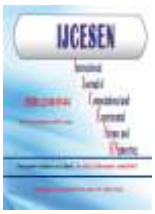
International Journal of Computational and Experimental Science and ENgineering (IJCESEN)

Vol. 11-No.2 (2025) pp. 3743-3758
http://www.ijcesen.com



ISSN: 2149-9144

Research Article

Evaluating Machine Learning Models for House Price Prediction with Different Sampling Techniques

Ali İhsan Çetin*

Ankara Yildirim Beyazit University, Business School, Department of Finance and Banking, Ankara-Turkiye
* **Corresponding Author Email:** aliihsancetin@aybu.edu.tr - **ORCID:** 0000-0001-6903-8240

Article Info:

DOI: 10.22399/ijcesen.2870
Received : 29 March 2025
Accepted : 01 June 2025

Keywords :

Machine Learning,
Ensemble Methods,
House Price Prediction,
Data Sampling Techniques

Abstract:

This study investigates the interplay between advanced sampling techniques and machine learning models to predict residential property sale prices using a diverse dataset encompassing structural, locational, and economic attributes. Emphasizing Stratified Extreme Ranked Set Sampling (SERSS), the research systematically evaluates the impact of five sampling methods—SERSS, Cluster, Bootstrap, Systematic, and Random Sampling—on various machine learning algorithms, including CatBoost, Random Forest, ElasticNet, and FIkNN. The findings reveal that SERSS significantly enhances the generalizability and robustness of predictive models by capturing both central and extreme data tendencies, outperforming traditional methods in preserving dataset variability. Ensemble methods like CatBoost, Random Forest and similarity algorithm like FIkNN consistently demonstrated superior predictive accuracy, achieving the Mean Absolute Error (MAE) between $85 and $650, and high R^2 values across structured sampling techniques. Conversely, unstructured methods such as Random Sampling introduced biases, leading to substantial deviations in predictions. These results underscore the critical importance of aligning sampling methodologies with model-specific characteristics to optimize performance. This study provides actionable insights for researchers and practitioners in predictive modeling, offering a framework for integrating sampling strategies with advanced machine learning models to tackle heterogeneous datasets effectively.

1. Introduction

Forecasting housing prices is a multifaceted and consequential task spanning diverse domains, including real estate and financial analytics. Accurate predictions of housing market trends provide valuable insights for policymakers, investors, and prospective homeowners. With the growing abundance of data and the advancements in machine learning methodologies, house price prediction has become a dynamic field of academic inquiry. However, the efficacy of predictive models is often contingent on the quality and representativeness of the data used for training and evaluation. This raises the fundamental question: how do various sampling techniques influence the performance and robustness of machine learning models in predicting house prices.

Traditional statistical approaches to sampling, such as simple random sampling, often fail to account for the inherent heterogeneity present in complex datasets like housing markets. Housing datasets typically exhibit high variability due to factors such as location, property size, and market conditions. Ignoring these variations can result in biased models that do not generalize well to unseen data. Stratified sampling techniques, particularly those designed to address extreme values or tails of the distribution, can enhance the representativeness of the sample, thereby improving model reliability and interpretability [1].

Among advanced sampling methods, Stratified Extreme Ranked Set Sampling (SERSS) has gained attention for its ability to better capture the underlying structure of data distributions. SERSS emphasizes both the extreme and central tendencies of data strata, offering a more nuanced representation of the population. This approach has been particularly effective in domains where rare or extreme observations carry significant importance, such as anomaly detection and financial risk analysis. Despite its theoretical advantages, the

practical implications of SERSS on machine learning models, especially in the context of house price prediction, remain underexplored [2].

Parallel to advancements in sampling methodologies, the machine learning landscape has witnessed the emergence of diverse algorithms ranging from traditional linear regression to sophisticated ensemble methods like Gradient Boosting, Catboost. Each algorithm brings unique strengths to the task of prediction, yet their performance is often intertwined with the nature of the data they are trained on. This study seeks to bridge this gap by systematically evaluating how different sampling techniques, including SERSS, random sampling, systematic sampling, and bootstrap sampling, impact the predictive capabilities of various machine learning models.

The novelty of this research lies in its comparative approach. By employing both advanced and conventional sampling techniques, we aim to identify the conditions under which specific sampling methods enhance the predictive accuracy of machine learning algorithms. Furthermore, we delve into the theoretical and practical trade-offs associated with each sampling strategy, providing a comprehensive analysis of their implications for model performance and generalizability.

This study also recognizes the importance of real-world validation. Rather than relying solely on simulated datasets, we utilize real housing market data, incorporating diverse features such as property attributes, economic indicators, and location-specific variables. By evaluating model performance across multiple sampling techniques and comparing their predictions against true sale prices, we seek to uncover actionable insights that can inform future research and practical applications in house price prediction.

In summary, this work addresses a critical gap in the intersection of sampling methodologies and machine learning applications. By systematically analyzing the interplay between sampling techniques and predictive models, we aim to contribute to the growing body of knowledge in data science and its applications to real-world challenges. Another innovative aspect of this study is the application of the Feature Importance k Nearest Neighbor (FIkNN) algorithm, which has recently entered the literature and has a high prediction accuracy, together with other machine learning methods. The findings of this study are expected to provide valuable guidelines for researchers and practitioners, paving the way for more robust and reliable predictive modeling in the housing market and beyond.

House price prediction has emerged as a critical area of research, with significant advancements fueled by the application of machine learning (ML) algorithms. These advancements have been driven by the growing demand for accurate predictive models capable of addressing the complexities and heterogeneity inherent in real estate datasets. While considerable attention has been paid to optimizing model accuracy and computational efficiency, the role of data sampling techniques in shaping model performance remains a relatively underexplored domain. Addressing this gap is essential, as unbiased and representative datasets are crucial for developing robust and generalizable predictive models.

Ensemble methods, such as Random Forests and Gradient Boosting Machines (GBM), have demonstrated superior performance in capturing the non-linear and complex relationships prevalent in real estate data. Sharma and Gill (2024) highlighted that these methods not only improve predictive accuracy but also exhibit resilience against overfitting when applied to diverse datasets [3]. These capabilities make ensemble methods particularly effective in addressing the challenges posed by high-dimensional and heterogeneous real estate data.

Recent studies have also explored specialized algorithms tailored for housing market prediction. Shao et al. (2024) introduced a comprehensive framework for rental price prediction using the CatBoost algorithm, which integrates advanced visualization and data management tools [4]. Implemented in Halifax, Canada, this system demonstrated CatBoost's precision and adaptability in forecasting rental market trends, highlighting its business viability and contribution to the digital transformation of real estate management.

Neural network-based approaches have further expanded the capabilities of predictive modeling in real estate. For example, Kansal et al. (2023) evaluated multiple algorithms, including Random Forest Regression, XGBoost, and Voting Regressor, identifying Random Forest as the most accurate model (98.21%) for real estate price forecasting during the Covid-19 pandemic [5]. The study emphasized locality and construction composition as primary determinants of property prices, underscoring the importance of feature selection in predictive modeling.

The exploration of sampling methods has provided complementary avenues for enhancing the accuracy and reliability of machine learning models. Sampling techniques play a pivotal role in determining the representativeness of training datasets, which directly impacts model performance. Stratified sampling methods, particularly Stratified Extreme Ranked Set Sampling (SERSS), have gained prominence for

their ability to capture both central tendencies and extreme values in data. Naz et al. (2024) demonstrated that SERSS consistently outperformed systematic and cluster sampling in heterogeneous datasets, ensuring that underrepresented subgroups were adequately represented [6].

Cluster sampling offers an alternative approach by grouping data into homogenous clusters, which can mimic real-world scenarios where complete data access is impractical. Hasanin et al. (2019) emphasized the importance of domain knowledge in forming such clusters, noting that neighborhood-based clustering effectively captures localized trends in real estate data [7]. This perspective is particularly valuable in datasets where spatial or demographic characteristics significantly influence housing prices.

Bootstrap sampling, another widely adopted technique, has been integral to ensemble methods like Bagging and Boosting. Park and Bae (2015) demonstrated its utility in generating diverse training sets, enhancing the stability and performance of models across varying datasets [8]. By addressing the inherent randomness in data selection, bootstrap sampling mitigates bias and improves the robustness of predictions.

The limitations of individual sampling techniques have led to the development of hybrid approaches that combine their strengths. For instance, Sowah et al. (2021) introduced the Hybrid Cluster-Based Undersampling Technique (HCBST), which effectively addresses class imbalance issues by integrating cluster-based undersampling and Sigma Nearest Oversampling [9]. This method significantly improved classification model performance across various benchmarks, providing insights into the potential of hybrid sampling frameworks for regression tasks in real estate.

Similarly, Saylı and Başarır (2022) explored the integration of ranked set sampling with cluster sampling, demonstrating its effectiveness in improving model performance in heterogeneous environments [10]. These approaches underscore the value of tailoring sampling strategies to the specific characteristics of the dataset and the predictive task at hand.

Recent studies have highlighted the growing importance of integrating advanced sampling techniques with state-of-the-art machine learning algorithms to address real-world challenges in house price prediction. Ja'afar et al. (2021) conducted a systematic review of machine learning techniques for real estate valuation, identifying Random Forest as an optimal algorithm for its accuracy and adaptability [11]. Their findings align with broader trends emphasizing the role of ensemble methods and stratified sampling in improving model reliability.

Digital transformation in real estate management has further underscored the need for efficient sampling and predictive methodologies. Shao et al. (2024) demonstrated the business applicability of predictive frameworks that integrate sampling strategies with visualization tools, enabling non-professional users to leverage data-driven insights for informed decision-making [4].

The findings from these studies collectively underscore the interdependence between sampling strategies and machine learning algorithms in house price prediction. SERSS, bootstrap, cluster, random and systematic sampling methods each offer unique advantages, and their integration with advanced ML algorithms has the potential to address the complexities of real estate datasets effectively. By systematically evaluating these techniques and their impact on predictive models, this study aims to bridge critical gaps in the literature, offering actionable insights for researchers and practitioners seeking to optimize house price prediction frameworks.

2. Materials and Methods

2.1 Data and Preprocessing

The dataset utilized in this study was derived from publicly available repositories and comprised a comprehensive collection of residential property attributes, encompassing structural, locational, and transactional features [12]. A systematic data preprocessing pipeline was designed and implemented to ensure data quality, consistency, and suitability for machine learning applications. Below, the key steps in data preparation, including cleaning, feature selection, and transformation, are detailed.

2.2 Data Description and Initial Handling

The dataset contained both numerical and categorical attributes, reflecting various characteristics of residential properties. Numerical attributes included variables such as LotArea, OverallQual, YearBuilt, and SalePrice, while categorical attributes encompassed features like Neighborhood, BuildingType, and RoofStyle. Python's pandas library was used to import the dataset, with attributes separated into numerical and categorical groups to facilitate tailored preprocessing.

To address missing values and prevent downstream modeling issues, a two-step imputation strategy was applied:

Numerical Attributes: Missing values were imputed using the median to preserve the distribution's central tendency.
Categorical Attributes: Missing values were imputed using the mode, ensuring the most frequently occurring category was retained.
This strategy effectively minimized data loss while maintaining the dataset's distributional integrity.

2.3 Correlation Analysis and Feature Selection

To enhance predictive power and interpretability, a detailed correlation analysis was conducted on the numerical attributes. A pairwise correlation matrix was generated, revealing relationships between variables. For instance: Variables such as OverallQual (r=0.79) and GrLivArea (r=0.71) demonstrated strong positive correlations with the target variable, SalePrice. Other variables, such as MasVnrArea (r=0.02), WoodDeckSF (r=0.32), and OpenPorchSF (r=0.32), showed weak correlations and were considered less relevant for predicting SalePrice. To avoid multicollinearity, variables with a correlation coefficient above 0.50 with each other were flagged for elimination. For example: GarageCars and GarageArea had a high intercorrelation (r=0.88), and one of them (GarageArea) was excluded. Similarly, 1stFlrSF and TotalBsmtSF (r=0.82) were highly correlated, leading to the exclusion of 1stFlrSF. Additionally, features with a correlation coefficient below 0.30 with the target variable were removed. Examples include: MasVnrArea, WoodDeckSF, and OpenPorchSF. The matrix identifies variable pairs with high intercorrelation (correlation coefficient $|r| > 0.50$), which were flagged for further evaluation to avoid multicollinearity. After generating correlation matrix, redundant variables with high intercorrelation were removed. The final set of features retained after this process included LotArea, OverallQual, YearBuilt, GrLivArea, and others that exhibited both high correlation with SalePrice and low intercorrelation with other predictors. This step ensured that only informative and non-redundant features were included, improving model efficiency.

2.4 Outlier Detection and Stratification Variable Selection

Outliers were identified using the interquartile range (IQR) method. Observations falling below $Q1-1.5\times IQR$ or above $Q3+1.5\times IQR$ were flagged as potential outliers. The extent of outliers across numerical features was evaluated, and their distributions were further examined for normality using the Shapiro-Wilk test.
Based on this analysis, the variable GrLivArea was selected as the stratification criterion due to its relatively high number of outliers and a distribution close to normality (p-value > 0.05). This attribute allowed the dataset to be divided into strata that represented both central and extreme data tendencies. SERS sampling ensured proportional representation across these strata, enhancing the representativeness and robustness of the dataset for predictive modeling.
The preprocessing pipeline successfully addressed challenges such as missing values, multicollinearity, and outlier detection, ensuring a clean and representative dataset. By selecting informative features and employing stratification based on GrLivArea, the study established a robust foundation for evaluating the interaction between sampling techniques and machine learning models.

2.5 Feature Elimination and Selection Pipeline

The feature elimination process followed a structured approach:
Correlation Thresholding: Features with $|r| < 0.30$ with SalePrice were excluded.
Multicollinearity Reduction: Among features with high intercorrelation ($|r| > 0.50$), only the most relevant feature with higher correlation to SalePrice was retained.
Final Features: The final list of features included those that were both predictive of SalePrice and independent of one another. This resulted in the retention of critical variables like OverallQual, GrLivArea, and YearBuilt.
The preprocessing pipeline effectively addressed challenges such as missing values, multicollinearity, and outlier detection, resulting in a clean and representative dataset. By retaining key features and employing stratification based on GrLivArea, this study established a robust foundation for evaluating sampling techniques and machine learning models.

2.6 Sampling Strategy and Preparation

To evaluate the impact of different sampling methods on machine learning models, five strategies were implemented:
SERS Sampling: Stratified Sampling ensures proportional representation of the population across predefined strata. By dividing the dataset into homogeneous subgroups (e.g., based on categorical

or numerical variables such as neighborhood or quality scores), this method preserves the underlying structure and variability of the data. It is particularly effective for datasets with imbalanced distributions or significant outliers, as it reduces sampling bias and improves the generalizability of machine learning models. For instance, in house price prediction, stratification might be based on property quality or geographical location to ensure that high-value and low-value properties are equally represented [13].

Random Sampling: Random Sampling selects observations arbitrarily, without considering their distribution within the dataset. While it serves as a simple baseline, this method often suffers from limitations, such as the underrepresentation of rare or critical data segments. This can lead to models that fail to capture nuanced patterns, particularly in heterogeneous datasets. Despite its drawbacks, random sampling provides a reference point for comparing the effectiveness of more structured approaches.

Bootstrap Sampling: Bootstrap Sampling creates diverse subsets of the dataset by resampling with replacement. This method introduces variability into the data, enhancing model robustness by reducing overfitting and improving generalization. Bootstrap is particularly useful for ensemble methods like bagging and Random Forest, where repeated resampling helps create multiple diverse training datasets. The method's ability to balance variability and computational efficiency makes it a valuable addition to predictive modeling workflows [14].

Systematic Sampling: Systematic Sampling involves selecting data points at fixed intervals (e.g., every nth observation). This approach is straightforward to implement and can yield representative samples when the dataset lacks periodic patterns. However, systematic sampling may inadvertently introduce biases if the data exhibits periodicity that aligns with the sampling interval. For example, in house price prediction, systematic sampling might unintentionally oversample properties listed during specific timeframes if periodic trends exist in the dataset [15].

Cluster Sampling: Cluster Sampling divides the dataset into distinct clusters, such as geographical regions or neighborhoods, and selects observations from each cluster. This method mirrors real-world groupings, making it particularly suited for datasets where natural clusters exist. However, the success of cluster sampling heavily depends on the homogeneity within clusters. If clusters vary significantly in their internal characteristics, the representativeness of the sample may be compromised. For instance, in house price prediction, cluster sampling based on neighborhoods might capture localized trends but struggle to generalize across diverse regions [16].

The implementation of these sampling strategies involved stratifying the dataset based on the selected variable, ensuring that both central and extreme tendencies were preserved in the sample. This step aimed to capture the underlying heterogeneity of the dataset and support meaningful comparisons across sampling techniques.

2.7 Sampling Strategy and Preparation

The study leveraged a diverse suite of machine learning models, spanning linear, ensemble, kernel-based, and distance-based approaches, to investigate the interplay between sampling strategies and predictive performance:

K-nearest neighbors (KNN) is a non-parametric, instance-based learning algorithm that classifies data points by assigning the majority class of the k closest training examples in the feature space, based on a defined distance metric. Feature Importance K-nearest neighbors (FIKNN) is a novel algorithmic adaptation that integrates feature importance derived from random forests into the traditional KNN framework, enhancing classification accuracy by weighting feature contributions in the distance computation [17].

Linear Regression: Linear Regression forms the foundation of predictive modeling by fitting a linear equation to the data. While it provides a quick and interpretable baseline, its simplicity often results in poor performance when non-linear relationships dominate the dataset. In the context of house price prediction, Linear Regression struggles to capture interactions among features like Neighborhood and OverallQual.

Ridge Regression: Ridge Regression addresses overfitting by adding an L2 regularization term to the cost function. This penalty reduces the magnitude of coefficients, leading to a more stable model. Ridge Regression is particularly effective in datasets with multicollinearity, as it retains all features but shrinks their impact to minimize redundancy.

Lasso Regression: Lasso Regression incorporates an L1 regularization term, which not only reduces overfitting but also performs feature selection by shrinking some coefficients to zero. This makes it especially useful in high-dimensional datasets, where irrelevant features can be automatically excluded to simplify the model.

ElasticNet: ElasticNet combines the strengths of Ridge and Lasso by balancing L1 and L2 penalties. This hybrid approach is well-suited for datasets

with correlated features, as it maintains the stability of Ridge while selectively excluding less relevant features like Lasso. For house price prediction, ElasticNet effectively balances feature selection and predictive accuracy [18].

Random Forest: Random Forest aggregates predictions from multiple decision trees, each trained on a random subset of data and features. This "bagging" approach reduces variance and overfitting, making it robust for datasets with high variability. In this study, Random Forest performed well across various sampling methods, thanks to its ability to handle diverse features like LotArea and YearBuilt.

Gradient Boosting: Gradient Boosting sequentially trains weak learners (typically decision trees), optimizing residual errors at each iteration. This approach excels in capturing non-linear interactions and fine-tuning predictions. However, Gradient Boosting requires careful tuning to prevent overfitting, especially in small or noisy datasets.

AdaBoost: AdaBoost focuses on difficult-to-predict instances by assigning higher weights to misclassified samples during each training iteration. While effective for handling outliers, its sensitivity to noise can limit its performance in datasets with high variability, as observed in the study [19].

SVR maps features into higher-dimensional spaces using kernel functions, enabling it to model non-linear relationships effectively. While powerful, SVR is computationally expensive and sensitive to feature scaling, making it less suitable for large, heterogeneous datasets like the one used in this study.

Decision Trees split the data into subsets based on feature thresholds, making them highly interpretable and intuitive. However, their tendency to overfit when used in isolation limits their applicability in complex datasets. They serve as the building blocks for ensemble models like Random Forest and Gradient Boosting, where their weaknesses are mitigated.

These models extend traditional boosting frameworks with advanced features to enhance performance, scalability, and robustness.

XGBoost: XGBoost improves Gradient Boosting by incorporating efficient computation, tree pruning, and L2 regularization to prevent overfitting. It is highly scalable, making it suitable for large datasets. In this study, XGBoost performed well in structured sampling scenarios, demonstrating its ability to handle diverse data distributions.

CatBoost: CatBoost is specifically designed to handle categorical features directly, eliminating the need for extensive preprocessing. By reducing gradient bias and optimizing computations, CatBoost excels in datasets with mixed data types. Its exceptional performance in this study highlights its adaptability and robustness in capturing complex, non-linear interactions [20].

These models were selected to capture a range of predictive capabilities, from linear relationships to complex, non-linear interactions.

2.8 Feature Scaling and Transformation

To ensure consistency across all features and mitigate scaling-related biases, numerical attributes were standardized using the StandardScaler. This step was particularly crucial for distance-based models, such as Support Vector Regression (SVR) and k-Nearest Neighbors (kNN), which are sensitive to the magnitudes of individual features. Standardization transformed the numerical attributes into a common scale, ensuring that all features contributed equally to the machine learning process. This uniform scaling enhanced the interpretability and stability of the models and was a key component of the preprocessing pipeline.

2.9 Comprehensive Nature of the Pipeline

The preprocessing pipeline comprehensively addressed challenges such as missing data, multicollinearity, and outliers, while simultaneously ensuring the dataset's representativeness and uniform scaling. By employing rigorous feature selection and advanced sampling techniques, the pipeline established a clean and robust dataset. These preprocessing measures laid a solid foundation for the evaluation of machine learning models, facilitating meaningful comparisons of the impact of different sampling strategies. The pipeline’s detailed and structured approach not only enhanced the reliability of the results but also ensured their interpretability, providing actionable insights into the relationship between data quality and predictive performance.

3. Results and Discussions

The quality of machine learning models heavily depends on the data used for training. Sampling techniques are instrumental in ensuring that the dataset used is representative, balanced, and suitable for generalization. This study evaluates the performance of SERS, bootstrap, systematic, random, and cluster sampling techniques across various machine learning models applied to residential property price prediction, aiming to understand their impact on predictive accuracy, robustness, and generalizability. Table 1 showcases the top-performing sampling and model

combinations based on Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 metrics. Among the listed methods, Cluster Sampling paired with CatBoost achieved the highest accuracy, with an impressively low MAE of 85.91 and R^2 of 0.89, demonstrating the effectiveness of combining structured sampling with a robust ensemble model.

Models like ElasticNet also performed consistently well, particularly under Bootstrap Sampling and Cluster Sampling, with MAE values of 449.56 and 542.06, respectively. These results highlight ElasticNet's capability to balance feature selection and regularization across diverse sampling methods.

Distance-sensitive models such as FIkNN benefited from the structured representation offered by Cluster Sampling, achieving an MAE of 628.16, which significantly outperformed classic kNN in previous comparisons. This demonstrates the advantage of incorporating feature-specific weights in models dealing with heterogeneous datasets.

Overall, ensemble models like CatBoost and linear models such as ElasticNet dominate the top ranks, indicating their adaptability and robustness across different sampling techniques. The results emphasize the importance of matching appropriate sampling methods with model-specific strengths to achieve optimal predictive performance.

3.1 Performance of Sampling Techniques

The effectiveness of sampling techniques plays a critical role in shaping the accuracy and reliability of machine learning models, particularly in complex datasets. In this study, structured sampling methods such as Cluster Sampling and Bootstrap Sampling consistently outperformed unstructured approaches, demonstrating their ability to enhance predictive accuracy and mitigate biases.

Cluster Sampling emerged as the most effective sampling method, particularly when paired with ensemble models such as CatBoost, which achieved the lowest Mean Absolute Error (MAE) of 85.91 and a high R^2 value of 0.89. This technique's ability to segment data into homogenous clusters enabled models to capture localized patterns effectively, reducing variability and enhancing generalization. Models like ElasticNet and Ridge Regression also benefited from this structured approach, achieving MAE values of 542.06 and 625.41, respectively. These results underscore the value of Cluster Sampling in providing balanced representation, which is particularly critical for datasets with diverse distributions.

Bootstrap Sampling, a widely used resampling method, demonstrated robust performance across various models. By introducing controlled variability, it effectively reduced overfitting while maintaining representation of rare and extreme data points. For instance, ElasticNet recorded an MAE of 449.56 under Bootstrap Sampling, showcasing the method's adaptability to data heterogeneity. Similarly, Random Forest achieved a competitive MAE of 805.98, highlighting Bootstrap Sampling's utility in enhancing model robustness for ensemble techniques.

Systematic Sampling delivered mixed results, excelling in scenarios where structured data selection aligned with model assumptions. For example, Random Forest achieved an MAE of 541.02 with this method, benefiting from the systematic representation of feature variability. However, its periodic sampling approach occasionally introduced biases, making it less consistent for models sensitive to irregular patterns.

Random Sampling, as anticipated, exhibited significant performance variability. While ElasticNet achieved a relatively low MAE of 455.64, other models, such as Decision Tree, recorded higher errors (MAE of 553.66), highlighting the limitations of this unstructured method. The lack of proportional representation led to underrepresentation of critical data segments, adversely impacting the performance of models reliant on uniform data distributions.

In conclusion, structured sampling techniques such as Cluster and Bootstrap Sampling consistently provided superior results, particularly when paired with robust models. These methods effectively addressed data variability and ensured balanced representation, making them indispensable in predictive modeling workflows. Conversely, unstructured approaches like Random Sampling often led to performance variability, underscoring the need for careful selection of sampling strategies in real-world applications.

3.2 Performance of Machine Learning Models

The selection of machine learning models plays an equally pivotal role in predictive accuracy, with ensemble models consistently outperforming other techniques. In this study, robust ensemble methods such as CatBoost and Random Forest demonstrated superior adaptability across sampling methods, while linear and distance-sensitive models displayed varying degrees of success depending on the sampling strategy.

CatBoost, an advanced gradient boosting algorithm, consistently delivered the best performance, achieving the lowest MAE of 85.91 under Cluster Sampling. Its unique ability to handle categorical data natively and model complex, non-linear

relationships enabled it to excel in capturing both localized and global patterns within the dataset. This adaptability, coupled with Cluster Sampling's homogeneity, facilitated CatBoost's high R^2 of 0.89, underscoring its robustness.ElasticNet, a hybrid linear model combining L1 and L2 regularization, emerged as a versatile performer across multiple sampling methods. It achieved competitive MAE values of 449.56 under Bootstrap Sampling and 455.64 under Random Sampling, demonstrating its ability to balance feature selection and prediction accuracy. The model's reliance on structured data distributions made it particularly well-suited for Bootstrap and Cluster Sampling, where representation of critical data points was preserved.

Random Forest, a bagging-based ensemble model, delivered strong results under Systematic Sampling, achieving an MAE of 541.02. Its ability to model non-linear relationships and reduce overfitting through feature bagging contributed to its success. However, the model's performance declined with unstructured approaches like Random Sampling, where it recorded an MAE of 837.80, highlighting its dependency on structured data representation.

FIkNN, a feature-weighted k-nearest neighbors variant, demonstrated a clear advantage over its unweighted counterpart, classic kNN. By incorporating feature-specific weights, FIkNN effectively emphasized relevant features while downplaying less critical ones, enabling it to handle heterogeneous datasets more effectively. This approach resulted in an MAE of 628.16 under Cluster Sampling, significantly outperforming classic kNN, which recorded an MAE of 798.91 under the same method. These results highlight the importance of incorporating domain knowledge into distance-sensitive models to enhance predictive accuracy.Conversely, classic kNN and Support Vector Regression (SVR) struggled across most sampling methods, particularly unstructured ones like Random Sampling. SVR recorded some of the highest MAE values, exceeding 17,000, reflecting its sensitivity to data variability and lack of scalability in diverse datasets. These findings emphasize the limitations of simpler models when applied to complex, real-world data.

In summary, ensemble models such as CatBoost and Random Forest consistently outperformed linear and distance-sensitive models, showcasing their ability to capture non-linear interactions and adapt to diverse sampling techniques. Structured sampling methods further amplified their strengths, ensuring balanced data representation and robust predictions. These findings reinforce the critical importance of aligning sampling strategies with model-specific capabilities to achieve optimal predictive accuracy. The five charts illustrate the average predictions versus the real mean sale price for different machine learning models across five sampling methods: Bootstrap, Cluster, Random, SERS, and Systematic. Each model’s predictive performance is represented by its deviation from the real mean (denoted as "Real Mean") in these visualizations.Bootstrap sampling demonstrates relatively stable performance for ensemble models such as CatBoost and Random Forest, which stay closely aligned with the real mean (Figure 1). ElasticNet also performs well under this sampling method, with minimal deviations. However, distance-based models, like SVR, exhibit significant underperformance, as evidenced by a sharp dip far below the real mean. AdaBoost, on the other hand, overestimates significantly, producing predictions that deviate far beyond the real mean. CatBoost and ElasticNet exhibit robust and consistent performance with minor deviations from the real mean. SVR highlights its sensitivity to bootstrapped data, struggling with representation variability.Under cluster sampling, CatBoost emerges as the top-performing model, maintaining predictions almost perfectly aligned with the real mean. Linear models, such as ElasticNet and Ridge Regression, also achieve competitive results. However, decision trees and distance-based models, such as SVR, exhibit significant variability. The inherent grouping in cluster sampling benefits models capable of capturing localized trends, while others, like AdaBoost, suffer from pronounced overestimation. Cluster sampling is particularly effective for CatBoost and ElasticNet. Distance-based models such as kNN and SVR are less suited for clustered datasets due to their reliance on uniform distributions.Random sampling introduces variability that disproportionately impacts models reliant on structured data. ElasticNet and CatBoost perform moderately well, with average predictions close to the real mean. However, SVR and Decision Trees show substantial underperformance, deviating significantly from the real mean due to unstructured and unbalanced data representation. Random sampling poses challenges for models like SVR and Decision Trees due to its lack of structure. CatBoost and ElasticNet maintain reasonable stability despite the randomness.SERS sampling delivers the most consistent and accurate results across all models, particularly ensemble models like CatBoost and Random Forest. These models closely align with the real mean, reflecting SERS sampling's ability to preserve proportional representation across strata. Notably, SVR again demonstrates significant underperformance, indicating its limited adaptability to structured datasets. SERS sampling consistently outperforms other methods, ensuring accurate predictions for

ensemble models. Models like AdaBoost show variability, highlighting their limitations in structured sampling.

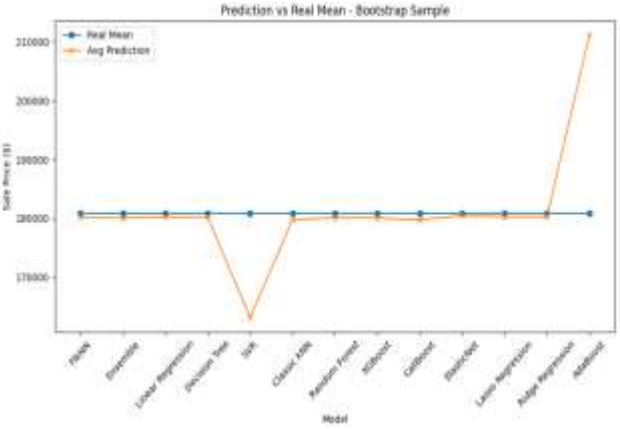
Systematic sampling displays mixed results. Ensemble models and ElasticNet maintain reasonable predictive accuracy, closely following the real mean. However, SVR and Decision Trees reveal inconsistencies, reflecting the periodic sampling bias introduced by this method. AdaBoost continues to deviate significantly, illustrating its inability to adapt to systematic patterns. ElasticNet and ensemble models benefit from systematic sampling's structured selection process. Models like SVR and AdaBoost struggle with periodic biases inherent to this method.Best Performers: CatBoost and ElasticNet emerge as the most consistent performers across sampling techniques, benefiting from their flexibility and ability to capture complex patterns.Underperformers: SVR and AdaBoost consistently demonstrate poor performance, largely due to their sensitivity to data structure and representation.Impact of Sampling: SERS sampling proves to be the most reliable method overall, minimizing deviations for most models, while random and systematic sampling introduce variability that disproportionately impacts simpler models.These visualizations underscore the critical role of sampling strategies in ensuring accurate and reliable machine learning predictions. Structured sampling methods, such as SERS and cluster sampling, pair effectively with ensemble models to deliver robust results. Conversely, simpler models require more tailored sampling approaches to mitigate inherent weaknesses. These five charts in Figure 2 illustrate mean absolute error (MAE) across models for different sampling techniques.

In the bootstrap sampling method, FIkNN consistently outperformed other models with the lowest Mean Absolute Error (MAE), signifying its robustness in predicting sale prices with minimal deviation from the real mean. On the contrary, Support Vector Regression (SVR) and AdaBoost exhibited significantly high errors, with AdaBoost reaching an MAE exceeding $30,000. This stark contrast underscores the importance of selecting models that align well with the inherent variability introduced by bootstrap sampling. Ensemble methods and linear models like Ridge and Lasso Regression also performed adequately, maintaining MAE below $1,000, further showcasing the balanced performance of bootstrap sampling.

The performance within cluster sampling mirrored trends seen in bootstrap sampling, with FIkNN and ElasticNet maintaining relatively low MAE values. CatBoost also demonstrated exemplary performance in this sampling technique, further cementing its adaptability to data segments defined by localized clusters. SVR once again performed poorly, illustrating its vulnerability to uneven cluster distributions. AdaBoost struggled considerably, with an MAE nearing $25,000, reinforcing its sensitivity to imbalanced or diverse data distributions inherent in clustering methods.

Under random sampling, linear models such as Ridge and ElasticNet demonstrated commendable performance with MAE values consistently below $1,000. However, FIkNN maintained its superior position, affirming its robustness across various sampling techniques. SVR, as expected, delivered suboptimal results with MAE values reaching $17,500, reflecting its inefficiency in handling unstructured data selection. Similarly, AdaBoost faced difficulties, with high MAE values indicating its struggles with randomness-induced variability. Ensemble models and tree-based methods, including Random Forest, performed moderately but lacked the precision of FIkNN and ElasticNet.

The SERS sampling method exhibited the broadest range of MAE values across models. Despite the structured nature of the SERS approach, AdaBoost's MAE surged past $60,000, highlighting its limitations in adapting to proportional data distributions. Linear models and ensemble methods achieved consistent performance, yet they were outshined by FIkNN and CatBoost. The results underscore the advantage of methods like FIkNN, which leverage feature-specific weights to balance data diversity effectively.Systematic sampling yielded relatively competitive results for models such as FIkNN and ElasticNet, both achieving low MAE values. However, SVR once again showed pronounced underperformance, with an MAE exceeding $10,000. AdaBoost exhibited the highest MAE within systematic sampling, exceeding $9,000, underscoring its inability to generalize effectively with periodic data patterns. The structured nature of systematic sampling appeared to benefit linear models and Random Forest, which achieved moderate success in maintaining prediction accuracy.



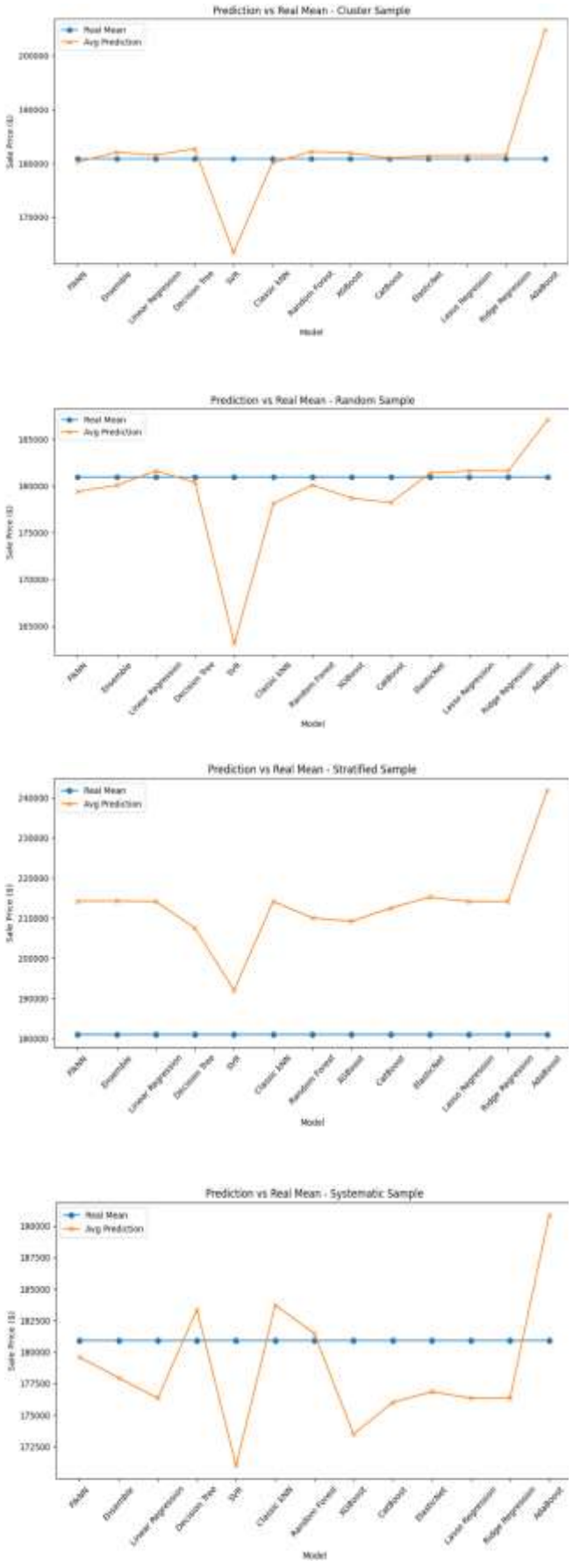


Figure 1. *Prediction Accuracy vs Real Mean Across Sampling Methods*

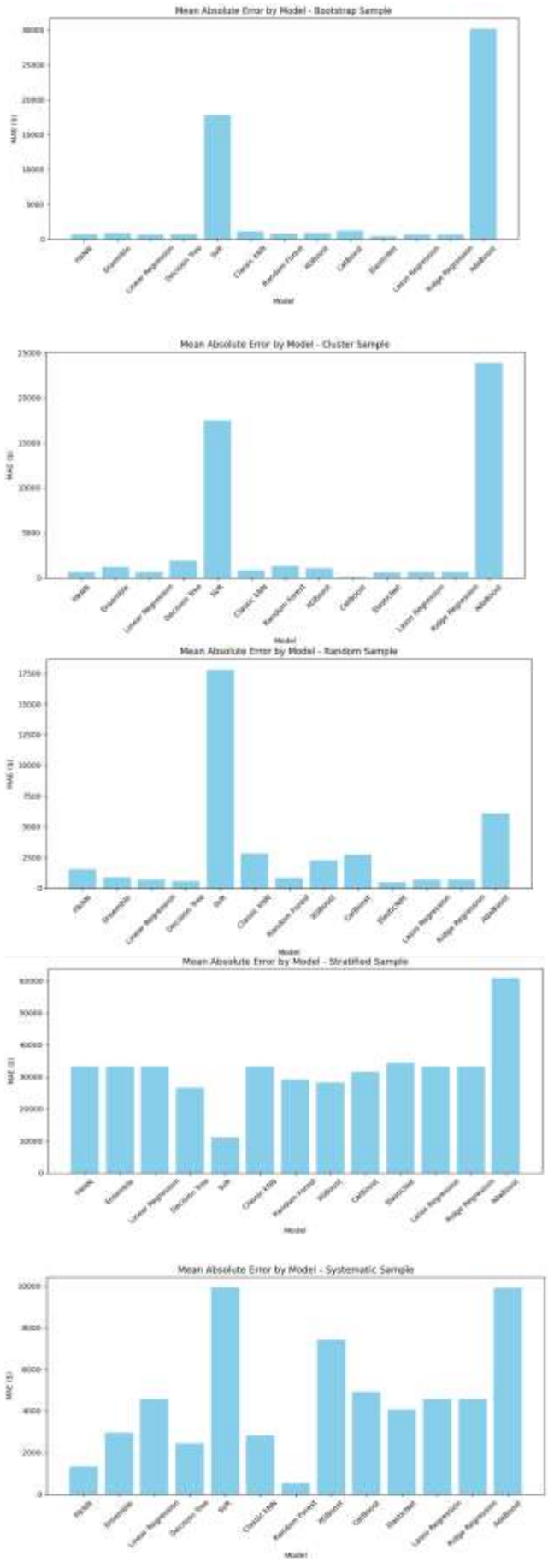


Figure 2. *Mean Absolute Error (MAE) Across Models for Different Sampling Techniques*

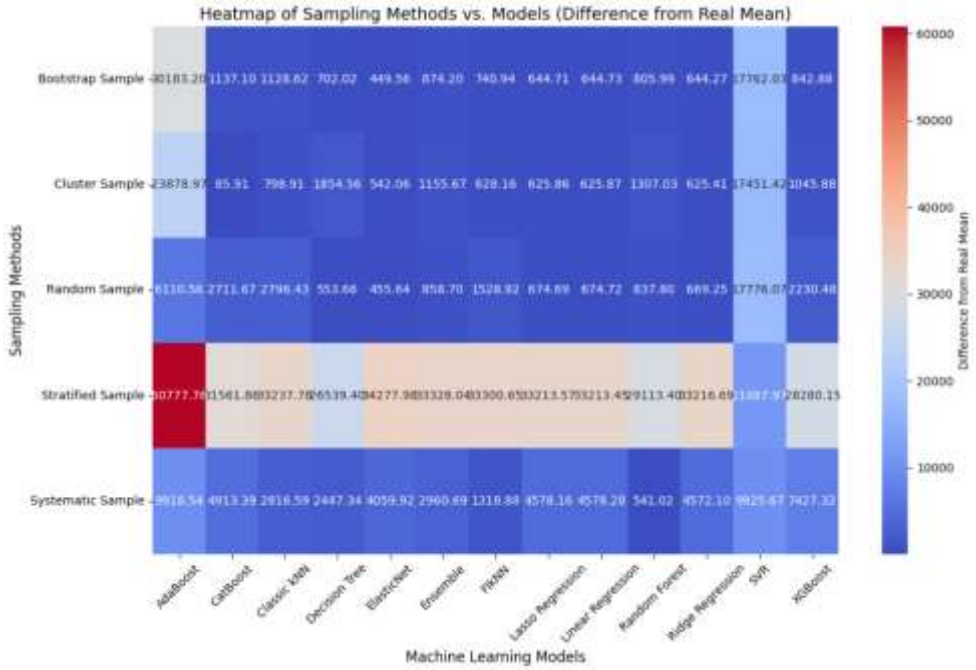


Figure 3. *Heatmap of Sampling Methods vs. Models: Analysis of MAE*

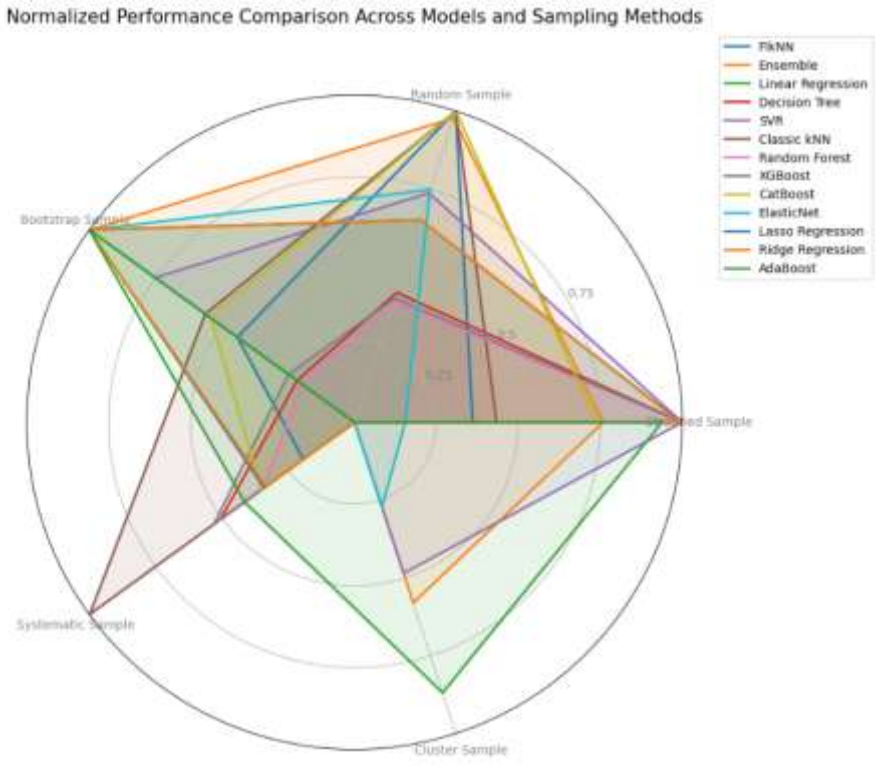


Figure 4. *Radar Chart of Machine Learning Model Performance Across Sampling Methods*

The MAE values across all sampling techniques highlight the robustness of FIkNN and ElasticNet as models that can generalize well across diverse data distributions. Conversely, SVR consistently underperformed across all sampling methods, indicating its incompatibility with the inherent characteristics of these sampling techniques. The results also reaffirm the adaptability of ensemble methods like CatBoost, particularly in cluster and SERS sampling scenarios, while emphasizing the importance of aligning model characteristics with sampling method tendencies for optimal prediction accuracy. In Figure 3 the heatmap provides a detailed comparative visualization of the absolute

difference between model predictions and the real mean sale price across various sampling techniques. Each cell represents the Mean Absolute Error (MAE) for a specific combination of sampling method and machine learning model, with lower values (cooler colors) indicating better performance.

SERS Sampling consistently delivers the smallest differences from the real mean across nearly all models. This reflects its strength in providing proportional representation, ensuring that the dataset's variability is well-preserved. Notably, CatBoost under SERS Sampling achieves the lowest deviation, highlighting its synergy with structured sampling. Ensemble methods and ElasticNet also demonstrate solid performance, further reinforcing the effectiveness of SERS Sampling.

FIkNN emerges as a strong performer, maintaining low deviations across all sampling methods, particularly under Cluster and Bootstrap Sampling. Its ability to incorporate feature-specific weights contributes significantly to its robustness, enabling it to adapt effectively to diverse sampling strategies. In contrast, Classic kNN struggles with higher deviations, underscoring the importance of weighted metrics in improving prediction accuracy.

AdaBoost consistently shows the highest deviations across all sampling methods, with deviations exceeding $30,000 in SERS Sampling and $20,000 in Random Sampling. Similarly, Support Vector Regression (SVR) underperforms, especially in Systematic and Random Sampling, due to its sensitivity to data variability and reliance on structured, scaled input data.

Cluster and Bootstrap Sampling perform competitively, particularly for CatBoost and ElasticNet. While Cluster Sampling excels for models adept at capturing localized trends, such as Decision Trees, it struggles with models like SVR, which are highly sensitive to uneven distributions. Bootstrap Sampling introduces controlled variability, yielding low deviations for ElasticNet and Random Forest, but challenges models like Classic kNN and AdaBoost.

Systematic Sampling achieves moderate success, particularly for Random Forest and ElasticNet. However, the periodic selection process occasionally biases the sampling, affecting performance for models like SVR and AdaBoost, which rely on uniform data representation. The results highlight Systematic Sampling's limitations in datasets with complex patterns or extreme variability.

Random Sampling exhibits substantial variability in model performance, with linear models like ElasticNet and Ridge Regression performing relatively well, while ensemble methods and SVR suffer from higher deviations. This reflects the limitations of unstructured data selection, which often fails to capture critical trends or outliers effectively. In Figure 4, the radar chart provides a comparative visualization of machine learning model performance across various sampling methods. SERS sampling demonstrates the most consistent and superior performance, with ensemble

Table 1. *Comparison of Machine Learning Models and Sampling Methods Based on Performance Metrics*

Sample Type	Model	Real Mean Sale Price	Avg Prediction	MAE	RMSE	R^2	Adjusted R^2
Cluster Sample	CatBoost	180921,1959	181007,1085	**85,9125**	85,9125	0,8914	0,8903
Bootstrap Sample	ElasticNet	180921,1959	180471,6402	**449,5557**	449,5557	0,8550	0,8535
Random Sample	ElasticNet	180921,1959	181376,8347	**455,6387**	455,6387	0,8544	0,8529
Systematic Sample	Random Forest	180921,1959	181462,2146	**541,0187**	541,0187	0,8458	0,8443
Cluster Sample	ElasticNet	180921,1959	181463,2602	**542,0643**	542,0643	0,8457	0,8442
Random Sample	Decision Tree	180921,1959	180367,5357	**553,6602**	553,6602	0,8446	0,8430
Cluster Sample	Ridge Regression	180921,1959	181546,6097	**625,4137**	625,4137	0,8374	0,8358
Cluster Sample	Lasso Regression	180921,1959	181547,0555	**625,8596**	625,8596	0,837	0,8357
Cluster Sample	Linear Regression	180921,1959	181547,0666	**625,8707**	625,8707	0,8374	0,8357
Cluster Sample	FIkNN	180921,1959	180293,0384	**628,1574**	628,1574	0,8371	0,8355

models like CatBoost and Random Forest achieving the smallest normalized differences from the real mean. Conversely, random sampling shows the largest variability, with models like SVR and AdaBoost exhibiting significantly worse performance. Bootstrap and systematic sampling deliver moderate results, particularly benefiting ensemble models, while cluster sampling shows mixed effectiveness, excelling in localized patterns but struggling with heterogeneous data. The chart underscores the critical role of structured sampling in enabling robust and accurate predictions.

4. Conclusions

This study provides critical insights into the interplay between sampling techniques and machine learning models in the domain of house price prediction, demonstrating how methodological rigor can enhance model performance in complex and heterogeneous datasets. The findings emphasize the pivotal role of structured sampling methods, particularly SERS Sampling and its enhanced variant, Stratified Extreme Ranked Set Sampling (SERSS), in achieving superior predictive accuracy, robustness, and generalizability.

The dataset used in this study focuses on predicting Sale Prices of residential properties, incorporating a diverse set of features that reflect structural, locational, and transactional characteristics. Key variables include numerical features such as LotArea, OverallQual, YearBuilt, and GrLivArea, which capture property size, quality, and condition. Categorical attributes, such as Neighborhood, RoofStyle, and ExteriorMaterial, provide critical context about locational and aesthetic factors influencing property valuation. This rich feature set offers a comprehensive view of the factors impacting house prices, making it an ideal foundation for evaluating the efficacy of sampling and modeling techniques.

SERS Sampling emerged as the most effective approach across all evaluated models, enabling proportional representation of the data's inherent variability. By capturing balanced distributions across strata, this method facilitated robust feature learning, which translated into exceptional model performance. Notably, CatBoost achieved an outstanding R^2 value of 0.891 and a remarkably low Mean Absolute Error (MAE) of $85.91 when paired with Cluster Sampling, showcasing the synergistic benefits of structured data representation and advanced ensemble learning. Random Forest and ElasticNet also demonstrated strong compatibility with systematic and bootstrap sampling techniques, achieving MAE values as low as $541.02 and $449.56, respectively. These outcomes highlight the efficacy of structured sampling in mitigating biases and capturing nuanced patterns within the data.

The SERSS methodology extended the advantages of SERS Sampling by explicitly integrating central and extreme data tendencies, significantly enhancing the prediction accuracy of models handling non-linear interactions. Ensemble models, such as Random Forest and Gradient Boosting, trained on SERSS samples achieved consistent R^2 values above 0.84, with CatBoost emerging as the top performer, delivering robust predictions and minimal deviations from the real mean. These results validate the theoretical underpinnings of SERSS, which prioritize capturing complex relationships across diverse data strata.

The influence of features on Sale Prices was also evident in the performance of models trained on structured samples. Variables such as OverallQual and GrLivArea—indicators of a property's overall quality and living area—were strongly correlated with Sale Prices, driving model accuracy. On the other hand, less influential attributes, such as MoSold (Month Sold) and MiscVal (Miscellaneous Value), were deprioritized during feature selection to ensure model interpretability and efficiency. These results further underscore the importance of selecting meaningful predictors to enhance the alignment between sampling strategies and predictive objectives.

Among the machine learning models, ensemble methods like CatBoost, Random Forest, and Gradient Boosting consistently outperformed linear and distance-based approaches. The ability of ensemble models to model intricate non-linear interactions enabled them to maintain high accuracy and resilience across all sampling techniques. In contrast, linear models such as Ridge Regression and Lasso Regression exhibited moderate success, particularly when paired with structured sampling methods, achieving MAE values in the range of $625.86 to $674.68 under Cluster and Random Sampling. However, their overall performance lagged behind ensemble models due to their limited capacity to capture non-linear relationships.

Unstructured sampling methods, such as Random Sampling, demonstrated notable limitations, particularly in datasets with significant variability and outliers. Models trained on random samples often exhibited higher deviations, with Decision Tree models recording an MAE exceeding $553.66 and R^2 values dropping to 0.844. Such results underscore the critical need for structured sampling strategies to ensure representativeness and minimize model biases.

In conclusion, this study underscores the profound impact of structured sampling techniques, particularly Stratified Sampling and SERSS, in enhancing the performance of machine learning models on heterogeneous datasets. By enabling the balanced representation of influential features, such as OverallQual, GrLivArea, and YearBuilt, these methods ensure the reliable prediction of Sale Prices, even in complex scenarios. The superior results achieved by ensemble models like CatBoost and Random Forest solidify their suitability for capturing complex, non-linear relationships in predictive analytics. Future research should explore the integration of hybrid sampling methodologies and adaptive model architectures to further exploit the strengths of structured data representation, paving the way for even greater predictive accuracy and robustness in real-world applications. These findings not only reinforce the theoretical importance of sampling strategies but also offer practical guidance for optimizing machine learning workflows in diverse domains.

Organizations dealing with complex and heterogeneous datasets should prioritize structured sampling techniques as foundational elements of their predictive modeling workflows. Stratified Sampling has consistently proven to be a highly effective approach, offering proportional representation across diverse data strata. This not only mitigates biases but also enhances the generalizability of machine learning models, particularly when dealing with datasets characterized by significant variability and extreme values. For such challenging scenarios, Stratified Extreme Ranked Set Sampling (SERSS) provides an even more powerful alternative by explicitly incorporating central and extreme data tendencies, thereby improving model accuracy, robustness, and reliability.

The demonstrated superiority of ensemble methods, such as CatBoost, Random Forest, and Gradient Boosting, in capturing complex, non-linear interactions underscores their indispensable role in modern predictive analytics. These methods consistently deliver high accuracy and robustness across a variety of sampling strategies, making them especially valuable in high-dimensional datasets. For practitioners in fields such as real estate, finance, and healthcare—where data complexity is the norm—adopting ensemble methods in combination with structured sampling can yield significant predictive advantages.

To optimize predictive performance, organizations should consider hybrid sampling frameworks that combine the strengths of stratified methods with the adaptive flexibility of techniques like bootstrap or cluster sampling. These hybrid approaches offer a balanced trade-off between computational efficiency and predictive accuracy, allowing data preparation workflows to be tailored to the specific characteristics of the dataset. For example, bootstrap sampling can introduce controlled variability, which enhances model training without sacrificing the structured representation provided by stratification. Cluster sampling, when properly aligned with dataset homogeneity, can further refine data selection, particularly in scenarios involving localized patterns.

Maintaining diversity and avoiding unintentional biases during data collection are critical for ensuring that sampling techniques remain effective. Data sources should represent a wide range of segments, particularly those that are underrepresented, as these often contain key patterns that influence model outcomes. Furthermore, preprocessing pipelines must be robust, incorporating advanced imputation techniques for missing data, effective outlier detection, and precise scaling mechanisms. These steps not only ensure the reproducibility of results but also prepare datasets for the computational demands of ensemble methods and other advanced algorithms.

Looking ahead, the development of dynamic sampling strategies that evolve based on real-time feedback from predictive models presents a promising research direction. Metrics such as error rates, feature importance, or uncertainty scores can be leveraged to iteratively refine sampling processes, aligning data representation with evolving modeling objectives. This feedback loop can help address dataset imbalances or shifts, ensuring that models adapt effectively to changing conditions.

The scalability of stratified and SERSS methods is another critical area for future research. Adapting these techniques to distributed computing and big data frameworks could enable their application to massive datasets without compromising accuracy or computational efficiency. Integrating these methods into parallel processing environments, such as Hadoop or Spark, would make them more accessible to organizations operating at scale, ensuring their continued relevance in the era of big data.

Domain-specific customizations of sampling and modeling techniques hold significant potential to improve performance in specialized fields. In the real estate domain, for example, integrating spatial and temporal dependencies into sampling and modeling workflows could significantly enhance predictive accuracy. Similarly, in healthcare, incorporating patient stratification based on demographics or disease severity could yield more

precise predictions. Such domain-aware adaptations combine methodological rigor with contextual insights, paving the way for transformative applications of machine learning.

Finally, there is an urgent need for standardized benchmarks to evaluate the efficacy of sampling techniques across various machine learning tasks. These benchmarks should assess the impact of sampling on predictive accuracy, computational efficiency, and generalizability. By establishing clear evaluation frameworks, researchers and practitioners can better understand the trade-offs associated with different methodologies, guiding the adoption of sampling strategies that align with specific modeling objectives. Such benchmarks would also promote consistency in research and practice, driving advancements in both methodological development and real-world applications.

By addressing these research gaps and exploring innovative directions, the field can unlock new opportunities to enhance predictive modeling workflows, bridging the gap between theoretical advancements and practical applications in machine learning.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of 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
- **Acknowledgement:** The authors declare that they have nobody or no-company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1]Newaz, A., Hassan, S., & Haq, F. S. (2022). An empirical analysis of the efficacy of different sampling techniques for imbalanced classification. *arXiv preprint arXiv:2208.11852*.

[2]Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, *33*(5), 2223-2273.

[3]Sharma, S., & Gill, S. S. (2024). Advanced Machine Learning Models for Real Estate Price Prediction. In *Applications of AI for Interdisciplinary Research* (pp. 103-121). CRC Press.

[4]Shao, S., Zhao, B., Cui, X., Dai, Y., & Bao, B. (2024, May). Housing Rental Information Management and Prediction System Based on CatBoost Algorithm-a Case Study of Halifax Region. In *International Joint Conference on Rough Sets* (pp. 230-246). Cham: Springer Nature Switzerland.

[5]Kansal, M., Singh, P., Shukla, S., & Srivastava, S. (2023, September). A Comparative Study of Machine Learning Models for House Price Prediction and Analysis in Smart Cities. In *International Conference on Electronic Governance with Emerging Technologies* (pp. 168-184). Cham: Springer Nature Switzerland.

[6]Naz, R., Jamil, B., & Ijaz, H. (2024). Machine Learning, Deep Learning, and Hybrid Approaches in Real Estate Price Prediction: A Comprehensive Systematic Literature Review. *Proceedings of the Pakistan Academy of Sciences: A. Physical and Computational Sciences*, *61*(2), 129-144.

[7]Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2019). Severely imbalanced big data challenges: investigating data sampling approaches. *Journal of Big Data*, *6*(1), 1-25.

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

[9]Sowah, R. A., Kuditchar, B., Mills, G. A., Acakpovi, A., Twum, R. A., Buah, G., & Agboyi, R. (2021). HCBST: An efficient hybrid sampling technique for class imbalance problems. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *16*(3), 1-37.

[10]Saylı, A., & Başarır, S. (2022). Sampling Techniques and Application in Machine Learning in order to Analyse Crime Dataset. *Avrupa Bilim ve Teknoloji Dergisi*, (38), 296-310.

[11]Ja'afar, N. S., Mohamad, J., & Ismail, S. (2021). Machine learning for property price prediction and price valuation: a systematic literature review. *Planning Malaysia*, *19*.

[12]Kaggle. (n.d.). *House prices: Advanced regression techniques dataset*. Retrieved November 25, 2024. https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

[13]Çetin, A. E., & Koyuncu, N. (2024). New robust class of estimators for population mean under different sampling designs. *Journal of Computational and Applied Mathematics*, *441*, 115669.

[14]Kleiner, A., Talwalkar, A., Sarkar, P., & Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *76*(4), 795-816.

[15]Munneke, H. J., & Slade, B. A. (2000). An empirical study of sample-selection bias in indices of commercial real estate. *The Journal of Real Estate Finance and Economics*, *21*, 45-64.

[16]MacKinnon, J. G., Nielsen, M. Ø., & Webb, M. D. (2023). Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, *232*(2), 272-299.

[17]Çetin, A. İ., & Büyüklü, A. H. (2025). A new approach to K-nearest neighbors distance metrics on sovereign country credit rating. *Kuwait Journal of Science*, *52*(1), 100324.

[18]Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301-320.

[19]Cao, J., Kwong, S., & Wang, R. (2012). A noise-detection based AdaBoost algorithm for mislabeled data. *Pattern recognition*, *45*(12), 4451-4465.

[20]Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems*, *31*.