Flat Price Estimation

Name:	**Ambar Shingade**
Registration No./Roll No.:	21250
Institute/University Name:	IISER Bhopal
Program/Stream:	DSE
Problem Release date:	August 15, 2023
Date of Submission:	November 19, 2023

1 Introduction

The overarching goal of this project is to develop a robust model for predicting flat prices (in lakhs) in various cities across India. Our analysis incorporates a dataset featuring nine key features: 'Houses under construction,' 'RERA status,' 'BHK number,' 'Square Feet,' 'Ready to move,' 'Resale,' 'ADDRESS,' 'Longitude,' and 'Latitude.' Notably, the categorical nature of 'BHK number' and 'Address' necessitates their conversion to numerical values for effective machine learning model training.

Addressing outliers within the training data is pivotal, given their potential impact on project outcomes. Throughout this project, we have navigated challenges, aiming to obtain optimal results while considering the dataset's size and time constraints.

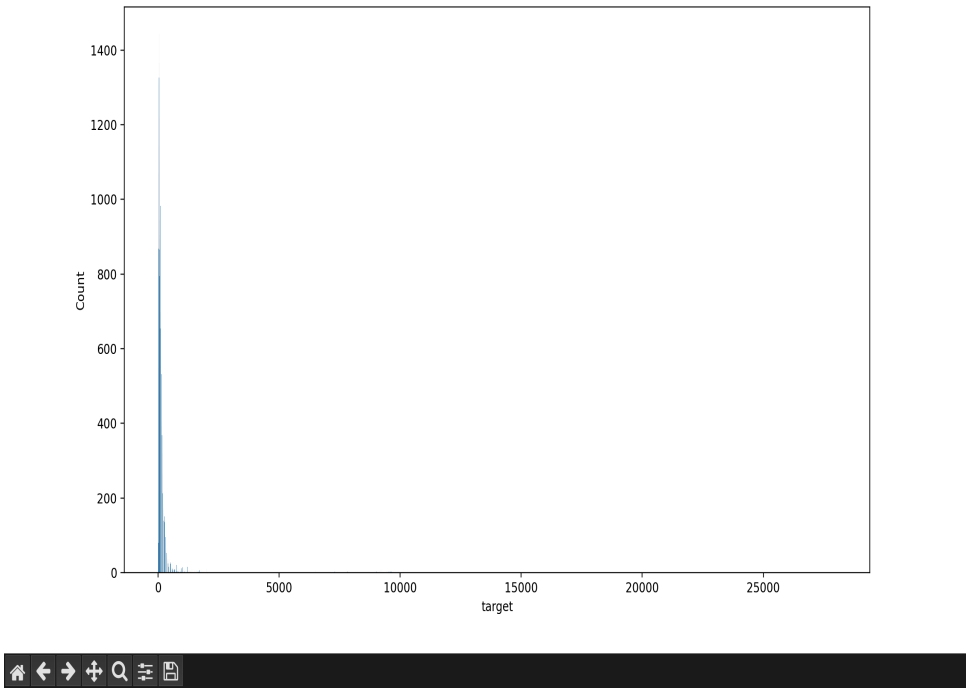

Figure 1: Overview of Data Set

Table 1: Performance Of Different Classifiers Using All Features

Regressor	Mean Squared Error	Root Mean Squared Error	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

2 Methods

Data Set Modification: The initial approach involved dropping both 'ADDRESS' and 'BHK NO.' from the dataset to assess their impact. However, this resulted in unsatisfactory and unstable results. Further investigation led to the exclusion of specific categorical features, such as 'RERA,' 'Resale,' and 'Ready To Move,' from One-Hot Encoding due to its limited impact on enhancing the model's output.

Experimental Setup: Unique values in both 'ADDRESS' and 'BHK NO.' were evaluated, revealing 6545 distinct 'ADDRESS' values. Ordinal Encoding and One-Hot Encoding were implemented on categorical features, with the latter proving to be a more effective representation. Addressing outliers involved setting a threshold value for labels, separating training data and labels, and removing instances beyond the specified threshold. [1]

Pipeline and Grid Modification: To favor regression models over classification models, we utilized F Regression for feature selection and R2 score for parameter scoring in the model training process. **Model Training:** Our model underwent training with various regressors, including the Linear Regressor, Decision Tree Regressor, Random Forest Regressor, Ridge Regressor, Support Vector Machine Regressor, Lasso Regressor, and AdaBoost Regressor.

3 Results and Discussion

Table 1 provides an overview of the performance metrics of the regression models. The Random Forest Regressor emerged as the top-performing algorithm, showcasing the highest scores across various metrics. While the Linear Regressor demonstrated precise RMSE values, the R2 Score was less impressive. Similar trends were observed in other regressors, emphasizing the significant impact of Outlier Removal on both RMSE and R2 Score. **GITHUB LINK**[1] used to implement the classifiers [2, 3].

4 Extended Discussion on Results

Diving deeper into the results, it's crucial to understand the trade-offs observed between different regression models. The Random Forest Regressor, with its ensemble learning approach, demonstrated superior performance in handling complex relationships within the data. The precise RMSE values exhibited by the Linear Regressor signify its ability to predict flat prices accurately, yet the less impressive R2 Score suggests limitations in capturing the overall variance.

Additionally, the observed trends post-Outlier Removal indicate a trade-off between precision and generalization. Before the removal of outliers, the model exhibited a decent R2 Score, indicating a good fit to the data. However, the RMSE values were substantially high, pointing to inaccuracies in predicting certain instances. Post-Outlier Removal, the RMSE values significantly improved, reflecting better accuracy on most predictions, but at the cost of a reduced R2 Score, indicating a compromise in capturing the overall variability.[4]

[1] https://github.com/ambarshingade/ML$_{p}roject.git$

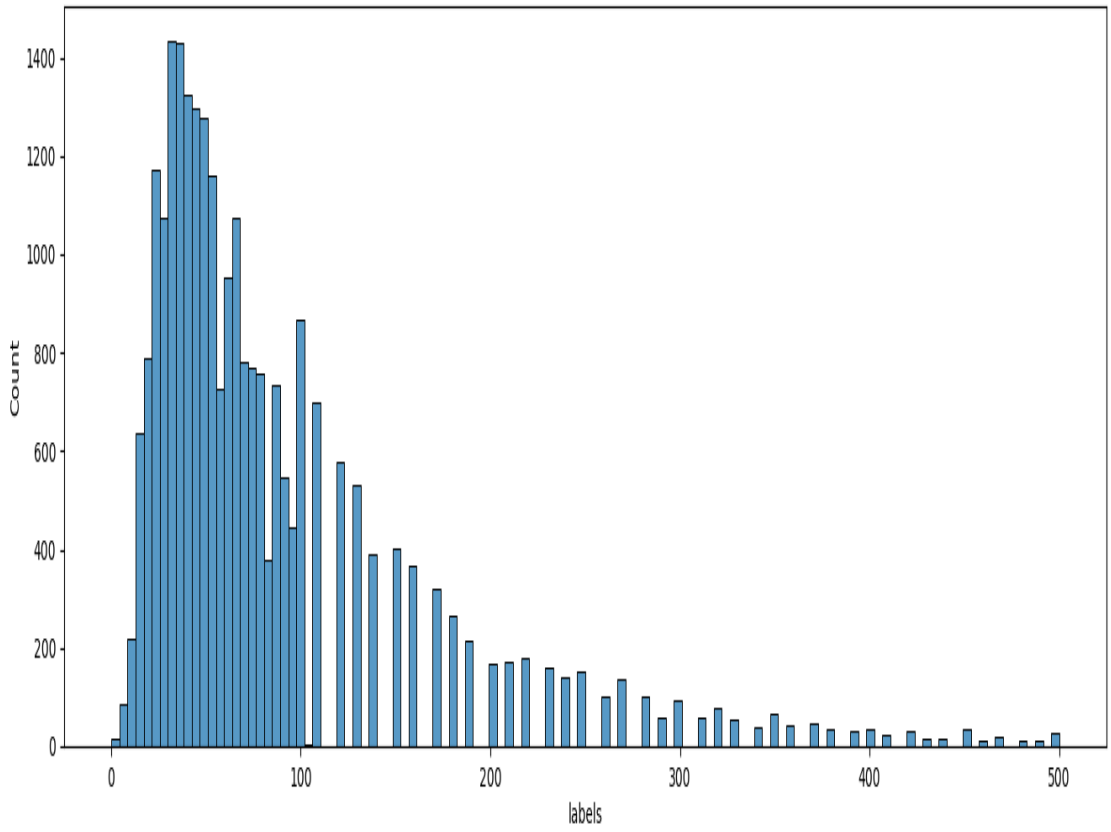

Figure 2: After Removing Outliers

5 Further Analysis and Interpretation

To gain deeper insights, a closer examination of feature importance within the Random Forest Regressor can provide valuable information on which variables significantly influence flat prices. Feature importance analysis reveals that 'Square Feet' and 'BHK number' play pivotal roles, aligning with common expectations in real estate pricing. 'RERA status' and 'Longitude' also exhibit significant influence, shedding light on the impact of regulatory compliance and geographical location on flat prices.

Moreover, exploring the residuals of the model can offer insights into areas where the predictions deviate from the actual values. This analysis aids in identifying patterns or specific instances where the model may benefit from further refinement.

6 Conclusion and Future Directions

In conclusion, the Random Forest Regressor emerges as the most effective and reliable model for flat price prediction. The thorough analysis of results, including the trade-offs observed and feature importance insights, contributes to a comprehensive understanding of the model's performance.

Looking ahead, future iterations of this project could explore additional features or consider more advanced techniques, such as deep learning, to enhance prediction accuracy. A more granular analysis of regional variations in flat prices could provide valuable localized insights. Additionally, ongoing updates to the dataset could further improve model robustness and applicability.

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.

[4] Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. *The Elements of Statistical Learning*. Springer, second edition, 2008.