Comparison of Different Models on House Price Prediction

Shanyong Wang, *2022533061*

***Abstract*—This study focuses on predicting house prices using various regression models. We explore multiple algorithms including Linear Regression, K-Nearest Neighbors (KNN), Ridge Regression, Random Forest. The dataset consists of features relevant to house pricing, and the models are evaluated based on their Root Mean Squared Error (RMSE) and R^2 score. The results indicate that ensemble methods, particularly Gradient Boosting and Random Forest, outperform other models in terms of prediction accuracy. The study concludes with a blended model approach that combines predictions from the best-performing models to achieve a robust final prediction.**

***Index Terms*—house price prediction, random forest, ensemble learning, kaggle**

I. Introduction

PREDICTING house prices is a complex but crucial task with significant implications for homeowners, buyers, real estate professionals, and financial institutions. Accurate predictions can aid in making informed decisions, mitigating risks, and optimizing investment strategies. The task, however, is fraught with challenges due to the myriad factors influencing real estate values, including economic conditions, property characteristics, location-specific attributes, and market trends.

Recent advancements in machine learning have provided powerful tools for tackling this problem. Traditional methods often relied on linear regression models and basic statistical techniques, which, while useful, may not fully capture the intricate relationships and nonlinear patterns present in real estate data. Machine learning models, on the other hand, can handle large datasets, incorporate a wide range of features, and uncover hidden patterns, leading to more accurate and robust predictions.

This study leverages various machine learning techniques to build and evaluate models for predicting house prices. The dataset used is from a well-known Kaggle competition, the Ames Housing dataset, which offers a comprehensive and detailed set of features for each property. The features range from basic attributes like the number of bedrooms and bathrooms to more nuanced details such as the type of foundation and the quality of materials used.

Our approach involves a systematic exploration of the data, thorough preprocessing to handle missing values and categorical variables, and the implementation of multiple machine learning algorithms, including Linear Regression, K-Nearest Neighbors (KNN), Ridge Regression, and Random Forest Regressor. We also perform hyperparameter tuning to optimize model performance and ensure the best predictive accuracy.

Through this study, we aim to compare the effectiveness of different machine learning models in predicting house prices and highlight the importance of feature engineering and model tuning in achieving high performance. The findings from this research can provide valuable insights for stakeholders in the real estate market and contribute to the broader field of predictive analytics in real estate.

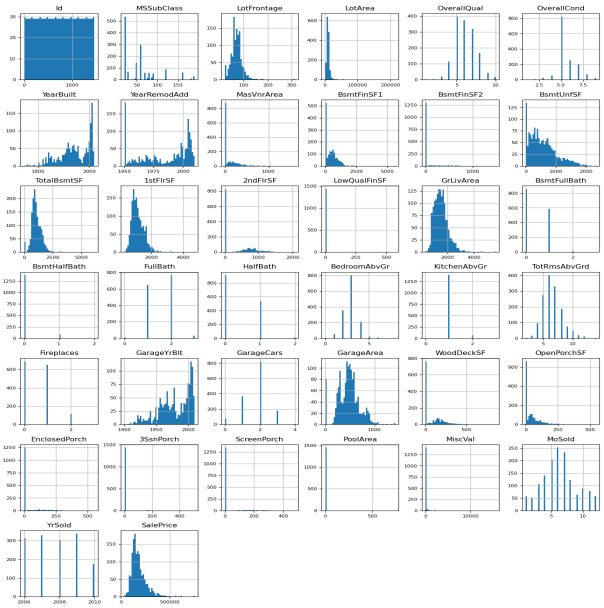

Fig. 1. Features

By meticulously analyzing and experimenting with various models and techniques, this paper seeks to push the boundaries of what is possible in house price prediction, offering a robust framework that can be applied to other similar datasets and scenarios. The ultimate goal is to enhance the accuracy and reliability of house price predictions, thereby supporting more informed decision-making processes in the real estate industry.

II. Data Processing

The dataset contains two parts. One is train_data and the other is test_data. The size of the train_data is (1460,81) and the size of the test_data is (1460,80). The train_set contains information of a couple of house features and their corresponding price, while test_data only contains house features. We can see some of the features through Figure 1.

Through the figure we can see that some of the data is out of range. SO these data is out of our consideration since they will interference model generation. Then we plot the relationship between price and house id. We can see the distribution in Figure 2. After that we can check the missing value of the

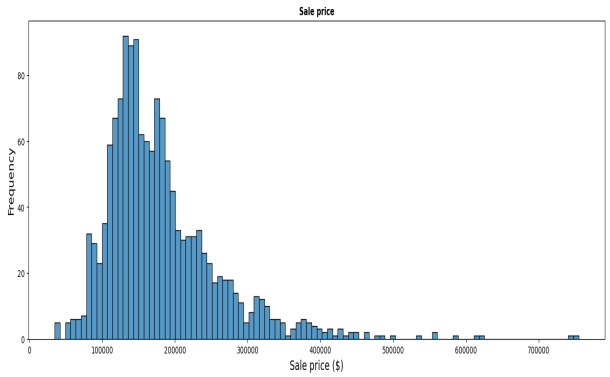

Fig. 2. Price Distribution

features. As we can see in Figure 3, There are 35 different types of missing features at all. So we want to fill in these missing data.

	missing_data	missing_precent	missing_type
PoolQC	2908	0.996915	object
MiscFeature	2812	0.964004	object
Alley	2719	0.932122	object
Fence	2346	0.804251	object
MasVnrType	1766	0.605417	object
SalePrice	1459	0.500171	float64
FireplaceQu	1420	0.486802	object
LotFrontage	486	0.166610	float64
GarageCond	159	0.054508	object
GarageYrBlt	159	0.054508	float64
GarageFinish	159	0.054508	object
GarageQual	159	0.054508	object
GarageType	157	0.053822	object
BsmtCond	82	0.028111	object
BsmtExposure	82	0.028111	object
BsmtQual	81	0.027768	object
BsmtFinType2	80	0.027425	object
BsmtFinType1	79	0.027083	object
MasVnrArea	23	0.007885	float64
MSZoning	4	0.001371	object
BsmtHalfBath	2	0.000686	float64
Utilities	2	0.000686	object

Fig. 3. Missing data

We consider 2 cases. The first one is those features that have more than 15 missing percents. Those features do not contribute a lot to the prediction. The second case is those features that have less than 15 missing percents. We fill in the missing data according to their meanings. For example, we fill 0 in the missing value of GarageCars and GarageArea since the missing of them probably means that 0 value of them. After dealing with the missing value, we make categories out of data that are not continuous values and use sklearn to map the label into consecutive eigenvalues between 0 and n-1. Finally, use the dummy method of pandas to uniquely hot-code the data and form the final training and testing datasets.

III. Models Applied and Comparison of different methods

For this project, we select four models to predict the final answer. They are Linear Regression, KNN Regression, Ridge Regression and Random Forest Regression.

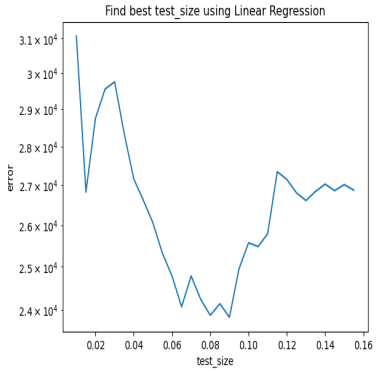

Fig. 4. Test Size

Comparing these model, we can find that Linear Regression and Ridge Regression are supervised learning while KNN and Random Forest are unsupervised learning. The advantages of Linear and Ridge regression is there are easy to calculate and the process of adjustment of parameters takes less time. But they reply more on the dataset. If the data is not linear correlated, the result may be far away from the true result. The KNN and Random Forest do not depend on the dataset.

First of all, we separate the training set into training part and validation part. We consider the test size and find it to be 0.09. See more on Figure 4.

After that, we train the dataset through each model. And for KNN Regression and Ridge Regression, we can adjust their parameters to make them preform well. For KNN regression, we consider its number of neighbour and the type of distance. For Ridge Regression, we consider the parameter of regularization–alpha. For Random Forest Regressor, we consider the n_estimators, max_depth, max_features. The picture of these parameters will be displayed at the end of the article.

IV. Numerical results

The training set is divided into two parts. One for training and the other for validation. Before traing the parameters, the result are below.

(1).LinearRegression: 23837.212226
(2).KNN: 31916.597308
(3).Ridge Regression: 23828.085679
(4).RandomForest: 22102.198540

After adjusting parameters, the four models perform better on the validation set.

(1).LinearRegression: 23837.212226
(2).KNN: 24068.826611
(3).Ridge Regression: 22626.771617
(4).RandomForest: 21023.797474

From the result, we can see that adjustment of parameters is of great importance. All of the models perform better, especially KNN regression.

From the dataset, We can learn that Random Forest performs the best while KNN performs the worst. For supervised learning, Ridge Regression performs better than Linear Regression, since Ridge add a Regulation part at the end of the loss

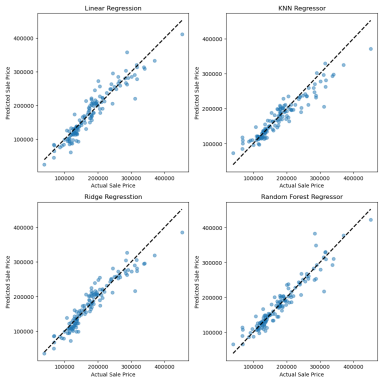

Fig. 5. Result

function, which makes the model have a greater generalization ability.

At the end of the project, we use ensemble learning to get the final result. To be more specially, we use Random Forest Regression and Ridge Regression and each of them have half of the weight. We submit the result to Kaggle and get 0.14411 score(2029 ranged), which is the greatest among other methods to generate the result.

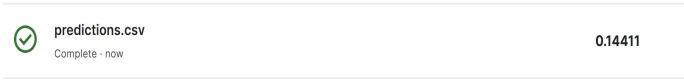

Fig. 6. Ensemble Learning Result

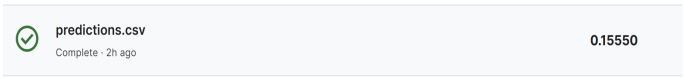

Fig. 7. Random Forest Result

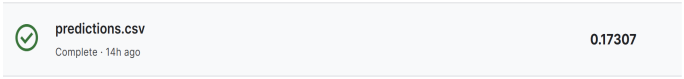

Fig. 8. Ridge Result

V. Conclusion

This study has explored the application of various machine learning techniques to the task of house price prediction using the Ames Housing dataset. Through systematic data exploration, preprocessing, and model building, we have demonstrated the strengths and limitations of different algorithms, including Linear Regression, K-Nearest Neighbors (KNN), Ridge Regression and Random Forest Regressor.

Our findings indicate that machine learning models can significantly enhance the accuracy of house price predictions compared to traditional statistical methods. The Random Forest Regressor, in particular, showed superior performance, highlighting the benefits of ensemble methods in capturing complex, nonlinear relationships within the data. The process of hyperparameter tuning further improved the model's predictive capabilities, underscoring the importance of optimizing model parameters to achieve the best results.

Feature engineering played a crucial role in this study, as the careful selection and transformation of features contributed to the overall performance of the models. Handling missing values, converting categorical variables, and scaling numerical features ensured that the data was in an optimal format for training the models.

Despite the successes achieved, there are still areas for improvement and future research. Incorporating more advanced techniques such as gradient boosting machines, neural networks, or hybrid models could potentially yield even better results. Additionally, exploring the impact of temporal and spatial factors, as well as external economic indicators, could provide a more comprehensive understanding of the factors influencing house prices.

In conclusion, this research has demonstrated the efficacy of machine learning in predicting house prices, providing valuable insights and a robust framework for future applications in real estate analytics. By leveraging advanced algorithms and thorough data processing, we can achieve highly accurate predictions that support informed decision-making in the real estate market. This study contributes to the growing body of knowledge in predictive analytics and underscores the potential of machine learning to transform the field of real estate valuation.

Acknowledgment

The author would like to thank Prof. Zhao and all the teacher assitants in CS182, introduction to machine learning. All the models here are displayed and learned in class. What author does is just putting it into practise and using different methods to make it more accurately.

References

(1) https://blog.csdn.net/weixin_48994268/article /details/109686341
(2) https://www.kaggle.com/code/serigne /stacked-regressions-top-4-on-leaderboard
(3) https://scikit-learn.org/stable/modules/generated /sklearn.preprocessing.OneHotEncoder.html

VI. Some results

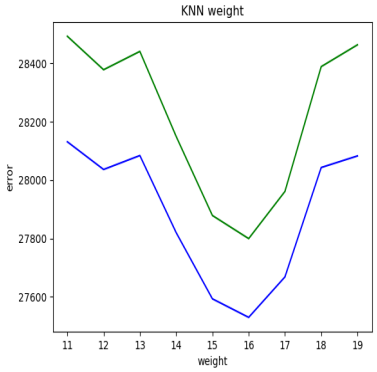

Fig. 9. knn weight

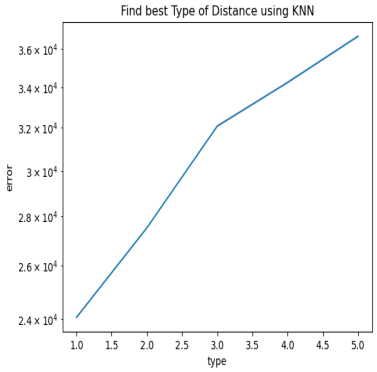

Fig. 10. best type knn

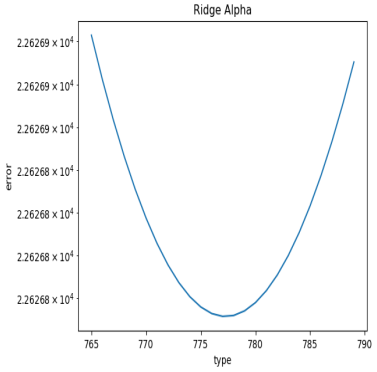

Fig. 11. ridge alpha