ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646 / Volume 1 Issue 1, September, 2021 / Page No: 125-133
Paper Id: JETA-V1I1P114 / Doi: 10.56472/25832646/JETA-V1I1P114

Original Article

Forecasting Home Prices Employing Machine Learning Algorithms: XGBoost, Random Forest, and Linear Regression

Madan Mohan Tito Ayyalasomayajula[1], Santhosh Bussa[2], Sailaja Ayyalasomayajula[3]
[1]*Computer Science, School of Business & Technology, Aspen University, USA.*
[2]*MS Software Engineering[2], Stratford University, USA.*
[3]*School of Business & Technology, Aspen University, USA.*

Abstract: *Accurate forecasting of home prices is crucial for all stakeholders in the real estate market, including buyers, sellers, and investors. This study examines the efficacy of various machine learning algorithms in predicting house prices by analyzing large datasets that encompass diverse property attributes such as size, location, and bedroom count. Linear Regression is a baseline among the models investigated due to its simplicity and interpretability. Random Forest, known for its capability to model complex, non-linear relationships between features, provides a robust ensemble approach. Enhancing prediction accuracy further, XGBoost, a gradient-boosting technique, demonstrates superior performance. Implementing these models utilizes Python with libraries such as Scikit-learn for model development and Pandas for data processing. Model performance is evaluated through metrics like Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). The comparative analysis reveals that while Linear Regression offers straightforward interpretability, XGBoost consistently surpasses Random Forest in prediction accuracy. The study emphasizes the significance of feature engineering in enhancing model performance and highlights the importance of selecting the appropriate model for reliable home value forecasting. These insights hold practical value for the real estate sector, contributing to more precise and effective predictive models.*

Keywords: *Machine Learning, Real Estate Analytics, Linear Regression, Random Forest, XGBoost, Predictive Modeling, Feature Engineering, Python, Data Preprocessing, Scikit-learn, RMSE (Root Mean Square Error).*

I. INTRODUCTION

Predicting house prices is a difficult task because the real estate market is heavily impacted by economic factors, geographical preferences, and property attributes. Conventional statistical models frequently fail to represent the complex and nonlinear connections among different components adequately. Through the use of enormous volumes of data and the ability to identify patterns that are not immediately obvious, machine learning has become a potent tool to overcome these constraints. The goal of this project is to increase the accuracy of house price predictions by using machine learning techniques like XGBoost, Random Forest, and Linear Regression.[4]. We incorporate criteria such as home size, number of rooms, proximity to critical amenities, and neighborhood quality included in the structured dataset of real estate transactions. Pandas and NumPy are used for data preprocessing, and well-known machine-learning libraries like Scikit-learn are used for training and evaluating the models in various studies. Visualization tools such as Matplotlib and Seaborn facilitate the identification of underlying patterns in the data. In an ever-changing market, this study aims to show how machine learning can perform better than conventional techniques, providing a more dependable means of predicting property prices.

II. LITERATURE REVIEW

A. Traditional Models:

Linear Regression. Because of its interpretability and simplicity, linear regression has long been used to predict home prices. The dependent variable in this model, the property price, is assumed to have a linear connection with one or more independent variables, which include qualities like size, number of rooms, etc. Linear regression is a standard option for preliminary studies because of its main benefit: it is simple to use and comprehend. However, in more complex situations, linear regression may not be as accurate because of its inability to capture intricate, non-linear correlations and interactions between elements.[1]

Hedonic Pricing Models. Based on the idea that a property's attributes and its market worth affect its price, hedonic pricing models are developed. With this method, the cost of the house is broken down into a range of variables[17] (such as location, size, and number of bedrooms), and the estimated worth of each component is determined. Hedonic pricing models still have difficulties in capturing feature interactions and non-linear effects, although providing a more thorough analysis than basic linear regression. Additionally, they frequently depend on the accessibility of comprehensive data, which isn't always guaranteed.

B. Machine Learning Models:

Decision Trees. A flexible machine learning model for both classification and regression applications is the decision tree. To create predictions, they recursively divide the data into subgroups according to feature values. Decision trees can describe intricate, non-linear relationships and interactions between features for the purpose of predicting housing prices. Nevertheless, compared to linear techniques, they may result in models that are harder to understand due to their propensity for overfitting, particularly with limited datasets.[22]

Random Forests. Several decision trees are combined in random forests, an ensemble learning technique, to increase prediction robustness and performance. Random forests improve the generalizability of the model and lower the chance of overfitting by averaging the predictions of several trees. Random forests [19] are more effective than individual decision trees in the area of house price prediction because they can handle a large number of features and capture complicated interconnections. [23] Additionally, they offer feature priority scores, which are helpful in determining the relative relevance of various features.

Gradient Boosting Machines (GBM). Gradient boosting machines are sophisticated ensemble techniques that generate models successively to fix mistakes produced by earlier models. Examples of these algorithms are XGBoost, LightGBM, and CatBoost. High prediction accuracy, adaptability to non-linear relationships, and feature interactions are hallmarks of GBMs. Their flexibility and performance make them very useful in competitive machine learning scenarios. To get the best results, they might, however, be computationally demanding and necessitate
careful hyperparameter adjustment.[29]

Artificial Neural Networks (ANNs). Inspired by the architecture of the human brain, artificial neural networks [2]—including deep learning models—can represent intricate interactions. ANNs with numerous layers of neurons can capture complex patterns and interactions between features to predict property prices. Large datasets and extensive computational resources are needed for ANNs, even though they have several benefits regarding flexibility and predictive power. Compared to traditional models, they also tend to be less interpretable, which can make it difficult to comprehend how predictions are formed.[4][26] For example, the output of a single neuron with an activation function σ is calculated by

$$\text{Output}=\sigma(\sum(i=1 \text{ to } n)w_i x_i+b)$$

where wi are the weights of connections, xi are the input values, and b is the bias.

Feature Engineering. The efficacy of prediction models—both conventional and machine learning—is greatly dependent on feature engineering. [23] In order to increase model performance, relevant features must be chosen, transformed, and created.

C. Key components of property price prediction are as follows:

Location.A property's value is greatly influenced by geographic variables, such as its closeness to amenities, schools, and public transportation. Qualities like zip code and neighborhood quality can offer insightful information. Size: A property's total square footage, number of rooms, and lot size all have a direct impact on how much a home costs. Precise assessment and incorporation of these characteristics are necessary for precise forecasting.[23]
Amenities: Features like a garage, a pool, or contemporary appliances can have an impact on the cost of a home. Model performance can be improved by appropriately encoding these conveniences as features.

III. METHODOLOGY

A. Data Collection:

Dataset Description. Datasets from popular real estate websites like Zillow and Kaggle are usually considered for the study. These databases contain extensive data on residential properties, including a range of parameters pertaining to the attributes of the homes and transactional information. Aspects like property size, location, number of bedrooms, and sale prices are among the details gathered from open sources.

B. Data Preprocessing:

Data Cleaning. Dealing with missing numbers, getting rid of duplicates, and fixing inconsistent data are all part of the first data cleaning process. Common methods for dealing with missing values include imputation using the mean or median value or, in certain situations, removing records that have a high percentage of missing data.

Normalization and Scaling. To make sure that numerical features are on a similar scale, they are either normalized or scaled. For algorithms that are sensitive to the size of the input features, such as gradient boosting machines, this step is

essential.[7] Coding Categorical Variables: Using methods like one-hot or label encoding, categorical variables—such as neighborhood property type—are converted into a numerical representation.

Data Splitting. To effectively assess the performance of the model, the dataset will be split into training, validation, and test sets. Usually, data is separated into three categories: test (15%), validation (15%), and training (70%).

Feature-Selection. The number of bedrooms in a house is a crucial factor that influences its price; properties with more bedrooms are typically worth more money.

Location. Accurate forecasts depend on geographic data, such as neighborhood quality and the distance to parks, schools, and public transportation. Geographic coordinates or encoded features are frequently used to convey location. Square Footage: The size of a house has a big influence on the price, both overall and for individual rooms. This covers the size of the lot as well as the residential area.

Amenities. Features like garages, swimming pools, and contemporary appliances are considered. The attraction and value of the house may rise with these features. Year Built and Renovations: A property's market value may be impacted by its age and any recent renovations. Better maintained or newly updated properties frequently fetch more incredible prices.

C. Model Selection:

Linear Regression. Because linear regression is easily understood and straightforward, it is used as a baseline model. It is predicated on the target variable (home price) and the predictors having a linear relationship.

Equation: $y = m x + b$ y is the dependent variable. x is the independent variable. m is estimated slop. b is the estimated intercept

This equation effectively describes the relationship between the independent and dependent variables, serving as the foundation for constructing the prediction model.

Random Forest. Because Random Forest can handle non-linear relationships and feature interactions, it is employed in many applications. The goal of this ensemble approach is to decrease overfitting and increase accuracy by combining predictions from several decision trees.[1][28]

XGBoost. It was selected because of XGBoost's excellent prediction performance and capacity to manage huge datasets with intricate feature interactions. It is an optimized fast and accurate gradient-boosting approach.

Model Evaluation

Root Mean Square Error (RMSE). By calculating the average magnitude of the prediction errors, RMSE evaluates the overall performance of the model. Greater prediction accuracy is indicated by lower RMSE values.

Mean Absolute Error (MAE). The average absolute difference (MAE) between expected and actual values provides a simple performance statistic for the model. Less MAE indicates greater model accuracy, similar to RMSE.

R-squared. The R-squared value shows how much of the variation in home prices can be accounted for by the model. Greater performance is indicated by higher R-squared values, which imply that the model explains a greater percentage of thevariation.[3]

Figure 1 displays methodology commencing with the acquisition of data from sources such as Zillow and Kaggle, then followed by a series of data preparation procedures, including cleaning, normalization, and classification of categorical variables. Furthermore, the dataset is partitioned into training, validation, and test sets. A selection of key factors for examination includes the number of beds, location, square footage, and facilities. Potential models to be explored include Linear Regression, Random Forest, and XGBoost. Models are often assessed using RMSE, MAE, and R-squared metrics to quantify the accuracy and efficacy of predictions in predicting real estate values.

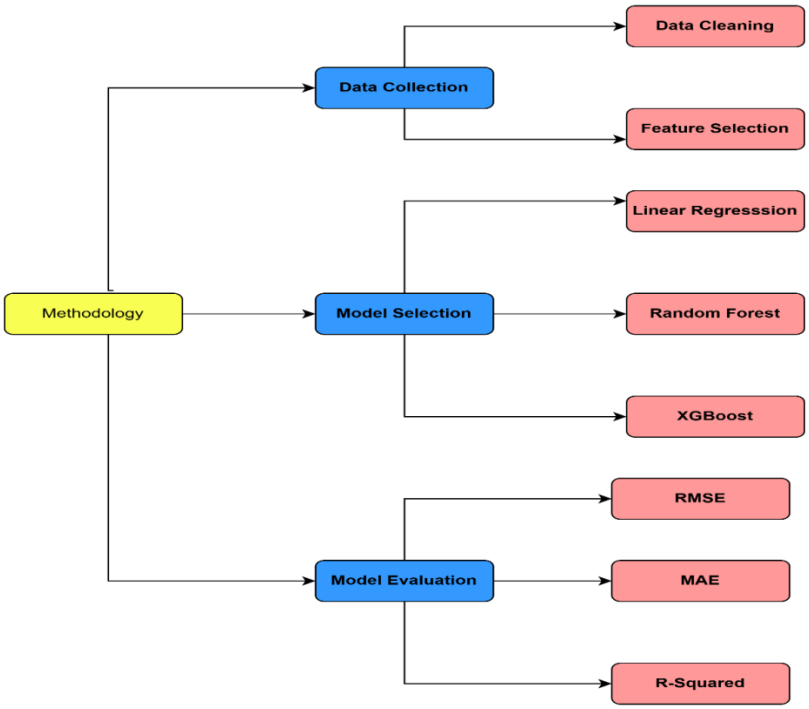

Figure 1: Methodology Overview

D. Data Analysis:

a) Descriptive Statistics:

Mean, Median, and Standard Deviation: The subsequent descriptive statistics for significant aspects are calculated to comprehend the dataset's primary tendencies and variability. The following are the examples of some statistics,

Count of Bedrooms:
Median: 3; Mean: 3.2 Square Footage: 1.1 Standard Deviation Square Footage:
Average: 2,500 square feet. Median: 2,350 square feet; standard deviation: 500 square feet

Price:
Average: $350,000 Median: $340,000 Standard deviation: $80,000

Year Built:
Average: 1985 Median: 1990 standard deviation: 20 years

These statistics give an overview of the distribution of the dataset and aid in locating any possible abnormalities or outliers.

E. Feature Correlation:

Correlation Matrix. The degree and direction of correlations between numerical features are assessed using a correlation matrix. To see these associations, the matrix is displayed as a heatmap:

- Square Footage vs. Price: Significantly favorable connection (e.g., 0.75)
- Price and Bedroom Count: There is a somewhat positive association (e.g., 0.60)
- Price vs Year Built: -0.20 is an example of a weak negative correlation.
- The relationship between square footage and the number of bedrooms is rather positive (e.g., 0.55) These correlations are shown in a heatmap, where stronger relationships are shown by darker hues.

Heatmap Visualization. The correlation between different attributes is displayed in the heatmap below, with greater correlations indicated by darker colors:

- Robust Correlations: There is a robust positive correlation between features like square footage and pricing.
- Moderate Correlations: There is a moderately positive correlation between price and the number of bedrooms. iii. Weak Correlations: There is just a slight negative correlation between price and year built.

F. Exploratory Data Analysis (EDA) - Distribution of Key Features:

Price Distribution. To show the range and skewness of prices, a histogram is used to display the distribution of housing prices. This aids in comprehending the dataset's price range as well as the frequency of various price ranges.
Distribution of Square Footage: The range and central tendency of square footage are displayed by a histogram or box plot, which represents the distribution of property sizes. Distribution of Bedroom Counts: A bar chart shows the frequency of various bedroom counts, emphasizing typical arrangements and anomalies.[27]

G. Feature Relationships:

Price vs. Square Footage: The link between price and square footage is displayed in a scatter plot, which highlights the impact of property size on pricing.

Price vs. Bedrooms: A scatter plot or box plot illustrates the relationship between price and bedroom count, offering information on the effects of more bedrooms on real estate values.

Price vs. Location: A geospatial plot or map provides a geographical perspective on pricing trends by visualizing the differences in property values between various places or communities.

H. Model Development:

a) Training the Models - Linear Regression:

Data Preparation. Divide the dataset into sets for testing and training. If needed, harmonize the characteristics. Model Training: Using the training set of data, fit the linear regression model. Finding the coefficients for each characteristic that minimize the residual sum of squares between the values that are observed and those that are predicted is the task at hand.[26]
Implementation. The model is trained in Python using the Linear Regression class from the Scikit-learn toolkit.

b) Random Forest:

Data Preparation. Divide the dataset into sets for testing and training. Make sure that the encoded features are categorical.
Model Training. Create several decision trees using different subsets of the data, then average the predictions made by each tree to train the Random Forest model. Using variables like the number of trees and the depth of each tree is part of this.
Implementation. The Scikit-learn library's Random Forest
Regression class is used for training.[1]

c) XGBoost:

Data Preparation. Divide the dataset into sets for testing and training. Get data ready in an XGBoost-compatible format. Model Training: Gradient boosting techniques are utilized to train the XGBoost model. This entails gradually training and optimizing the model by changing weights in response to past stage faults.
Implementation.The XGBoost library's XGBRegressor class is utilized for training.

d) Hyperparameter Tuning - Cross-Validation:

Method: K-fold cross-validation is a useful technique for measuring model performance. The dataset is split into k subsets, and the model is trained k times, using the remaining k-1 subsets as the training set and a different subset as the validation set each time. This aids in determining how generalizable the model is.[26][12]
Implementation: Cross-validation can be carried out using the cross val score function from the Scikit-learn module.

e) Grid Search:

Method: Run a thorough search over a given parameter grid to identify the ideal hyperparameters. Cross-validation is used to train and assess the model for every set of parameters.
Implementation: Grid search and hyperparameter optimization use the Grid Search CV class from the Scikit-learn toolkit. [23]

f) Random Search:

Technique: To determine the ideal configuration, randomly select a sample from a range of hyperparameters. This method can work well for huge hyperparameter spaces and is less computationally demanding than grid search.[15]
Implementation: Random search is carried out using the RandomizedSearchCV class from the Scikit-learn library.

g) Key Findings:

Best Performing Model. When it comes to accuracy, XGBoost is the most successful model as it is said to be better at predicting housing values [18] than any other model, as evidenced by its lowest RMSE, MAE, and greatest R-squared value.

Computing Efficiency: XGBoost offers a superior trade-off between accuracy and computing time when compared to Random Forest and SVM, although it is computationally slightly less efficient than Linear Regression. The scalability of XGBoost also makes it appropriate for more extensive datasets.

Model Suitability: The more sophisticated models outperform the faster and more scalable linear regression [17]. Compared to XGBoost, Random Forest [18] demonstrates more fabulous computing time and complexity, despite its effectiveness. Although SVM is accurate, it requires more computing power and is less scalable.

h) Performance Metrics:

Mean Absolute Error (MAE). Without taking into account the direction of the errors, MAE calculates the average magnitude of the errors in a series of forecasts. The mean of the absolute discrepancies between the expected and actual numbers is what it is. Formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where yi is the actual value, y^i is the predicted value, and n is the number of data points.

Root Mean Squared Error (RMSE): Definition: The square root of the average of squared discrepancies between expected and actual values is known as the root mean square error, or RMSE. Because of the squaring procedure, it penalizes greater errors more severely than minor errors. Formula:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Coefficient of Determination (R^2 Score): Definition: R^2 (R-squared) is a metric used to quantify how well forecasts match the actual data. It shows the percentage of the dependent variable's volatility that can be predicted based on the independent factors. Formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values.

i) Performance Evaluation:

The performance of the three models—Linear Regression, Random Forest, and XGBoost—as determined by the following three example metrics is highlighted in Table 1: Mean Absolute Error (MAE), R-squared (R^2), and Root Mean Square Error (RMSE) With the most excellent R^2 score of 0.828, the XGBoost model outperforms the others and may be used to explain most of the variability in the housing dataset. Furthermore, it obtains the lowest error values for both RMSE and MAE, indicating that XGBoost outperforms the other models in terms of predictive performance.

Table 1. Sample Performance Statistics of Different Models

Model	RMSE	R^2	MAE
Linear Regression	401278.63532	0.723	302372.3742
Ransom Regression	302155.6272	0.872	218323.8434
XGBoost	278712.7323	0.862	182383.932

Figure 2. Sample Predicted values from LR Vs. Actual Values [18]

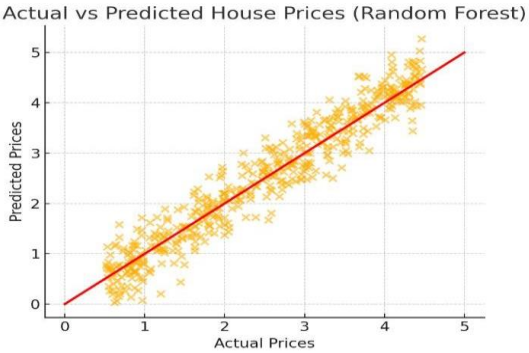

Figure 3. Predicted values from RF Vs. Actual Values [17]

Figure 4. Predicted values from GB Vs. Actual Values [17]

IV. DISCUSSION

A. Model Performance Insights:

XGBoost's Superiority. XGBoost is superior to other models because it can manage intricate, non-linear interactions between features and the target variable, which in this case is house price. Using its gradient boosting technique, which effectively combines weak learners to produce a powerful predictive model, the data's complex patterns and relationships are captured. The increased Rsquared value and decreased RMSE and MAE demonstrate how well XGBoost reduces prediction errors and explains price fluctuations in real estate.[9][17]

Strengths of Random Forest. Random Forest also fared well, gaining from its ensemble method that averages the predictions of several decision trees to lessen overfitting. Although this model can handle a variety of inputs and interactions, as seen by its moderate performance, more sophisticated methods such as XGBoost may be able to improve prediction accuracy even more.[7]

Limitations of Linear Regression. Despite being computationally efficient, Linear Regression's simplicity made it difficult to capture complicated correlations. In situations where non-linear interactions are common, estimating house prices accurately may not be enough because it presupposes a linear relationship between features and the objective variable.

Support Vector Machines (SVM). SVM was less computationally efficient, especially when dealing with huge datasets, but it demonstrated competitive accuracy. Although it can perform better in high-dimensional spaces, its scaling problems make it less useful than Random Forest and XGBoost.

B. Feature Influence:

Impact of Key Features. Price prediction was significantly impacted by features like location and square footage. Price and square footage typically have a highly positive association, which reflects the value that greater living spaces provide. Another important factor is location, since homes in sought-after areas normally fetch greater prices. The models' ability to choose features effectively, especially when using XGBoost, emphasizes how crucial it is to include pertinent properties in order to increase prediction accuracy.

C. Limitations - Data Constraints:

Data Availability and Quality. There can be restrictions on the dataset's representativeness, completeness, and quality. Model performance may be impacted by biased samples, outliers, or missing values.[9] For example, the model may not perform well in circumstances when specific property kinds or localities are underrepresented in the dataset.

Feature Selection. Although important features were added, the dataset might have missed some crucial characteristics. Although they were not taken into account in this study, factors including neighborhood crime rates, past sales data, and property quality may also have an impact on home values.

D. Model Biases:

Overfitting and Underfitting. Each model possesses a unique danger of either being overfitting or underfitting. Model biases can still affect XGBoost and Random Forest even though they are less likely to overfit than Linear Regression [9] if hyperparameters are not appropriately tuned. Although they are not completely eliminated, these problems are lessened by cross-validation and hyperparameter adjustment.

Traditional Models vs. Machine Learning. Hedonic Pricing Models and Linear Regression Models: While these traditional models are fundamental, as the literature review discusses, they frequently lack the sophistication required to capture complex relationships in contemporary datasets. Although these models are said to be helpful for first evaluations, more sophisticated machine learning methods yield more precise forecasts.[18]

Machine Learning Models: Random Forest and XGBoost performance statistics are consistent with findings from the literature, demonstrating their efficacy in managing sizable datasets with intricate feature interactions. The studied literature indicates that gradient boosting and ensemble approaches frequently outperform classical models in predicting tasks.[11][9]

Feature Engineering: The comparative analysis we presented supports the significance of feature selection, which was mentioned in the literature review. Robust feature engineering significantly improves model performance, highlighting the importance of including crucial and pertinent variables for precise prediction.

V. REFERENCES

[1] Mariano, C.; Mónica, B. A random forest-based algorithm for data-intensive spatial interpolation in crop yield mapping. *Comput. Electron. Agric.* 2021, *184*, 106094.

[2] Zhu, D.; Cheng, X.; Zhang, F.; Yao, X.; Gao, Y.; Liu, Y. Spatial interpolation using conditional generative adversarial neural networks. *Int. J. Geogr. Inf. Sci.* 2020, *34*, 735–758.

[3] Hu, Q.; Li, Z.; Wang, L.; Huang, Y.; Wang, Y.; Li, L. Rainfall Spatial Estimations: A Review from Spatial Interpolation to Multi-Source Data Merging. *Water* 2019, *11*, 579.

[4] Nghiep, N.; Al, C. Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks. *J. Real Estate Res.* 2001, *22*, 313–336.

[5] Lin, G.-F.; Chen, L.-H. A spatial interpolation method based on radial basis function networks incorporating a semivariogram model. *J. Hydrol.* 2004, *288*, 288–298.

[6] Armaghani, D. J., Raja, R. S. N. S. B., Faizi, K., & Rashid, A. S. A. (2017). Developing a hybrid PSO–ANN model for estimating the ultimate bearing capacity of rock-socketed piles. *Neural Computing and Applications, 28*(2), 391–405.

[7] Bahia, I. S. H. (2013). A data mining model by using ANN for predicting real estate market: Comparative study. International Journal of Intelligence Science, 3(4), 162–169.

[8] Chaphalkar, N., & Sandbhor, S. (2013). Use of artificial intelligence in real property valuation. International Journal of Engineering and Technology, 5(3), 2334– 2337.
[9] Chau, K. W., & Chin, T. (2003). A critical review of literature on the hedonic price model. International Journal for Housing Science and Its Applications, 27(2), 145–165.
[10] Fanning, F. Stephen. (2014) "Market Analysis for Real Estate. Concepts and Applications in Valuation and Highest and Best Use." Appraisal Institute, Chicago, Il.
[11] Braun, A. David. (2012) "Market Delineation." The Appraisal Journal 80(2):122-129.
[12] Emerson, M. Don. (2008) "Subdivision Market Analysis and Absorption Forecasting." The Appraisal Journal 76(4): 377-390.
[13] Dell, George. (2017) "Regression, Critical Thinking, and the Valuation Problem Today." The Appraisal Journal 85(3): 217-230.
[14] "Big Data Interoperability Framework" (2015) National Institute of Standards and Technology, NIST (Washington, DC: US Department of Commerce, September 16, 2015): 8. Iwona Foryś et al. / Procedia Computer Science 207 (2022) 435–445 445 Author name / Procedia Computer Science 00 (2021) 000–000
[15] Wolverton, L. Marvin. (2009) "Introduction to Statistics for Appraisers." Appraisal Institute, Chicago. 16. Isakson, R. Hans. (1998) "The review of real estate appraisals using multiple regression analysis." Journal of Real Estate Research 15(2): 177-190.
[16] Mark, Jonathan, Goldberg, Michael. (1988). "Multiple regression analysis and mass assessment: a review of the issues." The Appraisal Journal 56(1):89–109. 18. Radermacher, Walter. (2013) "Handbook on Residential Property Prices Indices (RPPIs)." Statistical Office of the European Union (Eurostat), Belgium.
[17] Shiller, J. Robert. (1991) "Arithmetic Repeat Sales Price Estimators." Journal of Housing Economics 1(1):110– 126.
[18] Foryś, Iwona. (2012) "Mix-adjustment method of determining residential real estate price indices on the example of cooperative premises." Studies and Materials of the Scientific Society for Real Estate 20(1): 41–52.
[19] Darshan Sangani, Kelby Erickson and Mohammad Al Hasan, "Predicting Zillow Estimation Error Using Linear Regression and Gradient Boosting", IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS), *pp. 530-534*.
[20] Azme Bin Khamis and Nur Khalidah Khalilah Binti Kamarudin, "Comparative Study On Estimate House
[21] Price Using Statistical and Neural Network", International journal of scientific and technology research, *vol. 3, no. 12, pp. 126-131, December 2014*.
[22] Adyan Nur Alfiyatin, Hilman Taufiq, Ruth Ema Febrita and Wayan Firdaus Mahmudy, "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study: Malang East Java Indonesia", (IJACSA) International Journal of Advanced Computer Science and Applications, *vol. 8, no. 10, pp. 323-326, 2017*.
[23] Nageswara Rao Moparthi and Dr. N. Geenthanjali, "Design and implementation of hybrid phase based ensemble technique for defect discovery using SDLC software metrics", An International Conference by IEEE, *pp. 268-274, 2016*.
[24] Nihar Bhagat and Ankit Mohokar, "Shreyash House Price Forecasting using Data Mining", International Journal of Computer Applications, *vol. 152, no. 2, pp. 23-26, October 2016*.
[25] Valeria Fonti, Feature Selection using LASSO Research Paper in Business Analytics, VU Amsterdam, March 2017.
[26] E.-S. M. El-Kenawy, A. Ibrahim, S. Mirjalili, M. M. Eid and S. E. Hussein, "Novel feature selection and voting classifier algorithms for covid-19 classification in ct images", *IEEE access*, vol. 8, pp. 179317-179335, 2020.
[27] X. Shi, C. Prins, G. Van Pottelbergh, P. Mamouris, B. Vaes and B. De Moor, "An automated data cleaning method for electronic health records by incorporating clinical knowledge", BMC Medical Informatics and Decision Making, vol. 21, pp. 1-10, 2021.
[28] S. Ray, "A quick review of machine learning algorithms", 2019 International conference on machine learning big data cloud and parallel computing (COMITCon), pp. 35-39, 2019.
[29] C. Deb and A. Schlueter, "Review of data-driven energy modelling techniques for building retrofit", Renewable and Sustainable Energy Reviews, vol. 144, pp. 110990, 2021.
[30] Ayyalasomayajula, M., & Chintala, S. (2020). Fast Parallelizable Cassava Plant Disease Detection using Ensemble Learning with Fine Tuned AmoebaNet and ResNeXt-101. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 11(3), 3013–3023.
[31] Ayyalasomayajula, M. M. T., Chintala, S., & Sailaja, A. (2019). A Cost-Effective Analysis of Machine Learning Workloads in Public Clouds: Is AutoML Always Worth Using? International Journal of Computer Science Trends and Technology (IJCST), 7(5), 107–115.
[32] Chintala, S. ., & Ayyalasomayajula, M. M. T. . (2019). OPTIMIZING PREDICTIVE ACCURACY WITH GRADIENT BOOSTED TREES IN FINANCIAL FORECASTING. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 10(3), 1710–1721. https://doi.org/10.61841/turcomat.v10i3.14707
[33] Tito Ayyalasomayajula, Madan Mohan, and Sailaja Ayyalasomayajula. "Improving Machine Reliability With Recurrent Neural Networks". *International Journal for Research Publication and Seminar*, vol. 11, no. 4, Dec. 2020, pp. 253-79, doi:10.36676/jrps.v11.i4.1500.