UNIVERSITEIT VAN AMSTERDAM

New Your City House Price Prediction with Regression Models

Presenter: Will Chien
University of Amsterdam
Fundamentals of Data Science - Assignment 3

1. Background

Estimating the value of a house would be valuable yet challenging for both potential sellers and buyers. While the real estate market can be unreliable and inefficient, this project aims to predict house prices based on the housing features in New York City in 2015 with regression analysis.

2. Methodology

2.1 DATA PROCESSING

Missing Values After importing the dataset, we first examine the dataframe's necessary information and locate the missing values. There is a significant proportion of missing values in columns [“easmnt”, “apt”], 0.99 for the former and 0.75 for the latter. Considering the ineligible amount of missing values, replacing them by mean or another most-frequent-seen feature might not help improve the prediction; in this step, excluding them would be a better choice.

Obvious Artefacts Aside from missing values, we also notice some zero value existing in columns [‘yr_built’, ‘tot_sqft', ‘price’, and ‘yr_built’], which seems bizarre artefacts and are consequently excluded.

Irrelevant Columns Besides, few columns subjectively deemed as less informative, including ["easmnt", "year", "address", "apt", “usable”], are also excluded in the phase.

2.2 DATA EXPLORATION & FEATURE ENGINEERING

2.2.1 Exploring response variable

Outliers Detection When we plot a histogram of house prices, some extreme outliers make the graph noticeable. We utilize the 2.5th and 97.5th quantile [500, 10628750] to discard the outliers to decrease the interference from them.

Skewness Normalization Next, the distribution looks observable, yet an evident right skewness can still be identified (Fig. 1A). Instead of removing all these extreme values, transformations of house prices might make the response value closer to Gaussian distribution without sacrificing too much information [1]. Consequently, both logarithm, square root, boxcox and inverse transformation are investigated and validated with a Shapiro test (base: 0.52; log; 0.91, square root: 0.79; boxcox(lambda = 0.2): 0.92; inverse: 0.04). Since the boxcox transformation yields the greatest statistic and symmetric distribution, we apply this transformed value of price for the following prediction. The distribution plot can be found in Fig 1B.

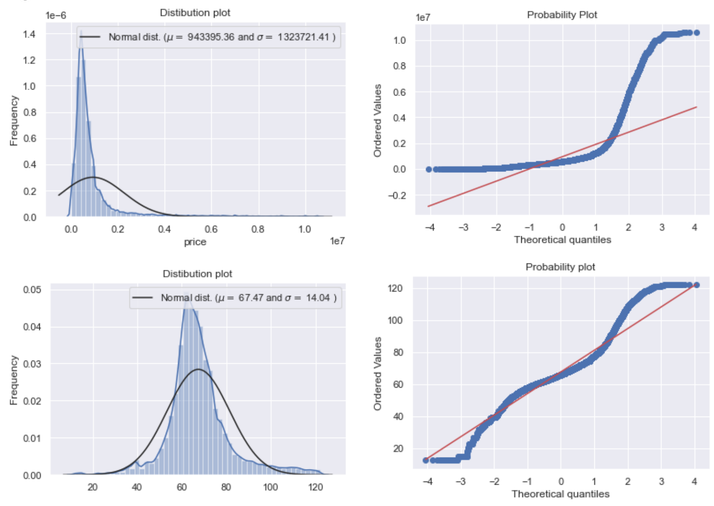

Fig. 1A
Distribution plot and QQ-plot in price before transformation

Fig. 1B
Distribution plot and QQ-plot in price after boxcox transformation

2.2.2. Numerical features

Outliers Detection While visually exploring numerical variables, some outliers are noticeable in specific attributes (Fig. 2). To decrease the noise from outliers, we chose the 97.5th quantile for “land_sqft” and “tot_unit”, which excludes the values beyond it. Furthermore, the 1st quantile for “yr_built” and 95th quantile for “tot_sqft” are applied, respectively.

Correlations The correlations between continuous variables and price are measured and outlined in the heatmap below (Fig 3). Both “tot_sqft” and “tot_unit” appear relatively correlated with the price, while the linear correlation coefficients in “lat” and “yr_built” are not remarkable. Considering the collinearity within “tot_unit” and “res_unit”, when selecting features among these two for a linear regression model, we should only pick the one with a higher correlation.

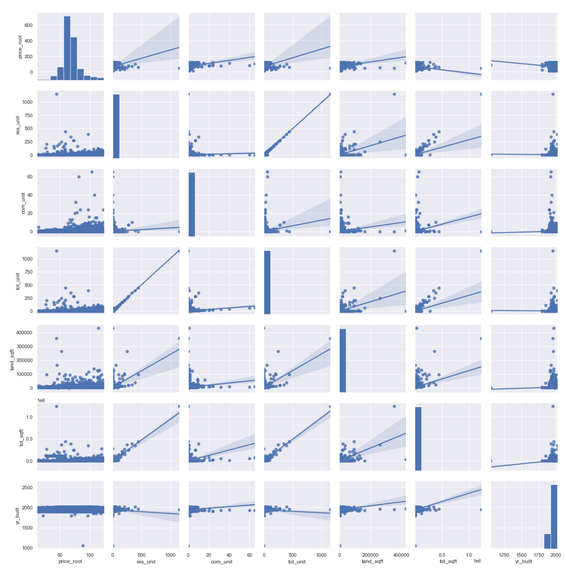

Fig. 3 Pair-plot for numerical

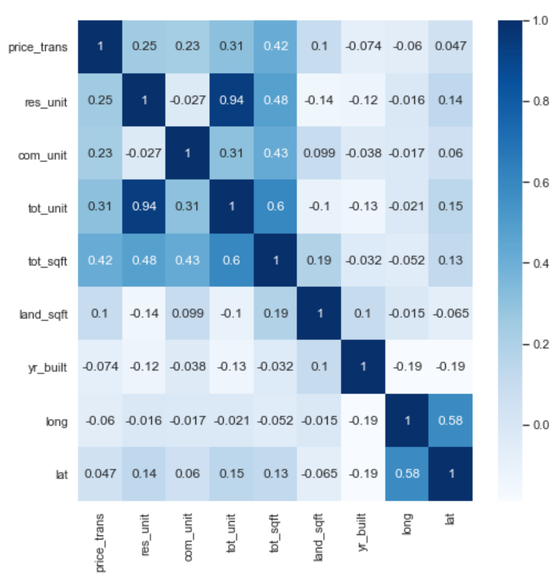

Fig. 4 Correlation heat-map regarding numerical

2. Methodology (cont.)

2.2.3 Categorical features

Correlations When assessing the associations between categorical features and response variable, both box-plot and scatter plot can roughly provide us a fundamental idea of the variance, as shown in Fig. 4. To further measure the correlations between them, we leverage contingency tables with the package `scipy.stats` and the result can be also seen in Fig. 5.

Encoding As categorical features require additional encoding to fit into regression models, several approaches are experimented, including one-hot encoding, label encoding, mean encoding, and group-by encoding. Noticing some remarkable variances of house price in building categories versus boroughs, the max, mean, and median house price is grouped by “borough” and “bldg_ctgy” for group-by encoding.

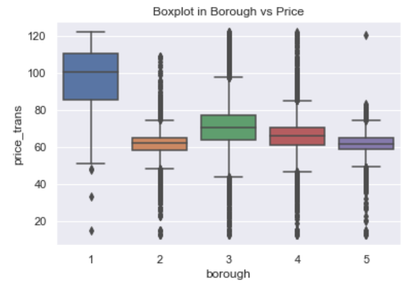

Fig. 4 Box-plot to examine the price

```
Correlations in categorical features:
bldg_ctgy : 0.405
tax_cls_p : 0.397
bldg_cls_p : 0.349
bldg_cls_s : 0.344
sale_date : 0.149
mth : 0.053
borough : 0.389
block : 0.0
lot : 0.0
zip : 0.265
tax_cls_s : 0.373
```

Fig. 5 Correlations in different

2.5 MODEL EVALUATION

2.5.1 Data Splitting

Before training different models, the processed data is split into two parts, the train/validation set and the test set, with a ratio of 80/20. We use the former to iteratively select features and fine-tune models by cross-validation, whereas the latter is held untouched until the model is finalized to prevent overfitting.

2.5.2 Cross Validation

Importing the function cross_val_score() from the sklearn package, we can efficiently perform k-fold cross-validation, which splits the training set into k smaller subsets. Using k-1 subsets to train and 1 subsets to validate, this approach can measure the average model performance in each loop without wasting data [2].

Also, despite its popularity in linear regression, the r-squared statistic not applicable to a nonlinear model. Instead, Mean Absolute Errors (MAE) and Mean Square Errors (MSE) would be more appropriate in terms of comparing accuracy [3]. A linear regression model is applied as a baseline and named “Model 0” in the first attempt. After that, we try to improve the model accuracy in price prediction by choosing different features combinations and algorithms. Finally, aside from Model 0, four additional models from 1 to 4 are developed, showing some improvement in reducing errors. Comparison of test values is summarized in Table 1.

Model		Model 0 (baseline)	Model 1	Model 2	Model 3	Model 4
Algoritm		Linear Regression	Polynomial kernel (n=2)	Support vector regression (kernel = “rbf”)	Gradient Boosting Regression	Gradient Boosting Regression
Selected Features	Numerical	“tot_sqft”, “tot_unit”	“com_unit”, “tot_unit”, “tot_sqft”, “land_sqft”,			
	Categorical	“bldg_ctgy”	"tax_cls_p", "bldg_ctgy", "bldg_cls_p", "borough", "zip", "mth"	“bldg_ctgy”, “tax_cls_p”, “bldg_cls_p”, “bldg_cls_s”, “sale_date”, “borough”, “zip”, “tax_cls_s”		
	Encoding method	One-hot encoding	Label encoding	Label encoding	Mean encoding	Group-by & mean encoding
Perfromace	R squared	0.081	0.136	0.297	0.485	0.606
	MAE	7.378	7.231	6.720	4.972	4.702
	MSE	115.176	120.665	98.250	61.1750	56.504

Table 1 Models comparison

2.6 ADDING FEATURES FROM AERAL PHOTOS

After cross-validation, Model 4 is chosen as the final model candidate. Before applying it to the test set, we are also curious about how the primary features extracted from aerial photos could improve model performance. Therefore, we merge the two dataframes by matching “sales_id” and exclude the rows with missing values. Applying cross-validation, we discern a marginal increase in MAE (4.997) and MSE (61.748) in this new model. Therefore, we could not state visual features are useful for determining the house price and would keep Model 4 as the optimal one.

2.5 RESULTS IN TEST SET

When finally applying Model 4 to the test set, it ends up rising marginally in errors (MAE: 4.810; MSE: 60.276), which indicates a less accuracy in actual house price prediction. Also, while plotting the predicated value against the real ones, we can notice some discrete data points scattering around the regression line (Fig 6), which implies the model’s predictions are improvable. In the residual plot, the pattern with a high-density cluster on the left is not symmetrically distributed.

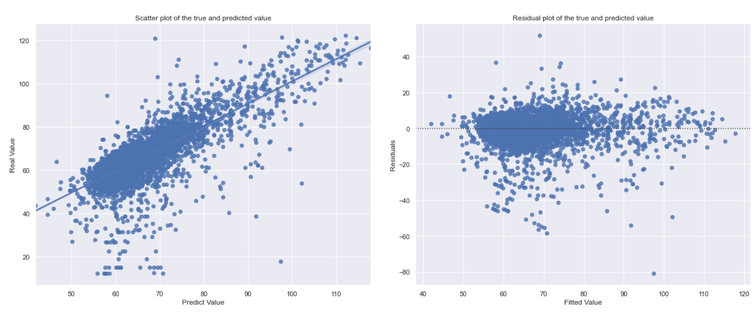

Fig 6. Scatter plot (left) and residual plot (right) of predicted and real values

4. Discussion & Conclusion

From the research, we can see that aggregations of features can improve housing price predictions in New York City with a robust model. More reliable features can be created through numerical-categorical combinations to optimize the current model further and obtain smaller error prediction values. Another tangible approach is introducing external datasets, such as distance to subways, matriculation rate, and demographics, etc. Besides, appropriate methods to examine over-fitting have not been thoroughly discussed and performed in this research. Lastly, to unveil the potential value of images, a higher-level of extracted features are required, such as combining neighborhood-level features from aerial photos or capturing city appearance semantics through street view, which seems tangible and worth the future investigations.

References:
1. Gupta, A.K., Nguyen, T.T. & Sanqui, J.A.T. (2004). *Ann Inst Stat Math* 56, 351–360
2. Raschka, S. (2018). *arXiv preprint arXiv:1811.12808*.
3. Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). Springer Science & Business Media.