Ames, Iowa pricing algorithm

Choosing a best-fit model for pricing homes in a new market

Why We Are Here

Company

Zillow is moving into a new market and needs a new algorithm to effectively price houses based on their data.

Context

Data:
Metadata about over 2,000 homes

80 columns of data for each home

Problem statement

Develop an algorithm to be used in a production application to estimate the price of a home based on metadata about the property & determine the potential for generalizability

Model Choice Process

Data Import, Cleaning & Design

Preparing for Prediction

Little data corruption.

The process of one hot encoding, ordinal variables encoding & filling null values went quickly.

Initial Modeling

Iteratively Choosing

Starting with high correlations and progressively shrunk the feature list.

Focus on simplest model with highest score.

Model Choice & Automation

Manually Chosen Features vs Adaptive Choice

Lasso, Ridge or Random Forest

Initial Correlations & Feature Choices

These features all had a positive correlation of greater than 0.3 to our target and thus were used in the first model.

overall_qual	0.805177
gr_liv_area	0.719053
exter_qual	0.717620
kitchen_qual	0.694905
total_bsmt_sf	0.667040
garage_area	0.658371

1st_flr_sf	0.651359
bsmt_qual	0.620888
year_built	0.578784
garage_yr_blt	0.569289
year_remod/add	0.551630
fireplace_qu	0.537147

full_bath	0.533397
mas_vnr_area	0.521920
totrms_abvgrd	0.508559
fireplaces	0.470045
heating_qc	0.462376
bsmtfin_sf_1	0.447577

lot_area	0.368378
open_porch_sf	0.360706
wood_deck_sf	0.337710

Initial Correlations & Feature Choices

These features all had a positive correlation of greater than 0.3 to our target and thus were used in the first model.

This initial model accounted for between 83% and 88% of the variation in the final sale price.

Narrowing Features

by coefficients, p-values & interactions

Not all the features in our initial model were impacting our accuracy. By pruning out those variables with a p-value less than 5% we simplified our model.

Some features were transformed or combined in interaction variables to reflect their impact on each other.

This improved our model to between 88 and 90% in accounting for variation.

Narrowing Features
by algorithmic detection of features

By using algorithmic approaches to parameter tuning and feature selection the hope is to develop a more effective model beyond what you can do by hand.

OLS	Lasso and Ridge	Random Forest Feature Selection for OLS
88 - 90% 28 features Scaling data did not affect the outcome	88-90% Identical features Scaling data did not affect the outcome	90-92% Over 125 features, was overfit and could not recorrect it.

Narrowing Features
algorithmic detection of features

By using algorithmic approaches to parameter tuning and feature selection the hope is to develop a more effective model beyond what you can do by hand.

OLS	This model contained 28 features, easily selectable and manipulated with a minimum of computing power and time. It resulted in nearly equivalent results to the Random Forest and Lasso/Ridge methodologies. Less over-fit and more generalizable.
88 - 90% 28 features Scaling data did not affect the outcome	

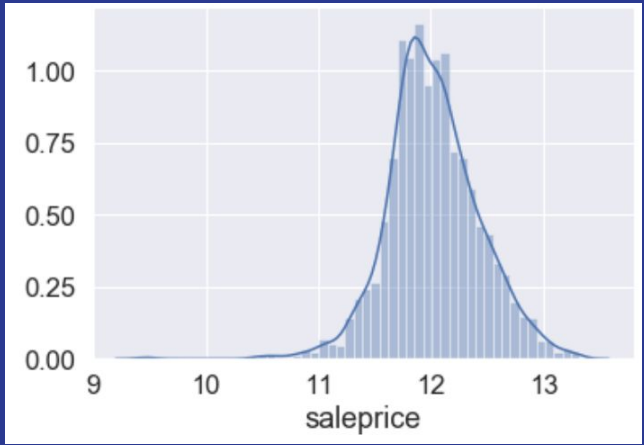
Combining
Polynomial Transformation & Random Forest

By applying Polynomial transformation to our original 28 features and then applying Random Forest the model was improved

Ridge Regression		
88 - 91% Scaling data did not affect the outcome	Polynomial transformation of 28 features. This model contains 107 features. Balanced bias and variation.	Further analysis would require more computing power to do stronger Polynomial + Random Forest searches.

Model Information

log() transformed our target for normal distribution



	OLS (88 - 90%)	**Ridge & Random Forest** (88 - 91%)
Untouched Features	21	4
Transformed Features	5	11
Interaction Features	2	92

Insights About Ames

By applying Polynomial transformation to our original 28 features and then applying Random Forest the model was improved

Desirable Neighborhoods	**Home Features to look out for and look into improving**	
Stonebrook Northridge Heights Northridge	Square Footage of the 1st Floor and Living area at Ground Level	The overall quality measure Central Air Condition improvements

Moving Forward into new markets

Our **model** is not generalizable.

But our **approach** is!

Future plans:

- Real time, iterative modeling based on most recent sales.
- Use of parcel and lot location information.
- Implement GridSearching and more in depth Random Forest techniques