Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/55/20241483

House price prediction using machine learning for Ames, Iowa

Qiongwei Ye

College of Computing, Georgia Institute of Technology, GA 30332, Atlanta, Georgia, USA

qye47@gatech.edu

Abstract. The real estate sector is pivotal to economic growth and plays a substantial role in the global GDP. In this technologically advanced age, the adoption of machine learning for accurate house price prediction is crucial. These models optimize decision-making for homeowners, sellers, and investors alike. This study represents a comprehensive exploration into the field of house price prediction within the context of Ames, Iowa, United States. The primary objective of this research is to construct a reliable and highly accurate predictive model, empowering individuals to estimate property values with unprecedented precision. The research encompasses three different machine learning algorithms: linear regression, random forest, and XGBoost. The use of the dataset from the reputed website helps improve the reliability of the result. Furthermore, the investigation extends to a detailed examination of the multifaceted determinants exerting a profound influence on house prices in the dynamic Ames real estate landscape, and determined that the factor that will influence the house price most is the total area of the house. Among all models, XGBoost produces the best result, which achieved an R-square of 0.8803. Moreover, the importance of each feature is also analyzed using the feature ranking algorithm in random forest, showing that the overall quantity of the house, the living area of the house, and the total area of the basement are the top three factors that influence the house price most.

Keywords: Machine learning, house price prediction, linear regression model, random forest, XGBoost.

1. Introduction

In the modern real estate landscape, accurate house price assessment has become a topic of paramount importance. The value of research in this domain extends beyond mere economic considerations, touching on broader societal and policy implications. Understanding the factors influencing house prices and developing robust evaluation methodologies can have profound significance.

Current research has made substantial strides in this area, employing advanced statistical techniques and computational models. However, there remain gaps and challenges that necessitate further exploration and refinement in the methodology, which needs more researchers to train the models and to increase the result's accuracy.

In the context of Ames, Iowa, the significance of this research becomes even more palpable. Ames, Iowa, though not as globally recognized as some cities, holds its charm and significance. Beyond its educational landmarks, primarily due to the presence of Iowa State University, Ames plays a pivotal

DOI: 10.54254/2755-2721/55/20241483

role in the region's economy. The consistent demand for housing, driven by a steady flow of students, faculty, and residents, has put a spotlight on housing prices in the area. Various factors, from location and number of bedrooms to amenities like garages or gardens, influence these prices, emphasizing the importance of detailed research.

Until 2020, Ames accommodated approximately 66427 residents and Iowa State University had a student population of 27854 as of spring 2023 [1-2] According to the study, the housing market in Ames, Iowa is a seller's market. The prices of homes in Ames have increased by 9.4% compared to last year [3]. This means the housing market is still very large in Ames. Therefore, the aim is to provide valuable and accurate insights about house prices to help people make wise decisions in this dynamic market.

This research embarks on a quest to address the pivotal question: Can one predict the price of a residence in Ames with effectiveness and great precision? The importance of precise price predictions is profound, as these forecasts hold the potential to transform the decisions of a multitude of stakeholders. These stakeholders encompass not only prospective homeowners seeking to realize their dream of settling in the city of Ames but also existing property owners eager to estimate the value of their assets and thoughtful investors seeking to capitalize on the opportunities inherent in the dynamic real estate ecosystem. Thus, an accurate model that has predicting techniques is indeed very important.

Machine learning can be categorized into three types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is the one that intends to identify an algorithm that is capable of giving accurate predictions. Previous research indicates that several models are commonly employed for predicting house prices, including linear regression, k-nearest neighbors, random forest, artificial neural networks, and XGBoost.[1]. Some of these methods will be implemented in this research to help predict the house price.

The organization of the paper is as follows: the methodologies that will be utilized in this paper in section 2, including the pre-processing of the data and three machine learning models. Section 3 will illustrate the results produced by the models. Finally, the discussion is in section 4 and the conclusion is in section 5.

2. Methodologies

In this research, initial data preprocessing occurs on the input data, followed by exploratory data analysis on the dataset. Subsequently, several machine learning models, including Linear Regression (LR), Random Forest (RF), and XGBoost Regression (XGB), are constructed and trained to procure results for further analysis.

2.1. Data

The model starts with the dataset about house price and features in Ames, Iowa from Kaggle [4], which contains 1121 records of data and 27 features that can use to train our model. The target feature for prediction is the house price. The independent variables are those 27 features, such as square meters, number of rooms, city part range, number of previous owners, the year it made, etc.

2.1.1. Data preprocessing

To guarantee the optimal performance of the machine learning models, it's imperative to commence with a clean, well-prepared dataset. Therefore, a thorough examination of the entire dataset is undertaken, specifically searching for any occurrences of null counts that could potentially impede or distort the training process. (Table 1.) Given that all the features within the dataset are intrinsically tied to house prices and are represented as numerical data, it is decided to utilize all of them. This comprehensive approach ensures that the richness of the data is preserved and harnessed when training the models, potentially leading to more accurate and robust predictions.

Table 1. Attributes in the Dataset

#	Feature Name	Non-Null Count	Data Type
0	LotFrontage	1121	Float64
1	LotArea	1121	Int64
2	OverallCond	1121	Int64

DOI: 10.54254/2755-2721/55/20241483

Table 1. (continued)

3	YearBuilt	1121	Int64
4	YearRemodAdd	1121	Int64
5	MasVnrArea	1121	Float64
6	BsmtUnfSF	1121	Int64
7	TotalBsmtSF	1121	Int64
8	1stFlrSF	1121	Int64
9	2stFlrSF	1121	Int64
10	LowQualFinSF	1121	Int64
11	GrLivArea	1121	Int64
12	BsmtFullBath	1121	Int64
13	BsmtHalfBath	1121	Int64
14	FullBath	1121	Int64
15	HalfBath	1121	Int64
16	BedroomAbvGr	1121	Int64
17	KitchenAbvGr	1121	Int64
18	TotRmsAbvGrd	1121	Int64
19	Fireplaces	1121	Int64
20	GarageYrBlt	1121	Float64
21	GarageCars	1121	Int64
22	GarageArea	1121	Int64
23	WoodDeckSF	1121	Int64
24	OpenPorchSF	1121	Int64
25	EnclosedPorch	1121	Int64
26	PoolArea	1121	Int64
27	YrSold	1121	Int64
28	SalePrice	1121	Int64

After the data pre-checking, a distribution graph (Figure 1.) of the house price is constructed using the sale price in the dataset, accompanied by a heat map (Figure 2.) implemented by using the features in the dataset.

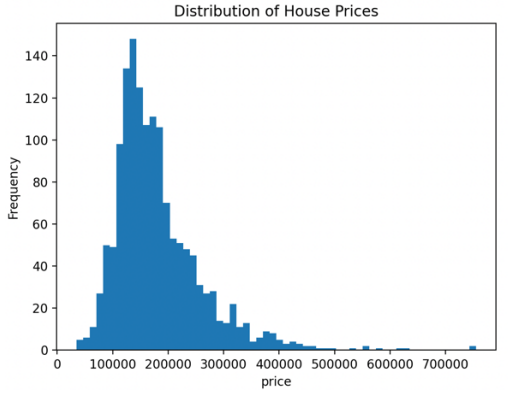

Figure 1. House price distribution

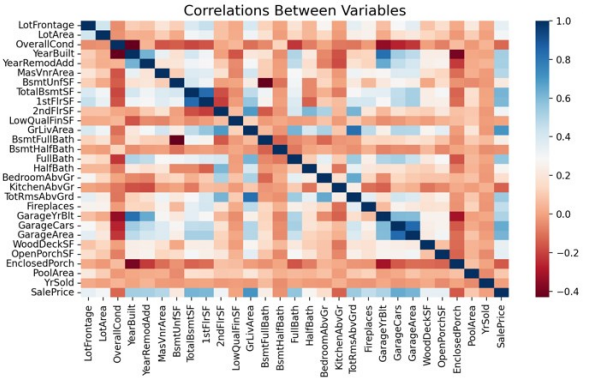

Figure 2. Heat map

The correlation graph between each feature and the sale price of the housing is also calculated and plotted in the figures below (Figure 3. 4. 5. 6.).

DOI: 10.54254/2755-2721/55/20241483

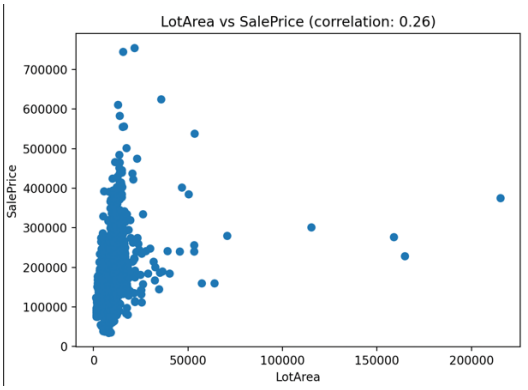

Figure 3. LotArea with SalePrice

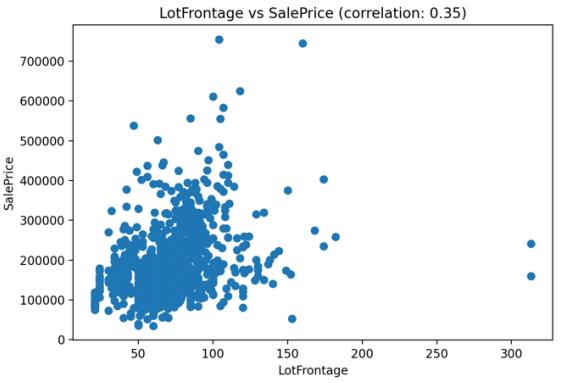

Figure 4. LotFrontage with SalePrice

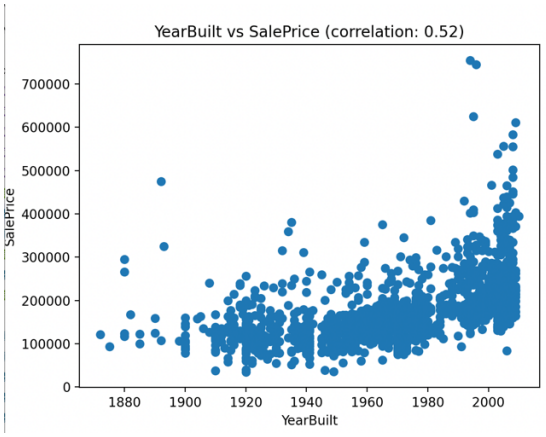

Figure 5. YearBuilt with SalePrice

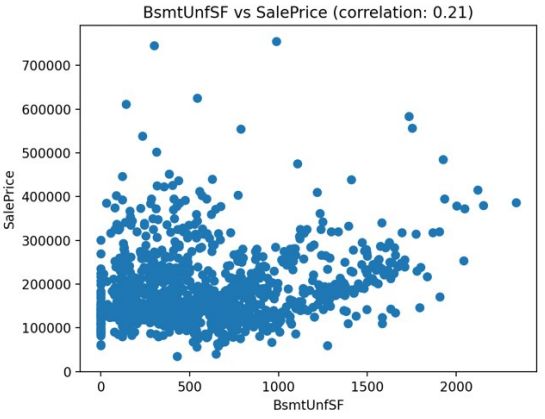

Figure 6. BsmtUnfSF with SalePrice

In order to train the models, the rows contained "NA" and duplicates were dropped from the dataset. The column "Id" was been popped as it is not needed for our training. Only the data with numerical data types in the data set were used. The dataset is partitioned into two parts: 20% of the data will be dedicated for training and the other 80% will be used for testing the model.

2.2. Linear regression

Linear regression is a technique used to determine the relationship between a dependent variable and one or multiple independent variables. The relationship calculated by the linear regression model can be used to predict the dependent variable by using different independent variables. It is a very efficient and effective model to help predict the relationships in machine learning.

Linear regression encompasses two principal forms: simple linear regression, applicable when there is only one independent variable, and multiple linear regression, employed when there are multiple independent variables. In the situation of house price prediction, because there are multiple features that may influence the house price, a multiple linear regression model will be deployed.

The target of this analysis will be the price of the house, which will be presented in the training dataset. The whole linear regression model aims to fit a curve to the provided dataset while minimizing errors [5].

The hypothesis of the linear regression is as follows:

$$h_\theta(X) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_n X_n = Y \quad (1)$$

where Y is the target, which is the price of the house, $X_1, X_2, \ldots, X_n$ are the independent variables or features, $\theta_0, \theta_1, \theta_2, \ldots, \theta_n$ are the weights of the independent variables $X_1, X_2, \ldots, X_n$.

DOI: 10.54254/2755-2721/55/20241483

2.3. Random forest

Random forest is a collective learning technique for both classification and regression tasks, aggregating predictions from numerous decision trees. The decision tree is a technique that finds the best way to divide the data and train the data through the Classification and Regression Tree (CART) algorithm. Single decision trees will have problems including bias and overfitting [6], but if we combine multiple decision trees, the algorithm will be secure and reliable, so random forest will provide more accurate results.

There are two types of ensemble methods: one is bagging and the other is boosting. Random forest uses bagging to combine each single decision tree. Bagging is a process that chooses a random subset from the dataset and generates each model from it. The final result of the random forest is coming from majority voting after combining the results from each subset [7]. Each subset is processed in parallel. This helps the random forest to have additional randomness when training the data and searching for the best features among the random subset it chose.

Random forest calculates the feature importance based on the decrease in the Gini impurity that each feature brings when it is used in the trees of the forest. In the Gini impurity, the significance of a variable is gauged by aggregating the impurity decreases at nodes where initiates a split, then dividing by the tree count. In classification contexts, Gini impurity is commonly employed:

$$\widehat{\Gamma(t)} = \sum_{j=1}^{J} \hat{\phi} j(t) \left(1 - \hat{\phi} j(t)\right) \tag{2}$$

with $\hat{\phi} j(t)$ denoting the frequency of class j at node t [8,9].

The steps of random forest in this paper are:

1. Data splitting: we split the data into two parts: training set and testing set.
2. Feature standardization: we utilize the StandardScaler from the sklearn library to make the features standardized, which set the value of each feature within the same range.
3. Implementing the random forest regression model.
4. Training the model using training dataset.
5. Using the test dataset to make the prediction of the house price.

2.4. XGBoost

XGBoost is a short form for extreme gradient boost regression. It is also an ensemble learning algorithm, which employs boosting to combine the multiple models it has. To make the final model more accurate, it combines the base learners into strong learns using sequential models [8]. It functions using the gradient boosting framework, a machine learning approach that generates a predictive model as a combination of several weak predictive models, usually decision trees.

A Gradient Boosting Decision Trees (GBDT) is a powerful ensemble learning method that leverages the concept of boosting to optimize a sequence of decision trees [10]. The objective function will be implemented to build the XGBoosts,

$$\text{Obj}(\Theta) = \sum_{i=1}^{n} l\left(y_i, y_i^t\right) + \sum_{i=1}^{t} \Omega(f_i) \tag{3}$$

where l is a convex loss function that can be differentiated and assesses the discrepancy between the observed and forecasted values. Ω serves as a regularization component that imposes a penalty on the model's complexity.This helps to keep the model general enough and prevent overfitting.

2.5. Evaluation Metrics

2.5.1. Mean Absolute Error (MAE)

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n} \tag{4}$$

where y_i denotes the true value, x_i represents the forecast value, and number of samples in the dataset is demonstrated by n. MAE takes the arithmetic average of the absolute errors $|e_i| = |y_i - x_i|$ by

DOI: 10.54254/2755-2721/55/20241483

calculating the sum of absolute errors divided by the sample size. The error is measured linearly in MAE, which makes it a useful metric for the model that has errors evenly distributed across the dataset.

2.5.2. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{N}(y_i - x_i)^2} \tag{5}$$

where y_i denotes the true value, x_i represents the forecast value, and number of samples in the dataset is demonstrated by n. RMSE is derived by taking the mean of the squared differences between the model's predicted values and the actual values from the dataset [11,12]. It serves as a metric to assess the efficacy of the regression model.

2.5.3. Root Mean Squared Logarithmic Error (RMSLE)

$$RMSLE = \sqrt{\left(log(y_i + 1) - log(x_i + 1)\right)^2} \tag{6}$$

where y_i denotes the true value and x_i represents the forecast value. RMSLE is calculated by taking the difference of the log of the actual and predicted values. It is a very advantageous metric because it is robust to outliers [13].

2.5.4. R-squared

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

where y_i denotes the true value, x_i represents the forecast value, and $\bar{y}$ is the sample mean of the actual value. $\sum_{i=1}^{n}(y_i - x_i)^2$ in the equation is the mean squared error, and $\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares. R^2 is the coefficient of determination and is calculated by one subtracted by unexplained variation divided by total variation [14].

3. Result

This research is conducted in the Python 3.10.9 environment, and all the models were implemented with Pandas, Scikit-Learn, Numpy, Seaborn, XGBoost packages imported.

3.1. Linear regression

The LinearRegression algorithm from the sklearn.linear_model package, which contains ordinary least square linear regression, is utilized. The intercept and coefficients are listed in Table 2 below. The intercept represents the predicted value of the dependent variable when all independent variables are set to zero. Each value in the coefficient array indicates the variation in the outcome variable for a unit increment in the respective predictor, keeping all other variables unchanged. The coefficients are corresponded to the features in the left column of the table.

Table 2. Intercepts and Coefficients for Linear Regression

Intercept	-110971.28643378298
Features	Coefficients array
['LotFrontage',	[4.91270670e+01,
'LotArea',	4.63038893e-01,
'OverallQual',	1.88648629e+04,
'OverallCond',	5.81632931e+03,
'YearBuilt',	2.69006708e+02,
'YearRemodAdd',	1.36207016e+02,
'MasVnrArea',	2.43016226e+01,

DOI: 10.54254/2755-2721/55/20241483

Table 2. (continued).

'BsmtUnfSF',	-7.43320240e+00,
'TotalBsmtSF',	1.23533039e+01,
'1stFlrSF',	1.49021345e+01,
'2ndFlrSF',	8.92993303e+00,
'LowQualFinSF',	2.31407519e+00,
'GrLivArea',	2.61461427e+01,
'BsmtFullBath',	1.00967067e+04,
'BsmtHalfBath',	-1.00974236e+03,
'FullBath',	3.86700730e+03,
'HalfBath',	-2.26869552e+03,
'BedroomAbvGr',	-7.70524231e+03,
'KitchenAbvGr',	-3.63390961e+04,
'TotRmsAbvGrd',	5.06107560e+03,
'Fireplaces',	5.47609529e+03,
'GarageYrBlt',	-4.21657938e+01,
'GarageCars',	1.91934094e+04,
'GarageArea',	8.99175842e+00,
'WoodDeckSF',	1.40710249e+01,
'OpenPorchSF',	7.08294909e+00,
'EnclosedPorch',	5.64916041e+00,
'PoolArea',	-6.93859218e+01,
'YrSold']	-3.38428348e+02]

3.2. Random Forest

The Random forest regression algorithm is implemented, which built 100 trees to calculate the result (Figure 7.). The feature importance feature in the model is utilized to build the feature ranking plot (Figure 8.).

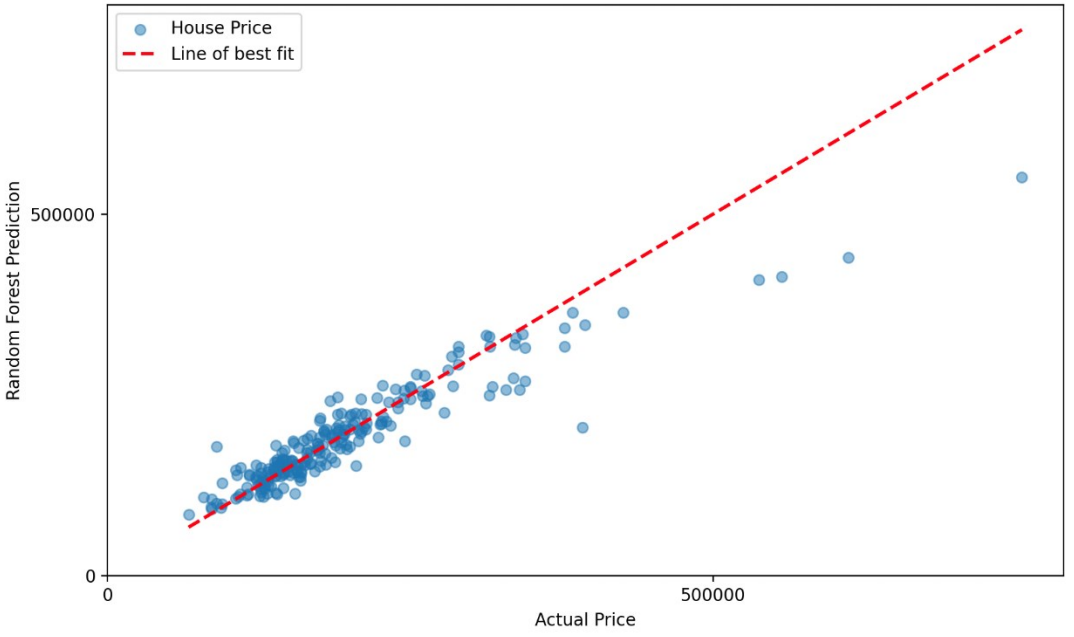

Figure 7. Random Forest Prediction

DOI: 10.54254/2755-2721/55/20241483

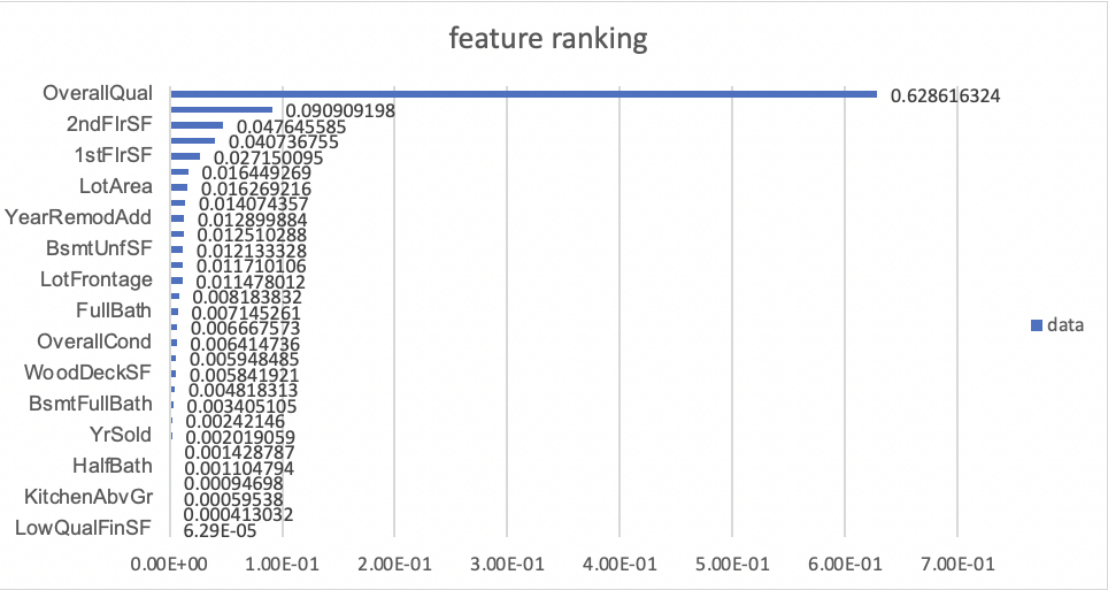

Figure 8. Feature importance with random forest

3.3. XGBoost

XGBRegressor model with 1000 estimators and 0.01 as learning rate was implemented to calculate the result. The plot importance algorithm is used to calculate the feature's impotance ranking for house price (Figure 9.).

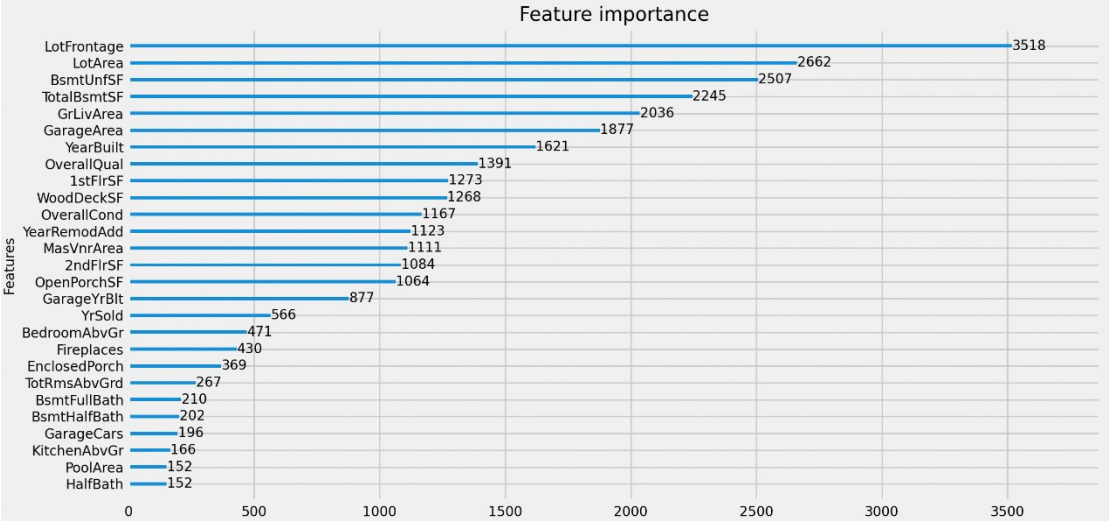

Figure 9. Feature importance with XGBoost

In examining the top three features—LotFrontage, LotArea, and BsmtUnfSF—it is evident from the correlation graphs presented in Section 2.1.1 that each of these features exhibits a positive relationship with the house price. Specifically, as the values of these features increase, there is a corresponding rise in the house's sale price. This observation underscores the idea that features with a strong correlation to the target variable are likely to emerge as significant when determining feature importance in the XGBoost algorithm.

DOI: 10.54254/2755-2721/55/20241483

3.4. Evaluation

The three models are assessed using the metrics MAE, MSE, RMSE, and R-squared score. The result is demonstrated in Table 3.

Table 3. Model Evaluation Results

Model	MAE	RMSLE	RMSE	R-squared
Linear Regression	23533.1938	0.1642	40356.3224	0.7935
Random Forest	20349.0903	0.1368	35241.1438	0.8502
XGBoost	18833.6568	0.1249	30718.0602	0.8803

4. Discussion

In the analysis of house price predictions in Ames, Iowa, there were noticeable differences in model performance. Specifically, the Linear Regression model was outperformed by more sophisticated algorithms, with XGBoost emerging as the most effective. As detailed in Section 3.4, the XGBoost model achieved the highest R-squared value, followed by Random Forest, while Linear Regression trailed with the lowest R-squared value.

Given the complexity of the dataset, comprising 27 independent features, the limitations of the Linear Regression model became apparent. This model inherently assumes linearity across the data distribution, which can introduce increased errors when the true relationship is nonlinear. In contrast, both Random Forest and XGBoost are well-equipped to handle nonlinearity in datasets. Although Random Forest shares certain characteristics with XGBoost, it was observed that XGBoost offers a computational advantage in terms of time complexity.

Further insights were derived from the feature importance rankings, as determined by both the Random Forest and XGBoost models. By combining the R-squared values from both models with their respective feature importance data, a comprehensive feature ranking was produced, depicted in Figure 10. The findings suggest that the overall quality of the house, its living area, and the total basement area (in square feet) are the three most influential factors affecting house prices in Ames, Iowa.

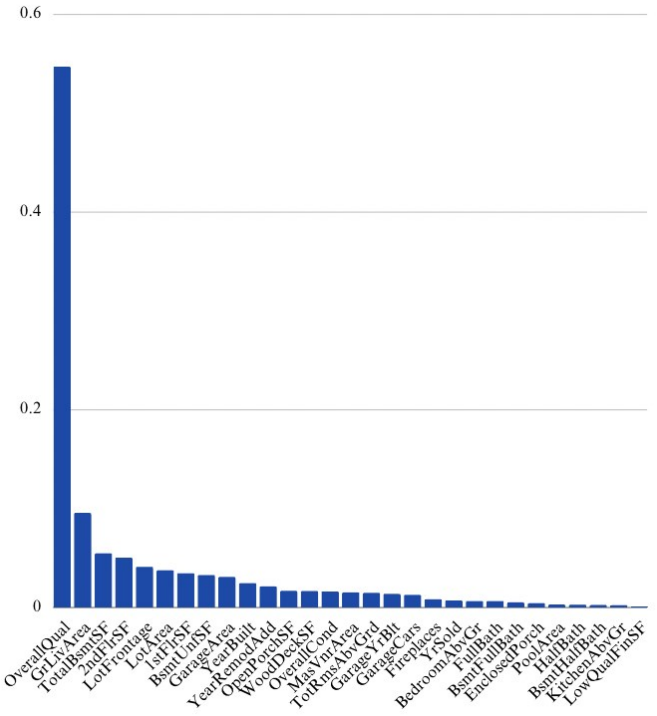

Figure 10. Feature importance

DOI: 10.54254/2755-2721/55/20241483

5. Conclusion

In conclusion, this paper explored the prediction of housing prices in Ames using three distinct machine learning models: Linear Regression, Random Forest, and XGBoost. Among these, Random Forest and XGBoost proved to be superior, achieving 0.8502 and 0.8803 R-squared values, respectively. The Random Forest was configured to use 100 trees, yielding satisfactory results with an RMSE of 35241 and an RMSLE of 0.1368. For XGBoost, it demonstrates better behavior than Random Forest, which has 30718 as RMSE and 0.1249 as RMSLE. A combined assessment of feature importance from these two models identified the overall quality of the house, the living area, and the square footage of the total basement area as the most influential factors determining house prices in Ames.

The further research could be conducted in the topics like ascertaining if XGBoost consistently emerges as the optimal model across varied datasets or if its performance is context-specific. Additionally, it would be valuable to delve deeper into understanding the inherent factors that drive the efficacy of tree-based models. Moreover, integrating more data sources or even exploring newer algorithms could refine the predictive capabilities further. The integration of temporal and spatial data might offer nuanced insights into housing price trends and the interplay of various factors over time and across different regions within Ames.

References

[1] Mora-Garcia, R.-T., Cespedes-Lopez, M.-F., & Perez-Sanchez, V. R. (2022). Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. Land, 11(11), 2100. MDPI AG.

[2] Wikipedia contributors. (2023, October 13). Ames, Iowa. In Wikipedia, The Free Encyclopedia. Retrieved from https://en.wikipedia.org/w/index.php?title=Ames,_Iowa&oldid=1179870443

[3] Ames Housing Market Report. (2023, September). Rocket. Retrieved from https://www.rockethomes.com/real-estate-trends/ia/ames

[4] Kaggle contributors. House Prices – Advanced Regression Techniques. Kaggle. https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

[5] N. N. Ghosalkar and S. N. Dhage, "Real Estate Value Prediction Using Linear Regression," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697639

[6] What is random forest? IBM. Retrieved from https://www.ibm.com/topics/random-forest

[7] Sruthi E. R. (2023, July). Understanding Random Forest Algorithms with Examples. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#Working_of_Random_Forest_Algorithm

[8] Jangaraj, Avanija & Sunitha, Gurram & Madhavi, Reddy & Kora, Padmavathi & Hitesh, R & Associate, Sai. (2021). Prediction of House Price Using XGBoost Regression Algorithm. Turkish Journal of Computer and Mathematics Education (TURCOMAT). 12. 2151-2155.

[9] Nembrini, S., König, I. R., & Wright, M. N. (2018). The revival of the Gini importance? Bioinformatics, 34(21), 3711-3718. https://doi.org/10.1093/bioinformatics/bty373

[10] NVIDIA., What is XGBoost?" NVIDIA Data Science Glossary, (2022). Retrieved from https://www.nvidia.com/en-us/glossary/data-science/xgboost/

[11] Computer Science Wiki contributors. (2023, January). Mean absolute error. Computer Science Wiki. Retrieved from https://computersciencewiki.org/index.php/Mean_absolute_error_(MAE)

[12] Computer Science Wiki contributors. (2023, January). Root-mean-square error. Computer Science Wiki. Retrieved from https://computersciencewiki.org/index.php?title=Root-mean-square_error_(RMSE)

[13] Padhma M. (2023, July). End-to-End Introduction to Evaluating Regression Models. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/#h-root-mean-squared-logarithmic-error-rmsle

DOI: 10.54254/2755-2721/55/20241483

[14] Chicco D, Warrens MJ, Jurman G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput Sci. 2021 Jul 5;7:e623. doi: 10.7717/peerj-cs.623. PMID: 34307865; PMCID: PMC8279135.