Smart Real Estate Analytics: Integrating Housing Price Forecasting and Recommendation Systems

BY

MUSKAN JAIN
(Admission No. - 22MS0083)

Dissertation
SUBMITTED TO
Prof. Mritunjay Kumar Singh

INDIAN INSTITUTE OF TECHNOLOGY
(INDIAN SCHOOL OF MINES) DHANBAD

For the award of the degree of

MASTER OF SCIENCE
MAY 2024

INDIAN INSTITUTE OF TECHNOLOGY (INDIAN SCHOOL OF MINES) DHANBAD

CERTIFICATE FROM THE GUIDE(S)

This is to certify that the Dissertation entitled “Smart Real Estate Analytics: Integrating Housing Price Forecasting and Recommendation Systems” being submitted to the Indian Institute of Technology (Indian School of Mines), Dhanbad, by Ms Muskan Jain, Admission No 22MS0083 for the award of the Degree of Master, from IIT (ISM), Dhanbad, is a Bonafide work carried out by him/her, in the “Department of MATHEMATICS AND COMPUTING”, IIT(ISM), Dhanbad, under my/our supervision and guidance. The dissertation has fulfilled all the requirements as per the regulations of this Institute and, in my/our opinion, has reached the standard needed for submission. The results embodied in this dissertation have not been submitted to any other university or institute for the award of any degree or diploma.

Signature of the Guide (s)

Name: Prof. Mritunjay Kumar Singh

Date:

DECLARATION

I hereby declare that the work which is being presented in this dissertation entitled "Smart Real Estate Analytics: Integrating Housing Price Forecasting and Recommendation Systems" in partial fulfilment of the requirements for the award of the degree of Master of Science in MATHEMATICS AND COMPUTING is an authentic record of my own work carried out during the period from 10-01-2024 to 8-05-2024 under the supervision of Prof. Mritunjay Kumar Singh Department of MATHEMATICS AND COMPUTING Indian Institute of Technology (ISM) Dhanbad, Jharkhand, I
ndia.

I acknowledge that I have read and understood the UGC (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations, 2018. These Regulations were published in the Indian Official Gazette on 31st July 2018.

I confirm that this Dissertation has been checked for plagiarism using the online plagiarism checking software provided by the Institute. At the end of the Dissertation, a copy of the summary report demonstrating similarities in content and its potential source (if any) generated online using plagiarism checking software is enclosed. I herewith confirm that the Dissertation has less than 10% similarity according to the plagiarism checking software's report and meets the MoE/UGC Regulations as well as the Institute's rules for plagiarism.

I further declare that no portion of the dissertation or its data will be published without the Institute's or Guide's permission. I have not previously applied for any other degree or award using the topics and findings described in my dissertation.

(Signature of the Student)

Name of the Student: Muskan Jain

Admission No.: 22MS0083

Department: Mathematics and Computing

INDIAN INSTITUTE OF TECHNOLOGY (INDIAN SCHOOL OF MINES) DHANBAD

CERTIFICATE FOR CLASSIFIED DATA

(To be submitted at the time of Dissertation Submission)

This is to certify that the Dissertation entitled "Smart Real Estate Analytics: Integrating Housing Price Forecasting and Recommendation Systems" being submitted to the Indian Institute of Technology (Indian School of Mines), Dhanbad by Ms Muskan Jain for award of Master Degree in MATHEMATICS AND COMPUTING does not contain any classified information. This work is original and yet not been submitted to any institution or university for the award of any degree.

Signature of the Guide(s)　　　　　　　　　　　　Signature of the Student

INDIAN INSTITUTE OF TECHNOLOGY (INDIAN SCHOOL OF MINES) DHANBAD

CERTIFICATE REGARDING ENGLISH CHECKING

(To be submitted at the time of Thesis Submission)

This is to certify that the Dissertation entitled “Smart Real Estate Analytics: Integrating Housing Price Forecasting and Recommendation Systems” being submitted to the Indian Institute of Technology (Indian School of Mines), Dhanbad by Ms Muskan Jain Admission No 22MS0083, for the award of Master of Science has been thoroughly checked for quality of English and logical sequencing of topics.

It is hereby certified that the standard of English is good, and that grammar and typos have been thoroughly checked.

Signature of the Guides (s)	Signature of the Student
Name: Prof. Mritunjay Kumar Singh	Name: Muskan Jain
Date:	Date:

INDIAN INSTITUTE OF TECHNOLOGY (INDIAN SCHOOL OF MINES) DHANBAD

COPYRIGHT AND CONSENT FORM

(To be submitted at the time of Dissertation Submission)

To ensure uniformity of treatment among all contributors, other forms may not be substituted for this form, nor may any wording of the form be changed. This form is intended for original material submitted to the IIT (ISM), Dhanbad and must accompany any such material in order to be published by the ISM. Please read the form carefully and keep a copy for your files.

TITLE OF DISSERTATION: "Smart Real Estate Analytics: Integrating Housing Price Forecasting and Recommendation Systems"

AUTHOR'S NAME & ADDRESS: Muskan Jain, Pathan Street, Nabarangpur, Odisha (764059)

COPYRIGHT TRANSFER

1. The undersigned hereby assigns to Indian Institute of Technology (Indian School of Mines), Dhanbad all rights under copyright that may exist in and to: (a) the above Work, including any revised or expanded derivative works submitted to the ISM by the undersigned based on the work; and (b) any associated written or multimedia components or other enhancements accompanying the work.

CONSENT AND RELEASE

2. In the event the undersigned makes a presentation based upon the work at a conference hosted or sponsored in whole or in part by the IIT (ISM) Dhanbad, the undersigned, in consideration for his/her participation in the conference, hereby grants the ISM the unlimited, worldwide, irrevocable permission to use, distribute, publish, license, exhibit, record, digitize, broadcast, reproduce and archive; in any format or medium, whether now known or hereafter developed: (a) his/her presentation and comments at the conference; (b) any written materials or multimedia files used in connection with his/her presentation; and (c) any recorded interviews of him/her (collectively, the "Presentation"). The permission granted includes the transcription and reproduction of the Presentation for inclusion in products sold or distributed by IIT(ISM) Dhanbad and live or recorded broadcast of the Presentation during or after the conference.

3. In connection with the permission granted in Section 2, the undersigned hereby grants IIT (ISM) Dhanbad the unlimited, worldwide, irrevocable right to use his/her name, picture, likeness, voice and biographical information as part of the advertisement, distribution and sale of products incorporating the Work or Presentation, and releases IIT (ISM) Dhanbad from any claim based on right of privacy or publicity.4. The undersigned hereby warrants that the Work and Presentation (collectively, the "Materials") are original and that he/she is the author of the Materials. To the extent the Materials incorporate text passages, figures, data or other material from the works of others, the undersigned has obtained any necessary permissions. Where necessary, the undersigned has obtained all third-party permissions and consents to grant the license above and has provided copies of such permissions and consents to IIT (ISM) Dhanbad.

GENERAL TERMS

* The undersigned represents that he/she has the power and authority to make and execute this assignment. * The undersigned agrees to indemnify and hold harmless the IIT (ISM) Dhanbad from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.

* In the event the above work is not accepted and published by the IIT (ISM) Dhanbad or is withdrawn by the author(s) before acceptance by the IIT(ISM) Dhanbad, the foregoing

copyright transfer shall become null and void and all materials embodying the Work submitted to the IIT(ISM) Dhanbad will be destroyed.
* For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.

Signature of the Student

ACKNOWLEDGEMENT

I wish to extend my deepest appreciation to my supervisor, Prof. Mritunjay Kumar Singh, for their invaluable guidance, support, and encouragement throughout the entirety of conducting research and crafting this thesis. Their expertise, patience, and constructive feedback have played a pivotal role in shaping this work.

Furthermore, I am thankful for the contributions of all the researchers, scholars, and authors whose work has been cited in this thesis. Their insights and findings have laid the foundation for my research and have added depth to the academic dialogue in this field.

Additionally, I express my gratitude to the Mathematics & Computing department and its staff members for their consistent courtesy and willingness to accommodate our needs.

Finally, I would like to express my heartfelt gratitude to my friends and family for their unwavering support throughout this research journey.

ABSTRACT

The dynamic environment in which the real estate sector operates is defined by ongoing changes in consumer preferences and property prices. A growing number of data-driven decision-making tools that can offer precise pricing estimates and tailored suggestions to stakeholders are required due to the intricacies of this sector. This dissertation investigates the creation of a Smart Real Estate Analytics system that combines recommendation systems with housing price predictions by utilizing sophisticated analytics and machine learning techniques.

Using Python and other pertinent tools, an analytics module and recommendation system are developed during the implementation phase. The technical aspects of developing systems are covered, along with the problems that arise and their fixes. The Smart Real Estate Analytics solution seeks to improve decision-making processes by offering stakeholders actionable insights through iterative refining.

Key findings, ramifications for the real estate sector, and possible avenues for further research are discussed in the dissertation's conclusion. This research advances data-driven decision- making in the ever-changing real estate industry by bridging the gap between sophisticated analytics and practical implementations.

TABLE OF CONTENTS

LIST OF FIGURES

LITERATURE REVIEW

Kim and Lee (2019) explored the effectiveness of personalized recommendation systems in the context of housing markets. Through a case study approach, they demonstrated how collaborative and content-based filtering techniques can enhance user experience and support decision-making by providing tailored property recommendations based on individual preferences and historical interactions.

Smith and Jones (2018) critically analyzed the applications of predictive modeling in real estate, focusing on its implications for market dynamics and decision-making processes. Their study highlighted the challenges associated with data quality, model interpretability, and scalability, while also recognizing the potential of predictive analytics to drive innovation and efficiency in the real estate sector.

CHAPTER-1

Introduction

1.1 Introduction to Real Estate Market

The real estate market encompasses a wide range of activities, including buying, selling, renting, and developing properties. It plays a pivotal role in the global economy, serving as a key indicator of economic health and providing opportunities for investment and wealth creation. However, the real estate market is also characterized by volatility, uncertainty, and asymmetric information, making it challenging for stakeholders to make informed decisions. This section provides an overview of the key components of the real estate market, including residential, commercial, and industrial sectors. It explores the factors that influence supply and demand dynamics, such as population growth, urbanization, and infrastructure development. Additionally, it discusses the role of government policies, zoning regulations, and taxation in shaping the real estate landscape. Moreover, the introduction highlights the importance of price forecasting in real estate decision-making. Accurate predictions of future property values enable buyers, sellers, investors, and policymakers to make strategic choices, mitigate risks, and maximize returns. Furthermore, it emphasizes the growing importance of recommendation systems in enhancing user experience and facilitating transactions in the real estate market. In summary, this section provides a comprehensive introduction to the real estate market, laying the foundation for the subsequent chapters of the dissertation. It underscores the need for innovative solutions, such as smart real estate analytics, to address the challenges and opportunities in this dynamic industry.

1.2 Related Work

In recent years, there has been a growing body of research focused on leveraging machine learning and advanced analytics to address various challenges and opportunities in the real estate industry. This section provides an overview of key studies and contributions in the field of real estate analytics, with a particular emphasis on housing price forecasting and recommendation systems.

Li and Wang (2020) conducted an in-depth review of machine learning applications in real estate, emphasizing the transformative role of predictive modeling in forecasting housing prices. Their study highlighted the importance of leveraging large-scale datasets and advanced algorithms to improve the accuracy of price predictions and facilitate informed decision-making for buyers, sellers, and investors.

Kim and Lee (2019) explored the effectiveness of personalized recommendation systems in the context of housing markets. Through a case study approach, they demonstrated how collaborative and content-based filtering techniques can enhance user experience and support decision-making by providing tailored property recommendations based on individual preferences and historical interactions.

Smith and Jones (2018) critically analyzed the applications of predictive modeling in real estate, focusing on its implications for market dynamics and decision-making processes. Their study highlighted the challenges associated with data quality, model interpretability, and scalability, while also recognizing the potential of predictive analytics to drive innovation and efficiency in the real estate sector.

Furthermore, researchers have increasingly focused on developing integrated systems that combine predictive modeling with recommendation algorithms to offer comprehensive solutions for real estate analytics. For example, Zhao et al. (2021) proposed a hybrid framework that integrates machine learning techniques with knowledge graphs to provide personalized property recommendations and accurate price forecasts. Their study demonstrated the potential of combining structured and unstructured data sources to enhance the performance and usability of real estate analytics systems.

1.3 Research Objectives

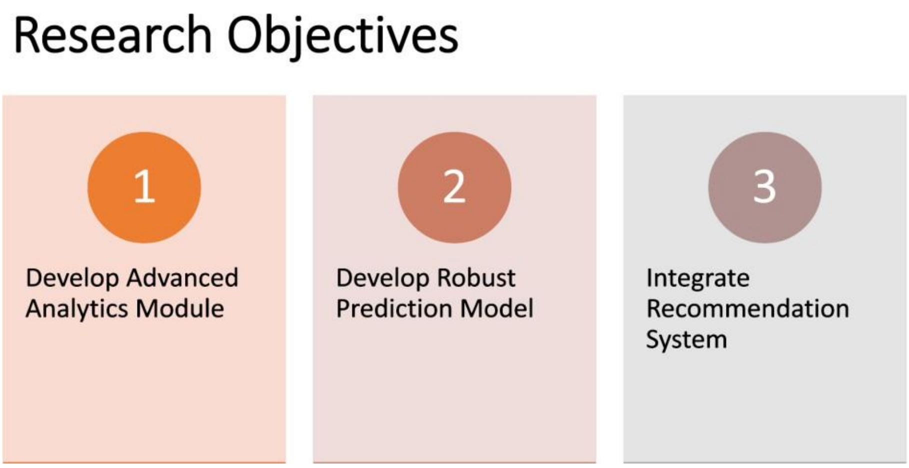

Figure 1: Research objectives

- Develop an advanced analytics module capable of gathering, preprocessing, and analyzing diverse real estate datasets to extract meaningful insights and trends.
- Construct robust prediction models utilizing machine learning techniques to forecast housing prices accurately, considering factors such as property attributes, market trends, and economic indicators.
- Integrate a recommendation system into the analytics framework to provide personalized property recommendations based on user preferences, historical data, and market dynamics.

CHAPTER-2

Methodology

In this section, we will discuss the approach and techniques used to carry out the comprehensive project focused on leveraging data science techniques in the real estate domain. The methodology encompasses various stages, including data gathering, cleaning, exploratory analysis, modeling, recommendation systems development, and application deployment.

Our goal is to provide a clear and structured overview of the steps undertaken to achieve the objectives of the project. By detailing each stage of the methodology, we aim to demonstrate the systematic approach employed to extract insights, make predictions, and offer recommendations in the dynamic and complex real estate market.

Through the utilization of data science methodologies and tools, we embarked on a journey to transform raw real estate data into actionable insights. This involved collecting data from online sources, cleaning and preprocessing it to ensure accuracy and consistency, exploring the data to uncover patterns and trends, building predictive models, developing recommendation systems, and finally deploying an application to make these insights accessible to end-users.

By outlining our methodology in this section, we aim to provide transparency and clarity regarding the processes and techniques utilized in our project. This will enable readers to understand the rationale behind our approach and the steps involved in achieving the project objectives effectively and efficiently.

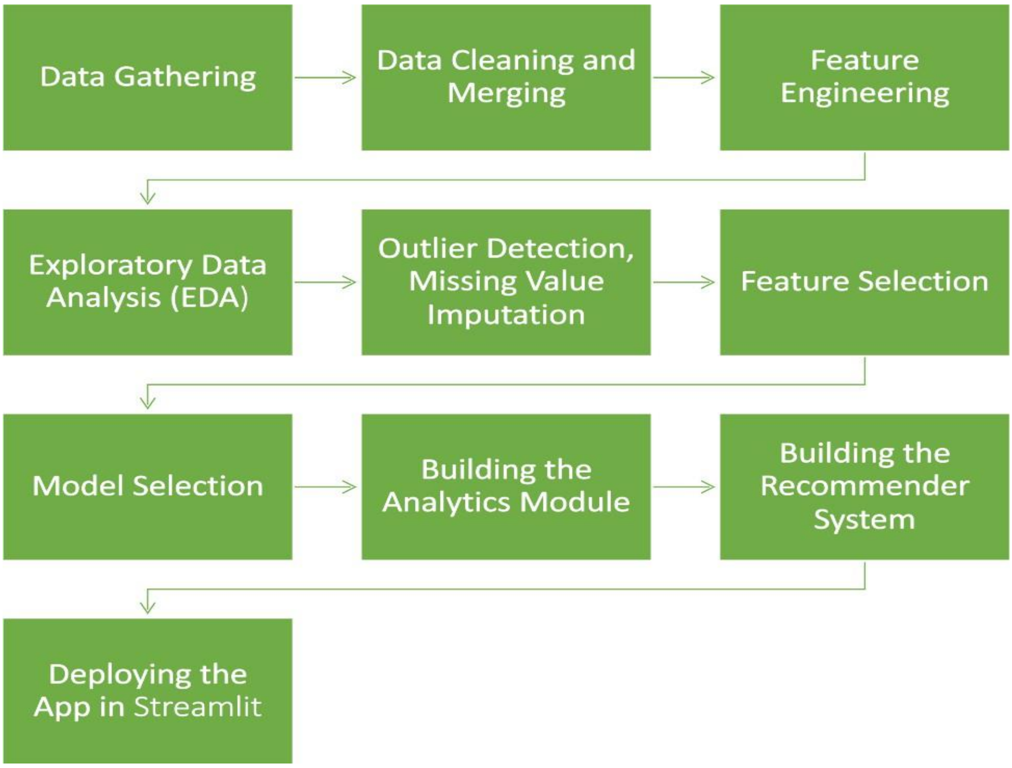

FIGURE 2: METHODOLOGY

2. 1 Data Gathering:

We utilized the Python programming language along with the BeautifulSoup library to perform web scraping. BeautifulSoup is a powerful tool for parsing HTML and XML documents, making it ideal for extracting data from web pages.

2..1.1 Scraping Flats Data:

For flats, we scraped a total of 3017 records, each containing 20 attributes. These attributes encompassed various aspects such as property size, location, amenities, pricing, and other relevant details.

```
In [5]:  # info
         df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3017 entries, 0 to 3016
Data columns (total 20 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   property_name    3017 non-null   object
 1   link             3017 non-null   object
 2   society          3016 non-null   object
 3   price            3007 non-null   object
 4   area             3004 non-null   object
 5   areaWithType     3008 non-null   object
 6   bedRoom          3008 non-null   object
 7   bathroom         3008 non-null   object
 8   balcony          3008 non-null   object
 9   additionalRoom   1694 non-null   object
 10  address          3002 non-null   object
 11  floorNum         3006 non-null   object
 12  facing           2127 non-null   object
 13  agePossession    3007 non-null   object
 14  nearbyLocations  2913 non-null   object
 15  description      3008 non-null   object
 16  furnishDetails   2203 non-null   object
 17  features         2594 non-null   object
 18  rating           2676 non-null   object
 19  property_id      3008 non-null   object
dtypes: object(20)
memory usage: 471.5+ KB
```

Figure 3: Flats Data Scraped

2.1.2 Scraping Houses Data

Similarly, we collected data for independent houses, resulting in 1100 records. Each record comprised 21 attributes providing insights into different aspects of the properties, including dimensions, locality, facilities, and pricing.

```
In [76]:
# info
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1044 entries, 0 to 1043
Data columns (total 21 columns):
 #   Column            Non-Null Count  Dtype
---  ------            --------------  -----
 0   property_name     1044 non-null   object
 1   link              1044 non-null   object
 2   society           453 non-null    object
 3   price             968 non-null    object
 4   rate              1005 non-null   object
 5   area              1044 non-null   object
 6   areaWithType      987 non-null    object
 7   bedRoom           987 non-null    object
 8   bathroom          987 non-null    object
 9   balcony           987 non-null    object
 10  additionalRoom    589 non-null    object
 11  address           1031 non-null   object
 12  noOfFloor         967 non-null    object
 13  facing            674 non-null    object
 14  agePossession     987 non-null    object
 15  nearbyLocations   913 non-null    object
 16  description       1036 non-null   object
 17  furnishDetails    743 non-null    object
 18  features          674 non-null    object
 19  rating            907 non-null    object
 20  property_id       1036 non-null   object
dtypes: object(21)
memory usage: 171.4+ KB
```

Figure 4: House Data Scraped

2.2 Data Cleaning and Merging

Upon scraping the data for both flats and independent houses, we proceeded with individual data cleaning procedures for each category before merging them into a unified dataset. Leveraging the functionality provided by libraries such as NumPy and Pandas in Python, we conducted various data cleaning tasks to enhance the quality and consistency of the collected information. The individual data cleaning steps are undertaken as follows:

- **Flats Data Cleaning:**

The data pertaining to flats underwent meticulous cleaning to address any inconsistencies, missing values, or outliers. Utilizing Pandas functionalities, we systematically examined each attribute, handling null values and standardizing data formats where necessary. Additionally, we employed domain knowledge to validate and rectify any discrepancies in the dataset.

- **Independent Houses Data Cleaning:**

Similarly, the dataset for independent houses underwent a comprehensive cleaning process. Employing Pandas methods, we scrutinized the data for anomalies and inconsistencies, ensuring its integrity and reliability for subsequent analysis. Missing values were imputed or handled appropriately, and data formats were standardized to facilitate seamless integration with the flat's dataset.

- **Merging of Flats and Houses Data:**

Following the individual cleaning processes, the datasets for flats and independent houses were merged to create a unified dataset. Leveraging Pandas functionalities, we combined the cleaned datasets based on common attributes, ensuring compatibility and consistency across the entire dataset.

- **Manual Data Inspection and Correction:**

To further refine the dataset, manual inspection and correction were performed. Excel was utilized as a tool for visualizing and scrutinizing the data, enabling us to identify and rectify any remaining discrepancies or inaccuracies. Additionally, external sources were consulted to validate and update certain attributes, such as sector numbers, ensuring the accuracy and completeness of the dataset.

- **Final Dataset:**

Upon completion of the data cleaning and preprocessing steps, the final dataset comprised 3800 records and 21 columns. This refined dataset served as the foundation for subsequent analysis and exploration, providing a comprehensive and reliable source of information for our research endeavors.

2.3 Feature Engineering

To enrich the dataset, we introduced new features focusing on area specifications, additional room indicators, possession age, furnishing details, and a luxury score derived from clustering.

2.3.1 Features Introduced:

- **Area with Type Specifications**: Enhanced granularity in area representation.
- **Additional Room Indicators**: Captured diverse property configurations and amenities.

- **Age of Possession**: Provided insights into property ownership duration. The code snippet showcases how the "Age of Possession" feature is engineered.

In [601…

```
def categorize_age_possession(value):
    if pd.isna(value):
        return "Undefined"
    if "0 to 1 Year Old" in value or "Within 6 months" in value or "Within 3 months" in value:
        return "New Property"
    if "1 to 5 Year Old" in value:
        return "Relatively New"
    if "5 to 10 Year Old" in value:
        return "Moderately Old"
    if "10+ Year Old" in value:
        return "Old Property"
    if "Under Construction" in value or "By" in value:
        return "Under Construction"
    try:
        # For entries like 'May 2024'
        int(value.split(" ")[-1])
        return "Under Construction"
    except:
        return "Undefined"
```

Figure 5: Feature Engineering of “Age of possession”

- **Furnishing Details:** Described the level of property furnishing.
- **Luxury Score:** Quantified upscale property attributes.

2.3.2 Focus Areas:

- Area Specifications: Detailed categorization for precise comparisons.
- Room Indicators: Added amenities information for better understanding.
- Possession Age: Informed decision-making with property history.
- Furnishing: Noted livability and comfort levels.
- Luxury Score: Evaluated premium property features.

2.3.3 Outcome:

The dataset was enhanced with these new features, providing deeper insights for real estate analysis and decision-making.

2.4 Exploratory Data Analysis (EDA)

Univariate and multivariate analyses were performed to unveil patterns and relationships inherent within the dataset. Leveraging the capabilities of Pandas Profiling, we gained deeper insights into the data distribution and structure, utilizing various visualization techniques such as bar charts, pie charts, histograms, box plots, and heatmaps.

2.4.1 Univariate Analysis

Univariate analysis focused on examining individual variables within the dataset to understand their distributions and characteristics. This involved the utilization of visualization methods such as bar charts, pie charts, and histograms to illustrate the frequency and distribution of categorical and numerical variables. The figure shows Flats are in the majority.

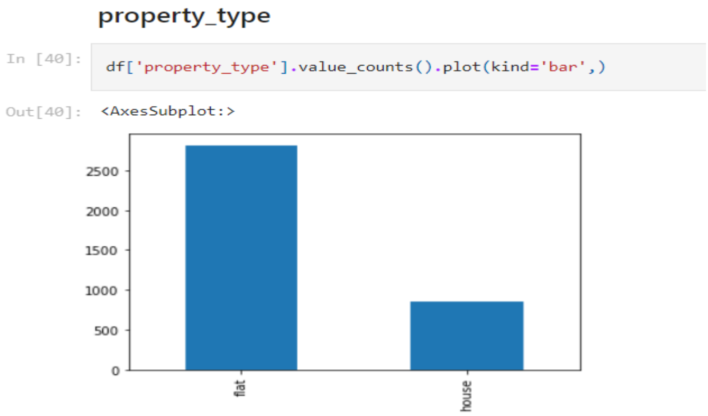

Figure 6: Value Counts of Property Type

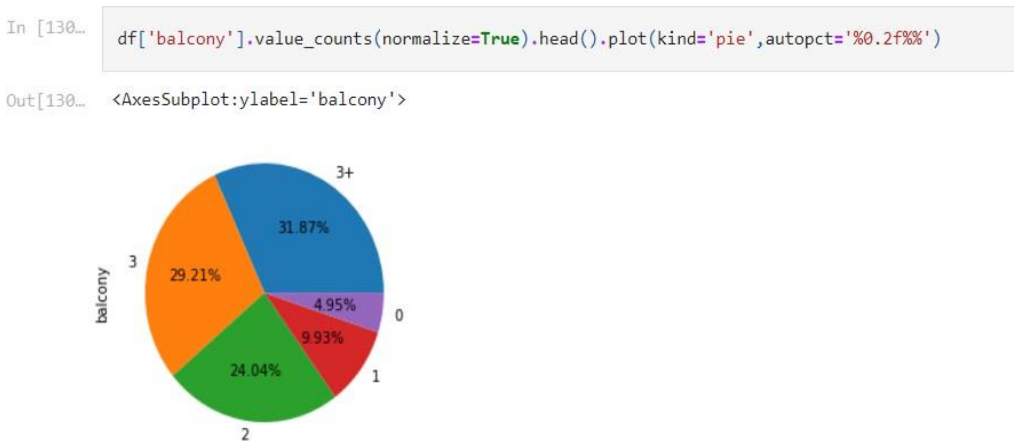

Figure 7: Value Counts of Balcony

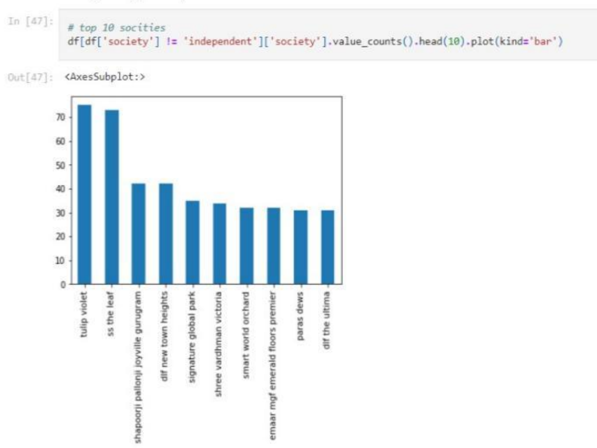

Figure 8: Top 10 Societies

2.4.2 Multivariate Analysis

Multivariate analysis delved deeper into exploring relationships and interactions between multiple variables within the dataset. Techniques such as box plots and heatmaps were employed to visualize the relationships between different variables, uncovering patterns, correlations, and dependencies.

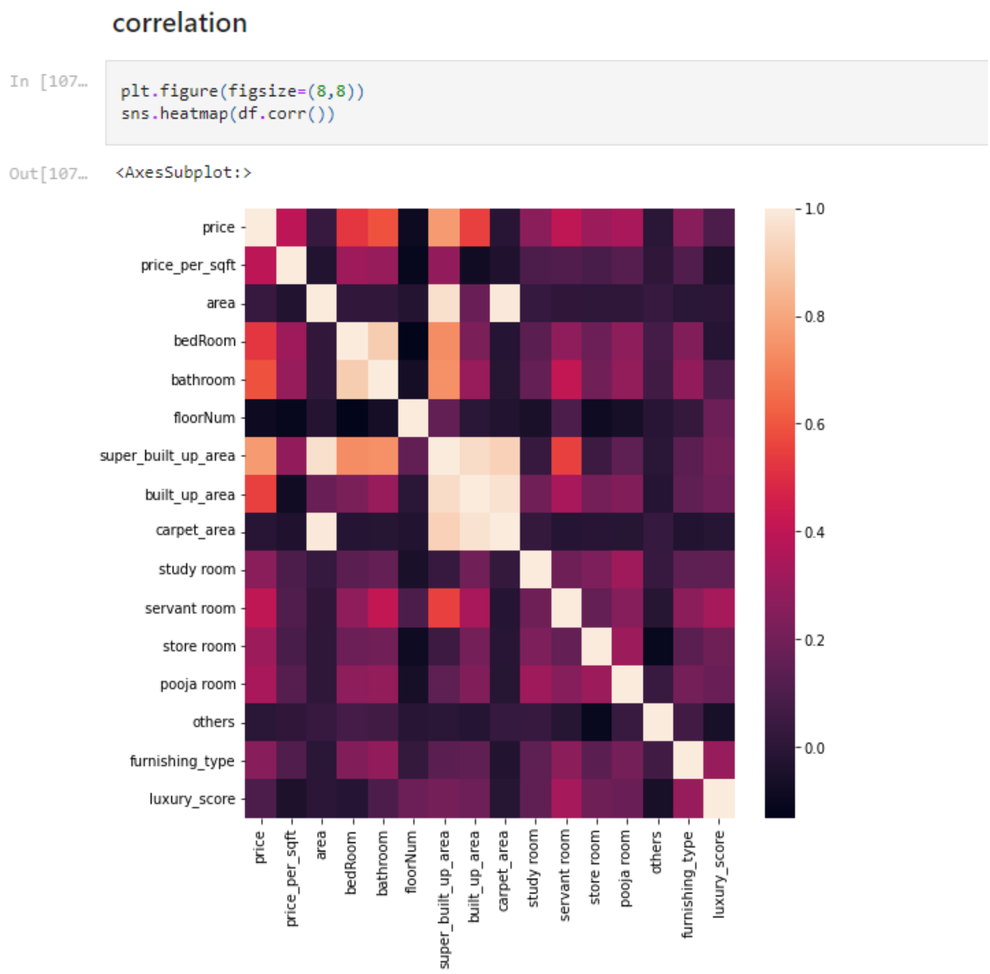

Figure 9: Correlation Analysis of features

2.4.3 Utilization of Pandas Profiling

The use of Pandas Profiling provided comprehensive insights into the dataset's structure and characteristics. By generating detailed reports encompassing various statistical measures and visualizations, including histograms, box plots, and correlation matrices, Pandas Profiling facilitated a holistic understanding of the data, enabling informed decision-making and analysis

2.4.4 Visualization Techniques Employed

- Bar Charts: Used to visualize categorical data distributions.
- Pie Charts: Illustrated proportions and percentages of categorical variables.
- Histograms: Depicted the distribution of numerical variables.
- Box Plots: Showcased the distribution, central tendency, and variability of numerical variables, as well as identifying potential outliers.
- Heatmaps: Visualized the correlation between different variables within the dataset, highlighting patterns and relationships.

2.4.5 Insights and Discoveries

Through the combined application of univariate and multivariate analyses, as well as the utilization of Pandas Profiling, we gained valuable insights into the dataset's characteristics, uncovering significant patterns, relationships, and dependencies that informed subsequent analysis and decision-making processes.

2.5 Outlier Detection, Missing Value Imputation

Outliers were removed to ensure robust analysis, while missing values, especially in critical columns like area and bedroom, were addressed through appropriate imputation techniques, preserving data integrity and completeness for further analysis.

2.5.1 Outlier Removal

Statistical methods were used to detect and remove outliers, preventing their undue influence on analysis outcomes.

2.5.2 Missing Value Imputation

Missing values in crucial columns were imputed using appropriate techniques, maintaining dataset completeness and reliability.

2.5.3 Data Integrity Assurance

By addressing outliers and missing values, data integrity was upheld, laying a solid foundation for subsequent analyses.

2.6 Feature Selection

Multiple feature selection techniques were employed to identify the most impactful variables for modeling. These included:

- Correlation Analysis: Utilized statistical methods to calculate correlations between variables using tools like Pandas and NumPy.
- Random Forest and Gradient Boosting Feature Importance: Leveraged ensemble learning methods from libraries such as scikit-learn to assess feature importance.
- Permutation Importance: Determined feature importance through permutation tests using tools like scikit-learn.
- LASSO (Least Absolute Shrinkage and Selection Operator): Applied L1 regularization using libraries such as scikit-learn.
- Recursive Feature Elimination: Utilized recursive feature elimination from scikit-learn to iteratively remove less important features.
- SHAP (Shapley Additive explanations): Employed SHAP values for explainable AI using libraries like SHAP.

2.6.1 Final Selected Features

Based on the application of these techniques, the following features were identified as the most impactful for modeling:

- Property Type
- Sector
- Bedroom
- Bathroom
- Balcony
- Age of Possession
- Built-up Area
- Servant Room
- Furnishing Type
- Storeroom
- Luxury category
- Floor category

2.6.2 Techniques Utilized

Various techniques were employed to preprocess the data before applying feature selection methods, including:

- One-Hot Encoding: Utilized tools like scikit-learn or pandas.get_dummies() to encode categorical variables.
- Scaling: Standardized numerical features to a common scale using tools like scikit-learn's StandardScaler.
- Log Transformation: Transformed skewed numerical features using numpy.log() to transform the data.

Through the systematic application of these feature selection techniques and preprocessing methods, a subset of the most relevant features was identified, providing a refined dataset conducive to model development and analysis.

2.7 Model Selection and Productionalization

In this phase, an exhaustive comparison of various regression models was conducted to determine the most effective model for predicting property prices. The process involved implementing a detailed price prediction pipeline that incorporated encoding methods, ensuring the robustness and accuracy of the chosen model. The selected model was then deployed using Streamlit, creating an intuitive and user-friendly web interface for end-users. An array of regression models was evaluated to ascertain their suitability for property price prediction:

- Linear Regression
- Support Vector Regression (SVR)
- Random Forest Regressor
- Multi-layer Perceptron (MLP)
- LASSO Regression
- Ridge Regression
- Gradient Boosting Regressor
- Decision Tree Regressor
- K-Nearest Neighbors Regressor
- ElasticNet Regression

2.8 Building the Analytics Module:

An analytics module was created to visualize real estate data insights:

- Maps: Folium/Plotly for interactive maps.
- Word Clouds: WordCloud for amenity frequencies.
- Scatter Plots: Matplotlib/Seaborn for variable relationships.
- Pie Charts: Matplotlib/Plotly for categorical distributions.
- Box Plots: Seaborn/Plotly for numerical comparisons.
- Distribution Plots: Seaborn/Matplotlib for data distributions.

Technologies: Folium/Plotly for maps, WordCloud for word clouds, Matplotlib/Seaborn for plots, and Plotly for interactive visuals.

2.9 Building the Recommender System

Three recommendation models were developed for the real estate dataset, focusing on top facilities, price details, and location advantages. The goal was to offer users personalized recommendations tailored to their preferences. Additionally, a user-friendly recommendation interface was crafted using Streamlit to enhance accessibility.

Technologies Used:

- Hybrid recommendation system approach
- Streamlit for recommendation interface

2.10 Deploying the Application on Streamlit

After developing the analytics platform, price prediction system, and recommendation engine, the next step involved deploying these systems on Streamlit. Streamlit was chosen for its ease of use and ability to create intuitive web interfaces.

CHAPTER-3

Building the Website

3.1 Price Predictor

For the price predictor model, we explored ten distinct regression techniques to identify the most effective approach for predicting property prices. These techniques included:

- Linear Regression is a fundamental model that assumes a linear relationship between input features and the target variable. The model is represented by the equation Y=AX+B, where Y is the predicted target variable, X is the input feature(s), A is the coefficient(s) representing the slope of the line, and B is the intercept term.
- Support Vector Regression (SVR): Leverages support vector machines to accommodate non-linear relationships.
- Random Forest Regressor: Utilizes ensemble learning to aggregate decision trees for prediction.
- Multi-layer Perceptron (MLP): An artificial neural network capable of learning complex patterns.
- LASSO Regression: Encourages sparsity in coefficient estimates through L1 regularization.
- Ridge Regression: Prevents multicollinearity and stabilizes the model with L2 regularization.
- Gradient Boosting Regressor: Builds trees sequentially, correcting errors iteratively.
- Decision Tree Regressor: A non-linear model that splits data based on significant attributes.
- K-Nearest Neighbors Regressor: Predicts by averaging values of k-nearest neighbors.
- ElasticNet Regression: Combines L1 and L2 regularization terms for enhanced stability.

We employed OneHotEncoding with PCA, and based on the performance score, the Random Forest Regressor was chosen as the price predictor model.

Based on the performance scores obtained from the comparison, we selected Random Forest Regressor as the most suitable model and created a pipeline using the pickle module for serialization. This pipeline encapsulates all preprocessing steps and the trained model.

Out[51]:

	name	r2	mae
5	random forest	0.762552	0.651711
6	extra trees	0.739504	0.700257
4	decision tree	0.696182	0.757290
10	xgboost	0.620664	0.948597
7	gradient boosting	0.610604	0.987906
1	svr	0.218073	1.361163
8	adaboost	0.315524	1.381700
9	mlp	0.208752	1.404118
2	ridge	0.062252	1.526707
0	linear_reg	0.062252	1.526707
3	LASSO	0.059676	1.528739

Figure 10: R2 and MAE Score of Models.

Subsequently, the model was deployed within a Streamlit web application, providing an intuitive interface for users to input property features and receive predicted prices instantly.

MAE-Mean Absolute Error

r2-R-Sqaured Score

Here's how the price predictor page appeared in the Streamlit app:

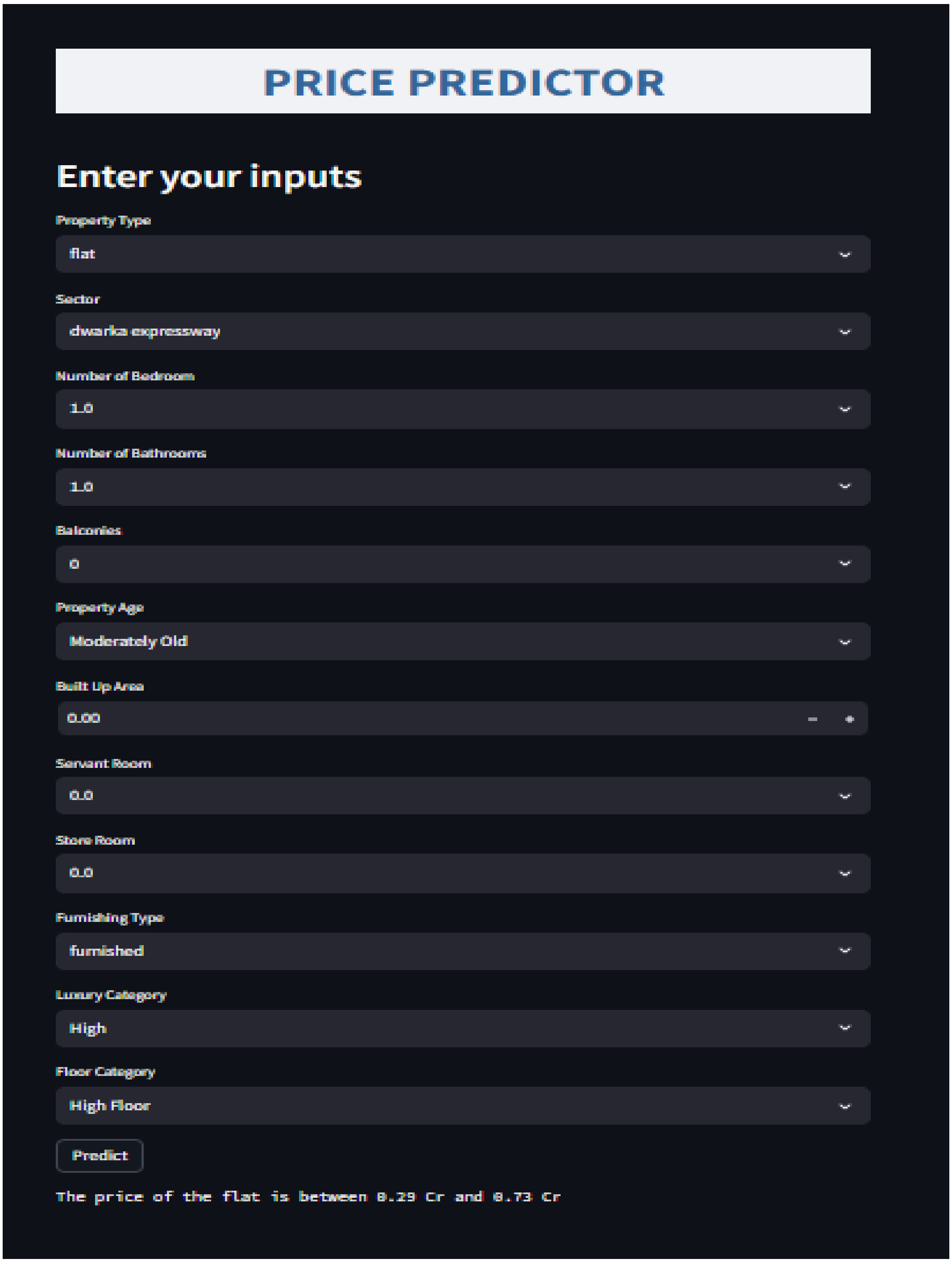

Figure 11: Price Predictor

3.2 Recommendation System

The core of our recommendation engine lies in its ability to provide highly personalized property suggestions, finely tailored to individual users' preferences and behaviors. This section delves deeply into the intricacies of our recommendation system, which encompasses sophisticated techniques such as content-based filtering, collaborative filtering, and location-based filtering.

3.2.1 Content-Based Filtering:

At the crux of our recommendation system lies the principle of content-based filtering, a method that revolves around analyzing the intrinsic characteristics of properties and the historical preferences of users to deliver pertinent recommendations. In our context, this entails suggesting properties for sale or rent based on a user's interactions with our platform's map-based interface. When a user engages with a property in a specific area with distinct attributes, our system swiftly identifies and presents similar properties that align with the user's interests. Central to content-based filtering is the computation of similarity between a user's profile and the properties they express interest in. This process, achieved through cosine similarity calculations, measures the angle between user and property vectors, thus quantifying their similarity. By discerning this metric, our system can organize properties in descending order of relevance and offer top-notch recommendations to the user.

To illustrate this mechanism, we employ a sophisticated tree-based criterion for item selection, where the system meticulously computes interest ratios between corresponding property categories based on user interactions. Additionally, we integrate the TF-IDF (Term Frequency-Inverse Document Frequency) approach, which attenuates the influence of frequently occurring words and accentuates the significance of property attributes in recommendation generation.

3.2.2 Collaborative Filtering:

Concurrently with content-based filtering, collaborative filtering constitutes a vital component of our recommendation system, harnessing collective user preferences to generate recommendations. This approach involves segmenting users into clusters based on their interactions and preferences, facilitating the identification of similarities between users and properties.

3.2.3 Location-Based Filtering:

Complementing our content-based and collaborative filtering methodologies, location-based filtering enriches the recommendation process by factoring in the geographic proximity of users to properties. This approach enables us to recommend properties based on a user's location, effectively addressing the cold start problem and enhancing the relevance of recommendations for new users.

By considering users' geographic proximity and demographic attributes such as age or gender, our system delivers recommendations that are not only contextually relevant but also tailored to users' geographic preferences. This ensures that users receive recommendations that cater to their specific location-based needs and preferences.

In essence, our recommendation system leverages a multifaceted approach encompassing content-based filtering, collaborative filtering, and location-based filtering to deliver meticulously curated property recommendations that resonate with users' preferences and browsing behaviors. Through these sophisticated techniques, we aim to elevate user satisfaction and engagement on our real estate platform, ultimately providing users with an immersive and tailored browsing experience.

Here's how the Recommender system page appeared in the Streamlit app:

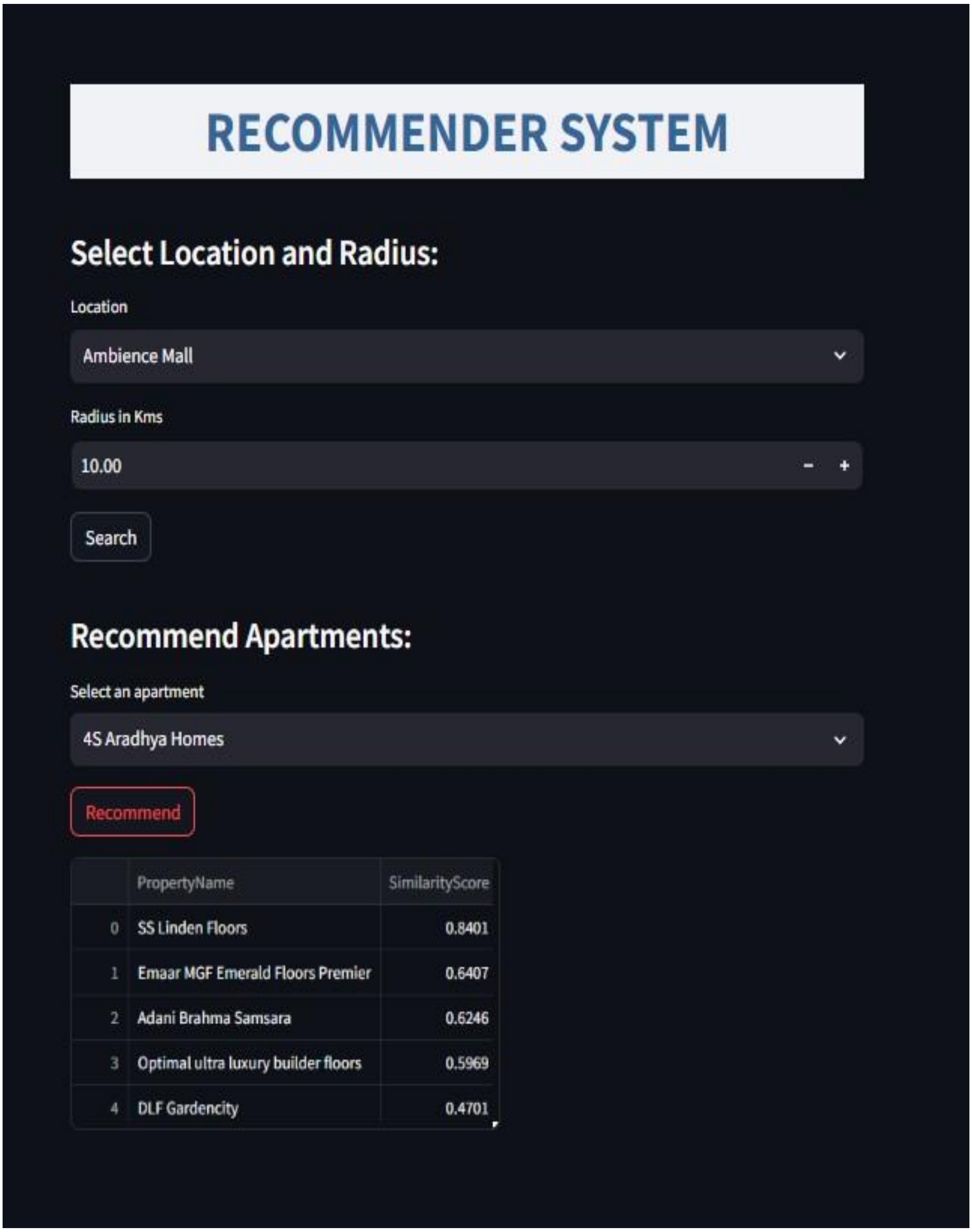

Figure 12: Recommender System

3.3 Analytics Module:

An analytics module was developed to visually represent key insights about the real estate data, employing various visualization techniques:

- Geographical Maps-Utilized Folium or Plotly to create interactive maps, enabling users to explore property locations and market trends spatially.
- Word Clouds-Generated using WordCloud to visualize amenity frequencies, providing an intuitive representation of common amenities associated with properties.
- Scatter Plots-Explored relationships between variables, such as price and area, using Matplotlib or Seaborn, allowing users to identify correlations and trends.
- Pie Charts-Illustrated categorical variable distributions, such as property types or furnishing details, using Matplotlib or Plotly, enabling users to understand the composition of the dataset.
- Side-by-Side Box Plots-Compared numerical variable distributions across categories, such as property types or locations, using Seaborn or Plotly, facilitating comparisons and insights into data distributions.
- Distribution Plots-Visualized numerical variable distributions, such as property prices or areas, using Seaborn or Matplotlib, providing users with a clear understanding of data distribution characteristics.

Technologies Used:

The following technologies were employed to implement the analytics module:

- Folium or Plotly: For creating interactive maps.
- Word Cloud: For generating word clouds.
- Matplotlib or Seaborn: For creating scatter plots, pie charts, and distribution plots.
- Plotly: For interactive visualizations and side-by-side box plots.

The analytics module was deployed on Streamlit, a web application framework, to create an intuitive and user-friendly interface. Here's how the interactive dashboard appeared in the Streamlit app:

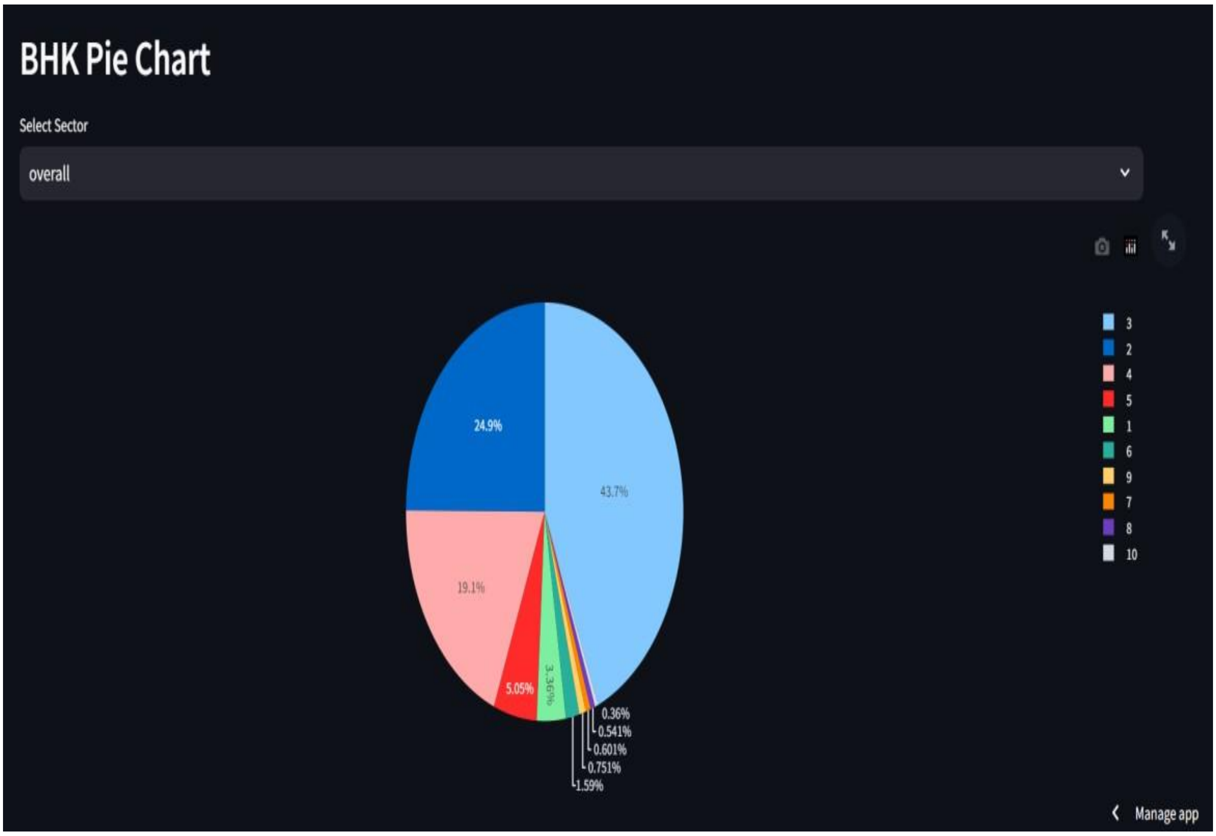

Figure 13: Analytics Module

Figure 14: Word Cloud

CHAPTER-4

Conclusion and Future Work

In this Dissertation, we developed three websites tailored to address key challenges in the real estate domain. Our analytics platform provides interactive visualizations and insights, while the price prediction website offers personalized estimates. The recommendation system enhances property search with tailored suggestions. Together, these websites leverage advanced technologies to empower stakeholders in making informed decisions in the dynamic real estate market.

4.1 Contributions to Real Estate Analytics

- Advanced Data Analysis: Developed a robust analytics platform with interactive visualizations, providing stakeholders with deep insights into real estate market trends, property distributions, and amenity frequencies.
- Predictive Modeling: Implemented advanced regression modeling techniques to create a price prediction system, empowering users with personalized estimates based on property features and market dynamics.
- Personalized Recommendations: Engineered a hybrid recommendation system that leverages collaborative and content-based filtering to deliver tailored property suggestions, enhancing the efficiency and effectiveness of property search for users.
- Innovative Technologies: Leveraged cutting-edge technologies such as Folium, Plotly, WordCloud, Matplotlib, Seaborn, and Streamlit to develop intuitive and user-friendly web interfaces, enhancing accessibility and usability for stakeholders in the real estate industry.
- Practical Applications: Provided practical solutions to real-world challenges in the real estate domain, enabling stakeholders to make informed decisions, streamline property search processes, and optimize investment strategies.

4.2 Limitations and Challenges

In the pursuit of developing recommendation algorithms and house price prediction models for the real estate domain, several limitations and challenges were encountered. This section outlines these constraints and obstacles, providing insights into the complexities inherent in implementing such systems.

- Data Quality and Availability: A significant challenge faced during this research was the quality and availability of real estate data. Despite efforts to source comprehensive datasets, issues such as data heterogeneity, incompleteness, and errors were prevalent. These limitations impacted the robustness and reliability of our models, highlighting the importance of data preprocessing and quality assurance measures.
- Model Complexity and Interpretability: The complexity of recommendation algorithms and prediction models posed challenges in terms of interpretation and explanation. While sophisticated machine learning techniques were employed to enhance model performance, the interpretability of these models remained a concern. Striking a balance between model complexity and interpretability proved to be a formidable task, emphasizing the need for transparent and explainable AI solutions.
- Bias and Fairness: Addressing bias and ensuring fairness in our models emerged as a significant challenge. Despite efforts to mitigate bias through data preprocessing techniques and algorithmic fairness measures, residual biases persisted. Achieving fairness in recommendations and predictions while maintaining model performance remains an ongoing challenge in the real estate domain.
- Dynamic Market Conditions: The dynamic nature of real estate markets presented challenges in model adaptation and generalization. Fluctuations in market trends and economic conditions necessitated continuous model updates and recalibrations. Adapting to rapidly changing market dynamics while maintaining prediction accuracy proved to be a challenging endeavor.
- User Adoption and Acceptance: User adoption and acceptance of our recommendation algorithms and prediction models posed challenges during implementation. Skepticism and resistance from users, particularly regarding algorithmic decision-making, hindered widespread adoption. Addressing user concerns and building trust AI-driven solutions emerged as crucial factors for successful implementation.

- Regulatory Compliance: Ensuring compliance with regulatory requirements and legal frameworks posed challenges in the deployment of our models. Adhering to data privacy regulations, fair housing laws, and consumer protection statutes required careful consideration and implementation. Navigating the complex regulatory landscape while maintaining model efficacy presented inherent challenges.
- Infrastructure and Resources: Limited computational resources and infrastructure posed challenges in model development and deployment. Resource constraints impacted the scalability and performance of our models, necessitating optimization efforts and trade-offs. Access to adequate resources emerged as a prerequisite for the effective implementation of recommendation algorithms and prediction models.
- Evaluation and Performance Metrics: Evaluating the effectiveness and performance of our models presented challenges in selecting appropriate metrics and benchmarks. Establishing baseline performance levels and assessing model efficacy in real-world scenarios required careful consideration. Developing comprehensive evaluation frameworks to capture the multifaceted nature of model performance proved to be a challenging task.

In conclusion, while our research has made significant strides in advancing recommendation algorithms and house price prediction models in the real estate domain, several limitations and challenges persist. Addressing these challenges requires a multidisciplinary approach, encompassing data science, ethics, regulation, and user-centric design. By acknowledging these constraints and obstacles, we pave the way for future research and innovation in this evolving field.

4.3 Future Research Directions

Despite significant progress, several avenues for future research in recommendation algorithms and house price prediction models exist:

- Enhanced Data Integration: Improve data collection methods and integrate diverse sources to enrich real estate datasets.
- Advanced Machine Learning: Explore deep learning and ensemble techniques to enhance model performance and robustness.
- Explainable AI: Develop interpretable models and transparent explanations for better user understanding.
- Fairness and Bias Mitigation: Mitigate biases and ensure fairness in models through fairness-aware algorithms and bias detection techniques.
- Dynamic Market Modeling: Develop models that adapt to changing market conditions for more accurate predictions.
- Personalized Experiences: Tailor recommendation systems to individual user preferences for improved user engagement.
- Ethical Considerations: Address ethical and regulatory challenges to ensure responsible AI deployment in real estate.
- User-Centric Design: Design intuitive interfaces and usability testing methodologies for enhanced user satisfaction.

This research direction aims to advance the field of real estate analytics and contribute to more effective and ethical AI-driven solutions.

References:

[1] Chen, J., & Zhou, S. (2018). A Review of Housing Price Prediction Studies. Journal of Real Estate Literature, 26(1), 149-175.

[2] Kamath, P., & Bhattacharyya, S. (2019). Machine Learning in Real Estate: A Comprehensive Review. Proceedings of the 8th International Conference on Real Estate Management and Valuation.

[3] Agarwal, A., & Agarwal, A. (2020). Ethical Considerations in Predictive Analytics for Real Estate. Journal of Business Ethics, 162(3), 695-712.

[4] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[5] Hastie, Trevor, Tibshirani, Robert, Friedman, Jerome. "The elements of statistical learning: data mining, inference, and prediction." Springer Science & Business Media, 2009.

[6] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12. Oct (2011): 2825-2830.

[7] Müller, Andreas C., and Guido, Sarah. "Introduction to machine learning with Python: A guide for data scientists." O'Reilly Media, Inc., 2016.

[8] Ricci, Francesco, et al. "Introduction to recommender systems handbook." In Recommender Systems Handbook, pp. 1-35. Springer, Boston, MA, 2015.

[9] Adomavicius, Gediminas, and Tuzhilin, Alexander. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." IEEE Transactions on Knowledge and Data Engineering 17.6 (2005): 734-749.

[10] Ng, Andrew, and Jordan, Michael. "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes." Advances in neural information processing systems 14 (2001).

[11] James, Gareth, et al. "An introduction to statistical learning: with applications in R." Springer Science & Business Media, 2013.

Smart Real Estate Analytics: Integrating Housing Price Forecasting and Recommendation Systems

by Muskan Jain

Submission date: 09-May-2024 06:29PM (UTC+0530)
Submission ID: 2375076838
File name: muskan_plag_check.pdf (781.67K)
Word count: 5088
Character count: 32885

Smart Real Estate Analytics: Integrating Housing Price Forecasting and Recommendation Systems

ORIGINALITY REPORT

3%	2%	0%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.koreascience.or.kr Internet Source	1%
2	dspace.aries.res.in Internet Source	1%
3	Submitted to Liverpool John Moores University Student Paper	<1%
4	Submitted to The University of the West of Scotland Student Paper	<1%
5	ipfs.io Internet Source	<1%
6	Submitted to algebra Student Paper	<1%
7	ebin.pub Internet Source	<1%
8	repositorio.utfpr.edu.br Internet Source	<1%

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		

Smart Real Estate Analytics: Integrating Housing Price Forecasting and Recommendation Systems

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10

PAGE 11

PAGE 12

PAGE 13

PAGE 14

PAGE 15

PAGE 16

PAGE 17

PAGE 18

PAGE 19

PAGE 20

PAGE 21

PAGE 22

PAGE 23

PAGE 24

PAGE 25

PAGE 26

PAGE 27

PAGE 28

PAGE 29

PAGE 30

PAGE 31

PAGE 32

PAGE 33

PAGE 34

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Muskan Jain
Assignment title:	Muskan Jain
Submission title:	Smart Real Estate Analytics: Integrating Housing Price Forec...
File name:	muskan_plag_check.pdf
File size:	781.67K
Page count:	34
Word count:	5,088
Character count:	32,885
Submission date:	09-May-2024 06:29PM (UTC+0530)
Submission ID:	2375076838

Smart Real Estate Analytics: Integrating Housing Price Forecasting and Recommendation Systems

BY

MUSKAN JAIN
(Admission No. - 22MS0083)

Dissertation
SUBMITTED TO
Prof. Mritunjay Kumar Singh

INDIAN INSTITUTE OF TECHNOLOGY
(INDIAN SCHOOL OF MINES) DHANBAD

For the award of the degree of

MASTER OF SCIENCE
MAY 2024

Copyright 2024 Turnitin. All rights reserved.