International Journal of Computer and Information System (IJCIS)
Peer Reviewed – International Journal
Vol : Vol. 06, Issue 03, August 2025
e-ISSN : 2745-9659
https://ijcis.net/index.php/ijcis/index

Evaluating Machine Learning Algorithms for Predictive Modeling of Large-scale Event Attendance

Deni Kurnianto Nugroho[1] , Marwan Noor Fauzy[2], Kardilah Rohmat Hidayat[3]
Department of Information System, Faculty of Computer Science
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
deni@amikom.ac.id[1] , marwannoorfauzy@amikom.ac.id[2], kardilah.rh@amikom.ac.id[3]

Abstract*—Predicting attendance at large-scale public events is a critical task to support better resource planning, logistics, and safety management. This study investigates the performance of various machine learning models in forecasting event attendance using metadata features such as event type, venue, location, date, and duration. The dataset comprises over 19526 event records obtained from a U.S. government open data repository, covering multiple years and diverse event categories. Model performance was evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Among the models tested, ensemble methods particularly Gradient Boosting Regressor and XGBoost outperformed others, achieving the lowest MAE (61.37 and 59.52, respectively) and the highest R^2 values (0.22 and 0.15). These results suggest superior generalization capability in capturing complex nonlinear patterns in the data. In contrast, linear models and simpler non-parametric methods such as Decision Trees and K-Nearest Neighbors (KNN) exhibited relatively weaker predictive accuracy, with R^2 scores close to or below 0.14. While the R^2 values indicate that metadata alone provides a limited view of attendance dynamics, the relatively low MAE across models implies that reasonable point predictions are still achievable. These findings highlight the potential of ensemble-based methods for baseline forecasting tasks. Furthermore, the study underscores the importance of incorporating richer feature sets such as pricing, weather, promotional activity, and social sentiment for future model improvement. This research provides a foundational benchmark for data-driven attendance forecasting and offers practical implications for event organizers seeking scalable, automated prediction tools to support strategic planning.*

Keywords *: forecasting, attendance, prediction, regression, optimization*

I. INTRODUCTION

Accurate forecasting of event attendance is essential for the successful planning and execution of large-scale public gatherings. Whether it involves concerts, sports competitions, festivals, political rallies, or exhibitions, predicting attendance levels has direct implications for operational efficiency, safety planning, resource allocation, and the overall attendee experience. Overestimating attendance can lead to underutilized venues, wasted resources, and financial losses, whereas underestimating can cause overcrowding, logistical failure, and potential safety hazards [1].

Traditionally, event organizers have relied on qualitative judgment and past experience to estimate attendance. These methods, however, are often subjective, inconsistent, and difficult to scale especially when dealing with novel events, changing audience behaviors, or external shocks (e.g., pandemics, extreme weather, or socio-political changes) [2]. As events become increasingly complex and data-driven decision-making becomes a norm across industries, there is growing demand for more robust and automated forecasting systems.

With the advancement of machine learning (ML) and the increasing availability of public data from ticketing platforms, event listings, social media, and transportation networks, it is now feasible to adopt predictive analytics for attendance forecasting. These methods can leverage structured event metadata such as event title, date, time, category, location, venue capacity, and historical context to identify underlying patterns and make informed predictions about future attendance [3].

Several studies have shown promising results in this domain. For instance, Li et al. developed a deep learning model to predict stadium attendance for football matches, incorporating both historical attendance and contextual features such as time, venue, and competing events [4]. Similarly, Ahmed et al. proposed a hybrid machine learning framework combining metadata and social signals for concert attendance prediction, demonstrating improved accuracy over traditional regression models [5]. Another study by Liu et al. employed gradient boosting trees to forecast demand for citywide public events, emphasizing the importance of location and weather variables [6].

Despite these advances, many existing models are either domain-specific or rely heavily on proprietary or real-time features (e.g., social media buzz, ticket prices), limiting their applicability across broader contexts. This research proposes a simpler yet scalable approach: to investigate how effectively attendance can be predicted using only basic, publicly available event metadata. The core research questions are: (1) How accurate are different machine learning models in forecasting attendance using minimal inputs? (2) Which features contribute most to the predictive power? and (3) How does model performance vary across event types, locations, and timeframes?

By answering these questions, this study contributes toward developing generalizable, low-cost predictive tools that can be applied across various domains, especially in

contexts where rich historical or behavioral data is unavailable. The findings can benefit city governments, event organizers, ticketing platforms, and public safety agencies by improving planning precision and minimizing uncertainty. This study aims to compare the performance of several machine learning models in predicting event attendance, and contributes a large-scale benchmark using public datasets for practical deployment in event planning systems.

II. RESEARCH METHODS

This research aims to develop and compare various regression models to predict attendance at large-scale events based on contextual features such as time, location, and event category. The approach used is quantitative and based on historical data available in tabular form. The methodological process involves several stages, as follows:

2.1 Data Collection and Preprocessing

This study uses the Parks' Special Events dataset, which is publicly available through Data.gov and provided by the U.S. Department of the Interior. The dataset was selected due to its availability, coverage of diverse event types, and inclusion of key metadata fields. However, it does not include real-time user behavior data or promotional campaign variables, which may affect attendance in real-world scenarios.

The dataset contains curated information about one-time special events facilitated by NYC Parks' Public Programs division. The dataset was cleaned by removing rows with empty or null Attendance values. Entries with invalid time values were also filtered. Time features such as Month, DayOfWeek, Hour, and weekend indicator (IsWeekend) were extracted from the date column. Categorical features such as location, event category, and event type were encoded into numeric form using one-hot encoding, resulting in a numeric dataset ready for use in model training.

2.2 Exploratory Data Analysis

Data exploration was conducted to understand emerging patterns in event attendance based on time, location, and category. Several visualizations were used, such as histograms of attendance distribution, boxplots by day of the week, event location (borough), and popular event categories. This analysis provides an initial overview of how these variables influence attendance.

2.3 Feature Engineering and Encoding

After data exploration, feature engineering was performed to prepare the data for input into the machine learning model. Categorical variables such as DayOfWeek, Borough, and Category were converted into numeric representations using one-hot encoding. The binary variable IsWeekend was also converted to values 0 and 1.

Furthermore, a Pearson correlation analysis was performed to identify the extent to which each feature correlated with the target Attendance variable. The correlation visualization results were presented in the form of a heatmap to aid in selecting relevant features for the modeling process.

2.4 Prediction Model Development

To build an event attendance prediction system, ten different regression algorithms were used to compare performance and identify the most appropriate model for ticket and capacity prediction. Model selection was based on the diversity of approaches they represent, ranging from simple linear models to complex ensemble-based and nonlinear algorithms. The following is a list of the models used, along with their explanations:

1) Linear Regression

Linear Regression is the most basic regression method, assuming a linear relationship between features and the target. This model is often used as a baseline due to its high interpretability [7].

2) Ridge Regression

Ridge Regression adds an L2 penalty to the linear regression loss function, which is useful for addressing multicollinearity and overfitting [8].

3) Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) uses L1 regularization to produce a slimmer model with automatic feature selection [9].

4) ElasticNet Regression

ElasticNet combines L1 and L2 penalties, making it suitable for use when there is correlation between features and the model requires double regularization [10].

5) Decision Tree Regression

This model divides the data based on features that maximize information and is very effective at capturing non-linear relationships without the need for feature transformation [11].

6) Random Forest Regression

Random Forest is an ensemble of many decision trees trained on subsets of data and features, resulting in a model that is robust against overfitting [12].

7) Gradient Boosting Regressor

Gradient Boosting Regressor gradually builds a predictive model by minimizing the error of the previous model. This technique has proven highly efficient in various machine learning competitions [13].

8) XGBoost Regression

XGBoost is a sophisticated implementation of Gradient Boosting Regressor optimized for computational efficiency and additional regularization, making it particularly superior in big data scenarios [14].

9) Support Vector Regression (SVR)

SVR works by finding the optimum margin in a higher-order feature space and is known to handle cases with

significant noise or outliers [15].

10) K-Nearest Neighbors Regression (KNN)

KNN predicts a target value based on the average of its k nearest neighbors. Although simple, it can be effective if the data distribution is sufficiently dense and uniform [16].

Each model was trained on the training subset (80%) and tested on the test subset (20%). The training process was performed using default parameters from the scikit-learn or xgboost libraries, without further hyperparameter tuning, to ensure a fair and objective baseline comparison. The primary goal of this approach was to evaluate the baseline performance of each model on the event attendance prediction problem before considering further optimization. To evaluate the prediction performance, we used standard regression metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2). These metrics capture both the average magnitude of errors and the goodness of fit.

III. RESULT AND ANALYSIS

After conducting experiments on the 10 selected methods, the following results were obtained:

3.1 Mean Absolute Error (MAE) Evaluation

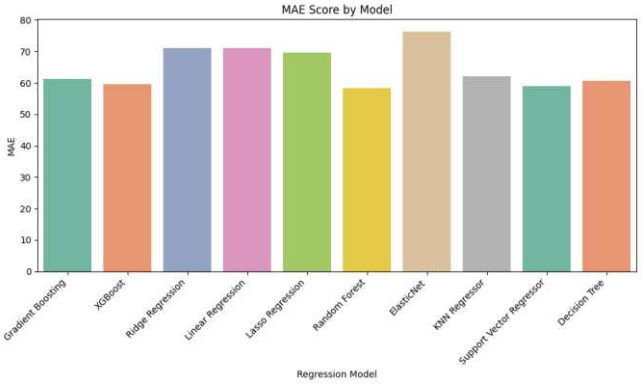

Figure 1. MAE score comparison across different regression models for event attendance prediction

Figure 1 illustrates the performance of various regression models in predicting event attendance, evaluated using the Mean Absolute Error (MAE) metric. Lower MAE values indicate better model accuracy in estimating actual attendance.

As shown in the figure, Random Forest and XGBoost achieved the lowest MAE scores approximately 58.7 and 59.3 respectively indicating superior predictive accuracy. These models are both ensemble-based tree learners that can capture non-linear relationships and complex feature interactions effectively. Their dominant performance suggests that ensemble techniques are well-suited for attendance forecasting, particularly when dealing with diverse event metadata such as category, venue, or scheduled time.

The Gradient Boosting Regressor and Support Vector Regressor (SVR) also demonstrated relatively strong performance, with MAE values in the range of 61–62. These models, while slightly less accurate than Random Forest and XGBoost, still manage to capture non-linearity in the data. SVR, in particular, is known for its robustness against outliers and overfitting in high-dimensional settings.

On the other hand, traditional linear models such as Linear Regression, Ridge Regression, and Lasso Regression performed moderately, with MAE values between 70–72. These results indicate that linear methods may struggle to model the complexities in attendance patterns, especially when important non-linear effects or feature interactions are present.

The ElasticNet model produced the highest MAE, approximately 76, making it the least accurate among the evaluated models. This could be due to the double-penalty mechanism (L1 and L2 regularization), which may oversimplify relationships between features and outcomes when applied to sparse or non-linearly distributed data.

Overall, the trend observed across models shows a clear performance gap between linear and non-linear approaches. Models capable of learning complex patterns, particularly tree-based ensembles, consistently outperform those relying on linear assumptions. This suggests that future work in event attendance prediction should emphasize non-linear modeling approaches, especially when using limited but structured input data.

3.2 RMSE-Based Model Evaluation

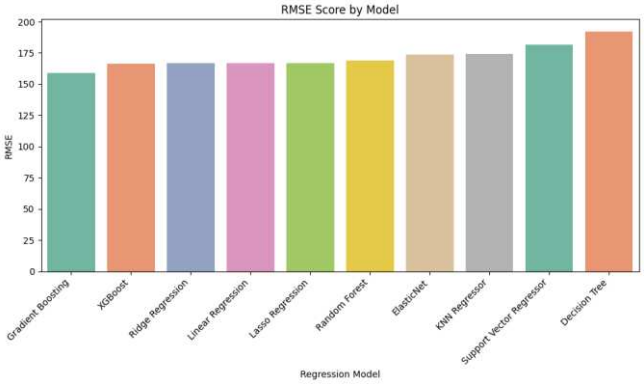

Figure 2. RMSE scores of different regression models for predicting event attendance.

Figure 2 compares the predictive performance of all evaluated regression models based on the Root Mean Squared Error (RMSE). RMSE is particularly useful for emphasizing large prediction errors due to its squared error formulation. A lower RMSE value indicates that a model generates predictions with smaller deviations from actual values, especially in high-variance scenarios.

As observed in Figure 2, the Gradient Boosting Regressor achieves the lowest RMSE (approximately 160),

underscoring its ability to minimize both minor and major deviations in prediction. This result complements its MAE-based performance and confirms the model's robustness in handling the non-linear relationships embedded in event metadata. The model's dominance across multiple evaluation metrics reinforces the findings of prior studies that favor ensemble learning in structured prediction tasks [13].

Other models with competitive RMSE values include XGBoost, Random Forest, and Lasso Regression, all ranging between approximately 165-170. This suggests a consistent performance advantage among ensemble-based models and regularized linear regressors when dealing with high-dimensional tabular data.

Conversely, the Decision Tree Regressor records the highest RMSE (approximately 193), indicating a tendency to overfit the training data while failing to generalize effectively. Support Vector Regressor and K-Nearest Neighbors Regressor also exhibit relatively poor RMSE values (approximately 180-185), signaling their sensitivity to feature scaling and potential limitations in modeling heterogeneous metadata features Interestingly, traditional linear models such as Ridge Regression, Linear Regression, and ElasticNet fall within a middle range (approximately 170-175), confirming their moderate capacity to generalize while lacking the expressiveness of more complex architectures.

A visible trend in this evaluation is that ensemble-based models consistently outperform single estimators or simpler regressors across all metrics. Moreover, models that tend to overfit or rely on distance-based assumptions (like Decision Tree and KNN) show elevated RMSE, highlighting the challenge of generalizing from sparse metadata alone. These results substantiate the earlier MAE findings and suggest that RMSE, while more sensitive to outliers, reaffirms the dominance of Gradient Boosting Regressor as the most suitable model under current feature constraints.

3.3 R^2-Based Model Evaluation

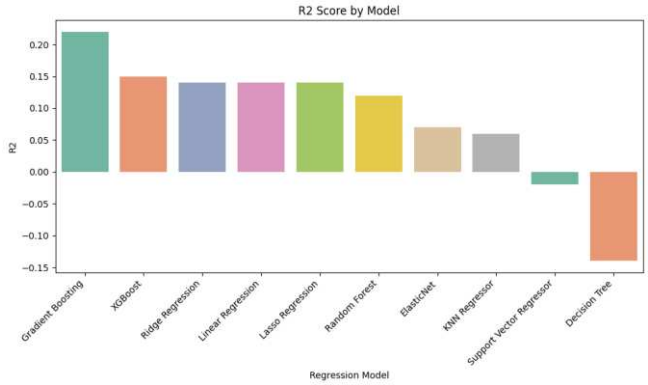

Figure 3. R^2 scores across different regression models for event attendance prediction.

Figure 3 reports the coefficient of determination (R^2) for all evaluated regression models, providing insight into their ability to explain the variance in event attendance based solely on event-level metadata. In contrast to MAE and RMSE which quantify prediction error magnitude R^2 represents the proportion of variance in the dependent variable that is accounted for by the model. A value of 1.0 denotes perfect fit, 0.0 indicates no explanatory power beyond predicting the mean, while negative values suggest worse performance than a baseline mean predictor.

Among the models tested, Gradient Boosting Regressor achieves the highest R^2 score (approximately 0.22), indicating its superior capability to extract relevant patterns from the feature set and explain approximately 22% of the variance in attendance outcomes. This performance reinforces its status as the most consistently effective model across all evaluation metrics, as evidenced by its leading MAE and RMSE scores in prior analyses.

Models such as XGBoost, Ridge Regression, Linear Regression, and Lasso Regression show relatively comparable R^2 scores (approximately 0.13-0.15). These values suggest moderate performance, potentially stemming from their reliance on linear assumptions that limit their ability to fully capture the non-linear dynamics often present in real-world attendance behavior. Random Forest exhibits slightly lower R^2 (approximately 0.12), despite competitive error-based scores (MAE/RMSE). This suggests a scenario where the model generates accurate point predictions without effectively explaining the overall variance, likely due to its ensemble architecture averaging out individual feature contributions.

On the lower end of the spectrum, models such as ElasticNet, K-Nearest Neighbors Regressor, and Support Vector Regressor yield R^2 values below 0.1, indicating minimal explanatory utility. The Decision Tree Regressor, notably, reports a negative R^2 value (approximately –0.14), which reveals that it performs worse than a naive model predicting the mean attendance across all events. This result is a clear indication of severe overfitting and lack of generalization to unseen data.

Overall, the generally low R^2 scores suggest that event metadata alone including categorical and temporal features is insufficient to model the full behavioral complexity underlying attendance decisions. This limitation points to the potential benefit of incorporating additional data sources, such as ticket pricing, artist popularity, promotional activity, social sentiment, or competing events, to increase model informativeness and predictive utility.

Despite these constraints, Gradient Boosting Regressor remains the most reliable model in this study, offering the best trade-off between error minimization and variance explanation. These findings align with the broader literature on machine learning for tabular data, which consistently favors tree-based ensemble methods in scenarios involving sparse, heterogeneous, or weakly predictive input features [13].

3.4 Model Performance Evaluation

Table 1. Performance Metrics for Attendance Prediction Models

Model	MAE	RMSE	R2
Grad Boost Reg	61.37	158.71	0.22
XGBoost	59.52	166.24	0.15
Ridge Reg	71.10	166.86	0.14
Linear Reg	71.12	166.87	0.14
Lasso Reg	69.57	167.01	0.14
Random Forest	58.31	168.97	0.12
ElasticNet	76.42	173.92	0.07
KNN Regressor	62.11	174.09	0.06
SVR	59.05	181.75	-0.02
Decision Tree	60.66	192.32	-0.14

The Gradient Boosting Regressor outperformed other models across multiple metrics, achieving the highest R^2 score (0.22), indicating that it explains 22% of the variance in the attendance data. It also achieved competitive MAE (61.37) and RMSE (158.71), showing better accuracy and error minimization compared to the rest. Interestingly, while Random Forest and XGBoost had slightly lower MAE values (58.31 and 59.52, respectively), their R^2 values were lower (0.12 and 0.15), suggesting that although their average prediction errors were smaller, they were less effective at capturing overall variance.

Classical linear models such as Linear Regression, Ridge, and Lasso all yielded similar performances, with R^2 around 0.14 and RMSE above 166. This implies that these models have limited ability to model complex nonlinear patterns in the data, likely due to the simplicity of the features (e.g., date, location, and category).

Models such as ElasticNet, K-Nearest Neighbors (KNN), and Support Vector Regressor (SVR) demonstrated weaker performance, with lower R^2 scores and higher RMSE. The SVR and Decision Tree models, in particular, showed negative R^2 values (–0.02 and –0.14), indicating poor generalization and worse performance than a naive mean predictor. Overall, ensemble-based models (Gradient Boosting Regressor, XGBoost, and Random Forest) showed superior performance compared to individual regressors, supporting findings from prior studies which emphasize their robustness in handling heterogeneous and sparse feature spaces [14][15].

The findings highlight the practical advantages of utilizing Gradient Boosting Regressor in predicting event attendance using metadata features. With the highest R^2 score among all tested models, Gradient Boosting Regressor demonstrates a superior capability to learn from complex patterns, non-linear relationships, and subtle feature interactions—characteristics often inherent in real-world event data.

From an operational perspective, this insight is especially valuable for event organizers, ticketing platforms, and digital marketing teams. By deploying Gradient Boosting Regressor based predictive models early in the event lifecycle, stakeholders can make data-informed decisions on resource allocation, marketing budget distribution, and ticket inventory management. For instance, events flagged as high-demand by the model can trigger earlier promotional campaigns or dynamic pricing strategies.

Additionally, the consistently poor performance of simpler models such as k-Nearest Neighbors or Decision Tree regressors underscores the importance of model selection in production environments. While those models offer ease of interpretation and lower computational cost, their inability to generalize well to the event attendance problem renders them less suitable for practical deployment. Furthermore, the performance gap between Gradient Boosting Regressor and linear models (like Ridge and Lasso) implies that non-linear modeling approaches should be prioritized in domains where feature interactions and saturation effects are likely such as when predicting human behavior influenced by time, location, pricing tiers, and genre tags.

VI. CONCLUSION

This study investigated the feasibility of predicting event attendance using machine learning models trained on publicly available event metadata. The models were evaluated based on three primary performance metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the Coefficient of Determination (R^2). Our findings reveal that among the tested models, Gradient Boosting Regressor and XGBoost achieved the best overall performance, with MAE scores of 61.37 and 59.52, and R^2 scores of 0.22 and 0.15, respectively. These results suggest that ensemble-based models are more effective at capturing the underlying patterns in event attendance data, likely due to their ability to model complex nonlinear relationships.

Traditional linear models, including Ridge, Lasso, and ElasticNet regression, yielded significantly lower R^2 values (around 0.14 or below), indicating limited capacity to explain the variance in attendance based solely on basic event metadata. Similarly, non-parametric models like K-Nearest Neighbors and Decision Trees exhibited suboptimal performance, further reinforcing the importance of advanced ensemble methods in this context. Despite the modest R^2 scores across all models, the relatively low MAE indicates that the models were still able to generate reasonably accurate point estimates of attendance. However, the overall predictive performance also reflects the limitations of using metadata alone (e.g., event date, location, category) without incorporating richer contextual features such as ticket price, social media engagement, weather conditions, or promotional campaigns.

From a practical standpoint, the findings imply that organizers can rely on ensemble-based models like Gradient Boosting to obtain baseline attendance forecasts using minimal input features, particularly when real-time or granular data is unavailable. This approach can support

preliminary decision-making in event logistics, including early resource allocation, crowd management strategies, and staff scheduling.

Nevertheless, the study has some limitations. The current model is trained exclusively on structured metadata without accounting for external or dynamic factors that may strongly influence attendance. Consequently, the relatively low R^2 values suggest that much of the variance remains unexplained. Future work should incorporate richer contextual and temporal information such as real-time weather, social sentiment, pricing strategies, and historical attendance trends to improve predictive accuracy and generalizability.

Additionally, model transparency and fairness should be considered, particularly if such predictive tools are to be deployed in high-stakes settings like public safety planning or ticket pricing optimization. Future studies could enhance prediction accuracy by integrating richer contextual data, such as weather, event marketing efforts, or historical attendance trends.

THANK-YOU NOTE

The authors would like to express their deepest gratitude to Universitas Amikom Yogyakarta for the continuous support, research facilities, and academic environment that enabled the successful completion of this study. Special thanks also go to the Department of Information Systems, Faculty of Computer Science, for their guidance, technical input, and encouragement throughout the research process. We extend our sincere appreciation to the U.S. Department of the Interior for providing access to valuable open data through the Parks & Special Events Dataset, which significantly contributed to our modeling and analysis of public event attendance patterns.

We also acknowledge the contributions of fellow researchers, academic peers, and anonymous reviewers whose constructive feedback helped improve the quality and clarity of this paper. Furthermore, we are grateful to the broader data science and open-source community whose tools, libraries, and platforms greatly facilitated the data processing, modeling, and visualization tasks throughout this research. The availability of reproducible and well-documented software frameworks was instrumental in ensuring the robustness and transparency of our findings.

REFERENCES

[1] A. Smith and R. Stewart, "Attendance demand: Past, present and future," *Sport Management Review*, vol. 2, no. 1, pp. 13–33, 1999.
[2] G. Zaman, M. A. Shah, and A. Wahid, "Crowd estimation for large-scale events using machine learning techniques: A review," *IEEE Access*, vol. 8, pp. 197503–197520, 2020.
[3] T. Petukhina et al., "Event attendance prediction using machine learning and heterogeneous data sources," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2497–2505.
Y. Li, B. Yu, and H. Gao, "Forecasting sports attendance using neural networks: A study of football matches," *Knowledge-Based Systems*, vol. 191, 2020.
[4] Y . Li, B. Yu, and H. Gao, "Forecasting sports attendance using neural networks: A study of football matches," *Knowledge-Based Systems*, vol. 191, 2020.
[5] M. Ahmed, F. A. Khan, and H. Malik, "Ensemble-based attendance prediction for live music events using hybrid data sources," in *Proceedings of the 2022 ACM Web Conference (WWW '22)*, pp. 1856–1865.
[6] Y. Liu, J. Chen, and X. Ma, "Predicting public event attendance with urban sensing and gradient boosted models," *Journal of Big Data*, vol. 6, no. 1, 2019.
[7] A. Montgomery, G. J. G. et al., "Linear regression analysis," *Journal of Statistical Software*, vol. 8, no. 2, pp. 1–20, 2003.
[8] H. Hoerl and R. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.
[9] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.
[10] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.
[11] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*. Wadsworth, Belmont, CA, 1984.
[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
[13] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[14] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.
[15] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, no. 3, pp. 199–222, 2004.
[16] C. Zhang and Y. Ma, *Ensemble Machine Learning: Methods and Applications*, Springer, 2012, ch. 3 (KNN regression).