TITLE: REAL ESTATE PROJECT

Rituj Upadhyay (2020570)
Hardik Sachdeva (2022193)

PROBLEM STATEMENT AND MOTIVATION

In Gurgaon, a rapidly expanding urban center in India characterized by dynamic economic growth and population influx, there exists a critical need for the development of residential properties that cater to the diverse needs and preferences of its inhabitants. With rising urbanization and limited availability of land, there is a pressing demand for innovative real estate solutions that not only provide comfortable living spaces but also integrate sustainable practices, modern amenities, and connectivity to urban infrastructure.

The real estate industry is characterized by complexity, uncertainty, and inefficiency, making it challenging for stakeholders to navigate and make informed decisions. This project aims to develop a comprehensive real estate data science application that includes modules for price prediction, insights generation, and personalized recommendations. By leveraging advanced data analysis techniques and machine learning algorithms, the application seeks to provide users with accurate price predictions, valuable insights, and personalized recommendations tailored to their preferences and requirements. The project unfolds through various stages, covering data gathering, cleaning, EDA, modeling.

LITERATURE REVIEW

Data science and machine learning techniques are increasingly applied in the real estate domain to enhance decision-making processes and improve user experiences. Various studies have investigated the application of machine learning algorithms for tasks such as price prediction, property valuation, and demand forecasting. For example, Smith et al. (2018)

developed a machine learning model to predict property prices based on features such as location, size, and amenities. Researchers have employed various methodologies, including hedonic pricing models, spatial analysis, and time-series forecasting, to analyze real estate data. For instance, Li and Brown (2019) conducted a comprehensive analysis of factors influencing property prices, including neighborhood characteristics, school quality, and transportation accessibility. Their study highlighted the importance of incorporating spatial and temporal dimensions into real estate market analysis to capture local variations and trends. Wang et al. (2020) proposed a hybrid recommendation system that combines user preferences with property features to deliver personalized recommendations. Their research demonstrated the effectiveness of incorporating contextual information and user feedback in improving recommendation accuracy and user satisfaction.

Data science applications in real estate offer opportunities to improve decision-making processes and user experiences, yet challenges persist regarding data quality and scalability, necessitating further research and innovation

DATASETS

The project commenced with the collection of real estate data, which was srapped from 99acres website. Similar dataset from other property listing websites were also explored, ensuring a diverse and representative dataset.

* To prepare the dataset for analysis, a meticulous data cleaning process was undertaken , handling missing values and ensuring consistency. The data was then merged , bringing together information on houses and flats into unified database.

The FLAT dataset has 3018 rows and 17 columns i.e features and HOUSES dataset has 1045 rows and 18 columns. After all the process of data cleaning we get 3555 rows and 13 reliable and important features for the unified dataset.

METHODOLOGY

The dataset underwent feature engineering to enhance its richness and informativeness. New features, such as additional room indicators , area with type specification, age of possession, furnish details and luxury score, we introduced to provide a more detailed representation of the properties.

Univariate and multivariate analysis were conducted to uncover pattern and relationships within data. The use of Pandas Profiling facilitated a deeper understanding of data distribution and structure.

Outliers were identified and removed to ensure the robustness of subsequent analysis. Missing values particularly in critical columns like area and bedroom, were addressed using appropriate imputation technique.

Multiple Feature Selection techniques were employed to identify the most impactful variables for modelling . These included correlation analysis, random forest and gradient feature importance, recursive feature elimination etc.

In model selection phase, an exhaustive comparison of various model 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 different models considered in the comparison included:

1.Linear Regression

2.Support Vector Regressor

3.Random Forest

4.Multi Layer Perceptron

5.Lasso Regression

6.Ridge Regression

7.Gradient boosting

8.Decision Tree

The comparison involves assessing the performance of each model on evaluation metrics , considering factors such as accuracy, precision, and recall. Model evaluation is essential to assess the performance and generalization capabilities of the selected algorithms. We split the dataset into training and testing sets to evaluate model performance on unseen data. Additionally, we employed techniques such as k-fold cross-validation to obtain more robust estimates of model performance. We compared the performance of different algorithms based on metrics such as mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). After careful evaluation, the chosen model was then integrated into a comprehensive price prediction pipeline , which included preprocessing steps , encoding methods and handling various features to ensure optimal performance.

BUILDING RECOMMENDATION SYSTEM

In the process of building the recommendation system, three distinct recommendation models were developed, each focussing on different aspects of real estate: top facilities, price details, location advantages. The goal was to provide users with personalized recommendations tailored to their preferences and priorities.

BUILDING ANALYTICAL MODEL

An analytics module was developed to visually represent the key insights about the real estate data. Histogram , scatter-plot, bar-chart, top plot etc were employed to offer users a comprehensive understanding of the market.