HOUSE PRICE PREDICTION

1st Shivanshu Vinay Singh
Mathematical Sciences
Stevens Institute of Technology
Hoboken, United States
ssingh75@stevens.edu

2nd Visharad Ravi
Electrical & Computer Engineering
Stevens Institute of Technology
Hoboken, United States
vravi4@stevens.edu

3rd Uday Kurella
Mathematical Sciences
Stevens Institute of Technology
Hoboken, United States
ukurella@stevens.edu

***Abstract*—A dynamic and complex system, the real estate market is impacted by several variables, including property attributes, location, and economic indicators. To improve real estate appraisal accuracy and efficiency, this research article investigates the use of machine learning approaches for house price prediction. The research makes use of an extensive dataset that includes a variety of factors, including square footage, the number of bedrooms, neighborhood attributes, and economic indicators. Preprocessing and feature engineering are used in the methodology to handle missing data and standardize variables. Various machine learning algorithms including Linear Regression, XGB Regression, and Random Forest are utilized in the process of creating prediction models. The models are assessed for accuracy and generalization potential using measures like Mean Absolute Error, Mean Squared Error and Root Mean Squared Error, which are derived from training them on historical data.**

I. Introduction

Accurately estimating home prices in the current dynamic real estate market is a difficult undertaking that necessitates a deep comprehension of the numerous aspects affecting property values. The complex patterns and non-linear relationships included in the data are frequently too complex for traditional approaches to fully capture. However, the development of machine learning algorithms has completely changed the industry, giving us strong instruments to more accurately assess and forecast home prices. The goal of this research is to create a strong and accurate housing price forecast model by utilizing machine learning techniques. We want to improve our forecasting skills to assist investors, real estate agents, and homeowners in making better decisions by utilizing the power of sophisticated algorithms.

Importance of House Price Prediction in Informed Decision-Making by a Buyers can make informed decisions about whether a property fits their budget.Sellers can set competitive prices based on market trends, maximizing their chances of a successful sale.Market Analysis on Real estate professionals and investors use house price predictions to analyze market trends and identify potential investment opportunities. The model will provide a lot of information and knowledge to home buyers, property investors, and house-builders, such as the valuation of house prices in the present market, which will help them determine house prices.

A wide range of elements, such as property qualities, location-based attributes, economic indicators, and historical sales data, are included in the dataset that was used for this project. Our goal is to use machine learning to find hidden patterns and connections among these variables so that the model may learn from the data and become more accurate in its predictions. We'll examine the a. Model Development: Create a predictive model that can estimate home prices based on pertinent features including location, size, number of bedrooms, amenities, and market trends by using machine learning algorithms. b. Data Analysis: To determine the main variables affecting home values, thoroughly examine real estate data. Investigating correlations, anomalies, and patterns that advance our knowledge of market dynamics as a whole is part of this. c. Feature engineering: It is the process of carefully choosing and designing features to maximize the performance of the model. This entails managing missing data, standardizing variables, and applying domain expertise to improve the algorithm's predictive power. d. Model Evaluation: Employing suitable metrics and validation procedures, carefully evaluate the model's performance. Ensuring the predictive model's dependability and capacity to generalize to new, untested data requires completing this stage. A deep understanding of the prediction capabilities and limitations will be provided by assessing the model's performance and interpreting the findings.

The objective of this machine learning-based house price prediction getaway is to not only develop a highly accurate model but also to add to the expanding body of knowledge that is shaping the future of machine learning and real estate. We want to give stakeholders a useful tool for navigating the complex world of real estate transactions by utilizing state-of-the-art technology.

II. Related Work

A literature survey has been conducted for the related title of this project. Various papers were taken into consideration and in that, the methodology used, the results obtained and the research gaps have been found.

In the paper [1], a classification algorithm-based house resale price prediction was presented. This study uses various classification algorithms, including K-Means, Random Forest, Decision Tree, and Linear Regression, to forecast the house's selling price. The price of a home is influenced by a wide range of factors, including location, economic conditions, and physical characteristics. Here, they used RMSE as the performance matrix for various datasets, and we applied these techniques to get the most accurate model that can forecast better outcomes.

A hybrid regression technique for predicting housing prices was developed in the paper [2]. This research examines the creative feature engineering method with a limited dataset and data features. Recently, the suggested method was used as the main framework for the Kaggle challenge ”House Price: Advance Regression Techniques.” Predicting suitable pricing for customers based on their objectives and budgets is the aim of the article. This study examined the ability to forecast home prices through feature analysis. A predictive model is constructed using several Machine Learning methods, such as Random Forest, Decision Tree, and Linear Regression. They have taken a methodical approach, starting with data collection and moving through pre-processing, data analysis, and model building. The results of each model evaluation are then saved in a ”.txt” file. Following this, Random forests provide the greatest outcome with the training set. Random forest was shown to have the best accuracy, at roughly 87%.

In the paper [3] House Price Prediction using Machine Learning Algorithms: A Case Study of Melbourne, Australia published by T. D. Phan. This comprehensive case study examines the dataset to provide insightful information about the housing sector of the Australian city of Melbourne. They have employed several models of regression. Beginning with the reduction of data to using the Principal Component Analysis (PCA) procedures to obtain the dataset’s best possible answer. After that, they applied Support vector machines, or SVMs, which were used in competing arrive. Consequently, the various techniques used to obtain the greatest outcomes from it.

III. Our Solution

A. Description of Dataset

Data analysts have an intriguing chance to examine and forecast the direction of property values in real estate markets such as those in Sydney and Melbourne. Property price prediction is becoming a more useful and significant skill. Real estate values are a reliable gauge of a nation’s economic health as well as the state of the market as a whole. We are organizing a sizable collection of real estate sales records that are kept in an unknown format and have unknown data quality concerns based on the information supplied. The unwanted columns are removed and the necessary columns such as price bedrooms, bathrooms, sqft_living, sqft_lot, floors, waterfront, condition, yr_built, and yr_renovated.

A heatmap is a crucial tool for examining data. A heatmap displays the relationship between various apartment characteristics and pricing in a city. The number of bedrooms, bathrooms, living rooms, basements, views, living area square footage, above-ground square footage, and basement square footage are all included in the amenities in Fig. 2. A positive association between the feature and price is indicated by red squares. For instance, higher-bedroom apartments are typically more expensive. A negative link between the feature and price is indicated by blue squares. For instance, basement flats are typically less expensive. White squares show that the feature and price do not correlate.

The square footage of the living area and price, the number of bedrooms and price, and the price and view have the biggest positive relationships. The basement’s square footage and price have the largest negative correlation.

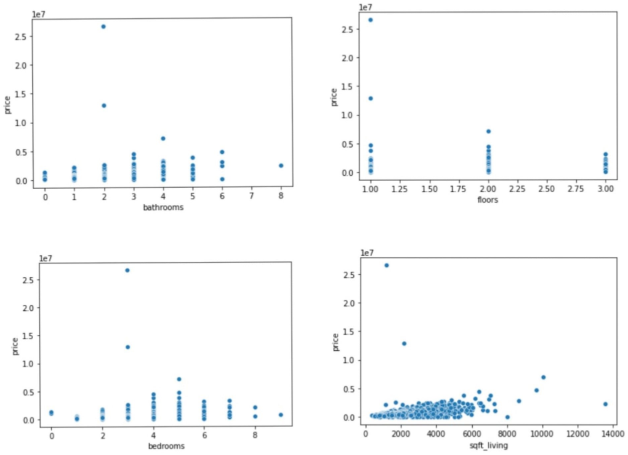


Fig. 1: Feature Selection of using the price (main feature) vs other features.

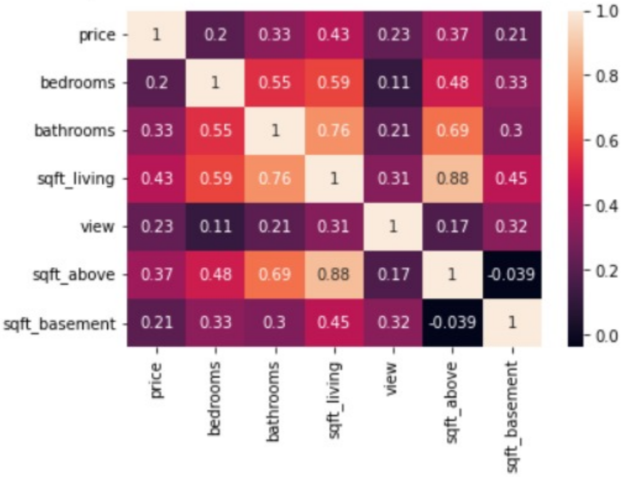


Fig. 2: Correlation between different features of apartment prices.

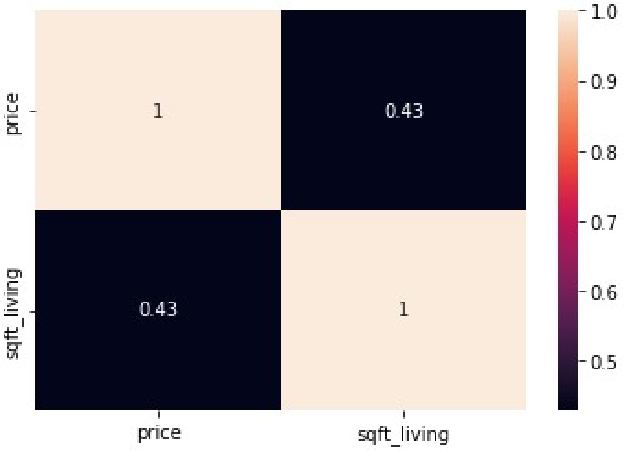


Fig. 3: Correlation between sqft_living and prices.

B. Machine Learning Algorithms

In machine learning, Random Forest is an ensemble learning method that is applied to both regression and classification problems. During the training phase, it builds a large number of decision trees, from which it outputs the mean prediction (regression) or the mode of the classes (classification). A random subset of the training data is used to build each tree in the forest, and each node's decision-making feature is chosen at random from a subset of features. Through this process, the trees become more diverse, which lowers overfitting and improves overall predictive accuracy.

Decision trees involve the greedy selection of the best-split point from the data set at each step. This algorithm makes decision trees susceptible to high variance if they are not pruned. This high variance can be harnessed and reduced by creating multiple trees with different samples of the training data set (different views of the problem) and combining their predictions. This approach is called bootstrap aggregation or bagging for short.

A limitation of bagging is that the same greedy algorithm is used to create each tree, meaning that it is likely that the same or very similar split points will be chosen in each tree making the different trees very similar (trees will be correlated). This, in turn, makes their predictions similar, mitigating the variance originally sought.

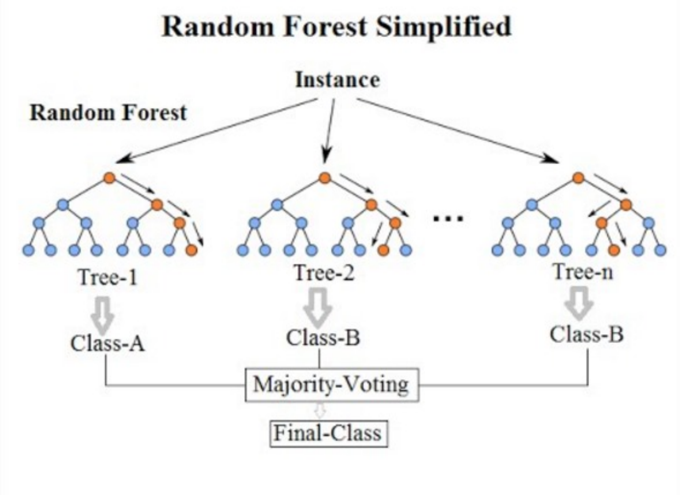


Fig. 4: Random Forest Algorithm

We can force the decision trees to be different by limiting the features (rows) that the greedy algorithm can evaluate at each split point when creating the tree. This is called the Random Forest algorithm. Like bagging, multiple samples of the training dataset are taken and a different tree is trained on each. The difference is that at each point a split is made in the data and added to the tree, only a fixed subset of attributes can be considered.

The basic tenet of Random Forest is that a robust and accurate predictive model can be produced by combining several weak learners, in this case, decision trees. When compared to individual decision trees, the algorithm is less prone to overfitting, has a high degree of flexibility, and performs well with large datasets. It also offers a feature importance metric, which aids in determining the variables that have the greatest influence during the prediction process. Random Forest's ability to handle complex relationships in data and produce accurate predictions has led to its widespread use in various domains, including image analysis, finance, and healthcare. Hence using random forest algorithm for this problem is one of the best solutions.

In machine learning, The linear relationship between a dependent variable and one or more independent features is calculated using the supervised machine learning algorithm known as linear regression. Multivariate linear regression is used when there are multiple independent features; univariate linear regression is used when there is just one independent feature. Finding the best Fit Line equation that can forecast the values based on the independent variables is the algorithm's main goal. Two primary categories of linear regression exist: Simple Linear Regression With just one independent variable and one dependent variable, this type of linear regression is the most basic. Multiple independent variables and one dependent variable are used in Multiple linear regression. Regression analysis uses a set of records containing X and Y values to learn a function. This function can then be applied to predict Y from an unknown X. To obtain the value of Y in a regression given X as independent characteristics, a function that predicts continuous Y is needed.

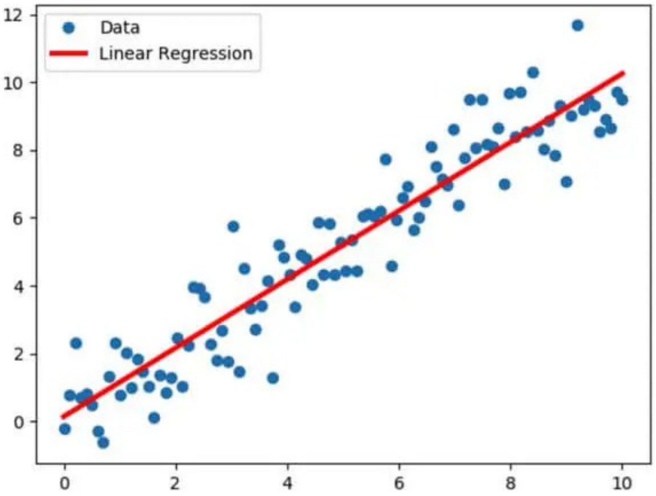


Fig. 5: Slope of the graph using Linear Regression

Finding the best-fit line is our main goal when using linear regression, which suggests that the error between the predicted and actual values should be as small as possible. The best-fit line will have the least amount of inaccuracy. The relationship between the dependent and independent variables is represented by a straight line in the best Fit Line equation. How much the dependent variable varies for a unit change in the independent variable(s) is indicated by the slope of the line.

Since linear regression is a comparatively simple technique, it is straightforward to comprehend and use. Understanding the links between variables can be gained by interpreting the coefficients of the linear regression model as the change in the dependent variable for a one-unit change in the independent variable. Large datasets can be handled with ease via linear

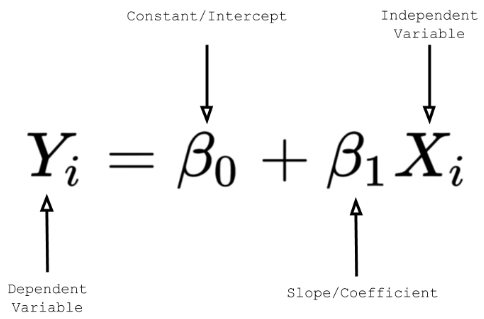


Fig. 6: Formula for Linear Regression

regression, which is also computationally effective. It can be trained on big datasets quickly, which qualifies it for real-time applications. In comparison to other machine learning techniques, linear regression is comparatively resistant to outliers. The effect of outliers on the performance of the model as a whole might be less. When comparing the approach to more intricate machine learning algorithms, it frequently provides a useful starting point model.

A strong open-source machine learning tool is XGBoost. It combines decision trees and gradient boosting to assist in creating better models. After training a model with training data, it assesses the model using fresh data. Until the model stops becoming better, this process is repeated. Model performance and execution speed are the two main uses of XGBoost. Execution speed is critical since handling big datasets requires it. You can deal with datasets that are larger than what would be feasible to utilize with other algorithms when you use XGBoost because there are no limitations on the size of your dataset. Because it enables you to build models that can outperform other models, model performance is also crucial. XGBoost has been contrasted with other algorithms, including Linear Regression and Random Forest.

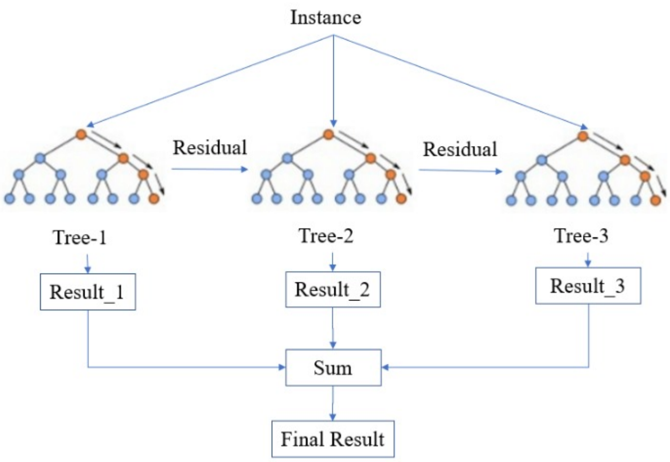


Fig. 7: XGBoost Flowchart

An ML approach called gradient boosting builds a sequence of models and then aggregates them to produce a final model that is more accurate than any of the individual models. It supports challenges involving predictive modeling for both classification and regression. Gradient boosting, a gradient descent algorithm, is used to add new models to an existing one. The XGBoost package, sometimes referred to as multiple additive regression trees, stochastic gradient boosting, or gradient boosting machines, implements gradient boosting.

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2$$

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}[l(y_i, \hat{y}^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2}h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

Fig. 8: XGBoost Formula

XGBoost has a proven track record of delivering excellent outcomes for a range of machine learning workloads. XGBoost is very adjustable because to its vast variety of hyperparameters that may be changed to maximize performance. In contrast to certain machine learning algorithms that may be challenging to comprehend, XGBoost offers feature importances, facilitating a more comprehensive grasp of the elements that hold the greatest significance when forming predictions.

C. Implementation Details

The implementation process commences with the crucial step of loading and exploring the dataset, laying the foundation for a comprehensive understanding of its structure. Leveraging the versatile Pandas library, the dataset is loaded, and preliminary Exploratory Data Analysis (EDA) is conducted through the use of functions like head(), describe(), and dtypes. This initial exploration offers valuable insights into the dataset's characteristics. To enhance the dataset's suitability for regression models, irrelevant columns are systematically removed, and categorical features undergo label encoding. This preprocessing step is pivotal for preparing the data for subsequent modeling stages.

A pivotal visualization step follows, wherein a correlation heatmap is generated using Seaborn, providing a graphical representation of the relationships between different features and the target variable ('price'). This heatmap serves as a navigational tool for identifying potential predictors crucial for the subsequent regression models. The feature selection process further refines the dataset by analyzing correlation indices. Features with correlations exceeding predefined thresholds, such as 0.2 and 0.4, are strategically chosen. Additional visualizations, such as pair plots, are employed to glean deeper insights into the interrelationships between the selected features. The data robustness is ensured by addressing outliers through the application of the Interquartile Range (IQR) method. This step aims to maintain the integrity of the dataset by identifying and filtering out anomalous data points.

Further enhancing the dataset's readiness for machine learning models, one-hot encoding is applied to the 'city' feature, transforming categorical data into a format conducive to model training. With the dataset appropriately preprocessed, the subsequent steps involve splitting it into independent (x) and dependent (y) variables. This division lays the groundwork for model training and evaluation. The dataset is then further partitioned into training and testing sets using the train_test_split function, facilitating an effective evaluation of the models' performance.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \qquad \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

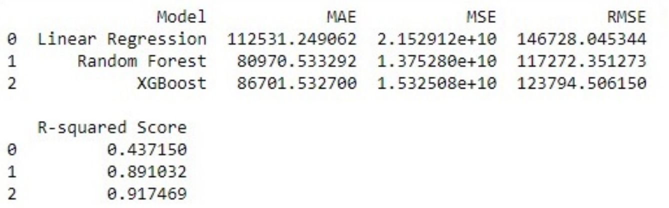
$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \qquad \text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Fig. 9: Mathematical interpretation of RMSE, MAE, R-squared, and MSE

The linear regression model, utilizing sklearn's built-in function, is first applied. To gauge its effectiveness, key performance metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and R-squared (Coefficient of Determination) are calculated. Subsequently, the analysis extends to the Random Forest regressor, employing n_estimators=300 to optimize performance. Similarly, the XGBoost model is implemented with n_estimators=5000, learning rate=0.05, and max_depth=4 to maximize its predictive capabilities. The holistic approach to model implementation, encompassing data preprocessing, feature selection, outlier handling, and model evaluation, underscores a meticulous and systematic strategy. This comprehensive process ensures that the final models, particularly the Random Forest and XGBoost regressors, are well-optimized for accurate and robust predictions. The utilization of key performance metrics facilitates a nuanced understanding of each model's strengths and weaknesses, allowing for informed decisions in selecting the most suitable model for house price estimation.

IV. Comparison

The efficacy of the three machine learning algorithms examined in this analysis—Linear Regression, Random Forest, and XGBoost—in forecasting apartment prices varied. XGBoost outperformed the other two models based on multiple important parameters, even though all three models produced statistically significant results.



```
               Model            MAE           MSE           RMSE
0  Linear Regression  112531.249062  2.152912e+10  146728.045344
1      Random Forest   80970.533292  1.375280e+10  117272.351273
2            XGBoost   86701.532700  1.532508e+10  123794.506150

   R-squared Score
0         0.437150
1         0.891032
2         0.917469
```

Fig. 10: Comparison of Models

The comparative analysis of XGBoost, Random Forest, and Linear Regression models for predicting apartment prices revealed intriguing insights into their respective performances. Notably, XGBoost emerged as the frontrunner, exhibiting the lowest Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) values. This outcome implies that, on average, XGBoost's forecasts were more accurate compared to Random Forest and Linear Regression. The superior predictive capabilities of XGBoost can be attributed to its capacity to handle intricate non-linear correlations in the data—a feature that simpler or linear tree-based models may struggle to capture as effectively. Beyond mere error metrics, the evaluation extended to the R-squared score, where XGBoost once again outshone its counterparts. A higher R-squared score indicates that XGBoost accounts for a greater share of the variation in apartment prices. This underscores the model's robustness in capturing complex relationships within the data, further affirming its efficacy in predictive accuracy.

However, despite its prowess, it's crucial to acknowledge potential shortcomings associated with XGBoost. The model's complexity, stemming from its ensemble nature, might pose challenges in terms of interpretability. Understanding the underlying factors influencing predictions could be intricate, limiting its applicability in scenarios where transparency and interpretability are paramount. In contrast, Linear Regression, while highly interpretable, falls short in modeling complex relationships, leading to larger error values. It provides clarity in understanding the linear relationships between variables but may lack the sophistication needed to grasp intricate patterns in the data.

Random Forest, positioned as a middle ground between interpretability and accuracy, demonstrated better accuracy compared to Linear Regression. However, it still trailed behind XGBoost on various measures. Random Forest strikes a balance by offering fair interpretability while delivering enhanced accuracy over simpler models like Linear Regression. Ultimately, the choice of the most suitable model hinges on the specific goals and constraints of the study or application. If interpretability is of paramount importance, Linear Regression or Random Forest may be preferred. However, if the primary objective is to achieve superior forecast accuracy, XGBoost stands out as the most efficient choice among the three algorithms scrutinized in this analysis. This nuanced evaluation underscores the trade-offs involved in model selection, emphasizing the need to align the chosen approach with the particular requirements and priorities of the given task. The careful consideration of interpretability, complexity, and predictive accuracy ensures that the selected model aligns optimally with the overarching goals of the analysis.

V. Future Directions

In the future, the project aims to expand its scope by using advanced machine learning methods, such as deep learning models, to improve prediction accuracy is one line of inquiry. A more thorough knowledge of the factors impacting home prices may also be possible by taking into account the dynamic nature of real estate markets, combining real-time data, and utilizing cutting-edge data sources like satellite imagery or sentiment analysis of regional economic trends. Moreover, the use of ensemble learning techniques, which combine the advantages of several models, can provide better generalization and robustness. It will be essential to continuously assess and compare these improved models against the backdrop of changing statistics to guarantee their efficacy in capturing the nuances of the housing market and producing more accurate and trustworthy predictions. Further research into neural networks, ensemble approaches, and advanced regression algorithms will shed insight on various approaches to home price prediction. Feature engineering and other relevant data sources will be integrated, and the model's resistance to temporal fluctuations will be assessed, increasing its predictive potential. Interpretability issues will also be addressed by including model-agnostic interpretability tools and balancing prediction accuracy and flexible decision-making. All of these actions are intended to improve and modify the prediction model to address the changing dynamics and problems of the housing market.

VI. Conclusion

In conclusion, we aimed to estimate house prices using three distinct regression models: XGBoost, Random Forest, and Linear Regression. Our analysis revealed notable differences in their performance indicators. Linear Regression exhibited a significant margin of error in both mean absolute error (MAE) and root mean square error (RMSE), despite a reasonable R-squared score of approximately 0.44. In contrast, Random Forest outperformed Linear Regression, showcasing enhanced accuracy with lower MAE, MSE, and RMSE values, along with an impressive R-squared score of about 0.89.

However, the standout performer was XGBoost, boasting an exceptional R-squared score surpassing 0.92 and demonstrating balanced accuracy across all evaluation measures. These findings emphasize the significance of employing advanced ensemble learning techniques for precise and robust predictions in intricate tasks such as house price estimation. The superiority of XGBoost suggests its efficacy in handling the intricacies of real estate data and capturing nuanced relationships, making it the preferred choice among the evaluated regression models. This project not only advances our understanding of machine learning applications in real estate but also underscores the practical importance of choosing the right algorithm for accurate and reliable house price predictions.

References

[1] P. Durganjali and M. V. Pujitha, ”House Resale Price Prediction Using Classification Algorithms,” 2019 International Conference on Smart Structures and Systems (ICSSS), Chennai, India, 2019, pp. 1-4, doi: 10.1109/ICSSS.2019.8882842.

[2] 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.

[3] Phan, The. (2018). Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia. 35-42. 10.1109/iCMLDE.2018.00017.

[4] GeeksforGeeks, XGBoost - Introduction to Machine Learning Algorithm, https://www.geeksforgeeks.org/xgboost/.

[5] Analytics Vidhya, Understand Random Forest Algorithms With Examples, https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/.

[6] House price prediction, SHREE, https://www.kaggle.com/datasets/shree1992/housedata.