W207 Final Project

AMES Housing
Mrinal Chawla, Thomas Gao, Fengyao Luo

The Inference Problem

GIVEN

- A vector of features about the house
  - Neighborhood
  - Quality
    - Overall, Pool, etc.
  - Sale Condition
    - Normal, Abnormal
  - Amenities
    - Alley, Garage, Basement, etc.
  - Size
    - Sqft, # of rooms, # of bathrooms etc

PREDICT

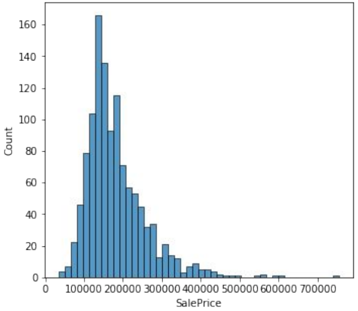
- Sale Price of the House

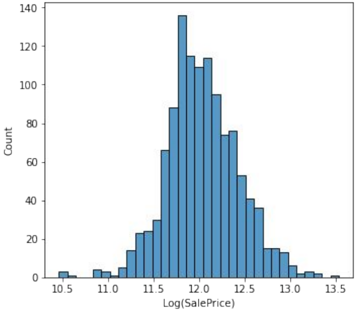
WHY

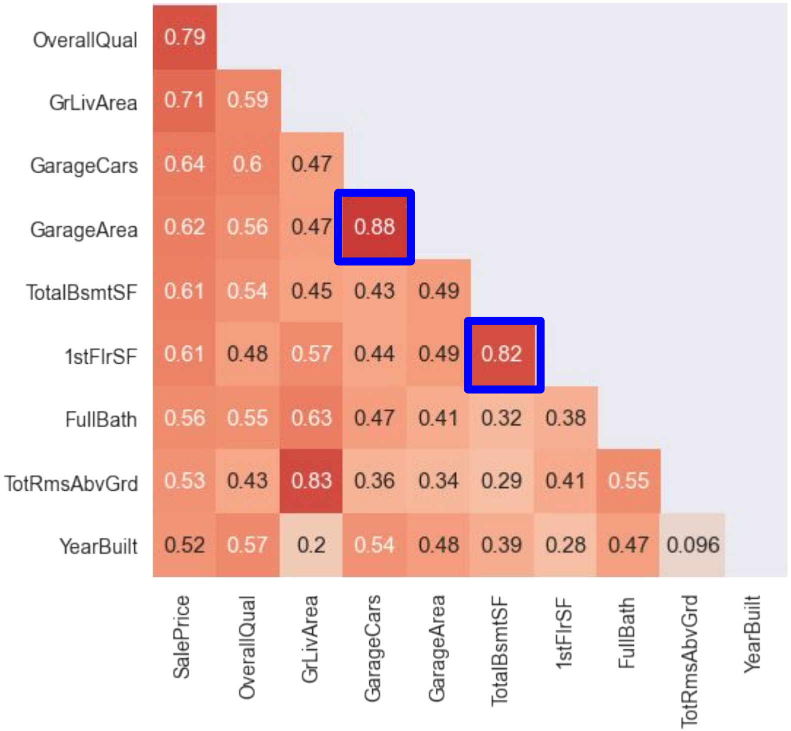
- Helps housing market for both sellers and buyers
- Help plan renovations
- Help plan infrastructure improvements
- Etc.

Exploratory Data Analysis

- SalePrice slightly right-skewed
  - Log transformation to fix
- Sizing variables highly correlated
  - Bsmt sqft vs 1st floor sqft
  - GarageArea vs GarageCars



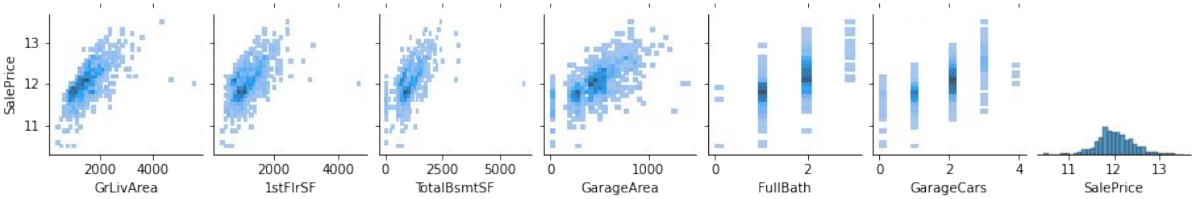




	SalePrice	OverallQual	GrLivArea	GarageCars	GarageArea	TotalBsmtSF	1stFlrSF	FullBath	TotRmsAbvGrd	YearBuilt
OverallQual	0.79									
GrLivArea	0.71	0.59								
GarageCars	0.64	0.6	0.47							
GarageArea	0.62	0.56	0.47	0.88						
TotalBsmtSF	0.61	0.54	0.45	0.43	0.49					
1stFlrSF	0.61	0.48	0.57	0.44	0.49	0.82				
FullBath	0.56	0.55	0.63	0.47	0.41	0.32	0.38			
TotRmsAbvGrd	0.53	0.43	0.83	0.36	0.34	0.29	0.41	0.55		
YearBuilt	0.52	0.57	0.2	0.54	0.48	0.39	0.28	0.47	0.096	

Exploratory Data Analysis

- Generally linear relationships
- Few outliers for expensive homes
- Top indicators
  - Overall Quality
  - Living Area
  - Neighborhood
- Scaling and encoding



Baseline Models

PREDICT MEAN

Predict average sale price for every house

RMSE: 0.419

LINEAR REGRESSION

Use top 2 features and neighborhood

RMSE: 0.169

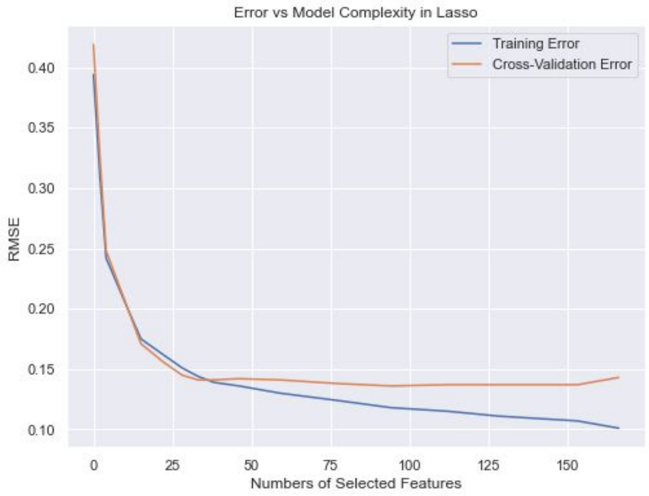
LASSO

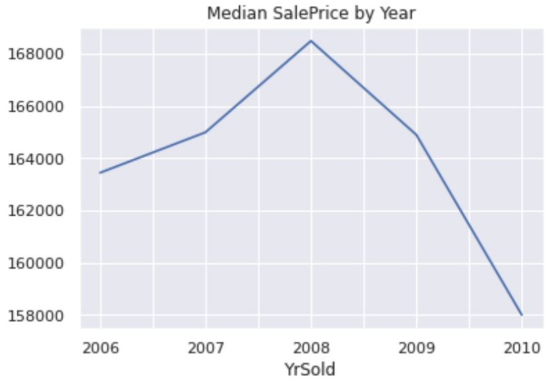
Linear Regression with L1 Regularization

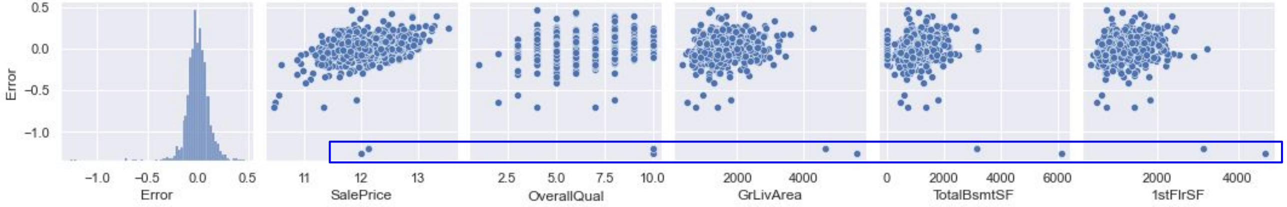
RMSE: 0.137

Error Analysis

- Complexity vs Performance
  - Number of Features
  - RMSE
- Outliers
  - Living Area
  - Quality
- Non-Linear Relationship
  - YrSold
  - SalePrice

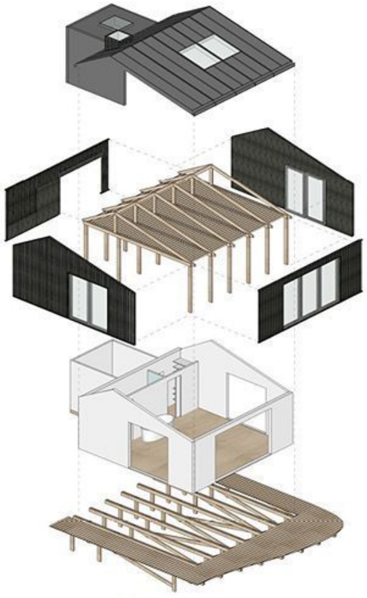

Error vs Model Complexity in Lasso


Median SalePrice by Year



Feature Engineering

- Aggregate size features
  - Total Sqft
  - Average Room Sqft
  - Total bathrooms
  - Total porch sqft
- Presence of amenities (binary)
  - Alley
  - Garage
  - Basement
  - Pool
- Years Since Remodelled (Year Sold - Year Remodelled)
- Seasonality (Month Sold → Season)
- Skewness (np.log)
- Neighborhood bins



Final Models

LASSO Linear Regression with L1 Regularization RMSE: 0.121	SPLIT LASSO Linear Regression with L1 Regularization + Separate Model for each Neighborhood Bin RMSE: 0.126
RANDOM FOREST Nonlinear Model RMSE: 0.135	ENSEMBLE 0.75 Lasso + 0.25 Random Forest RMSE: 0.119

Final Performance

- Model: Ensemble
- RMSE: 0.1217
- Kaggle: 10th percentile