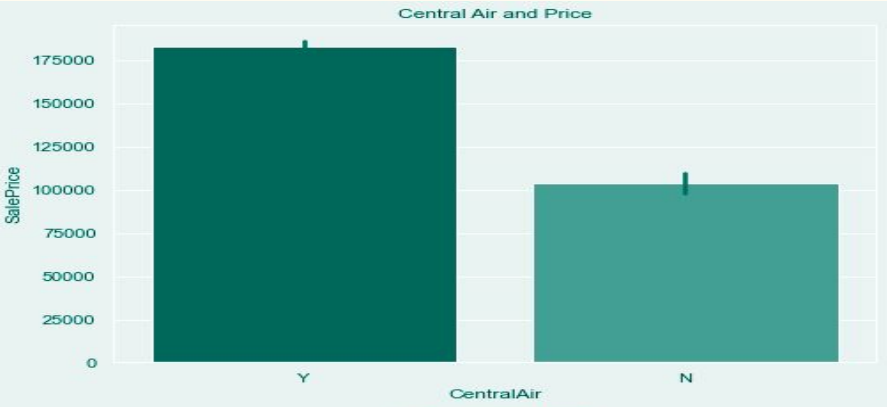


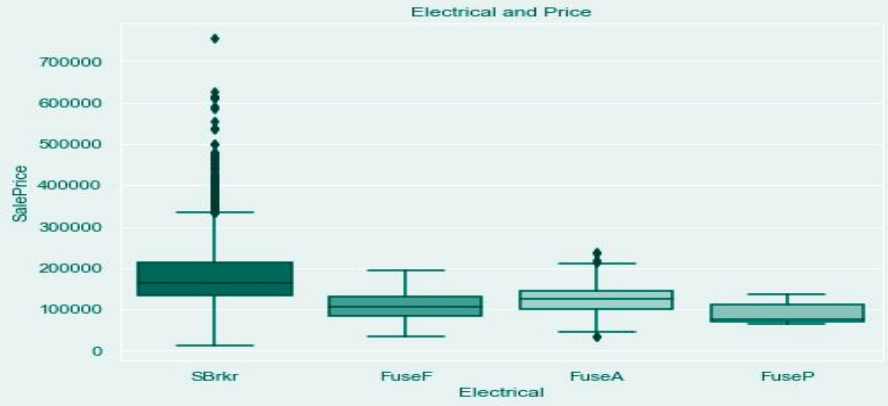
Garage Type vs Sales Price

Number of bathrooms vs Sales Price

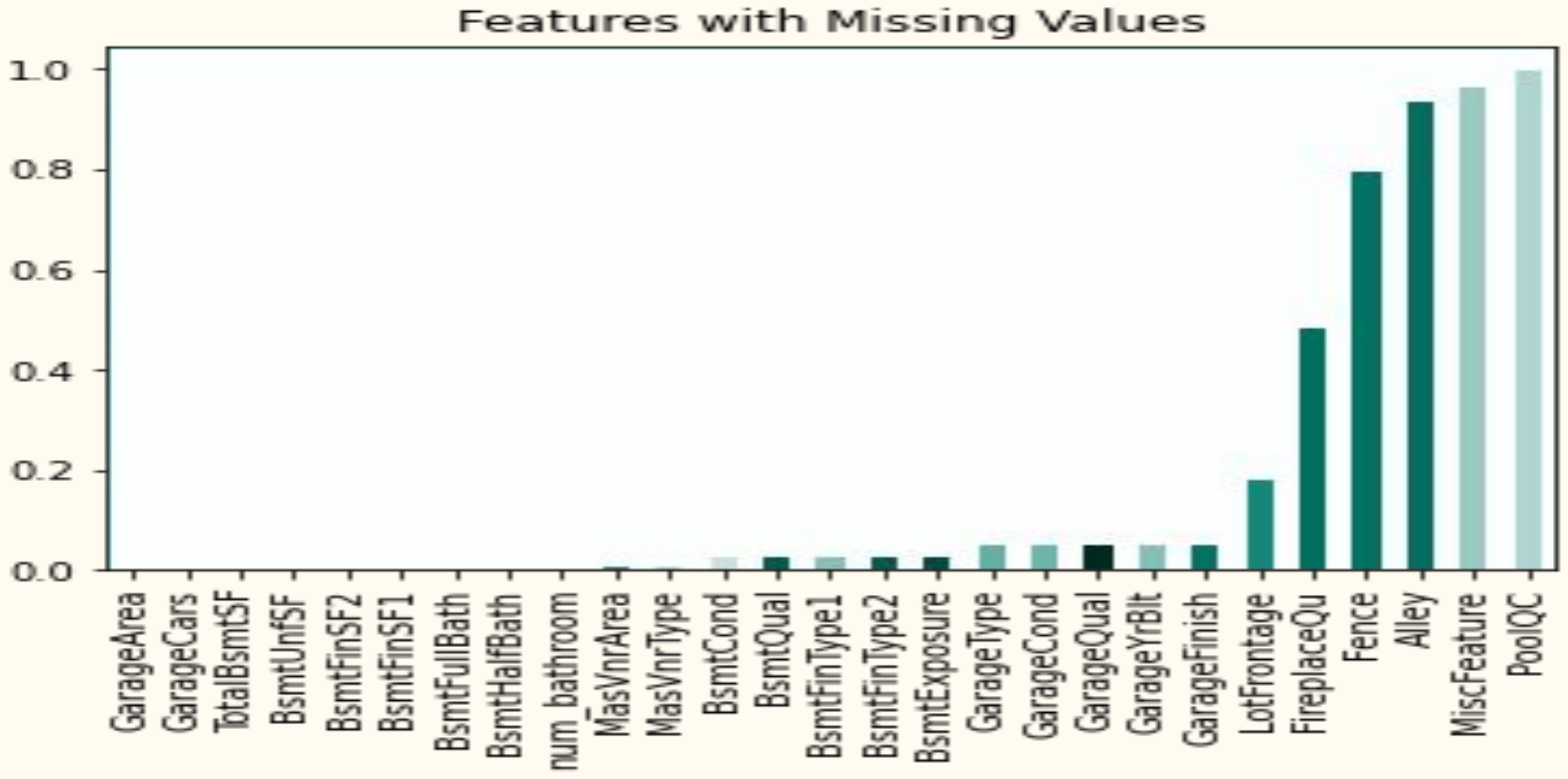


Heating Features vs Sales Price



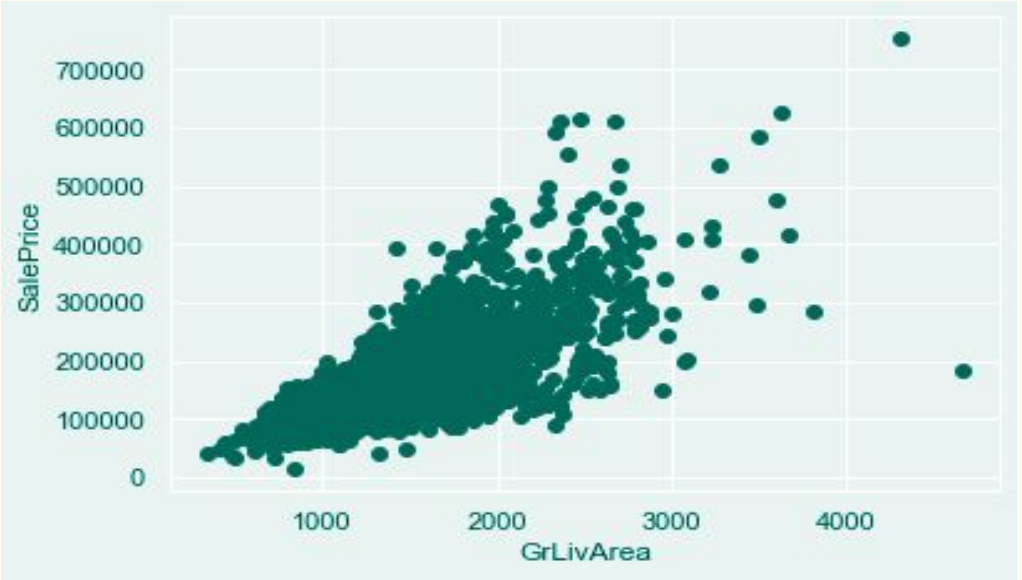


Features with missing values:

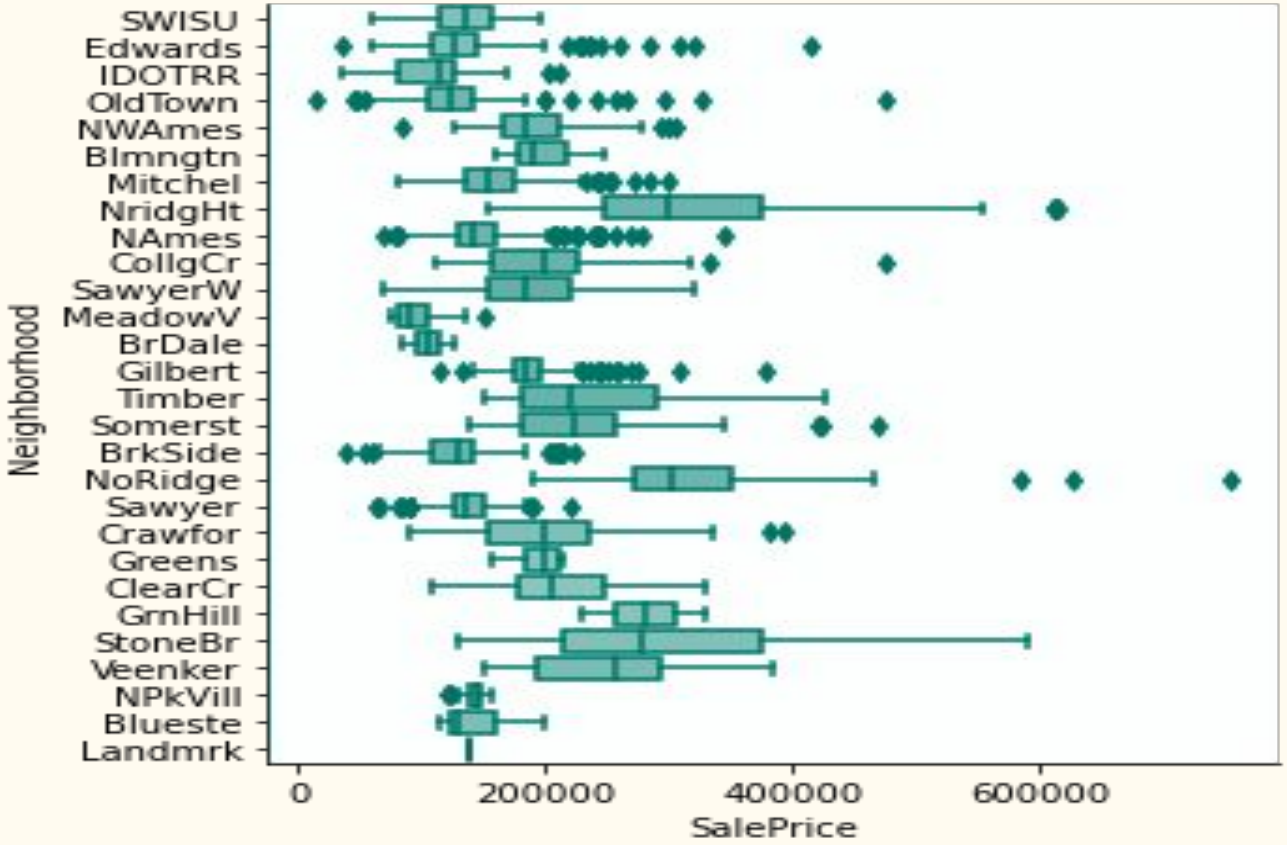


Outliers: Isolating the values that may skew results

- Outliers such as record with GrLivArea $>$ 4000 were removed from the data set.
- In the log transform of SalePrice,
  the values $<$ 4.67 were removed
  to improve normalization



Neighborhood vs Sale Price: Easy to identify Outliers



Feature engineering

- With numerous features available to us, many were replaced after certain feature engineering measures were implemented.

- For example YearRemodAdd and YearBuilt were replaced by the binary feature of Remodeled. Also BldgAge was created from (YrSold - YearBuilt) and EffAge was created from (YrSold - YrRemod).

- Fence category was binarized to the HasFence feature.

Feature Selection:

- Attempted a merge with the Ames_Real_Estate file and used the Zip code as an added feature. This resulted in a lower score for both train and test sets.
- The feature set was adjusted based on the model to maximize the R^2 value for both train and test data. Each model contains a different subset of the original feature list.
- Categorical data was dummified for the Linear Regression models. For for the tree based model, the OneHotEncoder method was applied.
- This resulted in a data set with 146 features.

Predictive Analysis:

- Model Selection (refer to Annexure)
- Final Score Comparison
- Recommendations and Conclusions
- Future work and analysis

Model Result Summary

Model	SLR	MLR (Ridge)	Decision Tree post HPT and Gridsearch	Random Forest	Gradient Boosting	Lasso post Normalisation
Train R^2 %	91.88	93.80	91.39	98.28	94.12	91.10
Test R^2 %	91.66	93.40	82.40	87.59	91.30	92.10

Conclusion and Recommendations

- Linear models and the gradient boosting model offer the highest level of accuracy in predicting Ames home sale prices.
- The decision tree and random forest models suffer from overfitting and little improvement is noticed after enhancements. The models are also computationally expensive.
- From our analysis it is our conclusion that investors and home buyers alike should look at the overall quality (informed by high quality features and finishes) in their decision to purchase a house. Focusing on improving these features will likely result in higher profits when flipping houses as well as sustainable value for long-term buyers.

Future work and analysis

- Apply unsupervised methods like PCA and Cluster Analysis to enhance feature selection and improve on the accuracy of our predictive models.

- We would like to enrich the data set with more up to date information and re-perform our analysis.

- Given more data around population growth, demand for student accommodation and rental yields, we could assess whether there is value in converting lower end properties into student accommodation.

- Productionalize the results of our model into an application (using for example Tableau) which will allow investors and potential buyers to access recommendations in an interactive way. For example selecting a house, enhancing quality features, afterwich the model will project an expected sales price.

Annexure: Model Selection

- Simple Linear Regression (OLS)
- Multiple Linear Regression (Ridge)
- Lasso Regression
- Decision Tree
- Random Forest
- Gradient Boosting

Simple Linear Regression (OLS)

Num of features: 146

AIC: -2946.635794574734

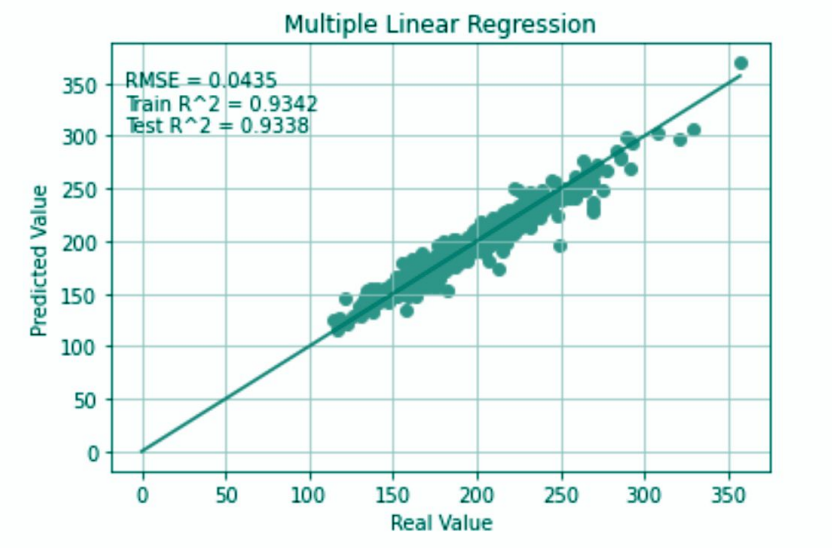
BIC: -2158.1674827447587

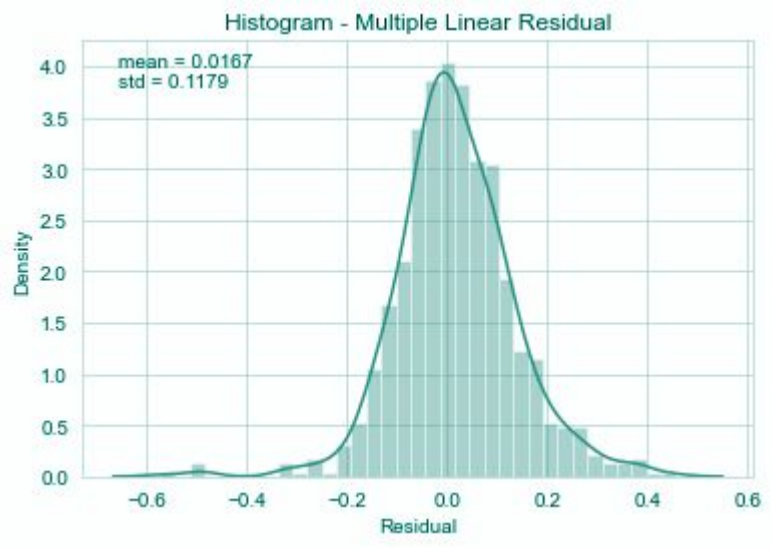
train score: 0.9188320945175004

test score: 0.9166251216071358

Multiple Linear Regression

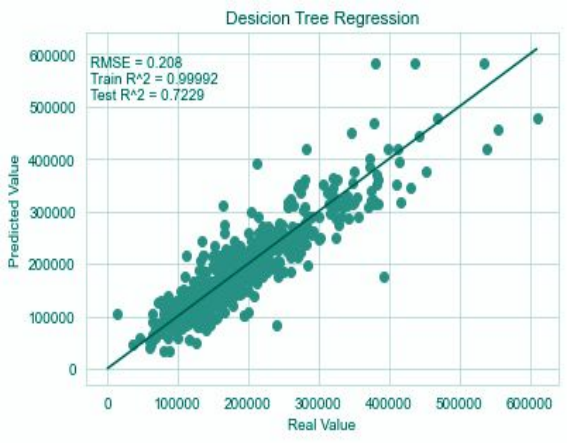
Multiple linear regression refers to the statistical technique that is used to predict the outcome of a variable based on the value of two or more variables.

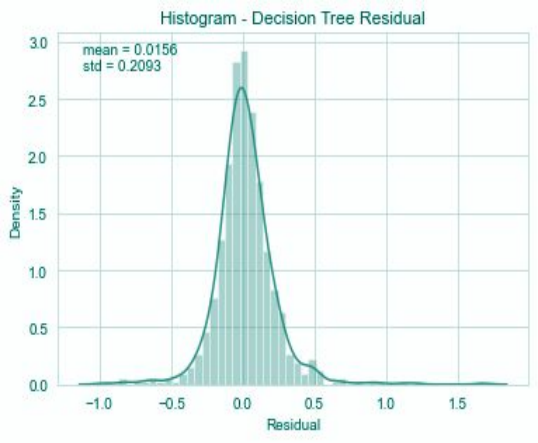


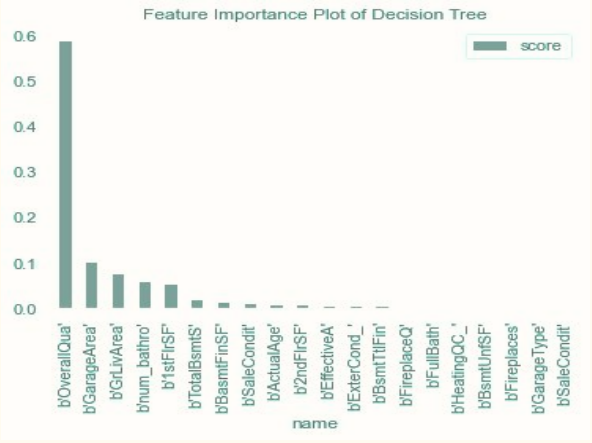


Decision Tree

Decision Tree is a supervised learning technique for regression and classification prediction model. It provides a simple and clear guide for decision making progress. However, this simplicity comes with some disadvantage: overfitting, Bias error and Variance error.

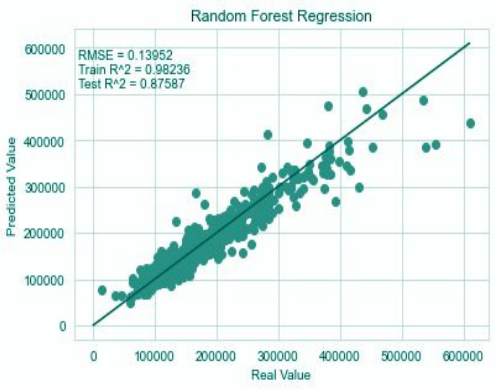


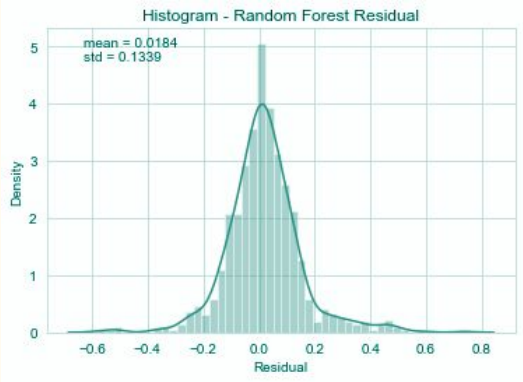


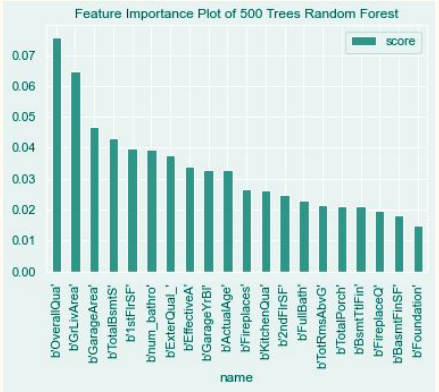


Random Forest

Random Forest is a ML algorithm that use the combination of multiples random decision trees each trained on a subset of data. The use of multiples trees gives stability to the algorithm and reduce variance.







Gradient Boosting

ML algorithm builds one tree at the time and updates the sequence of weak learners into a strong learners. Gradient Boosting combine results during the whole process of building trees.