



Predicting Housing Prices with Machine Learning

JOHN KOSMICKE, CHERIE WANG, TYLER KIM

MADE WITH

beautiful.ai

Introduction

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.

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

In addition, we also look use some neighborhood distance-based aggregation methods to get a sense of the demographic history of Ames housing development.

Description of Dataset:

Objectives



PREDICT

Predict home prices in Ames, Iowa.



IDENTIFY

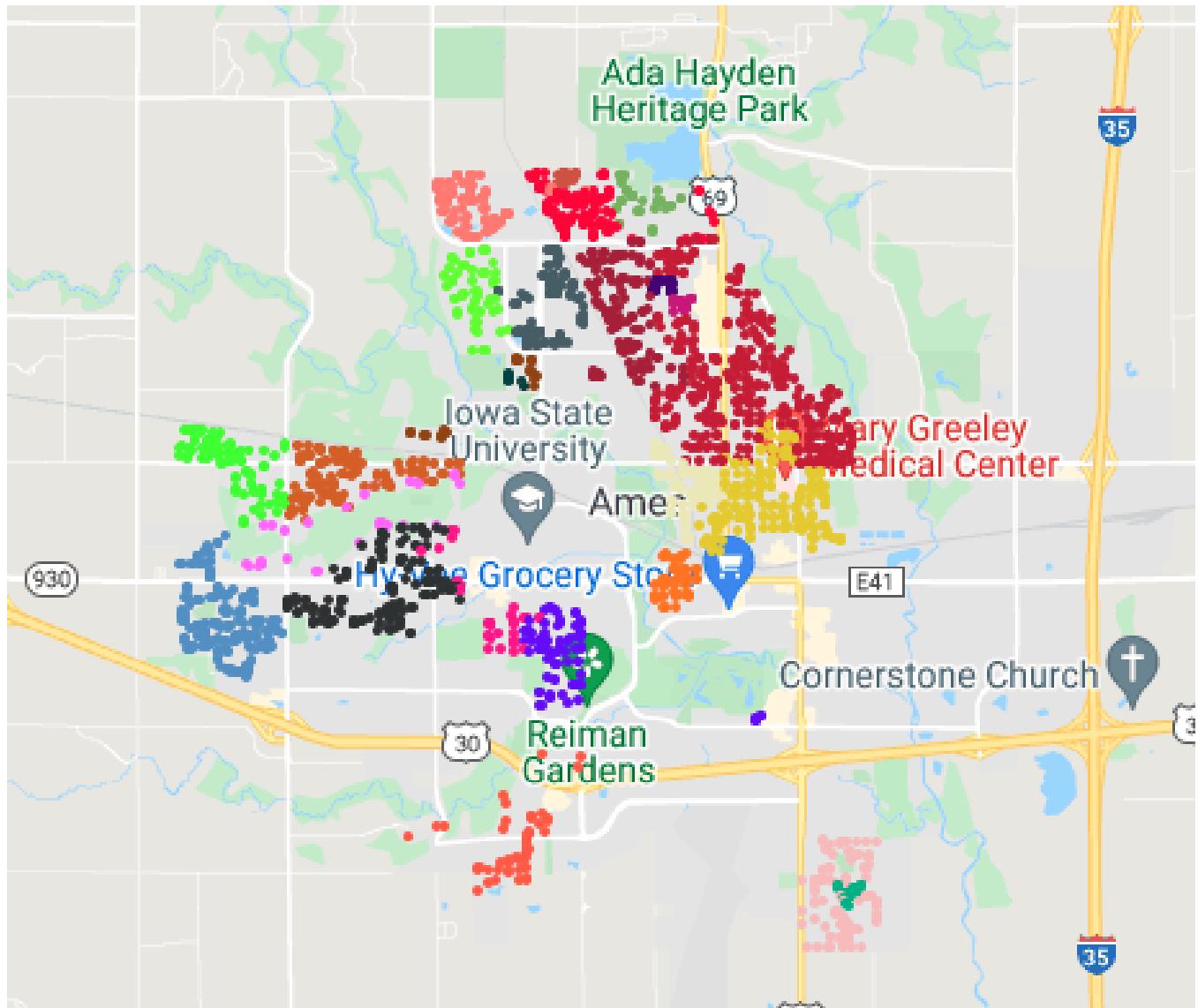
Identify the best methodologies.



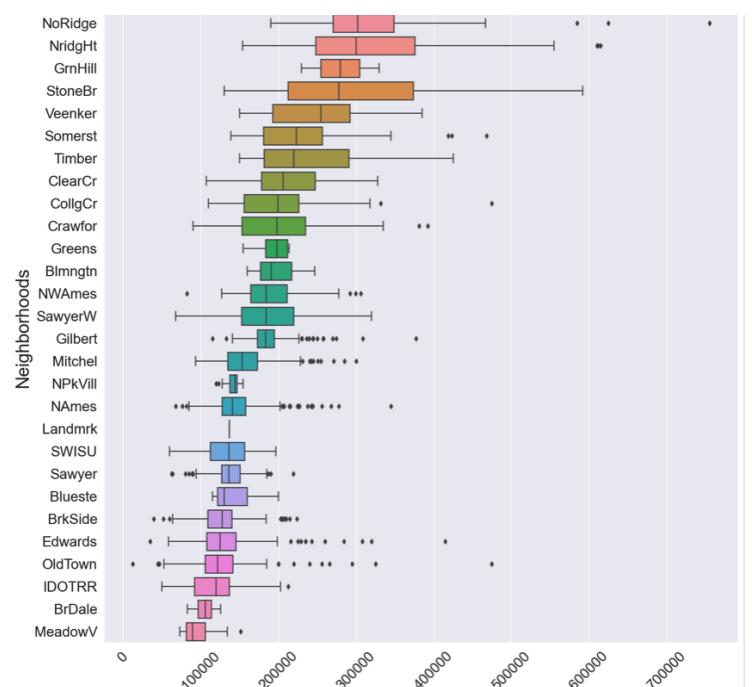
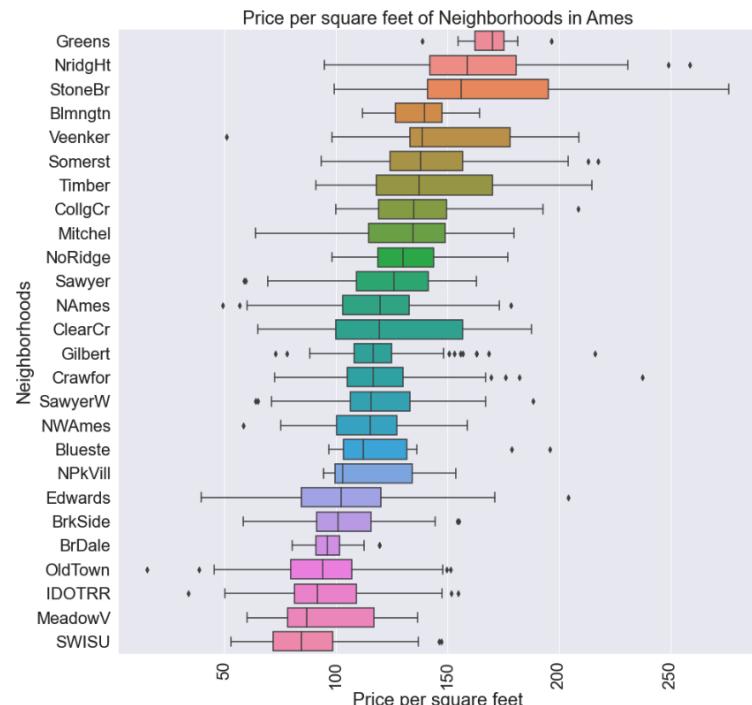
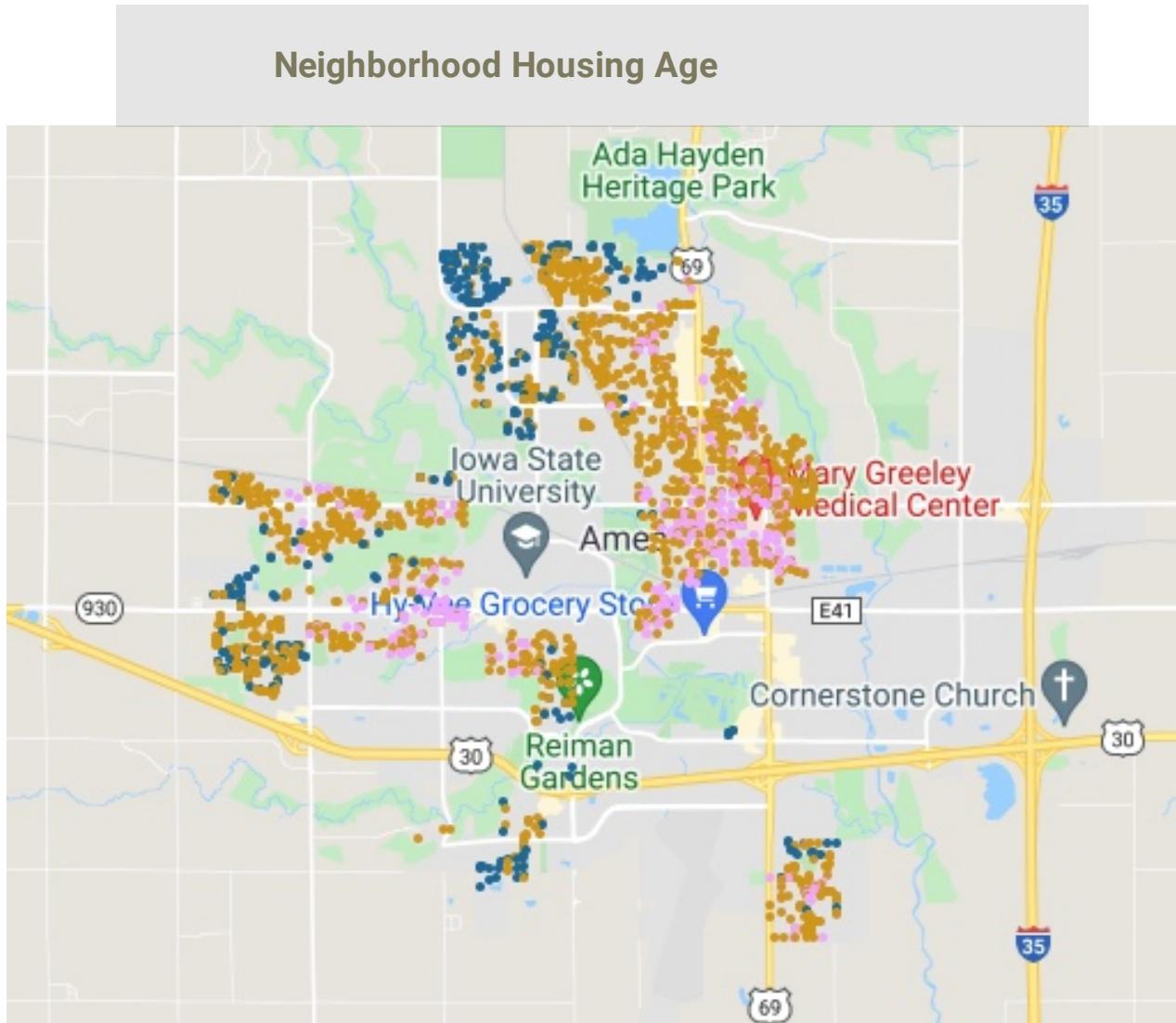
IMPLEMENT

Implement machine learning models.

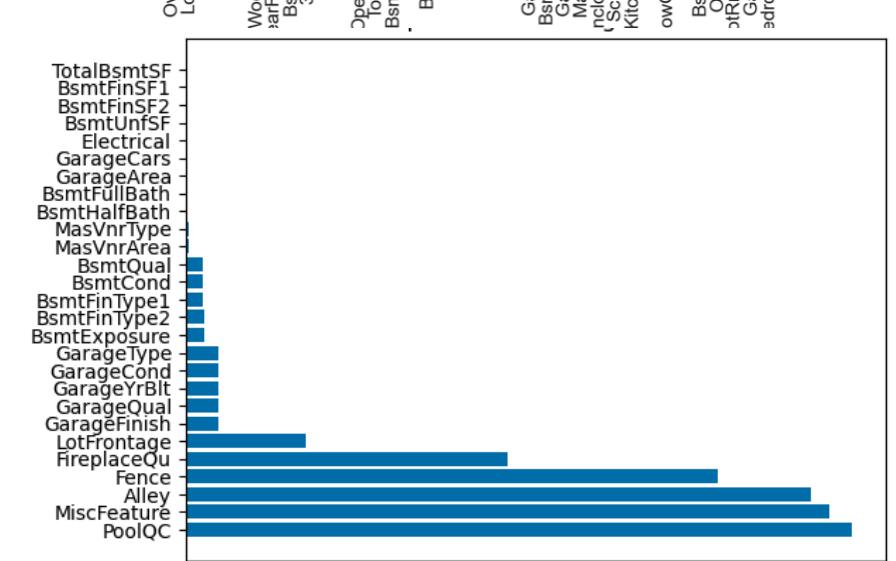
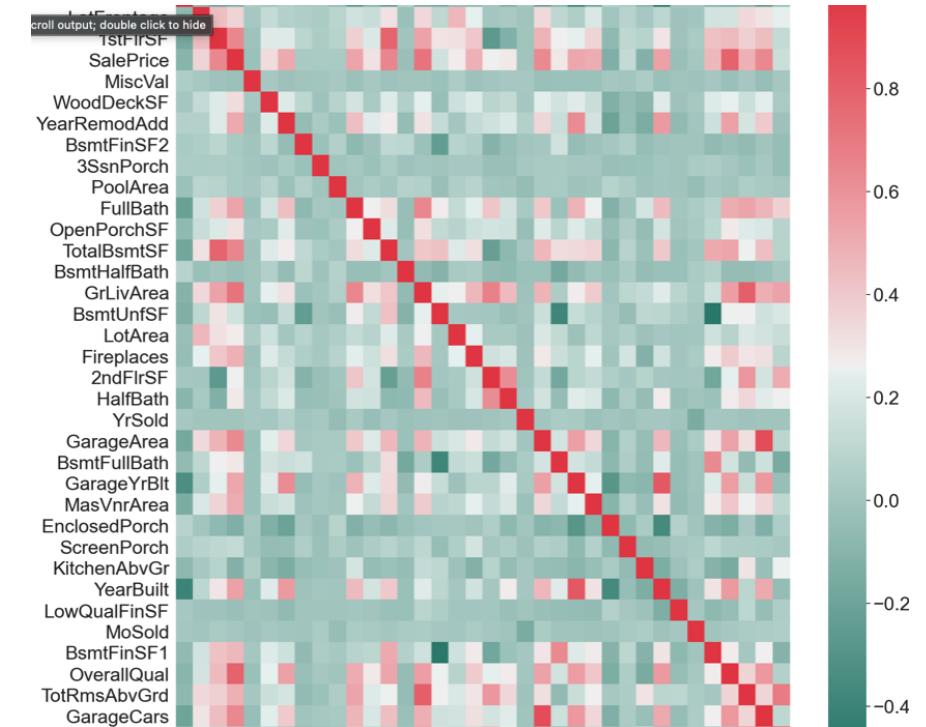
Ames Neighborhood Demographics



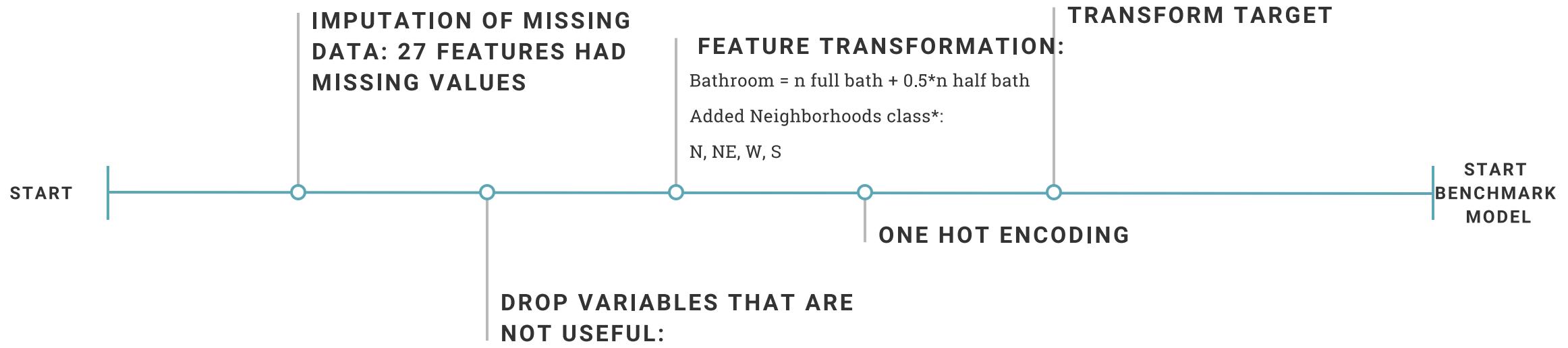
Exploratory Data Analysis



Price vs. Above Ground Living Area Aquare Feet (log scale)



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
- Lasso Regularization

Tree-Based Models

- Random Forest
- Gradient Boosting

Linear Models

Ordinary Least Squares

(Benchmark Model)

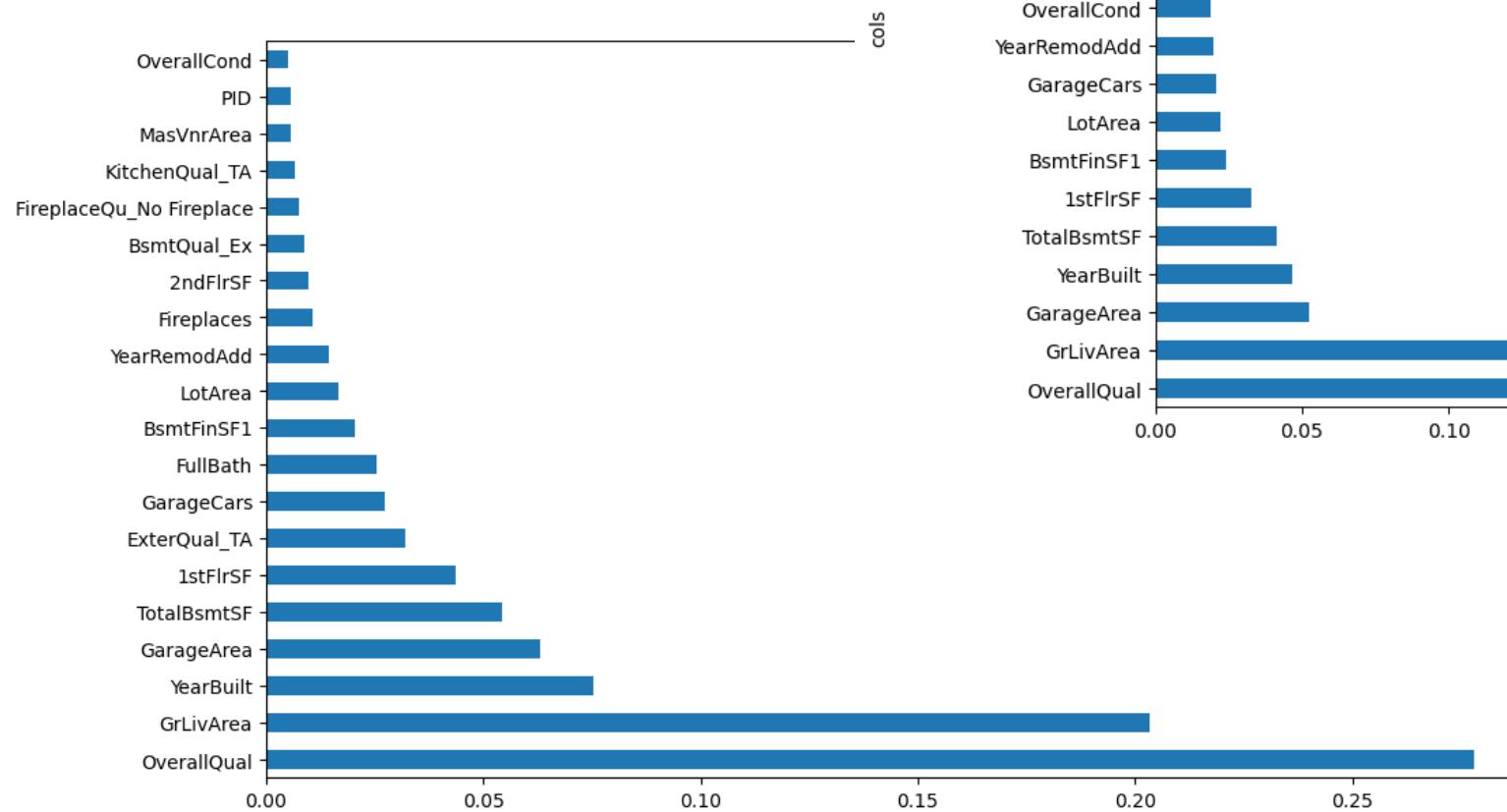
Penalized Linear Model: Lasso

TREE MODELS

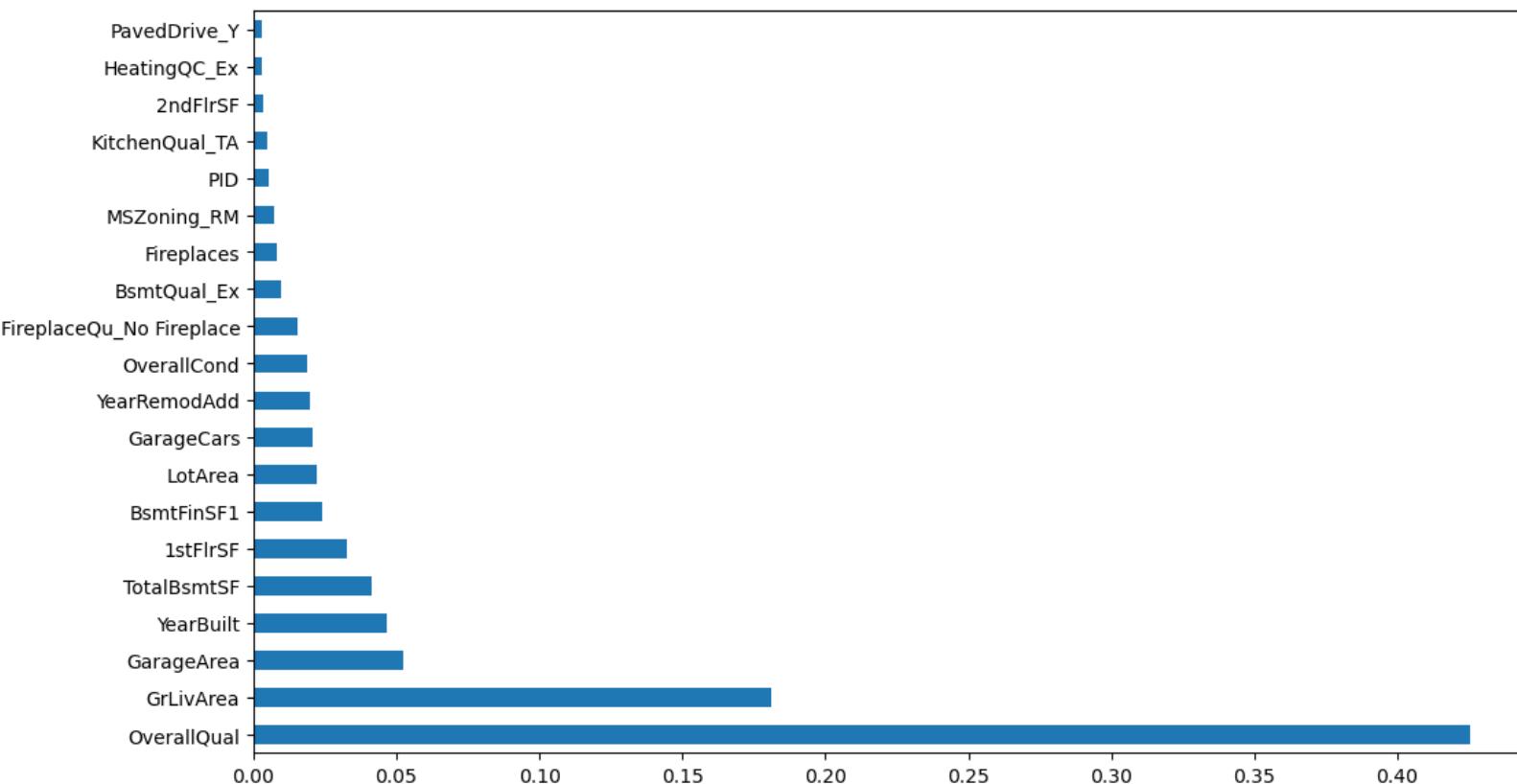
Random Forest

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

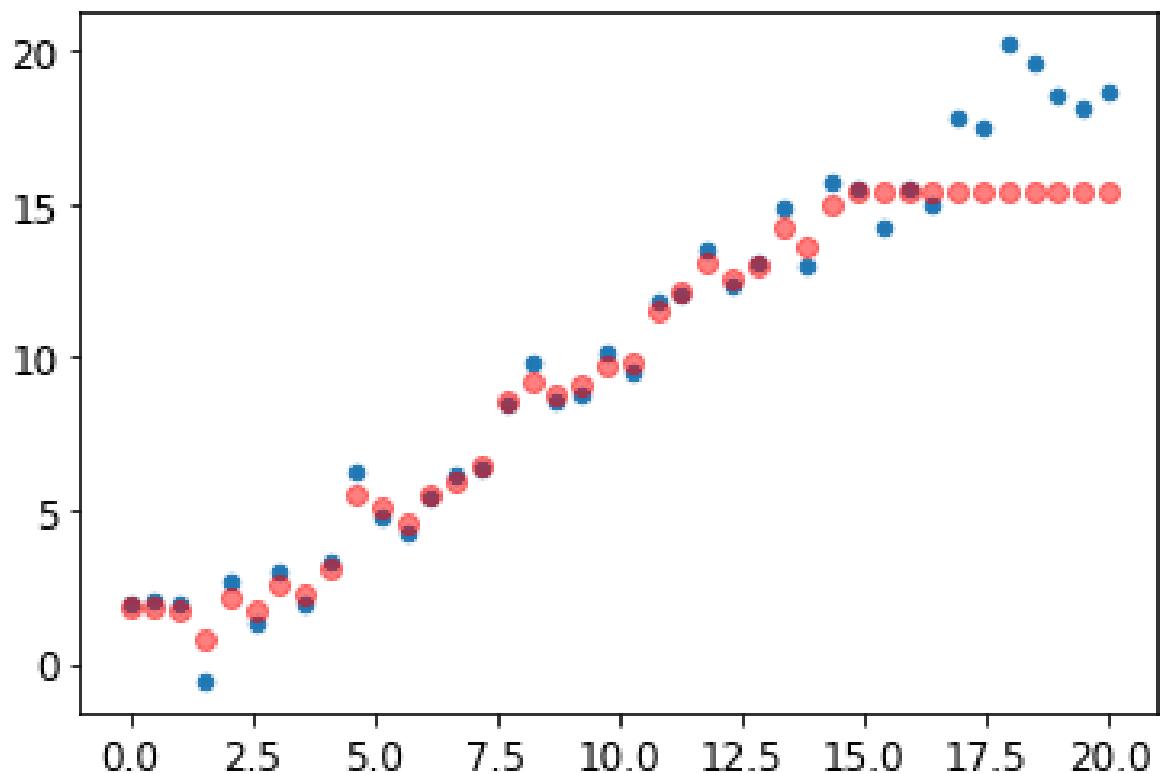
- Address Structural RF Problem
- Fit Linear Coefficient to Date-Only Data
- Transform Validation Set To One Year Earlier
- Evaluate Transformed Data Set
- Apply Term Structure Adjustment

Log Term Structure Adjustment = 0.0031734

Term Structure Adjustment = 1.007334

Log Root MSE (Train) = 0.05639

Log Root MSE (Valid) = .107604



Simulated Data - Deep Learning for Coders with fastai & Python

Conclusions



Future Work



ELIMINATE ONE-HOT ENCODING FOR
THE CATEGORICAL FEATURES IN
FAVOR OF LABEL ENCODING-ish
APPROACH



APPLY ADDITIONAL MACHINE
LEARNING ALGORITHMS



IMPLEMENT MORE SPECIFIC MODELS
TARGETING DEMOGRAPHICS AND
AGE GROUPS TO GET A MORE
ACCURATE ASSESSMENT OF HOME
VALUES IN AMES, AND OTHER
GEOGRAPHICAL REGIONS

Thank You



CHERIE WANG



TYLER KIM



JOHN KOSMICKE