5G-ENERGY CONSUMPTION MODELLING

Julius Maina

The author is a research student in data analytics at KCA University in Kenya

October 25th, 2023

Abstract *– 5G, the fifth-generation radio technology, has brought a wave of innovation with new services and networking paradigms, but it also raises concerns about increased energy consumption in network deployments. Despite 5G's potential to be more energy-efficient, the need for a larger number of cells and higher processing for wider bandwidths result in a substantial energy footprint. A major portion of operational costs is attributed to energy bills, with the radio access network (RAN) and base stations being significant consumers. Accurate modeling of base station energy consumption is crucial for optimizing network energy efficiency. This report presents an approach using a CatBoost model to estimate energy consumption, promoting generalization across various base station products and configurations, ultimately enhancing energy-efficient network deployments.*

Keywords – 5G Energy Consumption, Base Station Optimization, Machine Learning Model, Dataset Preprocessing, Feature Engineering, Time-based Features, Model Selection, LightGBM Regressor, Hyperparameter Tuning, Generalization Capacity.

1. INTRODUCTION

The advent of 5G, the fifth generation of radio technology, has ushered in a new era of communication services, technological advancements, and networking paradigms. These innovations have brought about significant societal benefits and transformed the way we connect. However, alongside the promises of 5G, there is a growing concern surrounding the energy consumption associated with these advanced network deployments.

While 5G networks are hailed as being approximately four times more energy-efficient than their predecessors, the sheer magnitude of these networks has led to an energy consumption increase of approximately threefold. This energy surge is primarily attributed to the need for a greater number of cells to provide coverage at higher frequencies and the intensified processing requirements necessitated by wider bandwidths and a multitude of antennas.

Notably, the operational expenditure (OPEX) of network operators already accounts for a substantial portion, approximately 25 percent, of their total costs, with an overwhelming 90 percent directed towards energy bills. Remarkably, more than 70 percent of this energy is devoured by the radio access network (RAN), with a specific focus on the energy-hungry base stations (BSs), while data centers and fiber transport systems claim a relatively smaller share.

The energy consumption of these base stations hinges on an intricate interplay of factors, including architectural variations, configuration parameters, traffic conditions, and the activation of energy-saving techniques. To mitigate this substantial energy footprint, optimizing base station parameters and energy-saving methods is paramount. This optimization demands a profound understanding of how these factors impact the energy consumption of diverse base stations. Thus, achieving accurate modeling of energy consumption is of utmost importance in the quest for more energy-efficient network deployments.

This report presents a design of a machine learning-based solution that can be trained on a diverse dataset of scenarios, showcasing robust generalization to uncharted territories. It delves into the development of a model capable of estimating the energy consumption of various base station products, achieving generalization across different base station products and configurations. The report outlines the efforts made to pave the way for energy-efficient network deployments in the era of 5G.

2. DATASET

The dataset provided for the problem statement was a comprehensive collection of cell-level traffic statistics from 4G and 5G sites, recorded on different days.

The dataset was organized into three distinct segments, each offering valuable insights into various

aspects of network performance:

- Base Station Basic Information (BSinfo.csv):

This dataset encompassed essential configuration parameters and hardware attributes of base stations. It provided in-depth information, including:

- ☐ Configuration settings.
- ☐ Hardware attributes.

These details were instrumental in understanding the structural underpinnings of base stations.

- Cell-Level Data (CLdata.csv):

In this dataset, hour-level counters that encompassed a range of critical parameters:

- ☐ Service compliance counters, such as load statistics.
- ☐ Energy-saving method counters, such as the duration of energy-saving mode activation.

These counters shed light on how base stations handle network traffic and energy-saving strategies.

- Energy Consumption Data (ECdata.csv):

This dataset offered hour-level specifications related to energy consumption. Key features included:

- ☐ Total energy consumption of the base stations.
- ☐ Insights into the energy usage patterns of network elements.

This data was pivotal in understanding the energy footprint of the network and therefore formed the target variable.

These datasets provide a comprehensive view of base station configurations, network traffic, and energy consumption, making them invaluable resources for designing machine learning models and data-driven strategies to enhance network efficiency and sustainability.

2.1 Data Preprocessing

Data pre-processing is a fundamental step in preparing the dataset for machine learning. It involves various operations to ensure that the data is in a suitable format and quality for model training. One critical aspect of pre-processing in this project was the transformation of the target variable. The target variable, which represented energy consumption, exhibited right-skewed (as seen in Fig 1) behavior, which can lead to modeling issues. To address this, a natural logarithm transformation (np.log1p) was applied to the target variable. This transformation is particularly useful when dealing with skewed data, as it reduces the impact of extreme values and promotes a more symmetrical distribution. By transforming the target variable, we aimed to improve the model's ability to capture underlying patterns in the data and enhance the accuracy of predictions.

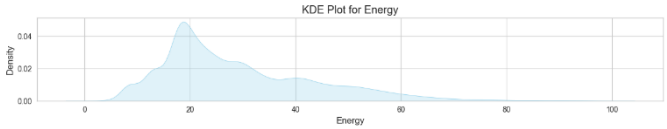

Fig. 1 – Distribution of 'Energy' Variable in Train

Categorical variables in the dataset were encoded to numerical form using the LabelEncoder. This process assigned a unique integer to each distinct category within the categorical features. By converting these categorical variables into numerical representations, the dataset became compatible with machine learning algorithms. This essential step prepared the data for subsequent feature engineering and model training, ensuring that valuable information within the categorical features was preserved.

The dataset was divided into training, validation, and test sets to facilitate robust model evaluation. Repeated K-Fold cross-validation with 10 splits and 5 repetitions was employed, ensuring that each data point was included in both training and validation sets at different times. This approach helped assess the model's performance under varying conditions and provided a reliable estimate of its generalization capabilities. The training data were used for model parameter estimation, the validation data for model assessment, and the test data for final performance evaluation. This separation allowed for comprehensive evaluation and ensured the model's ability to generalize to unseen data.

2.2 Data Cleaning

In the process of data cleaning, we identified that the ESMode4 column in the Cell-level CSV file contains a consistent value of zero throughout the dataset. Consequently, as this column did not provide any meaningful variation in the data, it was deemed unnecessary and was subsequently removed to streamline the dataset and enhance its overall quality

The dataset underwent a harmonization process to rectify inconsistencies in the number of antennas across Base Stations (BS). Initially, the unique count of 'Antennas' within each 'BS' group was calculated, determining whether it equaled 1, signifying uniformity in antenna numbers. Subsequently, rows with non-uniform antenna counts were isolated. To ensure data consistency, a grouping approach was employed based on the 'BS' column, facilitating the

identification of BS entries with inconsistent antenna counts. These entries were then flagged and adjusted for harmonization, ensuring uniformity in antenna numbers across the dataset.

2.3 Feature Engineering

To avoid any potential data leakage caused by using future values, we intentionally excluded features with shifted values in our engineering process. Additionally, it's worth noting that we did not treat this problem as a time-series issue, and consequently, we refrained from engineering any time-series-related features.

To start with active power saving modes features were engineered. The active power saving modes features engineering step focused on creating interactions between different activation modes and involved the combination of two modes at a time. These interactions were constructed by concatenating the mode values, which allows the model to capture relationships between the various power-saving modes and their joint effects on the data. This step significantly enriches the dataset by introducing features that provide insights into the combined influence of power-saving modes, potentially revealing more nuanced patterns and relationships in the data.

Next time features were engineered. In the time features engineering step, several temporal aspects were incorporated into the dataset. The 'hour' feature captures the hour of the day, and 'cosine_hour' and 'sine_hour' provide the cosine and sine of the hour, respectively, allowing the model to understand daily patterns. Additionally, 'day_of_week' indicates the day of the week, and 'day_of_week_cos' and 'day_of_week_sin' represent the day of the week as trigonometric values to capture weekly trends. The 'is_weekend' feature distinguishes between weekdays and weekends, enabling the model to account for different behaviors on these days. Finally, 'period_of_day' categorizes hours into four periods (morning, afternoon, evening, and night), and an 'hour_period_interaction' feature is created to account for interactions between the hour and period of the day. These time-based features offer valuable temporal context for the model to detect patterns and dependencies related to time.

In the final step, two transformations were applied to the dataset. Firstly, the 'bs_en' feature was created by extracting numeric values from the 'bs' column, removing the 'B_' prefix. This transformation simplifies the representation of base station IDs. Secondly, one-hot encoding was performed on the 'rutype' and 'mode' columns to convert categorical variables into binary values, allowing the model to work with these categorical features effectively. These engineering steps ensure that the data is in a format suitable for machine learning algorithms, making it easier for the model to understand and utilize these features during training and prediction.

A pictorial representation of our models' feature importance ranking is shown in figure 2

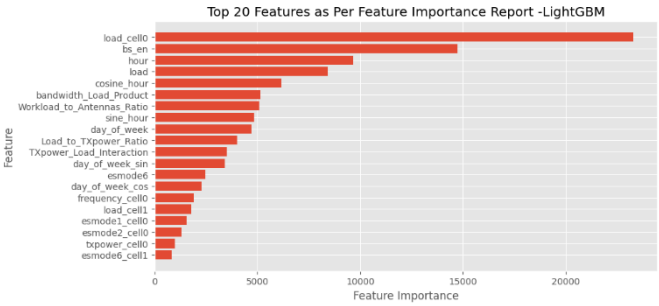

Fig. 2 – Summary of Feature Importance in LightGBM for top features

3. MODEL SELECTION

In this section, we elaborate on our approach to model selection, parameter tuning, and ultimately utilizing the LightGBM regressor for the task at hand.

3.1 Model Comparison

To ensure the efficacy of our predictive model, we commenced with a comprehensive comparison of various machine learning algorithms. The following models were considered for evaluation:

- ☐ XGBoost
- ☐ HistGradientBoosting
- ☐ LightGBM
- ☐ CatBoost

These models were trained and tested to gauge their generalizability capacity on our dataset. The evaluation was based on accuracy, a fundamental metric for classification tasks. The results of this initial comparison are summarized below:

Table. 1 – Performance of the models after close of the competition

Model	WMAPE Public	WMAPE Private
XGBoost	0.0918	0.0909
LightGBM	0.0917	0.0906
HistGradientBoosting	0.1061	0.099
CatBoost	0.0989	0.0935

It is evident from the comparison that both XGBoost and LightGBM outperformed other models in terms of accuracy.

3.2 Model Building

Final LightGBM regressor model was trained on the entire training dataset using the following hyperparameters:

- Objective: regression,
- Boosting_type: gbdt,
- n_jobs: -1,
- max_depth: -1,
- metric: mape,
- num_boost_round: 10000,

We employed repeated k-fold cross-validation with 10 folds and 5 repeats to ensure robust model evaluation and selection of the best hyperparameters. This approach allows for thorough testing and validation of the model's performance.

4. CONCLUSION

In conclusion, the advent of 5G technology has brought about transformative advancements in communication services, but it has also raised concerns about the surging energy consumption in network deployments. This report detailed the development of a machine learning-based solution to estimate the energy consumption of diverse base station products, paving the way for more energy-efficient 5G networks. The dataset, comprising base station configurations, network traffic, and energy consumption, served as a valuable resource. Data preprocessing, cleaning, and feature engineering were crucial steps in preparing the dataset for modeling. The introduction of active power saving modes and time-based features enriched the dataset, enhancing its ability to capture intricate patterns. Model selection led to the adoption of the LightGBM regressor for robust energy consumption estimation. This approach offers a promising solution for optimizing base station energy consumption, reducing operational costs, and advancing the sustainability of 5G networks.