01

ERICK GAITÁN | 18.04.20

House Price Prediction in Bangalore, India

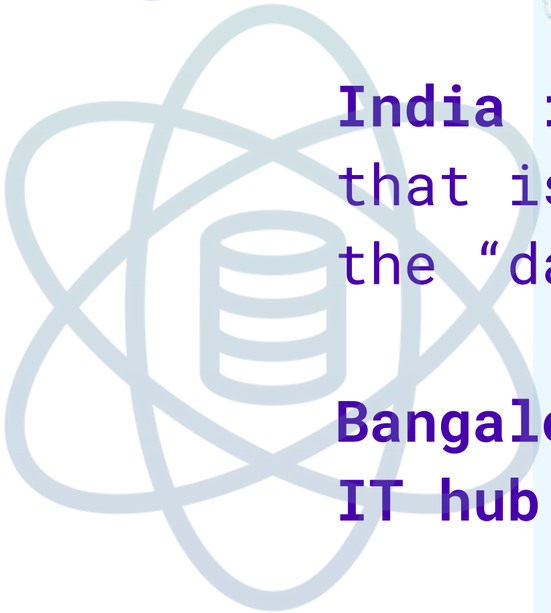
02

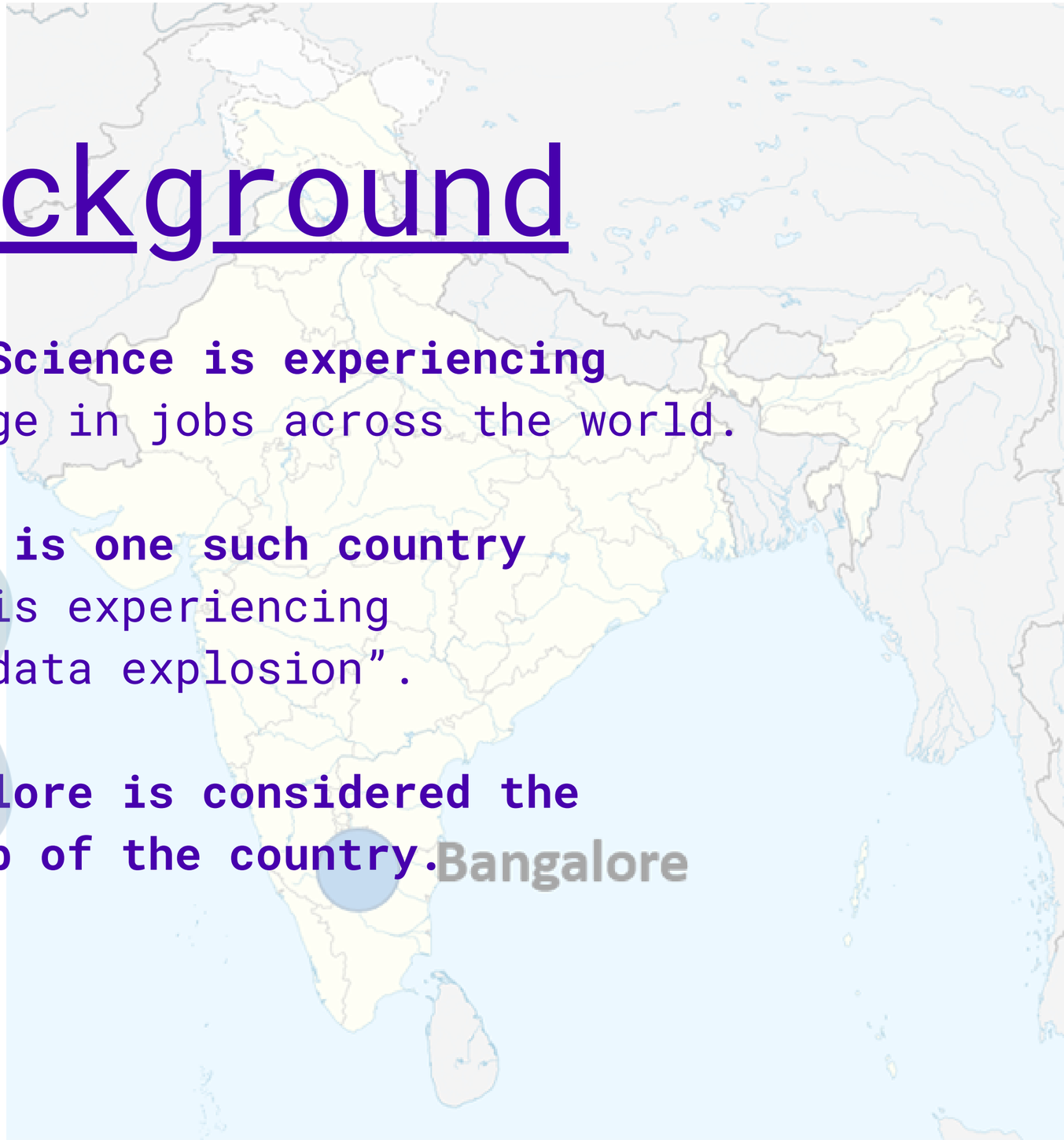
Background

Data Science is experiencing
a surge in jobs across the world.

India is one such country
that is experiencing
the “data explosion”.

Bangalore is considered the
IT hub of the country.





97,000

Data Science and Analytics job vacants in India on February 2019

Top Data Science companies

in Bangalore are hungry for Data Scientist

04

Top Companies hungry for Data Scientists in India

- LinkedIn
- Amazon
- Deloitte
- IBM
- Accenture
- Salesforce

ERICK GAITÁN | 18.04.20

05

ERICK GAITÁN | 18.04.20

The main purpose of this project is to give information about house prices in Bangalore, India

to any Data Scientist or aspirant to find interesting to live in this country. House prices are predicted by using different regression models.

Data

ERICK GAITÁN | 18.04.20

Source

CSV file obtained from Kaggle, with nine columns and 13,320 rows of data.

Cleaning

By eliminating incomplete observations, there are 7,396 rows remaining.

Encoders

Categorical variables are encoded using OneHotEncoder method.

Features

bath, area_type, balcony, location, size, society, total_sqft

Target

price

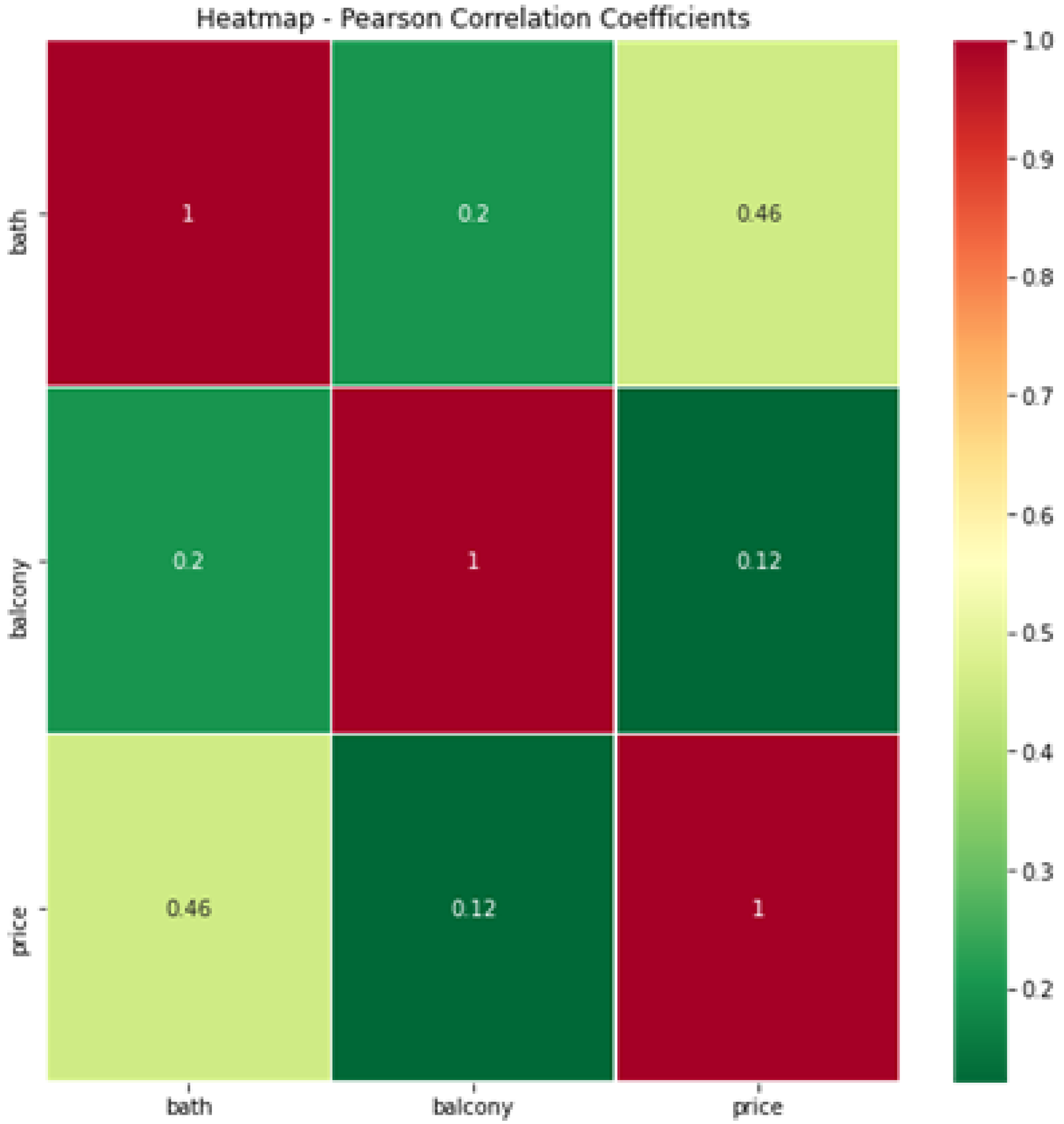
Source link here

07

PEARSON CORRELATION COEFFICIENTS

Since we are using regression algorithms, it is important to see if there is any high correlation between variables.

ERICK GAITÁN | 18.04.20



08

Regression models

- Random Forest
- XGBoost
- Gradient Booster
- Lasso Regression
- Ridge Regression

ERICK GAITÁN | 18.04.20

ERICK GAITÁN | 18.04.20

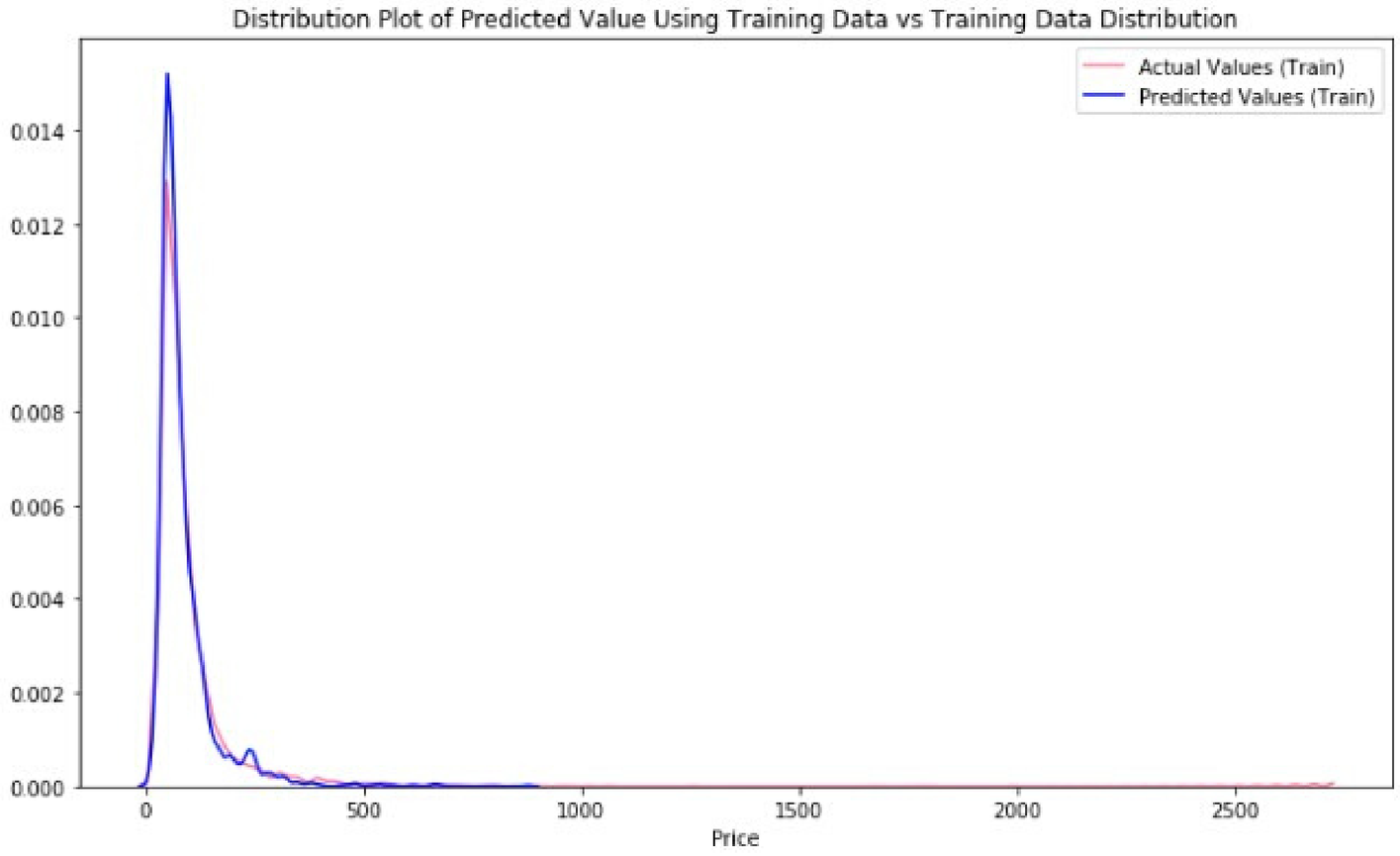
Results

Model Results	MAE	MSE	R^2
Random Forest Regressor	22.50	4,277.01	0.0814
XGBoost Regressor	26.24	4,095.41	0.2862
Gradient Booster Regressor	32.14	4,945.79	-0.1315
Lasso Regression	35.60	6,026.51	-0.4364
Ridge Regression	21.95	3,471.91	0.4165

The selected model is the **Random Forest Regressor.** It has the best R-squared score, and both MAE and MSE measures are not the best but actually good.

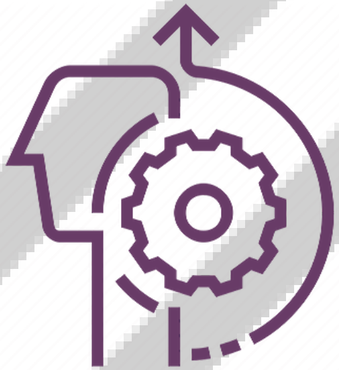
Distribution plot

Here is the distribution of the predicted values using the training dataset, and the distribution of this training data.

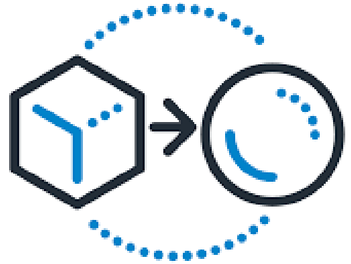


Model performances can improve by:

Applying Feature Selection methods



Applying methods for parameter optimization



Applying transformations

Erick Tadeo Gaitán González

Mail: erick.gaitan@udem.edu

LinkedIn: https://www.linkedin.com/in/erick-tadeo-gaitán-gonzález-5a5a1217a

Thank you for reading!