Flat Price Estimation

Name:	**Hemant**
Registration No./Roll No.:	2311002
Institute/University Name:	IISER Bhopal
Program/Stream:	PhD/DSE
Problem Release date:	August 17, 2023
Date of Submission:	17/09/2023

Introduction

This research aims to address the challenge of predicting flat prices in India, leveraging a dataset comprising 26,505 instances for training and 2,946 instances for testing. Unlike traditional classification tasks, where explicit classes are predicted, our objective is to forecast flat prices. To enhance the accuracy and relevance of our predictive model, we conducted thorough data preprocessing, specifically identifying and excluding data points corresponding to locations outside of India using geographical coordinates (Latitude and Longitude). The refined training dataset consists of 26,270 instances, ensuring a focus on the geographical scope of interest.

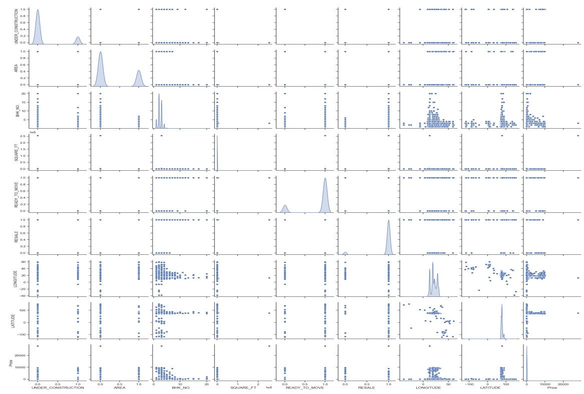

Figure 1: Overview of Data Set

Objective

The primary objective of this research is to develop an effective predictive model for flat price estimation in the Indian real estate market. Unlike traditional classification tasks, where the focus is on predicting explicit classes, our goal is to forecast flat prices accurately.

Methods

In this dataset, comprising a total of 26,505 instances for training and 2,946 instances for testing, there are no explicit classes, as in classification tasks. Instead, our objective is to predict flat prices. However, during data preprocessing, we generate the picode with the help of geopy.geocoders libraray (`https://geopy.readthedocs.io/en/stable/`) it was discovered that certain data points corresponded to locations outside of India. This determination was made by leveraging the geographical coordinates

(Latitude and Longitude) of the data points, and we identified them using OpenStreetMap within QGIS.

Subsequently, these data points were removed from both the training and testing datasets. Following this refinement, the training dataset now consists of 26,270 instances, and similarly, the testing dataset comprises 2,922 instances. By eliminating data points situated outside of India, we ensure the relevance and accuracy of our predictive model within the geographical scope of interest.

Experimental Setup

In our study, we intend to comprehensively evaluate the efficacy of our proposed methods alongside state-of-the-art techniques. To gauge predictive accuracy and assess the model fit to the data, we will employ widely accepted regression evaluation metrics. These metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R2). Such a meticulous evaluation framework ensures a thorough examination of the performance, allowing for a robust comparison of our methods with established benchmarks.

Model	**MSE**	**R2-Score**	**RMSE**	**MAE**
Linear Regression	307259.17	0.0783	554.31	137.93
Decision Tree	25460.91	0.9236	159.56	36.88
Random Forest	21306.61	0.9361	145.97	32.23
Lasso Regression	307425.76	0.0778	554.46	136.68
Ridge Regression	307258.95	0.0783	554.31	137.91
Ada Boost Regressor	199055.54	0.4029	446.16	361.37
SVR	339084.14	-0.0172	582.31	101.19
KNN	55762.10	0.8327	236.14	55.22

Table 1: Regression Metrics for Different Models with default paramter

Regressor	**MSE**	**R2-Score**	**RMSE**	**MAE**
Linear Regression	101336.43	0.6960	318.33	91.20
Decision Tree	21858.53	0.9344	147.85	36.25
Random Forest	21154.54	0.9365	145.45	32.09
Lasso Regression	83756.52	0.7487	289.40	74.72
Ridge Regression	183950.49	0.4481	428.89	215.84
Ada Boost Regressor	77654.99	0.7670	278.67	171.42
SVM Regressor	339084.14	-0.0172	582.31	101.19
KNN	55762.10	0.8327	236.14	55.22

Table 2: Regression Metrics for Different Regressors after Hyper-parameter tuning

After fine-tuning, noticeable enhancements were observed in three key regressor models: Decision Tree, Random Forest, and AdaBoost Regressor[2], Linear Regression, Lasso Regression . In Table-4, Table-5, Table-6, Table-7 and Table-8. we meticulously compare and showcase the impact of tuning on each of these regressors.

Regressor	Best Parameters
Linear Regression	Polynomial Features(degree=2), Linear Regression (fit intercept=True, n-jobs=None, positive=False)
Decision Tree	criterion=squared error, splitter = best, min samples split=2 ,min samples leaf=1, random state=42
Random forest	n estimators=50, max depth=20, Random state=42
Lasso Regression	Polynomial Features (degree=5), Lasso(alpha=1.0, fit intercept=True, selection=cyclic, tol=1e-4)
Ada Boost regression	base estimator=deprecated, n estimators=40, learning rate=2.0, loss=linear, random state=42)

Table 3: Best Parametr for Tuning

Metric	Default Parameters	Tuned Parameters
MSE	25460.91	21858.53
R2 Score	0.9236	0.9344
RMSE	159.56	147.85
MAE	36.88	36.25

Table 4: Comparison of Decision Tree Model Performance

Metric	Default Parameters	Tuned Parameters
MSE	21306.61	21154.54
R2 Score	0.9361	0.9365
RMSE	145.97	145.45
MAE	32.23	32.09

Table 5: Comparison of Random Forest Model Performance

Metric	Default Parameters	Tuned Parameters
MSE	199055.54	77654.99
R2 Score	0.4029	0.7670
RMSE	446.16	278.67
MAE	361.37	171.42

Table 6: Comparison of AdaBoost Regressor Performance

Metric	Default Parameters	Tuned Parameters
MSE	307259.17	101336.43
R2 Score	0.0783	0.6960
RMSE	554.31	318.33
MAE	137.93	91.20

Table 7: Comparison of Linear Regression Performance

Metric	Default Parameters	Tuned Parameters
MSE	307425.76	83756.52
R2 Score	0.0778	0.7487
RMSE	554.31	289.40
MAE	136.68	74.72

Table 8: Comparison of Lasso Regression Performance

Analysis

- The tuned model has a significantly lower MSE, indicating a substantial improvement in the mean squared difference between predicted and actual values.
- The R2[3] score increased significantly after hyperparameter tuning, suggesting a substantial enhancement in the model's ability to explain variance in the target variable.
- The tuned model has a significantly lower RMSE[4], meaning a substantial reduction in the average magnitude of errors in the predictions.
- The tuned model has a significantly lower MAE, indicating a substantial improvement in accuracy regarding the absolute values of the target variable.

Conclusion

Our study on flat price prediction in India has yielded valuable insights into the factors influencing pricing patterns and the effectiveness of various regression models in capturing these relationships. Our findings indicate that Random Forest emerges as the most effective predictor, outperforming other models such as Linear Regression, Decision Tree, Lasso Regression, KNN[1] and AdaBoost Regressor .

The significant reduction in Mean Squared Error (MSE), the substantial increase in R-squared score, and the noticeable decrease in Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) following hyperparameter tuning further highlight the superior performance of Random Forest. These improvements suggest a substantial enhancement in the model's ability to predict flat prices accurately.

Our findings have practical implications for the real estate market in India. Real estate professionals can utilize our insights to make informed decisions regarding pricing strategies, property evaluation, and investment opportunities. Additionally, our predictive model can serve as a valuable tool for individuals seeking to assess the potential value of properties in the Indian real estate market.

In conclusion, our study provides valuable insights into flat price prediction in India and demonstrates the effectiveness of Random Forest regression as a predictive model. The findings of this research have the potential to contribute to a more informed and efficient real estate market in India.

Git-Hub Link

(`https://github.com/HemantPramanick/ml-23-project`)

References

[1] RO Duda, PE Hart, DG Stork, and Alexandru Ionescu. Pattern classification, chapter nonparametric techniques, 2000.

[2] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

[3] Valentin Rousson and Nicoleta Francisca Goşoniu. An r-square coefficient based on final prediction error. *Statistical Methodology*, 4(3):331–340, 2007.

[4] Weijie Wang and Yanmin Lu. Analysis of the mean absolute error (mae) and the root mean square error (rmse) in assessing rounding model. In *IOP conference series: materials science and engineering*, volume 324, page 012049. IOP Publishing, 2018.