**Detailed Project Report House Price Prediction**

Table of Contents

[**Introduction /Problem Statement** 2](#_Toc140914044)

[**Data Engineering** 2](#_Toc140914045)

[**Data Analysis** 4](#_Toc140914046)

[**Data Science** 5](#_Toc140914047)

[**Results/Testing** 5](#_Toc140914048)

[**Explainable AI** 7](#_Toc140914049)

# Introduction /Problem Statement

The real estate industry is a crucial sector of the global economy, with house prices being influenced by a multitude of factors such as location, property features, neighborhood characteristics, and economic conditions. Accurate prediction of house prices is of paramount importance for property sellers, buyers, and investors to make informed decisions. In this project, we present an end-to-end data analytics and machine learning solution to predict house prices based on a diverse set of features.

Our project aims to leverage advanced data engineering, data analysis, and data science techniques to create a robust and accurate house price prediction model.

Accurate house price prediction can empower stakeholders in the real estate industry to make well-informed decisions regarding property investments, sales, and purchases. By leveraging a diverse set of factors and advanced machine learning algorithms, our project offers a reliable tool to estimate house prices, enabling users to explore the correlation between various features and property values.

[Link to dataset](https://www.kaggle.com/c/house-prices-advanced-regression-techniques/)

[Link to github repo](https://github.com/aditya699/Hackathon2)

# **Data Engineering**

Data engineering is a critical phase in the project, responsible for the collection, preprocessing, and efficient management of raw data to prepare it for further analysis and modeling. In this project, we focused on transforming raw real estate data into a structured format suitable for analysis and model training. The key steps in the data engineering process are outlined below:

1. Data Collection:

The first step in data engineering involves collecting the raw real estate data from diverse sources, such as CSV files or databases. The dataset typically consists of multiple attributes, including property features, neighborhood information, location details, and the target variable - house prices. We ensured that the data collection process is comprehensive and that the dataset covers a wide range of features to achieve a robust prediction model.

2. Data Cleaning:

Raw data is often noisy and incomplete, containing missing values, outliers, and inconsistencies. To address these issues, we performed data cleaning techniques to ensure data quality. Missing values were imputed using appropriate methods, and outliers were identified and either removed or transformed to mitigate their impact on the model's performance.

3. Data Transformation:

To enhance the data's suitability for analysis, we applied various data transformation techniques. Numerical features were scaled to bring them to a common range, ensuring no feature dominates the model training. Categorical variables were encoded to numerical form using techniques such as one-hot encoding to make them amenable for model training.

4. Data Ingestion with Kafka:

To handle large volumes of data efficiently, we employed Apache Kafka, a distributed streaming platform, for data ingestion and real-time data processing. The cleaned real estate data was ingested into Kafka as a producer, allowing us to publish the data as events to Kafka topics.

5. Data Consumption with Kafka Consumer:

The Kafka consumer played a pivotal role in consuming the data from Kafka topics, allowing us to access the data for further analysis and modeling. By consuming data in real-time, we ensured that any updates or new data were seamlessly incorporated into the downstream processes.

# **Data Analysis**

Data analysis is a crucial step in the project that involves the exploration and examination of ingested data to gain valuable insights and identify patterns that can influence the house prices. In this project, we employed PySpark, a powerful data processing framework, to conduct extensive data analysis.

After thorough data analysis and preparation, the dataset was transformed into a model-ready format and dumped into a CSV file. This CSV file served as the input data for model training, allowing us to develop predictive models based on the insights gained from the data analysis phase.

# **Data Science**

The data science phase in the project involved developing predictive models to forecast house prices based on the features identified during the data analysis phase. We utilized PyCaret, a powerful machine learning library, to streamline the model development process.

PyCaret offers an extensive collection of machine learning algorithms. We chose three popular models - Random Forest (RF), Gradient Boosting Machine (GBM), and Extra Trees (ET) - as candidates for predicting house prices. PyCaret's simple API allowed us to train multiple models simultaneously and automatically tune hyperparameters for each model.

After successfully developing and testing the predictive model, we moved towards deploying the project using Docker. Dockerizing the project involved creating a Dockerfile that specified the project's dependencies, including the necessary Python libraries and the trained machine learning model.

With the Docker image ready, the project could be easily deployed on any system that supports Docker. This simplified deployment process ensured that the project could be seamlessly shared and run by team members or stakeholders, eliminating compatibility issues and simplifying the setup process.

# **Results/Testing**

In the testing phase of our project, we have designed a seamless pipeline to predict house prices based on a provided test file. This phase involves the following steps:

1. Data Preprocessing:

When the client provides a test file, it goes through the data preprocessing stage to ensure that it aligns with the format and requirements of our model. This step involves handling missing values, encoding categorical variables, and scaling numerical features using the same preprocessing techniques applied during the data analysis phase.

2. Data Analysis :

Since the test file may contain unseen or new data, we can choose to perform data analysis on it to gain insights into the distribution of features, identify potential outliers, and assess its similarity to the training data. However, if the test data is expected to be similar in nature to the training data, this step may be skipped.

3. Model Prediction:

Once the test file is preprocessed, it is fed into the selected machine learning model that was previously trained using PyCaret. The model predicts house prices based on the provided features in the test data.

4. Price Column Addition:

The predicted house prices are added as a new column to the preprocessed test data.

5. Exporting the Results:

The final step involves exporting the test data, now augmented with the predicted house prices, to a CSV file. This CSV file can then be shared with the client or other stakeholders for further analysis, reporting, or decision-making.

# **Explainable AI**

In our project, we not only focus on achieving high model accuracy but also emphasize the importance of Explainable AI to gain insights into the model's decision-making process. Explainable AI refers to the ability of a model to provide transparent and interpretable explanations for its predictions, enabling stakeholders to understand how the model arrived at a particular result.

1. Model Accuracy:

With our data science efforts and the use of PyCaret, we have achieved a commendable model accuracy of approximately **86%.** This high accuracy indicates that our model is performing well in predicting house prices based on the provided features.

2. Importance of Explainable AI:

While accuracy is crucial, the interpretability of our model is equally important, especially in the domain of real estate where decisions can have significant financial implications. By using Explainable AI techniques, we can understand which features and factors influence the model's predictions the most. This transparency is vital for gaining trust and confidence in the model's results, both from clients and other stakeholders.

3. Feature Importance:

Through Explainable AI, we can identify the most influential features in our model. This information helps in understanding which attributes (e.g., location, property size, quality, etc.) have the most significant impact on house prices. By highlighting feature importance, we can provide actionable insights to clients, enabling them to make informed decisions based on the key drivers of property prices.

4. Model Transparency:

Explainable AI also ensures model transparency, which is crucial when making decisions related to real estate investments or market trends. Our stakeholders can review the underlying reasoning of the model's predictions, ensuring that it aligns with their domain knowledge and business requirements.

5. Enhanced Decision-Making:

By combining high accuracy with Explainable AI, we offer our clients the best of both worlds. They benefit from accurate predictions while being equipped with transparent insights into how those predictions were derived. This enhanced decision-making capability empowers them to strategize better, mitigate risks, and optimize their real estate investments effectively.