Flat Price Estimation

Name:	Prasanna Devendra Kharode
Registration No./Roll No.:	21144
Institute/University Name:	IISER Bhopal
Program/Stream:	e.g., DSE
Problem Release date:	Aug 17, 2023
Date of Submission:	Nov 19 2023

1 Introduction

The objective of this project is to predict the prices (in lakhs) of flats in various cities of India based on different factors. Precise estimation of flat prices within India's real estate sphere holds substantial importance for informed decision-making among stakeholders. This project is poised to meet this critical need by developing a resilient predictive model, employing diverse regression methodologies to ensure highly accurate estimations. The model aims to enhance the reliability of predictions regarding flat prices, catering to the intricate landscape of the Indian real estate market. If necessary, refer to Figure 1 for visual representations and further insights.

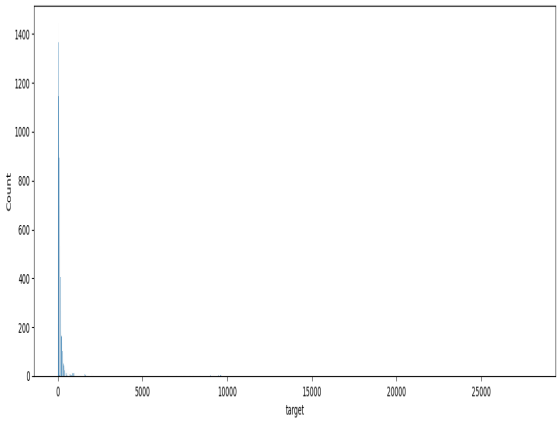

Figure 1: Overview of Data Set

Github Link: [1] used to implement the classifiers [1, 2].

2 Results and Discussion

3 Methods

Regression Techniques Employed Utilizing an array of regression models including Linear Regression, Ridge Regression, Decision Tree Regression, Random Forest Regression, Support Vector Machine Regression, and AdaBoost Regression. These models were selected for their adaptability to the complexities inherent in the Indian real estate domain.

[1] https://github.com/prasannakharode/ml-project.git

Table 1: Performance Of Different Classifiers Using All Features

Regressor	MSE	RMSE	R2 score
Adaptive Boosting	972941.114	986.3777	-1.5203
Decision Tree	357.2569	18.9012	0.376
Random Forest	517.113	22.7401	0.7058
Support Vector Machine	1816.174	42.6165	-0.0330
Linear Regression	821.2506	28.6574	0.5328

Data Preprocessing The dataset underwent meticulous preprocessing, encompassing treatment for missing values, encoding categorical variables, and scaling numerical features. This step aimed to fortify the models against potential data inconsistencies[3] In a flat project, handling outliers is essential for ensuring the accuracy and reliability of the predictive model. Outliers in this context might represent extreme values or irregularities in the features used for predicting flat prices, such as square footage, location details, or amenities.

4 Experimental Setup

The intricate feature selection process meticulously merged the SelectKBest and f regression methodologies, aiming to distill pivotal features crucial for optimal model performance.

Amidst model training and rigorous evaluation, the implementation of a robust k-fold cross-validation technique served as the cornerstone. Essential metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the esteemed R^2 score played the role of guiding stars. Lower MSE values signaled heightened precision in predictive prowess, while an elevated R^2 score attested to the model's adeptness in unraveling intricate dataset patterns and explaining variance.

These metrics aren't just numerical evaluations; they encapsulate the essence of model accuracy and its potential for widespread applicability. They stand as formidable benchmarks, articulating the model's adaptability to unforeseen data terrains and its resilience in real-world scenarios.

In this intricate realm of predictive modeling, the exclusion of outliers emerges as a linchpin, ensuring the sanctity and reliability of predictions in the context of flat price estimation. The pipeline method stands as an architectural blueprint, crafting organized pathways for the development of scalable, robust models with steadfast precision.

Performance Metrics on Training Data Employing k-fold cross-validation, the regression models exhibited commendable performance, particularly Ridge Regression and Random Forest Regression, showcasing promising predictive capabilities validated by favorable performance metrics (MSE, RMSE, R^2 score).

Github Link: [2] used to implement the classifiers [1, 2].

5 Results and Discussion

The Random Forest Regressor stands distinguished, showcasing superior performance metrics. Amidst its impressive accuracy in RMSE, the Linear Regressor reveals a shortfall in R2 Score, a recurring pattern observed in other models. This disparity may trace back to the dataset's preprocessing phase, notably the meticulous removal of outliers as previously discussed. The presented table encapsulates the RMSE, MSE, and R2 Score values, reflecting these nuances in model evaluation.

6 Conclusion

Summary of Findings The Random Forest Regressor emerges as the pinnacle choice for robust and dependable model training and deployment. The disparities in RMSE values and R2 Scores underscore

[2] https://github.com/prasannakharode/ml-project.git

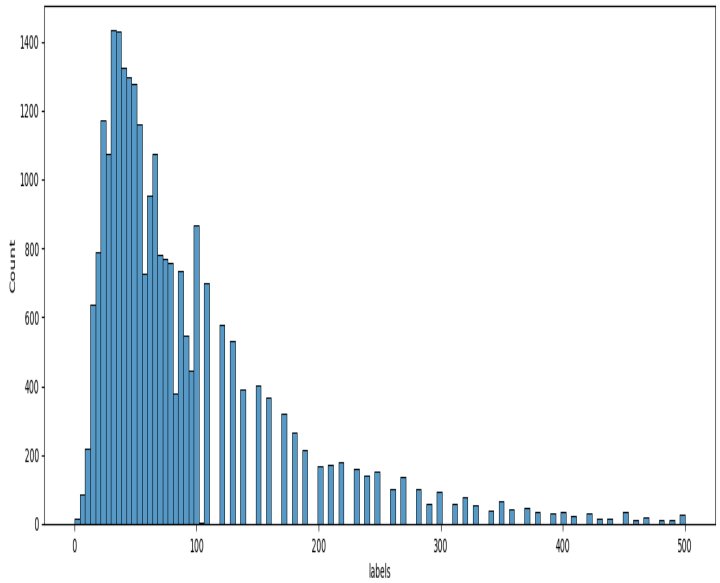

Figure 2: Removing outliners

the impact of imbalanced data points, offering insight into their influence on model performance. This underscores the significance of Outlier Removal, pivotal in yielding divergent metric outcomes.

Additionally, the meticulous process of Parameter and Feature Selection plays a pivotal role in achieving optimal results and enhancing precision within the model. It stands as a critical phase, pivotal for attaining superior performance and ensuring the model's efficacy in practical applications.

The analysis has successfully crafted predictive models to estimate flat prices in India's real estate landscape. To fortify their robustness and real-world relevance, validating these models with independent test data is paramount. This critical future phase ensures comprehensive validation beyond the training data's constraints.

Deploying the model in real-time scenarios constitutes a pivotal stride in affirming its practical utility and performance in predicting flat prices. Real-world applications provide invaluable insights into the model's behavior in dynamic, unpredictable settings, affirming its effectiveness and practical feasibility..

References

[1] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[2] I. H. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, third edition, 2011.

[3] James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. *Advances in neural information processing systems*, 24, 2011.