Housing Prices

Elsie He, Owen Ou, Woonsup Kim, Emmelyn Luveta

TABLE OF CONTENTS

1. Background
2. Data
3. Data Preprocessing
4. Feature Selection
5. Modeling
6. Evaluating the model
7. Conclusion & Recommendation

01 BACKGROUND

- Overall material and finish quality
- Above ground living area
- Basement area
- Size of garage in car capacity

02 DATA

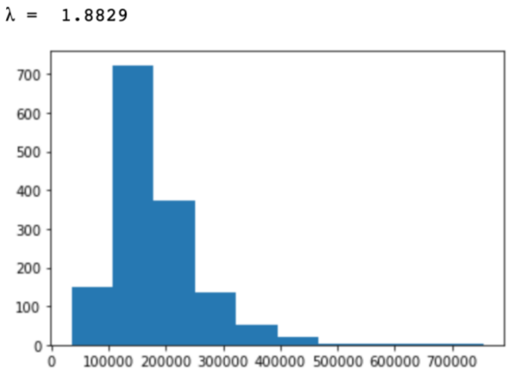
- Dataset of house prices in Ames, Iowa from 2006-2010.
- Dataset was collected by Bart de Cock in 2011.
- Raw data includes 1460 data points, 80 features, and 1 response variable, SalePrice.
- There are 38 numeric variables.
- NA columns.

03 DATA PREPROCESSING

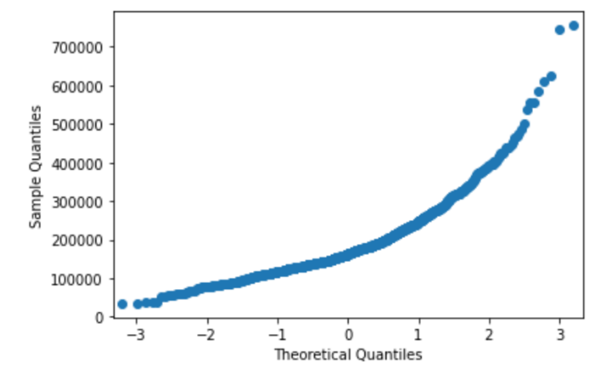
- Response Variable

Response

Skew: 1.8829

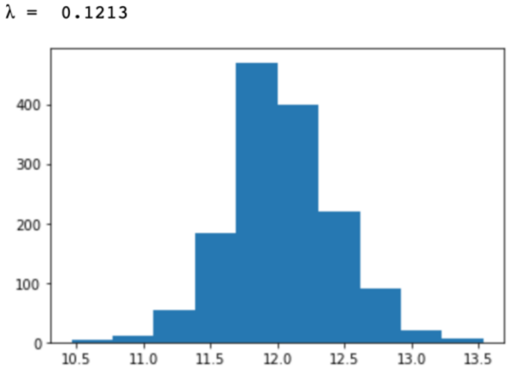


```
λ =  1.8829
```

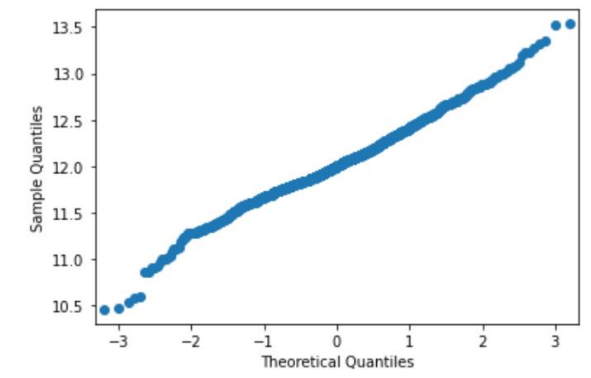


Log(response)

Skew: 0.1213

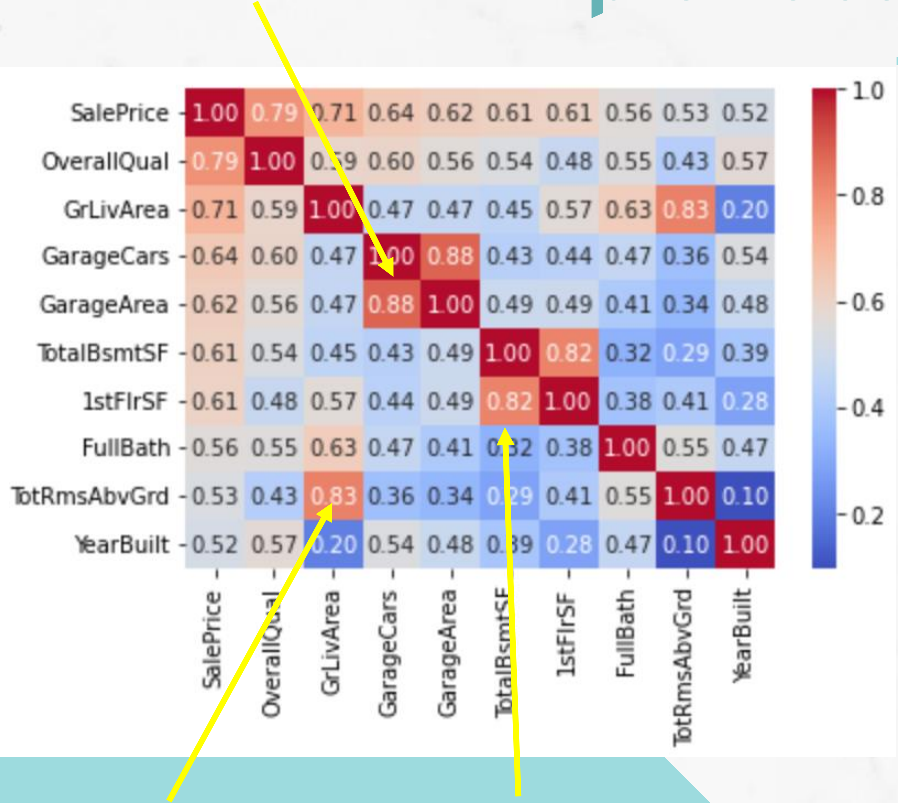


```
λ =  0.1213
```



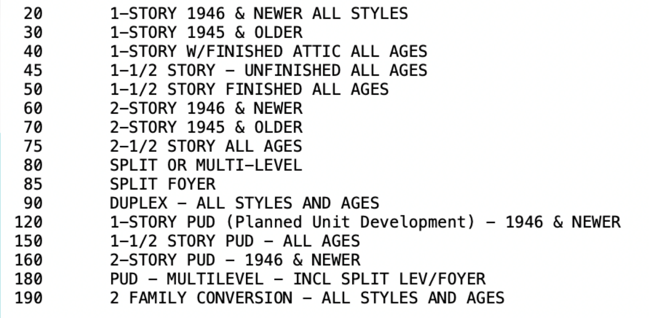
03 DATA PREPROCESSING

- Explanatory Variables

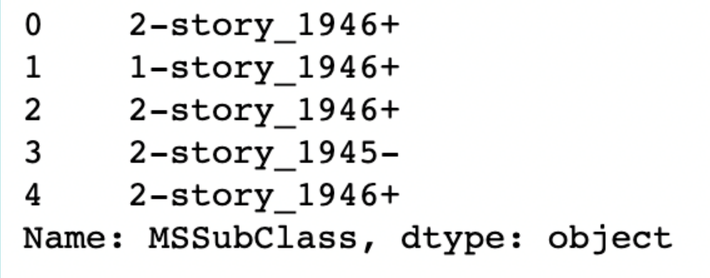


	SalePrice	OverallQual	GrLivArea	GarageCars	GarageArea	TotalBsmtSF	1stFlrSF	FullBath	TotRmsAbvGrd	YearBuilt
SalePrice	1.00	0.79	0.71	0.64	0.62	0.61	0.61	0.56	0.53	0.52
OverallQual	0.79	1.00	0.59	0.60	0.56	0.54	0.48	0.55	0.43	0.57
GrLivArea	0.71	0.59	1.00	0.47	0.47	0.45	0.57	0.63	0.83	0.20
GarageCars	0.64	0.60	0.47	1.00	0.88	0.43	0.44	0.47	0.36	0.54
GarageArea	0.62	0.56	0.47	0.88	1.00	0.49	0.49	0.41	0.34	0.48
TotalBsmtSF	0.61	0.54	0.45	0.43	0.49	1.00	0.82	0.32	0.29	0.39
1stFlrSF	0.61	0.48	0.57	0.44	0.49	0.82	1.00	0.38	0.41	0.28
FullBath	0.56	0.55	0.63	0.47	0.41	0.32	0.38	1.00	0.55	0.47
TotRmsAbvGrd	0.53	0.43	0.83	0.36	0.34	0.29	0.41	0.55	1.00	0.10
YearBuilt	0.52	0.57	0.20	0.54	0.48	0.39	0.28	0.47	0.10	1.00

03 DATA PREPROCESSING



```
 20        1-STORY 1946 & NEWER ALL STYLES
 30        1-STORY 1945 & OLDER
 40        1-STORY W/FINISHED ATTIC ALL AGES
 45        1-1/2 STORY - UNFINISHED ALL AGES
 50        1-1/2 STORY FINISHED ALL AGES
 60        2-STORY 1946 & NEWER
 70        2-STORY 1945 & OLDER
 75        2-1/2 STORY ALL AGES
 80        SPLIT OR MULTI-LEVEL
 85        SPLIT FOYER
 90        DUPLEX - ALL STYLES AND AGES
120        1-STORY PUD (Planned Unit Development) - 1946 & NEWER
150        1-1/2 STORY PUD - ALL AGES
160        2-STORY PUD - 1946 & NEWER
180        PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
190        2 FAMILY CONVERSION - ALL STYLES AND AGES
```



```
0     2-story_1946+
1     1-story_1946+
2     2-story_1946+
3     2-story_1945-
4     2-story_1946+
Name: MSSubClass, dtype: object
```

Dataset observation

- Map numerical features that are supposed to be categorical features
  - MSSubClass

MSSubClass: Identifies the type of dwelling involved in the sale.

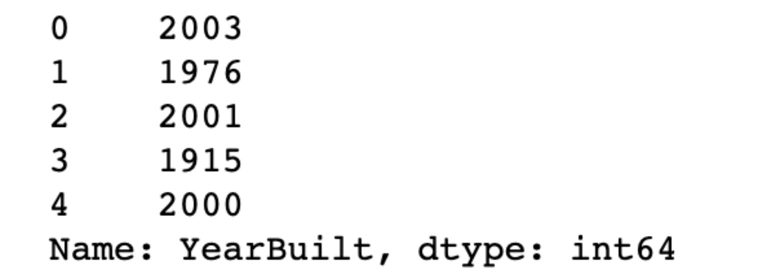
03 DATA PREPROCESSING

- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade

Dataset observation

- Combine 4 bathroom columns into 1
    - BsmtFullBath, BsmtHalfBath, HalfBath, FullBath

03 DATA PREPROCESSING



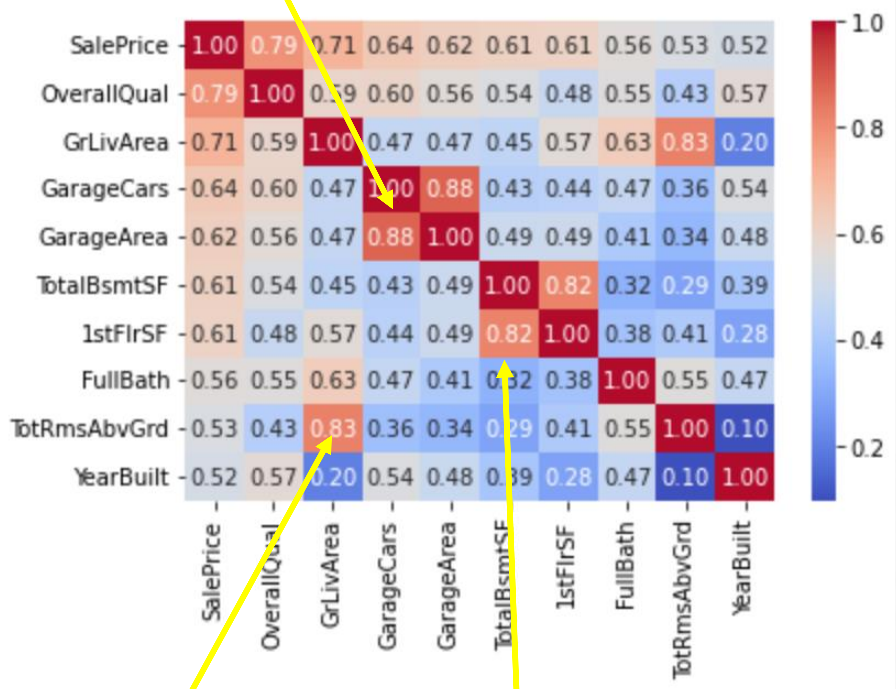
```
0     2003
1     1976
2     2001
3     1915
4     2000
Name: YearBuilt, dtype: int64
```

- Remod : indicates whether it has been remodeled
- HouseAge: YrSold - YearRemodAdd

Dataset observation

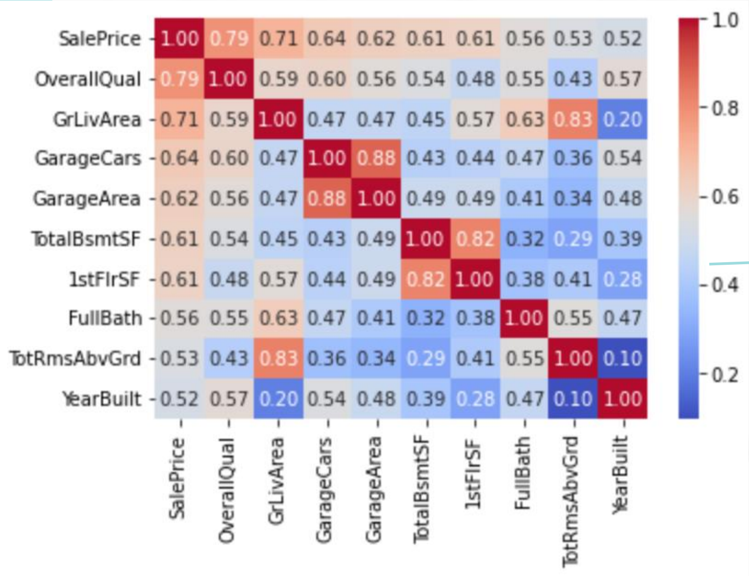
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
- YrSold: Year Sold

03 DATA PREPROCESSING

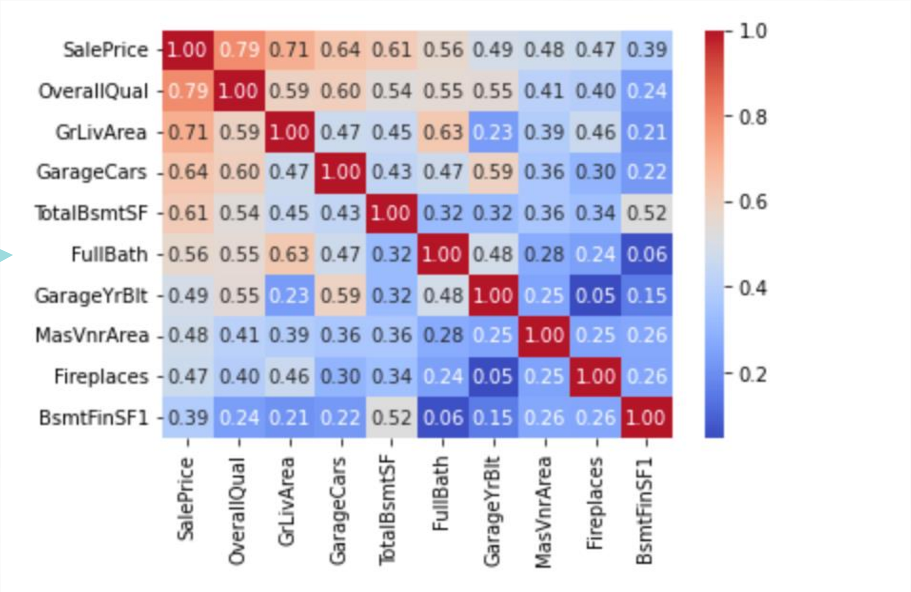


	SalePrice	OverallQual	GrLivArea	GarageCars	GarageArea	TotalBsmtSF	1stFlrSF	FullBath	TotRmsAbvGrd	YearBuilt
SalePrice	1.00	0.79	0.71	0.64	0.62	0.61	0.61	0.56	0.53	0.52
OverallQual	0.79	1.00	0.59	0.60	0.56	0.54	0.48	0.55	0.43	0.57
GrLivArea	0.71	0.59	1.00	0.47	0.47	0.45	0.57	0.63	0.83	0.20
GarageCars	0.64	0.60	0.47	1.00	0.88	0.43	0.44	0.47	0.36	0.54
GarageArea	0.62	0.56	0.47	0.88	1.00	0.49	0.49	0.41	0.34	0.48
TotalBsmtSF	0.61	0.54	0.45	0.43	0.49	1.00	0.82	0.32	0.29	0.39
1stFlrSF	0.61	0.48	0.57	0.44	0.49	0.82	1.00	0.38	0.41	0.28
FullBath	0.56	0.55	0.63	0.47	0.41	0.32	0.38	1.00	0.55	0.47
TotRmsAbvGrd	0.53	0.43	0.83	0.36	0.34	0.29	0.41	0.55	1.00	0.10
YearBuilt	0.52	0.57	0.20	0.54	0.48	0.39	0.28	0.47	0.10	1.00

03 DATA PREPROCESSING



	SalePrice	OverallQual	GrLivArea	GarageCars	GarageArea	TotalBsmtSF	1stFlrSF	FullBath	TotRmsAbvGrd	YearBuilt
SalePrice	1.00	0.79	0.71	0.64	0.62	0.61	0.61	0.56	0.53	0.52
OverallQual	0.79	1.00	0.59	0.60	0.56	0.54	0.48	0.55	0.43	0.57
GrLivArea	0.71	0.59	1.00	0.47	0.47	0.45	0.57	0.63	0.83	0.20
GarageCars	0.64	0.60	0.47	1.00	0.88	0.43	0.44	0.47	0.36	0.54
GarageArea	0.62	0.56	0.47	0.88	1.00	0.49	0.49	0.41	0.34	0.48
TotalBsmtSF	0.61	0.54	0.45	0.43	0.49	1.00	0.82	0.32	0.29	0.39
1stFlrSF	0.61	0.48	0.57	0.44	0.49	0.82	1.00	0.38	0.41	0.28
FullBath	0.56	0.55	0.63	0.47	0.41	0.32	0.38	1.00	0.55	0.47
TotRmsAbvGrd	0.53	0.43	0.83	0.36	0.34	0.29	0.41	0.55	1.00	0.10
YearBuilt	0.52	0.57	0.20	0.54	0.48	0.39	0.28	0.47	0.10	1.00

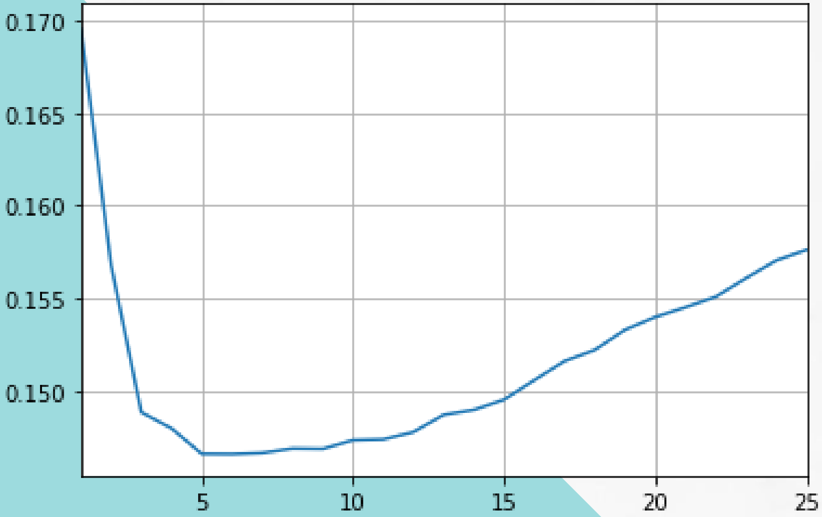


	SalePrice	OverallQual	GrLivArea	GarageCars	TotalBsmtSF	FullBath	GarageYrBlt	MasVnrArea	Fireplaces	BsmtFinSF1
SalePrice	1.00	0.79	0.71	0.64	0.61	0.56	0.49	0.48	0.47	0.39
OverallQual	0.79	1.00	0.59	0.60	0.54	0.55	0.55	0.41	0.40	0.24
GrLivArea	0.71	0.59	1.00	0.47	0.45	0.63	0.23	0.39	0.46	0.21
GarageCars	0.64	0.60	0.47	1.00	0.43	0.47	0.59	0.36	0.30	0.22
TotalBsmtSF	0.61	0.54	0.45	0.43	1.00	0.32	0.32	0.36	0.34	0.52
FullBath	0.56	0.55	0.63	0.47	0.32	1.00	0.48	0.28	0.24	0.06
GarageYrBlt	0.49	0.55	0.23	0.59	0.32	0.48	1.00	0.25	0.05	0.15
MasVnrArea	0.48	0.41	0.39	0.36	0.36	0.28	0.25	1.00	0.25	0.26
Fireplaces	0.47	0.40	0.46	0.30	0.34	0.24	0.05	0.25	1.00	0.26
BsmtFinSF1	0.39	0.24	0.21	0.22	0.52	0.06	0.15	0.26	0.26	1.00

04 FEATURE SELECTION

- Regularization Through Lasso Regression
- Removed unrelated features
- Reduced the effect of less dependent features
- Reduced the number of features from 334 to 117

05 MODELING



1. Linear Regression

    ○ Train RMSE = 0.1023

    ○ 10-CV RMSE = 0.1326

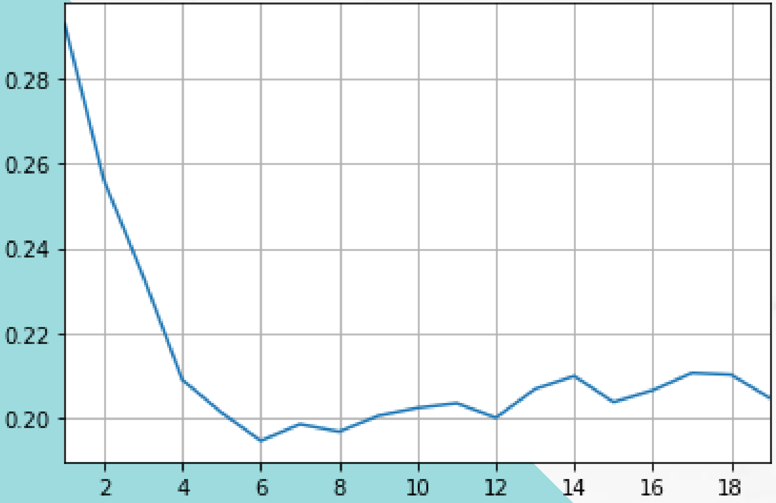
2. KNN

    ○ nn = 6

    ○ Train RMSE = 0.1334

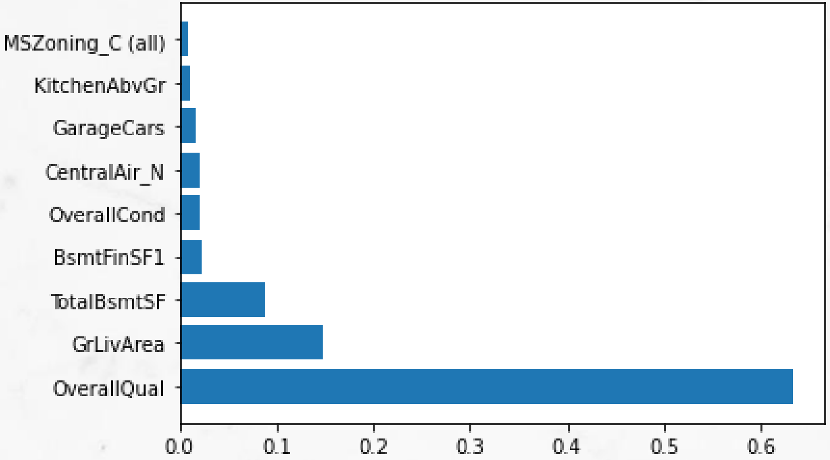
    ○ 10-CV RMSE = 0.1474

05 MODELING

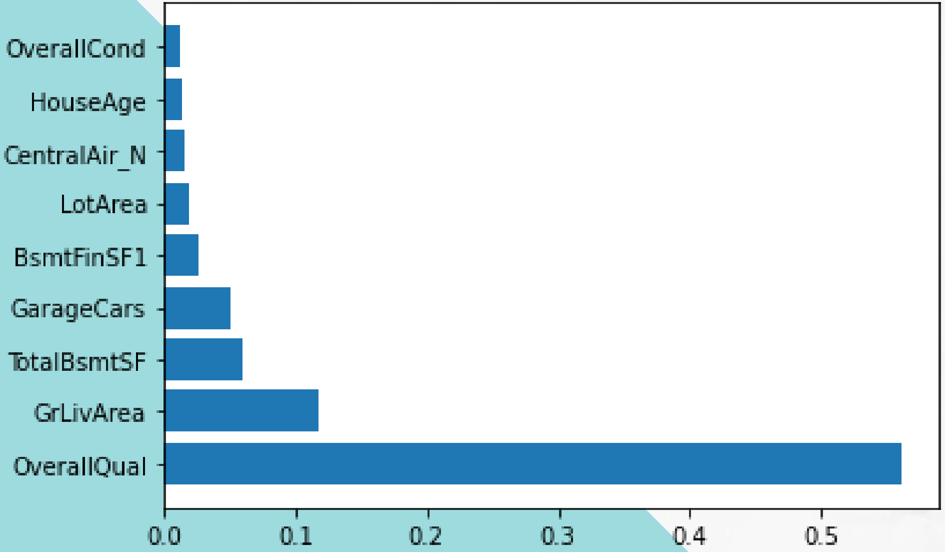


3. Decision Tree Regression

- ○ Max depth = 6
- ○ Train RMSE = 0.1343
- ○ 10-CV RMSE = 0.1924

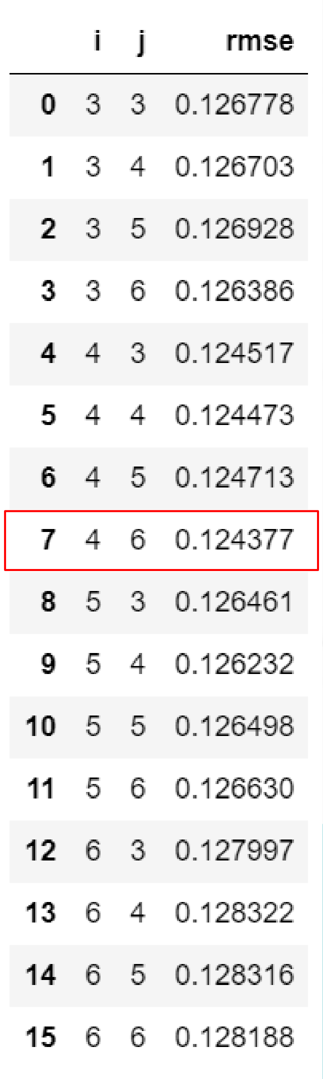


05 MODELING



4. Random Forest

- ○ Using 100 trees
- ○ No sub-sampling
- ○ Train RMSE = 0.05217
- ○ 10-CV RMSE = 0.1414

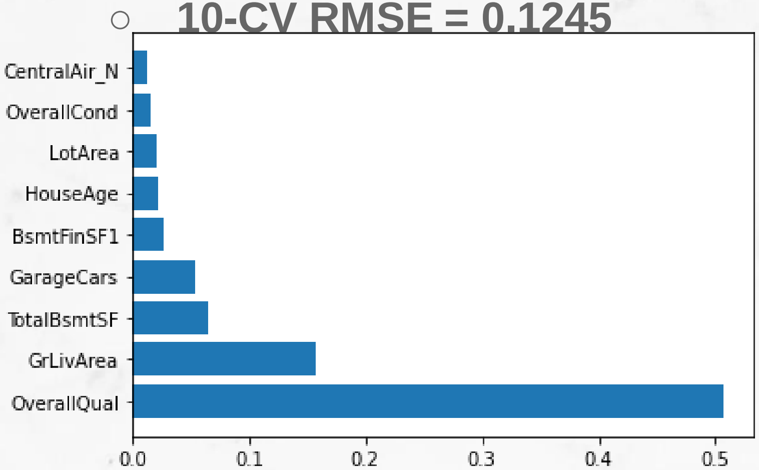


	i	j	rmse
0	3	3	0.126778
1	3	4	0.126703
2	3	5	0.126928
3	3	6	0.126386
4	4	3	0.124517
5	4	4	0.124473
6	4	5	0.124713
7	4	6	0.124377
8	5	3	0.126461
9	5	4	0.126232
10	5	5	0.126498
11	5	6	0.126630
12	6	3	0.127997
13	6	4	0.128322
14	6	5	0.128316
15	6	6	0.128188

05 MODELING

5. Gradient Boosting

- Max depth = 4
- Min samples split = 6
- Train RMSE = 0.06630
- 10-CV RMSE = 0.1245



06 EVALUATING THE MODEL

Lin Reg

Train RMSE
= 0.1023
10-CV RMSE
= 0.1326

KNN Reg

Train RMSE
= 0.1334
10-CV RMSE
= 0.1474

Dec Tree

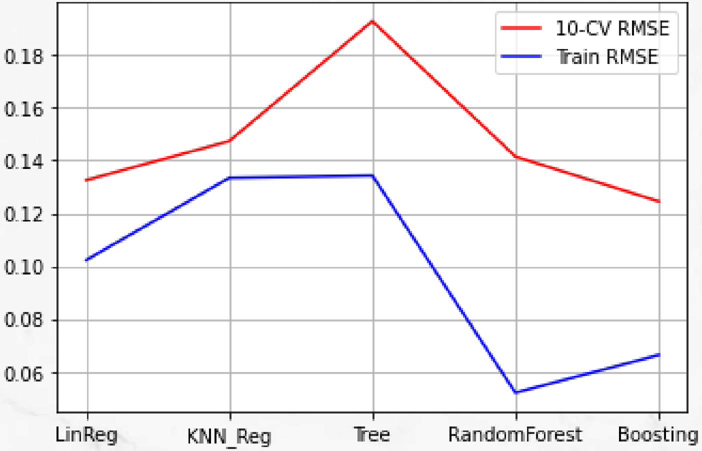
Train RMSE
= 0.1343
10-CV RMSE
= 0.1924

Random Forest

Train RMSE
= 0.05217
10-CV RMSE
= 0.1414

Gradient Boosting

Train RMSE
= 0.06630
10-CV RMSE
= 0.1245



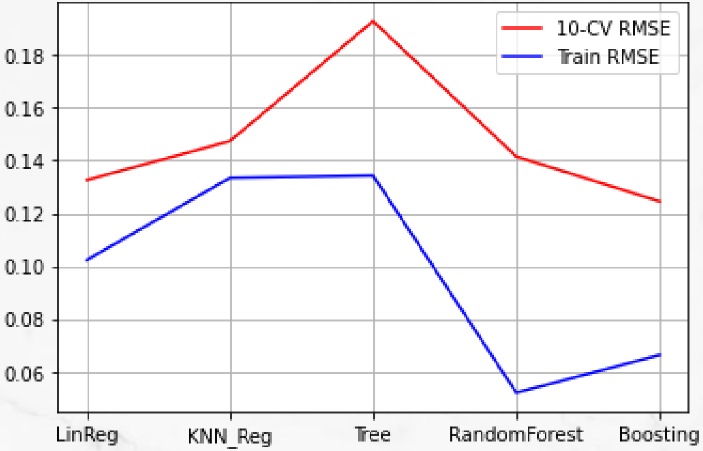
06 EVALUATING THE MODEL

Random Forest

- Lowest Train RMSE of **0.05217**
- High 10-CV RMSE suggests that there may be an overfit

Gradient Boosting

- Lowest 10-CV RMSE of **0.1245**
- Best performance



07 CONCLUSION & RECOMMENDATION

Based on our dataset, gradient boosting lists the following features to be most influential in housing prices

01 "OverallQual"

Overall material and finish quality

02 "GrLivArea"

Above ground living area square feet

03 "TotalBsmtSF"

Total square feet of basement area

04 "GarageCars"

Size of garage in car capacity

07 CONCLUSION & RECOMMENDATION

Rooms for improvement:

- PCA analysis when doing feature analysis
- Tune random forest parameters: number of estimators or sub sampling
- Try bagging models for regression

THANK YOU