Predicting House Prices – Bangalore, India

By Erick Gaitán

I. Background

It is common to find announces of data science jobs in Bangalore, India on websites like LinkedIn, indeed, naukri, monsterindia and many others.

It is known that Data Science is experiencing a surge in jobs across the world. India is one such country that is experiencing the "data explosion" [1]. The ones who started to learn about popular programming languages such as Python and R, about databases, and machine learning and deep learning algorithms in order to become a Data Scientist will found interesting that India is hungry for data scientists, the *"Sexiest Job of the 21st Century"*.

On February 2019, there was close to 97,000 data science and analytics jobs vacant in India, which represents the 45% over 2018 [2]. This country represents a strong option for career development for many emerging data scientists.

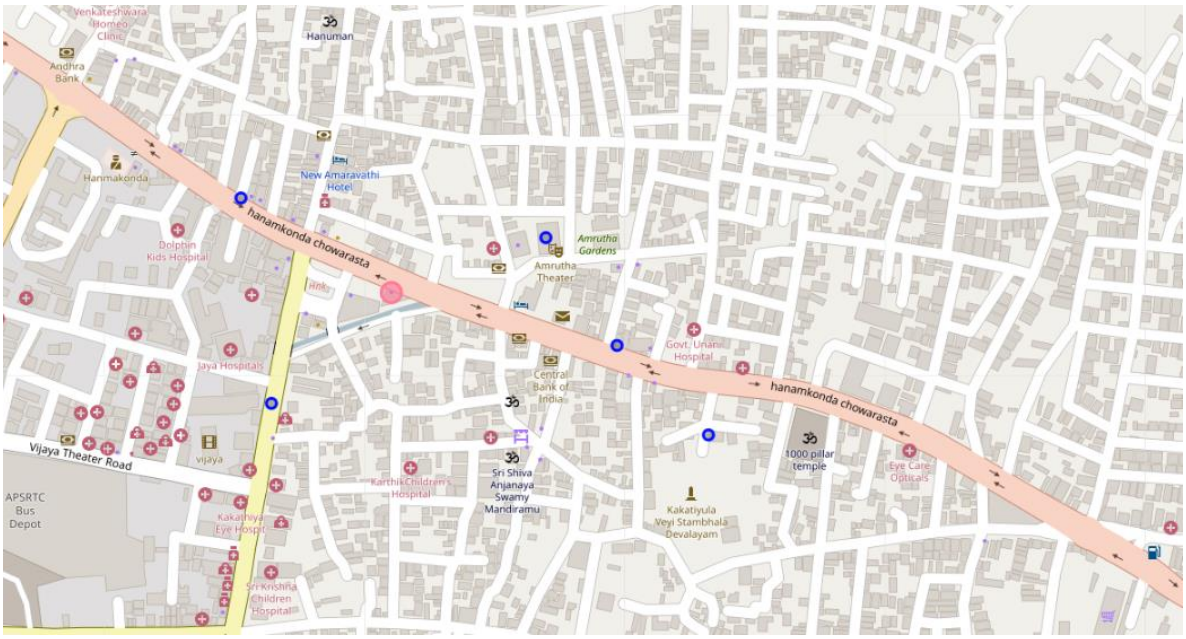
Bangalore is considered the IT hub of the country, it has some of the top Data Science companies in India [3] looking for: S

Data Science aspirants	**Data Science professionals**	**ML or AI aspirants**	**Cloud Computing professionals**
Equifax	Accenture	Braina AI	IBM
Gramener	Amazon	Niki.AI	Citrix
Fractal Analytics	Deloitte	Dell	Salesforce
	Flipkart	Juniper Network	Cisco
	IBM	ABB	Dell
	Citrix	Absentia VR	42Gears
	LinkedIn	Sig Tuple	Accenture

Bangalore has to be considered by those who want to start a career as a Data Scientist (or continue it). The main purpose of the project is to give information of house prices in Banglore by building and applying different regression models.

II. Data description

First, the Foursquare API is used for getting venues data of Bangalore, India. The dataframe contains five rows.



There are two Shopping malls, a snack place, a movie theater, and a historic site around 500 meters obtained from Foursquare; this map shows the State Bank of India, Reliance Mobile Store, the Govt Unani Hospital, a book store and more places around the Hanamkonda Chowarasta.

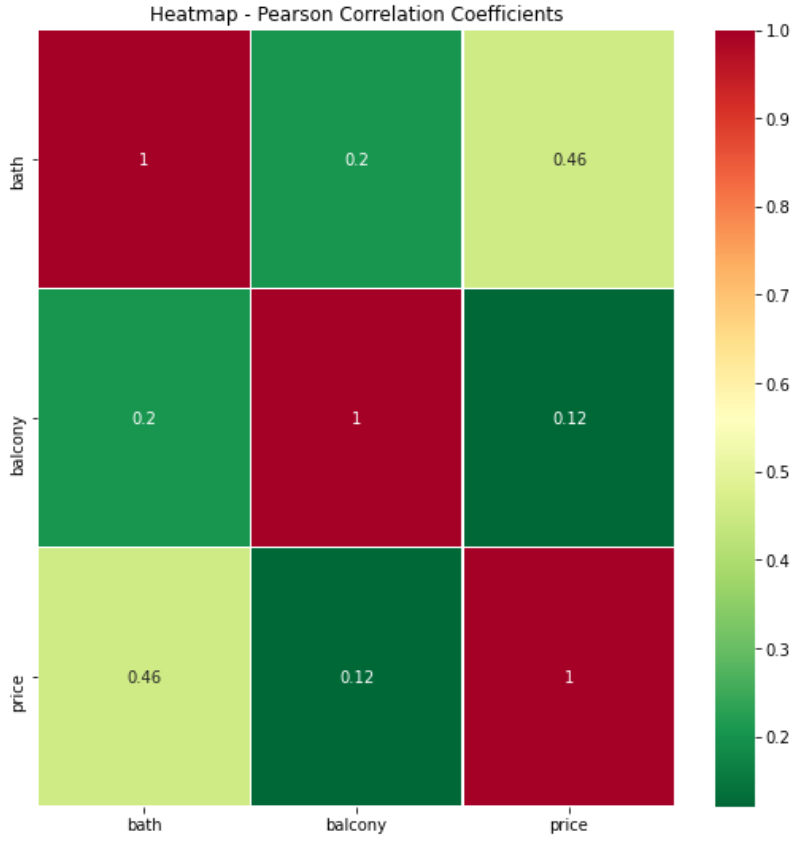
There are all kind of places near to this neighborhood, so no one has to be worried about having near hospitals or stores.

Then we have a dataset of house prices in Banglore obtained from Kaggle. It is a CSV file that contains 13,320 rows of data. You can find the file here.

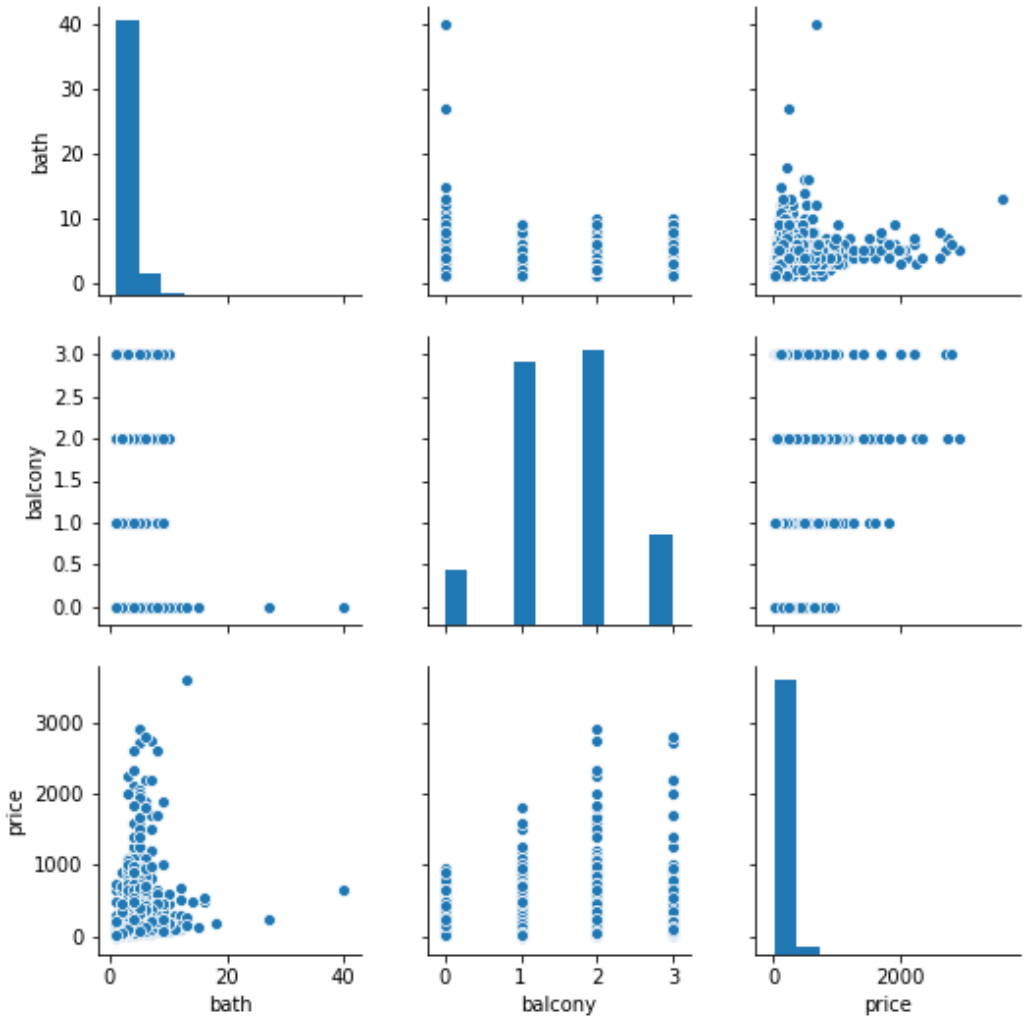
The dataframe contains nine columns, where the variable “price” will be the target. First, is necessary to remove empty values:

Column name	Data Type	#Distinct	NA Values
area_type	object	4	0
availability	object	81	0
location	object	1305	1
size	object	31	16
society	object	2688	5502
total_sqft	object	2117	0
bath	float64	19	73
balcony	float64	4	609
price	float64	1994	0

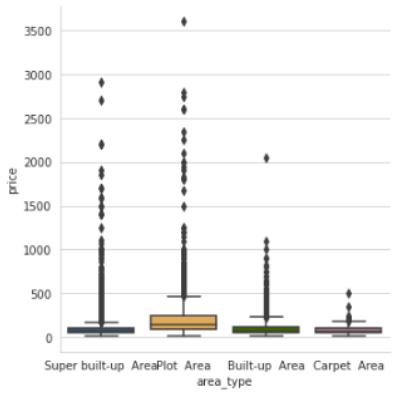
After removing those incomplete rows, the dataframe contains 7,496 rows. Now, it is moment to start describing those variables. Here is a heatmap that contains Pearson Correlation Coefficients, where it is shown that there are not strong correlations between variables.

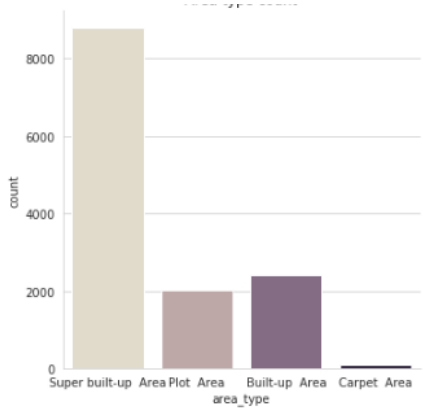


Here are histograms and scatter plots of continuous variables. There is a positive relationship between variables bath and price.



Then we have the categorical variables, one of them is area type, which contains four unique values. Below we can see price boxplos by area type and a barplot. We can see that there are many outliers.





Here are some basic statistics of each continuous and categorical variables:

Continuous variables

	bath	balcony	price
count	13247.000000	12711.000000	13320.000000
mean	2.692610	1.584376	112.565627
std	1.341458	0.817263	148.971674
min	1.000000	0.000000	8.000000
25%	2.000000	1.000000	50.000000
50%	2.000000	2.000000	72.000000
75%	3.000000	2.000000	120.000000
max	40.000000	3.000000	3600.000000

Categorical variables

	area_type	availability	location	size	society	total_sqft
count	13320	13320	13319	13304	7818	13320
unique	4	81	1305	31	2688	2117
top	Super built-up Area	Ready To Move	Whitefield	2 BHK	GrrvaGr	1200
freq	8790	10581	540	5199	80	843

The target variable is price as is mentioned before. The features are used to build different regression models to predict house prices in Banglore. The best one is selected by comparing various evaluation metrics of each model.

III. Data Preparation

The first step is removing rows with empty values, what is explained before. There are 7,496 remaining rows of data.

Then, it is necessary to transform object type variables to int type to apply model functions of scikitlearn library. The categorical variables area_type, availabilty, location, size, society and total_sqft are encoded using the OneHotEncoder method.

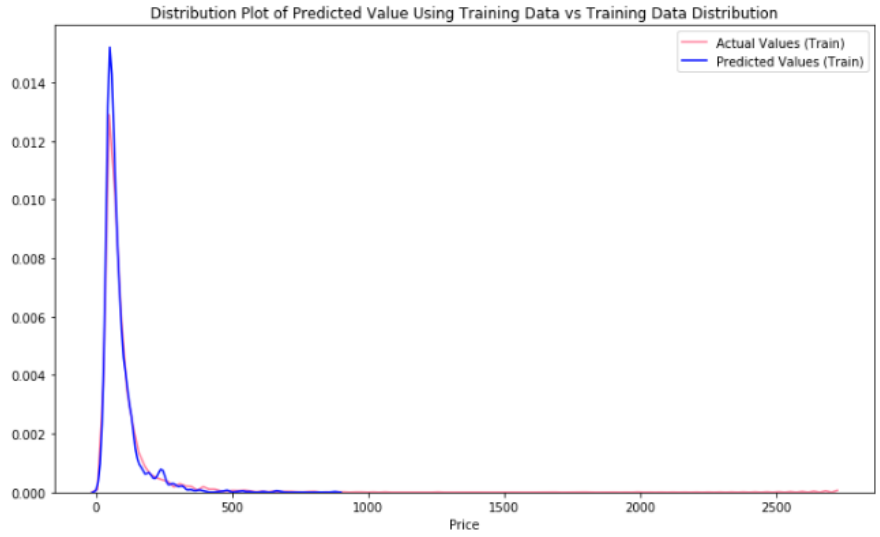
IV. Models and evaluation

For predicting the target variable price, these regression models are applied: Random Forest Regressor, XGBoost Regressor, Gradient Boosting Regressor, Lasso Regression and Ridge Regression.

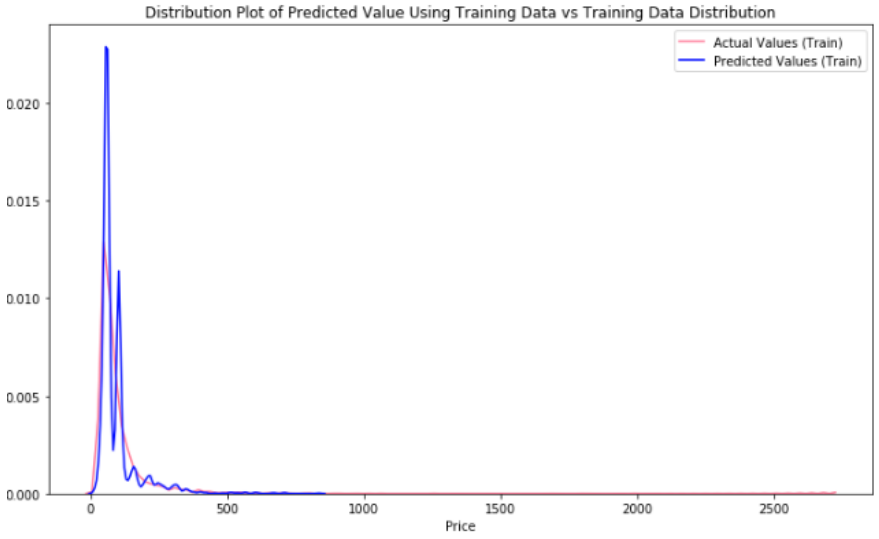
Each model is applied without any parameter definition, and all the features are included in each model.

The next table show the distribution plot of predicted values using the training dataset, and the distribution of this training dataset, for each of the models:

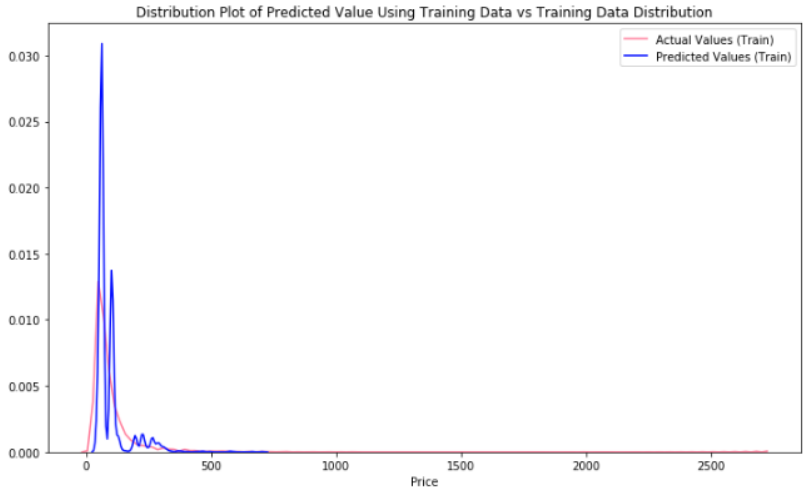
- Random Forest Regressor



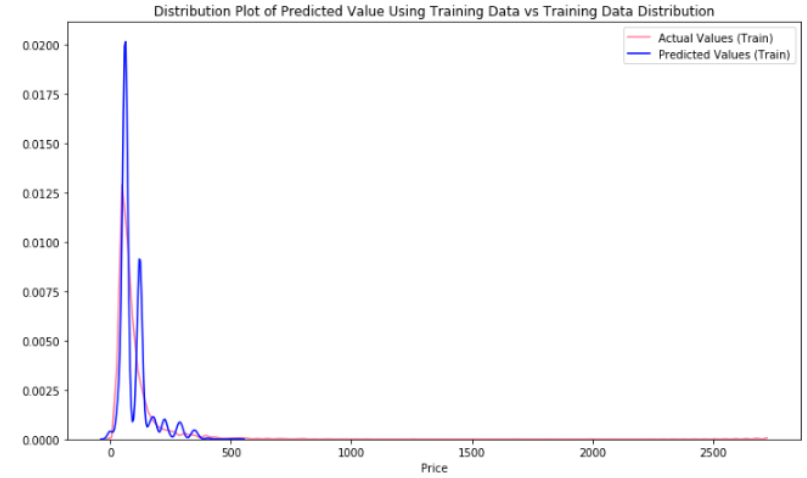
- XGBoost Regressor



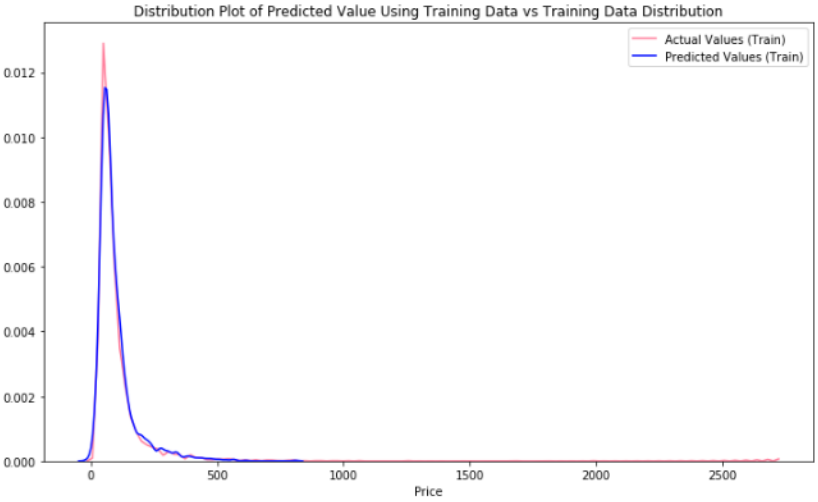
- Gradient Booster Regressor



- Lasso Regression



- Ridge Regression



Each model seems to be good, their predictions distribution are very similar to the distribution of the train dataset. There is just some variations that the model could not emulate.

V. Results comparison

The evaluation metrics used are Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R-squared score. Here are the results of each model.

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 Random Forest Regressor model has the R-squared score closest to zero and its MAE is the second lowest. For the Ridge Regression model, both MAE and MSE are the lowest, but its R-squared score is the second worst result.

There is a big difference between MSE of Random Forest and Ridge Regression, but their MAE results are close. The Random Forest's MSE value is not bad at all. On the other hand, the R-squared error of the Lasso model is actually bad. That's why the best model of those is the Random Forest Regressor.

The Lasso Regression model has definitely the worst performance. Its R-squared score, MSE and MAE results are the worst.

VI. Conclusions

The models shown here are “preliminar” models, because of the use of predetermined parameters, no data normalization applied, and every feature available is included for each model.

Those models can be better, by applying GridSearchCV method for parameter optimization, to find the best performance possible by defining values for each parameter.

Data transformations can also improve model performances, applying MinMaxcaler, RobustScaler, StandardScaler or Normalizer methods; the model performance will not necessarily improve by applying any of these methods, it is needed to compare results.

Also can be applied FeatureSelection methods, in order to know how good or bad is each variable for the model performance. The model will not necessarily by including more and more variable on its construction, so it is important to include just those most important variables to see how performance change.

References

[1] DATAFLAIR TEAM. (2019). Is there any Scope of Data Science in India? Take Expert’s Opinion. Retrieved from https://data-flair.training/blogs/scope-of-data-science/.

[2] Press Trust of India. (2019). 97,000 analytics, data science jobs vacant in India: Study. Retrieved from https://www.thehindubusinessline.com/economy/97000-analytics-data-science-jobs-vacant-in-india-study/article26399660.ece#.

[3] Manipal ProLearn. (n.d.) Data Science Companies In Bangalore. Retrieved from https://www.manipalprolearn.com/data-science-companies-in-bangalore.