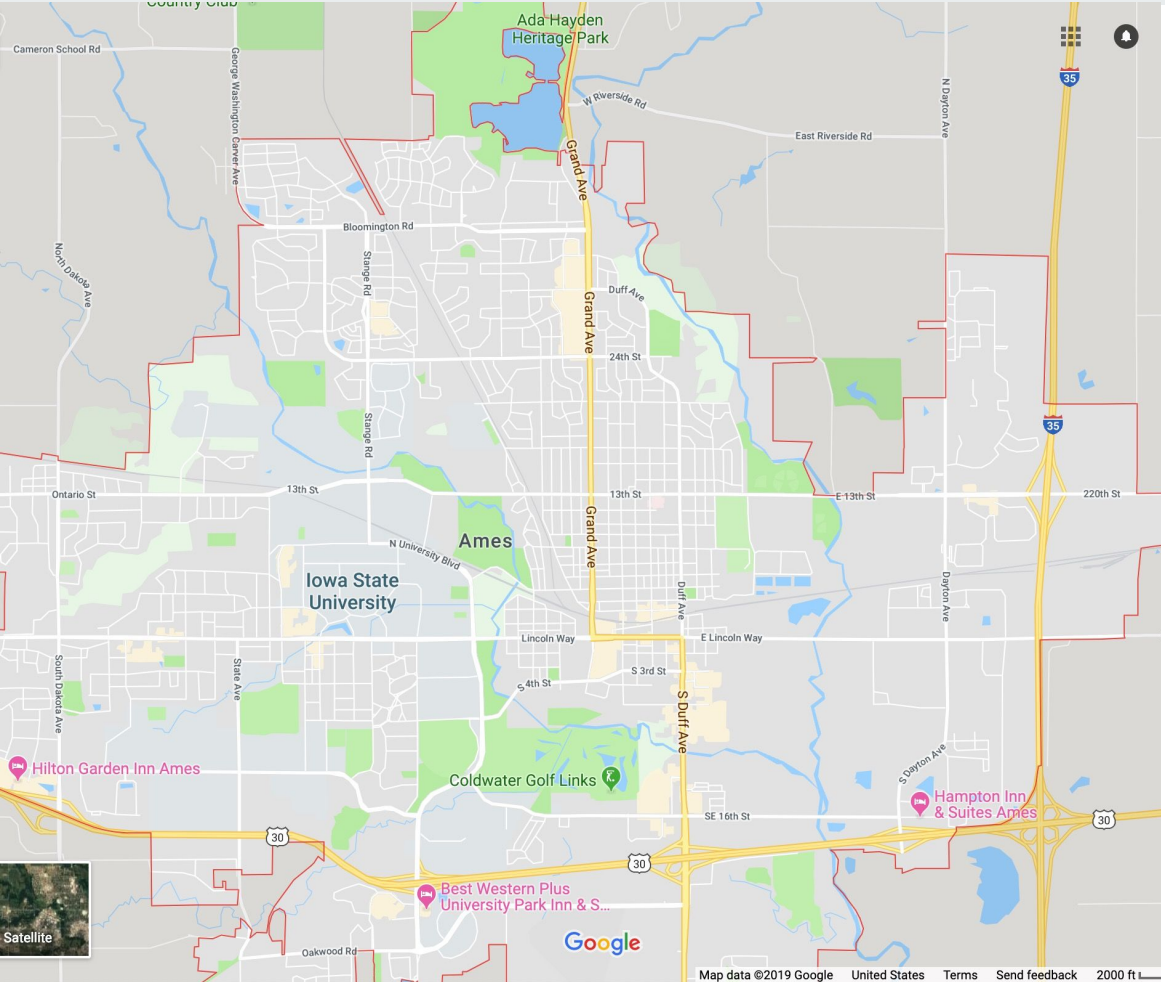
Predicting Housing Prices

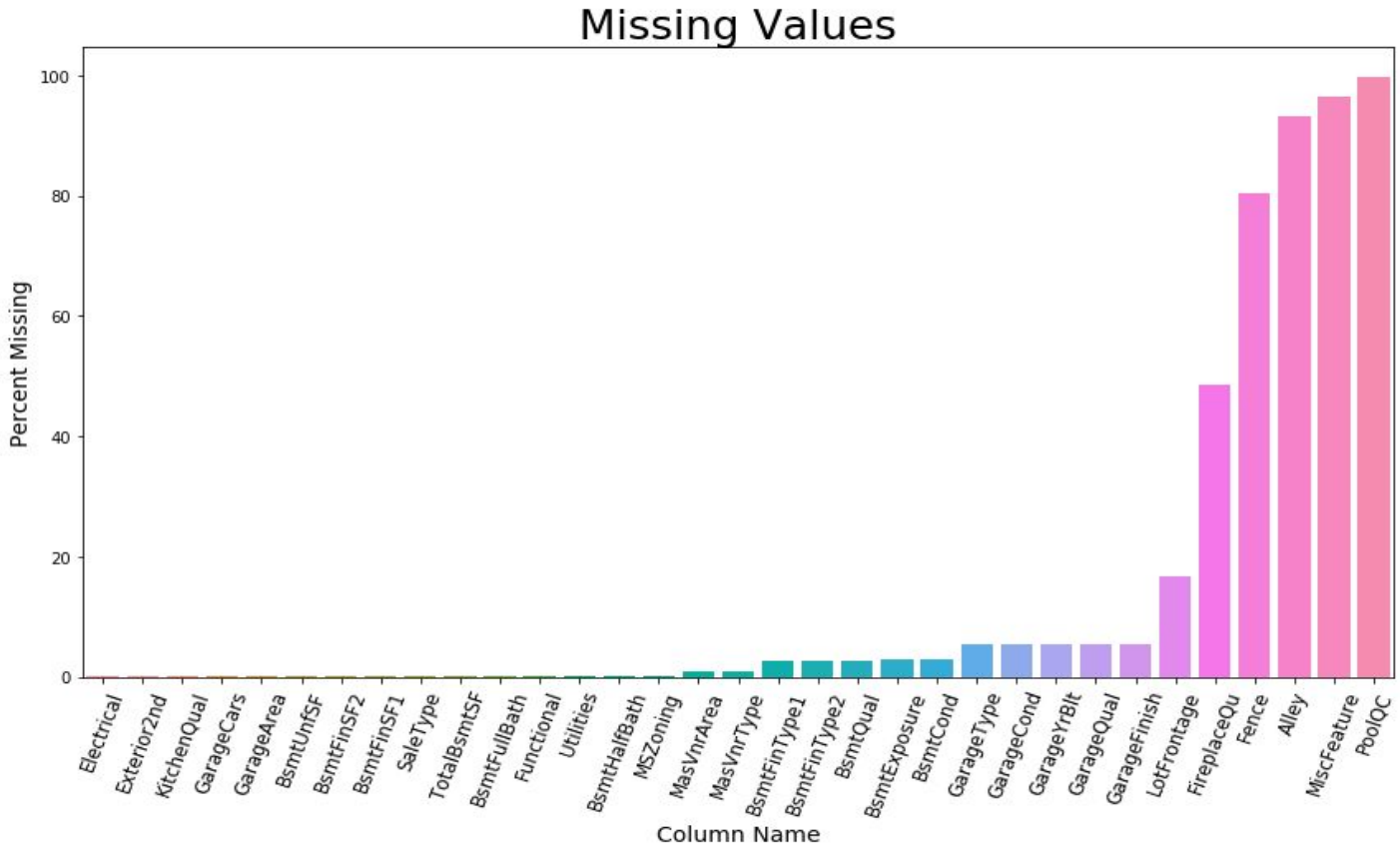
Oluwole Alowolodu, David Levy, Benjamin Rosen



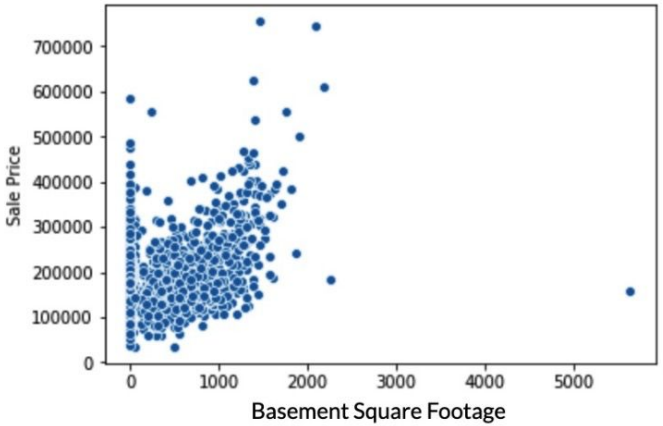
Understanding Ames, Iowa

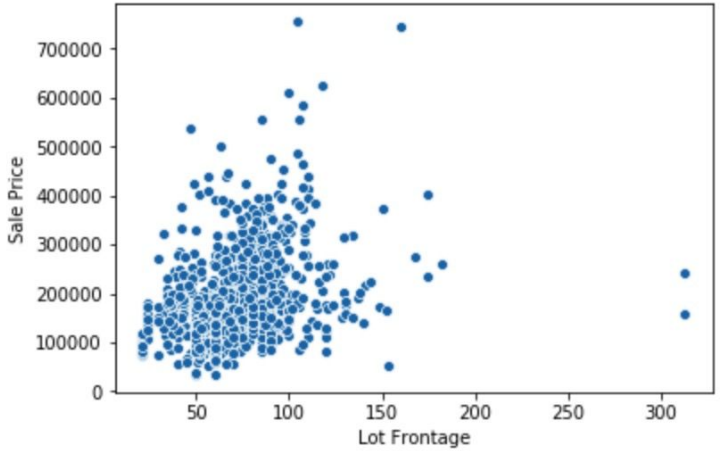
- Important factors other than household characteristics
- Neighborhood is important- parks, bodies of water, railroad, proximity to university, proximity to major streets

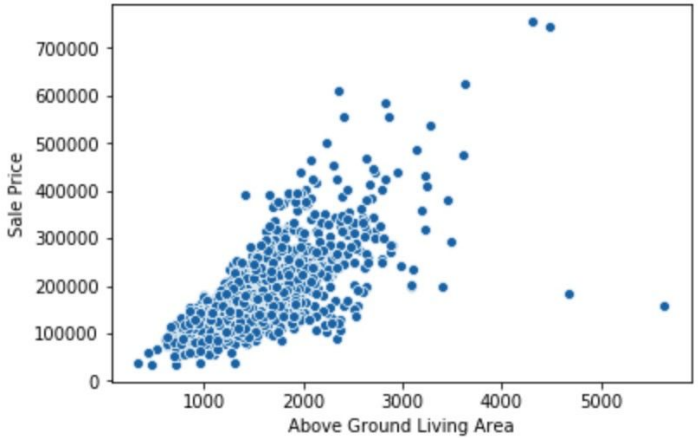
Missingness



Outlier Removal



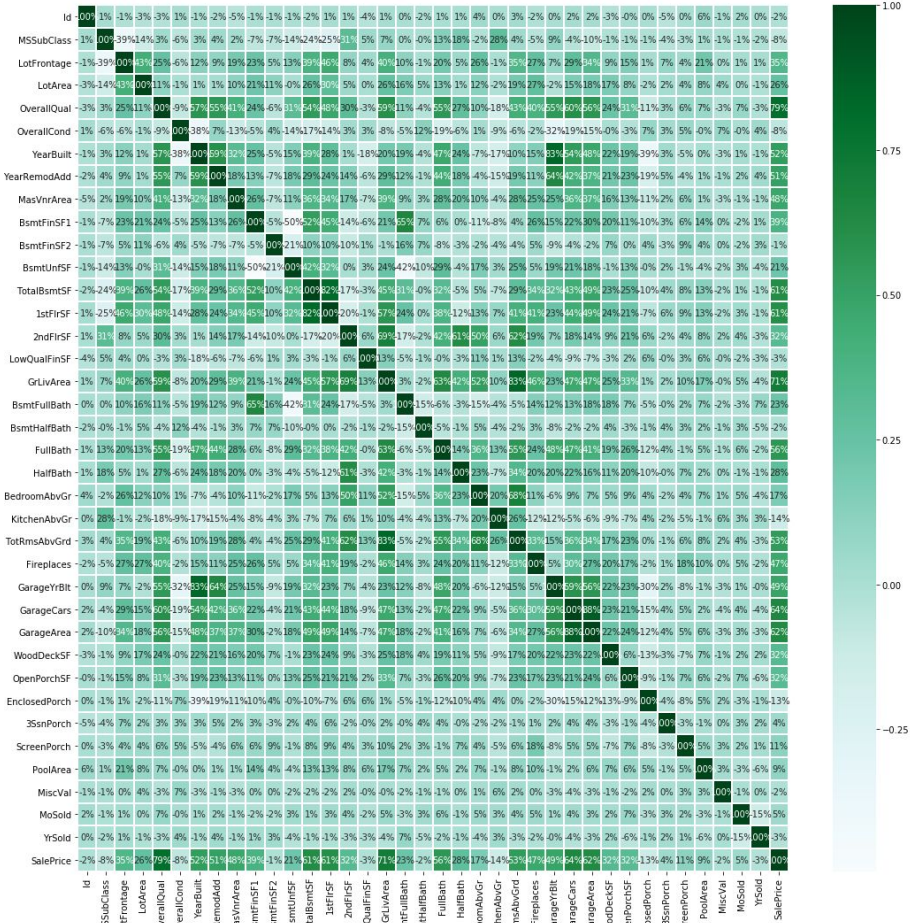




Correlation Matrix

Table showing relationship between features .

Each cell shows percentage correlation between the features.

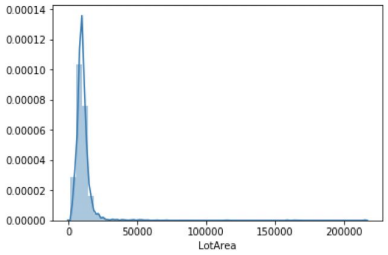


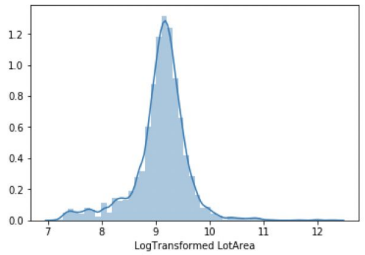
Feature Engineering

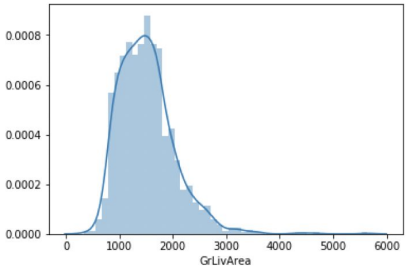
Added: Garage year difference (binary), remodeling(binary & numeric), house built after 1980 (binary), house built before 1960(binary), Lot Frontage (binary), Unfinished Basement (binary), Low Quality Square Footage (binary), Pool Area (binary)

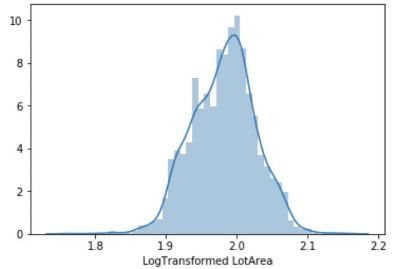
Transformed: Above Ground Living Area(log), First Floor Square Footage(log), Lot Frontage(boxcox), Lot Area (log)

Dropped: Garage Year Built

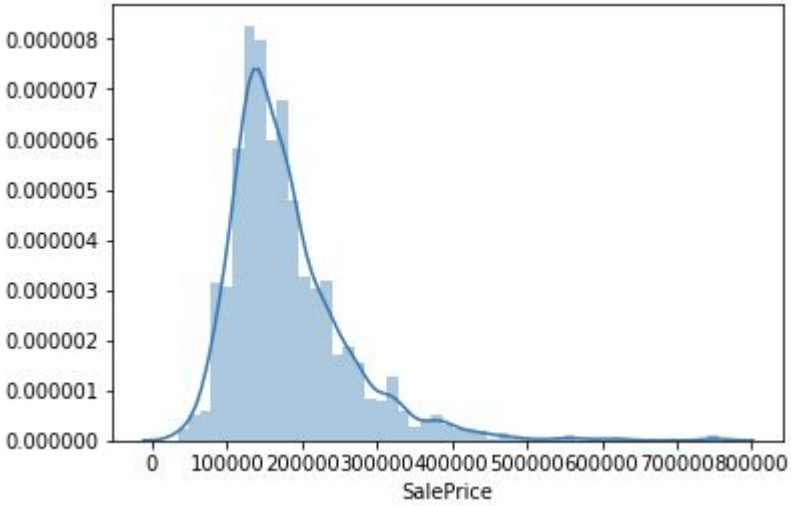


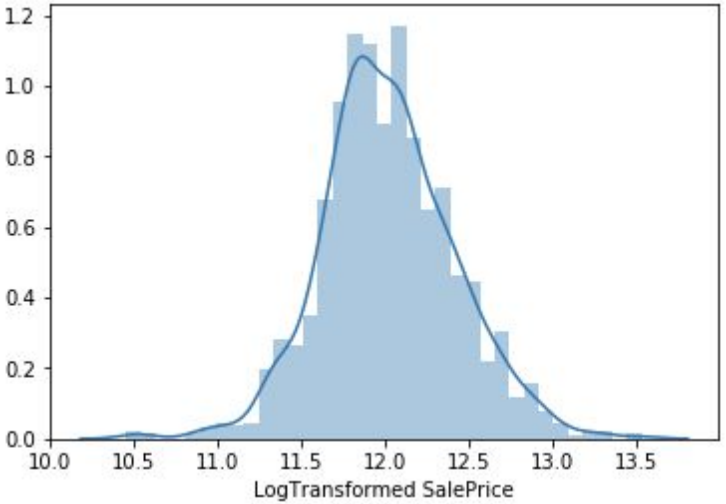




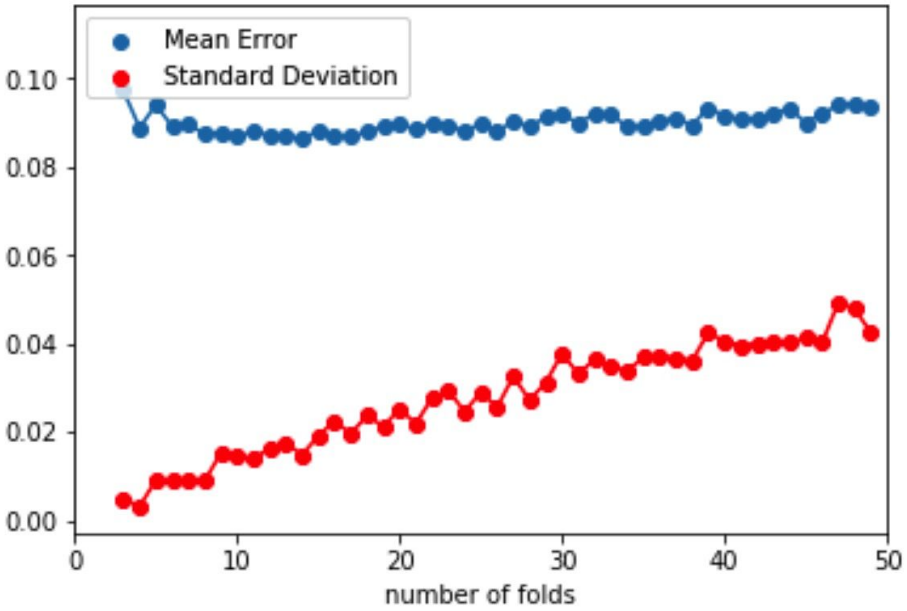


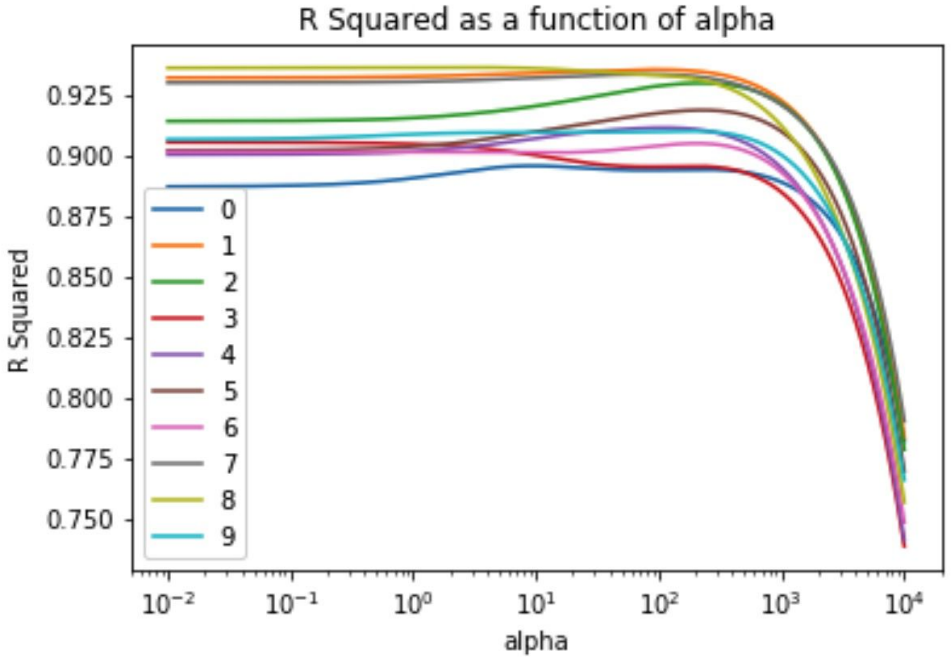
Transformation of target variable





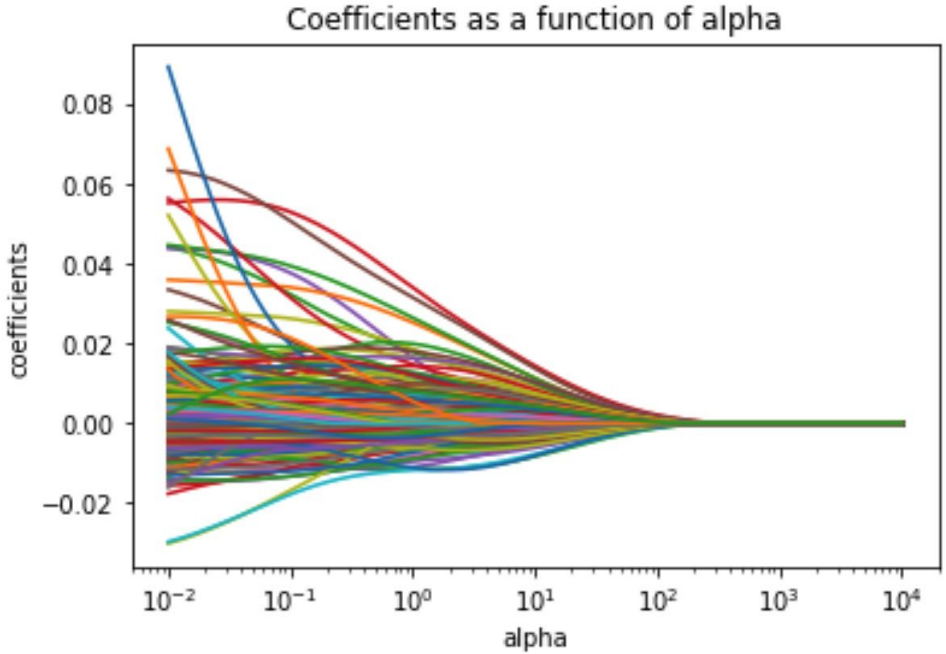
Ridge

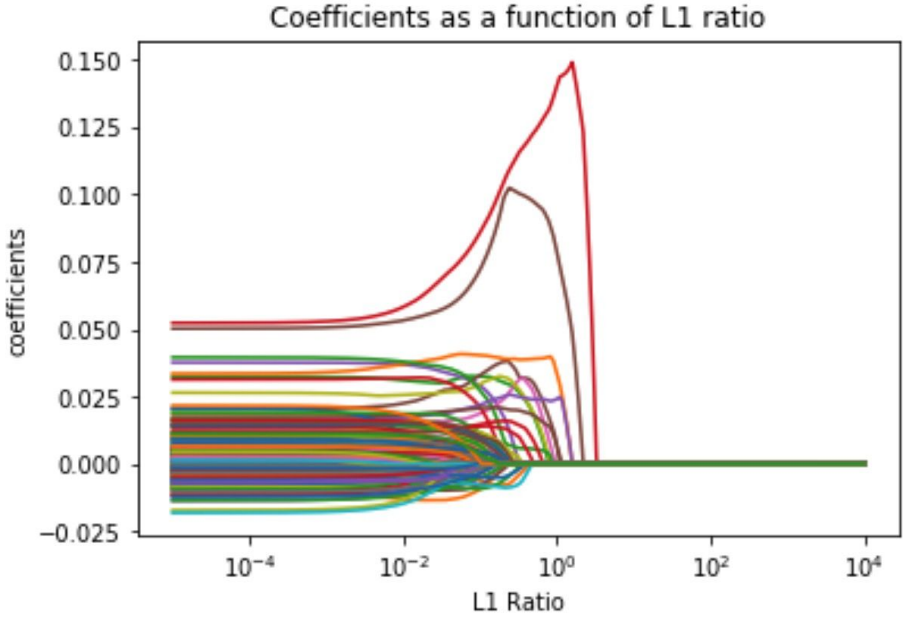

Mean Error
Standard Deviation
0.10
0.08
0.06
0.04
0.02
0.00
0
10
20
30
40
50
number of folds


R Squared as a function of alpha
R Squared
0.925
0.900
0.875
0.850
0.825
0.800
0.775
0.750
0
1
2
3
4
5
6
7
8
9
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
alpha

ElasticNet

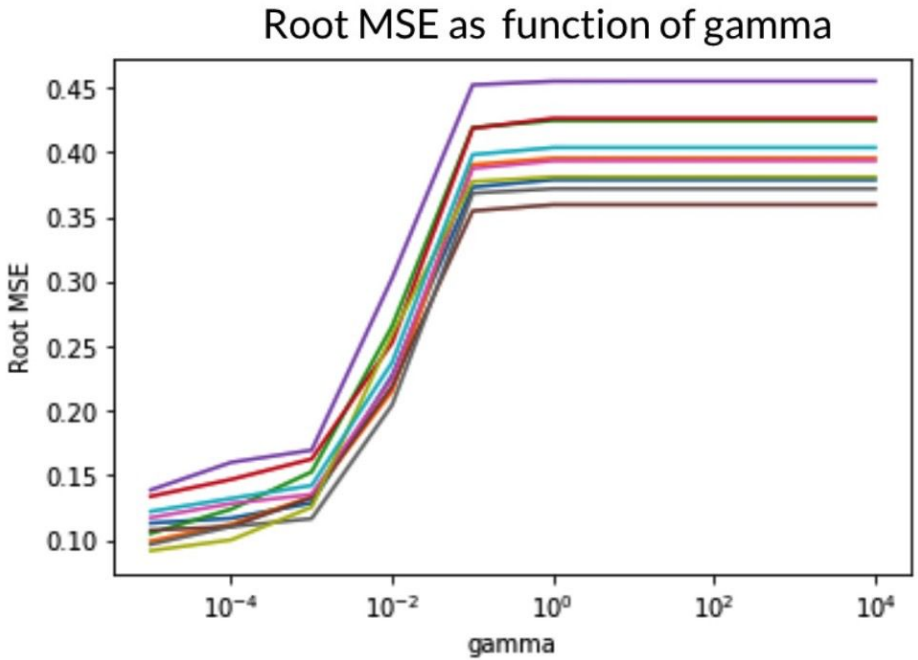
Fit intercept = True, l1_ratio = .001, alpha = .1

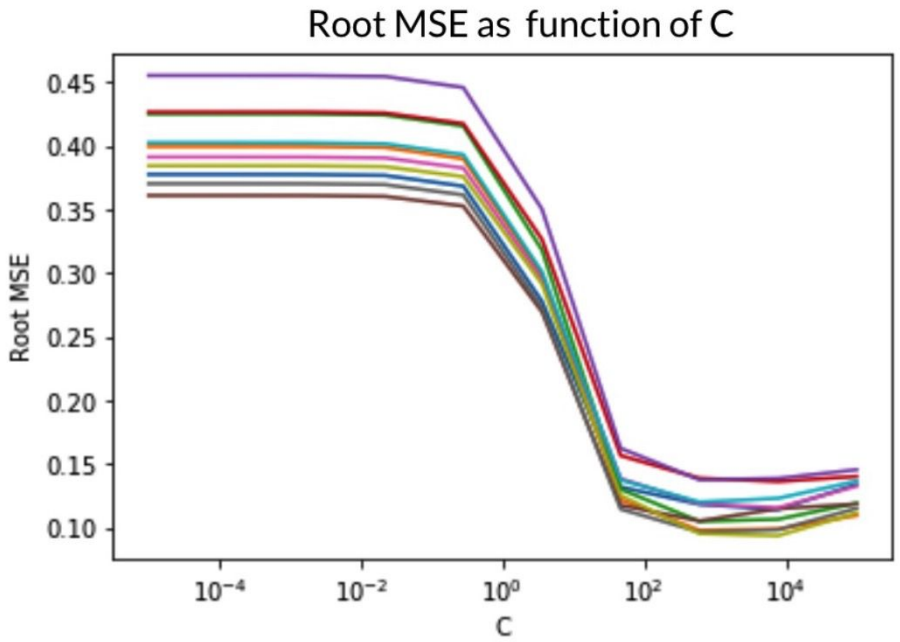




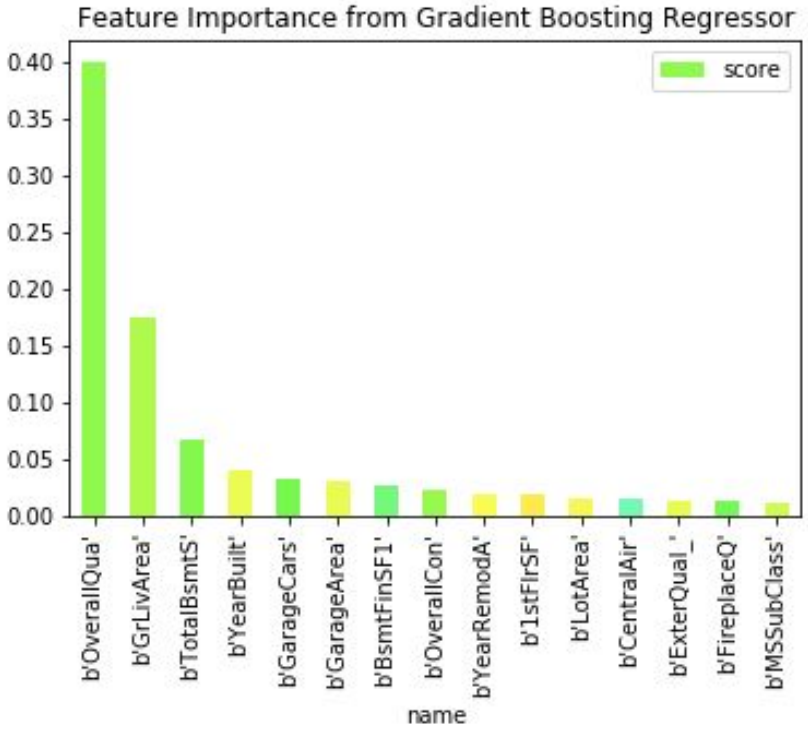
Support Vector Regression

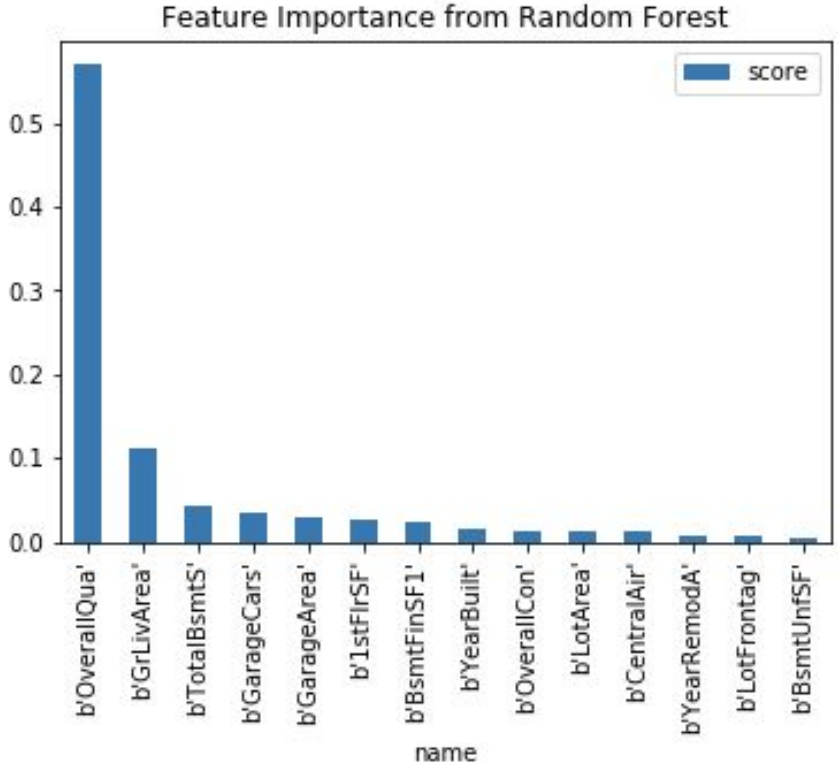
Gamma = .000001, Epsilon = 0, C = 1000





Feature Importance





Model	Parameters	Cross Validation Score	Explained Variance	Kaggle Score
Elastic Net	ALPHA = 0.1 L1 RATIO = 0.001	0.11204	.9209	.12286
Ridge	ALPHA = 10, TOL = 0.00001 NORMALIZE = FALSE, SOLVER = 'SVD'	0.113315	.9418	.11876
Lasso	ALPHA = 0.006 TOLERANCE = 0.107	0.11494	.9178	.12042
Gradient Boosting	ESTIMATORS = 400,SUBSAMPLE =0.7, MAX DEPTH = 2 , MIN LEAFS = 2 , MIN SAMPLE SPLIT = 3	0.12020	.9108	.12761
XG Boost	ESTIMATORS = 400, GAMMA = 0, MAX DEPTH = 3, SUBSAMPLE = .8, LAMBDA = 1, ALPHA = 0	0.11872	.9125	.12802
Support Vector Regressor	EPSILON = 0, GAMMA = .000001, C = 1000	0.11355	.9191	.12359