

Is it worth it?

Predicting house prices in Ames






Problem Statement

Purchasing a house is a big decision for both investors and those looking for a place to stay. One of the key considerations behind this decision is the price of the property is reasonable.


Goal

To build an accurate model to predict the expected sale price of a house in Ames. The key metric used will be root mean squared error.





Overview


1. Background research on factors which affect house prices
 2. Data processing
 3. Analysis/ model building
 4. Results
 5. Recommendations
 6. Future Directions
- 



Background research on factors which affect house prices

Based on online articles, several factors were known to influence house prices.

Home Related

1. Location, neighborhood, nearby features
 2. Home size, layout
 3. Home age
 4. Property condition
- 




Data processing

Dataset

Information on residential properties sold in Ames, Iowa from 2006 to 2010. Contains a mix of 82 variables with information on the properties.

Data cleaning and exploration

1. Processing NA values
 2. Exploring categorical features
 3. Exploring continuous features
- 

Data cleaning and exploratory data analysis (EDA)

Processing NA values

- Fill with non for categorical features
- Fill with 0 for continuous features
- Deductive imputation

bsmt_type_2	bsmtfin_sf_1	bsmt_type_2	bsmtfin_sf_1	bsmtfin_unf_sf	total_bsmtfin_sf
GLQ	1124	NaN	479	1603	3206

Data cleaning and exploratory data analysis (EDA)

Processing NA values

- Fill with non for categorical features
- Fill with 0 for continuous features
- Deductive imputation

bsmt_type_2	bsmtfin_sf_1	bsmt_type_2	bsmtfin_sf_1	bsmtfin_unf_sf	total_bsmtfin_sf
GLQ	1124	Unf	479	1603	3206

Data cleaning and exploratory data analysis (EDA)

Exploring categorical features

- Remove features with low counts in categories
- Change ordinal features to numeric
- Remove ordinal features that are unordered

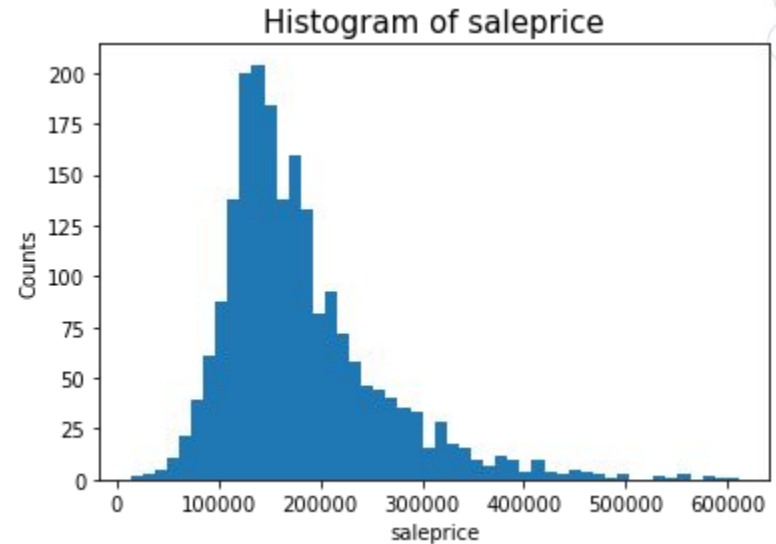
Utilities	Count
NoSewr	1
NoSewa	1
Allpub	2049

garage_cond	Count	Mean saleprice~
Poor	8	89,000
none	114	106,000
Fair	47	107,000
Excellent	2	124,000
Average	1868	188,000
Good	12	207,000

Data cleaning and exploratory data analysis (EDA)

Exploring continuous features

- Distribution of saleprice is approximately normal
-

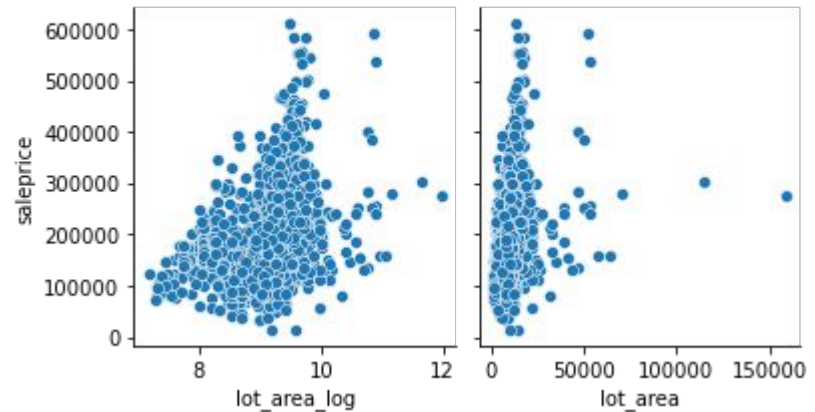


Data cleaning and exploratory data analysis (EDA)

Exploring continuous features

- Distribution of saleprice is approximately normal
- Transformation of features
-

Scatterplot of lot_area and its transformation

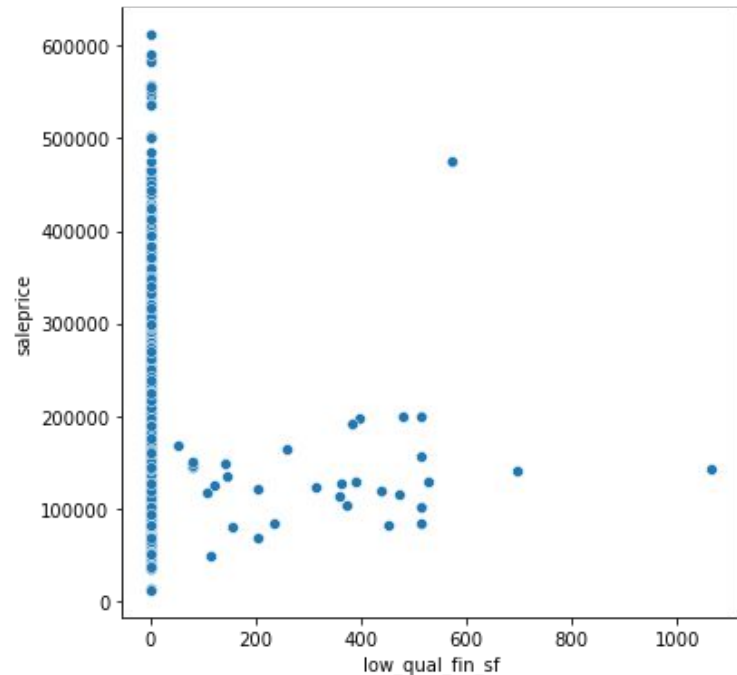


Data cleaning and exploratory data analysis (EDA)

Exploring continuous features

- Distribution of saleprice is approximately normal
- Transformation of features
- Removal of variables with poor linear relationship

Scatterplot of saleprice and low_qual_fin_sf

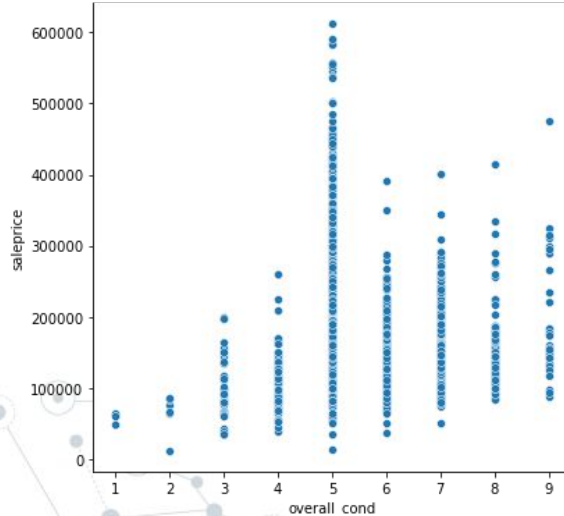


Data cleaning and exploratory data analysis (EDA)

Exploring continuous features

- *Special variables*

Scatterplot of saleprice and overall_cond

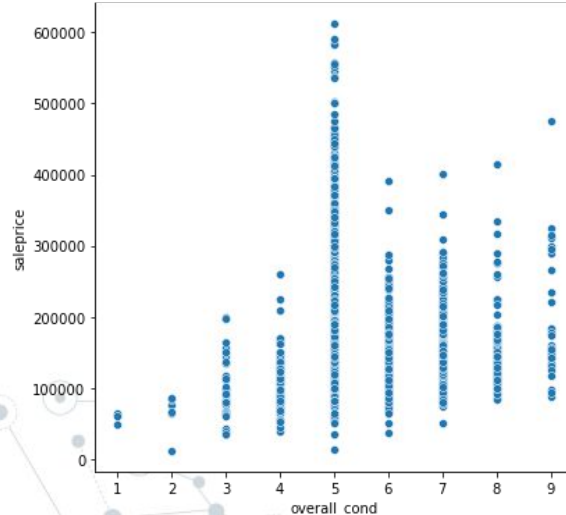


Data cleaning and exploratory data analysis (EDA)

Exploring continuous features

- Special variables*

Scatterplot of saleprice and overall_cond



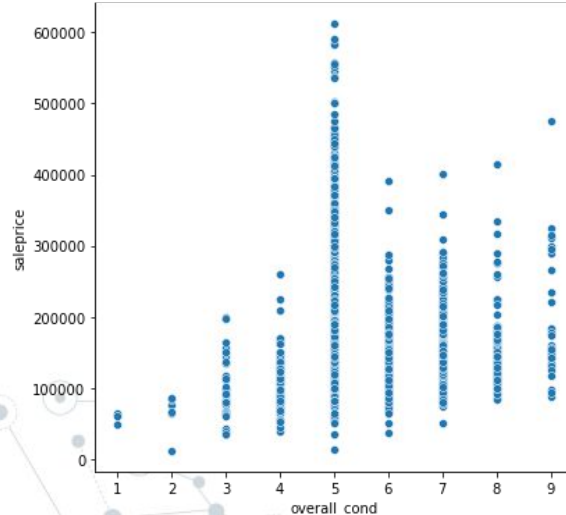
overall_cond	Count	Mean saleprice~
1	4	59,000
2	6	65,000
3	35	100,000
4	70	114,000
6	368	149,000
7	270	155,000
8	101	156,000
9	29	198,000
5	1168	207,000

Data cleaning and exploratory data analysis (EDA)

Exploring continuous features

- Special variables*

Scatterplot of saleprice and overall_cond



overall_cond	Count	Mean saleprice~
1	4	59,000
2	6	65,000
3	35	100,000
4	70	114,000
6	368	149,000
7	270	155,000
8	101	156,000
9	29	198,000
5	1168	207,000

Analysis/Model Building

Model building will be performed in a step-wise fashion. At each step, variables will be added and those performed poorly will be removed before the next step.

Model 1: Variables supported by background research

Model 2: To add the remaining continuous variables

Model 3: To add the remaining categorical variables

Model 4: Polynomial features

Analysis/Model Building

Model 1

Variables supported by
background research

+ new continuous
features

Model 2

- poor features
+ new categorical
features

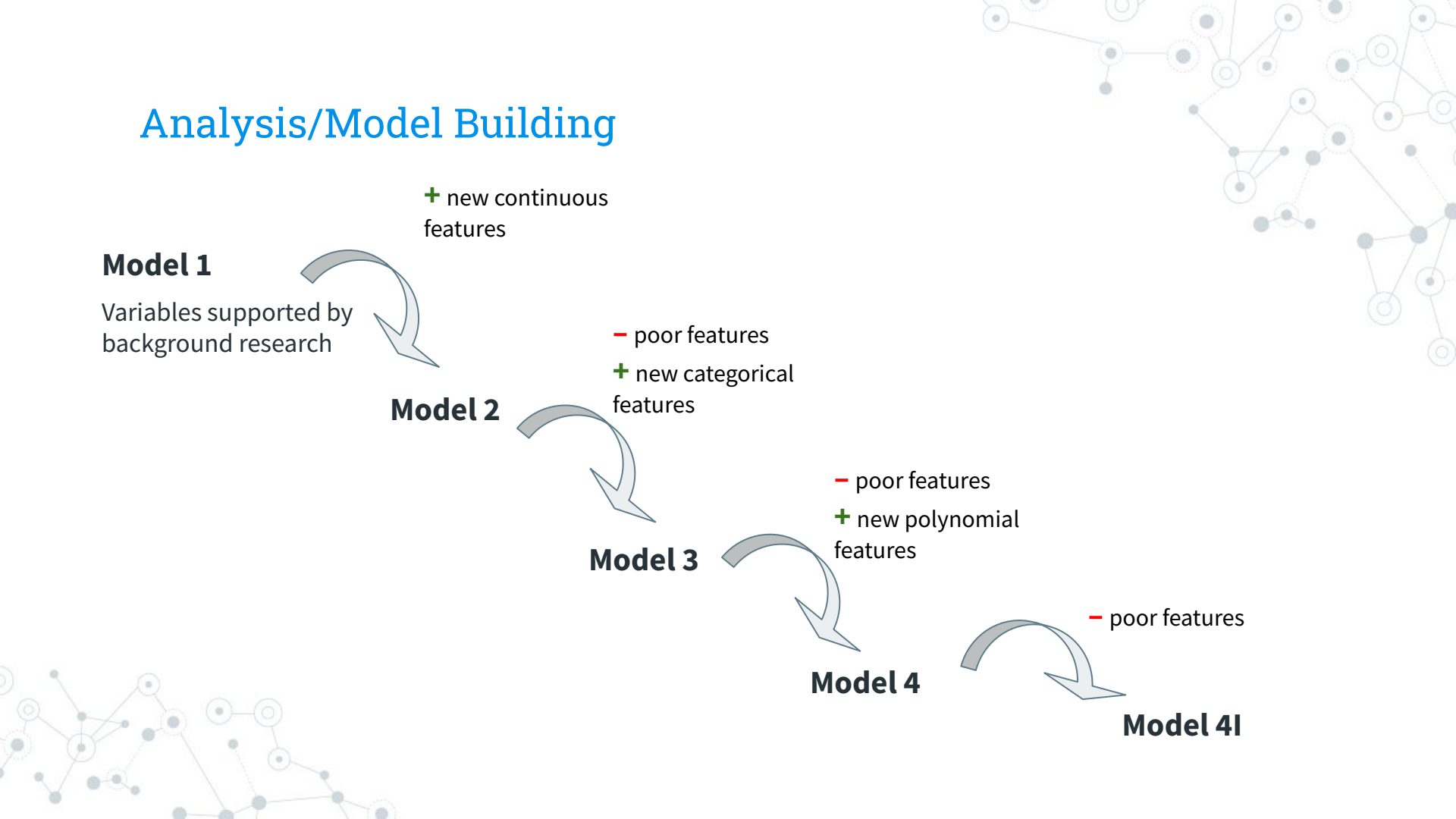
Model 3

- poor features
+ new polynomial
features

Model 4

- poor features

Model 4I



Analysis/Model Building

Analysis performed

1. Linear Regression
2. Ridge Regression
3. Lasso Regression
4. Elastic Net Regression

Metrics used

1. Mean Square Error
2. Root Mean Square Error

Model 1 (Base Model)

No.	Important factors based on research	Model Features (17 features)
1	Location, neighborhood, nearby features	Zone, Neighborhood Nearby streets
2	Home size, layout	Lot area,Ground living area, Garage area, Basement size
3	Home age	House age, Garage age
4	Property Condition	Overall Quality, Garage Quality Basement Condition, Basement Quality, Exterior Quality

Model 4I (Final Model)

57 features, 135 after dummifying

ms_subclass', 'ms_zoning', 'lot_frontage', 'land_contour', 'lot_config', 'neighborhood' 'condition_1', 'condition_2',	'bldg_type', 'overall_qual', 'overall_cond', 'house_age', 'year_remod/add', 'mas_vnr_area', 'exter_qual', 'bsmt_qual', 'bsmt_cond',	kitchen_qual', 'fireplaces', 'fireplace_qu', 'garage_age', 'garage_finish', 'garage_area', 'wood_deck_sf', 'open_porch_sf', 'lot_area_log'	'bsmt_exposure', 'bsmtfin_type_1', 'bsmtfin_sf_1', 'bsmtfin_type_2', 'total_bsmt_sf', 'heating', 'heating_qc', '1st_flr_sf', 'gr_liv_area',
--	---	--	---

Model 4I (Final Model)

57 features, 135 after dummifying

'Overall_qual * kitchen_qual',
'Overall_qual * total_bsmt_sf'
'Garage_area * overall_qual'
'Total_bsmt_sf * gr_liv_area'
'Overall_qual * bsmt_qual'
'Overall_qual * year_remod/add'
'Garage_area * gr_liv_area'
'1st_flr_sf * exter_qual'
'Total_bsmt_sf * kitchen_qual'
'1st_flr_sf * kitchen_qual'

Results

Model	Ridge	Lasso	
	holdout_data		test_data
Model 1	28580	28541	28273
Model 2	22878	23196	22777
Model 3	24347	24015	24848
Model 3l	24022	23897	22763
Model 4	18592	18739	22404
Model 4l	17470	17691	21252

cvs = cross validated score


All numbers represent root mean square error



Conclusions and Recommendations

The models above will be able to predict the sale price of houses to reasonable accuracy. The migration/relocation firm can confidently use this model. We recommend users to use **Model 4I using Ridge regression** to predict sale prices as it has the best accuracy.

If a scientist needs to build a new house price prediction model with with **limited time** and **resources**, known factors based on research should be incorporated as it provides the most value.



Future Directions

1. To explore variables deeper
 - Sale Type
 - Mas Vnr Type
 - House Style
2. Include an inflation factor

Recommendations and Insights (As an Analyst)

1. Perform background research well
2. Be more liberal in dropping features that are not convincing during data exploration
 - a. Save time
 - b. Could have explored more polynomial features



HAPPY
HOUR