Ames Undervalued Homes

By: The Outliers - Aleksey, Alex, Jesse, and Nixon

Objective & Target Audience

Objective: To find undervalued homes, determine the features importances that separate undervalued versus non-undervalued homes, and pick out features that have the biggest incremental impact for home improvements.

Target Audience: Potential homebuyers seeking to flip a home in Ames, Iowa, or a current homeowner looking to improve their home's value

Which features drive Sale Price Per GLA well below neighborhood average?

- We focused on the 'Bottom 80%' of GrLivArea as detailed in the Project Description

Data: Handling and Cleaning Overview

Features

Features were split into three groups based on type:

- Categoricals
- Ordinals
- Numericals

Null Values

Null Values were treated differently depending on the feature and feature's type

EDA

EDA was performed on all features to best understand feature values, distribution, & their effects on the target variable

Data: Features

- Categorical: 25 categorical columns
- Ordinal: 20 ordinal columns
- Numeric: 36 numeric columns
- Feature Engineering
  - SalePricePerGLA, which is defined as SalePrice divided by GrLivArea (Square feet)
  - Utilized the mean SalePricePerGLA for each Neighborhood and the std SalePricePerGLA for each Neighborhood to determine whether a home was undervalued

Data: Null Values

- Feature nulls as a percent to total
- Main Numerical Features of concern were:
  - “LotFrontage”, “GarageYrBlt”
- Many categorical features state that values are null when that property does not have the appropriate amenity for the feature
  - Ex. Homes without pools have a null value for their “PoolQC”

	Numerical %Null
LotFrontage	0.179139
GarageYrBlt	0.050019
MasVnrArea	0.005428
BsmtHalfBath	0.000775
BsmtFullBath	0.000775
TotalBsmtSF	0.000388
BsmtUnfSF	0.000388
GarageArea	0.000388
GarageCars	0.000388
BsmtFinSF2	0.000388
BsmtFinSF1	0.000388
PoolArea	0.000000
ScreenPorch	0.000000
3SsnPorch	0.000000
MiscVal	0.000000
MoSold	0.000000

	Categorical %Null
PoolQC	0.996510
MiscFeature	0.962389
Alley	0.934858
Fence	0.796433
FireplaceQu	0.481194
GarageCond	0.050019
GarageQual	0.050019
GarageFinish	0.050019
GarageType	0.049244
BsmtExposure	0.027530
BsmtFinType2	0.027142
BsmtFinType1	0.026755
BsmtCond	0.026755
BsmtQual	0.026755
MasVnrType	0.005428
Electrical	0.000388

Data Cleaning: Numerical Features

- LotFrontage (imputed from the Median by Neighborhood)
- GarageYrBuilt (if null, imputed from the Median by YrBuilt, or set equal to YrBuilt if not enough observations)
- Otherwise, nulls were treated as zeroes

Data Cleaning: Categorical Features

- Null values were changed to “NoneListed”
    - Allowed for ordinal features to be label encoded later on

- Fill 1 null “Electrical” value with “SBrkr” (Standard Breaker): Built in 2006 and all modern homes have a standard breaker system

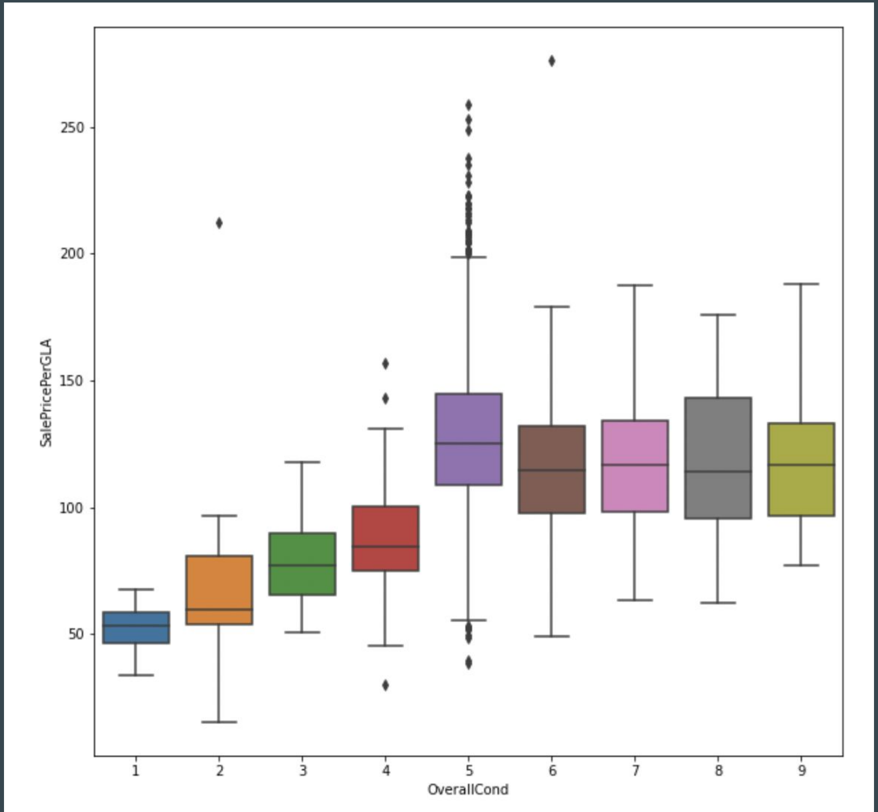
Data Cleaning: Removed Observations

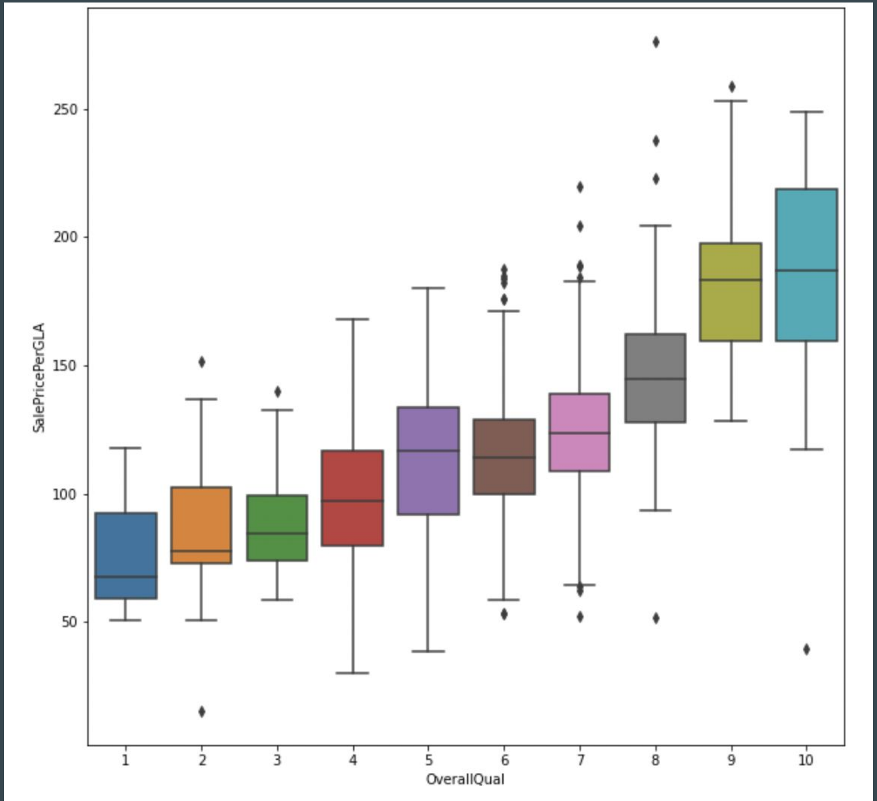
We removed 28 observations that had potential to skew the data (one was a duplicate)

- “MiscFeature" : "TenC", "Othr"
  - Only one home had “TenC” (tennis court)
  - Three homes with “Othr”, for which there was no information
- "Utilities" : "NoSewr"
  - Two homes used a septic tank, whereas the rest had all public utilities included
- "Functional" : "Sal"
  - One home was bought for salvaging materials
- "Heating" : "Floor"
  - One home used a floor furnace for heating
- "SaleCondition" : "Family", "AdjLand"
  - 17 homes were sold to family members, which could imply a lower sale price than at-market
  - Two homes were sold as part of a sale of adjacent land

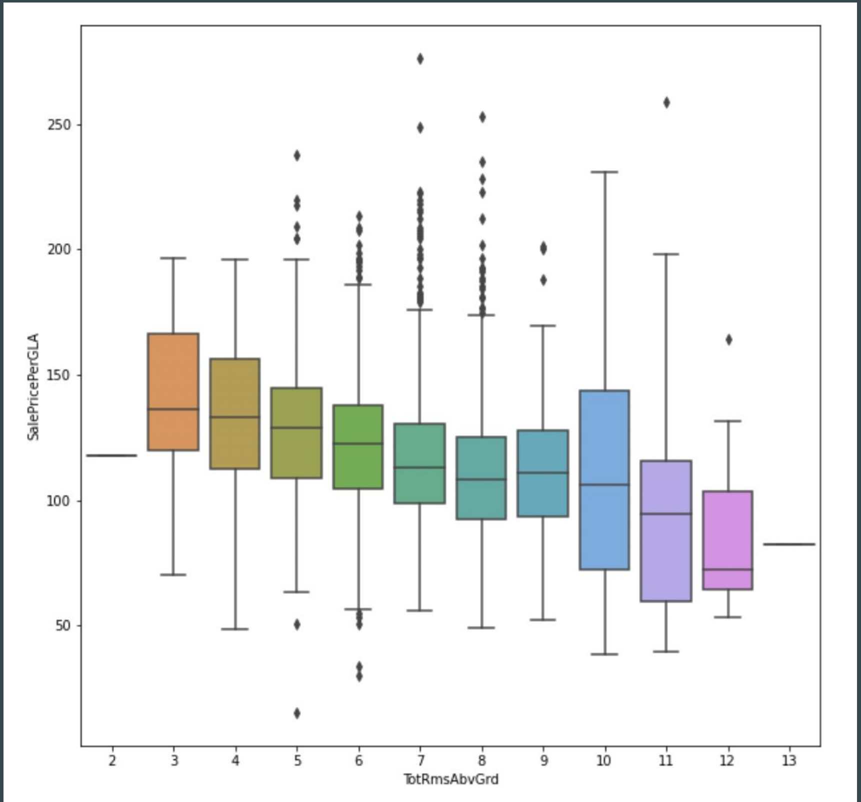
Exploratory Data Analysis

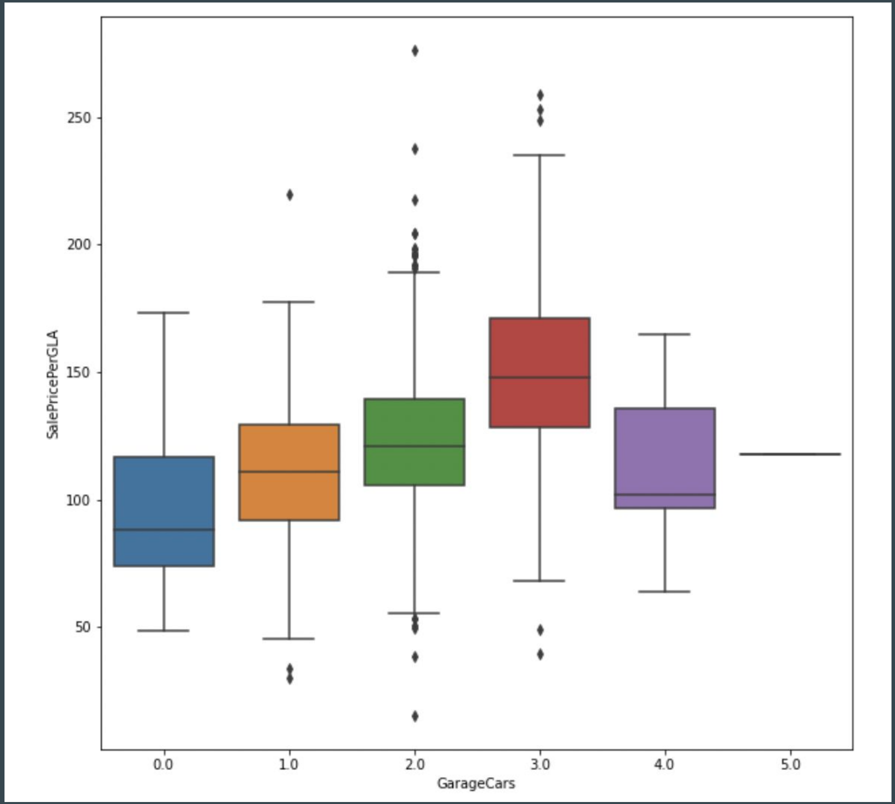
EDA of Ordinals



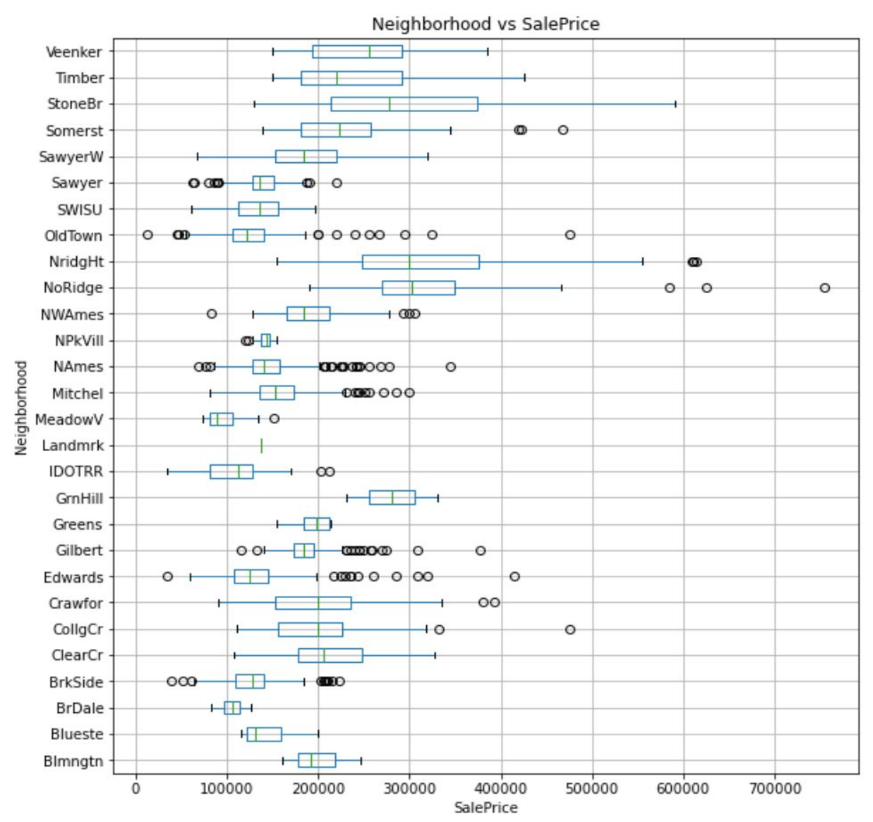


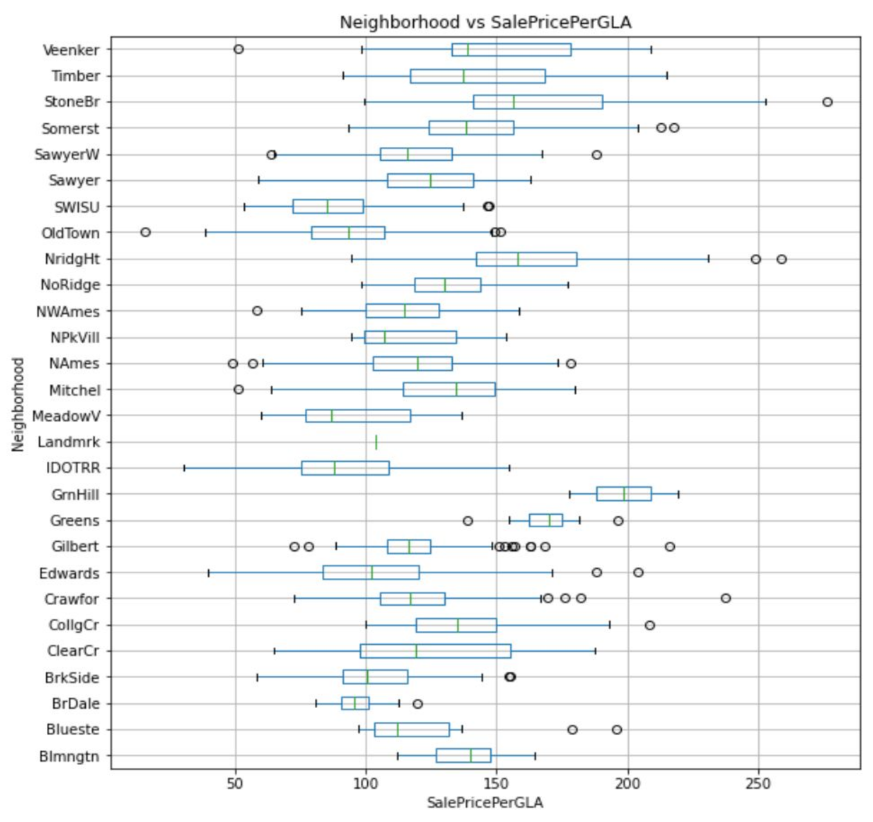
EDA of Numericals





EDA of Categoricals



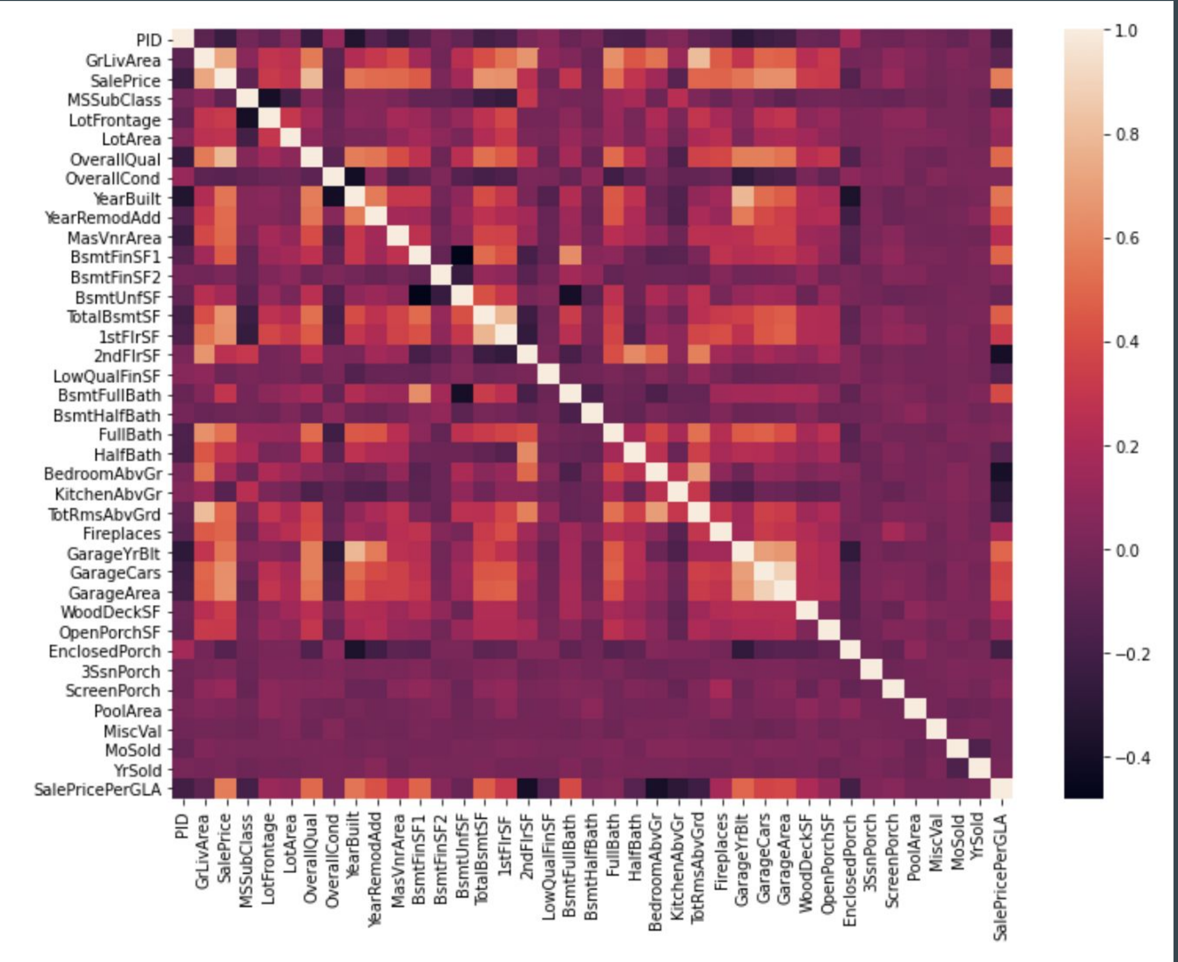


EDA Findings

SalePricePerGLA	SalePricePerGLA	1.000000
GarageArea	GarageCars	0.889129
YearBuilt	GarageYrBlt	0.835005
TotRmsAbvGrd	GrLivArea	0.806748
OverallQual	SalePrice	0.792510
TotalBsmtSF	1stFlrSF	0.783627
SalePrice	GrLivArea	0.723536
BedroomAbvGr	TotRmsAbvGrd	0.691686
2ndFlrSF	GrLivArea	0.663564
TotalBsmtSF	SalePrice	0.654793
FullBath	GrLivArea	0.645672
SalePrice	1stFlrSF	0.644983
GarageCars	SalePrice	0.640041
GarageArea	SalePrice	0.636669
BsmtFinSF1	BsmtFullBath	0.635159
2ndFlrSF	HalfBath	0.623710
YearRemodAdd	GarageYrBlt	0.623009
TotRmsAbvGrd	2ndFlrSF	0.585331
GarageYrBlt	GarageCars	0.580799
GarageCars	OverallQual	0.580630

dtype: float64

These are the top features that are heavily correlated to be aware of in avoiding multicollinearity.



Preprocessing

Preprocessing - Numericals

- Scaling numeric columns allowed for easier interpretability for coefficients when modeling with linear regression, also gave slightly better R^2 scores

BedroomAbvGr	YearBuilt	PoolArea	MasVnrArea	3SsnPorch	EnclosedPorch	KitchenAbvGr	2ndFlrSF	LotArea	SalePrice	TotalBsmtSF	LowQualFinSF	GarageCars
-1.032929	-1.058660	-0.05141	-0.562947	-0.099850	-0.361018	-0.199997	-0.794888	-0.271658	-0.698705	-0.432037	-0.091555	0.338441
-1.032929	0.455558	-0.05141	0.288960	-0.099850	-0.361018	-0.199997	-0.794888	-0.722263	-0.519002	0.029379	-0.091555	-1.015322
-1.032929	-1.361504	-0.05141	-0.562947	3.282513	0.294913	-0.199997	-0.794888	-0.497269	-0.713348	-0.477462	-0.091555	-1.015322
-1.032929	-2.370983	-0.05141	-0.562947	-0.099850	2.262705	-0.199997	-0.034531	-0.240097	-0.858442	-1.510268	-0.091555	-1.015322
0.186893	1.027595	-0.05141	-0.562947	-0.099850	-0.361018	-0.199997	1.224073	-0.208783	0.645743	-0.542012	-0.091555	0.338441
...	...	...	...	...	...	...	...	...	...	...	...	...
-1.032929	-1.832594	-0.05141	-0.562947	-0.099850	-0.361018	-0.199997	-0.794888	-0.152812	-0.765262	-0.202525	-0.091555	-1.015322
1.406715	-0.520272	-0.05141	-0.562947	-0.099850	-0.361018	-0.199997	-0.794888	0.442159	-0.517671	-2.478524	-0.091555	0.338441
1.406715	-0.722168	-0.05141	-0.562947	-0.099850	-0.361018	4.803848	1.568831	-0.471379	-0.445789	-0.085378	-0.091555	1.692204
0.186893	0.993946	-0.05141	0.260372	-0.099850	-0.361018	-0.199997	1.228795	-0.156264	0.519285	-0.123630	-0.091555	0.338441

Preprocessing - Ordinal Features

- Label encoding was used to prevent multicollinearity of ordinal features.
- A numeric rank was used when the feature had a clear rank of possible values.

Feature Values of External Quality (ExterQual)	Encoded Value
NaN (Missing Value)	0
Poor (Po)	1
Fair (Fa)	2
Typical (TA)	3
Good (Gd)	4
Excellent (Ex)	5

20 features encoded this way:

LotShape, LandSlope, OverallQual, OverallCond, ExterQual, ExterCond, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, HeatingQC, KitchenQual, Functional, FireplaceQu, GarageFinish, GarageQual, GarageCond, PoolQC, Fence

Preprocessing - Categorical Features

- Categorical features were encoded using *pd.get_dummies()* with drop_first = True
- The 25 categorical features were:
    - MiscFeature, LotConfig, Condition1, GarageType, SaleType, Electrical, Heating, Condition2, MasVnrType, LandContour, RoofMatl, Foundation, Exterior2nd, CentralAir, SaleCondition, MSSubClass, Neighborhood, RoofStyle, Alley, MSZoning, BldgType, Exterior1st, HouseStyle, Street, PavedDrive
- This resulted in 163 new features

Preprocessing: Linear Regression Feature Selection

Numerical:

- Removed Multicollinearity by measuring VIF and removing feature with the highest VIF in stepwise fashion
- Used Stepwise function to remove features that were not significant based on p-value

Categorical:

- Created dummies for categorical features, label encoded ordinal features and added them all in stepwise fashion until the Adj R Squared no longer increased

Final Data Frame contains 117 features of a maximum of 234 features

Models, Methods & Analysis

Models

- 75% / 25% Train/Test Splits were used for all modeling
  - 10-fold Cross Validation was done on the Train Split
- We used Linear (OLS, Ridge, Lasso, and ElasticNet), RandomForest, GradientBoost, Support Vector Regressor, and XGBoost to find the best predictor

Model	Training Scoring	Test Scoring
OLS	0.7358 (Adj.)	0.7207 (Adj.)
GS Ridge	0.7223 (Adj.)	0.7188 (Adj.)
GS Lasso	0.7112 (Adj.)	0. 7149(Adj.)
GS Elastic Net	0.7188 (Adj.)	0.7168(Adj.)
Random Forest	0.9519	0.7935
SVR (Linear)	0.8097	0.7983
XGBoost Regressor	0.815	0.8424

Linear Regression: Parameter Tuning

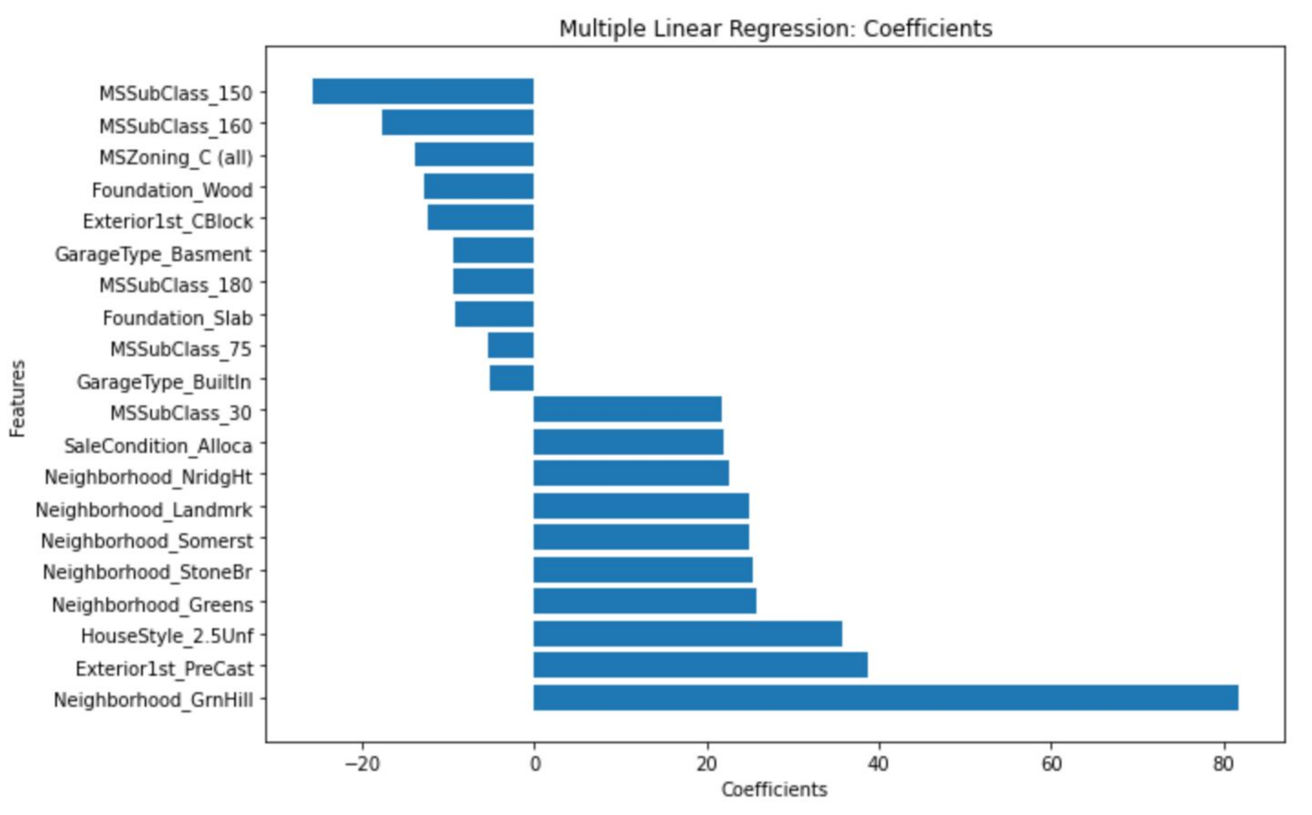
Ridge & Lasso

- Alpha: Tuned using grid search and affects the strength of the ridge or lasso penalty

Elastic Net

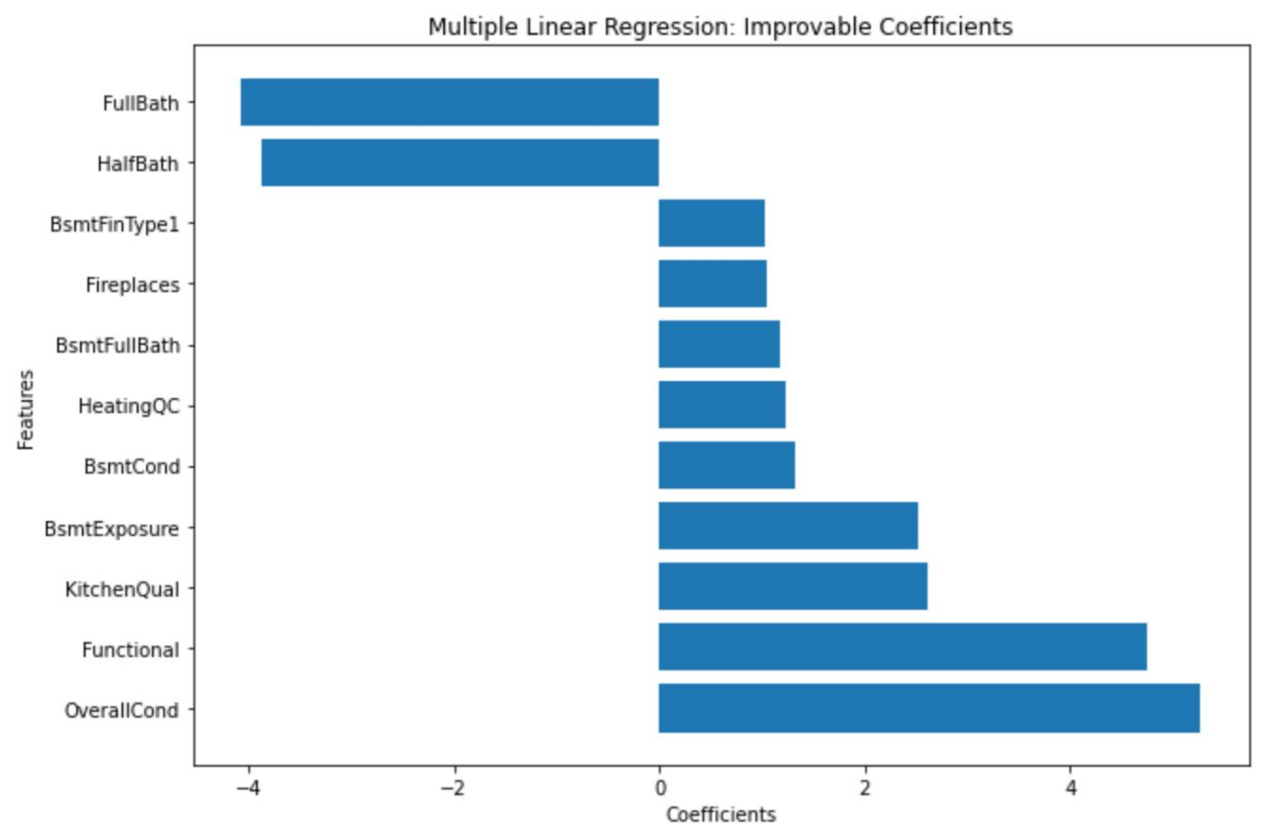
- Alpha: Tuned using grid search and affects the penalty strength
- L1 ratio: Tuned using same grid search and affects the ratio between the ridge and lasso penalties

Analysis: MLR Summary



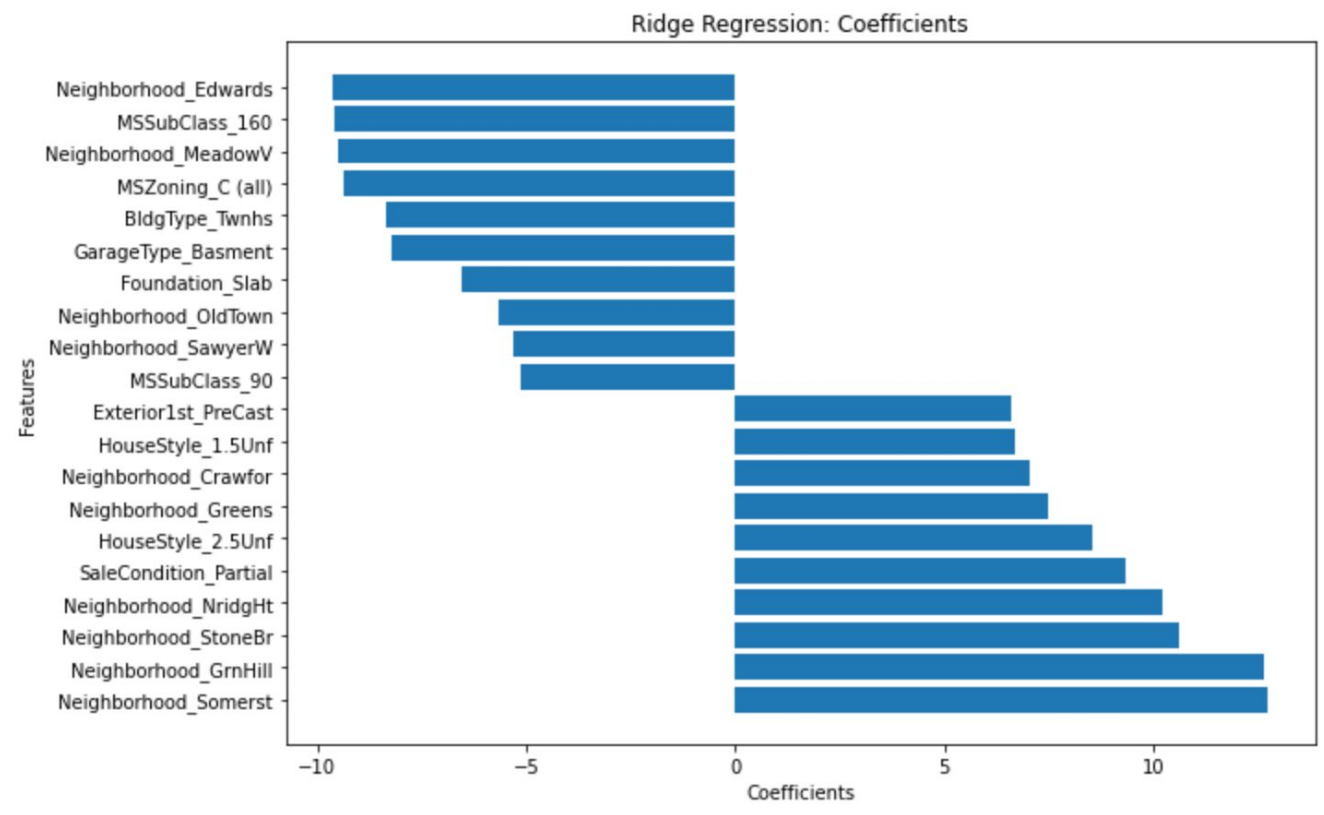
Features	Coefficients
Neighborhood_GrnHill	81.776110
Exterior1st_PreCast	38.828907
HouseStyle_2.5Unf	35.818923
Neighborhood_Greens	25.822369
Neighborhood_StoneBr	25.404627
Neighborhood_Somerst	25.015174
Neighborhood_Landmrk	25.004390
Neighborhood_NridgHt	22.549650
SaleCondition_Alloca	22.039118
MSSubClass_30	21.799691
GarageType_BuiltIn	-5.086861
MSSubClass_75	-5.353863
Foundation_Slab	-9.170288
MSSubClass_180	-9.320783
GarageType_Basment	-9.322064
Exterior1st_CBlock	-12.391358
Foundation_Wood	-12.865919
MSZoning_C (all)	-13.853768
MSSubClass_160	-17.659092
MSSubClass_150	-25.783014

Results: MLR Improvable Features



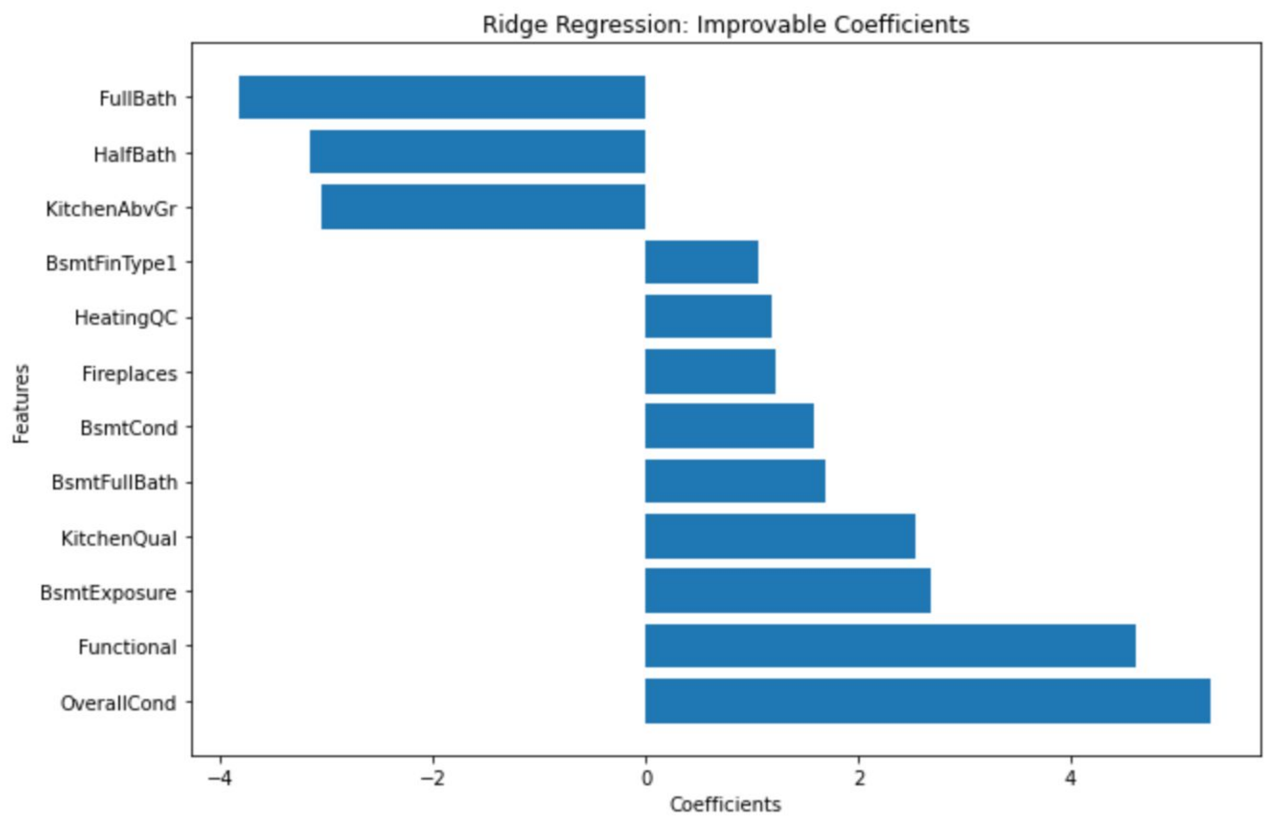
Features	Coefficients
OverallCond	5.270297
Functional	4.753592
KitchenQual	2.612149
BsmtExposure	2.517492
BsmtCond	1.328564
HeatingQC	1.219599
BsmtFullBath	1.175587
Fireplaces	1.036505
BsmtFinType1	1.024161
HalfBath	-3.872985
FullBath	-4.071589

Analysis: Ridge Summary



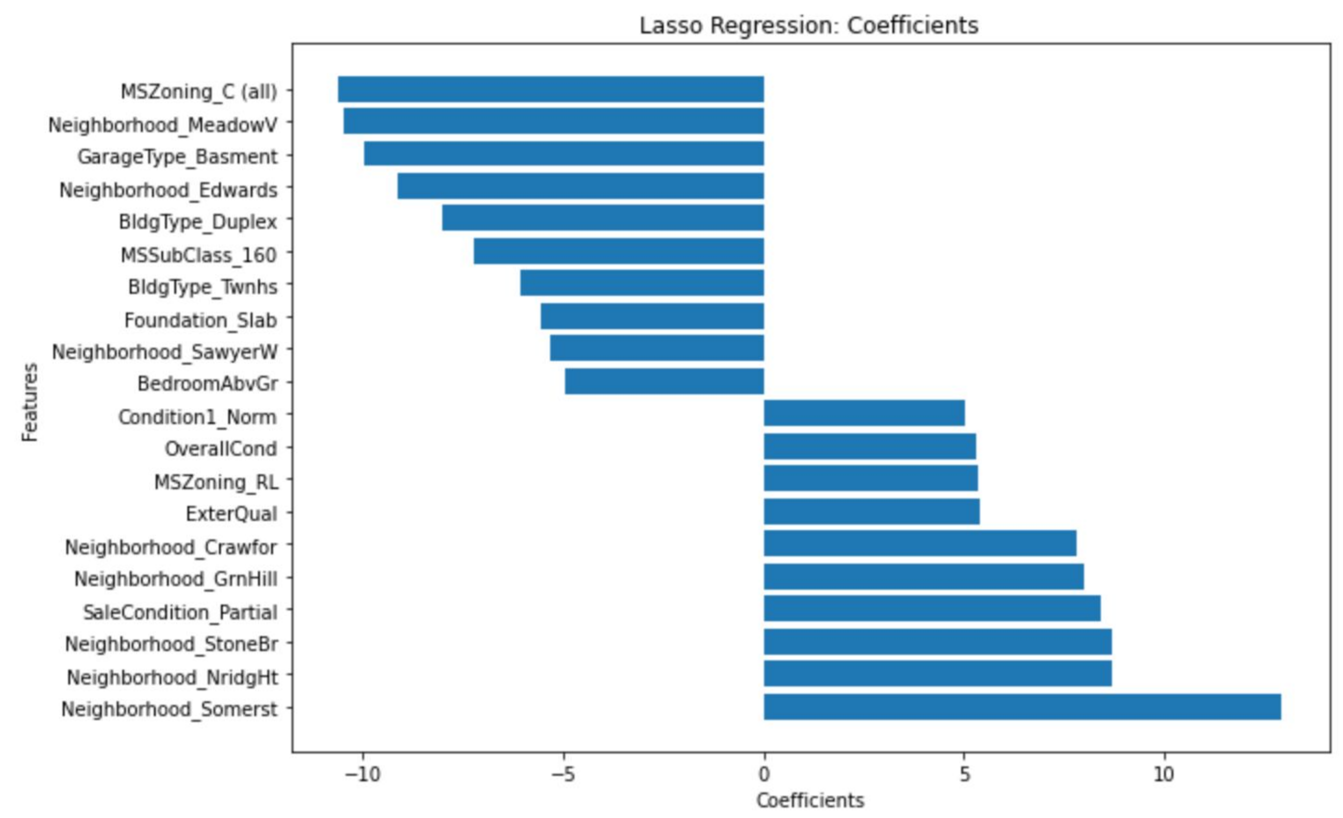
Features	Coefficients
Neighborhood_Somerst	12.755946
Neighborhood_GrnHill	12.641791
Neighborhood_StoneBr	10.629446
Neighborhood_NridgHt	10.225660
SaleCondition_Partial	9.359583
HouseStyle_2.5Unf	8.533000
Neighborhood_Greens	7.487013
Neighborhood_Crawfor	7.053668
HouseStyle_1.5Unf	6.674045
Exterior1st_PreCast	6.605180
MSSubClass_90	-5.109450
Neighborhood_SawyerW	-5.293967
Neighborhood_OldTown	-5.666453
Foundation_Slab	-6.525803
GarageType_Basment	-8.197030
BldgType_Twnhs	-8.357784
MSZoning_C (all)	-9.374796
Neighborhood_MeadowV	-9.485257
MSSubClass_160	-9.607845
Neighborhood_Edwards	-9.627113

Results: Ridge Improvable Features



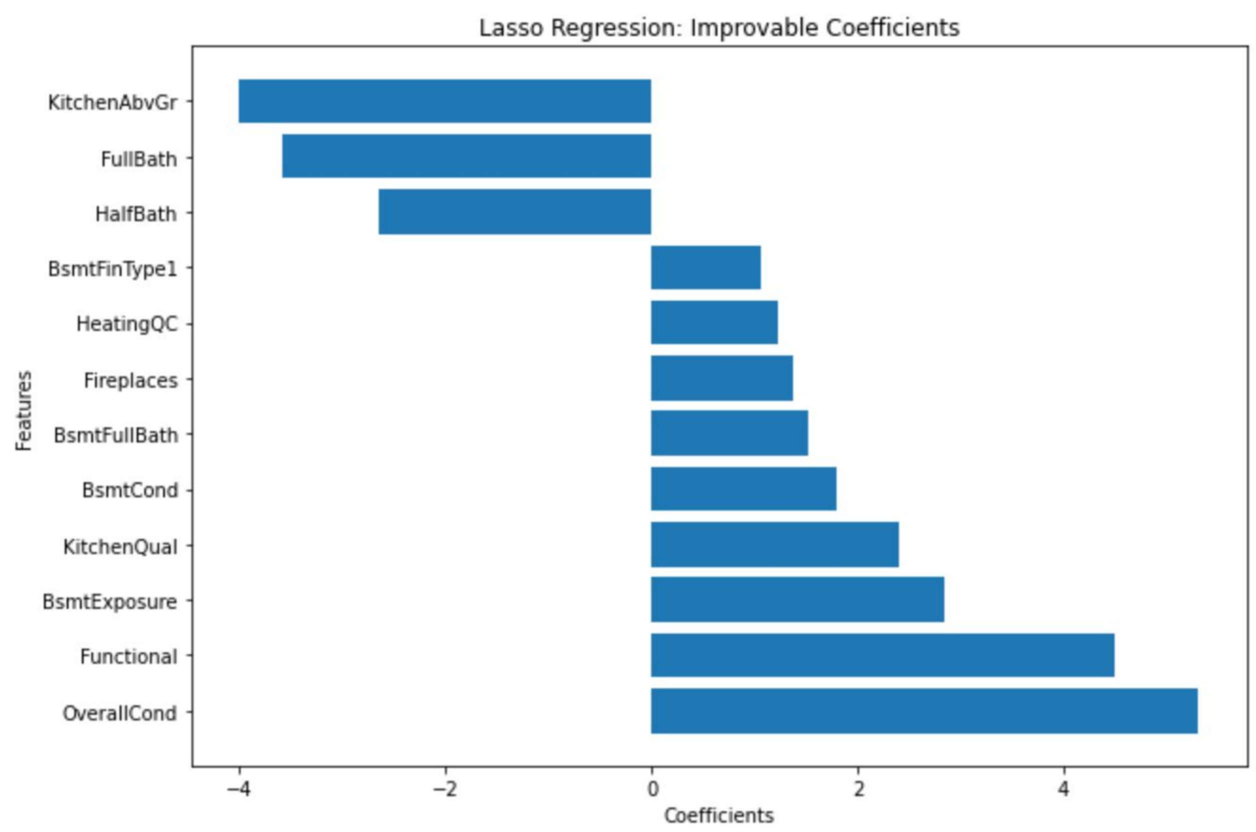
Features	Coefficients
OverallCond	5.320129
Functional	4.605067
BsmtExposure	2.675196
KitchenQual	2.527483
BsmtFullBath	1.686928
BsmtCond	1.574619
Fireplaces	1.216908
HeatingQC	1.191863
BsmtFinType1	1.057405
KitchenAbvGr	-3.058069
HalfBath	-3.156868
FullBath	-3.824261

Analysis: Lasso Summary



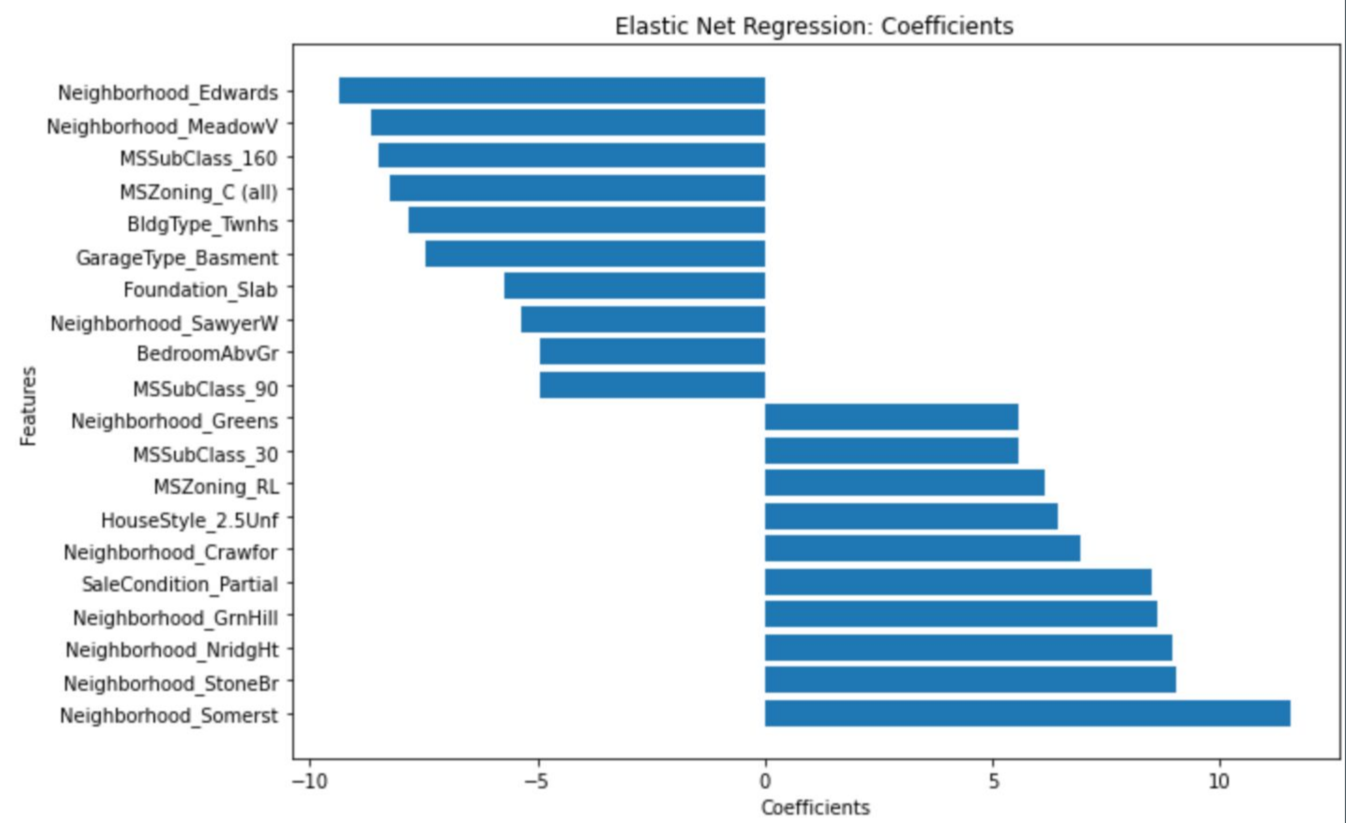
Features	Coefficients
Neighborhood_Somerst	12.927482
Neighborhood_NridgHt	8.682985
Neighborhood_StoneBr	8.668559
SaleCondition_Partial	8.389797
Neighborhood_GrnHill	8.003892
Neighborhood_Crawfor	7.784835
ExterQual	5.409972
MSZoning_RL	5.342809
OverallCond	5.314009
Condition1_Norm	5.007743
BedroomAbvGr	-4.958231
Neighborhood_SawyerW	-5.313066
Foundation_Slab	-5.552876
BldgType_Twnhs	-6.085709
MSSubClass_160	-7.227669
BldgType_Duplex	-8.019247
Neighborhood_Edwards	-9.112960
GarageType_Basment	-9.971147
Neighborhood_MeadowV	-10.491922
MSZoning_C (all)	-10.609242

Results: Lasso Improvable Features



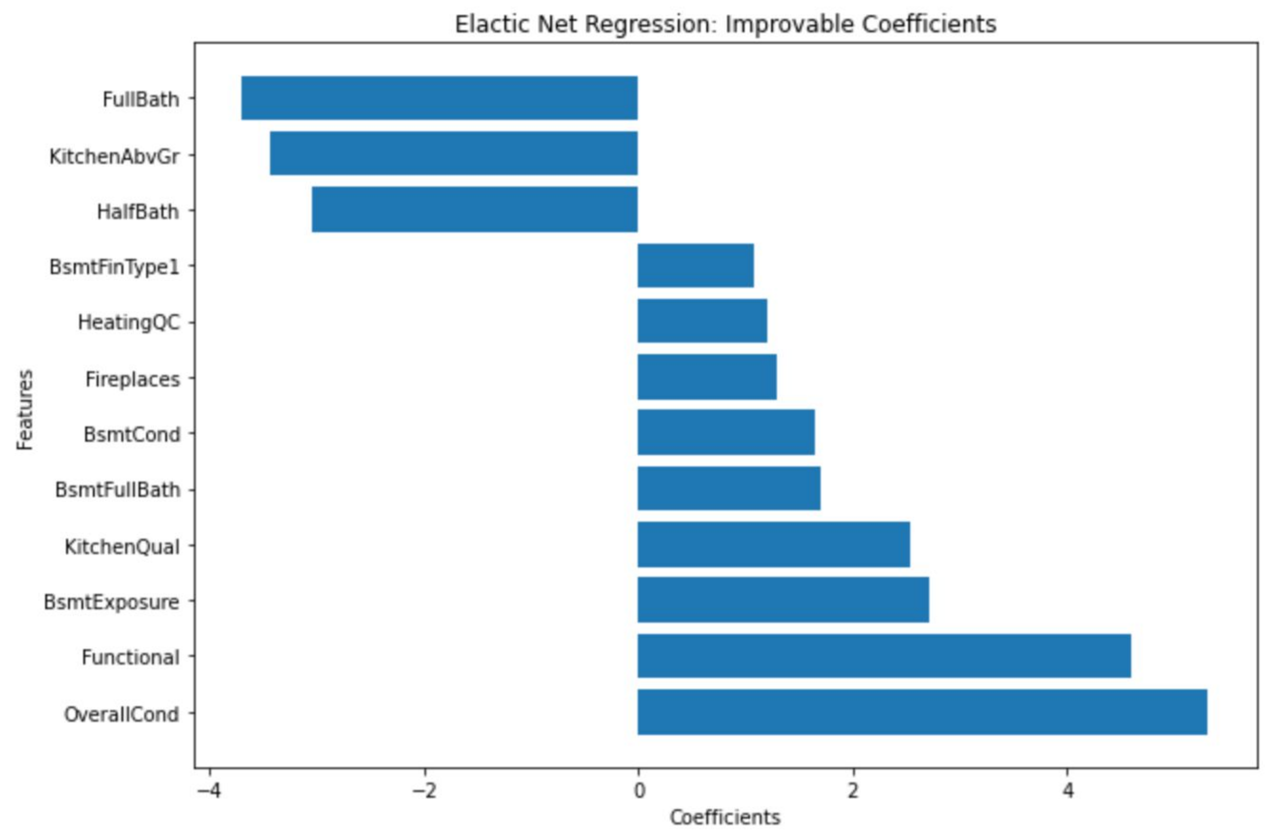
Features	Coefficients
OverallCond	5.314009
Functional	4.495283
BsmtExposure	2.842990
KitchenQual	2.408755
BsmtCond	1.805372
BsmtFullBath	1.514642
Fireplaces	1.374994
HeatingQC	1.221044
BsmtFinType1	1.066858
HalfBath	-2.645631
FullBath	-3.578070
KitchenAbvGr	-4.005470

Analysis: Elastic Net Summary



Features	Coefficients
Neighborhood_Somerst	11.576119
Neighborhood_StoneBr	9.061752
Neighborhood_NridgHt	8.957326
Neighborhood_GrnHill	8.652132
SaleCondition_Partial	8.513389
Neighborhood_Crawfor	6.942753
HouseStyle_2.5Unf	6.447519
MSZoning_RL	6.147991
MSSubClass_30	5.585174
Neighborhood_Greens	5.582697
MSSubClass_90	-4.944164
BedroomAbvGr	-4.950503
Neighborhood_SawyerW	-5.344775
Foundation_Slab	-5.717564
GarageType_Basment	-7.459996
BldgType_Twnhs	-7.844371
MSZoning_C (all)	-8.233326
MSSubClass_160	-8.487843
Neighborhood_MeadowV	-8.636713
Neighborhood_Edwards	-9.353469

Results: Elastic Net Improvable Features



Features	Coefficients
OverallCond	5.318558
Functional	4.597007
BsmtExposure	2.711707
KitchenQual	2.538005
BsmtFullBath	1.710271
BsmtCond	1.652357
Fireplaces	1.284620
HeatingQC	1.202317
BsmtFinType1	1.080270
HalfBath	-3.036817
KitchenAbvGr	-3.435780
FullBath	-3.699374

Random Forest Regression

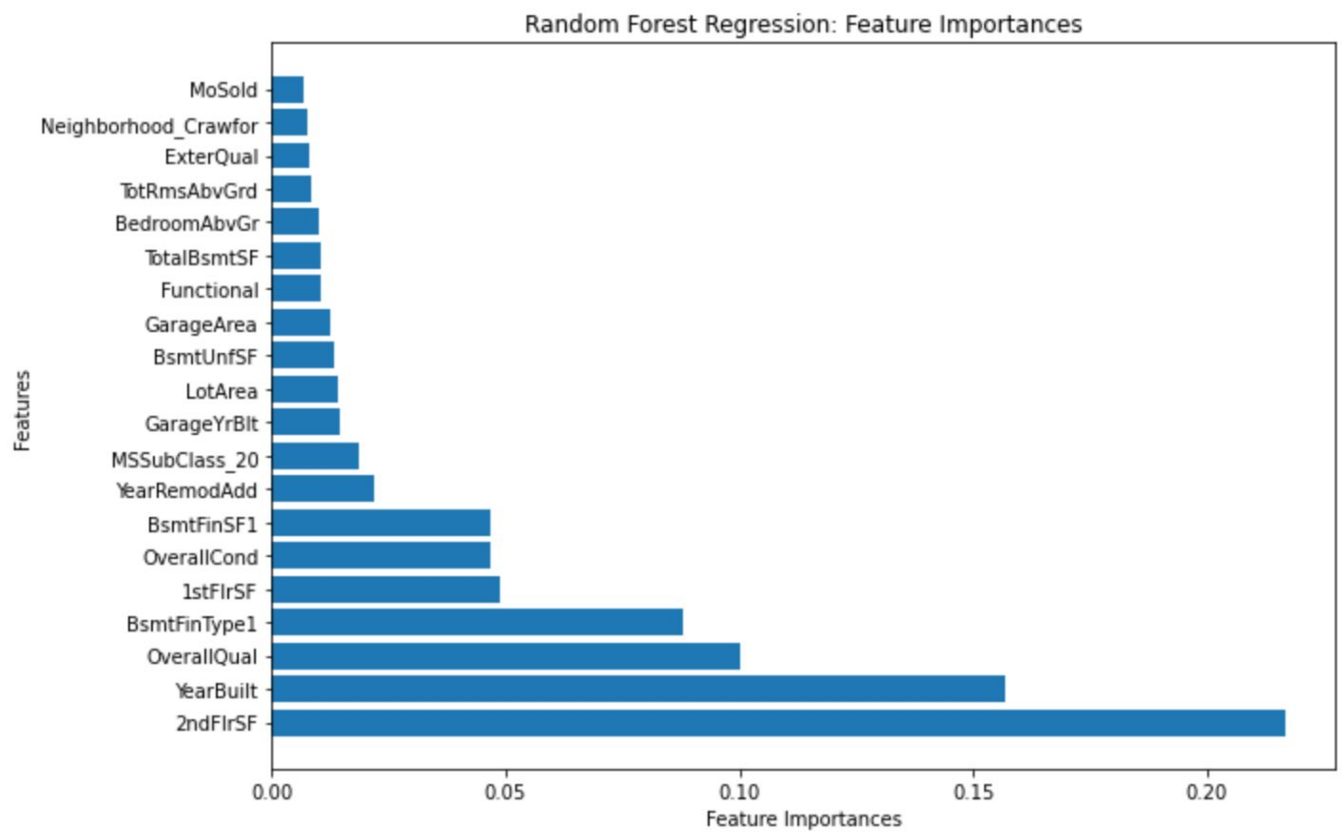
Parameter Tuning:

- N_estimators: Tuned using grid search and affects number of decision trees used
- Max_depth: Tuned using same grid search and affects the maximum depth of each tree

Feature Selection:

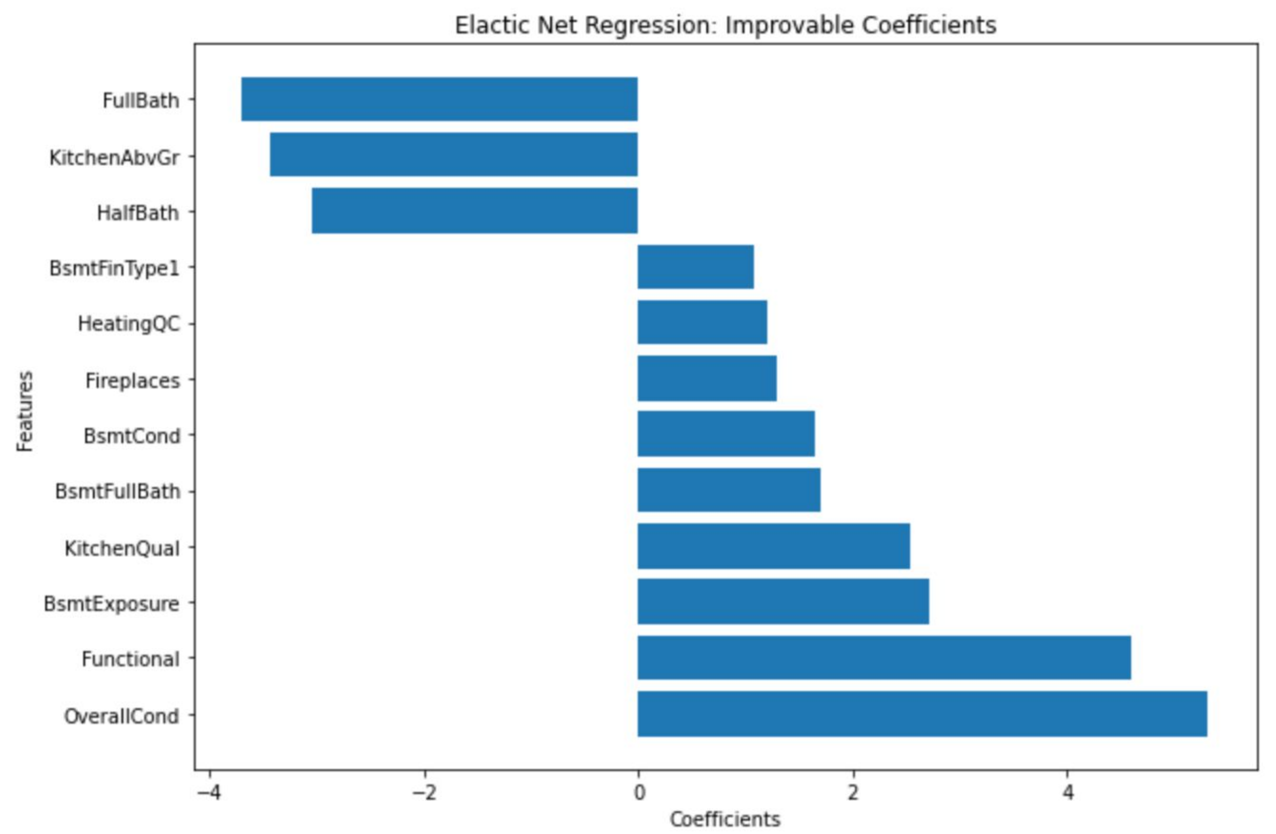
- Model uses all 234 features, with dummified categorical features and label encoded ordinal features

Analysis: Random Forest Summary



Features	Feature_Importances
2ndFlrSF	0.216807
YearBuilt	0.157041
OverallQual	0.100212
BsmtFinType1	0.088017
1stFlrSF	0.048939
OverallCond	0.046907
BsmtFinSF1	0.046684
YearRemodAdd	0.021980
MSSubClass_20	0.018694
GarageYrBlt	0.014367
LotArea	0.014120
BsmtUnfSF	0.013350
GarageArea	0.012318
Functional	0.010583
TotalBsmtSF	0.010246
BedroomAbvGr	0.010176
TotRmsAbvGrd	0.008402
ExterQual	0.007960
Neighborhood_Crawfor	0.007370
MoSold	0.006783

Results: Random Forest Improvable Features



Features	Feature_Importances
OverallQual	0.100212
BsmtFinType1	0.088017
OverallCond	0.046907
BsmtFinSF1	0.046684
YearRemodAdd	0.021980
BsmtUnfSF	0.013350
Functional	0.010583
TotRmsAbvGrd	0.008402
BsmtQual	0.006236
KitchenAbvGr	0.006195
WoodDeckSF	0.005315
OpenPorchSF	0.004338
BsmtExposure	0.004287
EnclosedPorch	0.002800
KitchenQual	0.002668
MasVnrArea	0.002636
Fence	0.002575
BsmtFullBath	0.002523
GarageCond	0.002516
SaleType_New	0.002399

Support Vector Regression: Parameter Tuning

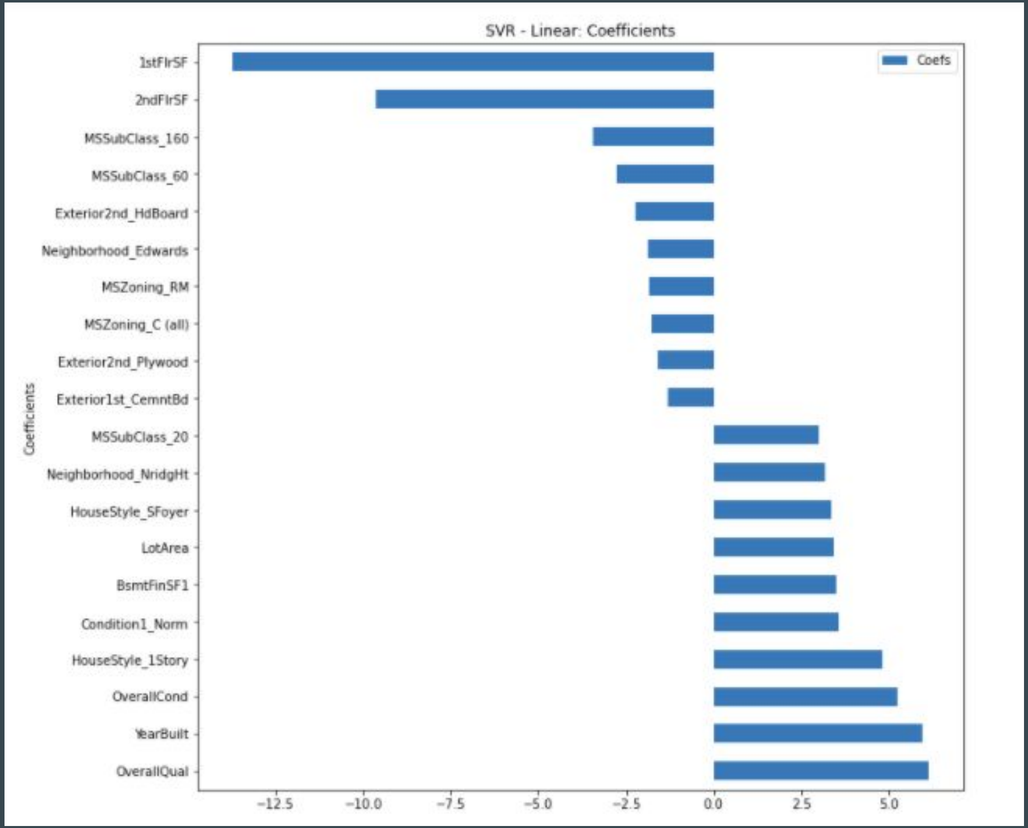
C:

- Tuned using Gridsearch to get the penalty parameter of the error term.

Epsilon:

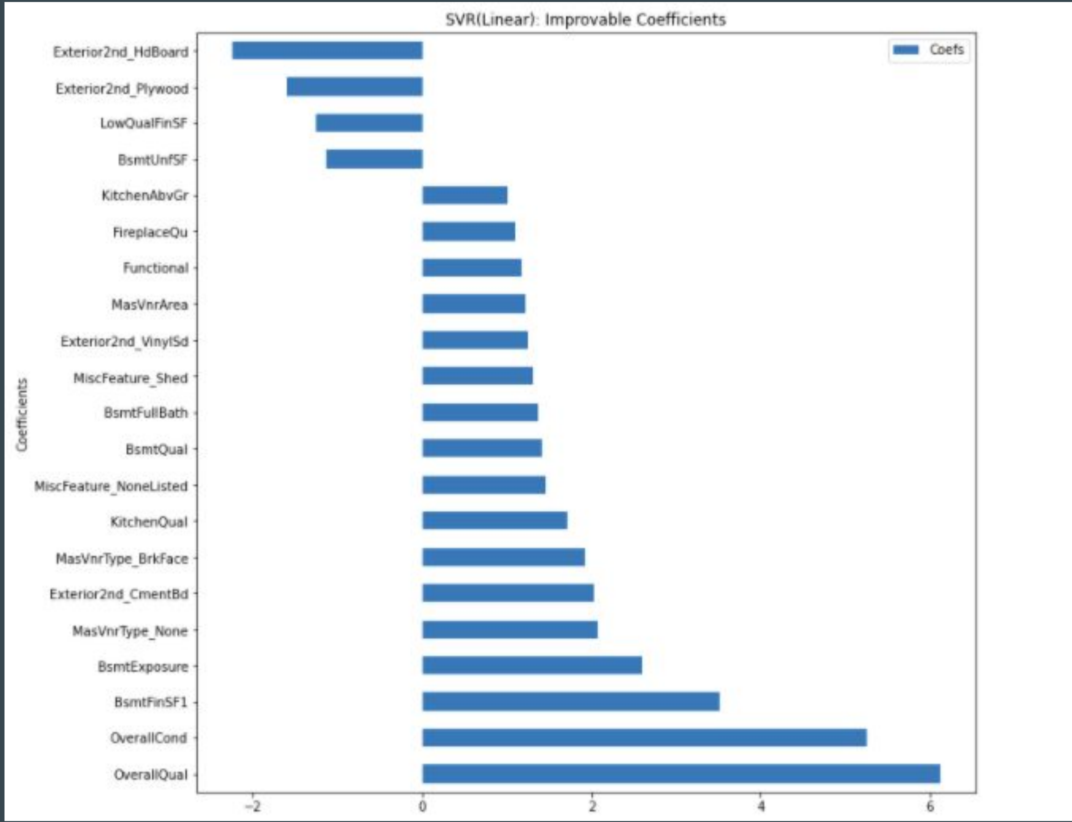
- Same Gridsearch to get the band within which the loss is zero

Analysis: SVR Linear



Features	Coeficients
OverallQual	6.123672
YearBuilt	5.962527
OverallCond	5.253630
HouseStyle_1Story	4.808384
Condition1_Norm	3.564337
BsmtFinSF1	3.511294
LotArea	3.434646
HouseStyle_SFoyer	3.340436
Neighborhood_NridgHt	3.181820
MSSubClass_20	3.000201
Exterior1st_CemntBd	-1.307272
Exterior2nd_Plywood	-1.606852
MSZoning_C (all)	-1.790347
MSZoning_RM	-1.834059
Neighborhood_Edwards	-1.867012
Exterior2nd_HdBoard	-2.243610
MSSubClass_60	-2.787500
MSSubClass_160	-3.446482
2ndFlrSF	-9.628135
1stFlrSF	-13.742777

Results: SVM-Linear Improvable Features



	Features	Coefs
0	OverallQual	6.123672
1	OverallCond	5.253630
2	BsmtFinSF1	3.511294
3	BsmtExposure	2.597847
4	MasVnrType_None	2.077273
5	Exterior2nd_CmentBd	2.027425
6	MasVnrType_BrkFace	1.917210
7	KitchenQual	1.706812
8	MiscFeature_NoneListed	1.459511
9	BsmtQual	1.412387
10	BsmtFullBath	1.363048
11	MiscFeature_Shed	1.304782
12	Exterior2nd_VinylSd	1.245269
13	MasVnrArea	1.220365
14	Functional	1.178834
15	FireplaceQu	1.098790
16	KitchenAbvGr	1.012028
17	BsmtUnfSF	-1.142101
18	LowQualFinSF	-1.259495
19	Exterior2nd_Plywood	-1.606852
20	Exterior2nd_HdBoard	-2.243610

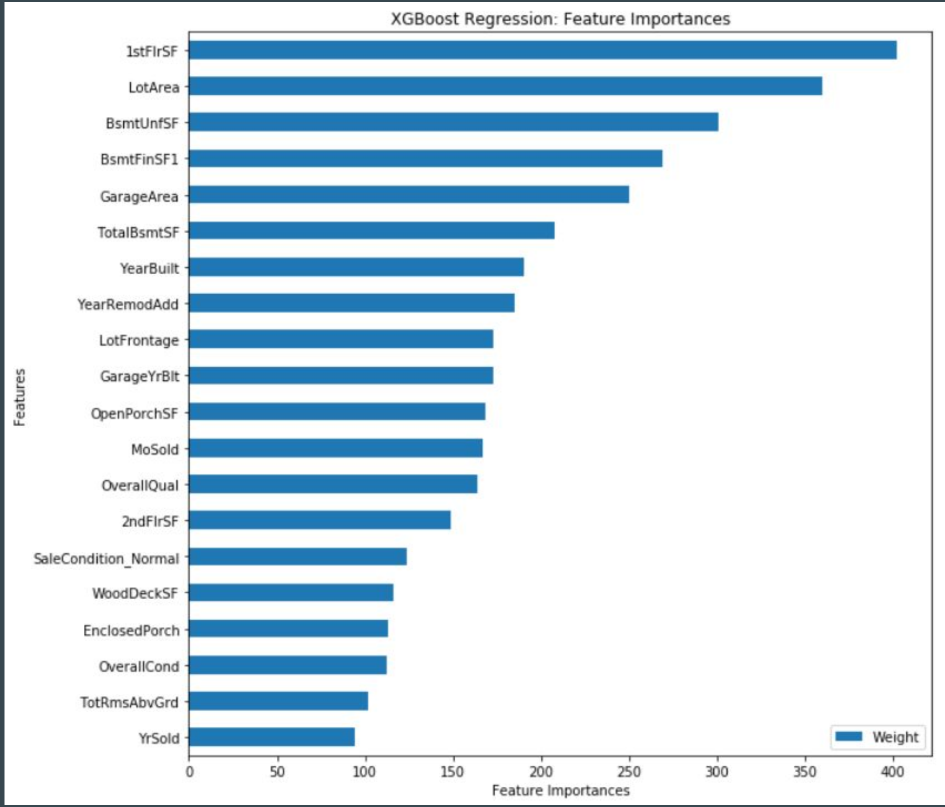
XGBoost Regressor Parameter Tuning

Utilized GridSearchCV to tune the parameters of gbtree*:

- "learning_rate": [0.05, 0.5, 0.75, 1],
- 'n_estimators': [100, 500, 1000]
- 'gamma': [0.5, 1, 1.5, 2, 5],
- 'max_depth': [3, 6, 9],
- 'min_child_weight': [1, 5, 10],
- 'subsample': [0.4, 0.75, 1.0],
- 'colsample_bytree': [0.6, 0.8, 1],
- 'reg_alpha': [1, 2, 3],
- 'reg_lambda': [1, 2, 3]

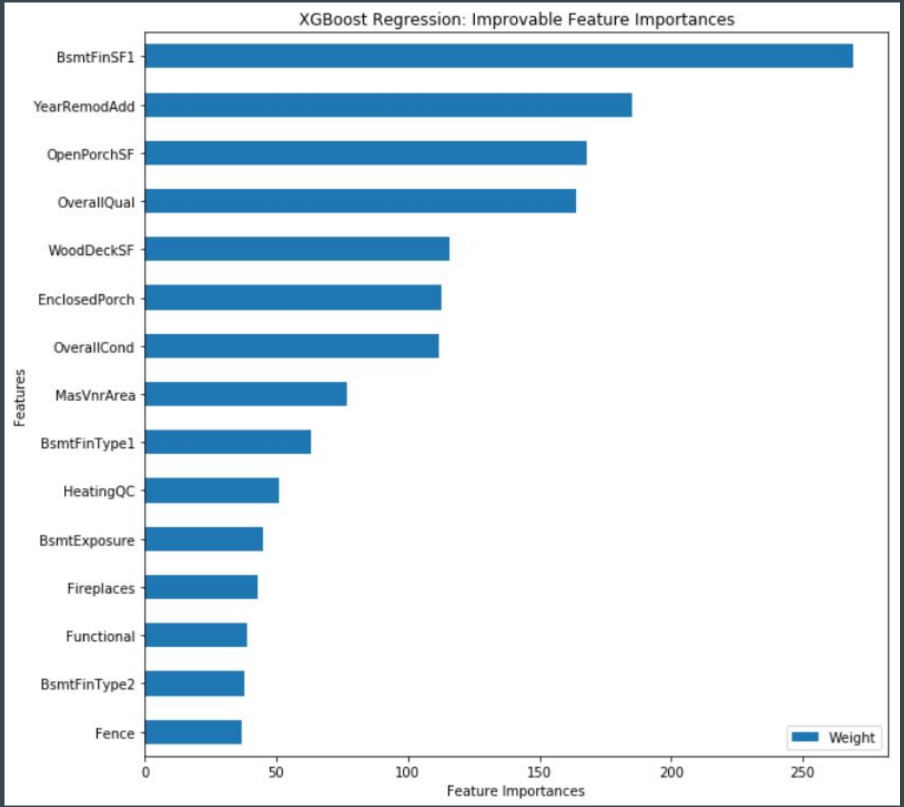
*gblinear and dart had much lower scores with default parameters; were not used

Analysis: XGBoost Feature Importance



	Weight
1stFlrSF	402
LotArea	360
BsmtUnfSF	301
BsmtFinSF1	269
GarageArea	250
TotalBsmtSF	208
YearBuilt	190
YearRemodAdd	185
LotFrontage	173
GarageYrBlt	173
OpenPorchSF	168
MoSold	167
OverallQual	164
2ndFlrSF	149
SaleCondition_Normal	124
WoodDeckSF	116
EnclosedPorch	113
OverallCond	112
TotRmsAbvGrd	102
YrSold	94

Results: XGBoost Improvable Features



	Weight
BsmtFinSF1	269
YearRemodAdd	185
OpenPorchSF	168
OverallQual	164
WoodDeckSF	116
EnclosedPorch	113
OverallCond	112
MasVnrArea	77
BsmtFinType1	63
HeatingQC	51
BsmtExposure	45
Fireplaces	43
Functional	39
BsmtFinType2	38
Fence	37

Results Summary

Linear Models (MLR, Ridge, Lasso, Elastic Net)

- Test scores are below 0.75
- Uses coefficients and is linear
- Provides a clearer picture of the data

Supervised Learning Models (Random Forest, SVR, XGBoost)

- Test scores are above 0.75
- More difficult to describe, but more accurate

Best Features to Improve

We chose the five features that all models valued positively to change so that that home’s SalePricePerGLA would rise, and thereby increase the SalePrice:

- 'OverallQual' : set to 9 (out of 10)
- 'OverallCond': set to 9 (out of 10)
- 'BsmtFinSF1': add 1/2 of BsmtUnfSF
    - Divide BsmtUnfSF by 1/2
- 'Functional': set to 6 (equivalent of Min1)
- 'BsmtExposure': set to 3 (equivalent of Av)

Predicting Improved Sale Price

Predicting Improved Sale Price Overview

1. Gather all the undervalued homes from the bottom 80% (of all homes' GrLivArea)
   a. This bottom 80% represented a more linear relationship between total Sale Price and GrLivArea, as compared to using the entire dataset (as detailed in the original project proposal)
2. Set the top improvable features of all models to near-max value
   a. Assume that half of the home's Unfinished Basement is turned into a Finished Basement
   b. Keep all other features the same
3. Predict a new SalePricePerGLA for the "improved" homes
4. Calculate the new SalePrice using each home's GrLivArea

Defining an Undervalued Home

- We used a naive threshold for determining whether a home was undervalued
    - This threshold would likely be different if we had a domain expert to consult

1. Compute the mean (and stddev) SalePricePerGLA for each neighborhood
    a. We assumed a standard distribution of SalePricePerGLA for each neighborhood
2. Compute the threshold for each neighborhood, which was that neighborhood's mean SalePricePerGLA minus the neighborhood's stddev SalePricePerGLA
3. Keep only the homes whose SalePricePerGLA was less than the threshold

“New” and Improved Homes by Model

From the top 50 homes (by gain in predicted Sale Price) of each model, we had the following overlaps of PIDs:

- XGBoost/Lasso Overlap: 21
- RF/Lasso Overlap: 3
- XGBoost/RF Overlap: 14
- SVR/Lasso Overlap: 15
- RF/SVR Overlap: 1
- XGBoost/SVR Overlap: 17

We believe the homes from the XGBoost/SVR overlap would be the best picks, since those two models had the highest test scores

Conclusion

Best Homes to Improve

These undervalued homes saw the greatest gains in Sale Price:

XGBoost Regression:

PID	GrLivArea	SalePricePerGLA	SalePrice	pred_SPPGLA	pred_SP	gained_SP
923125030	1600	50.94	81500	110.169998	176268.38	94768.38
534427010	1728	49.13	84900	96.230003	166293.32	81393.32
923202060	1771	64.94	115000	110.680000	196006.16	81006.16
905200290	1803	85.97	155000	130.660004	235584.03	80584.03
911102170	1317	30.37	40000	88.870003	117039.77	77039.77

SVR: