Evaluating the Performance of Random Forest Algorithm in Classifying Property Sale Amount Categories in Real Estate Data

Les Endahti[1,*], Muhammad Shihab Faturahman[2]

[1,2]*AMIK-YPAT Purwakarta, Indonesia*

(Received: June 7, 2025; Revised: July 18, 2025; Accepted: October 26, 2025; Available online: December 8, 2025)

Abstract

This study explores the use of machine learning algorithms to classify property sale categories in real estate data, focusing on the performance of the Random Forest (RF) algorithm. The dataset, comprising over one million records of property sales from 2001 to 2022, includes features such as sale amount, assessed value, sales ratio, property type, and residential type. The primary objective is to determine which algorithm better predicts property sale categories and to assess how these predictions can aid in market segmentation and property valuation. After preprocessing the data by removing irrelevant columns and handling missing values, we applied the RF classifier to predict five key property types: 'Single Family', 'Residential', 'Condo', 'Two Family', and 'Three Family'. The model achieved an accuracy of 82.98%, with high recall for categories like 'Single Family' and 'Condo', but struggled with 'Residential', which displayed a lower recall due to its diverse nature. The findings suggest that the RF algorithm performs well in predicting certain property types, but improvements are needed for categories with more variation. The study highlights the importance of selecting relevant features such as sale amount and assessed value, which were found to be the most influential in determining property type. Real estate professionals can leverage these machine learning models for more accurate market segmentation, leading to better pricing and marketing strategies. However, the study also acknowledges limitations, such as the complexity of the 'Residential' category and potential data imbalance. Future research could focus on incorporating additional features, such as location-specific data or detailed property descriptions, and testing alternative algorithms to further enhance classification accuracy.

Keywords: Random Forest, Property Classification, Machine Learning, Real Estate, Market Segmentation

1. Introduction

An overview of property sale amounts as critical indicators in the real estate market necessitates a multifaceted examination of the underlying factors that influence real estate prices. In recent years, various empirical studies and theoretical analyses have highlighted the intricate relationship between economic variables and real estate market dynamics. A significant assertion within the literature is that macroeconomic factors, such as unemployment rates, play a pivotal role in shaping consumer confidence and, subsequently, real estate prices. Ćetković et al. [1] demonstrate that lower unemployment typically correlates with increased consumer confidence, leading to price growth in the real estate sector; conversely, rising unemployment tends to depress property values, underscoring the sensitivity of the real estate market to economic fluctuations [1]. Moreover, the interconnectedness of global real estate markets can complicate local assessments of property values. Liow and Angela [2] research indicates that, over longer investment periods, public real estate markets exhibit substantial interdependence, particularly during times of economic distress, such as the global financial crisis. This interconnectedness suggests that local property sales are not just influenced by regional economic conditions but are also subject to the reverberations of global market dynamics, thereby complicating valuation processes and necessitating more robust analytical frameworks to capture these interactions adequately.

Further complicating the valuation landscape is the problem of information asymmetry prevalent within the real estate market. Paola [3] emphasizes that many impediments arise due to lack of standardization, incomplete data, and reluctance from private individuals to disclose transaction prices, which can adversely affect market understanding and investment decisions. Consequently, the need for effective valuation models that can account for these data deficiencies

[*]**Corresponding author: Les Endahti (endahti01@amikypat-purwakarta.ac.id)**
DOI: https://doi.org/10.47738/ijaim.v5i4.114

is imperative, indicating a gap in existing methodologies utilized in property evaluations. A noteworthy development in this arena involves the utilization of advanced statistical methods to enhance property price assessments. Giudice et al. [4] advocate for the application of semiparametric methods to perform hedonic analysis of housing sales, acknowledging that traditional models may inadequately account for the myriad qualitative and quantitative factors affecting property appreciation. The authors argue for the development of sophisticated models that can differentiate between various forms of appreciation, which is crucial for accurate market analysis [4]. This assertion is supported by evolving methodologies that aim to leverage machine learning techniques and large datasets, facilitating more nuanced insights into property valuations.

Classifying sale amounts into categories holds significant importance for effective market analysis and informed decision-making in the real estate industry. The classification of real estate sale amounts enhances the comprehension of market dynamics, allowing stakeholders to navigate the complexities of property transactions more effectively [5]. By delineating sale amounts into various categorizations, such as price brackets or geographic regions, analysts can leverage this structured data to identify patterns that inform market trends and consumer behavior, making it a critical component of strategic planning in real estate investments and developments. One of the primary advantages of classifying sale amounts is the facilitation of comparative analyses. Sobieraj and Metelski [6] illustrate the use of classification in understanding the sale-to-list ratio, a crucial metric in determining how well properties is priced in relation to market demand. This ratio simplifies the assessment of market conditions, enabling stakeholders to identify whether properties are overpriced or underpriced, thus providing actionable insights that can guide pricing strategies for real estate agents and investors. By using classification to break down these ratios, practitioners can also compare performance across different segments of the market, aiding in the identification of opportunities or risks inherent to specific categories.

Moreover, the classification aids in the transparency and accuracy of market valuations. Menghini et al. [7] discuss the significance of distinguishing between fair market value and judicial market value, highlighting how classifications can clarify the varied economic contexts within which properties are assessed. By providing clear categorizations, stakeholders can better understand the implications of different valuation methods on investment decisions and negotiate transactions with greater confidence. This clarity is essential in contexts such as foreclosure or debt recovery, where a precise understanding of asset values can greatly influence potential outcomes. Additionally, innovative methodologies utilizing machine learning and artificial intelligence have emerged, further reinforcing the importance of classification in property valuation [8]. By employing advanced computational techniques, such as Support Vector Machines (SVM) and multivariate regression analysis, researchers can categorize sale amounts based on an array of factors, including property characteristics and market trends. This analytical depth enables more dynamic pricing strategies and investment decisions, tailored to real-time market conditions, ultimately driving more efficient market operations.

The purpose of this study is to compare the performance of RF and SVM algorithms in the classification of real estate sale amounts. This comparison is rooted in the increasing importance of machine learning techniques in quantifying and predicting real estate prices, a field reliant on advanced computational methods to address the complexities of market dynamics. As real estate markets exhibit non-linear characteristics, choosing the appropriate classification algorithm becomes critical for accuracy and reliability in price predictions. RF is recognized for its robustness and ability to handle both numerical and categorical features effectively. Utilizing an ensemble approach, RF builds multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees, thus reducing overfitting problems often encountered in single decision tree models [9]. Studies have emphasized the superiority of RF over traditional methods, particularly its ability to improve prediction performance when faced with high-dimensional data or complex relationships among features [10].

On the other hand, SVM operates based on finding an optimal hyperplane that distinctly classifies data points from different classes. It performs particularly well in high-dimensional spaces and is efficient in cases where the number of dimensions exceeds the number of samples [11]. However, SVM can be sensitive to the choice of kernel functions and parameters, often requiring extensive tuning and cross-validation to enhance performance [12]. This makes the comparison with RF particularly relevant, as both algorithms utilize distinct methodologies that may yield different results in the classification of real estate sale amounts. Empirical studies have showcased the effectiveness of RF in

real estate contexts, with research indicating that RF frequently outperforms SVM in terms of predictive accuracy. For instance, a movement towards employing RF has been observed in various analyses concerning land value estimation and housing price predictions, demonstrating that it often yields better results compared to SVM and other regression approaches [10]. In their comparative analysis, Canaz and Aliefendioğlu [10] concluded that RF provided superior results in mass appraisal studies involving various predictive features from real estate sales. These findings resonate with earlier work by Wang et al. [12], which indicated that RF outperforms SVM, often providing longer-term insights into pricing trends based on multidimensional data [11], [13].

The research aims to address the question of which classification algorithm, between RF and SVM, better predicts property sale categories in the real estate sector. Specifically, it seeks to determine which model provides higher accuracy and reliability in classifying properties into price categories based on features such as sale amount, assessed value, property type, and location. By comparing the performance of these two widely used algorithms, this study will contribute to understanding their applicability in real estate data analysis and offer insights into optimizing property price predictions for better market segmentation and decision-making.

2. Literature Review

2.1. Previous Research on Property Price Prediction

The evolution of property price prediction methodologies has seen a significant transition from traditional valuation methods to sophisticated machine learning techniques. This study aims to provide an overview of both categories, highlighting their applications, strengths, and weaknesses in estimating property values. Understanding these methods is critical for real estate professionals and investors seeking reliable avenues for making informed decisions in an increasingly data-driven market. Traditional property valuation methods encompass approaches such as the comparative method, income method, residual method, and cost approach. Among these, the comparative method is the most widely adopted globally, as recognized by Abidoye et al. [4], who emphasize its prevalence in Australia and its effectiveness in leveraging market transactions for property valuation [15]. This method relies on the principle that similar properties should exhibit similar values, thus enabling valuers to draw comparisons based on similar property characteristics. However, while traditional models lay foundational practices in the field, they often lack the adaptability required to account for rapid market fluctuations and can be significantly influenced by subjective judgments in the selection of comparable properties.

In recent years, the advent of machine learning has broadened the horizons regarding property price estimation. Advanced algorithms capable of identifying non-linear relationships within the data challenge the boundaries defined by classical economic principles in property valuation. A considerable body of research has shown the efficacy of machine learning algorithms, such as RF, SVM, and deep learning models, which can analyze vast datasets with numerous variables to provide high-accuracy predictions [15], [16], [17]. These methodologies accommodate heterogeneous datasets, enabling a more nuanced analysis that captures the intricacies of property characteristics and market dynamics. Hedonic pricing models have emerged as a primary method for quantitative analysis in property price prediction. This approach decomposes property prices based on individual attributes such as location, physical characteristics, and neighborhood factors [18]. For instance, Yazdani [19] emphasizes that structural and spatial attributes significantly influence property prices, advocating for hedonic models to leverage these factors to enhance prediction accuracy. Hedonic models are inherently flexible, allowing for the integration of various property features and facilitating a more tailored analysis reflective of specific market conditions.

2.2. Classification in real estate data

The application of classification models in the categorization of property prices has garnered significant academic attention in recent years. Previous studies have explored various machine learning techniques and their effectiveness in predicting and classifying real estate sale prices. This overview synthesizes key findings from the literature to illuminate the methodologies and results obtained from different classification models used for property price categorization. One of the more comprehensive studies was conducted by Kuru et al. [8], who developed sale price classification models specifically for real estate appraisal. Their approach utilized regression analysis alongside

Artificial Neural Networks (ANN), demonstrating the effectiveness of these methods by factoring in attributes such as property characteristics, geographical position, and market trends. This research reveals that employing advanced algorithms significantly enhances the accuracy of predictions compared to simpler traditional models [8], [19]. The results highlight how machine learning approaches can address the complexities inherent in real estate datasets that traditional methods may overlook.

In Heli K'Akumu [5] work, the complexities of real estate as a discipline are explored, emphasizing the foundational role classification plays in business context applications of real estate management. Although K'Akumu [5] study is more philosophical in nature, it stresses the importance of rigorous methodology in the classification processes within the discipline. The implications of this work point to how epistemological foundations can shape understanding and application of classification in real estate economics, thus influencing subsequent analytical techniques. Pirogova et al. [20] introduced a systematic set of competitiveness indicators to assess retail real estate's performance, highlighting the use of classification systems in quantifying the level of competitiveness. This categorization is critical for landlords and investors as it allows for strategic decision-making based on the competitive landscape of retail properties [20], [21]. Their methodology provides a refreshing perspective on how classification can extend beyond mere price predictions to encompass broader market dynamics.

2.3. Random Forest and SVM Algorithms

RF and SVM are two prominent machine learning algorithms that have increasingly been applied in the field of real estate data analysis, particularly for the classification of property prices. Their robustness and efficacy in handling complex datasets make them well-suited for real estate applications, where numerous factors influence property valuation. RF is an ensemble learning method that operates by constructing multiple decision trees during training time and producing predictions based on the majority vote from these trees. This method is highly regarded for its accuracy and ability to mitigate overfitting, a common problem in machine learning where models perform well on training data but poorly on unseen data. According to Tanamal et al. [22], the RF algorithm improves the predicted accuracy of datasets by averaging the results from various trees, thus ensuring a more reliable classification process [22], [23]. Additionally, RF can handle large datasets with higher dimensionality, making it particularly effective in real estate, where numerous variables, such as location, square footage, and amenities, can influence property values.

Conversely, SVM is a supervised learning tool designed for classification and regression tasks. SVM identifies an optimal hyperplane that separates different classes in a high-dimensional space, making it suitable for classifying properties based on their features. As highlighted by Li and Chao [24], SVM is particularly powerful when dealing with high-dimensional data and is adept at generalization, thereby maintaining performance on unseen instances. Its strength lies in transforming data points into higher dimensions via kernel functions, which allows for more complex boundary formations between different property classes. The application of SVM in real estate has been documented by various researchers, including SVM can be highly effective in real estate evaluations, albeit it didn't outperform RF in their comparative studies [25]. SVM's sensitivity to the choice of kernel and parameters means that while it can deliver high accuracy, it often requires careful tuning and validation to reach optimal performance. Studies suggest that SVM performs exceedingly well when dealing with smaller datasets, which can often occur in niche markets or specialized classifications [26].

2.4. Relevant formulas

The RF algorithm uses the Gini impurity criterion to measure the effectiveness of a split in classifying data. The formula for calculating Gini impurity is:

$$\text{Gini} = 1 - \sum i = 1^{k} pi^{2} \quad (1)$$

where (p_i) represents the proportion of instances that belong to class (i) at a given node, and (k) is the total number of classes within that node. This formula helps to identify the split that results in the least impurity, thereby improving the model's predictive accuracy [22].

The SVM algorithm aims to find the optimal hyperplane that maximizes the margin between different classes. The objective function for SVM is defined as:

 ISSN 2776-8007

$$\min w, b \frac{1}{2}|w|^2 \quad \text{subject to} \quad yi(w^T xi + b) \geq 1 \tag{2}$$

In this formulation, (w) is the weight vector, (b) is the bias term, (yi) indicates the true class label for the instance (x_i) (which can be +1 or -1), and (w^T) represents the transpose of the weight vector. The objective function minimizes the magnitude of the weight vector to maximize the margin while ensuring all data points are correctly classified [25], [27].

Both RF and SVM algorithms are widely used in real estate data analysis and predictive modeling. Their strengths in classification tasks yield actionable insights that can guide investment and policy decisions within real estate markets. RF is a popular method for predicting property prices due to its capability to handle large datasets and its resistance to overfitting. Research Tanamal et al. [22] demonstrated that RF significantly enhances the accuracy of house price prediction models by averaging results from multiple decision trees, leading to robust performance even in complex datasets. Additionally, research indicates that RF often outperforms other machine learning techniques regarding accuracy while being interpretable regarding the importance of features impacting property values, such as location and structural characteristics [10], [28].

SVM also has noteworthy applications in real estate, particularly for classification tasks that require distinguishing between different property categories or pricing tiers. Wang et al. [12] research indicates that SVM effectively models complex relationships among property features, providing accurate price classifications across various market conditions. Importantly, SVM requires careful tuning of parameters, such as kernel functions and regularization terms, to achieve optimal performance, especially within high-dimensional real estate datasets [25], [29]. Empirical research, including studies by Canaz and Aliefendioğlu [10], highlights that while RF generally yields better results compared to SVM, both algorithms have unique strengths that can be leveraged in real estate data analysis. Their findings underscore the necessity for practitioners to assess which model to employ based on the specific characteristics of the dataset and the nature of the classification problem.

3. Methodology

This section outlines the detailed steps taken in this research, which are divided into six key sections: Data Loading, Exploratory Data Analysis (EDA) & Preprocessing, Feature Engineering & Selection, Model Training, Model Evaluation & Visualization, and Checkpoint Saving. Each section is explained with an emphasis on the parameters used throughout the process.

3.1. Data Loading

The first step involves loading the real estate sales data into the script from a local CSV file using pandas. The pd.read_csv() function is employed with the parameter low_memory=False to avoid warnings related to data type inference when dealing with large files. This ensures that all columns are correctly parsed, even those with mixed data types. The function reads the data into a DataFrame, which is then validated by checking its dimensions using data.shape to get the number of rows and columns. The first few rows are displayed with data.head() to provide a quick inspection of the dataset, allowing the user to check for any irregularities or unexpected values in the columns. Additionally, the script handles cases where the data file is not found. If the specified CSV file cannot be located, the script prints an error message and exits immediately using exit(). This prevents the script from attempting to process non-existent data. Once the data is loaded successfully, the next step is to prepare the dataset for analysis, ensuring that all columns are correctly interpreted and the dataset is clean and ready for preprocessing.

3.2. Exploratory Data Analysis (EDA) & Preprocessing

In this step, the data undergoes cleaning and basic analysis to prepare it for modeling. The data.drop() function is used to remove columns that are deemed unnecessary or irrelevant for the classification model, such as 'Serial Number', 'Address', 'Date Recorded', and 'Location'. These columns may contain too many unique values or be mostly empty, contributing little to model training. The parameter columns_to_drop holds the list of columns to remove, and the script ensures that only those columns present in the dataset are dropped using data.columns. This process reduces the dimensionality of the data, making it more manageable for the model. After dropping irrelevant columns, missing

values are addressed. For categorical columns such as 'Residential Type', missing values are replaced with a placeholder ('Not Applicable') using df['Residential Type'].fillna(). For numerical columns, the missing values are imputed with the median value of the respective column, which is calculated using df[col].median() and filled using df[col].fillna(). This ensures the dataset is complete, preventing errors during model training. Rows with missing values in the target variable ('Property Type') are dropped using df.dropna(subset=['Property Type']) to maintain data integrity. The process concludes with visualizations that display the distribution of the target variable, sale amounts, and assessed values, aiding in further understanding of the data before modeling.

3.3. Feature Engineering & Selection

The next step involves feature engineering, which prepares the data for classification. The target variable, 'Property Type', is selected for classification. To simplify the classification task, the dataset is filtered to only include the top five most frequent property types, which are identified using df['Property Type'].value_counts().nlargest(5). This ensures that rare property types, which could introduce class imbalance or sparse data, are excluded from the model. This selection is based on the frequency of property types, ensuring the problem remains tractable and balanced. Categorical features are then identified using df.select_dtypes(include=['object']), and one-hot encoding is applied using pandas' pd.get_dummies() function. The parameter drop_first=True is used to prevent multicollinearity by dropping one of the encoded categories. One-hot encoding transforms categorical variables into a series of binary features, making them suitable for machine learning models that require numerical input. After encoding, the script defines the feature matrix X (which contains all the features except for the target variable) and the target variable y (which contains 'Property Type'). This step ensures that the data is prepared for the next stage of model training, where these features will be used to predict the target.

3.4. Model Training

The dataset is then split into training and testing sets using scikit-learn's train_test_split() function. The parameters test_size=0.2 and random_state=42 ensure that 20% of the data is reserved for testing, while the random_state ensures that the split is reproducible across different runs. The stratify=y parameter is used to maintain the proportional distribution of the target variable in both the training and testing sets, ensuring that the class distribution is consistent. If the stratified split fails due to rare classes, the script falls back to a non-stratified split. Once the data is split, the RF Classifier from sklearn.ensemble.RandomForestClassifier is initialized with several parameters: n_estimators=100 specifies that the forest should consist of 100 trees, random_state=42 ensures reproducibility, and n_jobs=-1 enables the use of all available CPU cores for faster model training. The model is then trained using rf_model.fit(X_train, y_train), where X_train is the feature matrix for the training data and y_train is the target variable. After training, the model is saved to disk using joblib.dump() to allow for easy reuse without retraining in the future.

3.5. Model Evaluation & Visualization

After the model is trained, it is evaluated using the test dataset. The model’s predictions are generated using rf_model.predict(X_test), and the accuracy is calculated using scikit-learn's accuracy_score(). This metric provides an overall sense of how well the model is performing by comparing the predicted values to the actual values in the test set. A classification report is generated using classification_report(), which includes detailed metrics like precision, recall, and F1-score for each class in the target variable. These metrics help assess how well the model performs for each class and are essential for understanding the model’s behavior. To further evaluate the model, visualizations are created to better understand its performance. A confusion matrix is generated using confusion_matrix() and visualized with seaborn's sns.heatmap() to display how the model’s predictions align with the true labels. This matrix shows the number of correct and incorrect predictions for each class, providing insights into where the model is making errors. Additionally, the importance of each feature in the model’s decision-making process is visualized using the rf_model.feature_importances_ attribute. A bar plot is created using seaborn's sns.barplot() to display the top 20 most important features, highlighting the features that contribute most to the model's predictions.

3.6. Checkpoint Saving

Finally, after training and evaluation, the trained RF model is saved for future use. The script ensures that a directory (saved_models) exists using os.makedirs(model_dir, exist_ok=True), and the model is saved to this directory with joblib.dump(rf_model, rf_path). The parameter rf_path specifies the location where the model will be stored. Saving the model allows it to be easily loaded in the future for predictions or further analysis, avoiding the need to retrain the model. This checkpoint also ensures that the model can be deployed in a production environment, where it can be used to classify new real estate data without retraining.

4. Results and Discussion

4.1. Result

4.1.1. Results of EDA and Preprocessing

The dataset was successfully loaded, with a total of 1,097,629 rows and 14 columns. The dataset contains information about property sales, including sale amount, assessed value, sales ratio, property type, residential type, and other related features. After loading the data, the first five rows were displayed, showing various property details like sale amounts, assessed values, and property types. The dataset initially included columns that were later identified as irrelevant for the classification task, such as 'Serial Number', 'List Year', 'Date Recorded', and 'Address'. These columns were dropped to focus the model on the essential features for predicting property types.

The dataset underwent extensive cleaning during the preprocessing phase. Irrelevant columns were removed, and missing values in 'Residential Type' were handled by filling them with 'Not Applicable'. Additionally, rows with missing values in the target variable, 'Property Type', were dropped to maintain the integrity of the analysis. After these preprocessing steps, the dataset was reduced to 715,183 entries and 6 columns, consisting of 'Town', 'Assessed Value', 'Sale Amount', 'Sales Ratio', 'Property Type', and 'Residential Type'. The cleaned dataset, now ready for feature engineering, was visualized (Figure 1) to understand the distributions of key variables, such as sale amounts and property types. The visualizations helped in identifying the presence of outliers and informed subsequent data transformations, such as the use of logarithmic scaling for the 'Sale Amount' and 'Assessed Value' to handle skewed distributions.

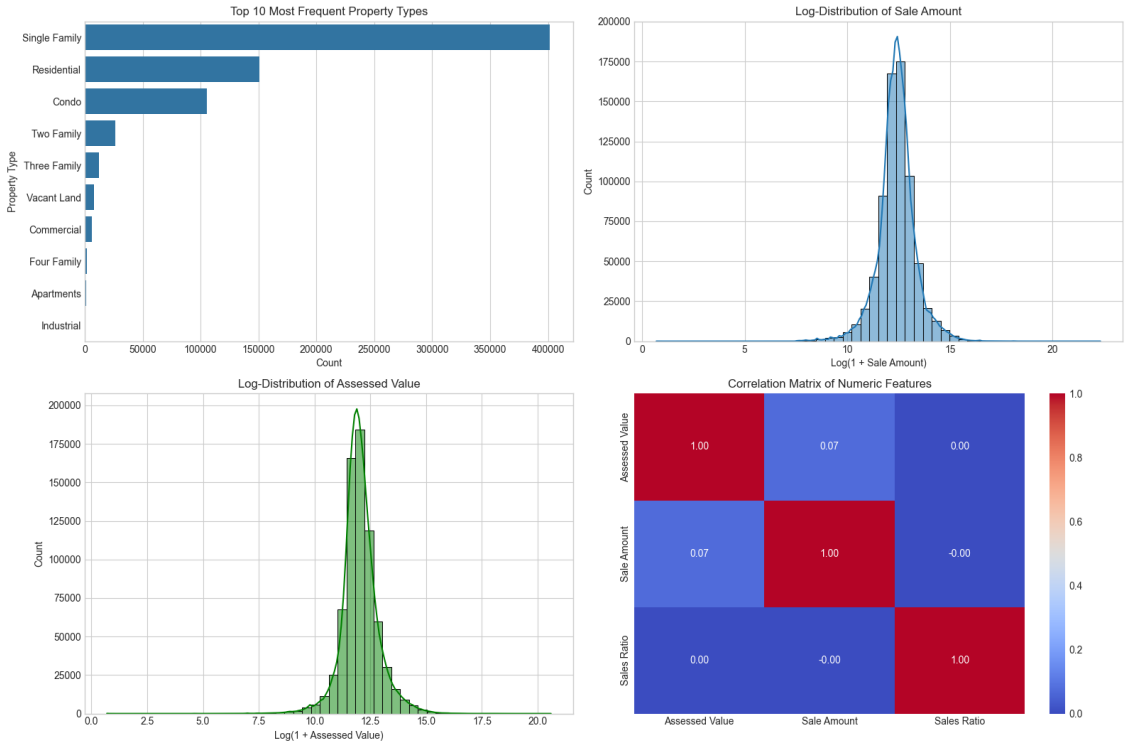

Figure 1. Exploratory Data Analysis Visualization

Feature engineering was carried out to prepare the dataset for modeling. The analysis focused on the top 5 most frequent property types: 'Single Family', 'Residential', 'Condo', 'Two Family', and 'Three Family'. These categories were selected to simplify the classification problem by excluding rare property types that could introduce noise. Categorical features, such as 'Town' and 'Residential Type', were one-hot encoded using pd.get_dummies() with drop_first=True to avoid multicollinearity. The dataset, now containing 177 columns, was transformed into a more suitable format for machine learning algorithms. The features included both the numeric columns (e.g., 'Assessed Value', 'Sale Amount', 'Sales Ratio') and the encoded categorical variables, ensuring that all data fed into the model was numeric and appropriate for training.

4.1.2. Results of Model Training and Evaluation

The data was split into training and testing sets using an 80/20 split, resulting in 557,672 rows for training and 139,418 rows for testing. This split ensured that the model was trained on a large portion of the data while reserving enough data for unbiased evaluation. The RF classifier was trained using the training set, and the model was configured with 100 trees (n_estimators=100) and a random state of 42 for reproducibility. The model was successfully trained and saved in the 'saved_models' directory using joblib for future use. The use of RF was chosen due to its robustness and ability to handle both categorical and numerical data efficiently. The model was trained to predict the 'Property Type', with the objective of classifying properties into one of the five selected categories.

The model's performance was evaluated using accuracy and a detailed classification report. The RF classifier achieved an overall accuracy of 82.98%. The classification report highlighted the precision, recall, and F1-score for each property type. For example, 'Single Family' achieved a high recall of 0.92, indicating the model's ability to correctly identify this class, while 'Residential' had a lower recall of 0.50, suggesting it was more challenging for the model to predict accurately. The F1-scores for each property type were generally high, with 'Condo' and 'Single Family' achieving F1-scores of 0.90, while 'Residential' had an F1-score of 0.56. This indicated that while the model performed well overall, it struggled with certain property types, particularly 'Residential'. The confusion matrix visualized the misclassifications, and a feature importance plot revealed which features were most influential in the model's predictions. This allowed for further refinement and understanding of the model's behavior.

Several visualizations were generated to support the evaluation of the RF model (Figure 2). A confusion matrix was plotted using seaborn's sns.heatmap() to illustrate the model's performance across different property types. The matrix clearly showed how well the model classified each property type, highlighting misclassifications between similar categories. Additionally, a feature importance plot was generated to display the most important features in the model. The feature importance plot showed that 'Sale Amount', 'Assessed Value', and 'Sales Ratio' were the most influential variables in predicting property types, as expected. These visualizations provided valuable insights into both the strengths and weaknesses of the model, suggesting potential areas for improvement, such as further tuning the model or adding more features.

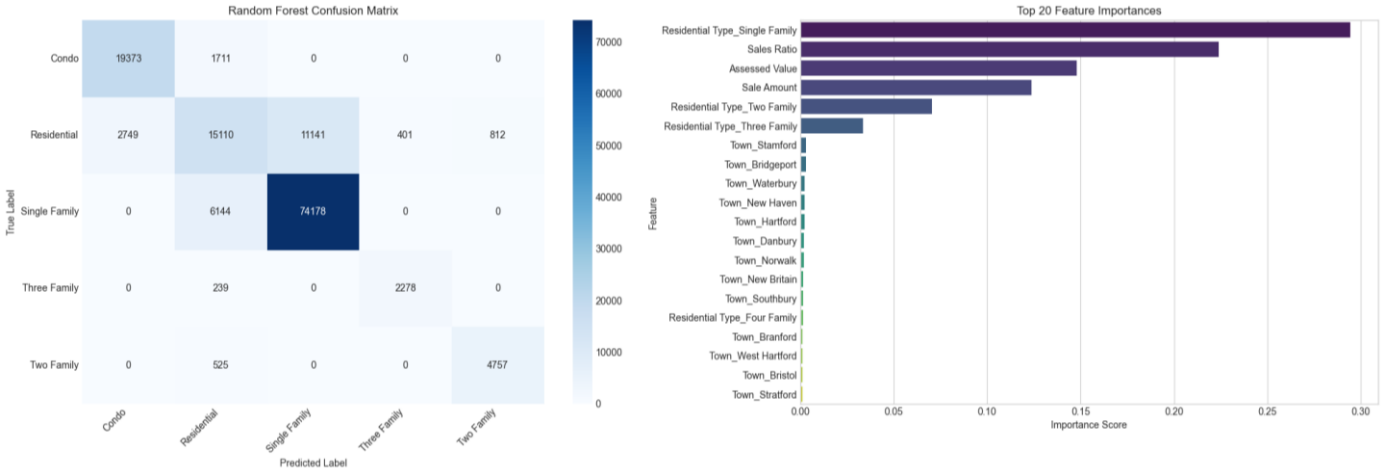

Figure 2. Random Forest Model Evaluation

4.2. Discussion

The RF model demonstrated solid performance, achieving an overall accuracy of 82.98% in classifying property types. This result is notable because RF is a robust ensemble learning method that handles both numerical and categorical data well. The model was able to correctly identify property types such as 'Single Family' and 'Condo' with high recall, suggesting that these categories were well-represented and easy to classify. The high performance in these categories can be attributed to the clear distinctions in features such as sale amount, assessed value, and sales ratio, which are often correlated with property type. These well-defined patterns likely allowed the RF algorithm to make accurate predictions.

However, the model struggled more with the 'Residential' category, which exhibited a much lower recall of 0.50. This lower performance can be explained by the variability within the 'Residential' category, which encompasses a wide range of property types, including single-family homes, multi-family units, and condos. The overlap between these property types might have made it more difficult for the model to draw clear boundaries. Additionally, the 'Residential' category may be influenced by factors such as location or specific residential characteristics, which the model may not have fully captured in the available features.

The feature importance analysis revealed that variables such as 'Sale Amount', 'Assessed Value', and 'Sales Ratio' were the most influential in determining property type classifications. This is expected, as these financial metrics are often strongly correlated with the property type. However, the performance could likely be improved by incorporating additional features, such as detailed geographical data or more granular property characteristics, which could better differentiate between property types within the 'Residential' category. Overall, the RF model performed well for most property types, but there is room for improvement, particularly with categories that have more complex or overlapping characteristics.

5. Conclusion

The RF algorithm performed well in classifying property sale categories, achieving an overall accuracy of 82.98%. This model demonstrated its strength in correctly classifying property types like 'Single Family' and 'Condo', where the distinctions between property types were clearer. However, it faced challenges with categories like 'Residential', which had a lower recall due to the inherent variability within this group. These results suggest that while RF is a robust algorithm for general classification tasks, improvements are needed for more complex property categories with overlapping characteristics. For real estate professionals, these findings have significant implications for market segmentation and property valuation. By utilizing machine learning models like RF, professionals can more accurately categorize properties, leading to better pricing strategies and targeted marketing efforts. However, the study's limitations, such as potential class imbalance and the lack of certain granular features (e.g., detailed property descriptions or geographic factors), should be addressed. Future research could explore incorporating additional features, such as neighborhood characteristics or property-specific data, and test alternative algorithms like SVM or Neural Networks to further refine property classification accuracy.

6. Declarations

6.1. Author Contributions

Conceptualization: L.E., M.S.F.; Methodology: L.E., M.S.F.; Software: L.E.; Validation: M.S.F.; Formal Analysis: L.E.; Investigation: L.E.; Resources: L.E.; Data Curation: L.E.; Writing – Original Draft Preparation: L.E.; Writing – Review and Editing: L.E., M.S.F.; Visualization: L.E.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. Ćetković, S. Lakić, M. Lazarevska, M. Žarković, S. Vujosevic, J. Cvijović, and M. Gogic, "Assessment of the Real Estate Market Value in the European Market by Artificial Neural Networks Application," *Complexity*, vol. 2018, pp. 1472957:1-1472957:10, 2018, doi: 10.1155/2018/1472957.

[2] K. H. Liow and S. Y. Angela, "Return and Co-Movement of Major Public Real Estate Markets During Global Financial Crisis," *J. Prop. Invest. Finance*, vol. 35, no. 5, pp. 489-508, 2017, doi: 10.1108/jpif-01-2017-0002.

[3] P. D. Paola, "Real Estate Valuations With Small Dataset: A Novel Method Based on the Maximum Entropy Principle and Lagrange Multipliers," *Real Estate*, vol. 10, no. 1, pp. 1-10, 2024, doi: 10.3390/realestate1010003.

[4] V. D. Giudice, B. Manganelli, and P. D. Paola, "Hedonic Analysis of Housing Sales Prices with Semiparametric Methods," *Int. J. Agric. Environ. Inf. Syst.*, vol. 8, no. 2, pp. 65-77, 2017, doi: 10.4018/IJAEIS.2017040105.

[5] O. A. K'Akumu, "What Is Real Estate? Five Ontological Questions For the Discipline," *J. Eur. Real Estate Res.*, vol. 16, no. 2, pp. 155-171, 2023, doi: 10.1108/jerer-08-2022-0022.

[6] J. Sobieraj and D. Metelski, "Machine Learning Insights: Exploring Key Factors Influencing Sale-to-List Ratio Insights From SVM Classification and Recursive Feature Selection in the US Real Estate Market," *Buildings*, vol. 14, no. 5, pp. 1471, 2024, doi: 10.3390/buildings14051471.

[7] S. Menghini, V. A. Sottini, and R. Fratini, "From Fair Market Value to Judicial Market Value of Real Estate," *Aestimum*, vol. 84, pp. 19-29, 2024, doi: 10.36253/aestim-15228.

[8] M. Kuru, O. Y. Erdem, and G. Çalış, "Sale Price Classification Models for Real Estate Appraisal," *Rev. la constr.*, vol. 20, no. 3, pp. 440, 2021, doi: 10.7764/rdlc.20.3.440.

[9] M. Nikou, G. Mansourfar, and J. Bagherzadeh, "Stock Price Prediction Using Deep Learning Algorithm and Its Comparison with Machine Learning Algorithms," *Intell. Syst. Account. Finance Manag.*, vol. 26, no. 4, pp. 164–174, 2019, doi: 10.1002/isaf.1459.

[10] S. C. Sevgen and Y. Aliefendioğlu, "Mass Appraisal with a Machine Learning Algorithm: Random Forest Regression," *Bilişim Teknol. Derg.*, vol. 20, no. 3, pp. 440, 2020, doi: 10.17671/gazibtd.555784.

[11] T. Hariguna and A. S. M. Al-Rawahna, "Unsupervised Anomaly Detection in Digital Currency Trading: A Clustering and Density-Based Approach Using Bitcoin Data," *J. Curr. Res. Blockchain*, vol. 1, no. 1, pp. 70–90, 2024, doi: 10.47738/jcrb.v1i1.12.

[12] G. Wang, C. Chen, Z. Jiang, G. Li, C. Wu, and S. Li, "Efficient Use of Biological Data in the Web 3.0 Era by Applying Nonfungible Token Technology," *J. Med. Internet Res.*, vol. 26, p. e46160, 2024, doi: 10.2196/46160.

[13] Z. Ouyang, "Research on the Diamond Price Prediction Based on Linear Regression, Decision Tree and Random Forest," *Highlights Bus. Econ. Manag.*, vol. 24, pp. 248-257, 2024, doi: 10.54097/13ccwv59.

[14] R. B. Abidoye, J. Ma, T. Y. M. Lam, T. Oyedokun, and M. Tipping, "Property Valuation Methods in Practice: Evidence From Australia," *Property Manag.*, vol. 37, no. 2, pp. 701-718, 2019, doi: 10.1108/PM-04-2019-0018.

[15] Hery and A. E. Widjaja, "Analysis of Apriori and FP-Growth Algorithms for Market Basket Insights: A Case Study of The Bread Basket Bakery Sales," *J. Digit. Market. Digit. Currency*, vol. 1, no. 1, pp. 63–83, 2024, doi: 10.47738/jdmdc.v1i1.2.

[16] D. Shi, J. Zurada, H. Zhang, Z. Chen, J. Guan, and X. Li, "Deep Learning in Predicting Real Estate Property Prices: A Comparative Study," *Proc. 2023 Hawaii Int. Conf. Syst. Sci.*, pp. 970-979, 2023, doi: 10.24251/hicss.2023.120.

[17] **N**. S. Ja'afar, J. Mohamad, and S. Ismail, "Machine Learning for Property Price Prediction and Price Valuation: A Systematic Literature Review," *Plan. Malaysia*, vol. 19, no. 3, pp. 411–422, 2021, doi: 10.21837/pm.v19i17.1018.

[18] A. B. Prasetio, B. bin M. Aboobaider, and A. bin Ahmad, "Assessing Geographic Disparities in Campus Killings: A Data Mining Approach Using Cluster Analysis to Identify Demographic Patterns and Legal Implications," *J. Cyber Law*, vol. 1, no. 1, pp. 1–22, 2025, doi: 10.63913/jcl.v1i1.1.

[19] M. Yazdani, "House Price Determinants and Market Segmentation in Boulder, Colorado: A Hedonic Price Approach," 2021, doi: 10.48550/arxiv.2108.02442.

[20] O. Pirogova, Y. Dovganeva, and N. Pogorelov, "Approach to the Assessment of the Competitiveness of Real Estate Objects in Retail," *E3S Web Conf.*, vol. 389, pp. 1-9, 2023, doi: 10.1051/e3sconf/202338909024.

[21] D. A. Dewi and T. B. Kurniawan, "Exploring Financial Trends through Topic Modeling and Time-Series Analysis: A Clustering Approach Using Latent Dirichlet Allocation (LDA) on Twitter Data," *J. Digit. Soc.*, vol. 1, no. 1, pp. 91-108, 2025, doi: 10.63913/jds.v1i1.5.

[22] R. Tanamal, N. Minoque, T. Wiradinata, Y. S. Soekamto, and T. Ratih, "House Price Prediction Model Using Random Forest in Surabaya City," *TEM J.*, vol. 12, no. 1, pp. 126-132, 2023, doi: 10.18421/tem121-17.

[23] T. Sangsawang and L. Yang, "Predicting Student Achievement Using Socioeconomic and School-Level Factors," *Artif. Intell. Learn.*, vol. 1, no. 1, pp. 20-34, 2025, doi: 10.63913/jcl.v1i2.4.

[24] J. Li and S. Chao, "A Novel Twin-Support Vector Machine for Binary Classification to Imbalanced Data," *Data Technol. Appl.*, vol. 57, no. 3, pp. 385–396, 2023, doi: 10.1108/dta-08-2022-0302.

[25] S. A. Ghaffar and W. C. Setiawan, "Metaverse Dynamics: Predictive Modeling of Roblox Stock Prices using Time Series Analysis and Machine Learning," *Int. J. Res. Metaverse*, vol. 1, no. 1, pp. 77–93, 2024, doi: 10.47738/ijrm.v1i1.6.

[26] Y. Tang, "The Impact of Brake Failure Rights Protection Event on Tesla Motors: Stock Prediction Based on ARIMA Model," *SHS Web Conf.*, vol. 188, p. 01011, 2024, doi: 10.1051/shsconf/202418801011.

[27] Y. Duan, K. Zhao, Y. Guo, and X. Wang, "Early Warning of Commercial Housing Market Based on Bagging-Gwo-SVM," *Comput. Syst. Sci. Eng.*, vol. 45, no. 2, pp. 2207-2222, 2023, doi: 10.32604/csse.2023.032297.

[28] J. Hong, H. Choi, and W. Kim, "A House Price Valuation Based on the Random Forest Approach: The Mass Appraisal of Residential Property in South Korea," *Int. J. Strateg. Prop. Manag.*, vol. 24, no. 3, pp. 140-152, 2020, doi: 10.3846/ijspm.2020.11544.

[29] S. Levantesi and G. Piscopo, "The Importance of Economic Variables on London Real Estate Market: A Random Forest Approach," *Risks*, vol. 8, no. 4, pp. 112, 2020, doi: 10.3390/risks8040112.