Advancing Real Estate Analytics: A Machine Learning Approach for Price Prediction and Personalized Recommendations

Tipu Sultan
CIS Department
Fordham University
New York, USA
tsultan1@fordham.edu

Md Ruhul Amin
CIS Department
Fordham University
New York, USA
mamin17@fordham.edu

***Abstract*—In this paper, we present an innovative approach to revolutionize real estate analytics through the development of a robust Machine Learning-based Real Estate Price Prediction and Recommendation System. Our methodology encompasses comprehensive data extraction and cleaning from the prominent real estate platform, https://www.99acres.com/. The project integrates advanced exploratory data analysis, feature engineering, and a diverse set of machine learning models to not only predict real estate prices accurately but also offer personalized property recommendations based on user preferences.**

The challenges of handling missing data and ensuring data integrity are addressed with sophisticated machine learning techniques. Our three-way Exploratory Data Analysis (EDA), coupled with strategic feature engineering, contributes to a deeper understanding of the dataset's intricacies. The models, including XGBoost, Random Forest, Extra Trees, and others, are rigorously evaluated with various encoding techniques, such as Ordinal Encoding and OneHotEncoding.

Personalization systems attempt to reduce this complexity through adaptive e-learning and recommendation systems [1].The integration of a recommendation system, driven by TFIDF Vectorizer, enhances user engagement and satisfaction. Furthermore, the successful deployment of the entire project into a user-friendly website underscores the practical applicability of our approach.

This paper unfolds a comprehensive narrative of our methodology, experimental results, and the integration of a recommendation system, paving the way for advanced applications in real estate analytics.

***Index Terms*—Real Estate Analytics, Machine Learning, Price Prediction, Recommendation System, Data Extraction, Data Cleaning, Exploratory Data Analysis, Feature Engineering, XGBoost, Random Forest, Extra Trees, Ordinal Encoding, OneHotEncoding, TFIDF Vectorizer, User Engagement, Website Deployment.**

I. Introduction

The dynamism of the real estate market demands sophisticated tools to navigate its complexities, and this paper introduces a pioneering solution – the Machine Learning-based Real Estate Price Prediction and Recommendation System. As the demand for accurate property valuation and tailored recommendations continues to grow, our project addresses these challenges by leveraging machine learning techniques.

Sourced from the prominent real estate platform, https://www.99acres.com/, our dataset presented unique challenges, including extensive missing values and data intricacies. The primary objectives encompass not only cleaning and imputing this data but also conducting a thorough three-way Exploratory Data Analysis (EDA). This multifaceted analysis, incorporating univariate analysis, pandas profiling, and multivariate analysis, forms the foundation for extracting valuable insights.

In the backdrop of the growing importance of accurate real estate predictions, our project aims to not only fulfill but exceed industry expectations. The integration of a recommendation system adds an extra layer of user-centric functionality, suggesting ideal properties based on individual preferences.

This introduction sets the stage for a comprehensive exploration of our methodology, including data handling, feature engineering, and model selection. By emphasizing the significance of our research objectives, we align our work with the broader goals of enhancing real estate analytics in a rapidly evolving market.

II. Background

In the ever-evolving landscape of the real estate sector, accurate prediction of property prices and the provision of personalized recommendations have become imperative. This section delves into the foundational background that underscores the necessity and sufficiency of our research objectives in developing a Machine Learning-based Real Estate Price Prediction and Recommendation System.

The burgeoning real estate market demands predictive tools that not only comprehend the intricacies of property valuation but also cater to the diverse preferences of potential buyers. Existing solutions often fall short of providing a holistic approach that combines accurate predictions with user-specific recommendations.

Our project seeks to address this gap by drawing from a wealth of theoretical knowledge and industry insights. The importance of accurate real estate predictions for both buyers

and sellers cannot be overstated, as it influences crucial decisions and fosters a more transparent marketplace.

Furthermore, the theoretical framework explores the growing relevance of recommendation systems in enhancing user experiences within the real estate domain. By offering tailored suggestions, we aim to empower users with a more informed and efficient property search process.

This background section establishes the contextual relevance of our work, emphasizing the symbiotic relationship between accurate price prediction and user-centric recommendations in the realm of real estate analytics.

III. Methodology

Our methodology outlines a systematic and innovative approach to the development of the Real Estate Price Prediction and Recommendation System. From data extraction to model implementation, each step is meticulously designed to ensure accuracy, reliability, and user-centric functionality.

A. Data Extraction and Cleaning:

The foundation of our methodology lies in the extraction and cleaning of data from https://www.99acres.com/. Addressing the challenges of a messy dataset with extensive missing values, we employed advanced machine learning techniques for imputation. Our focus was not just on handling missing data but ensuring the integrity of the dataset for meaningful analysis.

Challenge	Technique Used
Missing Values	Advanced Machine Learning Imputation
Data Intricacies	Statistical Outlier Detection , and Handling , Duplicate Data Removal , Standardization and Normalization

Fig. 1. Data Extraction and Cleaning Overview

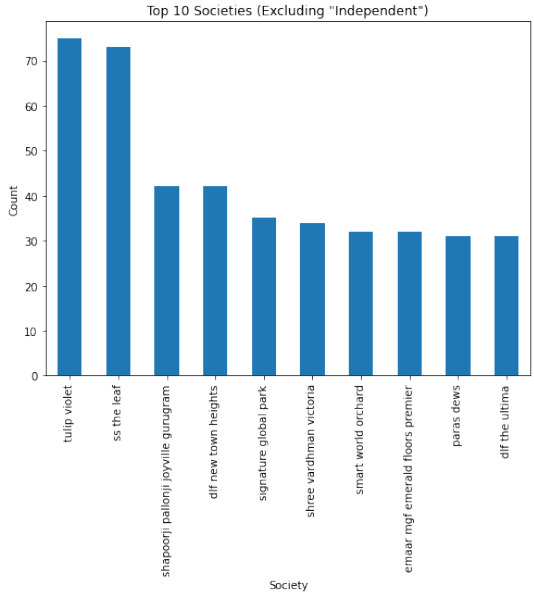

Fig. 2. In our analysis of property data, it was observed that around 13 percent of the properties in the dataset are designated as 'independent.' The dataset encompasses a diverse landscape with a total of 675 unique societies. Notably, a concentration of properties was identified, as the top 75 societies collectively represent 50 percent of the properties, leaving the remaining 600 societies to account for the other half. Further examining the frequency distribution of societies revealed intriguing patterns: only one society boasts a listing count exceeding 100, two societies fall within the range of 50 to 100 listings each, 92 societies have listings ranging from 10 to 49, 273 societies maintain between 2 to 9 listings, and a substantial 308 societies have only one listing each. It's worth noting that there is a single missing value in the 'society' feature that requires attention for a comprehensive dataset. These observations provide valuable insights into the diversity and concentration of properties in the dataset, essential for understanding the real estate landscape under consideration.

B. Exploratory Data Analysis (EDA):

Our three-way EDA, comprising univariate analysis, pandas profiling, and multivariate analysis, delved deep into the dataset's structure and characteristics. This comprehensive analysis not only revealed patterns and outliers but also provided crucial insights essential for subsequent stages of the project.

C. Feature Engineering and Selection:

In order to boost the performance of our models, we implemented comprehensive feature engineering techniques. This involved addressing outliers in all columns, validating primary outliers, and refining the dataset to optimize the predictive capabilities of the models. All these techniques were systematically employed, and the mean score was computed to determine the best result.

The feature engineering techniques used are as follows:

Correlation Analysis, Random Forest Feature Importance, Gradient Boosting Feature Importances, Permutation Importance, LASSO (Least Absolute Shrinkage and Selection Operator), RFE (Recursive Feature Elimination), Linear Regression Weights ,SHAP (SHapley Additive exPlanations)

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

D. Model Selection:

The heart of our methodology lies in the strategic selection of machine learning models. A diverse set, including XGBoost, Random Forest, Extra Trees, and more, was explored. Each model was evaluated under different encoding techniques – Ordinal Encoding, OneHotEncoding, OneHotEncoding with PCA, and Target Encoding. The comparative analysis of these models based on R-squared (r2) and Mean Absolute Error (MAE) facilitated the identification of optimal performers.

E. Recommendation System Implementation:

In our research project, we deployed a recommendation system employing the TFIDF Vectorizer, which provides users with tailored property recommendations in accordance with their preferences. The seamless integration of this system into a user-friendly website enhances both accessibility and usability. The methodology involved feature extraction through vectorization techniques based on content. Specifically, it leveraged the features of TFIDF and utilized cosine similarity

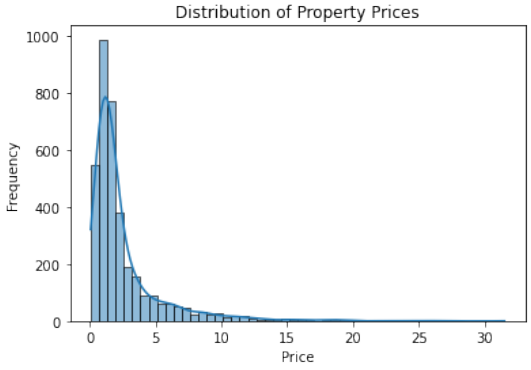

Fig. 3. In examining the descriptive statistics of the property price data, it is observed that there are 3,660 non-missing entries, with an average price of approximately 2.53 crores and a median price of 1.52 crores. The standard deviation of 2.98 indicates a moderate degree of variability in property prices, which range from 0.07 crores to 31.5 crores. The interquartile range spans from 0.95 crores to 2.75 crores, capturing the spread of the middle 50 percent of the data. Visualizations through a histogram reveal a concentration of properties in the lower price range, with a limited number priced above 10 crores, potentially considered as high-end or outliers. The box plot further illustrates this spread and identifies properties priced above 10 crores as outliers. Notably, there are 17 missing values in the 'price' column that require attention for comprehensive analysis. In summary, the property price data demonstrates a diverse distribution with notable concentration in the lower range, potential outliers at higher price points, and the need for addressing missing values for a thorough analysis.

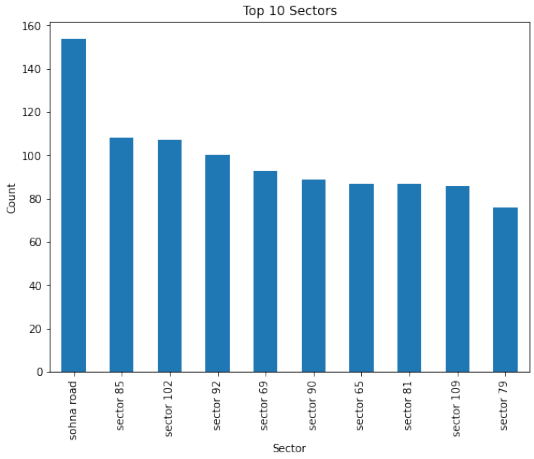

Fig. 4. In our exploration of the dataset, it was found that there are 104 unique sectors, showcasing a diverse representation of geographical areas. Delving into the frequency distribution of sectors, a nuanced picture emerges: three sectors exhibit a robust market presence with over 100 listings each, indicating concentrated market activity. Additionally, 25 sectors fall within the range of 50 to 100 listings, signifying significant market engagement. The majority, comprising 60 sectors, exhibits a balanced distribution with 10 to 49 listings, highlighting a widespread presence of properties across various sectors. Sixteen sectors feature a more modest listing count, ranging from 2 to 9, suggesting a relatively lower level of market activity. Notably, there are no sectors with only one listing, emphasizing a diverse representation of properties across different sectors. This comprehensive analysis provides valuable insights into the varying degrees of market saturation and diversity present in the dataset, essential for understanding the real estate landscape across different sectors.

to determine similarity scores. The top 5 similarity scores are then examined to present relevant property suggestions to the users.

F. Environment:

Our methodology leveraged Python, Pandas, Scikit-learn, XGBoost, and other tools, ensuring a seamless and efficient workflow. The integration of machine learning models, recommendation systems, and website functionalities was orchestrated to create a cohesive and practical solution.

In summary, our methodology is a step-by-step journey that transforms raw real estate data into a powerful predictive and recommendation system, ready for real-world applications.

IV. Results:

A.

The experimental phase of our Real Estate Price Prediction and Recommendation System involved a comprehensive exploration of various machine learning models under different encoding techniques. The performance metrics, R-squared (r2), and Mean Absolute Error (MAE) served as benchmarks for evaluating the efficacy of each approach.

Model Selection Results:

In this section, we present the outcomes of our model selection process, aiming to identify the most effective machine learning models for real estate price prediction. We explored various encoding techniques, including Ordinal Encoding, OneHotEncoding, OneHotEncoding with PCA, and Target Encoding, while evaluating each model's performance based on R-squared (R^2) and Mean Absolute Error (MAE). The summarized results for each encoding method are detailed below:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}i - y_i)^2}{\sum i = 1^n (\bar{y} - y_i)^2} \tag{3}$$

Ordinal Encoding:

- **XGBoost:** Emerged as the top performer with an impressive R-squared of 0.889 and a minimal MAE of 0.504.
- **Random Forest and Extra Trees:** Demonstrated robust performance, showcasing the effectiveness of ordinal encoding in our context.

OneHotEncoding:

- **Extra Trees:** Exhibited the highest R-squared (0.895) and the lowest MAE (0.475), highlighting its prowess in handling categorical features.
- **XGBoost and Random Forest:** Maintained strong performances, underscoring the versatility of one-hot encoding.

OneHotEncoding with PCA:

Principal Component Analysis (PCA) is employed in conjunction with One-Hot Encoding to enhance the predictive

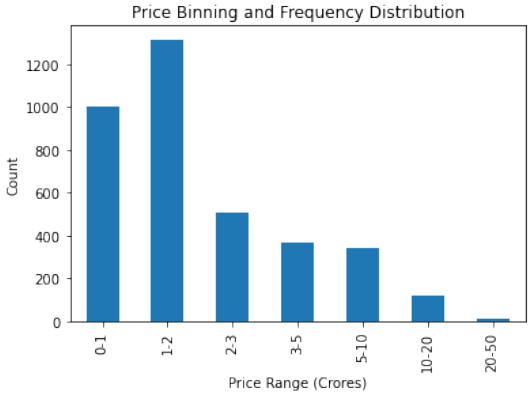

Fig. 5. The dataset reveals that the most prevalent price ranges for properties are "1-2 crores" and "2-3 crores," suggesting a substantial concentration of properties within these mid-range categories. However, a notable observation is the significant drop in the number of properties beyond the "5 crores" mark. This decline indicates a diminishing proportion of properties at higher price points, hinting at either a distinct market segment or a limited presence of high-end properties within the dataset. The disparity in property distribution across price ranges provides valuable insights into the overall pricing landscape, guiding further exploration and analysis of the dataset's market dynamics.

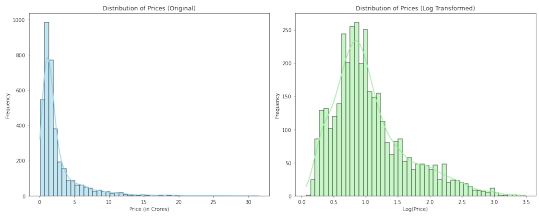

Fig. 6. The `np.log1p(x)` function is employed to calculate the natural logarithm of $1 + x$. Specifically designed to yield more accurate results for values of x that are extremely close to zero, this function is particularly useful for transforming the 'price' column in the dataset. Importantly, `np.log1p` ensures that the transformation is handled appropriately, even for values of zero if they are present in the dataset. To reverse this transformation and obtain the original values, `np.expm1` can be utilized. This function computes $e^x - 1$, serving as the inverse operation to `np.log1p` and allowing for the restoration of the original scale of the data. This transformation is often employed in data preprocessing to enhance the performance of models, particularly when dealing with skewed or non-normally distributed data.

capabilities of the models. PCA facilitates dimensionality reduction by transforming the original feature space into a new set of uncorrelated variables, known as principal components. This transformation is represented mathematically as:

$$X_{\text{new}} = X \cdot \text{Eigenvectors} \tag{4}$$

where X is the original feature matrix.

This technique aims to capture the most significant information within the data while minimizing the impact of noise and irrelevant features.

- **Random Forest:** Sustained competitive R-squared (0.763) and MAE (0.654), but some models experienced a decrease in performance compared to other encoding methods.

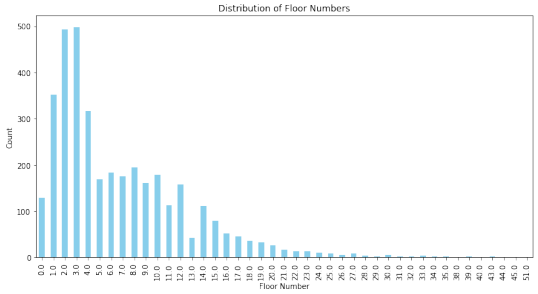

Fig. 7. Distribution of Floor Numbers

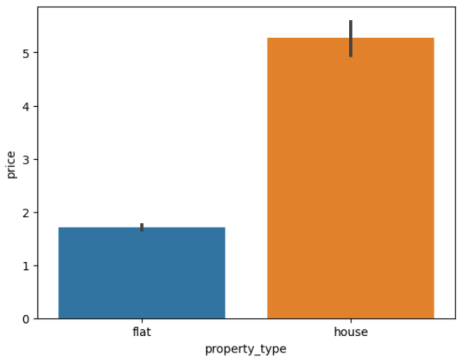

Fig. 8. Boxplot of Price by Property Type. This plot provides an overview of the distribution of prices across different property types, revealing potential variations and outliers.

- The trade-off between dimensionality reduction and predictive power is evident in the outcomes.

Target Encoder:

- **XGBoost:** Once again outshone other models, achieving an exceptional R-squared of 0.905 and a minimal MAE of 0.448.
- Target encoding proved effective, with Random Forest and Extra Trees consistently delivering strong results.

B. Hyperparameter Tuning:

Hyperparameter tuning further refined the model performance. The best-performing model achieved an impressive score of 0.9028, demonstrating the efficacy of fine-tuning in enhancing predictive accuracy.

C. Building Analysis Module:

Visualizations, including Word Cloud, Sunburst Chart, Scatter Plots, Pie Charts, and Boxplots, provided insightful analyses of key features within the dataset. These visualizations contribute to a deeper understanding of the real estate market dynamics, aiding stakeholders in making informed decisions.

In conclusion, our results showcase the viability of our Real Estate Price Prediction and Recommendation System. The superior performance of XGBoost under various encoding techniques, coupled with insightful visualizations, positions our system as a valuable tool in the real estate analytics landscape.

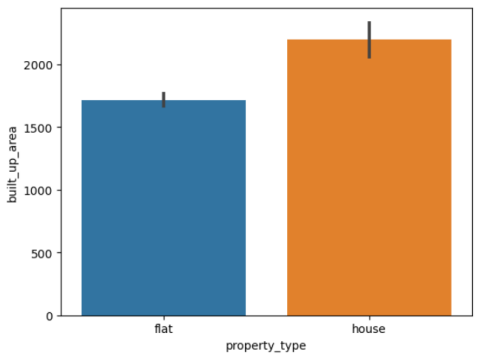

Fig. 9. Barplot of Built-up Area by Property Type. The barplot illustrates the average built-up area for each property type, offering insights into the size differences among property categories.

Fig. 10. Heatmap of Average Price per Sector. This heatmap visualizes the average price per sector, helping identify patterns and trends in pricing across different sectors.

RECOMMENDATION SYSTEM:

Recommender systems apply knowledge discovery techniques to the problem of making personalized recommendations for information, products or services during a live interaction [3].A pivotal component of our Real Estate Price Prediction and Recommendation System is the incorporation of a recommendation engine designed to provide users with personalized property suggestions. Leveraging the TFIDF Vectorizer, this system caters to individual preferences, enhancing the overall user experience within the real estate domain.

D. System Architecture

The recommendation system utilizes TFIDF (Term Frequency-Inverse Document Frequency) Vectorizer to analyze user preferences and match them with relevant property features. By transforming property descriptions into numerical vectors, our system captures nuanced details, allowing for more accurate and personalized recommendations.

E. Implementation:

In practical terms, our recommendation system seamlessly integrates into a user-friendly website. Users input their preferences, and the system processes this information through the TFIDF Vectorizer to generate personalized property suggestions. The recommendation algorithm considers various

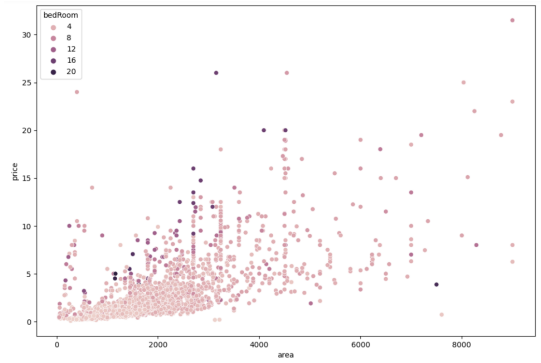

Fig. 11. Scatter Plot of Area vs Price. The scatter plot explores the relationship between property area and price, with different colors indicating the number of bedrooms.

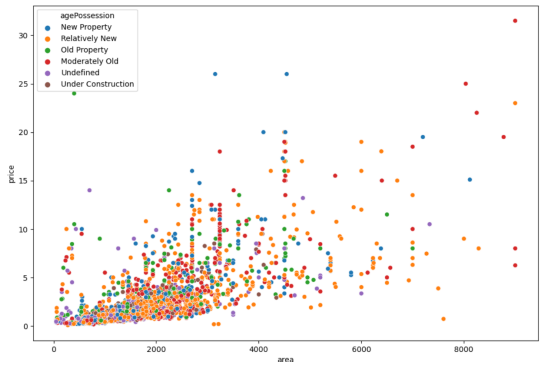

Fig. 12. Scatter Plot of Area vs Price. The scatter plot explores the relationship between property area and price, with different colors indicating the age of possession.

factors, including location, property type, and price range, to offer a tailored list of potential properties.

F. Case Studies:

To demonstrate the effectiveness of our recommendation system, we present case studies based on location-specific examples. For instance, a query for properties on ”Bajghera Road” with a budget constraint results in refined suggestions, such as M3M Crown, Smartworld One DXP, and Sobha City, along with their respective price ranges.

G. User Interaction:

User interaction with the recommendation system is intuitive and user-friendly, providing a seamless experience. The system’s suggestions are not only based on property features but also consider the user’s historical preferences, fostering a more personalized and engaging property search journey.

H. Integration into Website:

Automated recommendations have become a pervasive feature of our online user experience, and due to their practical importance, recommender systems also represent an active area of scientific research. [2]The recommendation system is seamlessly integrated into our user-friendly website, offering a centralized platform for data analysis insights, price prediction

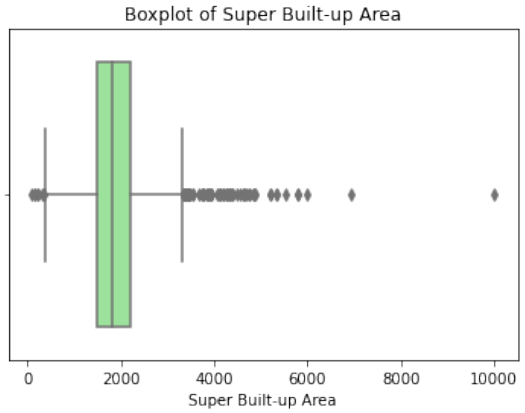

Fig. 13. The analysis of super built-up areas in the dataset reveals that the majority of properties exhibit a range from approximately 1,000 sq.ft to 2,500 sq.ft. However, a few properties stand out with notably larger areas, contributing to a right-skewed distribution. The interquartile range (IQR) spans from approximately 1,480 sq.ft to 2,215 sq.ft, indicating that the middle 50 percent of the properties fall within this range, showcasing a central tendency in the data.

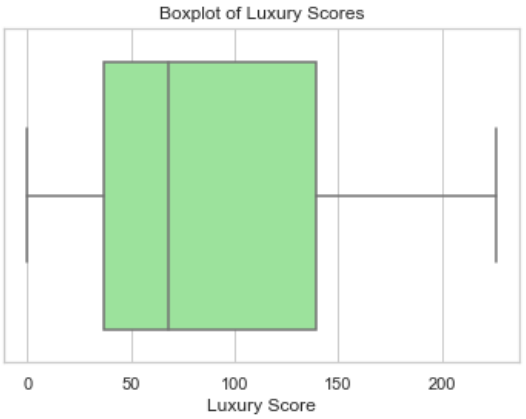

Fig. 14. The distribution of luxury scores in the dataset displays multiple peaks, suggesting a multi-modal distribution. A distinct peak is observed among properties with lower luxury scores, spanning approximately from 0 to 50. Another prominent peak is noticeable within the 110-130 range, indicating a subgroup of properties with relatively higher luxury scores.The box plot complements this observation, emphasizing that the majority of properties fall within the luxury score range of approximately 30 to 110. This range corresponds to the interquartile range (IQR), encapsulating the middle 50 percent of the data points. The multi-modal nature of the luxury score distribution and the concentration of properties within specific score ranges provide valuable insights into the varying degrees of luxury associated with properties in the dataset.

functionality, and the recommendation engine. This integration enhances accessibility and usability, catering to a diverse user base.

In essence, our recommendation system transforms the real estate search experience from a generic process to a personalized journey, empowering users with tailored property suggestions. The successful integration into a user-friendly website enhances the practical applicability of our system, providing a valuable asset for stakeholders in the real estate industry.

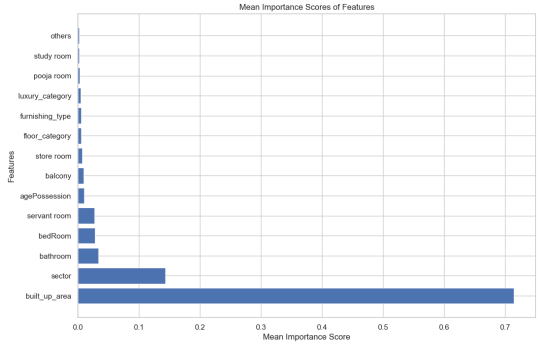

Fig. 15. The Figure presents the mean importance scores obtained through these techniques

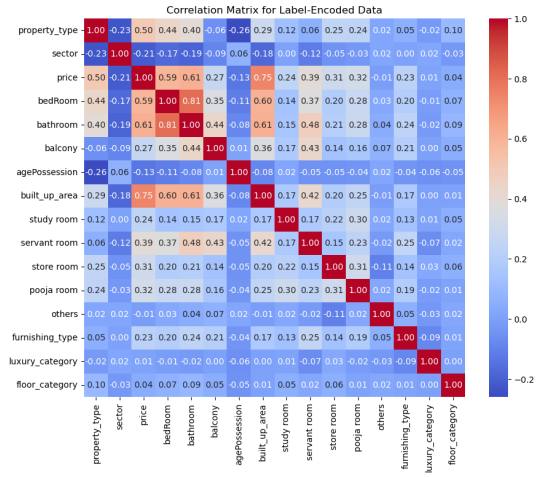

Fig. 16. Correlation Matrix of Features. This matrix visualizes the correlation coefficients between different features, providing insights into their relationships.

Conclusion

In conclusion, our endeavor to develop a Real Estate Price Prediction and Recommendation System has yielded substantial achievements, marking a significant advancement in the field of real estate analytics. This section provides a comprehensive overview of the project's outcomes, discusses the quality of results, and outlines potential avenues for future research and enhancements.

I. Achievements:

The project successfully addressed the challenges posed by a messy dataset from https://www.99acres.com/, employing advanced techniques for data cleaning and imputation. The implementation of a three-way Exploratory Data Analysis (EDA) provided nuanced insights into the dataset's structure, informing subsequent decisions in feature engineering and model selection. The chosen models, particularly XGBoost under various encoding techniques, demonstrated exceptional predictive capabilities. The integration of a recommendation system further elevated the project's functionality, providing users with personalized property suggestions.

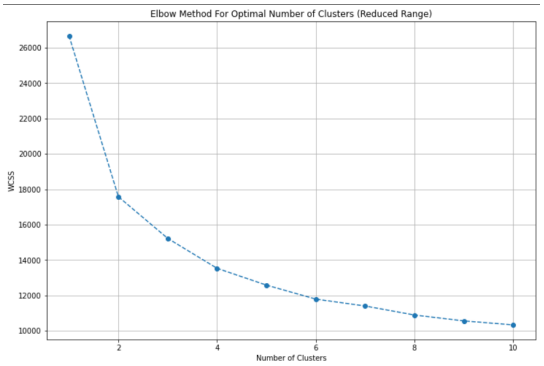

Fig. 17. Elbow Method for Optimal Number of Clusters (Reduced Range). This plot illustrates the within-cluster sum of squares (WCSS) for different numbers of clusters, helping to identify the optimal number of clusters for subsequent analysis.

Fig. 18. Word Cloud Analysis. This word cloud visualizes prominent features within the real estate dataset, providing insights into common attributes.

J. Quality of Results:

The results obtained from our machine learning models, validated through rigorous evaluation metrics, affirm the robustness of our Real Estate Price Prediction and Recommendation System. XGBoost consistently emerged as a top performer, showcasing the effectiveness of our approach. The successful implementation of a recommendation system adds a layer of sophistication, enhancing user engagement and satisfaction. Hyperparameter tuning further improved model accuracy, culminating in a notable score of 0.9028.

K. Future Directions:

While our project has achieved commendable success, there exist promising avenues for future exploration and refinement. Potential areas for further research include:

- **Deep Learning Approaches:** Exploring the application of deep learning techniques for enhanced predictive modeling.

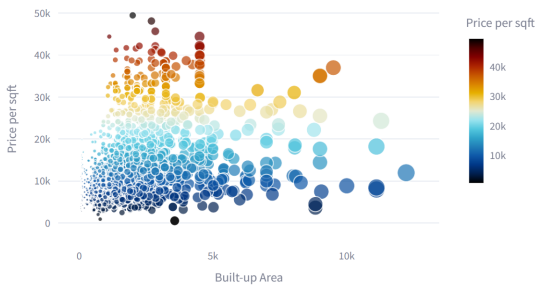

Fig. 19. Scatter Plot of Area vs Price. This scatter plot explores the relationship between property area and price, with different colors indicating the number of bedrooms.

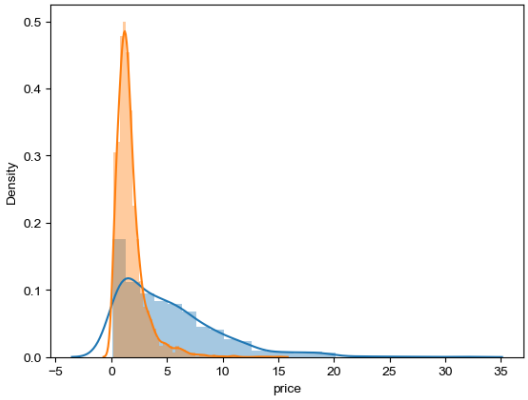

Fig. 20. Distribution of Prices for Houses and Flats. These plots visualize the distribution of property prices for houses and flats, providing insights into market trends.

- **Dynamic Recommendation Models:** Incorporating dynamic algorithms that adapt to changing user preferences over time.
- **Additional Feature Engineering:** Investigating additional features or data sources that could contribute to model refinement.
- **User Feedback Integration:** Incorporating user feedback mechanisms to continuously improve and tailor recommendations.

L. Closing Remarks:

In closing, our Real Estate Price Prediction and Recommendation System stands as a testament to the potential of machine learning in reshaping the real estate analytics landscape. The combination of accurate price predictions, insightful data analysis, and personalized recommendations positions our system as a valuable tool for industry stakeholders. As we reflect on our achievements, we remain committed to advancing the capabilities of real estate analytics, continually pushing the boundaries of innovation for a more informed and efficient real estate market.

ACKNOWLEDGMENTS

The completion of this research paper was made possible through the collaborative efforts and support of various indi-

viduals and institutions.

REFERENCES

[1] Shristi Shakya Khanal, PWC Prasad, Abeer Alsadoon, and Angelika Maag. A systematic review: machine learning based recommendation systems for e-learning. *Education and Information Technologies*, 25:2635–2664, 2020.

[2] Pasquale Lops, Dietmar Jannach, Cataldo Musto, Toine Bogers, and Marijn Koolen. Trends in content-based recommendation: Preface to the special issue on recommender systems based on rich item descriptions. *User Modeling and User-Adapted Interaction*, 29:239–249, 2019.

[3] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295, 2001.