School Of Computer Science

UNIVERSITY OF PETROLEUM & ENERGY STUDIES,

DEHRADUN- 248007. Uttarakhand

PROJECT REPORT

for

HOUSE PRICE PREDICTION

Submitted By

Name	SAP ID	Specialization/Batch
Abhishek Joshi	500090966	AI & ML (Hons) - B1
Nihar	500091867	AI & ML (Hons) - B1
Suraj Singh Bhandari	500094133	AI & ML (Hons) - B1

Submitted To:

Dr Rohit Srivastava

Assistant Professor

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Prologue

In an ever-evolving real estate landscape, the quest for precise house price prediction remains an enduring challenge. The dynamics of the housing market are shaped by a number of different factors. Amidst this complexity, stakeholders seek reliable methods to forecast house prices with accuracy, aiding informed decision-making and strategic planning.

1.2 Literature Review

We have performed a comprehensive literature review comparing and contrasting different research papers on house price prediction using different approaches:

House price prediction is a prominent research area in machine learning and artificial intelligence, where various methods have been explored. Deep learning models, including CNNs, RNNs, and LSTMs, have gained attention for their ability to learn complex patterns from data. Additionally, ensemble learning techniques like bagging, boosting, and stacking have been employed to improve prediction accuracy. By integrating deep learning with ensemble methods, researchers aim to achieve superior predictive capabilities for house price forecasting. This literature review sets the stage for further exploration into advanced deep learning-based ensemble models for accurate house price prediction.

A Hybrid Regression Technique for House Prices Prediction [1]

This paper discusses a hybrid regression approach for predicting house prices. The authors propose a composite data preprocessing and a creative feature engineering approach to enhance the linearity of the input features. They also propose a hybrid lasso and gradient boosting regression model to predict individual house prices. The proposed method was used as the core of the Kaggle Challenge "House Prices: Advanced Regression Techniques" and achieved promising results, reaching the top 1% of both teams and individual competitors. The book also discusses the challenges of forecasting home prices, with factors such as location, home size, home type, city, state, tax laws, economic cycles, population growth, interest rates, and other factors much that can affect demand as well as supply. The authors examine various regression algorithms, such as support vector machines, Lasso, Gradient boosting, Ridge, and Random forest, and examine their performance in predicting house prices The paper concludes with a discussion of the authors' creative feature engineering and the results of their hybrid regression techniques using lasso and gradient boosting. It is found that 230 features with Ridge regression produce the best score of the test data, while 160 features produce the lowest Root Mean Squared Error (RMSE) of the training data

House Price Prediction Using Machine Learning [2]

This paper discusses the use of machine learning to predict future housing prices. The authors compare and evaluate different forecasting methods and use Lasso regression as their model. They also discuss the importance of machine learning in improving safety awareness, improving medical diagnosis and providing quality customer service. The paper also includes a section on learning theory, using a pig model to avoid toxic tests, and developing machine learning algorithms to filter out spam emails. Finally, the authors note the importance of accurate estimates of house prices for the construction industry and financial planners.

House Price Prediction using Random Forest Machine Learning Technique [3]

This paper examines the use of random forest machine learning algorithms to predict house prices, highlights the limitations of the Home Price Index (HPI) and proposes a model based on the Boston Housing Dataset. The results show the accuracy of the model within an error of (+ -5). Home prices are influenced by physical condition, attitude and location, including property size, proximity to amenities and location. Understanding these policies is important for tenants, landlords, researchers and policy makers. Computer forecasting helps make informed decisions about property acquisition and timing.

Housing Prices Prediction with a Deep Learning and Random Forest Ensemble [4]

The paper includes a study that analyzes a data set of 12,223,582 housing advertisements collected from Brazilian websites from 2015 to 2018 to predict property prices The study uses two different machine learning algorithms, based on random forest and recursive neural networks so, and shows that by enriching the data set the combination of different ML methods can be a good way to forecast housing prices in Brazil The study received a small RMSLE, came first in the Data Science Challenge, a competition of Engineering Education for the Future (EEF) 2019.

Prediction of House Price Using XGBoost Regression Algorithm [5]

This paper discusses the use of machine learning algorithms, specifically XGBoost regression, to predict house prices. The authors emphasize the importance of considering various factors such as location, neighborhood, and acquisition factors when forecasting house prices. Preprocessing of data to remove anomalies and fill null values or remove data outliers is also discussed. The article also discusses the use of Scikit-learn, a module written in Python, to identify and cluster commonly used machine learning algorithms for supervised and unsupervised learning The authors integrate Gradient boost regression and Lasso model to predict the value of a house. The article concludes by highlighting the risks associated with manual methods of determining property market values and the importance of using machine learning for accurate forecasting.

Housing Price Prediction Using Support Vector Regression [6]

This paper examines the relationship between house prices and the economy and aims to predict unbiased housing prices to help buyers and sellers make informed decisions. The study uses an open dataset of 21,613 real estate transactions registered in King County, USA, and a comparison of different selection methods and feature extraction algorithms with support vector regression (SVR) de predict house prices The selection methods used in the carrot tests are recursive feature elimination (RFE), lasso, ridge, and random forest selector. The feature extraction method used is principal component analysis (PCA). After applying various reduction methods, SVR was regression modeled. Log transformation, feature reduction, and parameter tuning increased price prediction accuracy from 0.65 to 0.86, with a minimum MSE of 0.04. Experimental results show no difference in performance between PCA-SVR and feature selections-SVR in predicting housing prices in King County, USA. The advantage of using feature reduction is that it helps to identify the most important features, so the model cannot contain too many features.

Housing Price Prediction Based on CNN [7]

This paper proposes a new prediction algorithm based on Convolutional Neural Network (CNN) for predicting housing prices and product choices. The authors argue that the factors affecting the price of residential real estate are complex and the selection of effective factors is unclear, invalidating many traditional housing price forecasting methods The proposed CNN-based model is compared with other traditional methods, as well as with experiments using real data about the property. The work shows that the CNN model is superior in prediction accuracy and mean squared error. The authors conclude that their model can successfully predict housing prices in the context of real estate transactions and provide valuable insights into the real estate industry. Future work will consider more factors, such as government policy, and use more structured data.

House Price Prediction Using LSTM [8]

This paper focuses on the prediction of house prices in various provinces of Beijing, Shanghai, Guangzhou, Shenzhen etc. using machine learning models The authors use Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (2010). LSTM) network performance is compared in terms of Mean Squared Error (MSE) of the respective. Both stateful LSTM mesh and stacked LSTM mesh are used for improvement. The paper provides an overview of the previous steps and the design of the LSTM model. The authors conclude that the LSTM model has good properties of forecasting time series data and can be used to develop good forecasting models for house prices.

An Analysis of House Price Prediction Using Ensemble Learning Algorithms [9]

The aim of this paper is to predict house prices based on current market conditions using different machine learning and cluster learning methods. The evaluation evaluates the performance of the models based on four main parameters, namely RMSE, MSE, MAE, and R2 scores. The authors applied linear regression, KNN, SVM, XGB regressor, Adaboost, gradient boost regressor, random forest, and catboost to the Ames habitat dataset. The results show that ensemble learning methods, especially CatBoost, outperform machine learning methods in predicting house prices. The paper provides a detailed analysis of models' performance and scores.

Housing Price Prediction by Using Generative Adversarial Networks [10]

This paper discusses the use of generative adversarial networks (GANs) with long-term and short-term memory (LSTM) for housing price forecasting. The authors propose a new GAN model using LSTM for house price forecasting. The dataset used for training and analysis is House Sales in King County, USA. The paper also reviews related works on machine learning and deep learning methods for housing price forecasting. The authors compare their proposed model with other models and show that it gives lower prediction error and outperforms other models. The paper concludes with a methodological section describing the proposed modeling framework based on LSTM, time series neural networks, and GAN.

S. No.	Title	Method / Models	Dataset	Results
[1]	**A Hybrid Regression Technique for House Prices Prediction**	● Lasso Regression ● Gradient boosting ● Ridge Regression	Ames Housing Dataset	Ridge regression: ● For 160 features: RMSE = 0.112276 ● For 230 features: RMSE = 0.113627 ● For 280 features: RMSE = 0.114547 Lasso regression: ● For 160 features: RMSE = 0.113838 ● For 230 features: RMSE = 0.114974 ● For 280 features: RMSE: 0.115464 Note:- Best Result is achieved by using 65% Lasso and 35% Gradient Boosting.
[2]	**House Price Prediction Using Machine Learning**	● SVM ● Decision Tree ● Lasso ● XG Boost ● Random Forest ● Linear Regression	Dataset is from IRJET(International Research Journal of Engineering and Technology) Website	● Accuracy For SVM is 83% ● Accuracy For Linear Regression is Not Specific. ● Accuracy For Decision Tree is 99% ● Accuracy For XG Boost is 63%
[3]	**House Price Prediction using Random Forest Machine Learning Technique**	Random Forest	UCI Machine Learning Repository Boston Housing Dataset	● R-Squared: 0.90014311198457122 ● MAE: 1.9001315789473687 ● MSE: 6.702676631578947 ● RMSE:

				2.588952805977534 Note:- Here error margin is about 5%.
[4]	**Housing Prices Prediction with a Deep Learning and Random Forest Ensemble**	Enriched Random Forest	Dataset Composed of 12,223,582 Housing Advertisement Collections from Brazilian Website	RMSLE ● Enriched Random Forest: 0.30273 ● Ensemble Model: 0.23847
[5]	**Prediction of House Price Using XGBoost Regression Algorithm**	● XGBoost ● Ensemble Learning ● Base Learners	Kaggle Competition Dataset	Test Error:- ● XGBoost: 4.4% ● Ensemble Learning: 4.6% ● Base Learners: 4.8%
[6]	**Housing Price Prediction Using Support Vector Regression**	● SVM ● Lasso ● Ridge ● PCA ● RBF Kernel	Open Source Dataset of King County, USA	R-Square Score ● SVM without feature reduction is 0.65 ● Feature extraction with PCA and RBF Kernel is 0.86 ● Feature extraction with Lasso and Ridge is 0.86
[7]	**Housing Price Prediction Based on CNN**	● Gaussian Mixture (GM) Model ● XGBoost Model ● CNN Model	Dataset is from Land Resources and Housing InformationCenter for Dalian City, China	● MSE of GM Model is 0.43 ● XGBoost Model is 0.104 ● CNN Model is 0.01057
[8]	**House Price Prediction Using LSTM**	● ARIMA ● Stateful LSTM ● Stacked LSTM	House Price Dataset from January 2004 to September 2016 of Beijing, Shanghai, Guangzhou and Shenzhen	90% reduction in MSE compared to baseline Similar accuracy to Basic LSTM No significant improvement observed in terms of accuracy

[9]	**An Analysis of House Price Prediction Using Ensemble Learning Algorithms**	● Linear Regression ● KNN ● SVM ● XGB ● AdaBoost ● Gradient Boost ● Random Forest ● CatBoost	Ames Housing Dataset	Linear Regression: ● RMSE: 0.5037 ● MSE: 0.2537 ● MAE: 0.3873 ● R^2: 0.7463 KNN: ● RMSE: 0.4182 ● MSE: 0.1749 ● MAE: 0.3004 ● R^2: 0.8251 SVM: ● RMSE: 0.3826 ● MSE: 0.1464 ● MAE: 0.2712 ● R^2: 0.8536 XGB: ● RMSE: 0.3443 ● MSE: 0.1186 ● MAE: 0.2362 ● R^2: 0.8814 AdaBoost: ● RMSE 0.4901 ● MSE 0.2402 ● MAE 0.375 ● R^2 0.7598 Gradient Boost: ● RMSE: 0.3323 ● MSE: 0.1105 ● MAE: 0.2315 ● R^2: 0.8895 Random Forest: ● RMSE: 0.3499 ● MSE: 0.1225 ● MAE: 0.2414 ● R^2: 0.8775 CatBoost: ● RMSE: 0.3237 ● MSE: 0.1048 ● MAE: 0.2221 ● R^2: 0.8952

[10]	**Housing Price Prediction by Using Generative Adversarial Networks**	● Fully Connected Network ● Reduced Neurons to Model 1 ● Adding L2 Regularization to Model 1 ● Adding Drop Out to Model 1	Dataset is of House Sales in King County, USA	Model 1: ● MSE: 14748113.0 ● MAE: 2975.124755859375 ● Error Percentage: 13.57% Model 2: ● MSE: 418786.90625 ● MAE: 532.614013671875 ● Error Percentage: 15.32% Model 3: ● MSE: 418824.5 ● MAE: 532.665771484375 ● Error Percentage: 13.01% Model 4: ● MSE: 418858.5625 ● MAE: 532.5738525390625 ● Error Percentage: 14.60%

1.3 Problem Statement

Accurately predicting house prices is critical in real estate which impacts stakeholders such as sellers, real estate agents, investors, etc. This calls for a comparative study on different existing models which helps in house price prediction.

1.4 Objective

To implement a robust deep learning ensemble method for accurate house price prediction while documenting the entire workflow. At the end, evaluating our model based on evaluation metrics and validation techniques.

1.5 System Requirements

Hardware Requirements:

Processor:

a. Intel Core i5 or equivalent (or higher) for efficient language data processing.
b. Memory (RAM): Minimum 8 GB RAM for handling large datasets and computational tasks.
c. Internet Connectivity: Required for downloading datasets, libraries, and resources, as well as for collaboration and accessing online documentation.

Software Requirements:

a. Programming Language: Python
b. Development Environment: Google Collab, or Visual Studio Code for code development and debugging.
c. Libraries:
    - NumPy and pandas for data manipulation and analysis.
    - Matplotlib or Seaborn for data visualization.
    - sklearn for modelling

2. SYSTEM ANALYSIS

2.1 Motivation

Accurate house price prediction is vital for various stakeholders in real estate transactions. Our project aims to perform a comparative study on regression and random forest ensemble models for capturing market complexities, aiding individuals, organizations, policymakers, and urban planners in making informed decisions.

2.2 Proposed System

We aim to contrast between different house price prediction machine learning algorithms such as Regression and Random Forest. The system will be trained on relevant property data and market trends., ensuring accurate and reliable predictions. We will also perform validation techniques.

3. DESIGN

3.1 Diagrams

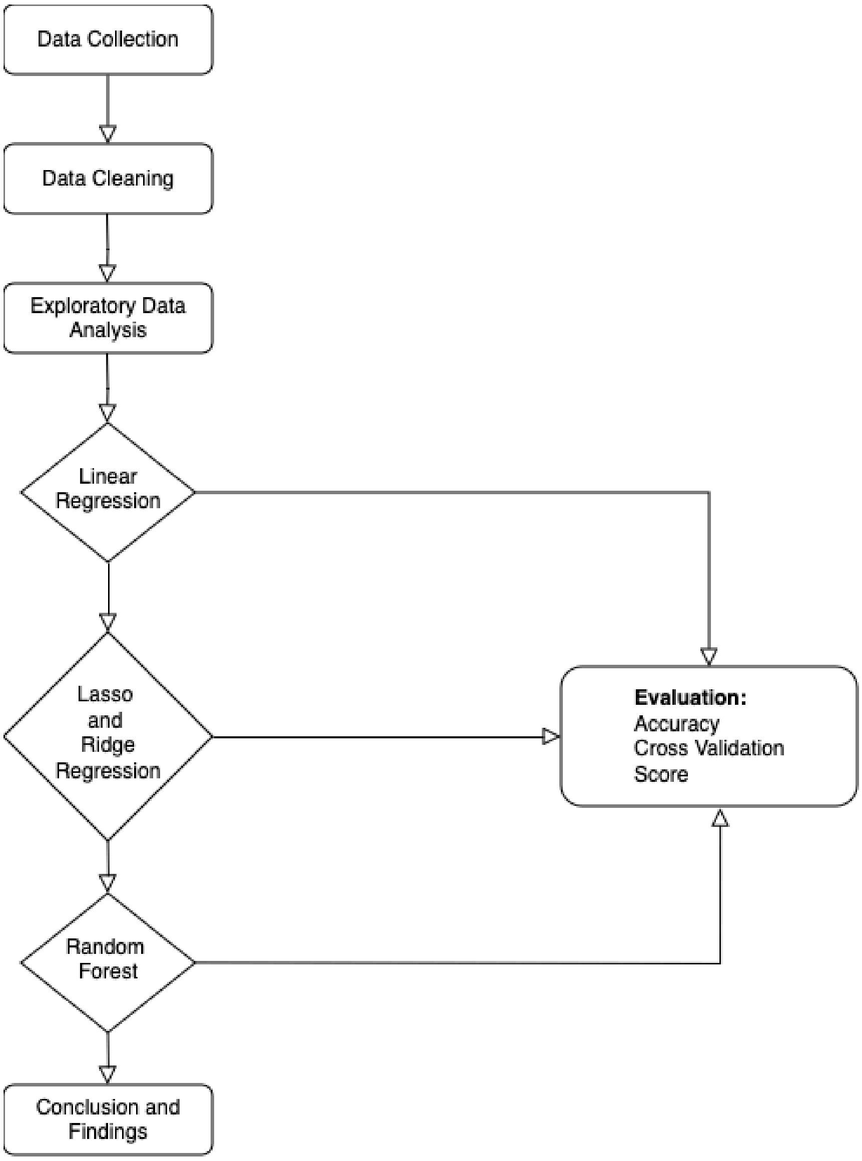

Fig. 1 Workflow Diagram

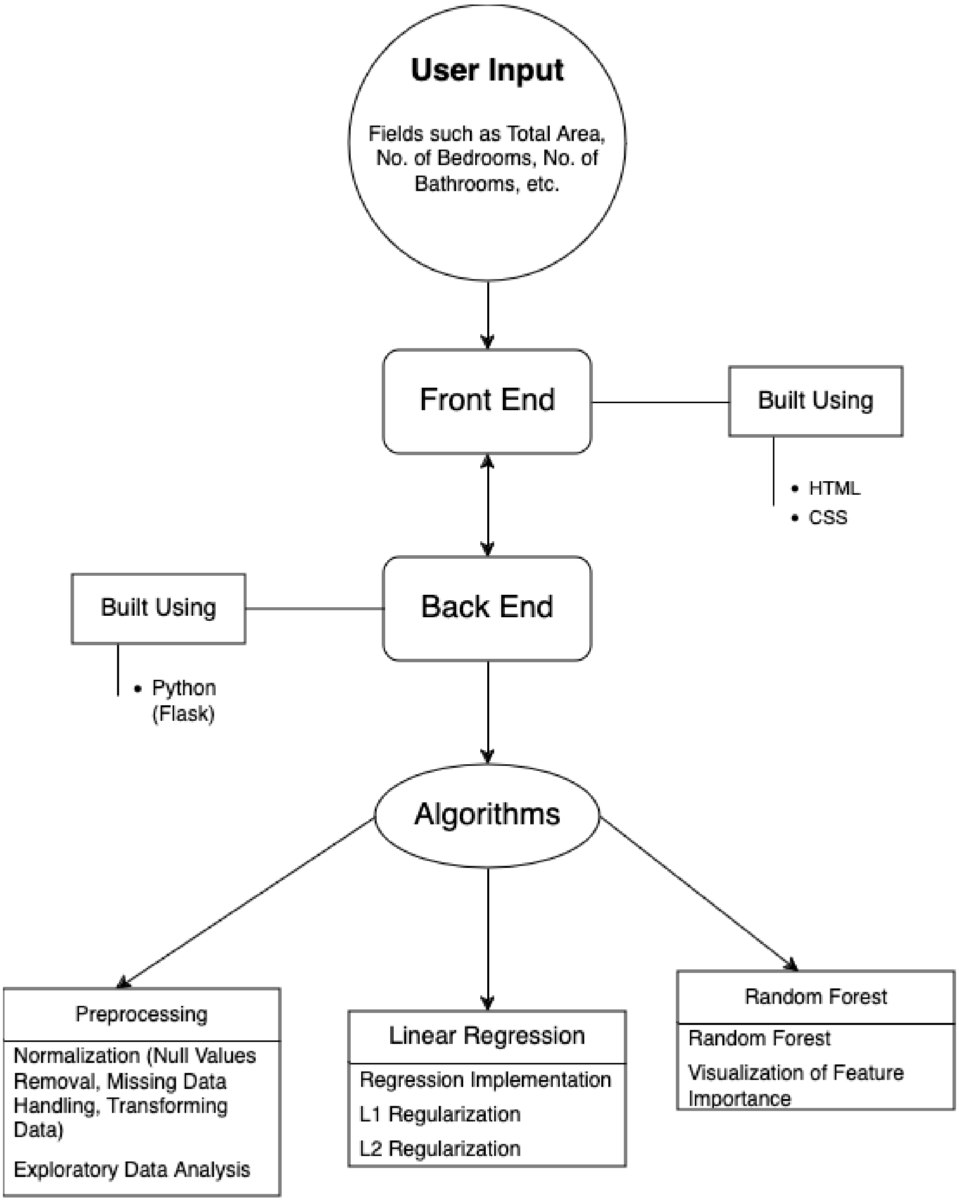
User Input
Fields such as Total Area, No. of Bedrooms, No. of Bathrooms, etc.
Front End
Built Using
• HTML
• CSS
Back End
Built Using
• Python (Flask)
Algorithms
Preprocessing
Normalization (Null Values Removal, Missing Data Handling, Transforming Data)
Exploratory Data Analysis
Linear Regression
Regression Implementation
L1 Regularization
L2 Regularization
Random Forest
Random Forest
Visualization of Feature Importance

Fig. 2 Architectural Diagram

4. IMPLEMENTATION

4.1 Methodology

The provided code outlines a methodology for predicting house prices using regression techniques (Lasso and Ridge regularization) and Random Forest. Below is a detailed methodology based on the code snippets provided:

1. Data Cleaning:

1) The dataset is loaded from a CSV file and initially contains 1460 observations and 81 features.
2) The data is inspected using df.head() and df.info() to understand the structure and missing values.
3) Features with a high percentage of missing values (e.g., PoolQC, MiscFeature, Alley, Fence, FireplaceQu) are removed.
4) The LotFrontage feature, with 16% missing values, is imputed with the mean of the column.
5) Rows with any remaining missing values are dropped, resulting in a cleaned dataset with 455 observations and 76 features.

2. Data Analysis:

1) The distribution of the target variable, SalePrice, is checked to ensure it follows a normal distribution.
2) As the SalePrice distribution is skewed to the right, the target variable is transformed using the natural logarithm to achieve a more normal distribution.

3. Model Building:

1) The dataset is split into features (X) and the target variable (y).
2) Linear Regression, Lasso Regression, Ridge Regression, and Random Forest models are imported from scikit-learn.
3) The dataset is further split into training and testing sets using train_test_split .
4) The models are trained on the training data and evaluated using cross-validation scores.
5) The Random Forest model is optimized using GridSearchCV to find the best parameters.
6) Feature importances are calculated for the Random Forest model to identify the most important predictors.

4. Feature Importance Analysis:

1) The feature importances obtained from the Random Forest model are mapped to the corresponding features.
2) The top 15 predictors are identified based on their feature importances.
3) The top predictors are printed to show the most promising features for predicting house prices.

4.2 Mathematical Model & Algorithms

1. Linear Regression:

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation to observed data.

Simple Linear Regression: Simple Linear Regression involves only one independent variable and one dependent variable.

$Y = a + bX + \varepsilon$

where,

- Y = Output/Dependent Variable
- a = Interceptor
- bX = Slope
- ε = Error

Multiple Linear Regression: Multiple Linear Regression involves more than one independent variable

and one dependent variable. The equation for multiple linear regression is:

$Y = b_0 + b_1x_1 + b_2x_2 + \ldots.. + b_nx_n + \varepsilon$

where,

- Y = Output/Dependent variable
- b_0 = Intercepter
- b_1, b_2, b_3 , b_4....= Coefficients of the model/slope
- x_1, x_2, x_3, x_4,...= Independent variable
- ε = Error

2. Ridge Regression:

The ridge regression model is similar to linear regression but with an additional regularization term:

$J(w) = (1/2) * \sum (y - h(y))^2 + \sum |w|^2$

where,

- y is the actual value
- h(y) denotes the predicted value
- w denotes the feature coefficient

3. Lasso Regression:

The lasso regression model also includes a regularization term but uses the L1 norm:

$J(w) = (1/2) * \sum (y - h(y))^2 + \sum |w|$

Lasso can perform feature selection by setting some coefficients to zero, effectively reducing the number of features used in the model.

where,

- where y is the actual value
- h(y) denotes the predicted value
- w denotes the feature coefficient

4. Random Forest:

- The random forest model is an ensemble method that uses multiple decision trees to make predictions.
- Each tree in the forest is trained on a random subset of the data and features.
- The final prediction is the average (regression) of all the individual tree predictions.

5. LIMITATIONS

1) **Data Quality:** One of the limitations of this analysis is the quality of the data. The dataset used for predicting house prices may contain missing values, outliers, or errors that could impact the accuracy of the predictions. Ensuring data quality through thorough cleaning and preprocessing is crucial for improving the reliability of the model.
2) **Feature Selection:** Another limitation is related to feature selection. The model's performance heavily relies on the features used for prediction. In this analysis, the selection of features may not have been optimal, leading to potential underfitting or overfitting issues. Exploring more advanced feature selection techniques could enhance the model's predictive power.
3) **Model Complexity:** The choice of models, such as Lasso, Ridge, and Random Forest, introduces complexity to the analysis. While Random Forest showed promising results, the complexity of the model may hinder its interpretability. Simplifying the model or exploring simpler alternatives could be beneficial.

6. FUTURE ENHANCEMENTS

1) **Advanced Algorithms:** To improve the accuracy and robustness of the predictions, future enhancements could involve exploring more advanced machine learning algorithms beyond linear regression and ensemble methods. Techniques like gradient boosting, neural networks, or support vector machines could be considered for more sophisticated modeling.
2) **Hyperparameter Tuning:** Fine-tuning the hyperparameters of the models can significantly enhance their performance. Utilizing techniques like GridSearchCV for hyperparameter optimization could help in achieving better results and improving the overall predictive power of the models.
3) **Feature Engineering:** Enhancing feature engineering techniques can lead to better model performance. Creating new features, transforming existing ones, or incorporating domain knowledge into feature selection could provide more meaningful insights and improve the model's ability to capture complex relationships in the data.

4) **Cross-Validation:** Implementing more robust cross-validation strategies, such as k-fold cross-validation, can help in assessing the model's generalization performance more accurately. Enhancing the validation process can lead to more reliable model evaluation and selection.

7. CONCLUSIONS

7.1 Interpretation and Comparative Study

1) **Performance:** Random Forest outperformed the regression models with an estimated accuracy of 85%, which was about 2% higher than the regression models.
2) **Interpretability:** Linear Regression is the most interpretable model, followed by Ridge and Lasso Regression. Random Forest, being an ensemble model, is less interpretable.
3) **Handling Complexity:** Random Forest is better at capturing complex relationships and interactions in the data compared to linear regression models.
4) **Feature Importance:** Random Forest provides feature importances, which can help in understanding the key predictors influencing house prices.
5) **Model Iteration:** The document emphasizes that machine learning is an iterative process, and further work is needed to build a high-performing prediction model.

8. REFERENCES

[1] S. Lu, Z. Li, Z. Qin, X. Yang and R. S. M. Goh, "A hybrid regression technique for house prices prediction," 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 2017, pp. 319-323, doi: 10.1109/IEEM.2017.8289904.

[2] A. P. Singh, K. Rastogi and S. Rajpoot, "House Price Prediction Using Machine Learning," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 203-206, doi: 10.1109/ICAC3N53548.2021.9725552

[3] Abigail Bola Adetunji, Oluwatobi Noah Akande, Funmilola Alaba Ajala, Ololade Oyewo, Yetunde Faith Akande, Gbenle Oluwadara, House Price Prediction using Random Forest Machine Learning Technique, Procedia Computer Science

[4] Afonso, Bruno & Melo, Luckeciano & Dihanster, Willian & Sousa, Samuel & Berton, L.. (2019). Housing Prices Prediction with a Deep Learning and Random Forest Ensemble.

[5] J. A. . et. al., “Prediction of House Price Using XGBoost Regression Algorithm”, TURCOMAT, vol. 12, no. 2, pp. 2151 –, Apr. 2021.

[6] Wu, Jiao Yang, "Housing Price Prediction Using Support Vector Regression" (2017). Master's Projects. 540.

[7] Y. Piao, A. Chen and Z. Shang, "Housing Price Prediction Based on CNN," 2019 9th International Conference on Information Science and Technology (ICIST), Hulunbuir, China, 2019, pp. 491-495, doi: 10.1109/ICIST.2019.8836731.

[8] Chen, X., Wei, L. and Xu, J., 2017. House price prediction using LSTM. arXiv preprint arXiv:1709.08432.

[9] Boyapati, S. V., Karthik, M. S., Subrahmanyam, K., & Reddy, B. R. (2023). An Analysis of House Price Prediction Using Ensemble Learning Algorithms. Research Reports on Computer Science, 2(3), 87–96.

[10] C. -F. Hsieh and T. -C. Lin, "Housing Price Prediction by Using Generative Adversarial Networks," 2021 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), Taichung, Taiwan, 2021, pp. 49-53, doi: 10.1109/TAAI54685.2021.00018.