Predicting Home Selling Prices with Historical Data

Derry Li

Problem Statement

In light of the *Zillow incident*, we will build a **predictive** model containing a multitude of input features to help robustly predict future housing prices.

Data

- Ames Housing Data
- 2930 observations, 82 features
- 20 features with nulls
- 22% observations have nulls

Feature types

categorical	23
ordinal	23
discrete	14
continuous	20
id	2

Data Cleaning

- drop Id’s
- column rename
  - Lot Area → lot_area
- more meaningful features
  -

yr_built
2001

→

yr_built
9

Base Model

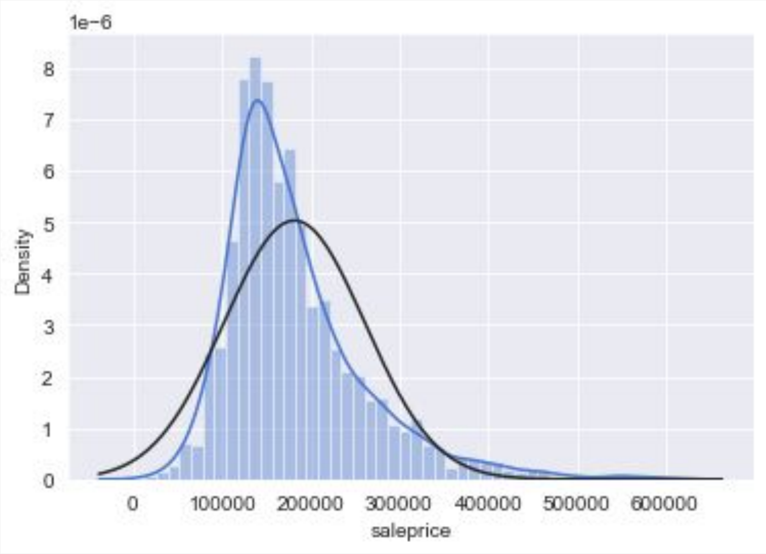
Base Model - Simple Linear Regression

- Input features
  - lot_area (ft ^{2})
  - overall_qual (1-10)
- Output feature : saleprice ($)
- StandardScaler

Base Model - Simple Linear Regression

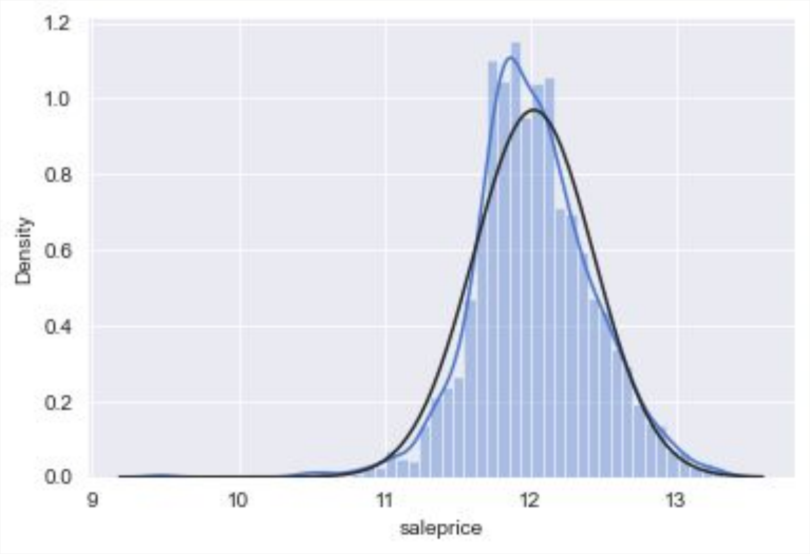
- Log transform

saleprice distribution



log
⟶

log**(saleprice)** distribution



Base Model - Simple Linear Regression

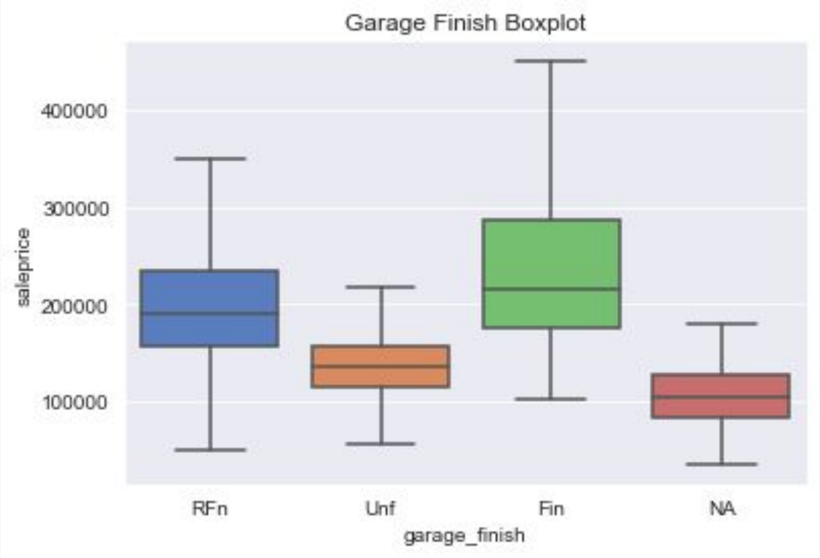
- Performance

	train	**test**
R^2 score	0.7247	0.7288
RMSE	0.2169	0.2111
CV score	0.7161	0.7341

Robust Model

Data Preprocessing

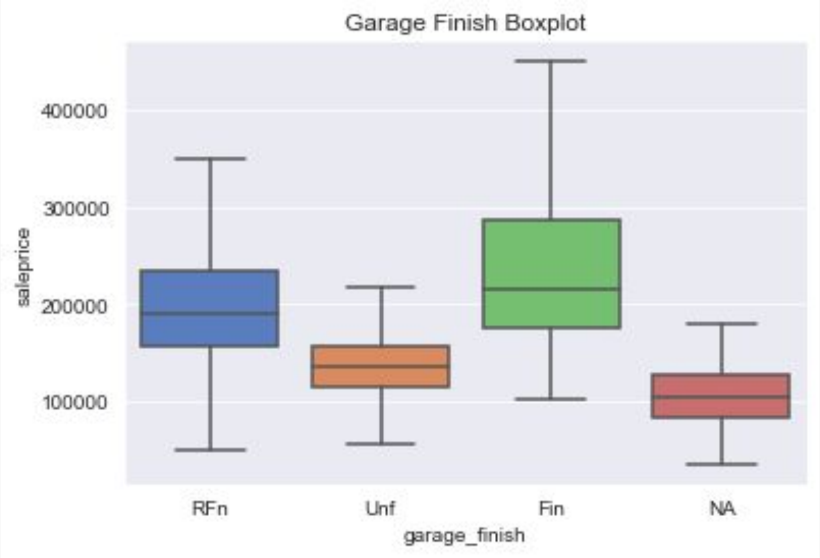
- **Ordinal Encoding**



Finished	3
Rough Finished	2
Unfinished	1
No Garage	0

Data Preprocessing

- **Ordinal Encoding**



Finished	3
Rough Finished	2
Unfinished	1
No Garage	0

Ex	5
Gd	4
TA	3
Fa	2
Po	1
NA	0

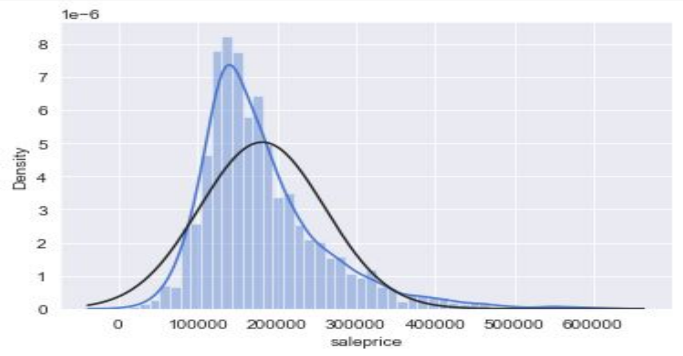
Data Preprocessing

- **Log transform**
    - on all numeric input features with skew > 0.6
    - on output feature (saleprice)

Data Preprocessing

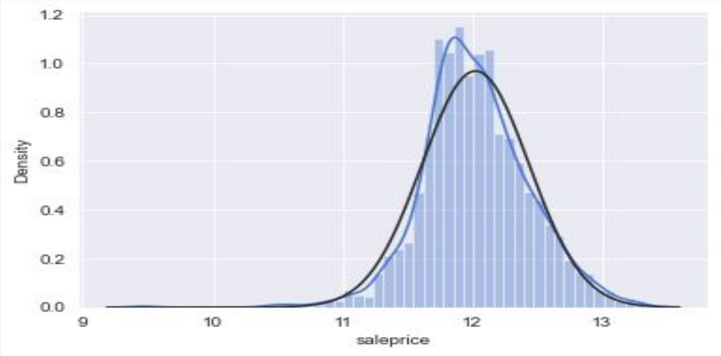
- **Log transform**
  - on all numeric input features with skew > 0.6
  - on output feature (saleprice)

saleprice distribution



log
⟶

log**(saleprice)** distribution



Data Preprocessing

- **Log transform**
  - on all numeric input features with skew > 0.6
  - on output feature (saleprice)
- **One Hot Encoding**
  - on all categorical feature
  - + selected ordinal features

Data Preprocessing

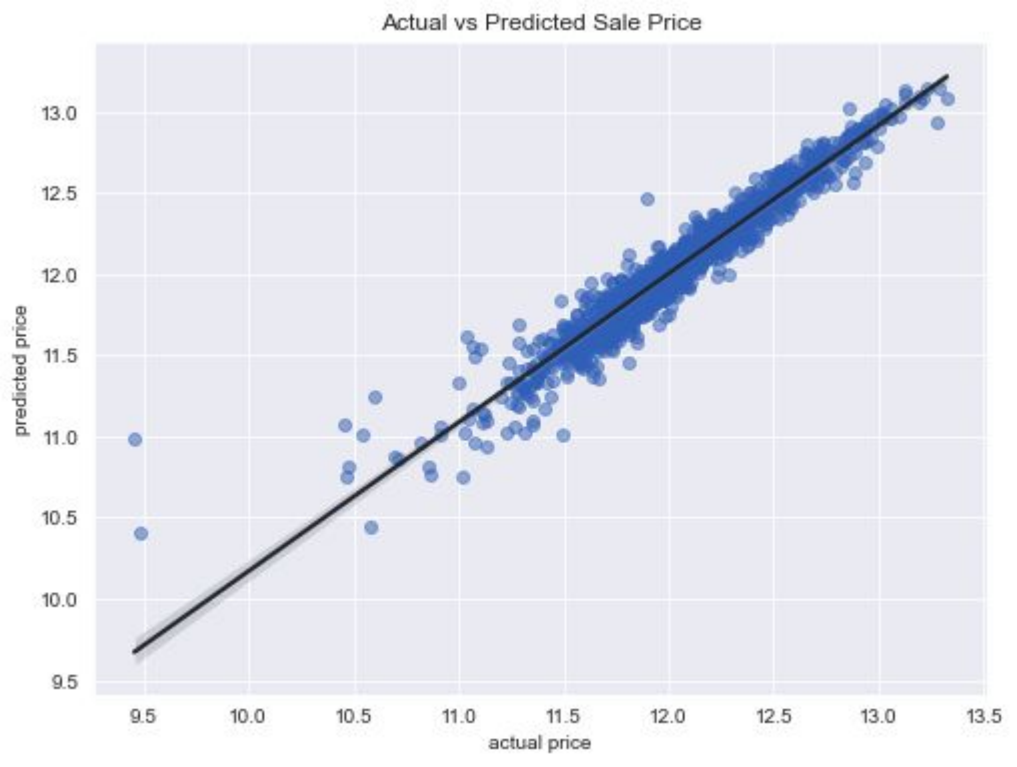
- **GridSearch**
  - Standard scaling
  - Missing values - **KNN Imputer** → $k = 5$
  - Feature selection - **Lasso** → $\alpha = 0.001$

Once again - more granular

- **GridSearch**
    - Standard scaling
    - Missing values - **KNN Imputer** → $k = 5$
    - Feature selection - **Lasso** → $\alpha = 0.0031$
        - feature reduction: 217 → **93**

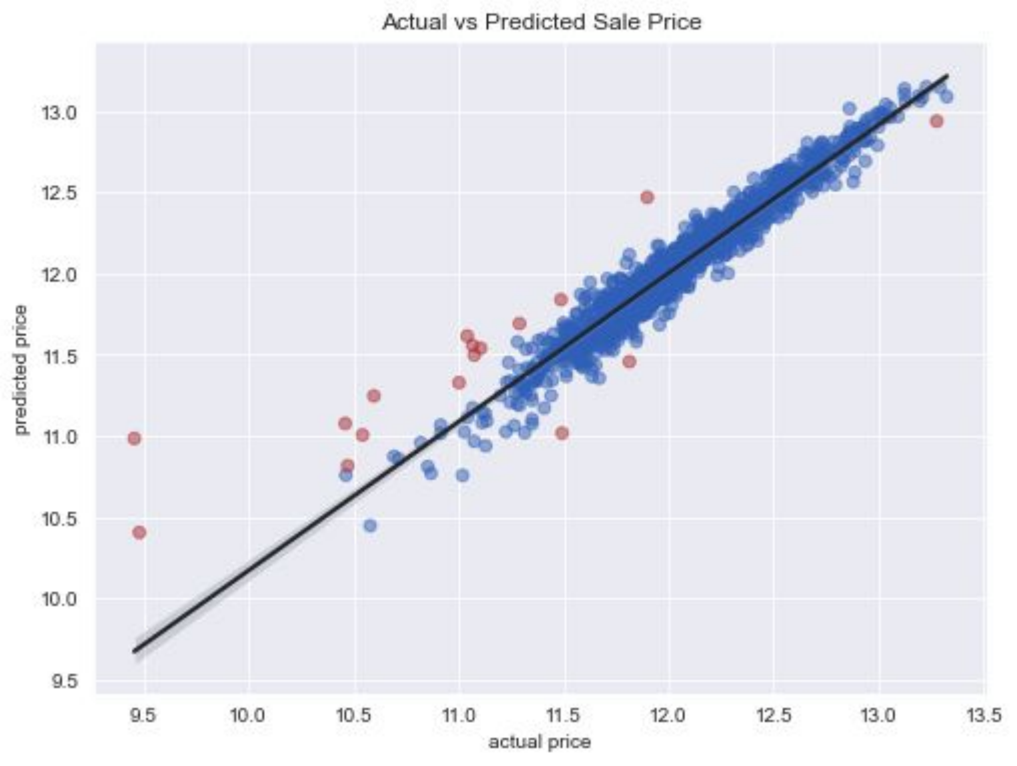
More Data Preprocessing

- **Outlier removal**
  - standardized residuals



More Data Preprocessing

- **Outlier removal**
  - standardized residuals
    - if > 3
    - if < -3
- Dropped 17 outliers



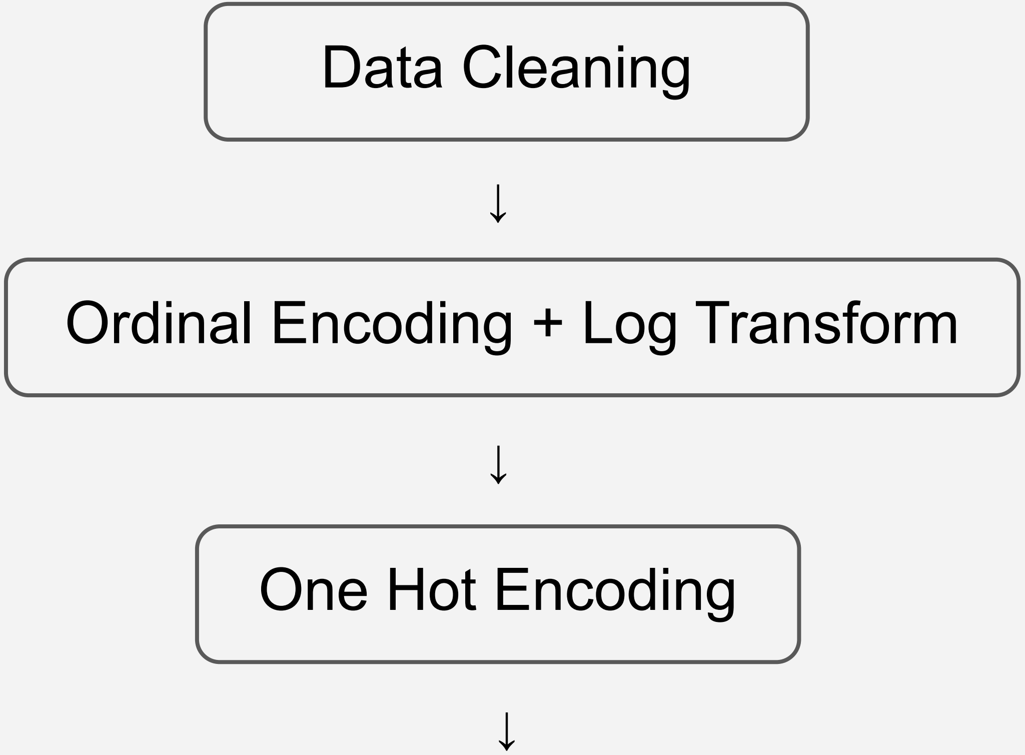
Modeling - LASSO, Ridge, ElasticNet, MLR

	LASSO		**Ridge**		**ElasticNet**		**MLR**	
	train	test	train	test	train	test	train	test
R² score	0.957	0.926	0.956	0.925	0.957	0.926	0.957	0.924
RMSE	15895	21053	15883	21163	15898	21057	15853	21132
CV score	0.949	0.913	0.949	0.909	0.949	0.912	negative	negative

Modeling - LASSO, Ridge, ElasticNet, MLR

	LASSO		**Ridge**		**ElasticNet**		**MLR**	
	train	test	train	test	train	test	train	test
R^2 score	0.957	0.926	0.956	0.925	0.957	0.926	0.957	0.924
RMSE	15895	21053	15883	21163	15898	21057	15853	21132
CV score	0.949	0.913	0.949	0.909	0.949	0.912	negative	negative

Recap



Data Cleaning

↓

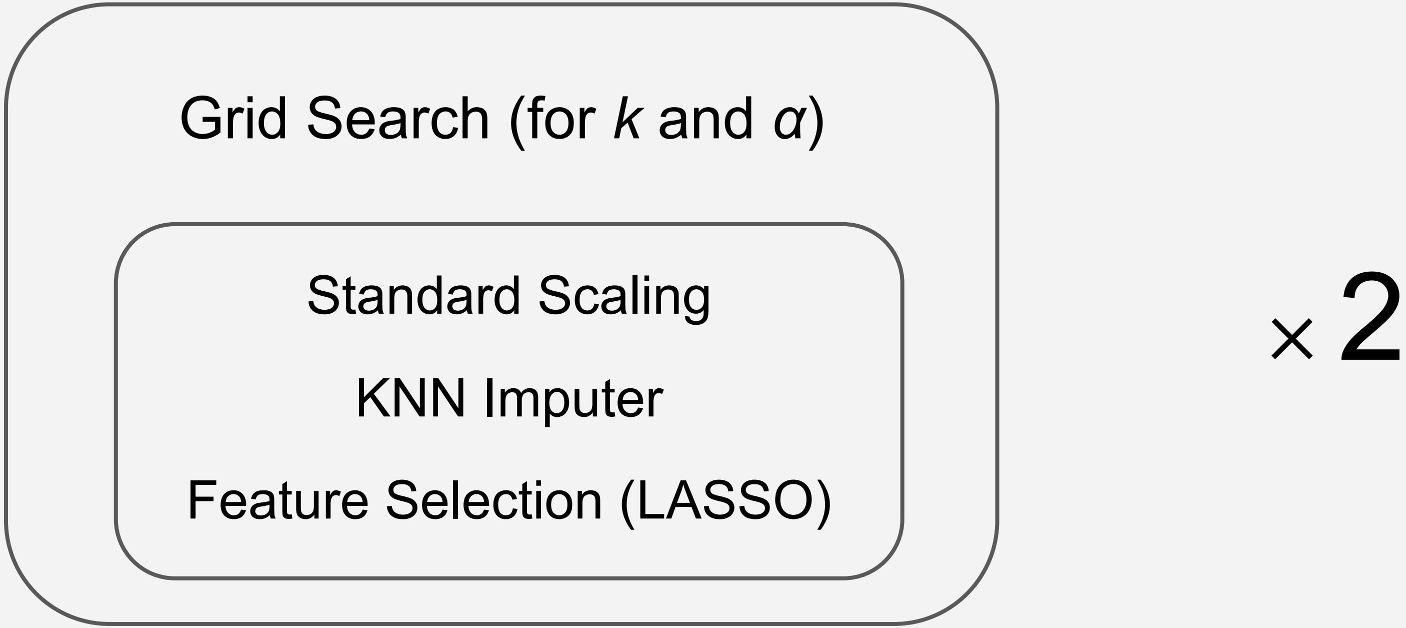
Ordinal Encoding + Log Transform

↓

One Hot Encoding

↓

Recap



Grid Search (for *k* and *α*)

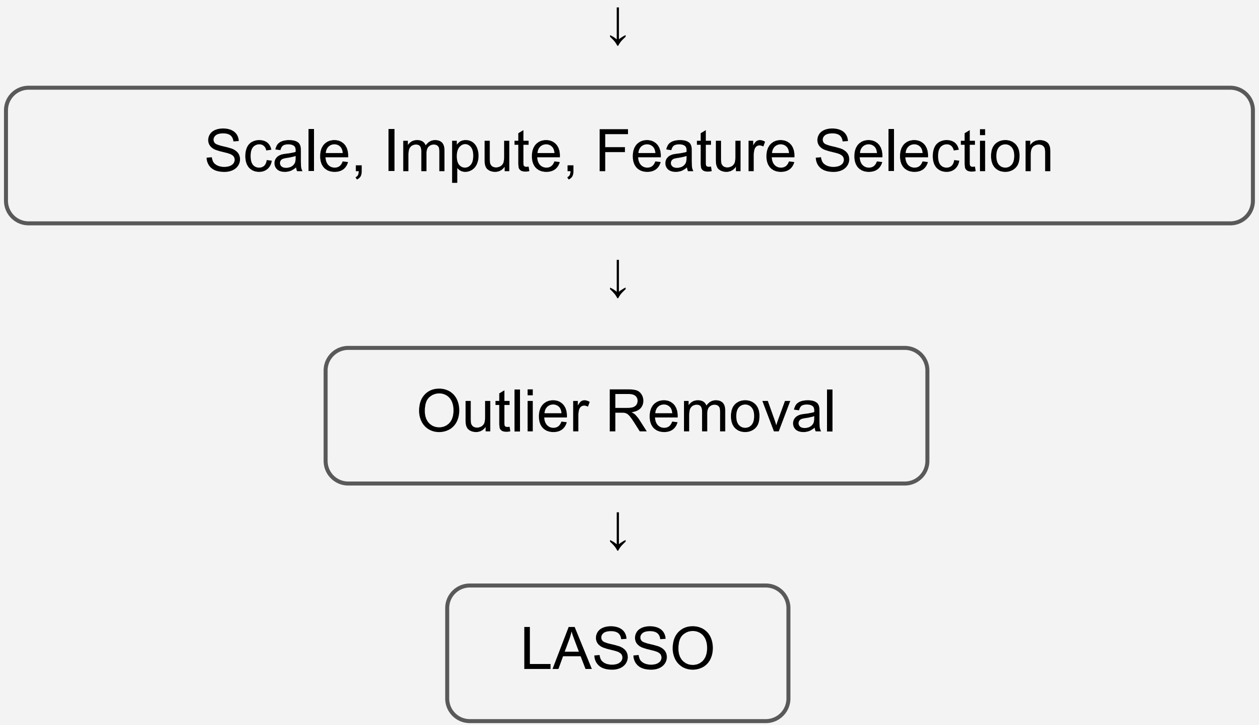
Standard Scaling

KNN Imputer

Feature Selection (LASSO)

×2

Recap



Scale, Impute, Feature Selection

Outlier Removal

LASSO

Results + Future Work

- 95% train, 92% test set can be explained (5-fold CV)

- Test the effect of train-test-split ratio
- Further feature reduction → clustering on OHE features
- Other advanced models…