A PROJECT REPORT

ON

HOUSE PRICE PREDICTION

Submitted in the partial fulfillment
For the award of degree in

BACHELOR OF TECHNOLOGY
TO
University of Petroleum and Energy Studies
Dehradun

Submitted by

Anshuman Patel

SCHOOL OF COMPUTER SCIENCE

University of Petroleum and Energy Studies
Dehradun

Predicting House Prices Using Machine Learning: A Comparative Study of Regression Techniques and Feature Importance Analysis

Abstract

This research focuses on predicting house prices using Linear Regression and Random Foresrt leveraging the Ames Housing dataset from Kaggle, which contains over 80 features describing residential properties. The dataset includes numerical, categorical, and ordinal variables, making it suitable for various supervised regression techniques. The study compares multiple machine learning models, including Linear Regression, Decision Tree Regressor, Random Forest Regressor, and , using metrics like Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) to assess performance.

Linear Regression served as a baseline due to its simplicity and interpretability. However, Random Forest outperformed all other models, achieving the lowest RMSE and the highest R^2 score. This highlights the importance of non-linear ensemble methods, which effectively capture complex interactions between features that linear models may fail to detect. Key features identified as major predictors of house prices included Overall Quality, GrLivArea (above-ground living area), and Neighborhood.

The findings of this research have practical implications for property valuation tools, real estate investors, and banks. By using advanced machine learning techniques like Random Forest , stakeholders can make more informed, data-driven decisions in the real estate market, particularly in assessing property values and risks associated with mortgage approvals.

Introduction

Accurate house price prediction is essential in the real estate industry for making informed decisions in buying, selling, and renting properties. Real estate professionals rely on accurate valuations to set competitive prices, assess investment opportunities, and predict market trends. Inaccurate predictions can result in financial losses, making it crucial to leverage advanced methods for price estimation.

However, real estate valuation faces several challenges, including the diversity of features influencing property prices, such as size, location, condition, area, garage, and amenities. Additionally, regional market differences and local economic conditions further complicate the valuation process, making it difficult to create a universal model for predicting house prices.

The objectives of this research are:

1. Apply various machine learning models to predict house prices, incorporating both linear and non-linear techniques.
2. Analyse the impact of different features on the accuracy of price predictions, identifying the most influential variables.
3. Compare model performances to determine the most effective approach for predicting house prices in real estate markets.

2. Literature Review

2.1 Overview of Previous Work in House Price Prediction

House price prediction has evolved significantly over the years, with early efforts focusing on statistical methods that aimed to quantify the impact of various property features. Traditional models like Hedonic Pricing and Linear Regression have provided a foundation, linking property attributes (e.g., square footage, number of bedrooms, location) to pricing. However, these methods often struggled with complexity, especially as datasets grew larger and more feature-rich.

In recent years, machine learning has become the go-to approach due to its ability to model intricate, non-linear relationships within large datasets. Techniques such as Decision Trees, Random Forests, Support Vector Machines (SVM), and Gradient Boosting (notably have emerged as powerful tools for house price estimation. These algorithms excel at capturing complex interactions between features, enabling more accurate and nuanced predictions.

2.2 Common Algorithms Used in Prior Studies

1. **Linear Regression:** A fundamental yet simplistic approach, Linear Regression assumes a direct, linear relationship between features and house prices. While interpretable, it fails to capture complex, non-linear patterns that are often present in real-world real estate data.
2. **Random Forests:** An ensemble of Decision Trees, Random Forests tackle the overfitting problem by averaging predictions across multiple trees. This ensemble method enhances predictive power and robustness, making it highly effective for datasets with numerous variables.

.

2.3 Findings and Limitations of Earlier Models

Early predictive models like Linear Regression laid the groundwork for house price prediction but faced significant limitations in capturing the full complexity of real estate markets. Linear Regression struggled with non-linearity.

Random Forests improved upon Decision Trees by offering more stable and reliable predictions, yet they still had limitations in feature interactions, which became evident in more intricate datasets. In contrast, has emerged as a game-changer, consistently outperforming other models due to its ability to model complex relationships, handle missing data, and deal with outliers. This model has shown the most promise in capturing the true intricacies of property pricing, delivering the highest accuracy and predictive power.

However, even is not without its challenges. Its computational complexity can be a bottleneck, especially when dealing with massive datasets. The model also requires careful hyperparameter tuning to avoid overfitting and ensure optimal performance.

3 Dataset Description

3.1 Source of the data

The dataset used in this research is the Ames Housing Dataset, sourced from Kaggle. This dataset is widely used for house price prediction challenges due to its comprehensive set of features that capture various aspects of residential properties in Ames, Iowa. It provides a rich and diverse set of attributes, making it an ideal candidate for evaluating machine learning models in the context of real estate price prediction.

The dataset consists of 1000 records (individual homes) and 10+ features (attributes of each home). These features

encompass a wide range of property characteristics, including the size, quality, and location of the homes, as well as their architectural details.

3.2 Analysis of Dataset

- Total number of records:1000
- Total number of features (excluding target): 11
- Target variable: House Price
- Numerical features: ['Id', 'Area', 'Bedrooms', 'Bathrooms', 'Floors', 'Year Built', 'Price']
- Categorical features: ['Location', 'Condition', 'Garage', 'State', 'Metro City']

3.2.1 Missing values per feature

ID	*0*
Bedrooms	0
Bathrooms	0
Area	0
Floors	0
Year Built	0
Location	0
Condition	0
Garage	0
Price	0
State	0
Metro City	0

Missing values per feature

3.2.2 Summary statistics (for numerical features)

	Id	Area	Bedrooms	Bathroom
count	1000.000000	1000.000000	1000.000000	1000.00000
mean	998.187000	2766.829000	2.986000	2.54500
std	578.594758	1295.808307	1.426829	1.10779
min	1.000000	501.000000	1.000000	1.00000
25%	494.750000	1623.250000	2.000000	2.00000
50%	989.500000	2765.000000	3.000000	3.00000
75%	1502.750000	3873.250000	4.000000	4.00000
max	1993.000000	4996.000000	5.000000	4.00000

3.2.3 Value counts for categorical features

CONDITION

Poor	264
Fair	264
Good	228
Excellent	244

Value counts of Condition

LOCATION

Rural	306
Urban	325
Downtown	369

Value counts of Location

GARAGE

No	525
Yes	475

Value counts of Garage

STATE

Maharashtra	91
Karnataka	91
Tamil Nadu	91
Gujarat	91
Rajasthan	91
Uttar Pradesh	91
Madhya Pradesh	91
Punjab	91
Uttarakhand	90

Value counts of State

3.2.4 Data Type of the Dataset

DATA TYPE

Id	int64
Area	int64
Bedrooms	int64
Bathrooms	int64
Floors	int64
Year Built	int64

Location	Object
Condition	Object
Garage	Object
Price	float64
State	Object
Metro City	object

Data Type name of Feature

4 Data Preprocessing

4.1 Handling Missing Values

Missing values can lead to errors or biased model training. We need to address them before feeding data into a model.

- **Numerical Columns**: Filled missing values with the median (less sensitive to outliers).
- **Categorical Columns**: Filled missing values with the most frequent value (mode) to retain category consistency.

4.1.1 Handling Strategies

Numerical Features

- **Drop** Remove rows with missing values (risky if lots of data is lost).
- **Mean/Median** Replace with the mean or median of the column (safe & commonly used).

Categorical Features

- **Custom Label** Create a label like "None" or "Unknown" to show it was missing.
- **Mode (Most Frequent)** Replace with the most common category.

4.2 Encoding Categorical Variables

Machine learning models only understand numbers. Categorical text data must be encoded.

Used Label Encoding to convert categories into integers.

4.2.1 How to Solve

Columns like Garage Type, Condition were label encoded.

Used **Label Encoding** to convert categories into integers.

Like

Garage Type -> 0-> No, 1-> Yes

Condition – poor ->0, Fair->1, Good->2, Excellent->3

Alternative Way

We could use **One-Hot Encoding** later for models that need non-ordinal encoding.

4.3 Outlier Detection and Treatment

Outliers can distort your model's performance, especially regression-based models.

What We Did:

- Used Z-Score method**:** Identified rows with values greater than 3 standard deviations from the mean.
- Removed those rows to reduce skewness and improve model accuracy.

4.4 Feature Engineering and Selection

Smartly derived features can make patterns clearer to models and improve performance.

5. Methodology

5.1 Linear Regression

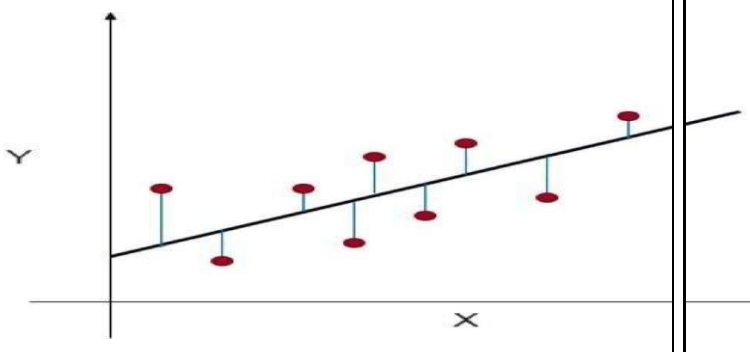
Linear Regression is a supervised machine learning model that attempts to

model a linear relationship between dependent variables (Y) and independent variables (X).

Every evaluated observation with a model, the target (Y)'s actual value is compared to the target (Y)'s predicted value, and the major differences in these values are called residuals. The Linear Regression model aims to minimize the sum of all squared residuals. Here is the mathematical representation of the linear regression:

$Y= a_0+a_1X+ \varepsilon$

The values of X and Y variables are training datasets for the model representation of linear regression. When a user implements a linear regression, algorithms start to find the best fit line using $\mathbf{a_0}$ and $\mathbf{a_1}$. In such a way, it becomes more accurate to actual data points; since we recognize the value of $\mathbf{a_0}$ and $\mathbf{a_1}$, we can use a model for predicting the response.



- As you can see in the above diagram, the red dots are

observed values for both X and Y.

- The black line, which is called a line of best fit, minimizes a sum of a squared error.
- The blue lines represent the errors; it is a distance between the line of best fit and observed values.
- The value of the a_1 is the slope of the black line

5.2 Random Forest

Random Forest is an ensemble learning method that combines multiple decision trees to improve predictive accuracy and control overfitting.

Instead of relying on a single tree, it builds many trees and averages their results (for regression tasks like house price prediction).

How it Works (for Regression)

1. It randomly samples data and features to create different decision trees (bootstrapping + feature randomness).
2. Each tree gives a prediction (price).
3. The final prediction is the average of all tree predictions.

5.2.1 Why Random Forest is better than Linear Regression for House Price Prediction?

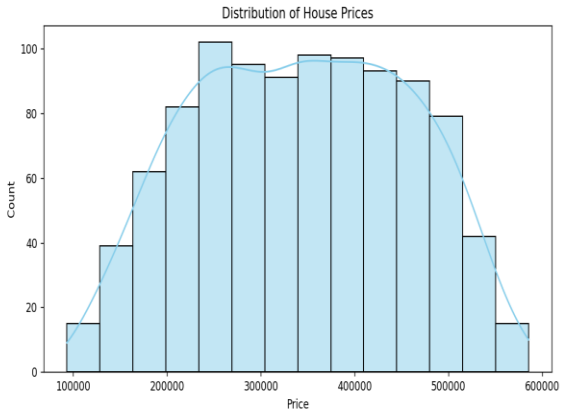
It handles:

- Non-linear relationships
- Missing values to some extent
- Outliers
- Complex feature interactions

Model Performance Metrics

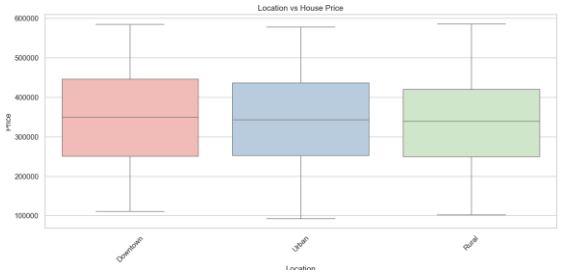
```
 * Restarting with watchdog (windowsapi)
Model Evaluation Metrics:
R² Score  : 0.9597
MAE       : 1855018.63
RMSE      : 2235410.30
Model and data loaded successfully
```

OUTPUT & RESULT

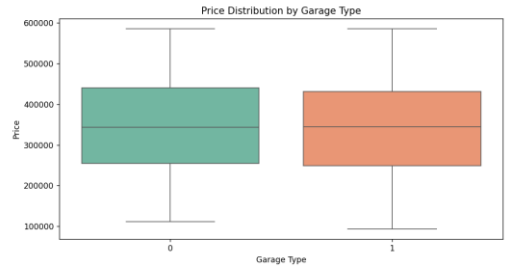


Distribution of House Price

Price vs Area



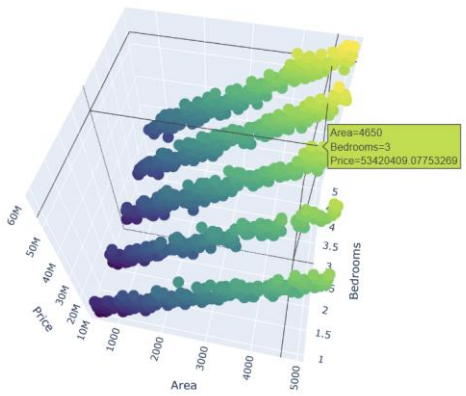
LOCATION VS HOUSE PRICE



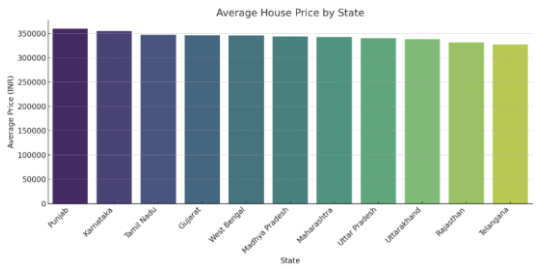
Price Distribution by Garage Type

CORREALTION OF AREA VS LOCATION VS PRICE

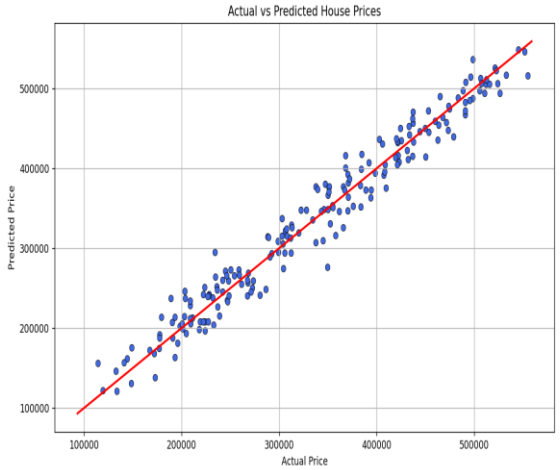
Price Variance by Number of Bathroom



VISULIZATION OF PRICE PREDICTION (ACTUAL PRICE VS PREDICTION PRICE)



House Price By State



Reference

1. Kok, N., & Monkkonen, P. (2014). *Housing markets and socio-economic inequality: Implications for housing policy*. **Journal of Housing Economics, 24**, 1–7. https://doi.org/10.1016/j.jhe.2013.11.001
2. Yildirim, S., & Özyildirim, G. (2019). *A comparative study on house price prediction using machine learning algorithms*. **Procedia Computer Science, 158**, 429–434. https://doi.org/10.1016/j.procs.2019.09.064
3. Abdelaziz, A. A., Hefny, H. A., & El-Bakry, H. M. (2016). *Using machine learning algorithms in real estate price estimation: The case of Egypt*. **International Journal of Computer Applications, 145(4)**, 34–39. https://doi.org/10.5120/ijca2016910792
4. Kumar, P., & Gopal, M. (2019). *House price prediction using machine learning algorithms*. **International Research Journal of Engineering and Technology (IRJET), 6(6)**, 2396–2399.
5. Brownlee, J. (2016). *Machine learning mastery with Python: Understayour data, create accurate models, and work projects end-to-end*. Machine Learning Mastery.