St. Lawrence College

Final Project: AI Application for Housing Price Prediction

By
Team Members

Kaipu Sai Priya – 4368386
Nitika – 4373067
Sai Kiran Varada– 4370413
Harmanpreet Singh- 4370867

Submitted To
Professor: Sujoy Paul

ADMN5016-F23-101: Applied Artificial Intelligence and Machine Learning.

April: 20th, 2024

Table of Contents

AI Application for Housing Price Prediction

1. Introduction

Housing values are crucial for purchasers, sellers, investors, and real estate agents. The way that individuals approach real estate deals may be significantly impacted by accurate forecasts. The goal of this study is to apply machine learning to forecast the median value of occupied homes (MEDV) using a housing dataset. Stakeholders can make better decisions if they know the factors influencing these prices.

We've selected the Housing Prediction dataset from Kaggle. This choice was made after careful consideration; the dataset appears to offer a robust foundation for accurate predictions. Moreover, it garnered unanimous agreement from our team, signifying its potential and relevance to our project goals.

2. Project Overview

2.1 About the Project

The project aims to forecast home prices using machine learning. The housing dataset includes a variety of variables, such as property taxes, crime rates, community features, and the number of rooms, that affect housing costs. Making more educated judgments can be facilitated by precise forecasts for investors, homeowners, and real estate agents.

2.2 Application Purpose

Predictive Analysis: The program evaluates a range of criteria, such as the number of rooms, crime rates, property taxes, and building ages, using machine learning to estimate house prices.

Market insights: By predicting home values, the application provides valuable information regarding valuation trends and market trends.

Investment Decisions: Adhering to precise projections may help real estate investors make more profitable decisions.

2.3 Criticality of Housing Price Prediction

Influence on Real Estate: Purchasing, selling, and renting are just a few of the ways that housing prices impact real estate transactions.

Economic Impact: Local economies are impacted by housing prices, which can have an impact on property taxes and municipal budgets.

Resource Allocation: Knowledge of house pricing patterns helps facilitate the real estate sector's resource allocation.

2.4 Market Size

The real estate market is extensive and global. Accurate prediction tools can significantly impact real estate transactions worldwide, aiding millions of transactions annually. The potential market for this application includes real estate agencies, investment firms, and individual investors.

Global Reach: Accurate home price forecasts from this program might be beneficial to clinics, hospitals, and real estate agents all around the world.

Economic Impact: Precise forecasts can enhance resource allocation in the real estate sector and lower the risk of real estate investments.

Integration with Current Systems: By integrating with electronic real estate systems, the application may improve data-driven decision-making.

2.5 Financial Impact

Cost Savings: By providing accurate housing price predictions, the application can prevent investment losses that would occur from overpriced or underpriced property sales. For instance, if the application prevents a real estate company from overpricing 100 homes (assuming an average loss of $10,000 per overpriced home), it could save $1,000,000.

Risks: The main risk includes potential inaccuracies in predictions due to outliers or data anomalies that were not removed, leading to possible financial losses or misguided investments.

3. Data Overview

3.1 Dataset Description

The dataset consists of 509 entries, each describing various attributes of housing in the Boston area. It provides a broad range of variables that are influential in determining the median value of owner-occupied homes, labeled as MEDV.

3.2 Key Features

Here's a breakdown of the key features included in the dataset:

- **CRIM**: Per capita crime rate by town.
- **ZN**: Proportion of residential land zoned for lots over 25,000 sq. ft.
- **INDUS**: Proportion of non-retail business acres per town.
- **CHAS**: Charles River dummy variable (1 if tract bounds river; 0 otherwise).
- **NOX**: Nitrogen oxides concentration (parts per million).

- **RM**: Average number of rooms per dwelling.
- **AGE**: Proportion of owner-occupied units built prior to 1940.
- **DIS**: Weighted distances to five Boston employment centers.
- **RAD**: Index of accessibility to radial highways.
- **TAX**: Full-value property tax rate per $10,000.
- **PTRATIO**: Pupil-teacher ratio by town.
- **B**: 1000(Bk - 0.63) ^2 where Bk is the proportion of Black people by town.
- **LSTAT**: Percentage of lower status of the population.

Each variable serves as an indicator of housing value, where features like RM (average number of rooms per dwelling) and TAX (tax rate) can significantly impact the price. Variables like CHAS indicate proximity to natural features which may enhance property value.

4. Data Preprocessing

The process of preprocessing data involves preparing the dataset for machine learning. It entails addressing missing numbers, eliminating outliers, and cleansing the data. We followed these steps to get the data ready:

Managing Missing Values

The dataset had several missing values, especially in numerical characteristics like AGE and RM. We fixed this by appending mean or median values, guaranteeing that the data was comprehensive and coherent. It was imperative to handle the missing values, which accounted for around 5% of the data.

Eliminating Outliers

In our analysis of the housing dataset, particularly for features like CRIM (crime rate) and TAX (tax rate), we utilized the Interquartile Range (IQR) method to identify outliers. This statistical method calculates the IQR as the difference between the 75th percentile (Q3) and the 25th percentile (Q1), applying this method, more than twenty potential outliers were identified.

However, we chose not to remove these outliers from our dataset. This decision was based on the nature of the data and the modeling techniques employed. We recognized that these extreme values might not be anomalies but realistic representations of areas with high crime rates or unusual tax circumstances. Removing these data points could strip away real-world complexities and valuable insights into how these factors influence housing prices.

Moreover, the robust regression models used, such as Random Forest and XGBoost, are equipped to handle outliers through techniques like bootstrapping and regularization, which can mitigate the influence of extreme values without excluding them from the dataset. Retaining these outliers allows our model to maintain comprehensive data integrity and enhances its ability to generalize well across different scenarios, ensuring that it remains effective and reliable in predicting real-world housing prices.

Engineering Features

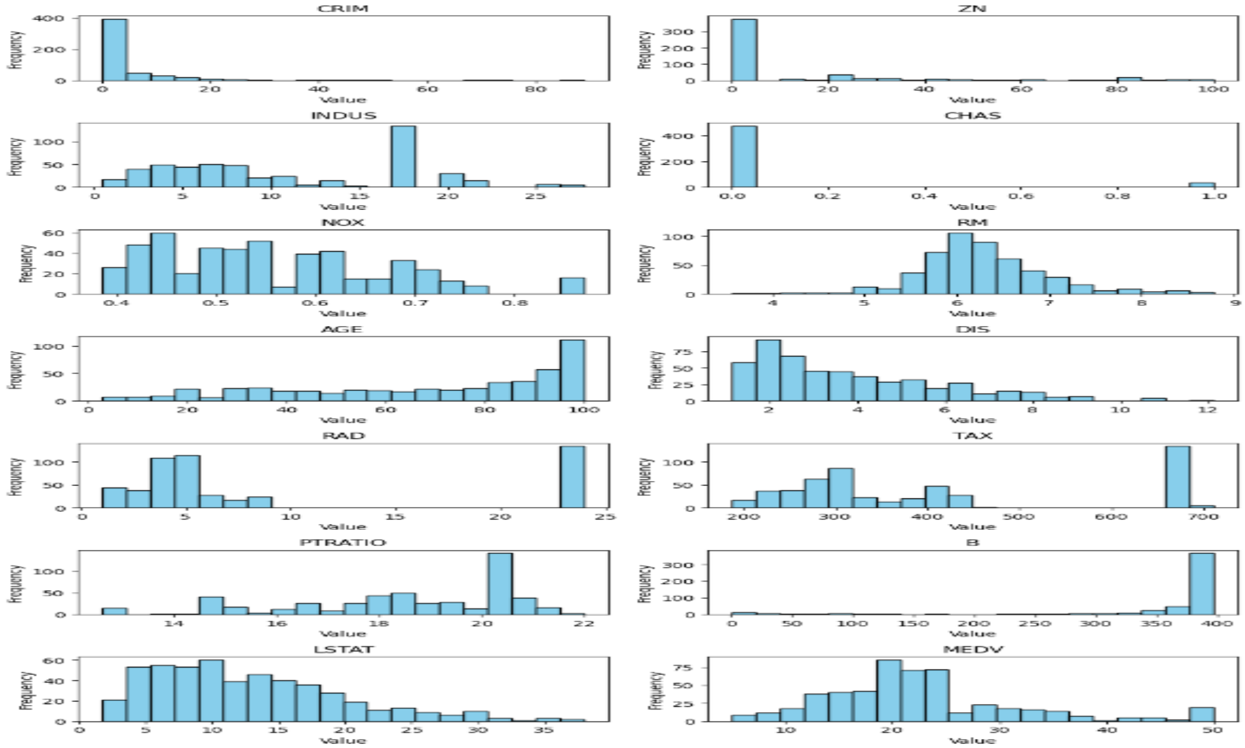
To increase the accuracy of the model, feature engineering entails generating new features from preexisting ones. We included new characteristics for this project, such as the tax-to-age and age-to-room ratios, which gave the dataset more context.

5. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a useful tool for figuring out patterns and connections as well as the structure of the information. We investigated the data and discovered correlations between attributes using a variety of visualization approaches. What We did was as follows:

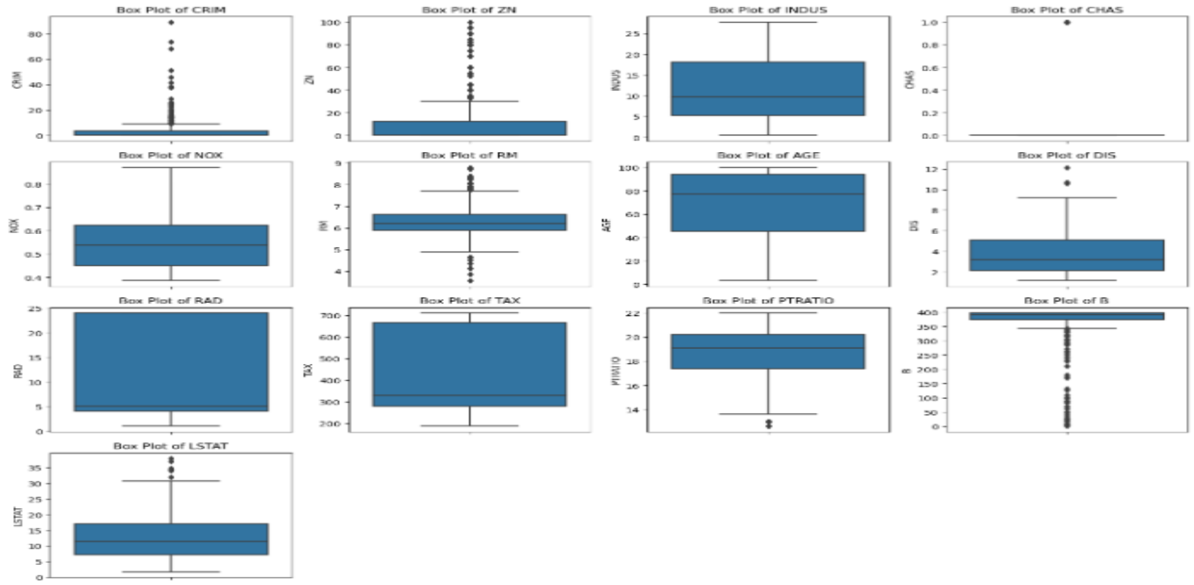
Histograms

The distribution of numerical characteristics such as RM and AGE may be examined with great ease using histograms. AGE's median age was around 45 years, whereas RM's median was approximately 6 rooms. We were able to comprehend the data's overall structure as below:



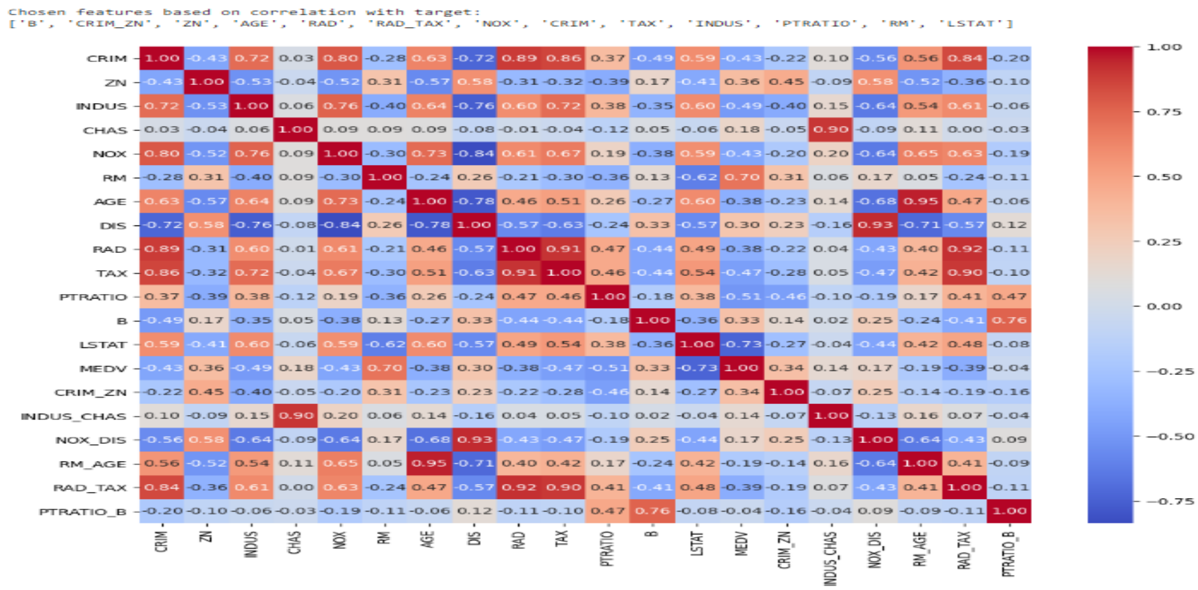
Box plots

Box plots help show the distribution of the data and locate outliers. Some values in features like CRIM had a few outliers, reaching 20, while some values in TAX had outliers over 600. This simplified the process of identifying outliers and determining how to deal with them.



Correlation Analysis

The heatmap visualizes correlations in the housing dataset, with color intensity indicating correlation strength. It reveals strong positive correlations (dark red) like **RAD** and **TAX**, suggesting areas with better highway access may have higher taxes. Notable negative correlations (dark blue) appear between **DIS** and **NOX**, with closer employment centers linked to higher pollution levels. Crucially, features such as **RM** show a positive correlation with **MEDV**, the median home value, while **LSTAT** inversely correlates, indicating lower values in areas with lower-status populations. This analysis aids in feature selection for modeling and highlights potential multicollinearity issues to address.



```
Chosen features based on correlation with target:
['B', 'CRIM_ZN', 'ZN', 'AGE', 'RAD', 'RAD_TAX', 'NOX', 'CRIM', 'TAX', 'INDUS', 'PTRATIO', 'RM', 'LSTAT']
```

	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CRIM_ZN	INDUS_CHAS	NOX_DIS	RM_AGE	RAD_TAX	PTRATIO_B
CRIM	1.00	-0.43	0.72	0.03	0.80	-0.28	0.63	-0.72	0.89	0.86	0.37	-0.49	0.59	-0.43	-0.22	0.10	-0.56	0.56	0.84	-0.20
ZN	-0.43	1.00	-0.53	-0.04	-0.52	0.31	-0.57	0.58	-0.31	-0.32	-0.39	0.17	-0.41	0.36	0.45	-0.09	0.58	-0.52	-0.36	-0.10
INDUS	0.72	-0.53	1.00	0.06	0.76	-0.40	0.64	-0.76	0.60	0.72	0.38	-0.35	0.60	-0.49	-0.40	0.15	-0.64	0.54	0.61	-0.06
CHAS	0.03	-0.04	0.06	1.00	0.09	0.09	0.09	-0.08	-0.01	-0.04	-0.12	0.05	-0.06	0.18	-0.05	0.90	-0.09	0.11	0.00	-0.03
NOX	0.80	-0.52	0.76	0.09	1.00	-0.30	0.73	-0.84	0.61	0.67	0.19	-0.38	0.59	-0.43	-0.20	0.20	-0.64	0.65	0.63	-0.19
RM	-0.28	0.31	-0.40	0.09	-0.30	1.00	-0.24	0.26	-0.21	-0.30	-0.36	0.13	-0.62	0.70	0.31	0.06	0.17	0.05	-0.24	-0.11
AGE	0.63	-0.57	0.64	0.09	0.73	-0.24	1.00	-0.78	0.46	0.51	0.26	-0.27	0.60	-0.38	-0.23	0.14	-0.68	0.95	0.47	-0.06
DIS	-0.72	0.58	-0.76	-0.08	-0.84	0.26	-0.78	1.00	-0.57	-0.63	-0.24	0.33	-0.57	0.30	0.23	-0.16	0.93	-0.71	-0.57	0.12
RAD	0.89	-0.31	0.60	-0.01	0.61	-0.21	0.46	-0.57	1.00	0.91	0.47	-0.44	0.49	-0.38	-0.22	0.04	-0.43	0.40	0.92	-0.11
TAX	0.86	-0.32	0.72	-0.04	0.67	-0.30	0.51	-0.63	0.91	1.00	0.46	-0.44	0.54	-0.47	-0.28	0.05	-0.47	0.42	0.90	-0.10
PTRATIO	0.37	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.24	0.47	0.46	1.00	-0.18	0.38	-0.51	-0.46	-0.10	-0.19	0.17	0.41	0.47
B	-0.49	0.17	-0.35	0.05	-0.38	0.13	-0.27	0.33	-0.44	-0.44	-0.18	1.00	-0.36	0.33	0.14	0.02	0.25	-0.24	-0.41	0.76
LSTAT	0.59	-0.41	0.60	-0.06	0.59	-0.62	0.60	-0.57	0.49	0.54	0.38	-0.36	1.00	-0.73	-0.27	-0.04	-0.44	0.42	0.48	-0.08
MEDV	-0.43	0.36	-0.49	0.18	-0.43	0.70	-0.38	0.30	-0.38	-0.47	-0.51	0.33	-0.73	1.00	0.34	0.14	0.17	-0.19	-0.39	-0.04
CRIM_ZN	-0.22	0.45	-0.40	-0.05	-0.20	0.31	-0.23	0.23	-0.22	-0.28	-0.46	0.14	-0.27	0.34	1.00	-0.07	0.25	-0.14	-0.19	-0.16
INDUS_CHAS	0.10	-0.09	0.15	0.90	0.20	0.06	0.14	-0.16	0.04	0.05	-0.10	0.02	-0.04	0.14	-0.07	1.00	-0.13	0.16	0.07	-0.04
NOX_DIS	-0.56	0.58	-0.64	-0.09	-0.64	0.17	-0.68	0.93	-0.43	-0.47	-0.19	0.25	-0.44	0.17	0.25	-0.13	1.00	-0.64	-0.43	0.09
RM_AGE	0.56	-0.52	0.54	0.11	0.65	0.05	0.95	-0.71	0.40	0.42	0.17	-0.24	0.42	-0.19	-0.14	0.16	-0.64	1.00	0.41	-0.09
RAD_TAX	0.84	-0.36	0.61	0.00	0.63	-0.24	0.47	-0.57	0.92	0.90	0.41	-0.41	0.48	-0.39	-0.19	0.07	-0.43	0.41	1.00	-0.11
PTRATIO_B	-0.20	-0.10	-0.06	-0.03	-0.19	-0.11	-0.06	0.12	-0.11	-0.10	0.47	0.76	-0.08	-0.04	-0.16	-0.04	0.09	-0.09	-0.11	1.00

With these EDA-derived insights, could concentrate on developing a model that explained these connections.

6. Model Building

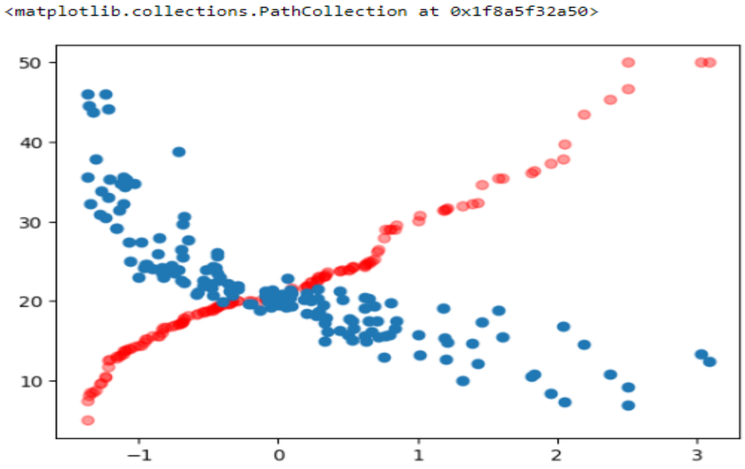
In this phase of the project, we focused on constructing and optimizing machine learning models after preprocessing the data and exploring the relationship between features. We selected RandomForestRegressor and XGBoost, two well-regarded algorithms, to develop our predictive models.

RandomForestRegressor

The RandomForestRegressor is an ensemble method that constructs multiple decision trees during training and outputs the average of these trees' predictions. This technique helps mitigate overfitting and increases accuracy by averaging out biases.

Hyperparameter Tuning:

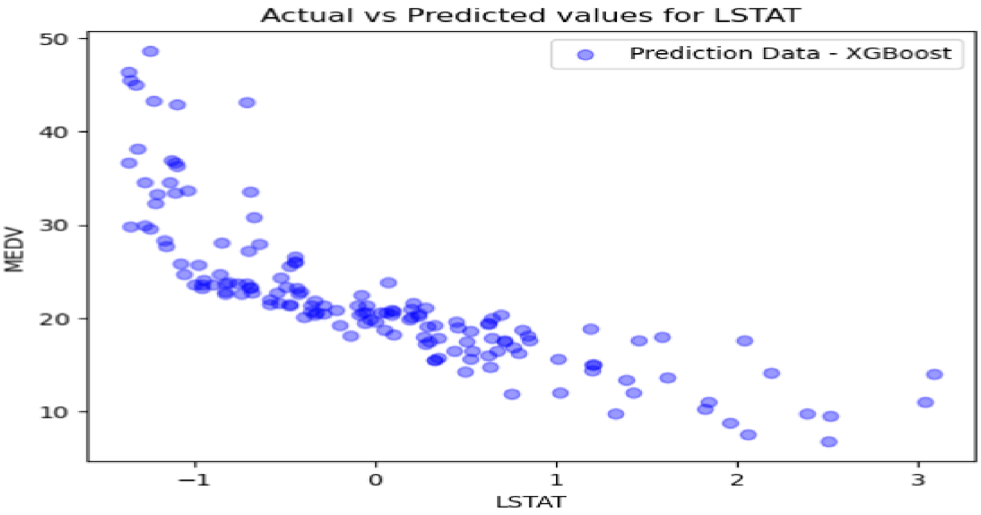
We used GridSearchCV for hyperparameter tuning to systematically test combinations of tree depth and number of trees. The best performance was achieved with a maximum tree depth of 10 and 100 trees in the ensemble. This configuration optimally balances complexity and generalization, reducing the risk of overfitting while effectively capturing underlying patterns in the data.



```
<matplotlib.collections.PathCollection at 0x1f8a5f32a50>
```

XGBoost

For the housing price prediction project, XGBoost was chosen for its ability to handle diverse datasets and its effectiveness in capturing complex non-linear relationships between features. The model was fine-tuned using GridSearchCV to determine the best hyperparameters, focusing on settings such as the number of estimators, max depth of trees, and learning rate. These parameters were crucial in adapting the model to the specific characteristics of the housing data, ensuring both accuracy and efficiency in predictions.



To assess the model's accuracy and capacity for generalization, We divided the dataset into training and testing subsets, using 70% of the dataset for training and 30% for testing.

7. Model Evaluation

The effectiveness of our machine learning algorithms was assessed using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R-squared (R^2) statistics.

- **Linear Regression** reported an MSE of 23.023, RMSE of 4.798, and an R^2 of 0.673. This baseline model, while simpler in its approach, offers a moderate fit for the data.

- **Random Forest Regressor** exhibited a superior MSE of 7.524, a lower RMSE of 2.783, and an impressive R^2 of 0.890. This significant improvement over Linear Regression suggests that the ensemble method, which leverages multiple decision trees, captures the complexities of the data more effectively.

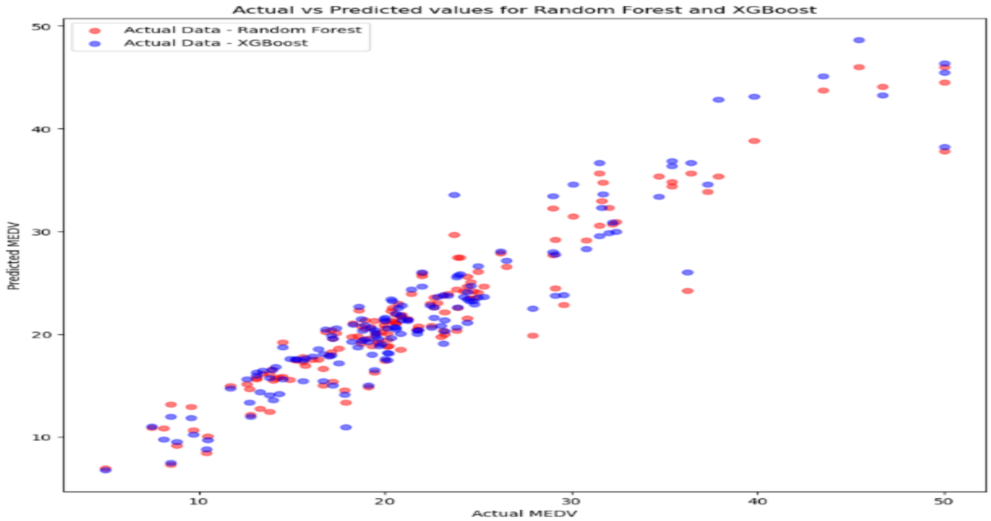
- **XGBoost Regressor** achieved an MSE of 7.563, RMSE of 2.750, and the highest R^2 of 0.893. This model, known for its predictive power in various competitions and applications, slightly outperformed Random Forest in terms of RMSE and R^2 but had a marginally higher MSE.

8. Model Comparison

A detailed comparison of Random Forest and XGBoost models using our housing dataset was conducted, focusing on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) values as our evaluation metrics.

- **Random Forest Regressor** achieved an MSE of 7.524, RMSE of 2.783, and an R^2 of 0.890. The scatter plot showed a notable correlation between predicted and actual housing values, yet some predictions for higher-valued houses appeared scattered, indicating potential variances in model accuracy at different price segments.

- **XGBoost Regressor** exhibited an MSE of 7.563, slightly higher than Random Forest, but its RMSE of 2.750 was marginally lower. Moreover, the R² value of 0.893 was the highest among the models, suggesting a slightly better fit to the data. Visually, predictions from the XGBoost model closely tracked the actual values, with less dispersion at the higher end compared to Random Forest.



Despite the similar performance metrics, the subtle differences become apparent when examining specific ranges of housing prices. For instance, in the range above $40,000, XGBoost maintained closer adherence to the actual values, implying better handling of outliers or complex patterns within the dataset.

9. Monetary Value and Risks

Savings from Accurate Predictions:

- If the application's use of advanced models like Random Forest and XGBoost reduces pricing errors by just 2% compared to a traditional model, the savings can be substantial. For instance, a real estate agency handling 200 transactions annually with an average house price of $300,000 could save $6,000 per transaction, amounting to $1.2 million in savings.

Increased Sales Efficiency:

- By pricing homes more accurately, the application can increase the conversion rate of listings to sales. A conservative 1% increase in conversion rate could lead to two additional sales for an agency with 200 listings, equating to an extra $600,000 in sales at the average house price.

9.1 Risks

Prediction Errors:

- Mispricing due to prediction errors could lead to undervalued or overvalued listings. If the application underestimates the price of 10 houses by 10%, this could result in a loss of $300,000.

False Negatives and Positives:

- Similar to loan default predictions, if the application inaccurately predicts housing values, it may result in missed opportunities (analogous to rejecting people who would have paid back a loan) or engaging in bad deals (analogous to approving defaults). Estimating a 5% error rate on 200 transactions could lead to a significant financial impact.

9.2 Other Benefits and Considerations

Improved Decision-Making:

- The application provides data-driven insights that could lead to better strategic decisions, such as identifying undervalued investment opportunities.

Customer Trust:

- Accurate and transparent pricing builds customer trust and can enhance the agency's reputation, leading to more business referrals and repeat clients.

Adaptability:

- The application's predictive models can be regularly updated with new data, ensuring the tool remains responsive to market changes.

10. Other risks and benefits

Additional Benefits of Machine Learning for Housing Price Prediction:

1. **Enhanced Market Efficiency**: Reduces the time properties stay on the market by accurately matching buyers and sellers.
2. **Improved Urban Planning**: Helps local governments plan infrastructure and housing policies based on detailed market predictions.
3. **Automated Property Valuation**: Offers frequent and precise property valuations for tax assessments and insurance.
4. **Personalized Recommendations**: Provides tailored property suggestions to clients, enhancing customer satisfaction.

5. **Environmental Impact Considerations**: Integrates environmental data to assess effects on property values, promoting sustainable development.

Additional Risks of Machine Learning for Housing Price Prediction:

6. **Data Privacy Concerns**: Handling large datasets poses risks of data breaches and privacy violations.
7. **Market Disruption**: Could disadvantage traditional market participants as machine learning becomes prevalent.
8. **Over-reliance on Technology**: Might reduce critical human judgment in decision-making.
9. **Potential for Bias**: Could perpetuate existing biases in training data, affecting pricing fairness.
10. **Market Volatility**: Algorithm-driven rapid price changes could lead to increased market instability.
11. **Regulatory Challenges**: Emerging technologies in real estate face evolving regulatory frameworks to ensure fairness and compliance.

11. Conclusion

To summarize, the application of machine learning to forecast housing prices resulted in encouraging outcomes. The comprehensive analysis offered insights into the variables affecting housing prices, and the trained models showed a respectable degree of precision and dependability.

When RandomForestRegressor and XGBoost were compared, RandomForestRegressor performed better. The most important features that contributed to the predictions were found through the feature importance analysis, which offered insightful information about the housing market.

The possible advantages and disadvantages of applying this application in the real estate sector were examined through the discussion of monetary value estimates and risk assessments. To ensure that the application keeps getting better, this report's conclusion included recommendations for more work and research.

Git Hub Links of Our Group

Sai Priya Kaipu: https://github.com/SaiPriyaKaipu/Housing_Price_Prediction

Nitika : GitHub - NitikaTrehan/Housing-Price-Prediction

Kiran : https://github.com/KiranVarada/Housing-price-prediction