Predicting Housing Prices with Machine Learning

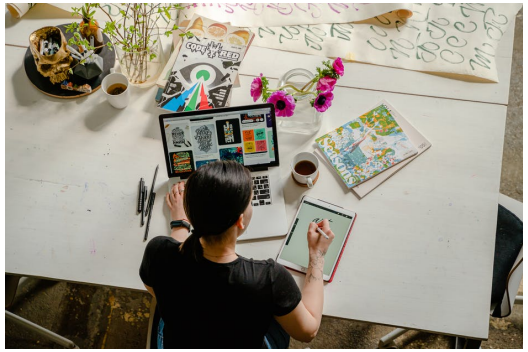
JOHN KOSMICKE, CHERIE WANG, TYLER KIM

Introduction

1. Buying a home is likely the largest and most important purchase a typical American would make in their life. Using data science and machine learning techniques, we will make predictions of home prices in Ames, Iowa according to several different methodologies.
2. Specifically, we consider a couple different linear methods and tree-based methods, and evaluate and adjust them based on the particulars of each methodology.
   - We do L1-regularization of the standard multiple linear regression (LASSO), with the intention of preventing model overfitting.
   - We also add a term structure adjustment to a standard Random Forest again hoping to improve the model accuracy.
   - We also compare accuracy and feature importance of a Gradient Boost tree model with a standard Random Forest
3. In addition, we also look use some neighborhood distance-based aggregation methods to get a sense of the demographic history of Ames housing development.

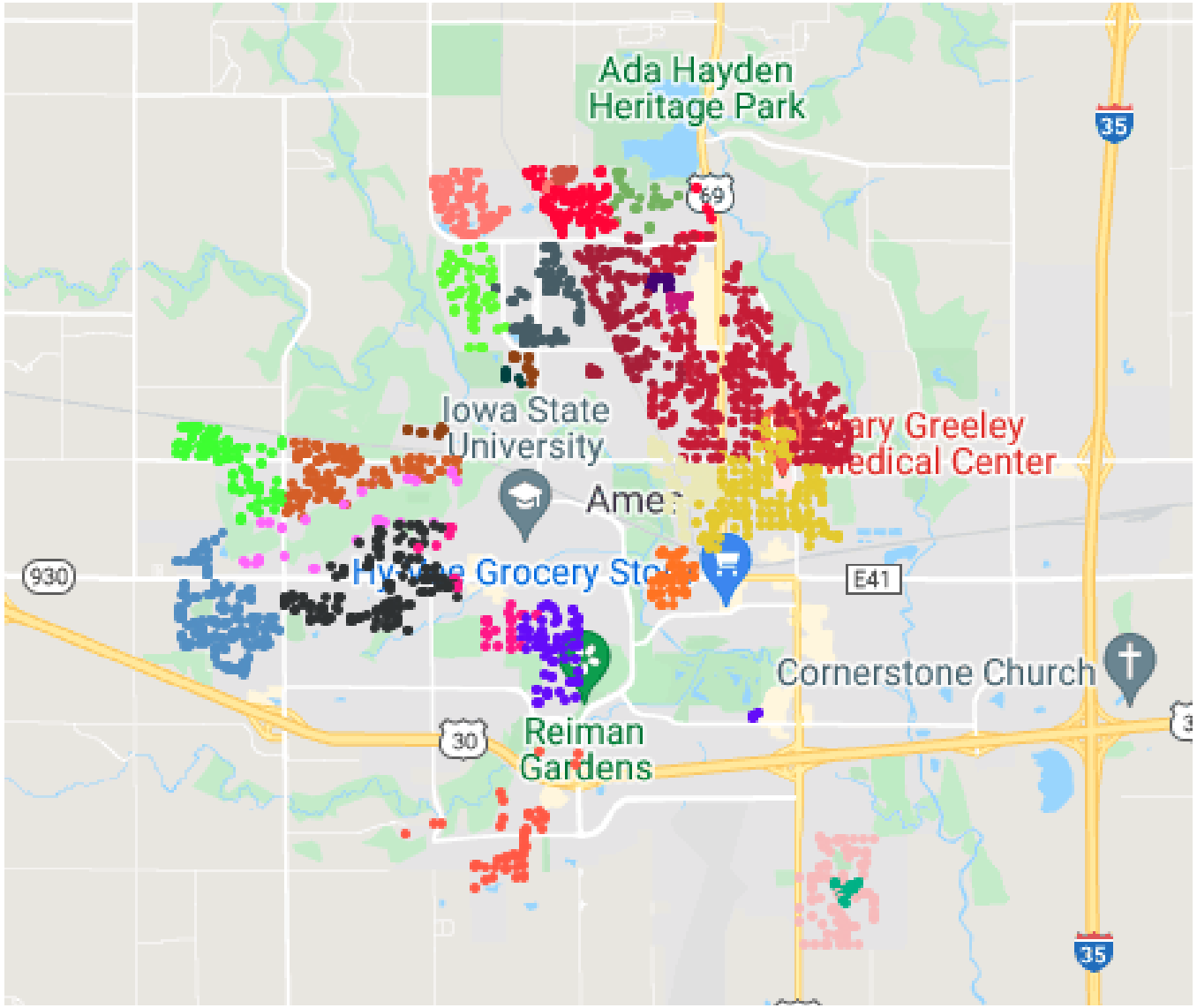
Predict home prices in Ames, Iowa

Identify the best methodologies



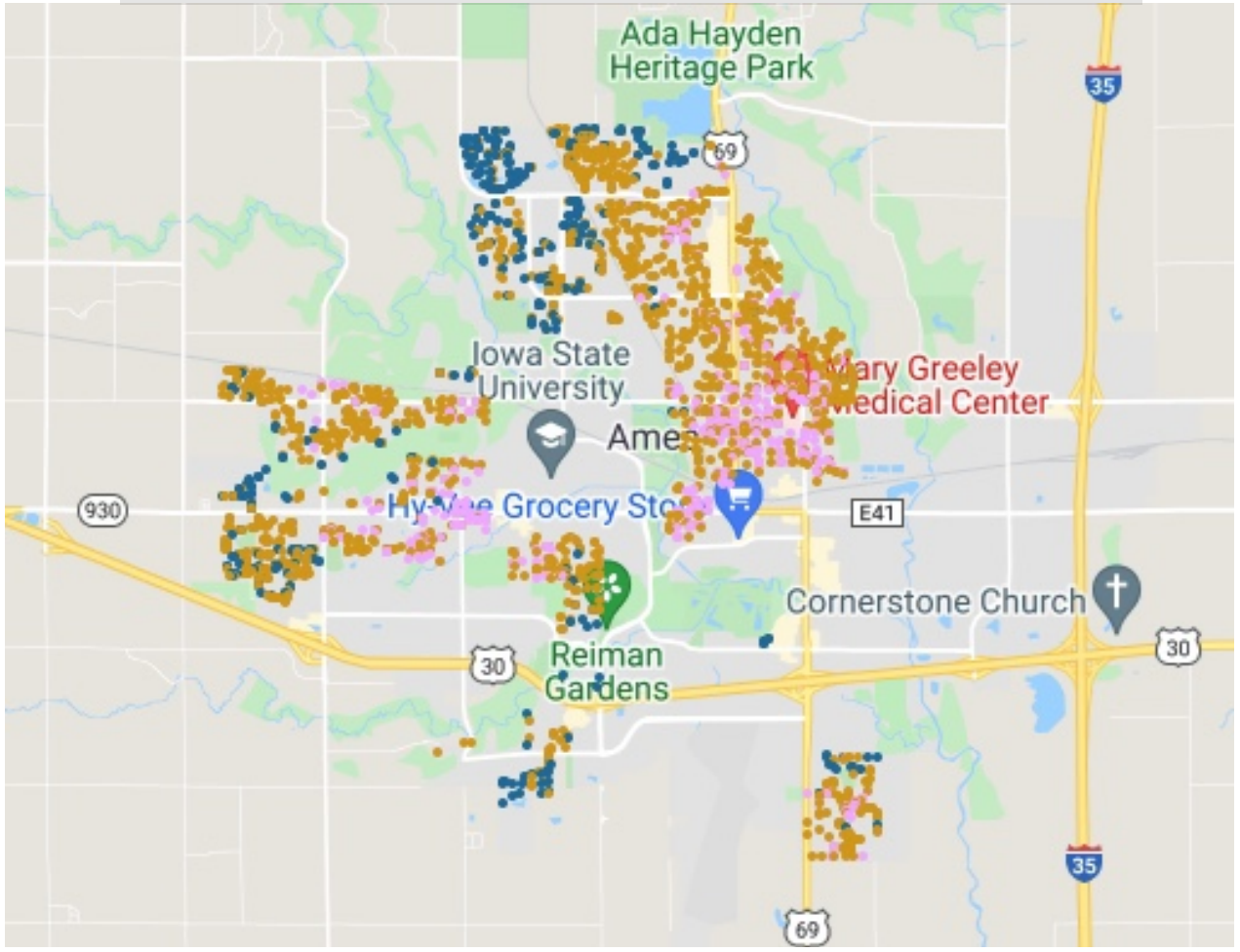
Implement machine learning models

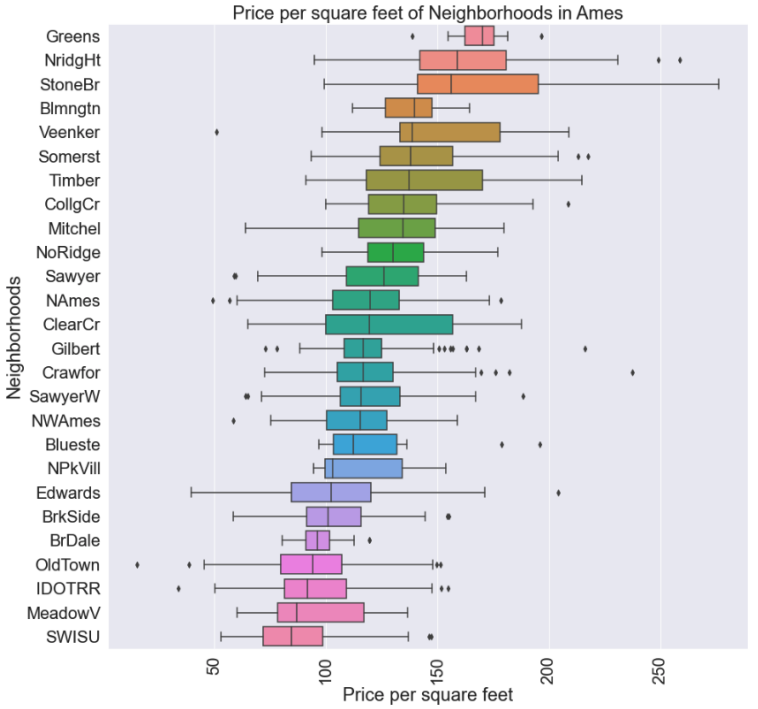
Ames Neighborhood Demographics

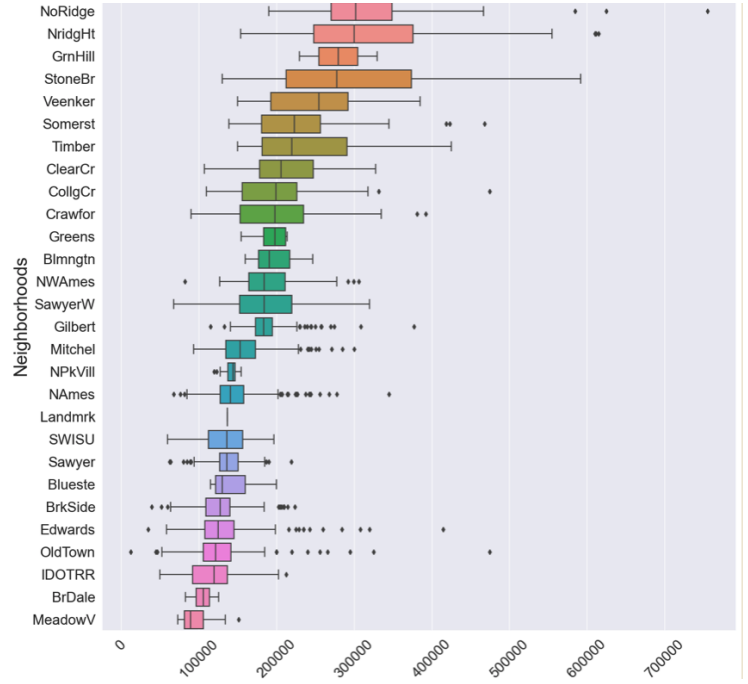


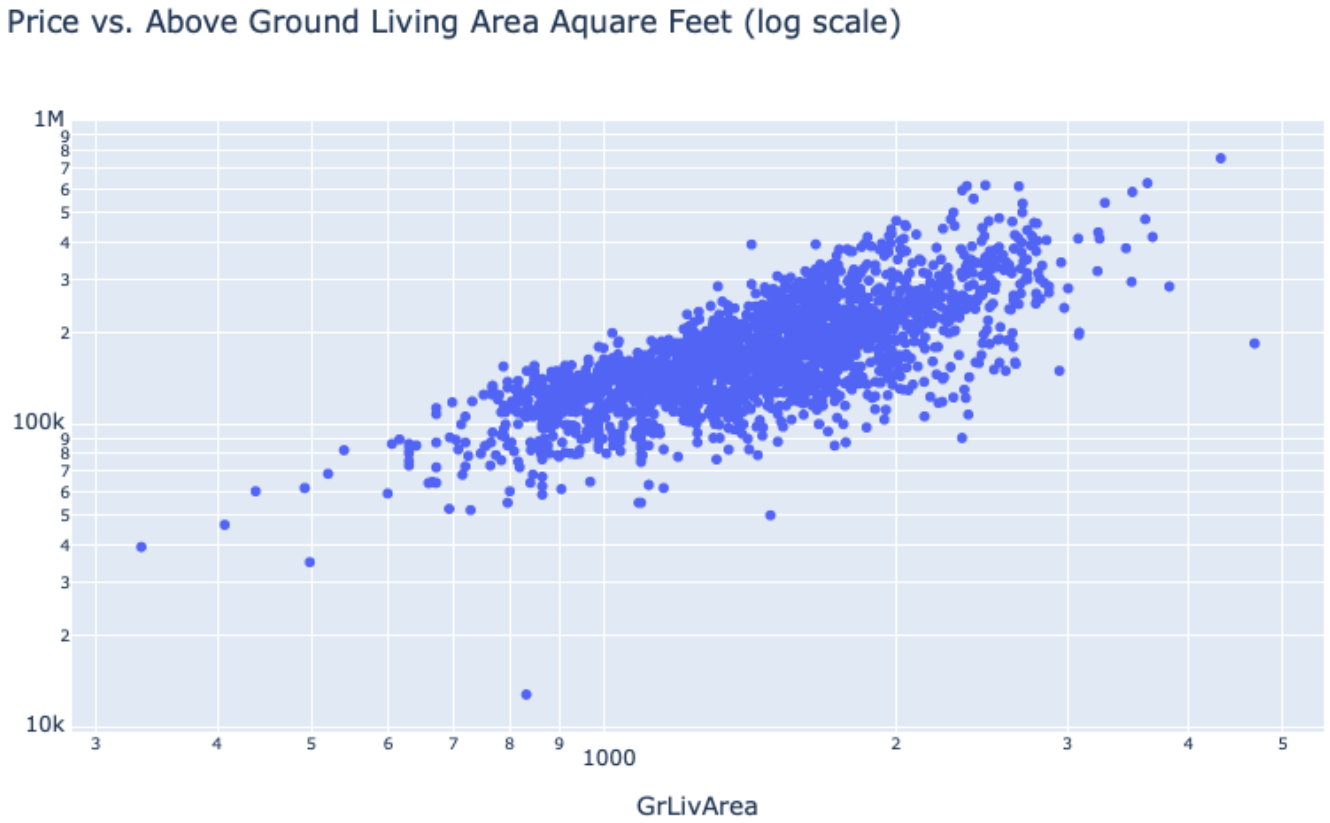
Exploratory Data Analysis

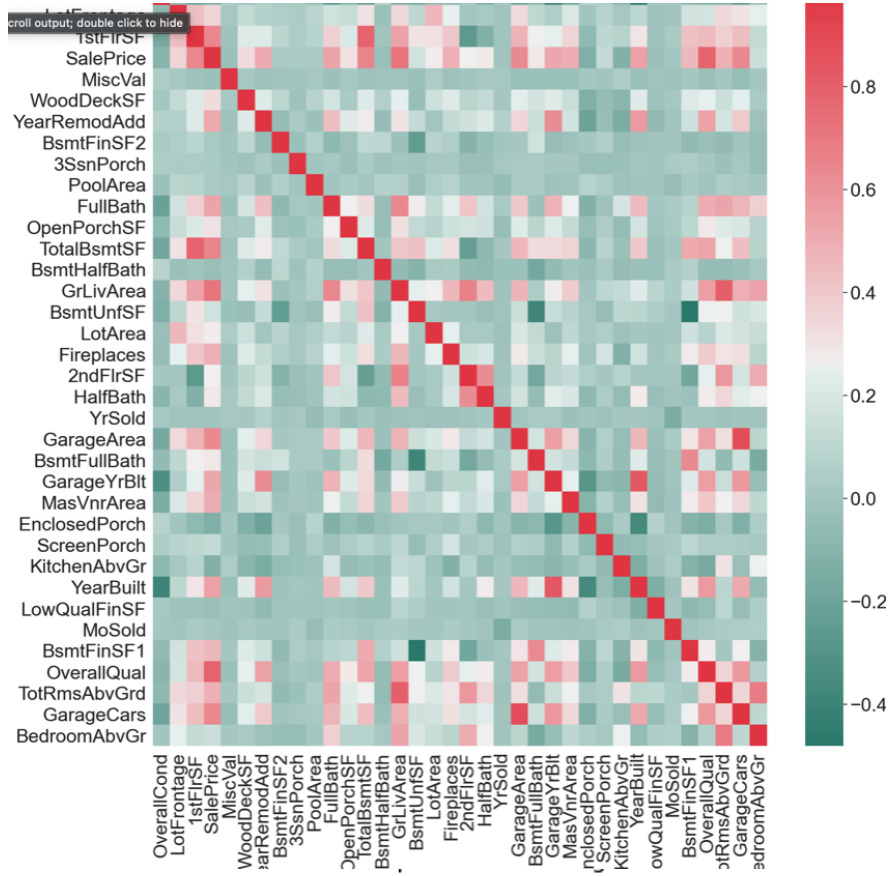
Neighborhood Housing Age

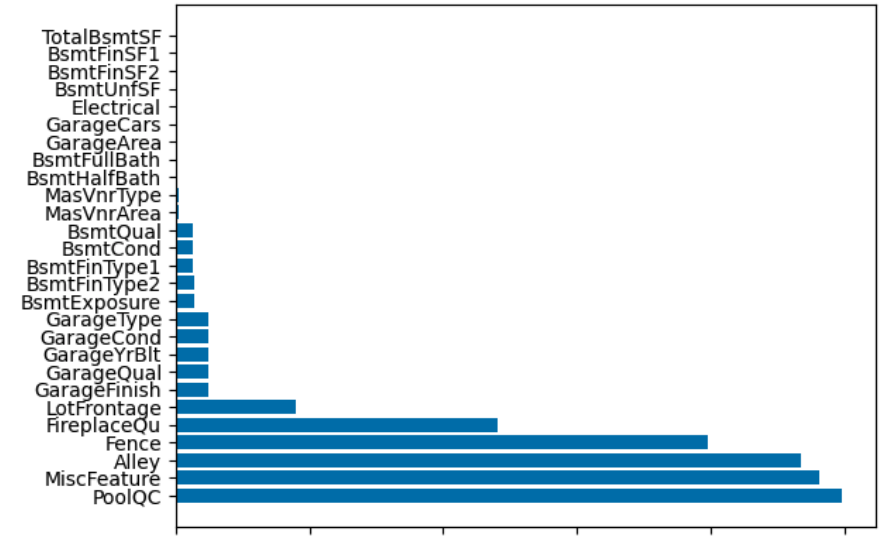




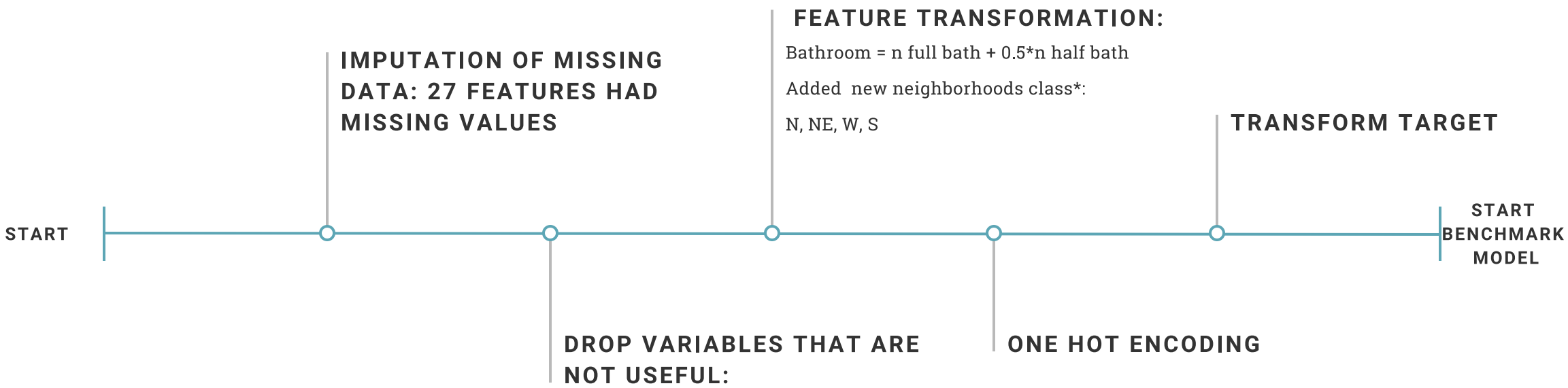



Price vs. Above Ground Living Area Aquare Feet (log scale)
1M
100k
10k
1000
GrLivArea


scroll output; double click to hide
1stFlrSF
SalePrice
MiscVal
WoodDeckSF
YearRemodAdd
BsmtFinSF2
3SsnPorch
PoolArea
FullBath
OpenPorchSF
TotalBsmtSF
BsmtHalfBath
GrLivArea
BsmtUnfSF
LotArea
Fireplaces
2ndFlrSF
HalfBath
YrSold
GarageArea
BsmtFullBath
GarageYrBlt
MasVnrArea
EnclosedPorch
ScreenPorch
KitchenAbvGr
YearBuilt
LowQualFinSF
MoSold
BsmtFinSF1
OverallQual
TotRmsAbvGrd
GarageCars
BedroomAbvGr
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


TotalBsmtSF
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
Electrical
GarageCars
GarageArea
BsmtFullBath
BsmtHalfBath
MasVnrType
MasVnrArea
BsmtQual
BsmtCond
BsmtFinType1
BsmtFinType2
BsmtExposure
GarageType
GarageCond
GarageYrBlt
GarageQual
GarageFinish
LotFrontage
FireplaceQu
Fence
Alley
MiscFeature
PoolQC

Pre-Processing



* N: NridgHt, NoRidge, Veenker, StoneBr;
NE: Blmngtn,Gilbert,OldTown, NWAmes, NAmes, Blmngtn,BrDale,Somerst, BrkSide, Greens, NPkVill;
W: CollgCr,SWISU, Edward,SawyerW,SawyerClearCr;
S: Crawfor,MeadowV,Timber,Mitchel,IDOTRR

Machine Learning Models

Linear Regression Models

- Multiple linear regression
  (Benchmark Model)
- Lasso Regularization
  - Feature Selection
  - Improves Model fit

Tree-Based Models

- Random Forest
- Gradient Boosting

Random Forest

- Ensemble of Decision Trees
- Random Selection of Rows and Columns
- Creates Statistical Independence of Trees
- Mitigates Overfitting

Gradient Boosting Machine

- Sequentially Adds Trees, Does Not Average
- Converts Weak Learners to Strong Learners

  Requires Aggressive Hyperparameter Tuning to Prevent Overfitting

Linear Models

Ordinary Least Squares

(Benchmark Model)

R^2 train set: 0.886059

R^2 test set: 0.960320 CV

R^2 train set: 0.90715 RMSE: 0.03249

We Need to Address Multi Colinearity Issues to improve the model from the Benchmark.

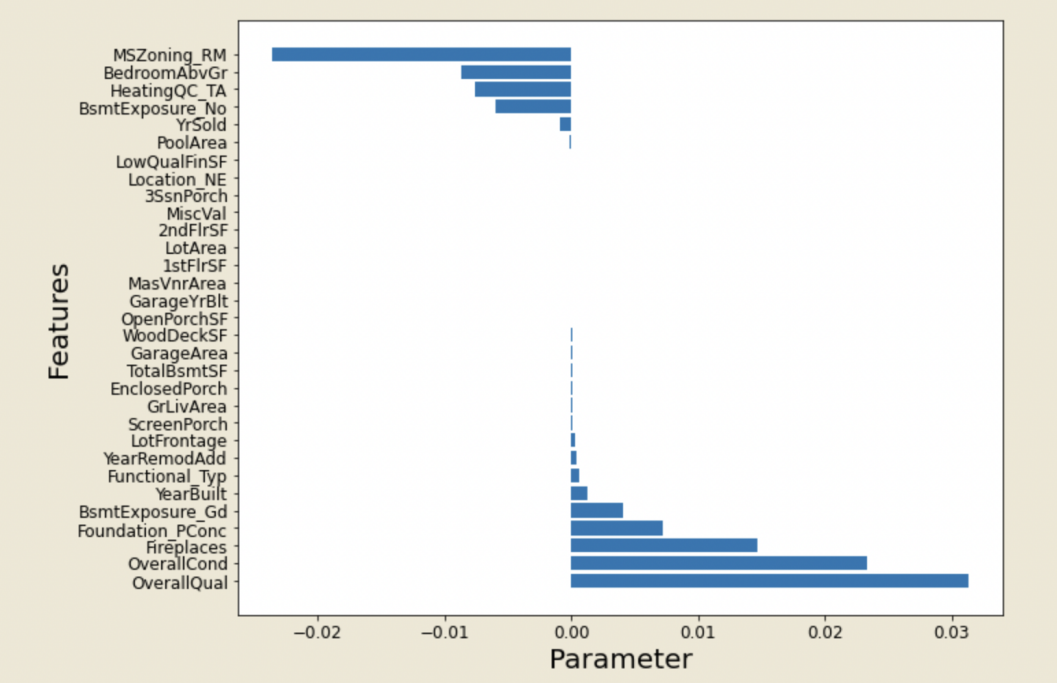
Penalized Linear Model: Lasso

- Lasso Regression with regularization terms

  Grid search to find an optimal hyp parameter lambda.

Lasso Training Set Score: 0.905440

Lasso Test Set Score: 0.913455



- Refit Multiple Linear Regression with Selected Features:
  - R^2 train set: 0.918851
  - R^2 test set: 0.932031
  - CV R^2 train set: 0.90888
  - RMSE: 0.0453

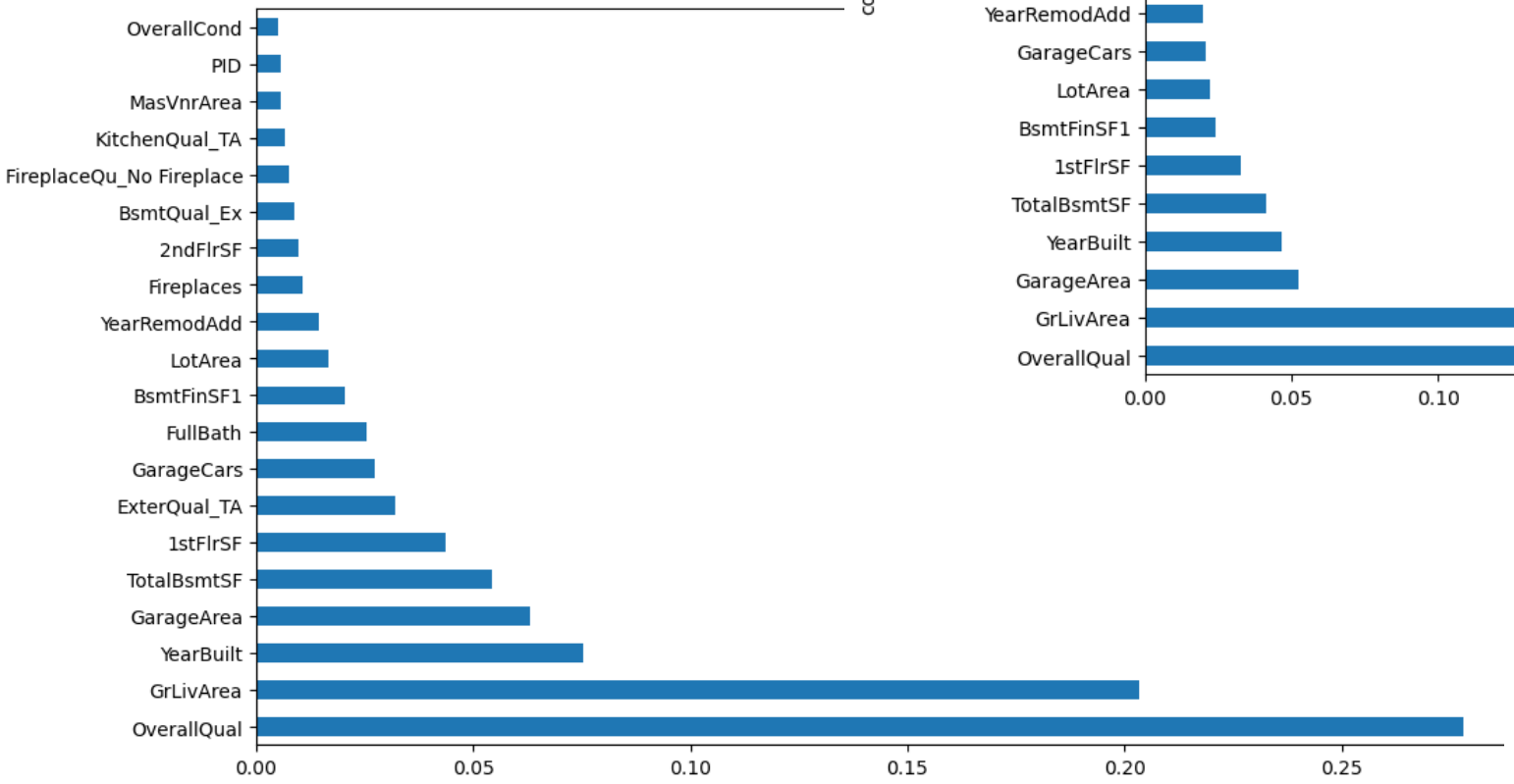
- Refit Multiplea Linear Regression with Neighborhood re-classed
  - R^2 train set: 0.913069
  - R^2 test set: 0.924962
  - CV R^2 train set: 0.90888
  - RMSE: 0.04467

TREE MODELS

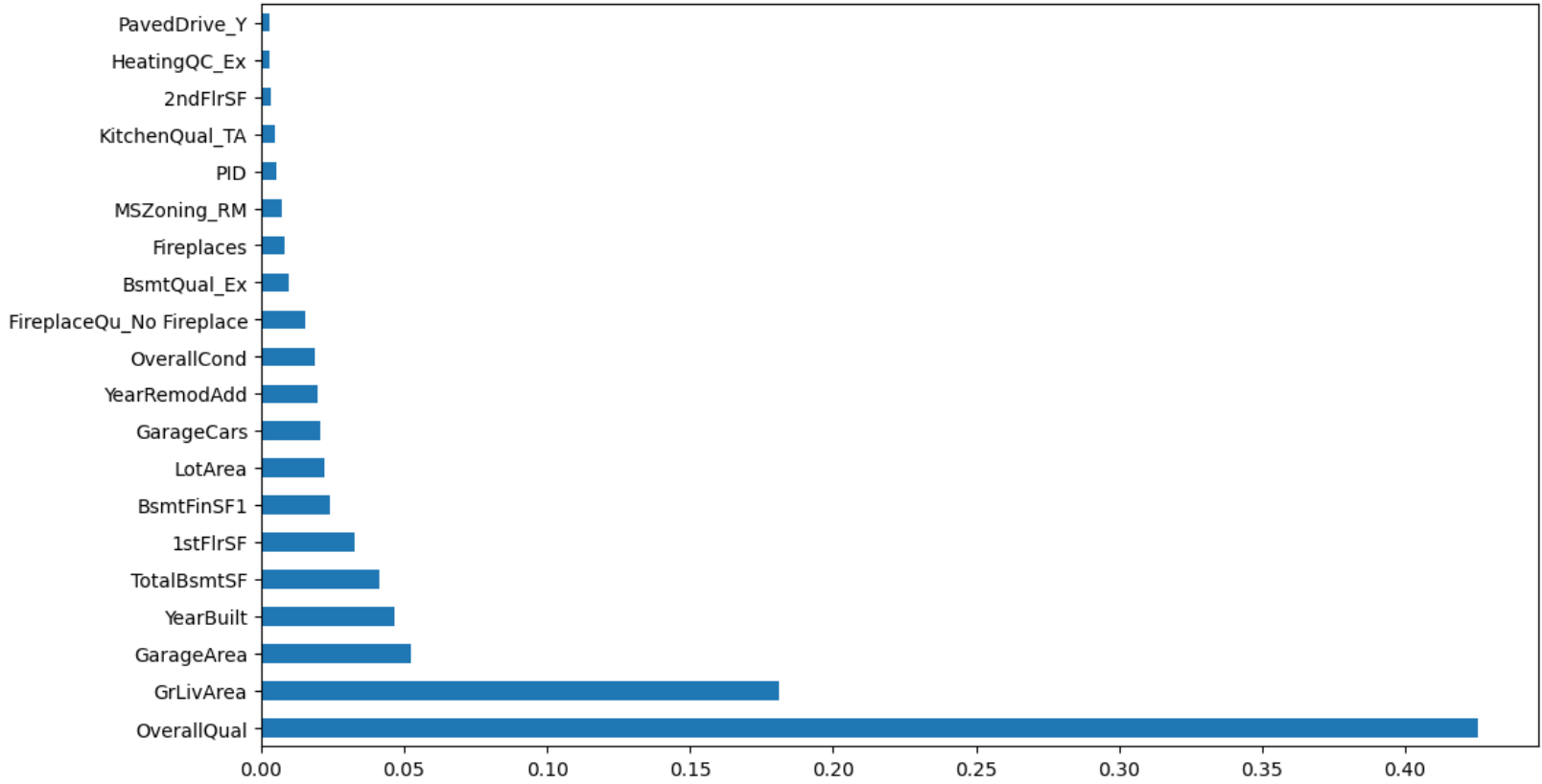
Random Forest

Log Root MSE (train) = .05639

Log Root MSE (valid) = .10619



Feature Importances



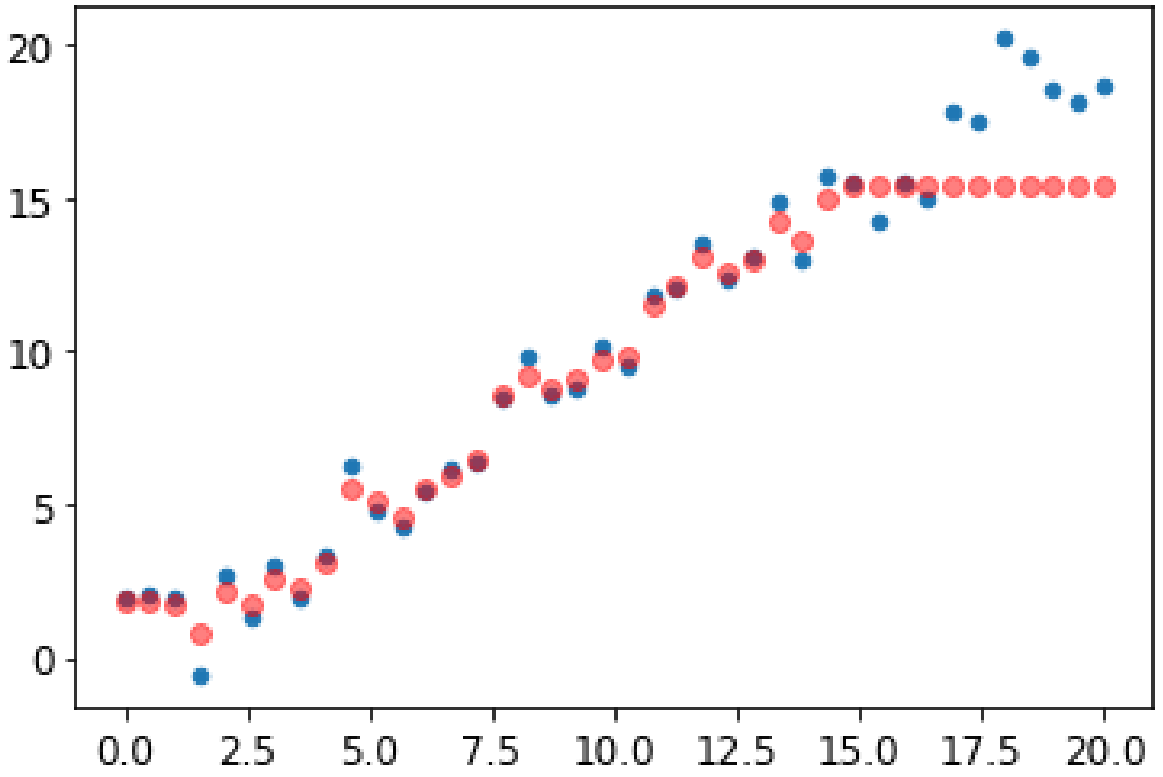
Gradient Boosting Model

Log Root MSE (train) = .003419

Log Root MSE (valid) = .131497

Term Structure of Tree Model

- Address Structural RF Problem
- Fit Linear Coefficient to Date-Only Data
- Transform Validation Set To One Year Earlier
- Evaluate Transformed Data Set
- Apply Term Structure Adjustment



Simulated Data - Deep Learning for Coders with fastai & Python

(Howard and Gugger, O'Reilly, 2020)

Log Term Structure Adjustment = 0.0031734

Term Structure Adjustment = 1.007334

Log Root MSE (Train) = 0.05639

Log Root MSE (Valid) = .107604

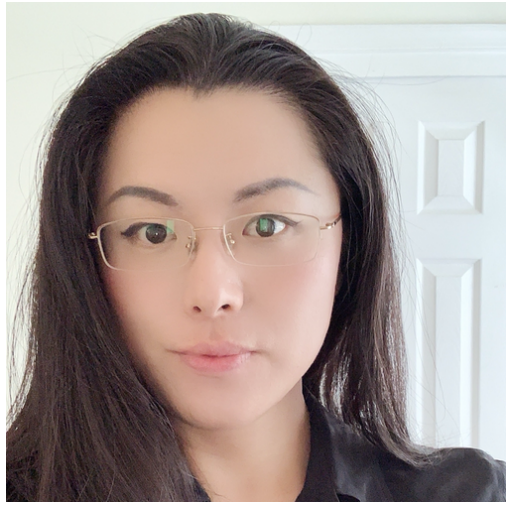
Conclusions and Future Work

- Term structure is not a crucial factor in the Ames data
- GBM has a *very* steep feature importance structure
- Feature importances between the Random Forest and GBM had high overlap
- Linear Models show that our models has slightly better fit in the testing set consistently, which is something worth investigating further. Leakage?

&

- Eliminate one-hot encoding for tree models
- More aggressive hyperparameter tuning for GBM
- Eliminate the test split for the tree models & implement cross-validation
- Use the untransformed SalePrice for sanity check and interpretability
- Try dimension reduction techniques to further increase the prediction power of our model
- Try adding polynomial terms to help Lasso regression fit

Thank You



CHERIE WANG

TYLER KIM

JOHN KOSMICKE