| Academic Year | Module | Assessment Number | Assessment Type |
|---|---|---|---|
| 2024 | 5CS037/HJ1: Concepts and Technologies of AI (Herald College, Kathmandu, Nepal) | ⭐ Final portfolio Project | An End- to- End Machine Learning Project on Regression and Classification Task |

# Real Estate Regression Report

**Student Name: Mohammad Rashid Siddiqui Student**

**ID: 2413748**

**Module Leader: Mr. Siman Giri Tutor:**

**Ms. Durga Pokharel**

## Abstract

This report deals with predicting house prices using regression techniques. The data treats different characteristics of each home, such as square footage, number of bedrooms, and location; all of which influence the sale prices. The study included a structured approach of Exploratory Data Analysis (EDA), model building with linear regression, a decision tree, and KNN which is supported by hyperparameter tuning techniques, and feature selection. The performance of the models was evaluated in terms of their predictive power using R-squared ($R^2$) and Mean Squared Error (MSE). After testing out multiple models, the KNN optimized model was found to provide the most accurate price forecast. Results confirm the belief that square footage and location were important in determining house prices, and there are further developments possible if additional real-time factors were taken into consideration.

## 1. Introduction

### 1.1 Problem Statement

The prices of houses change according to many factors, such as their size, location, and supply and demand in the market. Governments, real estate agents, and sellers need accurate price predictions in order to make informed decisions. This project aims to come up with a machine-learning model that predicts house prices based on various attributes of the properties. Historical real estate data will be analyzed so as to come up with the most significant features that affect house prices, which will then be used in a predictive model.

### 1.2 Dataset

The dataset in this study originates from properties for sales with detailed information on houses like the number of bedrooms, square footage, location, and year built. Since the location is a source of great influence on house prices, the data sheds light on the various contributions of other features into the values of properties. It also aligns itself with the United Nations Sustainable Development Goals (UNSDG) for Sustainable Cities and Communities because it will provide data-driven insights into housing affordability.

### 1.3 Objective

The project's primary aim is to create an accurate regression model for predicting house prices based on various property characteristics. The study is going to establish the most important factors in affecting the house value and tuned the model for optimum performance.

## 2. Methodology

### 2.1 Data Preprocessing

The training process for the model began with conducting several preprocessing steps to ensure data consistency and accuracy. Missing values were then treated adequately by replacing them with mean or, if that was not possible, with the median. Categorical features, such as neighborhood location, were converted into numerical values with Label Encoding, which enabled the models to interpret them appropriately. In addition, StandardScaler was employed to normalize the numerical features, making sure all variables existed on a similar scale.

### 2.2 Exploratory Data Analysis (EDA)

In order to develop an insight into the dataset and its important connections, exploratory data analysis was carried out. This involved the construction of various visualizations, including scatter plots and heat maps, to analyze the correlation of the features. The analysis revealed that square footage is one of the factors most correlated with price-since larger homes can generally be sold at higher values. Another very important part was the location: houses in prime areas generally sold at a higher price.

### 2.3 Model Building

To create a good predictive model,
The following three different types of regression were used:

- Linear Regression-Particularly simple, presumes a linear relationship between house features and prices.
- Decision Tree Regressor-According to this, a tree-based model captures the complicated relationships in the data.
- K-Nearest Neighbor (KNN) Regressor-This can also be used that predicts house prices based on the prices of similar homes.

The dataset was split into 80% training and 20% testing, ensuring that the models learned from historical data before making predictions on unseen data.

### 2.4 Model Evaluation

Of all models developed, two basic metrics were used to evaluate their performances:

R-squared ($R^2$): Measures how well the model explains variations in house prices. The higher the $R^2$ value, the better it would be.

Mean Squared Error (MSE): The average of the squared difference between actual and predicted prices, where low values indicate better performance.

### 2.5 Hyperparameter Optimization

Although hyperparameter tuning using GridSearchCV was done to increase model accuracy and optimize the KNN model, the process goes as follows-A value for k number of neighbors and distance metric was chosen based. This is how the tuning was done to get the- -best value of k and the most .suitable distance metric to predict price. After tuning, the optimized KNN model outperformed all other models, achieving the best predictions.

### 2.6 Feature Selection

To identify the most important factors influencing house prices, Recursive Feature Elimination (RFE) was applied. The analysis confirmed that square footage and location were the most crucial features, while factors like the number of floors and year built had less impact on price predictions.

## 3. Conclusion

### 3.1 Key Findings

The study revealed that the optimized KNN model performed better than Linear Regression and Decision Tree models, achieving the highest $R^2$ score. The results confirmed that square footage and location are the strongest predictors of house prices. This suggests that when estimating property values, buyers and sellers should focus on these features the most.

### 3.2 Final Model

The final model selected for house price prediction was the optimized KNN Regressor, as it demonstrated the highest accuracy. After hyperparameter tuning, it provided the most reliable predictions compared to other models.

### 3.3 Challenges

The study faced several challenges. Missing data was by far one of the most important problems to account for, which meant that there was a need for careful imputation for these values to ensure their results do not seem far-fetched. The other biggest challenge was probably external economic environment factors like interest rates and inflation that were not included in the dataset but could severely affect house prices.

### 3.4 Future Work

To further enhance the accuracy of house price predictions, future work could involve testing advanced models such as Random Forest or Gradient Boosting. Additionally, incorporating economic indicators like inflation rates, mortgage rates, and housing demand trends could lead to better predictions.

## 4. Discussion

### 4.1 Model Performance

Among all the models tested, the KNN model performed the best after hyperparameter tuning, achieving the highest $R^2$ score and the lowest prediction error. This suggests that KNN is well-suited for real estate price prediction when optimized correctly.

### 4.2 Impact of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning significantly improved the performance of the KNN model, demonstrating the importance of selecting the right parameters for price prediction. Additionally, feature selection using RFE confirmed that square footage and location are the two most important features, allowing the model to focus on the most relevant factors.

### 4.3 Interpretation of Results

The results confirmed that larger homes and properties in desirable locations command higher prices, which aligns with real estate market trends. This study shows that machine learning models can provide valuable insights into property valuation, helping buyers and investors make better decisions.

### 4.4 Limitations

While the model performed well, some limitations remain. The dataset did not account for market fluctuations, economic conditions, and seasonal changes, which can also affect house prices. Additionally, the dataset was limited to specific locations, meaning the model may not generalize well to all regions.

### 4.5 Suggestions for Future Research

Future research could explore deep learning techniques such as Neural Networks, which can capture more complex relationships in house pricing data. Additionally, including external economic factors like interest rates and inflation could further improve the model’s predictive power.

## Final Thoughts

This study demonstrated that machine learning regression models can accurately predict house prices based on key property attributes. The optimized KNN model provided the best results, reinforcing the importance of hyperparameter tuning and feature selection. By leveraging these models, real estate professionals, investors, and homebuyers can make data-driven decisions, leading to more informed property valuations. In the future, integrating additional economic indicators and testing more advanced models could further refine the accuracy of house price predictions.