A COMPARATIVE STUDY OF PREDICTING REAL ESTATE PRICES USING MACHINE LEARNING APPROACHES

RAMASONDRANO Andriamanjaka[1], Ramafiarisona Hajasoa Malalatiana [2],

[1] PhD, Telecommunication, Automatic, Signal and Images
[2] Professor, Telecommunication, Automatic, Signal and Images
Doctoral School of Science and Technics of Engineering and Innovation
University of Antananarivo, Madagascar

ABSTRACT

Research into real estate price prediction using artificial intelligence has been the subject of several empirical studies. The purpose of this project is to identify the most accurate model for predicting real estate prices using artificial intelligence, based on the characteristics of the property and its locality. This study was conducted using an Indian real estate dataset, which includes 14,620 property sales samples and 23 features. To achieve this, various machine learning techniques were employed to compare selected supervised algorithms: Linear Regression,Ridge Regression, Lasso Regression, Support Vector Machine, Random Forest , K-Nearest Neighbors . The choice of the best model is based on the comparison of performance evaluation metrics such as: MAE, MSE, RMSE, R2 and R2 Cross-Validation. The prediction results showed that the optimized Random Forest algorithm provides the best overall performance, with lowest values of MAE, MSE, and RMSE, as well as high R^2 and R^2 cross-validation. It gives 87.44% prediction accuracy. This project also demonstrated the importance of taking into account the geographic context in the analysis and prediction of real estate prices.[1]

Keyword: - *Real estate prices, Machine Learning, Supervised Algorithm, Study comparative algorithms*.

1. INTRODUCTION

The real estate market is one of the most dynamic and complex sectors of the economy, characterized by rapid growth in recent decades, marked regional diversity, and fluctuations in supply and demand. However, price volatility and associated uncertainty pose significant challenges to market participants, ranging from individual buyers to large real estate developers.

Accurate prediction of real estate prices is of paramount importance to various stakeholders, as it informs decisions related to buying, selling, investing, and urban planning. However, due to the complex and multidimensional nature of the real estate market, price prediction remains a major challenge. Additionally, the available data are often limited, containing only a restricted number of features related to each property.

In this context, Artificial Intelligence (AI), and specifically Machine Learning (ML) techniques, offer new perspectives for analyzing market trends and predicting real estate prices with increased accuracy. By leveraging large datasets on property characteristics and their localities, ML models can identify key factors influencing prices and provide reliable estimates.[2]

The literature reveals two fundamental approaches to determining real estate prices. One approach is predicting real estate prices using macroeconomic variables in the country where the property is located, and the other involves price prediction models expressed as micro-variables, considering the characteristics of the property itself.

Several factors influence real estate prices (e.g., location, size, condition, construction date, etc.), raising the question: "Which model allows us to better predict real estate prices, with greater accuracy, using Artificial Intelligence based on property characteristics and locality?"

This project will consist of three main chapters: the first chapter, primarily theoretical, will provide an overview of the state-of-the-art and existing studies on real estate price prediction. The second chapter will focus on the methodology and modeling tools. The third chapter will present the results of the analysis and estimation of Machine Learning models for predicting real estate prices.

1.1 LITERATURE REVIEW

Fluctuations in the real estate market are governed by a number of factors, such as the country's economy, current interest rates, inflation, the behavior of financial organizations, demographics, government policy, inelastic supply and demand for housing or commercial property, which depends on household income, as well as the general real estate situation in a given region. Demand is closely linked to demographic trends, economic growth, consumer sentiment and mortgage interest rates. Other global influences can also interpenetrate and influence the real estate market [1]. According to existing theories, the real estate price is the result of two valuation methodologies: traditional appraisal methods and financial valuation methods, which are considered modern methods. The price must be the result of a comparison between the outcomes of these two methodologies, which ultimately reflect the confrontation between the seller's offer and the buyer's demand [3]. The various methods presented above have the disadvantage of systematically referring to past data and comparisons with goods that are, by nature, heterogeneous. Modern methods strive to use values anticipated for the future, or to homogenize heterogeneous elements. The result is two modern approaches to appraisal:

- The financial cash-flow method
- The hedonic method

Among these methods, the hedonic method is the best-known and most widely used by researchers to analyze, study or estimate the price of so-called “environmental goods” such as real estate or housing. It is therefore important to highlight this method before presenting some empirical literature on the real estate market [4].
The following table summarizes some empirical work on real estate price prediction.

Author - Year	Problem Studied	Methods	Results	Disadvantages
Satish et al. (2019)	Method to predict future house prices using machine learning.	XGBoost, Neural System, and Lasso Regression	Lasso regression algorithm, with high precision, reliably outperforms alternative models in executing real estate price predictions.	Lasso selects at most n variables before saturating. Lasso cannot perform group selection
Alfiyatin et al. (2017)	Discusses prediction, model based on regression analysis and Particle Swarm Optimization (PSO).	PSO is a stochastic optimization technique used to select and assign variables. Regression is used to determine the optimal internal prediction coefficient.	The results obtained: Minimum prediction error.	In a PSO system, it can be difficult to initially define the design parameters
Segnon et al. (2020)	Smooth transition, autoregressive fractional, integrated	Analyzed complex statistics, models based on the	The results of the measures provide satisfactory	It operates on high-frequency data.

	process model to predict the volatility of real estate prices in the United States.	hypotheses of the integrated process.	prediction accuracy	
Singh et al. (2020)	The concept of Big Data to predict sales data of housing in Iowa.	Linear Regression, Random Forest, Gradient Boosting	The Gradient Boosting model outperforms other prediction models in terms of prediction accuracy.	The Gradient Boosting algorithm is sensitive to outliers.
Kuvelekar et al. (2020)	Prediction of the market value of a property.	A decision tree regressor helps find a starting price for a property based on geographic variables. By breaking down trends and past market values, future costs are predicted.	It offers 89% accuracy.	Instability, longer time to train the model and complete calculations

Tableau 1. The following table summarizes some empirical work on real estate price prediction

1.2 Empirical Works on Real Estate Price Prediction Using AI

In their article titled "A Comparative Study of Urban House Price Prediction using Machine Learning Algorithms," the authors compared three Machine Learning algorithms to predict real estate prices: Linear Regression (RL), Random Forest (RF), and Gradient Boosting (GB). They evaluated the performance of the models using three metrics: Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R2 score). The models were tested on the Melbourne real estate dataset, which includes 34,857 property sales and 21 features [5].

Model	RMSE	R^2 Score	MAE
Linear Regression	372,291.53505	0.639	272,953.5787
Random Forest	261,166.3075	0.822	171,486.2709
Gradient Boosting	257,062.4425	0.828	169,009.3626

Tableau 2. Summary of Results on Real Estate Price Prediction Using AI

Discussion of Results

- Mean Absolute Error (MAE): Measures the average magnitude of the errors in a set of predictions, without considering their direction. The lower the MAE, the better the model's predictive accuracy.
- Root Mean Square Error (RMSE): A quadratic scoring rule that measures the average magnitude of the error. It gives a relatively high weight to large errors. The lower the RMSE, the better the model's performance.
- R^2 Score: Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. The closer the R^2 score is to 1, the better the model fits the data.

1.3. State of the Art on Real Estate Price Prediction in India

Introduction

The empirical analysis in this project is based on a dataset from the Indian real estate market. It is crucial to first present the state of the art of a previous study conducted with this dataset before diving into the empirical analysis of this project.

Previous Work: Study by SOURAV CHANDA

Sourav Chanda conducted a study on the Indian real estate market dataset, which is available on the website "www.kaggle.com". This study employed Machine Learning techniques using algorithms such as Linear Regression, Decision Tree Regressor, XGBRF Regressor, and XGB Regressor. The analysis in this study focused on comparing the performance of these models using five metrics: Mean Squared Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and the R^2 Score.

The results of Sourav Chanda's study can be summarized in the following table:

Performance Metrics	Linear Regression	Decision Tree Regressor	XGBRF Regressor	XGB Regressor
MSE	35691971489.951	42247969028.708	30885773957.173	17228400215.7
RMSE	188923.189	205543.107	175743.489	131257.001
MAE	123730.296	99027.13	93515.466	67954.111
MAPE	0.25	0.176	0.177	0.125
R^2 Score	0.718	0.667	0.756	0.864

Tableau 3. The results of Sourav Chanda's study can be summarized

The comparison of these results showed that the XGB Regressor is the most effective among the models used to predict Indian real estate prices, with the lowest values of MSE, RMSE, MAE, and MAPE, as well as the highest R2 Score. It provides an accuracy of up to 86.4% in predicting real estate prices in India
Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work Introduction related your research work.

2. TOOLS FOR ECONOMETRIC ANALYSIS AND MODELING

In this second part of our project, we will enumerate the various tools that will enable us to analyze our dataset and predict real estate prices using econometric models.

2.1 Tools and Methodology

- ❖ **Methodologies**

To predict real estate prices based on property characteristics and location using AI, we will introduce the basic principles of machine learning and supervised learning techniques. In this context, we will follow these steps:

- ✓ **Data Collection**
  - ➢ Sources of Data: Identify and utilize various data sources such as real estate databases, public records, and market reports.

- Description of Collected Data: Document the types of data collected, including property features (e.g., size, number of bedrooms, age of the property) and location information (e.g., neighborhood, proximity to amenities).
- Methods of Data Collection and Cleaning: Implement techniques for gathering data, such as web scraping or API integration, and perform data cleaning to handle missing values, remove duplicates, and correct errors.

✓ **Exploratory Data Analysis (EDA)**
- Data Visualization and Description: Use visual tools like histograms, scatter plots, and box plots to describe the characteristics of the data.
- Identification of Trends and Correlations: Analyze trends and correlations between features and real estate prices using correlation matrices and pairwise plots.
- Handling Outliers and Missing Data: Identify and manage outliers and missing values to ensure data quality and reliability.

✓ **Feature Engineering**
- Selection of Relevant Features: Identify and select key features that significantly impact real estate prices, using techniques like feature importance and correlation analysis.
- Creation of New Features: Develop new features from existing data to enhance model performance, such as interaction terms or aggregated metrics.

✓ **Modeling**
- Selection of Prediction Algorithms: Choose suitable algorithms for predicting real estate prices, such as linear regression, random forests, and XGBoost.
- Data Splitting: Divide the dataset into training and testing sets to evaluate model performance accurately.
- Model Training: Train the selected models using the training dataset, adjusting parameters to improve accuracy.

✓ **Model Optimisation**
- Hyperparameter Tuning: Optimize model hyperparameters using techniques like grid search or random search to enhance model performance.
- Model Comparison: Compare the performance of different models based on evaluation metrics.
- Final Model Selection: Select the best-performing model for further analysis and deployment.

✓ **Validation and Evaluation**
- Performance Evaluation: Assess the final model's performance on the testing dataset using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and R2 score.
- Result Analysis and Interpretation: Analyze the model predictions, interpret the results, and identify areas for improvement.
- Model Robustness: Evaluate the model's robustness by testing its performance on different subsets of the data.

✓ **Deployment**
- Model Integration: Integrate the predictive model into an application or user interface to make it accessible for end-users.
- Additional Features: Implement additional functionalities such as result visualization, search filters, and interactive maps to enhance user experience.
- Continuous Monitoring: Set up mechanisms for continuous monitoring and maintenance of the deployed model to ensure its accuracy over time.

- **Analysis and Model Estimation Tools**

To analyze our dataset and estimate predictive models, we will use several computational tools and techniques.
First, we will employ the programming language **Python** in our empirical study. For exploratory data analysis, we will use the following libraries:

- **Pandas**: For data manipulation and analysis.
- **NumPy**: For numerical computing.
- **Matplotlib**: For creating static, animated, and interactive visualizations.
- **Seaborn**: For statistical data visualization.

For the modeling part, we will primarily use the library **Scikit-learn**:

- **Scikit-learn**: For implementing machine learning algorithms, including regression, classification, clustering, and model evaluation.

All these steps will be carried out using **Jupyter Notebook** within the **Anaconda Navigator (Anaconda3)** environment. This setup provides a comprehensive suite of tools for data science and machine learning, enabling efficient data analysis and model development.

Tools Summary

- **Programming Language**: Python
- **Data Analysis Libraries**:
  - Pandas
  - NumPy
  - Matplotlib
  - Seaborn
- **Modeling Library**:
  - Scikit-learn
- **Development Environment**:
  - Jupyter Notebook
  - Anaconda Navigator (Anaconda3)

These tools will allow us to effectively conduct exploratory data analysis, build and evaluate predictive models, and visualize our findings in a clear and informative manner. By leveraging these technologies, we aim to develop robust and accurate models for predicting real estate prices.

2.2 Machine Learning Algorithms and Modeling

In this section, we will outline the algorithms that will enable us to predict real estate prices based on property characteristics and location. We will utilize mathematical tools and existing econometric models, primarily focusing on supervised algorithms and regression techniques :

1. Linear Regression

- **Linear Regression**: A foundational algorithm for predicting continuous values. It models the relationship between the dependent variable (real estate price) and one or more independent variables (property characteristics) by fitting a linear equation to observed data.

2. Decision Trees

- **Decision Tree Regressor**: A non-linear algorithm that splits the data into subsets based on the values of input features. It creates a tree-like model of decisions and their possible outcomes, suitable for regression tasks.

3. Random Forest

- **Random Forest Regressor**: An ensemble learning method that constructs multiple decision trees and merges their results to obtain a more accurate and stable prediction. It helps to reduce overfitting and improves model generalization.

4. Gradient Boosting

- **Gradient Boosting Regressor**: Another ensemble technique that builds models sequentially. Each new model attempts to correct the errors made by the previous models, making it highly effective for regression tasks.

5. Extreme Gradient Boosting (XGBoost)

- **XGBoost Regressor**: An optimized implementation of gradient boosting designed for speed and performance. It handles large datasets and complex models efficiently, providing robust predictions.

6. Support Vector Machines

- **Support Vector Regressor (SVR)**: A supervised learning model that uses regression techniques. It fits the best possible line within a predefined margin, performing well in high-dimensional spaces.

7. k-Nearest Neighbors

- **K-Nearest Neighbors (KNN) Regressor**: An instance-based learning algorithm where the prediction is based on the average value of the nearest k neighbors. It is non-parametric and does not assume any specific distribution for the data.

Multiple Linear Regression Model

The multiple linear regression model is formulated as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon \quad (1)$$

Where:

- y is the dependent variable (e.g., real estate price).
- β0is the intercept term.
- β1,β2,…,βn are the coefficients corresponding to each independent variable.
- x1,x2,…,xnx_1, are the independent variables (e.g., property characteristics such as size, number of bedrooms, location).
- £ is the error term, representing the deviation of the observed values from the predicted values.

In this model:

- The intercept term β0 represents the value of y when all independent variables are equal to zero.
- Each coefficient βi indicates the change in the dependent variable y for a one-unit change in the independent variable xi holding all other variables constant.
- The error term ϵ captures the effects of all other factors that influence y but are not included in the model.

By fitting this model to the data, we can estimate the values of the coefficients β0,β1,β2,…,βn , which allows us to make predictions about the dependent variable y based on new values of the independent variables x1,x2,…,xn.

Matrix Form

The multiple linear regression model can be expressed in matrix form as follows:

$$Y = X\beta + \epsilon \quad (2)$$

Where :

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \text{and} \quad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad (3)$$

2.2 Model Assumptions

The model is based on the following assumptions:

- The values $x_{i,p}$ are error-free.
- $E[\epsilon_i] = 0 \text{ for all } i = 1, \ldots, n,$ the expected value of the error is zero
- $\text{Var}(\epsilon_i^2) = E[\epsilon_i^2] = \sigma_\epsilon^2, \quad \text{Var}(\epsilon_i^2) = E[\epsilon_i^2] = \sigma_\epsilon^2,$ the variance of the error is constant for all $i = 1, \ldots, n.$
- $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \text{ and } E(\epsilon_i, \epsilon_j) = 0 \text{ for all } i \neq j,$, the errors are independent.
- $\text{Cov}(x_{i,p}, \epsilon_j) = 0,$, the error is independent of the explanatory variables.
- Absence of collinearity between the explanatory variables, which implies that the matrix (X′X) is regular and the inverse matrix $(X'X)^{-1}$ exists .
- $(X'X)/n$ tends towards a finite non-singular matrix
- $n > p + 1,$, the number of observations is greater than the number of explanatory variables.

2.3 Estimation of Regression Coefficients

Given the model

$$Y = X\beta + \epsilon. \quad (4)$$

To estimate the vector β composed of the coefficients β0,β1,…,βp we apply the method of ordinary least squares (OLS). This method aims to minimize the sum of the squared errors, or the mean squared error (MSE), between the values predicted by the model and the actual values of the dependent variable.

$$\min \sum_{i=1}^{n} \epsilon_i^2 = \min \epsilon'\epsilon = \min(Y - X\hat{\beta})'(Y - X\hat{\beta}) = \min S \quad (6)$$

To minimize the function S with respect to the vector β, we differentiate S with respect to β and set the derivative equal to zero.

$$\frac{\partial S}{\partial \beta} = 2X'Y + 2X'X\beta = 0 \quad (7)$$

Solving for β :

$$X'X\beta = X'Y$$
$$\beta = (X'X)^{-1}X'Y \quad (8)$$

Thus, the estimated vector of coefficients $\hat{\beta}$ is given by:

$$\hat{\beta} = (X'X)^{-1}X'Y \quad (9)$$

This is the formula for the ordinary least squares (OLS) estimator of β.
The estimated model is written as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_p x_{i,p} + e_i \quad (10)$$

Where:

$$e_i = Y_i - \hat{Y}_i \quad (11)$$

2.4 Properties of Estimators

In the context of multiple linear regression, the ordinary least squares (OLS) estimators have several important properties:

- **Unbiasedness**

The OLS estimators are unbiased, meaning that the expected value of the estimators equals the true parameter values. Mathematically, this is expressed as:

$$E(\hat{\beta}) = \beta \quad (12)$$

This property holds under the assumption that the errors have a mean of zero and are uncorrelated with the independent variables.

- **Efficiency**

The OLS estimators are efficient, meaning they have the smallest variance among all unbiased estimators. This property is a consequence of the Gauss-Markov theorem, which states that OLS estimators are the Best Linear Unbiased Estimators (BLUE) under the assumption of homoscedasticity (constant variance of the error terms)

- **Consistency**

The OLS estimators are consistent, meaning that as the sample size increases, the estimators converge in probability to the true parameter values. Mathematically, this is expressed as:

$$\hat{\beta} \xrightarrow{p} \beta$$

This property holds if the errors are independently and identically distributed with finite variance and if the independent variables are not perfectly collinear.

- **Normality (Under the Assumption of Normally Distributed Errors)**

If the error terms are normally distributed, then the OLS estimators are normally distributed. This property allows for the construction of confidence intervals and hypothesis tests.

- **Minimum Variance**

Among all linear unbiased estimators, the OLS estimators have the minimum variance. This property is also derived from the Gauss-Markov theorem.

Given the model:

$$Y = X\beta + \epsilon \quad (13)$$

- $E[\hat{\beta}] = \beta$: $\hat{\beta}$ is an unbiased estimator of β
- $\hat{\sigma}_\epsilon^2 = \frac{\epsilon'\epsilon}{n-p-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-p-1}$: Estimator of the error variance.
- $\mathrm{Var}(\hat{\beta}) = \hat{\sigma}_\epsilon^2 (X'X)^{-1}$: Variance-covariance matrix of the regression coefficients.
- $\lim_{n\to\infty} \mathrm{Var}(\hat{\beta}) = 0$: $\hat{\beta}$ is a consistent estimator

Gauss-Markov Theorem

Theorem : The OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$ is classified as BLUE (Best Linear Unbiased Estimator), because it is the best linear unbiased estimator in the sense that it provides the smallest variances for the estimators.

Once the model is fitted, it can be used to predict the values of the dependent variable for new values of the independent variables. [7]

3. Measurement of Quality, Performance, and Robustness of Models

When evaluating machine learning models, several metrics and techniques are used to measure their quality, performance, and robustness.

3.1 Analysis of Variance (ANOVA) and Goodness of Fit

ANOVA is a statistical method used to compare the means of three or more samples. It helps in determining if at least one of the sample means is significantly different from the others. For regression models, ANOVA is used to analyze the variance in the data explained by the model compared to the variance explained by residuals (errors).

$$\text{Total Sum of Squares (SST)} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$\text{Explained Sum of Squares (SSR)} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad (14)$$

$$\text{Residual Sum of Squares (SSE)} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

The relationship between these sums of squares is given by:

$$\text{SST} = \text{SSR} + \text{SSE} \quad (15)$$

ANOVA Table Components:

Degrees of Freedom (DF):

- Model (Regression): p (where p is the number of predictors)
- Error (Residual): n-p-1
- Total : n-1

Mean Squares:

$$\text{Model: MSR} = \frac{\text{SSR}}{p} \quad (16)$$

$$\text{Error: MSE} = \frac{\text{SSE}}{n-p-1} \quad (17)$$

F-Statistic:

$$F = \frac{\text{MSR}}{\text{MSE}} \quad (18)$$

The F-statistic is used to test the overall significance of the model. A higher F-value indicates that the model explains a significant portion of the variance in the dependent variable

3.2 Performance Metrics for Supervised Machine Learning Models

Performance metrics are crucial for evaluating supervised machine learning models. To ensure that your model performs well in its predictions, it must be thoroughly evaluated. The objective is to identify the model's performance on new data. Some evaluation metrics can help determine if the model's predictions are accurate for a certain level of performance. For loss functions, we aim to minimize them, while for the coefficient of determination, we aim to maximize it.[21]

- Coefficient of Determination

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \qquad (19)$$

R^2 indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. Values range from 0 to 1, with higher values indicating a better fit.

- Adjusted

$$\text{Adjusted } R^2 = 1 - \left(\frac{\text{SSE}/(n-p-1)}{\text{SST}/(n-1)}\right) \qquad (20)$$

Adjusted R^2 adjusts the R^2 value for the number of predictors in the model. It provides a more accurate measure of goodness of fit for models with multiple predictors.

Performance Metrics

- **Precision, Recall, and F1 Score (for Classification Models)**:
  - **Precision**: The ratio of true positive observations to the total predicted positives.
  - **Recall**: The ratio of true positive observations to all actual positives.
  - **F1 Score**: The harmonic mean of Precision and Recall, providing a single metric that balances both concerns.
- **Confusion Matrix (for Classification Models)**:
  - A table used to describe the performance of a classification model by displaying the true positives, false positives, true negatives, and false negatives.
- **Area Under the ROC Curve (AUC-ROC)**:
  - Measures the ability of the model to distinguish between classes. Higher AUC indicates a better model.

By using these metrics, we can comprehensively evaluate the performance of supervised machine learning models. This ensures that the models not only fit the training data well but also generalize effectively to new, unseen data, providing reliable predictions in real-world applications. The ultimate goal is to minimize loss functions (such as MSE, RMSE, MAE, and MAPE) while maximizing the coefficient of determination (R^2) to achieve the best model performance.

4. Results and Interpretation

The data used to predict housing prices with machine learning algorithms were collected online. This dataset, which contains information on Indian housing, was obtained from the website "www.kaggle.com".

Data Format and Transformation

- The dataset is in **CSV** (Comma-Separated Values) format.
- It was imported and transformed into a **DataFrame** for analysis.

The dataset contains various attributes and characteristics related to houses located in India, which define the real estate prices. Here are some typical attributes that might be included in such a dataset:

1. **Location**: The geographic location of the property.
2. **Size**: The size of the house (e.g., in square feet).
3. **Bedrooms**: The number of bedrooms in the house.

4. **Bathrooms**: The number of bathrooms in the house.
5. **Price**: The price of the house.
6. **Year Built**: The year the house was constructed.
7. **Parking**: Availability of parking space.
8. **Amenities**: Information on available amenities (e.g., swimming pool, gym).
9. **Type of Property**: Type of house (e.g., apartment, villa, independent house).

The dataset used for predicting housing prices contains **14,620 rows** and **23 columns**, which form the explanatory variables and the dependent variable in our models. The data primarily consists of quantitative values distributed across the 23 columns:

- **4 columns** are of type "integer" (int) representing whole numbers without decimal points.
- **19 columns** are of type "real" (float) representing floating-point numbers, which are numbers with a decimal point but no fixed position for the decimal

4.1 Exploratory Data Analysis (EDA)

♣ Descriptive Statistics

The descriptive statistics of the columns in our dataset can be summarized in the following table, which shows the sample size (count), mean, standard deviation (std), minimum value (min), maximum value (max), and quartile values (25%, 50%, 75%) for each column in our dataset:

	count	mean	std	min	25%	50%	75%	max
id	14620.0	6.762821e+09	8237.574799	6.762810e+09	6.762815e+09	6.762821e+09	6.762826e+09	6.762832e+09
number of bedrooms	14620.0	3.379343e+00	0.938719	1.000000e+00	3.000000e+00	3.000000e+00	4.000000e+00	3.300000e+01
number of bathrooms	14620.0	2.129583e+00	0.769934	5.000000e-01	1.750000e+00	2.250000e+00	2.500000e+00	8.000000e+00
living area	14620.0	2.098263e+03	928.275721	3.700000e+02	1.440000e+03	1.930000e+03	2.570000e+03	1.354000e+04
lot area	14620.0	1.509328e+04	37919.621304	5.200000e+02	5.010750e+03	7.620000e+03	1.080000e+04	1.074218e+06
number of floors	14620.0	1.502360e+00	0.540239	1.000000e+00	1.000000e+00	1.500000e+00	2.000000e+00	3.500000e+00
waterfront present	14620.0	7.660739e-03	0.087193	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
number of views	14620.0	2.331053e-01	0.766259	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	4.000000e+00
condition of the house	14620.0	3.430506e+00	0.664151	1.000000e+00	3.000000e+00	3.000000e+00	4.000000e+00	5.000000e+00
grade of the house	14620.0	7.682421e+00	1.175033	4.000000e+00	7.000000e+00	7.000000e+00	8.000000e+00	1.300000e+01
Area of the house(excluding basement)	14620.0	1.801784e+03	833.809963	3.700000e+02	1.200000e+03	1.580000e+03	2.240000e+03	9.410000e+03
Area of the basement	14620.0	2.964791e+02	448.551409	0.000000e+00	0.000000e+00	0.000000e+00	5.800000e+02	4.820000e+03
Built Year	14620.0	1.970926e+03	29.493625	1.900000e+03	1.951000e+03	1.975000e+03	1.997000e+03	2.015000e+03
Renovation Year	14620.0	9.082401e+01	416.216661	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	2.015000e+03
Postal Code	14620.0	1.220331e+05	19.082418	1.220030e+05	1.220170e+05	1.220320e+05	1.220480e+05	1.220720e+05
Lattitude	14620.0	5.279285e+01	0.137522	5.238590e+01	5.270760e+01	5.280640e+01	5.290890e+01	5.300760e+01
Longitude	14620.0	-1.144040e+02	0.141326	-1.147090e+02	-1.145190e+02	-1.144210e+02	-1.143150e+02	-1.135050e+02
living_area_renov	14620.0	1.996702e+03	691.093366	4.600000e+02	1.490000e+03	1.850000e+03	2.380000e+03	6.110000e+03
lot_area_renov	14620.0	1.275350e+04	28058.414467	6.510000e+02	5.097750e+03	7.620000e+03	1.012500e+04	5.606170e+05
Number of schools nearby	14620.0	2.012244e+00	0.817284	1.000000e+00	1.000000e+00	2.000000e+00	3.000000e+00	3.000000e+00
Distance from the airport	14620.0	6.495096e+01	8.936008	5.000000e+01	5.700000e+01	6.500000e+01	7.300000e+01	8.000000e+01
Price	14620.0	5.389322e+05	367532.380804	7.800000e+04	3.200000e+05	4.500000e+05	6.450000e+05	7.700000e+06

Fig -1: Descriptive Statistics of the Dataset

♣ **Correlation Between Variables**

The following figure is a correlation matrix that illustrates the correlations and linear similarities between the different columns in this dataset. Some columns have a strong correlation greater than 0.5 with each other, but most show a coefficient less than 0.5. The columns "living area" and "Area of the house (excluding basement)" have the strongest correlation, with a correlation coefficient equal to 0.88.

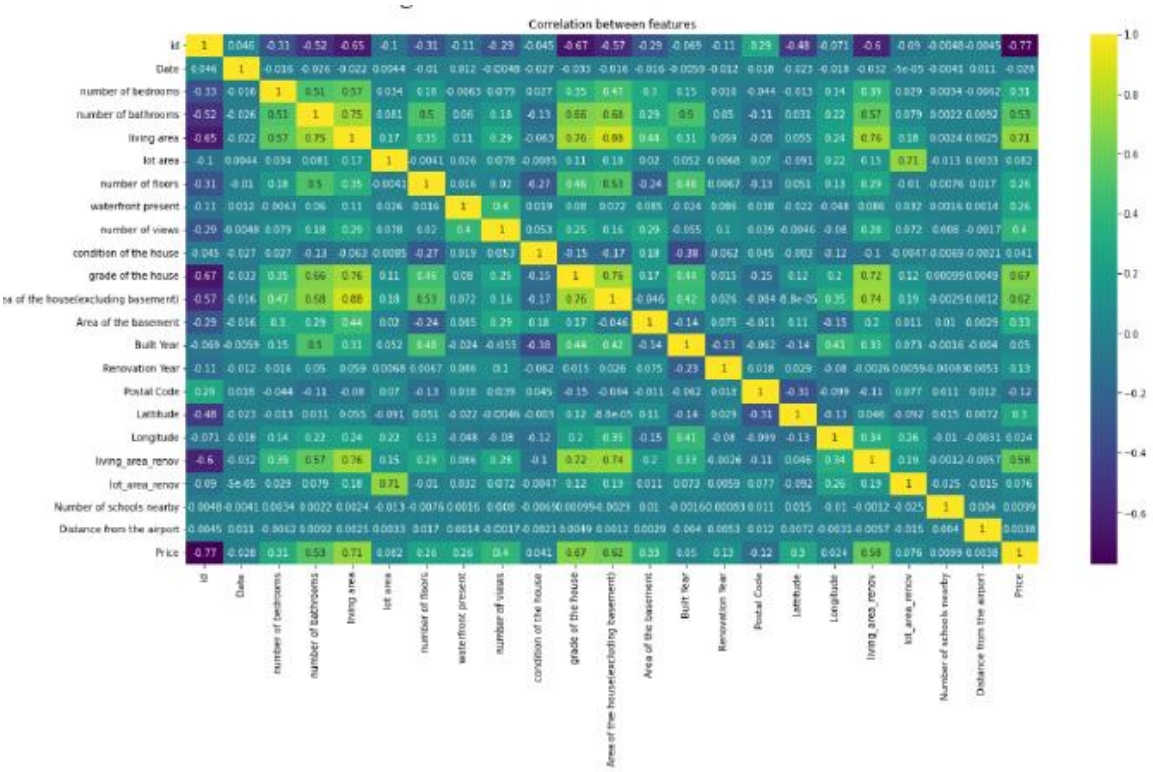

Fig 2 : Correlation Between Variables

4.2 Machine Learning Results

- **Modeling**

In the modeling phase, we applied various machine learning algorithms to predict housing prices based on the dataset attributes. Here, we describe the steps involved in building and evaluating the models.

In the principle of machine learning, solving a problem is essentially based on econometric modeling. The model is trained on historical data to identify critical features through the data and predict the required variable accordingly. The objective is to judge the model's performance on out-of-sample data.

Using the entirety of the data to train the model allows for a highly accurate model that memorizes almost all the characteristics of the training data and will perform extremely well on the training data. However, it may fail on test data because the model is too complex and cannot generalize properly. This leads to overfitting. Splitting the out-of-sample data into training and test data helps calculate the training and testing accuracy of the model, both of which need to be good for the model to be used for future prediction purposes. If the model is overly simplified, it will perform poorly on both training and test data (underfitting), and if we make it too complicated, it will perform very well on the training data but poorly on the test data (overfitting).

In this case, the size of the training and test data is divided using the train_test_split function, specifying the proportion of data to include in the test set with the argument test_size=0.2 and the argument random_state=101 to set a seed for reproducibility of the results. Thus, 80% of the data is used for training and 20% for testing.

- **Comparison of Model Performances**

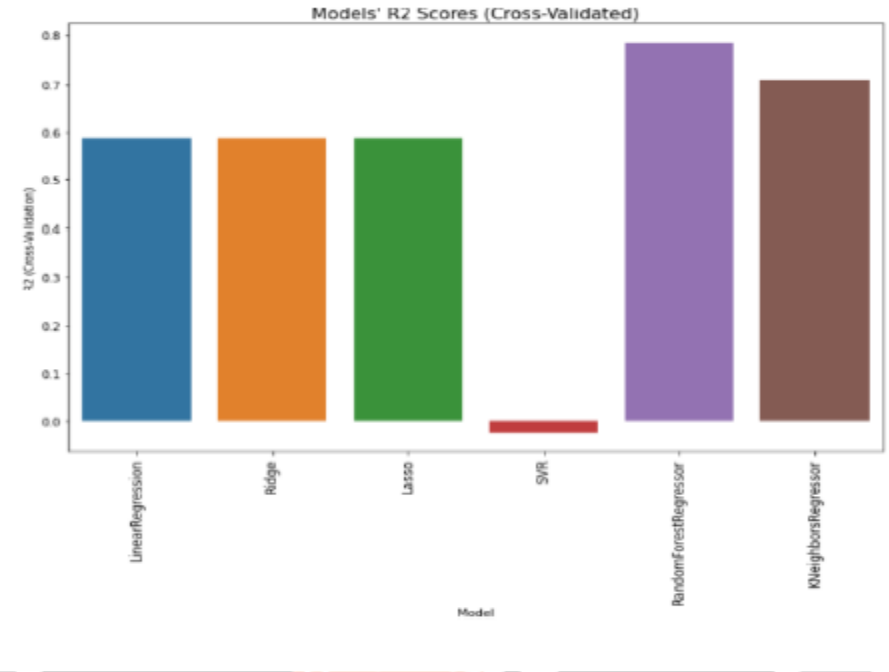

Fig 3 : Comparison of R^2 Cross-Validation for Regression Models

Based on the cross-validation R^2 scores, we can compare the performance of different regression models and determine which model is best suited for predicting real estate prices. The Random Forest Regressor shows the highest R^2 score, indicating that it generalizes well to unseen data and captures the underlying patterns in the dataset effectively.

- **Model Optimization**

Optimizing regression models involves fine-tuning hyperparameters, modifying model architectures, or applying optimization techniques to improve model performance. Here, we outline the steps to optimize the Random Forest Regressor and other regression models using hyperparameter tuning.

	Model_hyper	MAE	MSE	RMSE	R2 Score	R2 (Cross-Validation)
3	SVR	101547.234649	1.882148e+10	137191.415895	0.420694	0.395788
0	LinearRegression	88763.683351	1.387842e+10	117806.696102	0.572836	0.586722
2	Lasso	88763.247920	1.387823e+10	117805.891972	0.572842	0.586740
1	Ridge	88760.650514	1.387794e+10	117804.653090	0.572851	0.586744
5	KNeighborsRegressor	66419.214162	8.850680e+09	94078.055267	0.727585	0.729544
4	RandomForestRegressor	59100.255188	7.217382e+09	84955.177252	0.777856	0.784889

Fig 4 : Model Performance Evaluation Metrics

5. CONCLUSIONS

This thesis project explored the use of Artificial Intelligence (AI) for predicting real estate prices in India, based on property characteristics and location. Through in-depth data analysis and the application of advanced Machine

Learning techniques, significant insights into the distribution and trends in the Indian real estate market were obtained. Key Findings :

- Exploratory Data Analysis (EDA):
    - Understanding Data Distribution: EDA helped in understanding the distribution of the data and the correlation between different variables.
    - Identifying Key Determinants: The analysis identified the main determinants of real estate prices in India.
    - Feature Selection: Relevant features were selected to guide our analysis.
    - Handling Outliers: Outliers in the dataset were identified and addressed.
- Machine Learning Models:
    - High Performance: Machine Learning models, particularly the Random Forest Regressor and the KNN Regressor, demonstrated high performance in predicting real estate prices.
    - Evaluation Metrics: The models achieved competitive values for MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R^2 (Coefficient of Determination), and Cross-Validation R^2 on the test dataset.
    - Best Performing Model: The Random Forest Regressor was identified as the best-suited model for predicting real estate prices, with an accuracy of 78.44%.

This study contributes to a better understanding of the mechanisms of the Indian real estate market and offers valuable insights for future research in this area. By integrating advances in Artificial Intelligence with real estate industry expertise, we can create innovative tools and solutions to address the complex challenges of the real estate market in India and globally.

6. REFERENCES

[1]. Bernard THION, « VALEUR, PRIX ET METHODES D'EVALUATION EN IMMOBILIER », CEREG, Université Paris 9-Dauphine, 26.p

[2]. Deeba, F., Dharejo, F. A., Zawish, M., Memon, F. H., Dev, K., Naqvi, R. A., ... & Du, Y. (2021). A novel image dehazing framework for robust vision-based intelligent systems. International Journal of Intelligent Systems

[3]. Dharejo, F. A., Deeba, F., Zhou, Y., Das, B., Jatoi, M. A., Zawish, M., ... & Wang, X. (2021). TWIST-GAN: Towards wavelet transform and transferred GAN for spatio-temporal single image super-resolution. ACM Transactions on Intelligent Systems and Technology (TIST), 12(6), 1-20.

[4]. Du, Y., Wang, H., Cui, W., Zhu, H., Guo, Y., Dharejo, F. A., & Zhou, Y. (2021). Foodborne disease risk prediction using multigraph structural long short-term memory networks: algorithm design and validation study. JMIR Medical Informatics, 9(8), e29433.

[5]. Jean Cavailhès, « Le prix des attributs du logement », CONDITIONS DE VIE, ÉCONOMIE ET STATISTIQUE N° 381-382, 2005, p.91 - 123

[6]. Marchand Olivier, Skhiri Eric, « *Prix hédoniques et estimation d'un modèle structurel d'offre et de demande de caractéristiques* ». In : Économie & prévision, n°121, 1995-5. Comportements des ménages. pp. 127-140

[7]. G. (2022). Research on Prediction and Analysis of Real Estate Market Based on the Multiple Linear Regression Model. Scientific Programming, vol. 2022, pp. 1-8.