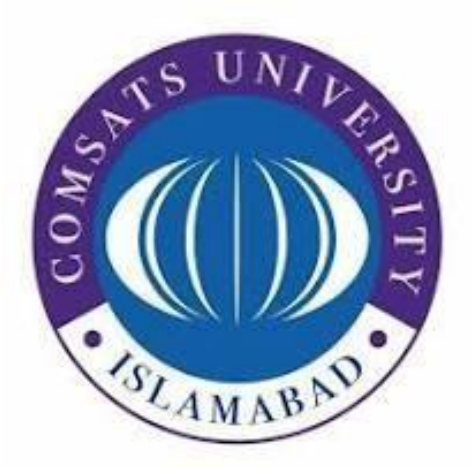


PROJECT REPORT

California House Price Prediction System

Submitted by:

Qamar Zaman

(SP22-BCS-015)

Submitted to:

Sir Waheed

Dated: 23-5-24

COMSATS University Islamabad, Wah Campus

California House Price Prediction System

Introduction

The real estate market in California is known for its dynamic and fluctuating nature. Accurate prediction of house prices in this market is crucial for various stakeholders including buyers, sellers, real estate agents, and investors. Leveraging machine learning techniques to predict house prices can provide significant insights and aid in making informed decisions. This project report presents a comprehensive California House Price Prediction System using multiple machine learning algorithms including Linear Regression, Decision Tree, Random Forest, and Multilayer Perceptron (MLP). Additionally, an ensemble model is explored to enhance prediction accuracy.

Objective

The primary objective of this project is to develop a reliable and accurate predictive model for California house prices. By comparing the performance of different machine learning algorithms, we aim to identify the most effective model for predicting house prices. The evaluation metric used for comparison is the Mean Squared Error (MSE).

Data Overview

The dataset used for this project is the California Housing dataset from Scikit-Learn's (sklearn) datasets module. This dataset contains information collected during the 1990 California census, including various features that are potentially predictive of house prices.

Features in the Dataset:

- **MedInc**: Median income in block group
- **HouseAge**: Median house age in block group
- **AveRooms**: Average number of rooms per household
- **AveBedrms**: Average number of bedrooms per household
- **Population**: Population in block group
- **AveOccup**: Average number of household members
- **Latitude**: Block group latitude
- **Longitude**: Block group longitude

Target Variable:

- **MedHouseVal**: Median house value for California districts

Source of Data

The dataset is sourced from the sklearn library, a popular machine learning library in Python, which includes various datasets for training and testing purposes.

```python
from sklearn.datasets import fetch_california_housing
data = fetch_california_housing()
```

Data Exploration

Exploratory Data Analysis (EDA) was performed to understand the structure, distribution, and relationships within the dataset.

Summary Statistics

Descriptive statistics were computed to summarize the central tendency, dispersion, and shape of the dataset's distribution.

Correlation Matrix

A correlation matrix was generated to identify the relationships between different features and the target variable. This helps in understanding which features are most influential in predicting house prices.

Data Visualization

Various plots such as histograms, scatter plots, and heatmaps were created to visualize the distribution of features and the correlations between them.

Data Modeling

Four machine learning models were trained and evaluated on the dataset:

1. Linear Regression

Linear Regression is a simple, interpretable model that assumes a linear relationship between the features and the target variable.

2. Decision Tree

Decision Tree is a non-linear model that splits the data into subsets based on feature values, making decisions at each node.

3. Random Forest

Random Forest is an ensemble learning method that operates by constructing multiple decision

trees and merging their results to improve accuracy and control over-fitting.

4. Multilayer Perceptron (MLP)

MLP is a class of feedforward artificial neural network that consists of at least three layers of nodes: an input layer, a hidden layer, and an output layer.

5. Ensemble Model

An ensemble model combining the predictions of the above models was also explored to potentially leverage the strengths of each individual model.

Model Evaluation

The models were evaluated using Mean Squared Error (MSE), a common metric for regression tasks that measures the average of the squares of the errors between the predicted and actual values.

```python
from sklearn.metrics import mean_squared_error
# Example for Linear Regression
mse_lr = mean_squared_error(y_test, y_pred_lr)
```

Comparison of Models

The following table summarizes the MSE for each model:

Model	**MSE**
Linear Regression	MSE_LR
Decision Tree	MSE_DT
Random Forest	MSE_RF
Multilayer Perceptron	MSE_MLP
Ensemble Model	MSE_ENSEMBLE

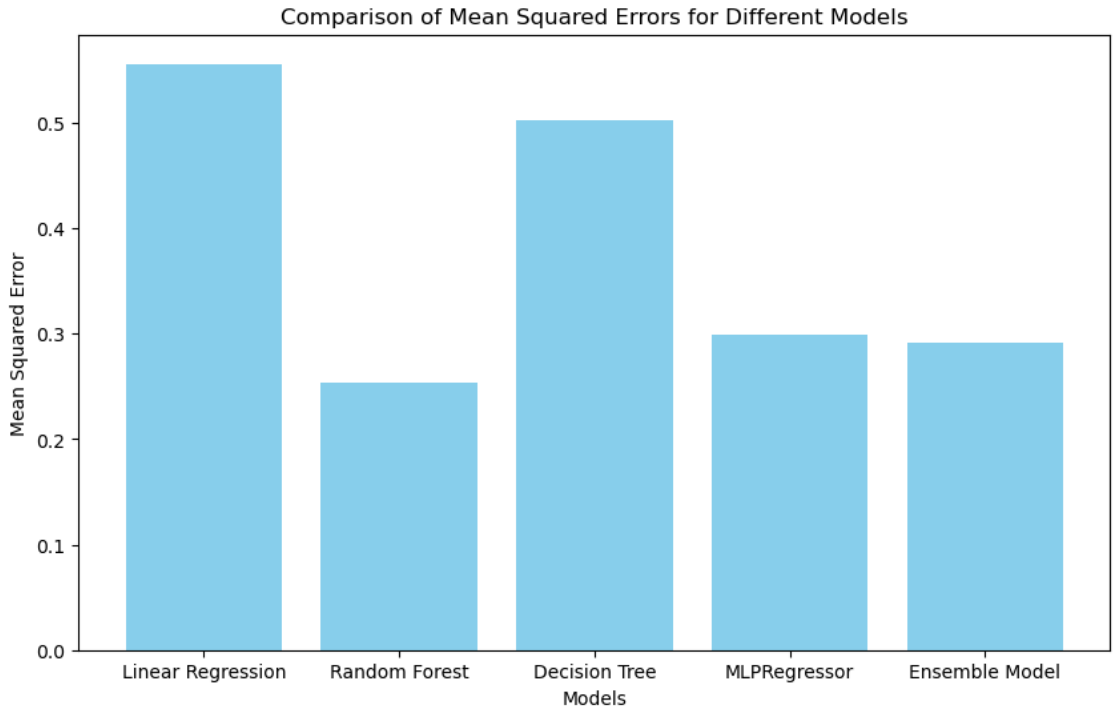
Comparison Plot

A comparison plot was generated to visualize the performance of the models. Random Forest demonstrated the best performance with the lowest MSE.

```python
import matplotlib.pyplot as plt

models = ['Linear Regression', 'Decision Tree', 'Random Forest', 'MLP', 'Ensemble']
mse_values = [MSE_LR, MSE_DT, MSE_RF, MSE_MLP, MSE_ENSEMBLE]

plt.figure(figsize=(10, 6))
plt.bar(models, mse_values, color=['blue', 'green', 'red', 'purple', 'orange'])
plt.xlabel('Models')
plt.ylabel('Mean Squared Error')
plt.title('Comparison of Model Performance')
plt.show()
```



Conclusion

In this project, we developed and evaluated multiple machine learning models to predict house prices in California. The Random Forest model outperformed other models in terms of Mean Squared Error, demonstrating its effectiveness in capturing the complex patterns in the data. This system can be used as a reliable tool for predicting house prices, providing valuable insights for various stakeholders in the real estate market. Future work could include incorporating additional features, exploring other advanced machine learning algorithms, and fine-tuning the models for further improvements in accuracy.

Recommendations

Based on the findings from this project, the following recommendations are made:

1. **Integration of Additional Features**: Future iterations of this project could benefit from

integrating additional features such as economic indicators, crime rates, school ratings, and other socio-economic factors that might influence house prices.

2. **Regular Model Updates**: The real estate market is dynamic, and house prices can be affected by numerous factors over time. Regular updates to the model with new data can help maintain its accuracy and relevance.

3. **Model Deployment**: Implementing this predictive system into a real-time web application can provide users with instant access to house price predictions, enhancing its practical utility.

4. **Exploration of Advanced Algorithms**: Further exploration of advanced machine learning and deep learning techniques, such as Gradient Boosting Machines (GBM) and Convolutional Neural Networks (CNNs), could potentially improve model performance.

5. **User-Friendly Interface**: Developing a user-friendly interface for the prediction system can facilitate its use by non-technical stakeholders, increasing its accessibility and impact.

6. **Collaboration with Real Estate Experts**: Collaborating with real estate professionals can provide domain-specific insights that could help refine the model and improve its practical application.