Machine Learning 2 Project

Sergio Cárcamo Jara (466116)

Overview

- Data Description
- Data Specificity
- Training/Test Data Division
- Comparison of Methods
- Summary and Conclusions
- Questions

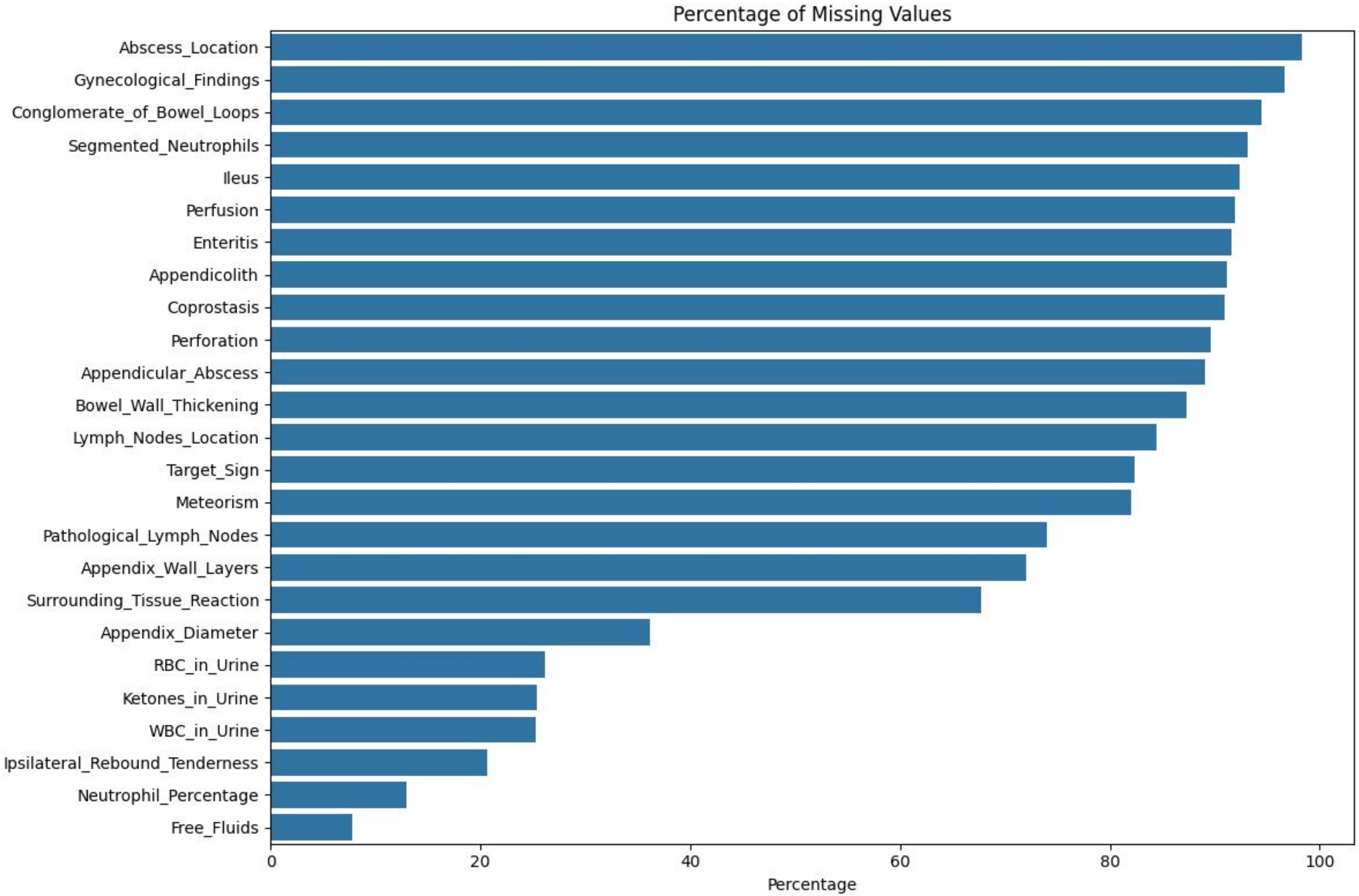
Pediatric Appendicitis

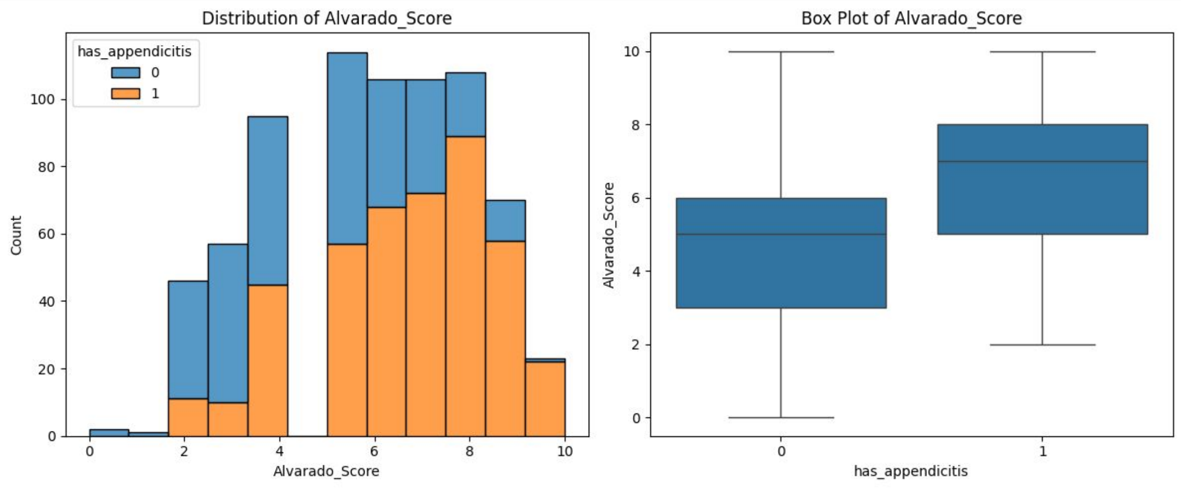
Classification

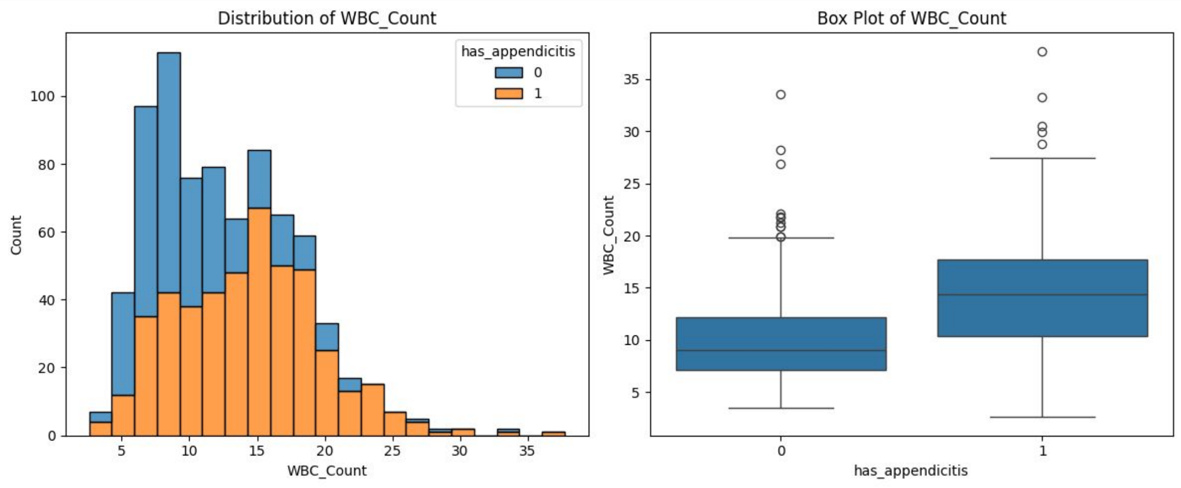
Data description

Data corresponds to pediatric patients with suspected appendicitis admitted with abdominal pain to Children's Hospital St. Hedwig in Regensburg, Germany, between 2016 and 2021.

The dataset includes ultrasound images, laboratory, physical examination, scoring results and ultrasonographic findings extracted manually by the experts, and the target variable is diagnosis, to predict if the patient has appendicitis or not.

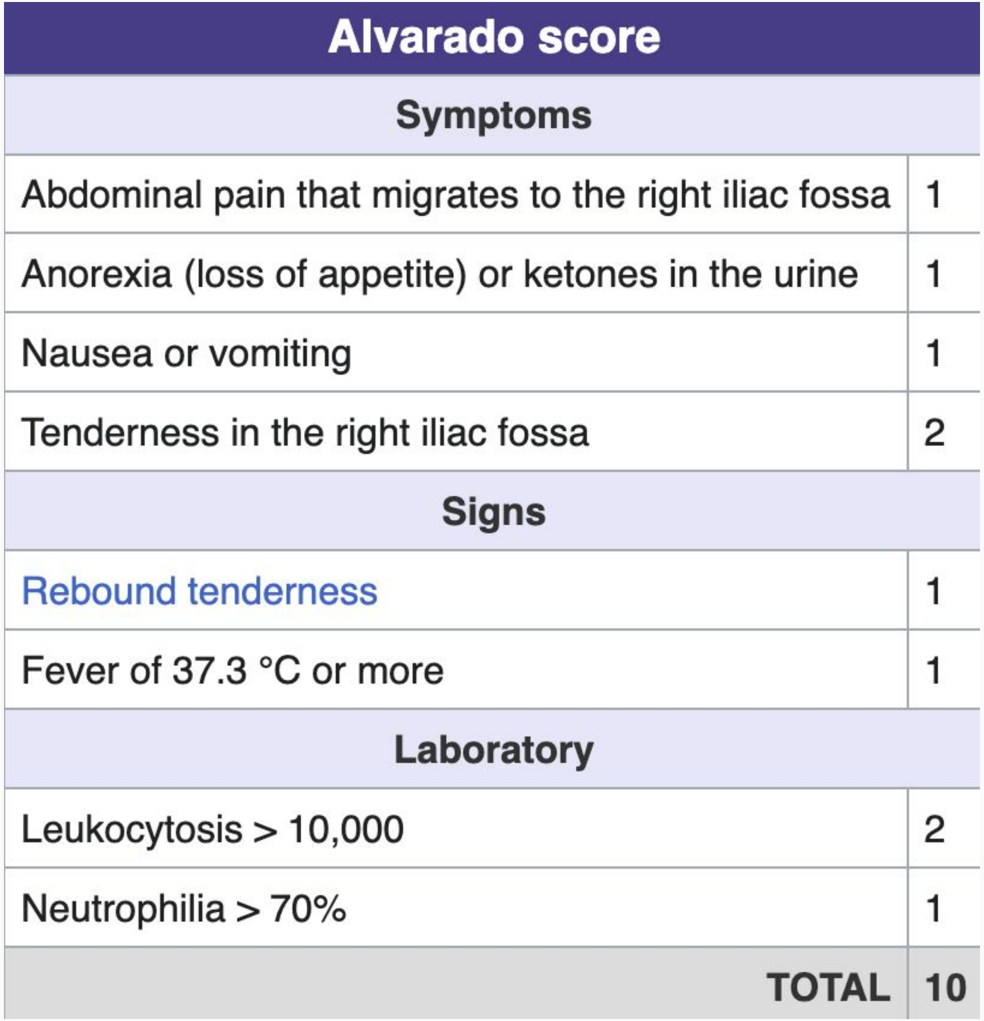

Percentage of Missing Values
Abscess_Location
Gynecological_Findings
Conglomerate_of_Bowel_Loops
Segmented_Neutrophils
Ileus
Perfusion
Enteritis
Appendicolith
Coprostasis
Perforation
Appendicular_Abscess
Bowel_Wall_Thickening
Lymph_Nodes_Location
Target_Sign
Meteorism
Pathological_Lymph_Nodes
Appendix_Wall_Layers
Surrounding_Tissue_Reaction
Appendix_Diameter
RBC_in_Urine
Ketones_in_Urine
WBC_in_Urine
Ipsilateral_Rebound_Tenderness
Neutrophil_Percentage
Free_Fluids
0
20
40
60
80
100
Percentage





The Alvarado score is a clinical scoring system used in the diagnosis of appendicitis. Based on symptoms and blood test results.

WBC: White blood cell count



Alvarado score	
Symptoms	
Abdominal pain that migrates to the right iliac fossa	1
Anorexia (loss of appetite) or ketones in the urine	1
Nausea or vomiting	1
Tenderness in the right iliac fossa	2
Signs	
Rebound tenderness	1
Fever of 37.3 °C or more	1
Laboratory	
Leukocytosis > 10,000	2
Neutrophilia > 70%	1
TOTAL	**10**

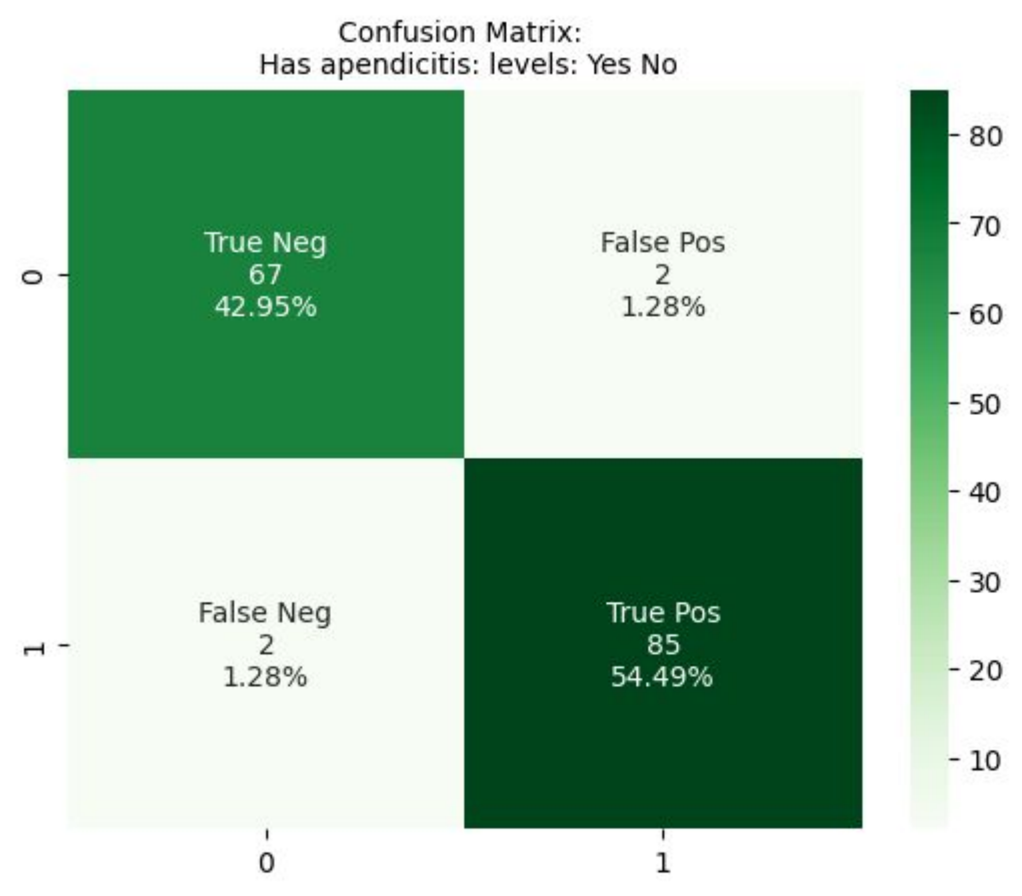
Training/Test Data Division

- Training/Test data split 80/20
- Randomized Search CV
  - parameters set for each model
  - 25 iterations
  - scoring F1
  - cv 5

Comparison of Methods

model	RandomForestClassifier	AdaBoostClassifier	XGBClassifier
accuracy	0.929487	0.974359	0.948718
precision	0.94186	0.977011	0.964706
recall	0.931034	0.977011	0.942529
F1	0.936416	0.977011	0.953488

AdaBoost Classifier



Summary and Conclusions

- Chosen model is AdaBoost Classifier (with Decision Tree Classifier as base estimator)
- Find variables that can be removed without sacrificing performance (such as ultrasound images)
- Improve imputation methods

- Best hyper parameters
  - algorithm: SAMME
  - learning_rate: 0.63435404
  - n_estimators: 111
  - imputer: mean
  - estimator criterion: entropy
  - estimator max_depth: 7
  - estimator min samples leaf: 3
  - estimator min samples split: 3

Real estate prices

Regression

Data description

For the regression problem, the data was scraped from otodom from september to december 2024. The scope is properties for sale in the Warsaw area, including flats, excludes houses.

The dataset includes data on the property itself, and the building if the property is a flat. The target variable is the price.

Data description

Property characteristics:

- estate: FLAT, HOUSE
- **area_m2**
- rooms_number
- floor_number
- windows_type
- heating
- **price**

Location

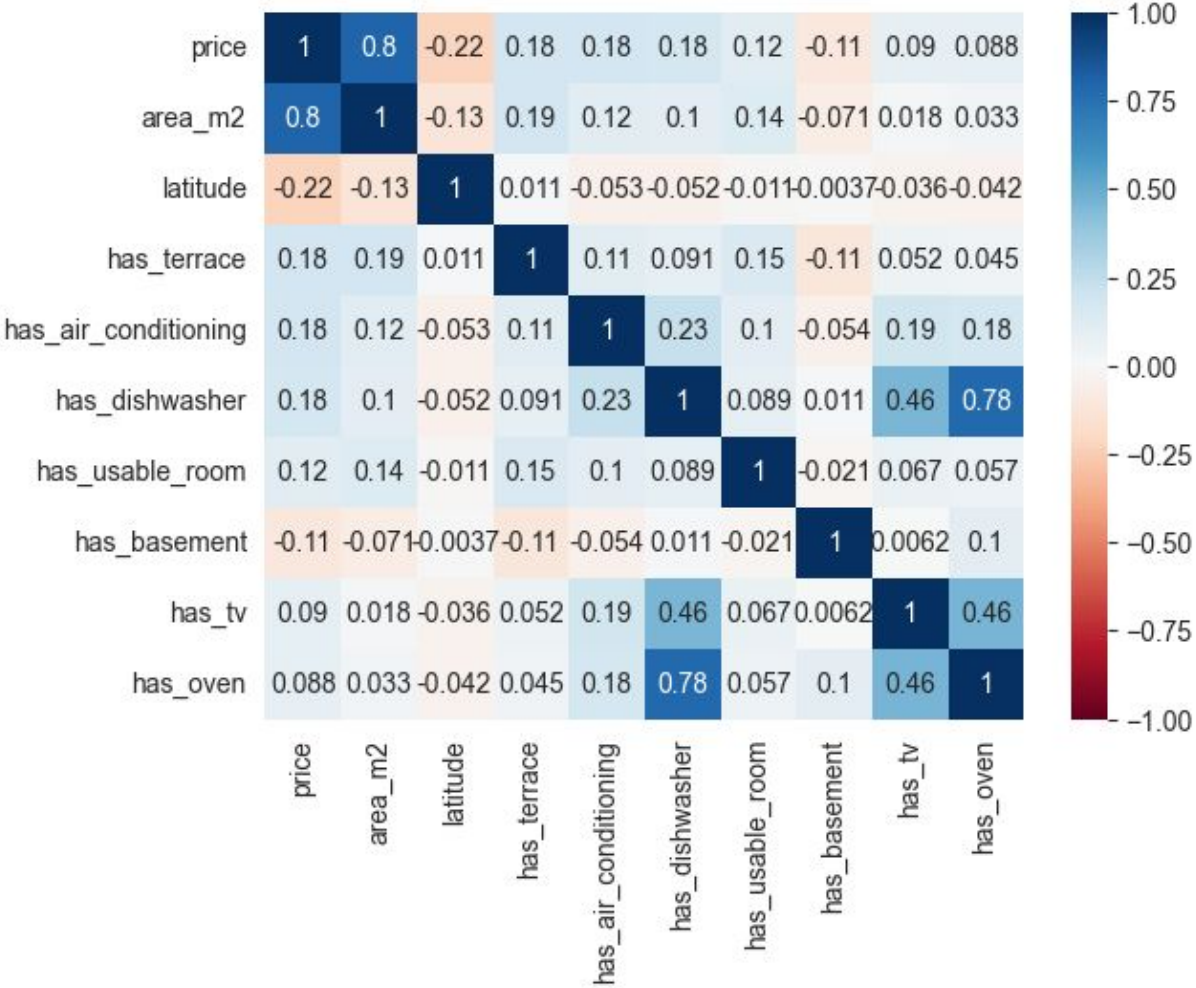
- **district**
- latitude
- longitude

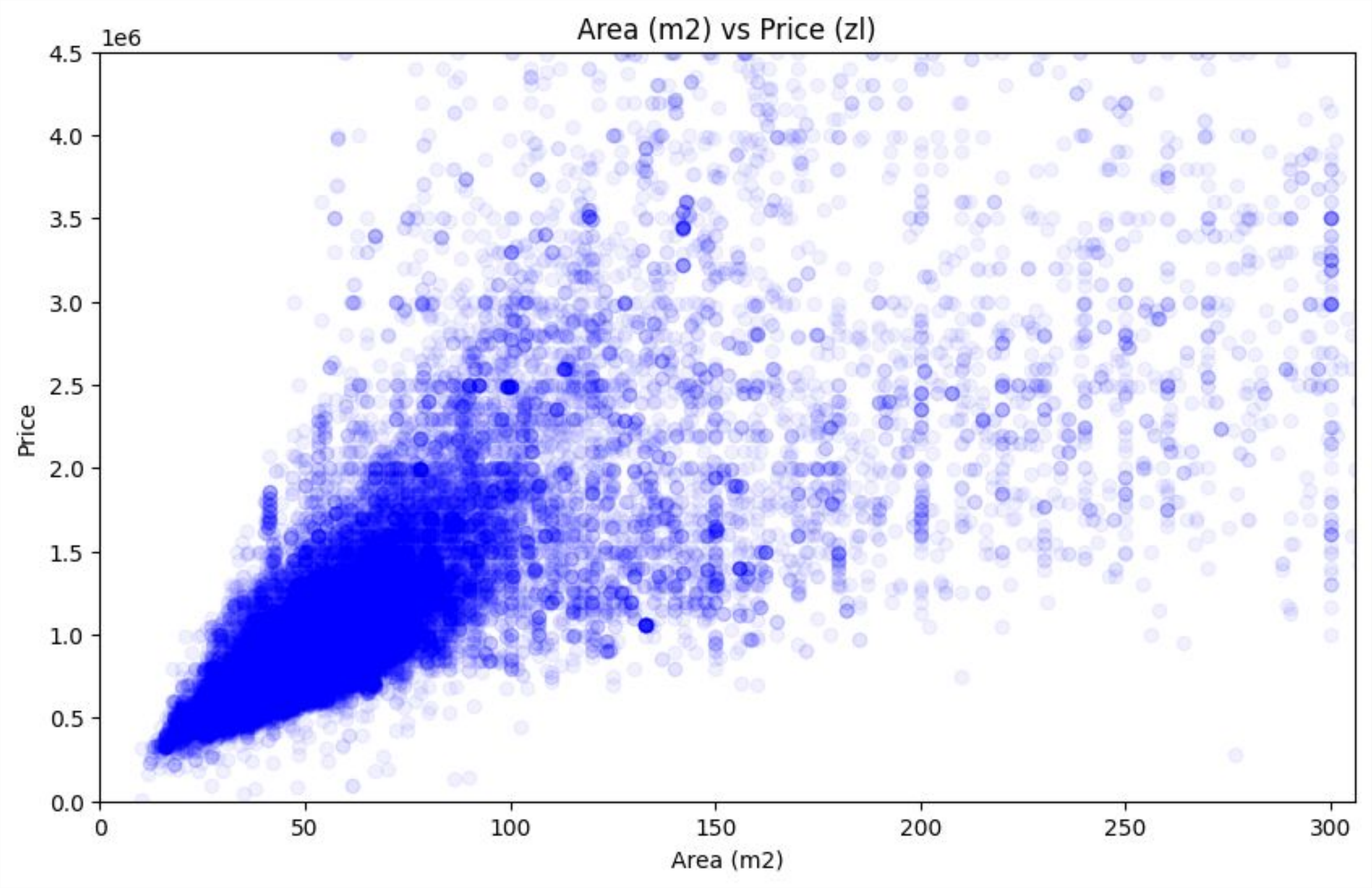
Building characteristics:

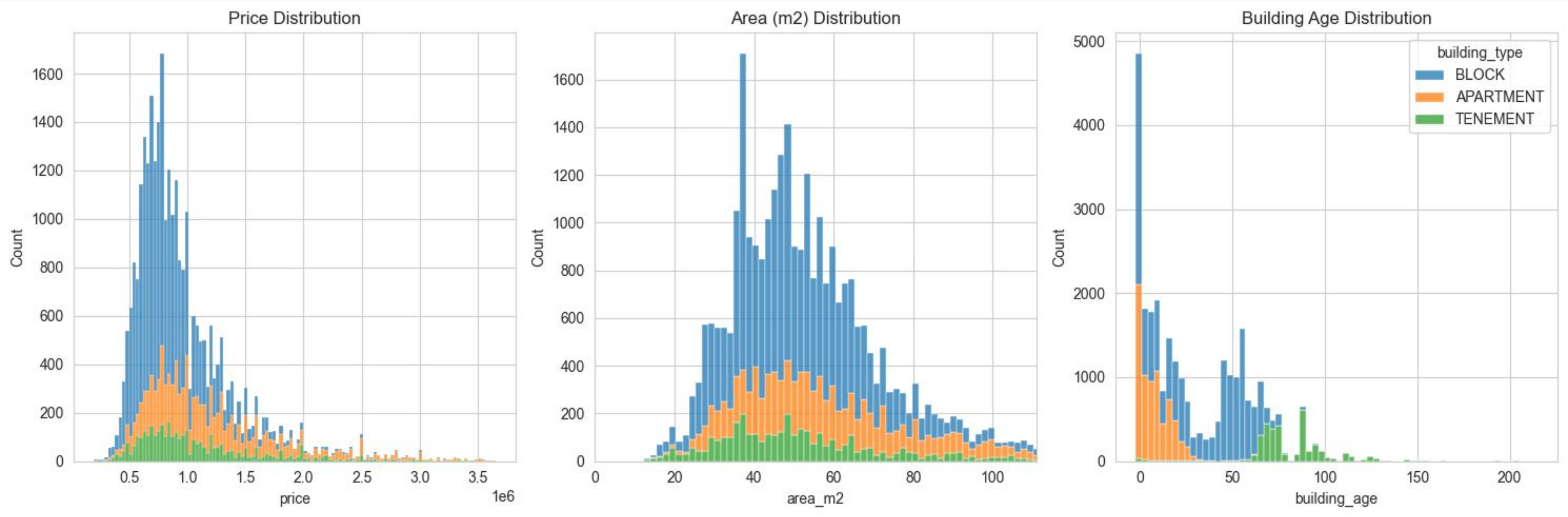
- building_year
- **building_age**
- building_type
- building_floors_num
- construction_status
- building_material
- building_ownership

Data description: amenities

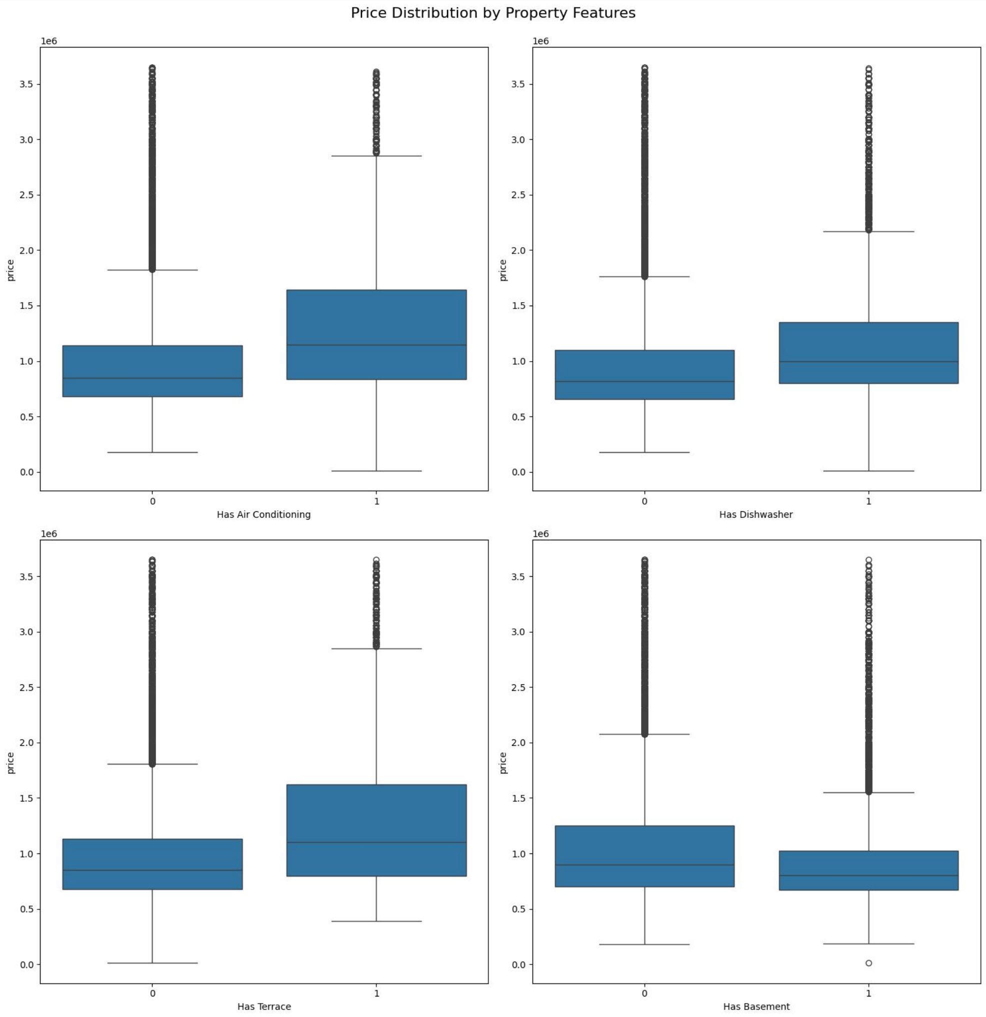
- has_lift
- has_internet
- has_furniture
- has_air_conditioning
- has_tv
- has_oven
- has_stove
- has_dishwasher
- has_fridge
- has_washing_machine
- has_separate_kitchen
- has_garage
- has_usable_room
- has_terrace
- has_balcony
- has_attic
- has_basement
- has_garden
- has_pool


price 1 0.8 -0.22 0.18 0.18 0.18 0.12 -0.11 0.09 0.088
area_m2 0.8 1 -0.13 0.19 0.12 0.1 0.14 -0.071 0.018 0.033
latitude -0.22 -0.13 1 0.011 -0.053 -0.052 -0.011 -0.0037 -0.036 -0.042
has_terrace 0.18 0.19 0.011 1 0.11 0.091 0.15 -0.11 0.052 0.045
has_air_conditioning 0.18 0.12 -0.053 0.11 1 0.23 0.1 -0.054 0.19 0.18
has_dishwasher 0.18 0.1 -0.052 0.091 0.23 1 0.089 0.011 0.46 0.78
has_usable_room 0.12 0.14 -0.011 0.15 0.1 0.089 1 -0.021 0.067 0.057
has_basement -0.11 -0.071 -0.0037 -0.11 -0.054 0.011 -0.021 1 0.0062 0.1
has_tv 0.09 0.018 -0.036 0.052 0.19 0.46 0.067 0.0062 1 0.46
has_oven 0.088 0.033 -0.042 0.045 0.18 0.78 0.057 0.1 0.46 1
price
area_m2
latitude
has_terrace
has_air_conditioning
has_dishwasher
has_usable_room
has_basement
has_tv
has_oven
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Area (m2) vs Price (zl)
1e6
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Price
0
50
100
150
200
250
300
Area (m2)

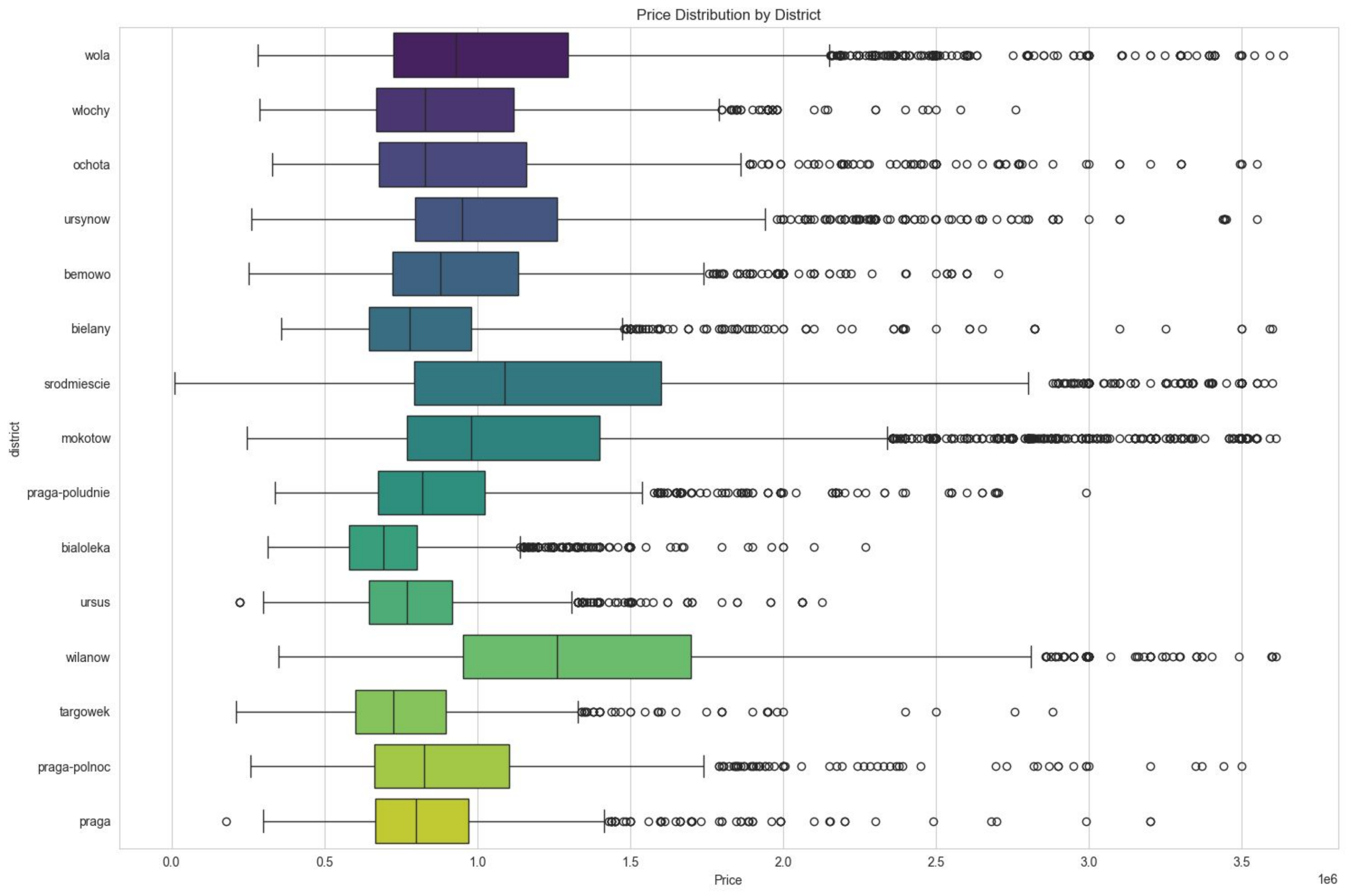


* tenement: kamienica



Highlights

- Air conditioning: correlated with newer and furnished flats
- Dishwasher: correlated with furnished properties
- Terrace: correlated with larger and newer properties
- Basement: correlated with older properties


Price Distribution by District
district
wola
wlochy
ochota
ursynow
bemowo
bielany
srodmiescie
mokotow
praga-poludnie
bialoleka
ursus
wilanow
targowek
praga-polnoc
praga
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Price
1e6

Training/Test Data Division

For splitting the data, I used a 80/20 split.

Top 30 features selected with highest correlation.

Randomized Search CV used for hyperparameter tuning.

- 50 iterations
- metric neg_root_mean_squared_error
- cv 5

Comparison of Methods

model	Decision Tree Regressor	XGBoost Regressor	Random Forest Regressor	Ensemble Learning*
RMSE	209411	120715	147654	131948
R2	0.80	0.932	0.90	0.92
MAPE	0.14	0.068	0.09	0.079

*Ensemble Learning: XGBoost, Random Forest Regressor, GradientBoostingRegressor

Summary and Conclusions

- Selected model is XGBoost Regressor due to lower RMSE and MAPE
- Improve feature selection methods
- Improve imputing methods
- Adjust ensemble pipeline
- Comparison of performance between training and test seems to indicate overfitting

Thank you!