All Sciences Proceedings

1st International Conference on Recent and Innovative Results in Engineering and Technology

August 16-18, 2023 : Konya, Turkey

ICRIRET

All Sciences Proceedings
http://as-proceeding.com/

https://as-proceeding.com/index.php/icriret

House Price Prediction Using Ensemble Learning Techniques

Hasan Ulutaş[1,*], M. Emin Sahin[1]

[1]*Department of Computer Engineering, Yozgat Bozok University, Turkey*
[*]*(hasan.ulutas@bozok.edu.tr) Email of the corresponding author*

Abstract – With the rapid increase in the world population, people's demands for houses are increasing day by day. Housing markets have an important place in the economy of a country, and this is also an important indicator of the welfare level of society. Accordingly, it is necessary to examine and change the prices of the houses in the housing sector in detail. From this point of view, in this study, price estimation will be made using machine learning techniques by making use of the public dataset containing information about houses with different characteristics in the district of California. The resulting house price estimation will be developed with ensemble learning techniques. According to the results obtained, the AdaBoost Regressor ensemble model is obtained with the best performance value of 0.118 RMSE. Finally, the project will be integrated into the web interface for house price estimation.

Keywords – Ensemble Model; Price Estimation; Machine Learning; Web Interface

I. INTRODUCTION

Today, the real estate sector and the housing sector are at the forefront of the industries in which individuals invest the most [1]. On the residential market, sellers seek the highest possible price, while buyers seek the most features at the lowest possible price. In this instance, a number of issues arise during the purchase and sale of the subject home, and the standard pricing policy cannot be applied. Numerous real estate agents and investment consultants in the housing industry charge their clients hefty commissions. This circumstance drastically diminishes the sector's confidence in and demand for real estate agents. In addition, the incorrect calculation of home prices is one of the leading issues in the housing industry.

In today's rapidly changing world, driven by technological advancements and improved living conditions, houses have evolved into personal havens where individuals can find happiness, tranquility, and the freedom to develop their personalities while freely expressing themselves in every way. As a result, it has become increasingly important for individuals to acquire a home that exceeds the traditional concept of a mere shelter and fully meets their needs. During the process of locating a home with desired features, it is crucial to accurately analyse the budget and prices in relation to the location and desired features. Those seeking real estate for investment purposes must also have access to applications that provide accurate price analysis. Also of great importance to insurance companies is the estimation of housing prices, highlighting the significance of applications that facilitate accurate price analysis for these organisations. In today's market, comprehensive tools and applications that enable individuals, investors, and insurance companies to conduct accurate price analysis are indispensable. Individuals can make informed property acquisition decisions with these tools, investors can evaluate potential investment opportunities, and insurance companies can accurately assess property values for insurance purposes.

With the successes achieved in recent years, machine learning methods have started to gain popularity in different fields and their application areas are gradually expanding. With this expansion, one of the main application areas of machine learning has been predictability and it has

been used on more optimized prediction models in order to obtain the closest prediction values with machine learning [2]–[7]. In addition, prediction studies made with machine learning techniques are also used in application development processes in many sectors. With the successful results obtained in these prediction studies, the use of machine learning techniques in the field of estimation increases and spreads. To predict housing prices, Quang et al. propose complex machine learning models. Experimenting with both conventional and advanced machine learning models, they examine the effects of complex models [1]. Sifei Lu et al. developed a hybrid Lasso and gradient increment regression model to forecast individual house prices [8]. For the proposed method, a dataset from the Kaggle platform was employed. The outcome indicated that hybrid regressions were superior to Ridge, Lasso, and Gradient boosting regressions. Using a combination of 65% Lasso and 35% gradient boosting, the best hybrid regression result for the test data is 0.11260. In addition, it provides a comprehensive validation of multiple techniques in the application of regression models and an optimistic result for home price forecasting. Chen and colleagues introduced a novel approach to address the bankruptcy prediction challenge through a machine learning framework. In their proposed method, the problem is approached from a learning standpoint. The available training data is organized into distinct bags, each containing various instances. Notably, the individual instances within these bags lack specific labels, and the only information provided is the overall ratio of samples associated with a particular class at the bag level [9]. The authors then present two new estimation techniques, Bagged-pSVM and Boosted-pSVM, which are based on proportional support vector machines and ensemble strategies including bagging and amplification. Extensive testing on benchmark datasets demonstrates that the proposed ensemble methods can effectively solve the problem of predicting corporate bankruptcy.

The aim of this study will be to estimate the price of a house in California city where different properties are entered by hybridizing machine learning techniques with community learning techniques. However, before the hybridization process, various machine learning algorithms are used and prediction results are obtained. Prediction results are evaluated and models to be hybridized is selected. After the model training, the application will be transferred to a web interface and a house price estimation is made according to the given house features.

II. Materials and Method

A. Dataset

This study utilised a dataset comprised of publicly accessible training and test data downloaded from the Kaggle website. The dataset contains up-to-date information on houses with various properties in California city. Both the training and test datasets include 81 features and 1460 data. Here, 3 properties of the data set are of the float data type, 35 properties are of the integer data type, and 43 properties are of the object data type. An overview of data type is shown in Figure 1.

```
RangeIndex: 1460 entries, 0 to 1459
Data columns (total 81 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Id             1460 non-null   int64
 1   MSSubClass     1460 non-null   int64
 2   MSZoning       1460 non-null   object
 3   LotFrontage    1201 non-null   float64
 4   LotArea        1460 non-null   int64
 5   Street         1460 non-null   object
 6   Alley          91 non-null     object
 7   LotShape       1460 non-null   object
 8   LandContour    1460 non-null   object
 9   Utilities      1460 non-null   object
 10  LotConfig      1460 non-null   object
 11  LandSlope      1460 non-null   object
 12  Neighborhood   1460 non-null   object
 13  Condition1     1460 non-null   object
 14  Condition2     1460 non-null   object
 15  BldgType       1460 non-null   object
 16  HouseStyle     1460 non-null   object
 17  OverallQual    1460 non-null   int64
 18  OverallCond    1460 non-null   int64
 19  YearBuilt      1460 non-null   int64
show more (open the raw output data in a text editor) ...
 78  SaleType       1460 non-null   object
 79  SaleCondition  1460 non-null   object
 80  SalePrice      1460 non-null   int64
dtypes: float64(3), int64(35), object(43)
memory usage: 924.0+ KB
```

Fig. 1. An overview of data type

B. Feature Extraction

Feature extraction is a method used in the fields of machine learning and data analytics, and it aims to make the complex structures in data sets more meaningful and workable. In this method, different techniques are used to improve the representation of the input features in the data set. It is used in many areas such as feature extraction, size reduction, pattern recognition and data visualization. The purpose of feature extraction is that the results to be obtained from machine learning algorithms are closer to reality. The numpy, seaborn, pandas, matplotlib.pyplot, scipy

python libraries required for feature extraction, machine learning and modeling have been added to the environment. For the dataset used in this study, how many features the dataset consists of and the types of these features are checked. Then, a distribution is created for the 'SalePrice' feature in the dataset and the relationship between the probability state and the probability is evaluated graphically. It is observed in the graph that the distribution is skewed to the right. A logarithmic transformation is applied to the 'SalePrice' property to transform this distribution into a closer to normal form. Incomplete data will adversely affect the forecast score while making a price estimation. Therefore, missing data is filled with 'none'. In order to look at the correlation graph of the data set, only 38 numerical features are separated from 81 features. The non-informative 'Id' attribute is deleted and the correlation graph is examined. After feature extraction, the target variable 'SalePrice' is assigned to a new variable to use the data in prediction, and the target variable to be predicted is deleted from the data set. The dataset is divided into two as 80% training and 20% testing. Correlation graph of study is given in Figure 2.

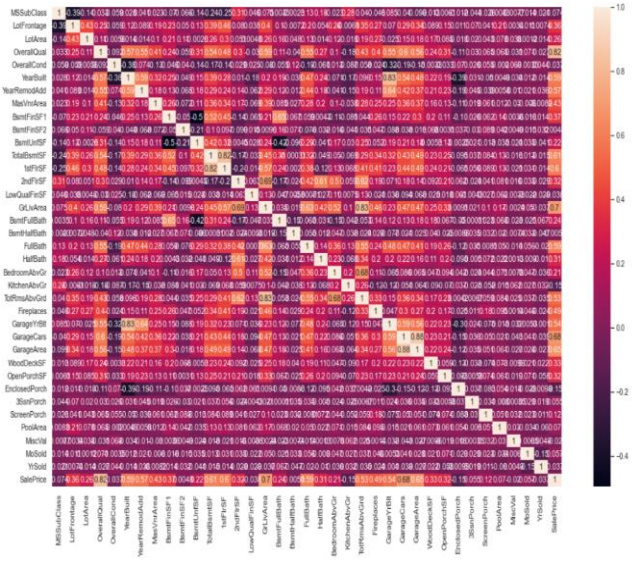


Fig. 2. Correlation graph of study

C. *Machine Learning*

Machine learning is a field of artificial intelligence where computer systems can learn from data and extract knowledge from experiences. This method allows algorithms to identify patterns and make future predictions by analyzing large data sets. By using techniques such as machine learning, statistical methods, deep learning and natural language processing, it achieves effective results in many tasks such as classification, regression, clustering and pattern recognition [10]. In the study, after completing the feature selection and extraction processes, home price estimation is performed with three different machine learning algorithms. These algorithms are Linear Regression, Random Forest algorithm and Gradient Boosting Regression algorithm. After these models, ensemble models are used and the results are analyzed.

D. *Ensemble Models*

Ensemble models are a method used in the field of machine learning and aim to create a stronger and more stable model by combining multiple learning algorithms. These models aim to obtain more accurate results by combining the predictions of different algorithms. Ensemble models are used with techniques such as bagging, boosting and stacking, and they both increase the stability of the model by reducing the variance and provide more accurate results by reducing the bias [11]. Ensemble models can be created in a variety of ways; bagging, boosting, voting and stacking.

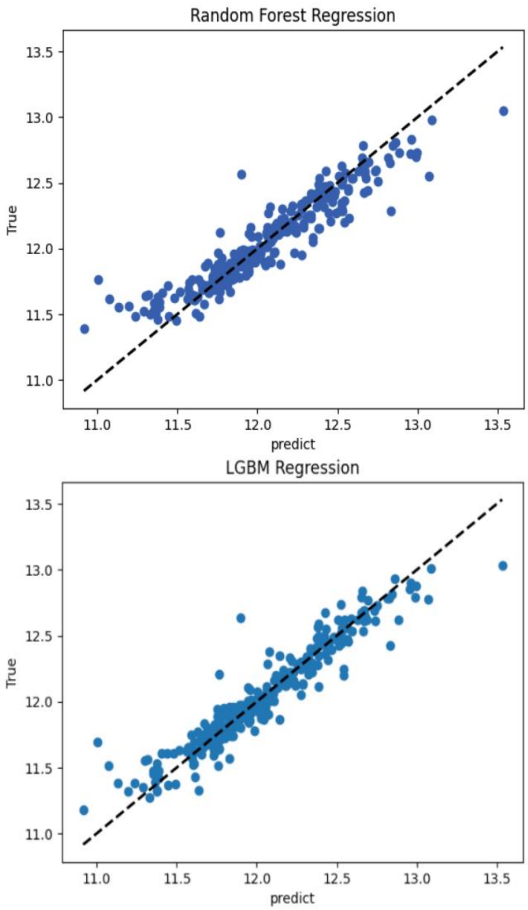
III. RESULTS

In this study, machine learning models and ensemble learning models created with these models are applied on the selected dataset. The models used and the evaluation metrics of these models are given in Table 1. The trained models are analyzed according to the coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) criterion values and are shown in Table 1.

Table 1. The evaluation metrics of these models

Algorithms	R^2	RMSE	MAE
Lineer Regression	0.887	0.137	0.104
Random Forest R.	0.863	0.151	0.102
LGBM Regressor	0.899	0.129	0.087
Gradient Boosting R.	0.913	0.120	0.084
XGB Regressor	0.902	0.127	0.088
Voting Regressor	0.914	0.119	0.081
Bagging Regressor	0.894	0.132	0.089
MLxtend-Stacking R.	0.914	0.119	0.084
AdaBoost Regressor	0.916	0.118	0.082

Among the models compared, it is observed that the best models are AdaBoostRegressor and VotingRegressor. When evaluated in terms of coefficient of determination (R2) values, it is seen that four algorithms have values of 0.91 and above. When examining these four algorithms in terms of mean square error (RMSE) and mean absolute error (MAE) criterion values, it is seen that the best algorithm is AdaBoostRegressor. The RMSE value, which is the most commonly used metric in terms of comparing the values of machine learning models in terms of some performance metrics, is shown in Figure 3. The comparison table of the RMSE values of the models is shown in Figure 4. The distribution chart of the ensemble models used in the study is given in Figure 5.



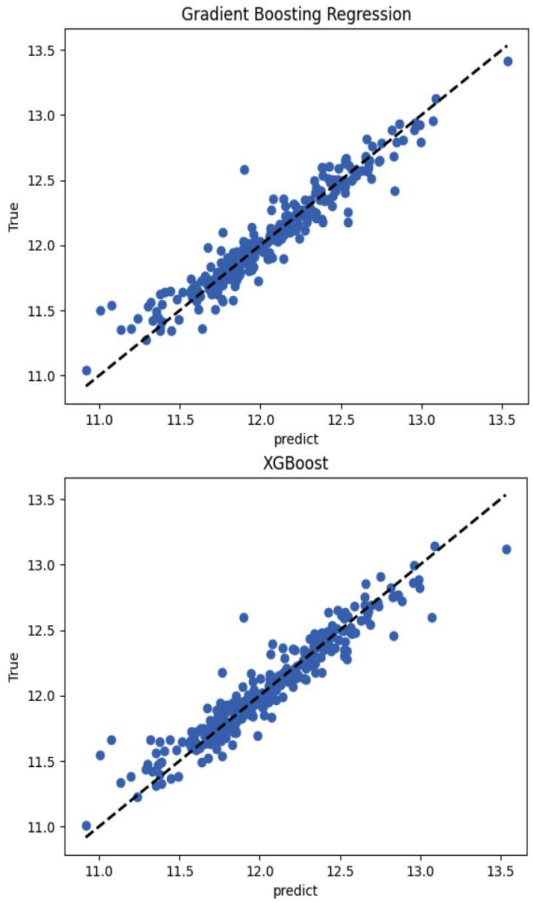


Fig. 3. Scatter plot of machine learning algorithms

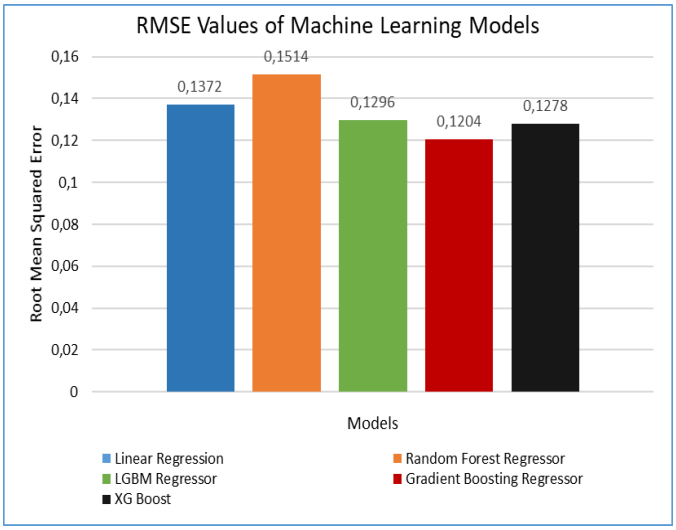


Fig. 4. The comparison table of the RMSE values of the models

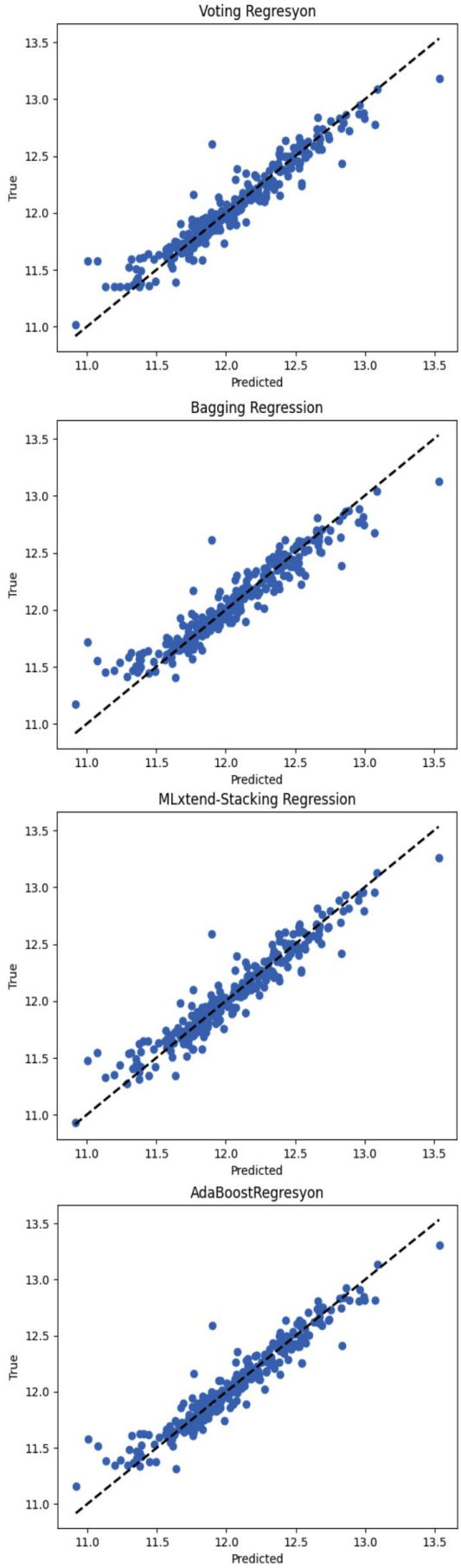


Fig. 5. Scatter plot of ensemble models

Within the scope of the study, web interface design is carried out for the home price estimation system. The model to be used for the interface program is selected as AdaBoostRegressor and the model is converted to ".pkl" file format so that the prediction process can be performed at any time. Then, the home features to be taken from the user are determined and the appropriate web interface is designed. In the interface shown in Figure 7, the user will select the desired house features and send this information to the "backend" side to work on the model. An average price estimate will be obtained by running the selected features on the model.

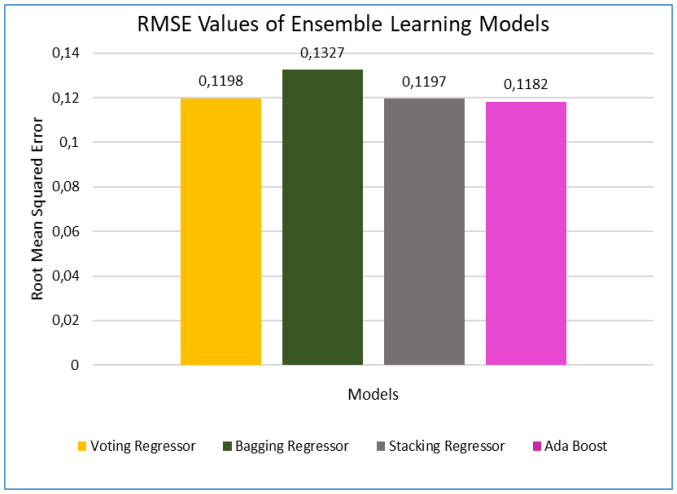


Fig. 6. Comparison of RMSE values of ensemble learning models

Fig. 7. Web user interface

IV. CONCLUSION

The rental housing market holds significant importance within the real estate sector and has garnered considerable focus. In practical terms, precise prediction models for rental prices play a crucial role in assisting property owners in effectively setting rental property prices, while also aiding tenants in locating affordable living spaces. In this study, a home rental price prediction model

was developed with machine learning algorithms. This research study analyzes home sales prices to predict customers' potential purchases. House price estimation for the application was carried out with machine learning algorithms and ensemble models. These proposed models were analyzed using a publicly available dataset in the literature. According to the results obtained, the AdaBoost Regressor ensemble model was identified with the best performance value of 0.118 RMSE. In the study, a website was also created to estimate the user's house price.

REFERENCES

[1] Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing price prediction via improved machine learning techniques. Procedia Computer Science, 174, 433-442.

[2] Joshi, I., Mudgil, P., & Bisht, A. (2022, November). House Price Forecasting by Implementing Machine Learning Algorithms: A Comparative Study. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3 (pp. 63-71). Singapore: Springer Nature Singapore.

[3] Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert systems with applications, 42(6), 2928-2934.

[4] Thamarai, M., & Malarvizhi, S. P. (2020). House Price Prediction Modeling Using Machine Learning. International Journal of Information Engineering & Electronic Business, 12(2).

[5] Ulutaş, H., & Aslantaş, V. (2023). Design of Efficient Methods for the Detection of Tomato Leaf Disease Utilizing Proposed Ensemble CNN Model. Electronics, 12(4), 827.

[6] Sahin, M. E. (2023). Image processing and machine learning-based bone fracture detection and classification using X-ray images. International Journal of Imaging Systems and Technology, 33(3), 853-865.

[7] Sahin, M. E. (2023). Real-Time Driver Drowsiness Detection and Classification on Embedded Systems Using Machine Learning Algorithms. Traitement du Signal, 40(3), 847.

[8] Lu, S., Li, Z., Qin, Z., Yang, X., & Goh, R. S. M. (2017, December). A hybrid regression technique for house prices prediction. In 2017 IEEE international conference on industrial engineering and engineering management (IEEM) (pp. 319-323). IEEE.

[9] Chen, Z., Chen, W., & Shi, Y. (2020). Ensemble learning with label proportions for bankruptcy prediction. Expert Systems with Applications, 146, 113155.

[10] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

[11] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.