Regression and Classification of Housing Prices using Machine Learning

Monowar Islam
Department of Computer Science and Engineering
University of Dhaka, Bangladesh
Email: dew.shishir5000@gmail.com

***Abstract*—House price prediction is a significant problem in real estate analytics. This paper applies machine learning techniques including Linear Regression, Logistic Regression, and Support Vector Machine (SVM) to analyze housing data and classify properties into high-priced and low-priced categories. The study compares model performances using accuracy and error metrics.**

I. Introduction

Predicting house prices plays an important role in real estate decision-making. Machine learning provides powerful tools for modeling complex relationships between housing attributes such as area, number of bedrooms, and facilities. This work uses supervised learning methods to perform both regression and classification on housing data.

II. Methodology

A. Dataset

The dataset contains housing attributes including area, bedrooms, bathrooms, parking, air conditioning, furnishing status, and price. The target variable is house price.

B. Preprocessing

Categorical attributes were converted into numerical values. The yes/no fields were encoded as 1 and 0. Furnishing status was encoded as furnished=2, semi-furnished=1, and unfurnished=0. The dataset contained no missing values.

C. Target Variables

For regression, the target is the continuous house price. For classification, houses were labeled as high price (1) or low price (0) based on the median price.

D. Train-Test Split

The dataset was divided into 70% training data and 30% testing data.

III. Models

A. Linear Regression

Linear Regression was used to predict continuous house prices based on multiple input features.

B. Logistic Regression

Logistic Regression was applied to classify houses into high-price and low-price categories.

C. Support Vector Machine

Support Vector Machine (SVM) was used with two kernels: linear and radial basis function (RBF), to perform classification.

IV. Results

A. Linear Regression

The Root Mean Squared Error (RMSE) obtained was:

$$RMSE = 435,304$$

B. Logistic Regression

The classification accuracy was:

$$Accuracy = 91.46\%$$

The confusion matrix is shown in Table I.

TABLE I
Confusion Matrix for Logistic Regression

	Predicted Low	Predicted High
Actual Low	80	6
Actual High	8	70

C. Support Vector Machine

TABLE II
SVM Kernel Comparison

Kernel	Accuracy
Linear	80.49%
RBF	81.10%

The RBF kernel performed better than the linear kernel.

V. Conclusion

Linear Regression successfully predicted housing prices with low error. Logistic Regression achieved the highest classification accuracy of 91.46%. The SVM with RBF kernel outperformed linear SVM, indicating that housing data contains non-linear patterns. Overall, Logistic Regression provided the best classification performance for this dataset.

Acknowledgment

The author thanks Skill Morph Research Lab for providing the dataset and guidance.