

Is it worth it?

Predicting house prices in Ames



Problem Statement

As a scientist working in a relocation/migration firm in Hong Kong, you have been tasked to build a model which can predict the expected sale price of houses

Since the potential buyers are NOT intending for these houses to be their forever homes, they will hope to get the houses at a good price so that they do not lose too much money when they eventually sell it.


Goal

To build an accurate model to predict the expected saleprice of a house in Ames.






Overview

1. Background research on factors which affect house prices
 2. Data cleaning and exploration
 3. Analysis/ Model Building
 4. Results
 5. Recommendations
 6. Future Directions
- 



Background research on factors which affect house prices

Home Related

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



Dataset for analysis


Information on individual residential properties sold in Ames, Iowa from 2006 to 2010.

Contains a mix of 82 variables with information on the properties.





Data cleaning and exploratory data analysis (EDA)

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)

Exploring categorical features

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

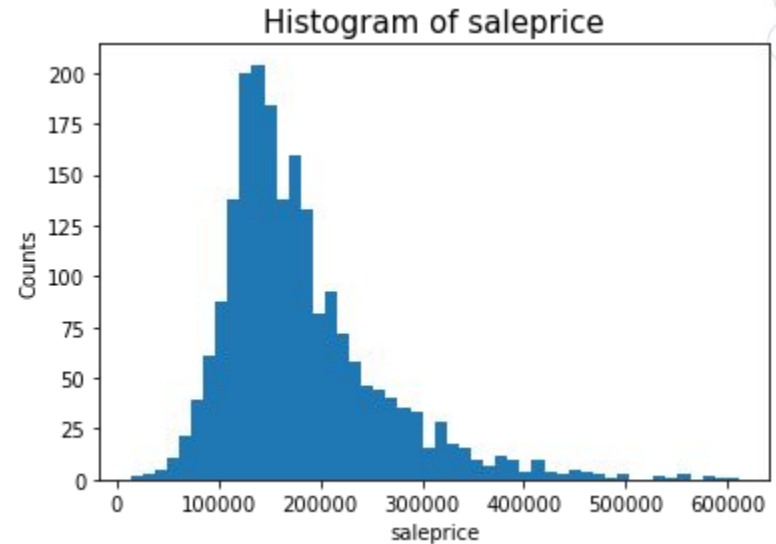
Utilities	Count
NoSewr	1
NoSewa	1
Allpub	2049

garage_cond	Count	Mean saleprice~
Po	8	89,000
non	114	106,000
Fa	47	107,000
Ex	2	124,000
Ta	1868	188,000
Gd	12	207,000

Data cleaning and exploratory data analysis (EDA)

Exploring continuous features

- Distribution of saleprice is approximately normal
-

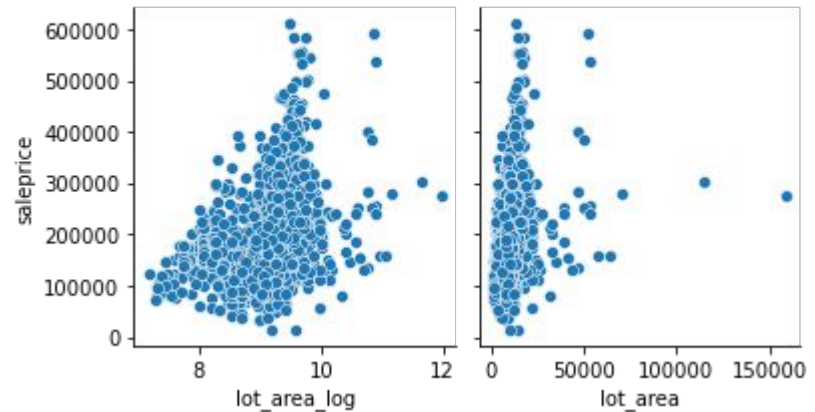


Data cleaning and exploratory data analysis (EDA)

Exploring continuous features

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

Scatterplot of lot_area and its transformation

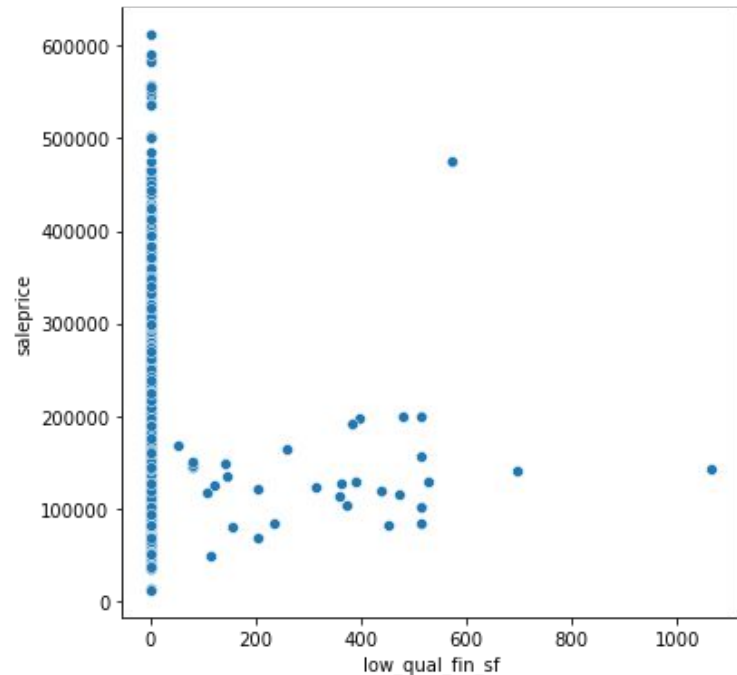


Data cleaning and exploratory data analysis (EDA)

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

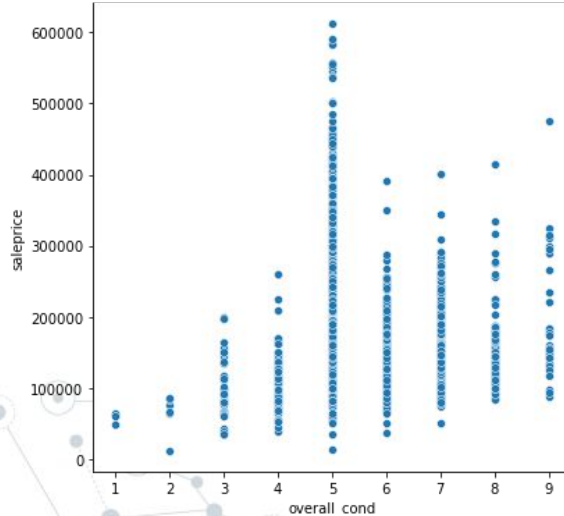


Data cleaning and exploratory data analysis (EDA)

Exploring continuous features

- *Special variables*

Scatterplot of saleprice and overall_cond

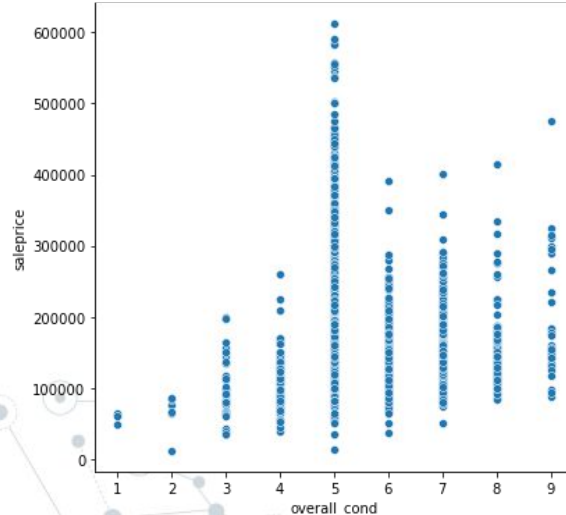


Data cleaning and exploratory data analysis (EDA)

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



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.

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

- poor features
- + new continuous
features

Model 2

- poor features
- + new categorical
features

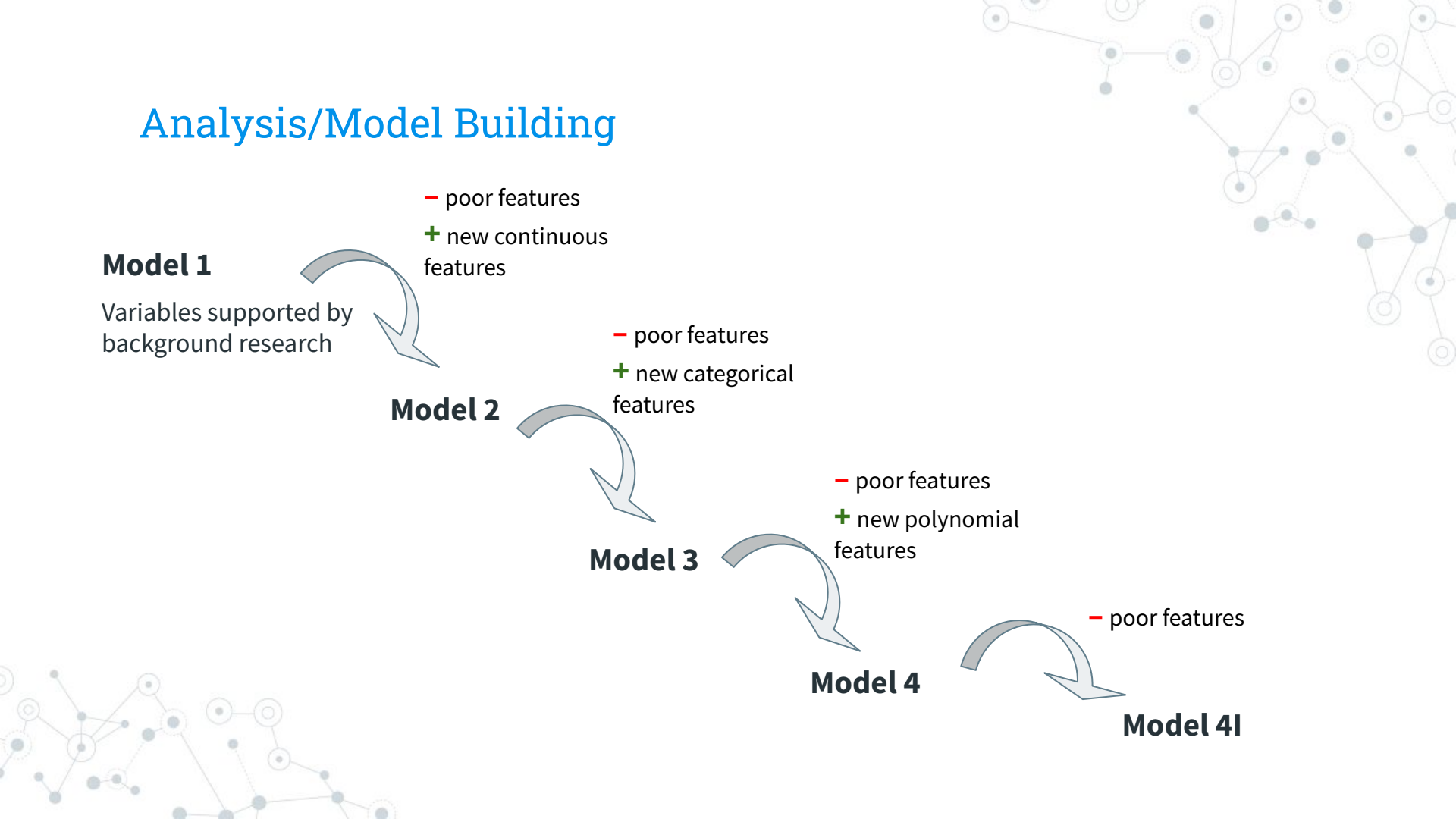
Model 3

- 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	Zone, Neighborhood Nearby streets
2	Home size, layout	Lot area, Lot Frontage 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(cvs)	Lasso(cvs)	Ridge	Lasso	Kaggle
	train_data		holdout_data		test_data
Model 1	26980	27431	28580	28541	28273
Model 2	25972	25906	22878	23196	22777
Model 3	21239	21272	24347	24015	24848
Model 3l	22867	22848	24022	23897	22763
Model 4	19948	19477	18592	18739	22404
Model 4l	21066	21152	17470	17691	21252



Conclusions and Recommendations

The models above will be able to predict the sale price of houses to reasonable accuracy. We recommend users to use Model 4I using enet regression to predict sale prices as it has the best accuracy.

It is recommended to use features similar to that in the base model if one needs to build a model with limited time and resources, as it provides a reasonable model.



Recommendations (As an Analyst)

1. Perform background research well.
2. Be more liberal in dropping features that will clearly not perform.
 - a. Save time
 - b. Could have explored more polynomial features
3. Unstable results, for rmse, but fall within a range

Future Directions

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



HAPPY
HOUR