Machine Learning Project:

Ames, Iowa Housing Prices

Ariani Herrera, Erin Dugan, John Nie, Won Kang
NYC Data Science Academy - August 27, 2018

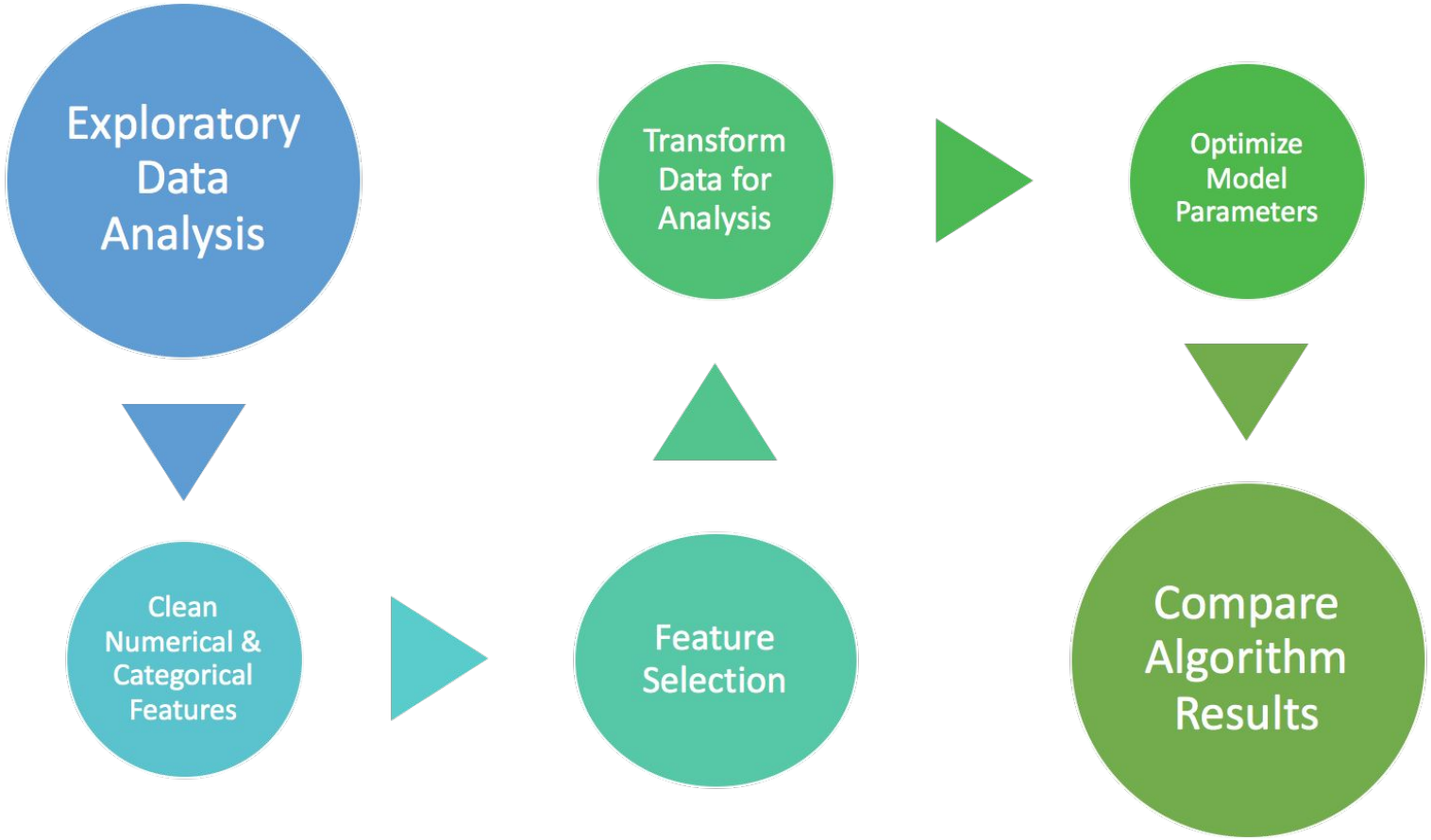
Agenda

1. Project Motivation
2. Background of Data Set
3. Data Transformation
4. Model Selection
5. Parameter Optimization
6. Model Validation & Results

Motivation

1. What are the key features that influence housing prices in Iowa?
2. What are the best models to predict housing prices?
3. Are there any ways to improve the models?

Process


Exploratory Data Analysis
Clean Numerical & Categorical Features
Feature Selection
Transform Data for Analysis
Optimize Model Parameters
Compare Algorithm Results

Background

Kaggle Competition to predict home sale prices

2930 sales in Ames, Iowa between 2006 - 2010

79 features describing the homes and sale conditions

38 numerical features

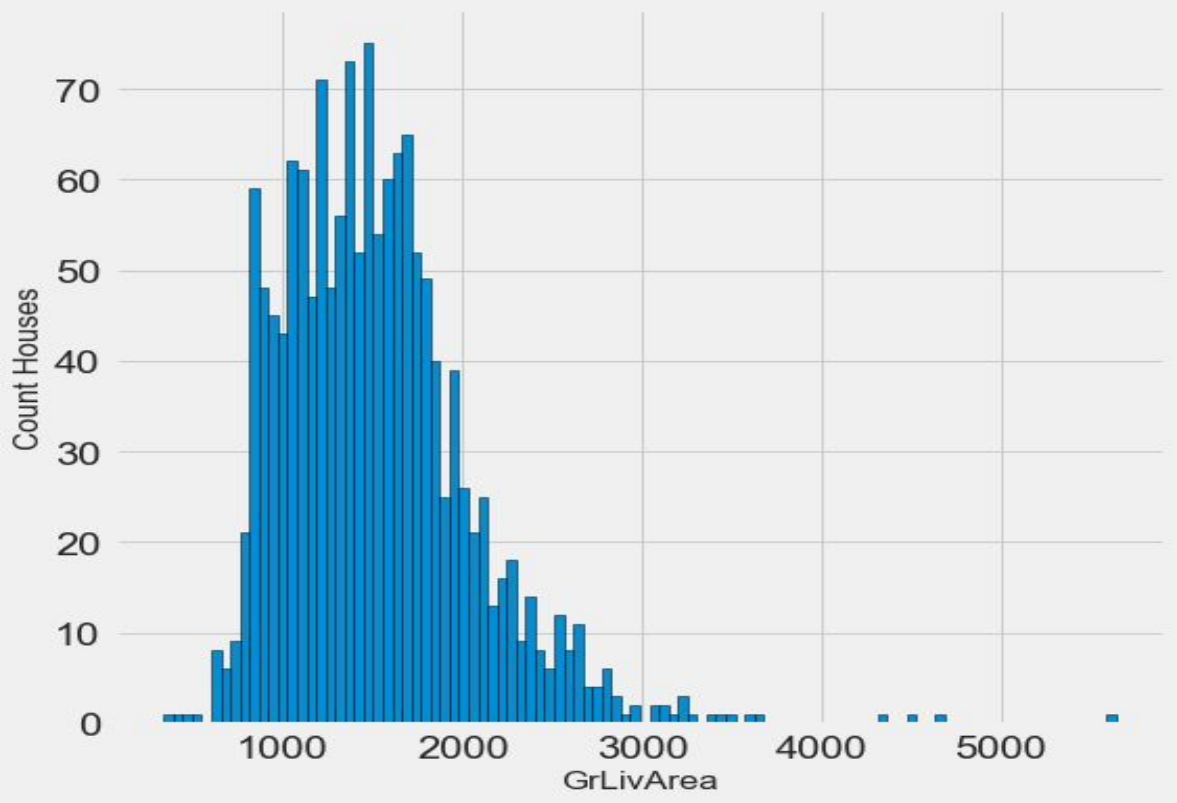
43 categorical features

10 = Very Excellent
9 = Excellent
8 = Very Good
7 = Good
6 = Above Average
5 = Average
4 = Below Average
3 = Fair
2 = Poor
1 = Very Poor

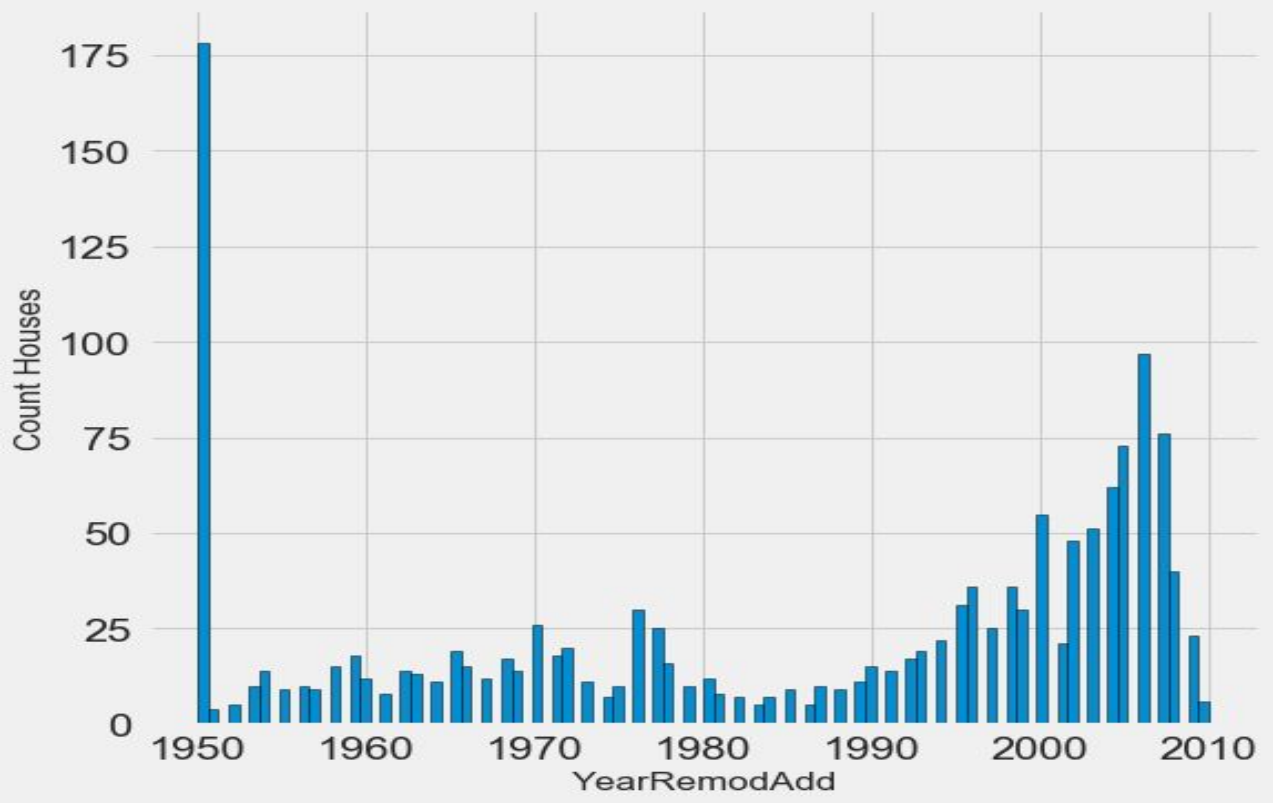
Ex= Excellent
Gd = Good
TA = Typical
Fa = Fair
Po = Poor
NA = No

GLQ = Good Living Quarters
ALQ = Average Living Quarters
BLQ = Below Average Living Quarters
Rec = Average Rec Room
LwQ = Low Quality
Unf = Unfinished
NA = No Basement

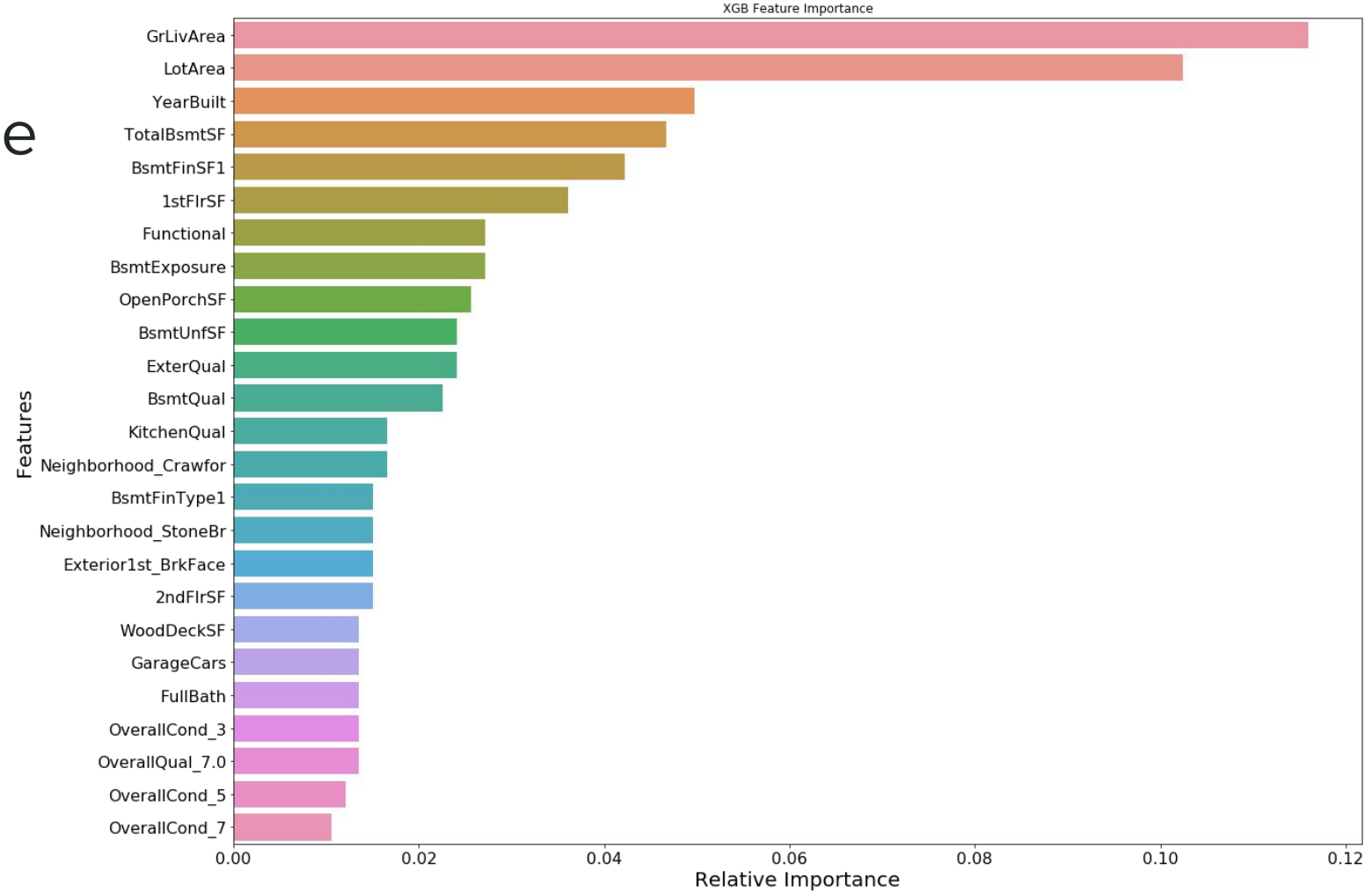
Above Ground Living Area SF



Remodel Date



Feature Importance with XGBoost



Data Transformation

To improve the accuracy of machine learning models, the training dataset needs to be changed and transformed, such as:

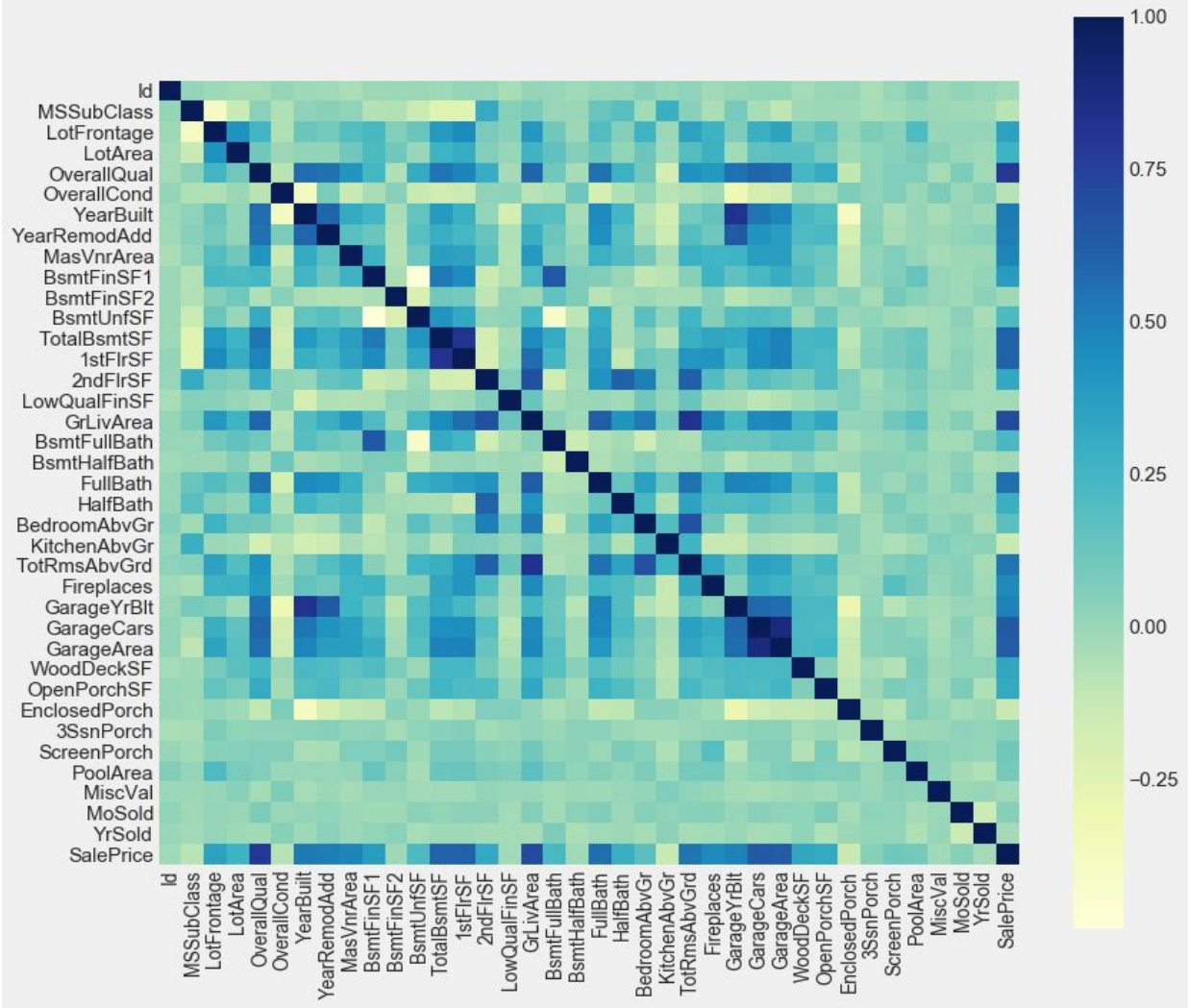
1. Check for multicollinearity
2. Remove outliers
3. Inspect missing values and impute missing values
4. Transform categorical columns with ordinal features
5. Transform numeric columns with categorical features
6. Transform skewed data

Check for multicollinearity

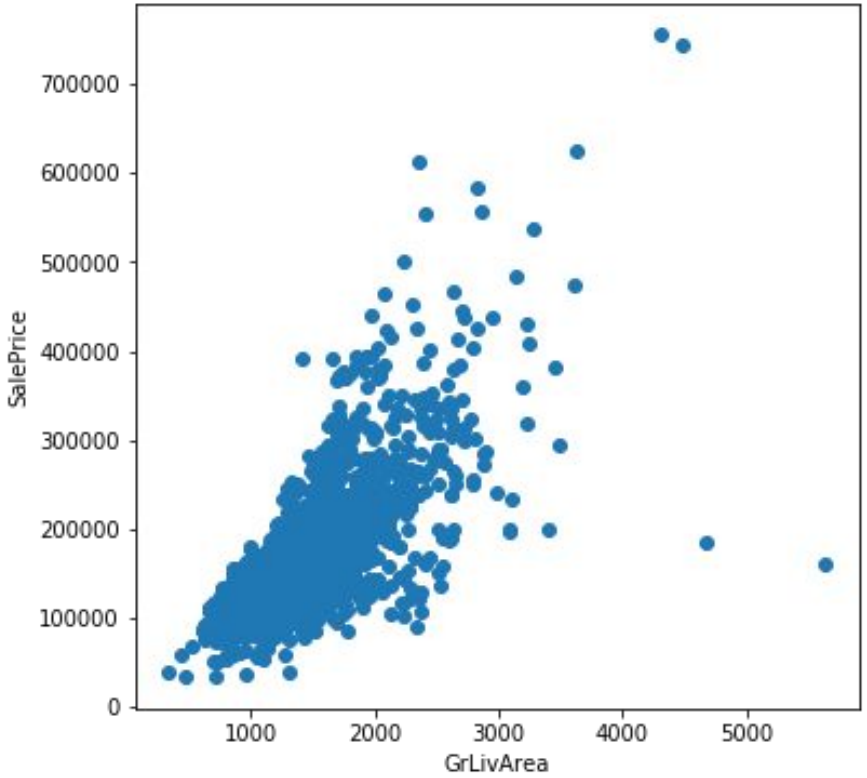
--Remove unnecessary

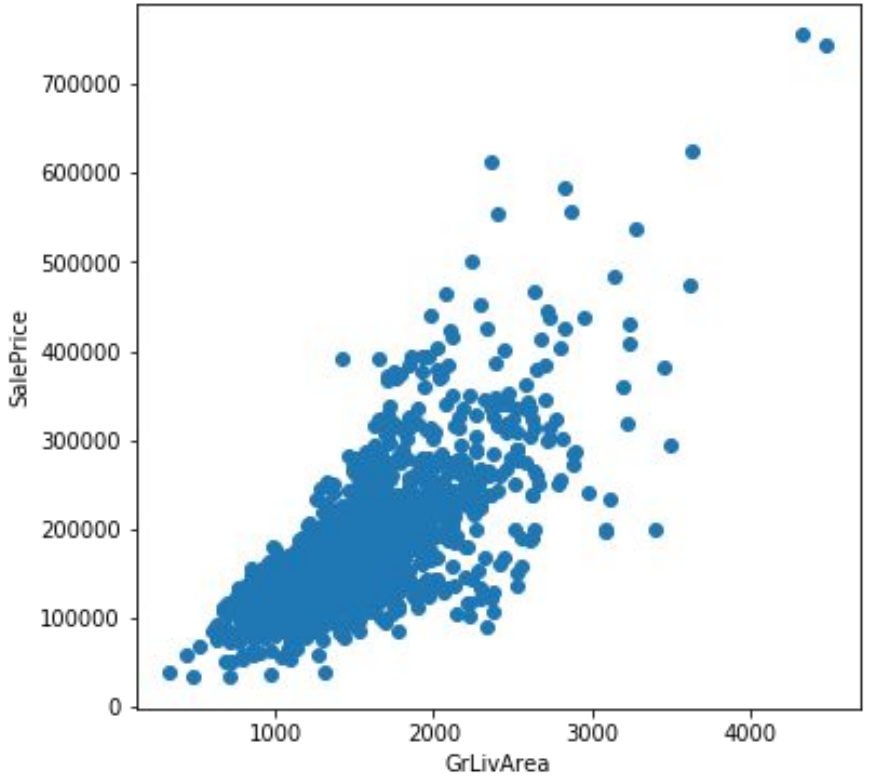
columns

For example,
GarageYrBlt (remove)
& YearBuilt (keep)

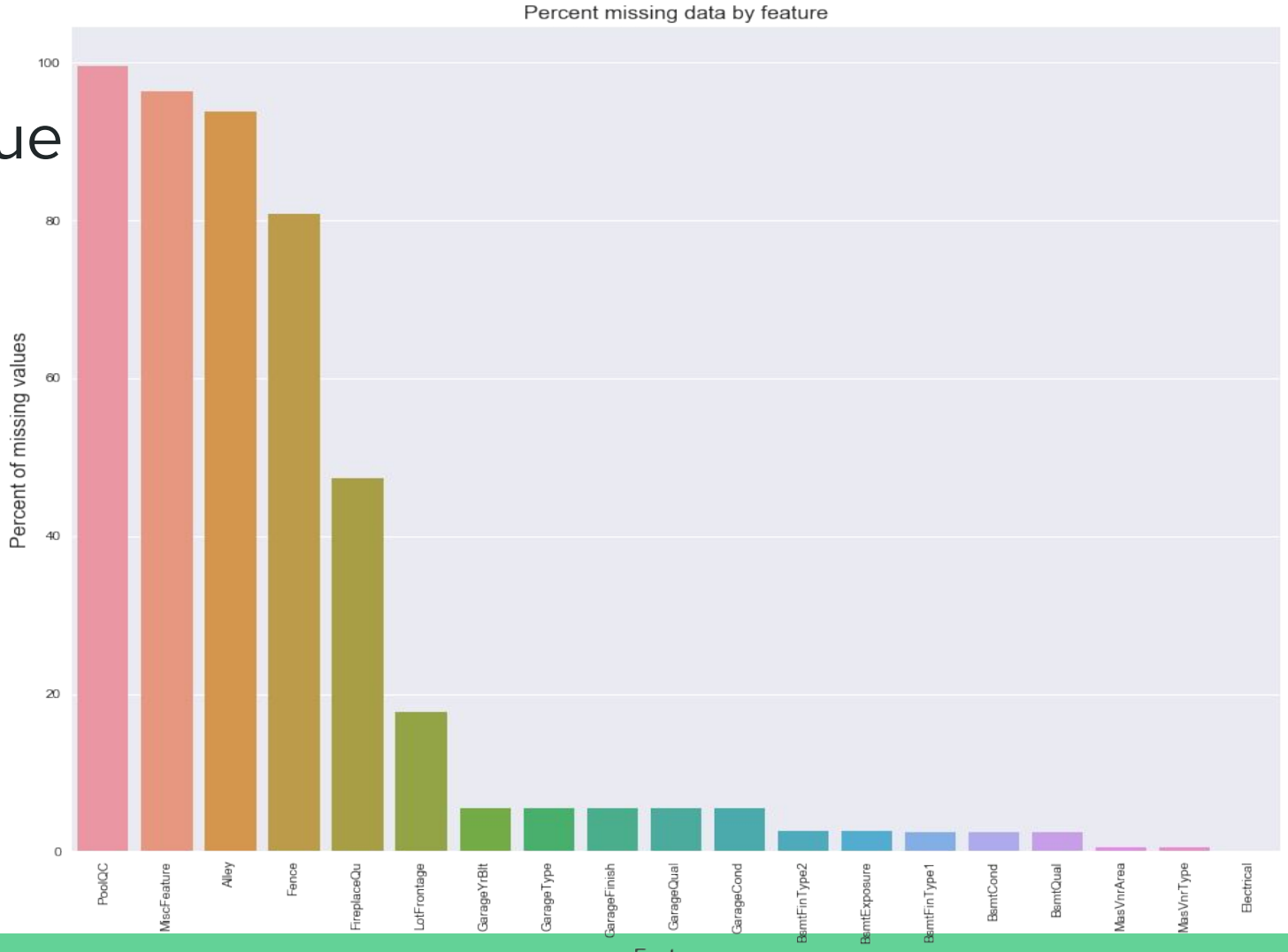


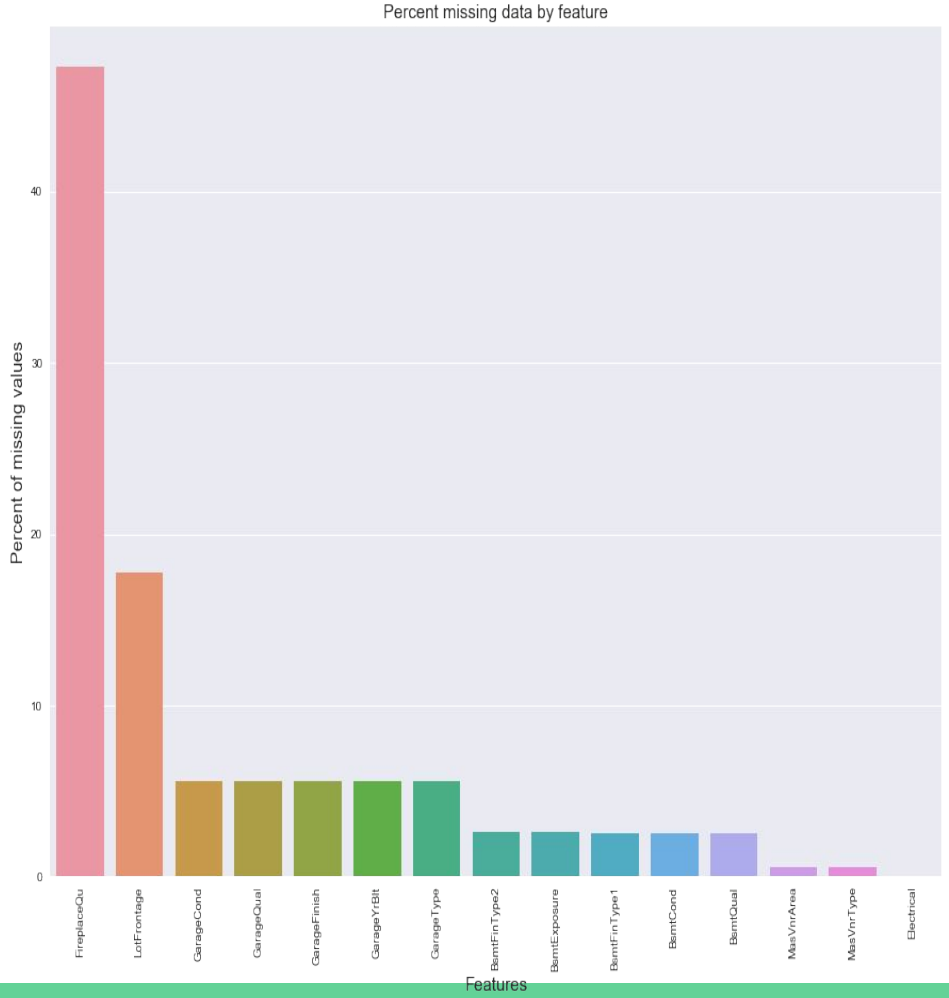
Exploratory Data Analysis - remove outliers





Missing Value By Feature:





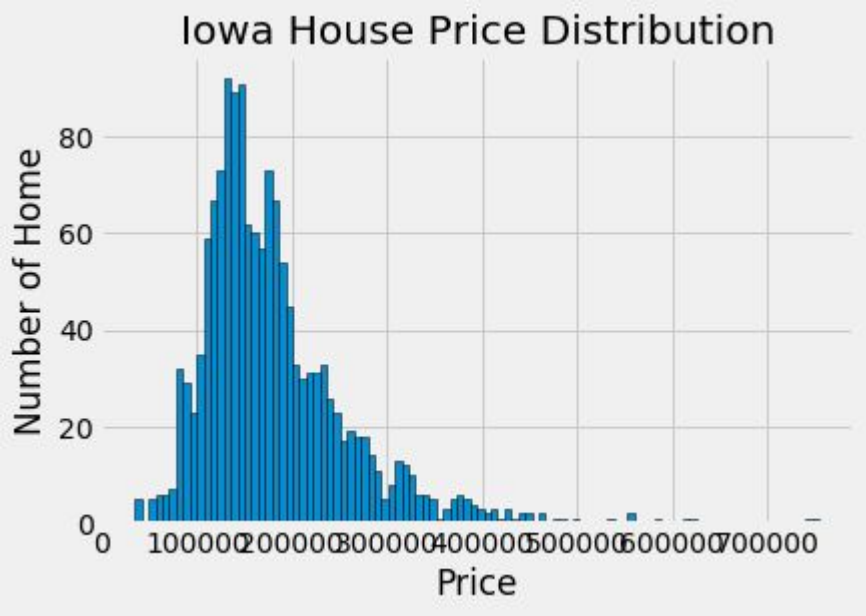
1. Features with “none” : Impute with “0”

2.Features with lost data: Impute with “average”

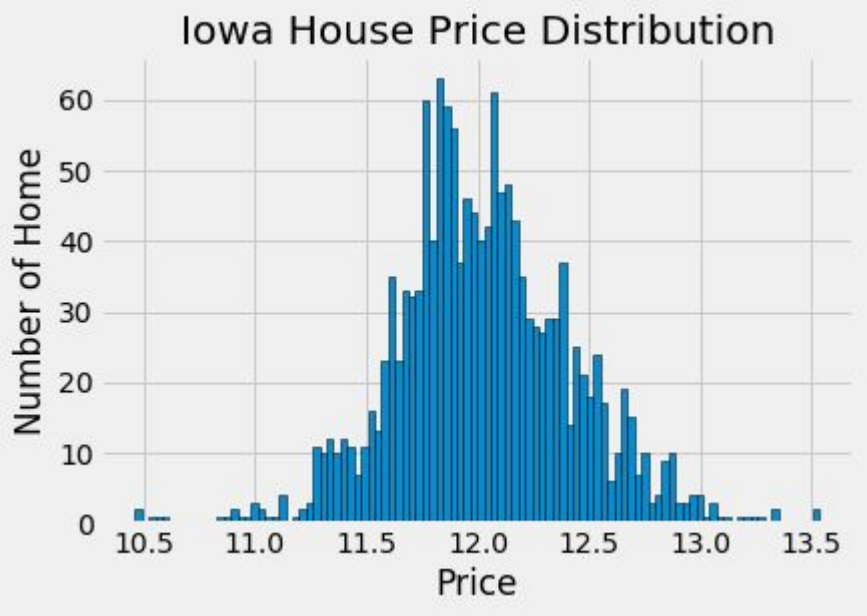
Imputation

Data Transformation

Price distribution with skewness

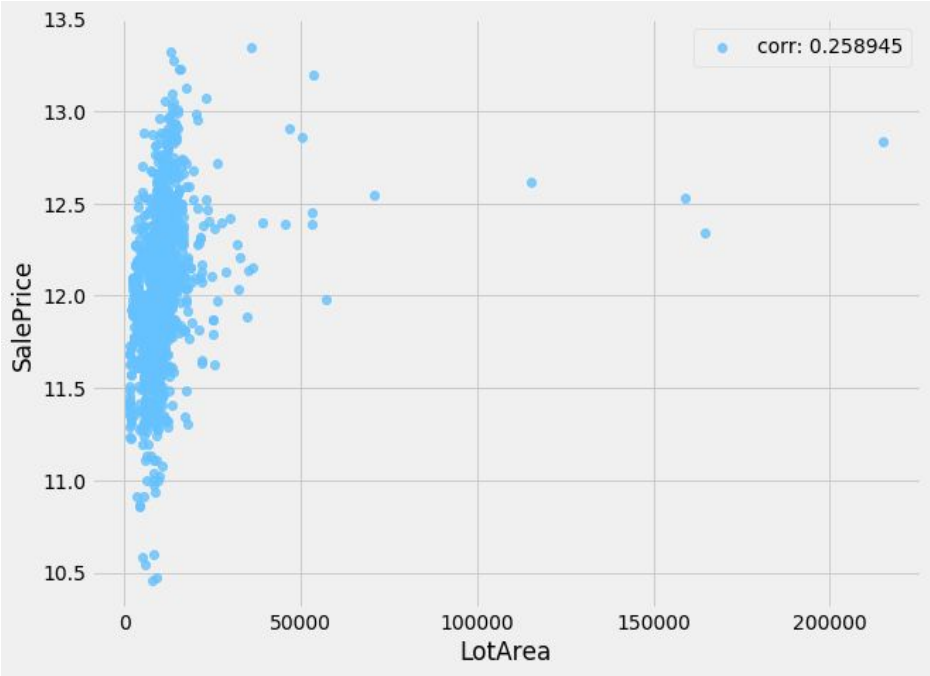


Price distribution after transformation

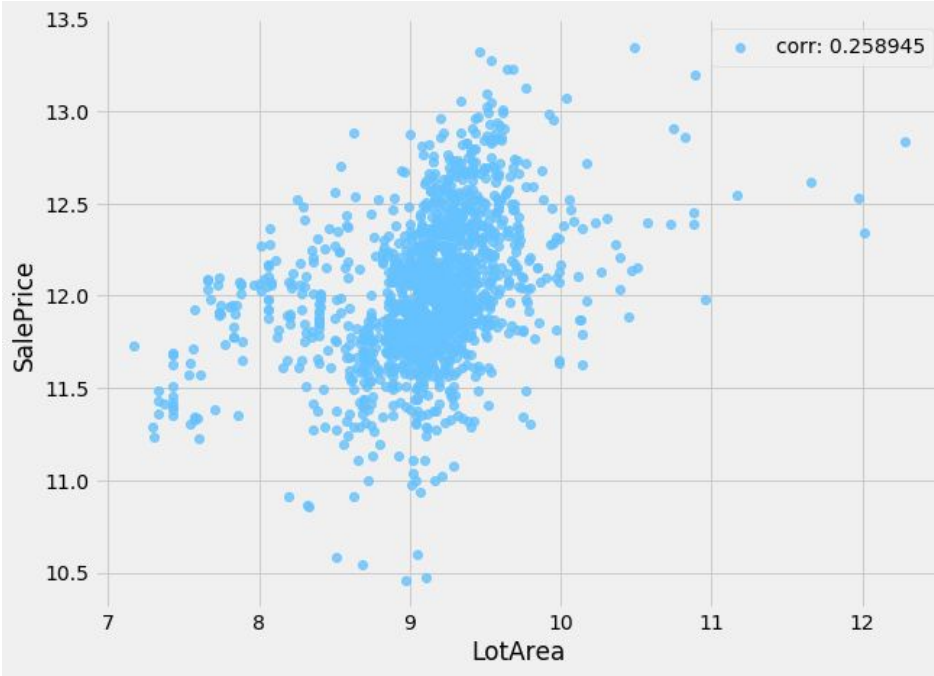


Handling Skewness

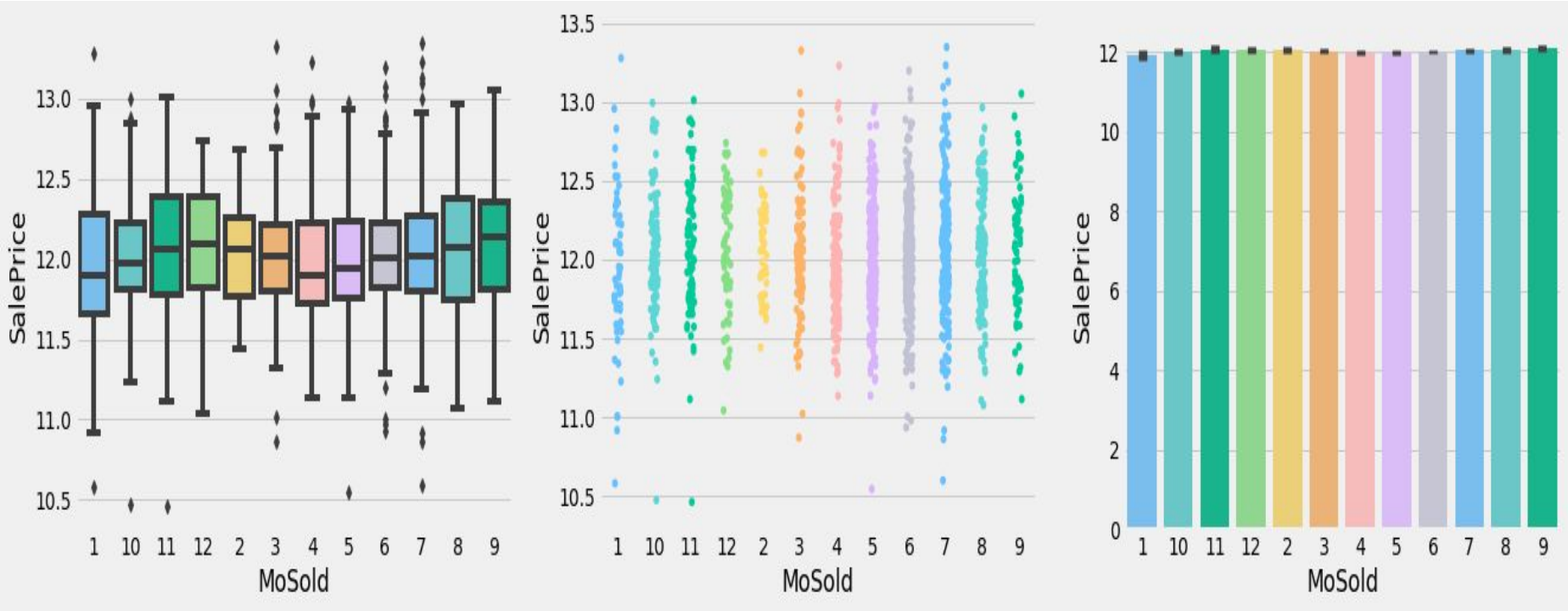
Lot Area with Skewness



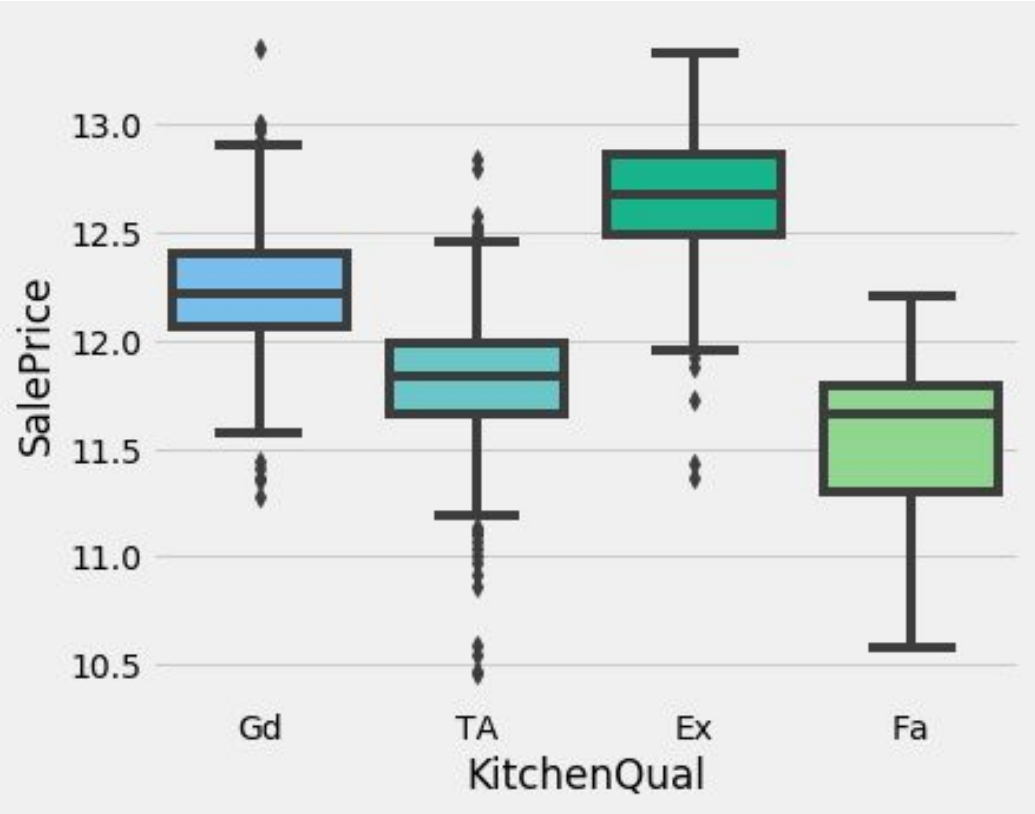
Lot Area after transformation



Numeric Variables with Categorical Feature



Categorical Variables with Label Encoding



Ex: Excellent - 4

Gd: Good - 3

TA: Typical/Avg - 2

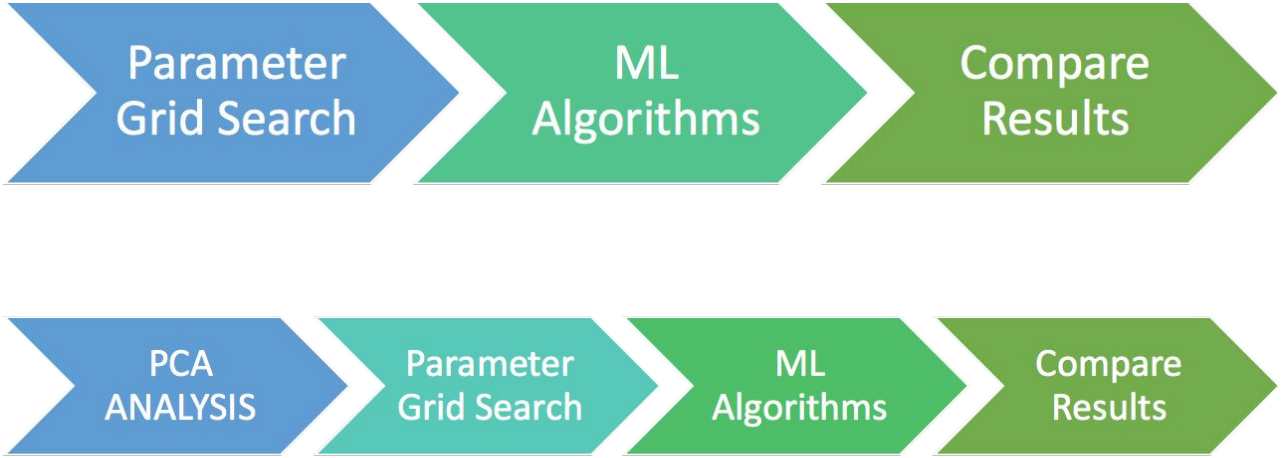
Fa: Fair - 1

Machine Learning Model Development

Set aside 20% for final validation of unobserved data & test for overfitting

Evaluated models with and without Principal Component Analysis

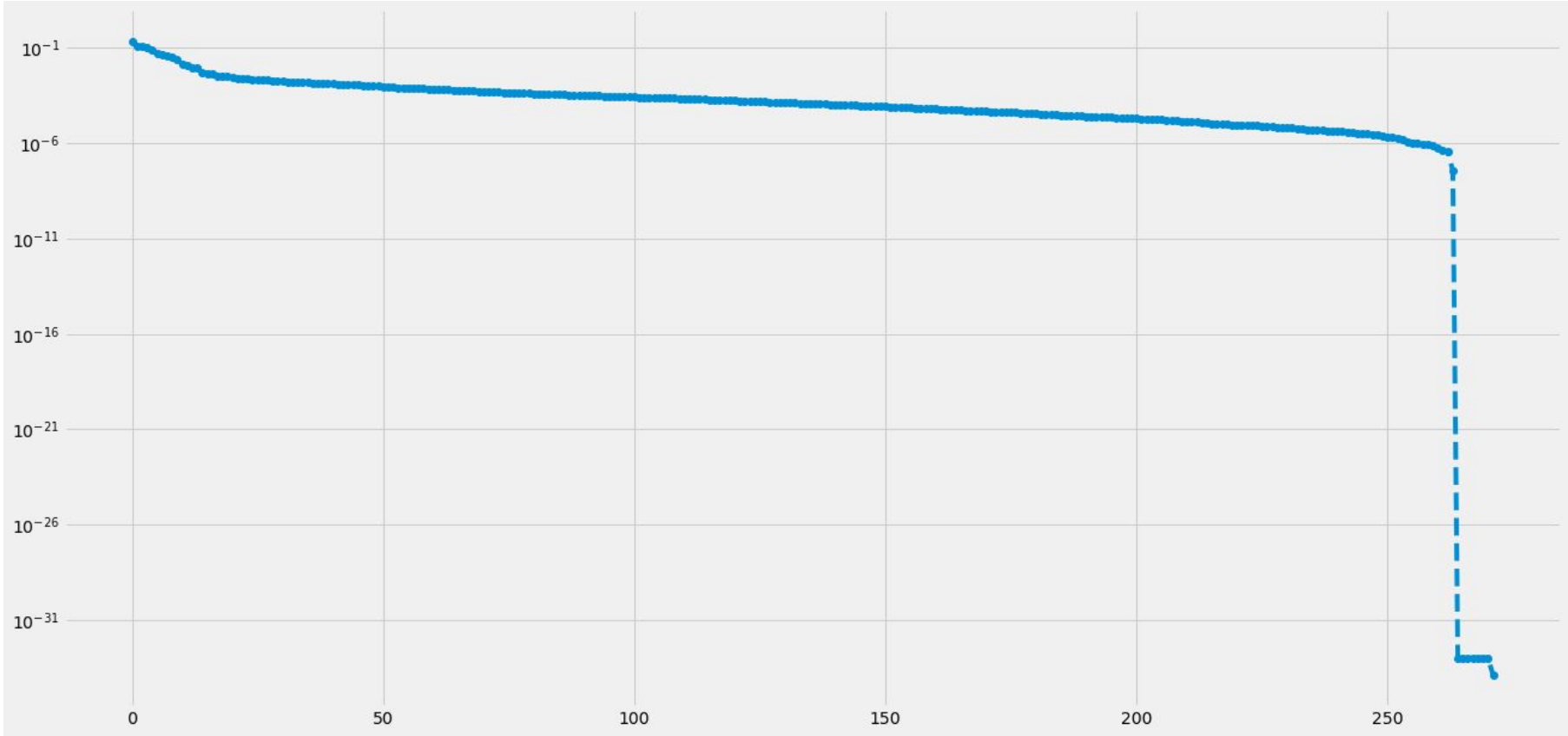
Algorithms optimized with parameter grid search and K-Fold cross validation



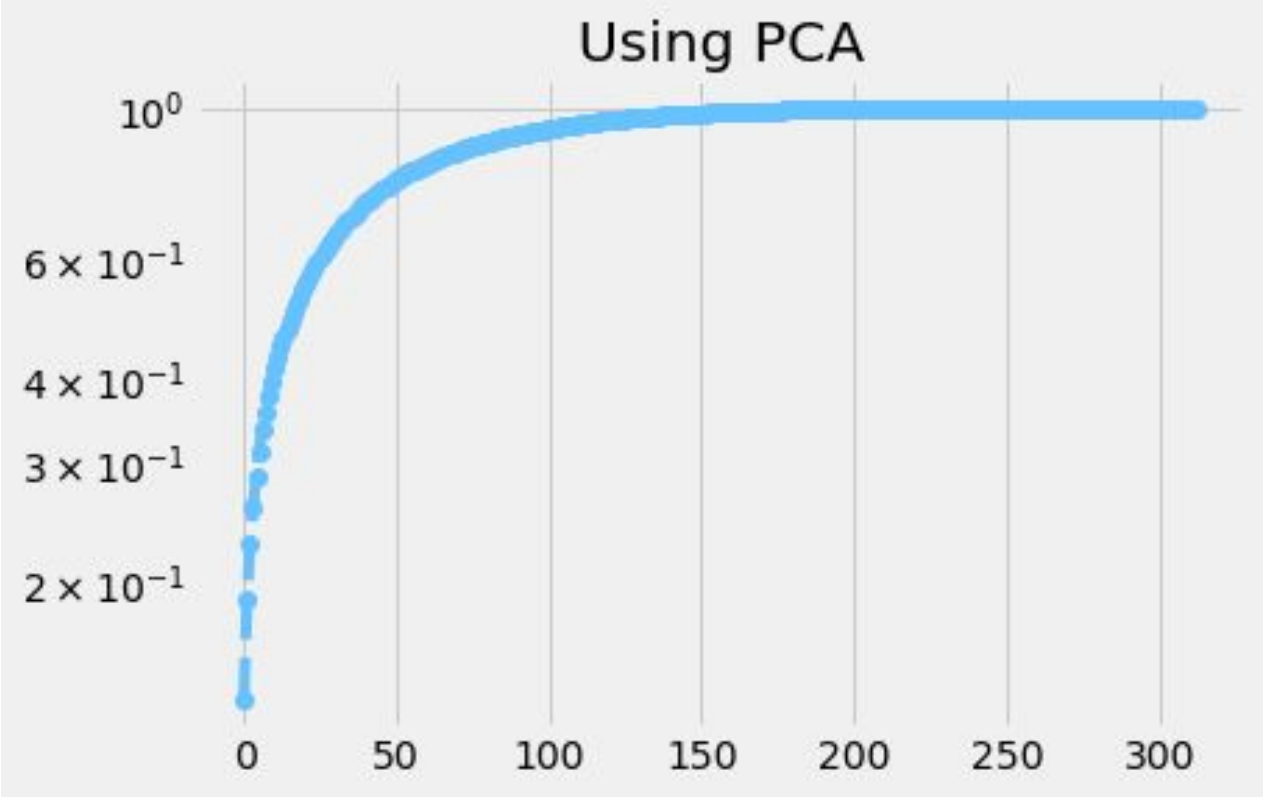
Methods of Testing for Accuracy

- To evaluate overfitting, we created a test set from 20 % of the training data.
- Once we fit a model we then scored it based on the unobserved test set.
- We concluded we were not overfitting from the increased R^2 score.

Using Principal Component Analysis

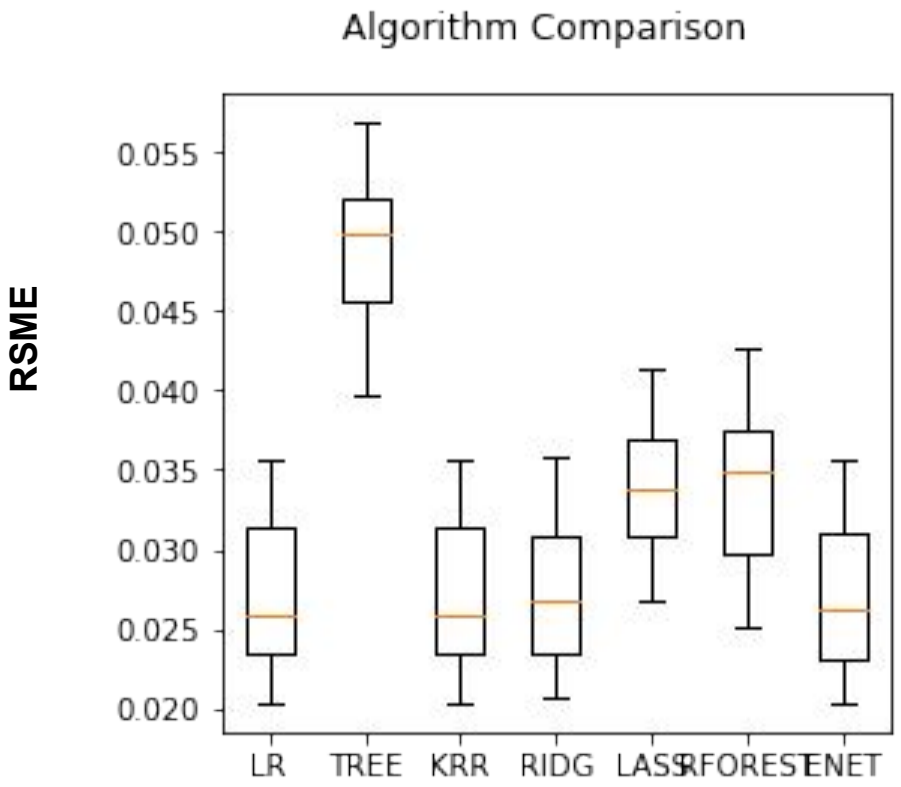


How many Principal Components should we use?

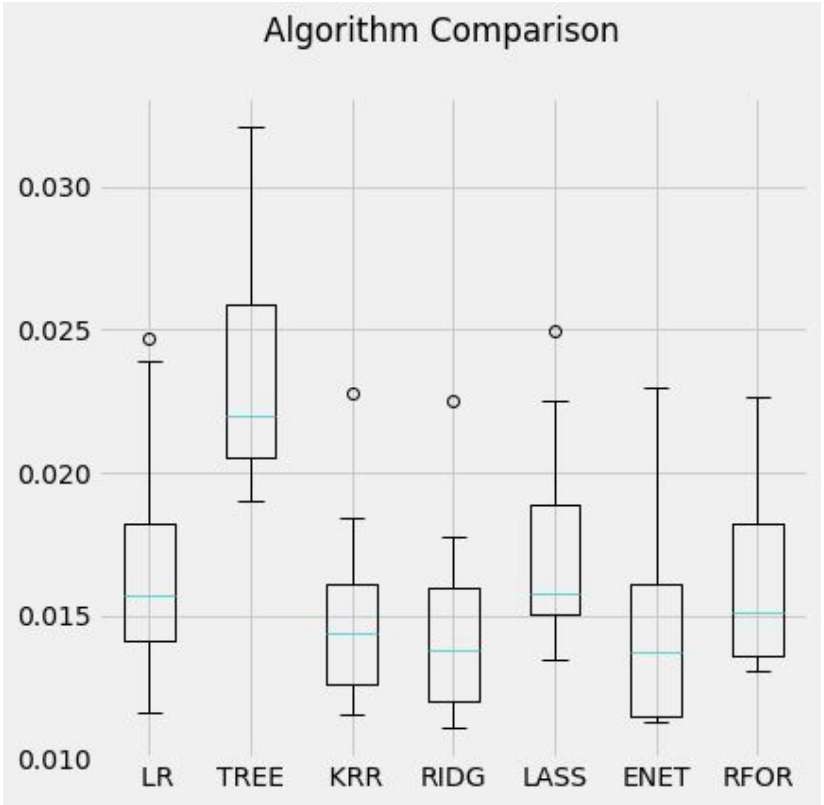


Machine Learning Models - Training Data Set

With PCA

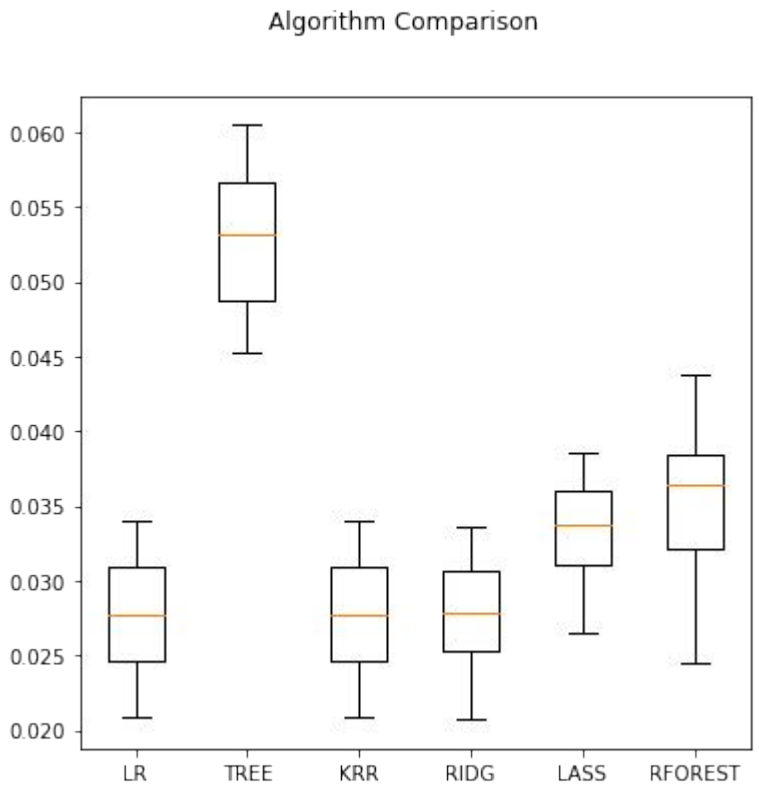


Without PCA

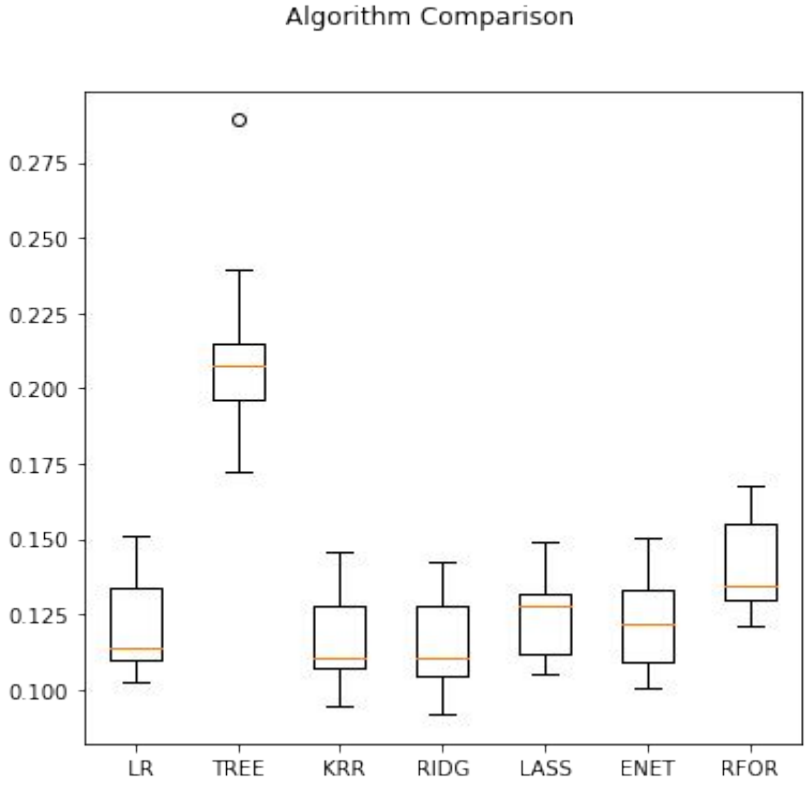


Machine Learning Models - Validation Data Set

With PCA



Without PCA



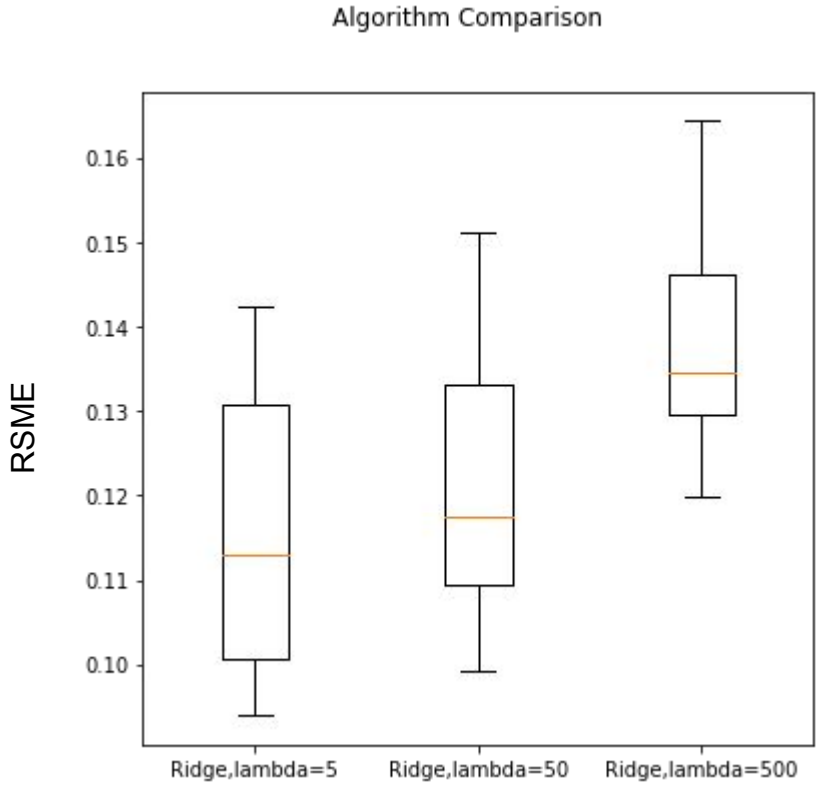
Comparing Results with the Validation Test Data

		Linear Regression	Linear Regression with PCA	Ridge	Ridge with PCA	Kernel Ridge	Kernel Ridge with PCA
Training Data	**RSME**	.018	.027	0.024	0.028	0.025	0.027
	R^2	.84	.81	0.86	0.84	0.84	0.84
Validation Data	**RSME**	.017	.026	0.030	0.026	0.015	0.026
	R^2	.85	.84	0.84	0.84	0.89	0.84

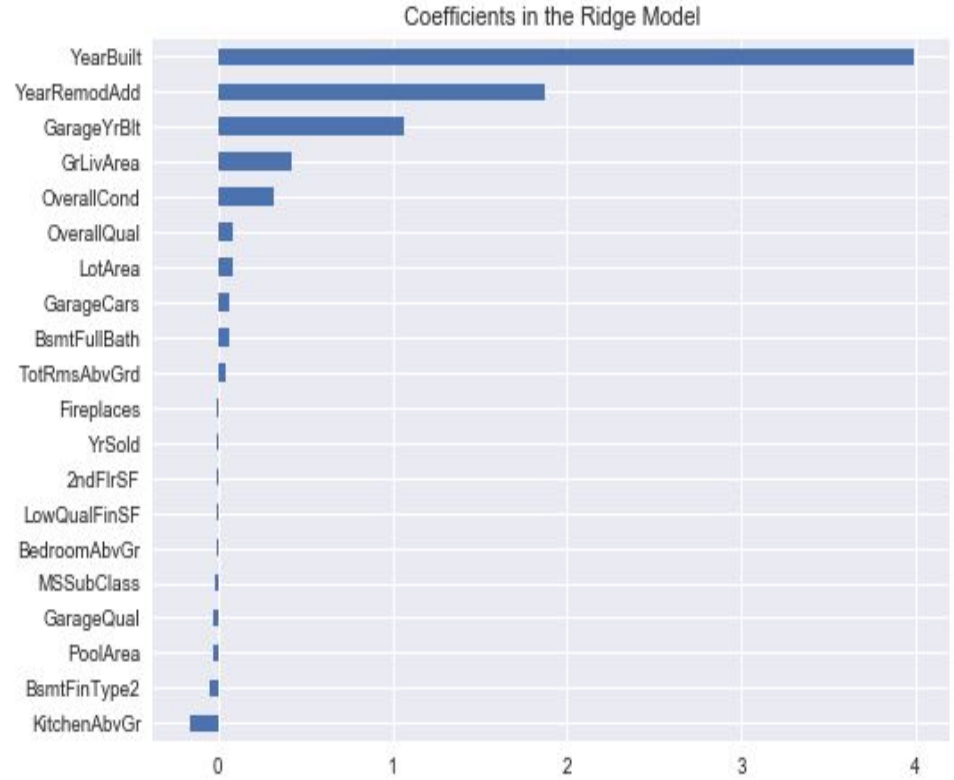
Optimize Model Parameters with Grid Search

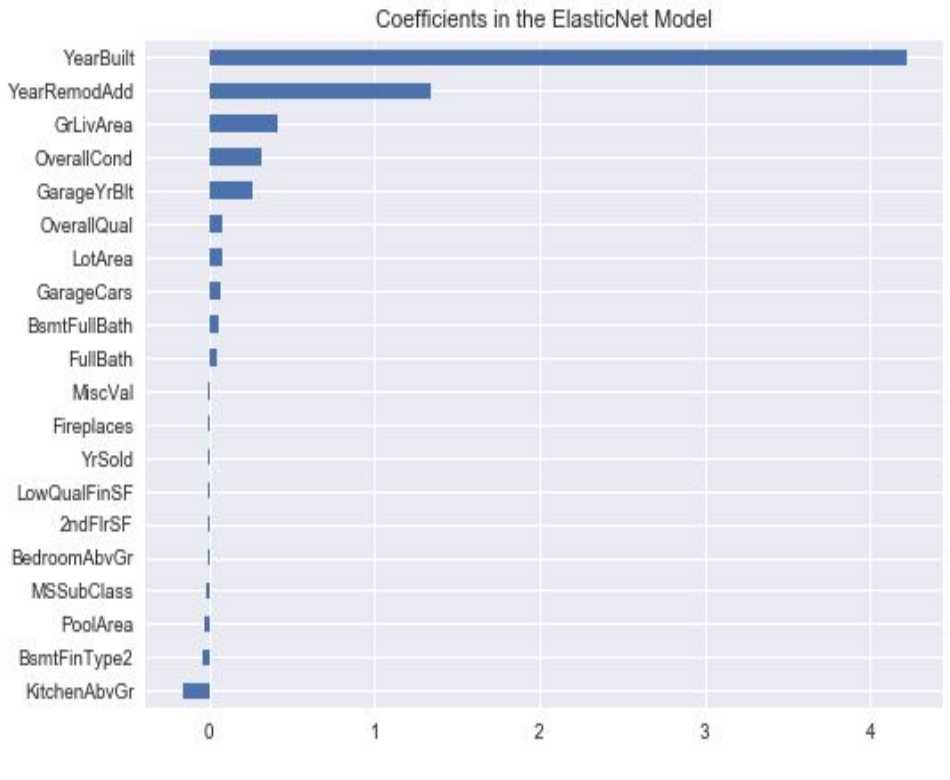
Machine Learning Models:

- Linear
- Lasso
- Ridge
- Kernel Ridge
- Elastic Net
- Random Forest Regression



Feature Importance Variations by Model





Conclusions & Key Insights

- Kernel Ridge Model provided best results
- Grid Search to tune model parameters was beneficial
- Stacking with PCA didn’t show significant improvement

Future Research

- Further parameter optimization with expanded grid search cross validation
- Additional machine learning models
- Ensembling or Stacking various models
- Examine how models work for later dates
- Consider additional data that would influence prices