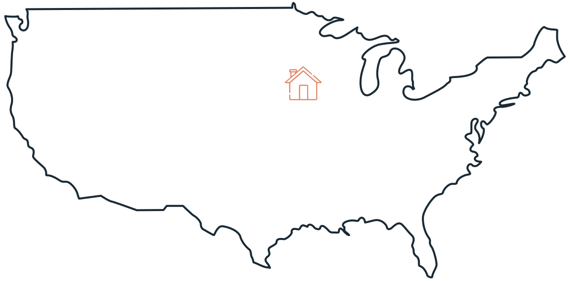
Predicting & Maximizing Home Value

NYCDSA Machine Learning Project

by Lucas Kim, Ryan Park, Sita Thomas

September 2020

Project Overview



Audience

- Fictional Data Mining Company

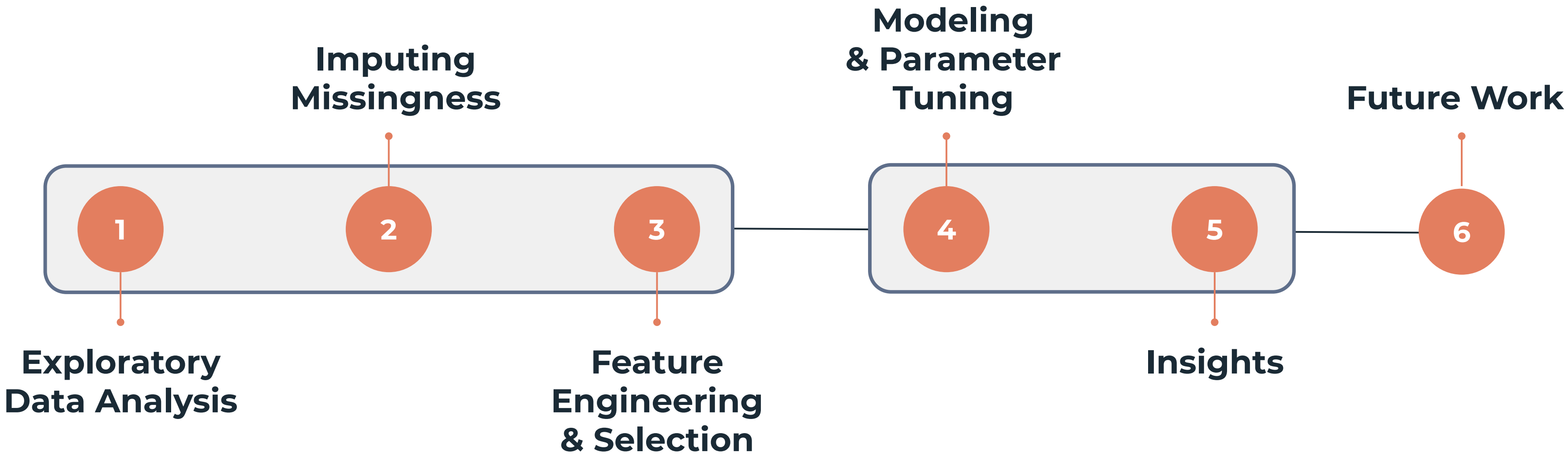
Dataset

- 1460 houses in Ames, Iowa
- 80 features

Objective

- Predict home sale prices
- Describe feature relationships

Workflow

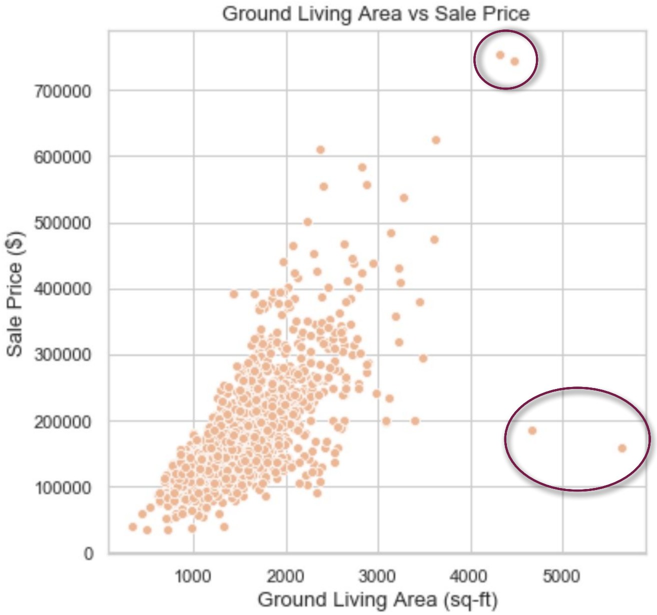

Imputing Missingness
Modeling & Parameter Tuning
Future Work
1
2
3
4
5
6
Exploratory Data Analysis
Feature Engineering & Selection
Insights

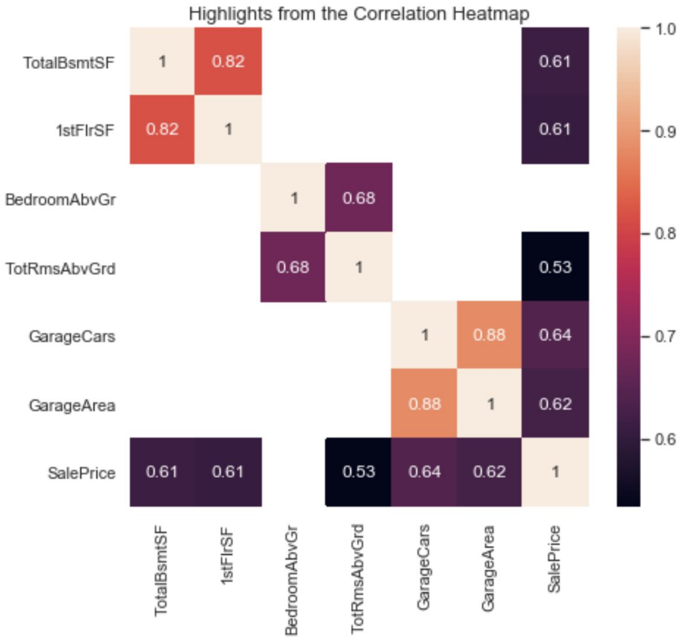
01.

Pre-Processing

Exploratory Data Analysis, Imputing Missingness, and Feature Engineering & Selection

Exploratory Data Analysis



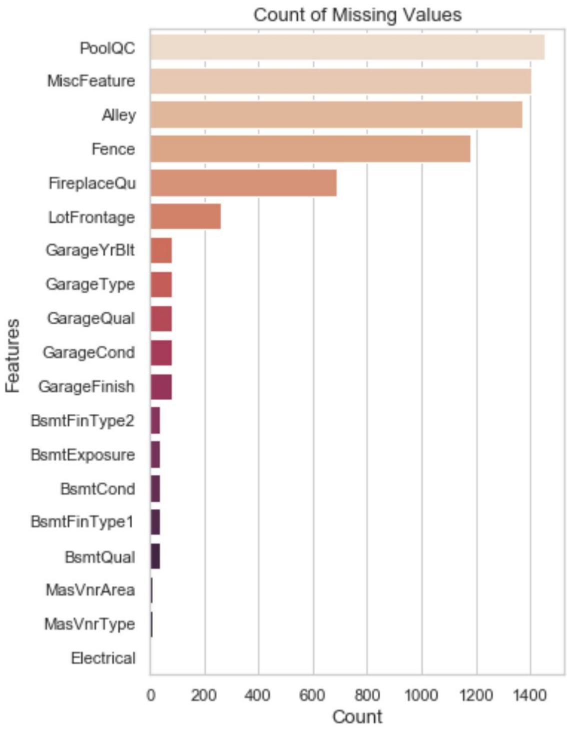


Imputing Missingness

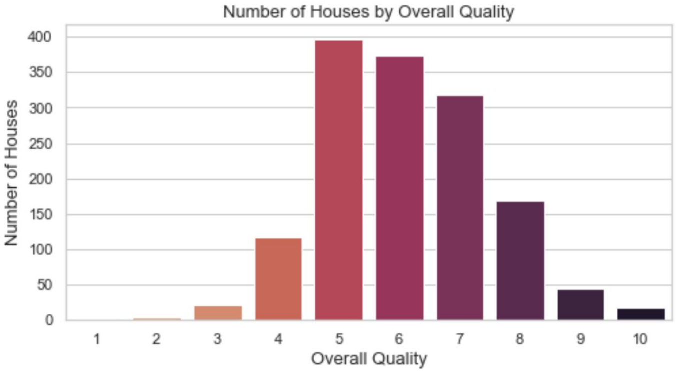
5.96% missingness

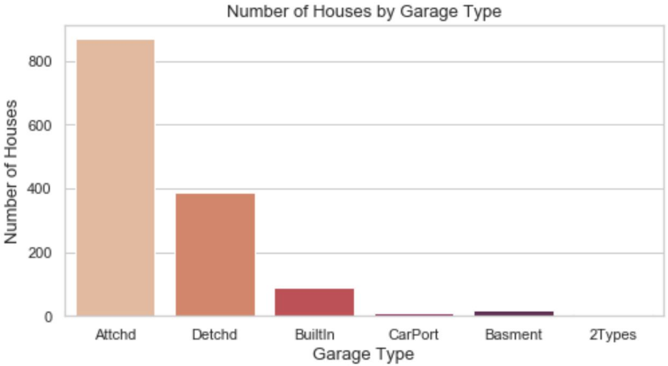
- ✓ **Missing Not at Random**
  Garage, Bsmt, PoolQC, etc.

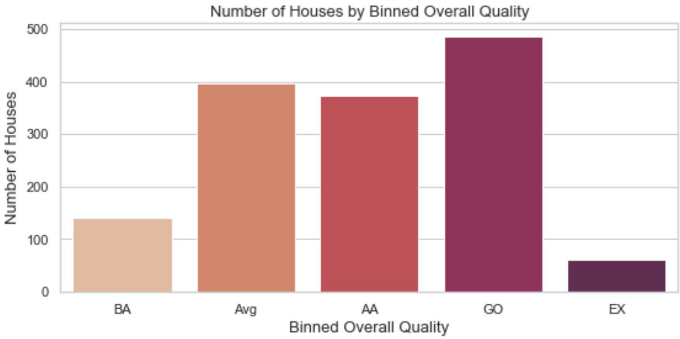
- ✓ **Missing at Random or Missing Completely At Random**
  LotFrontage, MiscFeature, etc.

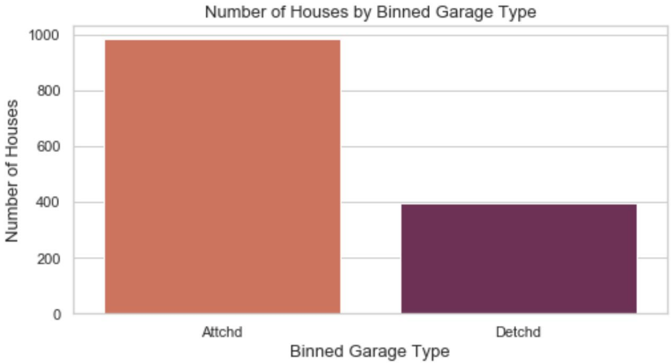


Feature Transformation

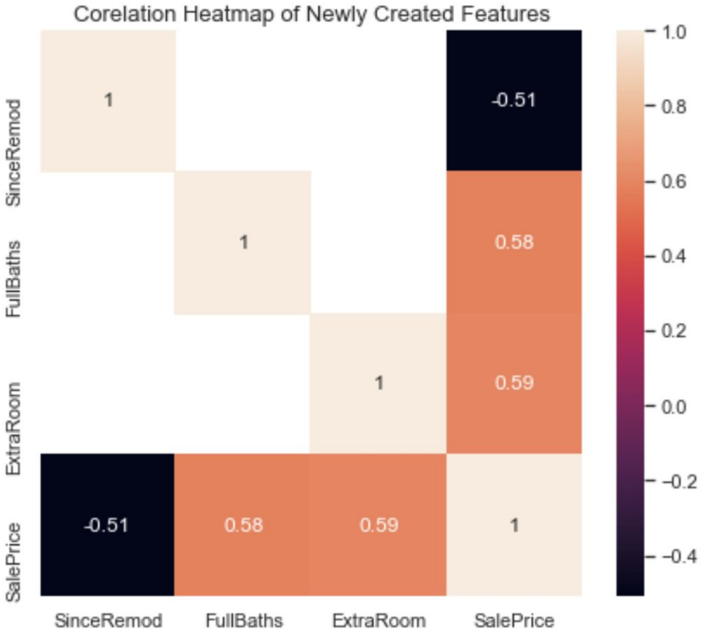








Feature Engineering



✓ **ExtraRoom**

Total Rooms - Bedrooms

✓ **FullBaths**

Above Ground Bathrooms + Basement Bathrooms

✓ **SinceRemod**

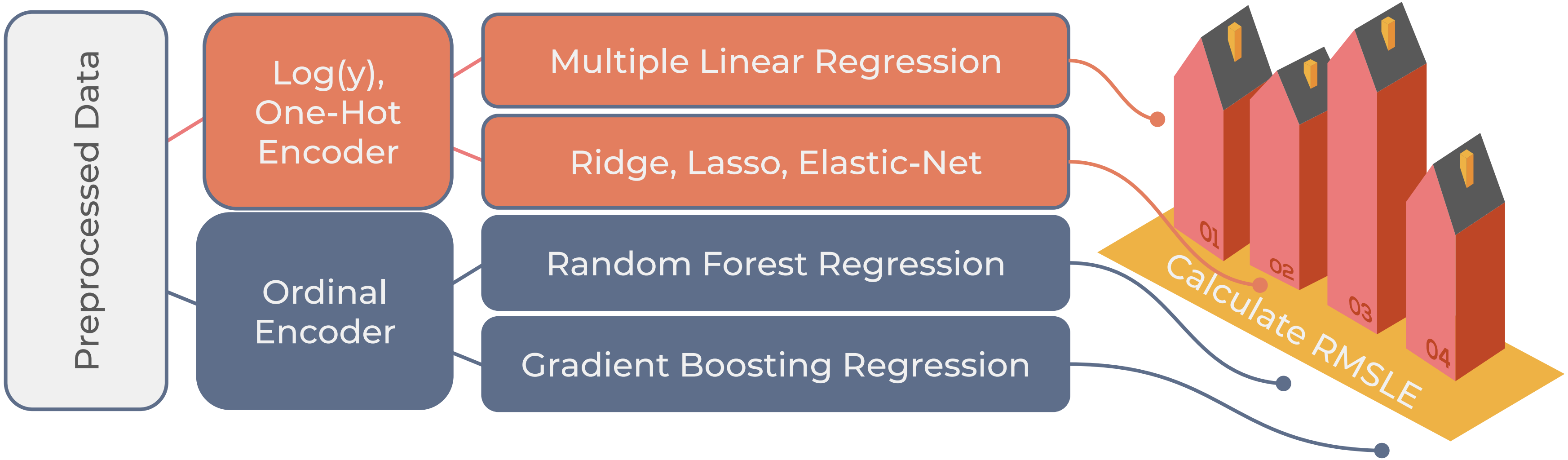
Year Sold - Year of Remodel

02.

Modeling

Regression
and Tree Based Models

Modeling Pipelines



Preprocessed Data

Log(y), One-Hot Encoder

Ordinal Encoder

Multiple Linear Regression

Ridge, Lasso, Elastic-Net

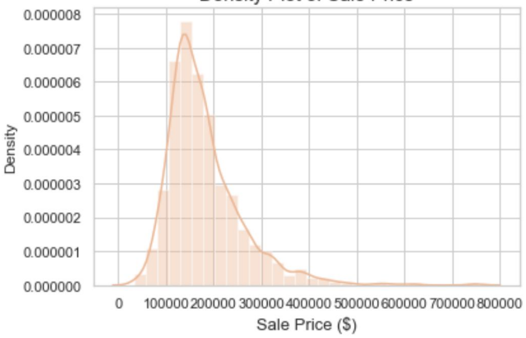
Random Forest Regression

Gradient Boosting Regression

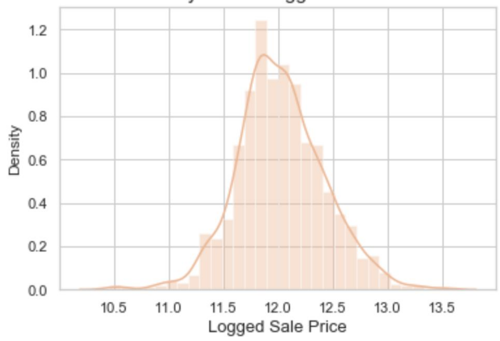
Calculate RMSLE

Log Transformation of Sale Price

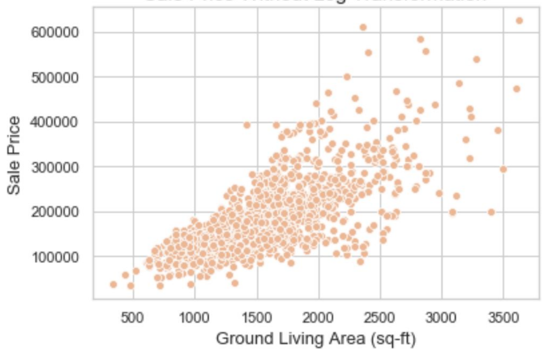
Density Plot of Sale Price



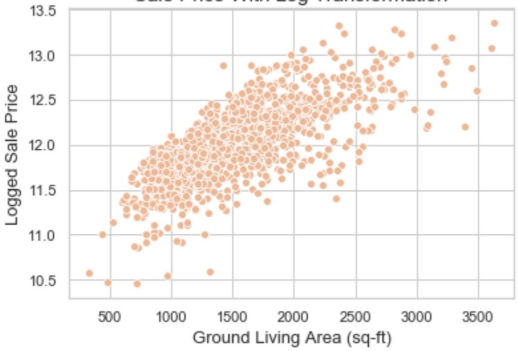
Density Plot of Logged Sale Price



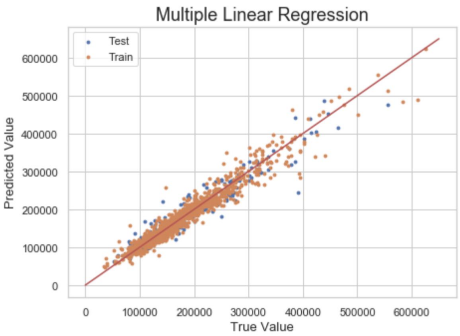
Sale Price Without Log Transformation

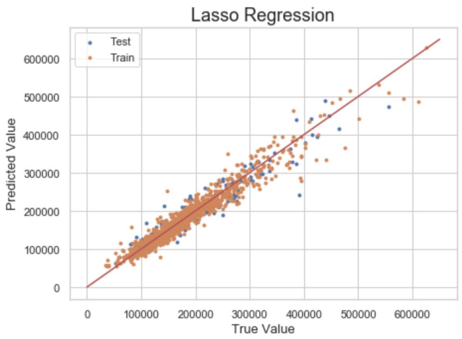


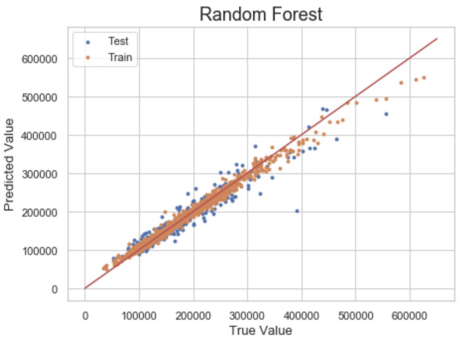
Sale Price With Log Transformation

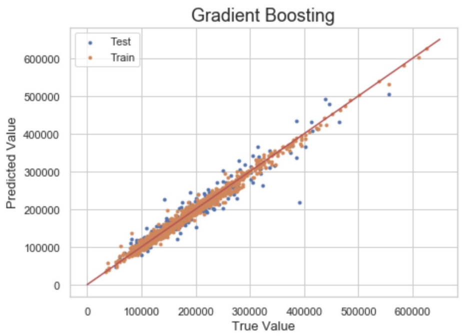


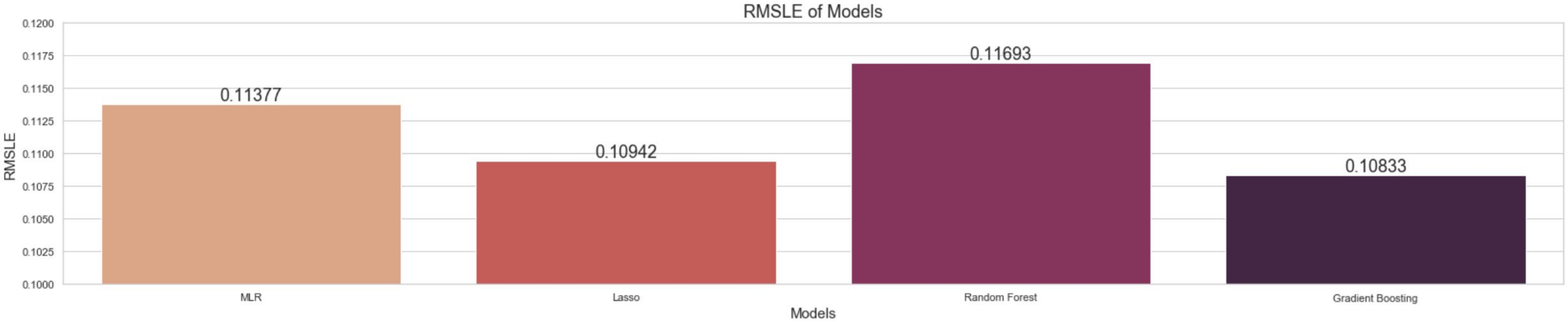
Model Performance


Multiple Linear Regression

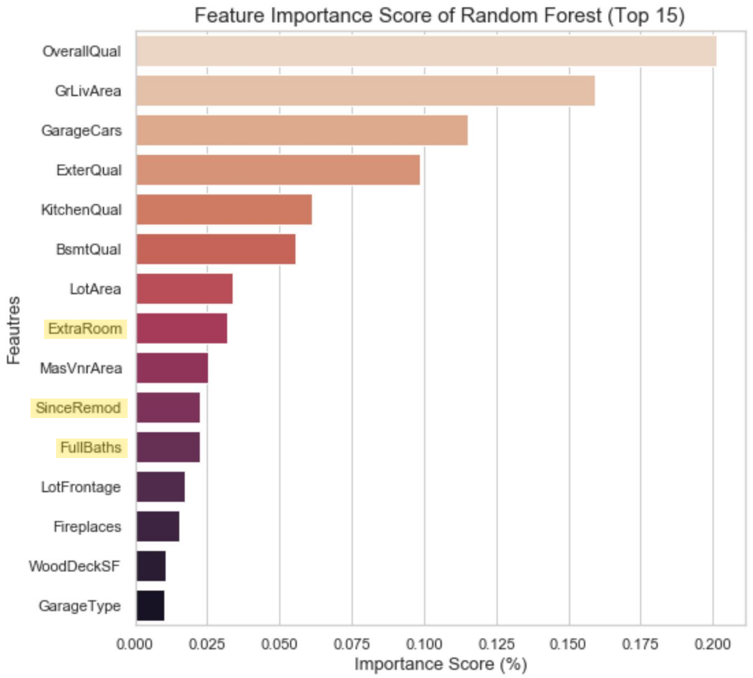

Lasso Regression

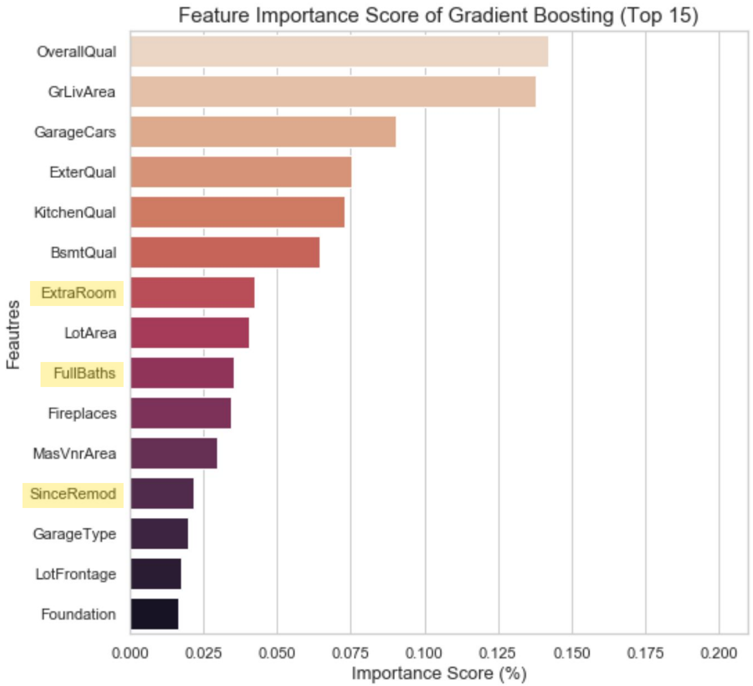

Random Forest


Gradient Boosting


RMSLE of Models

Feature Importances



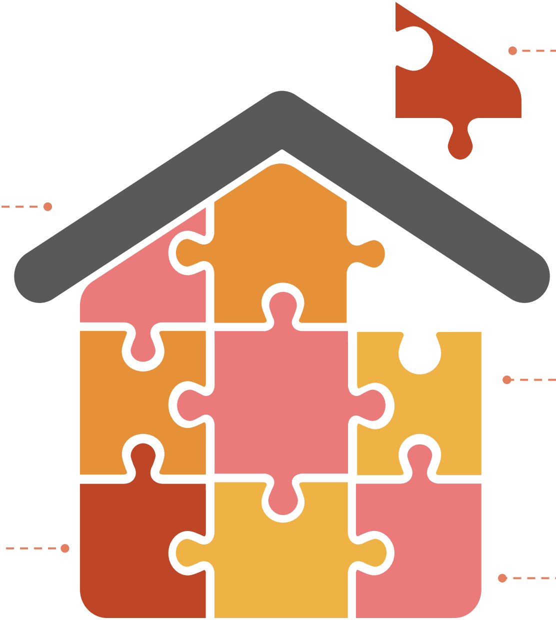


03.

Insights

Quality and Additions

Remodeling Insights



Overall Quality

Most added value but best when sold soon after updates.

Above Grade Living Area

Long-term value but a significant investment.

Garage Size by Car Capacity

Good return on investment for homes with no garage or carport.

Exterior Quality

May improve both home safety and utility costs.

Kitchen and Basement Quality

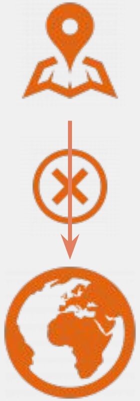
Relatively easy and inexpensive to update, but easily outdated.

04.

Future Work

Limitations

Not Enough Sample Points



Limited External Validity:
Hard to generalize

Yet, valuable insights for home buyers and sellers, to real estate companies, materials manufacturers, home goods retailers

Dimension Reduction

Automate feature engineering and selection

Outliers

Identify and remove outliers automatically

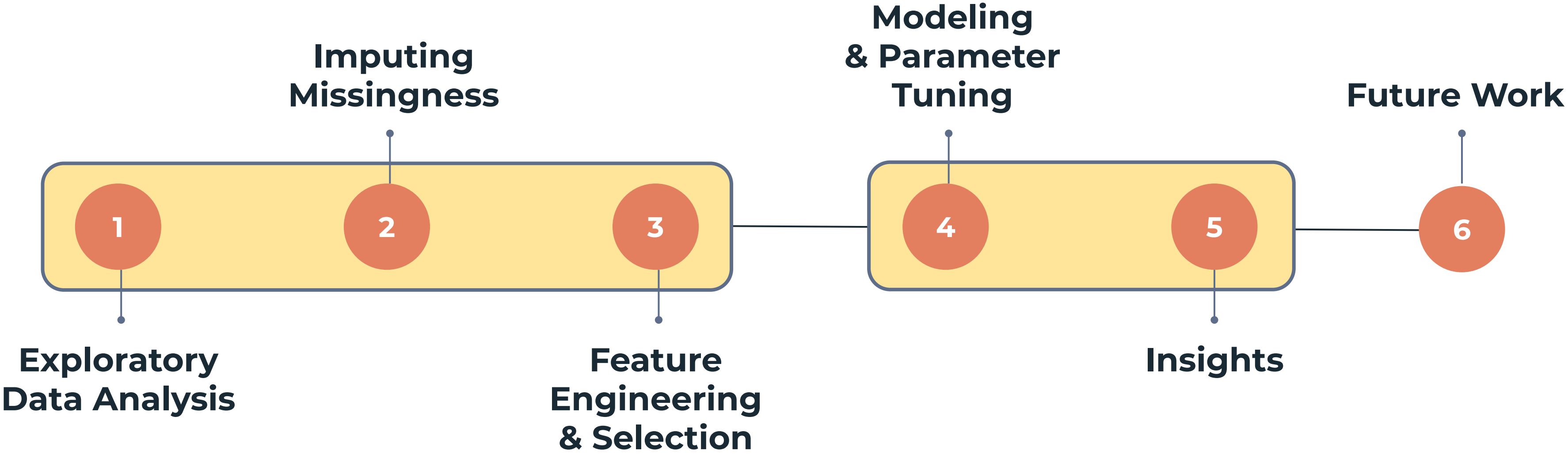
XGBoost, etc.

Try more advanced models to improve predictions

Stacking

Combine multiple models

Summary


Imputing Missingness
Modeling & Parameter Tuning
Future Work
1
2
3
4
5
6
Exploratory Data Analysis
Feature Engineering & Selection
Insights

Thank you!