Title Page

Project Title: Predicting Housing Prices in Melbourne Using Machine Learning

Student Information:

Name = Azahar Mahaboob

Shu_Id = 0929896

Email = shaika29@mail.sacredheart.edu

Executive Summary

TThe goal of this project is to predict housing prices in Melbourne by analyzing various factors such as location, house features, and neighborhood characteristics using machine learning models. The dataset for this project is sourced from Melbourne real estate data and contains variables like the number of rooms, lot size, property type, and location coordinates. Using machine learning algorithms such as Decision Trees and Random Forests, we will build predictive models capable of accurately forecasting housing prices. This project aims to provide insights into the factors most significantly affecting real estate prices and deliver a reliable model for predicting prices in Melbourne's competitive housing market.

Introduction

Background:

Housing prices are highly influenced by various factors, including location, property characteristics, and the broader economic environment. Accurate predictions of these prices are essential for real estate professionals, buyers, and investors to make informed decisions. With the increasing availability of housing data, machine learning offers the potential to analyze this data more effectively, enabling us to develop predictive models with high accuracy.

This project focuses on predicting housing prices in Melbourne, one of Australia's largest and most competitive real estate markets. The dataset includes key variables such as suburb, property size, number of rooms, and geographic coordinates, providing a comprehensive view of the factors influencing housing prices.

Problem Statement:

Traditional methods for predicting housing prices often rely on limited variables like location and property size. However, modern machine learning techniques can take into account a wider range of factors, enabling more accurate predictions. This project seeks to build a machine learning model that predicts housing prices based on multiple features extracted from Melbourne's housing data.

Objectives:

To develop a machine learning model that accurately predicts housing prices in Melbourne using a range of property and location-based features.

To identify the most important features affecting house prices and how they vary across different regions.

Significance:

Accurate predictions of housing prices are valuable for various stakeholders, including homebuyers, sellers, real estate agents, and policymakers. By leveraging machine learning, this project aims to contribute to the growing field of real estate analytics, offering a data-driven approach to understanding property values and trends in Melbourne.

Literature Review

Relevant Studies:

Past research has explored the application of machine learning models such as Linear Regression, Decision Trees, and Gradient Boosting for predicting housing prices. These models have demonstrated varying degrees of accuracy, depending on the data used and the specific characteristics of the housing markets in question. For example, studies based on the Ames Housing Dataset and Kaggle competitions have shown the effectiveness of machine learning techniques in improving predictive accuracy.

Gap Identification:

While several studies have been conducted on predicting housing prices, most focus on datasets from specific regions like the United States. There is a gap in the literature regarding the application of these models to localized real estate markets such as Melbourne. This project aims to address this gap by developing a machine learning model tailored to Melbourne's unique housing market.

Project Methodology

Approach and Framework:

The project will follow a structured approach, starting with data preprocessing, which includes handling missing data, encoding categorical variables, and scaling numerical features. Feature selection will be performed to identify the most relevant variables for predicting housing prices. The final model will be built using machine learning algorithms such as Decision Trees and Random Forests.

Research Methods:

Data Preprocessing: This will involve handling missing values (e.g., imputing missing values for BuildingArea and YearBuilt), encoding categorical variables (such as Type and Regionname), and normalizing numerical features like Landsize and Distance.

Feature Selection: Key features such as the number of rooms, property size, location (suburb, region), and proximity to amenities will be evaluated for their influence on house prices.

Model Development: Decision Tree and Random Forest algorithms will be employed due to their ability to handle complex relationships between features and outcomes. These models will be trained and validated using cross-validation techniques to ensure their robustness.

Model Evaluation: The models will be evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2). These metrics will help assess the models' accuracy in predicting housing prices.

Data Description:

The dataset contains 21 variables;

1. Suburb (object): The suburb where the property is located.

2. Address (object): The street address of the property.

3. Rooms (int64): The number of rooms (bedrooms and living areas) in the property.

4. Type (object): The property type (e.g., house, unit, townhouse).

5. Price (int64): The sale price of the property in Australian dollars (target variable).

6. Method (object): The sale method (e.g., auction, sold prior, vendor bid).

7. SellerG (object): The real estate agent or agency responsible for the sale.

8. Date (object): The sale date (format: day-month-year).

9. Distance (float64): The distance from the property to Melbourne's Central Business District (CBD) in kilometers.

10. Postcode (int64): The postal code of the property's location.

11. Bedroom2 (int64): The number of bedrooms.

12. Bathroom (int64): The number of bathrooms.

13. Car (float64): The number of car parking spaces available.

14. Landsize (int64): The land area of the property in square meters.

15. BuildingArea (float64): The size of the building in square meters.

16. YearBuilt (float64): The year the property was built.

17. CouncilArea (object): The local government area in which the property is located.

18. Lattitude (float64): The geographical latitude of the property.

19. Longtitude (float64): The geographical longitude of the property.

20. Regionname (object): The broader region in Melbourne where the property is located.

21. Propertycount (int64): The total number of properties in the suburb.

Deliverables

- A trained and validated machine learning model capable of accurately predicting housing prices in Melbourne.
- A feature importance analysis highlighting the most influential factors affecting housing prices.
- Visualizations comparing predicted versus actual prices, as well as feature importance rankings.

Team

Azahar Mahaboob

Conclusion

This project aims to develop a machine learning model to predict housing prices in Melbourne using a diverse range of property features. By analyzing key factors such as property size, location, and proximity to amenities, the project will offer valuable insights into what drives housing prices in Melbourne. The model developed will provide a data-driven tool to assist real estate professionals, buyers, and policymakers in making more informed decisions.

References

# (Kaggle) - Melbourne Housing Snapshot. **https://www.kaggle.com/datasets/dansbecker/melbourne-housing-snapshot/data**