Analysis of Housing price in California Using Machine Learning Methods

Group 14: Qianying Tang, Jiayi Xu

1. Introduction

Predicting housing price is crucial for various stakeholders, including homeowners, buyers, and policy makers, as it has far-reaching impacts on the economy and society. California has a vast urban center, beautiful coastal areas, and bustling metropolitan areas, presenting a unique housing dynamic. In California, the real estate market is vibrant and competitive, and accurately predicting housing prices is particularly challenging due to the state's vast geographical diversity and economic disparities. The rapid growth of population, diverse socio-economic factors, and geographical differences have created a multidimensional market, where there may be significant price differences between communities and regions. There are many factors that affect housing prices, such as median income, population density, and geographical location. Understanding these factors and their interactions are crucial for developing accurate predictive models. Due to reduced commuting time, houses near employment centers often have higher prices, while areas with higher middle incomes often experience an increase in housing demand, leading to price increases.

Machine learning algorithms are increasingly utilized in the mass appraisal of real estate and automated valuation models, employing standardized procedures to collect data from real estate offers.[1] This approach ensures that property valuations are conducted in a uniform and impartial manner.[1-4] Recently, machine learning methods have been adopted for estimating house prices, marking these technological approaches as relatively new innovations.[1] However, there is no consensus yet on which machine learning algorithm or algorithms are best suited for predicting house prices. In this project, machine learning methods are used to provide a way to recognize patterns, predict prices, and provide information for decision-making processes. A total of 4 different machine learning methods are selected into this project, linear regression model, KNN regression model, decision tree model, and random forest model. Therefore, the purpose of this study is in finding out the best machine learning algorithm for predicting house prices through a dataset of house prices in California.

2. Related work

Roy E. Lowrance developed and tested various designs for linear models of residential real estate prices in Los Angeles County from 2003 to 2009, focusing on optimizing features, training periods, and regularizers.[5] His comparison revealed that, despite

minimal design efforts, random forests outperformed the meticulously crafted linear models.[5] In addition, Wang and Wu analyzed 27,649 housing data from Arlington County, Virginia, finding that Random Forest surpassed Linear Regression in accuracy.[6] Researchers utilized 1,970 housing transaction records to estimate property values in Krasnoyarsk, using features like room count and construction year, and applying algorithms such as random forest, ridge regression, and linear regression.[7] Their findings indicated that random forest significantly outperformed the other models based on mean absolute error (MAE).[7] Using 89,412 housing transaction records from Los Angeles, the author analyzed home price log errors using machine learning methods like linear regression, decision tree, and random forest, among others, utilizing features such as bedroom count and tax value.[8] Despite different methods, all underestimated the Zestimate error, suggesting Zestimate's already accurate housing price predictions.[8] Alfaro-Navarro and his colleagues evaluated the performance of the boosting algorithm against two other ensemble algorithms.[9] They discovered that bagging and random forest yielded superior results compared to boosting.[9] The authors analyzed the accuracy of rental price predictions using four machine learning algorithms—Cubist, gradient boosting, MARS, and SVM—and a traditional hedonic model (GLM) across a comprehensive dataset of English rental properties from a property listings website.[10] Their findings revealed that machine learning algorithms, particularly the tree-based Cubist and gradient boosting, significantly outperformed the GLM and other regression-based methods.[10]

3. Methods

Our main objective is developing a machine learning model to predict the housing price in California based on data in 1990s. The outcome is continuous variable, the median housing price, therefore, frame the problem as a regression task. In this method part, we will discuss four popular methods, namely regression model, KNN regression model, decision tree, and random forest, would be used and compared in this project.

Regression model is a method of estimating the coefficients of multi-regression models to predict the independent variable median housing price based on the predictor variables. In our study of California housing prices, linear regression model was used as the baseline model. This model assumes a linear relationship between predictive variables and response variables. Linear regression is particularly valued for its simplicity, interpretability, and efficiency, as it can provide a preliminary understanding key influencing factors on housing prices.

KNN regression model provides a flexible method to capture more complex nonlinear relationships that linear regression may miss. This method can adaptively learn from data, allowing for local regression that is sensitive to patterns in specific regions of the feature space. This feature makes KNN particularly useful in real estate prediction, where the

proximity of features such as location, room number, and population characteristics significantly affects housing prices. The cross validation is used to find the value of K, which is the minimal cv error. Therefore, the selected k will be used in KNN regression model.

Decision tree are multifunctional models used for regression and classification tasks, capable of handling numerical and classification data. When predicting housing prices in California, the decision tree model divides the feature space into regions with similar values and makes decisions by learning simple decision rules inferred from data features. Decision trees provide clear visualization of the decision-making process. To avoid the overfitting, various depths were tested to find the optimal balance between bias and variance.

Random forest is an ensemble learning method that operates by constructing a large number of decision trees during training and outputting the average predictions of each tree. This model is very effective for predicting regression tasks such as housing prices, as it combines the simplicity and flexibility of decision tress, reduces the risk of overfitting, and improves prediction accuracy. Against the backdrop of California housing prices, random forests can utilize the collective power of multiple trees to handle the inherent complexity and non-linear relationships in the dataset, thereby making more accurate and stable predictions.

By launching those four models, training error and test error would be compared to determine the best model for predicting the housing prices in California based housing price dataset in California.

4. Data and Experiment setup

The dataset we plan to use is California Housing Prices Data in Kaggle. It provides a comprehensive snapshot of California's housing market in the 1990s, capturing various socioeconomic and demographic factors. The dataset includes 10 categories of predictors, namely median house value, median house age, total rooms, total bedrooms, households, population, median income, ocean proximity, longitude, and latitude. Histograms shows each feature's distribution is shown as followed. 20433 observations in California housing price were included, 80% of those observations are randomly selected as training sets for model building and the rest 20% of data are included in the test set as our initial settings. The training set is used to construct each model, and the test set is used to evaluate the model. To avoid overfitting, each training error and test error are important measurement criteria for selecting the best model.

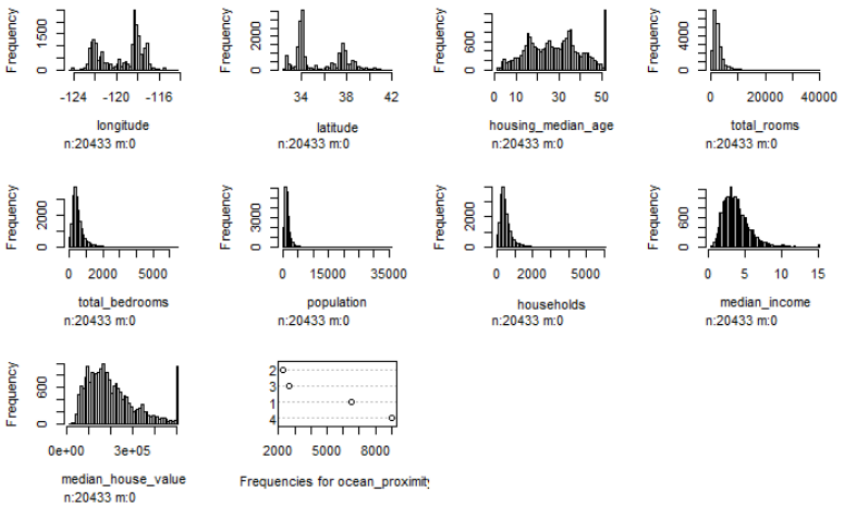

Figure 1: Feature Histogram Summary

5. Results

5.1 Linear regression model

The first model we used was the linear regression model. The results showed that all features used excepted households and ocean proximity "near ocean" were statistically significant. The house age, number of rooms, number of bedrooms, and income increased the price of housing. Our results showed negative coefficients for variables like population, longitude and latitude. The training error was 7.97e+13, and the test error was 2.19e+13. The big difference between training error and test error highlighted the challenges faced by the model in generalizing to new data, suggesting that there may be overfitting during the training phase.

```
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -2.578e+06  1.032e+05 -24.980  < 2e-16 ***
housing_median_age 1.126e+03  4.942e+01  22.784  < 2e-16 ***
total_rooms       -5.444e+00  8.727e-01  -6.238 4.53e-10 ***
households         5.350e+01  8.088e+00   6.614 3.85e-11 ***
total_bedrooms     9.295e+01  7.479e+00  12.429  < 2e-16 ***
population        -3.833e+01  1.167e+00 -32.848  < 2e-16 ***
median_income      3.824e+04  3.751e+02 101.954  < 2e-16 ***
ocean_proximity2   7.111e+04  3.213e+03  22.131  < 2e-16 ***
ocean_proximity3   3.808e+04  2.674e+03  14.243  < 2e-16 ***
ocean_proximity4   3.190e+04  2.090e+03  15.260  < 2e-16 ***
longitude         -2.986e+04  1.224e+03 -24.393  < 2e-16 ***
latitude          -2.786e+04  1.233e+03 -22.597  < 2e-16 ***
```

Figure 2. regression model summary

Model	Training error	Test error
Linear regression	7.97e+13	2.19e+13

5.2 KNN regression model

The second model we used was KNN regression model. Firstly, we used cross validation to confirm the value of k. This graph showed how the model's performance changed as we adjusted 'k'. Initially, when "k" was small, the model was highly sensitive to the noise present in the training data, leading to overfitting. However, as "k" increased, the model began to stabilize and better generalized the underlying trends in the data. Cross

validation error was minimum when k was 25. Therefore, in KNN regression model with k = 25 would be performed. The training error was 6.04e+10, and the test error was 3.91e+10.

Model	Training error	Test error
KNN regression	6.04e+10	3.91e+10

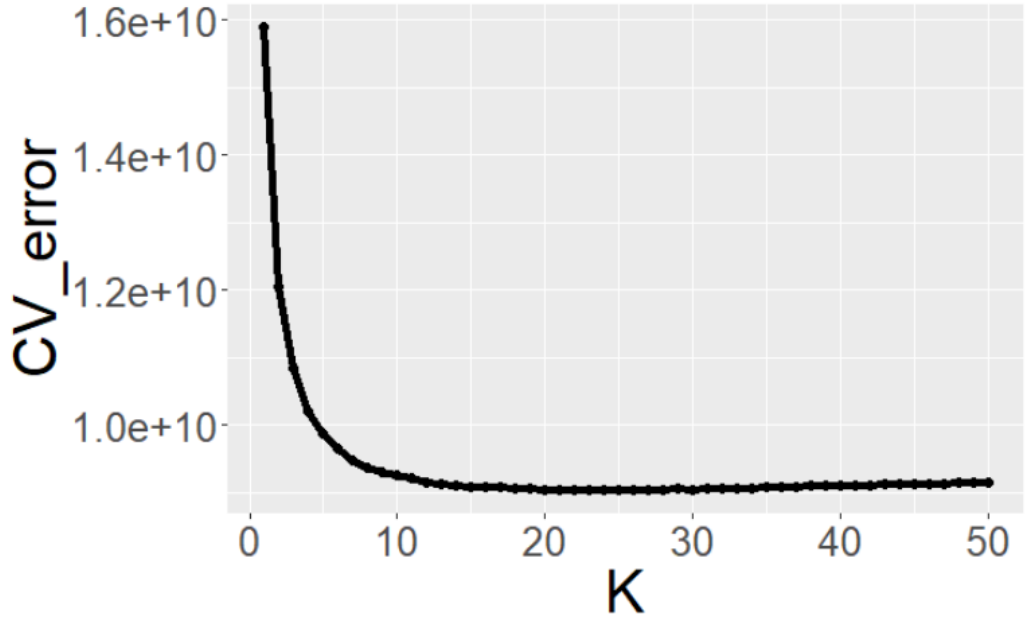

Figure 3. Cross Validation error

5.3 Decision tree model

The third method we used was decision tree model. Using cross validation, the best size of the tree turned out to be 10, and the pruned tree with 10 nodes was shown in Figure 4. According to the prune decision tree, it was intuitive that the predictor median income played a leading role in housing price, and ocean proximity was another essential indicator. Although the decision tree model was not complexly interpreted, its prediction performance was not that good. The training error was 5.42e+9, and the test error was 1.03e+10.

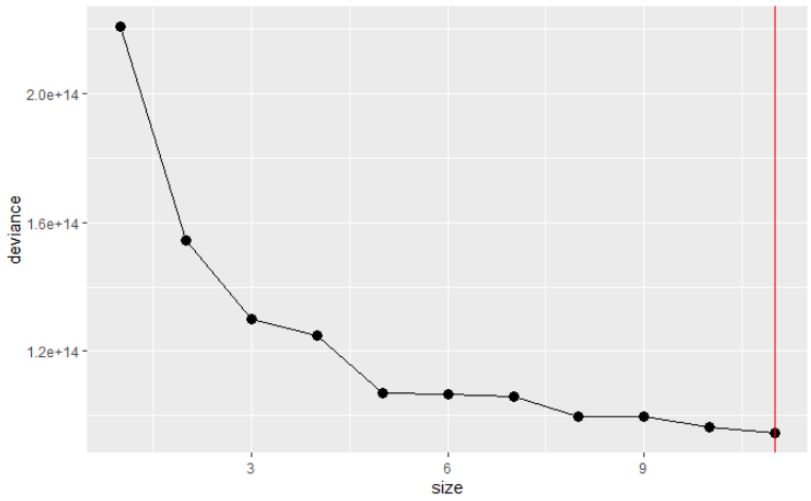

Figure 4. Tree size

Model	Training error	Test error
Decision tree	5.42e+9	1.03e+10

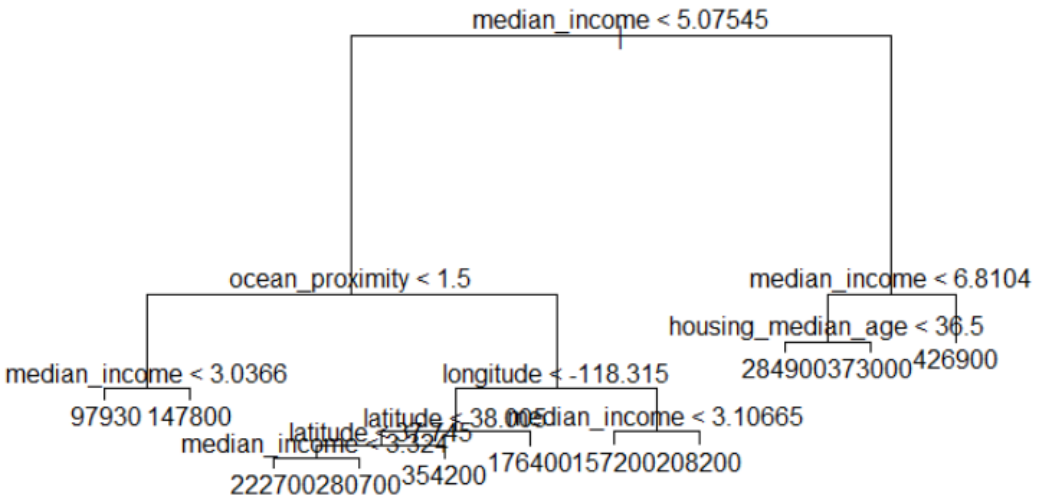

Figure 5. The pruned decision tree.

5.4 Random forest model

The last model we used was random forest model. When consider the variable importance indicators, median income, ocean proximity and latitude were the most important variables indicated by variable importance plot for the random forest model, which showed the relative importance of each feature in the model. For %Inc MSE and Inc Node Purity, the median income was the best fitted variable in random forest model.

According to the summary, the median income occupied 156.79 importance for this model. In additional, the training error was 4.93e+8, and the test error was 5.49e+9.

	%IncMSE	IncNodePurity
housing_median_age	87.18430	1.143941e+13
total_rooms	33.29776	9.096601e+12
households	45.44297	6.689727e+12
total_bedrooms	40.23563	7.126366e+12
population	59.31135	1.113788e+13
median_income	156.78609	8.313842e+13
ocean_proximity	86.38508	3.029890e+13
longitude	67.81794	3.151700e+13
latitude	59.49293	2.584337e+13

Figure 6. random forest summary

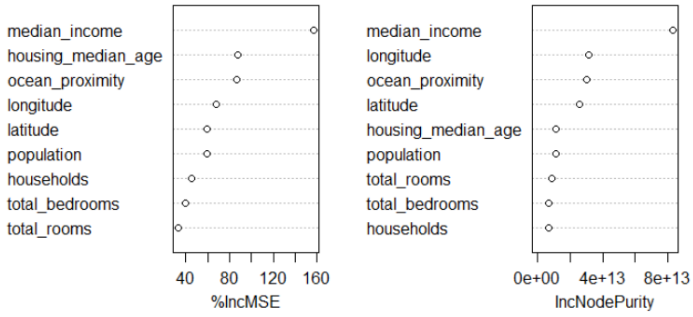

Figure 7. random forest variable importances summary

Model	Training error	Test error
Random Forest	4.93e+8	5.49e+9

6. Discussion

Below is the summary of model performance on the training and test set. According to the performance summary, the random forest was better than other models, with the smallest test error. So out of the four models, the KNN regression, decision tree, and random forest were recommended, they all had relatively small training error and test error. The linear regression model showed the basic relationships between housing features such as median income and housing prices, confirming established economic theories that link income levels to real estate values. More complex models like KNN and random forest demonstrated an ability to capture nuances, non-linear interactions between variables that simple models often missed. For instance, the random forest model identified specific location features and demographic characteristics as critical predictors of housing prices, providing a more detailed landscape of what drives market values in different regions.

Model	Train errors	Test error
Regression model	7.97e+13	2.19e+13
KNN Regression	6.04e+10	3.91e+10
Decision Tree	5.42e+9	1.03e+10
Random Forest	4.93e+8	5.49e+9

Our results demonstrated a progression in model complexity from linear regression to random forest, with random forest showing the lowest test error (5.49e+9). This aligns with Lowrance's finding that random forests, despite minimal design efforts, outperformed more traditional linear models due to their ability to capture non-linear patterns and interactions between variables.[5] Similarly, Wang and Wu noted that random forests surpassed linear regression in predicting housing prices in Arlington County, validating the superior performance of ensemble methods in handling large and complex datasets.[6] However, the observation in our study that all machine learning methods underestimated Zestimate prediction errors suggested a limitation in the models' ability to completely account for all influencing factors, despite their advanced algorithms.[8] This was an important consideration for future model improvements and indicated the sophistication of Zestimate's proprietary algorithms.[8]

The relationship between housing features, including location, demographic makeup, and public services, had direct implications for public health and urban planning. Understanding how these factors influenced housing prices ca guide policymakers in designing interventions that not only enhanced property values but also improved community health and well-being. Enhancing public transportation and access to parks could not only increase property values bot also promote healthier lifestyles.

The predictive accuracy of models like KNN and random forest can be sensitive to the choice of parameters and the representation of the dataset. Additionally, our study primarily focuses on quantitative data, potentially overlooking qualitative aspects such as neighborhood desirability and subjective perceptions of safety and community, which can also significantly affect housing prices. Meanwhile, the disparity between training and test errors in our linear regression model suggested possible overfitting, an issue that was less pronounced in the random forest model. This was reflective of the general trend seen in the reviewed studies, where more complex models like random forests tend to generalize better compared to simpler models like linear regression.

7. Conclusion

This study showed the effectiveness of multiple machine learning models in the prediction of housing prices, particularly emphasizing the superiority of ensemble methods like random forest over simpler models including linear regression and KNN. These models excelled not only in accuracy but also in managing overfitting. Identified key predictors such as median income and ocean proximity underscored the critical role of feature selection in the valuation of real estate. In the future, the housing price can combine with map coordinates to divide regions and determine where housing price have higher value. Additionally, exploring real-time data integration could provide more dynamic predictive capabilities, reflecting rapid market changes more accurately. Finally, the study also can explore the causal impacts of policy interventions on housing markets.

Authors' contributions:
Qianying Tang contributed to the thesis by writing the Introduction, Related Work, Conclusion, and Discussion sections, providing a comprehensive overview and synthesizing the findings into broader implications and future research directions. **Jiayi Xu** is responsible for the "Methods," "Results," and "Data and Experimental Setup" sections, detailing the technical methods used, demonstrating the data analysis, and outlining the experimental framework so that the study is grounded in empirical evidence and rigorous methodology.

Github Repository: https://github.com/Bernelq/machine-learning

Reference:

1. Mora-Garcia R-T, Cespedes-Lopez M-F, Perez-Sanchez VR. Housing price prediction using machine learning algorithms in COVID-19 times. *Land*. 2022;11(11):2100.
2. Kauko T, d'Amato M. Introduction: Suitability issues in mass appraisal methodology. *In: Mass Appraisal Methods: An International Perspective for Property Valuers*. 2008:1-24.
3. Grover R. Mass valuations. *J Prop Invest Financ*. 2016;34(2):191-204.
4. International Association of Assessing Officers. Standard on Mass Appraisal of Real Property. Kansas City, MI: International Association of Assessing Officers; 2019:22. Available at:https://www.iaao.org/media/standards/StandardOnMassAppraisal.pdf. Accessed August 22, 2022.
5. Lowrance RE. Predicting the market value of single-family residential real estate. New York University; 2015.
6. Wang CC, Wu H. A new machine learning approach to house price estimation. *New Trends Math Sci*. 2018;6(4):165-171. doi:10.20852/ntmsci.2018.327.
7. Koktashev V, Makeev V, Shchepin E, et al. Pricing modeling in the housing market with urban infrastructure effect. *In: Journal of Physics: Conference Series*. 2019;1353(1):012139. IOP Publishing.
8. Huang Y. Predicting home value in California, United States via machine learning modeling. *Stat Optim Inf Comput*. 2019;7(1):66-74.
9. Alfaro-Navarro JL, Cano EL, Alfaro-Cortés E , et al. A fully automated adjustment of ensemble methods in machine learning for modeling complex real estate systems. *Complexity*. 2020;2020:1-12.
10. Clark SD, Lomax N. A mass-market appraisal of the English housing rental market using a diverse range of modelling techniques. *J Big Data*. 2018;5:1-21.
11. Chen, Y. (2023). Analysis and Forecasting of California Housing. *Highlights in Business, Economics and Management*, *3*, 128-135. https://doi.org/10.54097/hbem.v3i.4704