Proceedings of ICFTBA 2024 Workshop: Finance in the Age of Environmental Risks and Sustainability
DOI: 10.54254/2754-1169/94/2024OX0208

A Study of Property Intrinsic Value Prediction Based on Machine Learning Modeling

Wentao Zhang[1,a,*]

[1]*Rotterdam School of Management, Erasmus University Rotterdam, Rotterdam, 3062 PA, Netherlands*
a. 645155wz@eur.nl
**corresponding author*

Abstract: When purchasing goods, buyers consistently prioritize the fairness of the price to make informed decisions, particularly for high-value items. This consideration becomes even more crucial in the real estate market, where verifying the fairness of a property's asking price through conventional methods can be challenging. This study addresses this issue by aiming to develop a functional predictive model that can estimate a property's intrinsic value using various machine learning techniques. By analyzing the California Census Data published by the US Census Bureau, this research investigates the potential for determining a property's intrinsic value based on readily available information about the block in which the property is situated. The study's objective is to create a model that provides accurate and reliable property valuations, helping buyers make more informed decisions and promoting transparency in the real estate market. This model leverages demographic, economic, and housing data from the census to predict property values, potentially transforming how properties are appraised and traded. By offering a method to verify asking prices, the research seeks to enhance fairness and efficiency in real estate transactions, providing a valuable tool for buyers, sellers, and real estate professionals alike.

Keywords: Property Intrinsic Value, Prediction Model, Machine Learning.

1. Introduction

In today's competitive real estate market, verifying the fairness of a selling price is harder than ever. Different properties, even locating right next to each other, share the exact same layout and have the same floorspace, can still be vastly different in selling price due to the different aspects of the properties. Some representative examples include different furniture, style of furnishment, construction materials, and the state of the property can significantly impact the market value and buyers' perception. With the presence of information asymmetry and the depicted scenario above, it is almost impossible for a regular buyer to objectively evaluate the fairness of a selling price without an appropriate baseline. This study focuses on tackling such phenomenon by utilizing multiple linear regression and random forest to provide buyers with an appropriate index to facilitate an objective and accessible decision-making process.

The objectiveness and generalizability of an index should be held at the most priority. Due to the individual differentiation nature of properties, it is reasonable to not include features that closely relates to an individual property to avoid reduced prediction accuracy. Therefore, the median house

DOI: 10.54254/2754-1169/94/2024OX0208

price within a block is selected as the prediction objective. Properties within a block typically shares the same underlying characteristics, for example, at the same location, sharing the same infrastructure, restrained by the same building codes, and sharing the same demographic profiles of residents including average income levels, education rates, and so on. These characteristics contributes to the overall objectivity when evaluating price of one specific property within the block, while at the same time improving the generalizability of the prediction model as the same scenario apply to most of the block in the US. The buyer is thus able to utilize the median house price of the block as a baseline, and then proceed smoothly to incorporate all the subjective elements mentioned previously to arrive at an informed evaluation of the fairness of the selling price.

2. Literature Review

Over the past years, efforts have been made to estimate the house prices with different machine learning techniques, these researches demonstrated the exceptional capabilities of machine learning. Thanks to machine learning's ability to provide accessible data handling, the automated weighing of different features, and its scalability, machine learning have become one of the most commonly used prediction approach. Several studies have previously uncovered the potential of machine learning algorithms on predicting house prices [1-4]. These studies have all focused on identifying significant property characteristics that influence house prices and enhancing the accuracy of prediction [5, 6]. Excepting for study, all other studies have implemented more than one machine learning technique and performed comparison between the different models, in which Milunovich performed a comprehensive comparison between 47 different algorithms. However, there is no sufficient evidence to summarize one specific model that outperform other models, though the results follow the general hierarchy of model performance that deep learning models outperforms other models such as multiple linear regression, boosting, random forest, and so on. Some researchers in the previously mentioned studies performed house price prediction with some interesting approach: Truong et al. and Lu et al. performs the prediction combining more than one models to enhance the prediction accuracy [7, 8]. Al-Sit et al. and Tchuente et al. focus their prediction with a combination of geo-coding and machine learning [9, 10]. Some studies both did not use property dependent characteristics in their prediction, Phan TD used property independent characteristics such as number of bedrooms, bathrooms, and land sizes [2, 8]. Sanjar used features such as median sale price in labeled neighborhood, but focused on the handling of missing data by filling in KNN-MCF based simulation results [6].

It is obvious to see that researchers have shown great interest in the estimation of house price with various different approaches and focuses on the different technical aspect aiming to establish robust prediction models. However, the current academic focus lies more in the exploration and enhancement of modeling the relationship between property dependent characteristics and house prices. This study aims at providing a new perspective on predicting intrinsic house price based solely on the block characteristics.

3. Methodology

3.1. Data Selection

The dataset used in this study is acquired from Kaggle uploaded by M. Shibu, it includes the California Census Data published by the US Census bureau. The original dataset includes 20640 rows of input describing blocks in CA, including longitude, latitude, house median age, total rooms, total bedrooms, population, households, median income, ocean proximity, and median house value. The detailed data description could be found in Table 1 and Figure 1 below.

DOI: 10.54254/2754-1169/94/2024OX0208

Table 1: Data description.

	Longitude	Latitude	Housing median age	Total rooms	Total bedrooms	Population	Households	Median income	Ocean proximity	Median house value
0	-122.23	37.88	41	880	129	322	126	8.3252	NEAR BAY	452600
1	-122.22	37.86	21	7099	1106	2401	1138	8.3014	NEAR BAY	358500
2	-122.24	37.85	52	1467	190	496	177	7.2574	NEAR BAY	352100
3	-122.25	37.85	52	1274	235	558	219	5.6431	NEAR BAY	341300
4	-122.25	37.85	52	1627	280	565	259	3.8462	NEAR BAY	342200

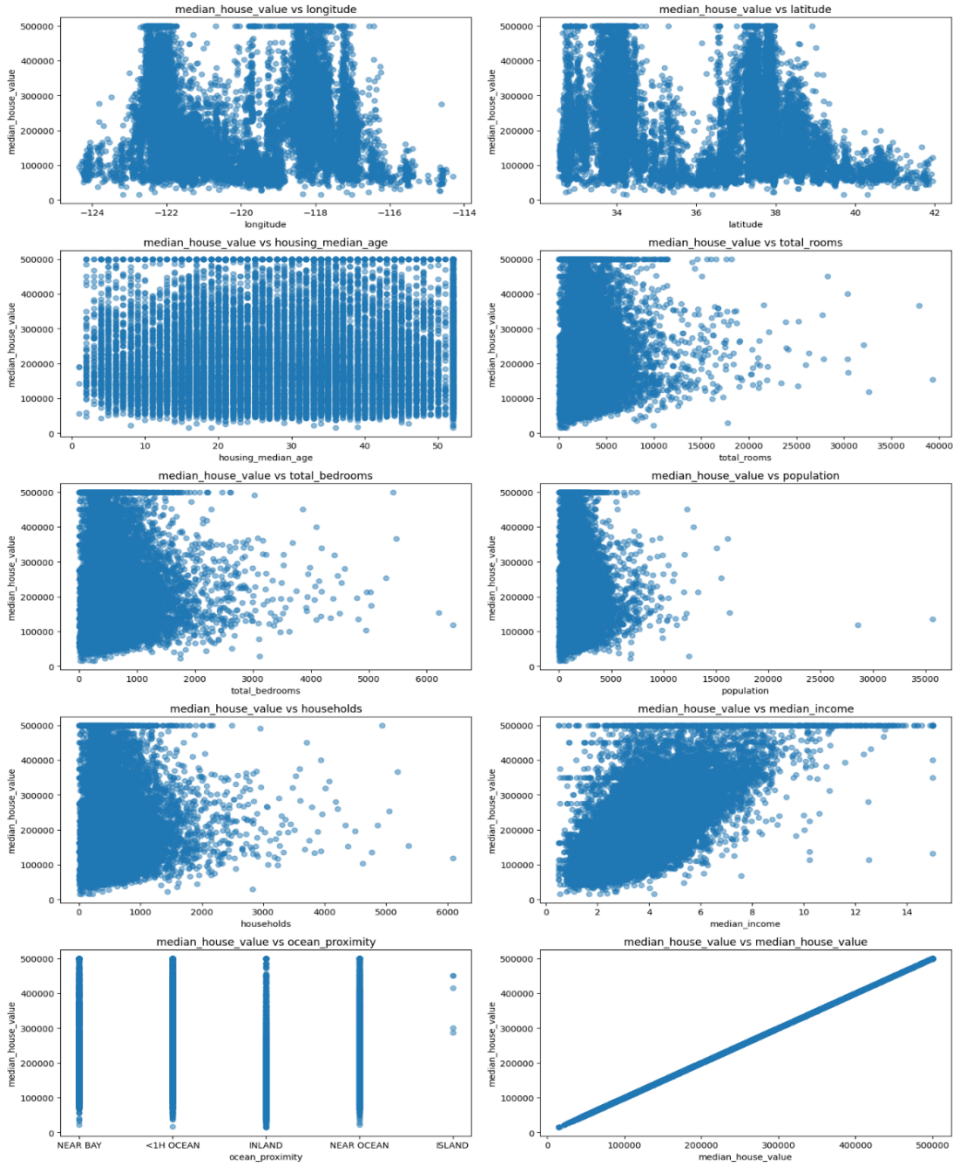

Figure 1: Data features plotted against median house value (Photo/Picture credit: Original).

DOI: 10.54254/2754-1169/94/2024OX0208

3.2. Data Processing

As previously mentioned, the original data set includes 20640 rows of input and 10 variables. Except for the object to predict, the rest of the nine features could be used to perform analysis. The data handling steps were:

- Handle missing data. No feature was removed due to excessive missing data. The original dataset includes 207 missing data in column total bedrooms
- Remove column ocean proximity
- Handle irregular data to purify data set. Looking at Figure 1 and Figure 2, it is clear there are irregularity in the variable median house value and housing median age, namely cluster of data in median house value of 500001 and housing median age of 52. A data cap was set at these two variables such that all data above these the values 50001 and 52 are stored as these numbers themselves. After consideration, all rows with the a median house value of 500001 and housing median age of 52 was deleted. After deletion, the data set reduced 14.5% in size, resulting in 17647 rows of data.

Figure 2: Irregularity in data set (Photo/Picture credit: Original).

- Handing outliers through identifying and deleting rows that contain z values over the 3 Inter-Quartile Rnage(IQR) range deviation from mean value

  Scatter plots of data set after data handling, as shown in Figure 3.

DOI: 10.54254/2754-1169/94/2024OX0208

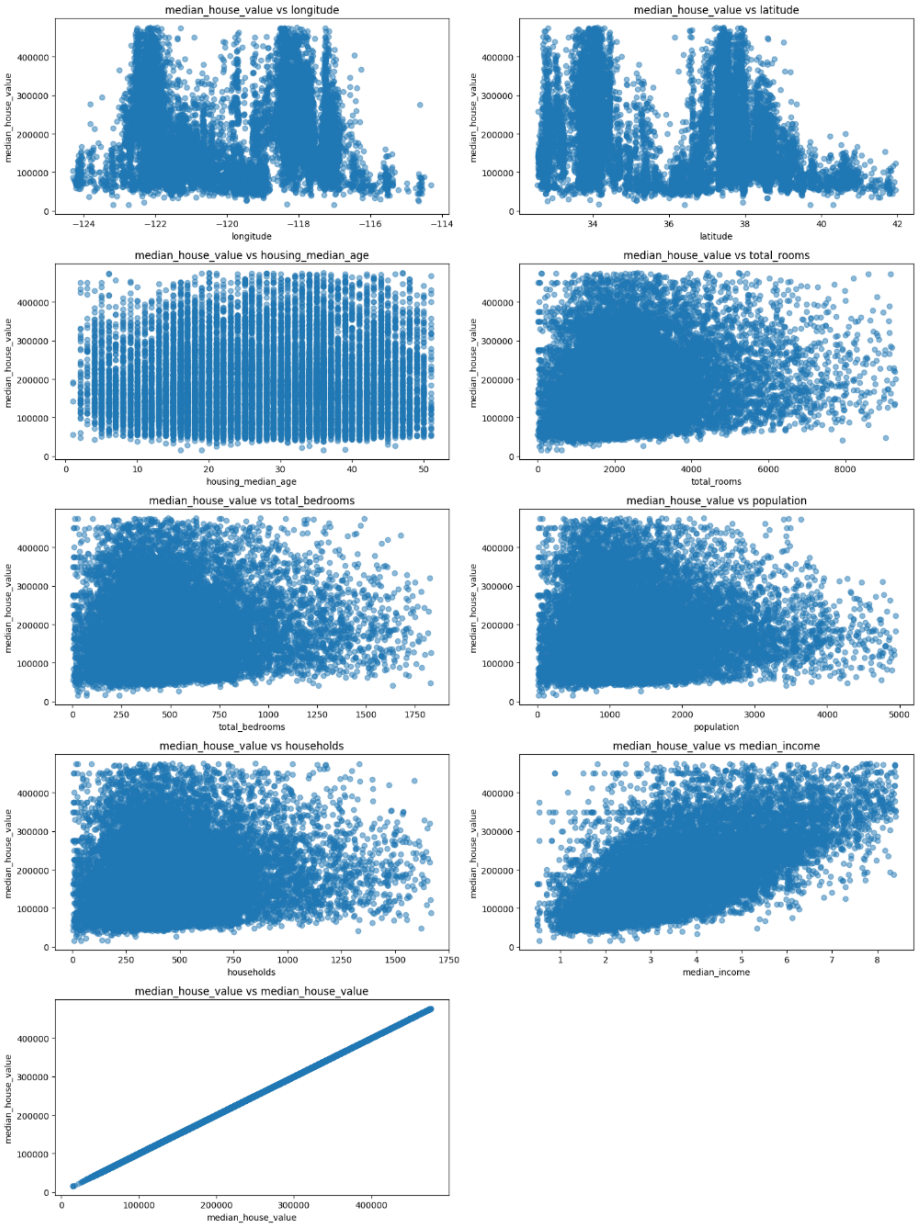

Figure 3: Data features plotted against median house value after data handling.
(Photo/Picture credit: Original)

3.3. Model Selection

The evaluaiton criteria for all the models are R^2, RMSE, MAE, and accuracy, below are the functions to illustrate the criterias:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

DOI: 10.54254/2754-1169/94/2024OX0208

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \quad (4)$$

$$Accuracy = 100\% - MAPE \quad (5)$$

3.3.1. Multiple Linear Regression

In this study the first model chosen is multiple linear regression where it models the relationship between the dependent variable(median house value) and the independent variables(the rest of the variables) by fitting a linear equation to the observed data.

The resulting figure and summary of actual vs predicted median house value using multiple linear regression is displayed in Figure 4 below, and Table 2 illustrates the results of linear regression.

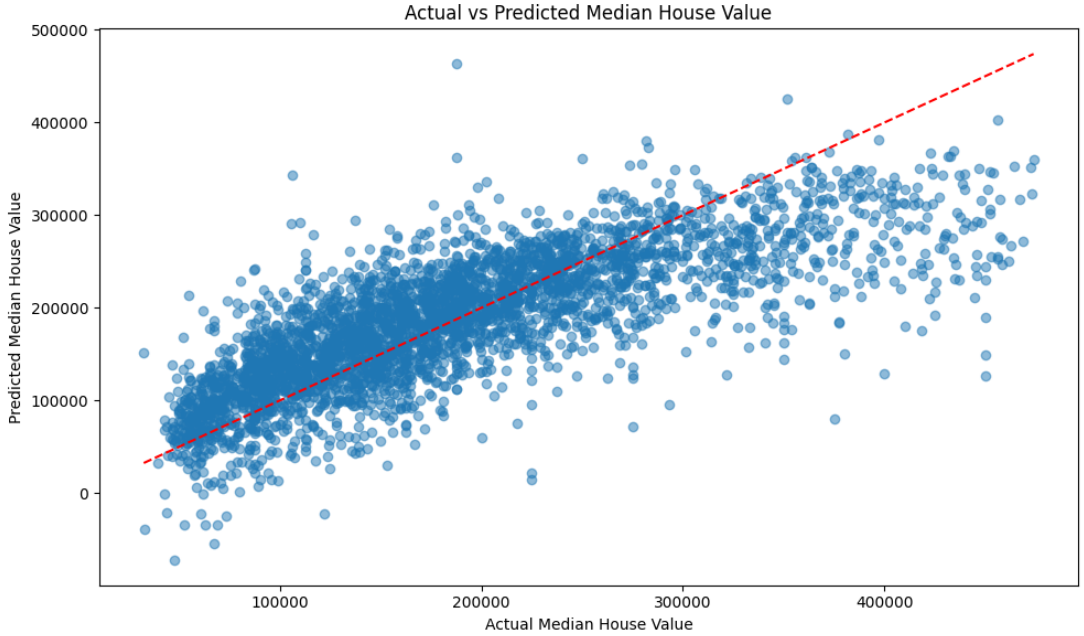

Figure 4: Actual vs predicted house value Multiple Linear Regression.
(Photo/Picture credit: Original)

Table 2: Summary of Multiple Linear Regression.

	Variable	Coefficient	Std. Error	t-value	$P > \|t\|$	[0.025	0.975]
1	Constant	-3159853	62604.93	-50.47	0	-3282564	-3037143
2	Longitude	-3758342	715.8	-52.51	0	-38986.47	-36180.36
3	Latitude	-37110.94	676.26	-54.88	0	-38436.5	-35785.37
4	Housing median age	752.82	47.86	15.73	0	659	846.64
5	Total rooms	-15.48	1.1	-14.04	0	-17.64	-13.32
6	Total bedrooms	155.57	8.61	18.07	0	138.69	172.45
7	Population	-42.2	1.37	-30.87	0	-44.88	-39.52
8	Households	51.36	9.25	5.55	0	33.21	69.49
9	Median income	40472.66	472.24	86.27	0	39817.01	41668.32

DOI: 10.54254/2754-1169/94/2024OX0208

3.3.2. Multiple Linear Regression Log Transformed

The study further implement log transformation to explore the impact it has on the prediction accuracy and in hope of enhance the model performance. Each data is transformed to a log number by performing this step: log(value). Such approach help to stabilize variance, facilitate prediction result when the data is skewed, and resuce the impact of outliers. Afterwards, the data is transformed back in order to interpret the results in the original scale for practical application and real-world understanding.

The resulting figure and summary of actual vs predicted median house value after performing log transformation is displayed in Figure 5 below, and Table 3 illustrates the results of linear regression after log transformation.

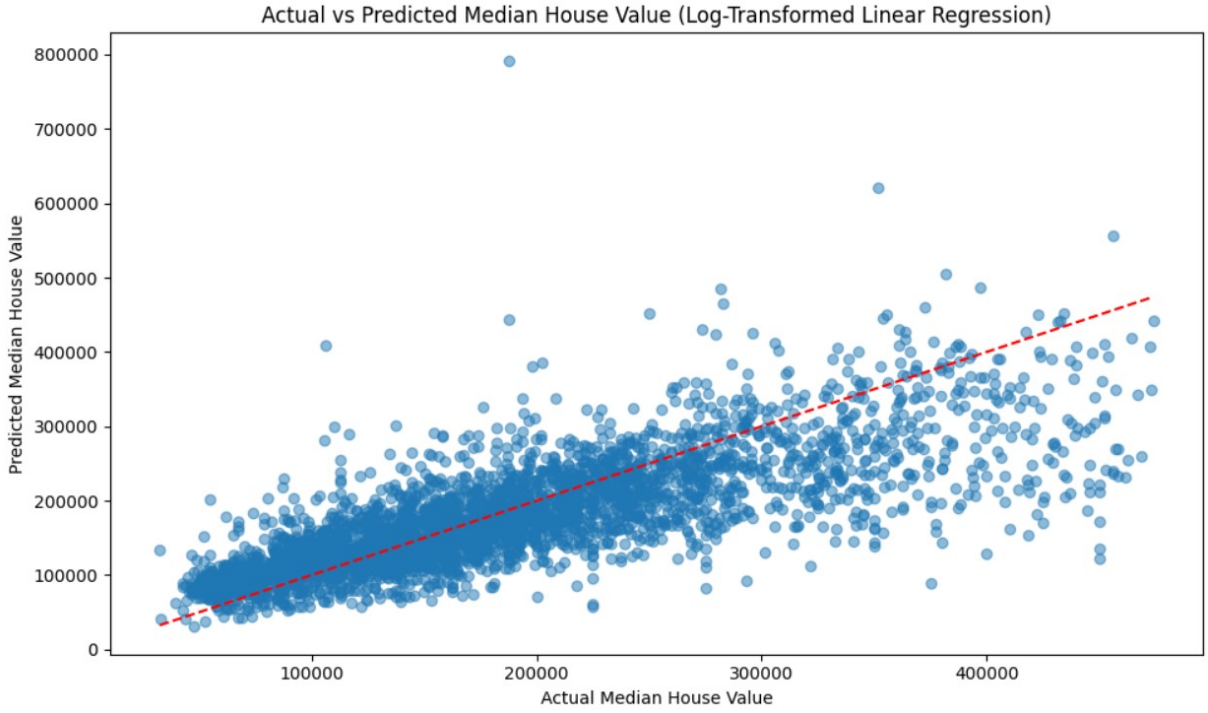

Figure 5: Actual vs predicted house value Multiple Linear Regression after log transformation. (Photo/Picture credit: Original)

Table 3: Summary of Multiple Linear Regression after log transformation.

	Variable	Coefficient	Std. Error	t-value	$P > \|t\|$	[0.025	0.975]
1	Constant	-9.2	0.34	-27.05	3.75e-157	-9.86	-8.53
2	Longitude	-0.24	0.0039	-62.44	0	-0.25	-0.23
3	Latitude	-0.25	0.0037	-66.97	0	-0.25	-0.24
4	Housing median age	0.002783	0.00026	10.7	1.22e-26	0.0023	0.0033
5	Total rooms	-0.000101	6e-06	-16.91	1.55e-63	-0.00011	-9-05
6	Total bedrooms	0.00088	5.3e-05	19.66	5.88e-85	0.00083	0.00093
7	Population	-0.0002	7e-06	-26.23	4.98e-148	-0.00021	-.00018
8	Households	0.000214	5e-05	4.26	2.01e-05	0.00012	0.00031
9	Median income	0.228	0.0026	88.89	0	0.22	0.23

DOI: 10.54254/2754-1169/94/2024OX0208

3.3.3. Random Forest

To evaluate the model accuracy, this study implements a second model random forest. Random forest is an ensemble learning method that build multiple decision trees during training process, it then merges their output into the model to further improve the accuracy and robustness of predictions. Implementing random forest instead of multiple linear regression helps to capture more complex and non-linear relationships between the predictors and the median house value. On top of that, random forests can also handle datasets with more dimensions by automatically manage interactions between features, therefore theoretically improves the model's prediction accuracy.

The resulting figure of actual vs predicted median house value after implementing random forest is displayed in Figure 6 below.

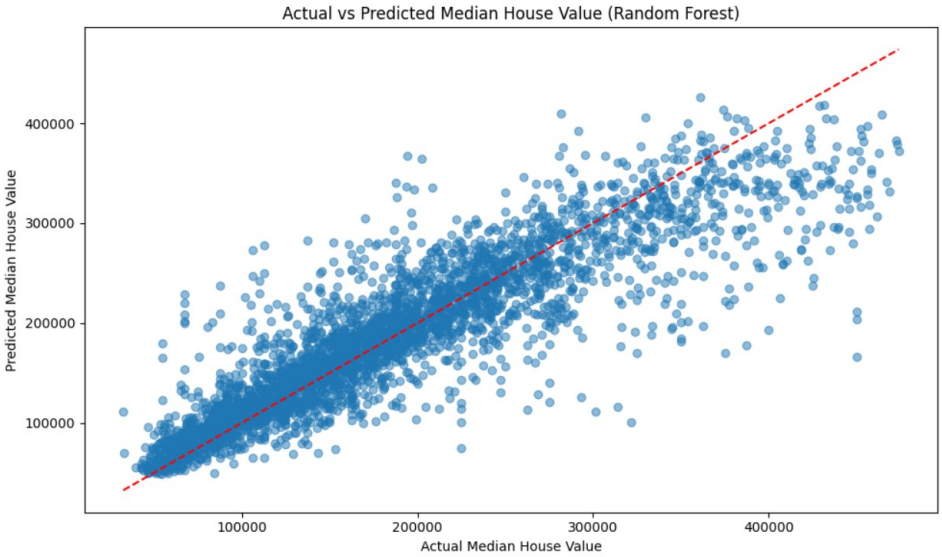

Figure 6: Actual vs predicted house value Random Forest (Photo/Picture credit: Original).

The chosen parameter for random forest:

$$max_{depth} = 30, Max\_features = None, \min\ \_samples\_leaf = 4, \min\ \_sample\_split = 10,$$
$$n_{estimators} = 400, rabdon_{State} = 42 \quad (6)$$

3.3.4. Random Forest Hyperparameter Tuned

A hyperparameter is a configuration that is set before the learning process begins, including the number of trees in a random forest or the maximum depth of each tree, these variables controls the model's complexity and behavior. Hyperparameter tunes a model systematically searching and optimizing the setting in order to enhance a model's performance, namely on improving accuracy and reduce overfitting. Implementing hyperparameter would tune the random forest to perform the prediction in a manner that maximize its performance on the data set, ensuring the model generalizes well to unseen data. This study implemented both randomized search as well as grid search.

The resulting figure of actual vs predicted median house value using random forest after implementing randomized and grid search is displayed in Figure 7 below.

DOI: 10.54254/2754-1169/94/2024OX0208

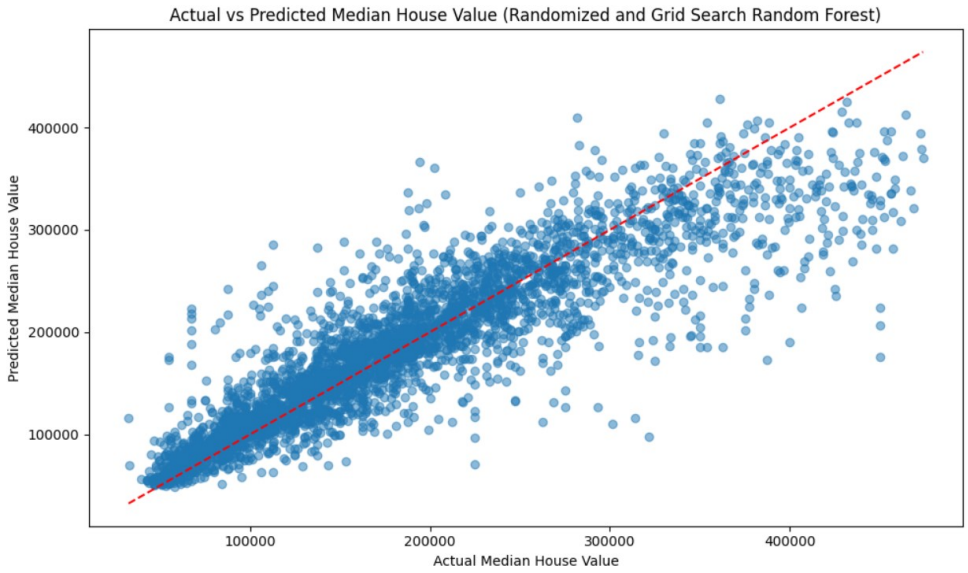

Figure 7: Actual vs predicted house value Random Forest after hyperparameter tuning. (Photo/Picture credit: Original)

After randomized and grid search, the best setting for this dataset is:

$$\max\ \_depth = 40, Max\_features = None, \min\ \_samples\_leaf = 2, \min\ \_sample\_split = 5, \quad n\_estimators = 500 \tag{7}$$

4. Results

The comprehensive comparison of results from different models, before and after the implementation of enhancement operations are displayed below in Table 4. To interpret the result section, evaluation criteria need to be understood.

- RMSE (Root mean squared error) is a quantified measure of average magnitude of prediction errors between the actual and predicted house prices and is used as a measure of model accuracy, lower RSME values indicate more precise predictions
- MAE (Mean absolute error) measures the average absolute difference between actual and predicted house prices, lower MAE values indicate more precise predictions
- R^2 (R-squared), which is also called coefficient of determination, is the proportion of variation in the dependent variable (house prices) can be explained by the independent variables (all other features), r sqaured value determine the reliability of a model, the higher the r squared value the better
- Accuracy describes the percentage of correctly predicted house values within the testing dataset, where a higher accuracy reflects a more reliable model as well as more accurate predictions

Table 4: Summary of results.

Metric	Multiple Linear regression	Multiple Linear Regression Log-transformed Value	Random Forest Value	Random Forest after Hyperparameter Tuning Value
RMSE	57969.99	58981.44	41895.93	41617.79
MAE	43016.38	41146.89	27912.11	27707.55
R^2	0.6040	0.5901	0.7932	0.7959
Accuracy (%)	85.24	83.75	84.32	86.53

DOI: 10.54254/2754-1169/94/2024OX0208

As Table 4 shown, the overall results of this study demonstrate the effectiveness of the above-mentioned machine learning models in predicting median house values within a block using various block related information. By applying multiple linear regression, log-transformed linear regression, standardized linear regression, and random forest models (with and without hyperparameter tuning), varying levels of prediction accuracy was achieved. It is reasonable to be concluded that the accuracy of results remain unfluctuating, however, there is significant difference between models observed in the rest of the indicators. Random forest has proven to be a more reliable model comparing to multiple linear regression as there is a noticeable drop in error indicators, namely 28.59% decrease in average RMSE and 52.31% decrease in average MAE. In addition, random forest model demonstrated superior performance on R-squared, achieving an astonishing result of 79.46% in average comparing to the 59.71% achieved by multiple linear regression model.

5. Conclusion

The application of machine learning techniques, particularly the random forest model with hyperparameter tuning, demonstrated superior predictive accuracy and robustness. This validates the potential of using such models to assess property values objectively, aiding buyers in making informed decisions. However, while the model of choice performed fairly well in this study, the objective limitation to the study cannot be overlooked.

The first limitation is the potential generalizability issue stemming from the dataset. As the dataset used in the study consists of only 17647 rows of block information solely acquired from California census bureau, further action to validate the real life application of the chosen model is required. Additional study should further examine the model performance when implementing to a larger dataset, a dataset with aggregated data from multiple regions, or both. The second limitation lies in the exclusion of temporal dynamics and market fluctuations in the predicted price. Although the intrinsic value defined in the study aims at eliminating all subjective factors, the timely variability of the price itself is not addressed under the current structure of study. The influence brought by temporal factors such as economic cycles, seasonal trends, interest rates, and policy changes should be addressed. Future research should investigate the possibility to incorporate such factors into the model using time-series data such as significant market trends, economic indicators and historical price changes as a backbone support to establish a more comprehensive prediction model.

Moreover, enhancements to the current model can be performed in the future after addressing the above limitations. Future researchers are encouraged to explore the effect on prediction accuracy by integrating more advanced machine learning models, such as Gradient Boosting Machines, XGBoost, and deep learning models. These models have shown desired capabilities such as better handling of non-linear relationships between features as well as better adaptability to larger and more diverse datasets, potentially further enhance robustness and reliability.

Overall, the models within this study effectively explain the data and offer a practical tool for evaluating property prices in the real estate market. In addition, the model demonstrate potential for further refinement by addressing the identified limitations and enhancing the current model, making them even more reliable and applicable in various contexts. This study lays the groundwork for ongoing improvements in the objective assessment of property values, benefiting future buyers, investors, and stakeholders in the real estate industry.

References

[1] *M. Thamarai, S P. Malarvizhi, "House Price Prediction Modeling Using Machine Learning," International Journal of Information Engineering and Electronic Business (IJIEEB), Vol.12, No.2, pp. 15-20, 2020. DOI: 10.5815/ijieeb.2020.02.03.*

DOI: 10.54254/2754-1169/94/2024OX0208

[2] *Phan T.D., "Housing Price Prediction Using Machine Learning Algorithms: The Case of Melbourne City, Australia," 2018 International Conference on Machine Learning and Data Engineering (ICMLDE), 2018. DOI: 10.1109/icmlde.2018.00017.*

[3] *Fan C., Cui Z., Zhong X., "House Prices Prediction with Machine Learning Algorithms," Proceedings of the 2018 10th International Conference on Machine Learning and Computing (ICMLC 2018), DOI: 10.1145/3195106.3195133.*

[4] *Mu J., Wu F., Zhang A., "Housing Value Forecasting Based on Machine Learning Methods," Abstract and Applied Analysis, 2014; 2014:1–7. DOI: 10.1155/2014/648047.*

[5] *Quang Truong, Minh Nguyen, Hy Dang, Bo Mei, "Housing Price Prediction via Improved Machine Learning Techniques," Procedia Computer Science, Vol.174, pp. 433-442, 2020. DOI: 10.1016/j.procs.2020.06.111.*

[6] *Lu S., Li Z., Qin Z., Yang X., Goh R.S.M., "A hybrid regression technique for house prices prediction," 2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), 2017. DOI: 10.1109/ieem.2017.8289904.*

[7] *George Milunovich. Forecasting Australia's real house price index: A comparison of time series and machine learning methods. International Journal of Forecasting, 36(1):100-112, 2020. DOI: 10.1002/for.2678.*

[8] *Sanjar Karshiev, Bekhzod Olimov, Jaesoo Kim, Anand Paul, and Jeonghong Kim. Missing data imputation for geolocation-based price prediction using KNN–MCF method. ISPRS International Journal of Geo-Information, 9(4):227, 2020. DOI: 10.3390/ijgi9040227.*

[9] *D. Tchuente and S. Nyawa. Real estate price estimation in French cities using geocoding and machine learning. Annals of Operations Research, 308:571–608, 2022. DOI: 10.1007/s10479-021-03932-5.*

[10] *W. T. Al-Sit and R. Al-Hamadin. Real estate market data analysis and prediction based on minor advertisements data and locations' geo-codes. International Journal of Advanced Trends in Computer Science and Engineering, 9(3):4077-4089, 2020. DOI: 10.30534/ijatcse/2020/235932020.*