House Price Prediction Project Report

Priyanshu bhatia

March 10, 2025

1 Introduction

This report outlines the steps taken in developing a machine learning model for predicting house prices, from data preprocessing to model deployment.

2 Data Preprocessing and Feature Engineering

- Loaded the California Housing Dataset from Scikit-learn.
- Performed exploratory data analysis to understand the dataset.
- Handled missing values (none were present in this dataset).
- Engineered new features:
  - Rooms per household: AveRooms/AveOccup
  - Bedrooms per room: AveBedrms/AveRooms
  - Population per household: Population/AveOccup
- Scaled all numerical features using StandardScaler.
- Visualized correlations between features and the target variable.

3 Model Selection and Optimization

- Trained and evaluated four regression models: Linear Regression, Decision Tree, Random Forest, and XGBoost.
- Evaluated models using RMSE, MAE, and R^2 scores.
- Optimized the Random Forest model using a simplified random search approach, tuning hyperparameters such as number of trees, max depth, min samples split, and number of features.

4 Deployment Strategy and API Usage Guide

- Deployed the optimized Random Forest model using a Flask application.
- Created a /predict endpoint that accepts JSON input and returns the predicted house price.
- Provided instructions for testing the API using CURL and Postman.
- Optionally containerized the application using Docker for easy deployment.

API Usage Guide:

```
curl -X POST \
  http://localhost:5000/predict \
  -H 'Content-Type: application/json' \
  -d '{"features": [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1]}'
```

5 Conclusion

The project successfully implemented a machine learning pipeline for house price prediction, from data preprocessing to model deployment. The optimized Random Forest model showed promising results and was effectively deployed as a REST API.