House Pricing Prediction

Yiming Lai, PhD

Mentored by Wayne Ang

Springboard Data Science Capstone Project

Problem statement

House buying or selling is a long and uncertain process especially for the first time buyers/seller.
In this capstone we try to answer the most critical and probably the first question we ask:

For the house buyers:

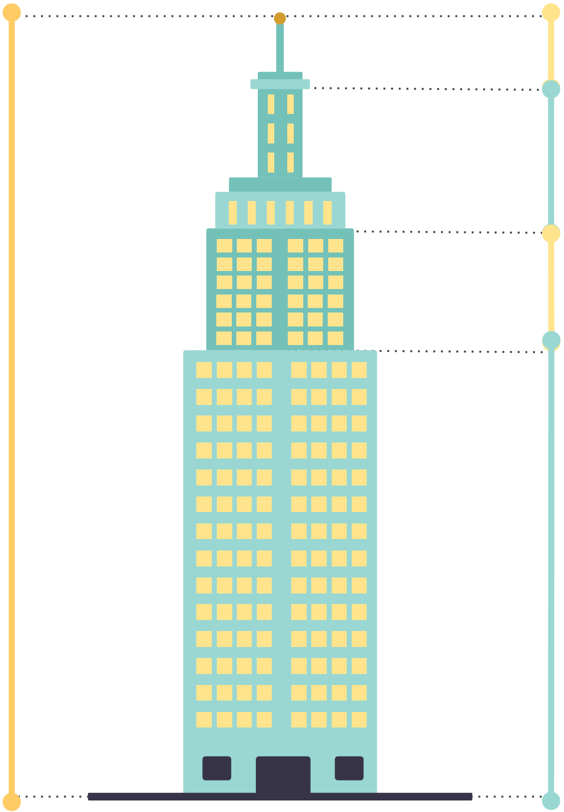
How much should I bid on the house without overpaying?

For the house sellers:

How much should I label the sale price for the house?

What factors may affect the house pricing?

$Total Budget



Others?
- Age of the house, interior/exterior qualities, utilities…etc

Size of the house?
- Living area, lot size

Type of the house?
- Single family, town house, condo…etc

Location, location, location?
- School district, downtown, rural area…etc

Data information

- Data content: Sale price of the individual residential property in Ames, Iowa from 2006 to 2010
- Number of record: 1460
- Number of features: 79
- Source: Kaggle (https://www.kaggle.com/c/house-prices-advanced-regression-techniques), compiled by Dean De Cock

Data engineering

Quick glance of the data set

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub

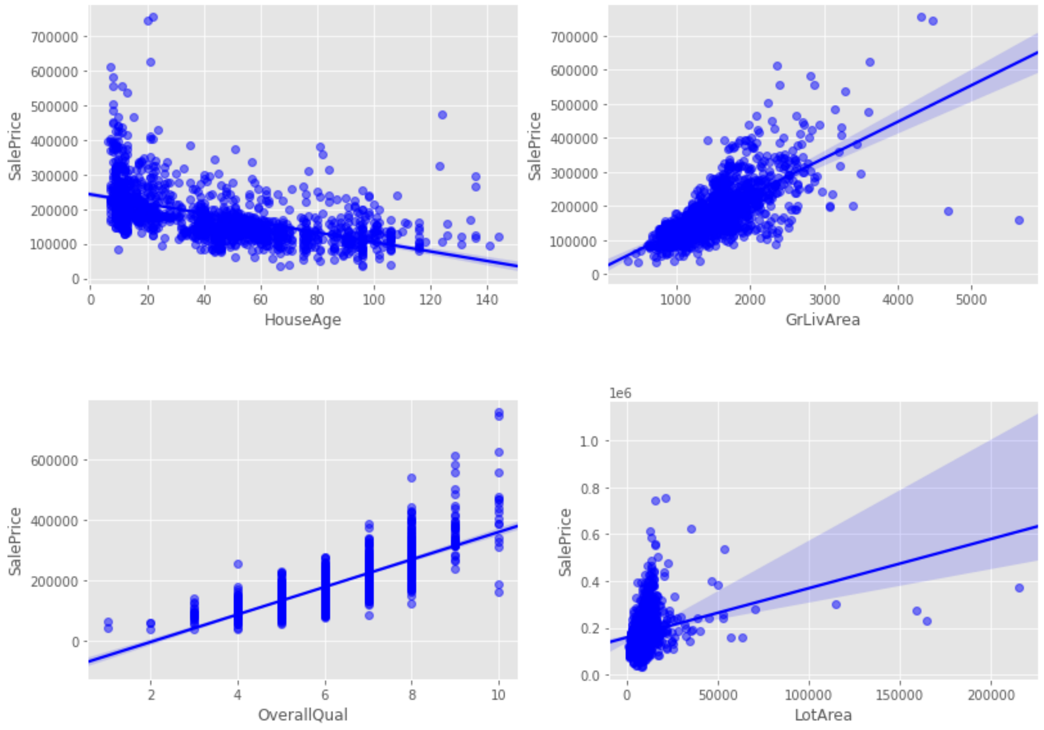
⋮

PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	SalePrice
0	NaN	NaN	NaN	0	2	2008	WD	Normal	208500
0	NaN	NaN	NaN	0	5	2007	WD	Normal	181500
0	NaN	NaN	NaN	0	9	2008	WD	Normal	223500
0	NaN	NaN	NaN	0	2	2006	WD	Abnorml	140000
0	NaN	NaN	NaN	0	12	2008	WD	Normal	250000

Key steps for data cleaning and wrangling

- Remove features PoolQC, MiscFeature, PoolArea, Alley and Fence since more than 80% of data in those features are missing
- Replace the missing values in categorical features such FireplaceQu, GarageFinish, etc with None to indicate the house doesn't have such feature
- Replace the missing values in numerical features such as GarageArea, GarageCars with 0 to indicate the house doesn't have such feature
- Replace the missing value in the LotFrontage by the mean value in the specific neighborhood the house belongs to
- Replace the rest of the missing values by the most common value in the corresponding neighborhood
- Convert the YearBuilt feature to the new feature HouseAge

Data exploratory and analysis – Numerical variables

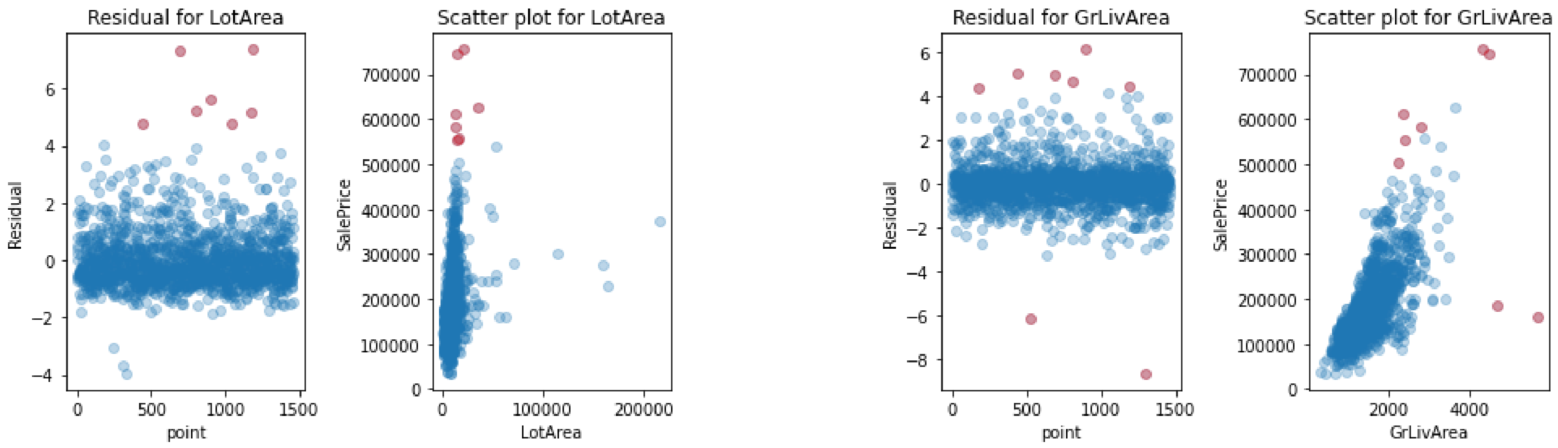


Observations

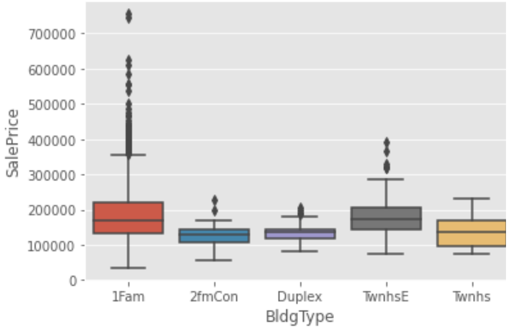
- The sale price is inversely proportional to the age of the house, but the correlation is not strong
- In general, the sale price is proportional to the size of the house. Some possible outliers were observed
- Better the quality of the overall material and finish of the house, higher the sale price

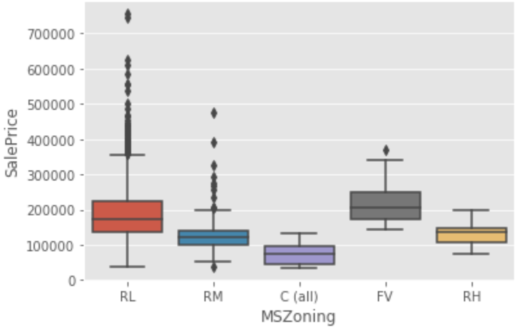
Data exploratory and analysis – Numerical variables

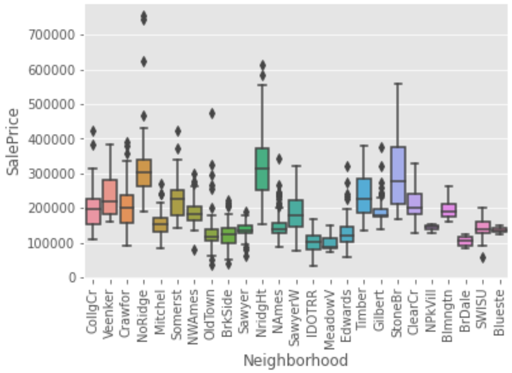
Calculate the studentized residuals and remove the data points that have the corrected p-values less than 0.05

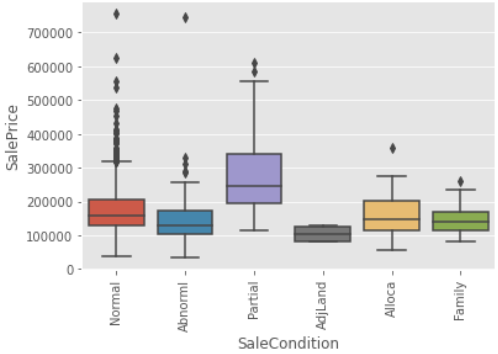


Data exploratory and analysis – Categorical variables









Observations

- Not much difference in the sale price between different building type but most of the high price houses are single family houses
- The houses in the low density residential area tends to have higher sale price which the houses in the commercial area tends to have lower sale price
- The sale price also depends on the neighborhood, which may be correlated to many different factors such as the safety index, nearby schools..etc
- Foreclosure and adjoining land purchase are less popular

Modeling – workflow

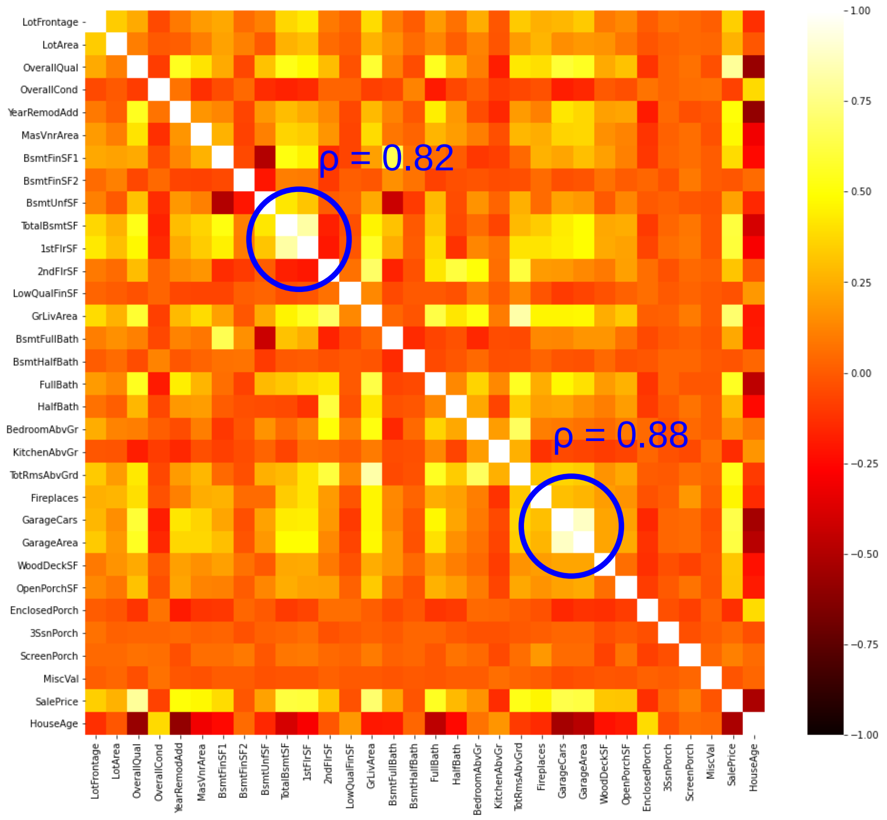
Feature selection

Log-transform/one hot encoding/scaling

Train-test split

Model training, hyperparamter tuning (scikit-learn), prediction

Feature selection – Numerical variables



- Remove 1stFlrSF and GarageArea
- Only keep the features that have > 0.1 correlation with the sale price

Feature selection – Categorical variables

Select categorical features

- Divide the sale price into 5 different categories: ['very low','low','medium','high','very high’]
- Perform Chi-Squared test and only keep the features that have p-values < 0.05

Modeling

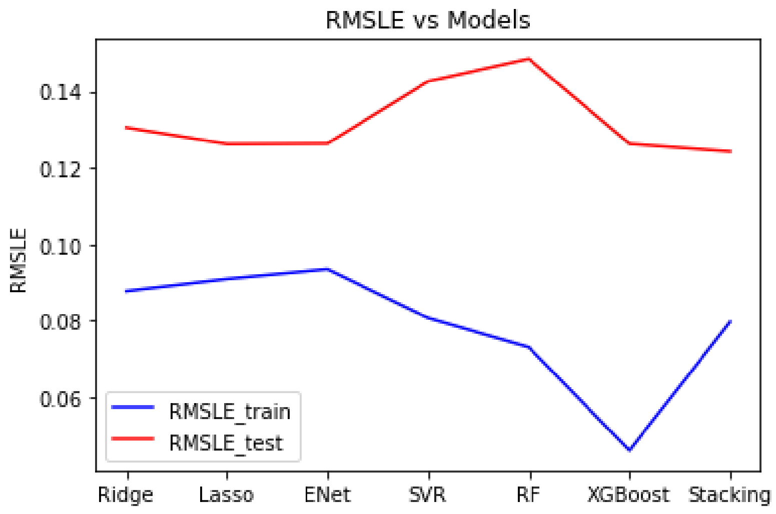
Base models

1. Ridge regression – Linear regression with L2 penalty
2. Lasso regression – Linear regression with L1 penalty
3. Elastic Net regression – combination of L1 and L2 penalties
4. Support Vector regression – Robust to outliers
5. Random Forest – Tree base regression
6. XGBoost – Gradient boosted tree base regression

Stacking model

Model comparisons

```
RMSLE(train) for Ridge: 0.08763077541934929
RMSLE(test) for Ridge: 0.1304877448414943

RMSLE(train) for Lasso: 0.09079906675875786
RMSLE(test) for Lasso: 0.12620327608023368

RMSLE(train) for ENet: 0.09334957829123673
RMSLE(test) for ENet: 0.1263293883413303

RMSLE(train) for SVR: 0.08058579443635039
RMSLE(test) for SVR: 0.1426133603346932

RMSLE(train) for RF: 0.07294982174601274
RMSLE(test) for RF: 0.14850367527465005

RMSLE(train) for XGBoost: 0.04587235763058956
RMSLE(test) for XGBoost: 0.12624786588997225

RMSLE(train) for Stacking: 0.07957217338030946
RMSLE(test) for Stacking: 0.12423559022005141
```

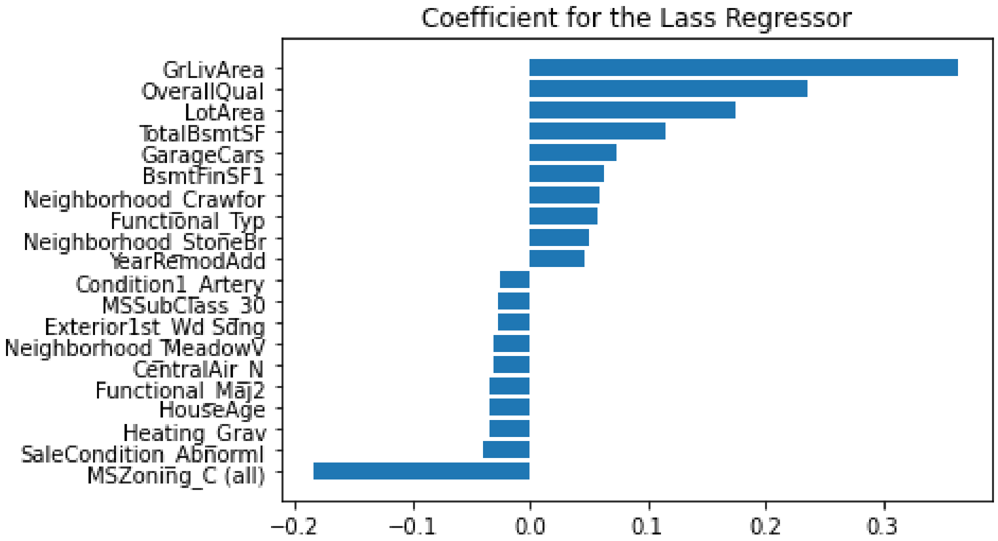


- Use the root mean square loss error as (RMSLE) the figure of merit for our model selection
- Stacking model showed the best performance on the training set with RMSLE ~ 0.124

More ideas to improve the model in the future

- Better outlier detection algorithm besides the simple linear model
- Implement the early-stop hyperparameter to prevent overfitting, especially in the tree-based algorithm
- Further feature selection such as discarding the features that have 0 importance in the Lasso regressor
- Continue updating the latest data into the dataset

Conclusion



- The important features that determine the sale price agree with our general consensus
- The most important features for the sale price is the size of the house with 4 out of top 5 positive contributors related to the size of the house
- The condition of the house is also important to the sale price
- To my surprise, the location of the house doesn't add too much value to the sale price but since our dataset was collected only in Ames city, the importance of the location might be under-estimated