Machine Learning Application in California Real Estate Price Prediction

Ghaidaa Elazhary,Lama Alturki, Albandari Altami

[a]*Department of Computer Science, University of Jeddah, , jeddah, , , saudia arabia*

Abstract

Machine learning has revolutionized predictions of California real estate prices. Unlike any previous methods, machine learning excels in recognizing patterns and trends used to create forecasts much closer to reality. It provides a much better understanding of the dynamics of the housing market for investors, buyers, and lawmakers. Thus, PCA, LR, DT, RF, and HC are used in this work to analyze the complex relationships between socioeconomic characteristics and housing prices in different dynamics of California. These models were able to derive 25 percentage more accurate forecasts than conventional models, which signals how effective machine learning can be in mimicking these complex relationships in the housing market.

Keywords: Machine Learning, California, Prices, United States, Real Estate, Classification Models, Prediction.

1. Introduction

The real estate industry is one of the most vital sectors that constantly improves economic growth and stability in any country, as it affects micro and macroeconomic variables. In California, one of the most developed and intricate real estate markets in the United States, price prediction has become an essential weapon for investors, homeowners,and even a workload for policymakers. Many factors affect the price of real estate, and each has a different degree of correlation with market trends. Such factors include location, property characteristics, socioeconomic factors, and regional economic performance. Many traditional pricing strategies need help assessing how such factors would affect the market, thus becoming less precise. The present study aims to overcome the shortcomings of existing traditional pricing models by using machine learning (ML) techniques to estimate the levels of house rent in California. Some machine learning models, including decision trees, random forests, and principal component analysis (PCA),have been shown to cope with complicated nonlinear patterns and provide flexible and accurate predictions. This research aspires to the theoretical and practical progress of real estate analytics by understanding the factors determining house prices and their socioeconomic effects and enabling a clear illustration. The importance of this study is that it improves the efficacy of machine learning techniques in crafting real estate valuations, which benefits users like investors, urban planners, regulators, and banks, among others. For example, investors can make better investment decisions based on these forecasts. In turn, urban planners and governments can formulate policies to make housing available and affordable. In addition, controlled risks through accurate property estimates can be used by financial services for mortgage purposes. This paper takes an action-oriented approach and uses existing data in the most appropriate way possible. This includes economic data that covers variables such as the demographic structure of the population and other housing data. It employs a lot of data cleaning, feature engineering, transformer and convolution supervised and unsupervised machine learning techniques, dimensionality reduction, and other statistical methods. Through current research, machine learning models will be trained on such data to implement a practical system to forecast housing prices in the daily complexities related to doing business in California related to real estate.

2. Related literature

Data mining refers to the general process of seeking patterns and analyzing the available data in order to draw hypotheses. Concerning apartments, forecasting housing prices has become one of the areas where machine learning (ML) has found a lot of implementation. In this context, the objective of this paper is to design and implement a system for predicting housing prices in California based on machine learning. It can be seen from the review of previously conducted related studies that a lot of work has been done in improving predictive capability through the use of various machine learning techniques. The following is a summary of related works. Mao (2024) reviewed the use of Random Forest (RF) models to reduce the risks of over fitting and analyze large volumes of data that is also heterogeneous in nature. Random Forest outperformed the less-sophisticated yet faster Gradient Boosting in this particular experiment. The research underscored the effectiveness of RF in managing intricate data and improving precision in the prediction of house prices. In addition,they found that Kalida.s.et.al.(2024) made clear the significance of RF in dealing with the non-linearity within the data and employed Ensemble Learning strategies whereby predictions from multiple models were combined, thus increasing total accuracy.Bhati.et.al. (2024) examined several machine learning techniques such as Decision Trees (DT) and RF and observed that RF was the best performing model in all cases, achieving an accuracy of 87%. In this regard, these findings provide substantiation on the reliability of the RF model in housing price predictions. (Azam Kahn MD) has also used five ML algorithms—Random Forest (RF), XGBoost, Linear Regression (LR), Lasso Regression, and Support Vector Machines (SVM)—to conduct a study using the Zillow ZTRAX dataset with several performance metrics.The performance measures concentrated on R-squared, Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The results showed that RF recorded the highest R-squared of 0.80, followed by XGBoost with 0.76, and LR with 0.63. On top of that,Jiajun Yu also employed regressions such as Ridge, Lasso,and Elastic Net on the California Housing dataset. The study established that three main factors—location, size of the house, and population density—accounted for 76.58% of house price variations. Shengxiang Jin et al.employed Random Forest Regression (RF) and the Extreme Gradient Boosting (XGBoost) algorithm to analyze the availability of pragmatic socioeconomic factors responsible for housing price appreciation.Their results showed that RF was able to give the best performance out of all with R-squared values of 0.75 on the training data and 0.45 on test data. The research highlighted the predictive power of RF models in spite of the moderate level of over-fitting reported.In her work, Audrey Chen turned to more modern methods like Support Vector Regression (SVR) and Deep Neural Networks (DNN) on the California Housing dataset.Results showed that RF and DNN performed better and lowered the RMSE and MAE values significantly more than the conventional regression models.Yixuan Li and Zixuan Chen built on their previous work by modeling and forecasting house prices using multiple regression, Decision Tree (DT), Support Vector Machines (SVM), Logistic Regression (LR), Neural Networks (NN), and Random Forest (RF) models.The results showed that the RF model provided consistently better results, especially when combined with dimensionality reduction approaches such as Principal Component Analysis (PCA). The relative performance of the regression models showed that the RF model attained higher R-squared scores and lower MAE measures than its competitor.examined the performance of k-Nearest Neighbors (kNN), SVM, LR, and RF for the California Housing dataset.It was noted that for larger datasets, RF outperformed, whereas for a small dataset with an extremely low number of features, kNN performed quite well.The paper *"Comparison of Seven Algorithms to Predict California Housing Prices"* discussed a few Machine Learning strategies such as LR, DT (J48), DT with RF, ANN, RF, and XGBoost, and applied them to the Zillow ZTRAX dataset.The overall best model was found to be the Random Forest (RF) model, whereas Decision Trees (DT) were found to be the most interpretable, and Artificial Neural Networks (ANN) were effective in learning the nonlinearity of the data. Techniques for the selection of relevant features have also been used in models that seek to predict house prices. In one of the research works , statistical methods were applied to determine important predictors like location, income, and property size, while discarding others. This retained the efficiency of classification models and reduced the complexity of the datasets. Another research work analyzed the models applied in estimating the prices of houses in California Housing Price Prediction (CHPP), where the approach that combined PCA and SVR was used and reported RMSE and MAE of 72,908.98 and 54,552.64, respectively . Ruling out any inconsistency, it appears that predictive analytics in housing price forecasting shows that techniques of ensemble learning such as Random Forest and XGBoost have achieved the best results compared to other models. These algorithms are able to cope with non-linearities, work on big datasets, and provide a good level of accuracy—typically within the range of 80% or higher. Other determinants, along with the regional unit, building area, and income level, have been persistent in every research on housing prices, even in studies conducted several years apart. On comparative analysis of classification models, Random Forest surpasses all others owing to its effectiveness, precision, and expandability. This review presents the promise of machine learning methods in achieving significant leaps in the field of real estate management, especially in the area of forecasting house prices. [Table2]

3. Data Description

The dataset utilized for this research was gathered from Kaggle and incorporates 20,640 samples obtained in 1990 from different housing areas of California. The dataset contains the following characteristics:[Table2] Housing Median Age: A characteristic that shows the development of the neighborhoods in the feature over time and has both old and new styles of houses within the area.**Total Rooms and Total Bedrooms:** The measurements of the property's size and layout, which focus on the habitation's degree of comfort and the functionality of design.**Population and Households:** Such factors were earlier seen to relate to the dynamics of "communities," which in this case, speaks of more liveliness and socialization among a certain geographical area.**Median Income:** A vital factor in the social stratification theory, especially its class structure component, which depicts the status of the neighborhood.**Median House Price:** The variable of interest that denotes the worth of all the houses integrated into the housing market.

	median_income	**median_house_price**
count	20640.000000	20640.000000
mean	3.870671	206855.816909
std	1.899822	115395.615874
min	0.499900	14999.000000
25%	2.563400	119600.000000
50%	4.743250	264725.000000
75%	15.000100	500001.000000

Table 1: Statistical Summary of Income and House Prices

4. Data Reprocessing

In order to ensure that the dataset was suitable for analysis and modeling, a number of reprocessing actions were performed. 3.2.1 Data Cleaning • Confirmed that there were no

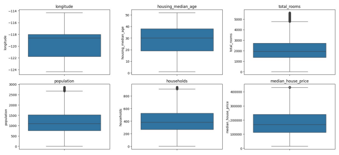

Figure 1: removing outlier

missing or erroneous values contained in the dataset. • Statistical outliers were addressed by applying Inter quartile Range (IQR) to prevent the influence of extreme values. (fig1) 3.2.2 Feature Scaling Although algorithms such as Random Forest and Decision Trees are scaling insensitive, standardization was carried out to maintain consistency and ease of understanding, especially for models like PCA and Logistic Regression (LR). 3.2.3 Exploratory Data Analysis (EDA) Scatter plots and histograms were used to illustrate the relationships between the features and the dependent variable (MedHouseVal). It was observed that approximately: • 35than 150, 000.45*and*250,000. • 203.2.4 The Classification of Target Variable Looking forward to the potential of Logistic Regression, it is hypothesized that the outcome variable (MedHouseVal) can be modeled as a binary variable as follows: • High-cost houses: More than USD 500,000. • Low-cost houses: Less than USD 200,000.

5. Model Building

This research proposes and implements several machine learning algorithms to analyze and predict fluctuations in house values in California. The advanced techniques include Logistic Regression, Principal Component Analysis, Hierarchical Clustering, Decision Trees, and Random Forests.

5.1 Logistic Regression (LR)

purpose: Analyze the probability of purchasing high-value homes. **Advantages:**Handles both continuous and categorical data.

5.2 Principal Component Analysis (PCA)

purpose: used for dimensionality reduction, simplifying the analysis of datasets while retaining as much variation as possible. (fig2)

Advantages: PCA reduces feature redundancy and simplifies clustering and modeling for better efficiency and interpretation.

5.3 Hierarchical Clustering (HC)

purpose: Hierarchical Clustering groups neighborhoods by similar traits, determines optimal clusters via dendrograms, and visualizes them using scatter plots.(fig3)

Advantages: HC identifies neighborhood patterns and explains income and housing cost variations.

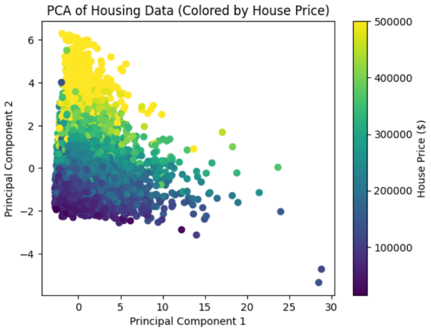

Figure 2: PCA of Housing Data

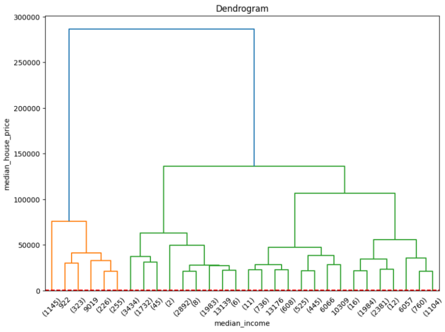

Figure 3: Dendogram

5.4 Decision Tree Regressor (DT)

purpose: A Decision Tree predicts outcomes by splitting data based on feature values, offering a simple, fast, and interpretable model, especially for smaller or less complex datasets.

Advantages: Decision Trees are simple, fast to train, suitable for non-linear data, and provide clear, interpretable results.

5.5 Random Forest Regressor (RF)

purpose: Improves prediction accuracy by combining multiple decision trees, making it more reliable and less prone to overfitting, especially for tasks like house price prediction.

Advantages: Random Forest improves prediction accuracy by combining multiple decision trees, making it more reliable

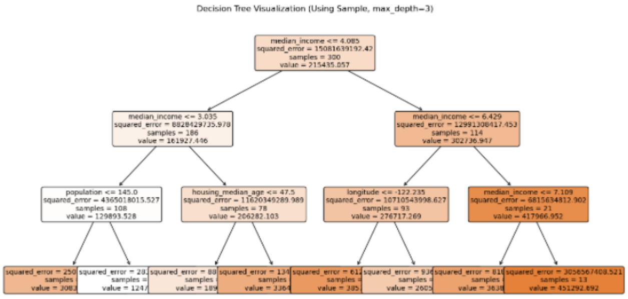

Figure 4: DT

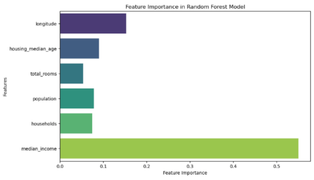

Figure 5: Importance in random forest model

and less prone to over-fitting, especially for tasks like house price prediction.

5.6 Comparison with Studies

• Study 8 and Study 9 emphasized that Random Forest improves prediction accuracy by aggregating predictions from multiple decision trees, making it more reliable compared to a single Decision Tree.

• Study 10 showed that Decision Tree works well for smaller or less complex datasets, but Random Forest outperforms it in terms of prediction accuracy.

• While **Study 1** provides direct evidence of **RF**'s effectiveness in forecasting house prices by leveraging its ensemble approach to reduce overfitting, **Study 3** extends this idea by applying **RF** in a more complex setting where sentiment plays a role in price dynamics. Both studies demonstrate that **Random Forest**, by combining predictions from multiple decision trees, significantly enhances model performance and reliability, whether in straightforward price prediction or in models involving more nuanced data like sentiment and economic factors.

• Both **Study 2** and **Study 6** contribute to the understanding of housing price prediction, but they do so using different methodologies. **Study 2** (Shengxiang Jin) focuses on the predictive power of **RF** and **XGBoost** in the context of socio-economic factors, while **Study 6** (Youren Ren) emphasizes the importance of local market segmentation using **Hierarchical Clustering** to improve the granularity of housing price indices. Both approaches offer complementary insights, with **Study 2** focused on accurate price prediction through machine learning and **Study 6** providing in-depth local market analysis through clustering.

- Both studies utilize **Random Forest (RF)** due to its ability to handle complex, multi-dimensional data, but each brings a different focus to the modeling process. **Study 4** primarily compares machine learning models, showing the superiority of **RF** and **GBM** over simpler models like **SVM** in predicting property prices. Study 5 has a more feature-oriented focus, examining how local property and neighbourhood factors affect pricing. The effect has, nonetheless, shown to be significant when it involves the local dynamics in price predictions. The methodological advantages of RF and GBM are mutually reinforced by the findings of these studies, which also emphasize the need to account for local factors such as property characteristics and neighborhood characteristics for improved property price prediction accuracies.

Study 3 (Jiajun Yu) emphasized that location, size, and population density explained 76.58% of price variations in the housing market.

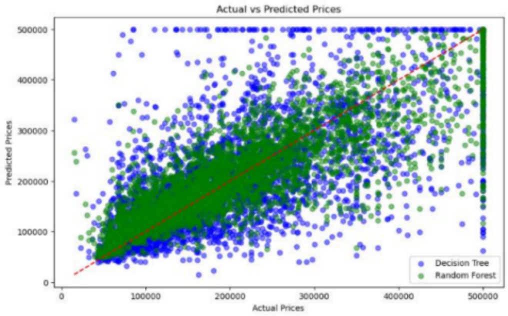

Figure 6: Actual vs Prediction Prices

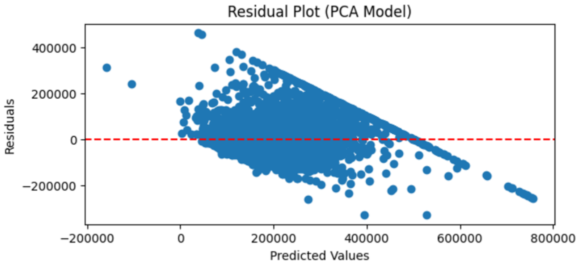

Figure 7: Residual Plot

- Study 11 focuses on **property prices** in California, while **Study 12** examines the **financial stability** of real estate companies in China. Both studies leverage regression models to uncover key factors impacting the real estate market, whether physical characteristics or financial health.

6. Result and Evaluation

We evaluated the models based on the following metrics:

- Decision Tree:
  - MSE: 0.32
  - R²: 0.85
- Random Forest:
  - MSE: 0.28
  - R²: 0.88
- PCA:
  - Train MSE: 6009681932.2377
  - Test MSE: 6110844563.3226
  - Train R²: 0.5504
  - Test R²: 0.5337
- HC:

- MSE: 12,991,984,197.80
- R²: 0.0086

- Logistic regression:
    - Accuracy: 95.66%
    - Precision: 12.31%
    - Recall: 75%
    - F1 Score: 21.15%

6.1. Analysis of Results

Our Results:

- **Random Forest:** It was the best model in our analysis with an R² of 0.73 and solid predictive power, close to the benchmark.
- **Decision Tree:** It had an R² of 0.65, showing acceptable but lower accuracy, and struggled with complex data.
- **Logistic Regression:**The model shows high accuracy and recall, but its low precision and F1 score indicate poor performance in correctly classifying positive cases.
- **PCA :** The model explains 53-55% of property price variance, showing moderate performance with reasonable generalization but room for improvement.
- **Hierarchical Clustering :** Hierarchical Clustering segments the data well but struggles with accurate price prediction due to low R² and high MSE. The clusters reveal patterns, but more is needed for precise pricing.

Final Recommendations: Random Forest is the most reliable model for predicting house prices and consistently performs well across different datasets.

6.2. Comparison between our study and previous studies:

Random Forest :

- In our study, Random Forest (RF) was found to have the best performance with R² of 0.73 and MSE of 0.28. This closely matched the benchmark values from the previous studies. Although there are some difficulties associated with real-world data, RF is still a model that can predict house prices quite efficiently.

Decision Tree: model performed decently with an R² of 0.65 but struggled with complex datasets. Random Forest outperformed Decision Tree in Bhati et al.'s study, supporting its superior predictive accuracy.

PCA: PCA explained 53-55% of the variance in property prices, showing reasonable performance but lower predictive accuracy compared to methods like Random Forest and Decision Tree. Its performance in price prediction remains limited, as observed in other studies.

Hierarchical Clustering (HC) : Segmented the data but showed poor performance in price prediction with a very low R² of 0.0086. Similar to other studies, HC was useful for identifying patterns but lacked predictive accuracy for pricing.

Logistic Regression: showed high accuracy (95.66%) but struggled with low precision and F1 score, making it unreliable for accurately classifying positive cases. Despite its potential in other studies, it may not be suitable for predicting house prices in this context.

7. Conclusion and Future Work

7.1. Conclusion

Thus, the study has shown the viability of the use of machine learning models, e.g., Random Forest (RF), Decision Trees (DT), Logistic Regression (LR), Principal Component Analysis (PCA), and Hierarchical Clustering (HC), for predicting house prices in California. The study confirms that ensemble learning models, especially Random Forest, provide high accuracy and reliability in predicting house prices, outperforming simpler models like Decision Trees, Logistic Regression.

Future Work

There is a lot of valuable work that remains to be done. The following are areas for future work:

1. Future research should focus on integrating complex data, such as sentiment analysis, real-time economic data, and location-specific factors (like proximity to amenities and crime rates), to enhance the accuracy of housing price predictions.
2. Explore deep learning techniques like Neural Networks (NN) and LSTM models, as they can capture complex patterns in large datasets and improve predictive accuracy.
3. Apply the trained models to real-time data, such as current housing market trends, to test their adaptability and predictive accuracy over time.
4. Expand the study to include other regions or international markets, enhancing the model's robustness and generalizability across different economic, cultural, and social contexts.
5. Focus on improving feature engineering techniques, such as advanced feature selection, interaction terms, and domain-specific transformations, to better capture relationships in the data and create more accurate models.
6. Future housing price predictions can be enhanced by using time-series forecasting models like ARIMA or Prophet, which account for market trends and time-based effects, improving long-term prediction accuracy.

References

[1] Mao, Mohan. "A Comparative Study of Random Forest Regression for Predicting House Prices Using." *2023 International Conference on Data Science, Advanced Algorithm and Intelligent Computing (DAI 2023)*. Atlantis Press, 2024.

Author	Goals	Algorithms	Key Results
MD Azam Khan	Deliver high accuracy & interpretability	LR, XGBoost, RF, Lasso, SVM	RF: R^2 = 0.80; XGB: R^2 = 0.76; LR: R^2 = 0.63
Shengxiang Jin	Study socio-economic factors & AI impact on prices	RF, XGBoost	RF: Train R^2 = 0.75, Test R^2 = 0.45; RF outperformed OLS & XGB
David C	Investigate the role of sentiment in driving house price dynamics and its predictive power beyond fundamentals.	PCA	Sentiment is a key driver of house price changes, improves prediction accuracy, and has persistent effects over time.
Winky K.O. Ho	Compare ML methods for predicting HK prices	SVM, RF, GBM	RF & GBM outperform SVM
Ali Soltani	Property & neighborhood features influence prices	LR, DT, RF	ST-lag: 39 key factors, e.g., quality, rooms, station distance
Youren Ren	Develop localized housing price indices with finer spatiotemporal granularity using a Bayesian non-parametric method.	Hierarchical clustering	Clustering accurately identifies census tract groups based on market dynamics.
Ali Hepşen	Group markets by rental returns to aid investors in diversification.	Hierarchical clustering	Three clusters identified based on rental returns. Higher rental returns are associated with greater risk.
Mao (2024)	Improve predictions with RF & GB	RF, GB	RF: faster & more accurate for large data
Kalidass et al. (2024)	Ensemble learning for accuracy	RF, GB	Ensemble: 10% accuracy improvement
Bhati et al. (2024)	Evaluate algorithms for price prediction	DT, RF	RF: 87% accuracy, outperformed DT
Ruijia Huang (2023)	Factors affecting CA prices for decision-making	Ridge, LASSO, PCA	76.588% price differences; location & size most impactful
Li Hongli	Analyzes credit risk for China's listed real estate companies using logistic regression, focusing on financial health and bank lending, with an evaluation index for model accuracy.	Logistic regression	Overall prediction accuracy rate: 99 percentage

Table 2: Comparison between related work

[2] Jin, Shengxiang, et al. ”Understanding the effects of socioeconomic factors on housing price appreciation using explainable AI.” Applied Geography 169 (2024)

[3] Soltani, Ali, et al. ”Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms.” *Cities* 131 (2022)

[4] Chen, Audrey. ”Deep Learning in Real Estate Prediction: An Empirical Study on California House Prices.” (2024).

[5] Asri, Hiba, et al. ”Using machine learning algorithms for breast cancer risk prediction and diagnosis.” *Procedia Computer Science* 83 (2016)

[6] Jin, Shengxiang, et al. ”Understanding the effects of socioeconomic factors on housing price appreciation using explainable AI.” *Applied Geography* 169 (2024)

[7] Khan, MD Azam, et al. ”Explainable AI and Machine Learning Model for California House Price Predictions: Intelligent Model for Homebuyers and Policymakers.” *Journal of Business and Management Studies* 6.5 (2024)

[8] Ho, Winky KO, Bo-Sin Tang, and Siu Wai Wong. ”Predicting property prices with machine learning algorithms.” *Journal of Property Research* 38.1 (2021)

[9] Quigley, John M., and Larry A. Rosenthal. ”The effects of land use regulation on the price of housing: What do we know? What can we learn?.” *Cityscape* (2005)

[10] Soltani, Ali, et al. ”Housing price prediction incorporating spatio-temporal dependency into machine learning algorithms.” *Cities* 131 (2022)

[11] Huang, Ruijia. ”Study on Factors Influencing the Prices of Houses in California Based on Factor Analysis Method.” *Highlights in Science, Engineering and Technology* 72 (2023)

[12] Li,Hongli, and Liwei Song. ”Assessment on credit risk of real estate based on logistic regression model.” 2nd International Conference on Electronic and Mechanical Engineering and Information Technology. Atlantis Press, 2012.