Zillow®

Finding the Zestimate

What is Data Science
Department working on?

Overview:

- ⌂ What is the Zestimate and why do we need it?
- ⌂ How did we come up with the Zestimate?
- ⌂ How can we implement it?
- ⌂ Steps to further improve the Zestimate

What is the Zestimate and why do we need it?

- **Estimate**s the price of a house on **Z**illow.com
- Important resource for customers
- More listings, more revenue
- More listings, more data

How did we come up with the Zestimate?

- Linear regression algorithm that can predict the prices of houses using historical data

- **Ames Housing Data**:
  - Residential properties sold in Ames, IA from 2006 to 2010
  - 2980 rows (houses)
  - 82 columns (features)

How did we come up with the Zestimate?

⌂ **Data Cleaning**:

- 26 feature columns with Null Values!!!
- Dropped outliers

⌂ **Feature Engineering and Data Transformations**:

- Log Transformation of sale price
- 40 feature columns transformed to dummies or ordinal data
- Build new features out of correlated features

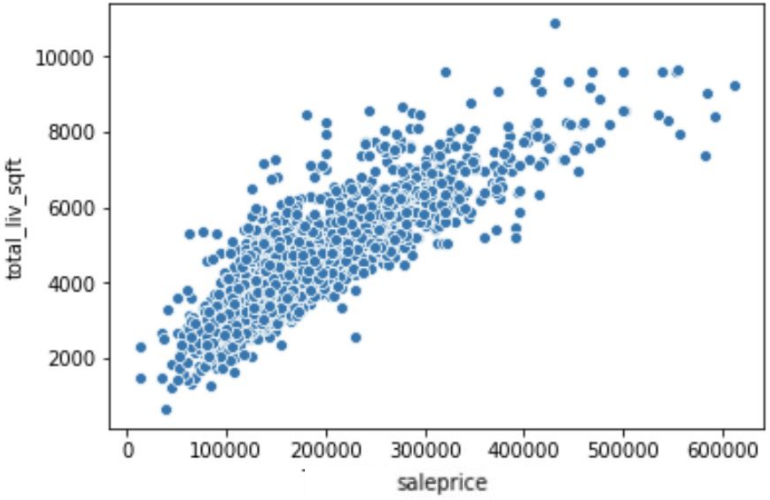
Metrics

- **R^2:** Percentage of variability in the data explained by Model
- **Cross-Val-R^2 :** R^2 for 5 fold cross validation within the train data
- **Mean Squared Error:** Mean of the squared residuals

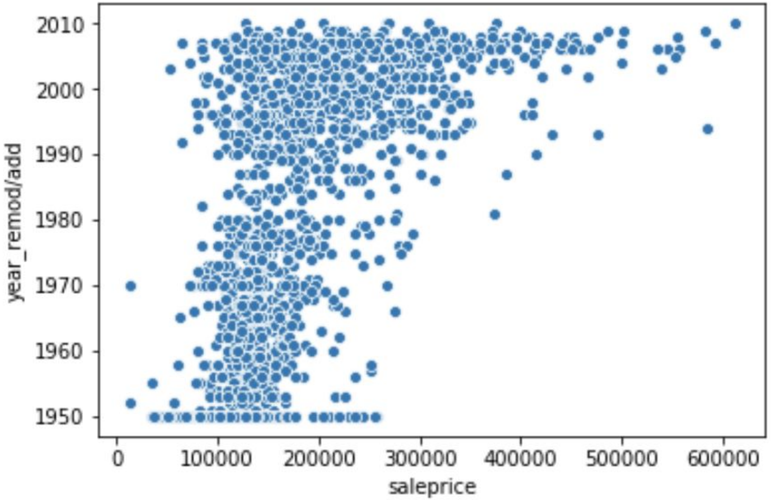
First Models: Linear Regression, no Regularization

⌂ Feature Selection:

Feature with a correlation coefficient with sale price

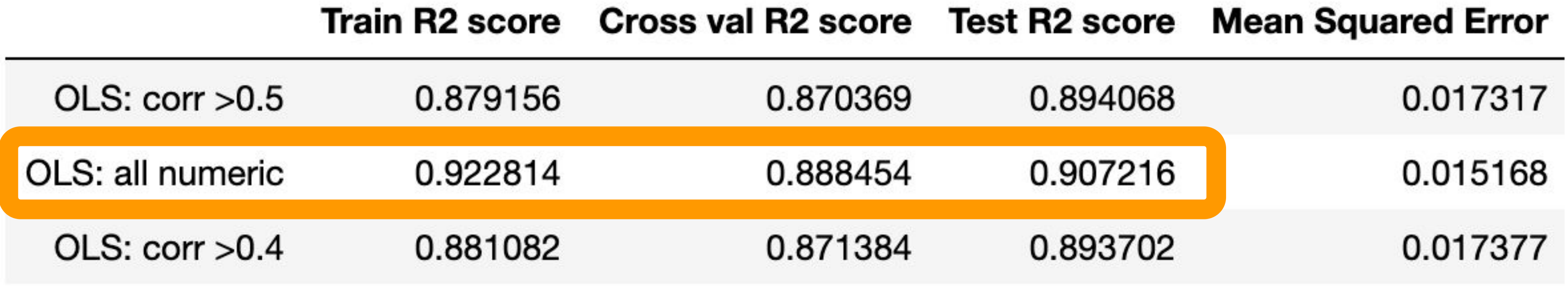


Total Living Sqft: corr 0.8



Year Remodeled: corr 0.55

First Models: Linear Regression, no Regularization

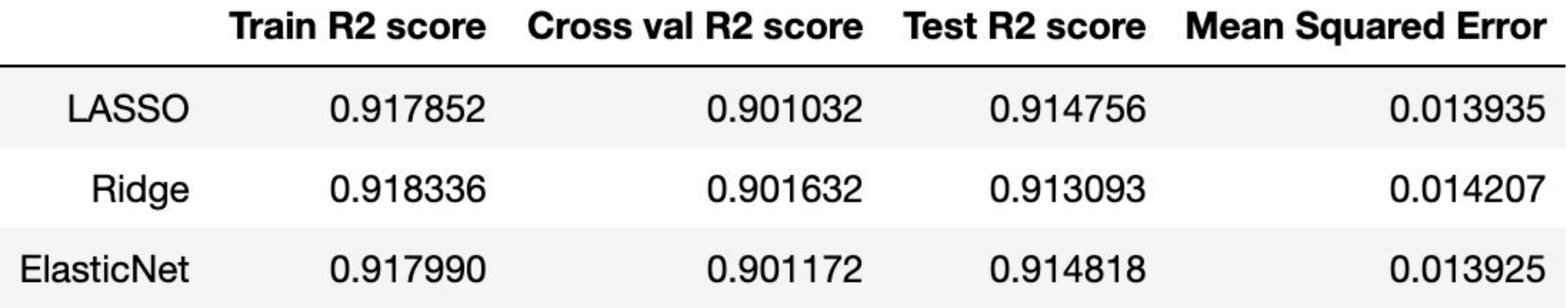


	Train R2 score	Cross val R2 score	Test R2 score	Mean Squared Error
OLS: corr >0.5	0.879156	0.870369	0.894068	0.017317
OLS: all numeric	0.922814	0.888454	0.907216	0.015168
OLS: corr >0.4	0.881082	0.871384	0.893702	0.017377

OLS corr > 0.5: 20 features
OLS corr > 0.4: 24 features

OLS all numeric: 106 features
→ Overfit!

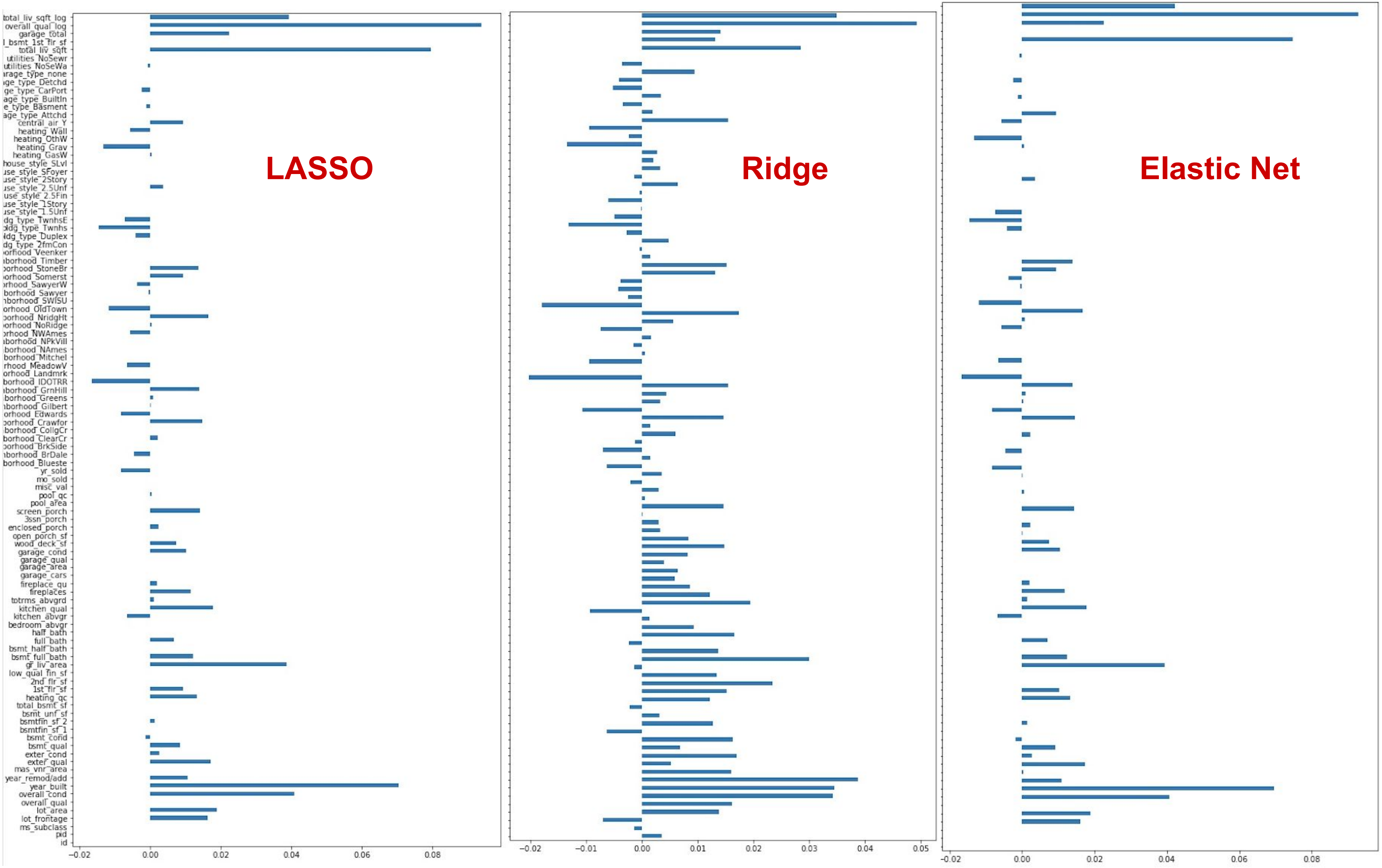
Improved Models: Linear Regression with Regularization



	Train R2 score	Cross val R2 score	Test R2 score	Mean Squared Error
LASSO	0.917852	0.901032	0.914756	0.013935
Ridge	0.918336	0.901632	0.913093	0.014207
ElasticNet	0.917990	0.901172	0.914818	0.013925

Performance against external Test Data:
Ridge > Elastic Net > Lasso

WARNING


LASSO
Ridge
Elastic Net
total_liv_sqft_log
overall_qual_log
garage_total
l_bsmt_1st_flr_sf
total_liv_sqft
utilities_NoSewr
utilities_NoSeWa
arage_type_none
ige_type_Detchd
ge_type_CarPort
age_type_BuiltIn
e_type_Basment
age_type_Attchd
central_air_Y
heating_Wall
heating_OthW
heating_Grav
heating_GasW
house_style_SLvl
use_style_SFoyer
use_style_2Story
use_style_2.5Unf
use_style_2.5Fin
use_style_1Story
use_style_1.5Unf
dg_type_TwnhsE
oldg_type_Twnhs
ldg_type_Duplex
dg_type_2fmCon
orhood_Veenker
iborhood_Timber
oorhood_StoneBr
orhood_Somerst
rhood_SawyerW
borhood_Sawyer
nborhood_SWISU
orhood_OldTown
oorhood_NridgHt
orhood_NoRidge
rhood_NWAmes
iborhood_NPkVill
borhood_NAmes
borhood_Mitchel
rhood_MeadowV
orhood_Landmrk
borhood_IDOTRR
iborhood_GrnHill
borhood_Greens
nborhood_Gilbert
orhood_Edwards
oorhood_Crawfor
borhood_CollgCr
borhood_ClearCr
oorhood_BrkSide
nborhood_BrDale
borhood_Blueste
yr_sold
mo_sold
misc_val
pool_qc
pool_area
screen_porch
3ssn_porch
enclosed_porch
open_porch_sf
wood_deck_sf
garage_cond
garage_qual
garage_area
garage_cars
fireplace_qu
fireplaces
totrms_abvgrd
kitchen_qual
kitchen_abvgr
bedroom_abvgr
half_bath
full_bath
bsmt_half_bath
bsmt_full_bath
gr_liv_area
low_qual_fin_sf
2nd_flr_sf
1st_flr_sf
heating_qc
total_bsmt_sf
bsmt_unf_sf
bsmtfin_sf_2
bsmtfin_sf_1
bsmt_cond
bsmt_qual
exter_cond
exter_qual
mas_vnr_area
year_remod/add
year_built
overall_cond
overall_qual
lot_area
lot_frontage
ms_subclass
pid
id
−0.02
0.00
0.02
0.04
0.06
0.08
−0.02
−0.01
0.00
0.01
0.02
0.03
0.04
0.05
−0.02
0.00
0.02
0.04
0.06
0.08

Next steps in the implement of the Zestimate

- Right now: only applicable for Ames, IA
- Future: Build similar model with the data from zillow.com
  - Predictions for the whole US
  - Include only features that are readily available

Steps to further improve the Zestimate in the future:

⌂ Experiment more with features and regularizations

⌂ Incorporate Location Data more strongly

⌂ Start a Data Science Competition on kaggle.com

Questions?