Table of Contents

Week 3 Project: Housing Price Prediction

Introduction

The goal of this project is to predict house prices based on various features such as area, bedrooms, bathrooms, stories, parking, and furnishing status. Since real-world data may contain irrelevant or noisy features, we use Linear Regression as the baseline model and then apply Regularization techniques (Ridge and Lasso Regression) to improve generalization.

Dataset

The dataset contains the following columns:

- price (Target variable)
- area
- bedrooms
- bathrooms
- stories
- mainroad (Yes/No → encoded as binary)
- guestroom (Yes/No → encoded as binary)
- basement (Yes/No → encoded as binary)
- hotwaterheating (Yes/No → encoded as binary)
- airconditioning (Yes/No → encoded as binary)
- parking
- prefarea (Yes/No → encoded as binary)
- furnishingstatus (Furnished, Semi-furnished, Unfurnished → encoded as categories)

Data Cleaning and Preparation

- Handling Missing Values
  - Median values were used for numerical columns like price and parking.
- Encoding Categorical Variables
  - Binary features such as mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea were encoded as 0 (No) and 1 (Yes).
  - furnishingstatus was label-encoded into numeric categories.
- Checking for Duplicates
  - Duplicate rows were checked and removed if necessary.

Train-Test Split

- The dataset was split into:
- Training set (70%) – used to train the model.
- Testing set (30%) – used to evaluate the model.
- A random_state (2) was set to ensure reproducibility.

Model Implemented

Linear Regression (Baseline)

- A simple linear regression model was trained.
- Training accuracy = 66%
- Testing accuracy = 61%
- This indicates a slight underfitting problem.

Ridge Regression (L2 Regularization)

- Adds a penalty proportional to the square of coefficients.
- Helps reduce the impact of irrelevant features.
- Controlled by alpha (λ) → higher alpha means stronger penalty.

Lasso Regression (L1 Regularization)

- Adds a penalty proportional to the absolute value of coefficients.
- Can shrink some coefficients to exactly zero → performs feature selection.
- Controlled by alpha → higher alpha means more coefficients set to zero.

Key Parameter – Alpha

- Alpha controls how much regularization is applied.
- Small alpha → behaves like normal Linear Regression.
- Large alpha → shrinks coefficients more aggressively, reducing overfitting.