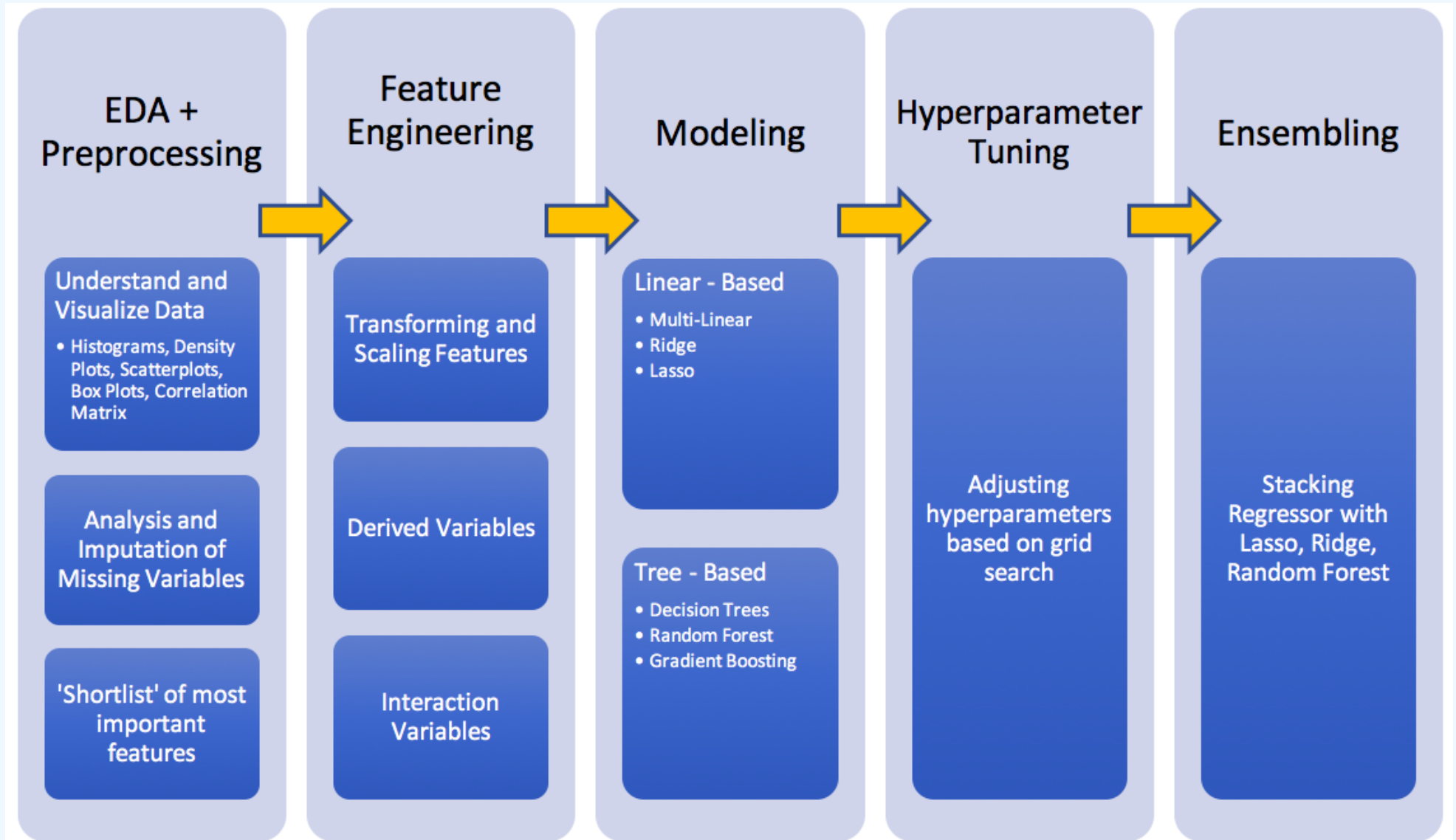# Kaggle House Price Challenge Team Least Squares
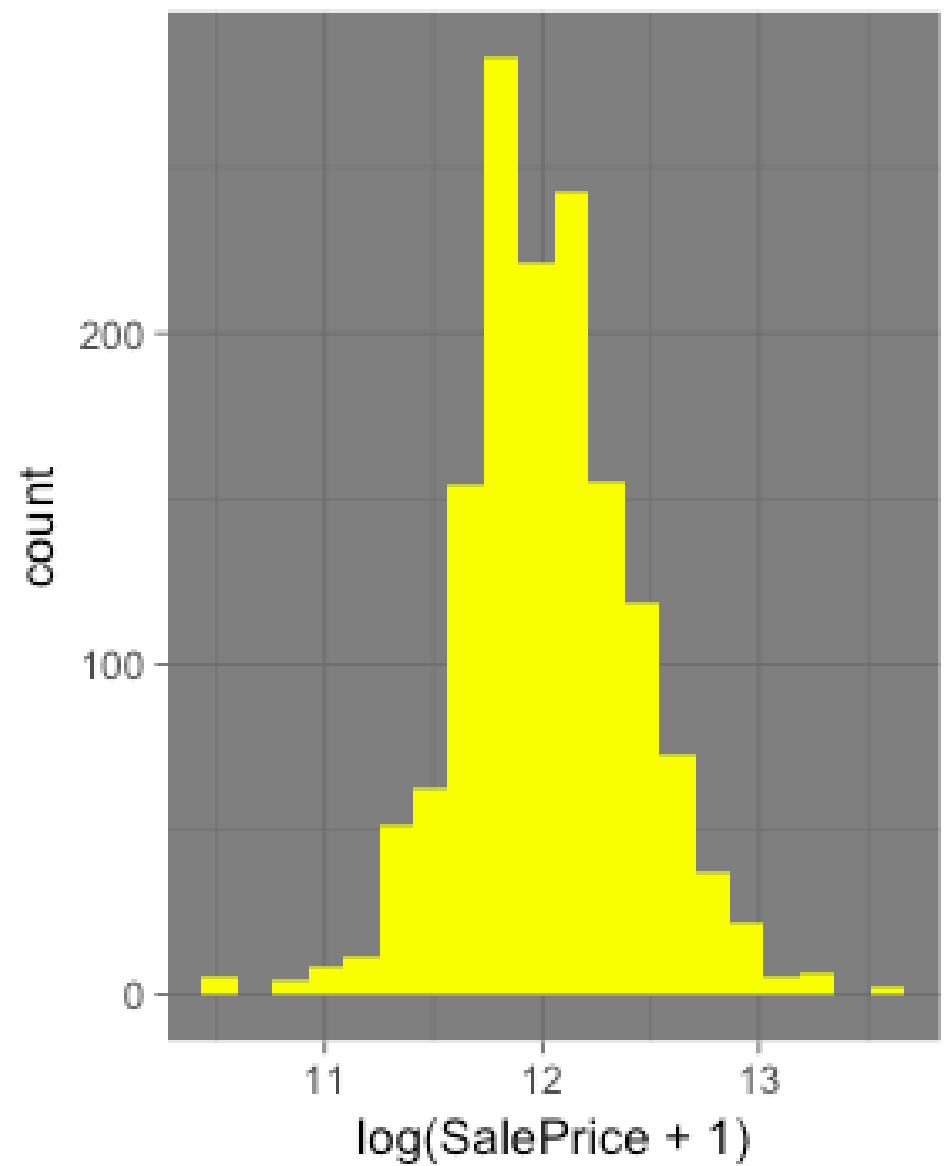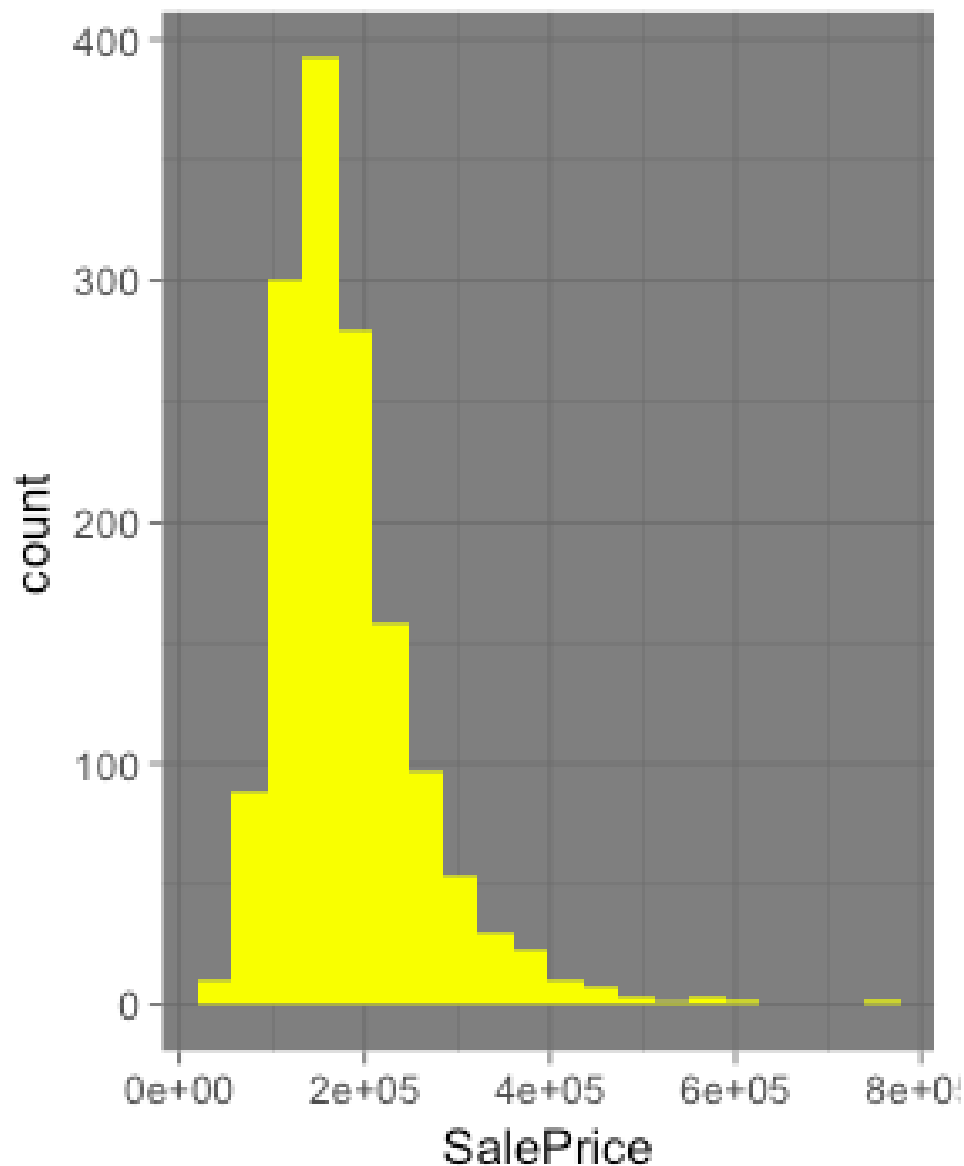
Wing, Iman, Chung, Theo

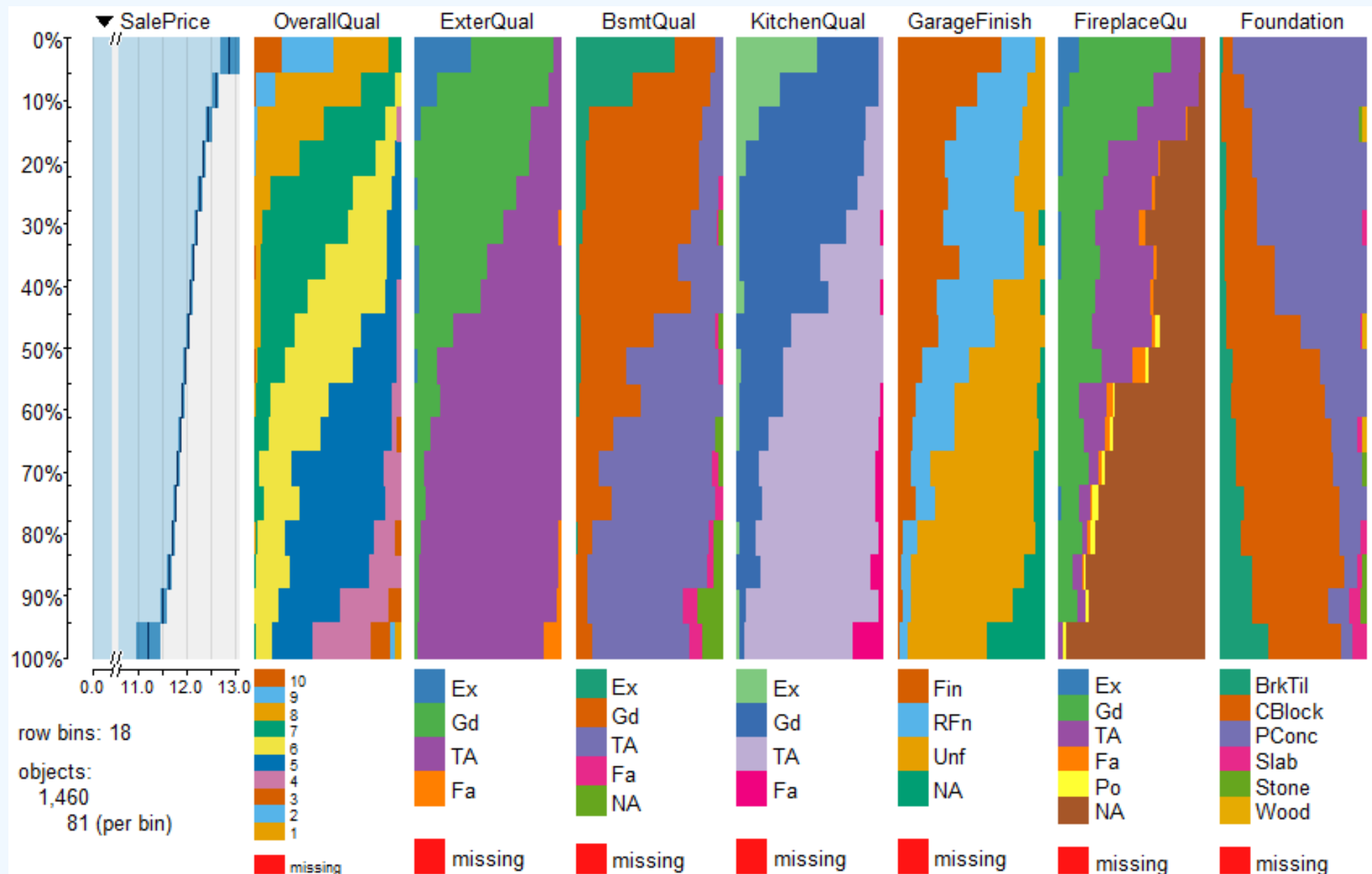# Workflow
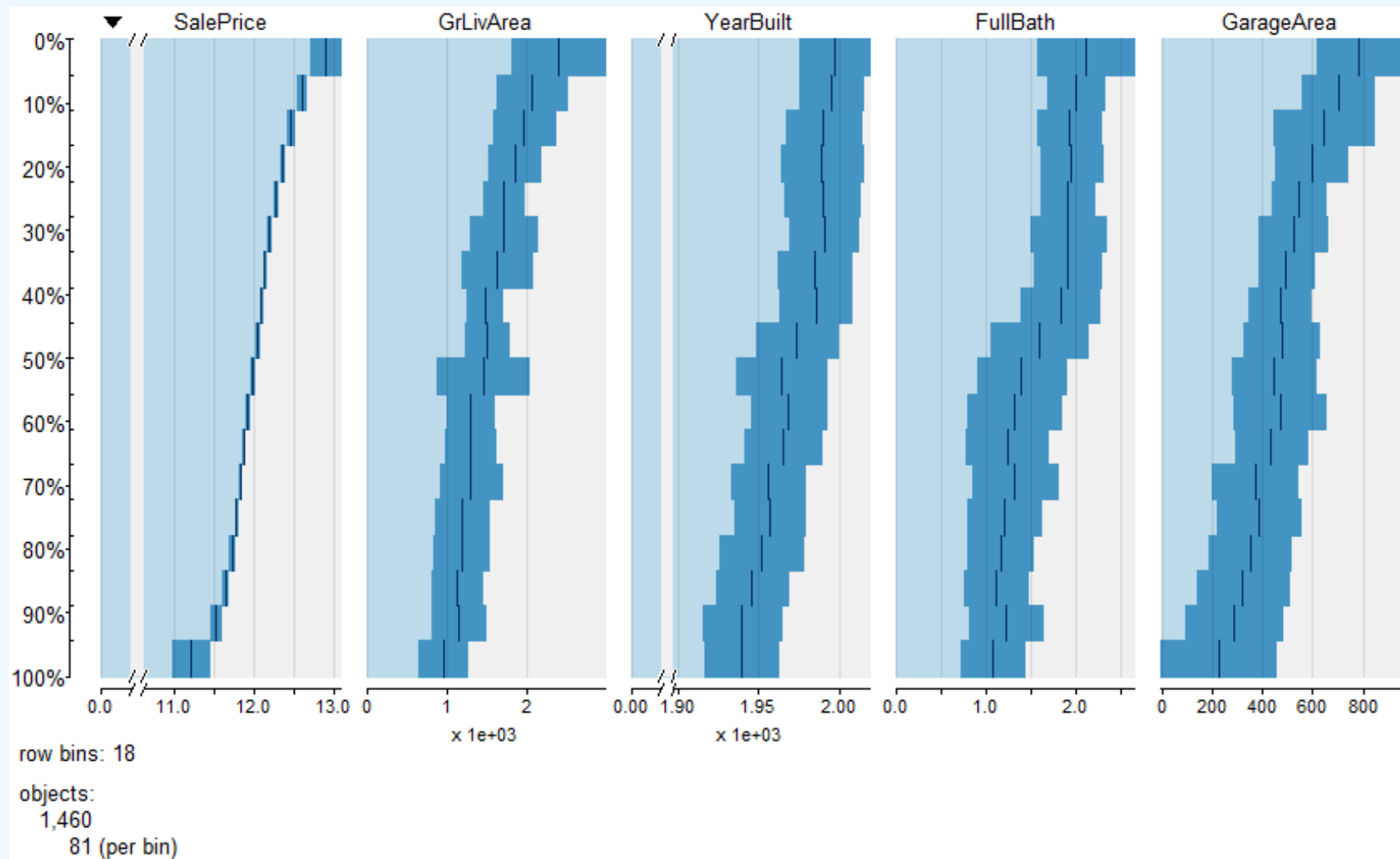
EDA + Preprocessing	Feature Engineering	Modeling	Hyperparameter Tuning	Ensembling

Understand and Visualize Data
- Histograms, Density Plots, Scatterplots, Box Plots, Correlation Matrix

Analysis and Imputation of Missing Variables

'Shortlist' of most important features

Transforming and Scaling Features

Derived Variables

Interaction Variables

Linear - Based
- Multi-Linear
- Ridge
- Lasso

Tree - Based
- Decision Trees
- Random Forest
- Gradient Boosting

Adjusting hyperparameters based on grid search

Stacking Regressor with Lasso, Ridge, Random Forest
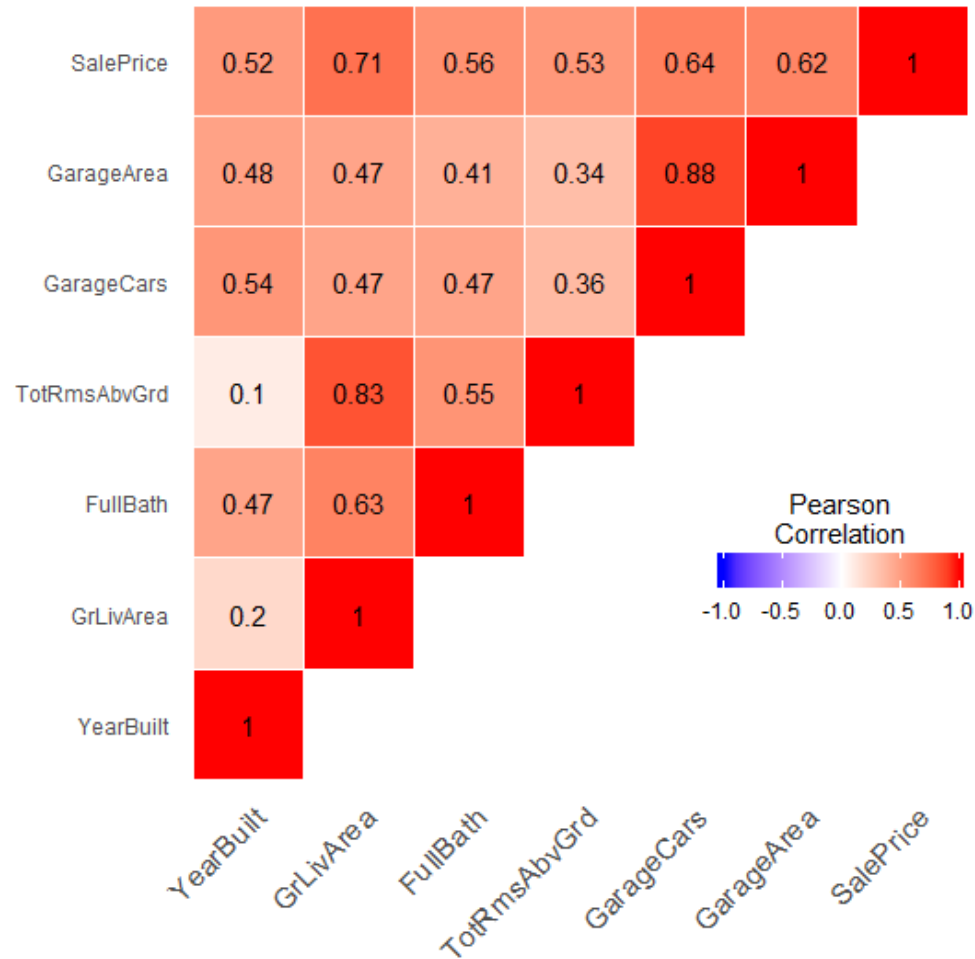
EDA

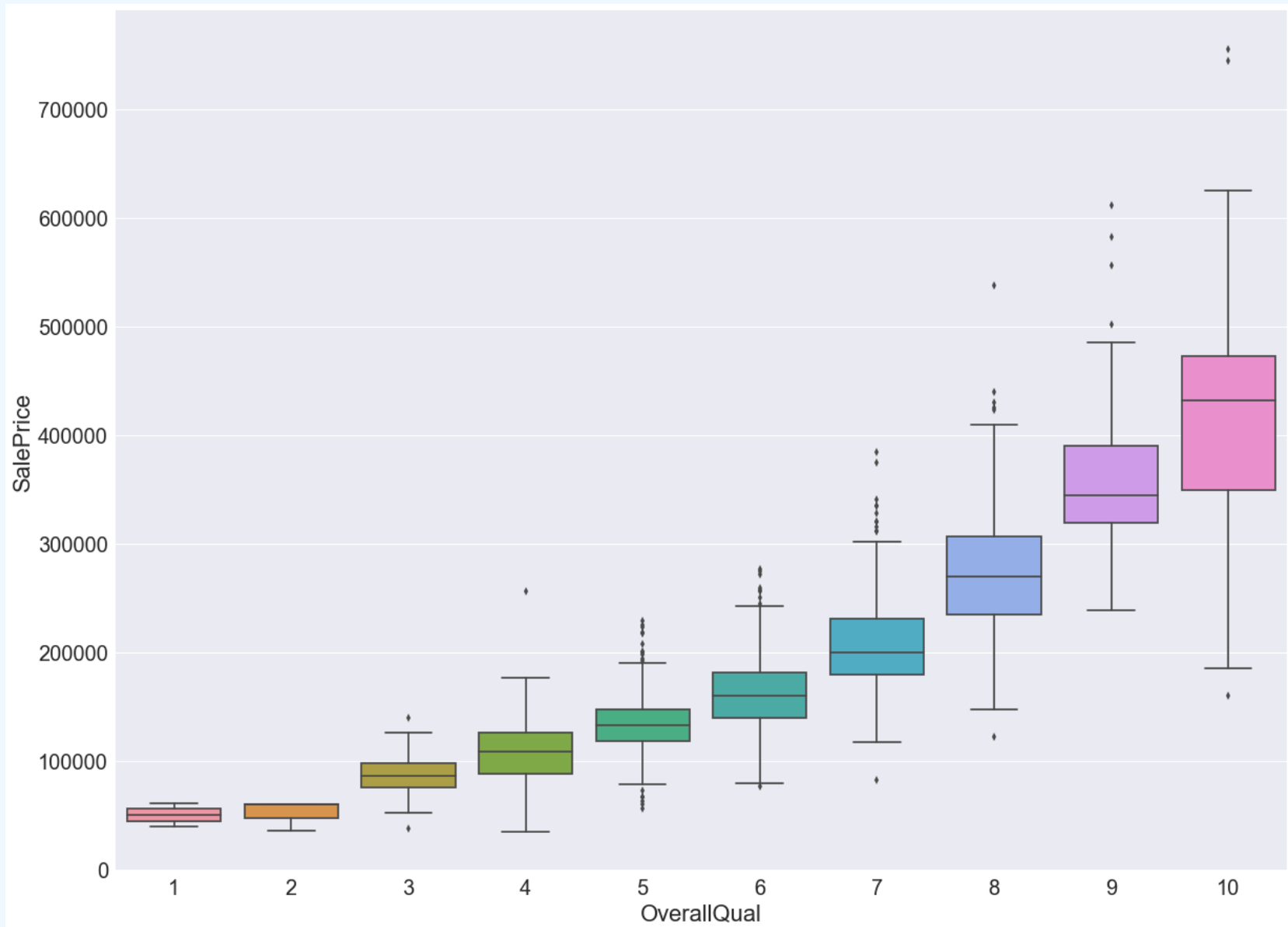# EDA - Sale Price Histogram
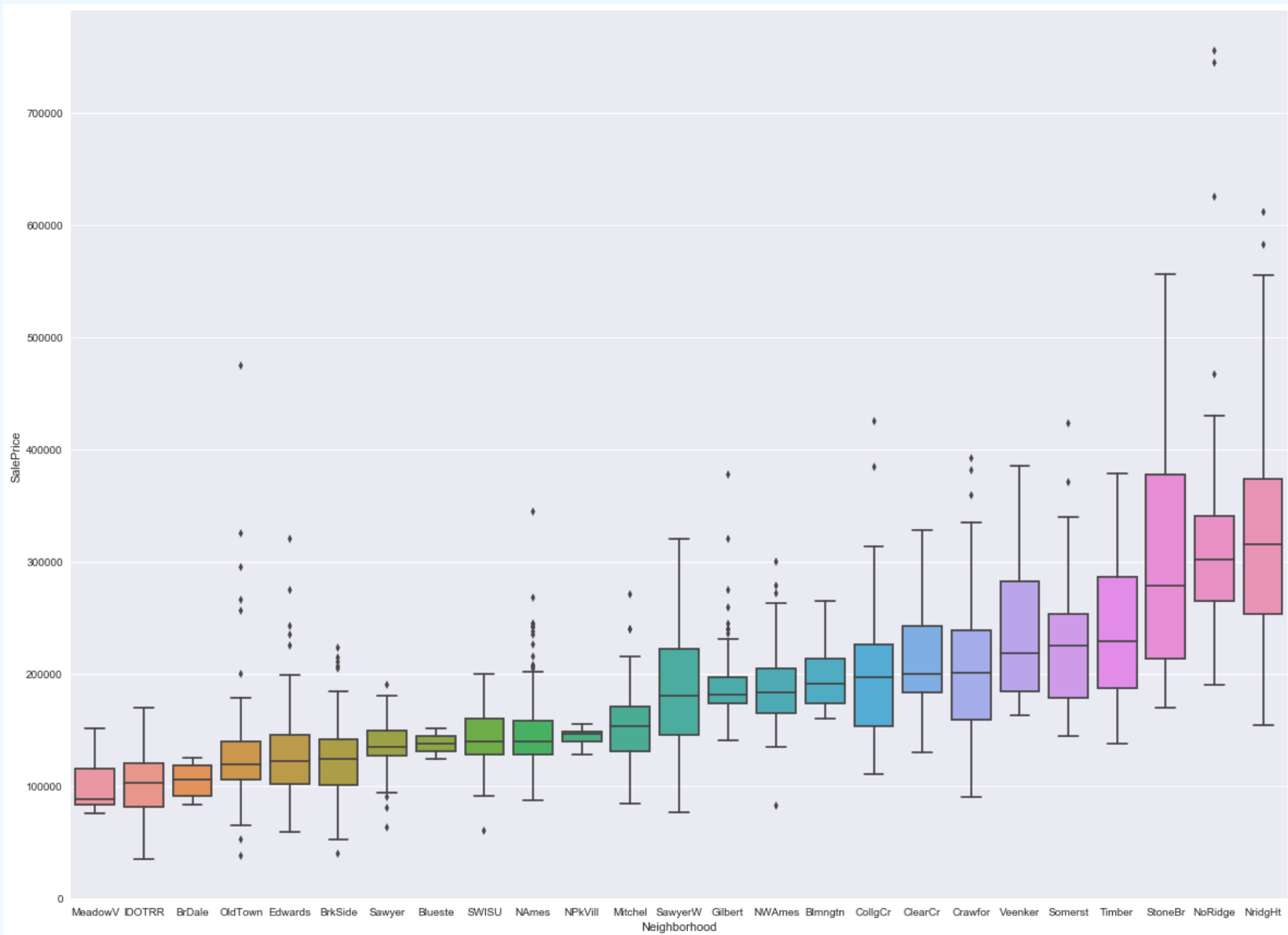
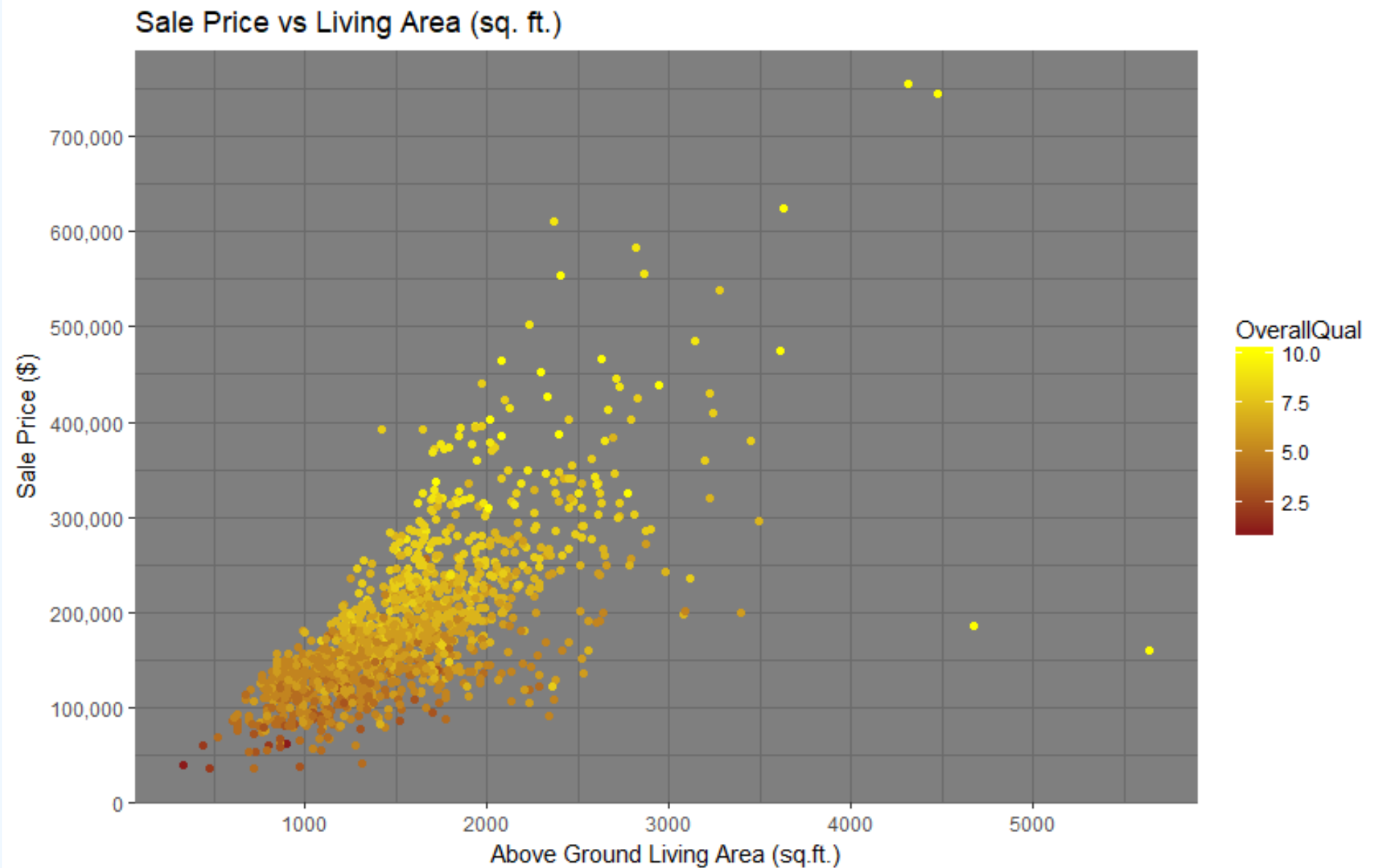# EDA - Categorical Vars

EDA - Continuous Vars

EDA – Correlation Plots

EDA - OverallQual

EDA - Neighborhood

EDA - Scatterplot

Feature Engineering

Imputation

Categorical Variables

Fill NAs : Mode
- MSZoning
- Electrical
- KitchenQual
- SaleType

Fill NAs : None
- PoolQC
- MiscFeature
- Alley
- Fence
- BsmtQual
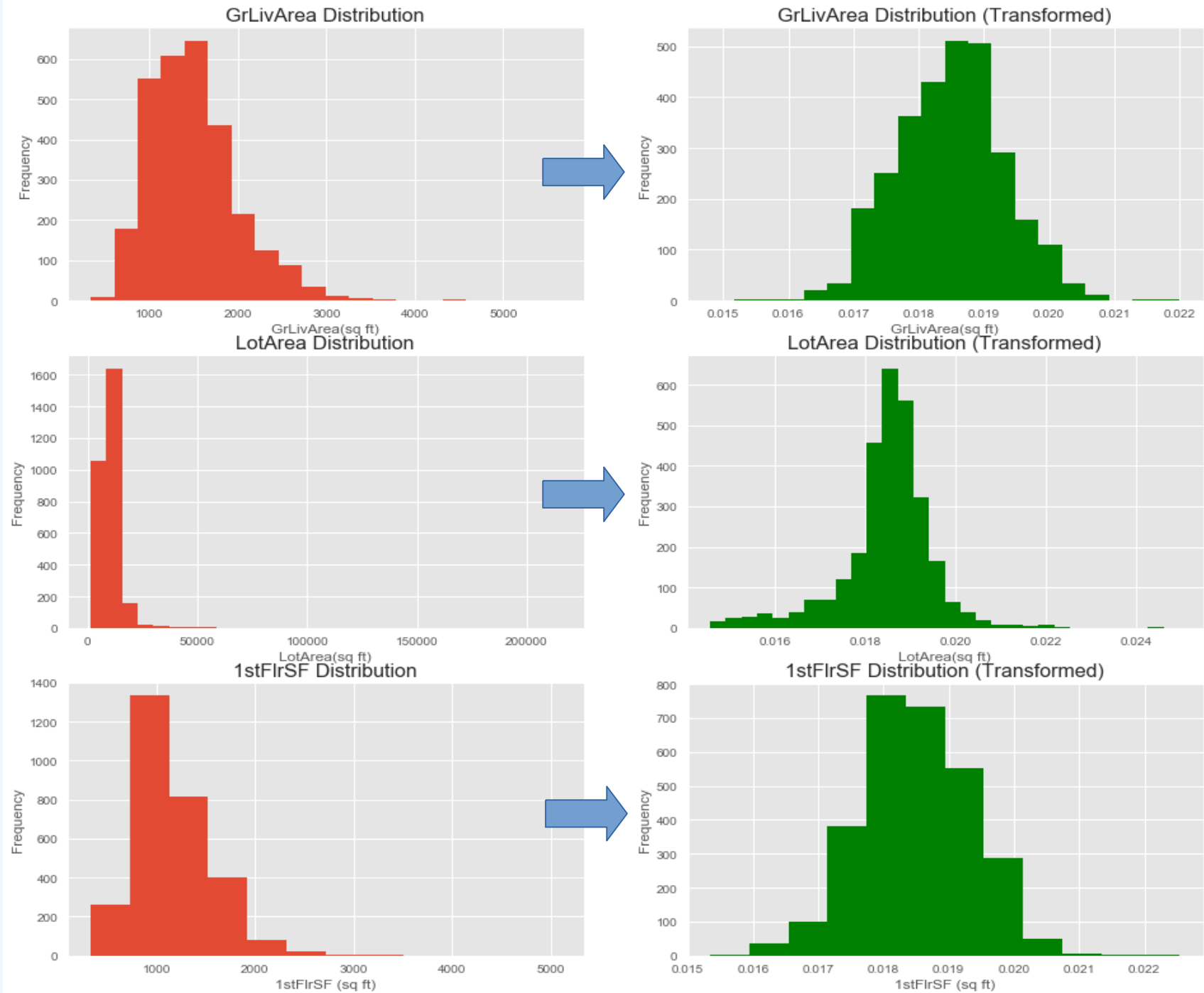- BsmtCond

Numerical Variables

Fill NAs : Median
- LotFrontage
- (Neighborhood Median)

Fill NAs : 0
- BsmtFullBath
- GarageCars
- GarageArea
- MasVnrArea

Feature Scaling & Skewness

Linear Models

- Multiple Linear Regression
  - Explored Initial shortlist of 20 Features
  - Narrowed list to 5 features
  - Performed Regression

- Ridge/Lasso
  - Explored Interaction of Features
 - Expanded Shortlist with Interaction Features

Multiple Linear Regression

Shortlist of 20 Features
(Red ones used in Initial Regression)

Location	Style	Condition	Size	Other
Neighborhood	HouseStyle	OverallQual	GrLivArea	SaleCondition
MSZoning	Foundation	OverallCond	1stFlrSF, 2ndFlrSF	SaleType
	GarageFinish	YrBuilt	FullBath	
	Paved Drive	ExterQual	TotRms	
		BsmtQual	GarageCars	
			GarageArea	

Multiple Linear Regression
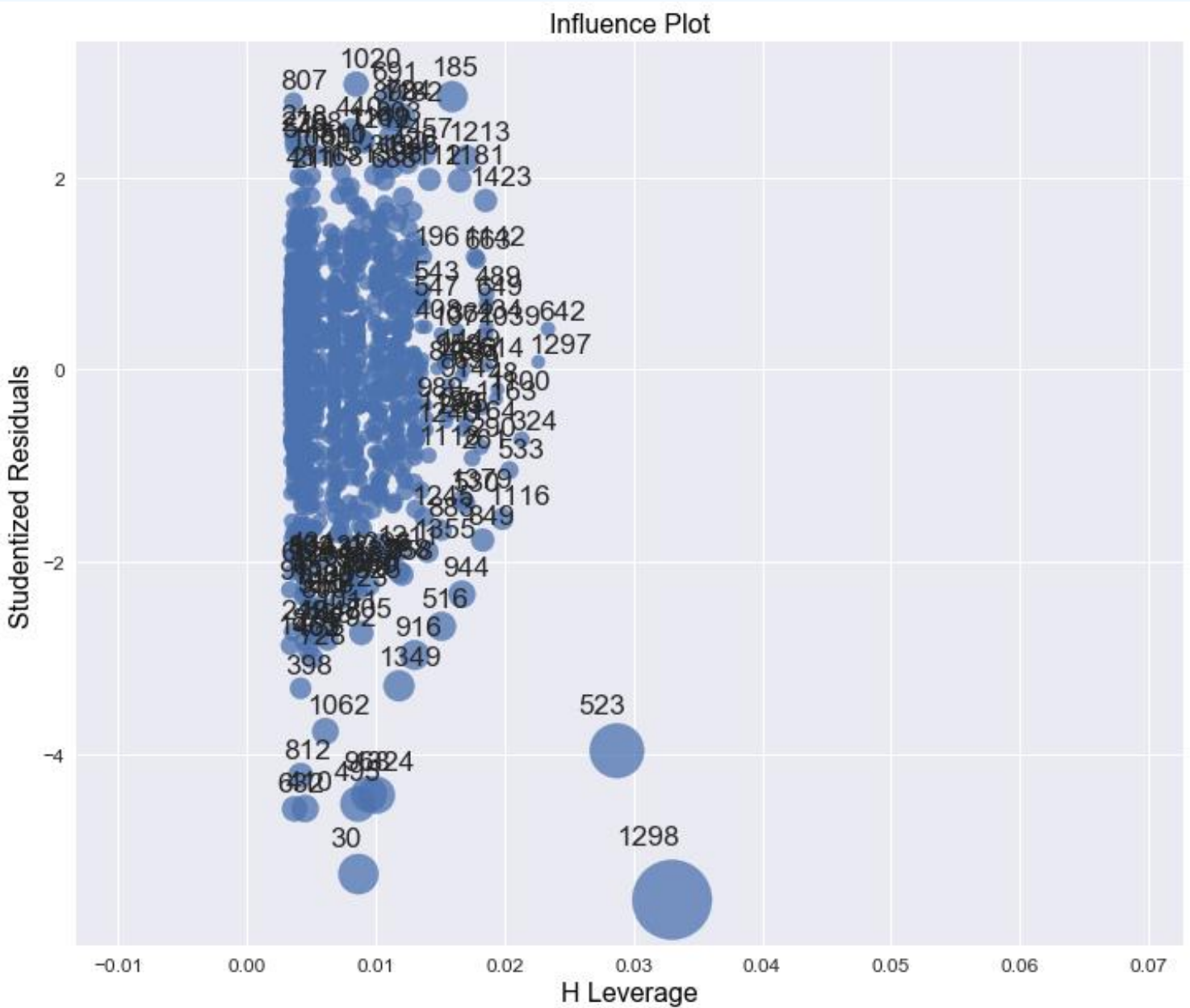
Multiple Linear Regression

	R^2 (Training Set)	RMSE (Training Set)
With Influential Points	0.800	0.178
Without Influential Points	0.807	0.176

Feature Engineering

Derived Features

- TotalSF = TotalBsmtSF + GrLivArea

- HighQualFinishedSF
-     = TotalSF - LowQualFinSF

- TotalBaths
    = Full Baths + BsmtFullBath +
      0.5 * (Half Baths + BsmtHalfBath)

Interaction Features

Examples of Numeric Interactions:
- TotalBaths * Total Square Footage
- OverallQual (as int) * Total Square Footage

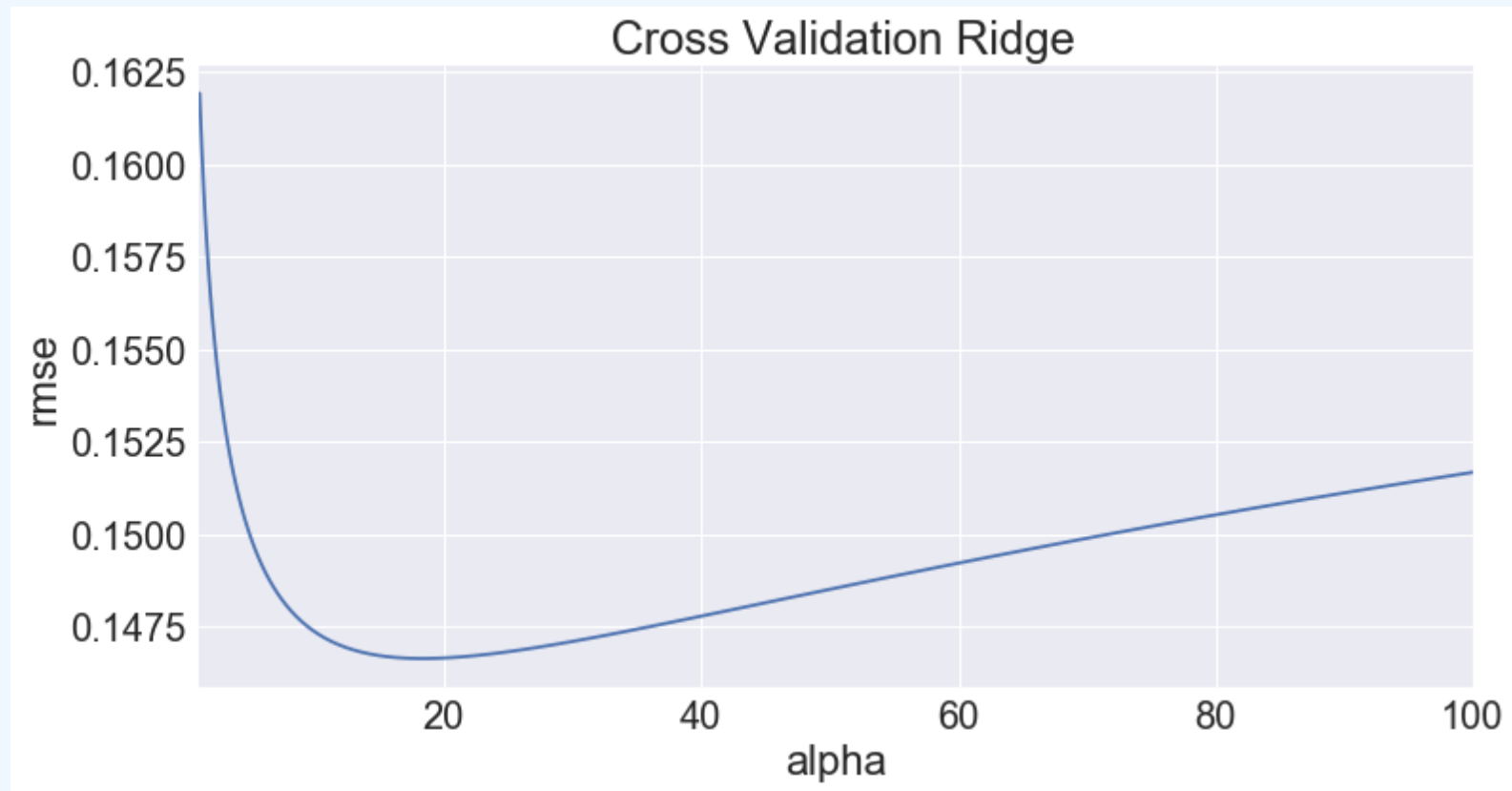Examples of Categorical Interactions:
- OverallQual + OverallCond
- Neighborhood + OverallQual + OverallCond

RIDGE/LASSO PROCESS

- Split training data 80%/20%
- Used grid search to tune hyperparameter
  - 5-fold cross-validation for each parameter value
- Tested against remaining 20%
- Trained against entire 100% of training set

# Ridge Regression

Best Alpha = 18.7

Ridge Regression

	R^2	RMSE
SUB-TRAINING SET (80%)	0.938	0.100
TEST SET (20%	0.913	0.111
TOTAL TRAINING SET (100%)	0.933	0.099

Ridge Regression

Top Features	Coefficient Value
OverallCond_3	-0.066691
Neighborhood + OverallQual + OverallCond_Edwards10+5	-0.066562
MSSubClass_160	-0.051320
Neighborhood_Edwards	-0.048231
Neighborhood + OverallQual + OverallCond_IDOTRR4+4	-0.048116
Neighborhood_Crawfor	0.048117
CentralAir_Y	0.049257
MSZoning_RL	0.053233
OverallQual	0.057236
Functional_Typ	0.057430

Lasso Regression

Cross Validation Lasso
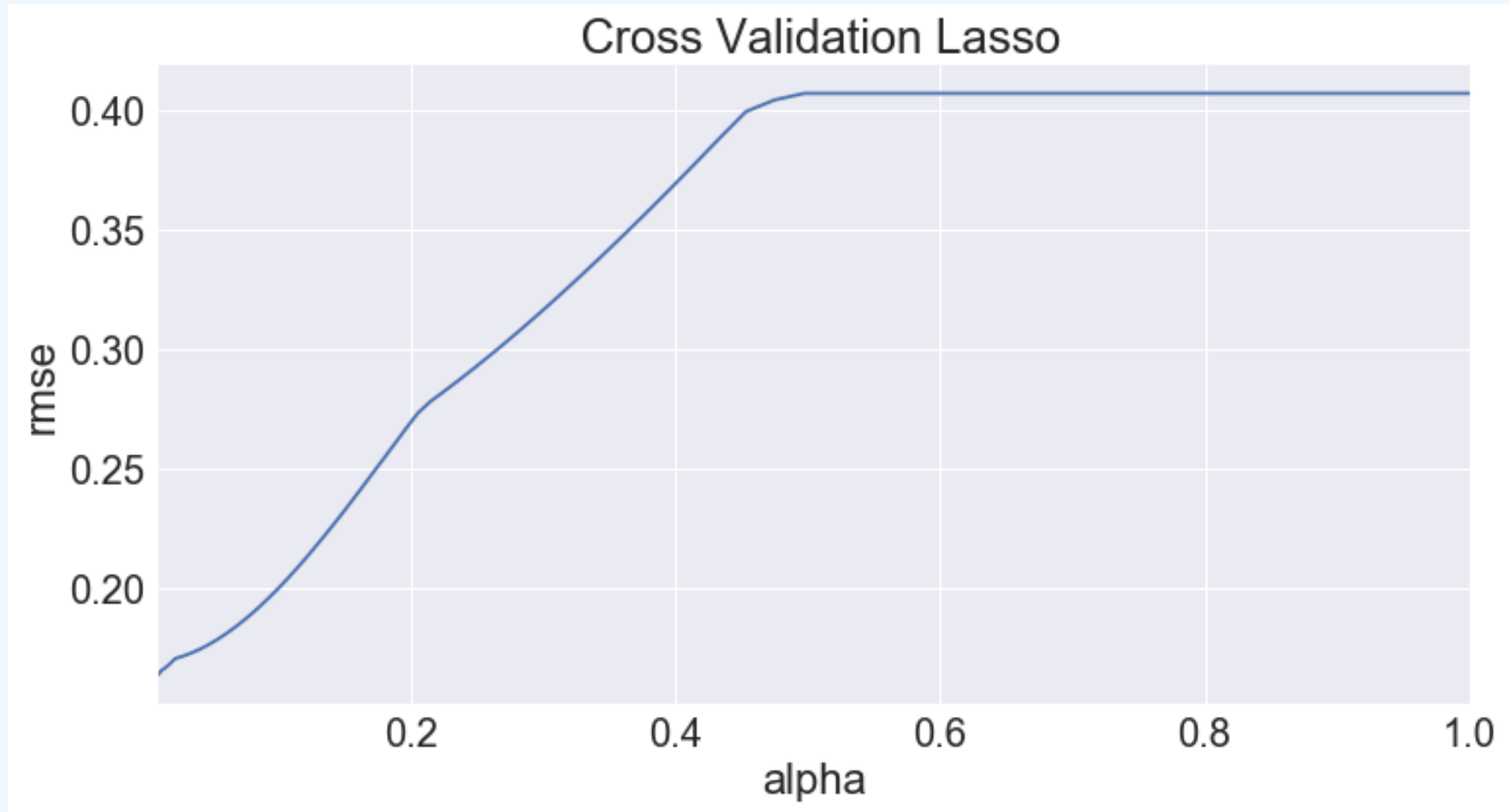
Lasso Regression

	R^2 (Training Set)	RMSE (Training Set)
SUB-TRAINING SET (80%)	0.839	0.163
TEST SET (20%)	0.838	0.137
TOTAL TRAINING SET (100%)	0.843	0.159

Lasso Regression

Top Features	Coefficient Value
Neighborhood_IDOTRR	-0.106394
Neighborhood_OldTown	-0.098231
OverallQual + OverallCond _5+3	-0.088836
Neighborhood_MeadowV	-0.075524
BsmtQual + BsmtCond_Ta+FA	-0.072136
Heating + HeatingQC_GasA+Ex	0.066236
Neighborhood_Crawford	0.068795
TotalFinishedSF	0.075379
OverallQual	0.098205
SaleType+SaleCondtion_New+Partial	0.128627

Tree Based Models

Decision Trees
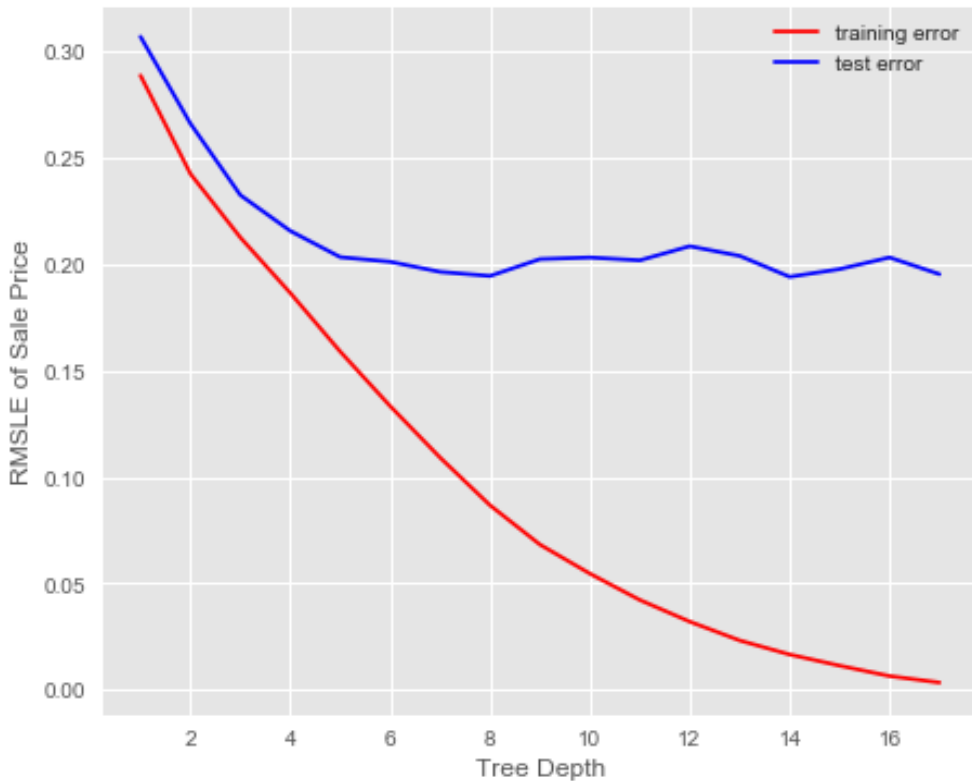
Decision Trees

Decision Trees
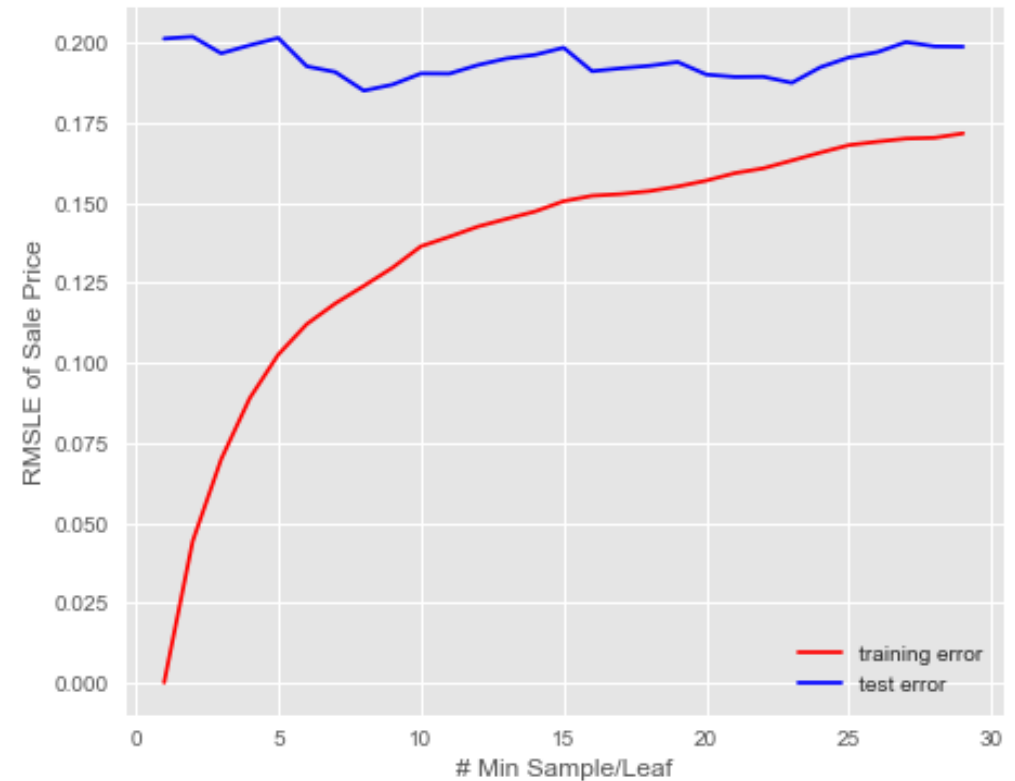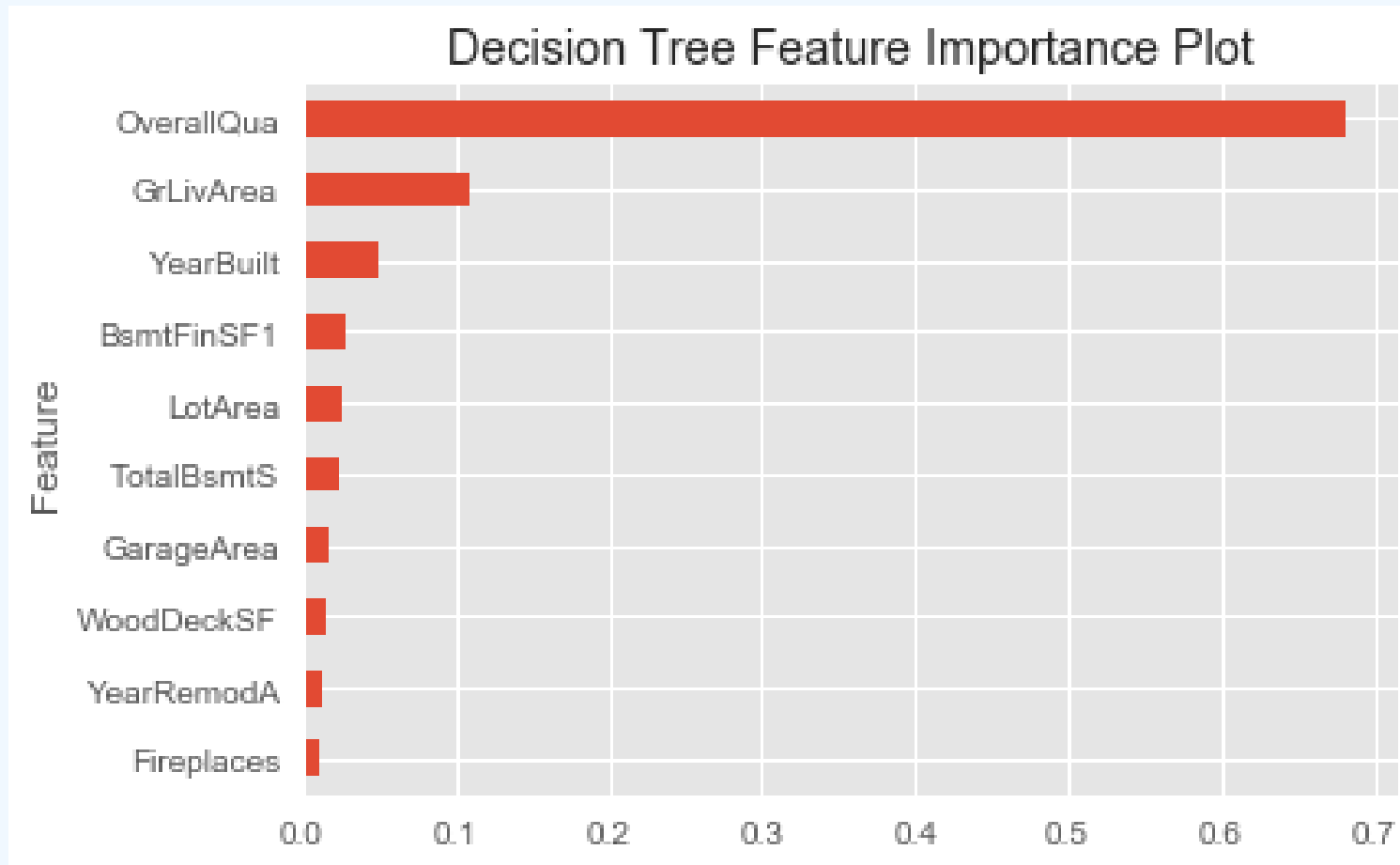
SalePrice Prediction Plots : Decision Trees
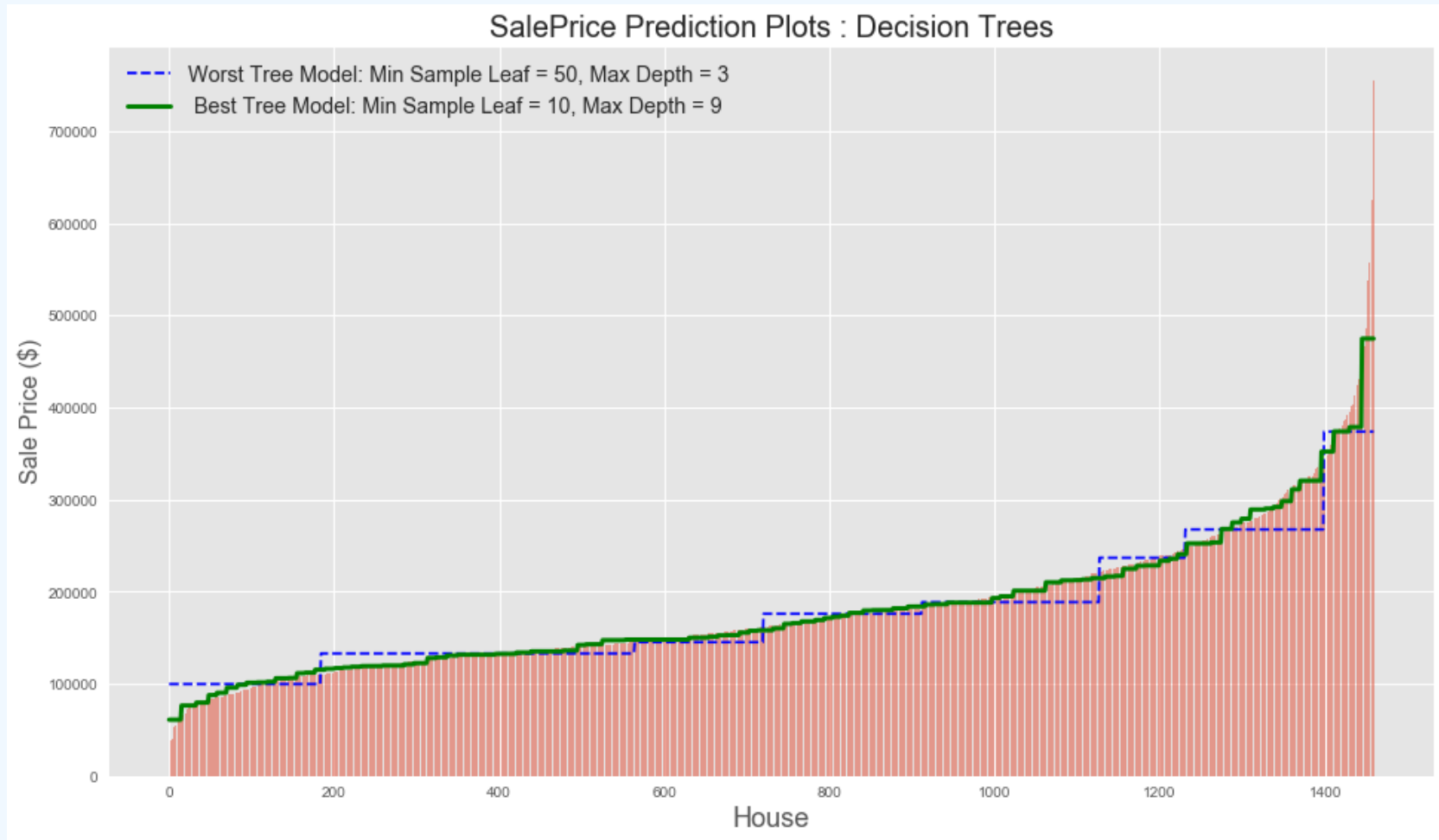
Random Forest

Random Forests

SalePrice Prediction Plots : Random Forests

- - - Worst Random Forests Model: N Estimators = 30
—— Best Random Forests Model: N Estimators = 300
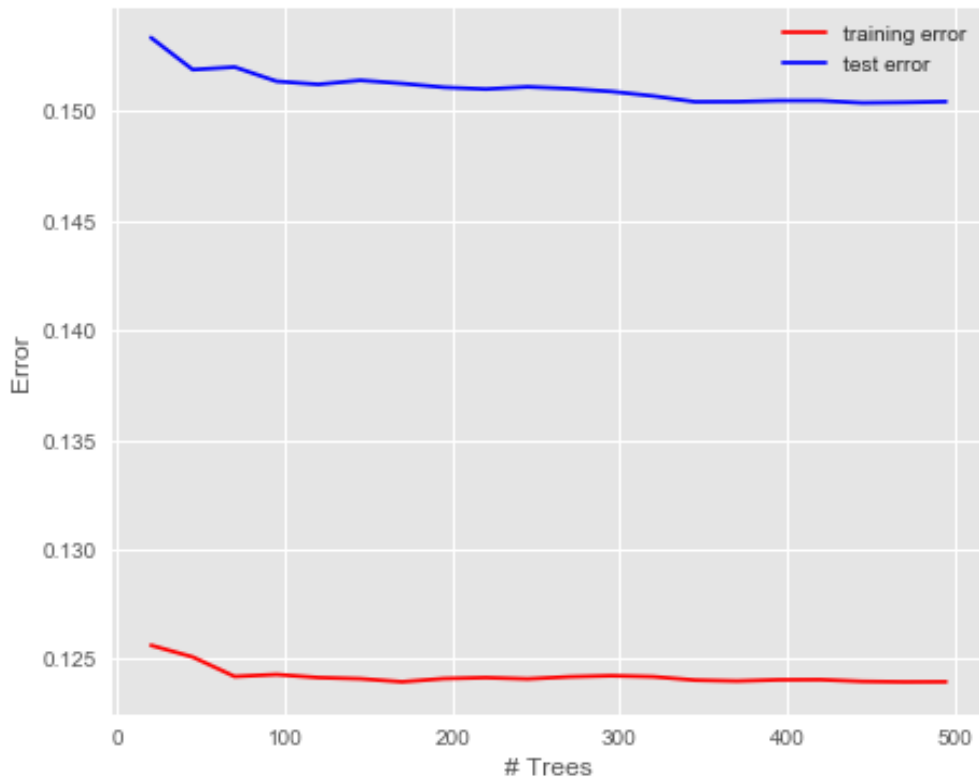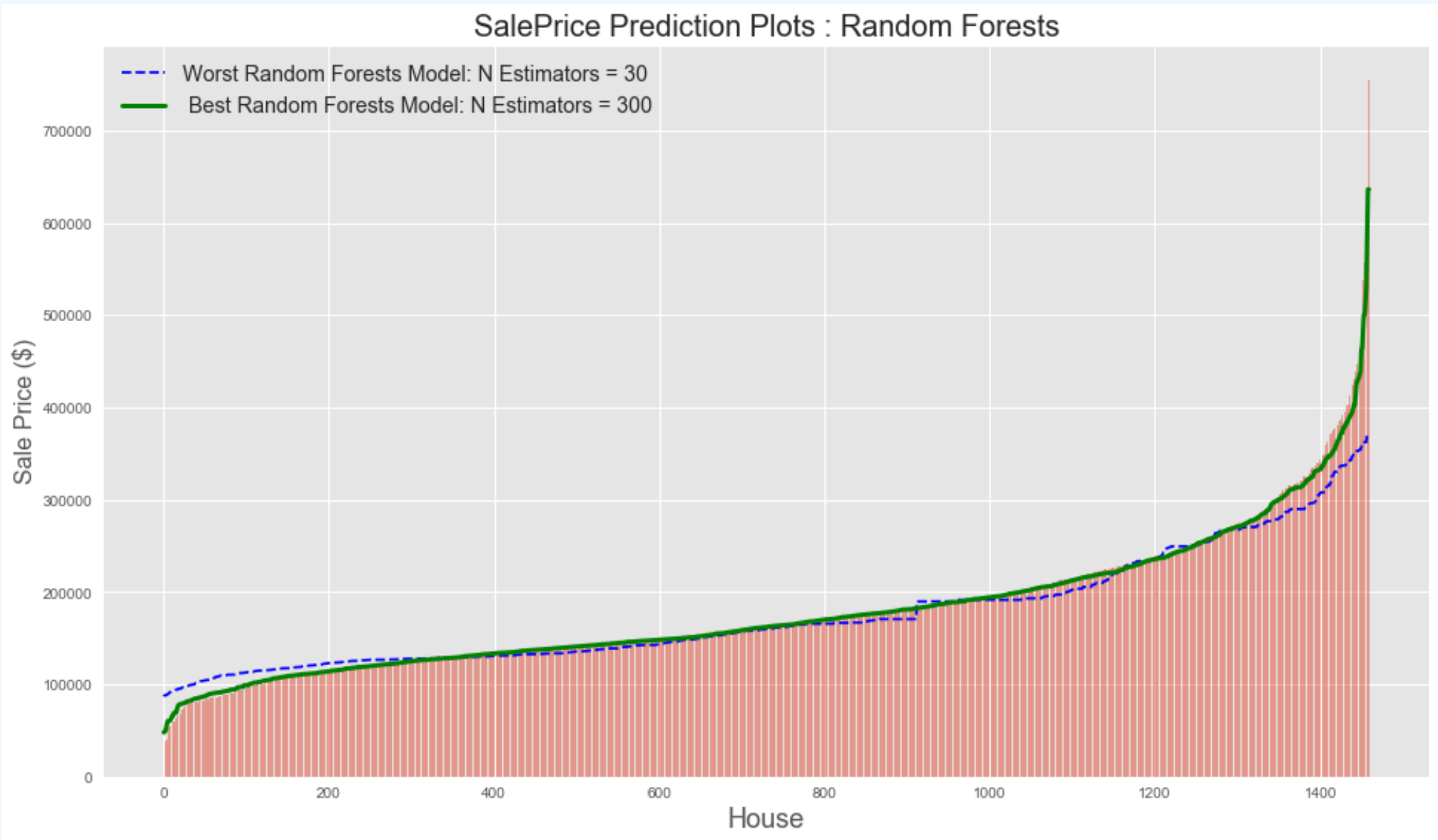
Gradient Boosting

Gradient Boosting

SalePrice Prediction Plots : Gradient Boosting

- - - Worst Gradient Boost Model: N Estimators=10, Learning Rate = 0.2
——— Best Gradient Boost Model: N Estimators = 300, Learning Rate=0.1
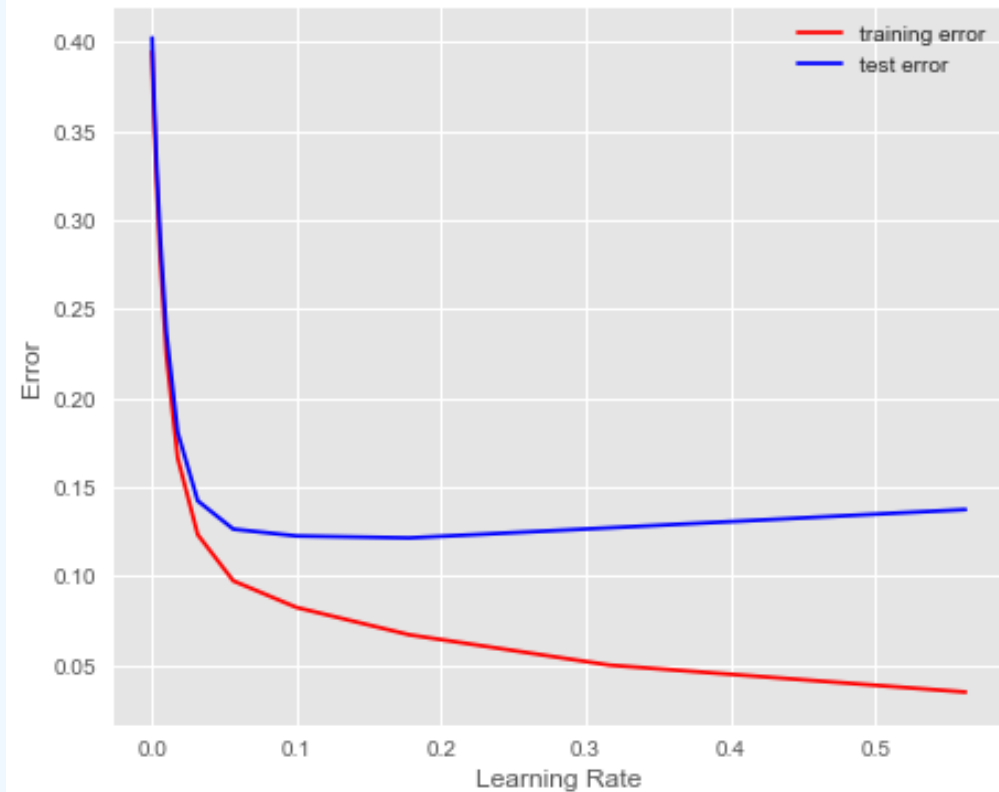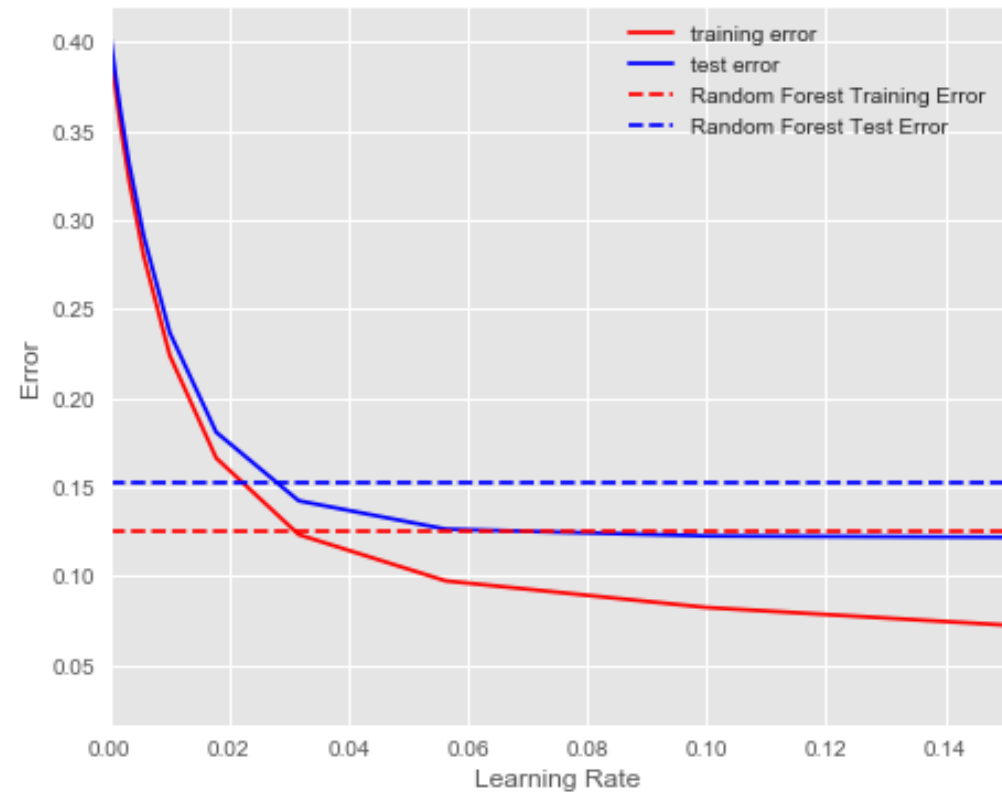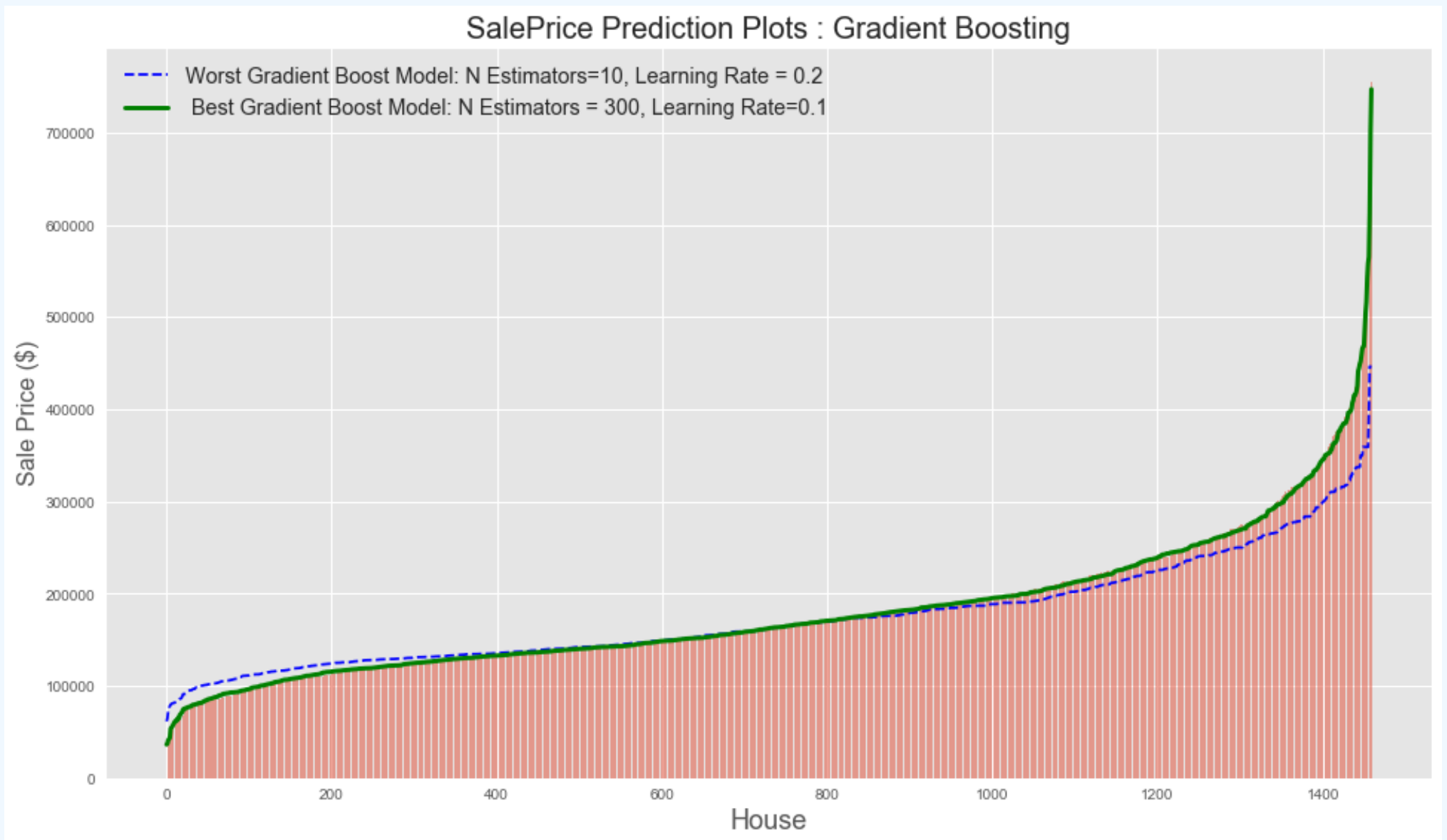
Tree Based Models
CV Grid Search

Models	Hyper Parameters		RMSLE		
	Type	Values	Training	Test	Kaggle
Decision Trees	Max Depth	9	0.1374	0.1910	0.1882
	Min Sample Splits	10			
Random Forests	Min Sample Leaf	2	0.0771	0.1358	0.1465
	N Estimators	300			
Gradient Boosting	Min Sample Splits	50	0.0746	0.1184	0.1245
	Learning Rate	0.06			
	N Estimators	300			

Ensemble Model
Stacking Regressor

Ridge Regression

Lasso Regression

Random Forests
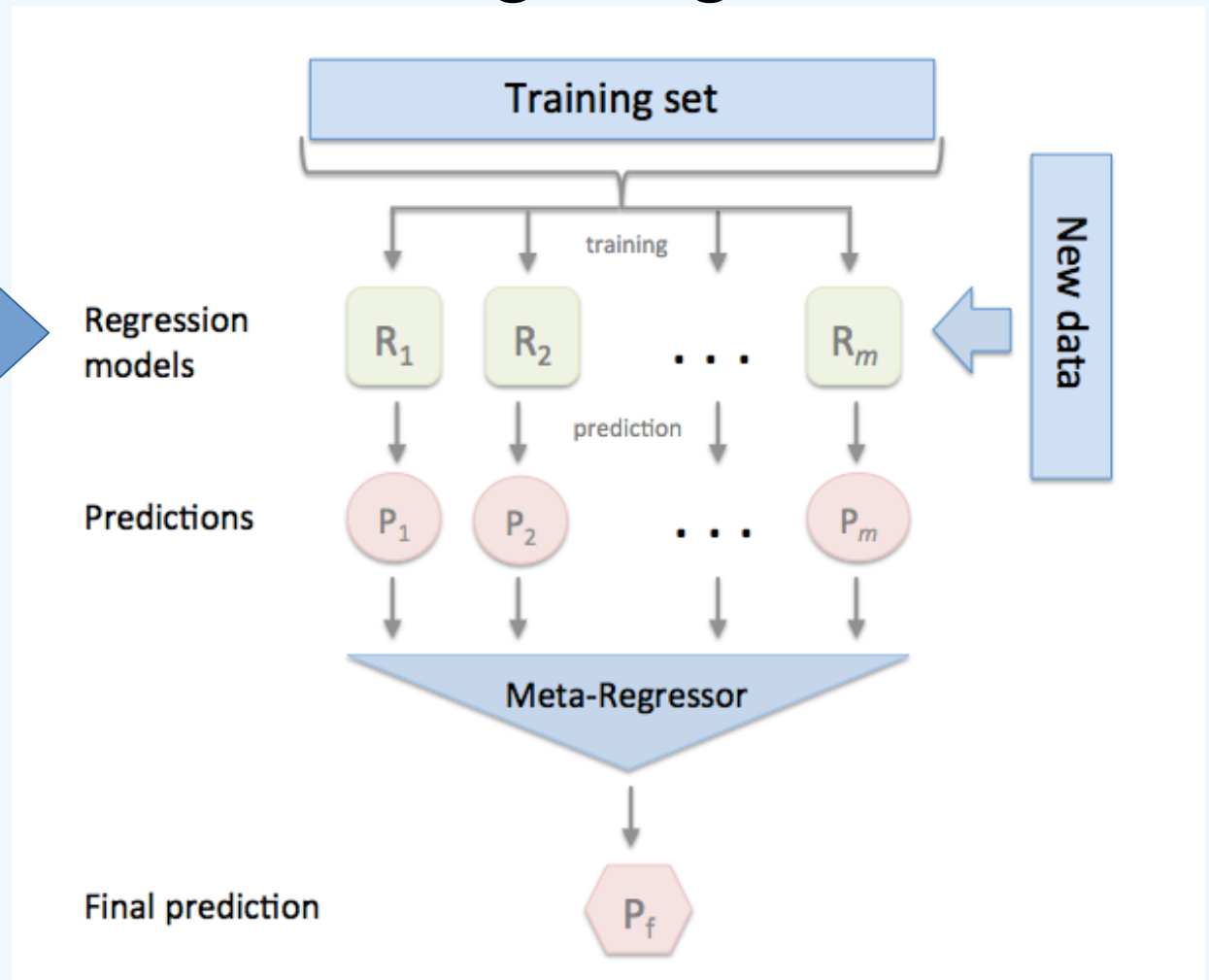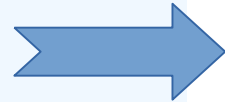
[Package MLXTEND]

Ensemble Model Results

	Models	Meta-Regressor	Alphas	RMSLE	
				Test	Kaggle
Ensemble 1	Lasso Regression Ridge Regression Random Forest	Lasso	0.001	0.0656	0.1492
Ensemble 2	Lasso Regression Ridge Regression	Lasso	0.0001	0.0949	0.1251

Models Results

	Models	RMSLE		
		Training	Test	Kaggle
Linear Based	Multi-Linear	0.1787	0.1641	0.1788
	Ridge Regression	0.1000	0.1110	0.1290
	Lasso Regression	0.1630	0.1370	0.1414
Tree Based	Decision Trees	0.1374	0.1910	0.1882
	Random Forests	0.0771	0.1358	0.1465
	Gradient Boosting	0.0746	0.1184	0.1245
Ensemble	Ensemble 1	0.0010	0.0656	0.1492
	Ensemble 2	0.0001	0.0949	0.1251

Conclusion

- ## Vital Steps

  - Feature Engineering
    - Surprising results with respect to most important features based on coefficients

  - Hyperparameters Tuning

- ## Model Performance:

  1. Ridge Regression

  2. Lasso Regression (within ensembling process)