Boston Housing Price Prediction using Machine Learning

NAME : DILLI GANESH B

CYBERNAUT PROJECT INTERN

MONTH 2 MINI PROJECT 2 (TEAM 11)

ABSTRACT

This mini project presents a complete machine learning workflow for predicting housing prices using the Boston Housing dataset. The study includes exploratory data analysis, feature correlation analysis, model building, hyperparameter tuning, and performance evaluation. Linear Regression and Ridge Regression models are implemented and compared using evaluation metrics such as R^2 score, RMSE, and MSE. Visualizations are used extensively to interpret feature importance and model performance, enabling clear and explainable conclusions.

INTRODUCTION

Housing price prediction is one of the most widely studied problems in the field of data science and machine learning. Accurate estimation of house prices is crucial for buyers, sellers, investors, and real estate companies, as it directly impacts financial decision-making. Traditional valuation methods are often time-consuming and subjective, whereas machine learning models can automatically learn patterns from historical data and provide consistent predictions.

In this project, the Boston Housing dataset is used to build predictive models that estimate housing prices based on socio-economic, environmental, and geographical factors. Regression-based machine learning algorithms are applied to understand relationships between variables and predict the target value. The project also emphasizes interpretability through visualization and statistical analysis, making the results understandable even for non-technical users

PROBLEM STATEMENT

The real estate market is influenced by multiple interdependent factors such as location, infrastructure, crime rate, pollution levels, and population demographics. Identifying how these factors affect housing prices is challenging using manual analysis. Therefore, the problem addressed in this project is to develop a machine learning-based regression system that can accurately predict house prices using available historical data while also identifying the most influential features.

OBJECTIVES

The primary objectives of this mini project are: - To study and understand the Boston Housing dataset in detail - To perform exploratory data analysis and feature correlation analysis - To preprocess data for effective model training - To build Linear Regression and Ridge Regression models - To evaluate model performance using standard metrics such as R^2, RMSE, and MSE - To compare different models and select the most suitable one - To visualize and interpret results for better understanding

SCOPE OF THE PROJECT

The scope of this project is limited to predictive analysis using regression techniques on the Boston Housing dataset. The system focuses on supervised learning and does not include real-time data integration or deployment. However, the methodology used in this project can be extended to other real estate datasets or enhanced with advanced algorithms. The project serves as a foundation for understanding regression models and model evaluation techniques in data science.

DATASET DESCRIPTION

The Boston Housing dataset contains 506 records with multiple explanatory variables related to housing conditions in Boston suburbs. The target variable represents the median value of owner-occupied homes.

KEY FEATURES

- **CRIM**: Per capita crime rate by town

- **ZN**: Proportion of residential land zoned for large lots
- **INDUS**: Proportion of non-retail business acres
- **NOX**: Nitric oxide concentration
- **RM**: Average number of rooms per dwelling
- **AGE**: Proportion of owner-occupied units built prior to 1940
- **DIS**: Distance to employment centers
- **RAD**: Accessibility to highways
- **TAX**: Property tax rate
- **PTRATIO**: Pupil–teacher ratio
- **B**: Proportion of African American population
- **LSTAT**: Percentage of lower status population
- **MEDV**: Median house value (Target Variable)

TECHNOLOGY STACK

PROGRAMMING LANGUAGE

- Python

LIBRARIES USED

- NumPy – Numerical computations
- Pandas – Data manipulation
- Matplotlib & Seaborn – Data visualization
- Scikit-learn – Machine learning models

SYSTEM ARCHITECTURE

The system architecture of the proposed housing price prediction model follows a structured machine learning pipeline:

1. **Data Collection**: The Boston Housing dataset is loaded from a standard library.

2. **Data Preprocessing**: Missing values are handled, and features are standardized.
3. **Exploratory Data Analysis**: Statistical summaries and correlation analysis are performed.
4. **Model Training**: Linear Regression and Ridge Regression models are trained using training data.
5. **Hyperparameter Tuning**: Ridge Regression alpha values are tuned to find the optimal model.
6. **Model Evaluation**: Performance is measured using R^2, RMSE, and MSE.
7. **Visualization & Interpretation**: Results are visualized for interpretability.

FUNCTIONAL REQUIREMENTS

DATA PROCESSING

- Handle missing values
- Normalize or standardize features
- Split data into training and testing sets

MODEL BUILDING

- Implement Linear Regression
- Implement Ridge Regression
- Train models using training data

MODEL EVALUATION

- Calculate Mean Squared Error (MSE)
- Calculate R-squared score

DATA VISUALIZATION

- Correlation heatmap
- Actual vs Predicted price plots

NON-FUNCTIONAL REQUIREMENTS

USABILITY

The system should present results and visualizations in a clear and interpretable manner. Even users without a strong technical background should be able to understand the outcome of the analysis.

PERFORMANCE

The prediction model should achieve reasonable accuracy with minimal error. Evaluation metrics such as R^2, RMSE, and MSE are used to assess performance.

RELIABILITY

The system should produce consistent results when executed multiple times on the same dataset.

REGRESSION MODELS USED

LINEAR REGRESSION

A basic regression technique that models the linear relationship between independent variables and the dependent variable.

RIDGE REGRESSION

An enhanced regression technique that applies regularization to reduce overfitting and improve prediction accuracy.

EVALUATION METRICS

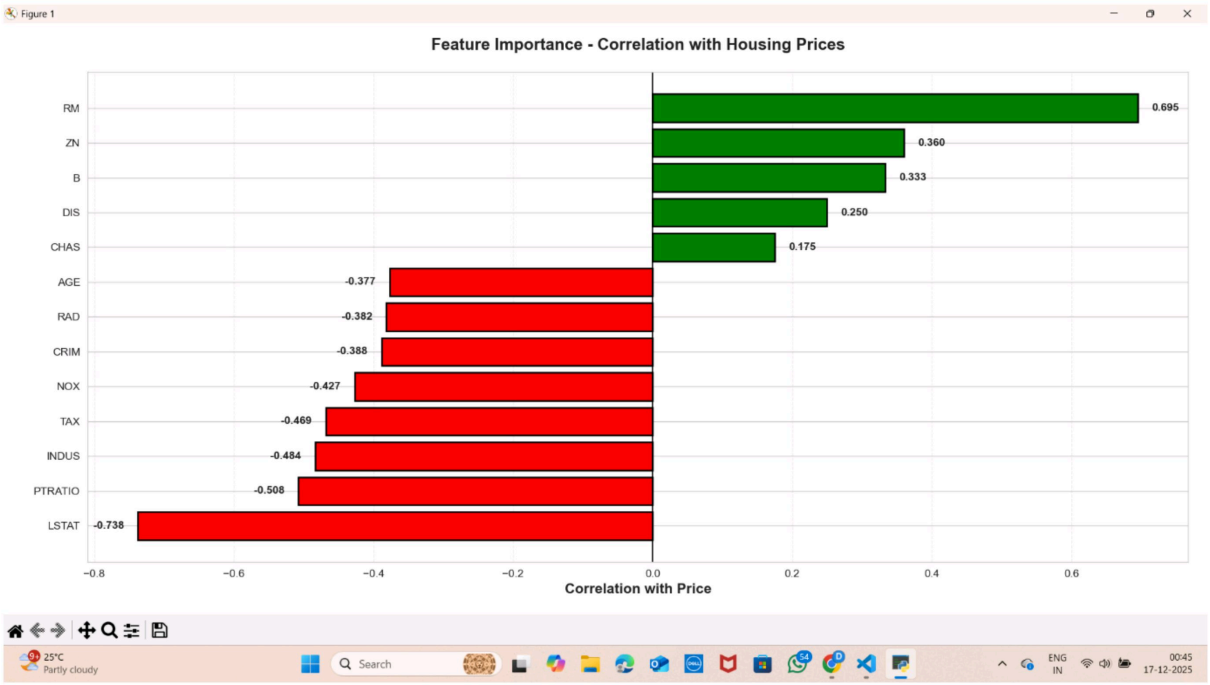
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared Score (R^2)

These metrics are used to evaluate and compare the performance of Linear Regression and Ridge Regression models.

RESULTS AND DISCUSSION

List of Figures

FEATURE IMPORTANCE ANALYSIS

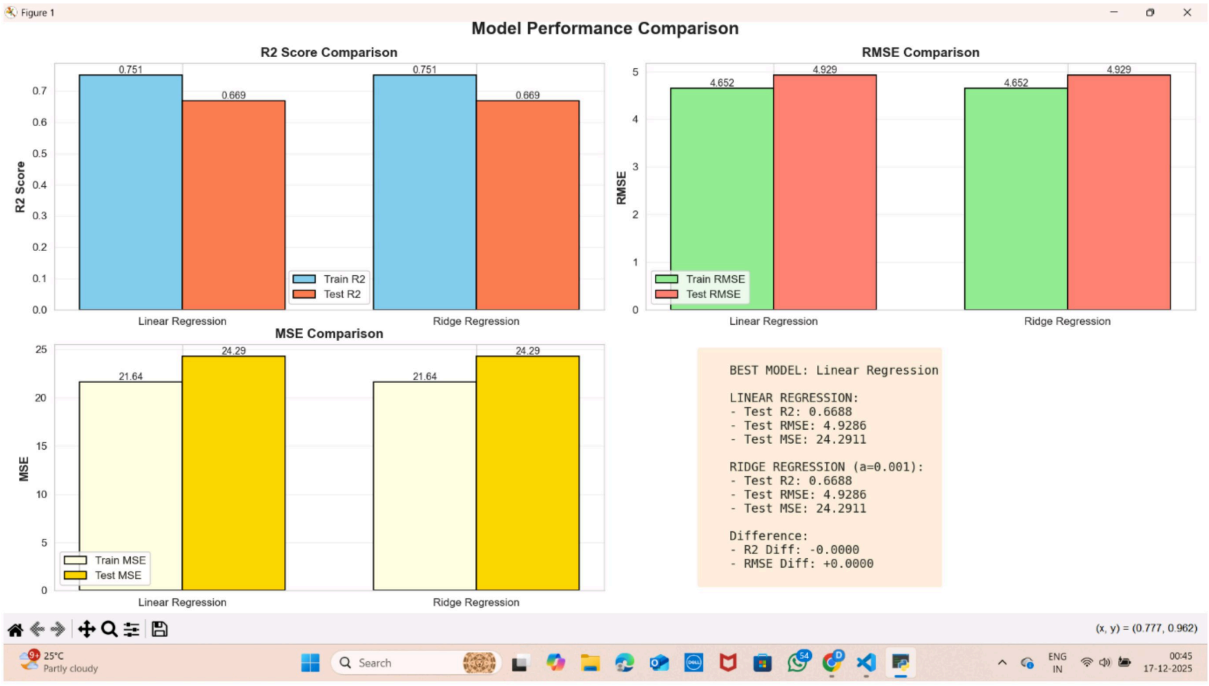


Feature importance analysis plays a crucial role in understanding the relationship between independent variables and housing prices. The correlation plot shows that the feature **RM (average number of rooms per dwelling)** has the highest positive correlation with housing prices, indicating that houses with more rooms tend to have higher values. On the other hand, **LSTAT (percentage of lower-status population)** exhibits a strong negative correlation, suggesting that areas with a higher lower-income population generally have lower house prices.

Other features such as crime rate (CRIM), pollution levels (NOX), and tax rate (TAX) also show negative correlations, indicating their adverse

effect on property value. This analysis helps in identifying key features that significantly influence housing prices and supports informed model building.

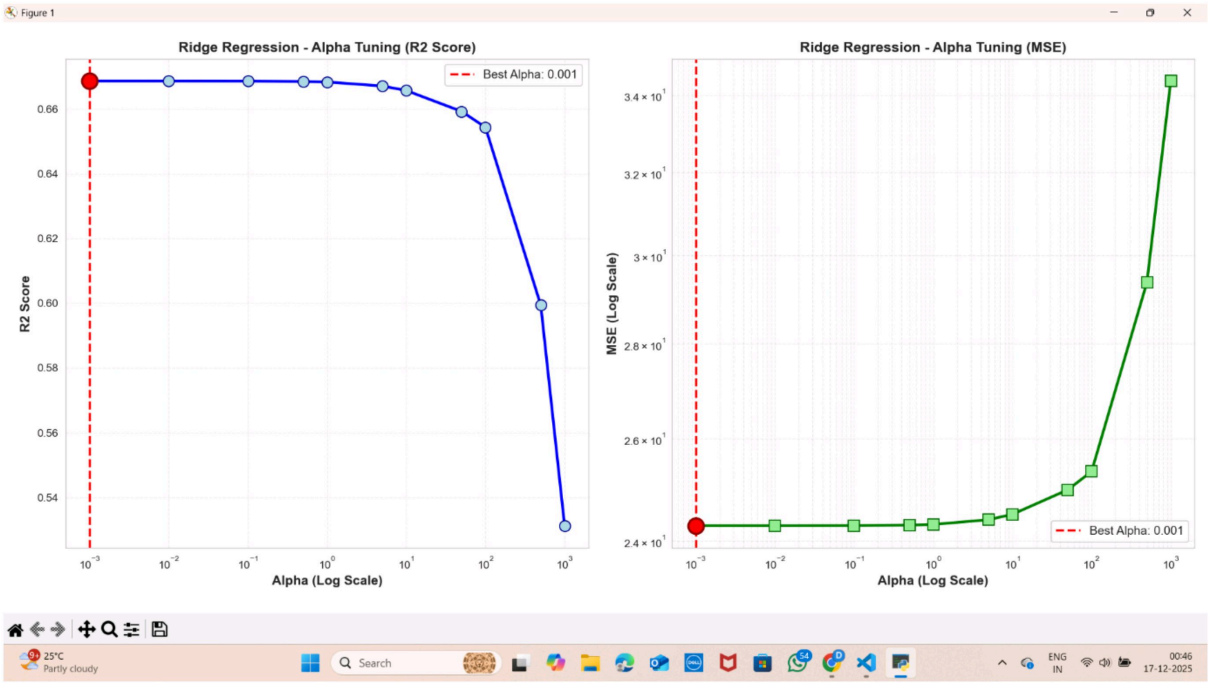
MODEL PERFORMANCE COMPARISON



The model performance comparison evaluates Linear Regression and Ridge Regression using multiple evaluation metrics. The R^2 score indicates how well the model explains the variance in the target variable, while RMSE and MSE measure prediction error. Both models achieve similar performance on the test dataset with an R^2 score of approximately 0.6688 and an RMSE of about 4.93.

Although Ridge Regression includes regularization to prevent overfitting, it does not significantly outperform Linear Regression for this dataset. Therefore, Linear Regression is selected as the best-performing model due to its simplicity and comparable accuracy.

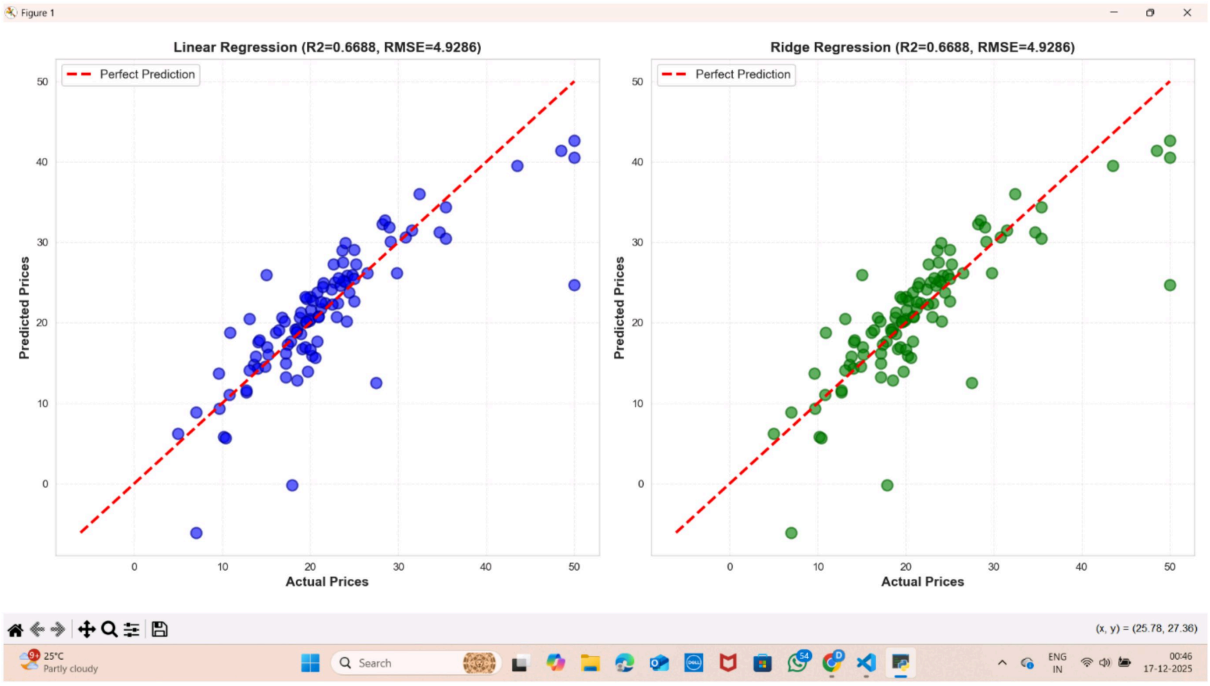
RIDGE REGRESSION HYPERPARAMETER TUNING



Hyperparameter tuning is an essential step in improving model performance. In Ridge Regression, the alpha parameter controls the amount of regularization applied to the model. The tuning graph shows that the best performance is achieved at an alpha value of **0.001**, where the R^2 score is maximized and the MSE is minimized.

As the alpha value increases beyond this point, the model becomes overly regularized, leading to underfitting and reduced prediction accuracy. This analysis highlights the importance of selecting an optimal regularization parameter.

ACTUAL VS PREDICTED PRICE COMPARISON



The scatter plots comparing actual and predicted housing prices provide a visual assessment of model accuracy. Points lying close to the diagonal line represent accurate predictions. Both Linear Regression and Ridge Regression models show similar distributions of points around the ideal prediction line.

Some deviations are observed for extreme values, which is expected in real-world datasets. Overall, the plots confirm that both models perform consistently and align with the quantitative evaluation metrics

CONCLUSION

This mini project successfully demonstrates the application of machine learning techniques for housing price prediction. Through extensive exploratory data analysis, feature correlation study, model training, and evaluation, valuable insights into the factors affecting housing prices were obtained. The results show that features such as number of rooms and socio-economic status play a significant role in determining house prices.

Both Linear Regression and Ridge Regression models were implemented and evaluated. Although Ridge Regression includes regularization, Linear Regression achieved equivalent performance with lower complexity, making it the preferred model. This project highlights

the importance of data preprocessing, feature analysis, and proper evaluation in building reliable predictive models.

FUTURE ENHANCEMENTS

- Use advanced models like Lasso or Random Forest
- Deploy the model using a web interface
- Apply the approach to real-time datasets

REFERENCES

- Scikit-learn Documentation
- UCI Machine Learning Repository
- Python Official Documentation