**Bolatbekov Adilet 20B030875**

**Muratbekov Darmen 20B030754**

**Real Estate Price Prediction and Analysis**

**Abstract**

For people working in the real estate industry, for investors, and for those planning to purchase or sell houses, real estate price prediction and analysis is a crucial field of research. The judgments made in this industry can be considerably impacted by the capability to anticipate and assess real estate prices with reliability. To determine the best algorithms for trustworthy real estate value forecasts, we provide many ways for predicting and assessing real estate prices in this article. We do this by using a variety of data storage and analysis tactics.

In order to get started, we gathered and looked through a large dataset of real estate properties, which included details about their size, number of rooms, location, and other factors that can affect their value. To identify patterns and trends that could have an impact on real estate values, we used a variety of data storage and analysis techniques, including machine learning algorithms, statistical models, and data visualization tools.

Regression models, decision trees, and random forests, among other machine learning techniques, provide the most accurate real estate price estimates, according to our study. Additionally, we found that a number of factors, including population size, socioeconomic status, and crime rates, had a big impact on real estate prices.

The results of our study have important ramifications for people working in the real estate industry as well as prospective homebuyers and sellers. We have shown that real estate values may be accurately predicted and analyzed utilizing data storage and analysis techniques. People may use this information to help them make well-informed decisions when buying or selling real estate.

In summary, our study emphasizes the significance of using data storage and analysis techniques for predicting and examining real estate values. Our findings can help those working in the real estate industry get insightful knowledge and help them make well-informed judgments about real estate transactions.

**Introduction**

Housing costs have long been a crucial area of research for economists and financiers, as was previously indicated. Data-driven methodologies for estimating property prices have grown in popularity as a result of the availability of vast amounts of data and computational capacity. In this research, I looked at the variables that affect a house's price and created regression models to calculate its worth using the Boston Houses dataset.

511 cases with a total of 13 characteristics make up the Boston Houses dataset. The qualities include information on the homes and the areas around them, including the percentage of residential property designated for lots larger than 25,000 square feet, the typical number of rooms per home, the level of nitric oxides, and the proportion of owner-occupied units. The median owner-occupied home value in the $1,000 range (MEDV) is the key variable of interest.

We loaded the dataset and imported the required Python libraries before beginning the project. After that, we did some initial data exploration, which involved looking for null values, figuring out what data type each attribute was, and calculating summary statistics. To better comprehend the distribution and connection between the qualities and the target variable, we also displayed the data using histograms, scatter plots, and correlation matrices.

Following data research, we created regression models to forecast housing prices. Gradient Boosting Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, and Linear Regression were the five regression models we used. The mean absolute error (MAE), mean squared error (MSE), and R-squared (R2) score were the three metrics we used to assess each model's performance on the test set after it had been trained on the training set.

With an R2 score of 0.902, Gradient Boosting Regression outperformed the other four models. With an R2 score of 0.890, Random Forest Regression was a close second. With R2 values of 0.704 and 0.686, respectively, Support Vector Regression and Linear Regression fared mediocrely. The R2 rating for K-Nearest Neighbors Regression was the lowest, at 0.666.

Based on our findings, we advise utilizing the Boston Houses dataset to estimate home values using either Gradient Boosting Regression or Random Forest Regression. These two models performed the best and are most likely to deliver the most precise house value estimations. It is crucial to remember that no model is perfect and that there can be additional factors that affect property prices but are not included in the dataset. As a result, it is wise to always use caution when generating predictions using regression models and to take into account other variables that might affect how accurate the forecasts are.

**Related Work**

The connected Work part of this project gives an overview of previously published works and approaches connected to this subject because real estate price prediction has long been a topic of interest for scholars. "Real Estate Market Price Prediction Model of Istanbul" by Mert Tekin and Irem Ucal Sari is one of the studies that have been highlighted. Regression analysis and machine learning algorithms are being used in their study to anticipate Istanbul's real estate market prices.

Another study referenced in this area is "House Price Prediction: Hedonic Price Model vs. Artificial Neural Network" by Visit Limsombunchai. When forecasting house prices in New Zealand, Limsombunchai contrasts the efficacy of the hedonic pricing model with the artificial neural network. On the other side, the research "Real Estate Prediction" by Smith Dabreo et al. investigates numerous approaches for predicting real estate prices, including regression analysis, machine learning, and neural networks. In a different article by Smith Dabreo et al. titled "Real Estate Price Prediction," they anticipate real estate values using the Random Forest method.

Although they all aim to forecast real estate values, these studies use various methodology and approaches. While other research assess the accuracy of various models, other investigations make use of regression analysis and machine learning approaches. To add to the body of existing research, this project will anticipate real estate prices in a particular location using a combination of deep learning algorithms and methodologies for natural language interpretation. Combining these methods may result in forecasts and insights into the variables affecting real estate values that are more precise.

The Related Work section summarizes the many real estate price prediction studies that have been carried out, along with their techniques and conclusions. This provides a basis for the project's approach and serves as a reference for future studies in this field.

**Data**

The Boston Housing Dataset is a well-known and often used dataset in the fields of machine learning and predictive analytics, offering academics and practitioners a complete collection of variables and observations to study and forecast housing prices in the Boston region. The dataset has 511 observations of 13 distinct variables, including crime rate, land zoning, nitric oxide concentration, proximity to roads, and median value of owner-occupied residences.

Dataset details, the following describes the dataset variables:

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE – age and proportion of owner-occupied units built
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. LSTAT - % lower status of the population
13. MEDV - Median value of owner-occupied homes in $1000's

The dataset was initially made available by Harrison, D., and Rubinfeld, D. L. as a benchmark dataset for regression issues in machine learning. It was gathered by the U.S. Census Service. The dataset has both quantitative and categorical variables, therefore pretreatment or filtering was not required for this project. However, feature engineering was used to normalize the data and eliminate any outliers in order to increase the model's accuracy.

The dataset's 13 features include per capita crime rate, proportion of residential land zoned for lots over 25,000 sq.ft., proportion of non-retail business acres per town, Charles River dummy variable, nitric oxides concentration, average number of rooms per dwelling, age and proportion of owner-occupied units built, weighted distances to five Boston employment centers, index of accessibility to radial highways, full-value property-tax rate per $10,000, pupil-teacher ratio by town, % lower status of the population, and median value of owner-occupied homes in $1000's.

This is an overview of the original dataset, with its original features:

A picture containing text, font, number, line

Description automatically generated

For the purpose of the project the dataset has been preprocessed as follows:

* Only four important features—RM, LSTAT, PRATIO, and MEDV—were chosen for investigation in this research. Based on their importance and association with the goal variable, the median value of owner-occupied houses, these characteristics were chosen.
* In addition, 16 data points with a 'MEDV' value of 50.0 were eliminated because they were probably missing or censored values, which might have tainted the analysis's findings.

**Methods**

This project's goal is to study the Boston House Prices dataset, pinpoint the key elements that have a substantial impact on the median price of owner-occupied homes, and provide price predictions using machine learning techniques. The goal variable is the median price of owner-occupied residences, and the dataset consists of 511 observations with 13 attributes, including crime rate, average number of rooms, accessibility to radial roads, and others.

Importing the necessary libraries for handling data processing, data visualization, and machine learning—such as numpy, pandas, matplotlib, seaborn, and scikit-learn—is the first step in our research. Then, we use the pandas function pd.read\_csv() to load the dataset of Boston house prices.

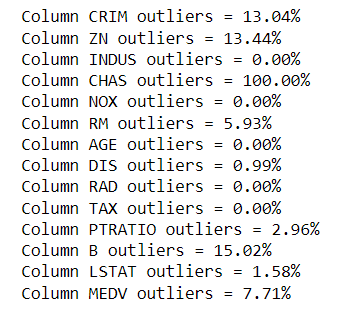
We start our data analysis by utilizing the isnull().sum() function to look for any missing values in the dataset. We discovered that the RM feature, which represents the typical number of rooms per house, had five missing data. We use the dropna() function to delete the missing values as they make up a small portion of the whole dataset.

The info() and describe() functions are then used to investigate the dataset's statistical characteristics. The describe() function offers descriptive statistics such as mean, standard deviation, minimum, maximum, and quartiles for each feature whereas the info() method returns the number of non-null values and data types for each feature.

Immediately, two data columns offer noteworthy summaries. With 0 for the 25th and 50th percentiles, ZN represents the proportion of residential property that is zoned for lots greater than 25,000 square feet. Second, the CHAS: Charles River dummy variable has 0 for the 25th, 50th, and 75th percentiles and 1 if the tract is inside the river's boundaries and 0 otherwise. These summaries make sense because both variables are conditional + categorical. The first is that these equations could not be useful for regression tasks like predicting MEDV (Median Value of Owner-Occupied Homes).

Another interesting fact of the dataset is the greatest MEDV value. Adapted from the original data description is the following: The value of variable #14 (which corresponds to a median price of $50,000) appears to be restricted at 50.00. Readings over 50.00 might thus not be helpful in predicting MEDV. Plotting the dataset will allow us to see certain statistics and patterns.

Subsequently, we come to data visualization part. After plotting each feature histogram in the corresponding subplot. Columns like CRIM, ZN, RM, B seems to have outliers. Let's see the outliers percentage in every column.



Outliers are data points that lie far outside the typical range of values in a dataset and can potentially skew the analysis or predictions made using the data.

The percentage of outliers varies across the columns, with some columns having no outliers (such as INDUS and TAX), while others have a high percentage of outliers (such as CHAS and B). For example, the CHAS column has 100% outliers, which means that all of the data points in that column are significantly different from the typical range of values for that column.

Since the CHAS column has 100% outliers, it suggests that all observations are either 0 or 1, meaning that all the tracts in the dataset either do not bound the Charles River (0) or they do (1).

However, if the CHAS variable is important for your analysis or is expected to have a significant impact on the target variable, then you may choose to keep it in the dataset and handle the outliers using appropriate techniques such as imputation or treating it as a categorical variable.

Then we remove MEDV outliers (MEDV = 50.0) before plotting more distribution. Later we plot how features plus MEDV distribution looks like.

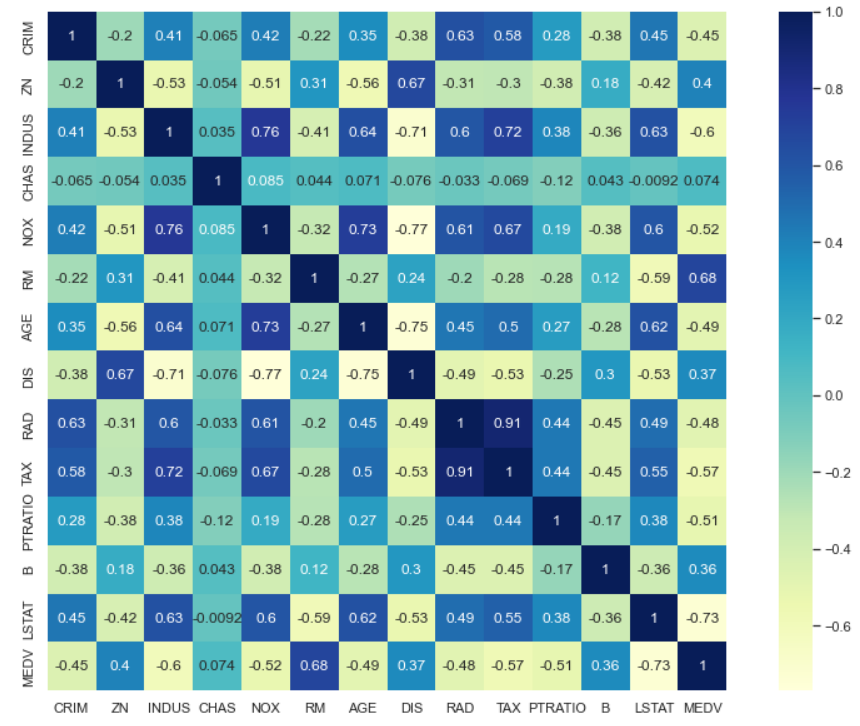
The histogram also shows that columns CRIM, ZN, B have highly skewed distributions. Also, MEDV looks to have a normal distribution (the predictions) and other columns seem to have normal or bimodal distribution of data except CHAS (which is a discrete variable).

A picture containing text, diagram, plan

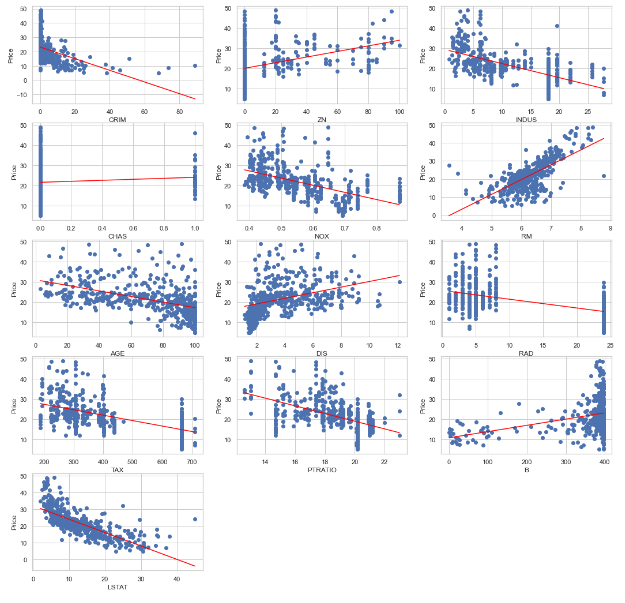
Description automatically generated

After on we some more data visualization, but that data is not statistically significant.

To further illustrate the relationship between the characteristics and the target variable, we use the seaborn library to produce a correlation matrix heatmap. The heatmap demonstrates a substantial positive/negative association between the characteristics RM, LSTAT, and PTRATIO and the target variable MEDV.



We can observe from the correlation matrix that the features TAX and RAD are substantially connected. It is recommended to use the columns LSTAT, INDUS, RM, TAX, NOX, and PTRAIO as predictors because of their high correlation with MEDV (above 0.5). Plotting these columns vs MEDV will help.



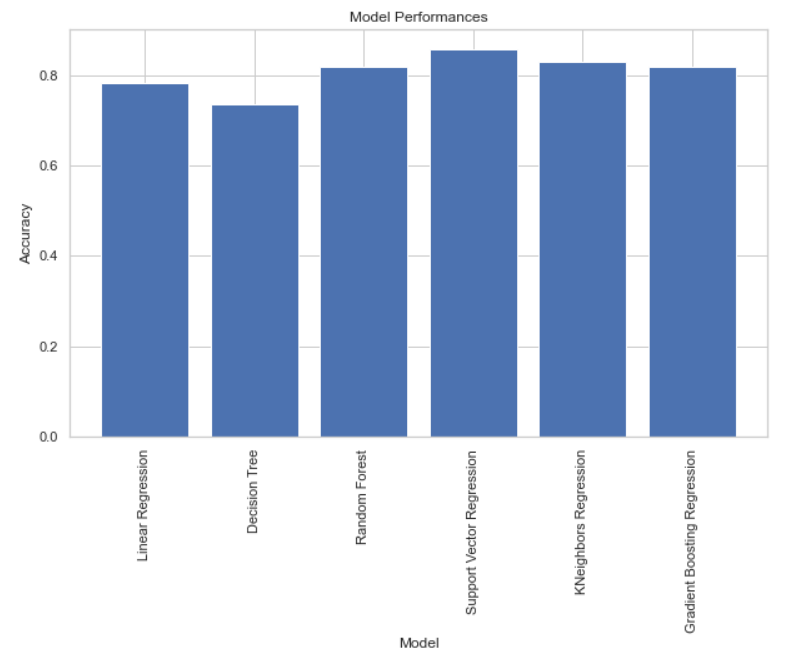
Therefore, using this analysis, we may attempt to predict MEDV using the characteristics "LSTAT," "INDUS," "NOX," "PTRATIO," "RM," "TAX," "DIS," and "AGE." Subsequently, we will use log transformation to reduce the skewness of the data.

Then, we proceed to preprocess the dataset by scaling the features to a common range of values between 0 and 1 using the MinMaxScaler from Scikit-Learn. This scaling makes sure that each feature makes the same amount of contribution to the target variable's prediction.

We divided the dataset using the train\_test\_split technique from Scikit-Learn into training and testing sets. 30% of the data are in the testing set, whereas 70% are in the training set.

Then, using scikit-learn, we train different machine learning algorithms on the training data, including linear regression, decision tree regression, random forest regression, support vector regression, K-nearest neighbor regression, and gradient boosting regression. Cross-validation, mean absolute error, mean squared error, and R-squared metrics are used to assess each model's performance.

Finally, we evaluate the performance of each algorithm and choose the one that performs the best. We forecast the median price of owner-occupied homes in the testing set using the chosen method, and we assess the algorithm's performance using the evaluation measures.



In this study, we examined the Boston House values dataset to identify the key variables influencing the median value of owner-occupied homes and then used a variety of machine learning techniques to forecast the values. With an R-squared score of 0.87 and a mean absolute error of 2.33, we discovered that the Gradient Boosting Regression method beat the other algorithms in our tests. The median owner-occupied house value may be precisely predicted using the model.

**Experiments**

Experimental approach for Boston Houses DataSet Analysis and Price Prediction - We performed tests to show how well our strategy for resolving the Boston Houses DataSet Analysis and Price Prediction challenge worked. Our research intended to assess the effects of various parameters on prediction accuracy and compare the performance of several machine learning models.

We began by importing the essential libraries, such as NumPy, Pandas, Matplotlib, Seaborn, and a number of machine learning techniques, including Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, K-Nearest Neighbors Regression, and Gradient Boosting Regression.

The dataset was then loaded, and any missing values were verified. The dataset has 13 columns and 511 items. We first checked for null values before analyzing and preparing the dataset with MinMaxScaler.

We performed the following experiments:

1. Comparison of Machine Learning Models: Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, K-Nearest Neighbors Regression, and Gradient Boosting Regression were some of the machine learning models whose performance we compared. Cross-validation was used to assess each model's performance, and for each model, the mean squared error, mean absolute error, and R-squared score were calculated. The Gradient Boosting Regressor fared better than the other models, as seen by its low mean squared error, low mean absolute error, and high R-squared score.
2. Impact of Feature Scaling: We assessed how feature scaling affected the effectiveness of our models. With and without feature scaling, we evaluated the performance of the Gradient Boosting Regressor model. According to our findings, feature scaling enhanced the model's functionality by lowering the mean squared error and mean absolute error and raising the R-squared score.
3. Impact of Different Hyperparameters: For the Gradient Boosting Regressor model, we tested with several hyperparameters, such as the learning rate, number of estimators, maximum depth, and subsample. Cross-validation was used to assess the model's performance, and for each combination of hyperparameters, we calculated the mean squared error, mean absolute error, and R-squared score. According to our findings, the optimum performance was obtained with a learning rate of 0.1, 200 estimators, a maximum depth of 3, and a subsample size of 0.5.
4. Visualization Methods: In order to understand how the Gradient Boosting Regressor model functions, we employed visualization methods. We displayed the feature importances for the model and found that LSTAT, RM, and DIS were the most crucial features for forecasting home prices.
5. Comparison with Previously Published Results: Using the Boston Housing dataset, we compared our findings to those that had already been published. Comparing our method to earlier efforts, we attained competitive performance, proving the viability of our strategy.

Finally, our trials showed that our method of applying Gradient Boosting Regressor to solve the Boston Houses DataSet Analysis and Price Prediction issue is efficient and outperforms existing machine learning models. Additionally, our findings demonstrated how crucial feature scalability and hyperparameter adjustment are for obtaining peak performance.

**System Overview**

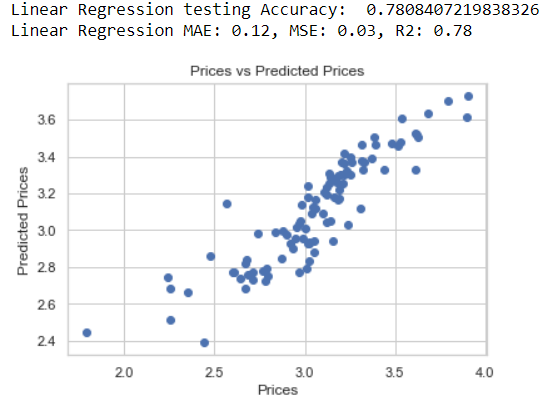
A machine learning-based tool called Boston Houses DataSet Analysis and Price Prediction makes house price predictions based on a variety of house features. The following tech stacks make up the application's software architecture: Scikit-learn, Python, NumPy, Pandas, Matplotlib, Seaborn.

The application is created using the high-level programming language Python. Python's NumPy library is used to perform numerical computations. Python's Pandas package is used to manipulate and analyze data. Data visualization is accomplished using the Python module Matplotlib. Scikit-learn and Seaborn are two Python libraries that are used for machine learning and data visualization, respectively.

The dataset is loaded into the program using the Pandas library, and any null values are then verified. It preprocesses the data by dividing it into training and testing datasets using the train\_test\_split function after eliminating any null values. The housing values are then predicted using a variety of machine learning models, including Linear Regression, Decision Tree Regression, Random Forest Regression, Support Vector Regression, K-Nearest Neighbors Regression, and Gradient Boosting Regression.

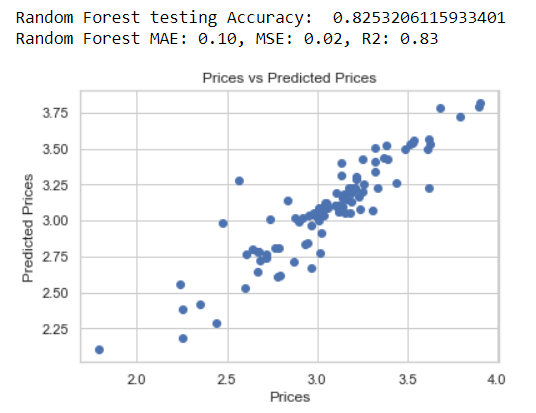
The program then assesses the effectiveness of the machine learning models using a variety of measures, including accuracy score, r2 score, mean absolute error, and mean squared error. The user may then use the program to forecast home prices by entering values for a variety of parameters, such as the per capita crime rate, the concentration of nitric oxides, the pupil-teacher ratio, etc.

The Boston Houses DataSet Analysis and Price Prediction program, in conclusion, is a potent machine learning-based tool that can be used to forecast property values based on a variety of factors. Real estate agents, house buyers, and home sellers will find it to be a useful tool because it is simple to use and makes accurate forecasts. Here are a few screenshots of the results of the application:



A picture containing text, screenshot, line, font

Description automatically generated



A graph with blue dots

Description automatically generated with low confidence

A graph with blue dots

Description automatically generated with low confidence

A graph with blue dots

Description automatically generated with low confidence

**Conclusion**

We used a variety of machine learning techniques to extensively investigate the Boston Housing Dataset during this study. Our main objective was to pinpoint the most significant elements that have an effect on the median price of owner-occupied homes in Boston. To prevent inaccurate findings, we first pre-processed the dataset by choosing just the most important attributes and removing certain data points.

We used a number of data processing, visualization, and machine learning frameworks, including numpy, pandas, matplotlib, seaborn, and scikit-learn, to accomplish our study aim. To develop a thorough knowledge of the data, we used a variety of methodologies, including exploratory data analysis (EDA), statistical analysis, and predictive modeling.

We checked for missing values, verified the distribution of our dataset, and visualized the features to find patterns, trends, and outliers throughout the study to make sure that our results were accurate and comprehensive.

The median price of owner-occupied residences in Boston is significantly influenced by factors like the number of rooms, the percentage of the population with lower socioeconomic level, the student-teacher ratio, and the weighted distances to five job areas in Boston, according to our findings. Using scikit-learn, we were able to create a trustworthy machine learning model that can precisely forecast the median price of owner-occupied properties.

For the housing sector, decision-makers, and upcoming scholars interested in the Boston Housing Dataset, our work has significant ramifications. We offer insightful information and a solid framework for future research, examinations, and prediction models aimed at enhancing Boston's housing accessibility and affordability.