Analysis and Prediction for House Sales Prices by Using Hybrid Machine Learning Approaches

S. Hossain, Jyoti Rawat, Doina Logofatu

▶ **To cite this version:**

S. Hossain, Jyoti Rawat, Doina Logofatu. Analysis and Prediction for House Sales Prices by Using Hybrid Machine Learning Approaches. 17th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI), Jun 2021, Hersonissos, Crete, Greece. pp.594-604, 10.1007/978-3-030-79150-6_47 . hal-03287678

HAL Id: hal-03287678

https://inria.hal.science/hal-03287678v1

Submitted on 15 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Analysis and Prediction for House Sales Prices by Using Hybrid Machine Learning Approaches

S M Soliman Hossain,[1] Jyoti Rawat[1] and Doina Logofatu[1]

Frankfurt University of Applied Sciences, Frankfurt Am Main , Germany
logofatu@fb2.fra-uas.de

Abstract. Over the past few years, machine learning has played an increasingly vital role in every aspect of our society. There are countless applications of machine learning, from tradition topic such as image recognition or spam detection, to advanced areas like automatic customer service or secure automobile systems. This paper analyzes a popular machine learning application, namely housing price prediction, by applying a full machine learning process: feature extraction, data preparation, model selection, model training and optimization, and last, but not least, prediction and evaluation. We experiment with different algorithms: linear regression, random forest, and gradient boosting. This paper demonstrates the comparison of effectiveness of these algorithms that may help sellers and buyers to have a fair deal of their respective businesses.

Keywords: Housing price · Linear Model · Random Forest · Gradient Boosting

1 Introduction

As the world population is increasing, the demand for affordable housing has soared like never before. Housing becomes a major concern of the society. The mismanagement in housing prices can have a negative impact on the economy of a country. That is why a scientific and deterministic approach should be considered to determine a fair price and ensure the benefits and fairness for both sellers and buyers. A good housing price predictive software can assist real estate developers in figuring out the selling price of the house and can guide the customer to decide the right time to make a purchase. A lot of research work has been conducted to untangle the secret of housing price prediction. For example, numerous researchers believe that the geographical position, cultural and socioeconomic situation of an area will decide the future increase or decrease in the demand of a house. In practice, those are just a small portion of the numerous relevant factors. The goal of this paper is to identify the deciding factors and generate a model that provides a good estimation of the market price of a given house.

In this study, our dataset is provided by a Kaggle competition, namely *House Price Prediction* [9]. The data is collected based on information from the city of Ames, USA. Each residential property is characterized by a set of up to 79 explanatory variables. The task is to train a model that can make predictions as close to the given prices as possible.

2 Related Work

In 2017, Wu [8] applies Support Vector Regression (SVR) on an open-source dataset which consists of 20 explanatory features and 21,613 housing entries. In his research, various methods for feature extract were used such as Recursive Feature Elimination (RFE), Lasso, Ridge, and Random Forest Selector and Principal Component Analysis (PCA). The test results showed that there was no significant distinction in performance among feature selection methods.

Towne (2016) [7] suggested a visualization process for estimating price for single-family properties. He analyzes a total of 5,142 online property listings between 2012 and 2015 in different areas. The selected features consist of mostly listed specifications, including interior square footage, lot size, number of bedrooms, number of bathrooms, the year the house was built, and date of sale. Various regression models such as multiple linear regression, k-nearest-neighbors, tree-based methods, and nonlinear regression techniques like splines are explored and compared to find an appropriate fit. His results showed that generalized additive models would perform best.

Pow, Emil and Liu (2014) [3] analyzed the real estate property prices in Montreal with respect to geological area, living region, and several rooms, and even geographical features such as the nearest police station and fire station. They apply and compared regression methods such as linear regression, Support Vector Regression (SVR), k-Nearest Neighbours (kNN), and Regression Tree/Random Forest Regression. In their paper, they predicted the asking price with an error of 0.0985 using an ensemble of kNN and Random Forest algorithms. Moreover, where applicable, the final price sold was also prophesy with an error of 0.023 using the Random Forest Regression. They presented the details of the analysis of the real estate listings, and the testing and validation results for the different algorithms in this paper. Besides, they were also discussing the significance of their approach and methodology.

3 Machine Learning Methodology

3.1 Data Collection

The process of data collection relies on the type of project. The dataset can be gathered from different sources such as a file, database, sensor, and many other sources. We can also use different datasets present on the internet like Kaggle and UCI Machine learning Repository. Kaggle is one of the most popular websites used for evaluating ML algorithms, and they also arrange competitions in which

different people can participate and get the opportunity to test their knowledge of ML. In this paper, we use a dataset from Kaggle. [9]

3.2 Data Pre-Processing

Data pre-processing is a technique of cleaning the raw data. When we collect the data from different sources in a raw format, and this data is not feasible for the analysis. Then some steps are executed to convert the data into a clean dataset, this part of the process is called as data pre-processing [5]. Messy data can be categorized as missing or noisy/inconsistent data. Some pre-processing methods can be applied to cleaning raw data such as:

- **Data conversion.** In this process, categorical and ordinal data must be somehow converted into numeric features while the missing values can be temporarily ignored. This process is sufficient if the data were missing in a row or column more than 70%.
- **Filling the missing values.** We can fill the lost data automatically, like by using the mean, median, or highest frequency value in the dataset.

3.3 Data Analysis

Data analysis is the technique of analyzing raw data in order to extract meaningful information. Any type of information can be used to optimize the overall efficiency of a system or human activity. Some steps like grouping, collecting and cleaning data can be helpful for the effective data analysis process. [2] A few basic types of data analysis can be mentioned as follows:

- **Descriptive analysis** describes what has happened in a given period.
- **Diagnostic analysis** focuses more on why something happened.
- **Predictive analysis** predicts what is likely going to happen in the near term.

3.4 Data Modelling

For data modelling, we summarize the shape of the dataset, then visualize it with summary statistics to get the mean, standard deviation, min, max, cardinality, quantile, and a preview of the dataset. Then we build unique views with window functions, filtering, binning, and derived columns. Finally, we design a random or stratified sampling plan to generate datasets for model training and scoring [4].

- **Model Selection** process where we can configure and train the model as a model selection. In each iteration, we deal with a new model we could choose to use or to modify, and the choice of a machine learning algorithm is part of that model selection process. All the possible existing models for a problem, a given algorithm and algorithm configuration on the chosen training dataset will provide a finally selected model.

– **Train Models** is a process depending on the dataset. For training a model, we initially divided the model into three sections, which are Training data, Validation data, and Testing data. For training, the classifier use the training dataset, tune the parameters using the validation set and finally test the performance of the classifier on an unseen test dataset. The important point is that during training, the classifier is available only for the training and for the validation set. The informational test index may be accessible during testing the classifier.

3.5 Analysing Model Performance

The model performance is an integral part of the model development process in a ML approach. It helps to find the best model representing the perfect result of our dataset and help us to choose a model working properly in the future.

4 Implementation Details

4.1 Analyze of the dataset

We can analyze the data in a couple of ways. We need to know the dimensions that mean how many instances(rows) and how many attributes (columns) contain a data file. In our train.csv file, we have 1259 rows and 81 columns. In our test.cvs file, we have 201 rows and 80 columns. As the goal is to predict the House Price we consider "Sales Price" as our independent variable and remove it from the dataset. The intended ML algorithm will predict the price and will compare it against this independent column.

Different variables can be analyzed using different visual means. Categorical variables like 'Neighbourhood' and 'Yearsold' are best visualized with a bar graph. However, numerical variables should be plotted in a line graph against Sales Price, which, in this case, shows that our data set is right-skewed.

4.2 Feature engineering

Real-world data can be messy and disorganized. A processed called feature engineering can reveal extra attributes or features from raw data by using data mining techniques.

– **Data Correlation** Correlation is a statistical process where we can see how strongly dependent variables are independent related in the dataset. That means how our features correspond with our output [1]. Now the following figure shows the Correlation variable with our target variable (sale price).
– **Scatter Matrix Plot** shows the correlation between pairs of variables. It can be helpful to find a predictable relationship between them.
– **Handling Outliers** As shown from the scatter plot, there are various outliers. For example, there is a two-house size more than 12000 sq with unusually low sale price. We have to remove all such abnormal data points. [6].

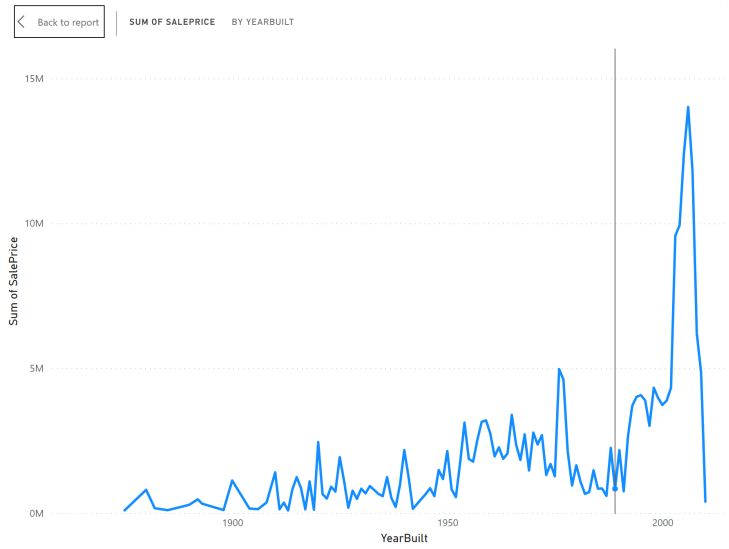

Fig. 1. Distribution of Sales Price as per Year Built

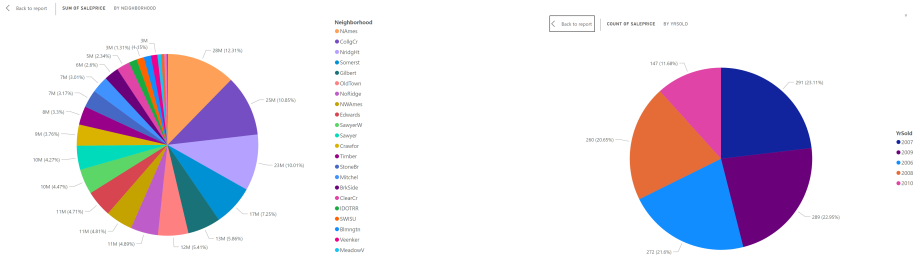

Fig. 2. Bar graph distribution of Neighbourhood and Yearsold columns

4.3 Handling missing values

As can be seen, this dataset contains a lot of missing values. Columns like 'PoolQC', 'Alley' and 'MiscFeature' where more than 99.5% values are missing can be dropped as they do not have any significant information. Column 'ID' contains the sequence number of each row and is also not useful. After that we can group the data by neighborhood and fill in the missing values of a certain attribute with its median per neighborhood.

4.4 Preparing Data for Prediction

Here we separate the features and the target variable for modeling. We will assign the features to X and the target variable (Sales Price) to y and take the target variable into Y. Next, we split data into train and test sets, accounting for 70% and 30% of the total data, respectively.

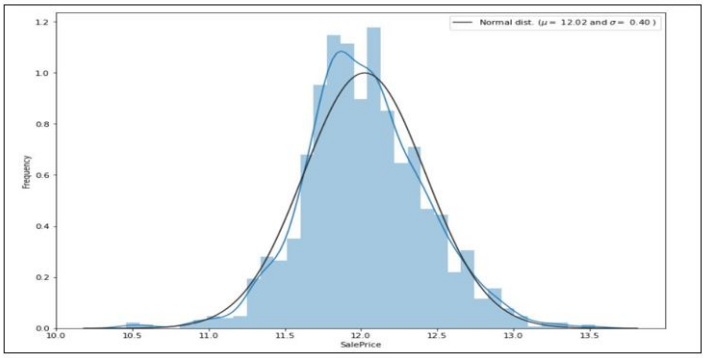

Fig. 3. Show the Distribution Plot after log transformation.

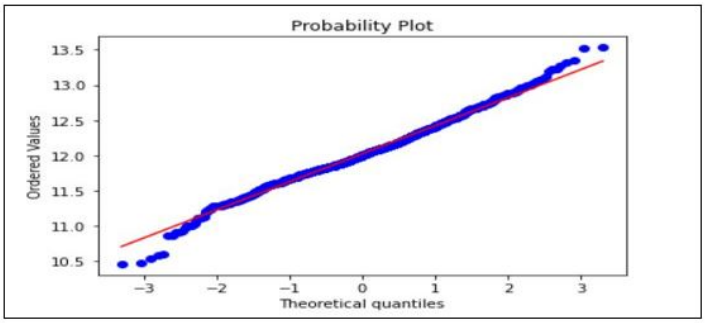

Fig. 4. Show the Probability Plot after log transformation.

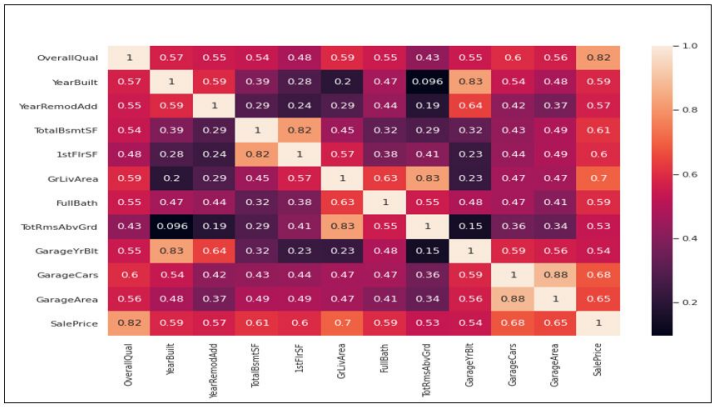

Fig. 5. Showing correlation where variable are correlated more than 50% with a target variable.

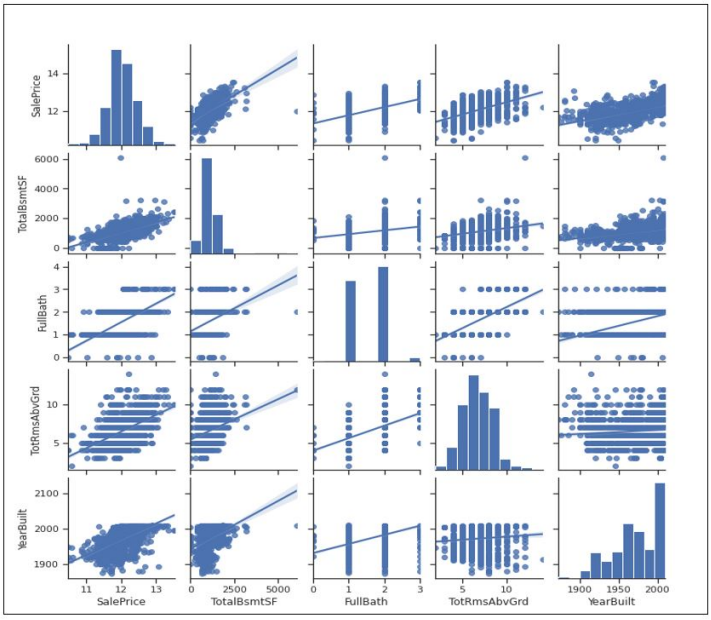

Fig. 6. Scatter Matrix Plot for some Input Variable.

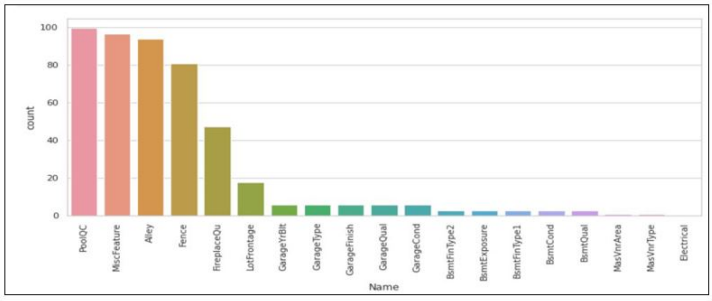

Fig. 7. Show Bar plot for every column which have missing value.

5 Experimental Results and Statically Analysis

Experimentation frequently produces various estimations of something very similar, for example, duplicate estimations. Statistical investigation can be utilized to sum up, those perceptions by evaluating the normal, which gives a gauge of the genuine mean. This section is based on the experimental data results and some statistical analysis.

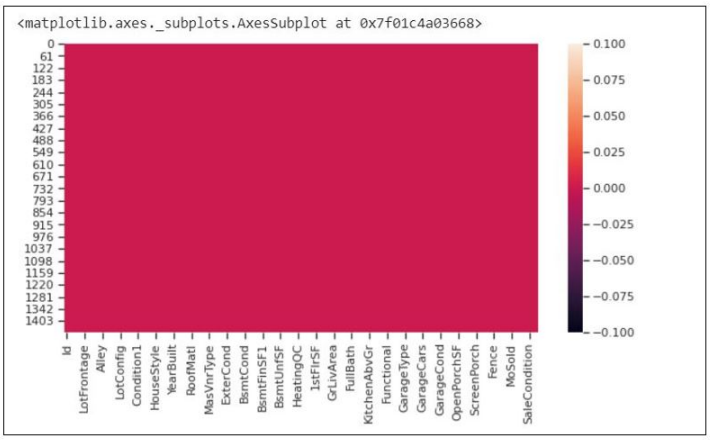

Fig. 8. Heatmap shows we have no missing value in dataset after handling.

5.1 Test of Goodness of Fit via Predicted vs. Actual Plot

The scatter plot is one of the more reliable forms of data visualization. Here it shows how accurate our model prediction is for the predicted price vs. the actual price and how the model is a good fit. As can be seen from all plots, the results are excellent, and most predicted values are close to actual ones. Hence, it can be concluded that = the model is a good fit and Gradient Boosting Regression is a slightly better than from Linear Regression and Random Forest Regression.

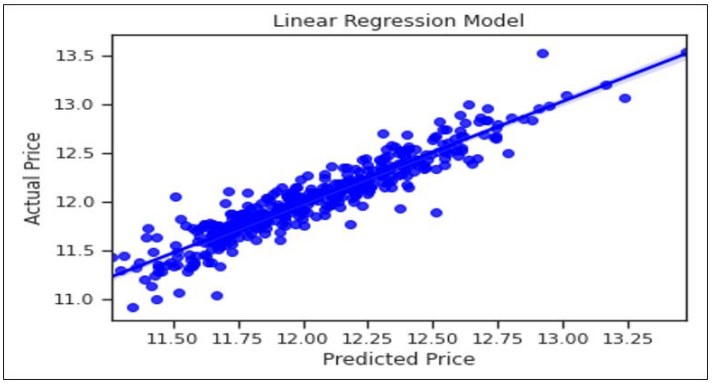

Fig. 9. Predicted vs. Actual Plot for Linear Regression.

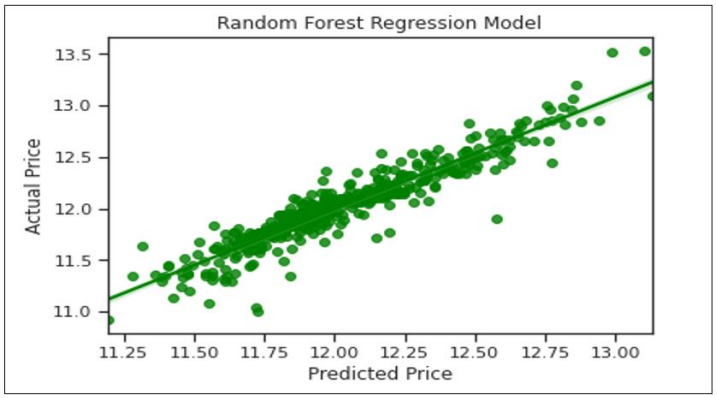

Fig. 10. Predicted vs. Actual Plot for Random Forest Regression.

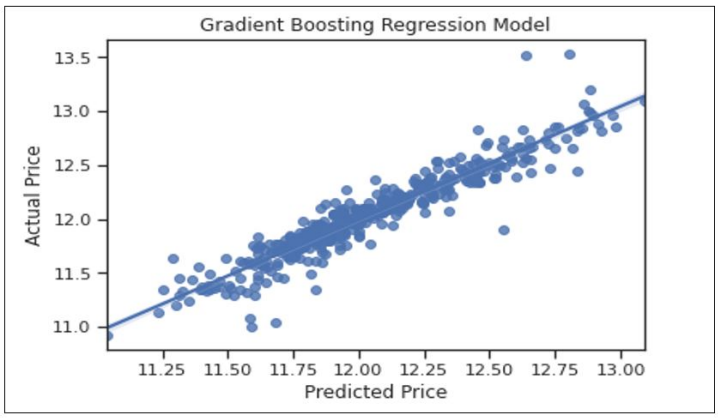

Fig. 11. Predicted vs. Actual Plot for Gradient Boosting Regression.

5.2 Experimental Score and result based on the Evaluation

The table below summarizes and compares our results based on the Accuracy score, R^2 score, MAE score, MSE score, and RMSE Score in demand forecasting using the Linear Regression, Random Forest Regression and Gradient Boosting regression algorithms. The results clearly show that Gradient Boosting outperform all others.

6 Conclusion and Future Work

In this paper, we have completed the goal of predicting house prices by using different machine approaches, including Linear Regression, Random Forest, and Gradient Boosting. We have provided a detailed process, starting with analyzing

Table 1. Experimental result based on the Evaluation

	Linear Regression	**Random Forest**	**Gradient Boosting**
Accuracy	87.56281245499873	88.09710339228504	89.71355497536405
R^2	0.8756281245499874	0.8809710339228504	0.8971355497536404
MAE	0.10410390943855012	0.0957662809858385	0.08787608733016163
MSE	0.02037090699516133	0.019495790257359046	0.016848199333612097
RMSE	0.14272668634548105	0.13962732632747446	0.12980061376438903

the dataset to generating the regression model. In addition, we visualize each model's performance using different performance metrics and compared them based on these metrics. We found that Gradient Boosting regression gives in our case the highest accuracy of 89.53% with an excellent performance.

For future work, we can try to deal with a considerably larger dataset. This would yield a better and genuine picture of the model and to identify a set of optimal hyperparameters for a learning approach. We have embraced just a few Machine Learning approaches regarding classifiers; however, we have to enhance various classifiers and comprehend their anticipating conduct for nonstop qualities as well. By improving the error esteems, this exploration work can be valuable for enhancing the utilization of different individual urban communities.

References

1. Bock, T.: What is correlation. In: Archived by online at `https://www.displayr.com/what-is-correlation/`. Displayr Blog (2019)
2. Frankenfield, J.: Data analytics. In: Archived by online at `https://www.investopedia.com/terms/d/data-analytics.asp`. Investopedia (2019)
3. Liu, Emil Janulewicz, N.P.: Applied ml project for a forecast of real estate property prices in the city montreal. In: Archived by online at `http://rl.cs.mcgill.ca/comp598/fall2014/comp598_submission_99.pdf`. Semantic Scholar (2014)
4. McLaurin, J.: Data modeling machine learning data sets. In: Archived by online at `https://dzone.com/articles/data-modeling-machine-learning-datasets`. AI Zone (2017)
5. Pant, A.: Workflow of a machine learning project. In: Archived by online at `https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94`. Towards Data Science (2019)
6. Santoyo, S.: A brief overview of outlier detection techniques. In: Archived by online at `https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561`. Towards Data Science (Sep 2017)
7. Towne, A.K.: Modles and visualizations for housing price prediction. In: Archived by online at `https://broncoscholar.library.cpp.edu/bitstream/handle/10211.3/185729/KomagometowneAnh_Thesis2016.pdf?sequence=4` (2016)
8. Wu, J.Y.: Housing price prediction using support vector regression. In: Archived by online at `https://scholarworks.sjsu.edu/cgi/viewcontent.cgi?referer=https://www.google.com/&httpsredir=1&article=1540&context=etd_projects` (May 2017)
9. ©2020KaggleInc: House prices: Advanced regression techniques. In: Archived by online at `https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data` (August 2016)