Housing Price Prediction Using Machine Learning

Chung-Hsuan Huang
Kate Treadwell
Laura Elliott

Outline

- Introduction
- Motivation (What do we want to know?)
- Preprocessing
- Model fitting
- Conclusion
- Future Work

Predicting the Real Estate Market of Ames, Iowa

Modeling the sales price of housing in Ames, Iowa:

- Exploratory Data Analysis (EDA)
  - What does our data look like (types, missingness, summary statistics, etc)
  - How do we better understand how our data works together (covariance, collinearity, duplicate information, etc.)
- Data Preprocessing
  - Cleaning and Imputing Data
  - Removal of outliers, feature selection, etc
- Model Selection
  - Determining best model
  - Lowering RMSE

Motivation

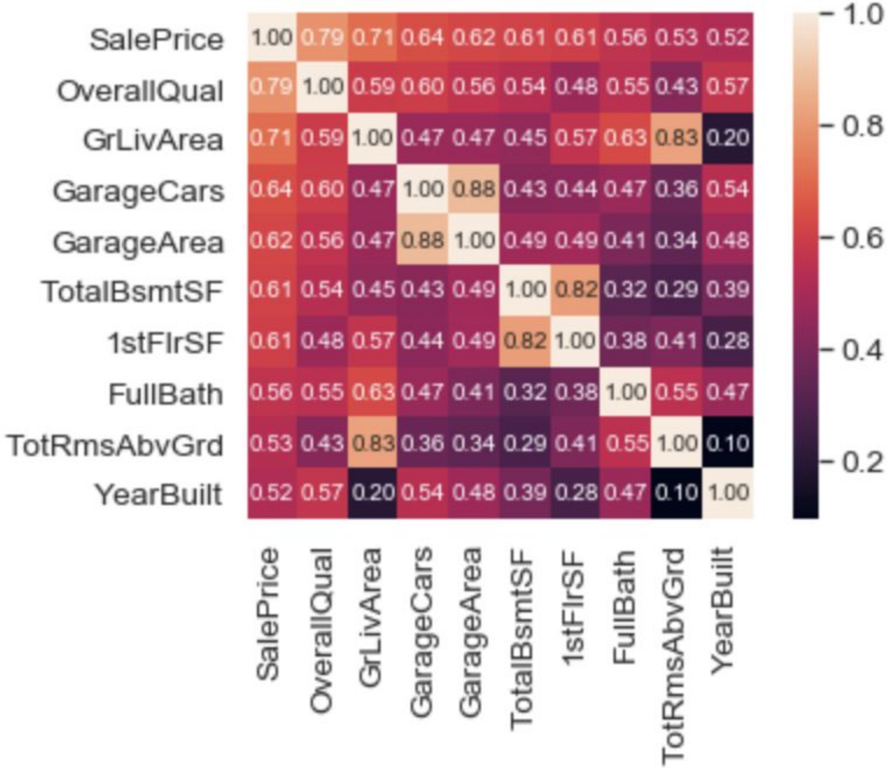
- Derive the lowest possible RMSE for housing sale prices
- Understanding Data Manipulation
    - How do we best manage and clean our data in order to drive the best models
    - How do we understand the results of our manipulation
- Understanding Grid Search and Model Selection
    - How can we go about choosing our best hyperparameters
    - How can we understand the effect of cross validation on our models
    - What do the results of our fitted models mean and how can we interpret them correctly
- Refining our Models
    - What do we change in order to decrease our RMSE
    - Revisiting data manipulation
    - Changing hyperparameters, grid search, and choice of actual models to compare results

Preprocessing - data exploration

Correlation Analysis Top 10

Note other high correlations between…

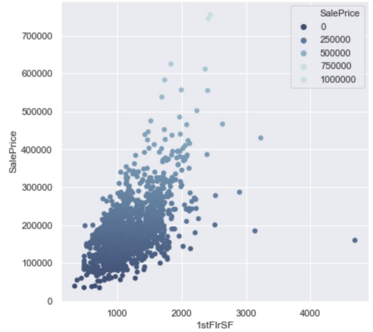
- **GarageArea** and **GarageCars** (r = 0.88)
- **1stFlrSF** and **TotalBsmtSF** (r =0.82)
- **TotalRmsAbvGrd** and **GrLivArea** (0.83)



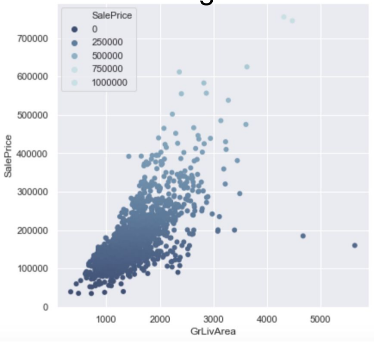
Relationship with Outcome

Top ten continuous variables

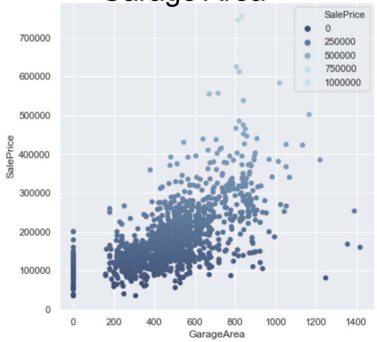
First Floor SF



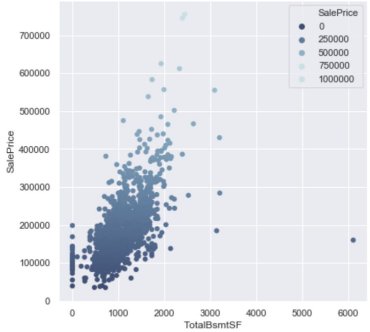
Gr Living Area



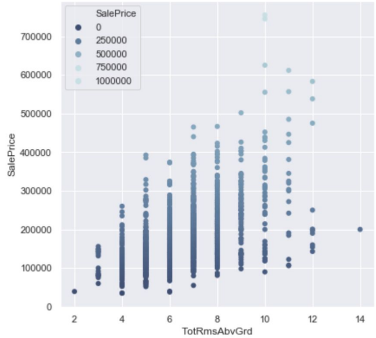
Garage Area



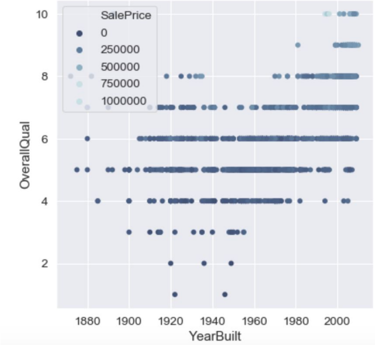
Total Basement SF



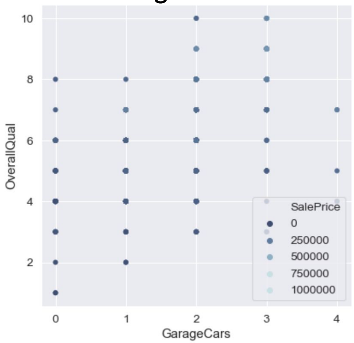
Total Rooms Above Ground



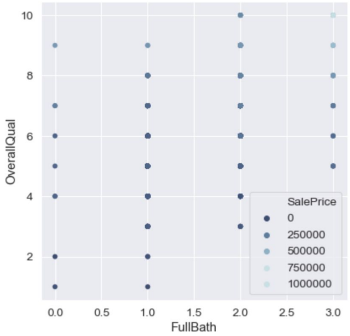
Year Built



Garage Cars



Full Baths

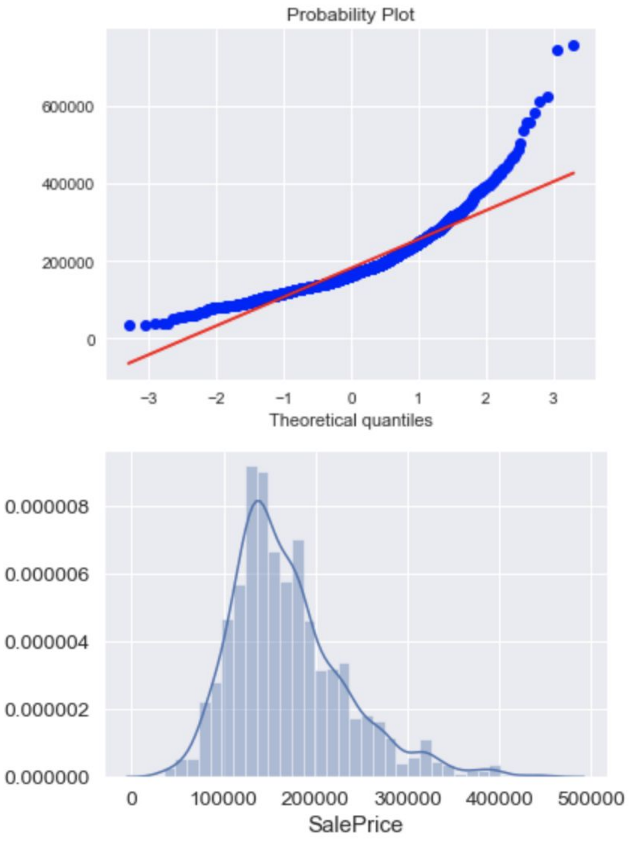


Preprocessing - remove outliers

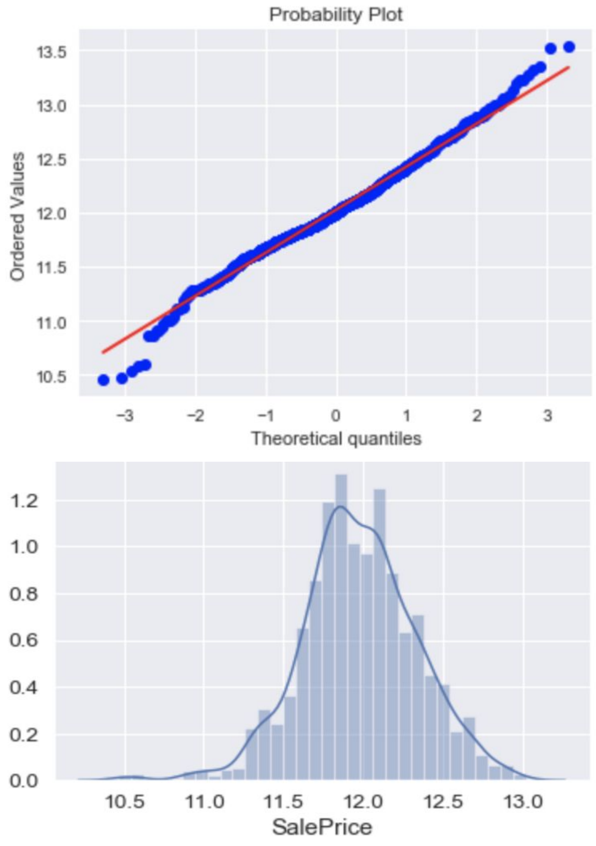
Other than those two points, we can use z-score to systematically remove outliers.

$$z = \frac{x - \mu}{\sigma}$$

Preprocessing - Normalize Sale Price



Logarithm



Preprocessing - Missing data

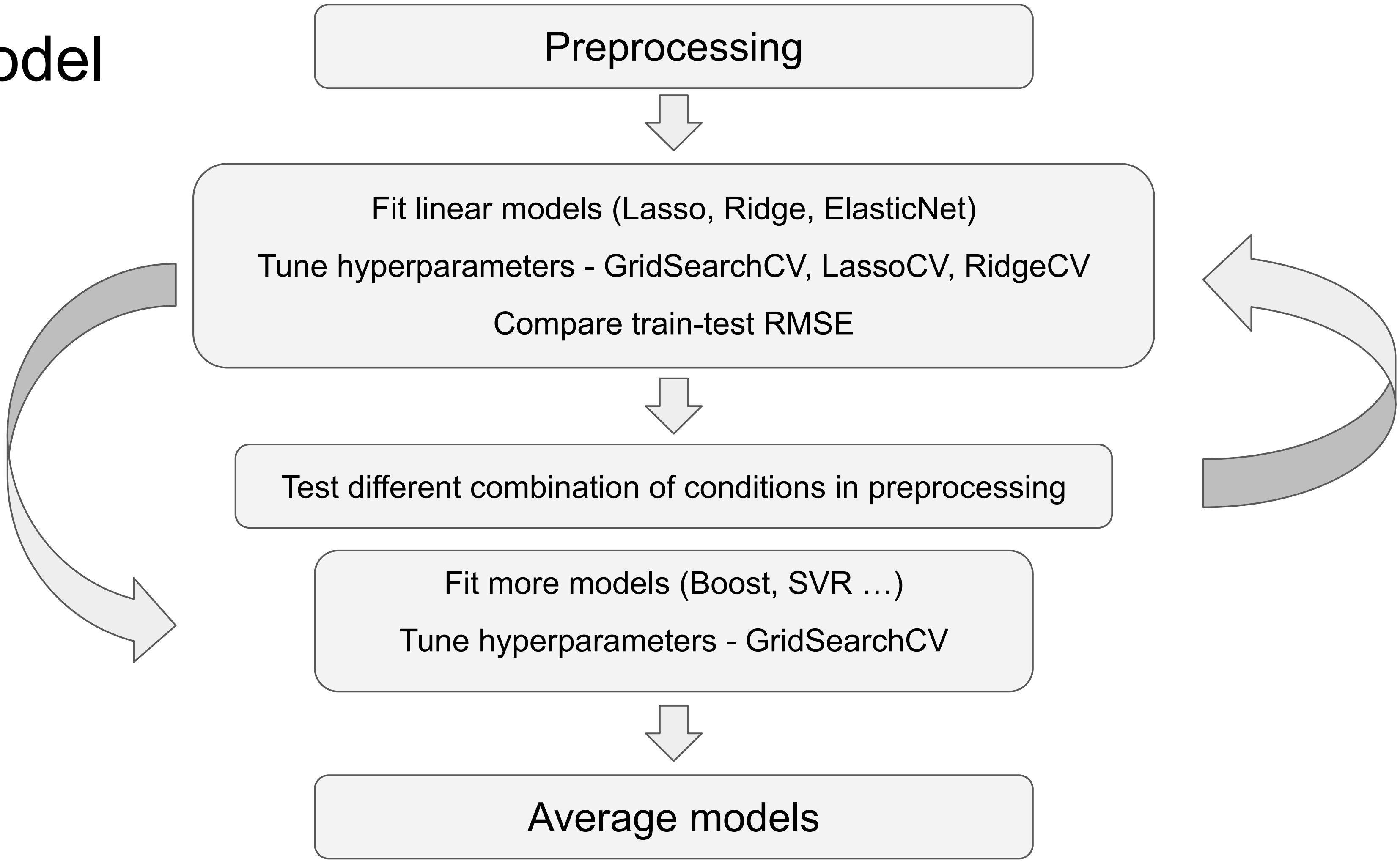
	Perc	Sum
PoolQC	0.997866	2806
MiscFeature	0.964083	2711
Alley	0.931366	2619
Fence	0.802987	2258
FireplaceQu	0.501778	1411
LotFrontage	0.166430	468
GarageYrBlt	0.055832	157
GarageFinish	0.055832	157
GarageQual	0.055832	157
GarageCond	0.055832	157
GarageType	0.055121	155
BsmtExposure	0.028805	81
BsmtCond	0.028805	81
BsmtQual	0.028450	80
BsmtFinType2	0.027738	78
BsmtFinType1	0.027738	78
MasVnrType	0.007824	22
MasVnrArea	0.007468	21
MSZoning	0.001422	4
Functional	0.000711	2

- Greater than 50% missing, however, the NaN values were meaningful
- Imputed ‘None’ for NaN

- ‘LotFrontage’ NaN was imputed with the median, grouped by Neighborhood
- ‘GarageYrBlt’ NaN imputed with mode

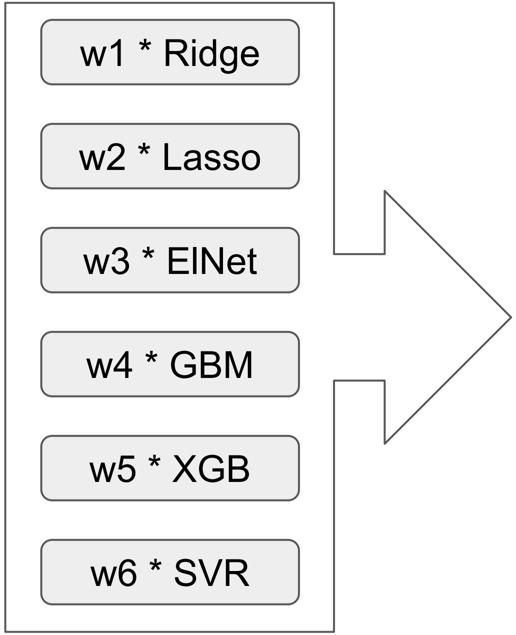
- Imputed with ‘None’

Model


Preprocessing
Fit linear models (Lasso, Ridge, ElasticNet)
Tune hyperparameters - GridSearchCV, LassoCV, RidgeCV
Compare train-test RMSE
Test different combination of conditions in preprocessing
Fit more models (Boost, SVR …)
Tune hyperparameters - GridSearchCV
Average models

Averaged Model

1. List[predicted price from models]
2. Average models



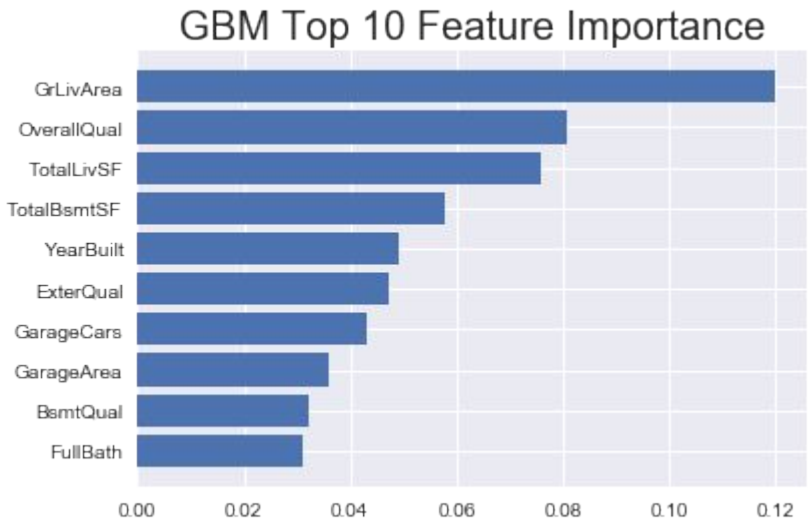
Calculated a new price

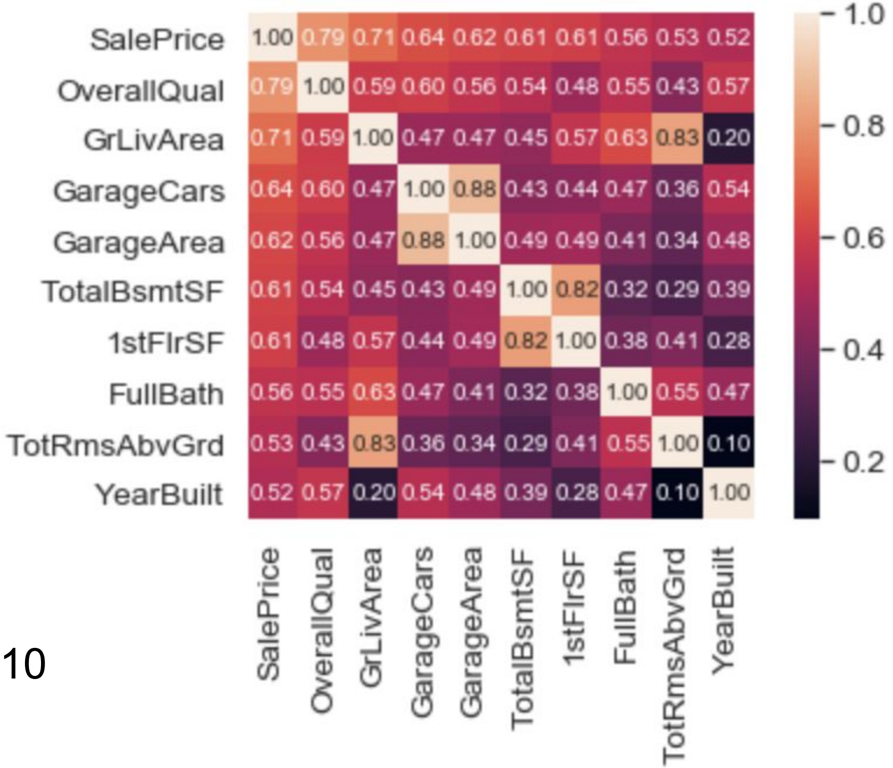
Results

	score_grid	RMSE	train_RMSE	test_RMSE	diff_RMSE	Kaggle_score
Ridge	0.940449	0.097509	0.097555	0.109320	-0.000415	0.11866
Lasso	0.939068	0.098633	0.098876	0.107920	0.009044	0.11938
ElNet	0.939095	0.098611	0.097866	0.109133	0.011266	0.11926
GBM	0.976308	0.061504	0.058831	0.115570	0.056739	0.12485
XGB	0.964267	0.075532	0.074912	0.115308	0.040396	0.12386
SVR	0.927498	0.107591	0.109351	0.112350	0.002998	0.12388

- GBM and XGB might require more parameter tuning to perform better
- The averaged model gave us best Kaggle score: 0.11653

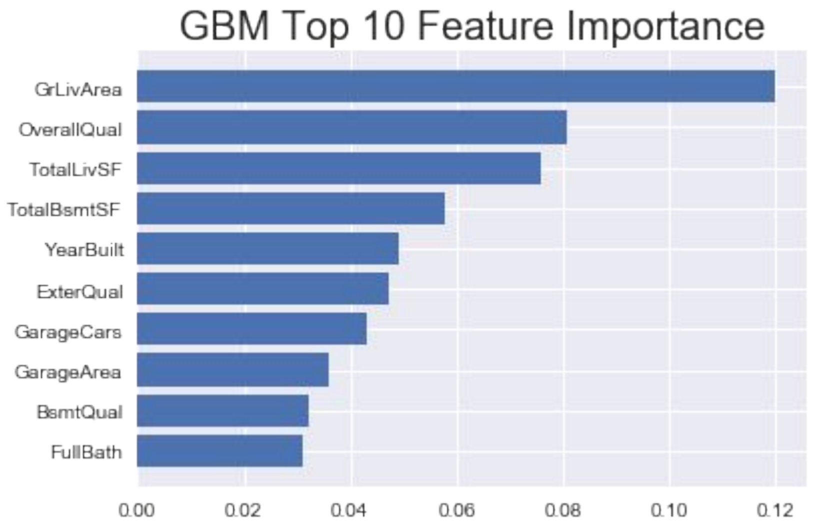
Feature importance

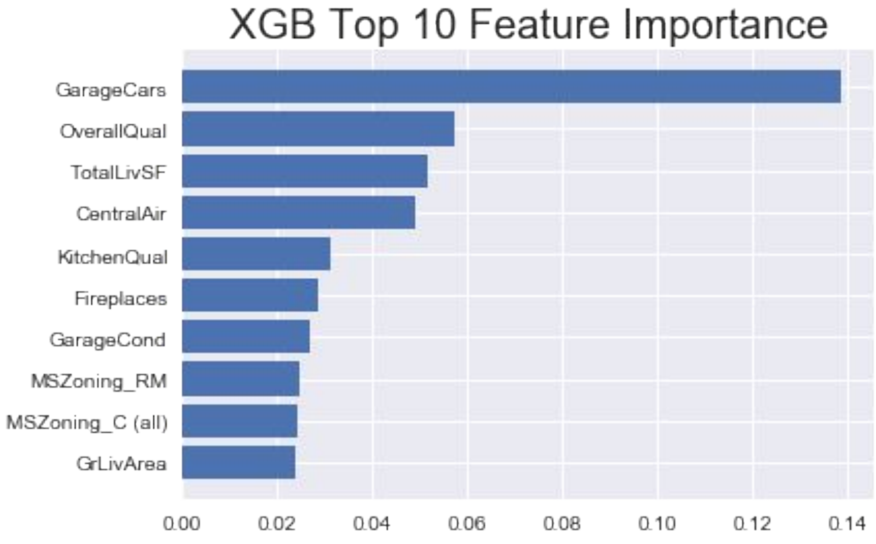




GBM top 10 features are consistent with the top 10 features from correlation analysis

Feature importance





- Some XGB top 10 features are different
- More hyperparameters tuning might be needed

Summary

- Our Best Score: .11653
  - Ensembling best models: Ridge, Lasso, ElasticNet, GBM, XGB, SVR
- Largest factors in reducing our model:
  - Imputation
  - Outlier removal
  - Normalization of data
  - Feature engineering
  - Hyperparameter tuning
- Best overall stand-alone model
  - Ridge

Future work

- Continued Tuning of Models
  - Testing additional models
  - Tuning hyperparameters
  - Feature engineering
- Adding additional data and incorporating into models
  - Economic data
  - Unemployment
  - Political unrest factors
  - Weather data
- Changing methods
  - Alternating different uses imputation
  - Ensembling different models