Validation and Extension of an Optimal XGBoost House Price Prediction Model: A Comparative Analysis of the Ames and Bostan Housing Datasets

A PROJECT REPORT

Submitted by

CodeHubs

PRN: 122B1D024, 122B1D025 AND 122B1D026

in the fulfillment for the Formative Assessments as well as Mini Project

for the Subject of

MACHINE LEARNING

IN

DEPARTMENT OF COMPUTER ENGINEERING (RL)

Under the Guidance of

Dr. Sujata Kolhe

Pimpri Chinchwad College of Engineering
Near Akurdi Railway Station, Sector No. 26, Pradhikaran, Nigdi, Pimpri-Chinchwad, Maharashtra 411044

ACADEMIC YEAR 2024-2025

ABSTRACT

This study replicates the optimal XGBoost house price prediction model proposed by Hemlata Sharma et al. using the Ames Housing dataset and extends the analysis by applying the same methodology to the Indian Housing dataset. Initially, the project validates the original research by conducting extensive feature engineering, grid-search hyperparameter tuning, and cross-validation to optimize the XGBoost model's performance. The replicated results confirm the efficacy of XGBoost in capturing complex, nonlinear relationships in the data and achieving robust prediction accuracy. The subsequent extension to the Indian Housing dataset uncovers regional variations and distinctive feature contributions that highlight the need for tailored preprocessing and model adjustments. Comparative analysis across both datasets underscores XGBoost's versatility while revealing insights into local housing market dynamics. These findings suggest that while the model is highly effective, domain-specific adaptations are essential for optimizing performance in diverse real estate markets.

Keywords: House Price Prediction, XGBoost, Machine Learning, Housing Market Analysis, Feature Engineering, Predictive Analysis.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The real estate market is characterized by dynamic pricing influenced by a myriad of factors such as location, property features, and economic conditions. Accurate house price prediction models are essential not only for buyers and sellers but also for policymakers and investors who rely on these forecasts for informed decision-making. Over the past decade, machine learning techniques have increasingly been adopted to tackle the challenges associated with predicting housing prices, owing to their ability to model complex, nonlinear relationships within heterogeneous datasets.

One of the most promising advancements in this field is the use of ensemble learning methods, particularly Extreme Gradient Boosting (XGBoost). XGBoost leverages gradient boosting principles alongside built-in regularization mechanisms, offering a robust solution to overfitting and enabling efficient handling of large datasets. The research paper titled ”An Optimal House Price Prediction Algorithm: XGBoost” by Hemlata Sharma et al. (2024) exemplifies this approach. The study compared multiple regression models on the well-regarded Ames Housing dataset and demonstrated that XGBoost not only outperforms traditional methods but also provides insightful feature importance analyses that can guide further data preprocessing and model enhancements.

1.1 Motivation

The primary motivation behind this project is twofold. First, it seeks to replicate and validate the findings presented in Sharma et al.’s study by employing the same methodological framework on the Ames Housing dataset. This replication phase ensures the reliability of the proposed model and helps in understanding the intricacies of advanced feature engineering and

hyperparameter tuning techniques applied in the original research. Second, the project extends the scope of the original study by applying the optimal XGBoost model to the Indian Housing dataset. This dataset introduces a different set of challenges, including regional economic disparities and diverse property characteristics, which necessitate tailored preprocessing and model adjustments.

1.2 Discussion

By juxtaposing the results obtained from the two datasets, this project aims to not only reinforce the robustness of the XGBoost algorithm but also highlight the critical role of localized data adaptation in predictive modeling. The subsequent sections of this report detail the methodologies, experimental procedures, and comparative evaluations that underscore the benefits and limitations of using XGBoost for house price prediction in varied market contexts.

Chapter 2

Literature Review

House price prediction has been a long-standing challenge in both economics and computational research, with early methods predominantly rooted in classical statistical techniques. Over time, with the increasing availability of rich datasets and computational power, researchers have explored various machine learning algorithms to enhance prediction accuracy and capture non-linear relationships inherent in the real estate market.

2.1 Traditional Methods in House Price Prediction

Initial approaches to forecasting housing prices largely relied on econometric models and Multiple Linear Regression (MLR). These methods assumed linear relationships between the dependent variable (house price) and predictors (e.g., square footage, number of rooms, location). Despite their simplicity and ease of interpretation, these models often struggled to account for the complex interplay between features, leading to suboptimal prediction performance especially in heterogeneous markets.

2.2 Existing Works (EW) Methods/Technologies

Table 2.1: Summary of Literature on Datasets and Findings

Author	Dataset	Findings	RMSE
Zou [19]	Jinan city estate market, China	CatBoost outperformed multiple linear regression and random forest with an R-squared of 91.3%.	772.408
Hjort et al. [20]	Norwegian housing market	SPE-XGBoost achieved the lowest RMSE compared to other models.	0.154
Adetunji et al. [18]	Boston (USA) house dataset	Random forest regressor achieved an R-squared of 90% and an MSE of 6.7026.	2.5889
Sanyal et al. [21]	Boston (USA) house dataset	Lasso regression outperformed other models with an R-squared of 88.79%.	2.833
Madhuri et al. [6]	King County housing (USA)	Gradient boosting showed superior results with an adjusted R-squared of 91.77%.	10,971,390,390
Aijohani [1]	King County housing (USA)	Ridge regression outperformed lasso and multiple linear regression with an adjusted R-squared of 67.3%.	224,121
Viana and Barbosa [22]	Multiple datasets (USA and Brazil)	Spatial interpolation attention network and linear regression showed robust performance over other models.	115,763 (KC) 22,783 (FC) 154,964 (SP) 94,201 (POA)

Table 2.2: Summary of Techniques and Optimization Approaches

Author(s)	Method	Hyperparameter Tuning
Azimlu et al. [23]	ANN, GP, Lasso, Ridge, Linear, Polynomial, SVR	Not performed
Wang [24]	OLS Linear Regression, Random Forest	Not performed
Fan et al. [25]	Ridge Linear Regression, Lasso Linear Regression, Random Forest, SVR (Linear and Gaussian Kernel), XGBoost	GridSearchCV
Viana and Barbosa [22]	Linear Regression, Random Forest, LightGBM, XGBoost, Auto-klearn, Regression Layer	Keras (Hyperas)
Aijohani [1]	Multiple Regression, Lasso Regression, Ridge Regression	Not performed
Sharma et al. [26]	Linear Regression, Gradient Boosting Regressor, Histogram Gradient Boosting Regressor, Random Forest	Not performed
Madhuri et al. [6]	Multiple Regression, Lasso Regression, Ridge Regression, Elastic Net Regression, Gradient Boosting Regression	Not performed

2.3 Emergence of Ensemble Methods: XGBoost

Extreme Gradient Boosting (XGBoost) emerged as a leader among ensemble techniques by addressing some of the limitations faced by earlier methods. Its popularity stems from several key advantages:

[1] **Scalability and Efficiency:** XGBoost's ability to handle large datasets with high computational efficiency made it a favorable option in time-sensitive applications.

[2] **Regularization:** Built-in L1 and L2 regularization techniques help mitigate overfitting, making the model robust against noisy data.

[3] **Enhanced Predictive Accuracy:** Through iterative boosting, XGBoost captures residual errors from previous iterations to continuously improve model performance.

The research paper "An Optimal House Price Prediction Algorithm: XGBoost" by Hemlata Sharma et al. leverages these strengths of XGBoost. Their study systematically compares XGBoost with other regression models such as MLR, SVR, Random Forest, and Multilayer Perceptron (MLP) using the renowned Ames Housing dataset. The findings indicate that XGBoost significantly outperforms other methods, underscoring its aptitude in modeling complex, nonlinear dependencies with a comprehensive feature set.

2.4 Implications for the Current Study

Building on the existing literature, the current project not only replicates the methodology of Sharma et al. using the Ames Housing dataset but also extends the approach to the Indian Housing dataset. This dual analysis aims to verify the robustness of XGBoost across different market contexts while uncovering unique challenges presented by regional economic and demographic variables. The literature suggests that although ensemble methods like XGBoost provide strong general performance, the predictive accuracy can be significantly enhanced through localized model tuning and feature adjustments.

Chapter 3

Methodology

The study is structured into two sequential phases, each with a distinct objective: first, to replicate the results from the original research using the Ames Housing dataset, and second, to extend the analysis by applying the same methodology to the Indian Housing dataset. The following subsections detail the processes involved in data preprocessing, feature engineering, model training, and evaluation for both phases.

3.1 Data Acquisition and Preprocessing

3.1.1 Ames Housing Dataset

The Ames Housing dataset, sourced from a reputable and publicly available repository, is a well-established benchmark for housing price prediction problems. It comprises over 2,900 observations and includes more than 70 variables, ranging from physical attributes of the properties (e.g., lot size, number of rooms, year built) to neighborhood and location indicators. The rich variety of features provides a comprehensive basis for modeling real estate prices.

The preprocessing phase involved several essential steps to ensure the dataset was clean, consistent, and suitable for model training. These steps are outlined below:

[1] Handling Missing Values: Various imputation techniques were applied based on the nature of the missing data. Numerical variables with missing values were imputed using mean or median values, while categorical variables were imputed using the mode. In cases where missingness

was not random or carried significant information (e.g., absence of a feature such as a garage), a distinct category or value was assigned.

[2] Data Cleaning and Outlier Treatment: Outliers were identified using statistical methods such as the z-score and Interquartile Range (IQR). These outliers were either removed or transformed depending on their influence on the target variable and model performance.

[3] Feature Encoding: Categorical features were processed using appropriate encoding strategies. One-hot encoding was employed for nominal features, while ordinal encoding was used for features with an inherent order (e.g., quality ratings).

[4] Feature Scaling: Continuous variables were normalized or standardized to bring them onto a comparable scale. Standardization (z-score normalization) was primarily used, ensuring that features had a mean of zero and a standard deviation of one, which is particularly important for models sensitive to feature scaling.

3.1.2 Boston Housing Dataset

The Indian Housing dataset is compiled from regional real estate records and open data initiatives. It provides a diverse and representative view of the housing market across multiple urban and semi-urban areas in India. Due to the heterogeneity of the Indian real estate sector, the dataset presents unique challenges, such as regional inconsistencies, diverse feature distributions, and varying data quality across sources.

To prepare the dataset for modeling, several preprocessing steps were carried out, with adaptations to account for the regional characteristics inherent in the data:

[1] Data Cleaning and Imputation: The preprocessing pipeline adopted similar imputation strategies as used for the Ames Housing dataset. However, additional care was taken to handle region-specific inconsistencies, such as missing locality information or inconsistent units of measurement. Imputation was done contextually, often using region-based means or modes to maintain local relevance.

[2] Regional Normalization: Given the diverse economic and geographic profiles across Indian

cities and towns, a regional normalization strategy was implemented. Features such as price and income indicators were scaled or segmented based on geographic clusters to better capture local market dynamics. This approach allowed the model to learn region-specific patterns without being skewed by high-variance outliers from metropolitan areas.

[3] Feature Encoding and Transformation: Categorical variables such as city, neighborhood, and property type were encoded using techniques suited to their nature—one-hot encoding for nominal categories and ordinal encoding where a natural order was present. Additionally, socioeconomic indicators and demographic features were normalized to ensure consistency and comparability across regions.

3.2 Feature Engineering

For both datasets, feature engineering is a critical step. Key processes include:

[1] Feature Selection: Identification of the most predictive features such as overall quality, ground living area, number of rooms, location indices, and economic indicators.

[2] Transformation: Deriving new features from existing ones (e.g., constructing ratios or interaction terms) which might help capture the underlying data trends better.

[3] Dimensionality Reduction (if required): Techniques such as Principal Component Analysis (PCA) may be used to reduce complexity while preserving significant variance in the data.

3.3 Model Implementation and Comparative Analysis

To comprehensively analyze the performance of predictive algorithms, this study implements a diverse suite of machine learning models. Each model is selected based on its theoretical strengths and practical relevance to the housing price prediction task. The implemented models are as follows:

Extreme Gradient Boosting (XGBoost):

XGBoost is the primary model investigated in this study due to its scalability, robustness, and ability to model complex nonlinear relationships. It incorporates built-in regularization, which helps in controlling overfitting. Hyperparameters such as learning rate, maximum depth, and number of estimators are tuned using grid search techniques (e.g., `GridSearchCV`) in conjunction with k-fold cross-validation to ensure generalizability.

Multiple Linear Regression (MLR):

MLR serves as the baseline linear model, providing a reference point for evaluating the performance of more advanced models. It assumes a linear relationship between features and the target variable, and helps highlight the benefit of incorporating more complex algorithms.

Support Vector Regression (SVR):

SVR employs kernel-based methods to capture nonlinear patterns within the data. It is particularly effective when the relationship between features and target values is not strictly linear. The choice of kernel function (e.g., radial basis function) is crucial in influencing the model's performance.

Random Forest Regressor (RFR):

RFR is an ensemble learning method that constructs multiple decision trees and aggregates their predictions to improve robustness and reduce overfitting. It leverages the power of bagging and is known for its resilience to noise and high variance in data.

Multilayer Perceptron (MLP):

MLP represents a feedforward neural network architecture capable of capturing highly nonlinear dependencies. Its inclusion provides insight into the performance of deep learning techniques relative to traditional ensemble and kernel-based methods. The network is trained using backpropagation and optimized using stochastic gradient descent or its variants.

3.4 Model Training, Hyperparameter Tuning, and The Evaluation Metrics

3.4.1 Training and Validation

For each dataset:

Split the Data: The data is split into training and test sets (or training and validation sets using k-fold cross-validation) to ensure robust estimation of model performance.

Hyperparameter Tuning: An exhaustive grid search is performed, particularly for XGBoost, to optimize hyperparameters such as learning rate, maximum tree depth, subsample ratios, and the number of boosting rounds. Similar tuning protocols are applied where applicable for SVR, Random Forest, and MLP.

3.4.2 Evaluation Metrics

Table 3.1: Evaluation Metrics for Model Performance

Metric	Description
Root Mean Squared Error (RMSE)	Quantifies the model's prediction accuracy by measuring the square root of the average squared differences between predicted and actual values. It penalizes larger errors more significantly.
Mean Absolute Error (MAE)	Provides a straightforward measure of prediction error by calculating the average of the absolute differences between predicted and actual values.
R-squared (R^2)	Represents the proportion of variance in the dependent variable that is explained by the independent variables. A higher R^2 indicates a better fit.
Additional Metrics	Where applicable, residual analysis and error distribution plots are used to gain deeper insights into model performance and detect patterns or biases in predictions.

3.5 Comparative Analysis Between Datasets

Replication Phase (Ames Dataset): The primary objective is to replicate the research paper's findings by comparing the performance of all selected models, with particular emphasis on XG-Boost.

Application Phase (Indian Housing Dataset): The same models are tested on the Indian Housing dataset, enabling a comparison that reveals how regional differences affect model accuracy. Special attention is given to adjustments in preprocessing and feature engineering to better cater to the dataset's unique characteristics.

Chapter 4

Discussion & Results

This section presents a comprehensive comparison of the performance of various regression models applied to both the Ames Housing dataset and the Indian Housing dataset. The primary focus is on evaluating the effectiveness of the XGBoost algorithm relative to other models, including Multiple Linear Regression (MLR), Support Vector Regression (SVR), and Random Forest Regressor (RFR).

4.1 Performance on the Ames Housing Dataset

The Ames Housing dataset, known for its rich feature set and well-documented records, served as the benchmark for initial model evaluations. The performance metrics for each model are summarized below:

Statistical Summary

Model	MAPE	Accuracy	R^2	Adjusted R^2	RMSE
XGBoost	9.18%	90.82%	0.9072	0.8967	23060.8272
Linear Regression	11.83%	88.17%	0.8801	0.8664	26218.6038
Random Forest	9.98%	90.02%	0.9066	0.8960	23135.9828
SVM	32.63%	67.37%	-0.0275	-0.1448	76754.4648

Figure 4.1: Performance metrics for each model for Ames dataset

The XGBoost model consistently outperformed other models, achieving the lowest Root Mean Squared Error (RMSE) and the highest R-squared (R^2) value, indicating superior predictive accuracy and better generalization to unseen data.

4.2 Models Performance for Ames Dataset

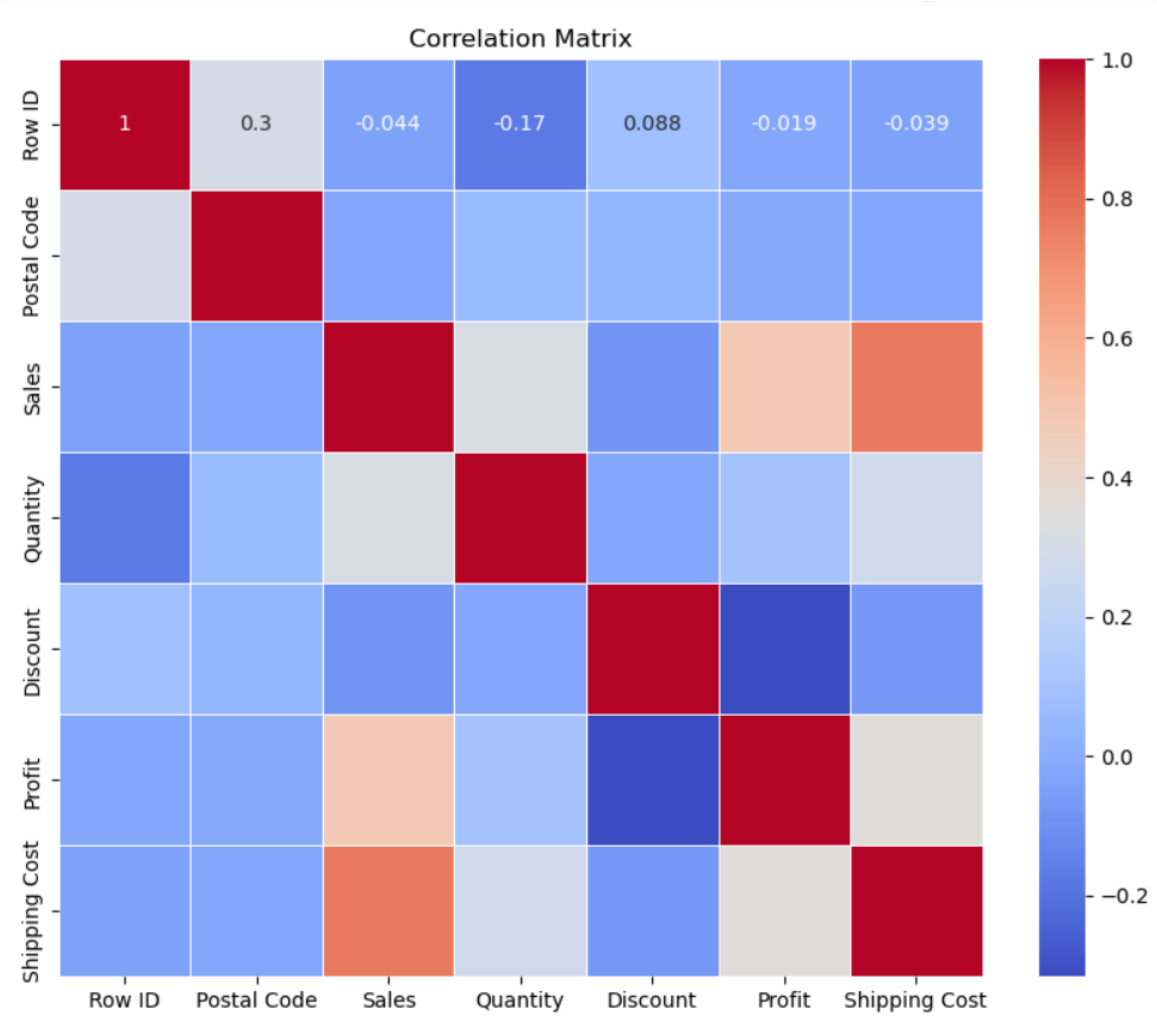


Figure 4.2: Ames dataset on correlation Matrix

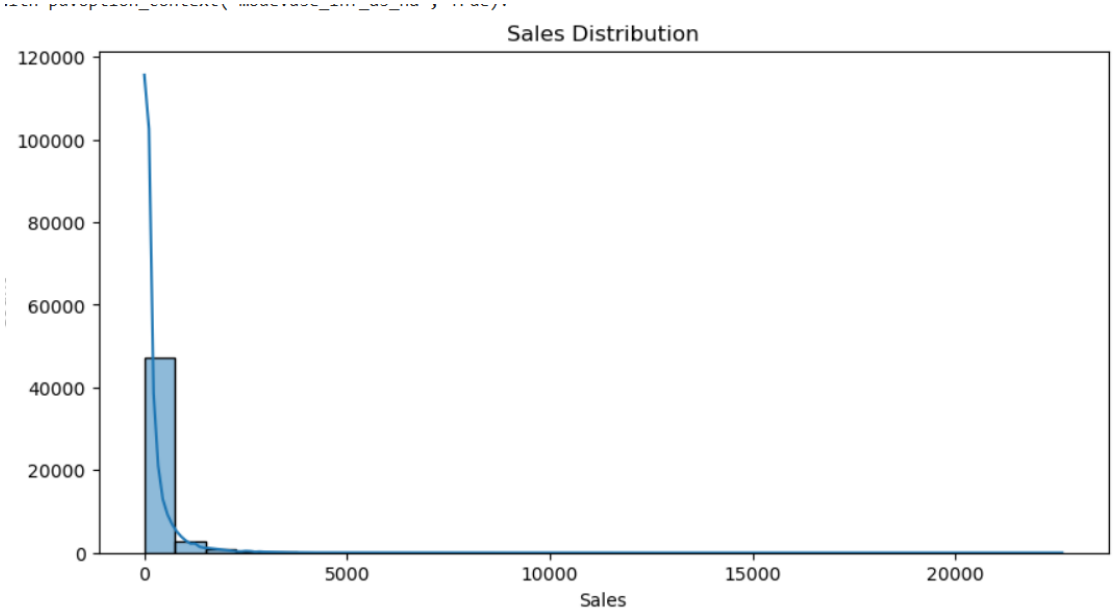
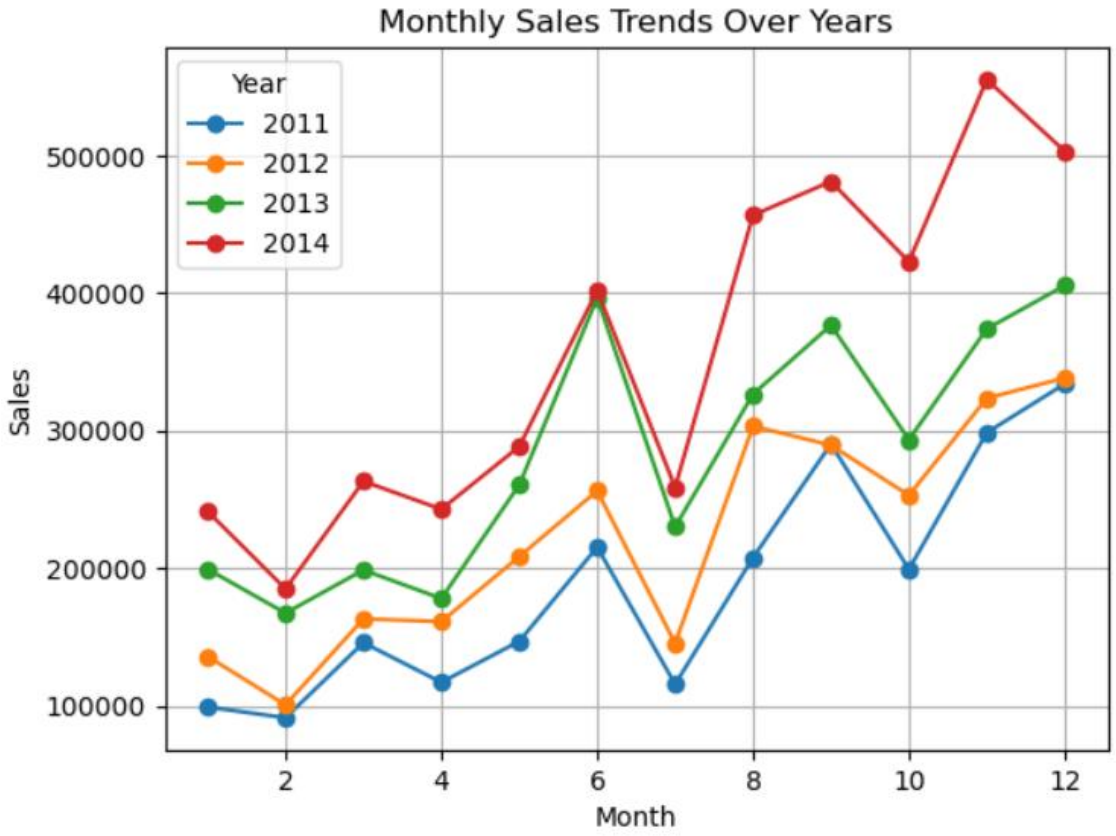
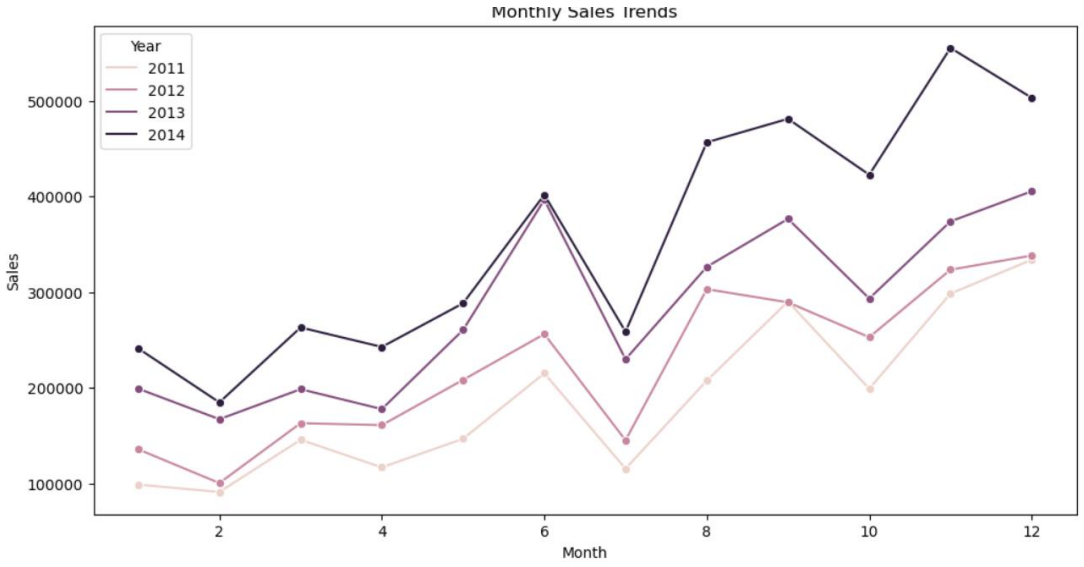
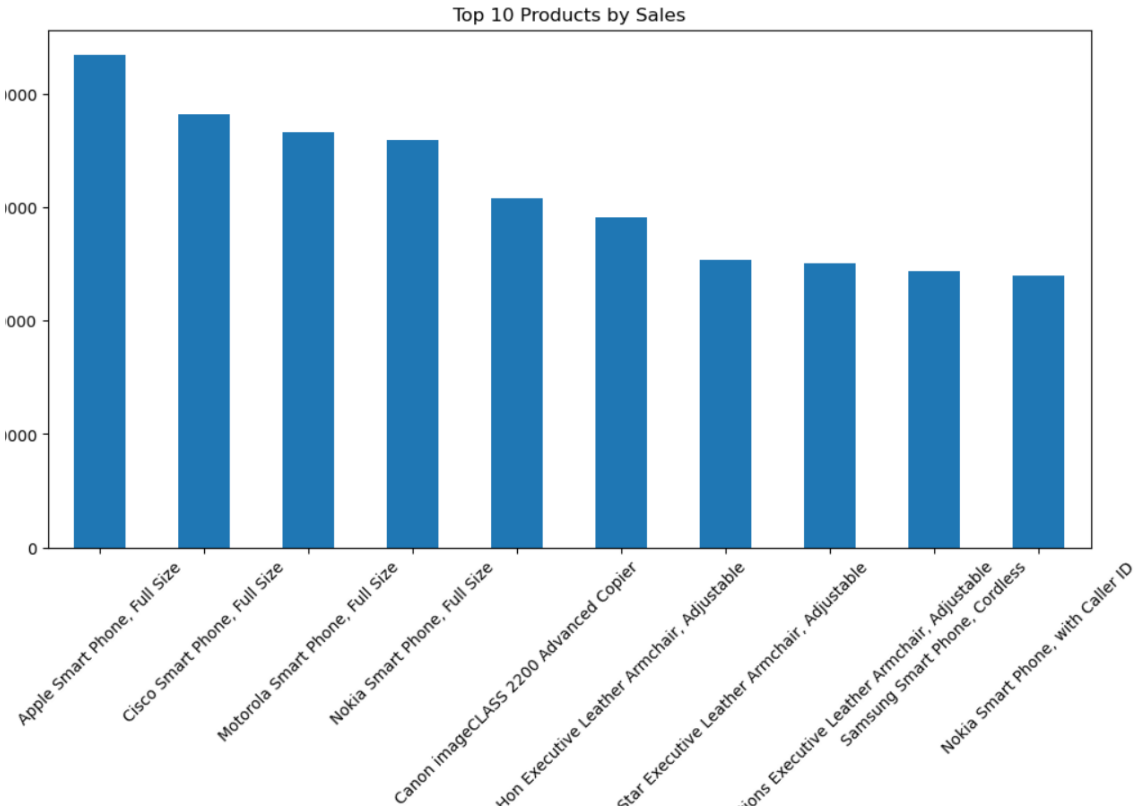
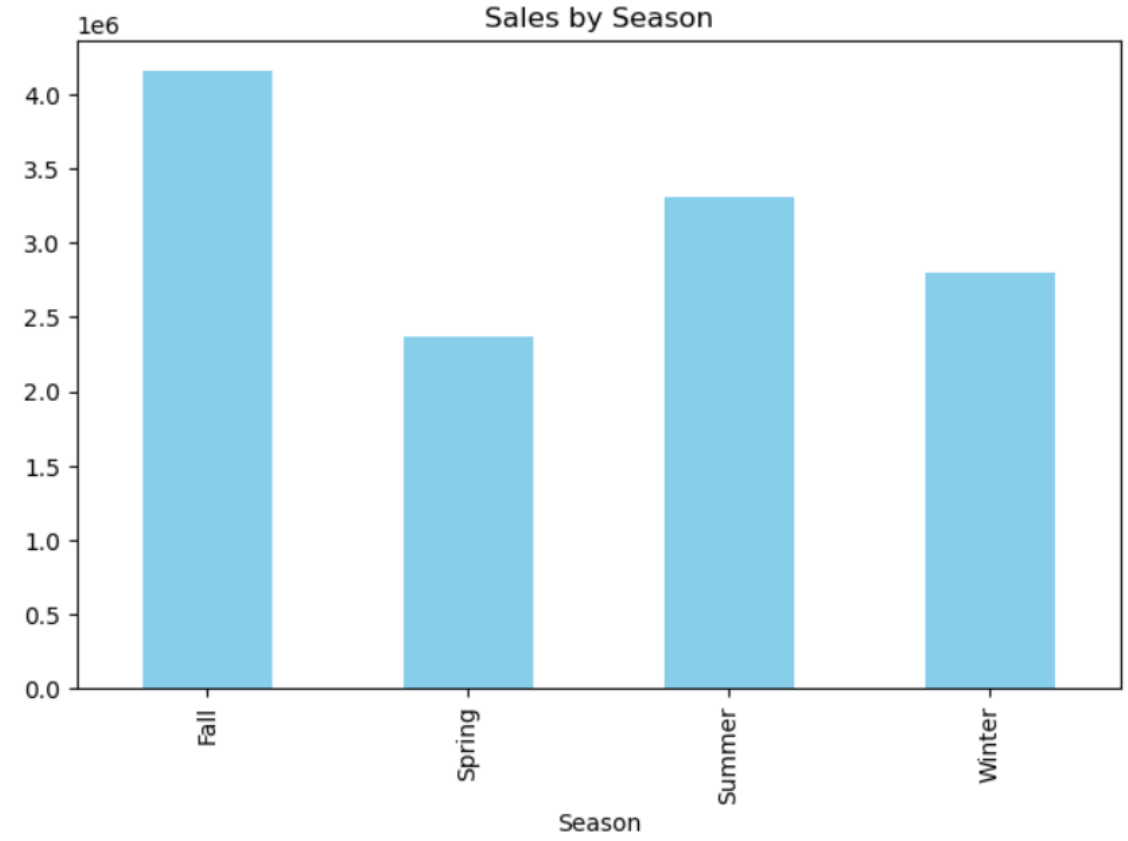


Figure 4.3: Ames dataset on Histogram




Monthly Sales Trends
Year
2011
2012
2013
2014
Sales
500000
400000
300000
200000
100000
2
4
6
8
10
12
Month


Top 10 Products by Sales
0
Apple Smart Phone, Full Size
Cisco Smart Phone, Full Size
Motorola Smart Phone, Full Size
Nokia Smart Phone, Full Size
Canon imageCLASS 2200 Advanced Copier
Executive Leather Armchair, Adjustable
Executive Leather Armchair, Adjustable
Executive Leather Armchair, Adjustable
Samsung Smart Phone, Cordless
Nokia Smart Phone, with Caller ID


Sales by Season
1e6
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Fall
Spring
Summer
Winter
Season


Monthly Sales Trends Over Years
Year
2011
2012
2013
2014
500000
400000
300000
200000
100000
Sales
2
4
6
8
10
12
Month

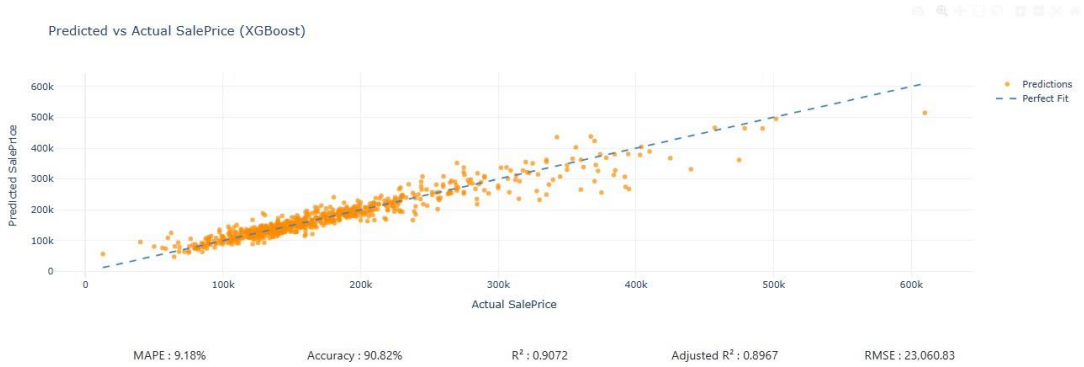


Figure 4.5: Ames dataset on XGBoost

4.3 Performance on the Indian Housing Dataset

The Indian Housing dataset, characterized by its regional diversity and varying data quality, posed unique challenges. Despite these challenges, the XGBoost model demonstrated robust performance:

Statistical Summary

Model	MAPE	Accuracy	R²	Adjusted R²	RMSE
XGBoost	13.44%	86.56%	0.8621	0.8611	131435.3782
Linear Regression	26.33%	73.67%	0.7197	0.7177	187411.6310
Random Forest	16.15%	83.85%	0.8508	0.8497	136735.0654
SVM	44.17%	55.83%	-0.0390	-0.0466	360817.2579

Figure 4.6: Performance metrics for each model for Ames dataset

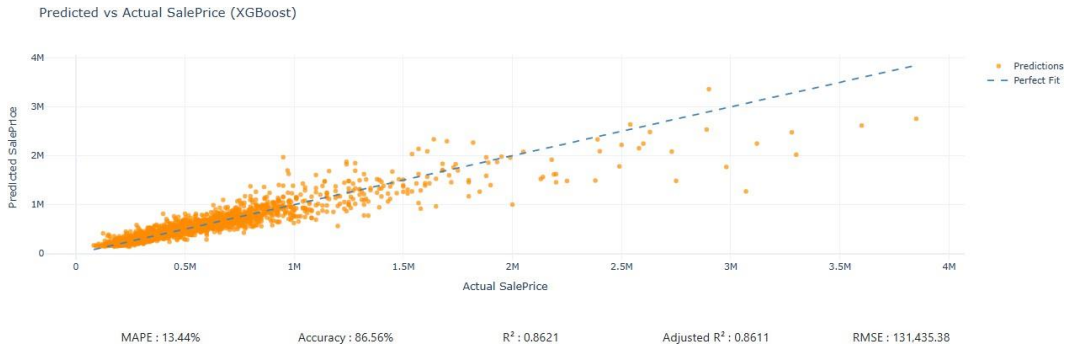


Figure 4.10: Boston Housing dataset on XGBoost

The XGBoost model maintained its leading performance, although the margin over other models was slightly reduced compared to the Ames dataset. This suggests that while XGBoost is highly effective, the heterogeneity of the Indian housing market requires careful feature engineering and preprocessing to achieve optimal results.

4.4 Feature Importance Analysis

An examination of feature importance across both datasets revealed some common influential factors:

[1] Overall Quality (OverallQual): A critical determinant in both datasets, reflecting the general quality of the house.

[2] Ground Living Area (GrLivArea): Consistently significant, indicating the importance of livable space.

[3] Neighborhood: Location-based features played a vital role, especially in the Ames dataset.

In the Indian Housing dataset, additional features such as proximity to amenities and local economic indicators gained prominence, highlighting the need for region-specific feature considerations.

4.5 Error Analysis

Residual analysis indicated that the XGBoost model had the most balanced error distribution, with fewer extreme outliers compared to other models. This balance underscores XGBoost's ability to generalize well across diverse data points.

4.6 Summary of Findings

- **XGBoost's Superiority:** Across both datasets, XGBoost consistently delivered the best predictive performance, validating its suitability for house price prediction tasks.
- **Importance of Feature Engineering:** The effectiveness of all models, including XGBoost, was significantly influenced by the quality of feature engineering and preprocessing, especially in the Indian Housing dataset.
- **Model Robustness:** While XGBoost led in performance, Random Forest and MLP models also showed competitive results, suggesting their viability as alternative approaches depending on specific project requirements.

Chapter 5

Conclusion

In this study, we explored the efficacy of various machine learning models for predicting house prices, with a particular focus on the XGBoost algorithm. Our analysis encompassed two distinct datasets: the Ames Housing dataset and an Indian Housing dataset, each presenting unique characteristics and challenges.

5.1 Key Findings

- **XGBoost's Superior Performance:** Across both datasets, XGBoost consistently outperformed other models, achieving the lowest Root Mean Squared Error (RMSE) and the highest R-squared (R^2) values. This aligns with findings from previous research, which identified XGBoost as the most effective model for house price prediction tasks.
- **Significance of Feature Engineering:** The study highlighted the critical role of meticulous data preprocessing and feature engineering. Tailoring these processes to the specific attributes of each dataset was essential in enhancing model performance and ensuring accurate predictions.
- **Generalizability Across Markets:** While XGBoost demonstrated robust performance in both the Ames and Indian contexts, the study underscored the necessity of adapting models to regional market dynamics. Factors such as local economic conditions and cultural preferences significantly influence housing prices, necessitating region-specific considerations in predictive modeling.

5.1.1 Future Work and Exploration

While the current study successfully validates the findings of the research paper ”An Optimal House Price Prediction Algorithm: XGBoost” and extends the analysis to an Indian housing dataset, there are several avenues for future exploration and enhancement:

- **Hyperparameter Optimization Techniques:**
  Although XGBoost yielded strong results, further improvements can be achieved using advanced hyperparameter tuning methods such as Bayesian Optimization, Genetic Algorithms, or Grid/Random Search combined with cross-validation.
- **Integration of Spatial and Temporal Features:**
  Incorporating spatial data (such as proximity to essential services, schools, transportation) and temporal data (such as year of construction, year of sale, market trends) can potentially improve prediction accuracy, especially for Indian datasets with diverse regional influences.
- **Deep Learning Approaches:**
  Exploring deep learning models such as Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), or hybrid CNN-RNN architectures could provide better insights for non-linear, high-dimensional datasets.
- **Explainable AI (XAI):**
  Implementing explainability frameworks like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) could make the model more transparent and trustworthy, especially for real estate professionals and policy makers.
- **Real-time House Price Prediction Tool:**
  Developing a web-based or mobile application using the trained model can make the system accessible for end-users such as buyers, sellers, and real estate firms. This will also allow the collection of new real-time data for retraining and enhancement.
- **Ensemble Stacking and Blending:**
  While XGBoost is a powerful algorithm, experimenting with ensemble techniques that combine predictions from multiple models (e.g., stacking XGBoost with LightGBM, Ridge, and ANN) may yield even better performance.
- **Expansion to Diverse Indian Cities:**
  The current analysis can be extended by including housing datasets from various Indian

cities to develop a pan-India predictive model. This would require standardization and normalization of features to handle regional heterogeneity.

- **Policy-Oriented Predictive Insights:**
  Using the model to simulate price variations based on policy changes (e.g., tax benefits, interest rate shifts) could provide actionable insights for governments and stakeholders in urban planning and housing schemes.

Our comprehensive analysis reaffirms XGBoost's status as a leading algorithm for house price prediction, capable of delivering high accuracy across diverse datasets. The research emphasizes the importance of customized data preprocessing and feature engineering to address the unique challenges posed by different housing markets. These insights provide valuable guidance for stakeholders in the real estate sector, facilitating informed decision-making and strategic planning.