Available online at www.ijournalse.org

Emerging Science Journal

(ISSN: 2610-9182)

Vol. 8, No. 6, December, 2024

Comparative Analysis of ARIMA, Prophet, and Glmnet for Long Term Evolution (LTE) Base Station Traffic Forecasting

Tutun Juhana [1], Hajiar Yuliana [1*], Hendrawan [1], Iskandar [1], Yasuo Musashi [2]

[1] *School of Electrical Engineering and Informatics, Bandung Institute of Technology, Bandung 40132, Indonesia.*

[2] *Research and Education Institute for Semiconductors and Informatics, Kumamoto University, Kumamoto 860-8555, Japan.*

Abstract

This study evaluates the performance of three forecasting models—ARIMA, Prophet, and Glmnet—with the primary objective of equipping the telecommunication industry with effective tools for cellular traffic forecasting. These tools lay the foundation for efficient resource management, cost optimization, and enhanced service delivery. The study begins with dataset description and preparation, followed by the selection of traffic forecasting models, and concludes with performance evaluation based on metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). The main contribution of this research is a comprehensive comparison of the three forecasting methods, aiding practitioners and researchers in identifying the best prediction model for specific contexts. The findings reveal that Glmnet consistently outperforms ARIMA and Prophet across all categories of traffic forecasting on the selected performance metrics. Its ability to handle complex data structures, manage multicollinearity, and deliver robust and accurate predictions makes it the preferred choice for forecasting cellular network traffic in the telecommunications domain.

Keywords:
Base Station Traffic;
4G/LTE; Forecasting;
GLMnet; ARIMA;
Prophet.

Article History:

Received:	21	July	2024
Revised:	08	November	2024
Accepted:	15	November	2024
Published:	01	December	2024

1- Introduction

In recent years, network traffic prediction has emerged as a critical topic in the telecommunications industry. The exponential growth in mobile device usage and data services has posed significant challenges in managing and optimizing network capacity [1]. Accurate prediction models enable network operators to efficiently manage resources, prevent congestion, and enhance service quality for end-users [2]. Traffic forecasting has become a cornerstone of telecommunication network management and optimization, particularly in the context of Long-Term Evolution (LTE), which underpins modern mobile data communications. Understanding traffic patterns and characteristics, such as seasonality, trends, and the continuous rise in mobile cellular traffic, is essential for effective resource allocation and ensuring quality of service for users [3]. The value of traffic forecasting lies in its ability to provide insights into future network demand, allowing operators to anticipate and adapt to traffic pattern changes. By analyzing historical data and identifying trends and seasonal variations, operators can make informed decisions regarding network expansion, capacity upgrades, and resource allocation. This proactive approach mitigates network congestion, minimizes service disruptions, and enhances the overall user experience.

* **CONTACT**: hajiar.yuliana@lecture.unjani.ac.id

DOI: http://dx.doi.org/10.28991/ESJ-2024-08-06-04

Time series prediction models are among the primary approaches for forecasting network traffic. By utilizing historical data to estimate future values, time series prediction provides a robust analytical tool for managing the complex and dynamic traffic fluctuations characteristic of telecommunication networks. Traffic prediction and time series forecasting are closely related concepts, both playing critical roles in telecommunication network management, particularly within Long-Term Evolution (LTE) systems. As data traffic continues to grow in LTE networks, effective forecasting methods are essential for ensuring high-quality service for end-users. Traffic prediction, specifically aimed at estimating the volume of data expected at a given time, is a direct application of time series forecasting techniques.

In recent years, numerous time series prediction models have been developed and evaluated to address the challenges posed by fluctuating and dynamic traffic in LTE networks. Time series forecasting plays a pivotal role across various domains, including telecommunications, healthcare, and environmental studies. Within the scope of LTE base station traffic forecasting, advanced methods have been proposed alongside traditional approaches, each offering distinct advantages.

Traditional models such as Autoregressive Integrated Moving Average (ARIMA) and Exponential Smoothing (ES) remain widely used for time series forecasting. ARIMA, combining autoregressive (AR), integrated difference (I), and moving average (MA) components, is recognized for its structured modeling approach and reliable forecasting performance [4-6]. It excels in capturing linear relationships and providing accurate predictions based on historical patterns. ARIMA has proven particularly effective for applications like traffic flow prediction [7] and energy usage forecasting in LTE networks [8]. However, its limitations include challenges in handling non-linear trends and complex seasonal variations. Exponential Smoothing (ES) models, such as Holt-Winters, are also popular for their ability to capture trends and seasonality in time series data [9]. While ARIMA models are well-suited for data with strong autocorrelation and consistent seasonal patterns, ES methods offer a simpler alternative for forecasting in situations where capturing underlying trends and seasonality is sufficient.

In addition to ARIMA, the Prophet algorithm has been widely explored for time series prediction. Prophet, a robust forecasting tool, excels at predicting network traffic patterns by using an additive approach to decompose data into trend, seasonal, and holiday components. One of Prophet's key advantages is its ability to handle missing data and outliers while providing reliable predictions with minimal manual parameterization. Its flexibility allows it to accommodate complex data variations and ensures ease of use with limited customization. Studies have demonstrated Prophet's effectiveness in predicting hourly traffic changes in LTE networks, achieving high accuracy as indicated by metrics like RMSE. For instance, Prophet has successfully forecasted weekly cell uplink and downlink traffic trends, capturing consistent patterns but varying amplitudes [10]. Additionally, a study by Jain & Prasad (2020) [11] highlighted that combining Prophet with XGBoost significantly improved telecom network traffic forecasting, aiding business planning and capacity allocation. When compared to ARIMA, Prophet has been found to deliver stable and accurate predictions, particularly when integrated with advanced models like LSTM [12, 13]. Its applications extend beyond telecommunications, as evidenced by its successful use in forecasting healthcare emergency department indicators alongside ARIMA [13].

Seasonality in traffic patterns refers to predictable, repetitive variations in traffic volume occurring at regular intervals, such as hourly, daily, weekly, or monthly [14]. Understanding and accounting for these seasonal fluctuations are critical for capacity planning, enabling operators to allocate resources efficiently during peak demand periods while avoiding over-provisioning during off-peak times. Identifying and forecasting seasonal patterns optimize network resource utilization and enhance cost efficiency. Similarly, analyzing trends in traffic data is essential for long-term network planning [15]. Trends reflect the overall directional movement of traffic volume over time, capturing gradual changes driven by user behavior, technological advancements, or market dynamics. By recognizing and forecasting these trends, operators can anticipate future traffic demand, strategically plan network expansion, and implement upgrades to address evolving user needs. The continuous increase in mobile cellular traffic presents significant challenges for network operators. The proliferation of smartphones, IoT devices, and high-bandwidth applications has resulted in a consistent surge in data consumption, exerting immense pressure on existing network infrastructure [1]. Addressing this growing demand requires innovative forecasting methods and strategic planning to ensure sustainable and efficient network operations.

Accurately forecasting mobile cellular traffic growth is critical to ensuring network scalability and maintaining quality of service. Advanced forecasting methods enable operators to predict future traffic levels and implement proactive measures to scale network capacity and optimize resource utilization. In the context of LTE Base Station traffic, accurate forecasting is foundational for optimizing performance and resource allocation. As LTE technology continues to support high-speed data services and multimedia applications, the importance of precise traffic predictions becomes increasingly significant [16]. By anticipating traffic patterns and demand dynamics, operators can equip LTE Base Stations to provide seamless connectivity, high data rates, and low latency, meeting the growing demand for superior mobile broadband services.

To identify the most effective prediction method for LTE network traffic, a comparative evaluation of ARIMA, Prophet, and Glmnet is essential. Each model offers distinct advantages in handling various data characteristics, and selecting the appropriate model can significantly enhance prediction accuracy and operational efficiency. Glmnet, with its capacity to manage complex data and deliver robust predictions, has emerged as a promising tool for network traffic

forecasting in telecommunications. However, further research and evaluation are needed to confirm Glmnet's superiority in diverse scenarios and to explore the potential benefits of integrating these models to improve overall forecasting performance. A comprehensive, data-driven approach will allow the telecommunications industry to continue optimizing network operations and delivering enhanced services to users.

Despite progress in mobile communications, traffic forecasting still faces significant challenges, with many proposed models having limitations in their applicability and effectiveness. Traditional methods, such as ARIMA, and common machine learning approaches have dominated existing research, while alternative models like Prophet have seen limited exploration in this domain. Meanwhile, the Glmnet model, known for its flexibility in handling high-dimensional data and providing deeper analytical insights, has yet to be applied to mobile network traffic forecasting. This research addresses this gap by evaluating the role of Glmnet in improving the accuracy and efficiency of traffic prediction in mobile communications. By introducing Glmnet into this context, the study aims to pave the way for innovative and effective models, offering new insights into traffic management in mobile networks.

This research is particularly significant given the rapid advancements in communication technology and the growing reliance on data services in cellular networks. The increasing complexity of network usage patterns presents significant challenges for telecommunications operators in managing network capacity and maintaining optimal service quality. Accurate mobile network traffic prediction is essential for anticipating demand, preventing congestion, and optimizing resource allocation. This study offers a comprehensive comparative analysis of three widely used time series prediction methods—ARIMA, Prophet, and Glmnet—highlighting their respective strengths and limitations in the context of LTE network traffic forecasting. By providing effective tools for cellular traffic prediction, this research aims to support the telecommunications industry in achieving efficient resource management, cost optimization, and improved service delivery. Furthermore, it facilitates strategic planning, ensuring networks can adapt to the evolving demands of modern digital services.

The main contributions of this paper can be summarized as follows:

- This research provides an in-depth and comparative analysis between ARIMA, Prophet, and Glmnet models in predicting LTE network traffic. This analysis provides insight into the advantages and disadvantages of each model in the context of complex time series data.
- This research utilizes various performance metrics such as Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2) to assess the prediction accuracy and reliability of each model. This ensures a comprehensive and objective assessment of model performance.
- The focus on LTE network traffic prediction provides high practical value for network operators. The results of this research can be directly applied in the context of capacity management and network planning to improve operational efficiency and service quality.
- This research helps identify the best prediction models to apply in a given context, providing guidance for practitioners and researchers in choosing the most suitable method for their needs.
- By providing a comprehensive comparative analysis and evaluation, this research adds to the wealth of scientific literature related to time series prediction in the context of telecommunications, opening up opportunities for further research and the development of improved models.
- This study uses actual mobile network traffic data, making the results more relevant and applicable in real-life situations. This increases the validity and reliability of the research findings.

2- Related Works

Traffic forecasting for cellular mobile communication is a critical area of research, with various methods developed to predict future traffic patterns [17]. Techniques such as clustered complex echo state networks [18], support vector regression models [19], and deep learning structures [20] have demonstrated potential in enhancing resource allocation and utilization. For instance, Nayak & Singh (2021) [21] proposed methods aimed at improving prediction accuracy to optimize cellular network operations. Broadly, traffic forecasting methods can be categorized into two main types: statistical and machine learning approaches.

Statistical methods, valued for their simplicity and interpretability, include exponential smoothing models, Autoregressive Integrated Moving Average (ARIMA) models, and clustering-driven approaches [22, 23]. These methods leverage historical traffic data under the assumption that past patterns will persist. ARIMA models, in particular, excel at capturing seasonality and trends in time series data, making them suitable for predicting variations in cellular network traffic. They provide clear insights into the factors driving traffic patterns, facilitating reliable forecasting of peak intervals and future network demands.

Studies have demonstrated the effectiveness of ARIMA models in various contexts. For example, Lv et al. (2021) [24] used the ARIMA (0,1,0) model to predict peak traffic intervals in the mobile industry, achieving a mean interval

error rate of 15.5%. Similarly, Lee et al. (2020) [25] reported an average error rate of 2.84% when using ARIMA models to forecast future network throughput, contributing to improved network protocols and reduced latency. Suarez et al. (2009) [26] also highlighted the utility of time series models like ARIMA for Wi-Fi traffic modeling, emphasizing their role in efficient planning, resource allocation, and timely control.

ARIMA methods and other statistical models are relatively easy to implement and require minimal computational resources. Designed to capture linear relationships in time series data, ARIMA is particularly effective for datasets with stationary and predictable patterns. However, telecommunication network traffic often exhibits complex non-linear behaviors due to factors such as evolving user behavior, the adoption of new technologies, and dynamic environmental conditions. Consequently, ARIMA struggles to model these non-linear patterns accurately, resulting in reduced forecasting accuracy in such scenarios [24].

Additionally, ARIMA relies heavily on historical data to predict future values. In cases of abrupt changes, such as the launch of new applications or shifts in network policies, ARIMA models may fail to adapt quickly, leading to less responsive and less accurate predictions [14]. Another limitation is ARIMA's reliance on stationarity, which assumes that the statistical properties of a time series, such as mean and variance, remain constant over time. While differencing techniques can address some aspects of non-stationarity, they may not fully capture the inherent complexity of traffic data, which often exhibits pronounced trends and seasonal variations [7].

Despite their utility, ARIMA models face significant challenges in data-scarce environments and in capturing the inherently stochastic and non-linear nature of traffic flow. Hybrid models and advanced techniques, such as Long Short-Term Memory (LSTM) networks, address these limitations by offering improved accuracy and robustness, particularly in handling anomalies and non-linear patterns. Studies consistently highlight the limitations of statistical methods like ARIMA, which often assume linear relationships and struggle to adapt to rapidly changing traffic behaviors or sudden shifts in data. As a result, there is a growing preference for machine learning approaches, which are better equipped to model complex, non-linear relationships in traffic data [27].

Machine learning techniques have become indispensable in forecasting base station traffic in mobile networks. By delivering more accurate predictions, these methods significantly enhance network performance, optimize resource allocation, and improve user experiences. Techniques such as Artificial Neural Networks (ANN) and LSTM have consistently outperformed traditional statistical models like ARIMA and Simple Moving Average (SMA) in terms of error reduction and reliability. These machine learning models enable network service providers to manage resources more effectively and prepare their networks to accommodate future growth in user demand [28].

Machine learning methods have gained significant prominence in base station traffic forecasting due to their ability to capture complex patterns and non-linear relationships in data [29-31]. These methods leverage advanced algorithms to predict traffic patterns with high accuracy, enabling network operators to optimize resource allocation and enhance overall network performance [32]. Ensemble learning, a popular machine learning technique, has been widely used in traffic forecasting to improve prediction performance by combining multiple weak base models [33]. Stacking ensemble learning models, which combine multiple machine learning algorithms (e.g., GDBT, XGBoost, LightGBM), can outperform single models in terms of prediction accuracy and generalization ability. This multi-model fusion approach, based on feature selection and stacking, improves base station traffic prediction accuracy and generalization compared to single machine learning models.

Additionally, deep learning approaches, such as deep reinforcement learning and attention-based graph bi-LSTM networks, have been applied to traffic forecasting tasks, highlighting the evolution of machine learning in this domain. Deep learning models, particularly those that treat traffic volume data as tensors and use convolutional neural networks, are effective at learning both temporal and spatial dependencies in cellular traffic data [34]. In a study by Sudhakaran et al. (2020) [35], deep neural networks demonstrated their ability to effectively predict cellular traffic volumes, thereby contributing to congestion control and energy efficiency in wireless networks. Techniques such as neural networks, support vector machines, and random forests have shown promise in predicting cellular mobile traffic [36]. These methods excel in handling large volumes of data and can capture intricate patterns that may be missed by traditional statistical approaches.

In addition to the machine learning methods mentioned, the Prophet and Glmnet methods are also utilized for time series traffic forecasting. Prophet, a technique developed by Facebook in 2017, is designed to forecast time series data with seasonal patterns (daily, weekly, and yearly), as well as holiday effects. It works well with time series data that exhibits strong seasonal influences and relatively few periods of missing data. Prophet employs an additive model to decompose the time series into trend, seasonal, and holiday components, and it is particularly effective when data contains clear seasonal trends and rate changes. One of the key advantages of Prophet is its ability to handle missing data, outliers, and provide reliable predictions with minimal manual parameterization. The model is flexible, easy to use, and accommodates complex variations in data without requiring extensive customization. Prophet is especially useful

for long-term forecasting (e.g., 30-day predictions) [11]. It has been shown to outperform many other forecasting models, including ARIMA and LSTM, in terms of accuracy and error rates. When combined with other techniques like LSTM and BPNN, Prophet can further enhance forecasting accuracy and reduce errors [37].

In the study by Wang et al. (2022) [38], the Prophet algorithm effectively predicts both cell uplink and downlink traffic for the upcoming week, showing consistent trends but varying amplitudes. The Prophet algorithm is particularly effective for predicting hourly traffic changes in base stations, providing consistent results with low RMSE values around 0.03, indicating high accuracy in short-term predictions [38]. When combined with other models like GRU, Prophet can decompose traffic flow data and achieve over 90% prediction accuracy, outperforming advanced models in real-time traffic flow prediction under complex conditions [39]. Integrating Prophet with CNN-LSTM attention models allows the capture of both local and global patterns in complex time series data, surpassing state-of-the-art models in accuracy for network traffic prediction [40]. The multivariate version of Prophet (M-Prophet) has also proven effective in predicting IP backbone network traffic, aiding in network planning and anomaly detection. Despite utilizing several statistical techniques, Prophet's flexibility in handling complex and volatile data categorizes it as a machine learning method. Thus, Prophet represents an example of combining robust statistical modeling with machine learning techniques for time series forecasting [41].

In addition, the use of Glmnet (Generalized Linear Models with Elastic Net Regularization) is an alternative method that has gained attention in traffic forecasting for cellular network systems and base stations. Glmnet was developed to address the limitations of traditional regression models by incorporating regularization techniques. As a regularization method combining Lasso and Ridge regression, Glmnet is widely used in predictive modeling, particularly for its efficiency in handling high-dimensional data [42]. By integrating Glmnet into predictive modeling processes alongside techniques like LSTM and XGBoost, researchers can enhance the accuracy and robustness of base station traffic forecasting models. Regularization, a common technique in machine learning, improves model generalization, and Glmnet excels with datasets containing a large number of features, a frequent occurrence in machine learning applications. This capability is particularly important for analyzing large, complex datasets. The model-fitting process in Glmnet involves sophisticated optimization algorithms to find the optimal combination of parameters, characteristic of machine learning approaches. However, the use of Glmnet in traffic forecasting remains relatively limited, particularly for base station traffic prediction. While not directly mentioned in the references provided, machine learning techniques such as LSTM and XGBoost are commonly employed alongside Glmnet for predictive modeling tasks. The relatively limited use of Glmnet, especially in traffic forecasting for base stations, is one of the reasons for the proposed exploration of Glmnet in this study, where it will be compared with other algorithms.

The advantages of machine learning methods include their ability to capture non-linear relationships and intricate patterns in traffic data, leading to more accurate forecasts. Another benefit is that machine learning methods can adapt to changing traffic behaviors and learn from new data without relying on predefined models. In terms of scalability, machine learning algorithms can handle large datasets, making them suitable for capturing the complexities of cellular network traffic. However, machine learning methods also have some disadvantages, such as the need for extensive parameter tuning and their complexity in interpretation compared to traditional statistical techniques. Overall, both statistical and machine learning methods have their advantages and disadvantages in the context of traffic forecasting for cellular mobile communication. While statistical methods provide interpretability and simplicity, machine learning methods excel at capturing complex patterns and adapting to evolving traffic behaviors. The selection of the most appropriate method should depend on the specific characteristics of the traffic data and the desired level of forecast accuracy.

There is limited research comparing the performance of ARIMA, Prophet, and Glmnet for predicting cellular network traffic. While a small number of studies compare pairs of these methods, none compare all three. A summary of related studies is presented below.

Azari et al. (2019) [43] evaluated the performance of ARIMA and LSTM models for predicting and classifying cellular network traffic. The study found that while ARIMA serves as a solid benchmark, LSTM models generally outperform ARIMA in terms of accuracy for complex datasets. ARIMA models are effective for simpler, linear traffic patterns, while LSTM models excel at capturing more complex, non-linear patterns. Santos Escriche et al. (2023) [44] conducted a survey of various methods for cellular traffic prediction, including ARIMA and Prophet. Their comparison emphasizes the strengths and weaknesses of these models in different scenarios. ARIMA is found to be reliable for short-term predictions, whereas Prophet is more user-friendly and better suited for handling seasonality and holidays. However, Prophet may not always match ARIMA's accuracy for certain datasets. Mehri et al. (2024) [45] explored the use of live prediction algorithms for real-time forecasting of cellular network traffic, comparing the performance of ARIMA and Prophet models. Their study highlights the applicability of both models in dynamic environments, with ARIMA models being suitable for real-time predictions due to their simplicity and efficiency, while Prophet offers flexibility in incorporating external factors, though it may require more computational resources. Perifanis et al. (2023) [46] investigated the application of federated learning for predicting traffic in 5G base stations. Federated learning enables collaborative training across multiple devices without centralizing data, thus maintaining privacy.

The study compares traditional ARIMA with LSTM-based models to analyze their effectiveness in a decentralized approach. The authors find that ARIMA is effective for simpler, linear patterns in time series data but less accurate for complex, non-linear patterns, while LSTM significantly outperforms ARIMA in capturing temporal dependencies and non-linear patterns in the traffic data. Wang (2022) [47] focused on predicting mobile traffic at base stations using ARIMA and LSTM models, aiming to evaluate the effectiveness of these models in capturing temporal patterns in LTE base station traffic. The study finds that ARIMA performs well for short-term forecasting with linear trends but struggles with more complex and non-linear data patterns, while LSTM provides better accuracy for long-term forecasting and non-linear trends, demonstrating superior performance in handling the intricacies of mobile traffic data compared to ARIMA. This paper highlights the need for more advanced models, such as LSTM, to capture detailed dependencies in mobile traffic, which ARIMA could not adequately address. Siami-Namini et al. (2018) [48] evaluated ARIMA and LSTM models in forecasting time series data, emphasizing their performance differences in capturing various data patterns. They found that ARIMA is adequate for short-term linear forecasts but limited in handling non-linear patterns and long-term dependencies, while LSTM is superior in modeling non-linear patterns and long-term dependencies, providing better accuracy and robustness in complex time series forecasting tasks. Overall, the study reinforces the need for advanced neural network models like LSTM to effectively manage the complexity of real-world time series data, which ARIMA cannot fully address.

Based on the above findings, the contribution of this study is to compare the performance of ARIMA, Prophet, and Glmnet—three models—in predicting cellular network traffic, specifically for the telecommunications industry and LTE base stations.

3- Research Methodology

Figure 1 illustrates the processes involved in the methodology of this work. Traffic forecasting follows a structured approach to predict future traffic patterns based on historical data. The first step is to collect relevant historical traffic data, which is then pre-processed to ensure its quality and consistency. Next, a model is selected according to the characteristics of the data. For instance, if the data exhibits strong seasonal patterns, Prophet may be the preferred model. Alternatively, if the data has a complex structure with many variables, Glmnet may be more suitable. ARIMA, on the other hand, is effective for modeling trends and variations in the data. Once the model is selected, it is trained using the historical data, with its parameters adjusted to best fit the data. The model is then evaluated using metrics such as Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2), which provide insights into the model's performance. These metrics help identify the most appropriate model for the task. After evaluation, the best model is selected to forecast future traffic patterns. The forecasted values are then post-processed to ensure quality and consistency. Finally, the results are visualized to facilitate interpretation and decision-making. This structured approach ensures that the most accurate and reliable traffic forecasts are generated, supporting informed decisions regarding traffic management and infrastructure planning.

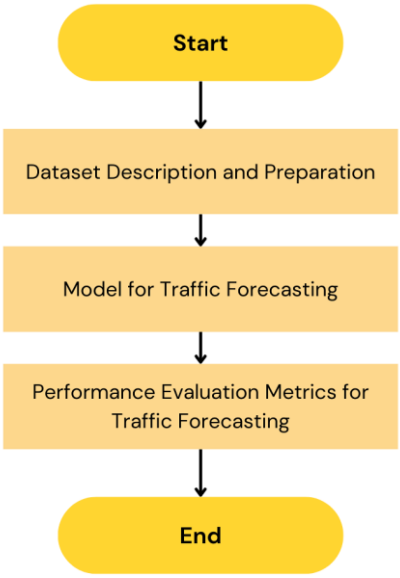

Figure 1. Flowchart of the research methodology

3-1- Dataset Description and Preparation

In this study, we compare three popular time series forecasting methods—ARIMA, Prophet, and Glmnet—for LTE base station traffic forecasting in Bandung, West Java, Indonesia. The traffic data was collected from a site of a national

telecommunications operator in Indonesia over a 30-day period, with hourly observations, resulting in a total of 721 hours of data. This dataset was used to analyze and forecast traffic patterns in the study area.

The data collection was crucial for understanding traffic flow dynamics and developing accurate predictive models. Bandung, shown in Figure 2, is a major city in West Java known for its vibrant urban environment and high traffic volume. The dataset includes insights into peak hours, off-peak hours, and seasonal variations, which are essential for effective traffic management strategies and predicting future demands. By capturing hourly traffic data over 30 days, this comprehensive dataset enables the training and evaluation of forecasting models, which can then be applied to optimize traffic infrastructure, signal timing, and other management strategies.

Figure 2. Area of the observed region

In this study, data preprocessing was carried out meticulously to ensure the quality of the data used in the forecasting models. The first step in preprocessing is data cleaning, which involves identifying and removing outliers and handling missing values. Outliers are detected using statistical methods such as the interquartile range (IQR) and z-scores. Extreme values outside a reasonable range are either removed or imputed. Missing values are addressed through imputation techniques, such as using the mean or median of the available data, or more advanced methods like interpolation.

Next, the data is normalized to ensure that all variables are on the same scale, which is crucial for algorithms like Glmnet that are sensitive to data scale. Normalization is performed using standard methods, such as min-max scaling or z-score normalization, depending on the data distribution.

In data splitting, the historical traffic data is divided into a training dataset and a test dataset using commonly used ratios such as 70:30 or 80:20. The splitting is done randomly while preserving the time distribution of the data to maintain the temporal nature of the traffic data.

The following assumptions were made during modeling:

- For the ARIMA model, it is assumed that the time series data is stationary or can be made stationary through differencing. The ARIMA (p, d, q) parameters are determined using ACF (Autocorrelation Function) and PACF (Partial Autocorrelation Function) analysis.
- For the Prophet model, it is assumed that the data includes trend, seasonality, and holiday effects, which can be modeled additively. Prophet also assumes an automatic trend change component, captured by a piecewise linear or logistic growth model.
- For the Glmnet model, the primary assumption is that the data exhibits a linear relationship between the predictor and response variables, with a penalty model (Lasso or Ridge) helping to regulate overfitting. The penalty and regularization parameters are selected through cross-validation to identify the optimal model.

The data was prepared and standardized for each of the three traffic categories (total, downlink, and uplink) to ensure uniformity and comparability between the forecasting models. A standard scale was used to standardize the traffic statistics, ensuring that the size of the data did not unduly impact the model's training. Normalization helped eliminate

the influence of data scale, allowing the models to focus on uncovering the underlying patterns. After normalization, the traffic data was decomposed into its trend and seasonal components. Time series decomposition techniques were used to separate the raw data from its overall growth trend and recurring seasonal patterns. By accounting for these components, the models were better equipped to capture the true behavior of the traffic data, leading to more accurate forecasts. Figures 3 to 5 illustrate total traffic, downlink traffic, and uplink traffic, respectively.

Figure 3 displays the total traffic data recorded over the 30-day monitoring period. This graph shows the combined volume of both uplink and downlink traffic handled by the base station. The x-axis represents time in hours, covering the entire 721-hour observation period, while the y-axis represents traffic volume in appropriate units (e.g., gigabytes or packets per hour). The plot highlights temporal fluctuations in total traffic, with periodic peaks and troughs corresponding to daily and weekly usage patterns.

Figure 4 illustrates the downlink traffic data over the same 30-day period, focusing on the volume of data transmitted from the base station to mobile devices. Like Figure 3, the x-axis represents time in hours, and the y-axis shows downlink traffic volume. The plot reveals distinct patterns in downlink traffic, indicating periods of high and low data transmission aligned with user behavior and network demand throughout the day and week.

Figure 5 presents the uplink traffic data for the 30-day observation period. This graph focuses on the volume of data sent from mobile devices to the base station. The x-axis represents time in hours, and the y-axis shows uplink traffic volume. The plot highlights variations in uplink traffic, showing periodic changes in data transmission volumes. These patterns reflect typical user behavior and network usage trends, with noticeable peaks during certain hours of the day when users are more active in uploading data.

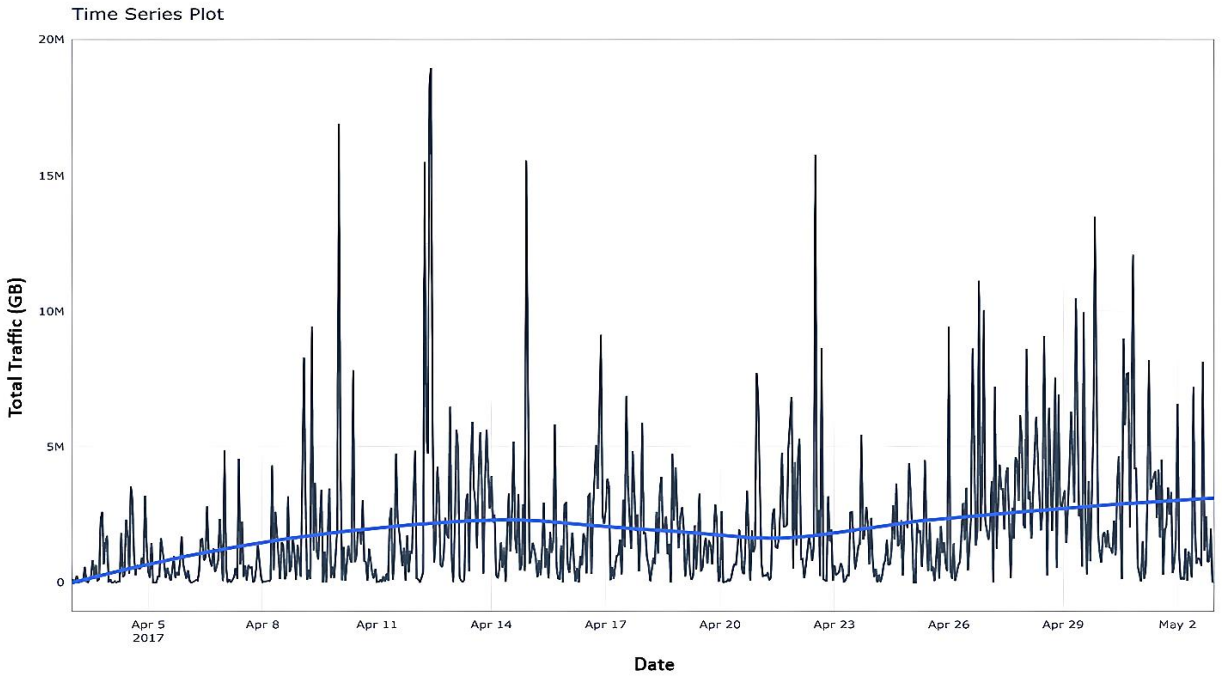

Figure 3. Total traffic data

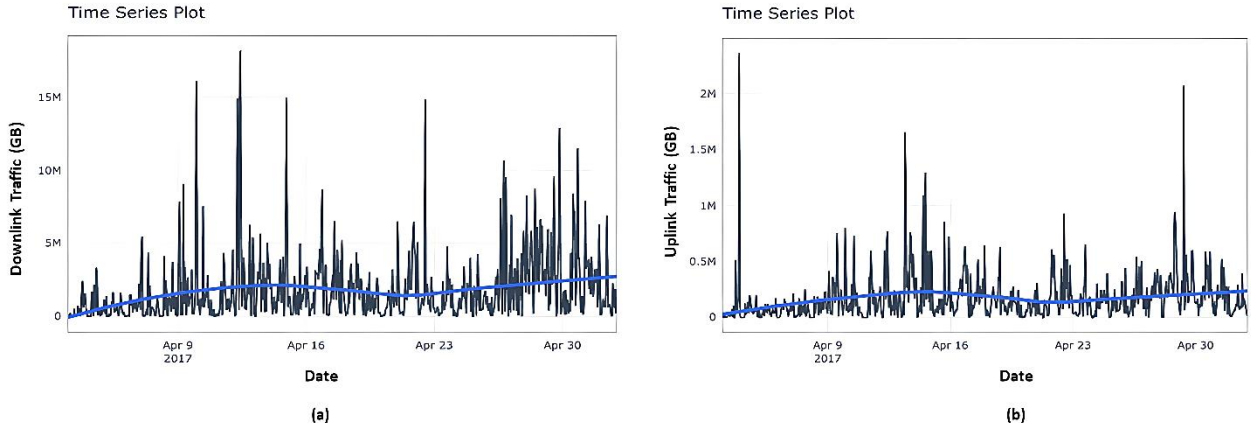

Figure 4. Downlink traffic (a) and uplink traffic (b)

The dataset includes three traffic data components: total traffic, downlink traffic, and uplink traffic. Downlink traffic measures the volume of data sent from the base station to mobile devices, while uplink traffic captures the volume of data sent from mobile devices to the base station. Total traffic represents the combined volume of both transmitted and received data by the base station.

To evaluate the forecasting models, the dataset is divided into training and testing sets. The training set consists of data from the first 27 days, while the testing set includes data from the last 3 days, as shown in Figure 5 for the total traffic dataset. The same data splitting method is applied to the downlink and uplink traffic datasets.

This approach ensures that the models are trained on a substantial portion of the data, providing them with sufficient information to learn, while the testing set remains untouched during training. This allows for an unbiased evaluation of the models' performance on unseen data. Any missing values or outliers in the data were identified and appropriately addressed. Missing values were imputed using methods such as linear interpolation or forward filling, while outliers were either removed or smoothed to avoid distorting the model training process. Additional features were created to enhance the models' predictive capabilities, including lagged variables representing previous traffic values, rolling averages, and other statistical measures. These features provided the models with more context regarding traffic patterns.

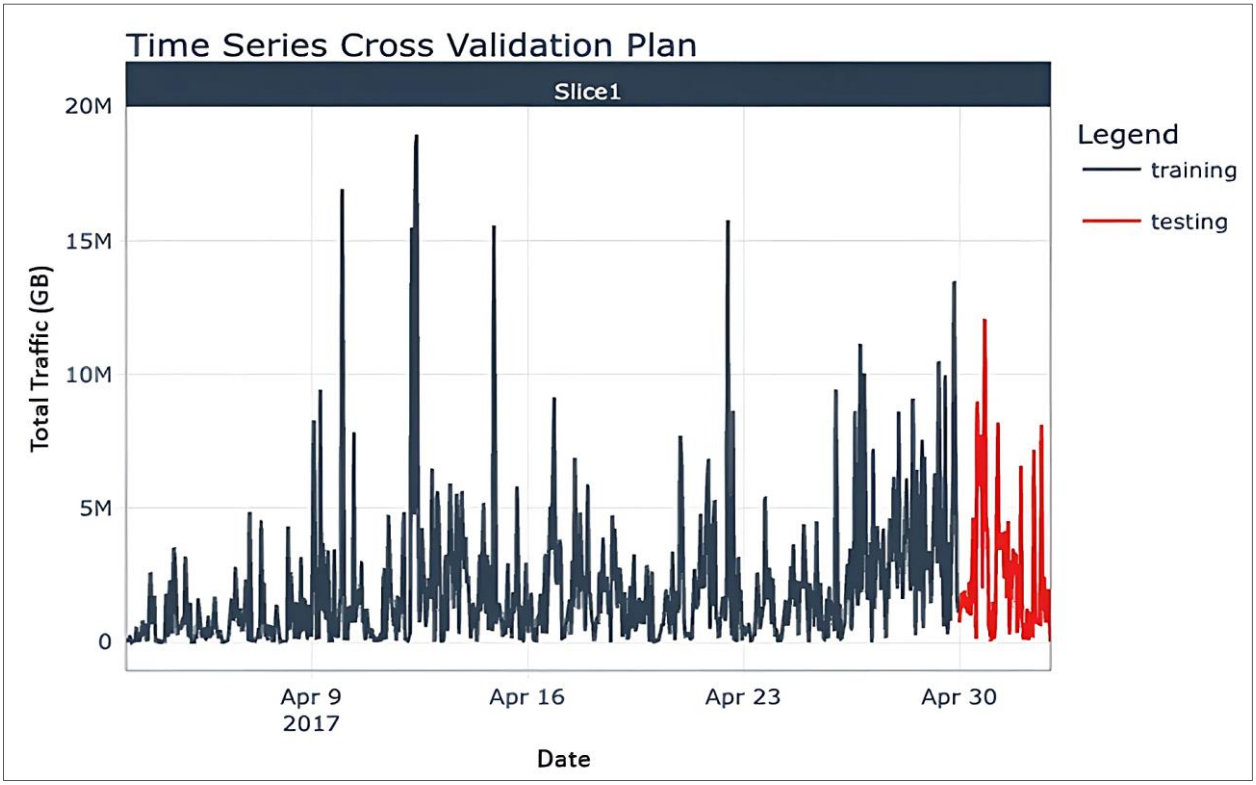

Figure 5. Splitting total traffic dataset for training and testing

Mobile cellular traffic exhibits distinct periodic patterns, including hourly, daily, and weekly cycles. These patterns are crucial for accurate forecasting, as they reflect natural fluctuations in traffic volume. Hourly traffic varies based on daily usage patterns of mobile users. Daily periodicity captures consistent changes in traffic volume throughout the day, influenced by factors such as peak hours, user behavior, and application usage. Weekly periodicity, on the other hand, reflects variations in traffic volume across different days of the week, influenced by factors like weekends, holidays, and special events.

Figures 6 to 8 show the seasonality, periodicity, and trend of the dataset. These figures decompose the traffic data into seasonal and trend components for total traffic, downlink traffic, and uplink traffic, respectively.

The time series data was tested for stationarity, and differencing was applied if necessary to make the series stationary. Stationarity refers to a time series having a constant mean and variance over time, which is a crucial assumption for many time series forecasting models. By differencing the data, trends and seasonality were removed, making the series more suitable for accurate forecasting.

Each forecasting model required specific preparations: for ARIMA, parameters such as the order of autoregressive, differencing, and moving average components were determined. For Prophet, seasonality modes and changepoints were identified and configured. For Glmnet, regularization parameters were tuned to optimize the model's performance. Analyzing these characteristics of the base station traffic provides valuable insights for predicting future traffic levels, facilitating proactive capacity management, network planning, and quality of service improvements. Additionally, this dataset was used to compare the accuracy of three time series forecasting methods: ARIMA, Prophet, and Glmnet.

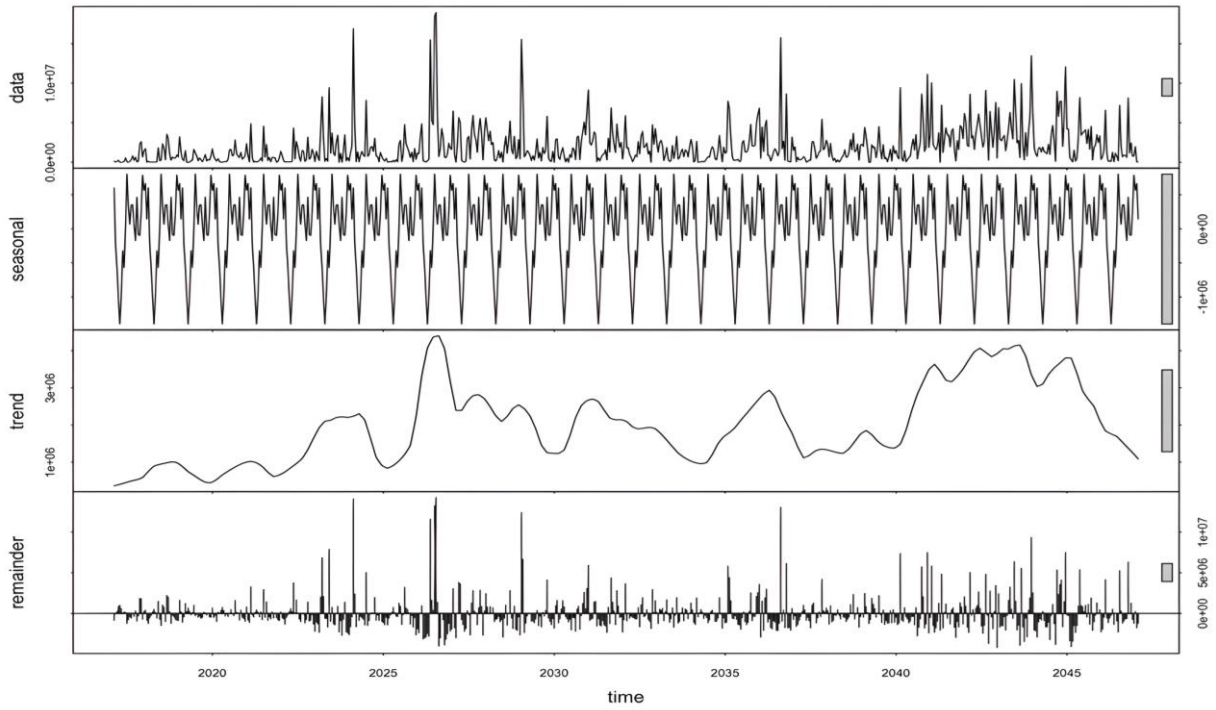

Figure 6. Decomposition total traffic data into sessional, trend a remainder

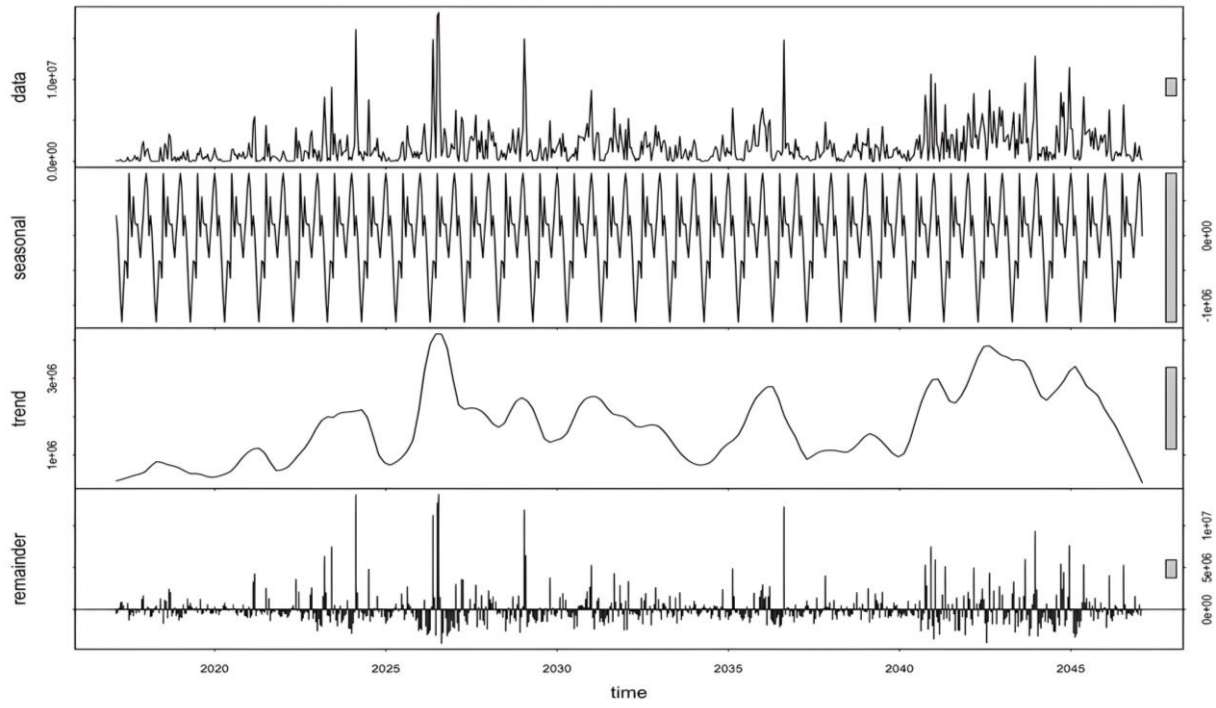

Figure 7. Decomposition downlink traffic data into sessional, trend a remainder

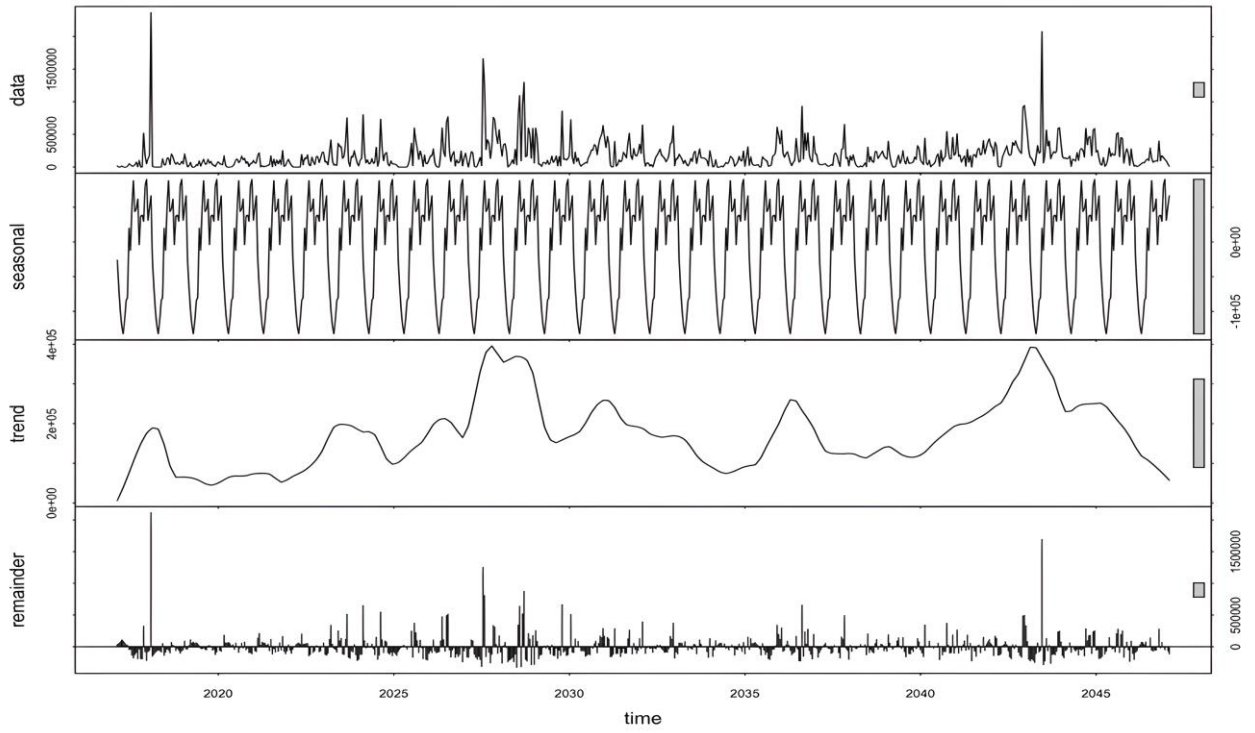

Figure 8. Decomposition uplink traffic data into sessional, trend a remainder

3-2- Models for Traffic Forecasting

Traffic forecasting in LTE base stations is crucial for efficient network management, resource allocation, and ensuring quality of service. Several models have been proposed and tested for this purpose, including traditional statistical models such as ARIMA, machine learning models like Prophet, and regularization techniques like Glmnet. This section presents a theoretical approach to traffic forecasting using these three models, emphasizing their strengths and application scenarios.

a) ARIMA (AutoRegressive Integrated Moving Average)

ARIMA (AutoRegressive Integrated Moving Average) models have long been used for time series forecasting in various domains, including telecommunications [14]. Based on the principles of time series analysis, ARIMA assumes that future values can be predicted from past observations. The model combines autoregressive (AR), differencing (I), and moving average (MA) components, making it effective for capturing linear trends and seasonal patterns in historical data. Its theoretical foundation stresses the importance of stationarity in time series data, which is crucial for accurate forecasting. While ARIMA is effective in many scenarios, it has limitations in handling non-linear patterns and sudden shifts in dynamic environments like telecommunications.

ARIMA is particularly suitable for non-stationary data, where differencing is applied to achieve stationarity. In the context of LTE base station traffic, ARIMA captures trends and seasonal patterns by adjusting three key parameters: p (autoregressive order), d (differencing order), and q (moving average order). The mathematical representation of the ARIMA model, expressed in Equation 1, illustrates how past values and errors contribute to forecasting future values. ARIMA models are well-suited for long-term forecasting, as they effectively capture traffic trends and seasonality, and have been shown to perform well in predicting peak traffic intervals and long-term traffic patterns. Additionally, ARIMA is relatively simple to implement and interpret, making it an effective tool for datasets with strong linear trends and seasonality [49].

$$Y_t = c + \sum_{i=1}^{p} \phi_i Y_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t \quad (1)$$

These models are particularly adept at capturing seasonality and trends in time series data, making them well-suited for predicting traffic variations in cellular networks. One of the key advantages of ARIMA models is their simplicity and interpretability, which allow for clear explanations of the underlying patterns and factors influencing traffic forecasts. However, ARIMA models may struggle to capture complex non-linear patterns in traffic data and may fail to adapt to rapidly changing traffic behaviors or sudden shifts in data.

ARIMA models are based on the premise that future values of a time series can be predicted by a linear combination of its past values and past forecast errors. The model has three main components:

1. The **autoregressive (AR) component**, which represents the correlation between the current time series value and its past values. It captures the relationship between the current observation and a set of previous observations.
2. The **integrated (I) component**, which involves differencing the raw observations to make the time series stationary. Stationarity refers to a constant mean and variance over time, and differencing removes trends to achieve this.
3. The **moving average (MA) component**, which accounts for the relationship between the current time series value and the residual errors from a moving average model applied to lagged observations. This component helps model the relationship between current values and past errors.

Together, these components form the ARIMA model, which is effective at capturing both linear and certain non-linear patterns in time series data. This makes ARIMA models well-suited for forecasting traffic variations in cellular networks. They provide valuable insights for addressing the seasonal, trend, and continuously increasing nature of mobile cellular traffic. However, ARIMA models may struggle with complex non-linear patterns and may not fully adapt to rapid changes in traffic behavior or sudden shifts in data.

b) Prophet

Prophet is a forecasting tool developed by Facebook that is designed to handle the challenges of forecasting data at scale [43]. The underlying theory of Prophet forecasting tool revolves around its ability to accommodate the key components of time series data – trend and seasonality. This tool is particularly tailored to address the common challenges faced in forecasting large-scale time series data, including the presence of outliers, shifts in trends, and changes in seasonality [43]. The Prophet model decomposes the time series into three main components, there are trend, seasonality, and holidays. The mathematical representation of the Prophet model can be expressed as follows in Equation 2. In this Equation 2, $Y(t)$ represents the observed value at time t, $g(t)$ denotes the trend component capturing the long-term progression of the time series, $s(t)$ represents the seasonal component, accounting for periodic fluctuations that occur at regular intervals, $h(t)$ reflects the holiday effects, which are specific to certain dates and can influence the traffic patterns, and $\in_t$ is the error term, capturing the noise or random fluctuations in the data. This equation effectively illustrates how Prophet combines these components to provide a comprehensive model for forecasting time series data, particularly in contexts where seasonality and holidays significantly impact the observed values.

$$Y(t) = g(t) + s(t) + h(t) + \in_t \quad (2)$$

Prophet is effective for short-term traffic forecasting and can handle irregularities and missing data in time series. It has been successfully applied to predict hourly traffic changes in LTE base stations. User-friendly and requiring minimal tuning, it is well-suited for real-world applications, as it effectively handles missing data and outliers [50].

Prophet introduces a flexible modeling framework that incorporates the non-linear growth trends often observed in real-world data. Using a piecewise linear model for the trend, it adapts to changing growth rates and captures the diverse growth patterns of different time series. Additionally, Prophet employs a robust approach to handle complex seasonal effects and multiple forms of seasonality, such as distinct hourly, daily, and weekly fluctuations in cellular mobile traffic.

Beyond trend and seasonality, Prophet can incorporate special events that significantly impact traffic patterns in cellular networks. This ability to account for known events enhances its forecasting capabilities, particularly for fluctuations caused by holidays, special events, or unforeseen anomalies.

One of Prophet's key strengths is its intuitive and user-friendly interface, which allows for easy parameter tuning and model interpretation. The transparency and interpretability of its models provide valuable insights into the underlying patterns and factors influencing traffic forecasts, aligning with the need for clear explanations in capacity management, network planning, and service quality enhancement.

c) Glmnet

GLMNET stands for Generalized Linear Models with Elastic Net regularization. It is a statistical modeling technique that has gained popularity for effectively handling high-dimensional data and complex relationships. Developed as an extension of traditional linear regression and logistic regression models, GLMNET addresses the limitations of these methods in managing high-dimensional datasets with multicollinearity. By combining ridge regression and lasso regression, it offers a flexible approach to variable selection and regularization [44].

The development of GLMNET was motivated by the need to improve predictive accuracy and interpretability in the context of complex data structures. The underlying theory is based on regularization, which adds a penalty term to the standard model fitting procedure. The elastic net penalty in GLMNET combines the strengths of both ridge and lasso penalties, allowing the model to select relevant variables while encouraging grouping and sparsity in coefficient estimates. This approach addresses challenges such as overfitting and multicollinearity, leading to more robust and accurate predictions for mobile cellular traffic forecasting.

GLMNET is highly effective in forecasting mobile cellular traffic, as it can capture the complex relationships and non-linear patterns often present in traffic data. It can also adapt to varying traffic behaviors and learn from new data, making it a valuable tool for dynamic forecasting in cellular networks. While GLMNET offers advanced capabilities for modeling complex relationships, it introduces some complexity in model interpretation compared to traditional statistical techniques. Nevertheless, its adaptability to changing traffic behaviors makes it a valuable asset for accurate and informed forecasting in telecommunications.

Glmnet is a machine learning technique that combines the properties of ridge regression and lasso regression. It is used for regression analysis and can handle multicollinearity in the data. The Glmnet model solves the following optimization problem, which is crucial for understanding its structure and functionality in forecasting. It shown in Equation 3, where β represents the coefficients of the model, N is the number of observations, y_i is the response variable for observation i, x_{ij} are the predictor variables for observation i and predictor j, β_0 is the intercept of the model, λ is the regularization parameter, α controls the mix of ridge and lasso penalties. Glmnet is suitable for scenarios where the traffic data has many predictors and potential multicollinearity issues. It can be used to predict traffic patterns by incorporating various features such as time of day, day of the week, and network conditions. Glmnet provides a balance between bias and variance, making it effective for complex datasets with many predictors. It can handle both linear and non-linear relationships in the data.

$$\min_{\beta}\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\left(\alpha\sum_{j=1}^{p}|\beta_j|+\frac{1-\alpha}{2}\sum_{j=1}^{p}{\beta_j}^2\right)\right\} \quad (3)$$

3-3- Performance Evaluation Metrics for Traffic Forecasting

To evaluate the performance of the models used in predicting LTE base station traffic, we employed several commonly used evaluation metrics in time series forecasting. These six metrics include MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), MASE (Mean Absolute Scaled Error), SMAPE (Symmetric Mean Absolute Percentage Error), RMSE (Root Mean Squared Error), and R^2 (Coefficient of Determination). These metrics were selected because they offer diverse insights into the quality of the forecasting models, aiding in the evaluation and optimization of predictions.

- **Mean Absolute Error (MAE)**. MAE measures the average absolute error between the actual value and the predicted value. MAE gives an idea of how large the forecasting error is in general and is expressed in the same units as the original data. MAE is also used to compare the performance of different models or algorithms in making predictions. The smaller the MAE value, the better the model or algorithm is at making predictions. MAE can be expressed as in the Equation 4, where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n_{sample} is the total number of samples.

$$MAE=\frac{1}{n_{sample}}\sum_{i=0}^{n_{sample}}|y_i-\hat{y}_i| \quad (4)$$

- **Mean Absolute Percentage Error (MAPE)**. MAPE calculates the average absolute percentage error between predicted and actual values. The smaller the MAPE value, the better the quality of the prediction model. This metric gives an idea of how much the error is relative to the actual value, making it easy to compare between datasets of different scales. It is useful for evaluating the error in percentage terms, which is important in situations where the prediction depends on the percentage error. MAPE can be expressed as in the Equation 5, where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the total number of samples.

$$MAPE=\frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y}\right| \quad (5)$$

- **Mean Absolute Scaled Error (MASE)**. MASE is a metric that measures the average absolute error produced by the model compared to the average absolute error of the model predicting the previous value SMAPE is a better version of MAPE as it is not affected by zero values and can reduce the bias caused by large values. It is useful for evaluating errors in percentages, which is more accurate and easier compared to MAPE. This metric helps assess the performance of the model compared to a simple model. MASE can be expressed as in the Equation 6, where n is the number of samples, y_i is i^{th} the actual value of i, and y_{i-1} is the previous actual value

$$MASE=\frac{1}{n-1}\sum_{i=2}^{n}|y_i-y_{i-1}|:\frac{1}{n-1}\sum_{i=2}^{n}|y_i-y_{i-1}| \quad (6)$$

- **Symmetric Mean Absolute Percentage Error (SMAPE)**. SMAPE is a variant of MAPE that is symmetric and less sensitive to small actual values. This is because SMAPE is not affected by zero values and can reduce bias caused by large values. SMAPE can be expressed as in the Equation 7, where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the total number of samples.

$$SMAPE=\frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{\frac{1}{2}(y_i+\hat{y}_i)}\right| \quad (7)$$

- **Root Mean Square Error (RMSE)**. RMSE measures the root of the mean square of the error between the actual value and the predicted value. RMSE calculates the average of the squares of the difference between the predicted and actual values. The smaller the RMSE value, the better the quality of the prediction model. It's useful for evaluating accuracy and error in prediction, which provides information on how close the predicted value is to the actual value. RMSE can be expressed as in the Equation 8, where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n_{sampel} is the total number of samples. The ML model's prediction is better when the RMSE values are less.

$$RMSE = \sqrt{\frac{1}{n_{sampel}} \sum_{i=0}^{n_{sampel}-1} |y_i - \hat{y}_i|^2} \quad (8)$$

- **R-squared (R^2)**. To assess the level of performance of the prediction models, we employed the coefficient of determination (R^2). R^2 measures the proportion of variability in the data that can be explained by the model. R^2 values range between 0 and 1, with higher values indicating a better model at explaining data variability. It is useful to evaluate how well the model can predict variations in the data, which is important to know how accurate the model is in predicting trends and variations in traffic data. R-squared can be expressed as in Equation 9.

$$R^2 = 1 - \frac{\sum_{i=1}^{i}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{i}(y_i - \bar{y}_i)^2} \quad (9)$$

where $\bar{y} = \frac{1}{i}\sum_{i=1}^{i} y_i$.

Using these evaluation metrics, we conducted a comprehensive comparative analysis to identify the most effective and efficient forecasting model for LTE base station traffic. The selected metrics—MAE, MAPE, SMAPE, RMSE, and R^2—were chosen because they offer different insights into the quality of the traffic forecasting model, aiding in performance evaluation and prediction optimization. These metrics assess absolute and percentage errors, accuracy, variation, and allow for comparison with a baseline model.

4- Result and Discussion

4-1- Total Traffic Forecast

Figure 9 illustrates the comparison between actual total traffic data and the forecasting results generated by the ARIMA, Prophet, and Glmnet models over the testing period. The x-axis represents time in hours, and the y-axis indicates the traffic volume. Each model's forecast is plotted alongside the actual traffic data to visualize their performance. The graph highlights discrepancies between predicted and actual values, providing a clear visual representation of each model's accuracy. From the plot, it is evident that Glmnet's forecast closely follows the actual traffic pattern, with smaller deviations compared to ARIMA and Prophet, indicating superior performance. This closer alignment suggests that Glmnet better captures the underlying patterns and trends in the data, as further supported by the performance metrics in Table 1.

Table 1 shows the total traffic forecast results in terms of MAE, MAPE, SMAPE, RMSE, and R^2 for ARIMA, Prophet, and Glmnet. Glmnet consistently outperforms the other models, with lower error rates and higher R^2 values, demonstrating its effectiveness in forecasting total base station traffic. Specifically, Glmnet achieves the lowest values for MAE, MAPE, SMAPE, and RMSE, indicating superior accuracy. The highest R^2 value reinforces Glmnet's better fit to the data, highlighting its ability to capture actual traffic dynamics.

This superior performance can be attributed to Glmnet's ability to handle multicollinearity and non-linear relationships in the data. While ARIMA is effective at modeling data with strong autocorrelation, it struggles with the complexities of LTE network traffic data, which exhibits seasonal patterns and non-linear trends. Prophet, though it provides better results than ARIMA in some categories, still lags behind Glmnet in prediction accuracy and the explanation of data variance. These findings have both scientific and practical implications. Scientifically, they confirm the superiority of Glmnet in complex time series prediction, especially in LTE network traffic forecasting. Practically, Glmnet's application by network operators can improve operational efficiency and capacity management, enabling more informed decision-making and better responsiveness to traffic fluctuations, which ultimately enhances the quality of service provided to end users.

Upon analyzing the forecast results using the specified performance metrics, it is clear that Glmnet outperforms both ARIMA and Prophet in forecasting accuracy. MAE and RMSE values for Glmnet are the lowest, indicating superior accuracy in predicting cellular network traffic. Glmnet's MAE is about 10% lower than Prophet's and 28% lower than ARIMA's. Its RMSE is about 7% lower than Prophet's and 16% below ARIMA's. Similarly, Glmnet demonstrates lower MAPE and SMAPE values, with MAPE being 40% lower than Prophet's and 50% lower than ARIMA's, confirming its better performance in forecasting accuracy. Although R^2 values should be interpreted cautiously in time series analysis, Glmnet exhibited the highest R^2, suggesting a better fit to the data compared to ARIMA and Prophet. Based on these performance metrics, Glmnet emerges as the most effective method for forecasting total base station traffic. Its ability to handle complex data structures, manage multicollinearity, and provide robust, accurate predictions makes it the preferred choice for telecommunications forecasting.

Table 1. Performance Evaluation of Total Traffic Forecast Results

Model	MAE	MAPE	MASE	SMAPE	RMSE	R^2
ARIMA	2,582,806	1,036	1.126	93.4	2,941,055	0.00692
PROPHET	2,240,593	970	1.10	87.1	2,710,276	0.0238
GLMNET	2,022,954	689	0.990	85.4	2,527,191	0.0892

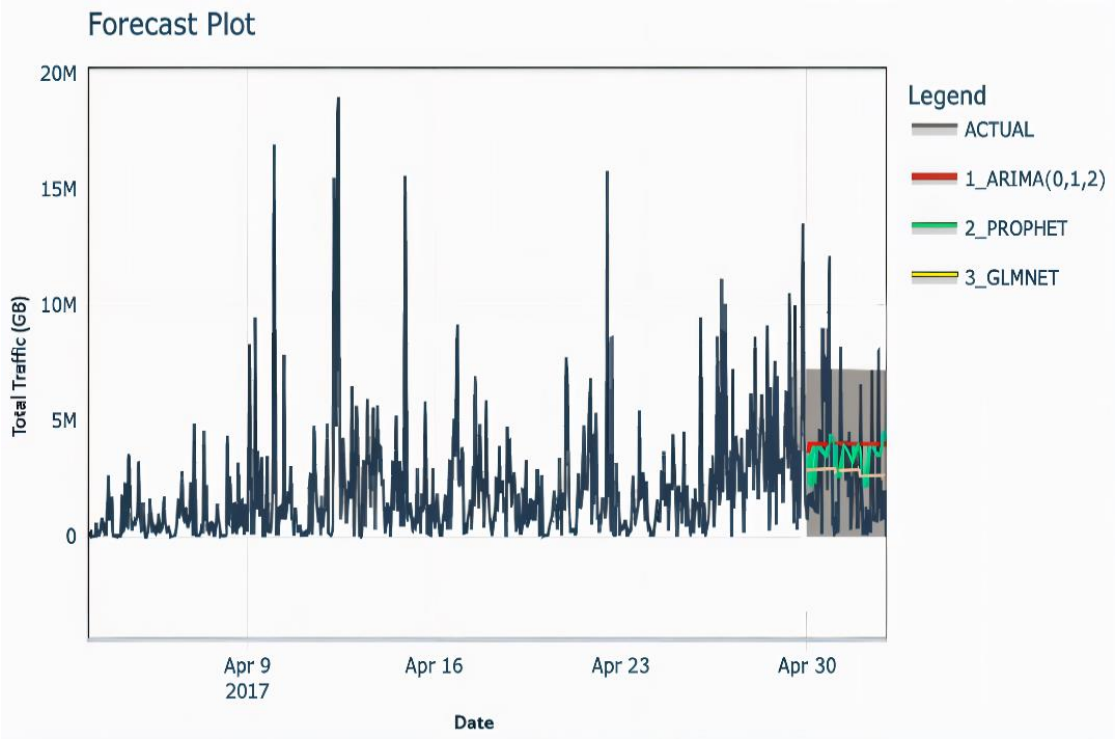

Figure 9. Actual total traffic and forecasting results

4-2- Downlink Traffic Forecast

The Downlink Traffic forecast results, in terms of Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2), for ARIMA, Prophet, and Glmnet are shown in Table 2.

Figure 10 compares the actual downlink traffic data with the forecasting results from ARIMA, Prophet, and Glmnet models. The x-axis represents time in hours, and the y-axis represents the downlink traffic volume. Similar to Figure 7, the forecasted values from each model are plotted alongside the actual downlink traffic data. This comparison is essential for evaluating how well each model predicts the volume of data transmitted from the base station to mobile devices. The Glmnet model provides a closer fit to the actual downlink traffic data than ARIMA and Prophet. The visual representation shows that Glmnet's forecast lines more closely align with the peaks and troughs of the actual data, indicating its higher accuracy. This observation is further supported by the performance metrics, with Glmnet achieving the lowest MAE, MAPE, SMAPE, and RMSE values, as well as the highest R^2 value, demonstrating a better fit and more reliable forecasts.

After analyzing the Downlink Traffic forecast results in Table 2, it is evident that Glmnet outperforms both ARIMA and Prophet in forecasting cellular network traffic. Specifically, Glmnet achieved MAE values that are 9% lower than Prophet's and 30% lower than ARIMA's, and MAPE values that are 28% lower than Prophet's and 60% lower than ARIMA's. These values signify better accuracy and performance in forecasting cellular network traffic. Furthermore, the higher R^2 value for Glmnet indicates a better fit to the data compared to the other methods. Based on the comprehensive analysis of these performance metrics, it is clear that Glmnet is the most effective method for forecasting cellular network downlink traffic among the evaluated models.

Table 2. Performance Evaluation of Downlink Traffic Forecast Results

Model	MAE	MAPE	MASE	SMAPE	RMSE	R^2
ARIMA	2,252,342	6,717	1.28	93.5	2,606,112	0.00684
PROPHET	1,885,628	5,393	1.07	85.9	2,336,109	0.0540
GLMNET	1,726,640	4,227	0.980	84.2	2,228,733	0.0840

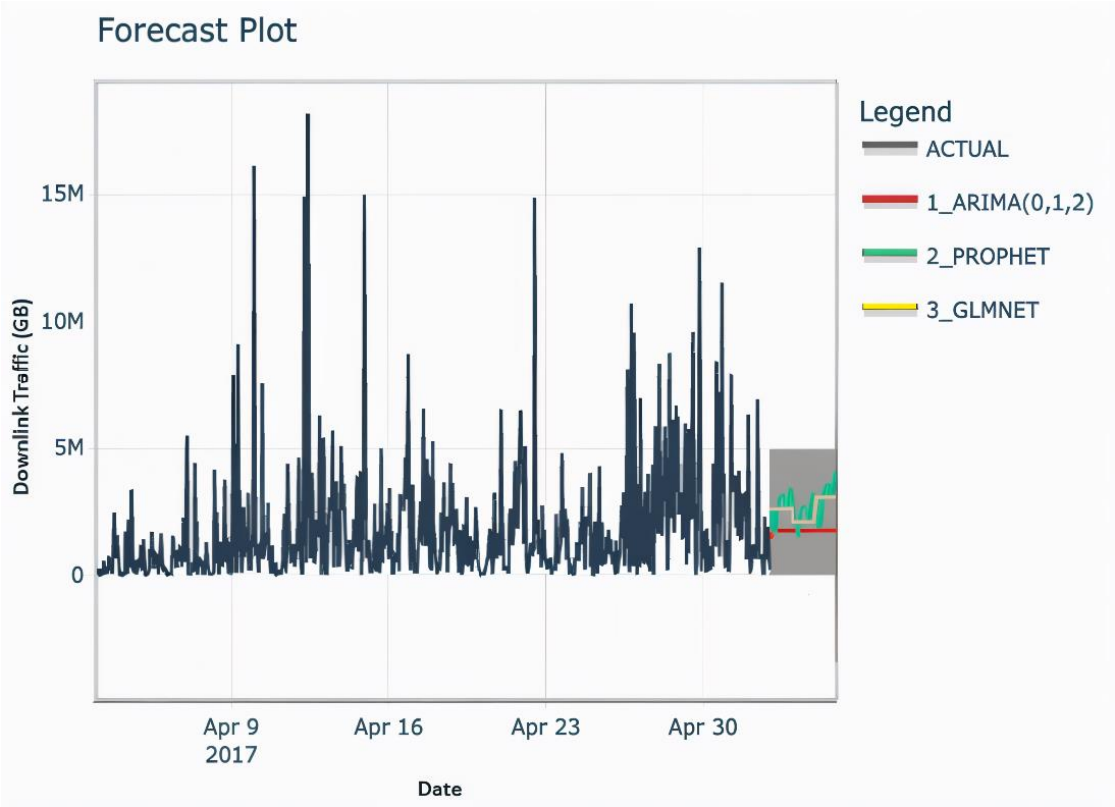

Figure 10. Actual downlink traffic and forecasting results

4-3- Uplink Traffic Forecast

Similar to the findings in Figures 9 and 10, Glmnet demonstrates superior forecasting performance for uplink traffic. The forecasted values generated by Glmnet are more closely aligned with the actual uplink traffic data, capturing the variability and trends more accurately than ARIMA and Prophet. The performance metrics in Table 3 further highlight Glmnet's effectiveness.

Figure 11 compares the actual uplink traffic data with the forecasts produced by ARIMA, Prophet, and Glmnet models. The x-axis represents time in hours, while the y-axis shows the uplink traffic volume. This comparison illustrates how well the models predict the volume of data sent from mobile devices to the base station.

The uplink traffic forecast results, in terms of Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2), are shown in Table 3. Glmnet's lower MAE, MAPE, SMAPE, and RMSE values, along with the highest R^2 value, confirm its superior ability to forecast uplink traffic. This model's better performance is consistent across all traffic types.

Upon analyzing these results, it is clear that Glmnet outperforms both ARIMA and Prophet in all performance metrics. Glmnet demonstrates lower values for MAE (21% lower than Prophet and 57% lower than ARIMA), MAPE (19% lower than Prophet and 63% lower than ARIMA), SMAPE (9% lower than Prophet and 24% lower than ARIMA), and RMSE (14% lower than Prophet and 45% lower than ARIMA). These improvements signify that Glmnet's forecasts exhibit smaller errors and better accuracy in capturing the variability of cellular network traffic. Additionally, the higher R^2 value for Glmnet further confirms its superior ability to explain the variance in the traffic data, indicating a stronger fit.

Therefore, based on the comprehensive analysis of the uplink traffic forecast results and performance metrics, we conclude that Glmnet is the most effective method for forecasting cellular network traffic among the three models evaluated.

Across all three figures and corresponding tables, Glmnet consistently outperforms ARIMA and Prophet in forecasting total, downlink, and uplink traffic. Both the visual comparisons and quantitative performance metrics demonstrate that Glmnet's ability to handle complex data structures and adapt to variations in traffic patterns makes it the most effective model for LTE base station traffic forecasting. These findings underscore the importance of selecting appropriate forecasting models to optimize network capacity management and improve service quality in telecommunications.

Table 3. Performance Evaluation of Uplink Traffic Forecast Results

Model	MAE	MAPE	MASE	SMAPE	RMSE	R^2
ARIMA	197,388	468	1.83	88.4	216,281	0.0318
PROPHET	151,945	340	1.41	77.5	170,567	0.147
GLMNET	125,882	287	1.16	71.3	149,646	0.158

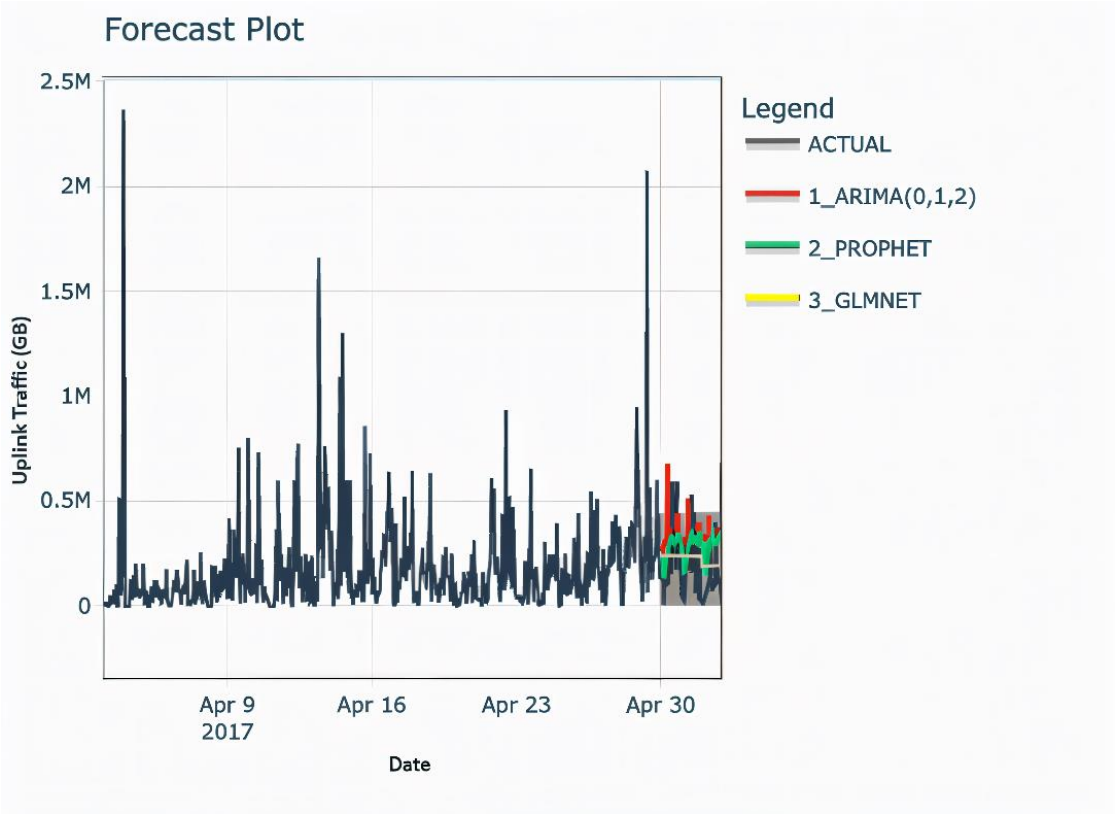

Figure 11. Actual uplink traffic and forecasting results

As previously shown, Glmnet outperforms Prophet and ARIMA, making it the most effective and accurate method for forecasting cellular network traffic. Other studies also highlight the superiority of Glmnet. For example, Garg et al. (2022) [51] examined various machine learning algorithms for time series forecasting, focusing on GLMnet and Prophet among others. Their study provides a comprehensive comparison of these models based on performance, usability, and accuracy in different forecasting scenarios. They found that Prophet is known for its simplicity and effectiveness in handling seasonality and missing data. It requires minimal tuning, making it accessible to users with limited expertise in time series forecasting. In contrast, Glmnet offers flexibility and robust regularization, making it ideal for high-dimensional datasets. Although it can achieve high accuracy, it requires extensive parameter tuning. The authors concluded that both models have their strengths: Prophet is more user-friendly, while Glmnet provides better accuracy when properly tuned.

Santos Escriche et al. (2023) [44] compared ARIMA with machine learning models, including Glmnet. They found that ARIMA is robust in handling historical traffic patterns, making it a good baseline for comparison. However, its performance diminishes as data complexity and non-linear patterns increase. On the other hand, Glmnet demonstrates strong performance with complex and high-dimensional datasets, thanks to its ability to manage multicollinearity and overfitting through regularization. The authors concluded that while ARIMA is effective for simpler, linear datasets, Glmnet offers improved accuracy and adaptability in more complex forecasting environments. Thus, Glmnet is suitable for traffic data with complex and high-dimensional datasets, but may be less effective in handling seasonality and missing data, especially if not properly tuned.

Based on the simulations and research results conducted in this study, several scenarios and conditions emerge in which the Glmnet algorithm is particularly effective, namely:

- Highly unbalanced data, where a single class or category predominates, can be problematic for Glmnet. This disparity may cause biased models to favor the dominant class, which could have the unintended consequence of having the minority class do poorly.

- Glmnet may experience multicollinearity in situations when predictor variables have substantial correlations with one another. Instable coefficients and subpar performance may result from this.
- While a large number of features can be handled by Glmnet, an excessively large number of features may cause computational problems and subpar performance.
- Non-stationarity of the data, which might arise from variations in the mean or variance over time, is not a feature of Glmnet. For these kinds of data, other models like Prophet or ARIMA would be more appropriate.
- For Glmnet to function properly, a sufficient quantity of samples are needed. Other models, such as decision trees or random forests, may perform better in situations where the sample size is very small because of their capacity to handle small datasets.
- Glmnet is mainly intended for use with generalized linear models, which work well with binary and continuous data. Other models like ARIMA or spatial autoregressive models may perform better if the data contains specialized categories like time series or spatial data.

Based on this research, we conclude that certain characteristics of traffic data make Glmnet more suitable and recommended compared to ARIMA and Prophet. Glmnet is known for its robustness and efficiency in handling large, complex datasets and optimization problems—critical factors in traffic forecasting, where data volumes can be vast and computational demands high. It is particularly effective at handling non-stationarity in traffic data, which often exhibits seasonal and trend changes. This capability allows Glmnet to capture the dynamic nature of traffic patterns, which can vary significantly over time.

Moreover, Glmnet excels in both short-term and long-term forecasting. It provides precise predictions for immediate decision-making in short-term applications, while also offering valuable insights into long-term trends and near-future scenarios. Its ability to manage highly correlated covariates, along with its robustness, efficiency, improved convergence properties, and capability to handle complex models, makes Glmnet a superior choice for traffic forecasting over ARIMA and Prophet.

5- Conclusion

In this study, we evaluated the performance of three forecasting models—ARIMA, Prophet, and Glmnet—for predicting LTE base station traffic, focusing on total, downlink, and uplink traffic. We used a dataset derived from continuous 24-hour monitoring over 30 days and analyzed the models based on several performance metrics, including Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Symmetric Mean Absolute Percentage Error (SMAPE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2).

Our findings show that Glmnet consistently outperforms both ARIMA and Prophet across all traffic categories. Specifically, Glmnet exhibited superior accuracy, as evidenced by its lower MAE, MAPE, SMAPE, and RMSE values, and higher R^2 values, indicating a better fit to the data. This performance was visually confirmed in Figures 8 to 10, where Glmnet's predictions closely matched the actual traffic data. Across total traffic, downlink traffic, and uplink traffic forecasts, Glmnet consistently showed lower errors and higher accuracy. These results highlight Glmnet's ability to generate more precise forecasts and capture the variability of cellular network traffic data. The higher R^2 value also confirms its stronger ability to explain the variance in the traffic data, making it a better fit overall.

Given the specific needs of the telecommunications industry and the characteristics of the traffic data, Glmnet proves to be an effective method for forecasting cellular network traffic. Its superior predictive performance, robustness, and efficiency support its use in telecommunications traffic forecasting. The high R^2 values, comparison with other models, and improved convergence properties all contribute to a strong recommendation for Glmnet. This model offers network operators a powerful tool for anticipating traffic demands, optimizing network capacity, and enhancing service quality.

This research highlights the importance of selecting advanced forecasting models that can adapt to the dynamic nature of telecommunications traffic. Future work could explore hybrid models that combine statistical and machine learning approaches to further enhance forecasting accuracy. Additionally, applying these models to other types of telecommunication data could offer broader insights and validate their generalizability in different contexts.

6- Declarations

6-1- Author Contributions

Conceptualization, T.J. and H.; methodology, T.J. and H.; software, H.Y.; validation, T.J. and H.; formal analysis, T.J. and H.; investigation, T.J. and H.; resources, I.; data curation, H.Y.; writing—original draft preparation, H.Y.; writing—review and editing, T.J. and H.Y.; visualization, H.Y.; supervision, T.J., H. and Y.M..; project administration, I.; funding acquisition, T.J. and I. All authors have read and agreed to the published version of the manuscript.

6-2- Data Availability Statement

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

6-3- Funding

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

6-4- Institutional Review Board Statement

Not applicable.

6-5- Informed Consent Statement

Not applicable.

6-6- Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancies have been completely observed by the authors.

7- References

[1] Thanawala, R., El-Sayed, M., Morawski, T., Mukhopadhyay, A., Zhao, J., & Urrutia-Valdés, C. (2010). The mobile data explosion and new approaches to network planning and monetization. Proceedings of 2010 14th International Telecommunications Network Strategy and Planning Symposium, Networks, Warsaw, Poland. doi:10.1109/NETWKS.2010.5624926.

[2] Jiang, W. (2022). Cellular traffic prediction with machine learning: A survey. Expert Systems with Applications, 201(April), 117163. doi:10.1016/j.eswa.2022.117163.

[3] Li, R., Zhao, Z., Zheng, J., Mei, C., Cai, Y., & Zhang, H. (2017). The Learning and Prediction of Application-Level Traffic Data in Cellular Networks. IEEE Transactions on Wireless Communications, 16(6), 3899–3912. doi:10.1109/TWC.2017.2689772.

[4] Capozzi, F., Piro, G., Grieco, L. A., Boggia, G., & Camarda, P. (2012). On accurate simulations of LTE femtocells using an open source simulator. Eurasip Journal on Wireless Communications and Networking, 1-13. doi:10.1186/1687-1499-2012-328.

[5] Zheng, Y. L., Zhang, L. P., Zhang, X. L., Wang, K., & Zheng, Y. J. (2015). Forecast model analysis for the morbidity of tuberculosis in Xinjiang, China. PLoS ONE, 10(3), 116832. doi:10.1371/journal.pone.0116832.

[6] Liu, Q., Liu, X., Jiang, B., & Yang, W. (2011). Forecasting incidence of hemorrhagic fever with renal syndrome in China using ARIMA model. BMC Infectious Diseases, 11(1), 1-7. doi:10.1186/1471-2334-11-218.

[7] Kumar, S. V., & Vanajakshi, L. (2015). Short-term traffic flow prediction using seasonal ARIMA model with limited input data. European Transport Research Review, 7(3), 1-9. doi:10.1007/s12544-015-0170-8.

[8] Hjort Kure, E. H., Engelstad, P., Maharjan, S., Gjessing, S., Zhang, X., & Zhang, Y. (2018). Energy Usage Forecasting for LTE: A Network-Wide Traffic Measurements Study. IEEE Globecom Workshops, GC Wkshps 2018 – Proceedings, Abu Dhabi, United Arab Emirates. doi:10.1109/GLOCOMW.2018.8644423.

[9] Lo Schiavo, L., Fiore, M., Gramaglia, M., Banchs, A., & Costa-Perez, X. (2022). Forecasting for Network Management with Joint Statistical Modelling and Machine Learning. In Proceedings - IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks, WoWMoM 2022, 60–69. doi:10.1109/WoWMoM54355.2022.00028.

[10] Zhu, X., Yang, Z., Liu, G., Li, Y., Xu, L., Cheng, X., & Dong, R. (2022, July). Cell Expansion Priority Recommendation Based on Prophet Algorithm. In Signal and Information Processing, Networking and Computers: Proceedings of the 8th International Conference on Signal and Information Processing, Networking and Computers (ICSINC), 1449-1457. doi:10.1007/978-981-19-3387-5_172.

[11] Jain, G., & Prasad, R. R. (2020). Machine learning, Prophet and XGBoost algorithm: Analysis of Traffic Forecasting in Telecom Networks with time series data. ICRITO 2020 - IEEE 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), 893–897. doi:10.1109/ICRITO48877.2020.9197864.

[12] Dou, R., & Meng, F. (2022). Prophet-LSTM combination model carbon trading price prediction research. Proceedings Volume 12474, Second International Symposium on Computer Technology and Information Science (ISCTIS 2022); 124741W. doi:10.1117/12.2653785.

[13] Duarte, D., & Faerman, J. (2019). Comparison of Time Series Prediction of Healthcare Emergency Department Indicators with ARIMA and Prophet, 123–133. doi:10.5121/csit.2019.91810.

[14] Kochetkova, I., Kushchazli, A., Burtseva, S., & Gorshenin, A. (2023). Short-Term Mobile Network Traffic Forecasting Using Seasonal ARIMA and Holt-Winters Models. Future Internet, 15(9), 1–15. doi:10.3390/fi15090290.

[15] Dongchen, Z., Shoufeng, W., Xiaoyan, X., Xingzheng, L., Wenwen, Y., & Tinglan, W. (2014). A novel long term traffic forecast algorithm and case study for china. Proceedings - 2014 IEEE 28th International Conference on Advanced Information Networking and Applications Workshops, IEEE WAINA 2014, 425–430. doi:10.1109/WAINA.2014.69.

[16] Gu, J., Ruan, Y., Chen, X., & Wang, C. (2012). A novel traffic capacity planning methodology for LTE radio network dimensioning. IET Conference Publications, 2011(586 CP), 462–466. doi:10.1049/cp.2011.0711.

[17] Cho, C., & Lee, S. (2012). Mobile broadband traffic forecasts in Korea. In DCNET 2012, ICE-B 2012, OPTICS 2012 - Proceedings of the International Conference on Data Communication Networking, e-Business and Optical Communication Systems, ICETE2012, 41–45. doi:10.5220/0004066200410045.

[18] Yu, L., Li, M., Jin, W., Guo, Y., Wang, Q., Yan, F., & Li, P. (2021). STEP: A Spatio-Temporal Fine-Granular User Traffic Prediction System for Cellular Networks. IEEE Transactions on Mobile Computing, 20(12), 3453–3466. doi:10.1109/TMC.2020.3001225.

[19] Rizwan, A., Arshad, K., Fioranelli, F., Imran, A., & Imran, M. A. (2018). Mobile Internet Activity Estimation and Analysis at High Granularity: SVR Model Approach. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC (Vols. 2018-September), Bologna, Italy. doi:10.1109/PIMRC.2018.8581040.

[20] Huang, C. W., & Chen, P. C. (2020). Joint demand forecasting and DQN-Based control for energy-aware mobile traffic offloading. IEEE Access, 8, 66588–66597. doi:10.1109/ACCESS.2020.2985679.

[21] Nayak, N., & R, R. S. (2021). 5G Traffic Prediction with Time Series Analysis. International Journal of Innovative Technology and Exploring Engineering, 10(12), 36–40. doi:10.35940/ijitee.l9555.10101221.

[22] Köppelová, J., & Jindrová, A. (2019). Application of exponential smoothing models and ARIMA models in time series analysis from telco area. Agris On-Line Papers in Economics and Informatics, 11(3), 73–84. doi:10.7160/aol.2019.110307.

[23] Mahdy, B., Abbas, H., Hassanein, H. S., Noureldin, A., & Abou-Zeid, H. (2020). A clustering-driven approach to predict the traffic load of mobile networks for the analysis of base stations deployment. Journal of Sensor and Actuator Networks, 9(4), 53. doi:10.3390/jsan9040053.

[24] Lv, T., Wu, Y., & Zhang, L. (2021). A Traffic Interval Prediction Method Based on ARIMA. Journal of Physics: Conference Series, 1880(1). doi:10.1088/1742-6596/1880/1/012031.

[25] Lee, D., Lee, D., Choi, M., & Lee, J. (2020). Prediction of Network Throughput using ARIMA. 2020 International Conference on Artificial Intelligence in Information and Communication, ICAIIC 2020, 1–5. doi:10.1109/ICAIIC48513.2020.9065083.

[26] Suarez, C. A. H., Parra, O. J. S., & Díaz, A. E. (2009). An ARIMA model for forecasting Wi-Fi data network traffic values. Ingenieria e Investigacion, 29(2), 65–69. doi:10.15446/ing.investig.v29n2.15163.

[27] Huang, C. W., Chiang, C. T., & Li, Q. (2017). A study of deep learning networks on mobile traffic forecasting. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC, 2017-October, 1–6. doi:10.1109/PIMRC.2017.8292737.

[28] Prajam, S., Wechtaisong, C., & Khan, A. A. (2022). Applying Machine Learning Approaches for Network Traffic Forecasting. Indian Journal of Computer Science and Engineering, 13(2), 324–335. doi:10.21817/indjcse/2022/v13i2/221302188.

[29] Zheng, J., & Huang, M. (2020). Traffic flow forecast through time series analysis based on deep learning. IEEE Access, 8, 82562–82570. doi:10.1109/ACCESS.2020.2990738.

[30] Kaur, G., Grewal, S. K., & Jain, A. (2024). Federated Learning Based Spatio-Temporal Framework for Real-Time Traffic Prediction. Wireless Personal Communications, 136(2), 849–865. doi:10.1007/s11277-024-11292-z.

[31] Yang, S., Wu, J., Du, Y., He, Y., & Chen, X. (2017). Ensemble Learning for Short-Term Traffic Prediction Based on Gradient Boosting Machine. Journal of Sensors, 2017, 1–15. doi:10.1155/2017/7074143.

[32] Wu, Q., Chen, X., Zhou, Z., Chen, L., & Zhang, J. (2021). Deep reinforcement learning with spatio-temporal traffic forecasting for data-driven base station sleep control. IEEE/ACM2021 Transactions on Networking, 29(2), 935–948. doi:10.1109/TNET.2021.3053771.

[33] Zhao, J., Qu, H., Zhao, J., Dai, H., & Jiang, D. (2020). Spatiotemporal graph convolutional recurrent networks for traffic matrix prediction. Transactions on Emerging Telecommunications Technologies, 31(11), e4056. doi:10.1002/ett.4056.

[34] Zhao, L., Huang, Y., Wang, Y., Xu, Y., Feng, Q., & Chen, E. (2023). Base Station Traffic Prediction based on Feature Selection and Stacking Ensemble Learning. ACM International Conference Proceeding Series, 113–117. doi:10.1145/3603781.3603800.

[35] Sudhakaran, S., Venkatagiri, A., Taukari, P. A., Jeganathan, A., & Muthuchidambaranathan, P. (2020). Metropolitan Cellular Traffic Prediction Using Deep Learning Techniques. 2020 IEEE International Conference on Communication, Networks and Satellite, Comnetsat 2020 - Proceedings, 106–110. doi:10.1109/Comnetsat50391.2020.9328937.

[36] Clemente, D., Soares, G., Fernandes, D., Cortesao, R., Sebastiao, P., & Ferreira, L. S. (2019). Traffic forecast in mobile networks: Classification system using machine learning. IEEE Vehicular Technology Conference, 2019-September, Honolulu, United States. doi:10.1109/VTCFall.2019.8891348.

[37] Fang, W. X., Lan, P. C., Lin, W. R., Chang, H. C., Chang, H. Y., & Wang, Y. H. (2019). Combine Facebook Prophet and LSTM with BPNN Forecasting financial markets: The Morgan Taiwan Index. Proceedings - 2019 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2019, 4–5. doi:10.1109/ISPACS48206.2019.8986377.

[38] Wang, H., Liu, S., Sun, C., Li, Z., & Jiang, X. (2022). Short term prediction of cell uplink and downlink traffic based on Prophet algorithm. ACM International Conference Proceeding Series, 419–423. doi:10.1145/3558819.3565118.

[39] Xiao, H., Zhao, Y., & Zhang, H. (2023). Predict Vessel Traffic with Weather Conditions Based on Multimodal Deep Learning. Journal of Marine Science and Engineering, 11(1), 39. doi:10.3390/jmse11010039.

[40] Tan, H. (2023). A Novel Approach to Wireless Network Traffic Prediction using CNN-LSTM Attention Model with Prophet Model. Proceedings - 2023 2nd International Conference on Machine Learning, Cloud Computing, and Intelligent Mining, MLCCIM 2023, MLCCIM 2023, 285–292. doi:10.1109/MLCCIM60412.2023.00047.

[41] Wei, X., Liu, Z., Li, M., & He, X. (2023). Application of Prophet Model in Traffic Matrix Prediction for IP Backbone Network. 2023 8th International Conference on Intelligent Computing and Signal Processing, ICSP 2023, 136–140. doi:10.1109/ICSP58490.2023.10248446.

[42] Mohammadjafari, S., Roginsky, S., Kavurmacioglu, E., Cevik, M., Ethier, J., & Bener, A. B. (2020). Machine Learning-Based Radio Coverage Prediction in Urban Environments. IEEE Transactions on Network and Service Management, 17(4), 2117–2130. doi:10.1109/TNSM.2020.3035442.

[43] Azari, A., Papapetrou, P., Denic, S., & Peters, G. (2019). Cellular Traffic Prediction and Classification: A Comparative Evaluation of LSTM and ARIMA. In P. Kralj Novak, T. Šmuc, & S. Džeroski (Eds.), Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 11828 LNAI, 129–144. doi:10.1007/978-3-030-33778-0_11.

[44] Santos Escriche, E., Vassaki, S., & Peters, G. (2023). A comparative study of cellular traffic prediction mechanisms. Wireless Networks, 29(5), 2371–2389. doi:10.1007/s11276-023-03313-9.

[45] Mehri, H., Chen, H., & Mehrpouyan, H. (2024). Cellular Traffic Prediction Using Online Prediction Algorithms. arXiv Preprint arXiv:2405.05239. doi:10.48550/arXiv.2405.05239.

[46] Perifanis, V., Pavlidis, N., Koutsiamanis, R. A., & Efraimidis, P. S. (2023). Federated learning for 5G base station traffic forecasting. Computer Networks, 235, 1389–1286. doi:10.1016/j.comnet.2023.109950.

[47] Wang, Y. (2022). Base Station Mobile Traffic Prediction Based on ARIMA and LSTM Model. Lecture Notes in Electrical Engineering: Vol. 797 LNEE, 164–175. doi:10.1007/978-981-16-5692-7_18.

[48] Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018, 1394–1401. doi:10.1109/ICMLA.2018.00227.

[49] Xu, M., Wang, Q., & Lin, Q. (2018). Hybrid holiday traffic predictions in cellular networks. IEEE/IFIP Network Operations and Management Symposium: Cognitive Management in a Cyber World, NOMS 2018, 1–6. doi:10.1109/NOMS.2018.8406291.

[50] Bangroo, R., Verma, S. R., Shivangi, & Shakuntala. (2023). Comparative Study of Elastic Net Regression, Naive Bayes & Lasso Regression. IEEE International Conference on Electrical, Electronics, Communication and Computers, ELEXCOM 2023, 1–6. doi:10.1109/ELEXCOM58812.2023.10370433.

[51] Garg, R., & Barpanda, S. (2022). Machine learning algorithms for time series analysis and forecasting. arXiv preprint arXiv:2211.14387. doi:10.48550/arXiv.2211.14387.