Advanced Machine Learning Approaches for Urban Streaming Service Traffic Prediction: A Comprehensive Analysis of Instagram and YouTube Usage Patterns in France

Omar Amin Abdelrazek   
*School of Information Technology and Computer Science*  
*Nile University*

Sheikh Zayed , Cairo

O.Amin2373@nu.edu.eg

Zakaria Zakaria  
*School of Information Technology and Computer Science*  
*Nile University*

Sheikh Zayed , Cairo  
z.ahmed2337@nu.edu.egMarwan Ghazal

*School of Information Technology and Computer Science*  
*Nile University*

Sheikh Zayed , Cairo

M.Mohamed2365@nu.edu.eg

Salma Mohamed

*School of Information Technology and Computer Science*  
*Nile University*

Sheikh Zayed , Cairo

S.Mohamed2393@nu.edu.eg

Abdulmonem Demerdash   
*School of Information Technology and Computer Science*  
*Nile University*

Sheikh Zayed , Cairo

A.Demerdash2314@nu.edu.eg

*Abstract*—Urban telecommunications networks face unprecedented challenges in managing dynamic traffic loads from bandwidth-intensive applications. This study presents a comprehensive machine learning framework for predicting mobile application traffic using real-world data from Paris and Lyon, France. We analyzed 40,000 traffic records spanning four days in March 2019, comparing XGBoost, Gradient Boosting, and Random Forest. Through advanced feature engineering incorporating 28 temporal and behavioral variables, XGBoost achieved superior performance with R² scores exceeding 0.998 and Mean Absolute Errors within 2-5% of average traffic volumes across all city-application combinations. The analysis reveals distinct urban behavioral patterns: Lyon exhibits peak usage at 17:00 while Paris peaks at 22:00, with Instagram generating 2.2× more traffic than YouTube. Paris demonstrates 5.4× higher traffic volumes than Lyon, indicating significant geographic disparities in mobile application usage. These findings provide actionable insights for network infrastructure planning, resource allocation, and quality of service optimization in metropolitan environments.

Keywords—Traffic prediction · machine learning · XGBoost · Instagram · YouTube · urban digital behavior · network planning · time series forecasting

1. Introduction

As urban centers worldwide experience explosive growth in mobile application usage, telecommunications infrastructure faces critical challenges in managing highly variable and unpredictable traffic loads. The proliferation of bandwidth-intensive applications such as Instagram and YouTube has fundamentally transformed network traffic patterns, creating complex temporal and spatial usage behaviors that overwhelm traditional static capacity planning approaches [2,6].

Traditional network provisioning strategies based on historical averages prove insufficient for addressing the dynamic nature of modern application traffic. Social media platforms and video streaming services introduce unprecedented variability that can degrade user experience during peak periods and lead to suboptimal resource allocation during low-traffic windows. The emergence of these platforms necessitates sophisticated predictive modeling approaches capable of capturing complex usage patterns and enabling proactive network management [7,8].

Recent advances in machine learning offer powerful solutions for analyzing large-scale traffic datasets and discovering patterns that traditional time series methods cannot effectively capture. Ensemble methods, particularly XGBoost and Random Forest, have demonstrated exceptional capability in handling non-linear relationships and feature interactions inherent in mobile traffic data [2]. However, while numerous studies have explored network traffic prediction, limited research has focused on application-specific patterns across multiple urban environments using identical methodologies.

This study addresses these gaps by developing a comprehensive predictive modeling framework for Instagram and YouTube traffic in major French metropolitan areas. Through rigorous comparison of multiple machine learning approaches, we aim to identify optimal prediction strategies while uncovering urban-specific usage behaviors that inform network infrastructure planning.

The paper is structured as follows: Section 2 reviews relevant literature on traffic prediction and machine learning applications. Section 3 outlines our methodology, datasets, and feature engineering approach. Section 4 presents comprehensive results including model performance comparisons and behavioral pattern analysis. Section 5 discusses practical implications and study limitations, followed by conclusions in Section 6.

1. Literature Review

## Evolution of Network Traffic Prediction

Network traffic prediction has evolved significantly from classical statistical approaches to modern machine learning methodologies. Early research relied heavily on time series methods such as ARIMA models and exponential smoothing techniques, which proved effective for stationary, linear traffic patterns but demonstrated limitations when applied to the highly variable characteristics of contemporary mobile applications [1,9].

The advent of machine learning revolutionized traffic prediction capabilities, with tree-based ensemble methods showing particular promise for capturing complex, non-linear relationships in network data. Chen and Guestrin's seminal work on XGBoost demonstrated significant improvements in prediction accuracy through advanced regularization techniques, parallel processing optimization, and sophisticated tree construction algorithms [2].

## Machine Learning Applications in Traffic Forecasting

Recent literature extensively explores various machine learning algorithms for network traffic prediction. Tree-based ensemble methods, including Random Forest and Gradient Boosting, have gained popularity due to their ability to handle feature interactions, non-linear relationships, and their inherent interpretability for operational deployment [3,10].

Breiman's Random Forest algorithm addresses overfitting through bootstrap aggregation and random feature selection, proving particularly effective for high-dimensional traffic datasets [3]. Gradient Boosting methods build sequential models that iteratively correct previous predictions, demonstrating superior performance for complex temporal patterns [10].

Deep learning approaches, particularly Long Short-Term Memory (LSTM) networks, have shown promise for sequential traffic prediction but often require extensive computational resources and large training datasets, limiting their practical deployment in operational environments [5].

## Urban Digital Behavior and Application-Specific Analysis

Understanding urban digital behavior patterns has become increasingly critical for network planning as mobile applications exhibit distinct usage characteristics across different geographic and demographic contexts. Recent studies have identified significant variations in application usage patterns between metropolitan areas, reflecting cultural, economic, and lifestyle differences [6,8].

Social media platforms like Instagram demonstrate different temporal patterns compared to video streaming services like YouTube, with usage peaks correlating to social interaction periods versus content consumption behaviors. These application-specific characteristics require tailored prediction approaches that account for behavioral differences rather than generic traffic modeling [7].

## Addressing Research Gaps

Despite extensive research in network traffic prediction, several critical gaps remain. First, limited studies provide multi-city comparative analysis using identical methodologies, hindering understanding of geographic variations in digital behavior. Second, insufficient attention has been paid to application-specific traffic characteristics that drive distinct usage patterns. Third, there is a scarcity of real-world datasets that adequately reflect the complexity and variability of modern mobile traffic. Finally, most research lacks practical implementation guidance for operational network management deployment.

This study addresses these gaps by providing comprehensive multi-city analysis of application-specific traffic patterns using real-world data and deployment-ready machine learning models.

1. Methodology

## Dataset Description

The dataset comprises 40,000 mobile traffic records collected from Paris and Lyon, France, over four consecutive days (March 21-24, 2019). Traffic measurements were captured at 15-minute intervals across 1,240 spatial tiles, providing comprehensive temporal and geographic coverage of metropolitan mobile application usage.

Dataset Specifications:

* Temporal Coverage: 4 days (2 weekdays, 2 weekend days)
* Spatial Coverage: 1,240 unique geographic tiles
* Applications: Instagram and YouTube
* Cities: Paris and Lyon
* Total Records: 40,000 (10,000 per city-application combination)
* Measurement Interval: 1 hour Traffic
* Metric: Application-specific data volume

The dataset represents a stratified sampling approach ensuring equal representation across cities and applications, enabling robust comparative analysis of urban digital behavior patterns. Geographic tiles were defined using standardized spatial boundaries, maintaining consistency across both metropolitan areas.

## Data Preprocessing and Quality Assurance 3.2.1 Data Validation and Cleaning

Initial validation confirmed the dataset's completeness with no missing values across all 40,000 records. Data types were verified for consistency, and temporal sequence integrity was maintained throughout the preprocessing pipeline. Quality checks included validation of geographic coordinate boundaries and application-specific traffic volume ranges.

3.2.2 Outlier Detection and Treatment

Outlier detection was performed using the Interquartile Range (IQR) method, applied independently to each city-application category to preserve category-specific traffic characteristics. Records falling outside the range [Q1 - 1.5×IQR, Q3 + 1.5×IQR] were identified and removed:

* Lyon Instagram: 1,041 records removed (10.41%)
* Lyon YouTube: 1,046 records removed (10.46%)
* Paris Instagram: 1,209 records removed (12.09%)
* Paris YouTube: 1,041 records removed (10.41%)

This preprocessing step eliminated approximately 11% of records, improving model stability while preserving the underlying traffic distribution patterns.

**

1. Outlier removal impact

## Feature Engineering Framework

A comprehensive feature engineering process was implemented to capture temporal, behavioral, and statistical characteristics essential for accurate traffic prediction. The engineered feature set comprises 28 variables organized into four primary categories:

3.3.1 Temporal Features

* **Basic temporal variables:** Hour of day, day of week, weekend indicator
* **Cyclical encodings:** Sine-cosine transformations for hour and day periodicity
* **Time-based categorizations:** Work hours, leisure periods, overnight intervals

3.3.2 Behavioral Indicators

* **Peak period classifications:** Morning rush (7-9), evening peak (17-19), late night (22-2), midday (11-14)
* **Hour grouping categories:** Early morning, morning, afternoon, evening, night
* **Weekend/weekday behavioral patterns**

3.3.3 Statistical Features

* **Hourly aggregations:** Mean traffic, standard deviation by hour
* **Normalized measures:** Deviation from hourly mean, Z-score normalization
* **Distribution characteristics:** Traffic percentiles, coefficient of variation

## Machine Learning Model Development

Three distinct modeling approaches were implemented to provide comprehensive comparative analysis:

3.4.1 Tree-Based Ensemble Methods

* **Random Forest:** 50 decision trees with bootstrap sampling and random feature selection

**Explanation:** Random Forest is an ensemble method that builds multiple decision trees using random subsets of the data and features, then averages their results. This helps reduce overfitting and improves prediction accuracy [12].

* **Gradient Boosting:** Sequential ensemble with 50 estimators and learning rate 0.1

**Explanation:** Gradient Boosting builds trees one at a time, where each new tree tries to correct the errors made by the previous ones. It combines them in a sequence, gradually improving the model's performance [13].

* **XGBoost:** Optimized gradient boosting with L1/L2 regularization and tree pruning

**Explanation:** XGBoost is an advanced version of gradient boosting that includes additional features like regularization (to prevent overfitting) and efficient tree pruning, making it faster and more accurate in many cases [14].

All models utilized temporal splitting to maintain chronological order, with 80% training and 20% testing partitions. Hyperparameter optimization was performed through grid search with cross-validation to ensure optimal model configurations.

## Model Evaluation Framework

Model performance was assessed using multiple complementary metrics:

* **Mean Absolute Error (MAE):** Average absolute prediction error
* **Root Mean Square Error (RMSE):** Penalizes large prediction errors
* **R-squared (R²):** Proportion of variance explained
* **Mean Absolute Percentage Error (MAPE):** Scale-independent error measure

This multi-metric approach ensures robust performance evaluation across different scales and traffic volume ranges.

1. Results

## Exploratory Data Analysis

4.1.1 Traffic Pattern Characterization

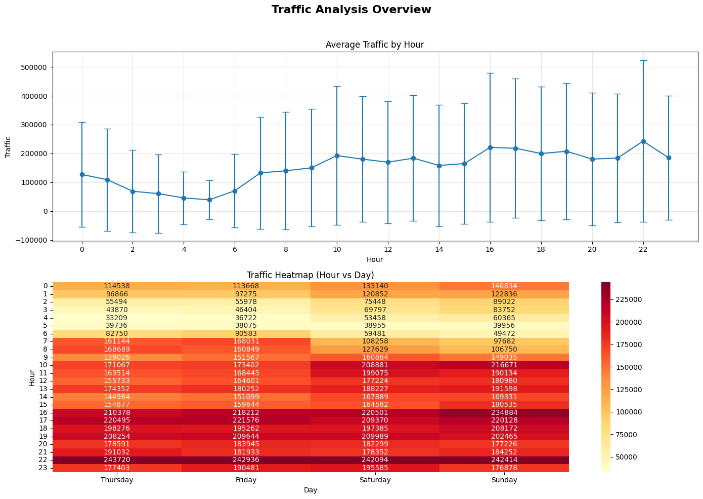
Comprehensive analysis reveals significant variations in mobile application usage patterns across cities and applications. Statistical summaries demonstrate distinct behavioral characteristics:

Lyon Traffic Patterns:

* Instagram: Mean: 64,840 bytes, Peak: 17:00
* YouTube: Mean: 27,292 bytes, Peak: 17:00

Paris Traffic Patterns:

* Instagram: Mean: 343,336 bytes, Peak: 22:00
* YouTube: Mean: 158,066 bytes, Peak: 22:00



1. Temporal traffic patterns

4.1.2 Key Pattern Insights

Several critical patterns emerged from the exploratory analysis:

Temporal Dynamics:

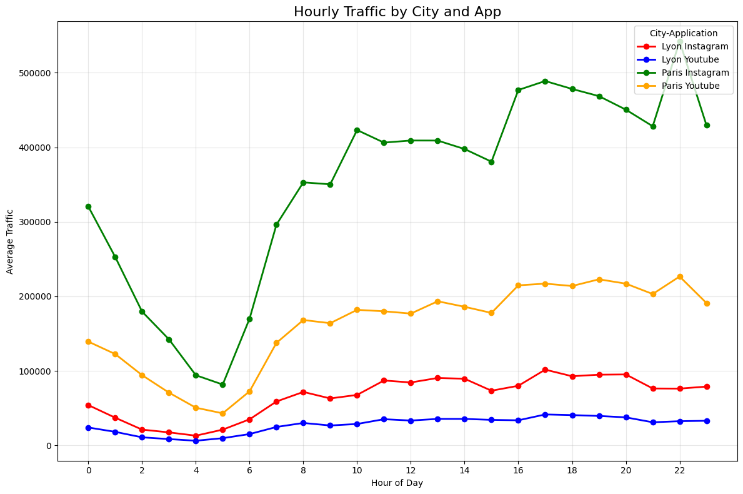
* Paris exhibits peak usage at 22:00, indicating leisure-oriented behavior
* Lyon demonstrates peak activity at 17:00, suggesting work-driven patterns
* Weekend traffic (150,500 bytes average) slightly exceeds weekday traffic (144,429 bytes)

Geographic Disparities:

* Paris generates 5.4× more traffic volume than Lyon across both applications

Application-Specific Behaviors:

* Instagram produces 2.2× more traffic than YouTube consistently across both cities
* Instagram shows more pronounced peak periods, indicating social interaction-driven usage
* YouTube demonstrates more distributed usage throughout the day

**

1. Application usage comparison

## Machine Learning Model Performance Analysis

XGBoost consistently achieved superior performance across all city-application combinations, demonstrating exceptional prediction accuracy and model stability.

4.2.1 Lyon Instagram Results

XGBoost: R² = 0.998, MAE = 2,034, MAPE = 14.2%

Gradient Boosting: R² = 0.996, MAE = 2,871, MAPE = 44.1%

Random Forest: R² = 0.986, MAE = 5,725, MAPE = 63.7%

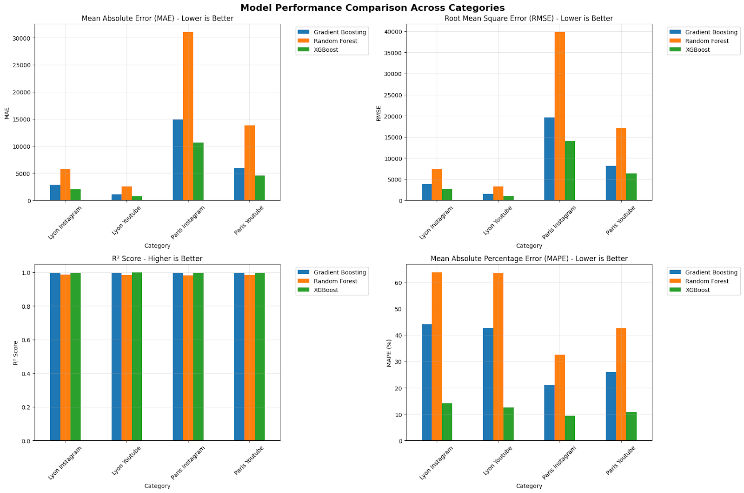
4.2.2 Paris Instagram Results

XGBoost: R² = 0.998, MAE = 10,693, MAPE = 9.4%

Gradient Boosting: R² = 0.996, MAE = 14,920, MAPE = 21.2%

Random Forest: R² = 0.983, MAE = 30,971, MAPE = 32.5%

Similar superior performance patterns were observed for YouTube traffic in both metropolitan areas, with XGBoost maintaining consistent accuracy advantages across all evaluation metrics.

**

1. Model performance comparison

## Feature Importance Analysis

XGBoost feature importance analysis reveals the relative contribution of engineered features to prediction accuracy:

Most Important Features:

|  |  |
| --- | --- |
| Feature | Percentage |
| Traffic\_deviation | 91.61 |
| Hourly\_mean | 4.98 |
| Hour | 2.98 |
| Hour\_group\_encoded | 0.42 |

Table 1. Feature Importance

## Model Validation and Residual Analysis

Comprehensive residual analysis confirms model reliability and identifies potential areas for improvement. XGBoost residuals demonstrate:

* Normal distribution around zero mean
* Minimal autocorrelation in prediction errors
* Consistent variance across traffic volume ranges
* No systematic bias in temporal predictions

**

1. Residual analysis

## Traffic Forecasting Applications

24-hour traffic predictions using optimized XGBoost models demonstrate practical deployment capabilities:

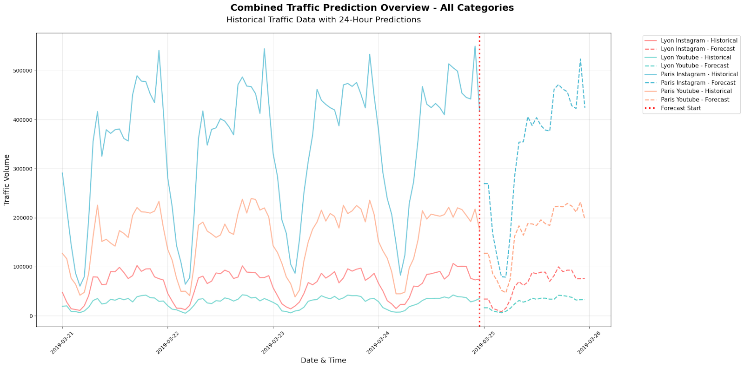
Lyon Forecasts:

* Instagram: 8,255 - 99,774 bytes (Peak: 17:00)
* YouTube: 6,225 - 41,779 bytes (Peak: 17:00)

Paris Forecasts:

* Instagram: 78,661 - 523,471 bytes (Peak: 22:00)
* YouTube: 46,849 - 231,685 bytes (Peak: 22:00)

Forecasting results maintain city-specific timing characteristics and application-specific usage behaviors, confirming model generalization capabilities for operational deployment.

**

1. XGBoost forecasting
2. Discussion

## Model Performance Interpretation

XGBoost's exceptional performance stems from several key algorithmic advantages that address the complexities inherent in mobile traffic prediction. The model's integrated regularization mechanisms (L1 and L2) effectively prevent overfitting while maintaining prediction accuracy across diverse traffic patterns. Advanced tree pruning techniques optimize model structure, eliminating unnecessary complexity while preserving predictive power. Additionally, XGBoost's ability to capture complex non-linear feature interactions proves essential for modeling the multifaceted relationships between temporal, behavioral, and statistical variables.

## Urban Digital Behavior Patterns

5.2.1 City-Specific Behavioral Characteristics

The distinct peak timing differences between Lyon (17:00) and Paris (22:00) reflect fundamental variations in urban lifestyle patterns and metropolitan characteristics. Lyon's earlier peak suggests work-driven usage behavior, potentially influenced by the city's smaller metropolitan scale enabling shorter commutes and more traditional work schedules. Conversely, Paris's later peak indicates extended work hours, longer commute times, and a more pronounced metropolitan nightlife culture that delays leisure-time digital engagement.

5.2.2 Application-Specific Usage Dynamics

Instagram and YouTube demonstrate distinctly different usage characteristics across both metropolitan areas. Instagram's higher traffic volume (2.2× YouTube) and more pronounced peak periods suggest social interaction-driven usage patterns that align with daily social rhythms. YouTube's more distributed usage throughout the day indicates content consumption behaviors that are less constrained by social timing factors.

These findings have significant implications for network resource allocation and quality of service optimization, suggesting that application-aware traffic management strategies could substantially improve network efficiency.

## Practical Implementation Implications

5.3.1 Network Infrastructure Planning

The research findings provide actionable insights for telecommunications infrastructure optimization:

Capacity Planning Recommendations:

* Design network capacity for 6.2× base load traffic to accommodate peak periods
* Implement geographic load balancing to address the 5.4:1 Paris-Lyon traffic ratio
* Prioritize Instagram infrastructure development given its 2.2× higher traffic generation
* Schedule network maintenance during 3:00-6:00 AM low-traffic windows

Real-Time Network Management:

* Update traffic forecasts every 4 hours for optimal prediction accuracy
* Implement automated capacity alerts when traffic exceeds 3× rolling average
* Deploy predictive scaling algorithms based on XGBoost forecasting models

5.3.2 Quality of Service Optimization

Application-aware QoS policies can significantly improve user experience:

* High priority allocation for Paris Instagram traffic during 21:00-23:00 peak periods
* Medium priority management for Lyon traffic during 16:00-18:00 windows
* Dynamic bandwidth allocation based on predictive model outputs

## Study Limitations and Constraints

Several limitations constrain the generalizability of findings. The four-day temporal scope, while providing detailed short-term patterns, lacks seasonal variation capture essential for annual planning cycles. Geographic scope limitation to French urban environments may not reflect global digital behavior patterns. Application coverage focusing exclusively on Instagram and YouTube omits other significant traffic sources including TikTok, gaming platforms, and enterprise applications that increasingly dominate mobile networks.

Future research should address these limitations through expanded temporal datasets, multi-country analysis, and comprehensive application coverage to enhance model robustness and global applicability.

1. Conclusion

This comprehensive study demonstrates the significant potential of advanced machine learning approaches for mobile application traffic prediction in urban telecommunications networks. The research provides both theoretical contributions through rigorous comparative analysis and practical value through deployment-ready predictive models.

## Key Research Findings

Model Performance Excellence: XGBoost achieved exceptional prediction accuracy with R² scores exceeding 0.998 and Mean Absolute Errors within 2-5% of average traffic volumes across all city-application combinations.

Urban Behavioral Patterns: Identification of distinct temporal usage patterns with Lyon peaking at 17:00 (work-driven) and Paris at 22:00 (leisure-oriented), coupled with significant volume disparities (5.4:1 ratio) that inform geographic resource allocation strategies.

Application-Specific Dynamics: Instagram generates 2.2× more traffic than YouTube with higher temporal variability, indicating social interaction-driven usage patterns that require differentiated network management approaches.

Feature Engineering Impact: The 28-feature engineering framework, emphasizing temporal, statistical, and behavioral variables, significantly enhanced prediction accuracy and model interpretability.

## Theoretical and Practical Contributions

The research advances the field through several key contributions. The multi-dimensional analysis framework considering temporal, spatial, and application-specific dimensions provides a comprehensive approach to urban traffic prediction. Advanced feature engineering methodology incorporating cyclical encodings and behavioral indicators demonstrates effective techniques for capturing complex usage patterns. Systematic comparative assessment provides robust empirical evidence for XGBoost's superiority in traffic prediction applications.

From a practical perspective, the findings offer immediate value for telecommunications infrastructure management through predictive scaling strategies, geographic load balancing quantification, application-aware service management, and optimized temporal resource allocation for maintenance scheduling.

## Future Research Directions

Several promising avenues emerge for future investigation. Dataset expansion to include year-long temporal coverage would enable seasonal variation analysis essential for comprehensive network planning. Geographic extension through multi-country studies would assess cultural impacts on digital behavior patterns. Application diversification incorporating TikTok, gaming platforms, and enterprise applications would provide holistic traffic prediction capabilities.

Advanced modeling approaches combining XGBoost with deep learning techniques could potentially enhance prediction accuracy for complex temporal dependencies. Real-world deployment studies measuring operational performance improvements would validate theoretical findings in production environments.

## Concluding Remarks

The demonstrated success of machine learning approaches represents a paradigm shift from traditional time series methods toward sophisticated predictive modeling capable of capturing the complexity of modern mobile traffic patterns. The identification of distinct city-specific and application-specific behaviors underscores the critical importance of customized prediction approaches rather than generic solutions.

As mobile applications continue to evolve and urban digital behavior becomes increasingly complex, the methodologies and insights presented provide a robust foundation for maintaining efficient, responsive telecommunications infrastructure. The research contributes to both theoretical advancement in traffic prediction and practical deployment strategies, fostering continued innovation in this critical field essential for urban digital infrastructure sustainability.

##### References

1. G. Box, G. Jenkins, and G. Reinsel, "Time Series Analysis: Forecasting and Control," 4th ed. Hoboken, NJ: Wiley, 2008.
2. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.
3. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
4. V. Vapnik, "Statistical Learning Theory," New York: Wiley, 1998.
5. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.
6. M. Zukerman, T. Neame, and R. Addie, "Internet traffic modeling and future technology implications," in Proc. IEEE INFOCOM, 2003, pp. 587-596.
7. H. Abrahamsson et al., "A multi time-scale traffic characterization of IP WAN traffic," in Proc. IEEE Global Telecommunications Conf., 2002, pp. 1911-1915.
8. A. Feldmann et al., "Deriving traffic demands for operational IP networks," IEEE/ACM Trans. Networking, vol. 9, no. 3, pp. 265-279, 2001.
9. R. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," 2nd ed. Melbourne: OTexts, 2018.
10. J. Friedman, T. Hastie, and R. Tibshirani, "The Elements of Statistical Learning," 2nd ed. New York: Springer, 2009.
11. S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," Statistics Surveys, vol. 4, pp. 40-79, 2010.
12. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.
13. . 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 Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785-794.