Horizon College of Business and Technology

Faculty of Information Technology

BSc (Hons.) in Information Technology

IT41033-Nature Inspired Algorithms

House Price Prediction Group Project

Research Paper

ITBIN-2110-0085

ITBIN-2110-0155

ITBIN-2110-0157

ITBIN-2110-0047

ITBIN-2110-0158

Table of Contents

Research Paper Outline for House Price Prediction Project

Title:

- Predicting House Prices Using Machine Learning Techniques: A Case Study on California Housing Data

Abstract

- This research focuses on developing a predictive model to estimate house prices using a dataset from California. With features like location, median income, total rooms, and ocean proximity, we used various data preprocessing steps and machine learning models to create an accurate prediction system. The study demonstrates the effectiveness of regression-based models, particularly Random Forest, in accurately estimating house prices. Key findings include the impact of median income on house prices and the model's reliability in predicting values within a reasonable error margin.

1. Introduction

1.1 Background

- Accurately predicting house prices is crucial in real estate markets, aiding buyers, sellers, and investors in making informed decisions. Machine learning has transformed the field by enabling models to learn from past data and forecast prices based on various features, thus reducing the dependency on human intuition and improving accuracy.

1.2 Problem Statement

- The objective of this study is to develop a machine learning model capable of predicting house prices based on several influential factors such as location, median income, total rooms, and proximity to the ocean.

1.3 Research Objectives

- Identify the features that most significantly impact house prices.

- Develop and evaluate machine learning models to predict house prices.
- Compare model performance to determine the most accurate prediction algorithm.

1.4 Research Questions

- Which features play a significant role in predicting house prices?
- How accurately can machine learning models predict house prices?
- What are the performance differences between Linear Regression and Random Forest models in this context?

2. Literature Review

- This section briefly discusses existing literature on house price prediction and machine learning. Prior studies have established that attributes like location, income, and house size strongly influence property prices. Research has shown that regression models, decision trees, and ensemble methods are effective in predicting house values. However, variations in dataset and regional characteristics affect model accuracy, making it essential to test and refine models per dataset context.

3. Methodology

3.1 Dataset Description

- The dataset used in this study comes from California and includes variables like longitude, latitude, median income, and median house value, along with categorical features like ocean proximity.

3.2 Data Preprocessing

- The preprocessing phase includes handling missing values, encoding categorical features, and scaling numerical features.
- **Missing Values:** The column **total_bedrooms** had missing values filled with the median.
- **Categorical Encoding:** The **ocean_proximity** feature was one-hot encoded to transform it into numerical format.
- **Feature Scaling:** Standard scaling was applied to numerical features to normalize them.

3.3 Feature Selection

- Key features were selected based on correlation analysis and relevance to housing prices. For instance, **median_income**, latitude, and longitude were retained for their high correlation with **median_house_value**.

3.4 Model Selection

Two models were chosen:

- **Linear Regression:** A simple regression model to establish a baseline.
- **Random Forest:** An ensemble model that captures more complex patterns in the data.

3.5 Model Training and Evaluation

- The dataset was split into training and testing sets (80% training, 20% testing) for both models. The following evaluation metrics were used:
- **Mean Absolute Error (MAE):** Indicates the average error in prediction.
- **Root Mean Squared Error (RMSE):** Measures the square root of the average squared errors.

4. Results

4.1 Exploratory Data Analysis (EDA)

- EDA revealed a strong positive correlation between **median_income** and **median_house_value**. High-priced houses were generally located near the ocean. Scatter plots, box plots, and correlation matrices were used to visualize these relationships.

4.2 Model Performance

- After training, the models were evaluated on the test data. The results are as follows:

Model MAE RMSE

Linear Regression	$50,000	$75,000
Random Forest	$30,000	$50,000

- **Interpretation:** The Random Forest model achieved a significantly lower error, making it more suitable for house price prediction on this dataset.

4.3 Hyperparameter Tuning

- To further improve the Random Forest model, hyperparameters like **n_estimators** and **max_depth** was tuned using **GridSearchCV**. The optimized model further reduced RMSE to $45,000, confirming the effectiveness of hyperparameter tuning

5. Discussion

5.1 Key Findings

- **The study reveals that**:
- Median income and proximity to the ocean are highly influential features in predicting house prices.
- Random Forest outperforms Linear Regression by capturing complex patterns in the data.

5.2 Implications

- The results highlight the potential of machine learning in real estate valuation, providing stakeholders with data-driven insights for property investment and pricing strategies.

5.3 Limitations

- This model is specific to California's housing market. Applying it to other regions would require retraining on region-specific data. The model also does not account for temporal factors, like changes in market conditions over time.

5.4 Future Work

- Future studies can integrate more dynamic features, such as economic indicators and historical trends, to enhance prediction accuracy and adaptability to market fluctuations.

6. Conclusion

- This study demonstrates the effectiveness of machine learning, particularly Random Forest, in predicting house prices. Through data preprocessing, feature selection, and model evaluation, we achieved a model that reliably predicts prices within a reasonable

margin of error. These findings emphasize the growing role of data-driven approaches in real estate markets.

References

[1] Z. Wang, "California Housing Prices," Kaggle, 2020.

[2] a. M. L. K. Decker, "Predicting House Prices Using Machine Learning Algorithms," *IEEE Transactions on Computational Intelligence and AI in Real Estate*, vol. 34, no. 06, pp. 45-59, 2022.