Predicting House Prices

Mikdad Kanbar - Lisa Verlare - Madelief Bresser

Introduction

Data set: House characteristics and price of residential homes in Ames, Iowa

Research questions:

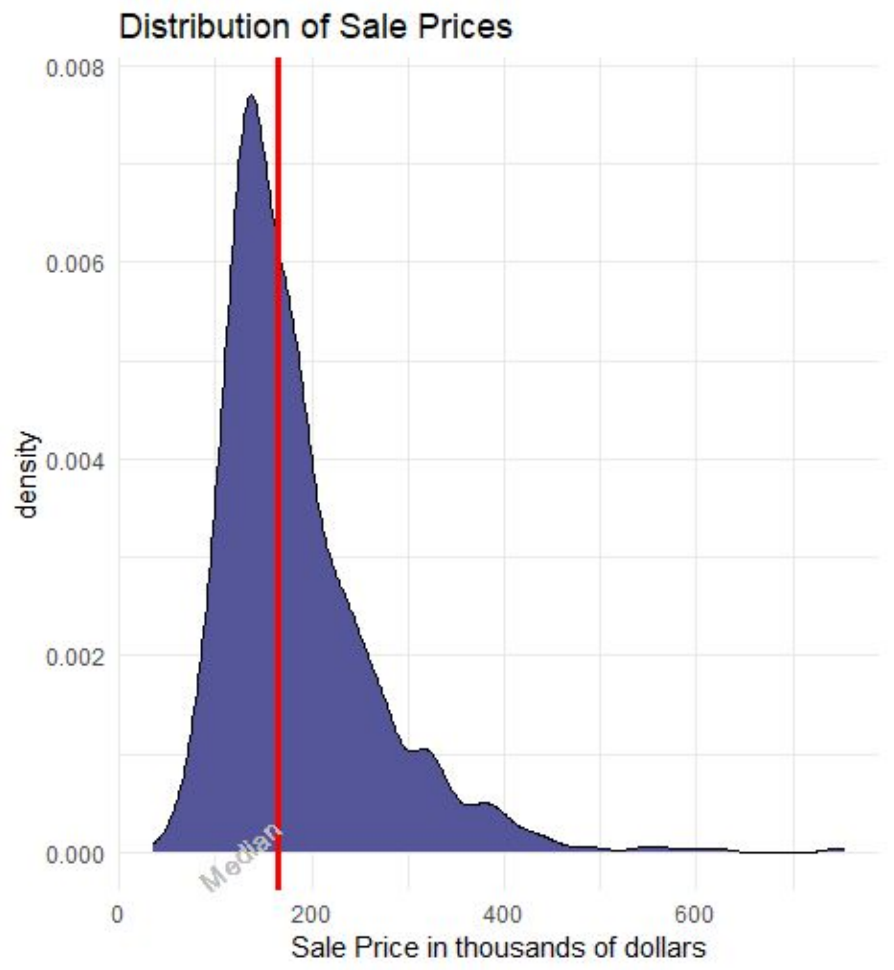
1. *What method is most accurate when predicting house prices based on available house characteristic variables?*
2. *Which house characteristics are most important in determining the sale price?*

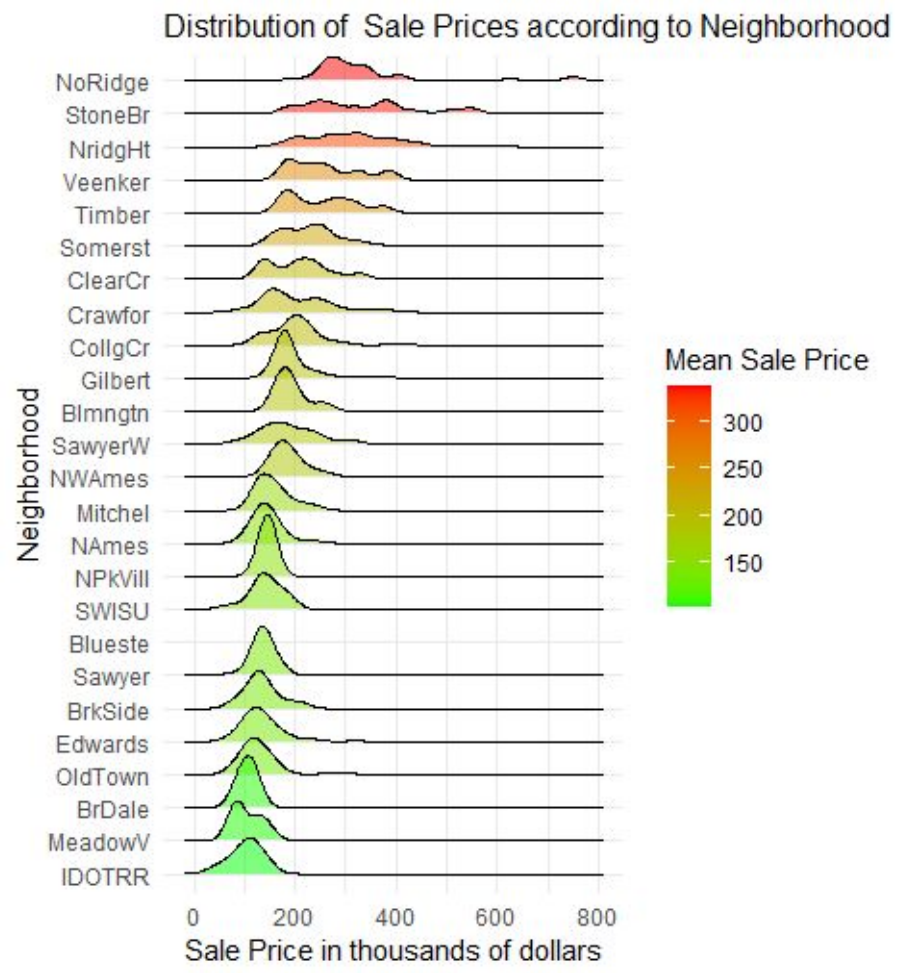
Models:

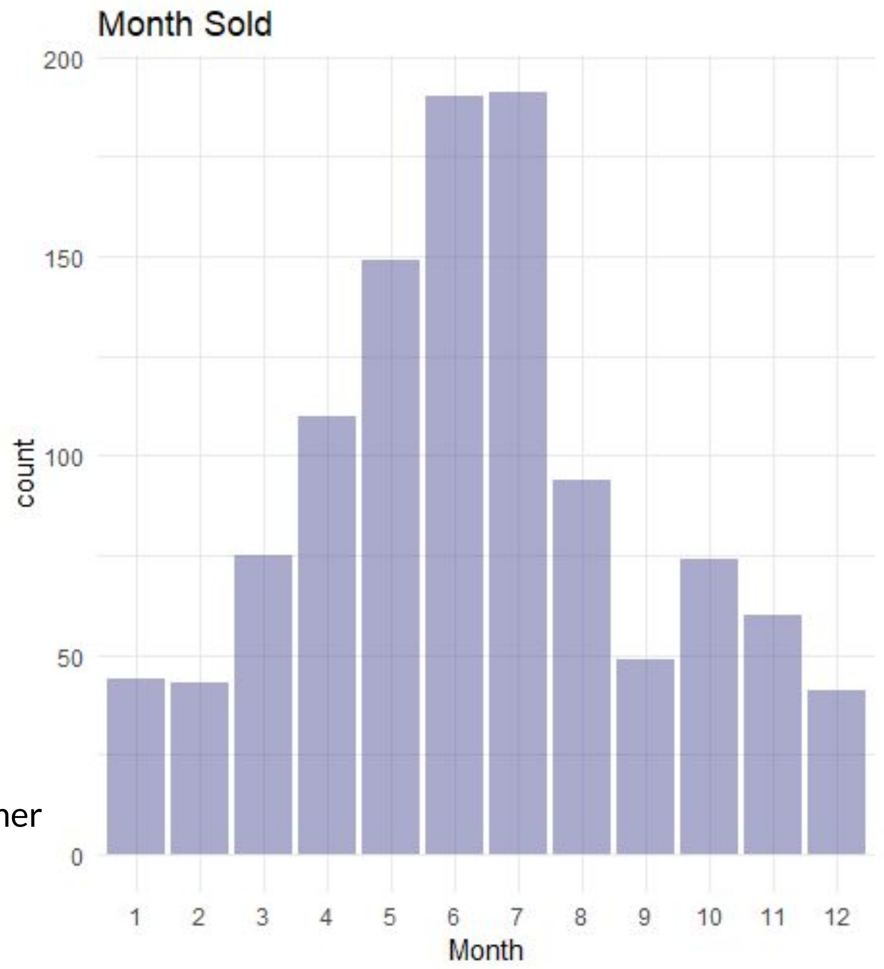
- Lasso (+ Ridge)
- Random Forest
- Gradient Boosting

Data Exploration :

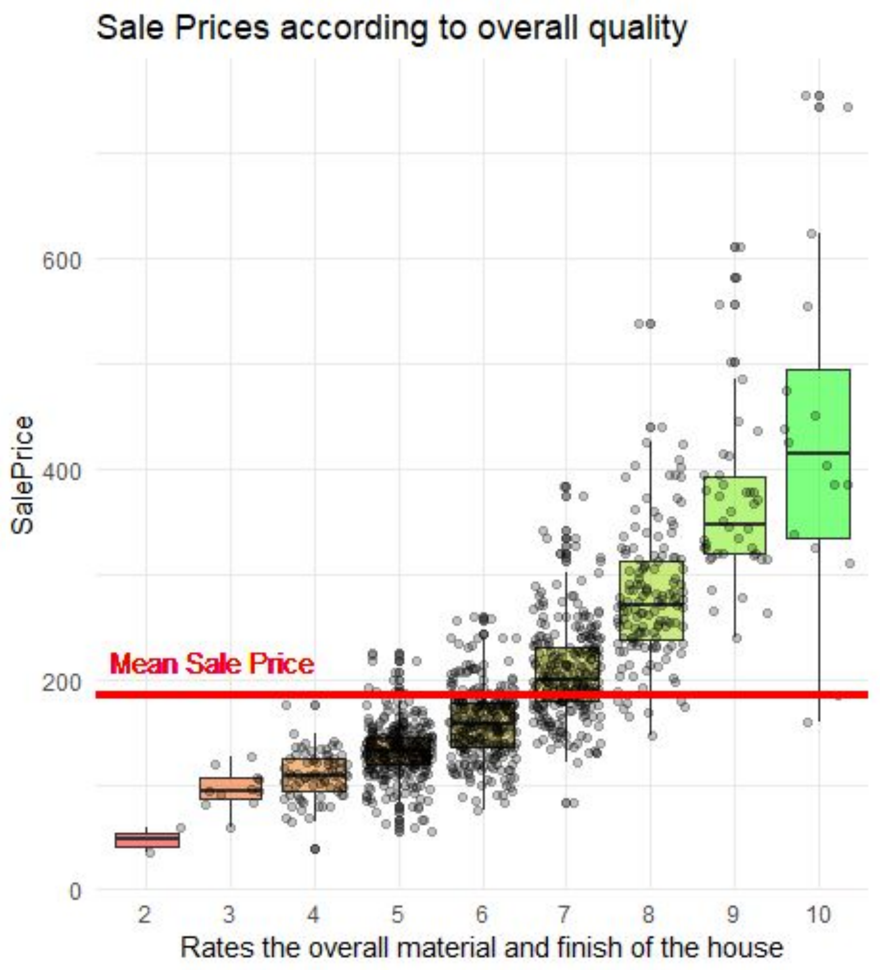
1120	80	~60%
Cases	Predictors	Categorical

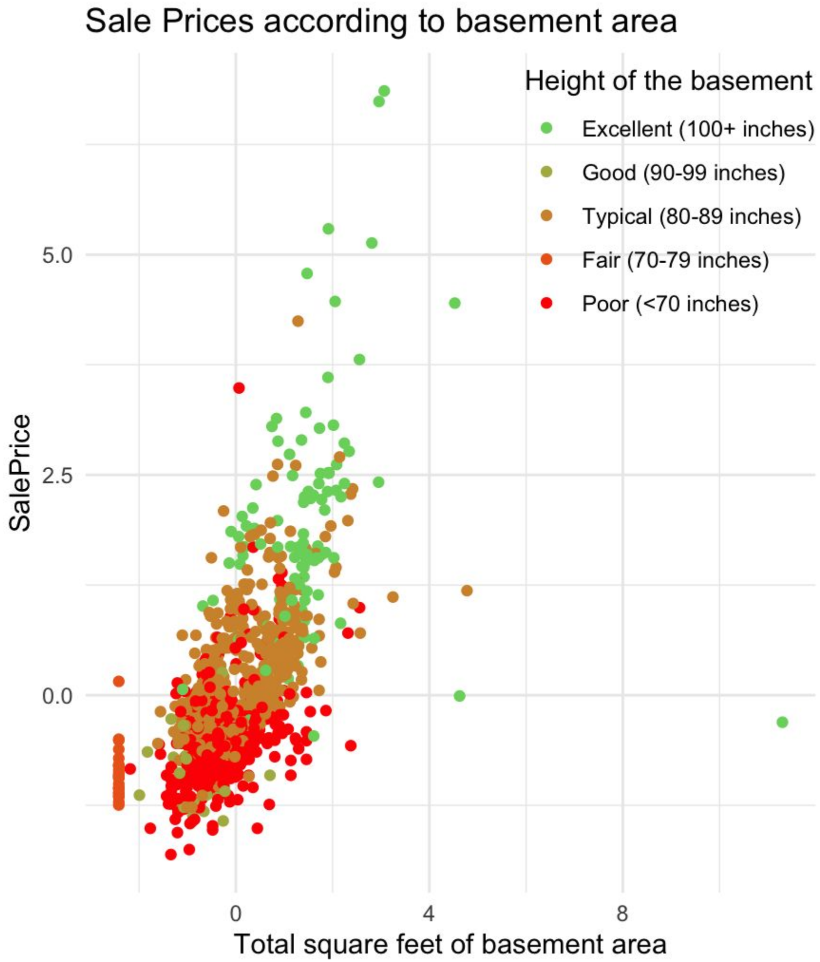

Distribution of Sale Prices
0.008
0.006
0.004
0.002
0.000
density
Median
0
200
400
600
Sale Price in thousands of dollars

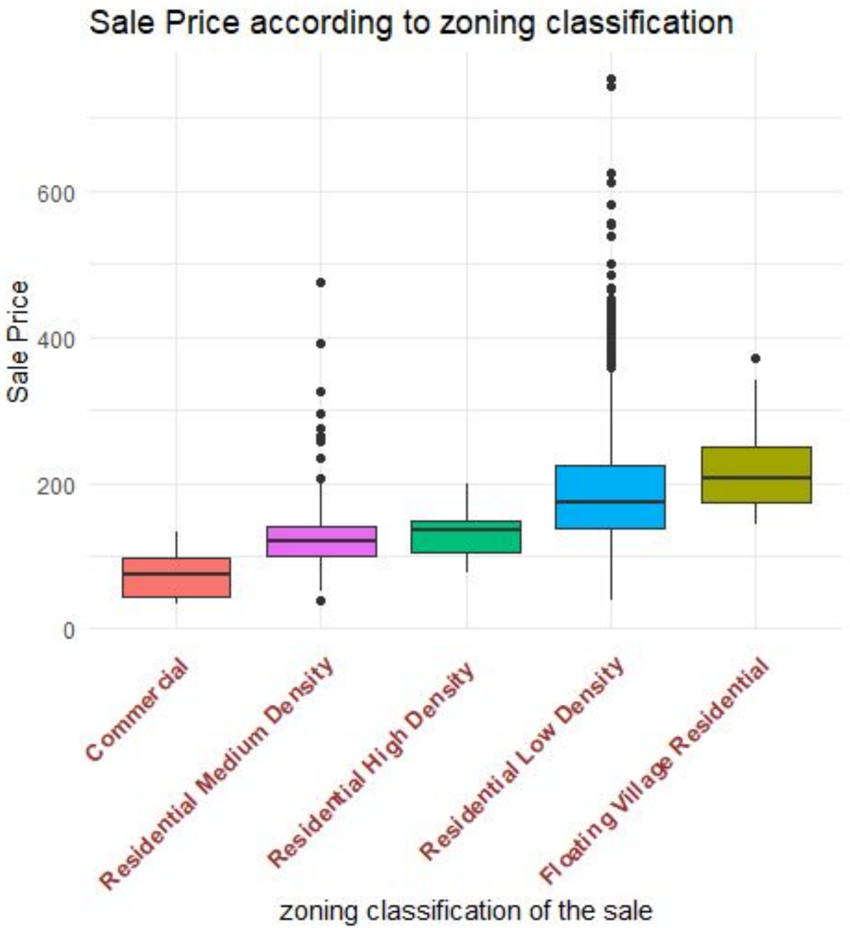

Distribution of Sale Prices according to Neighborhood
NoRidge
StoneBr
NridgHt
Veenker
Timber
Somerst
ClearCr
Crawfor
CollgCr
Gilbert
Blmngtn
SawyerW
NWAmes
Mitchel
NAmes
NPkVill
SWISU
Blueste
Sawyer
BrkSide
Edwards
OldTown
BrDale
MeadowV
IDOTRR
Neighborhood
0
200
400
600
800
Sale Price in thousands of dollars
Mean Sale Price
300
250
200
150



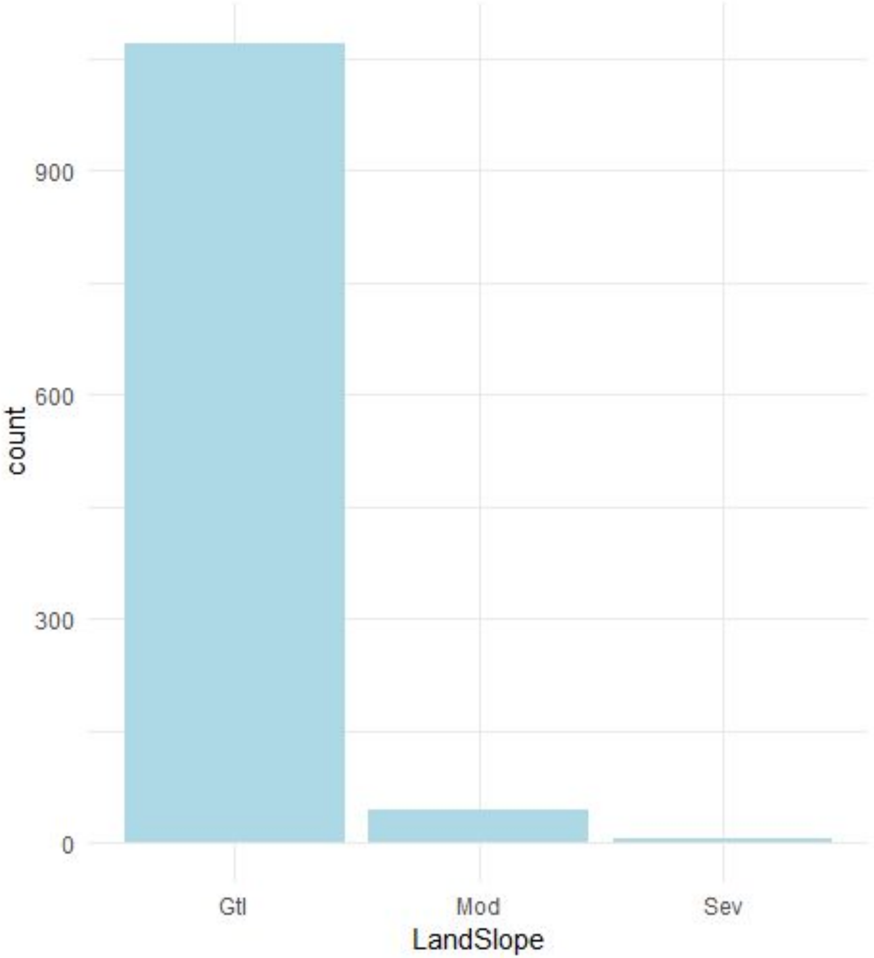
Most Sales occurred in Summer

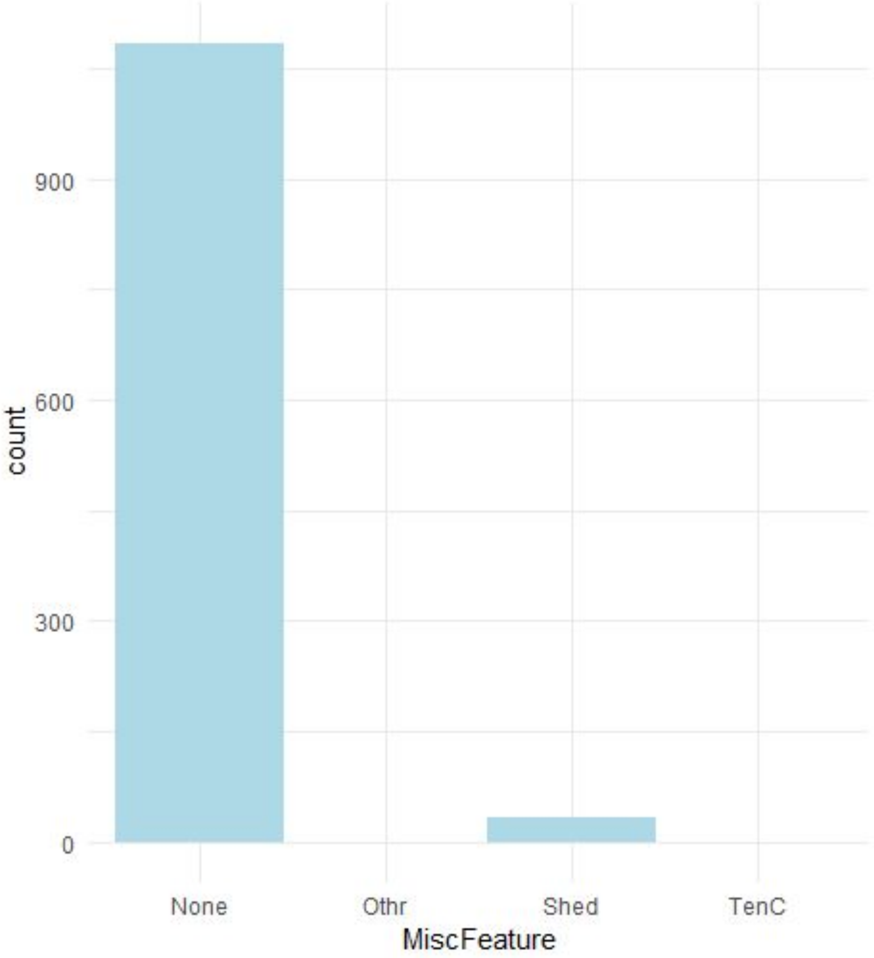

Sale Prices according to overall quality
SalePrice
600
400
200
0
Mean Sale Price
2
3
4
5
6
7
8
9
10
Rates the overall material and finish of the house


Sale Prices according to basement area
Height of the basement
Excellent (100+ inches)
Good (90-99 inches)
Typical (80-89 inches)
Fair (70-79 inches)
Poor (<70 inches)
SalePrice
5.0
2.5
0.0
0
4
8
Total square feet of basement area

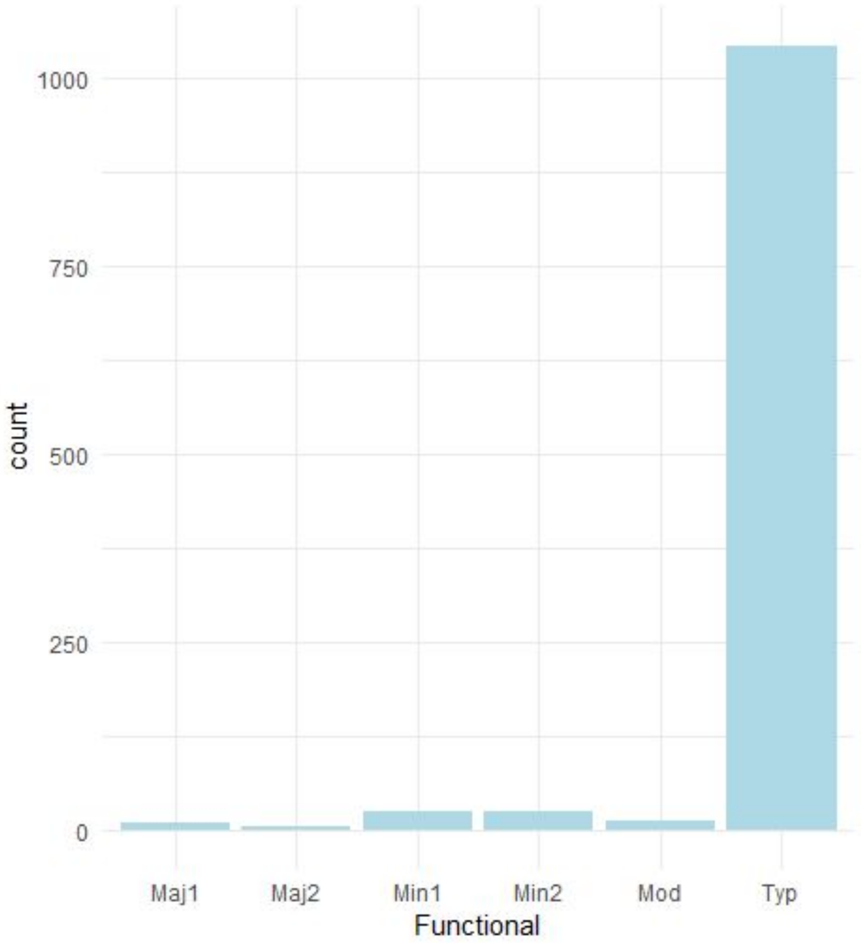

Sale Price according to zoning classification
Sale Price
600
400
200
0
Commercial
Residential Medium Density
Residential High Density
Residential Low Density
Floating Village Residential
zoning classification of the sale

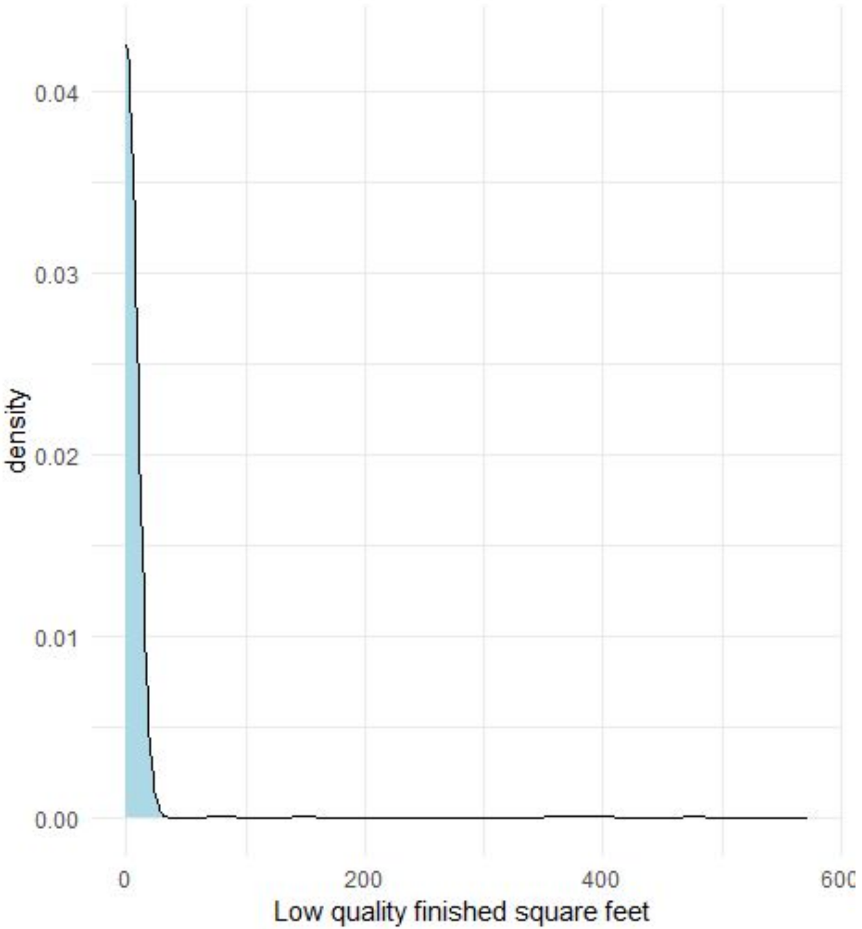
Examples of variables with one dominant category





Examples of variables with one dominant category: We drop the majority of them





Lasso

How does Lasso work?

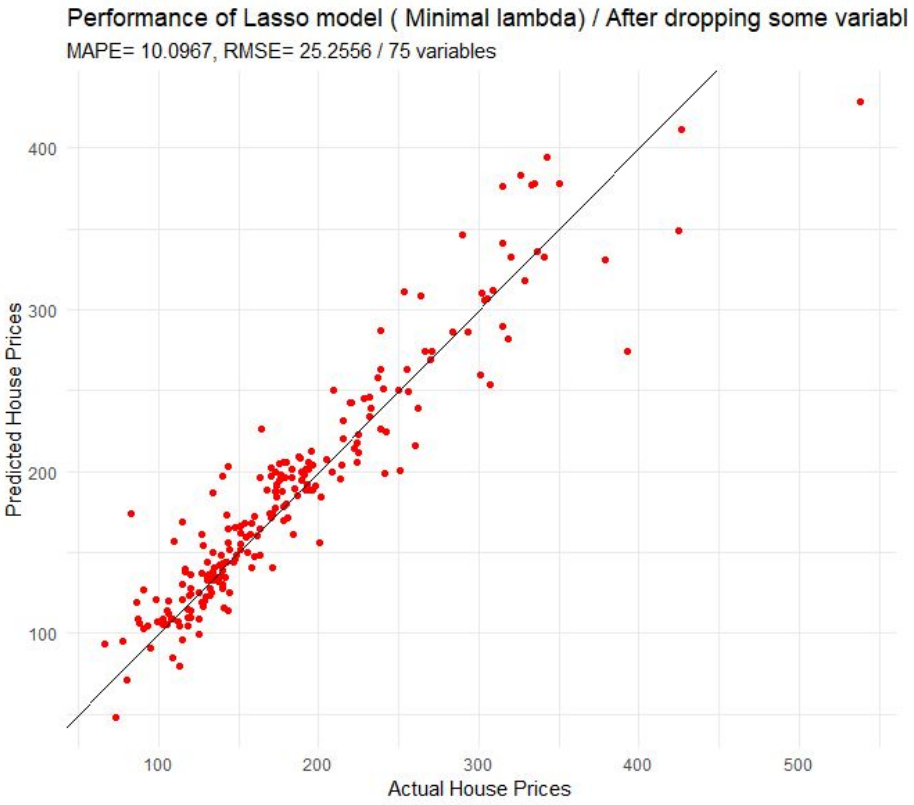
- L1 regularization, which adds a penalty term to the regression equation based on the absolute values of the coefficients.

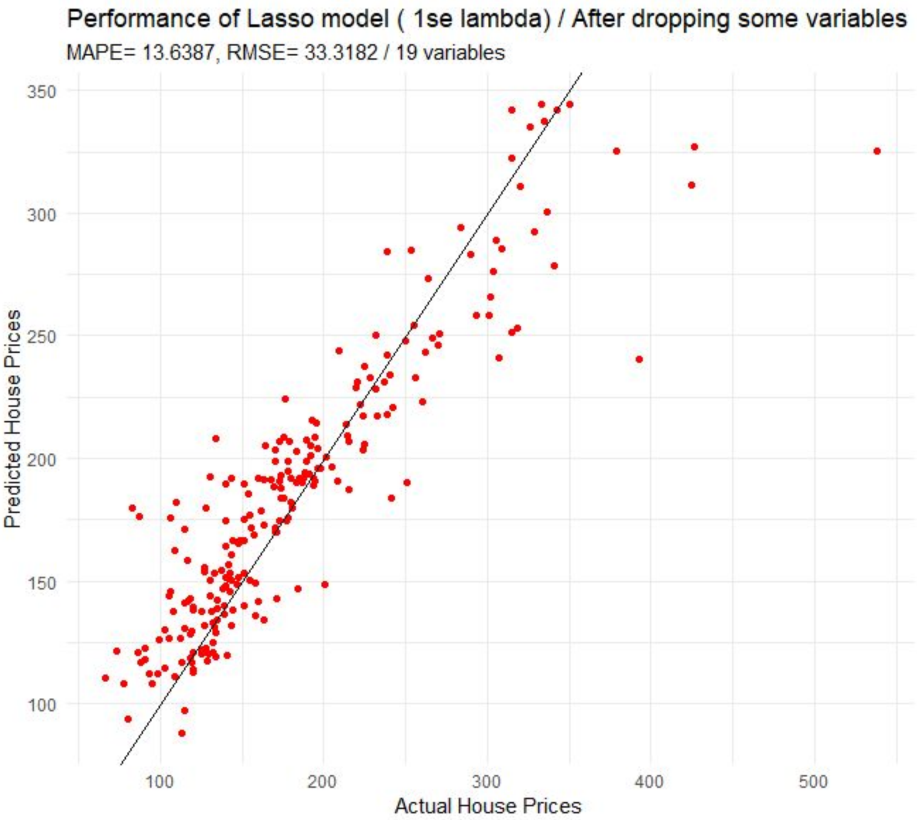
  > Shrinks irrelevant variables to 0

Why Lasso?

- Ability to handle high-dimensional
- Feature selection (by shrinking the coefficients of less relevant features)
- Lasso regression helps to reduce the risk of overfitting.

Comparison using minimal versus 1se lambda for Lasso

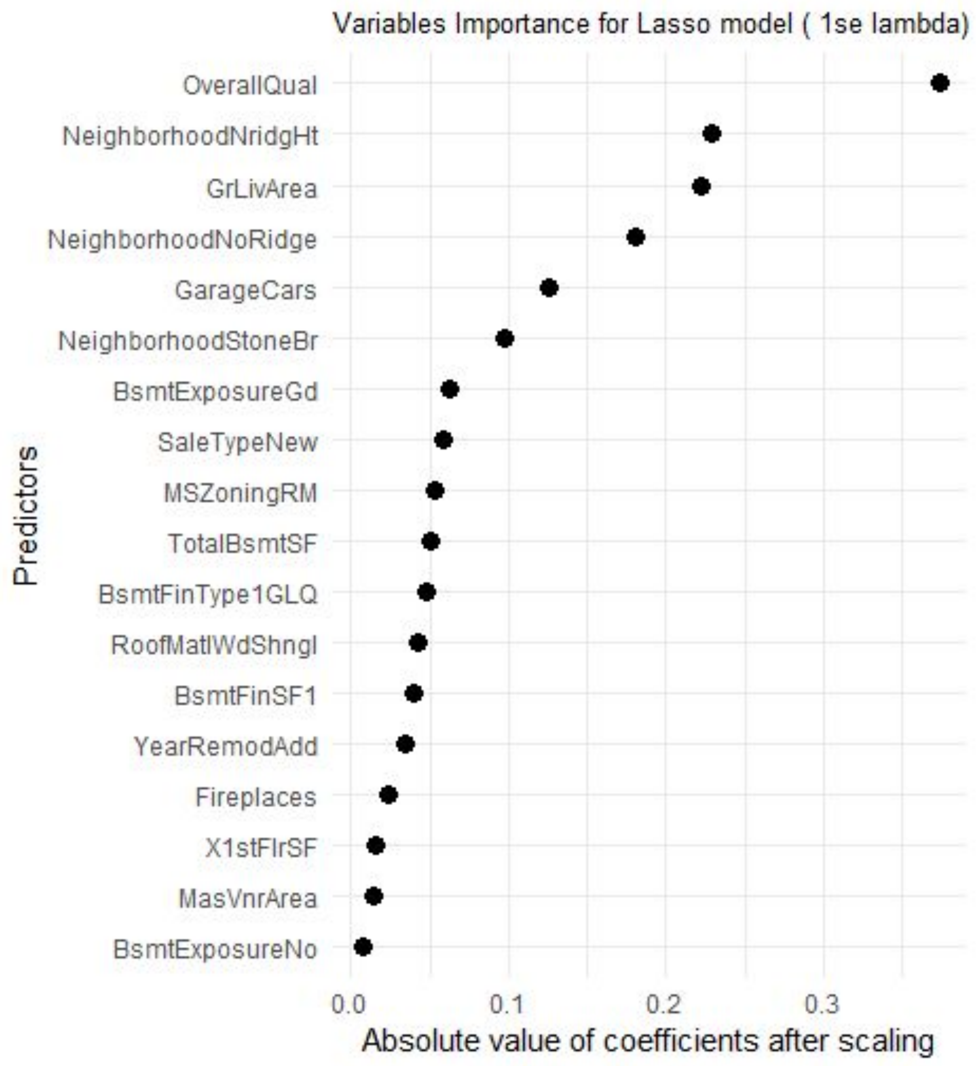




Variable Importance

We scaled the data and retrained the Lasso model. Then we sorted according to the absolute value of coefficients.

- Above ground living area in square feet
- Neighborhood Northridge Heights
- Above ground living area in square feet



Lasso

RMSE comparison: very similar results

Lasso	25.25
Ridge	25.27

Continue with **Lasso**

- Sparse - leave out irrelevant variables
- Improves prediction accuracy

Choose to use **lambda minimal**

- Goal is to maximize prediction accuracy
  - Provides lowest possible error
  - Interpretability is less important

Random Forest Ensemble

How does RF work?

- Ensemble Method that uses bagging.
- Can model nonlinear relationships and interactions.
- Ensemble increases prediction accuracy.

- Multiple decision trees trained on subsets of variables.
- Goal: minimizing information loss and node impurity.

Why Random Forest?

- Works well with complex interactions and non-linear relationships between the variables.
- Works well with high-dimensional data.

Random forest parameters

For stopping criterion:

- Minimum node size, we used 1
- Maximum depth of decision tree, we used default of random forest package in R

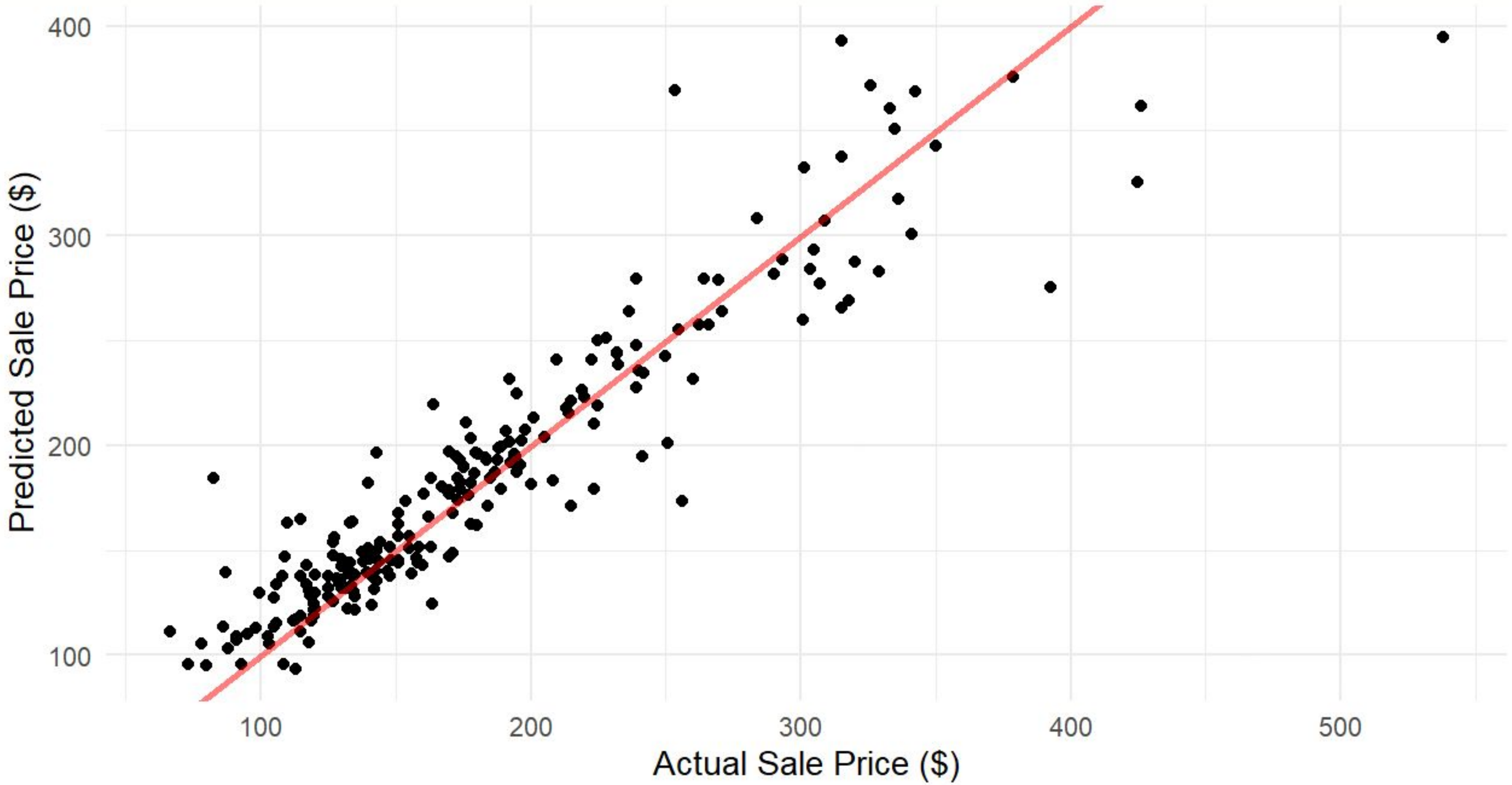
For feature selection:

- Amount of random features selected for each decision tree, we analysed a range of 1:40 and decided to go with 35.

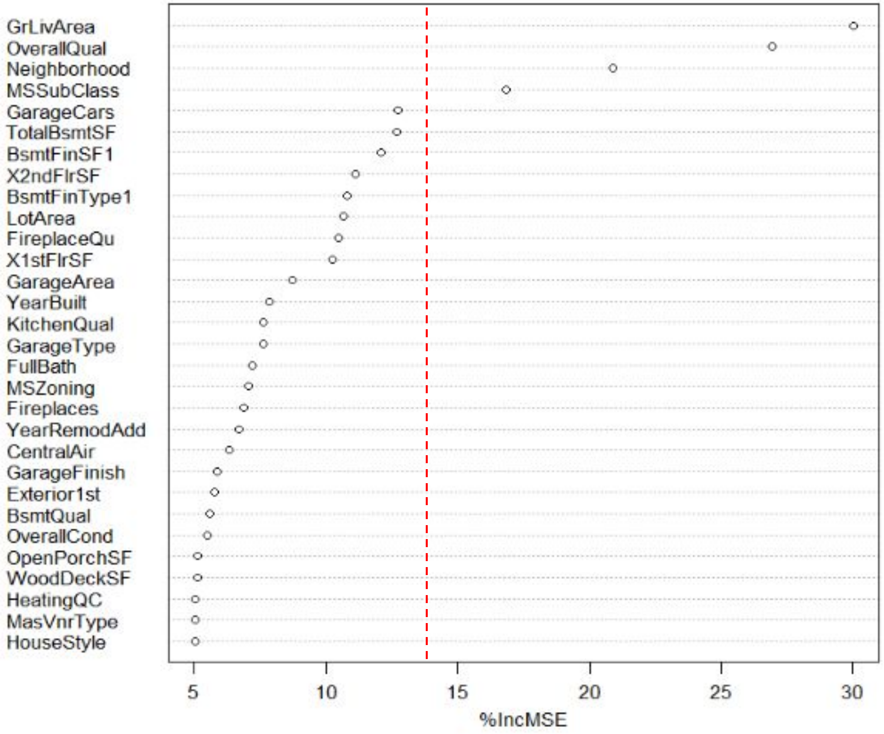
Results

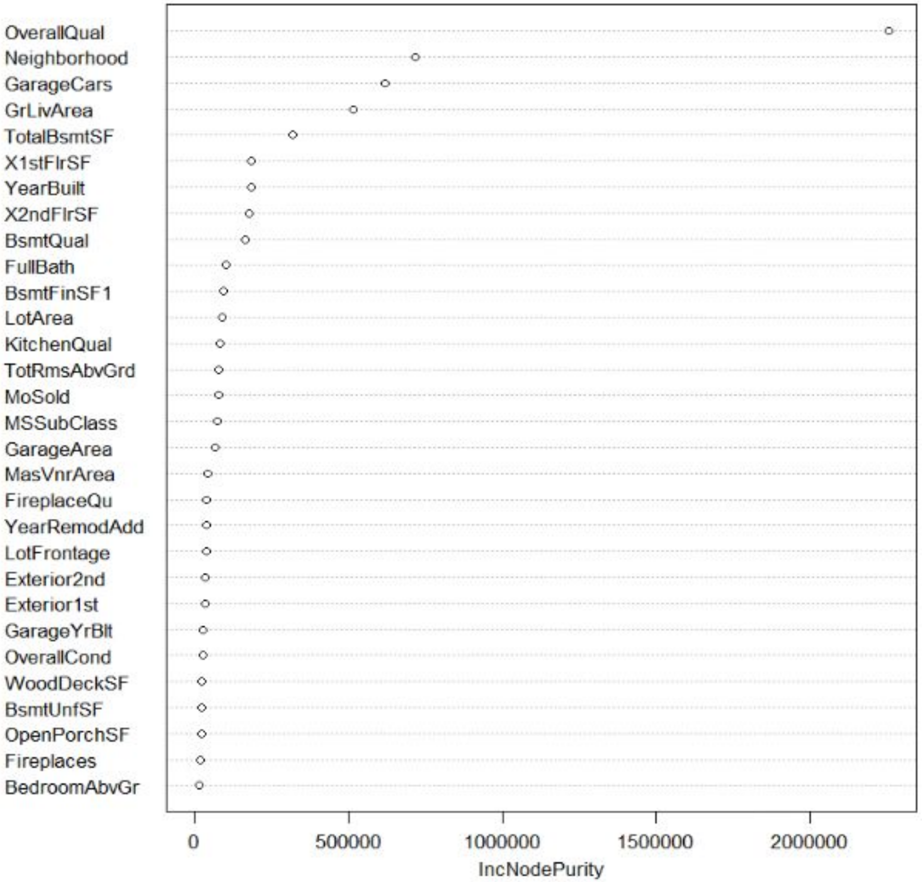
- RMSE: 27.68984

Predicted vs. true sale prices



Variable Importances for the Random Forest



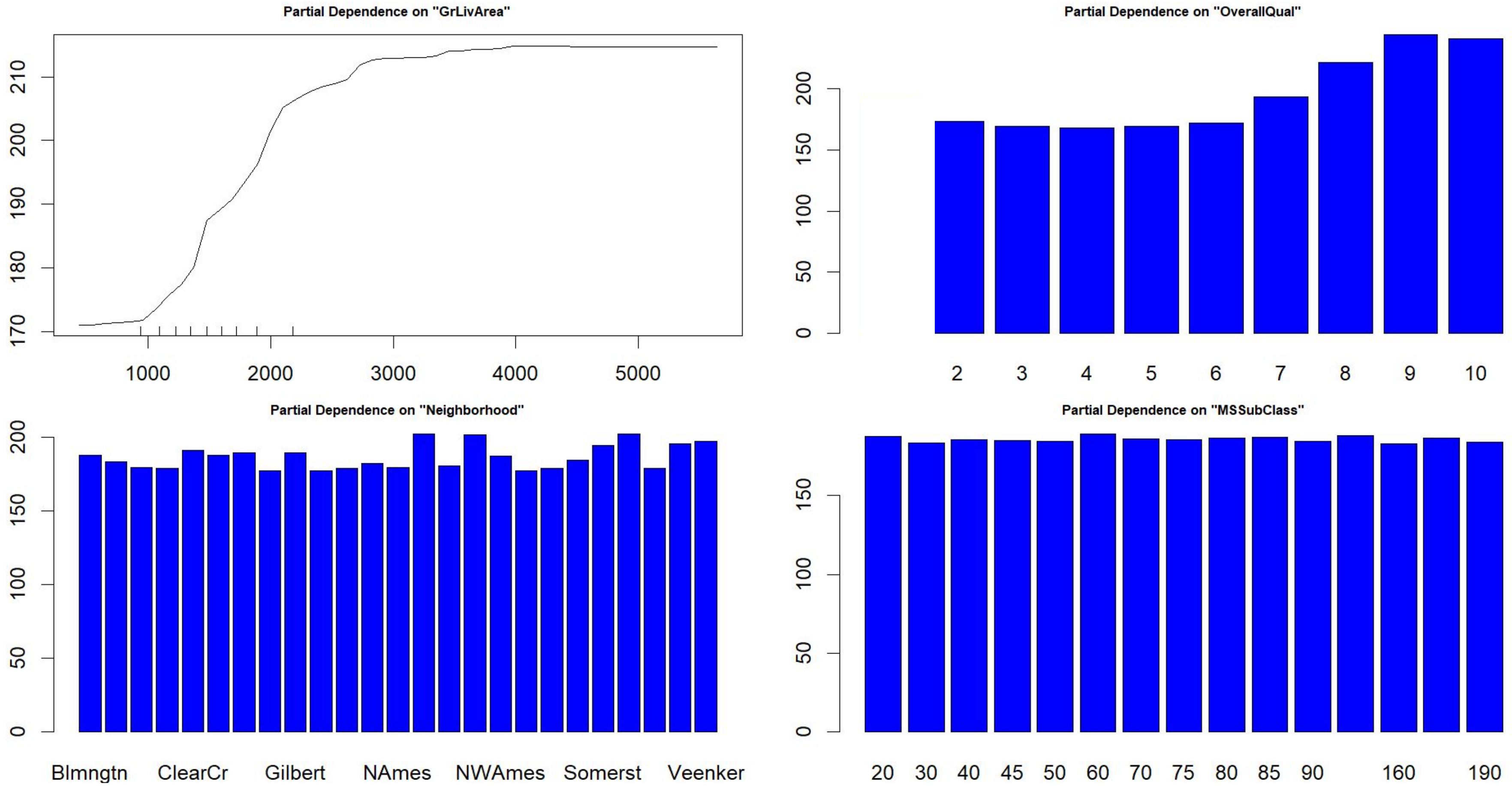


Most important values in Random Forest

Mostly size of areas

Variable	*Importance Score*
Above ground living area in square feet	30.017
Rating of overall material and finish of the house	26.951
Neighborhood	20.891
Type of dwelling/house	16.840
Car capacity of garage	12.737
Basement size in square feet	12.713
Finished basement square feet	12.093
Second floor size in square feet	11.127

Partial Dependence Plots



Gradient Boosted Ensemble

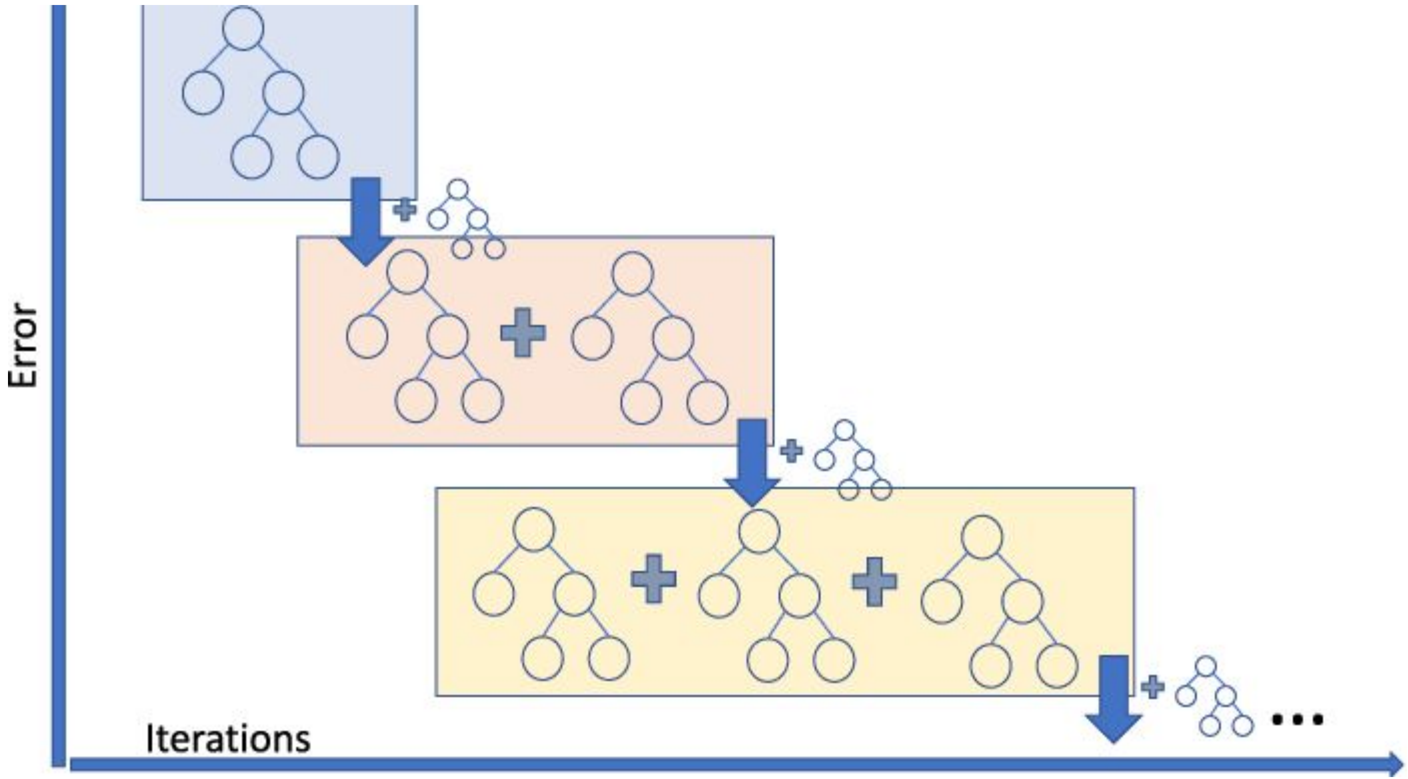
How does GBM work?

- Ensemble Method
- Ensemble increases prediction accuracy
- Models built **sequentially**
- Goal: minimize squared error loss
- Fix residuals in each sequential step

Why Gradient Boosting?

- Trees built on previous tree to correct errors
- Can model nonlinear relationships and interactions; do not assume relationship

Accuracy will improve



Source: https://www.researchgate.net/figure/Schematical-representation-of-gradient-boosting-regression-in-regards-to-algorithm_fig3_340524896

Gradient Boosted Ensemble

How does GBM work?

- Ensemble Method
- Can model nonlinear relationships and interactions
- Ensemble increases prediction accuracy

- Models built **sequentially**
- Goal: minimize squared error loss
- Fix residuals in each sequential step

Why Gradient Boosting?

- Trees built on previous tree to correct errors
- Do not assume relationship

Accuracy will improve

Tuning Parameters

❗ Overfitting could be an issue ❗

- 10-fold cross validation
- Tuning parameters:
  - **Shrinkage:** learning rate
    - Slower = better
  - **Number of trees** generated
  - **Interaction depth:** the maximum tree depth
    - How many splits per node?
    - Overfitting is possible
  - **Minimum number of observations**
    - Too little data can mean relationships are modeled that don't exist

Shrinkage: 0.001, 0.005, 0.01, 0.05, 0.1

Number of trees: 10, 100, 500, 1000

Interaction depth: 1, 2, 3, 4

Minimum number of observations (stopping criteria): 10

Tuning Parameters

❗ Overfitting could be an issue ❗

- 10-fold cross validation
- Tuning parameters:
  - **Shrinkage:** learning rate
    - Slower = better
  - **Number of trees** generated
  - **Interaction depth:** the maximum tree depth
    - How many splits per node?
    - Overfitting is possible
  - **Minimum number of observations**
    - Too little data can mean relationships are modeled that don't exist

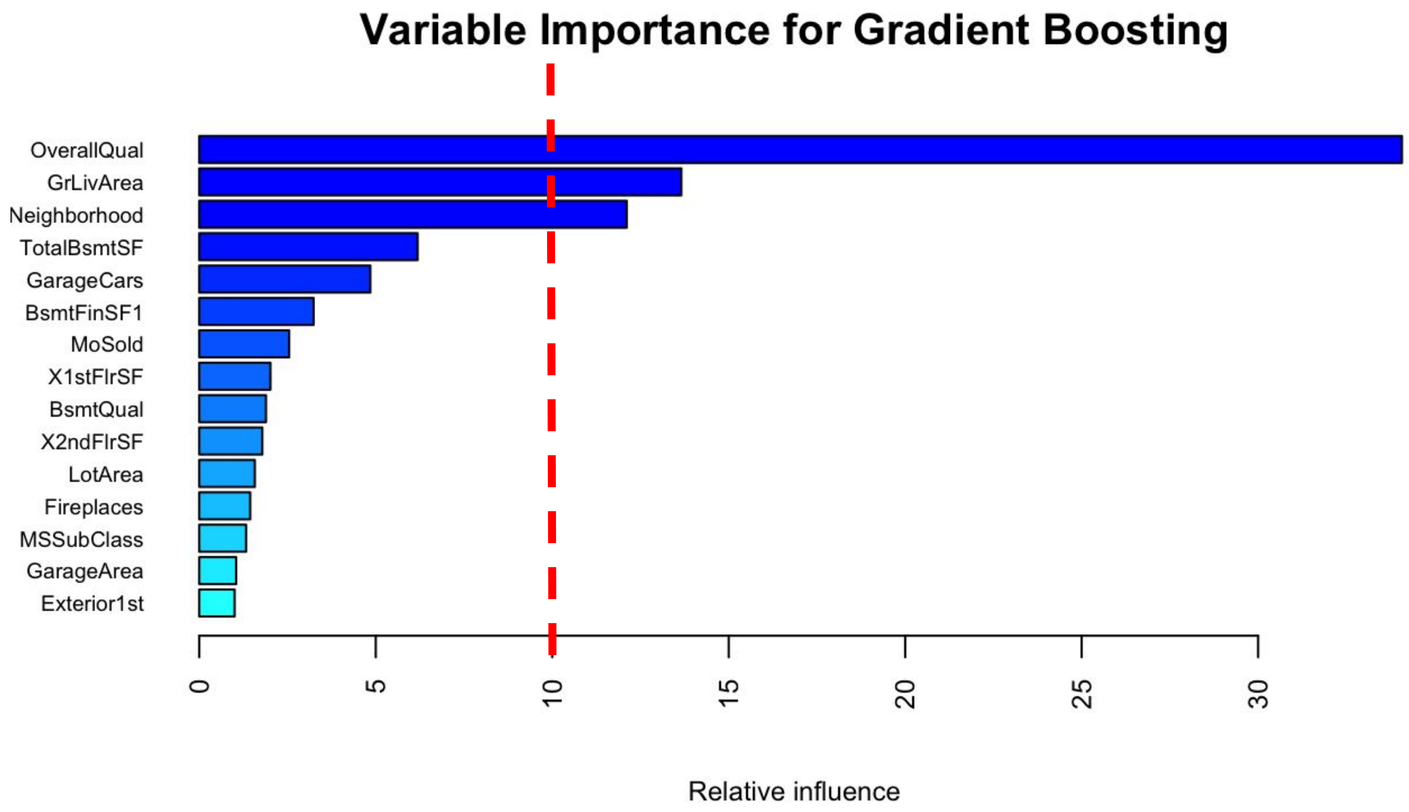
Shrinkage: 0.001, 0.005, 0.01, 0.05, **0.1**

Number of trees: 10, 100, 500, **1000**

Interaction depth: 1, 2, **3**, 4

Minimum number of observations: **10**

RMSE = 28.09


Variable Importance for Gradient Boosting
OverallQual
GrLivArea
Neighborhood
TotalBsmtSF
GarageCars
BsmtFinSF1
MoSold
X1stFlrSF
BsmtQual
X2ndFlrSF
LotArea
Fireplaces
MSSubClass
GarageArea
Exterior1st
0
5
10
15
20
25
30
Relative influence

Variable Importance

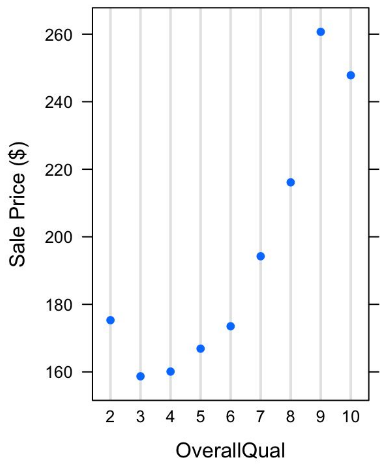
Most important factors to determine Sale Price:

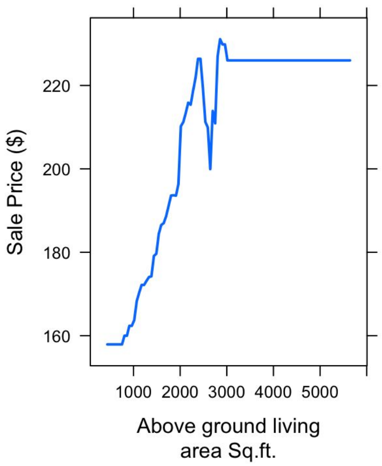
- Quality of the House
- Neighbourhoods
- Square feet of the house

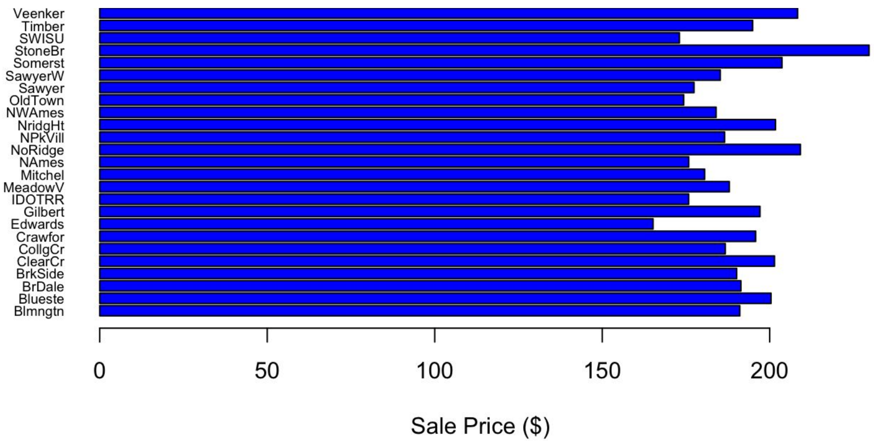
Variable	Relative Influence (%)
Rating of overall material and finish of the house	34.07
Above grade (ground) living area square feet	13.66
Physical locations within Ames city limits	12.11
Total square feet of basement area	6.19
Size of garage in car capacity	4.85
Type 1 finished square feet	3.24
Month Sold	2.55
First floor square feet	2.016318430

Partial Dependence Plots

- Higher quality increases Sale Price
  ** For Quality Rate = 2 only 2 observations are available, which explains the higher effect on the Sale Price*
- Larger square footage increases the Sale Price
- Some neighborhoods will increase the Sale Price more than other neighborhoods







Conclusion

Which model is most accurate in its predictions?

Lasso

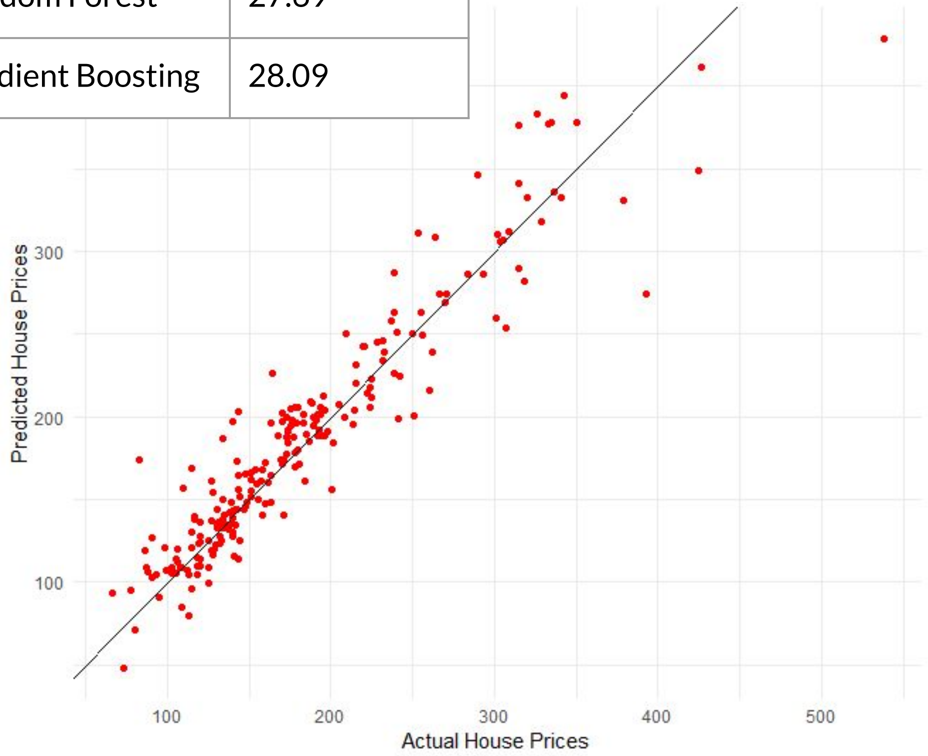
- Many data points = good prediction
- Few data points = bad prediction

Why?

- Feature selection
- Linear relationship
- Random forest and GBM slightly too complex

RMSEs could improve with more data

	RMSE
Lasso	**25.25**
Ridge	25.27
Random Forest	27.69
Gradient Boosting	28.09



Comparison

Which variables are most important to the Sale Price?

Lasso:

1. Rating of overall material and finish of the house
2. Neighborhood Northridge Heights
3. Above ground living area in square feet

---

4. Neighborhood Northridge
5. Car **capacity** of garage
6. Neighborhood Stone Brook
7. Basement: Good Exposure
8. Type of Sale: Home just constructed and sold

Random Forest

1. Above ground living area in square feet
2. Rating of overall material and finish of the house
3. Neighborhood

---

4. Type of dwelling/house
5. Car **capacity** of garage
6. Basement size in **square feet**
7. Finished basement **square feet**
8. Second floor size in **square feet**

Gradient Boosting

1. Rating of overall material and finish of the house
2. Above ground living area in square feet
3. Neighborhoods

---

4. Total **square feet** of basement area
5. Car **capacity** of garage
6. Type 1 finished **square feet**
7. Month sold
8. First floor **square feet**

Discussion & Challenges

Challenges

- Training grids yield high computational time.
- Variables with low variance.
- Missing values.

Suggestions

- Testing larger grids.
- Imputation techniques.
- Larger and more diverse data set.
- Unsupervised learning methods.