Group 40 Proposal

Peter Lehmann |Kasidit Muenprasitivej | Haoyuan Chen | Nai-jen Cheng | Yilun Zhou

In the real estate industry, where numerous factors influence property prices and rental rates, accurate forecasting is paramount for investors, developers, buyers, appraisers, and other stakeholders. Machine learning (ML) techniques have emerged as powerful tools to tackle this challenge. A review of the literature reveals a variety of ML algorithms utilized for housing price prediction, demonstrating their efficacy in providing insightful forecasts. For instance, Zhishang Huang's approach [3] utilizes K-Means clustering to segment housing data into distinct categories - economical, comfortable, and high-end houses - effectively differentiating between housing types based on key characteristics. Similarly, McCluskey's work [4] adapts regression models to fit Malaysian resident data, demonstrating reasonable fits for different features and corresponding rent prices. Decision tree learning, as exemplified by Priya [5], offers simplicity and efficacy in predicting housing prices, while Random Forest algorithms, as explored by Adetunji and Xue [1][6], excel in handling outliers and achieving high prediction accuracy.

Our motivation behind addressing the housing price prediction problem stems from the potential positive impacts on individuals, businesses, and the broader economy. By the outcome of this project, homebuyers would have a reliable prediction of property prices which aids them in making informed decisions in purchasing real estate. House sellers would be armed with accurate predictions and able to offer valuable advice to their clients, establishing trust and expertise. By leveraging supervised and unsupervised algorithms and evaluation metrics learned in class, this project aims not only to predict sale prices but also to contribute to the broader understanding of what factors drive property values.

The selected dataset, Ames Housing Dataset, provides a rich source of information with 2,930 instances and 79 features. These features encompass a wide range of variables, including numerical and categorical data, such as sale price, building class, construction date, house style, and neighborhood. This dataset serves as a valuable resource for developing and testing machine learning algorithms and techniques in the real estate domain. Hence, our project will focus on devising an ML algorithm to predict housing prices specifically in Ames, Iowa, based on property descriptions and locations within the city area analogous to the dataset [7].

In pursuit of these goals, the project will employ various data preprocessing methods, including handling missing data, outlier rejection, and dimensionality reduction, to prepare the dataset for analysis. Identified ML algorithms such as K-Means clustering, Regression Learning, Support Vector Machine (SVM), and Random Forest will be implemented, each selected for its suitability in addressing the housing price prediction problem. Among these models, we plan to implement techniques such as K-Means clustering combined with regression models, SVM, and

Random Forest algorithms as demonstrated by the literature reviews for their ability to handle diverse datasets and predict housing prices with notable accuracy. Quantitative metrics such as RMSE, R-Squared, and Mutual Information Score will be utilized for evaluating the accuracy housing price prediction. Through rigorous analysis and practical application, the project seeks to contribute to a deeper understanding of real estate dynamics and facilitate more equitable and informed transactions in the housing market.

REFERENCE:

1. A. B. Adetunji *et al.*, “House price prediction using Random Forest Machine Learning Technique,” *Procedia Computer Science*, vol. 199, pp. 806–813, 2022. doi:10.1016/j.procs.2022.01.100
2. Corsini, Kenneth Richard. "Statistical analysis of residential housing prices in an up and down real estate market a general framework and study of Cobb County, GA ." Thesis, Atlanta, GA: Georgia Institute of Technology, 2009.
3. G.G. Priya, “House price prediction using Machine Learning Techniques,” *International Journal for Research in Applied Science and Engineering Technology,* vol. 9, no. VI, pp. 3645–3650, Jun. 2021. doi:10.22214/ijraset.2021.35831
4. Huang, Zhishang & Lai, Guanren. (2023). A House Price Prediction Model Based on K-means Clustering and Random Forest in Guangzhou. Frontiers in Business, Economics and Management. 10. 377-381. 10.54097/fbem.v10i2.11077.
5. McCluskey, W. J., Daud, D. Z., & Kamarudin, N. (2014). Boosted regression trees: An application for the mass appraisal of residential property in Malaysia. Journal of Financial Management of Property and Construction.
6. G.G. Priya, “House price prediction using Machine Learning Techniques,” *International Journal for Research in Applied Science and Engineering Technology,* vol. 9, no. VI, pp. 3645–3650, Jun. 2021. doi:10.22214/ijraset.2021.35831
7. Thapa, S. (2023, June 20). Ames Housing Dataset. Kaggle. https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset
8. Xue, C., Ju, Y. F., Li, S. G., Zhou, Q. L., & Liu, Q. Q. (2020). Research on Accurate House Price Analysis by Using GIS Technology and Transport Accessibility: A Case Study of Xi'an,China. Symmetry-Basel, 12(8), Article 1329

Contribution Chart:

Name	Contribution
Peter Lehmann	• Literature Review • Presentation Slides • Motivation • Results & Expected Outcome
Kasidit Muenprasitivej	• Literature Review • Presentation Slides • Video Voiceover • Problem Definition • Results & Expected Outcome
Nai-jen Cheng	• Literature Review • Presentation Slides • Dataset Description • Evaluation Metrics • Results & Expected Outcome
Yilun Zhou	• Literature Review • Presentation Slides • Github Guy • Identify ML Packages • Results & Expected Outcome
Haoyuan Chen	• Literature Review • Presentation Slides • Problem Definition • Methods for ML • Data Preprocessing Methods • Results & Expected Outcome