BITS PILANI DIGITAL

FIRST TRIMESTER 2025-26

ADVANCED APEX PROJECT 1

Project Title	Indian Housing Price Prediction
Supervisor Name	Utkarsh Khare
Name of the Learner (with BITS ID)	<table><tr><th>Name of the Learner</th><th>BITS ID</th></tr><tr><td>SHIVANSH TIWARI</td><td>2025em1100502</td></tr><tr><td>HIMANSHU SONI</td><td>2025em1100506</td></tr><tr><td>PATHANENI GANGOTHRI</td><td>2025em1100507</td></tr><tr><td>SONKAR VEDANT RAJESH RANJEETA</td><td>2025em1100504</td></tr></table>

Sl. No.	Subject Name	State the relevance to Project
1	**Statistical Modelling & Inferencing**	Used in building regression models (Random Forest, Lasso) to predict house prices and performing feature selection using F-tests, p-values, and mutual information scores to identify significant predictors.
2	**Data Pre-processing**	Used for handling missing values through median/mode imputation, outlier detection, removing duplicate rows, and encoding categorical variables to prepare clean data for modeling.
3	**Feature Engineering**	Used in creating derived features like TotalSF, PropertyAge, QualityScore, and binary indicators (HasGarage, HasBasement) that enhance model performance by capturing complex relationships in housing data.
4	**Data Visualization & Storytelling**	Used to communicate insights through histograms, correlation heatmaps, distribution plots, and neighborhood price comparisons, making findings interpretable for stakeholders and identifying patterns like price skewness.
5	**Data Stores & Pipelines**	Used for downloading datasets via Kaggle API, implementing preprocessing pipelines with StandardScaler, automating workflows with scikit-learn pipelines (LassoCV), and organizing data transformation steps systematically.

Project Proposal

Project Title

Indian Housing Price Prediction

Problem Statement

Real estate pricing in India is influenced by a diverse range of factors including size, number of rooms, age of the property, location, neighborhood characteristics, and more. Buyers, sellers, and investors often struggle with inconsistent valuations due to a lack of data-driven insights. The goal of this project is to develop a predictive model that accurately estimates property prices based on these multiple features, helping reduce valuation uncertainty and enabling informed decisions.

Business Goal

The primary business objective is to build a robust regression model that can predict Indian housing prices based on relevant property and location attributes. This will assist real estate stakeholders—buyers, sellers, agents, and investors—in:

- Understanding key drivers of property prices
- Making evidence-based investment and pricing decisions
- Reducing reliance on subjective valuation methods

The insights and visualizations derived from the model will be deployed through an interactive dashboard to enable intuitive exploration and analysis.

Data Source

We will use the **"India House Price Prediction"** dataset available on Kaggle.

- **Source Platform:** Kaggle
- **Dataset Title:** Ames Housing Dataset,
- **Dataset URL:** House Prices - Advanced Regression Techniques | Kaggle

Tools & Technologies:

- **Programming Language:** Python
- **Libraries:**
  - **Data Analysis & Processing:** Pandas, NumPy
  - **Data Visualization:** Bokeh
  - **Modeling:** Scikit-learn, LightGBM
- **Development Environment:** Google Colab
- **BI Tools:** Bokeh as library

Project Workflow

The project will follow a structured Data Science workflow as outlined below:

- ➔ **Data Acquisition**
  - ◆ Fetch dataset from Kaggle using Python and Kaggle API.
  - ◆ Save raw data in a reproducible manner for team access.

- ➔ **Data Cleaning & Preprocessing**
  - ◆ Handle missing values, outliers, and inconsistent entries.
  - ◆ Outliers in numeric features were capped using winsorization at the 1st–99th percentile.
  - ◆ Convert categorical variables using encoding techniques.
  - ◆ Normalize/scale numerical variables as needed.
  - ◆ Applied log transformation to SalePrice to convert its right-skewed distribution into a normal distribution. This improved model stability, removed heteroscedasticity.

- ➔ **Exploratory Data Analysis (EDA)**
  - ◆ Use statistical summaries and visualizations to identify patterns and correlations.
  - ◆ Investigate geographic trends in pricing using location-based visualizations
  - ◆ The target variable SalePrice was highly right-skewed. Log transformation made it normally distributed, improving regression accuracy and reducing the effect of extreme high-price outliers..

- ➔ **Feature Engineering**
  - ◆ Generate new features such as price per square foot, location clusters, etc.
  - ◆ Reduce dimensionality or remove irrelevant features to improve model performance.
  - ◆ Used a multi-stage feature selection pipeline consisting of:
    1. Filter methods: F-test, p-values, Mutual Information
    2. Wrapper methods: RFECV (Random Forest cross-validation)
    3. Embedded methods: LassoCV for coefficient shrinking
  - ◆ Applied Principal Component Analysis (PCA) to reduce dimensionality while retaining 95% variance. Reduced feature space from 35 engineered features to 20 principal components, improving model generalization and reducing overfitting.

- ➔ **Model Building**
  - ◆ Linear Regression, Random Forest, Gradient Boosting, LightGBM,and Support Vector Regression (SVR) and Gradient Boosting performed best with 90.4% R^2.
  - ◆ Use cross-validation to ensure model generalization.

- ➔ **Model Evaluation**
  - ◆ Evaluate using metrics like RMSE (Root Mean Square Error), MAE (Mean Absolute Error), and R^2 score (Coefficient of Determination).
  - ◆ Select the most accurate and interpretable model for deployment.

➔ **Reporting & Visualization**

- ◆ Create a dashboard to present:
  - Key pricing influencers
  - Predicted vs. actual price trends
  - Location-wise pricing insights

Data Extraction

The dataset will be downloaded programmatically using the Kaggle API to ensure reproducibility. Steps include:

- Authenticate and connect using Kaggle API credentials
- Download and unzip dataset
- Load dataset into Pandas DataFrame
- Initial checks for missing values and data types

All steps will be documented in a Jupyter notebook titled:
final_phase4_submission.ipynb

Data Dictionary

Data Dictionary - House Prices Advanced Regression Technique