**Mini Project Report on**



**“A Framework for House Price Prediction**

**using Ensemble Learning”**



**Submitted in partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**(Specialization in Machine Learning)**

**Submitted by:**

**Name: Sanat Manas**  **University Roll No. 2024101**

***Under the Mentorship of***

**Mr. Priyank Pandey**

**Assistant Professor**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**January-2025**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“A Framework of House Price Prediction using Ensemble Learning”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Mr. Priyank Pandey, Assistant Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name: Sanat Manas University Roll no: 2024101

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | 1-2 |
| Chapter 2 | Literature Survey | 3-4 |
| Chapter 3 | Methodology | 5-9 |
| Chapter 4 | Result and Discussion | 10 |
| Chapter 5 | Conclusion and Future Work | 11 |
|  | References | 12 |

**Chapter 1**

**Introduction**

In the following sections, a brief introduction and the problem statement for the work have been included.

* 1. **Introduction**

This is a project on the House Price Prediction Model using Machine Learning which uses data (like area, bedrooms, bathrooms, stories, etc) and Machine Learning algorithm to predict the price of the house based on different parameters essential for the price of the house. This model can be used to get a rough idea of the price of any house just by entering some of the parameters that affect the pricing of a home. The model uses historical data and predictive algorithms to analyze patterns and trends, enabling it to forecast prices accurately. Regression techniques like Linear Regression, Decision Trees, or advanced models like Random Forest, XGBoost, or Neural Networks are used for this model, and using these techniques the model trains and tests, and forecasts the most appropriate pricing according to it.

**1.2 Future Applications**

The future application of this model is: -

1. **Market Analysis**- The model can easily predict future market trends which helps stakeholders to understand the downfall or rise of prices in specific areas. Insights of areas or types of properties that are gaining popularity.
2. **Real Estate Insights**- This model can be useful to get the justified price of the house based on the parameters that affect the pricing for buyers and facilitates accurate pricing strategies for a seller.
3. **Investment Decisions-** The model can assist in determining the optimal timing for selling or buying a property. Helps investors assess potential risks and returns on a real estate investment.

**1.3 Beneficiary**

Beneficiaries of this model are: -

1. Builders and Developers
2. Real Estate Agents
3. Sellers and Buyers
4. Government and Investors
5. Insurance companies and Banks

**1.4 Objective**

In this model, I have used ensemble learning to predict the output i.e. Prices of the houses which will be briefed in chapters 2 and 3.

The main objective of this model is: -

1. Data Cleaning and Data Preprocessing.
2. Model Training and Testing.
3. Evaluating the Model and Visualizing its performance.

**1.5 Problem Statement**

To create a House Price Prediction Model using Machine Learning techniques to predict prices somewhat accurately to the actual price according to the training dataset.

**Chapter 2**

**Literature Survey**

**2.1 Overview of House Price Prediction**

House price prediction models aim to find the current value or future value of a house by taking in different parameters that could affect its valuation. Such models consider structured parameters like the total area, location, and age of the house along with unstructured parameters like images and neighborhood reviews. The prediction process typically includes feature engineering, model development, and model validation to gain practical value for buyers, sellers, and investors.

**2.2 Previous Work**

Many of the methods used for estimating house prices have been available for quite a long time now: -

* Linear Regression- Undoubtedly, one of the simplest if not the outright simplest methods to begin with, linear regression presupposes a linear correlation between features (specifying type of space, number of rooms) and the dependent variable which is the market value of the house.
* Decision Trees- Decision trees serve a nonlinear structure in predicting house prices, these serve to enhance interpretability and performance for a wider range of datasets by partitioning the data into smaller and finer subsets depending on the value of the parameters.
* Random Forest- This is a type of ensemble decision tree learning that achieves high accuracy by establishing several trees each known as a forest, thereby combining the results of various trees and reducing the complexity.
* Support Vector Machines (SVM)-In various applications, SVM was employed for house price prediction by fitting a hyperplane through several price points and thereby obtaining separations between the different ranges of prices by maximizing the boundary margins.
* Gradient Boosting Machines (GBM)- Gradient Boosting Machines (GBM) will be covered including algorithms such as XGBoost, LightGBM, and Catboost that proved effective for house price prediction while iteratively improving weak learners and minimizing errors. The Iterative Predictive Modelling technique has proved to be effective in optimizing prediction errors over multiple iterations with weak predictors. In addition to the construction of one-sided rockets for teaching purposes.

**2.3 Modern Techniques**

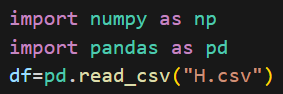
* Deep Neural Networks (DNN)- DNNs have come to the forefront in predicting house prices as they are capable of modeling complex interactions of features while being nonlinear, which is a missing piece in many other algorithms. These models function similarly to deep learning networks by stacking multiple hidden layers of the neural network to achieve optimal density from both structured and unstructured data, such as images and text.
* Stacking Ensembles- Stacking utilizes the predictive abilities of other base models, for example, XGBoost or Random Forest together with Linear Regression is used as a meta-model. This enables the cross-construction of predictions eliminating repetitive errors for real-time usage.
* AutoML- In general AI for automating house price prediction is highly demanded due to the complexity associated with selecting models, tuning high parameters, and evaluating them specifically designing a suitable architecture. These individuals and organizations utilize Automated machine-learning platforms that quickly identify and evaluate a prediction.

**Chapter 3**

**Methodology**

This model has been divided into various parts to make the model understandable and well-aligned. These are the several key points that combine to create a complete model: -

**3.1 Data Loading-** This is how we have imported the libraries and loaded the dataset named “H.csv” as shown in Figure 1.1.

****

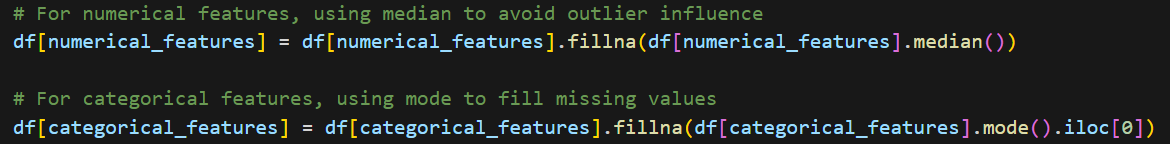
**Figure 1.1. Data loading**

**3.2 Data Analysis and Preprocessing-** In data analysis, we have separated numerical and categorical features by storing them in different variables as shown in Figure 2.1.

****

**Figure 2.1. Separation of Numerical and Categorical features**

In Data Preprocessing, Firstly I detected missing values in the dataset and treated them by using median for numerical features and mode for categorical features as shown in Figure 2.2.



**Figure 2.2. Handling Missing Values**

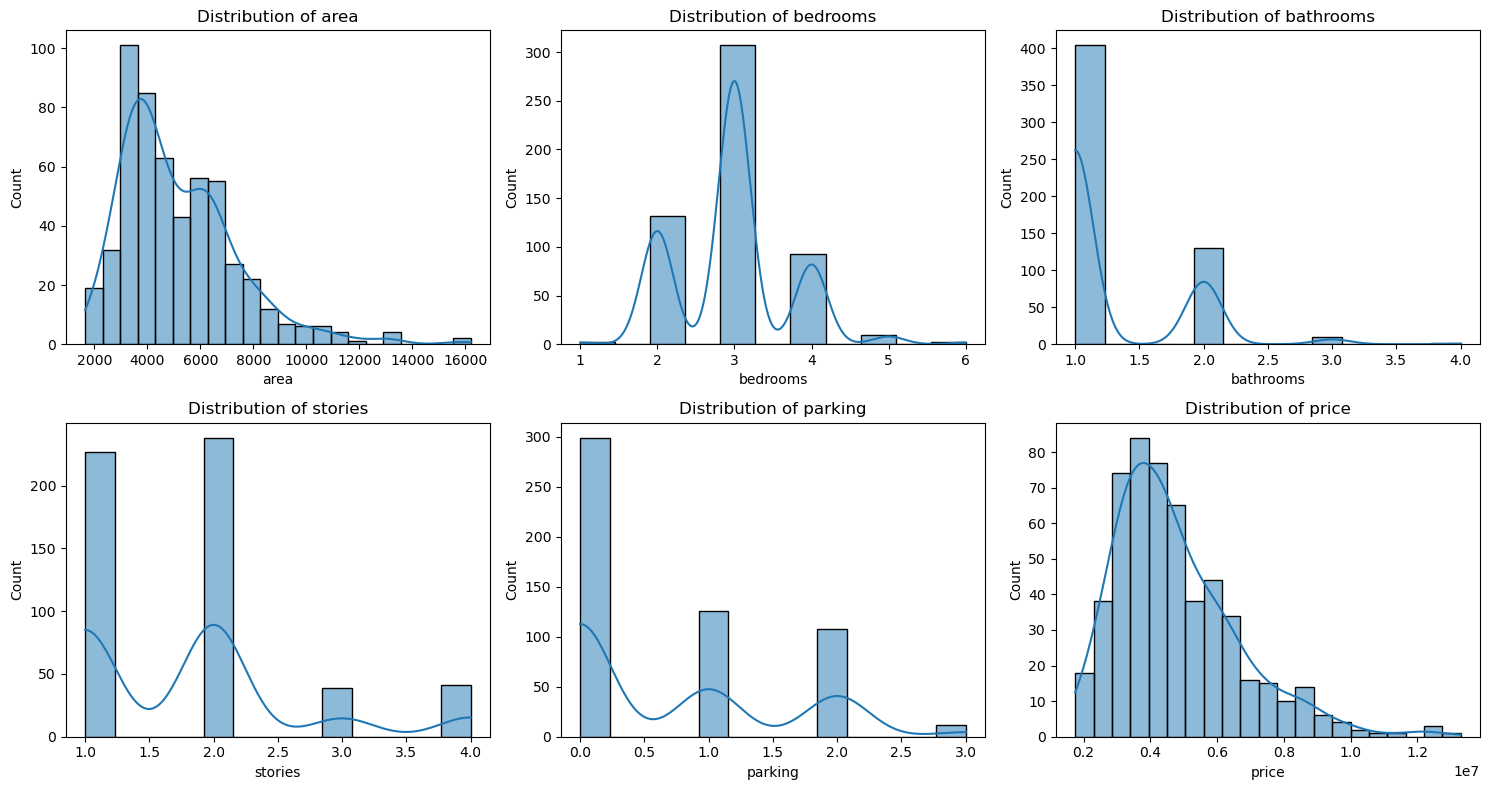
Then I analyzed outliers using boxplots for numerical features as shown in Figure 2.3. and treated the outliers in the dataset by creating a function to remove outliers.

A group of blue squares

Description automatically generated

**Figure 2.3. Outliers Visualization using Boxplots**

To understand the shape and skewness of each numerical feature, I visualized them using the Distribution Plot as shown in Figure 2.4.



**Figure 2.4. Distribution Plot of numerical features**

To understand the relationship between numerical features I have used a correlation heatmap as shown in Figure 2.5. It helps identify potential multicollinearity issues and informs feature selection decisions.

A screenshot of a graph

Description automatically generated

**Figure 2.5. Correlation Heatmap of numerical features**

This pair plot is shown in Figure 2.6. provides visualization of combined multiple aspects of our exploratory data analysis such as: -

* Shows the distribution of each variable on the diagonal.
* Displays relationships between all pairs of variables.
* Helps identify any non-linear relationships that might need special attention in modeling.
* Includes the crucial price vs. area relationship along with other important feature interactions.

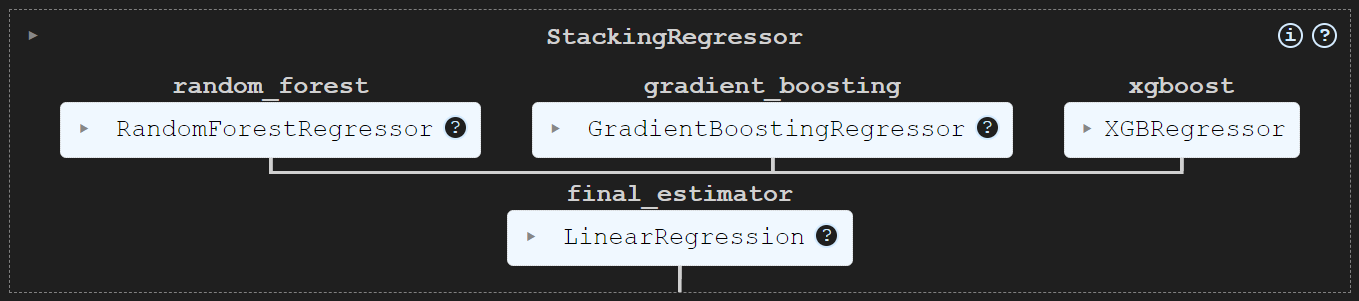
**A screenshot of a graph

Description automatically generated**

**Figure 2.6. Pair Plot of key features**

* 1. **Model**

There are different approaches for regression problems and one of them is ensemble learning which is used in this model. In ensemble learning, different models, often of the same type or different types, team up to enhance predictive performance. It's all about leveraging the collective wisdom of the group to overcome individual limitations and make more informed decisions. A Stacking Algorithm is one of the best techniques of ensemble learning which uses different models as a base model and then a final estimator to make a precise decision which has been used in this model as shown in Figure 2.7.

****

**Figure 2.7. Stacking Regressor Model**

* 1. **Training And Testing**

We train the model with a training set and then test the model. Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 scores are used to measure the performance of the Model as shown in Figure 2.8.

Median Absolute Percentage Error (MdAPE) measures the percentage deviation of predicted values from actual values, focusing on the median of absolute percentage errors to reduce the impact of outliers.

A black background with white text

Description automatically generated

**Figure 2.8.**

**3.5 Accuracy**

It is a measure of correct predicted values in percentage. It is also

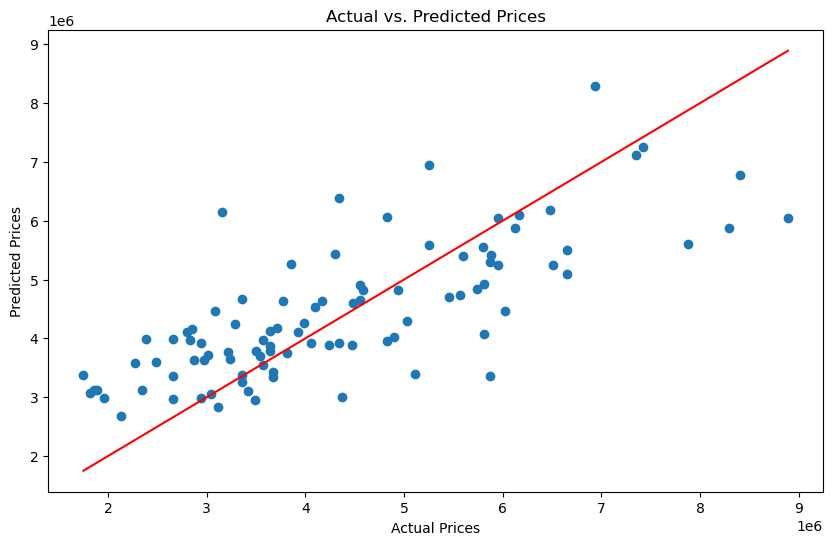
100 – MdAPE score. The accuracy of this model is: -



**Chapter 4**

**Result and Discussion**

Scatter plots of predictions versus actual values were generated, showing the correlation between the dependent and independent variables. The line of best fit indicates the model's ability to capture the relationship in the data. The result of this model is represented in Figure 4.1. The model performs adequately for the dataset used but shows some variance in predictions, suggesting potential underfitting or overfitting issues depending on the dataset's complexity.



**Figure 4.1. Visualization of Actual vs. Predicted Prices**

The linear regression model was successfully trained and evaluated. The evaluation metrics include metrics such as Median Absolute Percentage Error (MdAPE), Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared (R²). These metrics provide insights into the model's predictive accuracy and goodness-of-fit.

Limitations in model performance might be due to: -

* Insufficient feature engineering.
* Possible noise or outliers in the dataset.
* Non-linearity in the data that a simple linear model cannot capture.

**Chapter 5**

**Conclusion and Future Work**

This model works a bit well, but advancements can be made to make the model more accurate and precise. Advanced engineering can be done in the model. New potential data features can be added to the dataset that significantly improves the forecasting of the model.

More complex models like Decision trees or Polynomial Regression can be used in the model. We can train and test more models and compare their performances to select the best model for data training and testing. Additional metrics like Root Mean Squared Error (RMSE) can be included for more comprehensive evaluation. Implementation of regularized versions of linear regression, such as Ridge or Lasso, to address potential overfitting issues.

**References**

**[1] Dataset:**

* 1. H.csv: <https://drive.google.com/file/d/1B6dxFy65A1xFrObh2fWpGHoa9TN9RjAe/view?usp=drive_link>

**[2] Libraries:**

* 1. Pandas: McKinney, Wes. *"Data Structures for Statistical Computing in Python."* Proceedings of the 9th Python in Science Conference, 2010.
  2. NumPy: Oliphant, Travis E. *"A guide to NumPy."* USA: Trelgol Publishing, 2006.
  3. Matplotlib: Hunter, J. D. *"Matplotlib: A 2D Graphics Environment."* Computing in Science & Engineering, 2007.
  4. Seaborn: Waskom, M. L. *"Seaborn: Statistical Data Visualization."* Journal of Open-Source Software*,* 6(60), 3021, 2021.

**[3] Scikit-Learn Documentation:**

* 1. Scikit-learn: Machine Learning in Python. Pedregosa, et al. Journal of Machine Learning Research, 2011.
  2. Available at: Scikit-Learn Official Documentation

**[4] XGBoost Documentation**

1. Guides implementing XGBoost for predictive modeling tasks.
2. URL: <https://xgboost.readthedocs.io/>

**[5] Books and Tutorials:**

* 1. *"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow"* by Aurélien Géron, O'Reilly Media, 2019.
  2. Articles on ensemble techniques and house price prediction models using Python and popular libraries like Scikit-learn and TensorFlow.  
     URL: <https://towardsdatascience.com/>