Predicting Housing Prices in Ames Iowa through Machine Learning Techniques

Rajesh Arasada, Yung Chou, Nilesh Patel,
Pankaj Sharma, Tim Waterman

Purpose

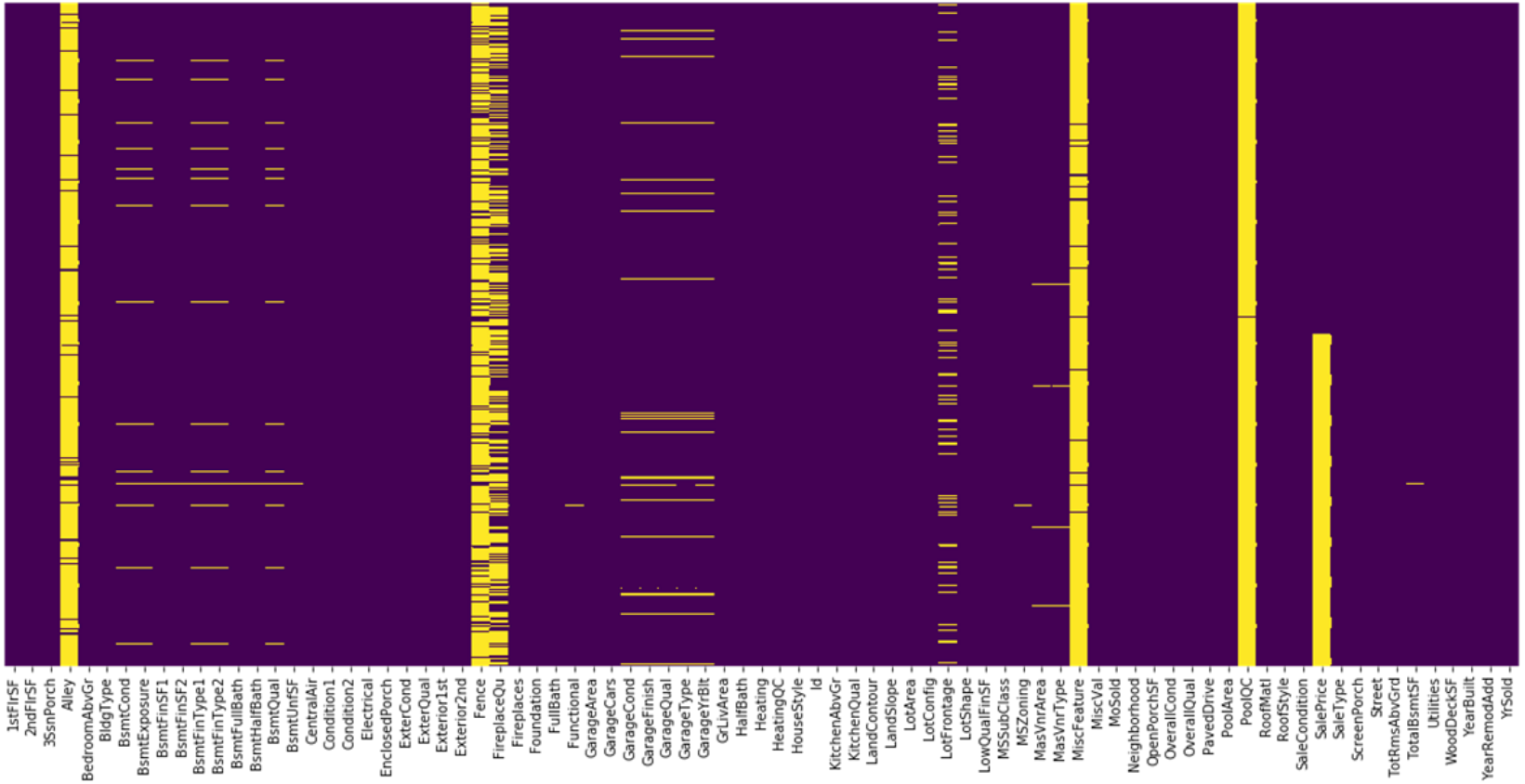
For Industry: Understand which housing aspects are valued by consumers, in order to accurately price assets, and evaluate undervalued/overvalued assets n the market.

For Consumers: Understand how much they might expect to pay for specific features/aspects of a new home.

Presentation Outline:

Visualizing Data
Missing Data
Feature Selection
Data Transformation
Linear Models
Tree Models
KNN

Missingness



Data cleaning

we drop few unwanted columns, which don't have any impact on sale price. Then we change the data type of the ‘MSSubClass’ features from numeric to string.

- Missing Values
- Flagging as 'No':
  BsmtFinType1, BsmtFinType2, GarageType, GarageFinish
  Values are Missing, Because there is no Garage and Basement.
- Impute the Mode:
  Electrical,Exterior1st, GarageCars, MSZoning, KitchenQual, MasVnrType, SaleType
- Impute Zero:
  GarageQual, GarageCond, BsmtCond, BsmtExposure, BsmtQual, FireplaceQu
- Changing categorical ranking into numerical scale values.
- Dummify remaining categorical features.

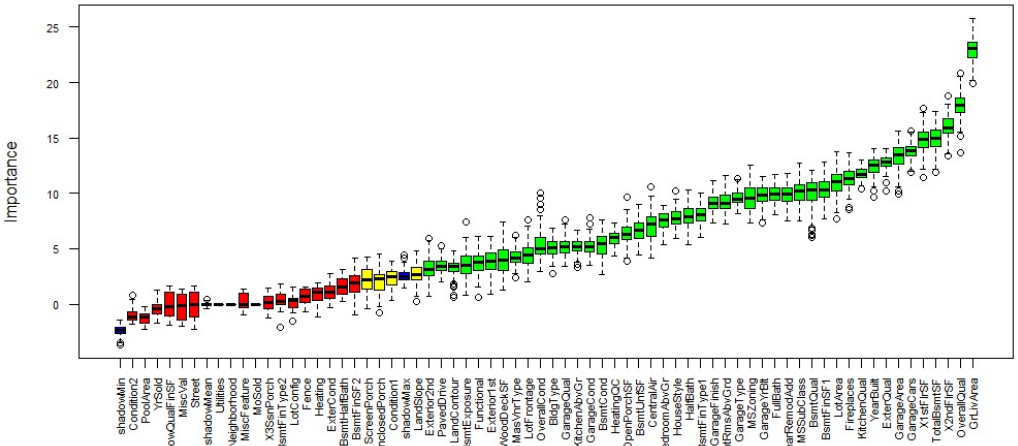
Features engineering

- Derived Features
- Feature:
    - TotalSF = TotalBsmtSF + GrLivArea + all other areas
    - TotalBaths = FullBath + BsmtFullBath + .5(HalfBath + BsmtHalfBath)
    - YearBuilt_Age = 2018 - YearBuilt
- Dropping columns which are repeated:
    - TotalBsmtSF = sum of(BsmtFinSF1, BsmtFinSF2, BsmtUnfSF')
    - GrLivAre = sum of(1stFlrSF, 2ndFlrSF)
- Taking log of ‘SalePrice’

Understanding and visualizing data

- Missing values
- Boruta as starting point
  - Random Forest

```
> missingness
          Id   MSSubClass     MSZoning  LotFrontage      LotArea       Street
        0.00         0.00         0.00        17.74         0.00         0.00
       Alley    LotShape  LandContour    Utilities    LotConfig    LandSlope
       93.77         0.00         0.00         0.00         0.00         0.00
Neighborhood   Condition1   Condition2     BldgType   HouseStyle  OverallQual
        0.00         0.00         0.00         0.00         0.00         0.00
 OverallCond    YearBuilt  YearRemodAdd    RoofStyle    RoofMatl  Exterior1st
        0.00         0.00         0.00         0.00         0.00         0.00
```



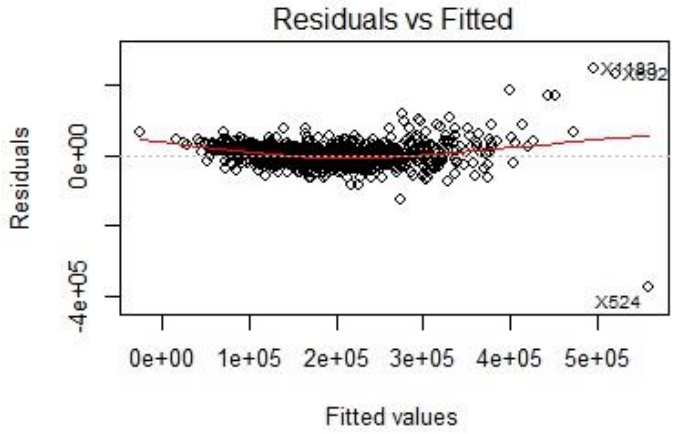

```
> train.boruta.selected.features.stats
                meanImp    medianImp       minImp      maxImp    normHits  decision
MSSubClass   10.23146182 10.19289121  7.0381690 13.4160559 1.000000000 Confirmed
MSZoning      9.54442241  9.54223872  6.6190851 13.5386744 1.000000000 Confirmed
LotFrontage   4.84000063  4.81041786  1.0591037  8.2433175 0.945891784 Confirmed
LotArea      11.06137068 11.07546492  7.6883059 14.6355573 1.000000000 Confirmed
Street        0.32034836  0.60605652 -2.2139295  1.6233930 0.000000000  Rejected
LandContour   3.29294944  3.33744953  0.8217137  5.0138084 0.825651303 Confirmed
Utilities     0.00000000  0.00000000  0.0000000  0.0000000 0.000000000  Rejected
LotConfig     0.01045448 -0.34702343 -1.6685811  2.2380066 0.000000000  Rejected
LandSlope     2.78879262  2.78509174 -0.1224927  6.3589709 0.597194389 Confirmed
Neighborhood  0.00000000  0.00000000  0.0000000  0.0000000 0.000000000  Rejected
Condition1    2.31997212  2.37257459 -0.7026733  4.6660047 0.452905812  Rejected
Condition2   -1.05018471 -1.39913091 -2.5243537  0.5650928 0.000000000  Rejected
BldgType      5.14919430  5.13505760  2.5850106  7.1341928 0.993987976 Confirmed
HouseStyle    7.80085816  7.90203825  4.2218717 10.1629154 1.000000000 Confirmed
OverallQual  17.83794999 17.97926215 13.8490065 20.3389953 1.000000000 Confirmed
OverallCond   5.19890016  4.97132116  2.8419043  9.9911439 0.993987976 Confirmed
YearBuilt    12.45174947 12.49480428  9.2918448 14.1892539 1.000000000 Confirmed
```

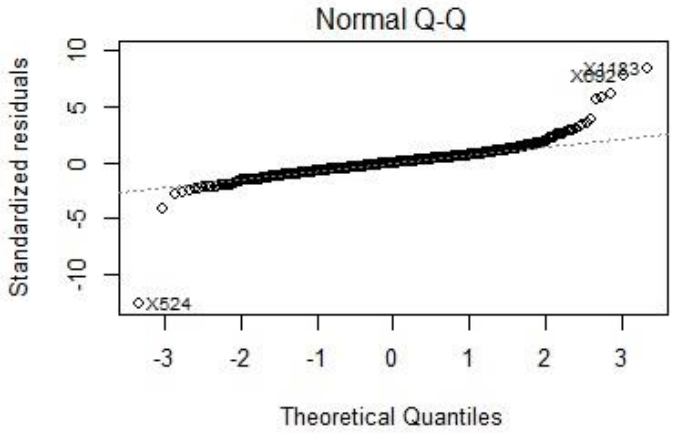
Linear models

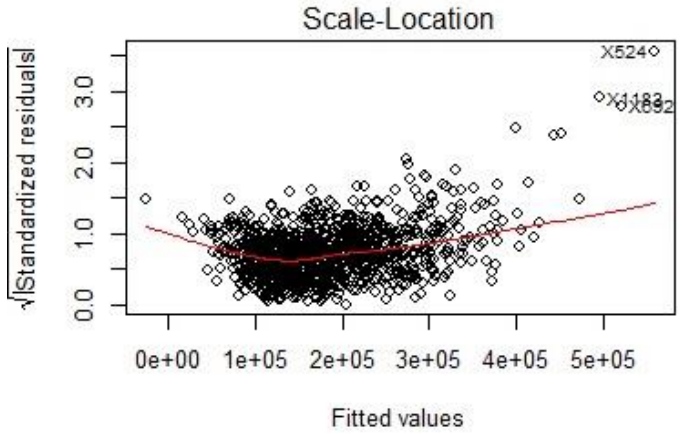
- Individual team models, best “crowdsourced” results selected
- Multivariable linear models
- Ridge
- Lasso
- Elastic

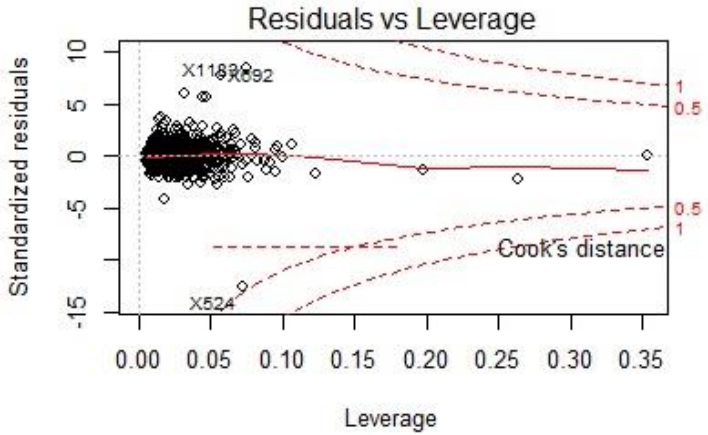
Plot different values for different attempts by different team members for different models

lm

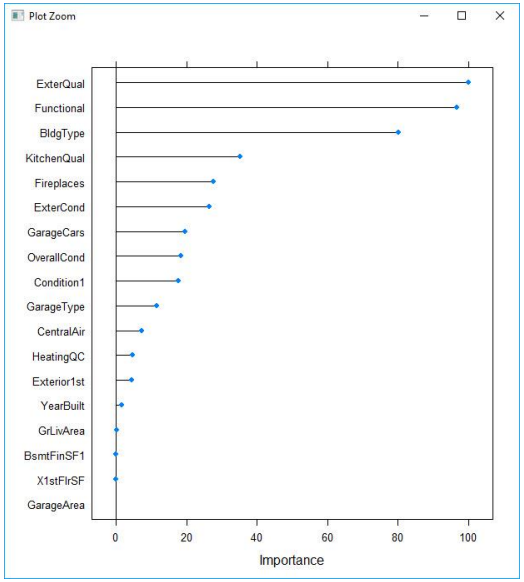

Residuals vs Fitted
Residuals
Fitted values
X524

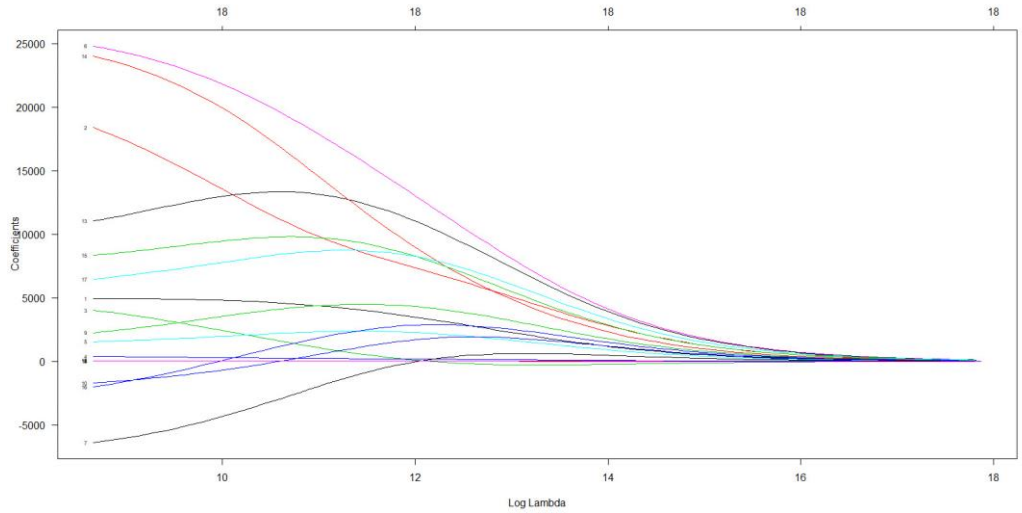

Normal Q-Q
Standardized residuals
Theoretical Quantiles
X524

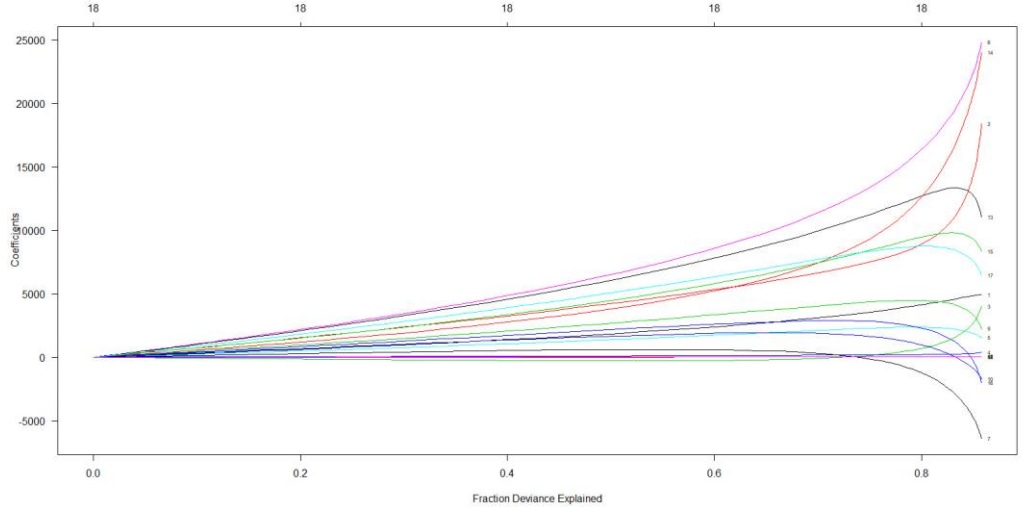

Scale-Location
√|Standardized residuals|
Fitted values
X524


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
X524

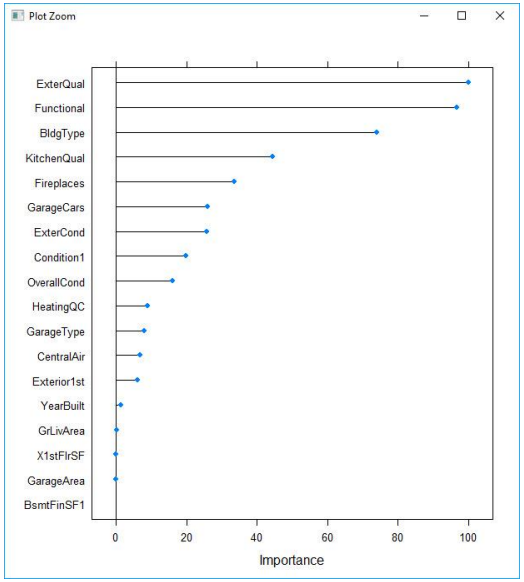
ridge

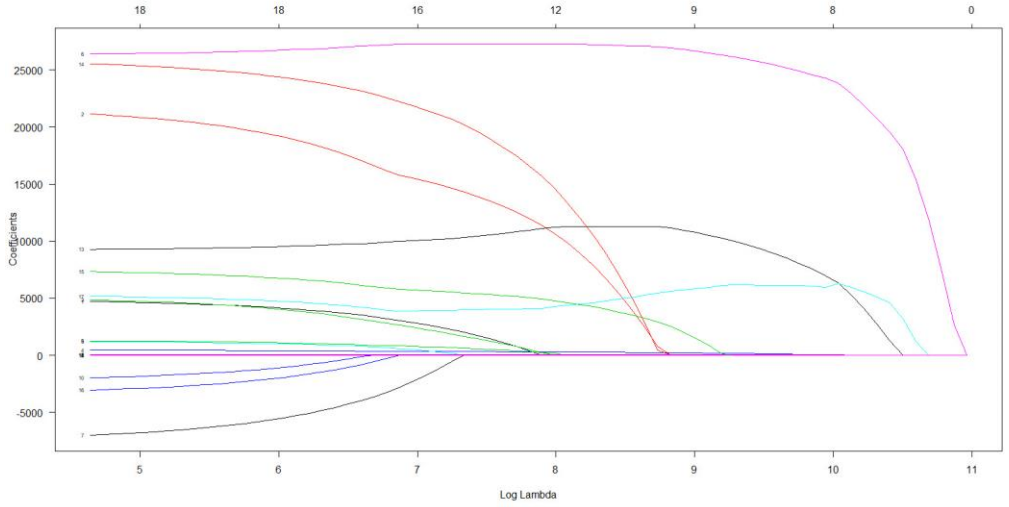

Plot Zoom
ExterQual
Functional
BldgType
KitchenQual
Fireplaces
ExterCond
GarageCars
OverallCond
Condition1
GarageType
CentralAir
HeatingQC
Exterior1st
YearBuilt
GrLivArea
BsmtFinSF1
X1stFlrSF
GarageArea
0
20
40
60
80
100
Importance

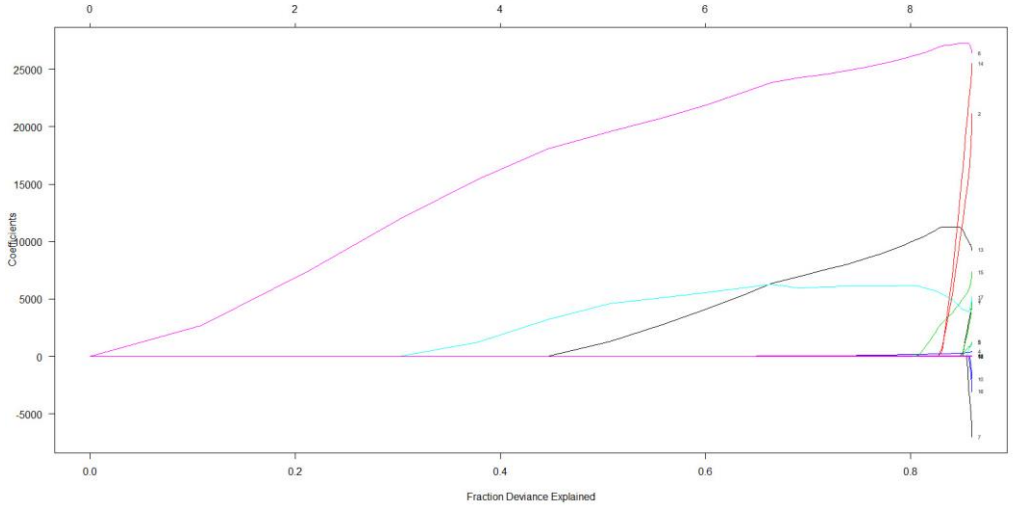

18
18
18
18
18
25000
20000
15000
10000
5000
0
-5000
Coefficients
10
12
14
16
18
Log Lambda


18
18
18
18
18
25000
20000
15000
10000
5000
0
-5000
Coefficients
0.0
0.2
0.4
0.6
0.8
Fraction Deviance Explained

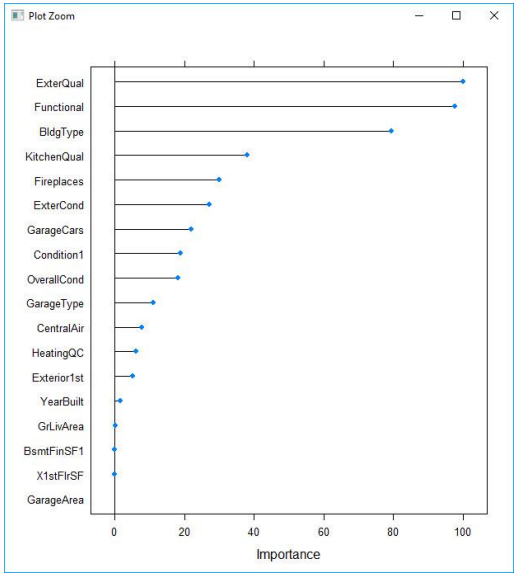
lasso

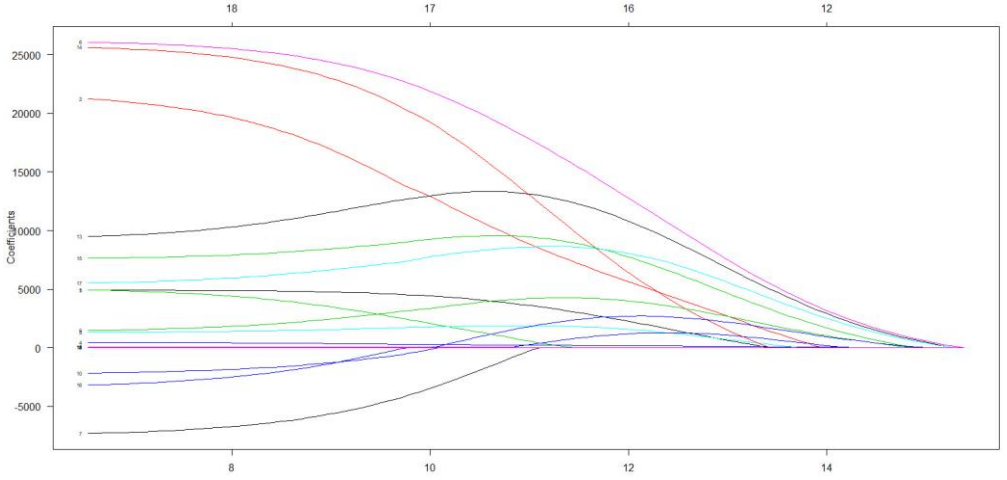

Plot Zoom
ExterQual
Functional
BldgType
KitchenQual
Fireplaces
GarageCars
ExterCond
Condition1
OverallCond
HeatingQC
GarageType
CentralAir
Exterior1st
YearBuilt
GrLivArea
X1stFlrSF
GarageArea
BsmtFinSF1
0
20
40
60
80
100
Importance

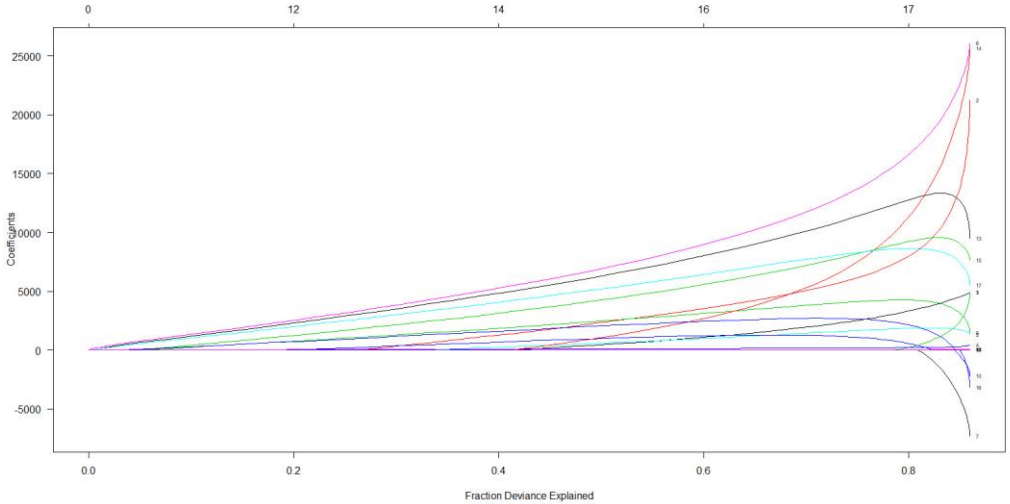

18
18
16
12
9
8
0
25000
20000
15000
10000
5000
0
-5000
Coefficients
5
6
7
8
9
10
11
Log Lambda


0
2
4
6
8
25000
20000
15000
10000
5000
0
-5000
Coefficients
0.0
0.2
0.4
0.6
0.8
Fraction Deviance Explained

elastic net

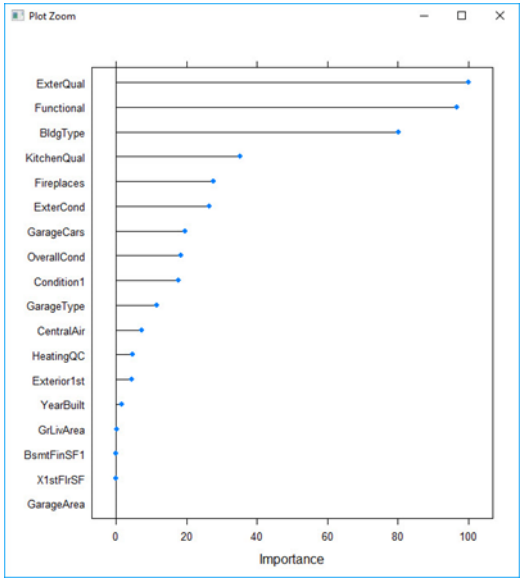

Plot Zoom
ExterQual
Functional
BldgType
KitchenQual
Fireplaces
ExterCond
GarageCars
Condition1
OverallCond
GarageType
CentralAir
HeatingQC
Exterior1st
YearBuilt
GrLivArea
BsmtFinSF1
X1stFlrSF
GarageArea
0
20
40
60
80
100
Importance


18
17
16
12
25000
20000
15000
10000
5000
0
-5000
Coefficients
8
10
12
14

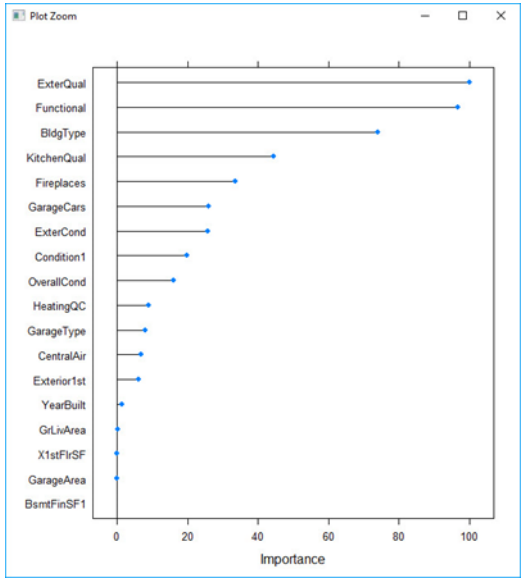

0
12
14
16
17
25000
20000
15000
10000
5000
0
-5000
Coefficients
0.0
0.2
0.4
0.6
0.8
Fraction Deviance Explained

Comparing variables importance

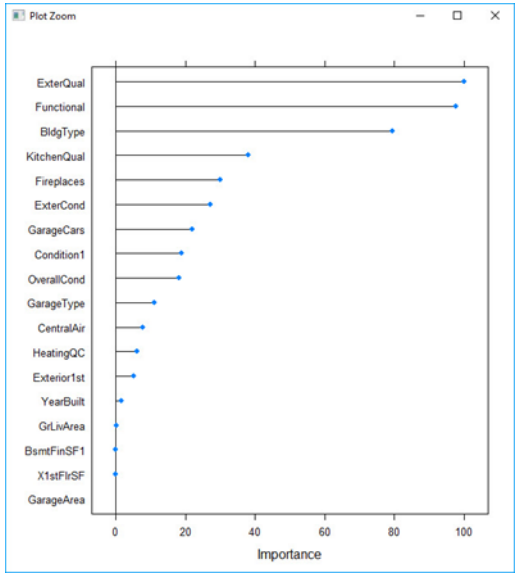
Ridge



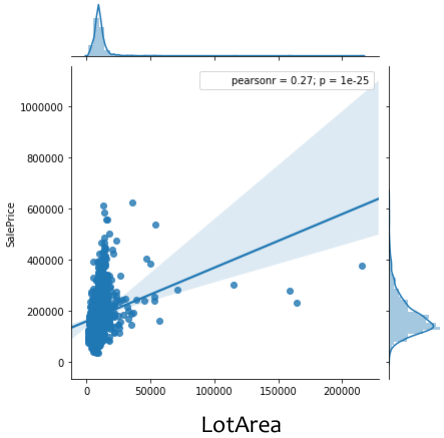
lasso

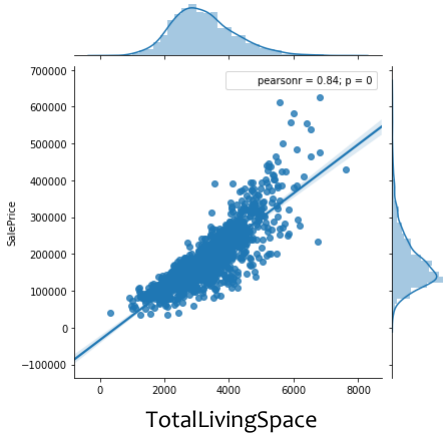


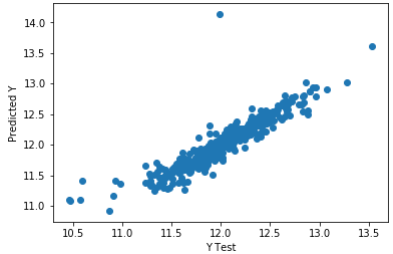
Elastic

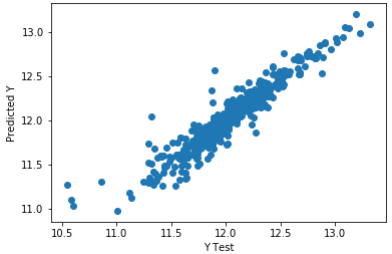


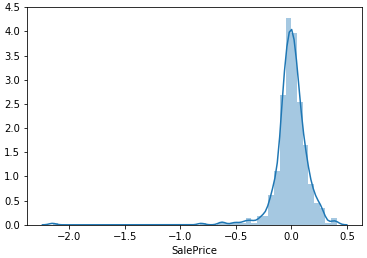
Predicting house prices using linear models

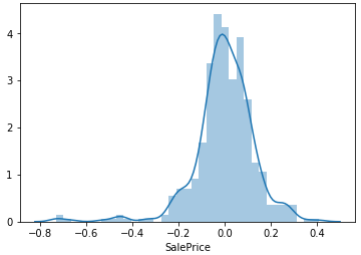












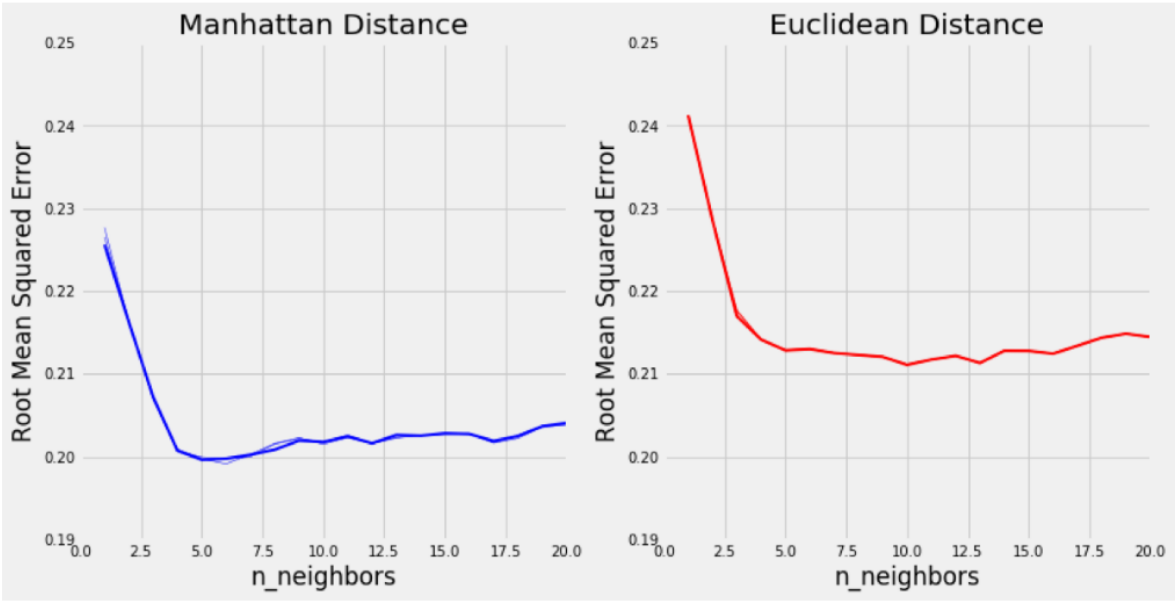
Model	RMSE	Comments
Linear Regression	0.1685	49 features, TotalLivingSpace
Linear Regression	0.1317	Squared # Bedrooms

Predicting house prices using k-nearest neighbors regression

New Features:
TotalLivingSpace instead of **LotArea**
Squared # BedRooms

Normalized the values: 0 - 1

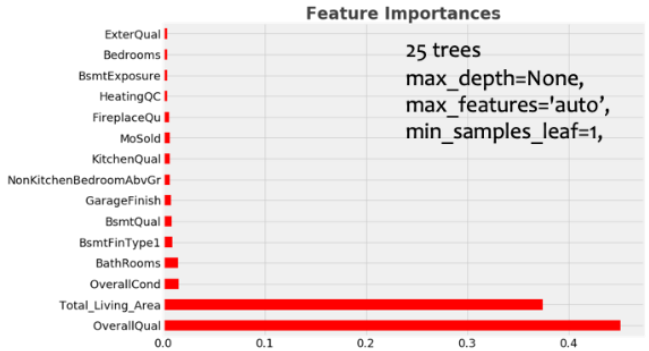
Metrics
Manhattan Distance
Euclidean Distance



Best **RMSE** obtained when K = 5 **0.19915480**

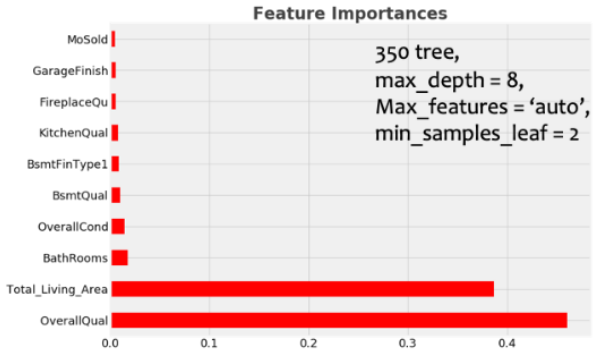
Predicting house prices using Tree-based models

Random Forest



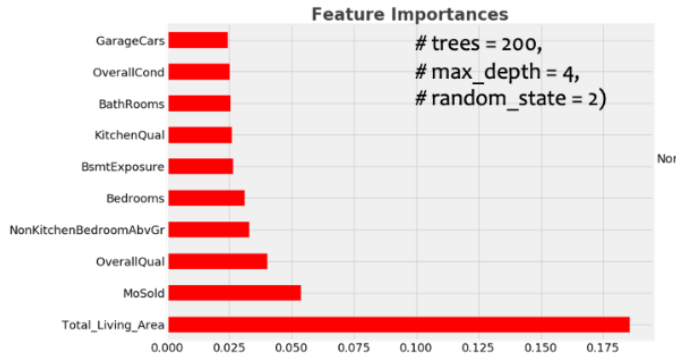
RMSE : 0.14987602771011657

Optimal Random Forest



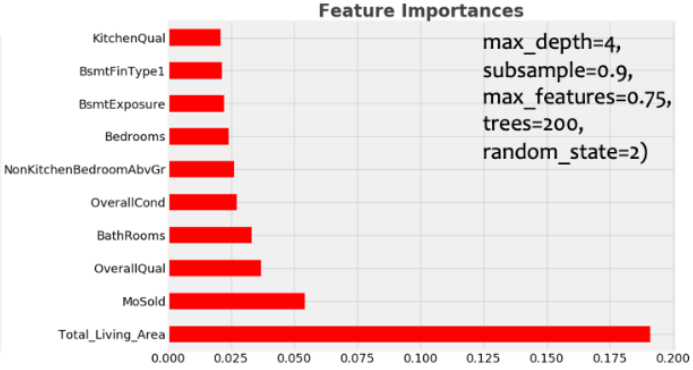
RMSE : 0.136551205992994

Gradient Boosting Regressor



RMSE :0.132472706

Optimal Gradient Boosting Regressor



RMSE: 0.121276024

Decision Tree

0.2383957875036891

Random Forest

0.14987602771011657

0.136551205992994

Boosting Regressor

0.132472706

0.121276024

Stochastic Boosting Regressor

0.13312094

0.122846123

XGBoost Regressor

1.963953

0.167252

0.1492399

Conclusions and lessons learned

- Feature selection is difficult and subjective!
- Heart on common sense, mind on statistics
- The selected features reveal similar results in all regression models.
- The following features appear to have much influence on a house price:
  - Total Living Area
  - Lot Area
  - Month Sold
  - Overall Quality
- For the tree-based models,
  - Total Living Area
  - Month Sold
  - Number of Bathroom
  - Kitchen Quality