House Price Prediction using Linear Regression (From Scratch)

Name: Abhishek
Program: Data Science & Machine Learning Internship
Project Title: House Price Prediction using Linear Regression (From Scratch)

A Beginner Friendly Machine Learning Guide Without Using Scikit-Learn

Prepared By:
Abhishek

Domain:
Data Science / Machine Learning

Objective:
To understand how Linear Regression actually learns from data by manually implementing the algorithm using Python, NumPy, and Pandas instead of prebuilt libraries.

1. Introduction

Machine Learning is a branch of Artificial Intelligence that allows computers to learn patterns from data and make predictions without being explicitly programmed. Instead of writing fixed rules, a model studies historical data and discovers relationships between input features and the target output.

In the real estate industry, estimating the price of a house is an important problem. Buyers want to know whether a property is worth its price, and sellers want to set a competitive and fair value. Traditionally, prices are estimated manually based on experience, location, and property features. However, manual estimation can be inconsistent and time-consuming.

This project aims to build a predictive model that estimates house prices based on property features such as area, number of bedrooms, bathrooms, parking availability, and facilities. To understand how machine learning works internally, Linear Regression is implemented from scratch using Python, NumPy, and Pandas without using scikit-learn.

The objective of this project is not only to predict house prices but also to understand the internal mathematical working of a machine learning algorithm, including preprocessing, training using gradient descent, and performance evaluation.

2. Problem Statement

House price prediction is a real-world regression problem where the goal is to estimate the selling price of a property based on its characteristics. The dataset contains various features such as area, bedrooms, bathrooms, number of floors, parking spaces, and facilities like air conditioning and basement.

The challenge is to develop a machine learning model that learns the relationship between these features and the house price. The model should be able to predict the price of a new house that it has never seen before.

The task is solved using Linear Regression implemented manually without machine learning libraries, so that the learning process of the algorithm can be clearly understood.

3. Dataset Description

The dataset used in this project is a housing price dataset containing 545 records and 13 attributes describing the characteristics of different houses. Each row in the dataset represents a single property, and the target variable is the selling price of the house.

The dataset contains both numerical and categorical features. Numerical features include area, number of bedrooms, bathrooms, stories, and parking spaces. Categorical features describe facilities and conditions of the property such as main road connectivity, guest room availability, basement availability, hot water heating, air conditioning, preferred area, and furnishing status.

Before training the machine learning model, categorical attributes were converted into numerical values. Binary categorical values such as “yes” and “no” were mapped to 1 and 0. The multi-category feature “furnishing status” was converted into multiple columns using One-Hot Encoding.

After preprocessing, the dataset was ready for training the Linear Regression model.

4. Methodology

The project follows a structured machine learning workflow consisting of data preprocessing, feature scaling, model training, and evaluation.

First, the dataset was analyzed using Exploratory Data Analysis (EDA). Histograms, scatter plots, box plots, and correlation heatmaps were used to understand feature distributions and relationships between variables. It was observed that house price has a positive relationship with features such as area, bathrooms, and number of stories.

Next, data preprocessing was performed. Categorical variables were converted into numerical form using mapping and One-Hot Encoding so that the machine learning algorithm could process the data.

After preprocessing, the dataset was divided into training and testing sets. The training set (80%) was used to learn the model parameters, and the testing set (20%) was used to evaluate performance on unseen data.

Feature scaling was then applied using standardization. The mean and standard deviation were calculated from the training data only and applied to both training and testing sets to avoid data leakage.

Linear Regression was implemented from scratch using gradient descent optimization. The model started with initial weights set to zero and iteratively updated them to minimize the Mean Squared Error loss function.

Finally, the trained model was evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R^2 score to measure prediction accuracy.

5. Results and Discussion

After training the Linear Regression model, its performance was evaluated on the test dataset containing unseen house records.

The model achieved a Root Mean Squared Error (RMSE) of approximately 1.1 million. RMSE represents the average difference between the predicted house price and the actual house price in real-world units. This means that on average, the model's predictions differ from the true price by about this amount.

The R^2 score obtained was approximately 0.64. The R^2 score indicates how well the model explains the variation in the target variable. An R^2 value of 0.64 means that around 64% of the variation in house prices is explained by the features used in the model, while the remaining variation may be due to factors not present in the dataset such as location quality, neighborhood development, and property age.

The loss curve showed a steady decrease during training, indicating that the gradient descent algorithm successfully minimized the cost function and converged to stable model parameters.

Residual analysis showed that prediction errors were distributed around zero, suggesting that the model does not show strong bias toward overestimation or underestimation for most houses.

Overall, the model captures the general trend of housing prices but cannot perfectly predict prices because real estate values depend on many external factors not included in the dataset.

6. Conclusion

In this project, a Linear Regression model was successfully implemented from scratch without using machine learning libraries such as scikit-learn. The project demonstrated the complete machine learning workflow including data preprocessing, feature encoding, feature scaling, gradient descent training, and model evaluation.

The model was able to learn meaningful relationships between housing features and price. Evaluation metrics indicated moderate predictive performance, showing that Linear Regression can effectively capture general pricing trends in the dataset.

However, prediction accuracy is limited due to the absence of important real-world factors such as geographic location, surrounding infrastructure, and property age. Including these features or using more advanced machine learning models could further improve prediction performance.

This project helped in understanding how machine learning algorithms work internally rather than relying only on library functions.

7. Limitations

Although the model provides useful predictions, it has several limitations.

The Linear Regression model assumes a linear relationship between input features and house price. In real-world housing markets, price is influenced by complex and non-linear factors, so a simple linear model cannot fully capture actual behavior.

The dataset does not include important variables such as geographic coordinates, distance to city center, nearby schools, crime rate, or age of the property. These missing features reduce prediction accuracy.

The model is also sensitive to outliers. Extremely high-priced houses can significantly affect the learned weights and increase prediction error.

8. Future Improvements

The project can be improved in several ways.

More features such as location data, property age, and neighborhood quality can be added to improve prediction accuracy. Feature engineering techniques can also be used to create more informative variables.

More advanced models such as Polynomial Regression, Ridge Regression, Lasso Regression, Decision Trees, Random Forest, or Gradient Boosting could be implemented and compared with Linear Regression.

Hyperparameter tuning and cross-validation can also be applied to obtain a more robust and reliable model.

9. References

1. Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow. O’Reilly Media.

2. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

3. Pandas Documentation: https://pandas.pydata.org/

4. NumPy Documentation: https://numpy.org/

5. Housing Dataset Source: Kaggle Housing Price Dataset.