Prediction on Housing Price Based on the Data on Kaggle

Jiachen Yu(✉)

College of Letters and Science, University of California, Davis, USA
hjcyu@ucdavis.edu

Abstract. People's lives have always relied on having a secure place to stay. As a result, housing prices have risen to people's top priority list. This paper uses a series of correlation tests, exploratory data analysis, and feature selection approaches to the training and testing datasets to find the most accurate model for forecasting housing prices. There is a dataset about housing prices on Kaggle. The author found that the variables PoolQC, MiscFeature, Alley, and Fence include almost 90% missing values during the data preparation process. After removing those variables, the author generated a correlation matrix to visualize the relationship between the rest variables. In addition, the exploratory data analysis on the dataset shows the overall quality of a home, the size of the living area, the total basement size, and the presence of newer homes contribute the most to a house's value. The author created seven different machine learning models and calculated the R-square and root mean square values. Among these models, the random forest algorithm has the highest R-Squared value and the lowest RMSE. As a result, the random forest algorithm is the best model for predicting the price of a house.

Keywords: Data preparation · Correlation · Exploratory data analysis · RMSE · Machine learning models

1 Introduction

Housing prices began to rise in response to the economy's continued growth and the impact of inflation. Many people have been put under financial strain as a result of increased housing prices. Housing is an essential and significant aspect of people's lives. Therefore, it is critical to determine if a house's price is appropriate and to understand how different housing characteristics affect the price of a house before choosing and purchasing one [9].

This paper will examine a housing pricing dataset and notebook from Kaggle.com. The notebook is based on a housing price prediction competition on Kaggle. The Housing Dataset contains observations of 1,460 residential properties sold in Ames, Iowa, between 2006 and 2010. There are 80 observation variables in total in the dataset, with 23 nominal variables, 23 ordinal variables, 20 continuous variables, and 14 discrete variables [3]. These variables encompass all aspects of a property listed on the market, which is convenient for us to predict housing prices. The model's dependent variable will

Z. Zeng et al. (Eds.): ECIT 2022, AHE 11, pp. 627–634, 2023.
https://doi.org/10.2991/978-94-6463-005-3_64

be the variable SalePrice which stands for the property's sale price in dollars. Variable SalePrice is the target variable that we will be forecasting in our machine earning models. All of the other variables represent a property's features. The variable LotArea, for example, denotes a property's lot size in square feet. The YearBuilt variable represents a property's original construction date. Variable Neighborhood is a property's physical location within Ames City limits. The author will first check the dataset and prepare the data for further analysis and modeling. In this part, the author will check the number of missing values of each variable in the dataset and delete all duplicated data and variables with a large number of missing values.

Furthermore, the author will conduct a correlation test to the dataset. After the first step, the author will perform exploratory data analysis on the housing price dataset to understand the correlations between the variables. Following that, the author will use the provided data to create many machine learning models, including Random Forest, Gradient Boosting, Linear regression, etc. Finally, by comparing the histograms of different models and the adjusted R-squared values, the author will determine the most accurate model for predicting future housing prices.

2 Method

2.1 Data Preparation

To begin with, the author first checked the number and percentage of missing values in each variable of the dataset. As Table 1 shows, there are 2909 missing values in variable PoolQC which means 99.66% of the values in the PoolQC column are missing values. For variable MiscFeature, there are 2814 missing values in total which accounts for about 96.40% of the values in the MiscFeature column [7]. The percentages of missing values of variable Alley and Fence are 93.22% and 80.44%. All four of these variables have an extremely high amount of missing values, leading to inaccurate results if we keep them in the dataset [2]. Thus, the author decided to remove these four variables from the dataset. Moreover, the author cleaned missing values using the Random Value Imputation method since it is the best technique to preserve distribution for each feature. The author then checked if there was data duplication in the dataset. There were no data duplicated.

Furthermore, Fig. 1 represents the author's correlation matrix using all the variables. The author observed that the variables TotRmsAbvGrd and GrLivArea have a correlation

Table 1. Number and percentage of missing values of each variable.

	Null Values	Percentage Null Values
PoolQC	2909	99.657417
MiscFeature	2814	96.402878
Alley	2721	93.216855
Fence	2348	90.438506
SalePrice	1459	49.982871

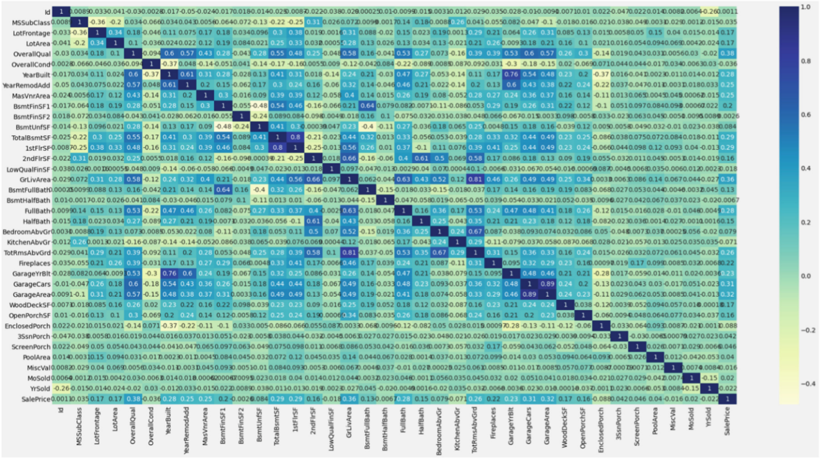

Fig. 1. Correlation matrix of all variables except the removed ones.

coefficient of 0.81. This value indicates that the total number of rooms above grade a house has is highly correlated with the living area measured square feet above the ground. Variable 1stFlrSF and TotalBsmtSF have a correlation coefficient of 0.8 which means that the square feet of the first floor of a house are highly correlated with the total square feet of the basement area. Additionally, the variables GarageArea and GarageCars have the highest correlation coefficient among all pairs, which is 0.89. This result indicates that the size of the garage of a house is highly correlated with the garage's car capacity. Variable GarageYrBlt and YearBuilt have a correlation coefficient of 0.76 which means the year the garage is built correlated with the original construction date of the house.

2.2 Exploratory Data Analysis

In order to have a more detailed understanding of the variables in the dataset, the author decided to perform Exploratory Data Analysis on the housing price dataset [8]. The author created box plots and histograms of all the numerical variables in the dataset. The author also created scatterplots for each pair between variables SalePrice, OverallQual, GrLivArea, GarageCars, TotalBsmtSF, FullBath, and YearBuilt. As Fig. 2 shows, variable SalePrice tends to be positively correlated with variables QverallQual, GrLivArea, TotalBsmtSF, and YearBuilt. There is no apparent trend between variable SalePrice and GarageCars and FullBath. Thus, characteristics such as higher overall quality of, larger living area above ground, a larger basement area, older age tends to increase the housing value. The variable OverallQual shows a positive linear relationship with variables SalePrice, GrLivArea, TotalBsmtSF, and FullBath. No apparent relationship was observed with GarageCars and YearBuilt. A larger living area above ground, larger basement area, and more full bathrooms tend to increase the overall quality of a house [4]. Variable GrLivArea also shows a positive linear relationship with variables TotalBsmtSF and FullBath. The larger the basement area and more full bathrooms in a house, the larger the living area above ground. The scatterplots of variable GarageCars show a

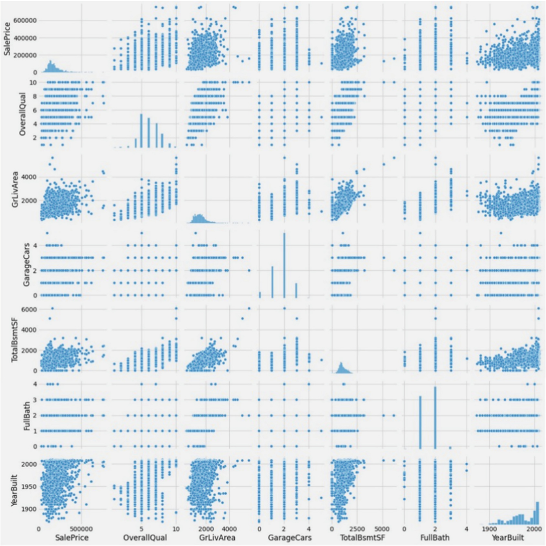

Fig. 2. Scatter plot of variables SalePrice, OverallQual, GtLivArea, GarageCars, TotalBsmtSF, FullBath, and YearBuilt.

positive linear relationship with variable TotalBsmtSF which means the larger the basement area of a house, the larger car capacity the house will have. The basement area tends to be larger for newer houses.

In addition, the author created kernel density estimator plots for variables SalePrice, YearBuilt, TotalBsmtSF, and GriLivArea. As Fig. 3 shows, a sale price of 150,000 dollars tends to have the highest probability density function among other prices. Because of the highest probability of sale, the most sold houses were constructed about 2006. The total square feet of the basement area tends to be about 900 square feet since it has the highest probability density function of 0.0012. Lastly, the most living area above ground is about 1,500 feet^2 with the highest probability density function of over 0.0008.

Besides, the author created violin plots for variables SalePrice, FullBath, BedroomAbvGr, KitchenAbvGr, and TotRmsAbvGrd. These plots can help to visualize the distribution of the variables better. As Fig. 4 shows, the sale price of 138.346 thousand dollars is the most popular price with a kernel density estimator probability of 1.00. For variable FullBath, most properties sold have two full bathrooms above ground with an estimated probability density function of 1.00. The probability of a property sold with three or four full bathrooms is low compared with those with one or two full bathrooms. However, properties sold mostly have three bedrooms with an estimated probability density function of 1.004. The probability density function of a house sold with one kitchen is 0.999. As of variable TotRmsAbvGr, the houses sold with six rooms above ground have a probability density function of 0.999; it is also the median value of the variable.

Finally, the author developed scatterplots for the variables TotalBsmtSF, EnclosedPorch, OpenPorchSF, and WoodDeckSF by pairing these four variables with our dependent variable SalePrice. Figure 5 shows a scatter plot showing a positive trend between

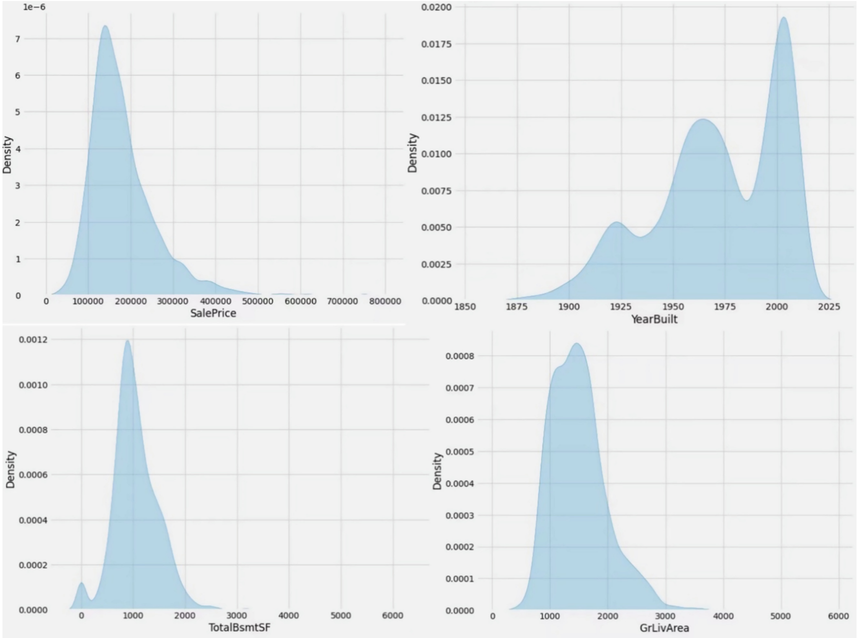

Fig. 3. Kernel density estimator plots for variables SalePrice, YearBuilt, TotalBsmtSF, and GriLivAre.

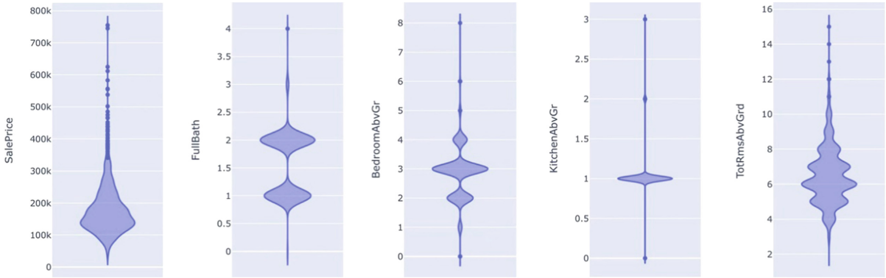

Fig. 4. Violon plots for variables SalePrice, FullBath, BedroomAbvGr, KitchenAbvGr, and TotRmsAbvGrd.

the variables TotalBsmtSF and SalePrice. The house's sale price will rise in lockstep with the total basement space. The scatter plots of OpenPorchSF and WoodDeckSF reveal that SalePrice has a negative relationship. This means that when the open porch and wood deck area increases, the sale price will fall. Finally, there is no real relationship between the variables SalePrice and EnclosedPorch [5].

2.3 Feature Selection

This model was built to figure out the features that contribute the most to the dependent variable. The author began by converting all categorical variables to dummy variables,

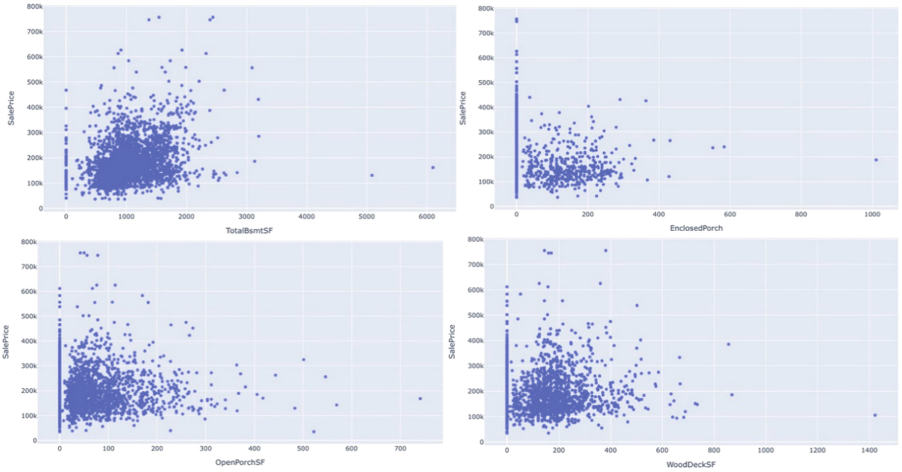

Fig. 5. Scatterplots of variables TotalBsmtSF, EnclosedPorch, OpenPorchSF, and WoodDeckSF paired with vairbale SalePrice indivisuallt.

Table 2. R-squared and RMSE of each model.

	R-squared	RMSE
Random Forest	0.3716	61604.55
Gradient Boosting	0.1949	69727.96
Linear Regression	−183240320.23	1051946840.7
Lasso	0.0677	75035.7
Ridge	0.1042	73550.42
K-Neighbor	0.0514	75687.56
Decision Tree	−0.1770	84308.62

then re-insert them into the dataset. The author extracted all of the data from the variable SalePrice in both train and test datasets and call it “y_train” and “y_test”. The author collected all of the data for the remaining variables in both datasets and name them “X_train” and “X_test”. Then, the author separated the dependent and independent features in both datasets. According to this research, the original X shape contains 2,919 rows and 276 columns. Finally, the author used the SelectKBest function to select features. The X shape includes 2,919 rows and 40 columns after feature selection.

2.4 Modeling

The author utilized the X train and y train datasets as the predictor and response in all of the prediction models. To compare, the author developed seven machine learning models. Random Forest, Gradient Boosting, Linear Regression, Lasso, Ridge, K-Neighbors, and Decision Tree algorithms are the seven models the author created. The author also computed each model's R-square, and root means square error value. Table 2 lists all of the numbers of each model. The Random Forest Algorithm results in the highest R-squared value and the lowest root mean square value among all the prediction models.

3 Conclusion

The author concludes that the random forest algorithm is the best machine learning model for forecasting housing prices based on the findings in Table 2. Throughout the author's investigation, he discovered that a house's overall quality, large living area, total basement size, and newer houses are all factors that contribute to a house's increased value. As a result, these four criteria should be considered while purchasing a house. These factors can be used to determine if a house's pricing is reasonable. This is really beneficial and efficient for most people when investing in property.

Furthermore, the violin plots in the study showed at the most common features of residences sold between 2006 and 2010. The most popular residences cost approximately $138,346 on average. These residences typically include two full bathrooms, three bedrooms, and a kitchen; the total number of rooms is around six. As a result, if homebuyers wish to sell their houses more swiftly in the future, they should look for these features while making their purchase.

Zillow recently stated that their house price forecasting program will be discontinued off. Because predicting housing values is exceedingly difficult and needs a significant amount of capital and labor to maintain. "We have been unable to properly anticipate future property prices at various times in both directions by a significant amount," Zillow claims [10]. Although precisely projecting future house prices is extremely difficult, the social and economic benefits of doing so are enormous. More welfare can be obtained by both the government and the public.

References

1. Dufitinema J (2021) Stochastic volatility forecasting of the Finnish housing market. Appl Econ 53(1):98–114
2. Ghodsi R, Boostani A, Faghihi F (2010) Estimation of housing prices by fuzzy regression and artificial neural network. In: 2010 4th Asia international conference on mathematical/analytical modelling and computer simulation. https://doi.org/10.1109/ams.2010.2
3. House prices - advanced regression techniques. Kaggle. (n.d.). https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data. Retrieved 3 Dec 2021
4. Kuvalekar A, Manchewar S, Mahadik S, Jawale S (April 2020) House price forecasting using machine learning. In: Proceedings of the 3rd international conference on advances in science & technology (ICAST)
5. Li W, Cheng Y, Fang Q (2020) Forecast on silver futures linked with structural breaks and day-of-the-week effect. North Am J Econ Financ 53:101192
6. Li Y, Xiang Z, Xiong T (2020) The behavioral mechanism and forecasting of Beijing housing prices from a multiscale perspective. Discrete Dynamics in Nature and Society (2020)
7. Mahrous AM (17 October 2021) House price prediction (99.5%ACC). Kaggle. https://www.kaggle.com/ahmedmohamedmahrous/house-price-prediction-99-5-acc/notebook. Retrieved 3 Dec 2021

8. Nazemi B, Rafiean M (2020) Forecasting house prices in Iran using GMDH. Int J Hous Mark Anal
9. Nur A, Ema R, Taufiq H, Firdaus W (2017) Modeling house price prediction using regression analysis and particle swarm optimization case study: Malang, East Java, Indonesia. Int J Adv Comput Sci Appl 8(10). https://doi.org/10.14569/ijacsa.2017.081042
10. Ray T (2 November 2021). Zillow says difficulty of forecasting, huge capital need felled home buying business. ZDNet. https://www.zdnet.com/article/zillow-says-difficulty-of-forecasting-huge-capital-need-felled-home-buying-business/. Retrieved 3 Dec 2021