Ahliya Journal of Business Technology and MEAN Economies [Published by Palestine Ahliya University]
Vol. 03 Issue 01 (2026)
ISSN: 3007-9691

California Housing Price Prediction Using Machine Learning: A Comparative Study Using Feature Engineering and Ensemble Methods

Mohammad Siwad[1], Mutaz Rasmi Abu Sara[2] and Ahlam Awwad[2]
[1] Palestine Ahliya University (Palestine)
✉ Mohd.siwad25@gmail.com
[2] Faculty of Engineering and Information Technology, Palestine Ahliya University (Palestine)
✉ moutaz.a@paluniv.edu.ps
✉ ahlam.awwad@paluniv.edu.ps

Received:13/03/2026 Accepted:26/04/2026 Published:31/05/2026

Abstract:

This study aimed to develop an accurate housing price prediction model for California by comparing multiple machines learning algorithms and evaluating the impact of feature engineering and ensemble techniques on predictive performance. The study utilized the California Housing Dataset, comprising 20,433 observations after data cleaning and preprocessing. A five-stage methodology was implemented, including the evaluation of seven baseline regression models, generation of engineered features, feature selection using the F-statistic method, hyperparameter optimization of the best-performing models through GridSearchCV, and the construction of Voting and Stacking ensemble models. The findings revealed that linear models achieved limited performance due to the nonlinear relationships among variables, whereas tree-based and ensemble methods demonstrated superior predictive capabilities. The Stacking Ensemble model achieved the highest performance with an R^2 value of 0.8431, an RMSE of $46,317, and an MAE of $30,150. Furthermore, the results confirmed that engineered features, particularly rooms per household, played a significant role in enhancing prediction accuracy. The scientific contribution of this study lies in proposing an integrated framework that combines feature engineering, feature selection, hyperparameter optimization, and advanced ensemble learning within a unified comparative environment. This approach improved predictive performance and surpassed the widely recognized benchmark by 3.21 percentage points in terms of R^2, while highlighting the importance of household-level feature normalization in housing price prediction.

Keywords: *California Housing Dataset; Ensemble Learning; Feature Engineering; Housing Price Prediction; Hyperparameter Optimization; Machine Learning.*

Ahliya Journal of Business Technology and MEAN Economies Vol. 03 Issue 01 (2026)
[Published by Palestine Ahliya University] ISSN: 3007-9691

التنبؤ بأسعار المساكن في كاليفورنيا باستخدام التعلم الآلي: دراسة مقارنة باستخدام هندسة الخصائص وأساليب التجميع

محمد سواد[1]، معتز رسمي أبو سارة[2]، أحلام عواد[2]

[1] جامعة أهلية فلسطين (فلسطين)

Mohd.siwad25@gmail.com ✉

[2] كلية الهندسة وتكنولوجيا المعلومات، جامعة أهلية فلسطين (فلسطين)

moutaz.a@paluniv.edu.ps ✉

ahlam.awwad@paluniv.edu.ps ✉

تاريخ الاستلام:2026/03/13 تاريخ القبول:2026/04/26 تاريخ النشر:2026/05/31

ملخص:

هدفت هذه الدراسة إلى تطوير نموذج دقيق للتنبؤ بأسعار المساكن في ولاية كاليفورنيا من خلال مقارنة مجموعة من خوارزميات التعلم الآلي وتقييم أثر هندسة الخصائص وتقنيات التجميع على أداء النماذج التنبؤية. اعتمدت الدراسة على بيانات (California Housing Dataset) التي تضم 20,433 سجلاً بعد معالجة البيانات وتنظيفها. وتم تطبيق منهجية مكونة من خمس مراحل شملت: تقييم سبعة نماذج انحدار أساسية، وإنشاء خصائص جديدة مشتقة، واختيار الخصائص الأكثر تأثيراً باستخدام اختبار (F-statistic)، وضبط المعلمات الفائقة لأفضل النماذج باستخدام (GridSearchCV)، ثم بناء نماذج تجميعية من نوع (Voting) و(Stacking). أظهرت النتائج أن النماذج الخطية حققت أداءً محدوداً بسبب الطبيعة غير الخطية للعلاقات بين المتغيرات، في حين تفوقت نماذج الأشجار والتجميع. وحقق نموذج (Stacking Ensemble) أفضل أداء بقيمة معامل تحديد بلغت (R^2=0.8431)، وخطأ جذر متوسط التربيع (RMSE=46,317 دولاراً)، ومتوسط الخطأ المطلق (MAE=30,150 دولاراً). كما أثبتت النتائج أن الخصائص المشتقة، وخاصة عدد الغرف لكل أسرة، أسهمت بصورة جوهرية في تحسين دقة التنبؤ. تتمثل الأصالة العلمية لهذه الدراسة في تقديم إطار متكامل يجمع بين هندسة الخصائص واختيارها وضبط المعلمات الفائقة وتقنيات التجميع المتقدمة ضمن بيئة موحدة للمقارنة، مما أسهم في تحسين الأداء التنبؤي وتجاوز المعيار المرجعي الشائع للدراسة بنسبة 3.21% في قيمة معامل التحديد، إضافة إلى إبراز أهمية تطبيع البيانات على مستوى الأسر في تحسين نماذج التنبؤ العقاري.

الكلمات المفتاحية: *مجموعة بيانات الإسكان في كاليفورنيا؛ التعلم الجماعي؛ هندسة الميزات؛ التنبؤ بأسعار المنازل؛ تحسين المعلمات الفائقة؛ التعلم الآلي.*

1. Introduction

Evaluation of properties and getting good results is not a simple task; The price heavily depends on location, structure, neighborhood, income, and proximity to local services. These features and the relationship between them is non-linear, which makes traditional methods have a hard time capturing them. It makes the process slow, expensive, and most of the time subjective when it comes to making an accurate prediction. The Machine learning methods that are used in this study offer a better alternative. With sufficient data, we can capture these non-linear relationships and build models that can estimate prices accurately without any manual intervention.

The California Housing Dataset (1997), which has revolutionized regression benchmarks, includes real house prices with multiple relevant features: geographical location, the age of the house, room count, approximate income, and the median house value as the target variable. Despite the history of this dataset and the amount of research, it is still limited due to the studies not utilizing robust feature engineering, feature selection, and ensemble stacking.

This study aims to address these issues with a resilient, five stage methodology: evaluation of seven regression models, construction of new features, feature selection of the 16 features in the dataset, hyperparameter tuning for the two best-performing models, and implementing Stacking ensemble models. The main contributions of this study incorporate a comparative analysis for multiple models under identical circumstances. Household and the interaction feature that are derived from aggregates (Géron, 2022), F-statistic feature selection, and hyperparameter tuning for Random Forest (Breiman, 2001). Afterwards, we will employ Gradient Boosting (Friedman, 2001) while utilizing the function GridSearchCV (Wu et al., 2019). Finally, voting and stacking ensembles ensure we provide a quality pipeline with good results and no subjectivity. (Pedregosa et al., 2011). The Stacking Ensemble yields $R^2 = 0.8431$, exceeding the benchmark of $R^2 = 0.8110$ (Géron, 2022) by 3.21 percentage points.

2. Related Work

The California Housing dataset was originally introduced as sparse recursive inputs. Pace and Barry's (1997) inventive objective was aimed at statistical modeling rather than prediction. Despite that fact, due to the simplicity of the dataset, it quickly turned into one of the most popular regression benchmarks.

On the other hand, Géron (2022) widely popularized this dataset by using machine learning. In his textbook, the Random Forest model achieved a groundbreaking R^2 of approximately 0.81. This outcome became the standard against which subsequent methods are compared. The most vital aspect of his approach included using feature normalization per household. He assembled new ratios such as bedrooms per house, and residents per household, which is the basis for the feature engineering that we will use in this paper.

Polanitzer (2022) used a similar approach but with different techniques on the same dataset. He achieved an R^2 score of 0.808 using Gradient Boosting, and 0.812 with RandomForest. These results are compiled without the use of any feature engineering or hyperparameter tuning. This is a confirmation of our previous statements that unoptimized methods do not achieve competent results and have a similar ceiling that cannot be broken.

A more recent study drove the ceiling higher by applying advanced algorithms instead of traditional ones. Sharma et al. (2024) proved the superiority of the XGBoost model over other simpler methods for predicting house prices. This highlights the aim of this paper, which is to show the importance of feature engineering and hyperparameter tuning for achieving the best accuracy possible.

Breiman (2001) and Friedman (2001) demonstrated the gains that come from ensemble models, which almost always outperform simpler models on prediction tasks. In their paper, they also highlight the significance of hyperparameter tuning, which is vital for boosting accuracy and building a better model, especially for gradient boosting.

3. Methodology

3.1 Dataset Description

The California Housing Dataset contains 20,640 inputs taken from the 1990 U.S. Census. We removed 207 observations that are labeled as missing in the feature total-bedrooms. After the cleaning process, 20,433 observations remain. The range of the independent variable, median_house_value starts at 15000 dollars with a max value of 500000 dollars. The mean of the house value prices is roughly 206000 dollars. Table 1 below summarizes all the features in the dataset.

Table 1. Dataset Features and Descriptions

Feature	Type	Description
Longitude	Continuous	Longitude coordinate of the district centroid
Latitude	Continuous	Latitude coordinate of the district centroid
housing_median_age	Continuous	Median age of houses within the block
total_rooms	Continuous	Total number of rooms in the block
total_bedrooms	Continuous	Total number of bedrooms in the block
Population	Continuous	Total population within the block
households	Continuous	Total number of households in the block
median_income	Continuous	Median household income (tens of thousands $)
ocean_proximity	Categorical	Location relative to the ocean (5 categories)
median_house_value (Target)	Continuous	Median home value for households in block ($)

The variable we want to predict is skewed to the right, and we observe a rise at the value of $500,001, reflecting the maximum limit imposed by the census on reported values; that is, it is the maximum price. Ocean proximity is dominated by the "<1H OCEAN" category. Median income is approximately normally distributed, peaking around $3–5 (in tens of thousands). These distributional characteristics are shown in Figure 1.

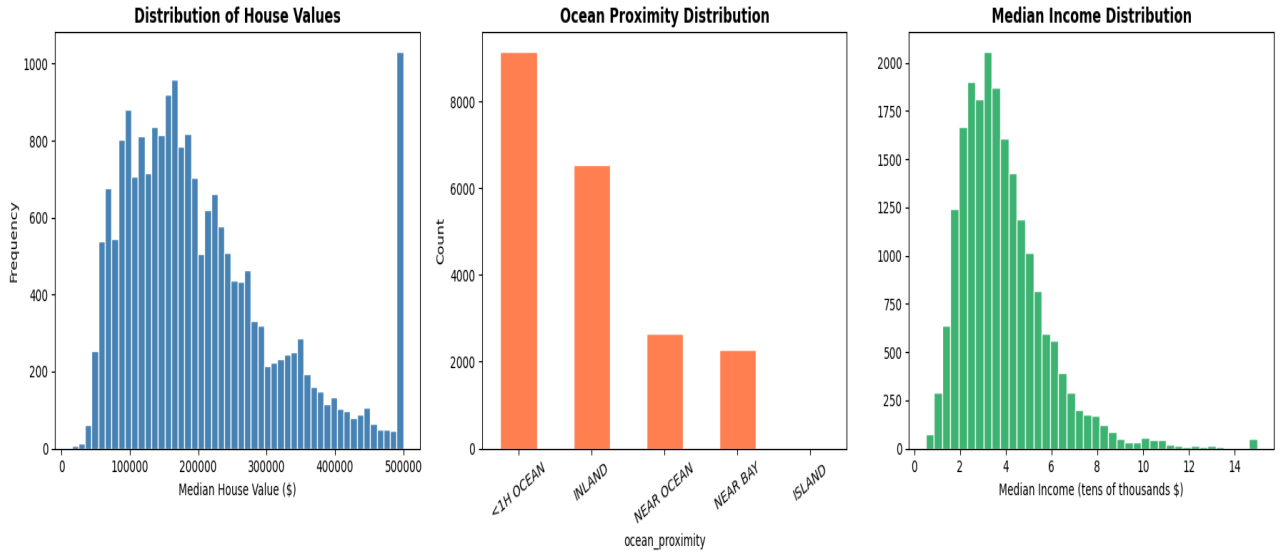

Figure 1. Distribution of house values, ocean proximity categories, and median income.

3.2 Data Preprocessing

The 207 records with missing total_bedrooms values were removed, yielding a clean dataset of 20,433 observations. The categorical feature ocean_proximity was converted to integer codes via label encoding. This choice is appropriate because all top-performing models in this study are tree-based and are therefore invariant to the ordinal assumptions implied by label encoding (Pedregosa et al., 2011). All numeric features were standardized to zero mean and unit variance. Standardization is required for distance-sensitive models (KNN) and regularization-based models (Ridge, Lasso), which penalize features according to their scale.

3.3 Exploratory Data Analysis

Prior to modeling, we examined feature correlations using the Pearson correlation matrix (Figure 2). Median_income is the only feature that correlates strongly with the target (r = 0.69), consistent with the established importance of income in housing markets. Total_rooms and households are highly collinear (r = 0.92), reflecting that both largely capture district size.

As Géron (2022) also published, standardizing these statistics and allocating them according to the number of households in detail leads to obtaining predictive indicators with greater significance than if the statistics were general for the region as a whole.

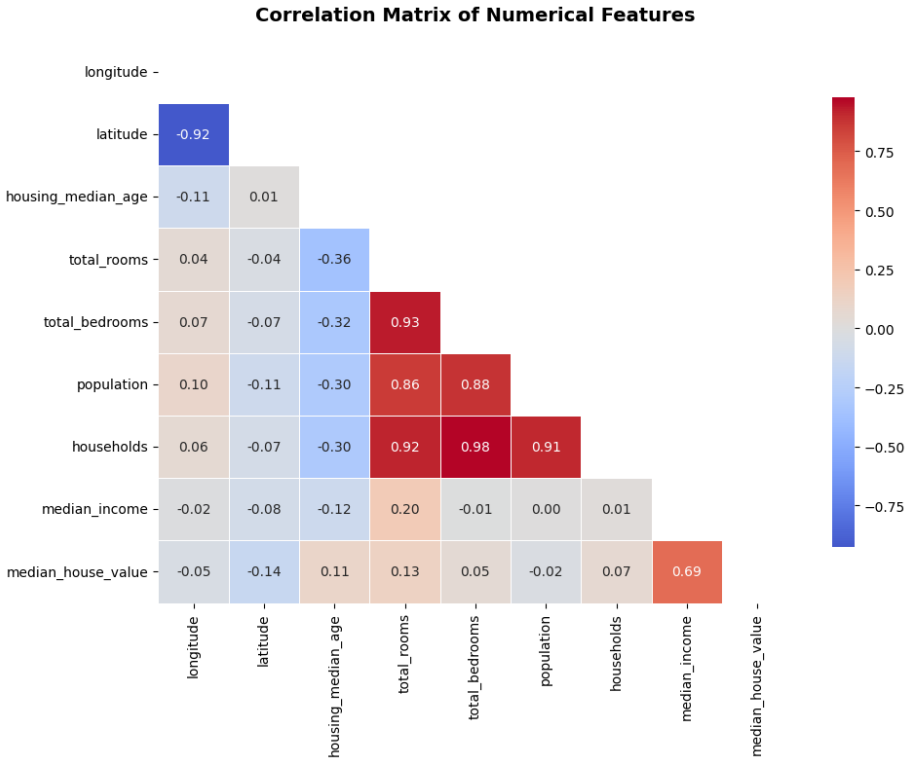

Figure 2. Correlation matrix of numerical features.

In Figure 3, looking at the left side, we see the geographical distribution of home values. We observe that homes with high prices are concentrated in neighborhoods around the San Francisco Bay Area and Los Angeles. At the same figure on the right, we see a direct relationship; there is a significant disparity with higher income levels. We conclude that income alone cannot fully explain the variation in home prices.

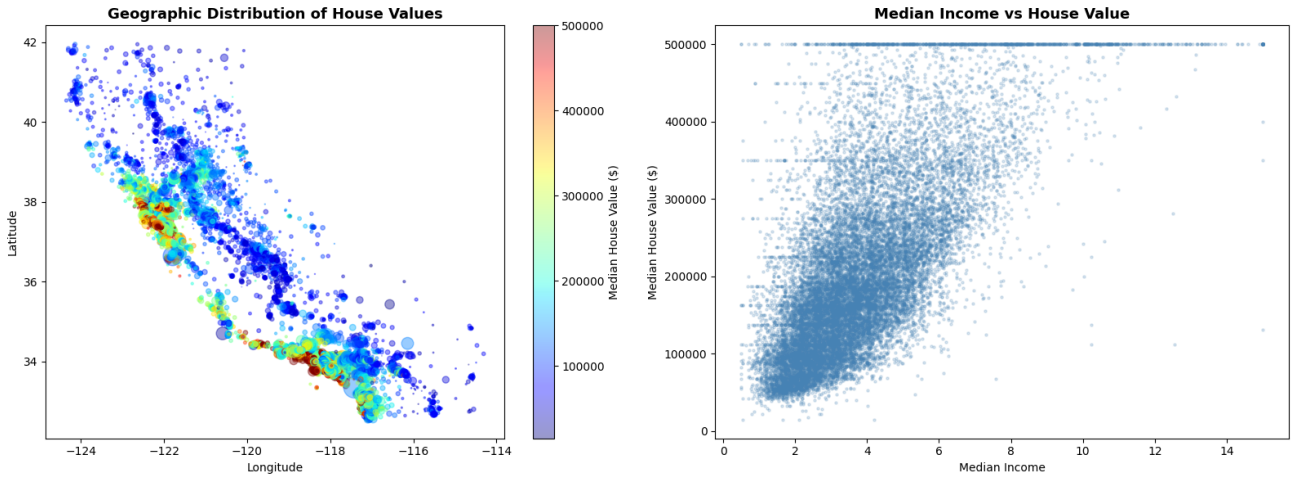

Figure 3. Geographic distribution of house values on the left. median income vs house value on the right

3.4 Feature Engineering

The initial dataset contained the total number of rooms and population at the neighborhood level without allocation. For example, a neighborhood with 10,000 rooms and 5,000 households differs substantially from another neighborhood with 10,000 rooms and 500 households, even though the first and second neighborhoods have the same total number of rooms. Following the standardization methodology for each household described by Géron (2022), we derived 7 new characteristics from the original characteristics, which were as follows:

- **rooms_per_household** = total_rooms / households: average dwelling size, a direct indicator of housing space per family.
- **bedrooms_per_room** = total_bedrooms / total_rooms: the bedroom fraction, distinguishing apartment-style from house-style stock.
- **population_per_household** = population / households: average occupancy, a proxy for crowding.
- **rooms_per_person** = total_rooms / population: space per resident.
- **income_x_age** = median_income × housing_median_age: an interaction capturing the joint effect of neighborhood wealth and housing vintage.
- **bedroom_ratio** = total_bedrooms / households: bedrooms per household, complementary to rooms_per_household.
- **income_per_room** = median_income / total_rooms: an affordability index relating income to room supply.

These additions expand the feature space from 9 to 16 variables.

3.5 Feature Selection

SelectKBest with the F-statistic (f_regression) was applied to all 16 features to score their linear association with the target. The top 12 features were retained. Of the seven engineered features, three were selected—rooms_per_household, bedrooms_per_room, and income_x_age—confirming that per-household normalization captures information absent from the raw aggregates (Géron, 2022). Figure 4 shows Random Forest feature importances across all 16 features.

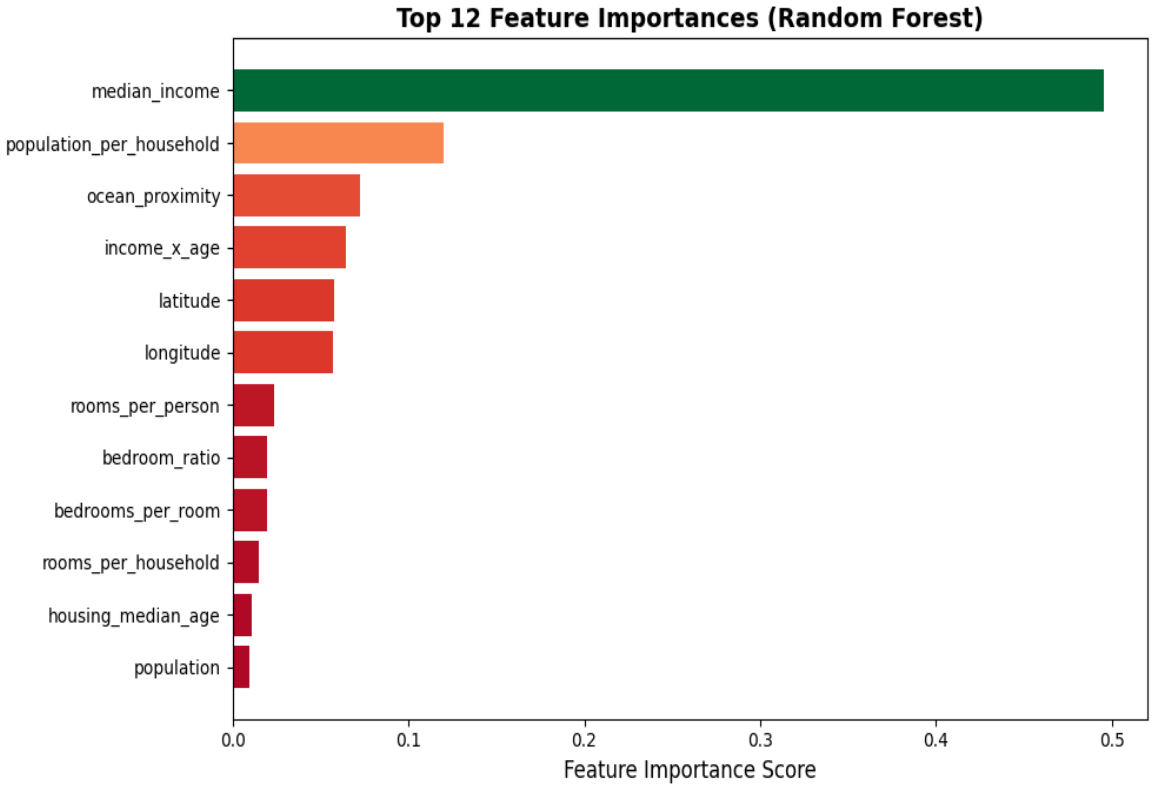

Figure 4. Random Forest feature importances — top 12 features retained for model training.

3.6 Model Training and Evaluation

Data were split 80/20 into training (16,346 samples) and test (4,087 samples) sets. Seven baseline regressors were trained: Linear Regression, Ridge, Lasso, KNN, Decision Tree, Random Forest (Breiman, 2001), and Gradient Boosting (Friedman, 2001). Each model was evaluated on R^2, RMSE, MAE, and 5-fold cross-validated R^2.

The two best-performing baselines were further optimized with GridSearchCV and 5-fold cross-validation (Wu et al., 2019). The search grids are detailed in Table 2.

Table 2. Hyperparameter Search Grids for the Two Best Models

Model	Parameter	Values Searched
Random Forest	n_estimators	[200, 300]
	max_depth	[None, 20, 30]
	min_samples_split	[2, 5]
	min_samples_leaf	[1, 2]
Gradient Boosting	n_estimators	[200, 300]
	max_depth	[4, 5, 6]
	learning_rate	[0.05, 0.1]
	subsample	[0.8, 1.0]

A Voting Regressor was built from the four tuned models (RF, GB, Ridge, KNN). A Stacking Regressor used the same four as base learners, with Linear Regression as the meta-model, implemented via scikit-learn (Pedregosa et al., 2011).

4. Results and Discussion

Table 3 reports performance across all eleven models.

Table 3. Comparative Performance of All Evaluated Models

Model	R^2	RMSE ($)	MAE ($)	CV R^2
Linear Regression	0.6593	68,262	49,409	0.6481
Ridge	0.6592	68,264	49,408	0.6481
Lasso	0.6592	68,263	49,409	0.6481
KNN	0.7405	59,569	39,640	0.7253
Decision Tree	0.6596	68,224	43,375	0.6281
Random Forest	0.8196	49,674	32,027	0.8135
Gradient Boosting	0.7885	53,785	37,217	0.7855
RF (Tuned)	0.8167	50,063	32,892	—
GB (Tuned)	0.8420	46,488	30,436	—
Voting Ensemble	0.8051	51,625	34,681	—
Stacking Ensemble	**0.8431**	**46,317**	**30,150**	—

Linear models (Linear Regression, Ridge, Lasso) plateau at $R^2 \approx 0.66$, demonstrating that the dataset's relationships are fundamentally nonlinear. The KNN algorithm reached 0.74 and of course it suffered from the well-known problem of increasing dimensions in the expanded feature space compared to the random forest algorithm (Breiman, 2001). It is considered a strong benchmark at $R^2 = 0.8196$, exceeding the Géron benchmark (2022) which reached 0.8110.

Adjusting the hyperparameters leads to vastly and noticeably different results for the two best models. Gradient Boosting (Friedman, 2001) improves substantially from $R^2 = 0.7885$ to 0.8420 with optimal settings (n_estimators=300, max_depth=6, learning_rate=0.1, subsample=0.8)—a gain of 5.35 percentage points. However, the performance of the random forest algorithm drops slightly from 0.8196 to 0.8167 after we adjusted it. This indicates that the default settings of the algorithm were better and achieved the best level for this set of data.

The Stacking Ensemble delivers the best overall performance: $R^2 = 0.8431$, RMSE = $46,317, MAE = $30,150. The superlearner assigns different weights to the underlying models across different feature space regions; for example, it favors gradient enhancement in some regions and a random forest in others. This diversity is precisely what makes the clustering more robust than the classic group voting model ($R^2 = 0.8051$), which simply averages the predictions.

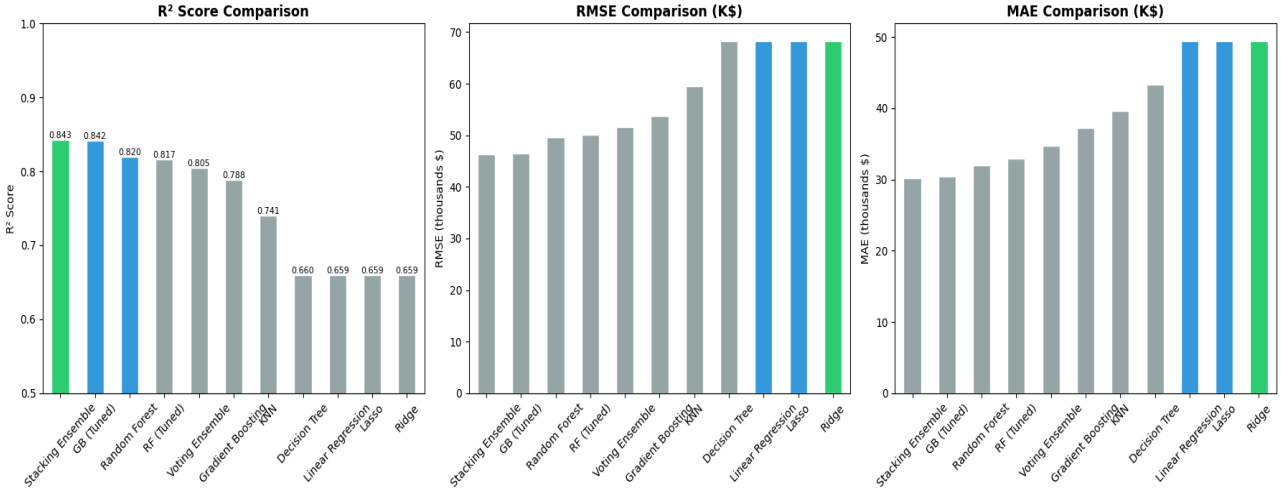

Figure 5. Performance comparison of all models across: R^2,RMSE،MAE.

Figure 6 illustrates the plot of actual values compared to expected values and the distribution of residuals for the pooled group. We observe that the expectations are in close agreement with the observed values across most of the price range. The right tail of the residuals shows a systematic undervaluation of properties valued at less than $500,001. This is a known limitation of census data and not a flaw in our model. Otherwise, the distribution of residuals is nearly symmetrical at zero, indicating that the model is unbiased.

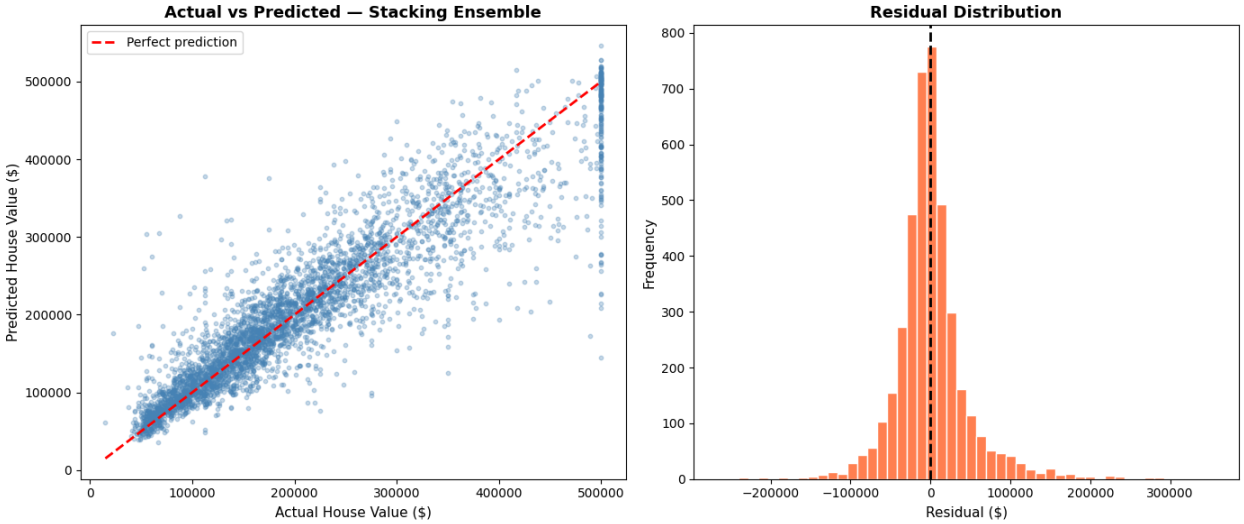

Figure 6. Actual vs predicted values. and residual distribution for the Stacking Ensemble.

Through engineered features, the number of rooms per household ranked second in the random forest algorithm, its importance primarily following average income. This underscores the value of standardizing data for each household (Géron, 2022). The total number of rooms (total number of rooms, total number of bedrooms) ranked considerably lower, confirming that normalization reveals housing density information that aggregation obscures.

6. Conclusion

Based on the research done in this paper, we assessed eleven regression models while using a robust pipeline of feature engineering, F-statistic feature selection, GridsearchCV tuning, and ensemble stacking. These methods proved to increase the accuracy and scalability of our model.

The Stacking Ensemble achieved $R^2 = 0.8431$, RMSE = $46,317, and MAE = $30,150 using only scikit-learn (Pedregosa et al., 2011)—a 3.21 percentage-point improvement over the Géron (2022) benchmark of $R^2 = 0.8110$.

Out of all the results, three outcomes stand out; First, feature normalization for households was the top contributor for increasing the performance of the models, specifically rooms_per_household, which acted as an outstanding predictor for the dataset. Second, Hyperparameter tuning was especially effective for Gradient Boosting, unlike the RandomForest method. Although, Gradient Boosting requires vigorous hyperparameter tuning due to it requiring optimal configuration to get good results. Third, stacking surpasses voting because it uses multiple base models and turns it into one robust model rather than taking a fixed average for the whole dataset.

When it comes to future work, we need to explore three approaches: utilizing log transformation to diminish the impact of the high ceiling of prices (500000 dollars). Second, we need to add more spatial features, such as the distance to local services like banks or schools to capture the correlation of geographical location and price. Lastly, we apply SHAP analysis so we can better interpret the prices and features so that companies can deploy them.

References

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc.".

Pace, R. K., & Barry, R. (1997). Sparse spatial autoregressions. *Statistics & Probability Letters*, *33*(3), 291-297.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Polanitzer, R. (2022, March 12). *Machine learning for California housing*. Medium. https://medium.com/

Sharma, H., Harsora, H., & Ogunleye, B. (2024). An optimal house price prediction algorithm: XGBoost. *Analytics*, *3*(1), 30-45.

Wu, J., Chen, X. Y., Zhang, H., Xiong, L. D., Lei, H., & Deng, S. H. (2019). Hyperparameter optimization for machine learning models based on Bayesian optimization. *Journal of Electronic Science and Technology*, *17*(1), 26-40.