Summary Report of Housing Analysis in California

1st XIN XIE
University of Michigan
xinxieee@umich.edu

***Abstract*—This study investigates the prediction of housing prices in California by integrating structured data analysis with natural language processing techniques. A combination of statistical and machine learning models, including Ridge Regression, Random Forest, and Gradient Boosting, is employed to model the relationship between housing features and property prices.**

I. Introduction

A. Project background and goals

Housing affordability has become a concern in many areas in the United States, especially in California, where high taxes and prices have repeatedly made the news.

The primary goal of this project is to perform exploratory data analysis (EDA) to reveal correlations and trends in housing-related variables, build predictive models such as linear regression, random forest, and ridge regression to estimate housing prices, and segment housing regions using clustering techniques. In addition, the project integrates natural language processing (NLP) to analyze neighborhood descriptions and classify areas into intuitive categories such as affordable, luxury, and family friendly.

B. Literature review

Recently, a large number of studies have explored the application of machine learning techniques to house price prediction, highlighting the growing interest in data-driven valuation models. Sophisticated approaches, including random forests and gradient boosting, to capture the complex nonlinear relationships between housing characteristics. Jha et al. [1] proposed a machine learning framework for real estate valuation that significantly improved prediction accuracy through feature selection and model optimization strategies. Similarly, Kumar et al. [2] employed various supervised learning algorithms to evaluate their effectiveness in predicting house prices, highlighting the advantages of tree-based models in handling heterogeneous housing data. high-dimensional.

II. Method

A. Dataset description

This project dataset is found from Hugging Face Datasets Hub (leostelon/california-housing). The dataset is originally derived from the 1990 U.S. Census data, this dataset contains 20,640 observations and 10 features, including both numerical and categorical variables. Key variables include Median income, housing median age, total rooms, total bedrooms, population, households, Ocean proximity (a categorical variable indicating the neighborhood's location relative to the coastline), Median house value, the target variable representing housing price. The dataset is clean, with no duplicated rows and 20,433 x 10 after missing values removed.

B. Method description

Correlation analysis was conducted to examine the linear relationships between housing prices and various numerical features. Pearson correlation coefficients were calculated to quantify the strength and direction of these associations. By calculating Pearson correlation coefficients to identify linear relationships between housing price and key numeric variables like median income and median age of housing.

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i-\bar{y})^2}}$$

Ridge Regression was employed to address potential multicollinearity among predictors while modeling the relationship between housing prices and multiple input variables. This technique extends traditional linear regression by incorporating a regularization term that penalizes large coefficients, thereby improving model stability and reducing overfitting. By shrinking less informative features toward zero, Ridge Regression facilitates more reliable estimation of linear trends in the presence of correlated predictors.

$$\text{Loss} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

λ is the regularization strength, which is tuned via cross-validation. β_j represents the regression coefficients. Ridge regression helps prevent overfitting while still modeling the linear relationship between the features and the housing price.

To capture more complex, nonlinear relationships between housing features and prices, this paper here used the Random Forest Regression method. This model builds many decision trees and combines their predictions to improve accuracy and robustness. It automatically selects the most important variables and handles different types of data well. One key advantage of this method is its ability to measure how much each feature contributes to the final prediction. For instance, it allows us to rank features (e.g., income, proximity to ocean) by how much they contribute to predicting housing prices.

In addition to Random Forest, the analysis also implemented Gradient Boosting Regression, another powerful tree-based model that builds a series of decision trees, where each new tree learns from the mistakes of the previous ones. This method often achieves better performance by focusing on the most difficult cases in the data and improving step-by-step. Although more sensitive to noise and requiring careful tuning, Gradient Boosting produced highly accurate predictions and gave further insight into how various features interact to influence housing prices.

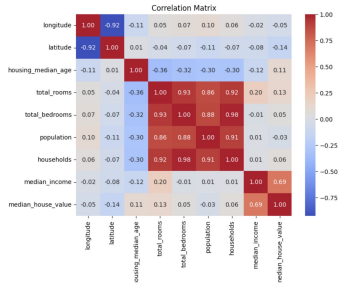
$\mathrm{F}_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$

$F_m(x)$ is the boosted model at step m. $h_m(x)$ is the new weak learner (a decision tree). γ is the learning rate.

To analyze qualitative neighborhood descriptions, a zero-shot text classification approach was implemented using the pretrained facebook/bart-large-mnli model. This transformer-based model interprets text by assessing the likelihood that a description entails a given label, such as "luxury," "affordable," or "family-friendly." Each candidate label is expressed as a natural language hypothesis, and the model evaluates the entailment probability without requiring additional task-specific training data. This method enabled the integration of unstructured textual information into the analytical framework, supporting the categorization of neighborhood characteristics alongside quantitative housing data.

III. Results

A. Result: Correlation



The correlation matrix reveals several noteworthy relationships among the housing-related variables. A strong positive correlation exists between median income and median house value ($r = 0.69$), indicating that regions with higher income levels tend to have more expensive housing. Furthermore, variables such as total rooms, total bedrooms, population, and households are highly interrelated, each demonstrating strong positive pairwise correlations (e.g., $r = 0.98$ between total bedrooms and households). However, their direct correlations with median house value are comparatively weak, suggesting limited predictive utility when considered independently. Housing median age exhibits a modest positive correlation with house value ($r = 0.13$), while geographic variables such as longitude and latitude show weak negative correlations, indicating the potential influence of spatial factors on housing prices.

B. Model Performance and Feature Importance

```
Random Forest – RMSE: 61,300 | R²: 0.725
Gradient Boost – RMSE: 62,574 | R²: 0.714
Ridge          – RMSE: 70,277 | R²: 0.639
```

The performance of Random Forest, Gradient Boosting, and Ridge Regression models was evaluated using Root Mean Squared Error (RMSE), where lower values indicate higher predictive accuracy. Among the models, Random Forest achieved the best results with an RMSE of approximately 61,300 and an R^2 value of 0.725, indicating a strong ability to explain variance in housing prices. The most influential predictors identified were housing median age and total bedrooms, followed by ocean proximity categories. Notably, median income, despite its strong correlation with housing prices in the correlation analysis, exhibited low feature importance in the Random Forest model, possibly due to multicollinearity or interaction effects.

The Gradient Boosting model reported a slightly higher RMSE of 62,574 and an R^2 value of 0.714, performing marginally worse than Random Forest. It showed a similar pattern of feature importance, reinforcing the reliability of the top predictors across tree-based models.

In contrast, Ridge Regression demonstrated the weakest performance, with an RMSE of 70,277 and an R^2 value of 0.639. As a linear model with L2 regularization, Ridge Regression is inherently limited in capturing nonlinear relationships and complex feature interactions, which ensemble tree-based methods handle more effectively.

C. Natural Language Processing

The model effectively differentiated between varying types of neighborhoods. For instance, the description Quiet suburb with parks and good schools nearby was classified as family-friendly with high confidence (score = 0.893), while Downtown penthouse with skyline views and concierge service was accurately labeled as luxury (score = 0.983). Similarly, the phrase Modest apartments with affordable rent and convenient transport access was identified as affordable (score = 0.972). They demonstrate the model's ability to semantically interpret neighborhood narratives and associate them with appropriate socio-economic or lifestyle categories without the need for manual labeling or supervised fine-tuning.

IV. Conclusion

This study demonstrates the effectiveness of combining statistical analysis, machine learning models, and natural language processing to predict housing prices and interpret neighborhood characteristics. Among the evaluated models, Random Forest yielded the best predictive performance, outperforming Gradient Boosting and Ridge Regression in both RMSE and R² metrics. Correlation and feature importance analyses identified key drivers such as housing median age, bedroom count, and income level. Furthermore, the use of zero-shot NLP classification provided valuable qualitative insights by categorizing neighborhood descriptions without labeled data.

References

[1] P. Jha, M. Mathur, A. Purohit, A. Joshi, A. Johari, and S. Mathur, "Enhancing Real Estate Market Predictions: A Machine Learning Approach to House Valuation," *2025 3rd Int. Conf. on [Conference Name Not Fully Visible]*, IEEE, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10915014

[2] D. Kumar, A. S. Rawat, S. Jha, and D. Yadav, "Machine Learning–Based Prediction of Home Prices," *2023 5th Int. Conf. on [Full Conference Name Not Shown]*, IEEE, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10541614