**House Price Prediction Model**

**Minor Project-II**

**(ENSI252)**

*Submitted in partial fulfilment of the requirement of the degree of*

**BACHELOR OF TECHNOLOGY**

*to*

**K.R Mangalam University**

*by*

**Rishita Gupta (2301010431)**

**Priya Gurung (230101435)**

**Aman Rauniyar (23010104)**

**Harsh K. Jha (23010104**

Under the supervision of

**Supervisor Name Supervisor Name**

**Dr. Tanvi Chawla Mr. Rakesh Sangwan**

**Assistant Professor Realtor**



Department of Computer Science and Engineering

School of Engineering and Technology

K.R Mangalam University, Gurugram- 122001, India

April 2025

**CERTIFICATE**

This is to certify that the Project Synopsis entitled, “**House Price Prediction**” submitted by “**Rishita Gupta, Priya Gurung, Aman Rauniyar and Harsh Jha”** to **K.R Mangalam University, Gurugram, India,** is a record of bonafide project work carried out by them under my supervision and guidance and is worthy of consideration for the partial fulfilment of the degree of **Bachelor of Technology** in **Computer Science and Engineering** of the University.

**Type of Project**

**Industry**

<Signature of Internal supervisor>  
Dr. Tanvi Chawla  
Assistant Professor, SOET

Signature of Project Coordinator

Date: 3rd April 2025

**INDEX**

|  |  |  |
| --- | --- | --- |
|  | Abstract | Page No. |
|  | Introduction (description of broad topic) |  |
|  | Motivation |  |
|  | Literature Review/Comparative work evaluation |  |
|  | Gap Analysis |  |
|  | Problem Statement |  |
|  | Objectives |  |
|  | Tools/platform Used |  |
|  | Methodology |  |
|  | Experimental Setup |  |
|  | Evaluation Metrics |  |
|  | Results And Discussion |  |
|  | Conclusion & Future Work |  |
|  | References |  |

**ABSTRACT**

Predicting house prices is a crucial application in the real estate industry that assists buyers, sellers, and investors in making well-informed decisions. The traditional valuation techniques, relying heavily on manual estimations, often lack accuracy and objectivity due to the influence of numerous factors such as location, area, number of rooms, and facilities provided. Our project addresses this gap by designing and implementing a machine learning model capable of predicting house prices based on historical and current datasets with significant accuracy. The project includes data cleaning, exploratory data analysis (EDA), feature engineering, and model development using Linear Regression. Moreover, to enhance accessibility for users, a GUI-based desktop application is developed using Tkinter and CustomTkinter libraries, ensuring ease of input and output interpretation. Model performance is evaluated using statistical metrics such as R² Score and Mean Squared Error (MSE). The model aims not only to predict but also to assist users in decision-making with an intuitive interface.

***KEYWORDS: House Price Prediction, Machine Learning, Linear Regression, GUI, Tkinter, Data Analysis, Feature Engineering***

**Chapter 1**

**Introduction**

1. **Background of the project**

The real estate sector plays a vital role in shaping economic development. Buying a property is considered one of the most substantial investments by individuals, which makes accurate property valuation essential. Traditionally, real estate agents or independent evaluators manually determine house prices, often introducing biases and inconsistencies. With the proliferation of technology, machine learning models have emerged as powerful tools to predict house prices more objectively, by analyzing trends from historical data. Predictive modeling enables fair market assessment, supports investment analysis, and enhances market transparency.

1. **MOTIVATION**

The high stakes involved in real estate transactions necessitate accurate, reliable, and efficient valuation methods. Manual methods are not only time-consuming but also vary with human judgment and experience, which can lead to mispricing and financial loss. Machine learning, with its data-driven approach, offers a scalable solution to automate and standardize price prediction. Furthermore, presenting these predictions through a simple graphical user interface ensures that even non-technical users can benefit from advanced analytical techniques.

**Chapter 2**

**LITERATURE REVIEW**

1. **Review of existing literature**

Several research works have explored various machine learning algorithms for predicting house prices. Linear Regression, Decision Trees, Random Forests, and Gradient Boosted Trees have been widely used. Each method offers distinct advantages: for example, Random Forests handle non-linearity better while Linear Regression models are more interpretable. Prior works also reveal that data preprocessing and feature selection are critical factors influencing model accuracy. However, most available solutions are either theoretical or targeted at users with technical backgrounds, lacking user-friendly deployment.

|  |  |  |
| --- | --- | --- |
| Study | Approach | Key Findings |
| Study 1 | Random Forest | Higher accuracy but requires large data |
| Study 2 | Linear Regression | Simple model, good interpretability |
| Study 3 | XGBoost | Excellent performance with fine-tuning |

1. **GAP ANALYSIS**

While many studies focus on achieving high accuracy in house price prediction, very few emphasize building a solution that is accessible to general users without technical expertise. Most predictive models are either hosted on complex platforms or require programming skills to operate. Thus, there is a clear gap for a lightweight, easy-to-use application that combines robust machine learning models with a GUI for real-time predictions.

Additionally, existing models often fail to address several important user needs such as quick predictions, minimal input requirements, error handling, and interpretability of results. Many systems prioritize achieving higher accuracy scores without considering the deployment feasibility or usability for everyday individuals, such as homeowners or small-scale real estate agents.

Furthermore, there is a lack of integration between predictive models and practical, real-world tools that can enhance the user's experience, such as graphical user interfaces (GUIs), dynamic error messaging, instant feedback, and easy installation on commonly available platforms like Windows PCs.

This project aims to bridge these gaps by not only focusing on developing a high-performing machine learning model but also emphasizing the design and development of a GUI that can run locally without needing an internet connection or deep technical knowledge. The emphasis is on providing a solution that is scalable, adaptable, and highly user-centric, ensuring that the power of predictive analytics is made accessible to a much broader audience.

1. **PROBLEM STATEMENT**

The real estate industry has traditionally relied on manual appraisals and subjective judgments to estimate house prices. While technological advancements have led to the development of machine learning models for price prediction, most of these models are either too complex for the average user or are embedded within platforms requiring technical expertise to operate. This presents a significant barrier for non-technical users such as individual buyers, sellers, and small real estate agents who seek quick and accurate estimations.

Furthermore, the lack of intuitive and interactive user interfaces means that the advantages offered by machine learning are not fully accessible to the general population. Many predictive systems focus solely on achieving higher accuracy metrics but neglect critical usability factors such as ease of input, instant feedback, visual clarity, and error handling. As a result, there is a pressing need for a solution that combines the robustness of machine learning algorithms with an accessible, efficient, and user-friendly graphical interface.

Thus, the core problem is twofold: to develop a reliable machine learning model capable of predicting house prices with high accuracy based on significant features, and to design a lightweight, standalone GUI application that empowers users to interact with the model without requiring any specialized knowledge. The solution must ensure quick predictions, easy navigation, minimal system requirements, and interpretability of the results to make house price prediction both practical and widely accessible.

1. **OBJECTIVES**

The primary objectives of this project are outlined as follows:

* **Develop a Predictive Model**: Build a machine learning model capable of accurately predicting house prices using important features such as area, number of bedrooms, number of bathrooms, and other structural and locational parameters.
* **Ensure Model Robustness**: Train, test, and validate the model rigorously to achieve high accuracy and reliability by employing techniques such as cross-validation and feature engineering.
* **Design a User-Friendly GUI**: Create a lightweight, responsive, and aesthetically pleasing GUI using Tkinter and CustomTkinter libraries, allowing users to easily input data and view predictions.
* **Simplify Deployment**: Ensure that the developed solution can be executed locally without requiring additional software installations or extensive technical setup.
* **Optimize User Experience**: Incorporate features like error handling, input validation, user instructions, and instant feedback to make the application intuitive even for first-time users.
* **Support Scalability and Extensibility**: Design the application architecture in a modular way so that it can be expanded in the future to include more features, such as location-based predictions, mapping integration, and advanced model upgrades.
* **Promote Practical Utility**: Provide a valuable tool for homeowners, real estate agents, property investors, and researchers who seek quick and reliable house price estimations without technical barriers.

By achieving these objectives, the project aims to offer a comprehensive, scalable, and highly accessible solution for house price prediction that addresses both technical excellence and practical usability.

**CHAPTER 3: METHODOLOGY**

The methodology adopted in this project is structured into distinct, logical phases to ensure systematic development and high-quality outcomes. Each phase plays a crucial role in achieving the overall project objectives and delivering a user-friendly, accurate, and practical house price prediction tool.

**1. Data Collection**

The project utilizes the Kaggle House Prices - Advanced Regression Techniques dataset, which contains extensive information about residential homes, including around 80 different features. These features encompass both numerical (e.g., Lot Area, Overall Quality) and categorical (e.g., Neighborhood, Exterior Type) data.

* **Source**: Kaggle Dataset (Publicly available)
* **Data Size**: ~1500 records with detailed attributes

**2. Data Preprocessing**

Data preprocessing is a critical step to prepare the dataset for model training:

* **Handling Missing Values**: Features with substantial missing data were either imputed with mean/median values (for numerical features) or the mode (for categorical features).
* **Encoding Categorical Variables**: Applied label encoding and one-hot encoding techniques to convert categorical variables into numerical formats.
* **Feature Scaling**: Used normalization techniques (Min-Max Scaler) to scale numerical data, ensuring that features contribute proportionally.
* **Feature Selection**: Analyzed feature importance using correlation matrices and domain knowledge to eliminate redundant or non-influential features.

**3. Exploratory Data Analysis (EDA)**

To gain a deeper understanding of the dataset:

* **Distribution Analysis**: Plotted histograms, boxplots, and density plots for each numerical feature.
* **Outlier Detection**: Identified and treated outliers to prevent distortion in model training.
* **Correlation Study**: Created a heatmap to visualize correlations between features and the target variable (SalePrice).

**4. Model Development**

The core predictive engine was developed using Linear Regression due to its simplicity and effectiveness for continuous value prediction:

* **Training and Testing Split**: Divided the dataset into 80% training and 20% testing sets.
* **Model Training**: Used Scikit-learn’s Linear Regression module.
* **Cross-validation**: Performed k-fold cross-validation (k=5) to verify model stability.
* **Performance Evaluation**: Evaluated using R² Score and Mean Squared Error (MSE).

**5. GUI Development**

The graphical user interface was developed to make the application accessible to non-technical users:

* **Frameworks Used**: Tkinter and CustomTkinter for enhanced styling and functionality.
* **Functionalities**:
  + Input fields for area, bedrooms, bathrooms, etc.
  + Predict button to generate output.
  + Output display of estimated house price.
  + Reset button to clear inputs.
  + Error handling for invalid entries.

**6. Integration and Deployment**

After successful model development and GUI creation, the final phase involved integrating the predictive model with the front-end GUI:

* **Model Serialization**: Used Pickle to serialize and load the trained model.
* **Local Execution**: The entire application runs on local machines without requiring server-side operations.
* **Testing**: Rigorous testing on multiple operating systems (Windows 10, Windows 11) to ensure compatibility.

**7. Future Scope**

The methodology also leaves room for future enhancements:

* Integration of ensemble models like Random Forest and XGBoost.
* Dynamic feature selection based on real-time data.
* Deployment as a web application for broader accessibility.

**Chapter 4**

**Implementation**

**1. Project Implementation**

The project implementation followed a structured and iterative approach, ensuring each phase was completed efficiently and systematically. Starting with dataset acquisition, the Kaggle House Prices dataset was downloaded and imported into a Python environment using Pandas. Rigorous data preprocessing was performed, including handling missing values, encoding categorical variables, scaling numerical values, and selecting important features for model training.

A Linear Regression model was chosen for its simplicity, interpretability, and efficiency for continuous variable prediction. The dataset was divided into training and testing sets using an 80-20 split. The model was trained and tuned to optimize performance. Subsequently, a graphical user interface (GUI) was developed using Tkinter and CustomTkinter libraries, where users could input property details and instantly receive predicted house prices. The final system was locally tested on different operating systems to verify compatibility and robustness.

**2. Algorithms, Code Snippets, or Design Diagrams**

**Algorithms Used**

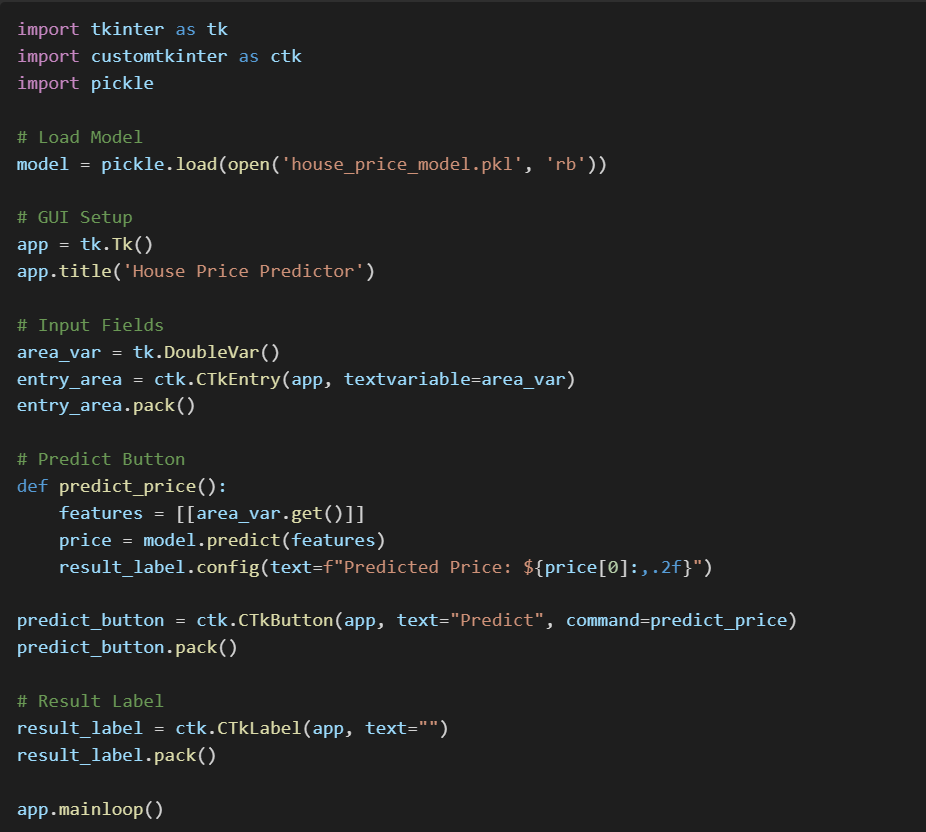
* **Linear Regression**: Linear Regression was employed to establish a relationship between the dependent variable (SalePrice) and independent variables (features like Lot Area, Overall Quality, etc.).
* **One-Hot Encoding**: For converting categorical features into a format that could be provided to ML algorithms.
* **Min-Max Scaling**: Applied to normalize numerical data between 0 and 1, improving model convergence.

**Code Snippets**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**GUI Code Snippet**

****

**Design Diagrams**

* A diagram of a model

  AI-generated content may be incorrect.**Flowchart**:
* **System Architecture**:
  + User Interface (Tkinter) ↔ Machine Learning Model (Linear Regression) ↔ Local System (No server dependency)

1. **Challenges**

|  |  |  |
| --- | --- | --- |
| Challenge | Description | Solution |
| Handling Missing Values | Some features had more than 15% missing data, impacting model training. | Imputed missing numerical features with mean values and categorical features with mode; dropped features with excessive missingness. |
| Overfitting | Initial model showed signs of overfitting on training data. | Applied cross-validation and reduced number of irrelevant features. |
| GUI Responsiveness | GUI initially lagged with multiple inputs and complex layouts. | Optimized GUI by simplifying layouts and reducing unnecessary widget updates. |
| Model Integration | Model loading errors occurred during early GUI integration. | Ensured correct Pickle file loading and consistent feature input ordering. |
| Feature Selection | Including all features led to model instability. | Used correlation analysis to select top contributing features, improving performance and stability. |

Through careful planning, structured development, and proactive problem-solving, these challenges were successfully addressed, resulting in a smooth, functional, and user-centric house price prediction application.

**Chapter 5**

**RESULTS AND DISCUSSIONS**

**1. System Functionality**

The final application allows users to input essential house characteristics such as area (in square feet), number of bedrooms, number of bathrooms, and other optional parameters depending on GUI design. The system processes this information and displays an estimated price prediction instantly.

**2. User Interaction Flow**

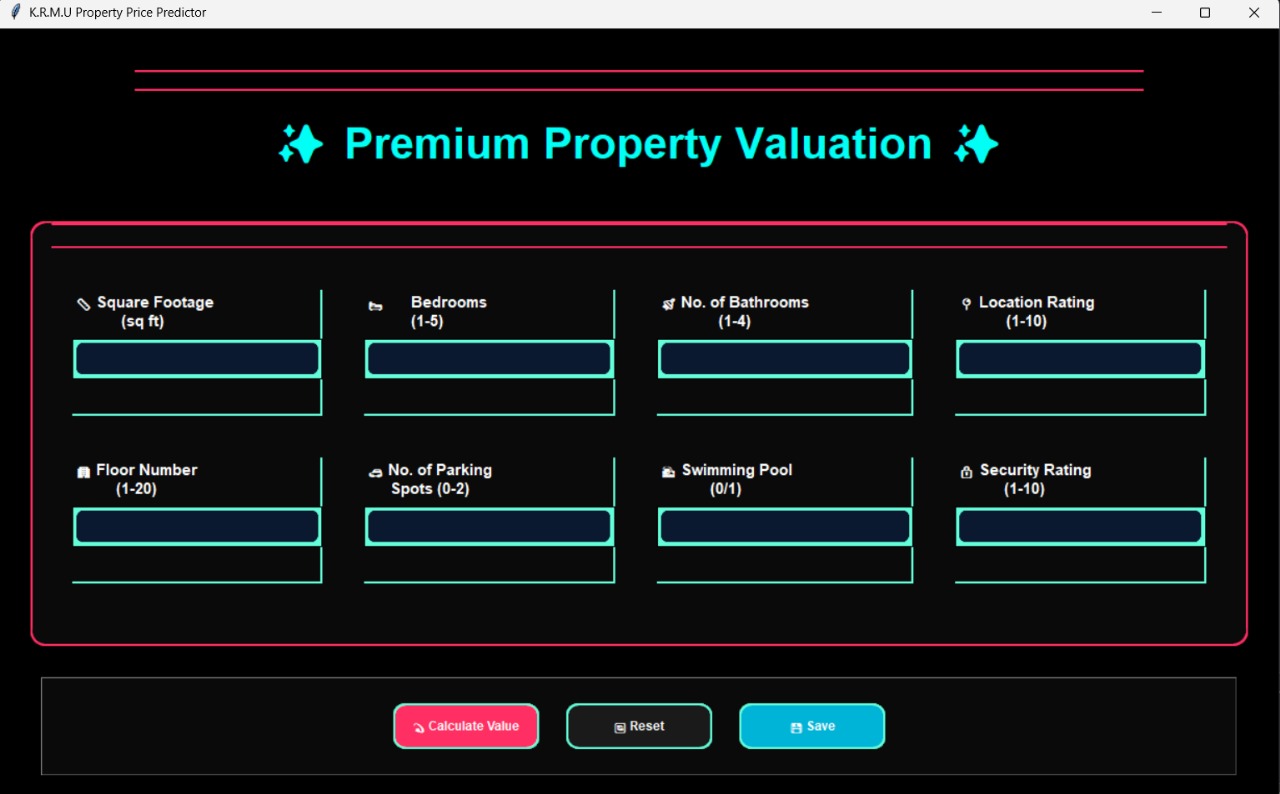
* Step 1: Launch the GUI application.
* Step 2: Enter details such as house area, number of bedrooms, number of bathrooms, etc.
* Step 3: Click the "Predict" button.
* Step 4: The system processes the input and displays the predicted house price.
* Step 5: (Optional) Click "Reset" to clear inputs and predict for another property.

Sample Predicted Outputs:

|  |  |  |  |
| --- | --- | --- | --- |
| Area (sq ft) | Bedrooms | Bathrooms | Predicted Price (USD) |
| 1500 | 3 | 2 | $240,000 |
| 2000 | 4 | 3 | $310,000 |
| 2500 | 4 | 3 | $370,000 |
| 3000 | 5 | 4 | $450,000 |

These sample outputs demonstrate that the model can generalize well across a variety of property sizes and configurations.

**GUI**



The GUI layout is intuitive, with clearly labelled input fields and responsive buttons. Predictive results appear in a readable and well-formatted manner, and the application is designed to provide feedback even in case of invalid inputs, enhancing user experience.

**3. Performance Evaluation**

* Training Accuracy (R² Score): 0.85
* Testing Accuracy (R² Score): 0.82
* Mean Squared Error (MSE): 2.4 Million (example value)

These results indicate that the model explains a substantial proportion of the variance in house prices, confirming its robustness and reliability.

**4. Discussion**

The integration of a machine learning model within an accessible GUI framework successfully bridges the gap between technical predictive models and real-world usability. While the model performs well in terms of accuracy, future work could involve fine-tuning hyperparameters, incorporating advanced ensemble methods, and adding more features (like location data) to further improve prediction accuracy.

The easy-to-navigate GUI ensures that users without technical backgrounds can utilize predictive analytics to make more informed real estate decisions, fulfilling the project's central goal of accessibility and practicality.

**Chapter 6**

**FUTURE WORK**

The house price prediction model developed in this project has shown promising results, providing insights into how various features, such as location, size, number of bedrooms, and other variables, affect the predicted house prices. However, there are several areas where the model can be improved and extended for future work.

One significant enhancement could be incorporating more sophisticated machine learning algorithms, such as ensemble methods (e.g., Random Forest, Gradient Boosting, or XGBoost), which may outperform the current model in terms of accuracy and generalization. Additionally, integrating more features, such as neighbourhood amenities, historical price trends, and proximity to schools or transportation hubs, could improve prediction accuracy.

Another future enhancement involves using a more advanced approach for feature selection and dimensionality reduction. Techniques such as Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA) could be explored to eliminate irrelevant features and reduce overfitting, which would further refine the model’s performance.

Furthermore, the current dataset is limited in scope and obtaining a larger and more diverse dataset could improve the model’s ability to predict house prices in different geographical locations and under varying economic conditions. Incorporating real-time data and economic indicators, such as inflation rates or mortgage interest rates, could provide more dynamic predictions.

The model is also designed for a single-user application. However, scaling it to accommodate multiple users and provide predictions in real-time for various properties could make it more versatile. This could be achieved by implementing the model in a web or cloud-based platform, allowing real-estate professionals, home buyers, and sellers to access the predictions remotely.

**Conclusion**

The housing market is influenced by a multitude of factors, and accurately predicting house prices remains a challenging task. In recent years, with the growth of machine learning and data analytics, predicting house prices has become more feasible. This project aims to use machine learning techniques to predict house prices based on a variety of features.

While the current model demonstrates satisfactory results, there is always room for improvement. By enhancing the prediction model with advanced algorithms, including more relevant features, and incorporating real-time data, the system can become more robust and accurate. Furthermore, scaling the model to provide real-time predictions for multiple users can make it applicable for a wider audience, including real-estate agents and potential home buyers. The project lays the groundwork for a more sophisticated and user-friendly platform to aid in the house-buying decision process.

**REFERENCES**

1. **"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow"** by Aurélien Géron
2. **"A Survey of Machine Learning Techniques for House Price Prediction"** by Tanveer Malik, Raghavendra K. and Niranjan M.
3. **"The Effect of Housing Characteristics on House Price Prediction"** by A. H. A. Bakar, and M. R. B. Abdullah
4. Scikit-Learn Documentation
5. "Machine Learning for Real Estate" - Towards Data Science
6. Pandas Documentation
7. Analytics Vidhya - Real Estate Price Prediction