Gievn Title of The Project

Name:	**Tushar Dhananjay Kale**
Registration No./Roll No.:	21134
Institute/University Name:	IISER Bhopal
Program/Stream:	e.g., DSE
Problem Release date:	February 02, 2022
Date of Submission:	The date you submitted

1 Introduction

The objective of this analysis is to predict flat prices in India using various regression techniques.The focal point of this study centers on accurately forecasting flat prices prevalent in the dynamic Indian real estate domain. Our approach integrates a comprehensive suite of regression methodologies, encompassing a diversified range of models. The selected arsenal includes Linear Regression, Ridge Regression, Decision Tree Regression, AdaBoost Regression, Random Forest Regression, and Support Vector Machine Regression. Each of these techniques brings unique strengths and adaptability to the table, promising a multifaceted analysis that can effectively capture the intricate nuances and complexities inherent in real estate pricing. This diversified toolkit empowers us to discern the most fitting model, leveraging the distinctive traits of each technique to craft a holistic and precise prediction framework for flat prices in India.

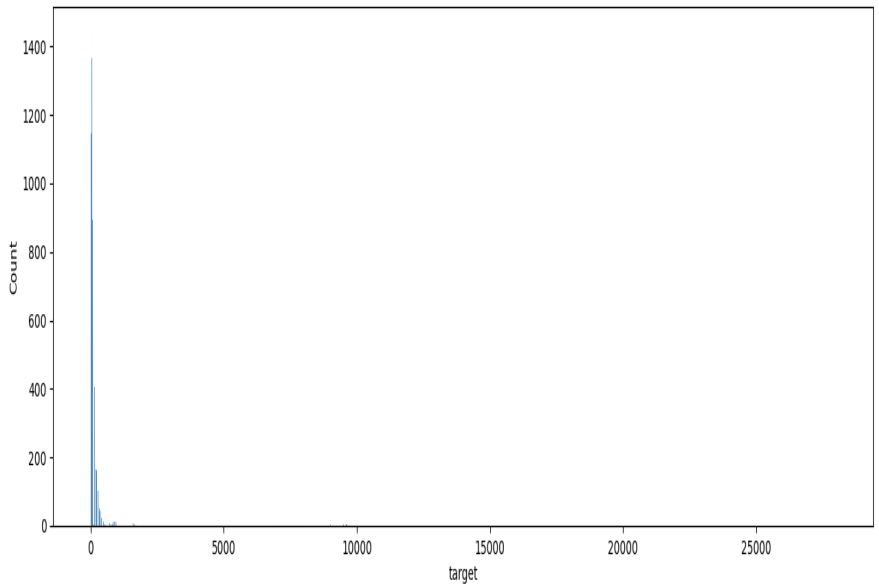

Figure 1: Overview of Data Set

2 Methods

The dataset, derived from real estate listings, underwent meticulous preprocessing steps:

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 Loading: The dataset was imported, encompassing features such as square footage, location details, room counts, and amenities. Data Cleaning: One-Hot Encoding was employed, and outliers were removed to enhance data quality and model performance. Regression Techniques The chosen regression techniques were tailored with specific parameters for optimal performance:

AdaBoost Regression: Employed Decision Tree as the base estimator. Decision Tree Regression: Tuned parameters including criterion, maximum depth, maximum features, and alpha. Ridge Regression: Configured with the solver for optimization. Random Forest Regression: Parameters tuned for criterion, number of estimators, maximum depth, and maximum features. Support Vector Machine Regression: Configured with the kernel and regularization parameter. We use grid search is a technique used for hyperparameter optimization in machine learning. In many machine learning algorithms, there are parameters that are not directly learned from the training data. These parameters, also called hyperparameters, [1] need to be set prior to training and can significantly impact the model's performance.

3 Experimental Setup

The feature selection process involved a hybrid strategy amalgamating SelectKBest and f regression, aimed at discerning pivotal features pivotal for model performance.

For model training and assessment, a k-fold cross-validation technique was adopted. Performance evaluation relied on key metrics like Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score. Lower MSE values indicate enhanced precision and reliability in predictive capabilities. A high R^2 score signifies a proficient model adept at capturing underlying patterns and explaining variance within the dataset. These metrics serve as vital benchmarks, portraying the accuracy and generalization prowess of the models. The utilization of these evaluation tools not only quantifies the predictive performance but also aids in determining the model's adaptability to unforeseen data, bolstering its reliability in real-world applications. in this outlier removal plays a crucial role in ensuring the accuracy and reliability of predictive models. In this context, the process involves identifying and handling data points that significantly deviate from the norm within the dataset used for predicting flat prices. Pipeline serves as a systematic approach, facilitating the development of robust and scalable models while maintaining organization and efficiency throughout the entire process.

4 Results and Discussion

The Random Forest Regressor stands out, demonstrating superior performance across various metrics. Although the Linear Regressor achieved a remarkably accurate RMSE value, its R2 Score fell short. Similar patterns emerged across other Regressors, highlighting a recurring trend of precise RMSE values coupled with less satisfactory R2 Scores. This discrepancy could be attributed to the dataset's preprocessing, particularly the removal of outliers as previously outlined.the above table shows the value of RMSE,MSE,R2 Score. Github Link: [1] used to implement the classifiers [2, 3].

[1] https://github.com/tushar200777/ML$_{Project.git}$

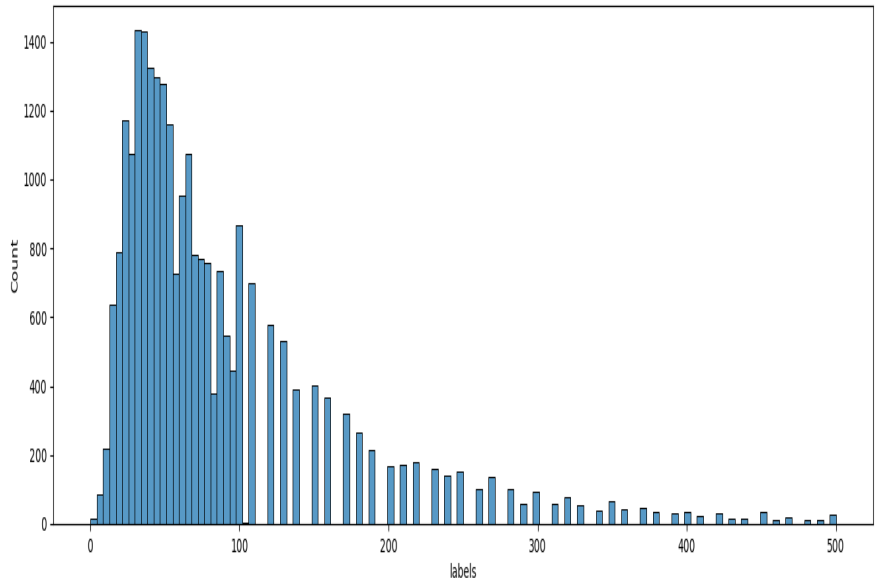

Figure 2: After removing Outliners

5 Conclusion

The analysis has effectively developed predictive models for estimating flat prices in the Indian real estate market. To bolster the models' reliability and real-world applicability, further exploration and evaluation on independent test data are essential. This future step will validate the trained models comprehensively and verify their efficacy outside the training set.

Moreover, enriching the model through advanced feature engineering approaches or the inclusion of additional relevant features could substantially elevate its performance. This aspect opens avenues to enhance the model's predictive capabilities, potentially capturing more nuanced patterns within the data.

Deploying the model in real-time scenarios constitutes another crucial step in validating its practical utility and performance in predicting flat prices. Real-world applications will offer invaluable insights into the model's behavior in dynamic, unpredictable environments, affirming its effectiveness and practical viability.

References

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

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

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