Predictive Analytics Term Project: Housing Price Prediction

Course: ISOM 835 — Predictive Analytics & Machine Learning
Instructor: Hasan Arslan
Student: Saumya Patel
Institution: Suffolk University
Date: December 2024

Executive Summary

This project applies predictive analytics and machine learning techniques to a housing dataset to predict median house values and identify key factors influencing housing prices. The analysis follows a complete data science workflow, including exploratory data analysis (EDA), data cleaning and preprocessing, feature engineering, model development, and model evaluation.

Three predictive models were implemented: Linear Regression, K-Nearest Neighbors (KNN), and Random Forest. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score. Among the models tested, Random Forest achieved the strongest predictive performance, demonstrating the lowest error and highest explanatory power.

The results highlight the importance of income-related and location-based features in determining housing prices. This project demonstrates how predictive analytics can be used to support data-driven decision-making in real estate and urban planning contexts.

1. Introduction

Housing prices are influenced by a complex interaction of economic, demographic, and geographic factors. Accurate prediction of housing values is essential for policymakers, real estate developers, financial institutions, and individual buyers. Predictive analytics enables the identification of key drivers of housing prices and supports informed decision-making.

The objective of this project is to develop predictive models that estimate median house values using demographic and housing-related attributes. The project also aims to compare different modeling techniques and evaluate their performance.

Business Questions

1. Which housing and demographic factors most strongly influence median house value?
2. How accurately can machine learning models predict housing prices?
3. Which predictive model provides the best overall performance?

2. Dataset Description

The dataset used in this project is the **California Housing Dataset**, which contains housing-related information collected from California census data.

- **Number of observations:** 20,640
- **Number of features:** 10 (before encoding)
- **Target variable:** median_house_value

Key Variables

- median_income: Median income in the block group
- total_rooms: Total number of rooms
- total_bedrooms: Total number of bedrooms
- population: Population in the block group
- households: Number of households
- ocean_proximity: Categorical variable describing proximity to the ocean

This dataset is well-suited for regression-based predictive modeling and exploratory analysis.

3. Exploratory Data Analysis (EDA)

Exploratory data analysis was conducted to understand the structure, distribution, and relationships within the data.

Key EDA Findings

- The dataset contained missing values in the total_bedrooms feature.
- Numerical variables showed varying distributions, with some exhibiting right skewness.
- Correlation analysis revealed a strong positive relationship between median_income and median_house_value.
- Location-related features also showed significant influence on housing prices.

Visualizations such as histograms, correlation heatmaps, and boxplots were used to identify trends, relationships, and potential outliers.

4. Data Cleaning and Preprocessing

Several preprocessing steps were performed to prepare the data for modeling:

Missing Values

- Missing values in total_bedrooms were imputed using the median to reduce the influence of outliers.

Categorical Encoding

- The categorical variable ocean_proximity was converted into numerical format using one-hot encoding.

Feature Scaling

- Standardization was applied to numerical features to improve model performance, particularly for distance-based models such as KNN.

Train-Test Split

- The dataset was split into training (80%) and testing (20%) sets to evaluate model performance on unseen data.

5. Feature Engineering

To enhance predictive performance, additional features were engineered:

- **Rooms per household**
- **Bedrooms per room**
- **Population per household**

These engineered features helped capture household-level density and utilization patterns, improving the explanatory power of the models.

6. Model Building

Three regression models were implemented:

1. Linear Regression

Used as a baseline model to establish a reference level of performance.

2. K-Nearest Neighbors (KNN)

A non-parametric model that predicts values based on similarity to neighboring observations.

3. Random Forest

An ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting.

7. Model Evaluation and Comparison

Model performance was evaluated using:

- **Mean Absolute Error (MAE)**
- **Root Mean Squared Error (RMSE)**
- **R² Score**

Model Comparison Summary

Model	**MAE**	**RMSE**	**R^2**
Linear Regression	50888.660016	72668.538379	0.597018
KNN Regression	41745.599855	62106.747677	0.705645
Random Forest	32250.569247	50268.150817	0.807168

Random Forest outperformed the other models, achieving the lowest error and highest explanatory power.

8. Interpretation of Results

Feature importance analysis from the Random Forest model indicated that **median income** was the most influential predictor of housing prices. Location-based variables related to ocean proximity also played a significant role.

These findings align with economic intuition, as income levels and geographic desirability are key determinants of housing value.

9. Limitations and Ethical Considerations

Limitations

- The dataset represents housing data from a specific geographic region and may not generalize to other markets.
- The models do not account for temporal changes in housing markets.

Ethical Considerations

- Predictive models may reinforce existing socioeconomic biases if used without caution.
- Responsible use of housing prediction models requires transparency and fairness considerations.

10. Conclusion

This project demonstrated the application of predictive analytics to housing price prediction. Through data cleaning, feature engineering, and model comparison, Random Forest emerged as the most effective model.

The project highlights the value of machine learning in understanding complex real-world problems and provides a foundation for further improvements, such as hyperparameter tuning and incorporation of additional data sources.