Rent Price Prediction with Advanced Machine Learning Methods: A Comparison of California and Texas

Hong Ru Chan *

Department of Statistics and Applied Probability, University of California, Santa Barbara, United States

* Corresponding Author Email: hchan@ucsb.edu

Abstract. The forecast of rent prices in dynamic housing markets is of fundamental importance to renters, landlords, investors, and politicians alike. Machine learning models offer flexibility, excel at modeling complex relationships, and provide outstanding forecast precision. This study compares advanced machine learning models, extreme gradient boosting regressor (XGBoost), light gradient boosting machine (LightGBM), random forest, ridge regression, and two ensemble approaches, to predict California and Texas rent prices. The two ensemble approaches include a hybrid regression of averaging base models and a 2-level stacked generalization model. The results revealed that a stacked generalization ensemble with base models random forest, XGBoost, LightGBM, and meta-model ridge regression achieved the best performance for the California dataset by generating the lowest MSE and highest R^2 of 46116.3 and 0.8858, respectively. In contrast, random forest outperformed both ensemble models with the lowest MSE and MAPE of 18401.93 and 9.7003%, respectively, and the highest R^2 of 0.7992. These methodologies can assess future rental property worth, serve as indicators for market dynamics, and aid in establishing real estate policies, thereby providing practical guidance to individuals, businesses, and policymakers.

Keywords: Rental price prediction; real estate; machine learning; stacked generalization; hybrid regression.

1. Introduction

The analysis and prediction of rent prices are significant in interpreting the housing dynamics and socioeconomic developments of an area. Rent prices give crucial insights into housing affordability, market trends, and prospective economic shifts, enabling individuals, businesses, and politicians to make educated decisions. For tenants and landlords, rent prediction enables more transparent negotiations in lease agreements. For property investors and developers, accurate rent prediction allows them to decide on proper property acquisition and development strategies to benefit from rental rates. For policymakers, rent prediction aids urban planning and housing policy development, as housing affordability issues, regulations, taxation, and development incentives can be addressed to work towards maintaining social equilibrium.

The most well-known current model in the real estate industry is the Hedonic Price Model (HPM) [1]. In this model, the price of a building or plot of land is influenced by property size, location, and amenities. However, traditional methodologies frequently fall short of reflecting the complexities of current housing markets, which are characterized by complicated linkages and non-linear dynamics. Their dependence on simplified linear models may result in erroneous predictions, especially when complex relationships and diverse data are present. Machine learning models thus appear as a viable option to solve these limitations, as they have adaptive learning abilities and the capacity to capture nuanced interactions among data.

Machine learning models have been implemented in the realm of real estate. Singh et al. predicted property sale prices in Iowa by utilizing three models: linear regression, random forest, and gradient boosting. The results demonstrated that the gradient boosting model surpassed other forecasting models in terms of predicting accuracy [2]. Abdul-Rahman et al. compared the accuracy of light gradient boosting machine (LightGBM)and extreme gradient boosting regressor (XGBoost) with multiple regression and ridge regression in forecasting real estate value in Malaysia, with XGBoost outperforming all other models [3]. Xu and Li included text descriptions using natural language

processing technology in their model and used GIS to examine the latitude and longitude coordinates of second-hand housing in China [4]. Akyüz et al conducted a hybrid approach that employs linear regression, clustering analysis, nearest neighbor classification, and support vector regression (SVR) on the Kadıköy district in Istanbul and Kaggle dataset [5].

This research focuses on California and Texas, two of the most economically significant states in the United States, due to their economic landscapes and housing issues. Texas is recognized for its busy cities and vast rural areas, whereas California has thriving technical hubs and scenic coastal cities. As for housing challenges, the two states both suffer from housing shortages. The housing shortage and high rent costs in San Francisco and Los Angeles are well-known throughout California. Population expansion in Texas, recognized for its affordable housing costs, has increased housing demand, raising concerns about supply and affordability. These issues highlight the need for predictive models to understand the relationship between property characteristics and rent price.

Prior research has investigated the prediction of home values using machine learning models, revealing their ability to capture complicated patterns and correlations. Nonetheless, the specific focus on rental costs, particularly in the context of two diverse yet economically relevant states, remains a relatively understudied topic. This study tries to overcome this gap by applying a variety of machine learning methods, such as ridge regression, random forest, extreme gradient boosting, light gradient boosting machine, a hybrid regression model, and a stacked generalization model of all previously mentioned models on the dataset to uncover its underlying tendencies. By incorporating both tree-based ensemble approaches and linear regression variations, the research considers several underlying links that lead to rent price fluctuations. By comparing the performance of different models, this study aims to determine which algorithms are the most competent at forecasting California and Texas rental costs.

2. Method

2.1. Data Source

This study utilized the “USA Housing Listings” Version 3 dataset, which was scraped from Craigslist in June 2020 and uploaded to Kaggle by Austin Reese [6]. Craigslist is a platform for posting and sharing classified advertisements and is the primary platform for rental home ads in the United States [7]. The dataset contains 384,977 data and 22 variables, with California and Texas subsets having 33085 and 31137 rows, respectively.

2.2. Data Preprocessing

A typical data cleaning process was employed. Duplicate listings and irrelevant variables were removed. Minimum and maximum values were set for “lat” and “long” to ensure all data in each subset were within state borders. Missing values of categorical variables “laundry_options” and “parking_options” were imputed with mode values of the listing’s respective property type. A minimum price of $100 and an area of 100 square feet were implemented to filter spam advertisements and rental parking spaces. Outliers were removed using the Inter-Quartile Range (IQR) method.

The cleaned data was further pre-processed so models could be trained more accurately. Synonyms in categorical features were merged; for instance, in variable “type”, “apartment” and “flat” were merged because they are British and American terms of the same property type. Both states have almost 30 values in “region”, so they were binned into 5 groups (0 to 4) based on median price and then ordinally encoded. Other categorical variables were one-hot-encoded and the first value in each variable was removed to decrease correlation between variables. Numerical variables were standardized.

For feature selection, both the Pearson correlation matrix and lasso regression (alpha=0.02) were employed. As observed in Fig. 1, “cats_allowed” and “dogs_allowed” have extremely high multicollinearity. After lasso regression, “dogs_allowed” was removed from the California dataset,

while "cats_allowed" was removed from the Texas dataset. While "beds", "baths", and "sqfeet" have high collinearity with each other (Fig. 1), they were not removed because lasso regression included them in selected features; in addition, in real life, an increase in the number of beds and baths impacts the value of property outside of total square footage.

After pre-processing, the California dataset contains 22946 rows and 26 variables, and the Texas dataset contains 19156 rows and 24 variables.

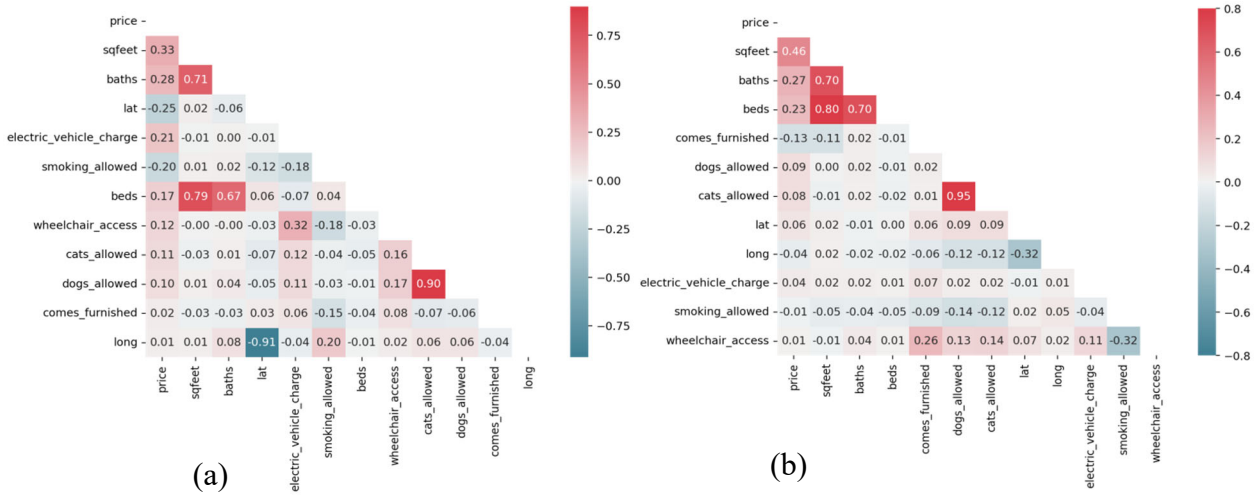

Fig. 1 Heat map of correlation matrix – numerical variables: **(a)** California; **(b)** Texas

2.3. Model Selection

2.3.1 Evaluation Metrics

In this experiment, the performance of the models was assessed using metrics of Mean Squared Error (MSE), coefficient of determination (R2), and Mean Absolute Percentage Error (MAPE).

MSE measures the average squared difference between the estimated and actual values. It calculates the distance between a regression line and a set of data points. R2 indicates the percentage of variance in the dependent variable that can be explained by the independent variable and ranges between 0 to 1. MAPE determines the average deviation from predictions, regardless of whether the deviations are positive or negative. The lower the MSE and MAPE, and the closer R2 is to 1, the more accurate the model predicts the data. Their equations are listed below:

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

$$\text{R}^2 = 1 - \frac{SS_{residuals}}{SS_{total}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

$$\text{MAPE}\ (\%) = \frac{1}{n}\sum_{i=1}^{n}\frac{|(y_i - \hat{y}_i)|}{y_i} \times 100 \tag{3}$$

Where y_i represents the actual value of y, $\hat{y}_i$ represents the predicted value of y, $\bar{y}$ represents the mean of the y values, and n represents the number of data points.

Initially, the data was fit on 12 models on an 80-20 train-split ratio with 10-fold cross-validation: linear regression, lasso regression, ridge regression, decision tree, random forest, extra trees regression, support vector machine (SVM), gradient boosting regressor, adaptive boosting regressor (AdaBoost), XGBoost, LightGBM, and K-nearest neighbors regressor (KNN). Random forest, XGBoost regressor, and LightGBM regressor were selected because they outperformed all other models for both California and Texas datasets in terms of mean MSE. Ridge regression achieved the lowest MSE out of all linear regression models and was selected as the meta-model for stacked generalization to increase the diversity of models in the ensemble.

2.3.2 Random Forest

Random forest constitutes an ensemble learning framework that combines several decision trees' predictions to improve predictive accuracy [8]. It incorporates bagging, random selection, and bootstrap. The process involves several steps: initially, a bootstrapped sample is drawn from the training data for each tree in the random forest. Subsequently, this sample is used to train an individual decision tree. During the construction of this tree, at each node, a subset of m features is randomly chosen to assess information gain, and the optimal feature is chosen. This iterative process continues until the tree reaches completion [9]. This methodology ensures that each tree contributes a distinct perspective, collectively bolstering the overall predictive performance of the random forest.

2.3.3 XGBoost

XGBoost is a scalable machine learning system for tree boosting [10]. Numerous machine learning and data mining challenges, such as on Kaggle, have acknowledged the significance of the method [10]. Scalability in all circumstances is the most important factor in XGBoost's success [10]. The process exhibits a speed improvement of over tenfold compared to existing methods when executed on a single machine and can handle billions of examples in various settings [10]. This scalability is achieved through crucial optimizations in both systems and algorithms, including a specialized tree learning algorithm for sparse data and a weighted quantile sketch technique for effective instance weight management [10]. Parallel and distributed computing speeds up the learning process, hence accelerating model exploration. Moreover, the framework involves the use of out-of-core computation and allows the processing of one hundred million samples on a desktop. After merging these techniques to achieve an end-to-end system, it is possible to accommodate more extensive datasets with the fewest cluster resources [8].

2.3.4 LightGBM

LightGBM represents a relatively recent addition to the gradient boosting decision tree (GBDT) algorithms. It distinguishes itself by incorporating innovative techniques such as gradient-based one-side sampling for handling extensive datasets and exclusive feature bundling to address a high number of features [11]. The experiment results in [11] have indicated that these two strategies aid LightGBM in achieving a faster computational speed and lower memory consumption, as compared to XGBoost [8].

2.3.5 Hybrid Regression

Hybrid regression consists of stacking two or more distinct regression models. This approach is proven in [8], in which the model combination of random forest, XGBoost, and LightGBM, each worth 33.33%, resulted in a lower test RMSLE than any of the models alone for predicting house prices in Beijing. It is essentially averaging the predictions of all three base models. By doing so, the ensemble of XGBoost, LightGBM, and random forest forecasts leverages the advantages of each model. As a consequence, the prediction improves in accuracy and robustness while also mitigating the inherent biases and errors associated with individual models.

2.3.6 Stacked Generalization

Stacked generalization enables a new model to discover the optimal way to combine the predictions of numerous existing models. It involves training base models using a k-fold approach, making predictions on the left-out fold. The meta-model is then trained on these out-of-fold predictions, with the new outputs guiding the base models in formulating final predictions [9]. This study implemented a 5-fold cross validation, in which the training data was divided into 5 parts. Base models were trained on four segments, and predictions for the fifth were obtained. This method was repeated five times, then out-of-fold predictions were generated from the entire dataset. These predictions serve as a new feature to the meta-model training process. During the prediction phase, the outcome is derived from averaging all forecasts [9]. This stacking approach capitalizes on the diversity of underlying models to enhance overall prediction performance.

For the experiment, a 2-level stacking architecture is implemented, with random forest, XGBoost, and LightGBM as base models to generate initial predictions, and ridge regression as the meta-model that takes the initial predictions as input and learns the optimal way to combine their outputs. A non-complex model was chosen as the meta-regressor so that it would not overfit the base model's results. Ridge regression was chosen over linear regression because it includes regularization parameters that can handle the correlation between base model predictions more effectively, and it achieved the lowest MSE among linear models during the model selection stage. While hybrid regression assigns the same weight to each model, the second layer ridge regression optimizes the weights given to each model.

3. Results

3.1. Data Analysis

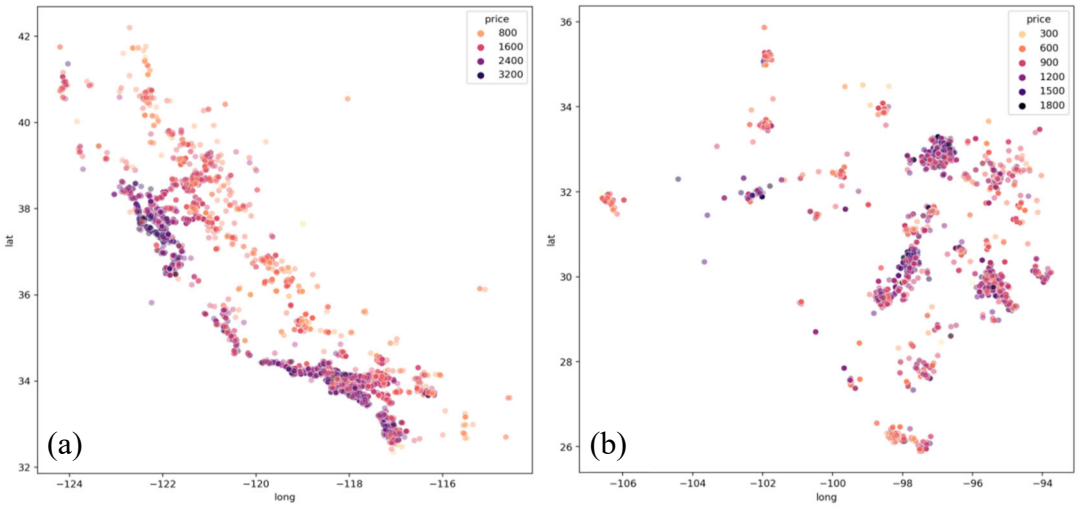

Fig. 2 Price distribution of rental properties: (**a**) California; (**b**) Texas

Exploratory data analysis was employed to uncover insights within the dataset, allowing for comprehensive understanding of variables and their interrelationships [12]. Fig. 2 depicts each property listing as data points on the map of California and Texas. The x-axis and y-axis represent the property's longitude and latitude coordinates, respectively. In California, it can be observed that locations along the coast, specifically the Bay Area and Los Angeles, have higher rental values compared to those of inland regions (Fig. 2a). In Texas, high-rent properties are more scattered, with clusters concentrated around Austin, Dallas, and Houston.

Fig. 3 illustrates the increase in price per square footage as the median rent of the region increases, where the x-axis represents the total square footage, and the y-axis represents the price. The "regions" variable was binned into 5 groups (region 0 to region 4) based on their median rents, with 0 being the most affordable and 4 being the costliest. It is evident that as square footage increases, regions with higher median rent (regions 3-4) have sharper increases in price compared to regions with lower median rent (regions 0-1) (Fig. 3).

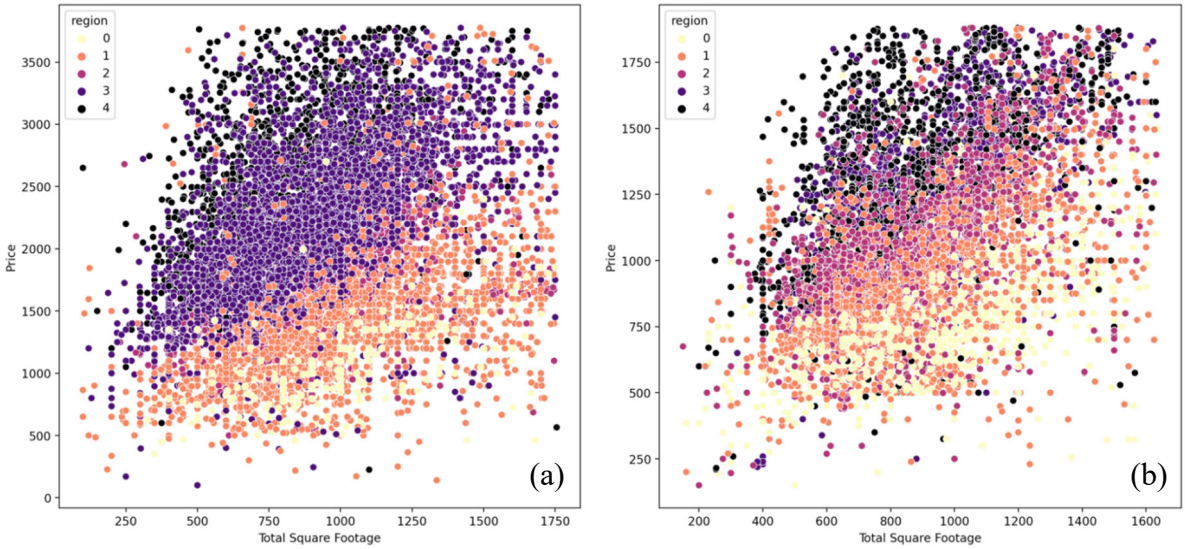

Fig. 3 Price per Total Square Footage by Region Median Value Rank: (**a**) California; (**b**) Texas

The individual distributions of each region also contribute to the model's performance. In Fig. 4, it can be observed that higher-priced regions in both datasets generally have higher IQR (box length) than lower-priced regions, meaning that the middle 50% of values have a larger spread. While we removed outliers based on the IQR method for the entire dataset, there are numerous high outliers for medium and lower-priced regions, which may impact the models' ability to predict accurately.

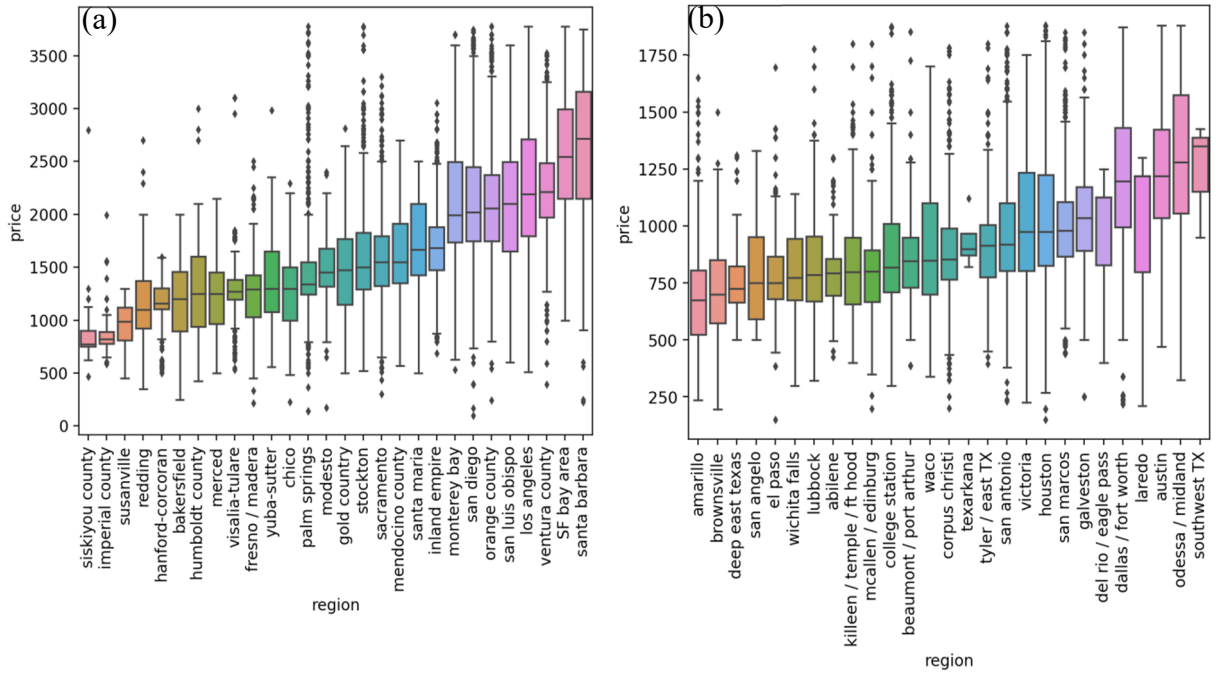

Fig. 4 Correlation between region and price: **(a)** California; **(b)** Texas

3.2. Evaluation

Many trials of hyperparameter tuning for each of the base models were conducted. To obtain the results in Table 1, XGBoost and LightGBM were meticulously optimized with Optuna, while the random forest was tuned with RandomSearchCV and GridSearchCV. Since both the hybrid regression and stacking methods were composites of the best regressions, further performance optimization was not employed.

Table 1. Prediction results

Model	California			Texas		
	MSE	R^2	MAPE (%)	MSE	R^2	MAPE (%)
Ridge	140890.60	0.6511	16.9581	36091.16	0.6062	15.1031
Random Forest	52842.38	0.8691	**7.6512**	**18401.93**	**0.7992**	**9.7003**
XGBoost	47947.73	0.8812	8.0460	21346.67	0.7671	10.9840
LightGBM	55541.83	0.8691	8.9059	21191.98	0.7688	10.6584
Hybrid Regression	46939.51	0.8837	7.8047	19363.49	0.7887	10.2116
Stacked Generalization	**46116.37**	**0.8858**	7.6843	19135.82	0.7912	10.0963

As shown in Table 1, both hybrid regression and stacked generalization models exceeded the base models for MSE and R2 for the California dataset. The stacked generalization model obtained the lowest MSE and highest R2 of 46116.3 and 0.8858, respectively, while the random forest model achieved the lowest MAPE of 7.6512% (Table 1). Given the significant improvements in MSE and R2 relative to random forest and other models, as well as the fact that the MAPE is only 0.0331% lower than that of random forest, it is reasonable to consider the stacked generalization model to be the best fit for this dataset.

For the Texas dataset, random forest outperformed both hybrid regression and stacked generalization in all evaluation metrics: it achieved the lowest MSE and MAPE of 18401.93 and 9.7003%, respectively, and the highest R2 of 0.7992 (Table 1). This is surprising because combining regressions is expected to have a coupling effect in which the various regression models aid each other in obtaining the closest predictions. Based on the first stacking level's prediction values, the second stacking level can train again and more accurately forecast house prices.

It can also be observed that for the California dataset, R2 is better but MSE is worse for all models when compared to evaluation results for the Texas dataset (Table 1). The MSE for California models is higher because we are predicting rent prices, which are generally higher in California.

Fig. 5 shows the predictions of each dataset's best models, with the x-axis as the true price and the y-axis as the predicted price. The line of best fit indicates when the true price and predicted price are equal. The training data overall fit much closer to the line of best fit, while the testing data are more scattered. Both models tend to overpredict rent costs when the true price is low and underpredict when the true price is high. Fig. 6 and 7 graph the predicted and true price of 100 samples of the California and Texas datasets, respectively. Stacked generalization more accurately predicted the extremely high or low true prices of California property compared to the random forest on the Texas dataset, which failed to capture some outlier prices. This could be due to the distribution of data in each dataset. The price difference among regions is more apparent in the California dataset, whereas the Texas dataset contains many regions with high price outliers near the upper bounds of the most expensive regions (Fig. 4).

From the results, the stacked generalization model is most appropriate for the California dataset and the random forest model is most appropriate for the Texas dataset. Ridge regression performed poorest on both datasets, with a MAPE of over 15% in both datasets. However, ridge regression proved effective as a meta-model on all the other base models in stacked generalization. In both datasets, multi-level stacking (stacked generalization) improves the accuracy of a model more than only stacking base models (hybrid regression) in terms of MSE, R2, and MAPE (%). However, stacking increases the performance accuracy of a model at the expense of computing time. In the case of the Texas dataset, random forest is a more optimal model as it outperforms all other models in all three metrics, and the computational time is less costly than the stacked models.

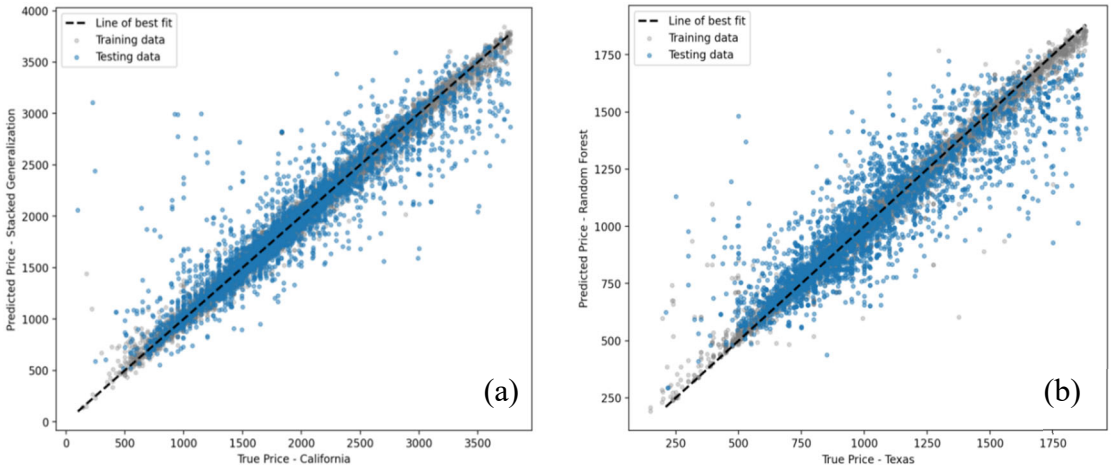

Fig. 5 Fit of best models: (a) California – stacked generalization (b) Texas – random

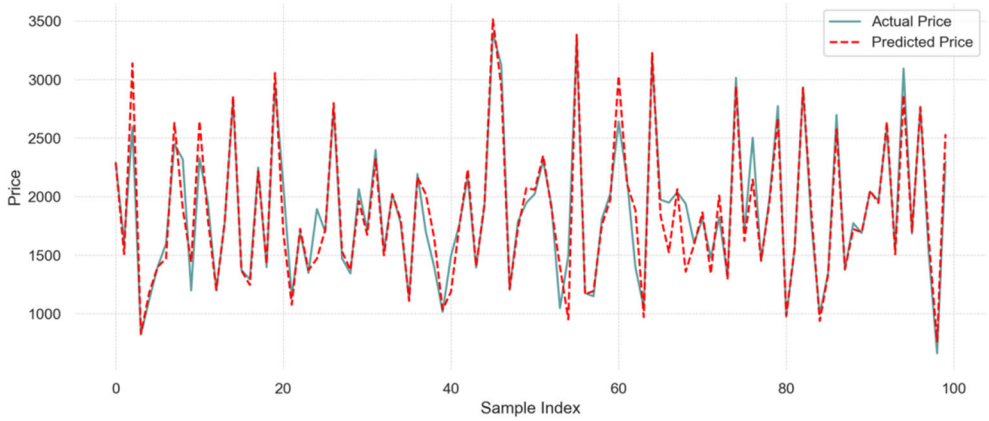

Fig 6. Comparison of stacked generalization's actual and predicted prices – California

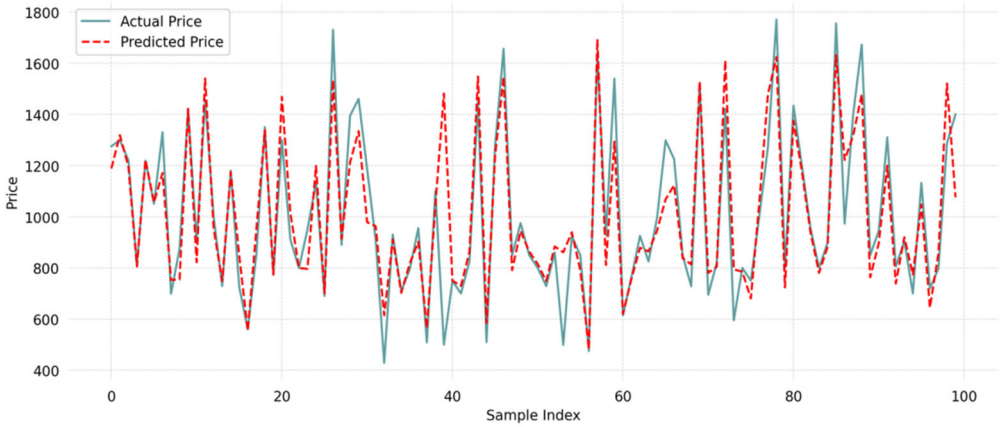

Fig 7. Comparison of random forest's actual and predicted prices – Texas

4. Discussion

During the initial training on 12 models to select the final models in Table 1, ridge regression obtained the lowest mean MSE among linear regression models for both datasets. However, compared to the models in Table 1, it performed the worst for all evaluation metrics on both datasets. This demonstrates that linear regression is indeed insufficient to account for the non-linear aspects and relationships in real estate data. However, when implemented as the meta-regressor in the stacked generalization model, ridge regression improved the MSE, R2, and MAPE of both datasets compared to the hybrid regression model, which only consists of random forest, XGBoost, and LightGBM – the

base models for stacked generalization. Therefore, while both the hybrid regression and stacked generalization models incorporate the combined influence of multiple regressions, the efficacy differs. Hybrid regression relies on a learning method that simply averages predictions, rendering it less effective. On the other hand, stacked generalization's second level allows for further model training and refinement, resulting in improved prediction accuracy [8].

The results indicate that a complex stacked model may not always perform better. Although hybrid regression and stacked generalization were trained on the Texas dataset to improve prediction accuracy, random forest, one of the base models, still outperformed both techniques. A reason for this may be due to the base models selected for stacking. Ensembles typically produce superior outcomes when there is a substantial amount of base model diversity, and when there are more base models available for training [13]. In this study, base models' random forest, XGBoost, and LightGBM were selected for having the lowest mean MSE individually out of 12 models; however, XGBoost and LightGBM share similar mechanisms and assumptions, which may not significantly assist the meta-model in determining the best predictions. More model diversity could be added, such as neural networks, KNN regressors, and SVM.

Despite using numerous variables and models to predict rent prices, there are still limitations in the predictions. Future work may be accomplished by expanding the dataset, the number of features, and the variety and number of models in stacked generalization. Expanding the dataset refers to more geographical locations within and outside the United States to find the best model for the generalization of rent prices in a larger scope. Additional numerical features such as distance from the city center, distance from closest school, ranking of closest school, and distance to closest mall can be added to potentially increase prediction accuracy. Non-numerical for further investigation include textual and image data that are usually included with rental listings on Craigslist and other data sources. By integrating these features, a wider range of models can be tested in stacked generalization, such as natural language processing on listing descriptions that may contain crucial keywords that impact rent prices and cannot be reflected by non-textual variables, and neural networks for image data, which can detect property conditions and quality.

5. Conclusion

This study compared the application of machine learning models for predicting rental prices using California and Texas property listing data. Ridge regression, random forest, XGBoost, and LightGBM were individually trained on each dataset, along with a hybrid regression model of the latter three models and a multi-level stacked generalization model with ridge regression as the meta-regressor. The findings demonstrate that the stacked generalization model served as the optimal choice for California, with an MSE of 46116.37, R^2 of 0.8858, and MAPE of 7.6843%, and the random forest model was most appropriate for Texas, with an MSE of 18401.93, R^2 of 0.7992, and MAPE of 9.7003%. Stacked generalizations, including the hybrid regression model, generally achieved better results than their model components obtained individually for the California dataset, but it is important to consider the computational cost of stacking model predictions. Future studies can incorporate real estate data from additional geographical regions to create a more versatile model. In addition, textual, image, and more numerical features can be integrated to determine if model improvement can be achieved. Lastly, for stacked generalization, additional model types and combinations of base and meta-models can be tested. This research can considerably contribute to forecasting future rent values, which benefits tenant-landlord negotiations, property investment decisions, and real estate policy formulations.

References

[1] X. Zhou, W. Tong, and D. Li. Modeling housing rent in the atlanta metropolitan area using textual information and deep learning. ISPRS International Journal of Geo-Information, 2019,8(8): 349. 2019.

[2] A. Singh, A. Sharma, and G. Dubey. Big data analytics predicting real estate prices. International Journal of System Assurance Engineering and Management, 2020, 11(S2):208–219.

[3] S. Abdul-Rahman, N. H. Zulkifley, I. Ibrahim, and S. Mutalib. Advanced machine learning algorithms for house price prediction: case study in kuala lumpur. International Journal of Advanced Computer Science and Applications, 2021, 12(12).

[4] L. Xu and Z. Li. A new appraisal model of second-hand housing prices in china's first-tier cities based on machine learning algorithms. Computational Economics, 2020, 57(2):617–637.

[5] S. Özöğür Akyüz, B. Eygi Erdogan, Ö. Yıldız, and P. Karadayı Ataş. A novel hybrid house price prediction model. Computational Economics, Sep. 2022.

[6] A. Reese, USA Housing Listings, Kaggle, Jun. 17, 2020. [Online]. Available: https://www.kaggle.com/datasets/austinreese/usa-housing-listings

[7] G. Boeing and P. Waddell. New insights into rental housing markets across the United States: web scraping and analyzing craigslist rental listings. SSRN Electronic Journal, 2016.

[8] Q. Truong, M. Nguyen, H. Dang, and B. Mei. Housing price prediction via improved machine learning techniques. Procedia Computer Science, 2020,174: 433–442.

[9] H. Zhang, K. Wang, M. Li, X. He, and R. Zhang. House price prediction with an improved stack approach. Journal of Physics: Conference Series, 2020,1693(1): 012062.

[10] T. Chen and C. Guestrin, XGBoost, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[11] G. Ke et al. LightGBM: A Highly Efficient Gradient Boosting DecisionTree, 31st Conference on Neural Information Processing Systems, 2017: 3149–3157.

[12] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani. Exploratory data analysis using python. International Journal of Innovative Technology and Exploring Engineering, 2019, 8(12):4727-4735.

[13] P. Sollich and A. Krogh. Learning with ensembles: how over-fitting can be useful. Proceedings of the 8th International Conference on Neural Information Processing Systems, 1995: 190–196.