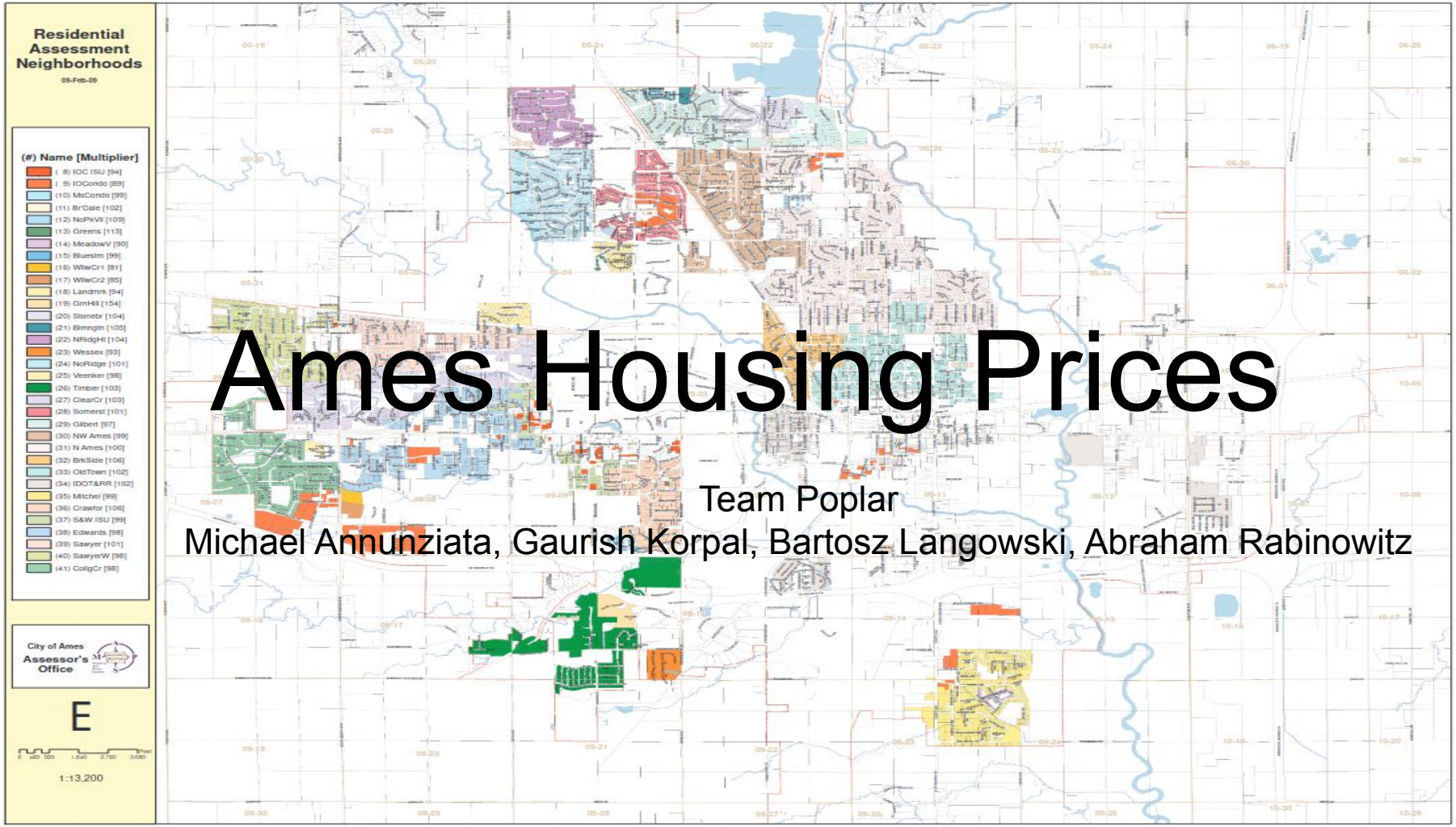


Ames Housing Prices

Team Poplar
Michael Annunziata, Gaurish Korpal, Bartosz Langowski, Abraham Rabinowitz

The Problem

1. Accurately predicting housing prices in Ames, Iowa.

2. Understanding what features of a home influence its value.

The Dataset

Dean De Cock (2011) Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project, Journal of Statistics Education, 19:3, DOI: 10.1080/10691898.2011.11889627

Anna Montoya (2016), House Prices - Advanced Regression Techniques. Kaggle. https://kaggle.com/competitions/house-prices-advanced-regression-techniques

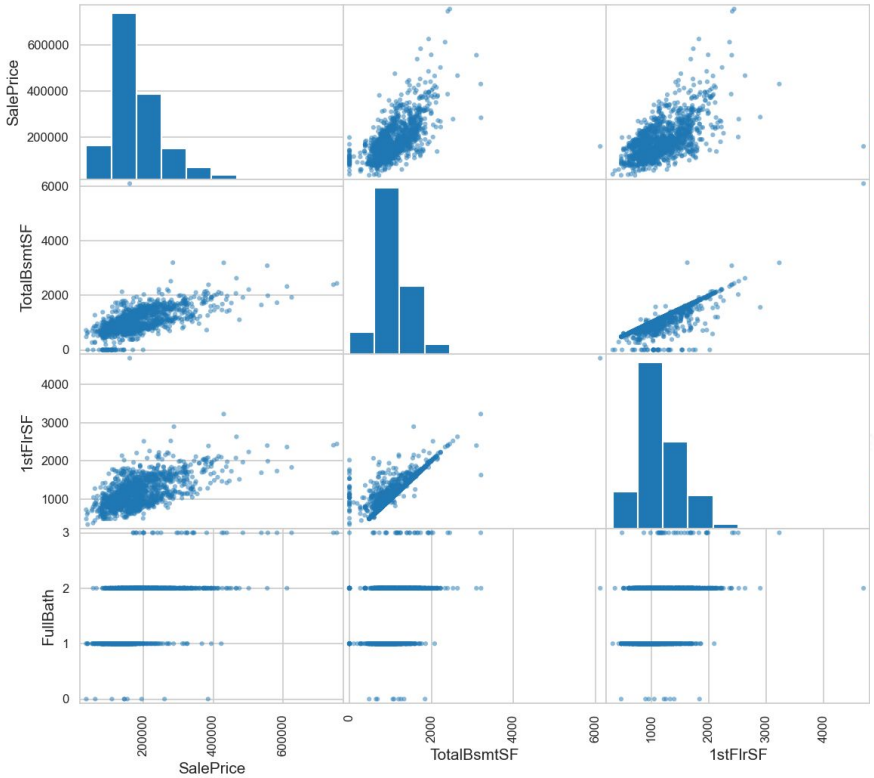
- Provides a description of the sale of 2930 individual residential property in Ames, Iowa from 2006 to 2010.

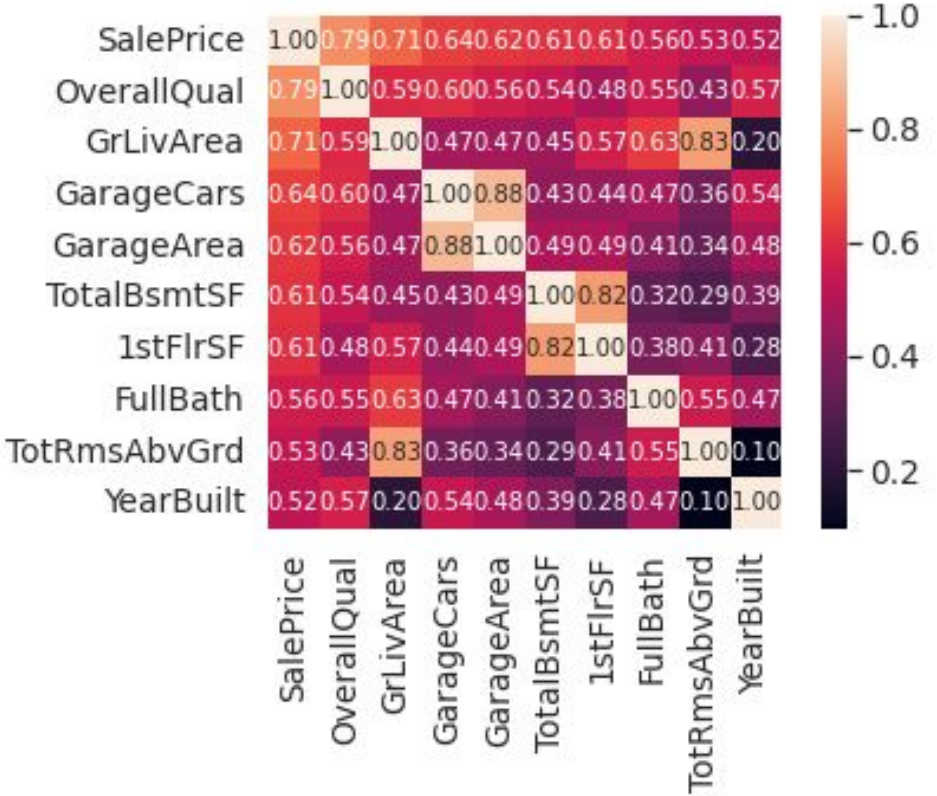
- Contains 80 explanatory variables with numerical and categorical data such as square footage, neighborhood, and overall house condition.

- The training data consists of 1460 observations.

Exploratory Data Analysis

- To study correlation between the **Sale Price** of the house and numerical features we used:
    - *Correlation matrix*
    - *Heat map*
    - *Scatter plots*


SalePrice
TotalBsmtSF
1stFlrSF
FullBath
600000
400000
200000
6000
4000
2000
0
4000
3000
2000
1000
3
2
1
0
200000
400000
600000
0
2000
4000
6000
1000
2000
3000
4000


SalePrice 1.00 0.79 0.71 0.64 0.62 0.61 0.61 0.56 0.53 0.52
OverallQual 0.79 1.00 0.59 0.60 0.56 0.54 0.48 0.55 0.43 0.57
GrLivArea 0.71 0.59 1.00 0.47 0.47 0.45 0.57 0.63 0.83 0.20
GarageCars 0.64 0.60 0.47 1.00 0.88 0.43 0.44 0.47 0.36 0.54
GarageArea 0.62 0.56 0.47 0.88 1.00 0.49 0.49 0.41 0.34 0.48
TotalBsmtSF 0.61 0.54 0.45 0.43 0.49 1.00 0.82 0.32 0.29 0.39
1stFlrSF 0.61 0.48 0.57 0.44 0.49 0.82 1.00 0.38 0.41 0.28
FullBath 0.56 0.55 0.63 0.47 0.41 0.32 0.38 1.00 0.55 0.47
TotRmsAbvGrd 0.53 0.43 0.83 0.36 0.34 0.29 0.41 0.55 1.00 0.10
YearBuilt 0.52 0.57 0.20 0.54 0.48 0.39 0.28 0.47 0.10 1.00
SalePrice
OverallQual
GrLivArea
GarageCars
GarageArea
TotalBsmtSF
1stFlrSF
FullBath
TotRmsAbvGrd
YearBuilt
1.0
0.8
0.6
0.4
0.2

Modeling Approach

- Linear Regression:
  - Took advantage of a strong linear relationship between sale price and numerical variables.
  - Elastic Net combined Lasso and Ridge regressions:
    - Performed better than ordinary least squares, lasso, or ridge regression alone.
  - Did not incorporate information from many categorical variables.

Modeling Approach

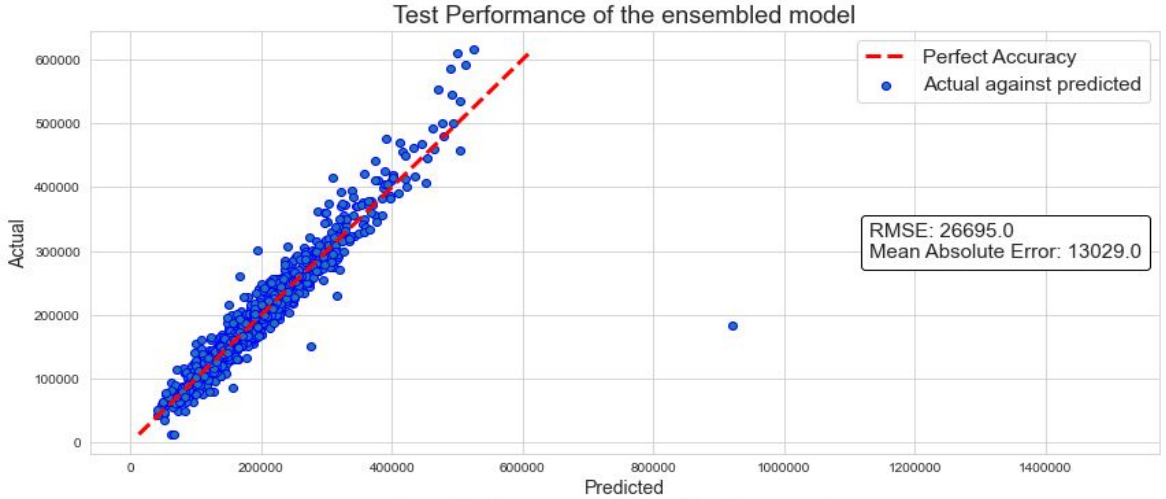
- Ensemble Learning:
  - Considered several models to better capture nonlinearity and incorporate categorical variables.
  - Final ensemble model is a stacked regression combining Elastic Net, XGBoost, and LightGBM.
  - Cross validation to tune hyperparameters. For example:
    - alpha = 0.0006
    - L_1 ratio = 0.9
    - Learning rate = 0.05
    - n estimators = 1000

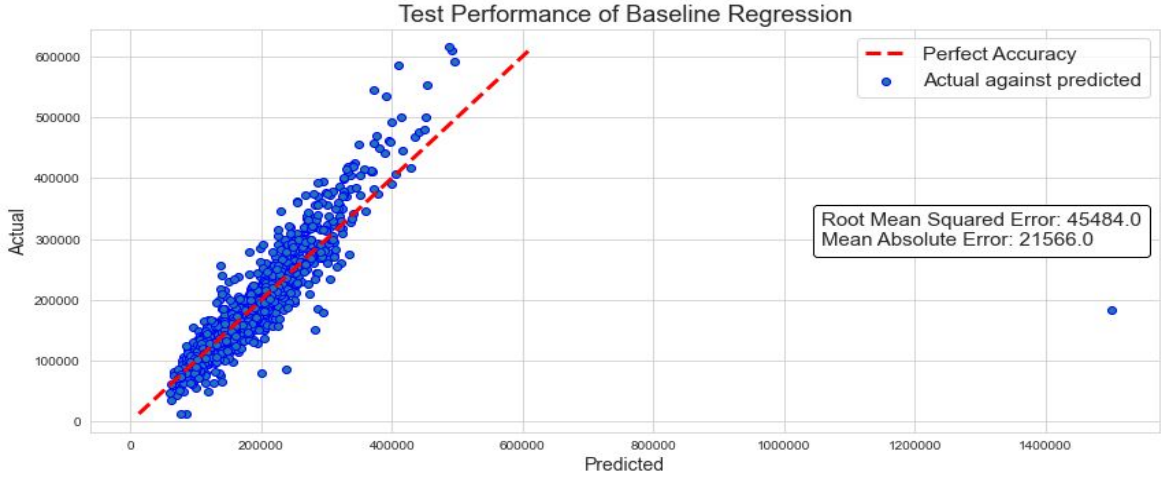
Modeling Approach

- Used cross validation to select final model:

Model	Log Mean Squared Error
Ridge Regression	.0125
Elastic Net Regression	.0122
LightGBM	.0142
XGBoost	.0136
Ensemble (Elastic Net, LGBM, XGBoost)	**.0117**

Results





Performance on Unseen Data:

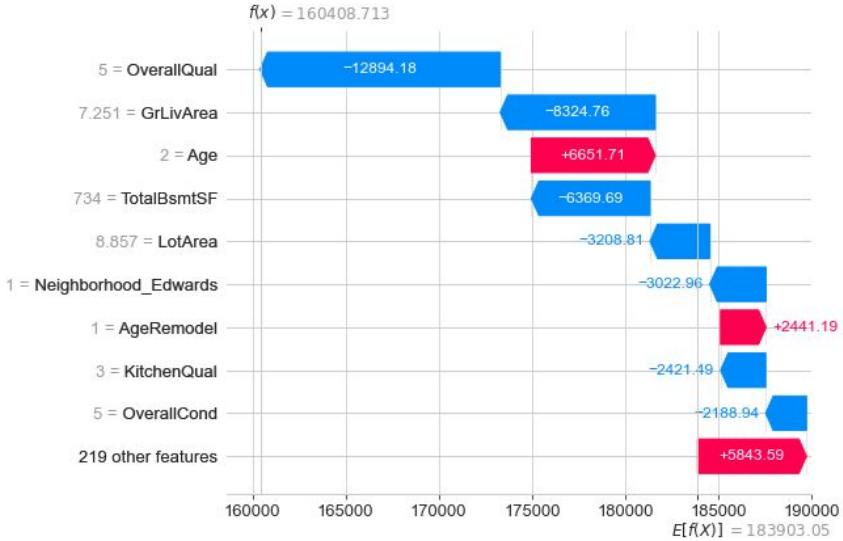
- More accurate than a baseline linear model.
- Better at pricing more expensive houses.
- Cannot handle extreme outliers.
- Further work can focus on handling such outliers or enlisting domain experts to help price them.

Results

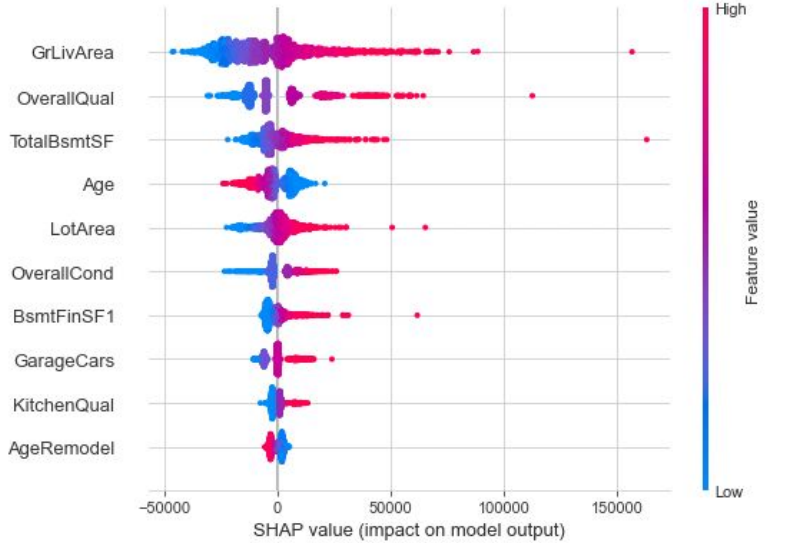
Interpreting our model with SHAP:

- SHAP values quantify the effect that changing values of specific features has on the way that the model predicts Sale Price.

SHAP values for a sample prediction



Distribution Summary of SHAP Scores



Appendix: Exploratory Data Analysis

Analysis of categorical features:

- Performed *one-hot encoding.*
- We used *violin plots* to understand the feature importance: