COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR TIME ON MARKET (TOM)

ALI RAHBARIMANESH

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2087663

COMMITTEE

dr. Javad Pourmostafa Roshan Sharami
dr. Koen Haak

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

January 22th, 2024

WORD COUNT

8546

ACKNOWLEDGMENTS

I want to thank Dr. Javad Pourmostafa Roshan Sharami for his unwavering support, valuable guidance, and scientific insights at the end of my master's thesis. Dr. Pourmostafa's expertise and encouragement have significantly contributed to this research's accuracy and scientific quality. His commitment to creating a conducive learning environment and commitment to the pursuit of knowledge has shaped my academic path. I am truly fortunate to have had the privilege of working under him.
Ali Rahbarimanesh

COMPARATIVE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR TIME ON MARKET (TOM)

ALI RAHBARIMANESH

Abstract

This study investigates the effectiveness of machine learning (ML) in predicting "Time on Market" (TOM) in the Netherlands. This approach significantly departs from traditional methods, such as the Ordinary Least Squares (OLS). The research utilizes an extensive dataset from the Funda-Sold database, encompassing Dutch house sales from 2020 to 2023. Advanced ML algorithms like XGBoost, CatBoost, LightGBM, and Random Forest are applied, focusing on using Recursive Feature Elimination with Cross-Validation (RFECV) for feature selection. This technique is measured against each algorithm's inherent feature importance methods, aiming to streamline the dataset and enhance the understanding of critical factors affecting TOM. The results demonstrate that ML algorithms, especially CatBoost, outperform OLS accuracy, signaling noteworthy progress in real estate analytics. By integrating ML, this research offers real estate professionals more refined tools for market analysis, facilitating improved decision-making and policy formulation in the evolving Dutch housing market.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

- The dataset for this study is sourced from "Kaggle.com" (2023), a renowned platform for hosting various open-source datasets. Regarding coding, ChatGPT-3.5 (OpenAI, 2023) was employed as a debugging tool to resolve coding errors in certain sections and sometimes for paraphrasing. Additionally, for assistance in academic writing and grammar, Grammarly (Grammarly, 2024) was utilized, while Quillbot (QuillBot, 2023) aided in paraphrasing some parts of the thesis.

- Data cleaning, preprocessing, and modeling tasks are conducted using Python Version 3.9.13 (Python Software Foundation, 2022). In this study, several key libraries played pivotal roles in data manipulation and machine learning. These libraries included Pandas (Version 2.0.3), NumPy (Version 1.24.3), Scikit-learn (Version 1.3.0), and RFECV (Version 1.3.0). Their contributions were instrumental in data preprocessing, feature engineering, and model development. The RandomForestClassifier from Scikit-learn and, specifically, the Recursive Feature Elimination with Cross-Validation (RFECV) technique was employed for feature selection (Python Software Foundation, 2022). All figures presented in the thesis were created by the author using Matplotlib (Version 3.7.2) and Seaborn (Version 0.12.2) libraries for Python.

- Notably, three popular gradient-boosting libraries—XGBoost, CatBoost, and LightGBM—were employed from their respective "GitHub" (2023) repositories, which are publicly available (Chen & Guestrin, 2023; Microsoft, 2023; Yandex, 2023).

2 INTRODUCTION

With its multifaceted nature, the real estate industry is a vital part of national and international economies and consistently attracts the attention of business managers and researchers (Potrawa & Tetereva, 2022). The rich and accessible data sources and their unique qualities make it an exciting and dynamic field of study. Researchers often use it as a proxy for investigating various unobserved phenomena, including environmental impacts, micro- and macroeconomic trends, and other social aspects. Furthermore, the sector is a vital indicator of overall economic health, reflecting the interplay between consumer behavior, governmental policies, and global economic shifts (Potrawa & Tetereva, 2022).

There are various criteria in the real estate industry, including the inherent liquidity challenges of real estate investments. Liquidity, the ease of selling and turning real estate into cash, is commonly measured by the metric "time on market" (TOM) (Zhu et al., 2016). Understanding TOM is critical for market participants in the Netherlands and elsewhere, as it represents how long a property stays on the market before being sold (Lippman & McCall, 1986; Zhu et al., 2016). Additionally, TOM is a valuable indicator for potential buyers that provides insights into the popularity and demand for a property (Taylor, 1999; Tucker et al., 2013).

More broadly, TOM represents real estate investment risk at the macroeconomic level (Zhu et al., 2016). Understanding and having a more accurate

approximation of TOM is essential to a real estate professional's business plan, as it is to individual sellers and investors, as the industry is highly dependent on dynamic cash flow. Through detailed analysis of TOM, marketers may create customized programs precisely tailored to the characteristics of asset types, price dynamics, and current market patterns. According to Cheng et al. (2008), the customized strategy ensures faster and more effective turnover, simplifying the sales process.

Prior Research on TOM in real estate, both in the Netherlands and globally, has predominantly utilized statistical methods like Ordinary Least Squares (OLS) and two-stage Least Squares (2SLS) models, which have been vital in understanding the impact of market conditions and property features on TOM (Awad & Fraihat, 2023; Reber, 2017). However, it is noteworthy that the existing research has primarily focused on understanding rather than predicting TOM. The reliance on these predominantly linear and traditional methods has underscored a significant gap which is the need for more advanced data analysis techniques in this domain.

In contrast, machine learning (ML) and artificial intelligence (AI) have significantly influenced real estate analytics. However, their application to TOM prediction still needs to be improved, especially compared to the extensive research on property price forecasting (Askarisichani et al., 2022; Kars, 2021). This underexplored area in the research landscape highlights the need for a comprehensive exploration of ML approaches to predict TOM and move beyond the traditional statistical methodologies, empowering real estate professionals to make informed decisions in an ever-evolving housing market. (Zhu et al., 2016).

To address the existing gap in real estate analytics, this study adopted a novel approach by employing three gradient boosting algorithms—namely, Extreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), and Light Gradient Boosting Machine (LightGBM)—for the prediction of TOM within the Dutch real estate industry. This study uses a dataset from "Funda", a leading real estate company in the Netherlands. While these three algorithms have proven their ability to predict real estate prices, their application to TOM remains largely unexplored in the current knowledge landscape. Their prowess in handling intricate, nonlinear datasets and their versatility, interpretability, and resilience against overfitting underscores their promise as robust candidates for this specific application.

In addition, this study includes random forest, an algorithm commonly used in real estate analysis, although not widely used for TOM prediction. OLS regression, a traditional statistical method commonly used in real estate analysis, is also included to provide a basis for comparison of a conventional method used in prior studies on TOM inspired by a study by Yang (2023). For identifying crucial features in real estate analytics,

especially with extensive datasets, this study employs Recursive Feature Elimination with Cross-Validation (RFECV) for feature selection and dimensionality reduction as explained in methodology 4.3.4 (Awad & Fraihat, 2023).

2.1 *Research Questions*

In pursuing the research objectives, a careful selection of advanced ML algorithms, including Random Forest, XGBoost, CatBoost, and LightGBM, has been made to predict TOM for residential properties within the dynamic Dutch real estate market. Additionally, the conventional statistical method, namely OLS, has been included as a baseline for comparison. The primary research question encapsulates the core of the inquiry:

M-RQ *How do the selected ML algorithms and OLS compare in predicting TOM for residential properties in the Dutch real estate market, and what insights can be gained from this comparative analysis?*

Complementary to the main research question, two subsidiary research questions have been designed to delve deeper into critical aspects of the investigation. The analysis will compare RFECV with each algorithm's built-in feature importance, providing a comprehensive understanding of TOM dynamics in the Dutch real estate market.

S-RQ1 *Given the application of RFECV to finding optimal features, what are the critical predictors identified as most influential for estimating TOM, and how does RFECV enhance the predictive accuracy of the ML models?*

S-RQ2 *What pivotal features do each of the selected ML algorithms identify as influential in predicting TOM, considering their inherent ranking mechanisms, and how do these rankings compare to each other and RFECV?*

2.2 *Societal And Scientific Relevance*

This research introduces the application of ML techniques for TOM prediction in the real estate sector, a new approach not previously explored in the existing literature. It represents a paradigm shift from traditional statistical methods to advanced ML algorithms, namely XGBoost, CatBoost, LightGBM, and Random Forest. The goal is to increase prediction accuracy and deepen the understanding of TOM dynamics, thereby addressing a significant gap in real estate analysis. By applying these sophisticated ML models, this study can provide powerful decision-support tools for various stakeholders in the real estate market, including sellers, buyers, investors,

and industry professionals. This innovative approach contributes to the theoretical understanding of TOM and creates new benchmarks in the methodological practices of real estate market analysis.

2.3 *Findings*

This comprehensive study employed ML algorithms, namely Random Forest, XGBoost, CatBoost, and LightGBM, to predict the TOM for residential properties in the Netherlands. Each model demonstrated a moderate performance but remarkable improvement in accuracy over the traditional OLS method, with CatBoost emerging as the standout performer. A significant element of this research involved using RFECV and the built-in feature importance metrics within each algorithm. Interestingly, while the specific ranking of important features varied between the models, the overall performance of these models on the reduced dataset showed striking similarities. This consistency underscores the robustness and reliability of these advanced ML techniques in accurately forecasting TOM in the real estate market, highlighting their potential to revolutionize market analysis and decision-making processes in the industry.

3 RELATED WORK

The following section provides a concise theoretical background for this research. It reviews previous studies on TOM and its relevance in real estate research. Additionally, it explores the incorporation of ML into real estate analysis, highlighting the evolving landscape of data-driven approaches in this field.

3.1 *Conventional Approaches and Their Limitations*

A statistical study by Luna Andonegui (2023) delved into the influence of market conditions and price on the TOM in the Swedish real estate market and found that properties tend to sell quickly with rising prices in a growing market. In contrast, the opposite occurs in a declining market. Another Academic work by Yu et al. (2021) employs statistical methods such as 2SLS and OLS to assess the impact of virtual tours on real estate transactions and TOM. The study found that properties with virtual tours take longer to sell and are associated with longer TOM. These studies highlight the importance of property market conditions, property characteristics, and TOM relationships while suggesting exploring alternative models for deeper insights and TOM prediction.

Another research on various design and architectural features and their impact on the TOM for properties and buildings listed in five cities in the Dutch real estate market was conducted in a statistical review by BuHamdan et al. (2022). They highlighted the challenge of incorporating qualitative sustainability aspects, particularly social and cultural dimensions, in the initial design phase of construction projects. Also, Aydin et al. (2019) in Their analysis focused on the influence of energy performance certification on TOM within the Dutch housing market. While their study revealed a significant reduction in TOM associated with energy labels, it primarily centered on energy performance certification (EPC), did not consider broader parameters and features, and had certain limitations, including potential biases in data collection.

In their research, Cajias et al. (2019) extensively analyzed the German housing market. They discovered that properties with energy-saving features appeal more to renters and buyers, resulting in shorter TOM. Furthermore, as highlighted by Benefield et al. (2014), many other studies have predominantly relied on statistical techniques such as OLS, Hazard Models, and 2SLS, often overlooking the potential of data science techniques and ML for TOM prediction and the identification of non-linear relationships within complex datasets. In summary, while these studies provide insights into property characteristics and TOM, they underscore a promising direction for future research. ML algorithms have the potential to uncover concealed insights beyond the reach of traditional statistical methods, particularly in terms of enhancing prediction accuracy (Potrawa & Tetereva, 2022).

3.2 *Transition to ML in Real Estate Studies*

According to Potrawa and Tetereva (2022), recent advancements in ML and AI methodologies hold significant promise for enhancing real estate and marketing studies. ML algorithms, including neural networks, boosted trees, random forests, and support vector machines (SVM), have demonstrated superior predictive capabilities compared to traditional models (Hong et al., 2020; Neloy et al., 2019). Navigating the complex realm of real estate prediction requires precise consideration of predictive features. Choosing an optimal feature selection method is crucial for enhancing interpretability and predictive accuracy. Numerous studies have explored different feature selection approaches, including wrapper, filter, and embedded methods, underscoring the critical role of feature selection in real estate research (Naotunna, 2023).

In a real estate research study by Jha et al. (2020), diverse ML algorithms were utilized for housing price prediction, including XGBoost, CatBoost,

and Random Forest. Remarkably, XGBoost demonstrated superior predictive accuracy compared to other models. The study incorporated Pearson correlation as a filter method for feature selection. However, it is crucial to note the limitations associated with this method, particularly its sensitivity to linearity and assumptions of normal distribution. Despite these considerations, the study's findings emphasize the effectiveness of XGBoost in housing price prediction within the specific context of the research (Jha et al., 2020).

In a distinct study conducted by Abut et al. (2023) for price prediction, they employed Support Vector Regression (SVR) and LightGBM, integrating a filter-based method, namely Minimum Redundancy Maximum Relevance (MRMR) for feature selection. In this study, LightGBM, in particular, outperformed SVR, achieving lower MAE and Root Mean Squared Error RMSE values. The study's results revealed a significant improvement in prediction accuracy, consistently favoring the hybrid approach. While acknowledging limitations related to dataset specifics and regional focus, the study provided valuable insights and proposed avenues for future research, including incorporating additional features and alternative feature selections and algorithms (Abut et al., 2023).

Several ML algorithms, including random forest models, have been employed by Potrawa and Tetereva (2022) to develop a robust pricing system for the Dutch real estate market. While random forest models may not be considered the most advanced ML algorithms, they have demonstrated their effectiveness in this research. In addition, Guliker et al. (2022) compared linear regression and XGBoost models to predict real estate prices in five cities in the Netherlands. XGBoost encompassed linear regression and spatially weighted regression. The results emphasized the outstanding accuracy of the XGBoost model and highlighted its remarkable potential as a powerful tool in real estate evaluation (Guliker et al., 2022; Potrawa & Tetereva, 2022).

Nonetheless, it is crucial to acknowledge that these studies had limitations, such as small sample sizes, potential language biases in data sources, and limited data collection periods during events like the first wave of COVID-19. Other challenges include difficulty modeling high-end property values influenced by individual preferences, heavy reliance on Dutch-specific variables, and the need to address monthly fluctuations in real estate prices. The need for further research and development in other real estate criteria in this area has been shown in their studies (Guliker et al., 2022; Potrawa & Tetereva, 2022).

In research in Sri Lanka by Naotunna (2023), they employed five ML algorithms - Multiple Linear Regression, Random Forest, SVR, Extra Trees Regression, and XGBoost for land price prediction. The performance of

each model was systematically enhanced through two key optimization strategies: feature reduction and hyper-parameter optimization. The study employed two distinct approaches for feature reduction: RFECV and SelectKBest. Among the five ML algorithms, the Random Forest outperformed the others. Model performance obtained through the SelectKBest method is similar to that obtained for the ML algorithms after feature reduction through RFECV (Naotunna, 2023).

In this study, the primary contribution to the literature is the introduction of ML algorithms to TOM research, a notable departure from the conventional statistical methods that have been predominantly used, as highlighted by Luna Andonegui (2023) and Potrawa and Tetereva (2022). This approach shifts the focus from traditional real estate topics like property price prediction to a more sophisticated exploration of TOM. The study innovates further by employing RFECV for feature selection, showcasing an advanced TOM analysis technique for identifying critical predictors. This methodological advancement can enhance the accuracy of TOM predictions and provide a more detailed understanding of the factors influencing TOM.

4 METHOD

This section explains the method used to achieve the research objectives. It begins with an overview of the dataset, including insights into its composition and characteristics. Following this, it covers data cleaning and preprocessing. This section then deals with the selected algorithms and techniques and provides a rationale and justification for their selection. The methodology concludes with a thorough discussion of the evaluation process, detailing the criteria used to evaluate the performance of models. This brief yet comprehensive review ensures clarity and precision in the research approach.

4.1 *Dataset*

The Funda dataset (2020-2023) by Yang (2023) in **ka<empty citation>** is an extensive and detailed aggregation of data from the sold properties in the Dutch real estate market, serving as the empirical foundation for this study. A 1 MB CSV file includes 9,884 unique real estate listings (instances), each detailed with critical information. These records are rich in data, presenting 112 distinct variables that comprehensively examine the properties. The variables cover crucial attributes such as listing prices, property sizes, number of rooms, and energy efficiency labels, offering a holistic view of the market's characteristics and trends. This depth of

information facilitates a thorough analysis of the real estate landscape in the Netherlands. Significantly, the range of the target variable spans from 0 to 1,209 days, covering scenarios from homes that sold immediately on the listing day to properties that remained on the market for nearly three years.

Given the contemporary focus on sustainability, including energy efficiency labels in the dataset is particularly pertinent, aligning the study with current environmental concerns and market dynamics. Initially compiled in Dutch, the dataset has been translated into English, ensuring clarity and broad accessibility for international research. This translation maintains the data's accuracy and expands its utility beyond the Dutch-speaking community. The comprehensive nature of the Funda dataset, provided by one of the leading real estate platforms in the Netherlands, makes it an invaluable tool for understanding property valuation, buyer preferences, and market behaviors in the context of evolving economic and environmental factors.

4.2 *Exploratory Data Analysis*

Exploratory data analysis (EDA), as introduced by Tukey et al. (1977), is a fundamental starting point and a necessary step for data scientists to understand and visualize a dataset. In the following, visualizations of key variables within the dataset are presented.

TOM - PRICE:

The distribution of the `Last_asking_price` in **Figure** 1 is demonstrated. On the x-axis, the latest asking price of the property is shown in Euros and is classified into bins, providing additional information through color gradients. The y-axis shows the `Number_of_days_until_sale`, the target variable, TOM.

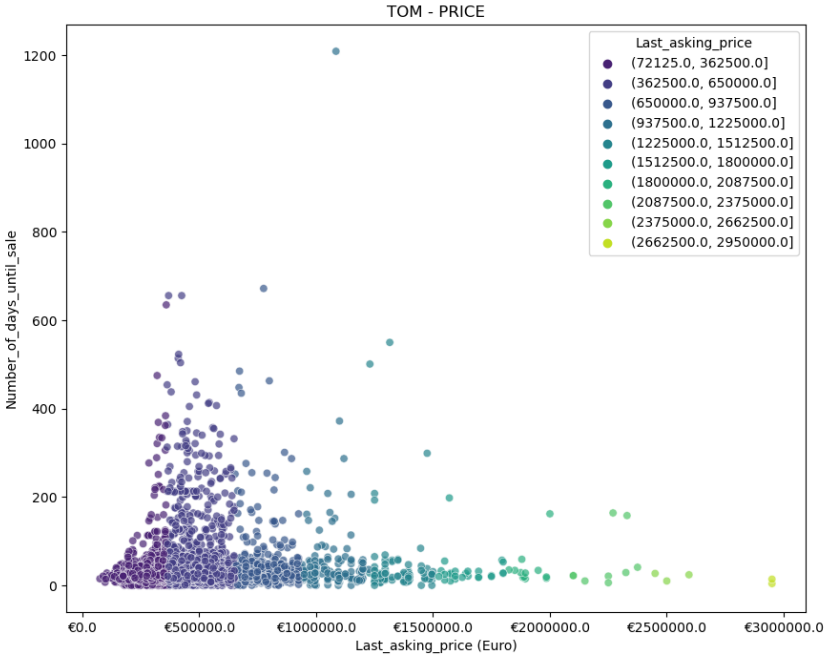

Figure 1: TOM - PRICE

The skewness of `Last_asking_price` variable is also calculated using the `skew` function from the `scipy.stats` library and obtained a value of 2.83. This positive skewness indicates a rightward skew in the distribution, indicating a longer right tail and the presence of relatively higher final asking prices.

Energy Labels:

The distribution of `Energy_label` as an essential indicator of a property's environmental friendliness and energy efficiency and nowadays is one of the most critical factors in the real estate industry, including observed and missing values, is shown in **Figure** 2. The bar plot provides an overview of each energy label category. Notably, the analysis indicates that 5.2% of the data set shows missing values for this attribute.

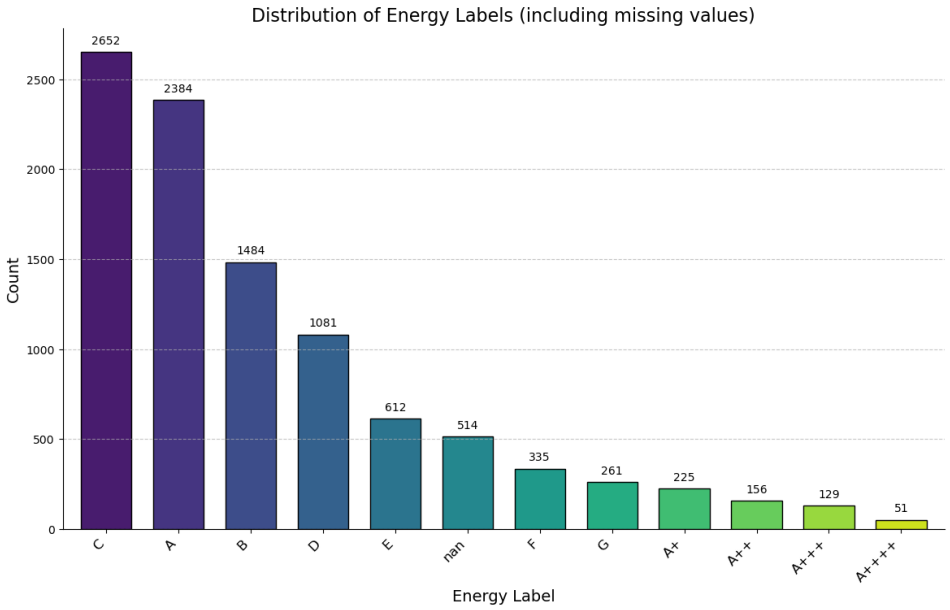

Figure 2: Energy Labels Distribution

The bar plot clearly illustrates the distribution of energy labels within the dataset, with 'C' and 'A' being the most frequently occurring categories, indicating a common standard of energy efficiency among the properties analyzed. In stark contrast, the 'A++++' label is the least represented, suggesting that such high levels of energy efficiency are exceptionally rare, perhaps due to higher costs or more advanced technology requirements.

Living Space - TOM:

The relationship between `Living_space_in_m2` and the target variable, TOM, represented In **Figure** 3. The skewness of `Living_space_in_m2` is 1.46. A positive skewness of this magnitude indicates that the life space distribution is skewed to the right. It suggests that relatively larger living spaces may contribute to a longer right tail in the distribution.

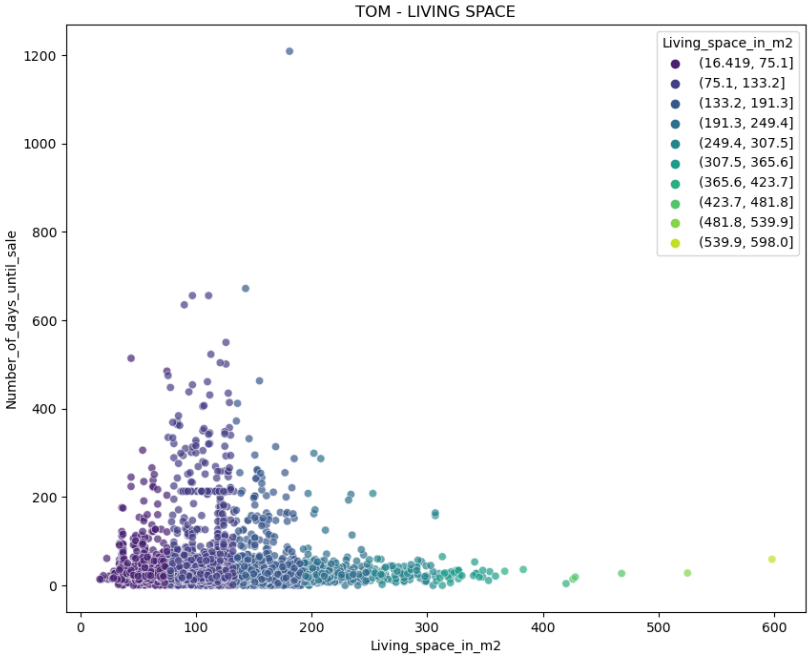

Figure 3: TOM - Living Space

4.3 *Data Cleaning and Preprocessing*

The following parts describe the basic procedures for dataset preparation in this study, including tasks such as handling missing data, deciding on outliers, encoding categorical features, and feature selection. These preprocessing and data cleaning steps are critical to ensure the dataset is ready for analysis and maintain the quality and accuracy of study results.

4.3.1 *Handling Missing Data*

According to Enders (2003) criteria, variables with more than 20% missing values were determined to be excluded from the analysis. In this data set, only the `Energy_label` that is demonstrated in **Figure** 2 showed a 5.2% missing values and, instead of deletion, to preserve all the records of the dataset was treated through an imputation method by mode. This approach involves replacing the missing values with the most frequent label in the variable (Mirzaei et al., 2022).

4.3.2 *Handling Outliers and Skewed Variables*

Another crucial and challenging part of the cleaning process involved dealing with outliers and skewed variables. The initial approach involved considering constraints on these outliers to moderate their influence by techniques such as Winsorization and Yeojohnson, a method supported by Jiang and Shekhar (2017). However, after a thorough analysis, the choice was not to apply specific alterations or correct outliers in the study. This decision aligns with the research conducted by Kars (2021), which emphasized preserving the original data distribution to ensure its relevance in real-world scenarios. This decision reflects an awareness that real-world data, especially in complex markets like real estate, frequently exhibits inherent variations that models should be capable of handling without any adjustments (Kars, 2021).

4.3.3 *Handling Categorical Features*

In this study, categorical variable preprocessing followed best practices in data science. Label encoding with `LabelEncoder` from Scikit-Learn was employed for the `Energy_label` variable, maintaining its ordinal relationships for improved pattern recognition and prediction capabilities (Ma & Zhang, 2020). For the inherently unordered feature `Type_of_construction`, one-hot encoding was utilized to create binary columns, preserving category distinctions without introducing artificial relationships (Ma & Zhang, 2020; Zeng, 2023). Due to the having 86 unique values in `Roof_type` and `Type_of_residence` with 104 unique values, frequency encoding was applied by mapping each category to its frequency or percentage in the dataset using `map` function and `value_counts` method (Uyar et al., 2009).

4.3.4 *Feature Selection*

In predictive modeling, especially in areas like real estate market analysis where datasets are extensive, the performance of an algorithm is deeply connected to the quality and relevance of its features. A crucial initial step in improving these models involves identifying and eliminating redundant features from the dataset. For instance, specific features, namely `offer_time`, `sale_date`, and `status`, were determined to be redundant and were subsequently removed. This selective pruning of the dataset is fundamental in refining the model's effectiveness and precision (Otchere et al., 2022).

After eliminating the redundant variables, a methodology inspired by Awad and Fraihat (2023) is employed. This methodology effectively utilized Scikit-learn's `RFECV` implementation in conjunction with a Decision Tree (DT) and used the Root Mean Square Error as the evaluation metric for

feature selection and the identification of optimal features. This approach not only ensures a consistently streamlined dataset but also enhances computational efficiency. Subsequently, in their study, this refined dataset, crafted through Recursive Feature Elimination with Cross-Validation and Decision Trees (RFECV-DT), was similarly applied to algorithms such as Random Forest and XGBoost for predictive modeling (Awad & Fraihat, 2023).

Following the implementation of RFECV-DT, the next step involves conducting a comparative analysis of the importance of TOP N features across various algorithms used in the study. This experiment will examine each algorithm's built-in feature importance rankings and compare these findings among themselves and with the feature selection outcomes derived from the RFECV-DT with the Decision Tree process (Lai et al., 2019). This comprehensive analysis is essential to validate the effectiveness of the RFECV-DT approach and gain insights into the subtleties of feature selection within different predictive models.

4.4 *Algorithms*

This study integrates conventional and innovative techniques to predict TOM in the real estate sector. Common in statistical research known for its focus on linearity and interpretability, OLS is a foundational baseline, providing a comparative benchmark for more complex models. It is worth noting that the primary objective of this selection is to compare the performance differences between the models similar to the approach by Yang (2023) rather than exploring specific details of OLS. Alongside OLS, ML algorithms considered for this study are explained in the following.

4.4.1 *Random Forest*

Random Forest, a supervised learning algorithm well-suited for housing market analysis, is selected for this study due to its robust predictive capabilities. It exhibits resilience to outliers to some extent and excels in handling complex, high-dimensional datasets, as evidenced by prior research (Breiman, 2001; Čeh et al., 2018). Additionally, Random Forest is skilled at mitigating overfitting and identifying nonlinear relationships among features, making it highly effective for TOM prediction, as supported by recent studies (Mostert, 2022). Its ensemble approach and Built-in feature importance ability are pivotal for the study's goals (Biggs et al., 2023; Rico-Juan & de La Paz, 2021).

4.4.2 *XGBoost*

This study selected XGBoost for its renowned high scalability and efficiency, as demonstrated in prior research (Chen & Guestrin, 2016; Senthilkumar, 2023). It excels in managing datasets with numerous variables, a crucial requirement for the comprehensive analysis needed in TOM predictions (Chen & Guestrin, 2016; Senthilkumar, 2023). Additionally, XGBoost's built-in feature importance analysis further enhances its suitability as a potent tool for this research, aligning to apply advanced ML algorithms to TOM prediction and analysis (Chen & Guestrin, 2016).

4.4.3 *CatBoost*

CatBoost is a valuable asset for predictive modeling in real estate, and its capabilities extend beyond categorical data handling. Its gradient boosting framework is a key feature, where it continuously builds decision trees to improve prediction accuracy. This aspect is crucial in real estate analysis, as each incremental improvement in accuracy can significantly enhance the model's reliability (Senthilkumar, 2023). Additionally, CatBoost's ability to process large and complex datasets efficiently makes it ideal for the varied and extensive data often encountered in real estate market analysis.(Senthilkumar, 2023). CatBoost's built-in feature importance also makes it, like other selected algorithms, a rational choice for this study.

4.4.4 *LightGBM*

This thesis includes LightGBM for its proficiency in processing complex real estate datasets and provides built-in feature importance similar to other selected ML algorithms. Its key feature is a histogram-based approach for decision tree construction, optimizing training efficiency and memory usage (Ke et al., 2017). LightGBM can effectively handle high-dimensional data and is particularly beneficial for real estate analysis. These characteristics and its built-in feature importance ability make LightGBM a valuable tool for predicting TOM in the real estate sector.

4.5 *Hyperprameter Tuning*

A data split strategy is employed in the research methodology, allocating 70% of the dataset for training and 3-fold cross-validation purposes. The remaining 30% of the data is reserved as unseen data, explicitly designated for separate testing following the hyperparameter tuning process. This approach ensures that the models undergo a rigorous evaluation of unseen data, reliably assessing their predictive performance (Kars, 2021). 3-fold

cross-validation is chosen over conventional data splitting for its ability to provide a more robust and reliable evaluation of predictive models. It helps mitigate overfitting, offers a more accurate estimation of model performance, and enhances the validity of study findings, particularly in cases of limited data (Anguita et al., 2012; Kars, 2021).

Based on Anguita et al. (2012)'s analysis and experimentation with various k values in Cross-Validation in this thesis, a spectrum of k values was explored, such as k=3, k=5, and k=10. Nevertheless, the results consistently suggested adopting a 3-fold cross validation approach, aligning with the findings reported by Anguita et al. (2012), which indicated that optimal k values typically fall within the range of 3 to 4. This choice ensures precise generalization error estimation, and it is rooted in the observed advantages of attaining accurate generalization error estimates while maintaining computational efficiency (Anguita et al., 2012).

There are numerous tuning methods, each using a distinct approach to navigating the search space. GridSearch, a time-consuming process that tests every conceivable combination of hyperparameter settings, may not be feasible for this research (D'Hoooghe, 2023). In contrast, randomized cross-validation, implemented using the `RandomizedSearchCV` from the Scikit-learn library, is a resampling technique that involves randomizing the dataset's partitions during the model evaluation process. It can provide a robust and unbiased algorithm performance assessment, and it has been selected for this study (Yennimar et al., 2023).

4.5.1 *Tuning Random Forest*

Several crucial parameters need to be considered for tuning the Random Forest. For instance, the n-estimators and the Max-depth control the number of trees and the depth of the decision tree, respectively (D'Hoooghe, 2023). It is important to experiment with rational values for each hyperparameter. Extremely high values for n-estiMators and Max-depth would likely result in an overfitting model, as the model becomes too complex (Jha et al., 2020). Tuning of Random Forest is shown in **Table** 1.

Table 1: Random Forest Hyperparameter Values and Descriptions

Hyperparameter	Values	Description
n_estimators	100, 200, 300, 400, 500	Number of trees in the forest.
max_features	'auto', 'sqrt'	Number of features for best split.
max_depth	10, 20, 30, 40, 50, None	Maximum depth of each tree.
min_samples_split	2, 5, 10	Min samples to split a node.
min_samples_leaf	1, 2, 4	Min samples at a leaf node.
bootstrap	True, False	Use of bootstrap samples in trees.

4.5.2 *Tuning XGBoost*

Since XGBoost is a tree-based model, its hyperparameters are similar to those of Random Forest. The n-estiMators and the Max-depth exist in both of them. However, XGBoost has some unique hyperparameters, one of them namely the learning rate considered here in **Table** 2 (D'Hoooghe, 2023; Zhang & Liu, 2023).

Table 2: XGBoost Hyperparameter Values and Descriptions

Hyperparameter	Values	Description
n_estimators	100, 200, 300, 400, 500	Number of boosting rounds.
max_depth	3, 4, 5, 6, 7, 8, 9	Maximum depth of trees.
learning_rate	0.01, 0.05, 0.1, 0.2, 0.3	Step size shrinkage for updates.
subsample	0.6, 0.7, 0.8, 0.9, 1.0	Subsample ratio of training instances.
colsample_bytree	0.6, 0.7, 0.8, 0.9, 1.0	Subsample ratio of columns when constructing each tree.
gamma	0, 0.1, 0.2, 0.3, 0.4	Minimum loss reduction for partition.
min_child_weight	1, 2, 3, 4	Minimum sum of instance weight (hessian) needed in a child.

4.5.3 *Tuning CatBoost*

CatBoost, like other tree-based models, has a set of hyperparameters that influence its performance and accuracy . Some parameters are similar to those in models like XGBoost and Random Forest and demostrated in **Table** 3, but there are also unique parameters. Adjusting these parameters

correctly can significantly improve the model's effectiveness (D'Hoooghe, 2023; Zhang & Liu, 2023).

Table 3: CatBoost Hyperparameter Values and Descriptions

Hyperparameter	Values	Description
iterations	100, 200, 300, 400, 500	Number of training iterations.
depth	4, 6, 8, 10	Depth of the tree.
learning_rate	0.01, 0.05, 0.1, 0.2, 0.3	Step size for weight updates.
subsample	0.6, 0.7, 0.8, 0.9, 1.0	Subsample ratio of the training data.
colsample_bylevel	0.6, 0.7, 0.8, 0.9, 1.0	Subsample ratio of features for each split.
l2_leaf_reg	1, 3, 5, 7, 9	Coefficient at the L2 regularization term of the cost function.

4.5.4 *Tuning LightGBM*

LightGBM is another robust gradient-boosting framework that uses tree-based learning algorithms. Its hyperparameters are similar to those in XGBoost and CatBoost, but each hyperparameter's influence on the model's performance can differ as shown in **Table** 4.

Table 4: LightGBM Hyperparameter Values and Descriptions

Hyperparameter	Values	Description
n_estimators	100, 200, 300, 400, 500	Number of boosting iterations.
max_depth	4, 6, 8, 10	Maximum depth of trees.
learning_rate	0.01, 0.05, 0.1, 0.2, 0.3	Shrinkage rate for weight updates.
subsample	0.6, 0.7, 0.8, 0.9, 1.0	Subsampling rate of the training data.
colsample_bytree	0.6, 0.7, 0.8, 0.9, 1.0	Ratio of subsampling of features.
reg_lambda	0, 1, 2, 3, 4	L2 regularization term on weights.

4.6 *Evaluation Metrics*

Model evaluation is performed by drawing on three evaluation metrics suited for regression tasks. The combination of these metrics is chosen to

provide a comprehensive overview of the model performance and allow for comparing outcomes.

- **R-squared (R^2):**

R^2, which measures the proportion of variance in the target variable that the independent variable can explain, $R^2 = \mathbf{1}$ indicates the predictions of a regression model are perfectly fitted (Ahn et al., 2023; Kars, 2021).

- **Root Mean Squared Error ($RMSE$)**

$RMSE$ represents the square root of the differences between the actual y and the predicted values $\hat{y}$. A higher $RMSE$ score shows a larger prediction error .(Ahn et al., 2023; Kars, 2021).

- **Mean Absolute Error (MAE)**

MAE represents a linear score that includes the absolute errors between the predicted values $\hat{y}$ and the actual values y. Low MAE scores indicate small prediction errors (Ahn et al., 2023; Kars, 2021).

4.7 *Experimental Setup*

This study begins with a comprehensive data preprocessing phase, encompassing tasks such as addressing missing values, encoding categorical variables, and eliminating redundant features. Subsequently, the primary focus shifts to evaluating the effectiveness of various ML models in predictive analytics, with a notable comparison to the OLS method, which serves as the baseline. The complete dataset is initially used to assess model performance without hyperparameter tuning. The subsequent phase involves applying feature selection methods, both with RFECV-DT and by considering each algorithm's top N feature importance.

Simultaneously, hyperparameter tuning with randomized cross-validation is conducted to investigate its impact on predictive accuracy. The algorithms under review include OLS, Random Forest, XGBoost, CatBoost, and LightGBM. These models are evaluated using the full dataset and RFECV-DT in all phases, except for the OLS algorithm, which is excluded only in the phase focusing on built-in features. The dataset is consistently split into two segments throughout the study: 70% for training and hyperparameter tuning with 3-fold cross-validation, while the remaining 30% is reserved for final evaluation. This structured approach comprehensively analyzes each model's predictive capabilities under varying conditions and feature selection techniques.

4.8 *Methodology Overview*

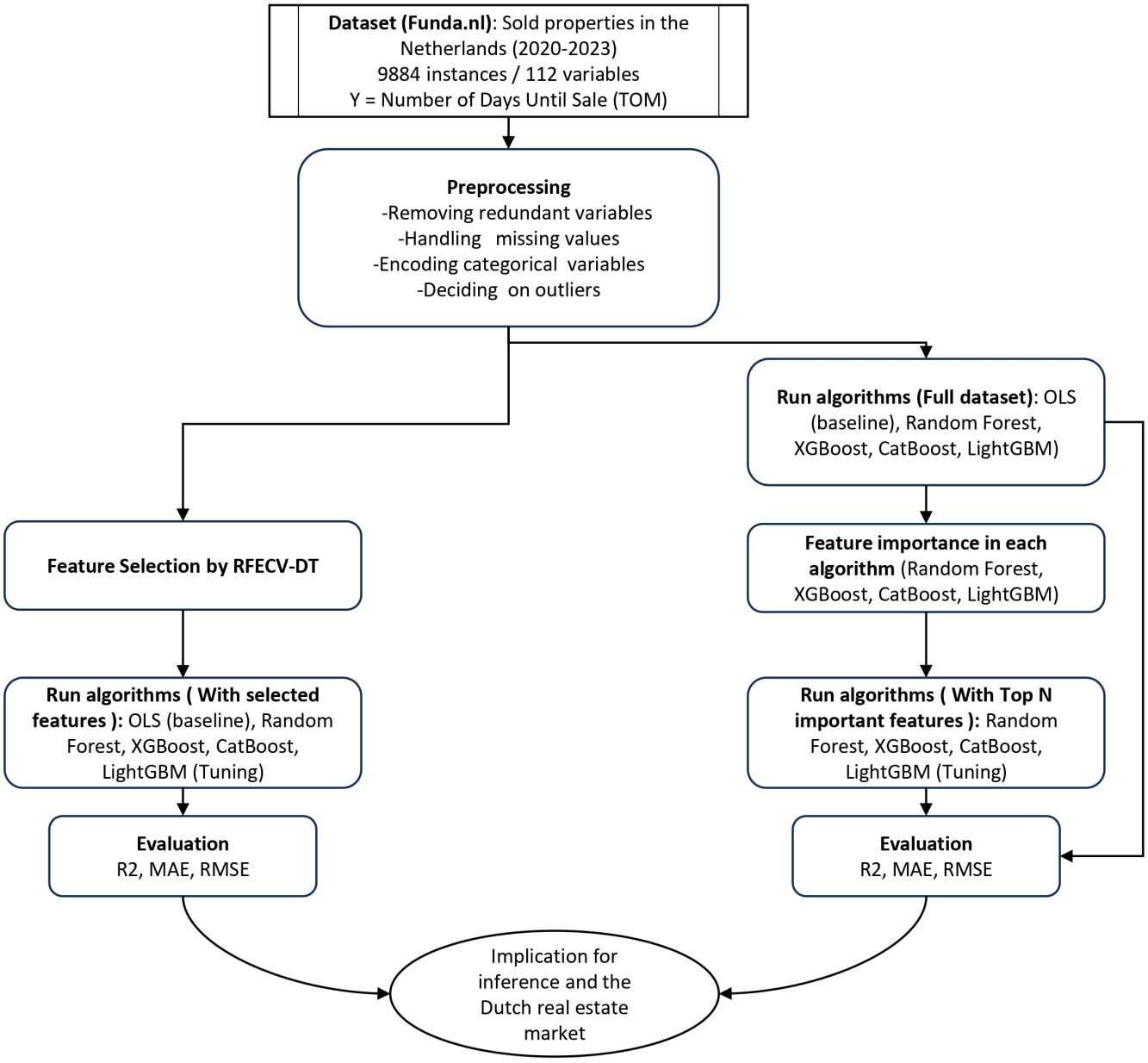

Figure 4: Systematic Methodology Overview

- The methodology overview is demonstrated in **Figure** 4

5 RESULTS

In the subsequent sections, the results are presented in a detailed and comprehensive manner. Some graphical representations and plots are included directly within this section to enhance readability and clarity. Additional plots and detailed visual data are provided in Appendix A (8) to maintain a balance between the thoroughness and readability of the main text.

5.1 *Feature Selection by RFECV-DT*

In the context of examining TOM in the real estate domain, RFECV-DT was applied and identified 18 optimal features. This data-driven feature selection method, based on RMSE, played a pivotal role in uncovering the most influential features for predicting TOM. Among the identified optimal features, three stood out with the highest ranks in their contribution to TOM prediction: `Construction_Existing`, `Roof_Type`, and `Plot_Area_in_m2` as demonstrated in **Figure** 5.

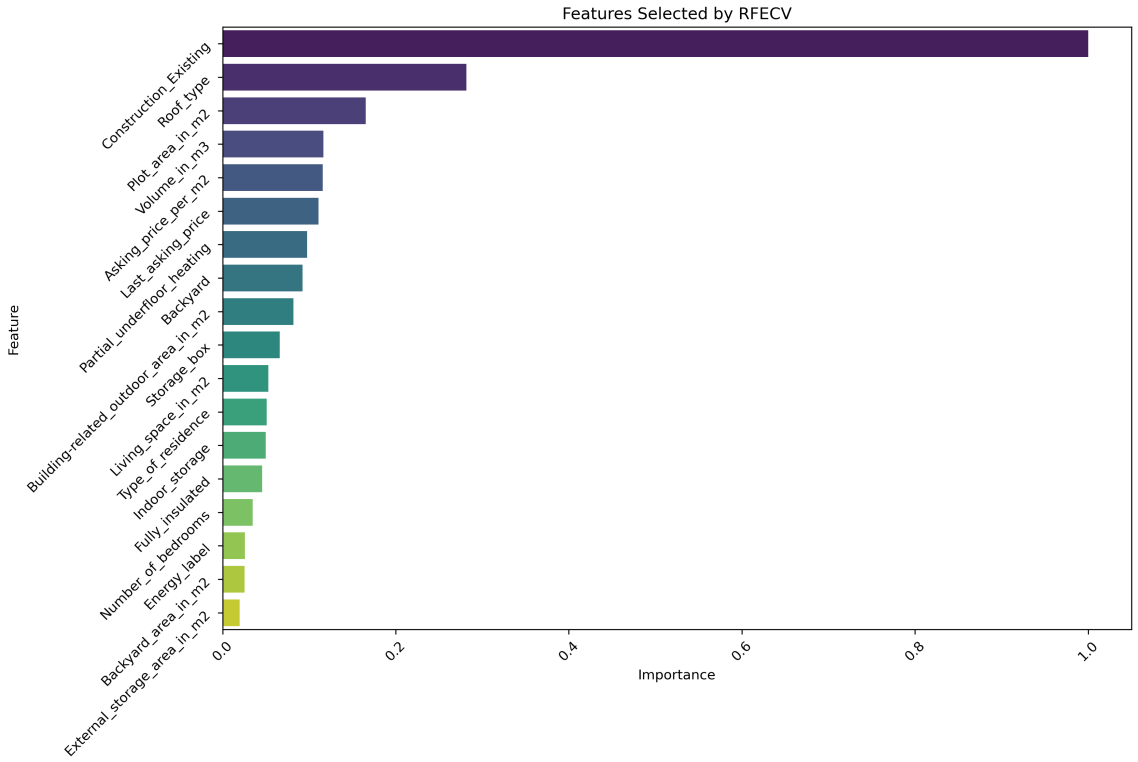

Figure 5: Optimal Features by RFECV

In the following part 5.2, these 18 optimal features will be compared with the top 18 features ranked by built-in feature importance in each algorithm. Additionally, the results of models that utilized this reduced dataset are presented in **Table** 7.

5.2 *Feature Importance by Each Algorithm*

The rationale for selecting a specific number of features for analysis in the study is grounded in empirical findings and consistency considerations. Initially, various top features were explored for each algorithm, including 18, 36, 54, and 72, based on their built-in feature importance rankings. It was observed that the top 18 features exhibited a higher resemblance to the 18 optimal features identified by RFECV-DT across all algorithms. However, each algorithm distinctly ranked these 18 features. The outcomes of employing the top 18 features for each algorithm are presented in **Table** 9.

Given this observation, and to maintain consistency in comparative analysis, it was decided to focus solely on the top 18 features from each algorithm. This approach allows for a more controlled and uniform comparison. In the feature importance analysis, it was evident that certain variables were consistently highlighted as significant in the RFECV-DT results and across the top 18 features of each algorithm. This consistency in feature selection across different methods reinforces the relevance of these variables in the predictive models, providing a solid rationale for their inclusion and analysis in the study.

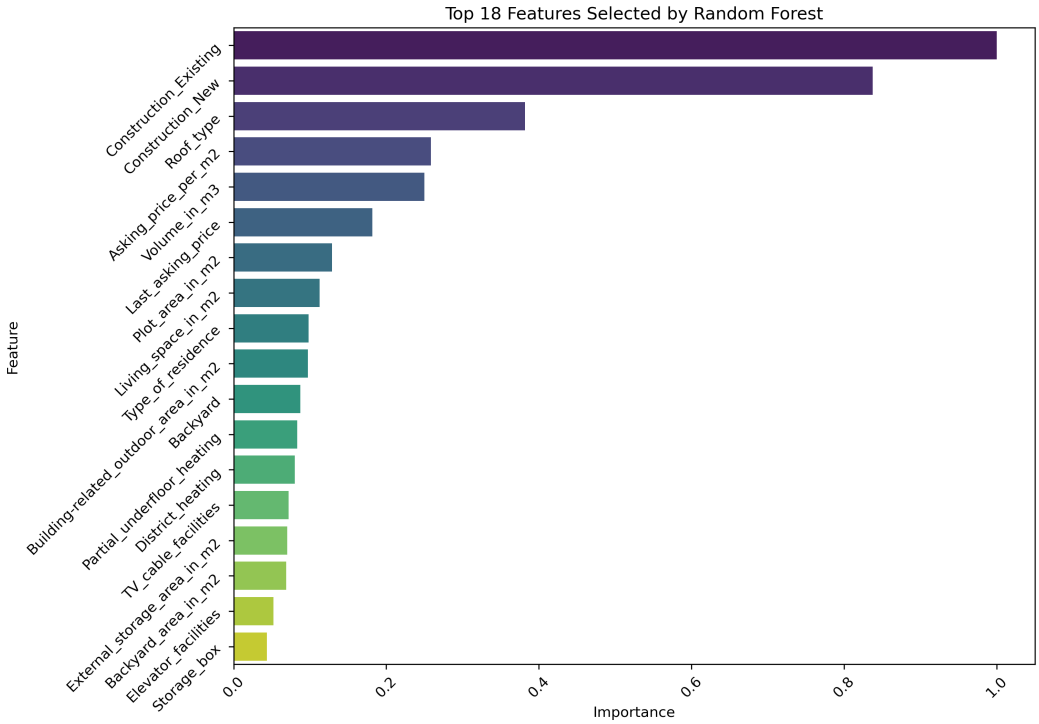

Figure 6: Feature Importance by Random Forest

The feature importance plot for Random Forest in **Figure** 6 illustrates the dominance of `Construction_Existing` and the consistent alignment with RFECV-DT. Most of the variables mutually exist in both RFECV-DT and top 18 important features recognized by Random Forest among them `Roof_type` again showcased their significance here.

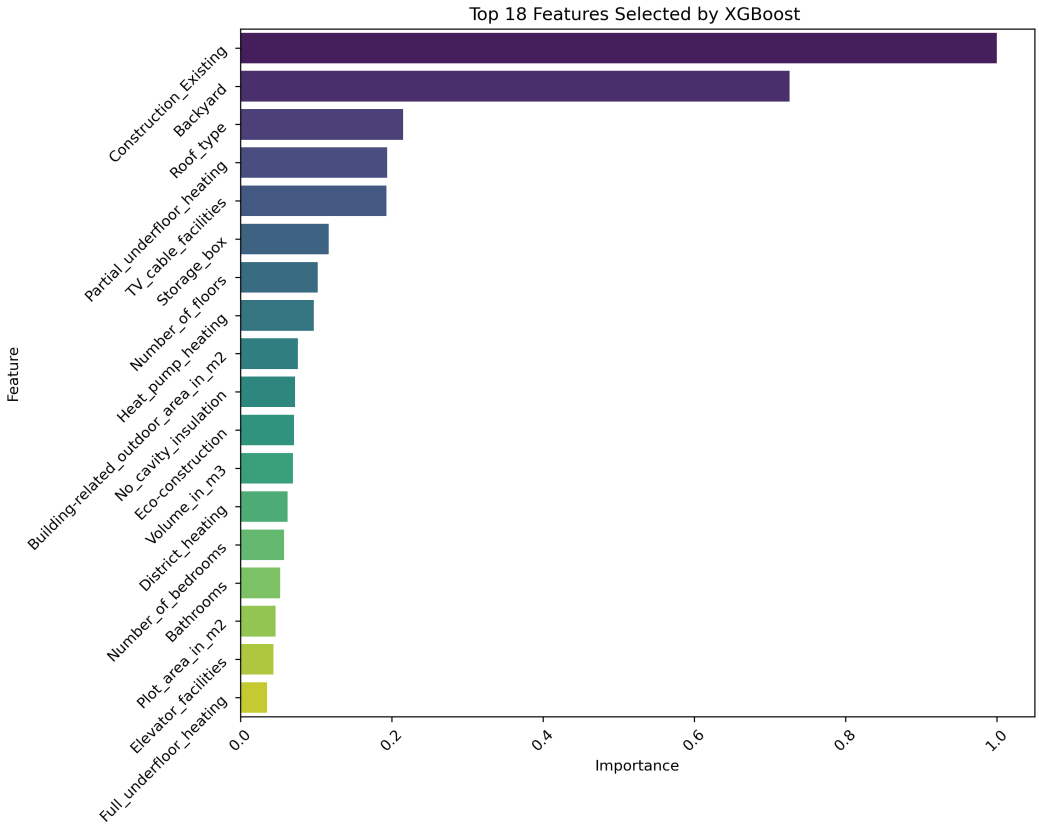

Figure 7: Feature Importance by XGBoost

Similarly, in **Figure** 7, the feature importance plot for XGBoost demonstrates the significance of `Construction_Existing`, maintaining harmony with both RFECV-DT and Random Forest again `Roof_type` is also among the highest ranked variables.

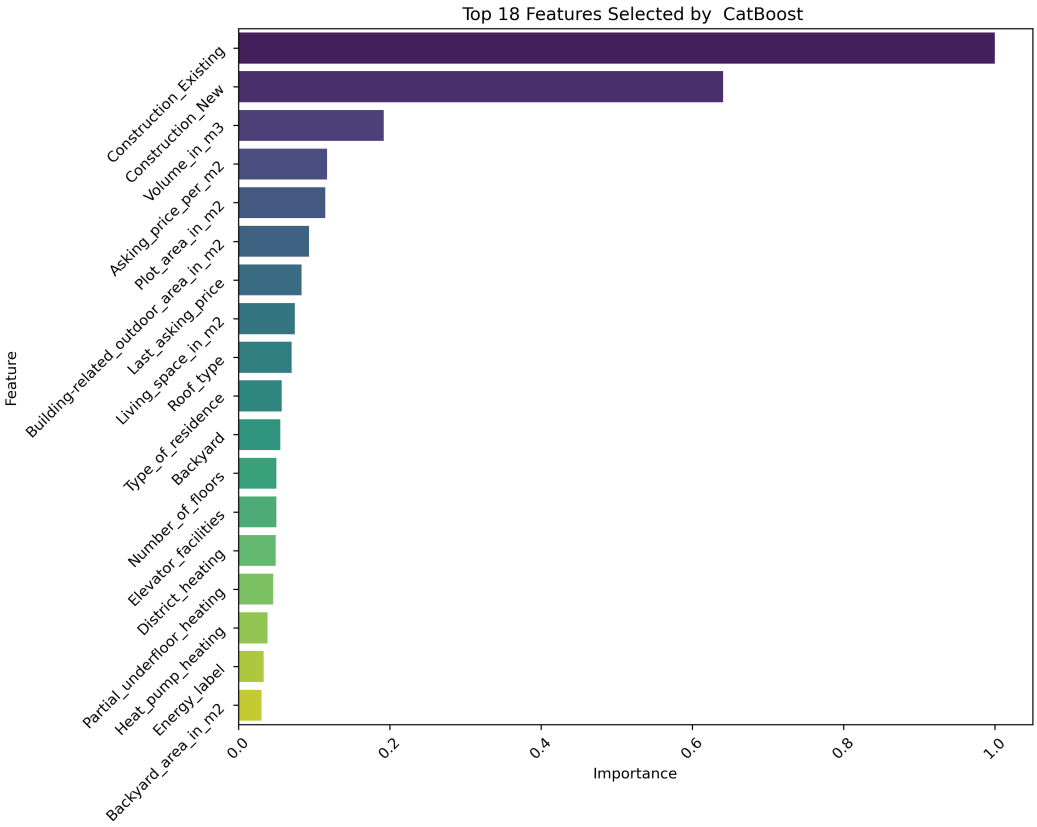

Figure 8: Feature Importance by CatBoost

In **Figure** 8, the feature importance plot for CatBoost demonstrated, again emphasizing the persistent relevance of `Construction_Existing` and other shared variables but `Roof_type` here is not among the highest ranked features as before.

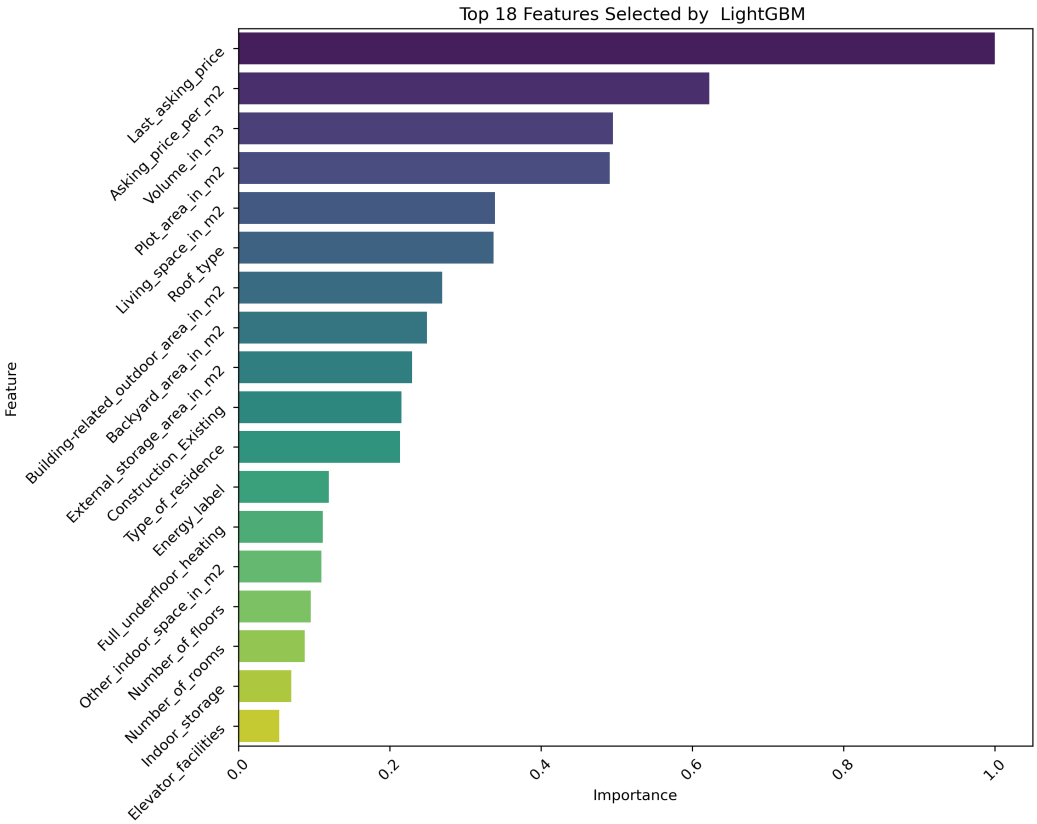

Figure 9: Feature Importance by LightGBM

Lastly, the feature importance plot for LightGBM in **Figure** 9 reveals a distinctive pattern, with `Last_asking_price` taking precedence over `Construction_Existing`. This variation highlights the unique considerations in LightGBM's feature importance ranking compared to other algorithms. Consequently, The variables `Construction_Existing`, `Roof_Type`, `Volume_in_m`^{3}, `Plot_Area_in_m`^{2}, and `BuildingRelated_Outdoor_Area_in_m`^{2} remain consistent, while other variables may vary in all of them.

5.3 *Full Dataset Results*

This section focuses on evaluating the performance of various ML models and OLS with a full dataset. The analysis specifically examines the performance metrics of these models on the test set. The models under comparison are OLS as a baseline, Random Forest, XGBoost, CatBoost, and LightGBM.

Table 5: Performance Metrics on Test Set (Full Dataset)

Model	R-squared	MAE	RMSE
OLS	0.435	14.957	32.413
Random Forest	0.609	11.670	26.966
XGBoost	0.587	12.191	27.713
CatBoost	0.632	11.356	26.176
LightGBM	0.609	11.839	26.952

In **Table** 5, the performance metrics of each model on the test set are presented. The OLS model, serving as the baseline, exhibits the least accuracy with an R^2 value of 0.435, along with the highest MAE and $RMSE$ values of 14.957 and 32.413, respectively. Among the ML models, which all performed moderately, XGBoost, despite its advanced features, shows lower performance compared to its counterparts, as indicated by its R^2 value and higher error metrics. In contrast, CatBoost outperforms the other models with the highest R^2 value of 0.632 and the lowest error metrics, recording 11.356 for MAE and 26.176 for $RMSE$. This underscores its effectiveness in accurately predicting TOM.

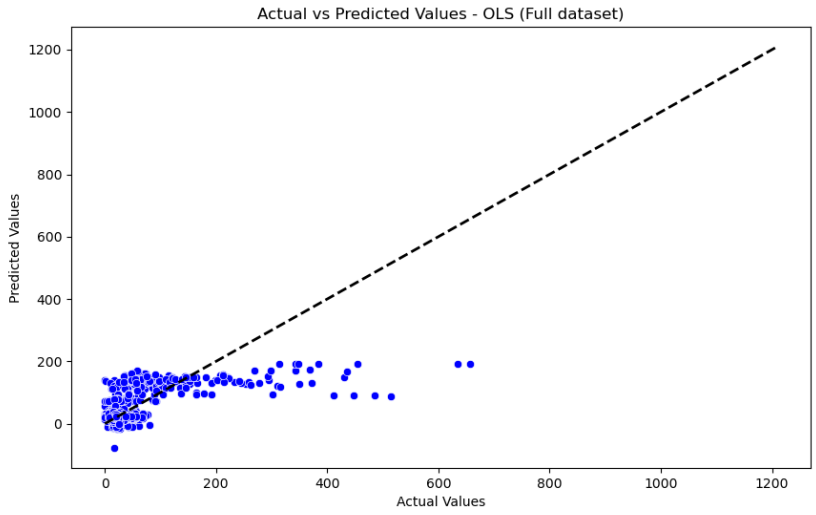

Figure 10: Actual vs Predicted - OLS (Baseline) (Full Dataset)

The plot in **Figure** 10 for the OLS model clearly shows a broad dispersion between the predicted and actual TOM values, visually supporting its lower accuracy as indicated by the test set performance metrics.

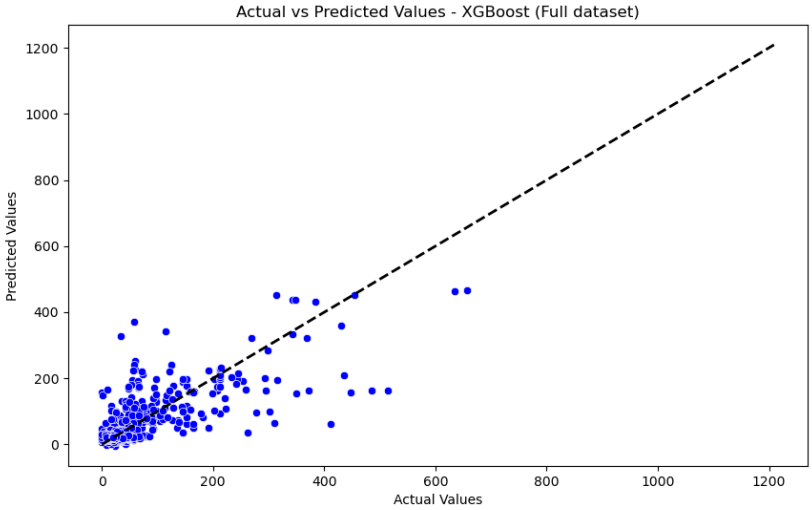

Figure 11: Actual vs Predicted - XGBoost (Full Dataset)

The XGBoost plot in **Figure** 11, while showing a comparatively tighter grouping than OLS, still reveals areas of deviation, especially in capturing extreme values, reflecting its comparatively lower performance on the test set.

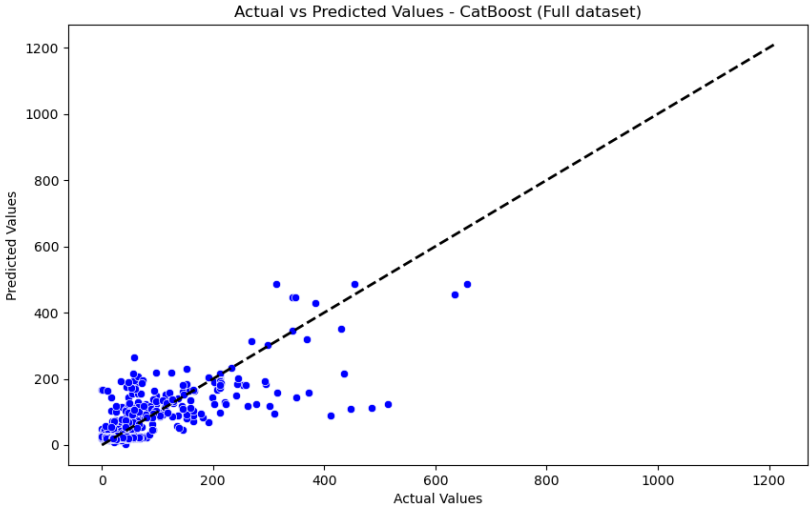

Figure 12: Actual vs Predicted - CatBoost (Full dataset)

On the contrary, the CatBoost plot in **Figure** 12 demonstrates a much closer alignment between actual and predicted values, visually highlighting

its effective performance as reflected in the test set metrics. In conclusion, this analysis emphasizes the varying degrees of effectiveness among the different models in predicting TOM on the test set. While the OLS model serves as a baseline, it is significantly outperformed by the ML models. Among them, CatBoost stands out with impressive performance, boasting an R^2 of 0.632, an MAE of 11.356, and an $RMSE$ of 26.176.

5.4 *Models with RFECV-DT Dataset*

Following the assessment of various ML models on a full dataset, this subsection shifts to evaluating their performance on the reduced dataset by using RFECV-DT which narrowed the dataset down to 18 optimal features. Alongside this feature reduction, an essential aspect of this phase was the application of hyperparameter tuning to enhance the models' performance. The best hyperparameters identified for each model are as follows in **Table** 6:

Model	**Hyperparameters**
Random Forest	{'n_estimators': 200, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 40, 'bootstrap': False}
XGBoost	{'subsample': 0.6, 'n_estimators': 300, 'min_child_weight': 3, 'max_depth': 4, 'learning_rate': 0.05, 'gamma': 0, 'colsample_bytree': 0.6}
CatBoost	{'subsample': 0.7, 'learning_rate': 0.2, 'l2_leaf_reg': 7, 'iterations': 400, 'depth': 4, 'colsample_bylevel': 0.6}
LightGBM	{'subsample': 0.6, 'reg_lambda': 3, 'n_estimators': 100, 'max_depth': 8, 'learning_rate': 0.1, 'colsample_bytree': 0.8}

Table 6: Best Hyperparameters for ML Models (RFECV-DT)

The effectiveness of this strategy is evident in the performance metrics detailed in the subsequent **Table** 7:

Table 7: Performance Metrics on Test Set (RFECV-DT)

Model	R-squared	MAE	RMSE
OLS	0.398	14.473	33.459
Random Forest	0.616	11.767	26.722
XGBoost	0.566	12.370	28.424
CatBoost	0.618	11.845	26.658
LightGBM	0.600	11.883	27.259

This table shows that all the ML models acted moderately like with the full dataset, and there is a decrease in the performance of OLS from the full dataset to the RFECV-DT reduced dataset, with a reduction in R^2 value from 0.435 to 0.398. CatBoost , despite a slight reduction R^2, maintains robust performance with MAE of 11.845 and $RMSE$ of 26.658 . Random Forest and LightGBM also exhibit minor declines in their metrics, while XGBoost shows a more noticeable decrease, suggesting a greater sensitivity to the change in the dataset from R^2 0.587 to 0.566.

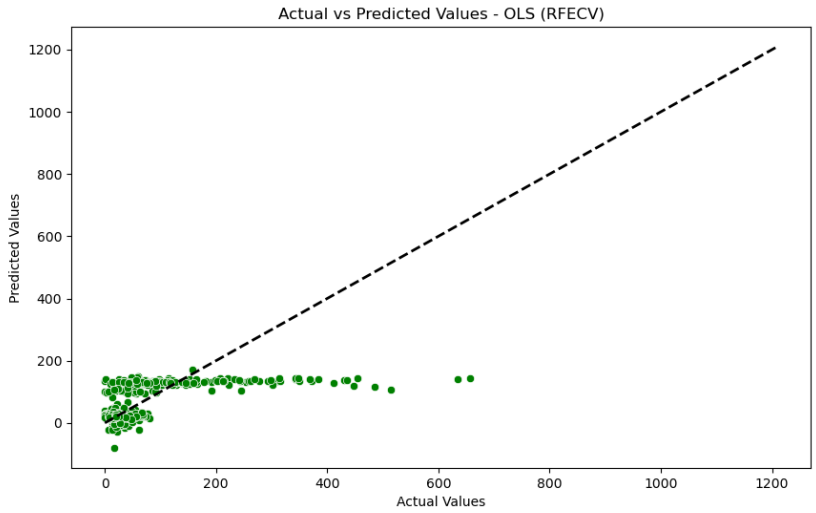

Figure 13: Actual vs Predicted - OLS (RFECV-DT)

The OLS model displays a wider spread between predicted and actual values, indicative of its diminished fit on the optimized dataset as shown in **Figure** 13.

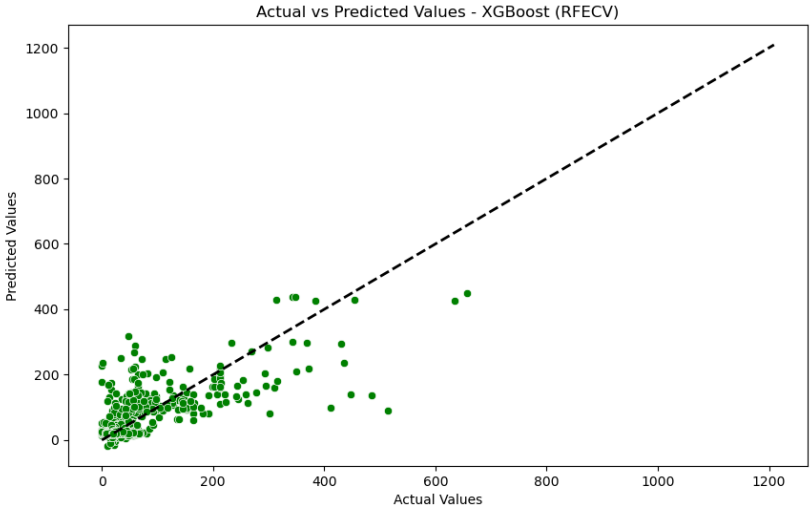

Figure 14: Actual vs Predicted - XGBoost (RFECV-DT)

The plot for XGBoost in **Figure** 14, even with optimized hyperparameters, shows a deviation from actual values, highlighting the model's reliance on a broader feature set for optimal performance.

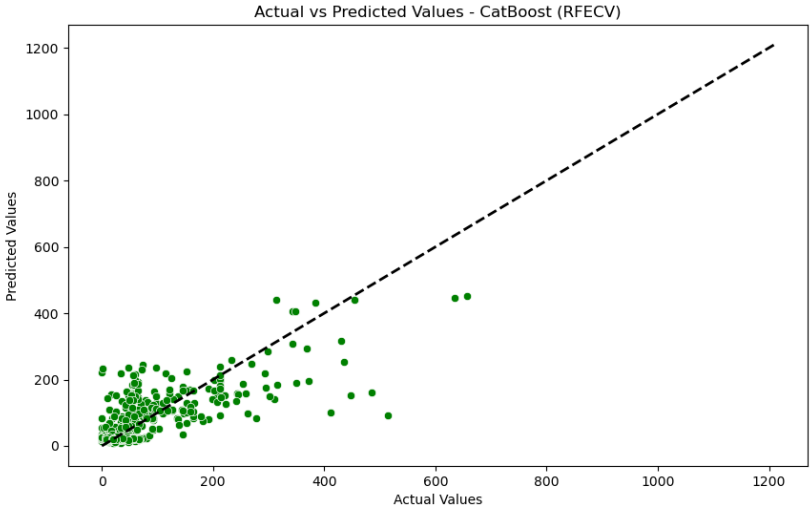

Figure 15: Actual vs Predicted - CatBoost (RFECV-DT)

In contrast, CatBoost's plot in **Figure** 15 demonstrates a more strong alignment between predicted and actual values, suggesting that it outper-

formed other algorithms both with full dataset and reduced dataset with RFECV-DT.

5.5 *Models with Built-In Top N Feature Importance*

This section delves into the analysis of ML models using their respective top 18 features identified through built-in feature importance, complemented by hyperparameter tuning which best hyperparameters are shown in **Table** 8. This evaluation is juxtaposed with previous stages, namely the full dataset analysis and the RFECV-DT dataset, to discern the effectiveness of each model's intrinsic feature selection capabilities.

Model	**Hyperparameters**
Random Forest	{'n_estimators': 100, 'min_samples_split': 2, 'min_samples_leaf': 2, 'max_features': 'sqrt', 'max_depth': 40, 'bootstrap': True}
XGBoost	{'subsample': 0.9, 'n_estimators': 300, 'min_child_weight': 3, 'max_depth': 3, 'learning_rate': 0.05, 'gamma': 0.2, 'colsample_bytree': 0.9}
CatBoost	{'subsample': 0.6, 'learning_rate': 0.2, 'l2_leaf_reg': 1, 'iterations': 100, 'depth': 8, 'colsample_bylevel': 0.8}
LightGBM	{'subsample': 0.6, 'reg_lambda': 3, 'n_estimators': 100, 'max_depth': 8, 'learning_rate': 0.1, 'colsample_bytree': 0.8}

Table 8: Best Hyperparameters for ML Models (Top 18 Features).

Firstly, the performance metrics of each model using the top 18 important features based on their built-in rankings are demonstrated in **Table** 9:

Table 9: Performance Metrics on Test Set (Top 18 Features)

Model	R-squared	MAE	RMSE
Random Forest	0.618	11.796	26.648
XGBoost	0.547	12.367	29.034
CatBoost	0.633	11.380	26.139
LightGBM	0.610	11.808	26.948

The results, as presented in the table, indicate varied effectiveness in feature selection across the models. The Random Forest model exhibits a

more strong performance with an R^2 value of 0.618, a MAE of 11.796, and a $RMSE$ of 26.648. Conversely, CatBoost stands out as the most effective, achieving an R^2 of 0.633, an MAE of 11.380, and the lowest $RMSE$ of 26.139. This suggests that CatBoost's built-in feature selection methods might be more proficient, particularly in this study also again all the models acted averagely.

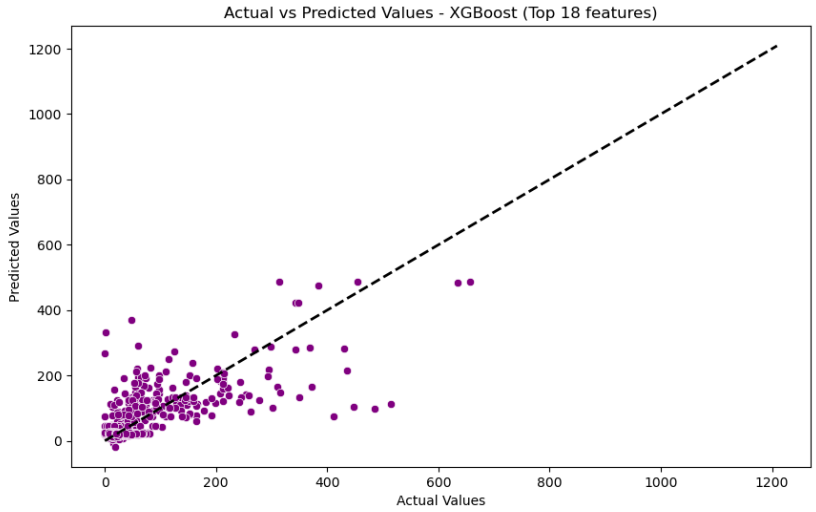

Figure 16: Actual vs Predicted - XGBoost (Top 18 features)

This plot in **Figure** 16 illustrates that XGBoost, despite its hyperparameter tuning, shows some discrepancies between the predicted and actual values, possibly due to its internal feature selection not aligning optimally with the dataset's predictive requirements.

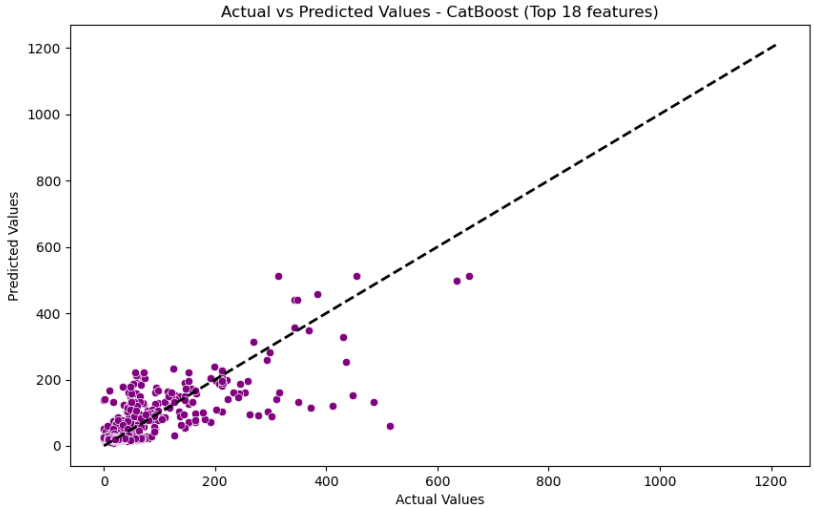

Figure 17: Actual vs Predicted - CatBoost (Top 18 features)

In contrast to XGBoost, CatBoost's plot in **Figure** 17 shows a much closer alignment between the predicted and actual values, indicating a more effective use of its selected features.

6 DISCUSSION

In this comprehensive discussion, the focus shifts to the profound insights gained from research on applying advanced ML models for predicting the TOM of residential properties in the Netherlands. The discussion thoroughly examines the findings, methodological insights, implications, limitations, and future research directions.

6.1 *Comparative Analysis of ML Models*

This study conducted an in-depth comparative analysis of four prominent ML models: Random Forest, XGBoost, CatBoost, and LightGBM, in contrast to the linear OLS method which is one of the main methods in TOM's prior studies. The results demonstrated the superiority of these ML models over OLS, with CatBoost emerging as the best performer in TOM prediction as demonstrated in **Table** 10. CatBoost's performance can be attributed to its ability to handle and manage complex data relationships. Notably, in both full and reduced datasets, including RFECV-DT and Top N built-in feature importances, CatBoost consistently outperformed all other models.

Conversely, XGBoost exhibited the least effective performance among the ML algorithms assessed.

Table 10: Performance Metrics for OLS and CatBoost

Phase	Model	R^2	MAE	RMSE
Full dataset	OLS	0.435	14.957	32.413
	CatBoost	0.632	11.356	26.176
RFECV-DT	OLS	0.398	14.473	33.459
	CatBoost	0.618	11.845	27.259

It is worth noting that in other real estate studies, XGBoost has shown strong performance in price prediction such as the study by Guliker et al. (2022). Nevertheless, the substantial enhancement in predictive accuracy highlights the transformative potential of ML in the realm of real estate analytics. It's important to note that real estate, particularly in the context of TOM, often involves extreme outliers. Despite this, ML algorithms demonstrate moderate performance, yet they remain valuable for achieving more accurate predictions compared to linear models.

6.2 *Methodological Insights*

An essential part of this research involved extensively utilizing RFECV-DT with each algorithm's native feature importance tools. This approach efficiently distilled the dataset to 18 optimal features, yielding valuable insights into the importance of individual features for TOM prediction. Remarkably, the results from both feature selection methods applied to the reduced datasets were closely aligned, highlighting their consistent performance. Although the rankings of these features varied somewhat across different models, the persistent overall improvement underscored the robustness and adaptability of these advanced algorithms. It is worth noting that all the algorithms, as demonstrated in **Section** 5, exhibited slightly better performance when applied to the entire dataset, but the feature reduction approach closely approximated the results of the entire dataset, offering computational efficiency advantages.

6.3 *Implications for Real Estate Analytics*

This research goes beyond predictive modeling, signifying a data-driven revolution in real estate. Embracing ML, especially models like CatBoost empowers professionals with precise TOM predictions. It can be helpful in

informed decision-making, optimizing pricing, and gaining a competitive edge. Accurate predictions enhance customer satisfaction, reduce holding costs, and maximize marketing ROI(Return on Investment). They also offer valuable market insights, mitigate investment risks, save time and resources, and foster trust. This research showcases the potential of data-driven approaches, positioning the real estate industry at the forefront of market analysis and strategic decision-making.

6.4 *Challenges and Future Research Directions*

Several challenges and limitations were encountered in this study. Notably, while modest in size, the dataset prompted consideration of the potential benefits of using a larger dataset for future research. Another significant challenge was the absence of previous studies using ML to predict TOM, which limited comparability while underscoring the pioneering nature of this work.

For future research, it would be valuable to explore several directions:

- **Broadening the Dataset:** To gain deeper insights and potentially enhance predictive accuracy, consider expanding the dataset to encompass a wider variety of real estate properties and variables.
- **Comparative Analysis:** Explore the possibility of conducting a comparative analysis involving alternative ML algorithms or delving into ensemble models to gain additional perspectives on TOM prediction.
- **Feature Engineering and Preprocessing:** Investigate the effects of various feature engineering techniques and data preprocessing methods on model performance, thereby contributing to the refinement of predictive models in the real estate domain.
- **Generalizability Assessment:** Assess the models' applicability across diverse geographic regions and property types to improve the practicality and relevance of research findings.
- **Continuous Adaptation:** In light of the evolving nature of the real estate market, consider the value of ongoing monitoring and adaptation to ensure the continued effectiveness of predictive models amid changing market conditions.

In conclusion, this research emphasizes the transformative potential of advanced ML models, with CatBoost emerging as a leading contender in reshaping real estate TOM prediction. This comprehensive analysis can pave the way for data-driven decision-making in the real estate sector, with implications that reach beyond the Dutch market.

7 CONCLUSION

This study embarked on a comprehensive analysis to understand the effectiveness of selected ML algorithms compared to a traditional statistical method namely OLS and with each other in TOM for residential properties in the Dutch real estate market. The primary research question and two sub-questions provided a structured approach to exploring this domain.

M-RQ *How do the selected ML algorithms and OLS compare in predicting TOM for residential properties in the Dutch real estate market, and what insights can be gained from this comparative analysis?*

A comparative analysis between ML models, namely Random Forest, XGBoost, CatBoost, LightGBM, and the conventional OLS method, showed significant findings. Therefore, the ML models performed moderately but consistently outperformed the OLS model in predicting TOM. This superiority became evident in the model's capacity to effectively handle complex and non-linear data while maintaining higher accuracy, even in the presence of outliers. In particular, CatBoost emerged as the most effective algorithm and showed the highest R^2 values and the lowest error measures in different test scenarios. This finding underscores the potential of advanced ML algorithms to enhance predictive analytics in the real estate sector, particularly in applications such as TOM prediction, where traditional methods may fall short.

S-RQ1 *Given the application of RFECV to finding optimal features, what are the critical predictors identified as most influential for estimating TOM, and how does RFECV enhance the predictive accuracy of the ML models?*

Utilizing RFECV-DT on a dataset with 112 variables, the study identified 18 optimal predictors for estimating TOM. Notable among these were features like `Construction_Existing`, `Roof_Type`, and `Plot_Area_in_m2`. The application of RFECV not only streamlined the feature set but also helped to keep the predictive accuracy of the ML models just by a few variables.

S-RQ2 *What pivotal features do each of the selected ML algorithms identify as influential in predicting TOM, considering their inherent ranking mechanisms, and how do these rankings compare to each other and RFECV?*

Each selected ML algorithm identified the top 18 features pivotal in predicting TOM, reflecting their inherent ranking mechanisms. Interestingly, while there was some overlap in the features identified as necessary across different algorithms, each algorithm also brought to light unique predictors.

This variance highlights the distinct analytical approaches inherent to each algorithm. When comparing these findings to the RFECV-DT results, it was observed that although there was considerable alignment in the influential features, the ML algorithms also valued certain features differently, providing a broader perspective on the factors influencing TOM.

Implications for Real Estate Industry:

- The outcomes of this research offer valuable insights for the real estate industry. By highlighting the efficacy of ML models in predicting TOM, the study opens avenues for more data-driven, accurate, and strategic decision-making in real estate operations. Real estate professionals, including sellers, investors, and market analysts, can leverage these insights for more precise market analysis, pricing strategies, and understanding market dynamics. The novel application of machine learning algorithms for TOM prediction, a relatively unexplored area in real estate analytics, marks a significant step forward in the field, offering both scientific and societal benefits.

In summary, this research contributes significantly to the field of real estate analytics by introducing and validating the use of advanced ML algorithms for TOM prediction. The findings not only reinforce the superiority of ML over traditional statistical methods in certain aspects of real estate analytics but also provide a comprehensive understanding of the influential factors affecting TOM in the Dutch real estate market.

REFERENCES

Abut, F., Arlı, H. Ş., Akay, M. F., & Adıgüzel, Y. (2023). A new hybrid approach for real estate price prediction using outlier detection, feature selection, and clustering techniques. *2023 8th International Conference on Computer Science and Engineering (UBMK)*, 1–6.

Ahn, J. M., Kim, J., & Kim, K. (2023). Ensemble machine learning of gradient boosting (xgboost, lightgbm, catboost) and attention-based cnn-lstm for harmful algal blooms forecasting. *Toxins*, *15*(10), 608.

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., Ridella, S., et al. (2012). The'k'in k-fold cross validation. *ESANN*, 441–446.

Askarisichani, O., Bullo, F., Friedkin, N. E., & Singh, A. K. (2022). Predictive models for human–ai nexus in group decision making. *Annals of the New York Academy of Sciences*, *1514*(1), 70–81.

Awad, M., & Fraihat, S. (2023). Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems. *Journal of Sensor and Actuator Networks*, *12*(5), 67.

Aydin, E., Correa, S. B., & Brounen, D. (2019). Energy performance certification and time on the market. *Journal of Environmental Economics and Management*, *98*, 102270.

Benefield, J., Cain, C., & Johnson, K. (2014). A review of literature utilizing simultaneous modeling techniques for property price and time-on-market. *Journal of Real Estate Literature*, *22*(2), 149–175.

Biggs, M., Hariss, R., & Perakis, G. (2023). Constrained optimization of objective functions determined from random forests. *Production and Operations Management*, *32*(2), 397–415.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5–32.

BuHamdan, S., Minayhashemi, S., Alwisy, A., & Bouferguene, A. (2022). The influence of design-related features on houses time-on-market: A statistical analysis. *International Journal of Housing Markets and Analysis*, *15*(5), 953–976.

Cajias, M., Fuerst, F., & Bienert, S. (2019). Tearing down the information barrier: The price impacts of energy efficiency ratings for buildings in the german rental market. *Energy Research & Social Science*, *47*, 177–191.

Čeh, M., Kilibarda, M., Lisec, A., & Bajat, B. (2018). Estimating the performance of random forest versus multiple regression for predicting prices of the apartments. *ISPRS international journal of geo-information*, *7*(5), 168.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Chen, T., & Guestrin, C. (2023). Xgboost: Scalable, portable and distributed gradient boosting library.

Cheng, P., Lin, Z., & Liu, Y. (2008). A model of time-on-market and real estate price under sequential search with recall. *Real Estate Economics*, *36*(4), 813–843.

D'Hoooghe, I. (2023, May). *Predicting (missing) energy labels: A comparative study on random forest, xgboost, and tabnet* [Master's thesis, Tilburg University] [Master's thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Data Science & Society].

Enders, C. K. (2003). Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychological methods*, *8*(3), 322.

Github [Accessed on [Dec-2023]]. (2023).

Grammarly. (2024). *Grammarly* [Accessed [Jan-2024]]. https://www.grammarly.com/

Guliker, E., Folmer, E., & van Sinderen, M. (2022). Spatial determinants of real estate appraisals in the netherlands: A machine learning approach. *ISPRS international journal of geo-information*, *11*(2), 125.

Hong, J., Choi, H., & Kim, W.-s. (2020). A house price valuation based on the random forest approach: The mass appraisal of residential property in south korea. *International Journal of Strategic Property Management*, *24*(3), 140–152.

Jha, S. B., Babiceanu, R. F., Pandey, V., & Jha, R. K. (2020). Housing market prediction problem using different machine learning algorithms: A case study. *arXiv preprint arXiv:2006.10092*.

Jiang, Z., & Shekhar, S. (2017). Spatial big data science. *Schweiz: Springer International Publishing AG*.

Kaggle.com. (2023, September). https://www.kaggle.com

Kars, J. C. (2021). Predicting neighborhood prices: Machine learning and hedonic pricing in the dutch housing market. *Yayımlanmamış Yüksek Lisans Tezi), Tilburg University*.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, *30*.

Lai, V., Cai, J. Z., & Tan, C. (2019). Many faces of feature importance: Comparing built-in and post-hoc feature importance in text classification. *arXiv preprint arXiv:1910.08534*.

Lippman, S. A., & McCall, J. J. (1986). An operational measure of liquidity. *The American Economic Review*, *76*(1), 43–55.

Luna Andonegui, C. (2023). Tom: Measuring liquidity in the swedish real estate market: Investigation of the influence of the market conditions on the time-on-market and price relationship.

Ma, Y., & Zhang, Z. (2020). Travel mode choice prediction using deep neural networks with entity embeddings. *IEEE Access*, *8*, 64959–64970.

Microsoft. (2023). LightGBM: A fast, distributed, high performance gradient boosting (GBDT, GBRT, GBM or MART) framework.

Mirzaei, A., Carter, S. R., Patanwala, A. E., & Schneider, C. R. (2022). Missing data in surveys: Key concepts, approaches, and applications. *Research in Social and Administrative Pharmacy*, *18*(2), 2308–2316.

Mostert, N. E. (2022). *Predicting house listing prices within the netherlands using environmental features* [Master's Thesis]. Tilburg University [Cognitive Science and Artificial Intelligence].

Naotunna, R. (2023). *A model for the estimation of land prices in colombo district using web scraped data* [Doctoral dissertation].

Neloy, A. A., Haque, H. S., & Ul Islam, M. M. (2019). Ensemble learning based rental apartment price prediction model by categorical fea-

tures factoring. *Proceedings of the 2019 11th International conference on machine learning and computing*, 350–356.

OpenAI. (2023). Chatgpt-3.5 [Accessed: [2023-2024]].

Otchere, D. A., Ganat, T. O. A., Ojero, J. O., Tackie-Otoo, B. N., & Taki, M. Y. (2022). Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering*, *208*, 109244.

Potrawa, T., & Tetereva, A. (2022). How much is the view from the window worth? machine learning-driven hedonic pricing model of the real estate market. *Journal of Business Research*, *144*, 50–65.

Python Software Foundation. (2022). *Python* (Version 3.9.13). https://www.python.org/

QuillBot. (2023). QuillBot: A paraphrasing tool.

Reber, B. (2017). Does mispricing, liquidity or third-party certification contribute to ipo downside risk? *International Review of Financial Analysis*, *51*, 25–53.

Rico-Juan, J. R., & de La Paz, P. T. (2021). Machine learning with explainability or spatial hedonics tools? an analysis of the asking prices in the housing market in alicante, spain. *Expert Systems with Applications*, *171*, 114590.

Senthilkumar, V. (2023). Enhancing house rental price prediction models for the swedish market: Exploring external features, prediction intervals and uncertainty management in predicting house rental prices.

Taylor, C. R. (1999). Time-on-the-market as a sign of quality. *The Review of Economic Studies*, *66*(3), 555–578.

Tucker, C., Zhang, J., & Zhu, T. (2013). Days on market and home sales. *The RAND Journal of Economics*, *44*(2), 337–360.

Tukey, J. W., et al. (1977). *Exploratory data analysis* (Vol. 2). Reading, MA.

Uyar, A., Bener, A., Ciray, H. N., & Bahceci, M. (2009). A frequency based encoding technique for transformation of categorical variables in mixed ivf dataset. *2009 annual international conference of the Ieee engineering in medicine and biology society*, 6214–6217.

Yandex. (2023). CatBoost: A high-performance open-source library for gradient boosting on decision trees.

Yang, T. (2023). Sales prediction of walmart sales based on ols, random forest, and xgboost models. *Highlights in Science, Engineering and Technology*, *49*, 244–249.

Yennimar, Y., Rasid, A., Kenedy, S., et al. (2023). Implementation of support vector machine algorithm with hyper-tuning randomized search in

stroke prediction. *Jurnal Sistem Informasi dan Ilmu Komputer Prima (JUSIKOM PRIMA)*, *6*(2), 61–65.

Yu, W., Ma, Z., Pant, G., & Hu, J. (2021). The effect of virtual tours on house price and time on market. *Journal of Real Estate Literature*, *28*(2), 133–149.

Zeng, G. (2023). On the analytical properties of category encodings in logistic regression. *Communications in Statistics-Theory and Methods*, *52*(6), 1870–1887.

Zhang, X., & Liu, C.-A. (2023). Model averaging prediction by k-fold cross-validation. *Journal of Econometrics*, *235*(1), 280–301.

Zhu, H., Xiong, H., Tang, F., Liu, Q., Ge, Y., Chen, E., & Fu, Y. (2016). Days on market: Measuring liquidity in real estate markets. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 393–402.

8 APPENDIX A

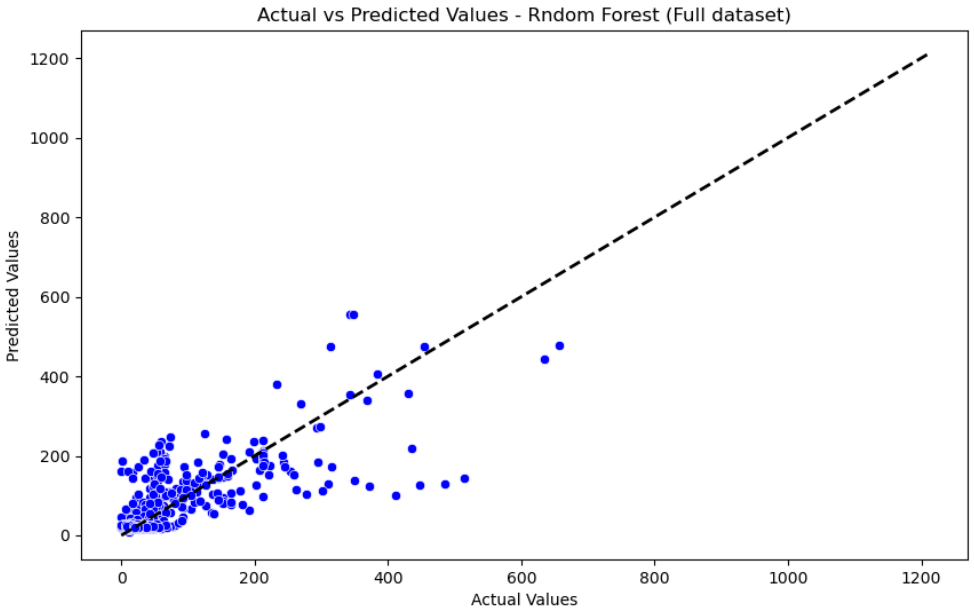

Figure 18: Actual vs Predicted - Random Forest (Full dataset)

Figure 18 - Actual vs Predicted - Random Forest (Full dataset): This plot displays the comparison between actual and predicted values using the

Random Forest model on the full dataset, highlighting the model's overall predictive accuracy and potential areas for improvement.

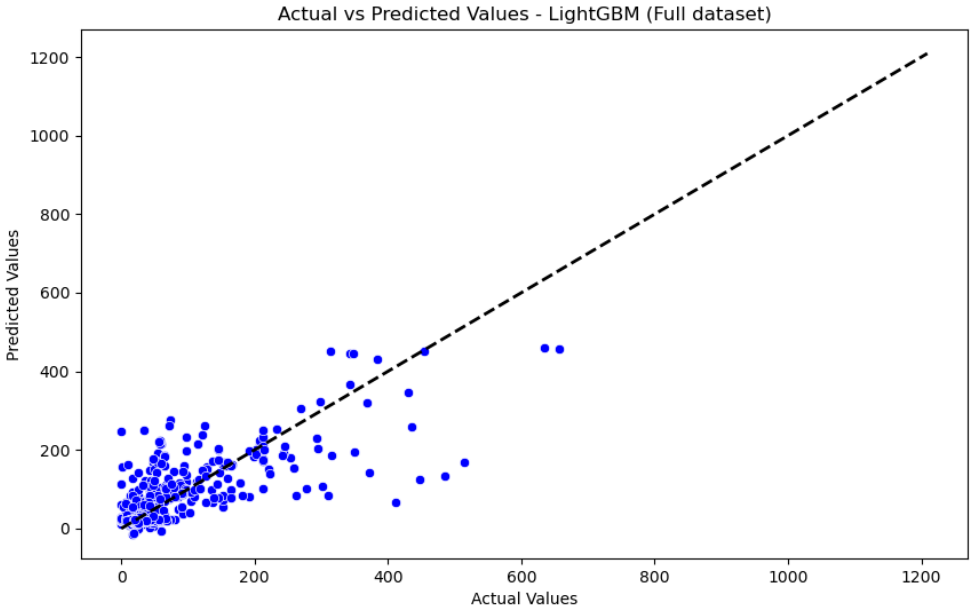

Figure 19: Actual vs Predicted - LightGBM (Full dataset)

Figure 19 - Actual vs Predicted - LightGBM (Full dataset): This figure illustrates the performance of the LightGBM model on the full dataset, revealing how closely the predicted values align with the actual ones, thus reflecting the model's effectiveness in this context.

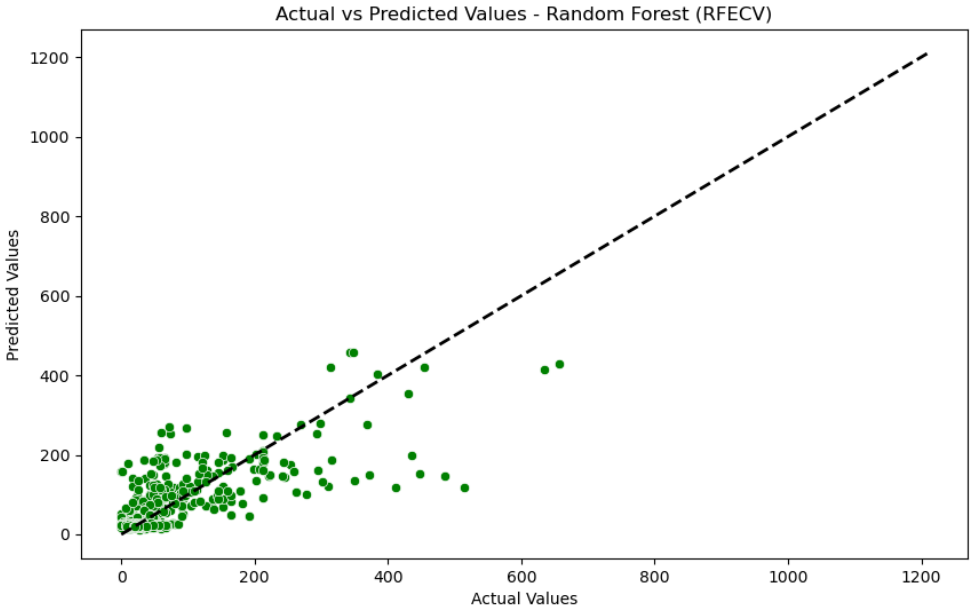

Figure 20: Actual vs Predicted - Random Forest (RFECV-DT)

Figure 20 - Actual vs Predicted - Random Forest (RFECV-DT): Here, the graph shows the efficacy of the Random Forest model using RFECV-DT feature selection, demonstrating how the model performs with a more refined set of features.

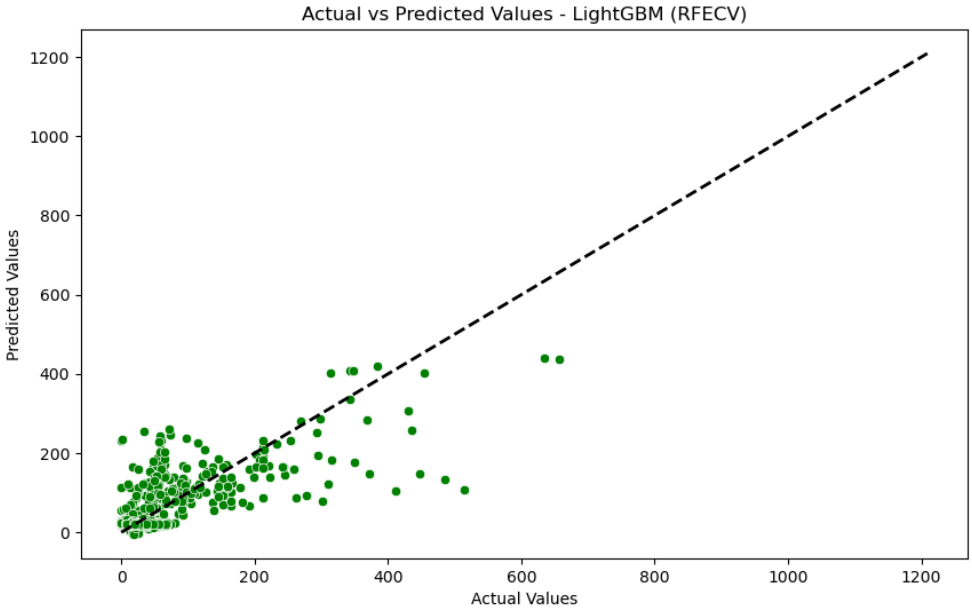

Figure 21: Actual vs Predicted - LightGBM (RFECV-DT)

Figure 21 - Actual vs Predicted - LightGBM (RFECV-DT): This plot represents the LightGBM model's predictions compared to actual values using RFECV-DT for feature selection, offering insights into the model's precision with optimized features.

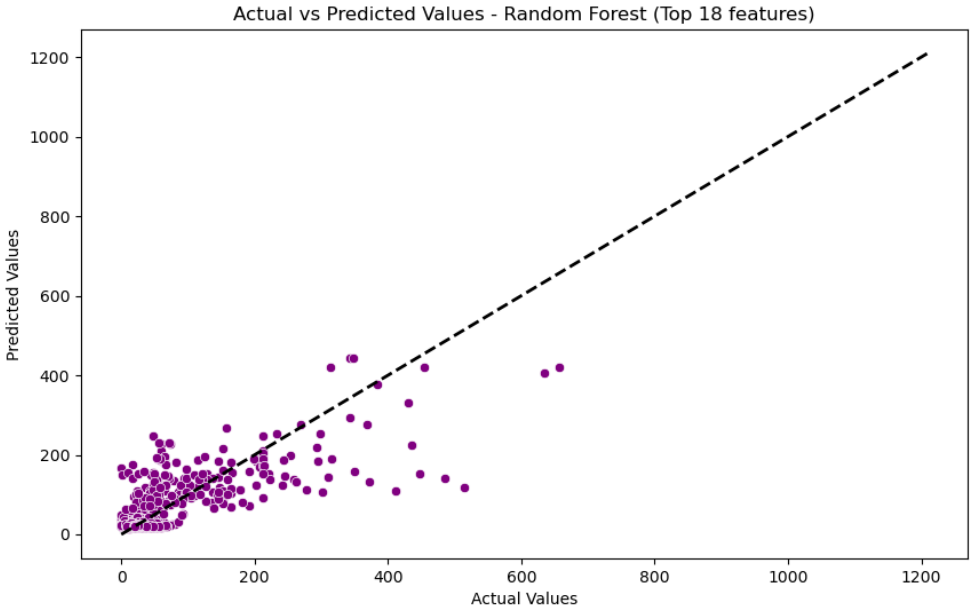

Figure 22: Actual vs Predicted - Random Forest (Top 18 features)

Figure 22 - Actual vs Predicted - Random Forest (Top 18 features): This visualization compares the actual and predicted values by the Random Forest model, specifically using the top 18 features, showcasing the impact of focused feature selection on model performance.

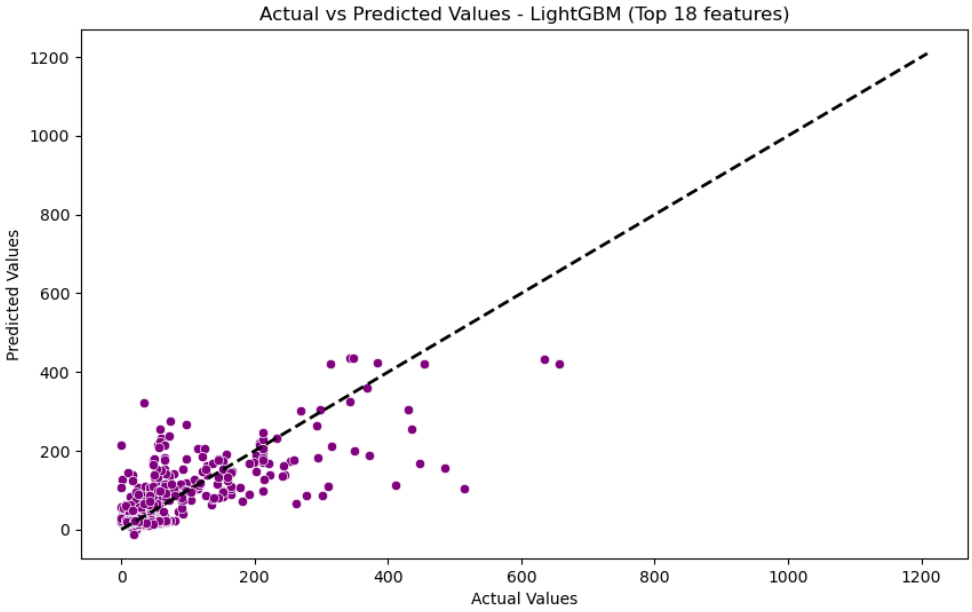

Figure 23: Actual vs Predicted - LightGBM (Top 18 features)

Figure 23 - Actual vs Predicted - LightGBM (Top 18 features): The figure depicts the predictive accuracy of the LightGBM model using the top 18 features, highlighting how targeted feature selection influences the model's predictions.