ISSN: 2769-545X

Research Article

Advances in Machine Learning & Artificial Intelligence

Machine Learning for Wireless Network Throughput Prediction

Gustavo A Fernandez*

School of Mathematical and Statistical Sciences, College of Sciences, The University of Texas Rio Grande Valley. USA

***Corresponding Author**
Gustavo A Fernandez, School of Mathematical and Statistical Sciences, College of Sciences, The University of Texas Rio Grande Valley, USA.

Submitted: 2023 Dec 06; **Accepted**: 2024 Jan 08; **Published**: 2024 Jan 31

Citation: Fernandez, G. A. (2024). Machine Learning for Wireless Network Throughput Prediction. Adv Mach Lear Art Inte, 5(1), 01-06.

Abstract

This paper analyzes a dataset containing radio frequency (RF) measurements and Key Performance Indicators (KPIs) captured at 1876.6MHz with a bandwidth of 10MHz from an operational 4G LTE network in Nigeria. The dataset includes metrics such as RSRP (Reference Signal Received Power), which measures the power level of reference signals; RSRQ (Reference Signal Received Quality), an indicator of signal quality that provides insight into the number of users sharing the same resources; RSSI (Received Signal Strength Indicator), which gauges the total received power in a bandwidth; SINR (Signal to Interference plus Noise Ratio), a measure of signal quality considering both interference and noise; and other KPIs, all derived from three evolved node base stations (eNodeBs). After meticulous data cleaning, a subset of measurements from one serving eNB, spanning a 20-minute duration, was selected for deeper analysis. The PDCP DL Throughput, as a vital KPI metric, plays a paramount role in evaluating network quality and resource allocation strategies. Leveraging the high granularity of the data, the primary aim was to predict throughput. For this purpose, I compared the predictive capabilities of two machine learning models: Linear Regression and Random Forest. Metrics such as Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were used to examine the models as they offer a comprehensive insight into the models' accuracies. The comparative analysis highlighted the superior performance of the Random Forest model in predicting the PDCP DL Throughput. The insights derived from this research can potentially guide network engineers and data scientists in optimizing network performance, ensuring a seamless user experience. Furthermore, as the telecommunication industry advances towards the integration of 5G and beyond, the methodologies explored in this paper will be invaluable in addressing the increasingly complex challenges of future wireless networks.

Keywords: Wireless Network, Machine Learning, Regression, Random Forest

1. Introduction

In today's digital age, telecommunications stand as a cornerstone of global connectivity. As the world becomes increasingly interconnected, cellular network operators grapple with the relentless challenge of accommodating escalating user demands. The explosion in media consumption, especially with the introduction of bandwidth-intensive applications, real-time media streaming on social platforms, and the rapidly evolving realm of connected and autonomous vehicles, has placed unprecedented pressure on network resources. To address these challenges, operators are in a continuous quest for cutting-edge solutions. One of the primary objectives is to refine resource allocation and load balancing mechanisms, ensuring that networks can handle the ever-growing data traffic without compromising on performance. The anticipatory approach to resource allocation and network management is a ground-breaking paradigm that offers a potential solution to these challenges. At the heart of this approach lies the ability to predict network connectivity fluctuations before they occur. By proactively identifying potential changes in connectivity, operators can take preemptive actions, ensuring that the user's Quality of Service (QoS) remains consistent and reliable. A prime example of this forward-thinking strategy is the concept of pre-buffering video content. By allocating additional resources in anticipation of a potential drop in future throughput values for a user, operators can guarantee uninterrupted streaming experiences. The need for such proactive measures stems from the paramount importance of delivering consistent and high-quality network connectivity. The proliferation of bandwidth-demanding applications and the exponential growth in media publishing and streaming on social platforms highlight the critical need for

innovative network management strategies.

Certain studies have made significant contributions to the field among the plethora of research dedicated to enhancing network QoS. For instance, Yue et al. embarked on a comprehensive correlation analysis, exploring the intricate relationships between Radio Signals (RSs) and throughput across various scenarios, from stationary settings to dynamic highway driving conditions [1]. Their findings emphasized the potential of the Random Forest machine learning model in predicting network performance based on metrics like RSRP, RSRQ, and CQI. In a similar vein, Raca et al. delved into the realm of predicting future throughput windows, evaluating the predictive prowess of diverse machine learning models, from Random Forest and Support Vector Machine (SVM) to Neural Networks (NN) [2]. Furthermore, a study by A.Y. et al. emphasized the significance of machine learning models in predicting downlink throughput on 4G-LTE networks [3]. Their research provides invaluable insights into the practical applications of these models in real-world network scenarios.

Building on the seminal work of these researchers and addressing the requirements of contemporary telecommunication networks, this paper presents a detailed analysis of a 4G LTE network dataset. Concentrating on key metrics such as RSRP, and RSRQ, the research aims to employ machine learning techniques, notably Linear Regression and Random Forest, to predict PDCP DL Throughput - a crucial metric in assessing network quality. Recent studies, including those by D. Minovski et al., and R. Zhohov et al., have further emphasized the importance and potential of machine learning in throughput prediction, underscoring the relevance and timeliness of the present research [4,5]. While many studies have ventured into network throughput prediction, the distinctiveness of this research manifests in several pivotal areas:

1.1. Real World Data

Grounded in data sourced from an operational 4G LTE network in Nigeria, this research offers a pragmatic vantage point often eclipsed in predominantly theoretical pursuits.

1.2. Granularity of the Data

The dataset vividly depicts real-time network dynamics with its intricate granularity captured within a mere 20-minute frame. Such granularity is instrumental in discerning intricacies that expansive datasets might inadvertently bypass.

1.3. Focused Predictors

Singular emphasis is placed on RSRP (Reference Signal Received Power) and RSRQ (Reference Signal Received Quality) as the chief predictors, enabling a meticulous probe into these pivotal metrics.

1.4. Temporal Feature Engineering

In addition to RSRP and RSRQ, a feature engineered with a lag of 1 is seamlessly integrated, infusing a temporal essence into the predictors. This strategic inclusion is geared towards encapsulating the inherent temporal interdependencies, bolstering the predictive prowess of the models.

1.5. Revisiting Linear Regression

Contrary to the prevailing trend of gravitating toward intricate machine learning architectures, this research reaffirms the merits of the foundational Linear Regression model. Enhanced with a temporal facet, it underscores its relevance and efficacy in specific contexts. This research heralds a pragmatic and astute approach to throughput prediction. As the ensuing pages unravel the findings and insights, it is imperative to acknowledge the meticulous strategies employed and fathom their ramifications for the telecommunications realm.

2. Materials and Methods

2.1. Data Collection

This research is centered on data derived from a functioning 4G LTE network in Nigeria [6]. The data, specifically the Key Performance Indicators (KPIs), were obtained from stationary transmitters, commonly referred to as evolved node base stations (eNodeB). These eNodeBs, with an average height of 25 meters, were equipped with commercial gear from a leading network provider in Nigeria. Throughout the study, specialized Drive Test (DT) equipment was employed to capture a range of metrics, including SINR, RSRP, RSRQ, and RSSI, among other vital KPIs, directly from the active sectors of the eNodeBs. My research specifically hones in on metrics associated with the Packet Data Convergence Protocol (PDCP) Downlink (DL) Throughput, particularly emphasizing radio measurements like RSRP and RSRQ. The data in focus was recorded at a 4G LTE frequency of 1876.6MHz, operating within a bandwidth of 10MHz.

2.2. Data Preprocessing

Wireless network datasets are inherently intricate, owing to their exposure to various fluctuating environmental and technical variables. Recognizing this complexity, rigorous data preprocessing was essential to ensure the robustness and reliability of the subsequent analysis. To achieve a consistent dataset free from site-specific anomalies, focus was narrowed to data sourced from a single site. This approach aimed to eliminate discrepancies or inconsistencies that might emerge from variations across different sites. To ensure a comprehensive understanding of the data's completeness, missingness heatmaps were used to represent missing values across features visually. Given the critical nature of certain columns, rows with missing values in these columns were eliminated, and they were determined to be missing at random. The result of this meticulous cleanse was a dataset with heightened integrity.

The importance of temporal features in the analysis became evident. The Date Time column was converted into a date time data type, laying the foundation for time series analysis. The data was then grouped by this temporal feature, and specific aggregations were applied to other columns to capture the mean within each time group. To further optimize the dataset, rows with specific

abnormal values in the Serving EARFCN column were removed. A lag feature was introduced based on the PDCP Throughput DL column to add depth to the analysis. This temporal aspect provides a time-shifted perspective, invaluable for forecasting and understanding patterns.

2.3. Model Selection

A combination of modeling strategies was employed to address the multifaceted challenge of forecasting PDCP throughput. Linear Regression, a foundational pillar of statistical modeling, was harnessed. Enhanced with a temporal feature, it adeptly discerned and accounted for the time-based patterns and trends intrinsic to the dataset. The essence of this model was to draw a linear relationship between the target variable, PDCP DL Throughput, and its array of predictors, prominently spotlighting the temporal aspect.

On the other hand, the Random Forest algorithm provided a more intricate perspective. This ensemble technique, by weaving together insights from numerous decision trees during its training phase, offers a profound depth of analysis. For classification, it delivers the mode of the classes, while in regression scenarios, it presents the mean prediction. Its inherent ability to grapple with non-linear complexities and unearth subtle patterns in the dataset proved invaluable in this forecasting endeavor. These methodologies, each with distinctive strengths, converged to form a robust and comprehensive forecasting framework.

2.4. Model Evaluation

The efficacy of the employed models was rigorously evaluated using an array of metrics, each uniquely tailored to gauge different facets of prediction accuracy and model reliability:

• **Mean Absolute Error (MAE):** A straightforward metric, the MAE computes the average of the absolute discrepancies between forecasted and actual outcomes. Mathematically, it is defined as:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

• **Root Mean Squared Error (RMSE):** Delving deeper into error magnitudes, the RMSE captures the square root of the mean of squared deviations between predictions and actual observations. Its formula is:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

• **R-squared:** Primarily associated with linear regression, the R^2 value elucidates the proportion of variance in the dependent variable that the independent variables in the model account for. It is computed as:

$$R^2 = 1 - \frac{\mathrm{SS_{res}}}{\mathrm{SS_{tot}}}$$

where SSres is the sum of squares of the residuals and SStot is the total sum of squares.

By harnessing these evaluation techniques, the aim was to measure the prediction precision of the models for PDCP DL throughput and to furnish insights that can illuminate pathways for subsequent research endeavors in this arena.

3. Results

3.1. Descriptive Analysis

An examination of the PDCP DL Throughput data over the specified 20-minute interval revealed its inherently dynamic nature. While no discernible pattern was immediately evident, the data vividly portrayed wireless networks' ever-fluctuating and volatile nature. Every passing second exhibited throughput alterations, underlining the network environment's non-static and rapidly evolving characteristics. This continuous oscillation in throughput underscores the challenges and intricacies of predicting such a metric, given its susceptibility to a multitude of factors that can change from moment to moment. A visual representation of this dynamic throughput over the interval can be seen in Figure 1.

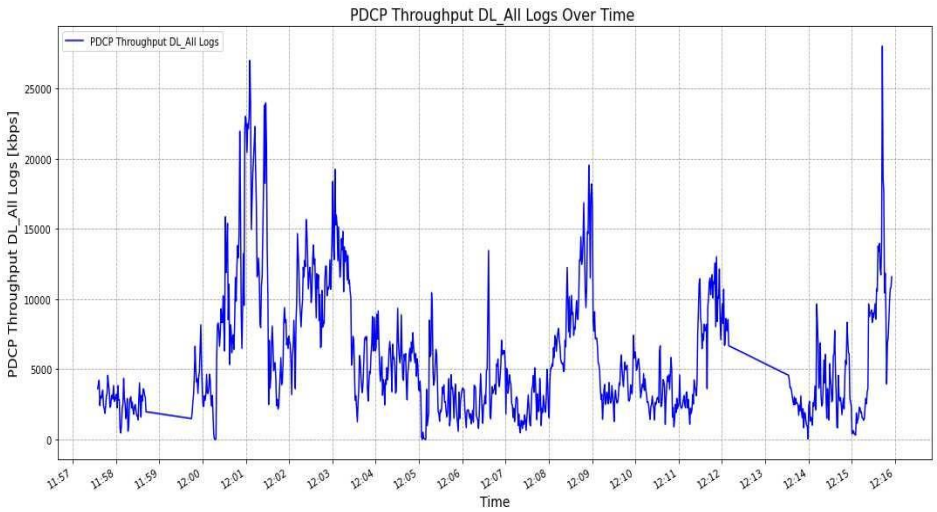

Figure 1: Dynamic PDCP DL Throughput over a 20-minute interval

To further comprehend the data's characteristics, summary statistics of the numerical features were computed. These statistics offer insights into the distribution, central tendency, and spread of the data for each feature.

	Serving RSRP	Serving RSRQ	Serving RSSI	PCC SINR	PHY Throughput DL	PDCP Throughput DL
count	1504.0	1504.0	1504.0	1504.0	1504.0	1504.0
mean	-85.27	-9.00	-62.02	9.08	7186.58	6026.89
std	7.88	1.10	7.52	6.65	5201.80	4696.75
min	-99.91	-14.25	-76.56	-7.62	128.0	0.0
25%	-91.28	-9.62	-67.51	3.79	3461.82	2676.51
50%	-86.75	-8.83	-63.95	7.95	5326.36	4454.01
75%	-79.72	-8.30	-56.82	13.7	9596.29	8297.02
max	-59.05	-3.7	-35.92	26.08	28890.94	28040.11

Table 1: Descriptive Statistics of the Dataset's Numerical Features

3.2. Model Performance

Central to this research was the task of gauging the forecasting aptitudes of two distinct models for PDCP DL throughput. Their performances, detailed in Table 2, offer a lens into their predictive strengths. The Linear Regression model, bolstered with a temporal aspect, showcased a commendable accuracy, as evidenced by its R^2 value. However, when benchmarked against metrics like MAE and RMSE, the Random Forest model, with its ensemble-based approach, proved to be slightly superior in its predictive accuracy.

Comparing the performance metrics, it becomes evident that the Random Forest model holds a slight advantage over the Linear Regression model in terms of predictive accuracy, as indicated by measures like MAE and RMSE. This comparison is further visualized in Figure 2, which plots predicted values against actual values. Through this figure, areas where each model excels or struggles are highlighted, offering a clear view of their respective prediction strengths and limitations.

Model	MAE	RMSE	R^2
Linear Regression with Temporal Feature	1,188.59	1,789.64	0.8218
Random Forest	1,100.69	1,736.90	0.8321

Table 2: Performance Metrics for Model Evaluation

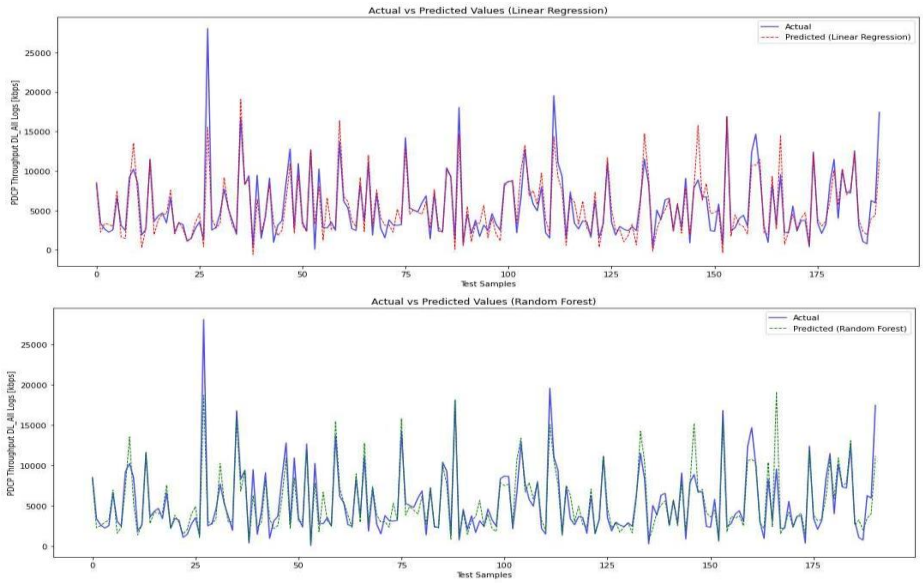

Figure 2: Comparison of Predicted Values against Actual Values for the Linear Regression and Random Forest Models

4. Discussion

Predicting PDCP DL Throughput in wireless networks is an intricate endeavor, laden with both challenges and avenues for the application of advanced predictive modeling. This study delved deep into these intricacies, utilizing both Linear Regression enhanced with a temporal feature and the Random Forest model to shed light on throughput predictability. A cornerstone in model evaluation, the Mean Squared Error (MSE) speaks volumes about prediction accuracy. Both models displayed admirable prowess. Yet, the Random Forest model slightly edged out its counterpart, registering an MSE of 3,016,817.89 against Linear Regression's 3,202,810.38. This edge can be attributed to the ensemble nature of Random Forest, adept at discerning non-linearities and subtle data patterns.

The □2 score, delineating the explanatory power of the models regarding the variations in PDCP DL Throughput, painted a congruent picture. Both models posted impressive R^2 scores exceeding 0.8. The Random Forest model, however, with an □2 of 0.8321, slightly surpassed the 0.8218 score of the Linear Regression model. Further insights were gleaned from the Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics. The close MAEs of 1,188.59 for Linear Regression and 1,100.69 for Random Forest, coupled with respective RMSEs of 1,789.64 and 1,736.90, reiterate the neck-to-neck performance of the two models. Yet, the slight superiority of the Random Forest model remained consistent across all metrics.

While the empirical data leans toward Random Forest, the virtues of each model in varied contexts cannot be understated. With its transparency, Linear Regression elucidates clear feature-target relationships priceless in situations where clarity supersedes sheer accuracy. With its nuanced handling of complex feature dynamics, Random Forest becomes the go-to when top-tier prediction accuracy is the order of the day. However, it's imperative to temper these findings with the understanding of the dataset's scope focused on a singular site over a 20-minute span. This dataset, albeit rich, captures a mere moment in the vast expanse of network operations. When faced with diverse conditions or prolonged durations, the true mettle of these models beckons further exploration.

To encapsulate, this investigation accentuates the significance of judicious model selection in the realm of throughput forecasting. While Random Forest clinched slightly superior metrics in this endeavor, the ultimate choice hinges on the unique demands of the task whether it's model interpretability, sheer accuracy, or computational nimbleness. As the tapestry of wireless communication grows more intricate, the tools we harness must evolve in tandem, propelling the field to new pinnacles of innovation and service par excellence.

4.1. Limitations and Future Directions

Regardless of its depth and rigor, every research endeavor has inherent limitations, and this study is no exception. One of the perennial challenges in machine learning is overfitting, where a model becomes too attuned to the training data, compromising its generalization to new, unseen data. Given the high granularity of our dataset and the complexity of the Random Forest model, there's a potential risk of overfitting. Regularization techniques or pruning might be needed to ensure that our models are robust and generalizable. Additionally, while offering a detailed snapshot, the study's focus on a singular site over a brief 20-minute span does not encompass the myriad dynamics of wireless networks over extended periods or across diverse geographical locales. Such constraints could influence the model's adaptability to broader network scenarios.

Future research could address these limitations by incorporating

data from multiple sites or by exploring longer timeframes, thus ensuring a more comprehensive representation of network dynamics. Moreover, experimenting with other advanced predictive models or ensemble techniques might further enhance the predictive accuracy and robustness against overfitting. The integration of additional features, perhaps from external datasets or newer technological advancements in the wireless domain, could also prove invaluable in refining throughput predictions. Ultimately, as the field of wireless communication continues to evolve, there will be an ever-growing imperative to adapt, innovate, and refine the methodologies employed, ensuring that research remains abreast of technological progress.

5. Conclusion

Wireless networks are the bedrock of our increasingly digitalized world. Ensuring their optimal performance is more than just a technical imperative; it's pivotal to the seamless integration of technology into our daily lives. In this study, the endeavor to predict PDCP DL Throughput via Linear Regression and Random Forest models cast light on the multifaceted nature of such a task. While the Random Forest model slightly edged ahead, showcasing the prowess of ensemble methodologies in deciphering complex data patterns, the Linear Regression's performance was not to be overshadowed. Its robustness, especially when bolstered with a temporal dimension, reiterated the lasting relevance of traditional statistical approaches.

The scope of the research, limited to a dataset from a singular location within a concise time window, serves as a snapshot a vignette of the grander tableau of challenges in wireless network predictions. A key takeaway is the absence of a one-size-fits-all solution. The choice of predictive model hinges on the nuanced requirements of the task at hand, be it sheer predictive accuracy, model transparency, or computational pragmatism. Looking ahead, as we stand at the cusp of a 5G-dominated world with whispers of 6G innovations, the imperative for refined, accurate, and adaptable forecasting tools grows exponentially. This study underscores the necessity for an adaptive research ethos one that is receptive to the swift currents of technological progress. By championing such a spirit of relentless innovation and introspection, we pave the way for wireless networks that are not just technically superior but also deeply resonant with the dynamic needs of their users.

References

1. Yue, C., Jin, R., Suh, K., Qin, Y., Wang, B., & Wei, W. (2017). LinkForecast: Cellular link bandwidth prediction in LTE networks. IEEE Transactions on Mobile Computing, 17(7), 1582-1594.
2. Raca, D., Zahran, A. H., Sreenan, C. J., Sinha, R. K., Halepovic, E., Jana, R., & Gopalakrishnan, V. (2020). On leveraging machine and deep learning for throughput prediction in cellular networks: Design, performance, and challenges. IEEE Communications Magazine, 58(3), 11-17.
3. Al-Thaedan, A., Shakir, Z., Mjhool, A. Y., Alsabah, R., Al-Sabbagh, A., Salah, M., & Zec, J. (2023). Downlink throughput prediction using machine learning models on 4G-LTE networks. International Journal of Information Technology, 15(6), 2987-2993.
4. Minovski, D., Ogren, N., Ahlund, C., & Mitra, K. (2021). Throughput prediction using machine learning in lte and 5g networks. IEEE Transactions on Mobile Computing.
5. Zhohov, R., Palaios, A., & Geuer, P. (2021, October). One step further: Tunable and explainable throughput prediction based on large-scale commercial networks. In 2021 IEEE 4th 5G World Forum (5GWF) (pp. 430-435). IEEE.
6. Imoize, A. L., Orolu, K., & Atayero, A. A. A. (2020). Analysis of key performance indicators of a 4G LTE network based on experimental data obtained from a densely populated smart city. Data in brief, 29, 105304.