Kinga Szarkowska, ID: 13359533

UNIVERSITY OF AMSTERDAM

Fundamentals of Data Science Assignment 3: Property Prices

1. Introduction

Multiple factors can influence the price of a property, starting with different property types, though their size, location, or even tax affiliation. Creating a robust prediction model for a property price is therefore challenging, and was examined using different models in the literature [1][2][3]. Using data provided by the New York City Finance Department I compared different types of machine learning models. Linear Regression, Support Vector Regression, Decision Tree Regression, and Random Forest Regression models were performed using information about a property such as their surface size, number of units, year of built, building category, and tax affiliation. Furthermore, I showed that adding additional features extracted from CNN pre-trained on aerial photos of those properties does not change results drastically.

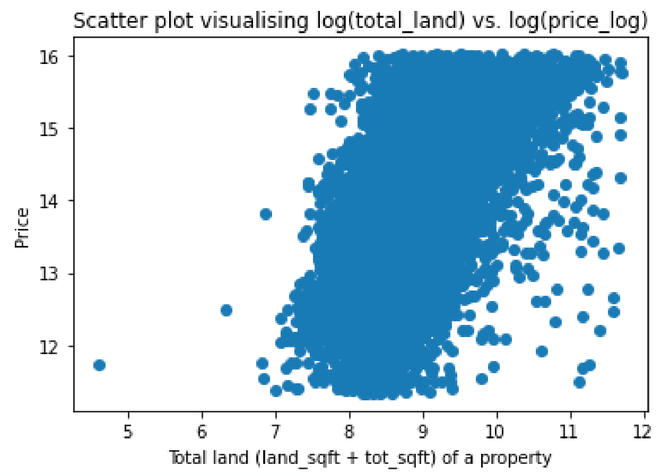
2. Methodology and dataset

METHODOLOGY

- **main goal:** analysis of whether models using additional features extracted from aerial photos are useful for determining property prices or not,
- **step 1:** analysis of four different regression models using dataset without aerial photos features, using **Mean Absolute Error(MAE)** as a baseline of analysis,
- **step 2:** analysis of above-mentioned models with additional features extracted from aerial photos.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

(n: test set; y_i: predicted vaue; $\hat{y}_i$: actual value)



DATASETS

- New York City property prices from 2015 gathered by the City of New York, Department of Finance,
- initially 84,768 records, with 26 variables.

- Features extracted using CNN from aerial photos gathered by the New York City Department of Information Technology & Telecommunications in 2018,
- initially 84,302 records, with 34 variables.

Data preparation:

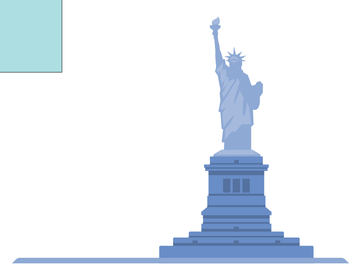
- selection of relevant variables using pair plots and correlation coefficient,
- removing outliers, duplicated values, houses without features from images, and houses that have been sold more than once,
- data transformation (adding new variable total_sqft = tot_sqft + land_sqft, transforming it using log function and transformation price to log - for obtaining a linear shape for total_sqft vs. price),
- columns: borough, blg_ctgy, tax_cls_s - transformed to "dummy" variables as those are categorical variables,
- 24,243 records in the final dataset.

Explanatory variables:

borough - name of the borough that property is located,
bldg_ctgy - building category, initially 28 different categories,
tot_unit - total number of commercial units and residential units within the property,
land_sqft - the land area in squared feet,
tot_sqft - total area of the building in squared feet,
tax_cls_s - one of four tax classes based on the use of a property,
yr_built - the year the structure on the property was built

Response variable:

price - the price paid for the property in $



land_sqft, tot_sqft, tot_unit

filtering values > 0

yr_built

filtering values after 1840

price

first filtering values > 5000$, later filtering values between 3 and 97 percentile

bldg_ctgy

selected five categories: one, two, three family dwellings, apartments and else

MODEL SELECTION

- four different models: Multivariate Linear Regression, Support Vector Regression, Decision Tree Regression, and Random Forest Regression,
- all of them are from scikit learn library in Python,
- all of them were trained using the same set of variables, were compared on the same test set (10% of final dataset) selected before splitting the dataset into training (80% of final dataset) and validation set (10% of final dataset),
- MAE is a baseline selected for comparison, and R^2.

Models characteristics:

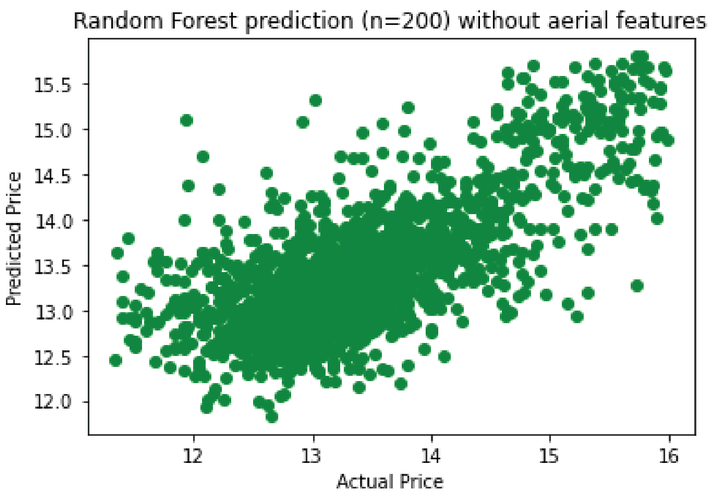
- **Multivariate Linear Regression** - the simplest form of regression, with transforming price and total_sqft to log values we might observe linear dependency,
- **Support Vector Regression** - by picking the right kernel function SVR[2] can considers the presence of non-linearity (I selected radial basis function), it has also a better tolerance than linear regression to outliers[7],
- **Decision Trees Regression** - gives good outcomes where data is mostly categorical[3],
- **Random Forest Regression** - capable of creating a nonlinear mapping between an input and an output vector[8].

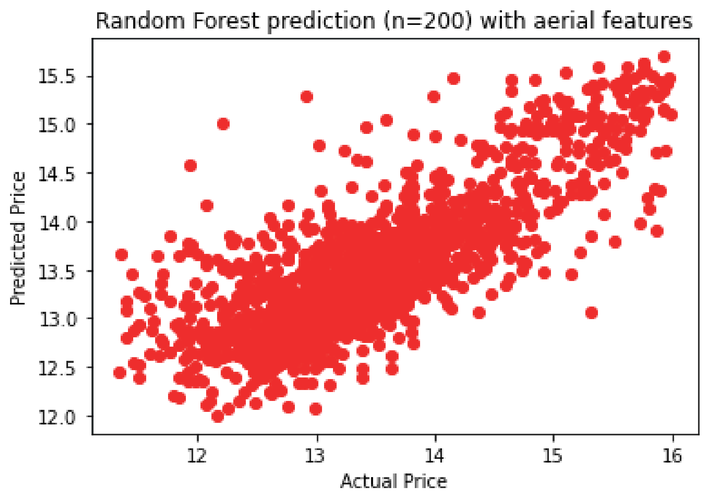
All of them have been successfully implemented so far in the property price prediction problem [4][5][6].

3. Results

	Without Features			With Aerial Features		
Regression Model	Hyperparameters	R^2	MAE	Hyperparameters	R^2	MAE
Linear Regression		0,53	$ 400 023,90		0,53	$ 400 806,90
Support Vector Machines	RBF, with c: 1000	0,36	$ 1 634 135,20	RBF, with c: 1000	0,42	$ 571 492,30
Decision Trees	max_depth: 400	0,22	$ 489 843,20	max_depth: 400	0,26	$ 455 774,70
	max_depth: 300	0,22	$ 486 805,20	max_depth: 300	0,24	$ 452 898,20
	max_depth: 250	0,21	$ 488 740,40	max_depth: 250	0,21	$ 459 113,60
	max_depth: 100	0,20	$ 492 995,90	max_depth: 100	0,22	$ 465 735,60
	max_depth: 200	0,20	$ 491 498,80	max_depth: 200	0,26	$ 460 054,90
	max_depth: 20	0,30	$ 467 806,50	max_depth: 20	0,33	$ 432 258,80
Random Forest	estimator no: 400	0,51	$ 373 956,30	estimator no: 400	0,64	$ 329 976,30
	estimator no: 300	0,52	$ 373 820,40	estimator no: 300	0,64	$ 330 888,20
	estimator no: 200	0,52	$ 373 351,30	estimator no: 200	0,64	$ 329 236,20
	estimator no: 250	0,51	$ 373 727,80	estimator no: 250	0,64	$ 330 034,40
	estimator no: 100	0,52	$ 373 787,30	estimator no: 100	0,64	$ 329 912,90
	estimator no: 20	0,51	$ 377 419,30	estimator no: 20	0,62	$ 330 907,10

Table 1: Summary of results using MAE and R^2 as a baseline.





4. Analysis and discussion

- using MAE as a baseline we see that the best model in both cases is the Random Forest Regression model with a number of estimators set to 200, noting that changing this parameter does not change the results drastically,
- for all of the models (except from Linear Regression) we notice an improvement in MAE and R^2 after adding aerial features, yet only in SVR we notice a substantial difference in MAE (-65%, in other cases difference between 5-12%),
- limited improvement might be because the large number of extracted features (32, which doubles the number of variables) and it does add more noise than clarity to the data.

5. Limitations and next steps

- only one set of variables was checked, hence accuracy might be improved by checking different sets of variables, especially testing a lower number of them,
- assumptions for bldg_ctgy should be checked with property experts, correlation between tax affiliation and bldg_ctgy, and impact of sale date.