Ames Housing Price Prediction

Wirach Leelakiatowng

Problem Statement

Presently, one of my relatives is gaining interested in a real estate trading business. What she's currently doing is searching for second-hand houses/condominiums that worth renovating for resale. One of the problems she has is the valuation of the property since it is very subjective, and there are plenty of factors affecting its price.

Therefore, this project aims to **study significant factors affecting the property value (which add the most value to a property and which hurt the price most?) and finally build a model that accurately predicts the property price** in a timely manner. So we do not miss an opportunity to get undervalued properties and maximize our profit!

Understanding the dataset

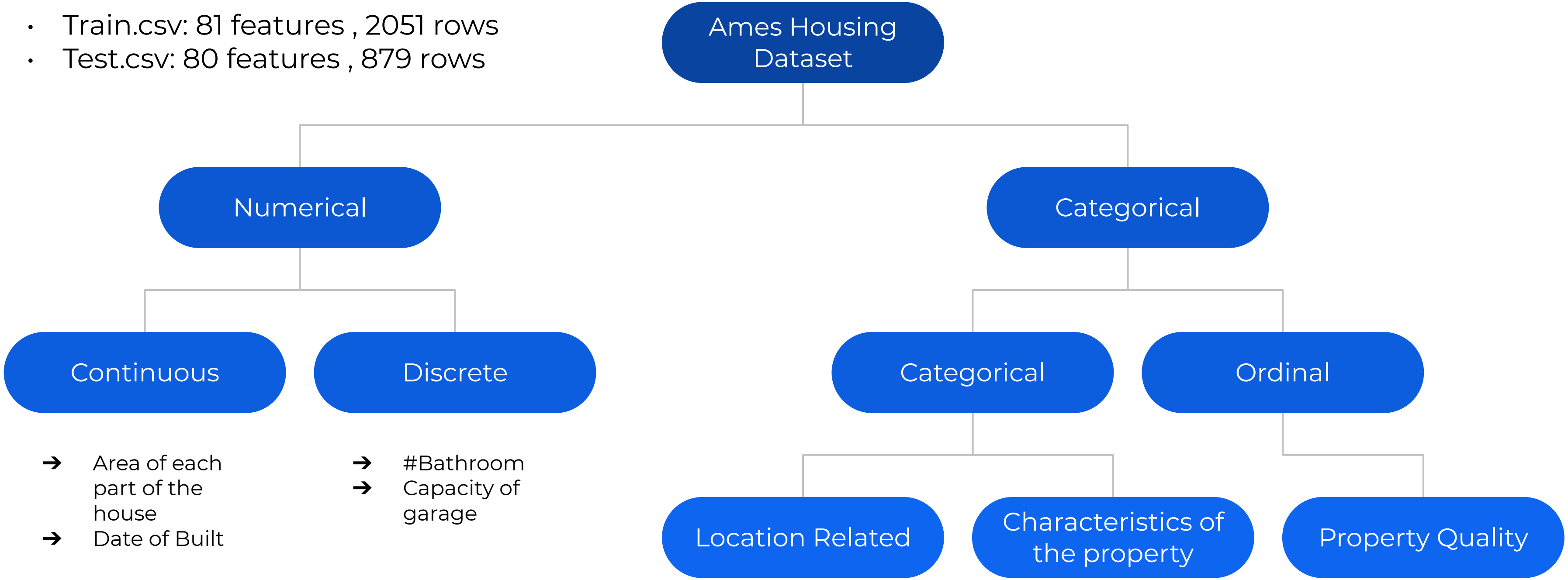
Ames Housing Dataset

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

The dataset contains **2051 observation of 80 features,** describing specific characteristics of residential property in Ames, Iowa sold between 2006 - 2010.

Data Dictionary

- Train.csv: 81 features , 2051 rows
- Test.csv: 80 features , 879 rows



Ames Housing Dataset

Numerical

Categorical

Continuous

Discrete

Categorical

Ordinal

Location Related

Characteristics of the property

Property Quality

Continuous:

- ➔ Area of each part of the house
- ➔ Date of Built

Discrete:

- ➔ #Bathroom
- ➔ Capacity of garage

Location Related:

- ➔ MS Zoning
- ➔ Neighborhood

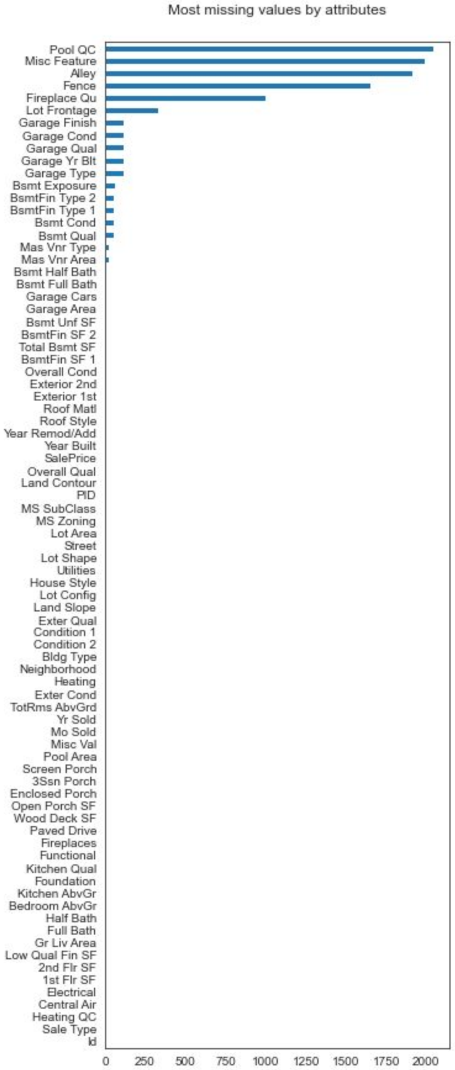
Characteristics of the property:

- ➔ Lot shape
- ➔ Building type
- ➔ Material

Property Quality:

- ➔ Quality/Condition of each part of the house

Data Cleaning and Feature Engineering



01 | Dropping top 5 most missing features

02 | Impute frontage by LinearRegression taking in (1st floor area and lot area)

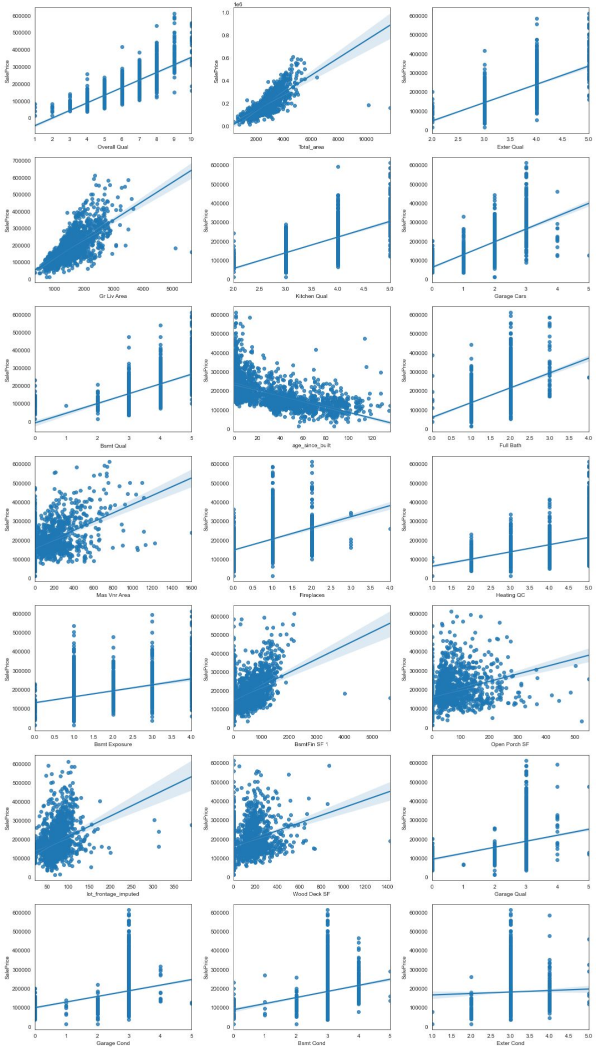
03 | Magna aliqua lorem ipsum dolor sit amet

04 | Impute garage and basement missing quality with ‘NA’

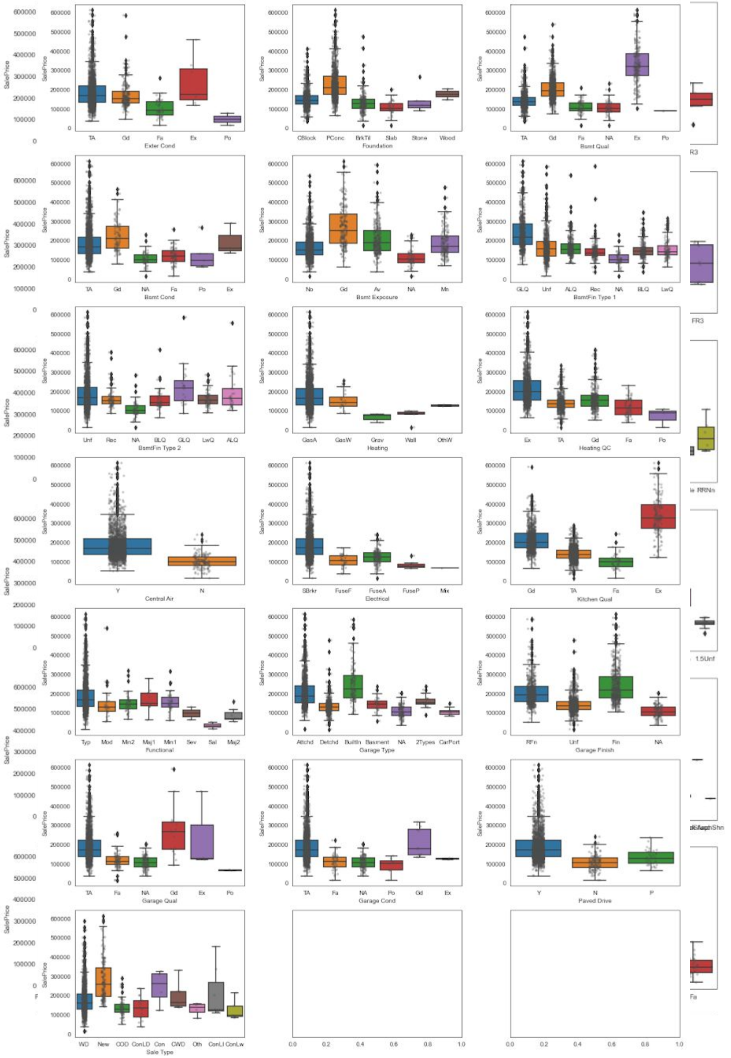
Numerical Features Selection

01 | Remove highly correlated feature (coefficient > 0.7)

02 | Taking the feature with coefficient > 0.3 in consideration



Categorical Feature Selection



01 | Manually selected from the boxplot

02 | cat_col = [

'MS Zoning', 'Street', 'Land Contour', 'Bldg Type', 'Mas Vnr Type',

'Foundation', 'Heating', 'Central Air', 'Electrical', 'Paved Drive', 'Neighborhood'

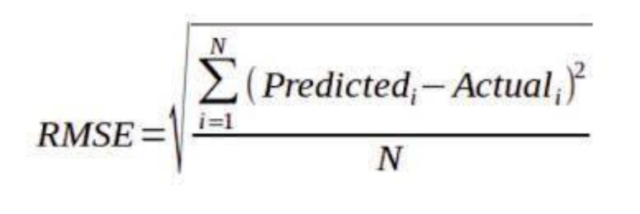
]

Model List :

1. Multiple Linear Regression
2. Regularized Linear Regression in multiple variance
   - Ridge Regression
   - Lasso Regression
   - ElasticNet Regression

Evaluation Metrics :

RMSE
(Square root of Mean Squared Error) will be used as an evaluation metrics



$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}}$$

Baseline Performance :

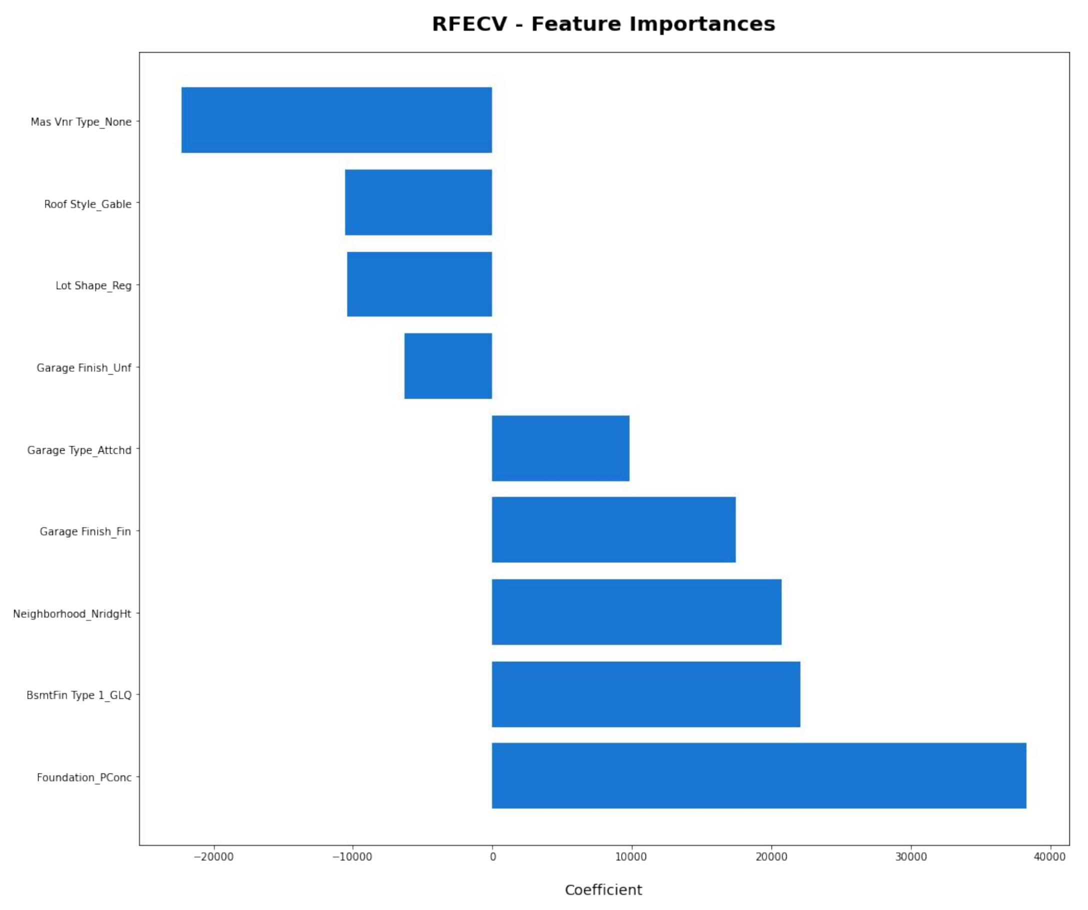
The performance of our model will be compared against the baseline model that make an average of property price as a prediction in every time.

Baseline RMSE = 79,229

Model	RMSE on Training set	Cross val RMSE
Baseline	79229.31	79229.31
Hand-selected features		
Linear Regression	30873.96	1.12 e17
Ridge Regression	31040	32141.44
Lasso Regression	30860	32422.76
ElasticNet Regression	31064	32136.14
Polynomial degree 2 of numerical features		
Linear Regression	18987.03	29876.54
Ridge Regression	19774.57	25865.26
Lasso Regression	19026.74	29783.91
ElasticNet Regression	19881.52	25875.36
Polynomial of numerical features + addition categorical features by RFE		
ElasticNet Regression	21164.30	25616.22

Additional Categorical Features from RFE

- Roof Style
- Lot shape
- Garage Type
- Garage Finish



Model development

Kaggle Score (RMSE):

30311

Model 1 :

X : {top 2o numerical features + Hand-selected categorical features}

- RMSE on training set = 31064
- RMSE on cross-val set = 32136

The model has a sign of underfitting, since the RMSE on training set still high

Kaggle Score (RMSE):

27952

(10% improvement)

Model 2 :

X : {top 20 numerical features with 2 degree polynomial + hand-selected categorical features}

- RMSE on training set = 19881.5
- RMSE on cross-val set = 25875

The bias was significantly reduced. However It's cleary overfitted

Kaggle Score (RMSE):

26915

(3% improvement!)

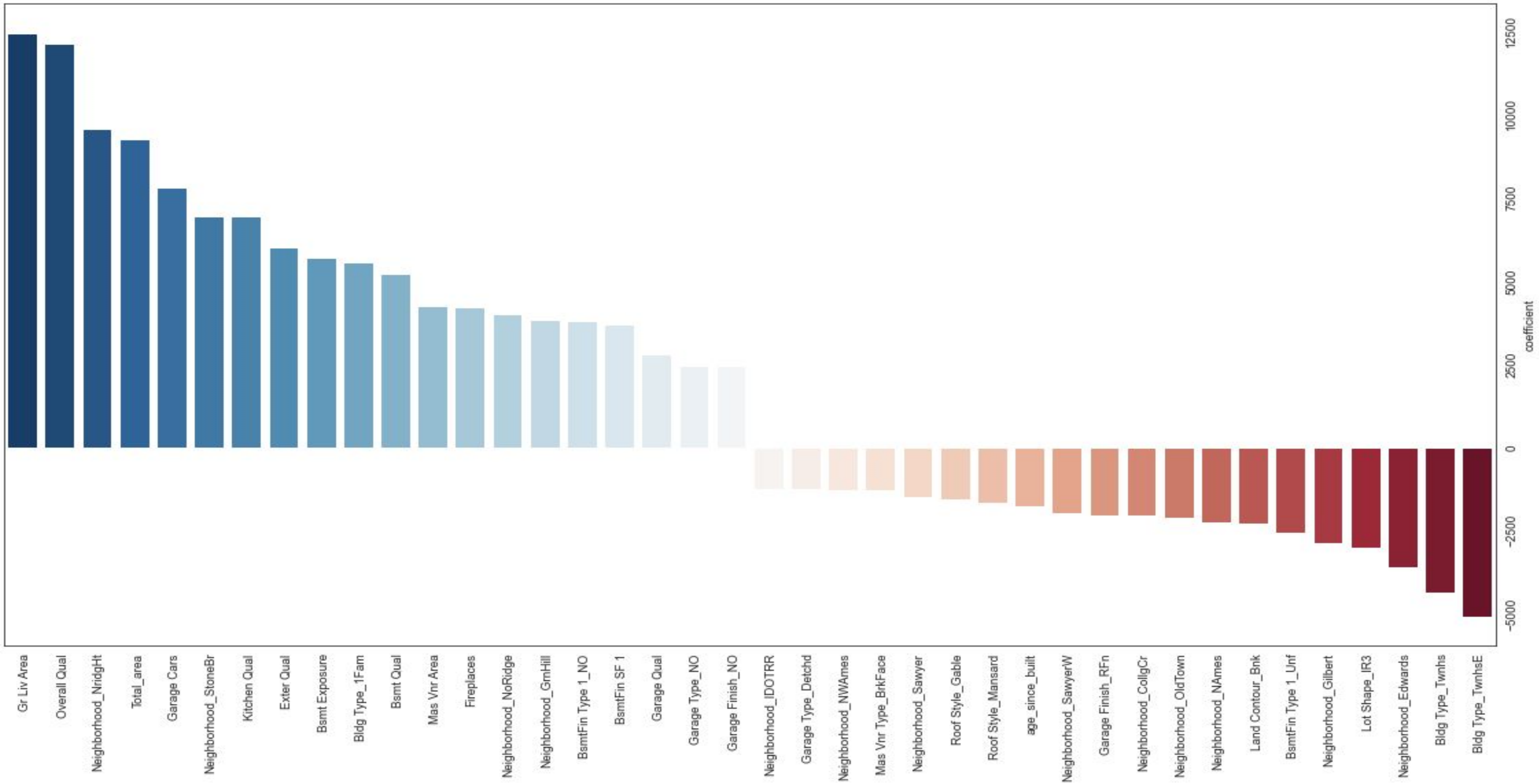
Model 3 :

X : {top 20 numerical features with 2 degree polynomial + hand-selected and RFE categorical features}

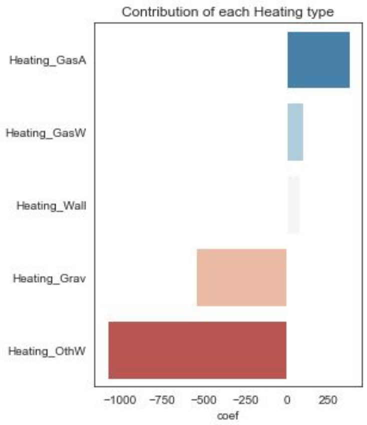
- RMSE on training set = 21164
- RMSE on validation set = 25616

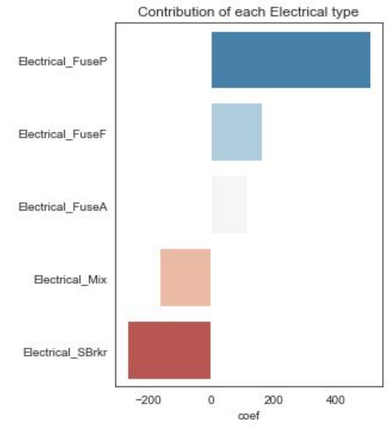
We can marginally reduce the overfitting problem. Nevertheless the model reach plateau performance around testing RMSE = 27,000

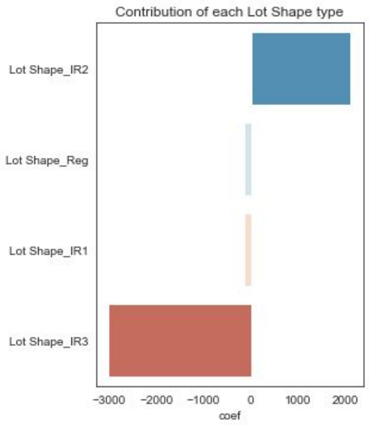
Top 20 most positively & negatively impact on residential property price

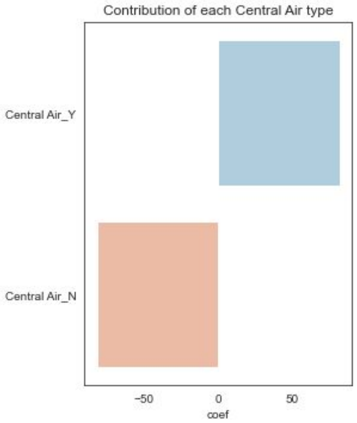


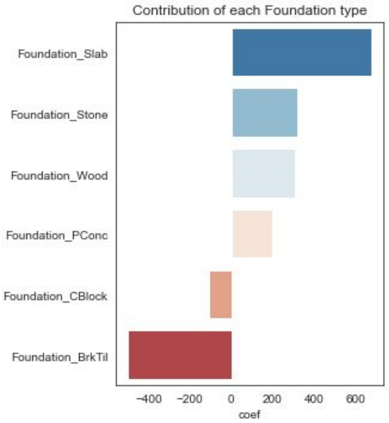
Effect of each Categorical features

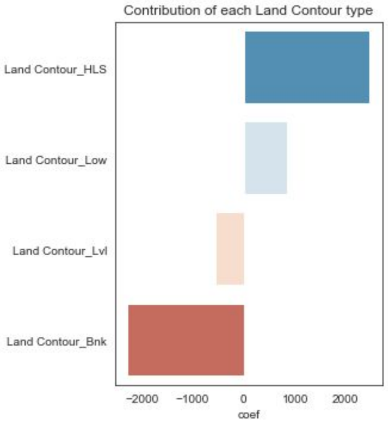

Contribution of each Heating type


Contribution of each Electrical type

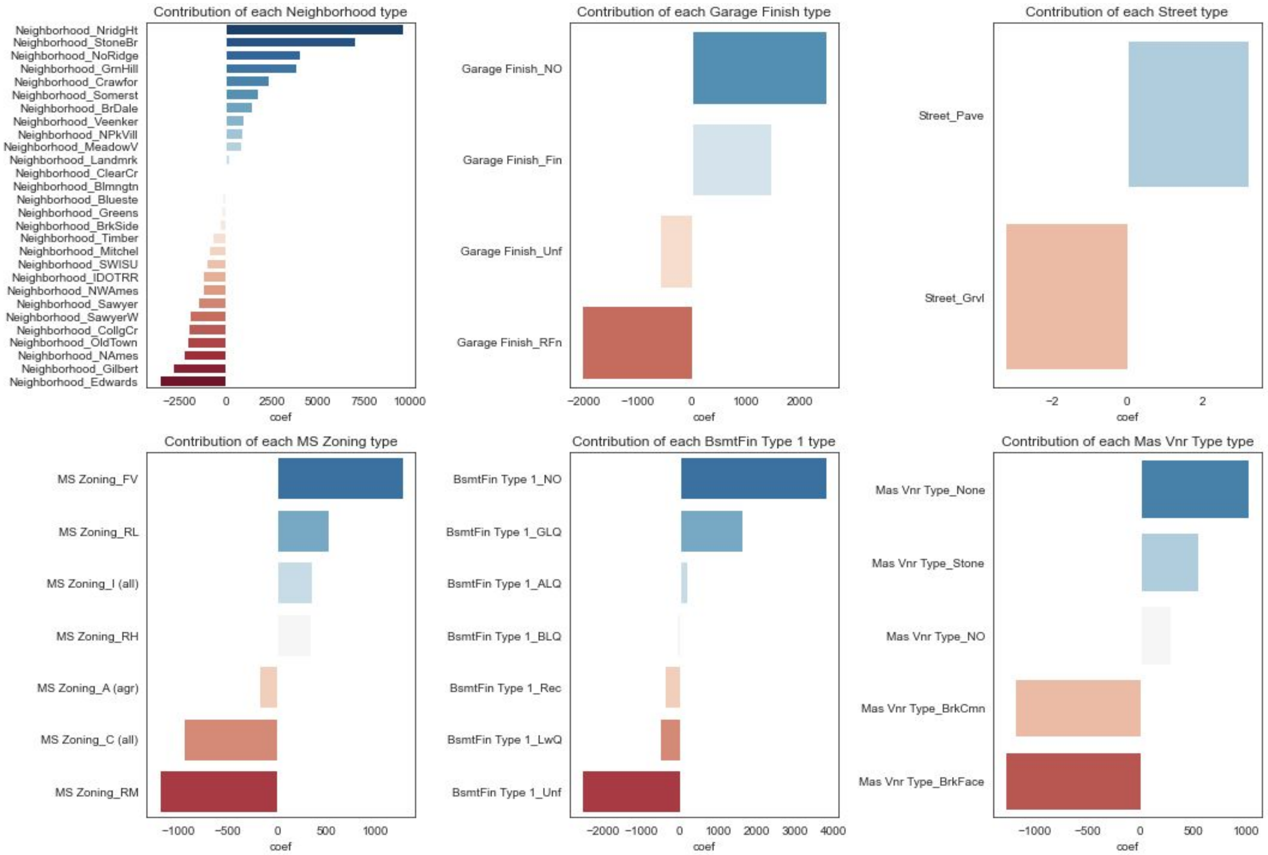

Contribution of each Lot Shape type


Contribution of each Central Air type

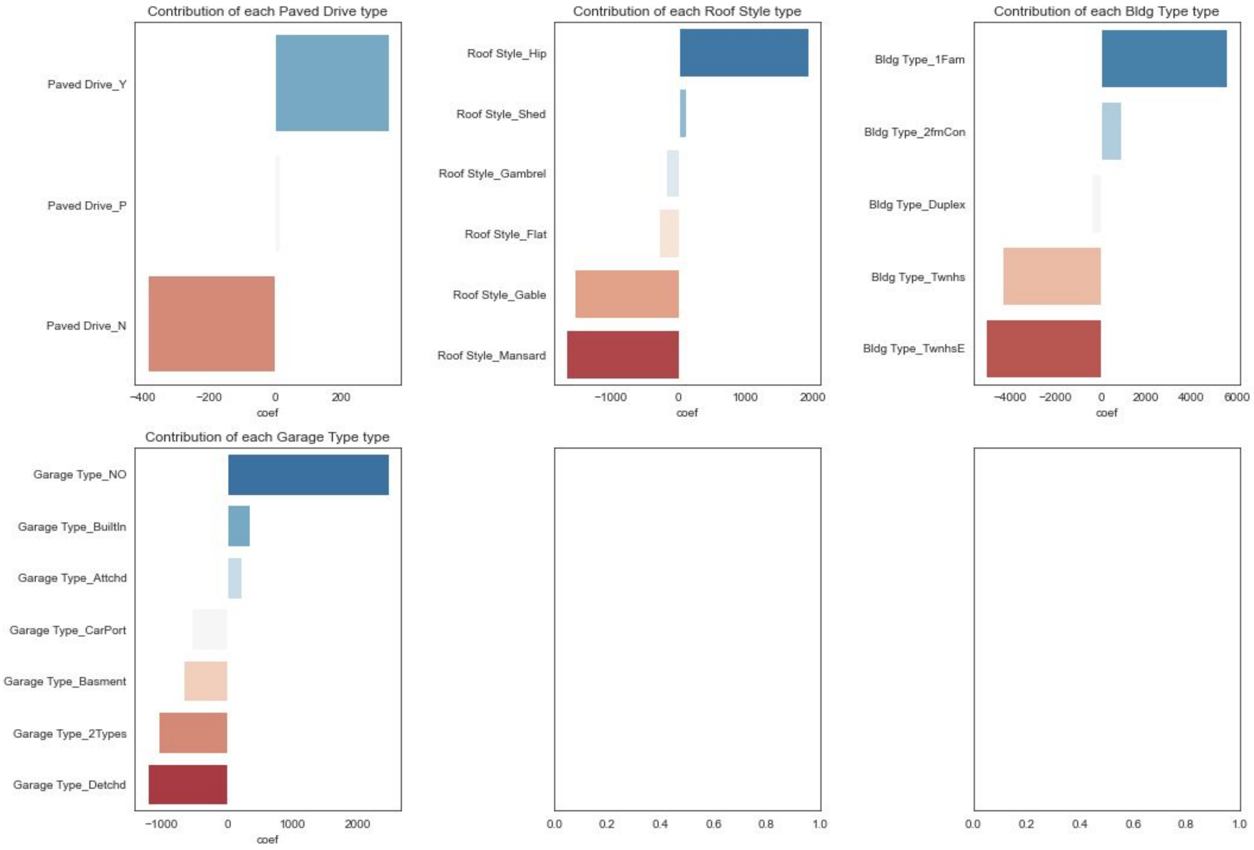

Contribution of each Foundation type


Contribution of each Land Contour type

Effect of each Categorical features(2)



Effect of each Categorical features(3)



Contribution of each Paved Drive type

Contribution of each Roof Style type

Contribution of each Bldg Type type

Contribution of each Garage Type type

Conclusion and Recommendation

Among all models in our study, an ElasticNet regressor with polynomial has the best predictive performance (evaluated by RMSE). Although numerous factors impacted the property value, some of the factors worth mentioning are ...

Area

Although increasing in area directly increasing the price

“The living area is most expensive”

Quality

As our common sense tell us, the better quality, the higher the price

“The Kitchen quality is the key to increase the price”

Garage

If there is no garage, do not waste your time building one.

“It won’t add much value to the house”

Detailed Recommendation

- Although price tend to increase with the total area of the house, living area is the most expensive part of it. → Try renovating the house to maximize the living area before sell it!
- From the perspective of property reseller, while overall quality directly means high resell price, we could consider buying an average quality house and focusing in renovation of important areas to boost its price.
- With a limited budget, focus on kitchen renovation and refurbishing external appearance by repainting.
- If there is no garage, don't waste your time building one. It won't add much value to the house.
- People love wiring inside the tube which is the most easiest approach to do. Don’t over complicated it.
- If you got irregular lot shape, try modifying it by gardening to make it as rectangular as possible

Limitation and Improvement ideas

Limitation

- Our model doing well in property valuation in Ames area. However it might not be very accurate to be applied for data from other city or country.
- As a matter of fact, residential demand and preference vary in every single country due to the variety in cultures and social norm.
- The data is acquired between 2006 - 2010, hence it might not accurately reflect the present price.

Improvement ideas

- taking in to account most updated data
- try including some of important features when we are looking to buy a new house. For example in Thailand,
  - proximity to infrastructure such as hospital, shopping mall, school, tollways, public transportation
  - number of nearby restaurants
  - Was it suffer from a flood?

Thank you.