A

Project Report

on

Big Data Analytics

Forecasting Energy Production in Turkey and Romania: A Comparative Analysis of Machine Learning Models

Submitted by:

Godavarthi Sai Nikhil (210214)

Under guidance of

Dr. Yogesh Gupta

Department of Computer Science Engineering

SCHOOL OF ENGINEERING AND TECHNOLOGY

BML MUNJAL UNIVERSITY,

GURUGRAM (INDIA)

May 2024

ACKNOWLEDGEMENT

I am highly grateful to **Dr. Yogesh Gupta,** Professor at BML Munjal University, Gurugram, for providing supervision to carry out the Big Data Analytics from February-May 2024.
We extend our sincere thanks to him for his continuous encouragement throughout the project. Additionally, I am grateful to my friends who generously dedicated their time and assistance, contributing significantly to the successful completion of our group project.

Godavarthi Sai Nikhil

List of Figures

Table of Contents

Abstract

Electricity production forecasting is crucial for effective energy management and planning. This study investigates the effectiveness of various machine learning models for predicting hourly electricity production in two countries: Romania and Turkey. We explore the application of advanced time series models to enhance forecasting accuracy. Our project implements Long Short-Term Memory (LSTM) networks, Random Forest (RF), and XGBoost models, comparing their performance across different time granularities, including seasons and years. Through this comparative analysis, we identify the strengths and weaknesses of each approach, with a particular focus on the LSTM's ability to capture different time dependencies in the data. The insights gained from this analysis are then leveraged to implement the best-performing model on a new dataset for Turkey, resulting in more accurate electricity production forecasts. This work contributes to the field by providing a comprehensive evaluation of machine learning techniques in the context of energy production prediction, offering valuable guidance for future applications in similar domains.

Motivation

The motivation for this project arises from the importance of accurate and timely electricity production forecasts, which is crucial for the planning and management of electricity usage. Timely and precise predictions of energy demand ensure that consumers receive the right amount of energy from the service provider and significantly minimize the use of reserve power. It is especially important for renewable energy sources as they are stochastic by their nature. The main benefits of increased predictability of energy production are the ability to better manage business resources and operational expenses, as well as the overall grid reliability.

Another important reason for this research is to achieve economic efficiencies. Reduced energy costs mean that companies can forecast the amount of energy required for a given period, and avoid costs incurred by overproduction or underproduction. This economic benefit is very important to keep competitive the prices of energy and guarantee the sustainability of the energy providers. In addition, accurate forecasting enables the incorporation of renewable forms of energy such as solar and wind energy into the grid, thereby adopting a sustainable approach to energy supply. The optimization of renewable energy sources in power production decreases the dependency on fossil fuels and has a positive impact on greenhouse gas emissions, which are in line with international environmental standards.

Additionally, this project builds upon several modern techniques in machine learning including LSTM, Random forest, and XGBoost to improve the forecasting. These models are most suitable for time series forecasting since they can identify intricate relationships and dependency in data. Thus, besides enhancing the accuracy of energy forecasting in Romania and Turkey, this research seeks to enrich the general body of energy literature with advanced models and methods. These models can be implemented in other regions, and hence, the application of these models contributes towards the enhancement of both the energy sector and the field of machine learning.

1.INTRODUCTION

The demand for energy is increasing, and due to climate change, there are shifts in the global energy mix, and technological development is also playing a major role. Forecasting energy production with precision and reliability has become crucial for ensuring energy security, optimizing energy infrastructure, and facilitating the transition toward a sustainable energy future. In this project, the emphasis is made on the prediction of the hourly energy production in Turkey and Romania by using advanced machine learning models.

Accurate forecasting models are essential for informed decision-making in several key areas:

- Resource Planning and Allocation include energy production estimates, which can be used in determining power infrastructure development, equipment scheduling and supply of resources in different energy types.
- Grid Stability and Reliability ensures a stable and reliable power grid hinges on the ability to accurately predict energy production, especially from intermittent renewable sources like wind and solar.
- Energy Market Operations forecasts Energy production as an essential process as it helps in determining the price of energy, trading energy, and managing risks in energy markets.
- Policy Development helps governments in making policies related to energy security, sustainable development goals, and emission control plans.

In this project, we delve into the intricacies of forecasting energy production by focusing on two distinct European countries Turkey and Romania. These countries can be said to be having different energy profiles thus making it easier to understand their different energy prospects. Turkey is a young country with a quickly developing economy and, accordingly, a growing need for energy, which is why it is trying to increase the share of renewable energy sources in its energy mix. While Romania has a large hydropower potential and is still working to incorporate wind and solar power to its grid.

By studying these two contrasting cases, we aim to develop and evaluate robust forecasting models capable of capturing the complex dynamics of energy production in different contexts. This will be done using deep learning and other conventional methods of machine learning to determine how well they perform in terms of accuracy and reliability.

1.1 Project Scope and Objectives: Leveraging Machine Learning to Forecast Energy Production

The core objective of this project is to develop and compare multiple machine learning models for forecasting energy production in Turkey and Romania.

Our analysis will encompass two distinct datasets, each providing valuable insights into the specific characteristics of energy production in each country:

- Energy Consumption and Pricing in Turkey dataset gives a complete picture of Turkey's energy mix as well as the consumption, prices, and production of energy of all types. This dataset will then be utilized as the benchmark for analysing the accuracy and efficacy of the models in predicting energy production in a constantly shifting market.
- Hourly Electricity Consumption and Production in Romania dataset provides detailed information about the electricity generation in Romania daily. It will help us evaluate the ability of the models to predict short-term production variation, which is essential for maintaining the stability of the grid and incorporating renewable energy sources into the power system.

To achieve our objective, we will implement and compare the performance of three different machine learning models LSTM, Random Forest, XGBoost

In this study, we evaluate the proposed models against each other and with the benchmark models using both datasets to determine the best forecasting approach in each scenario. This comparative analysis will also be useful in identifying the strengths and weaknesses of each model towards enabling the choice of the most suitable tool for use in energy production forecast under various conditions.

Furthermore, our project will delve into the following aspects:

- Feature Engineering: We will explore the impact of different features on model performance, including historical production data, weather variables, economic indicators, and calendar information.
- Model Optimization: We will fine-tune the hyperparameters of each model to achieve optimal performance on the respective datasets.
- Evaluation Metrics: We will employ robust evaluation metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared to assess the accuracy and reliability of the forecasting models.

Ultimately, this project aims to contribute to the advancement of energy production forecasting methodologies, providing a valuable resource for researchers, policymakers, and energy stakeholders seeking to navigate the complexities of the evolving energy landscape.

2. PROBLEM STATEMENT

2.1 State the problem or problems that motivated or required a solution provided by this project:

The energy sector faces multiple challenges that forecasts energy production. These challenges include:

- The increasing adoption of renewable energy sources like wind and solar introduces significant variability and uncertainty into energy production due to their intermittent nature. This variability poses challenges to grid reliability and stability by presenting challenges to forecasting models that may be used to predict production.
- Developing countries like Turkey experience rapid economic growth, leading to surging energy demand. Accurately forecasting this demand is crucial for infrastructure planning, resource allocation, and ensuring a secure energy supply.
- Global energy markets are subject to price fluctuations influenced by geopolitical events, economic trends, and supply-demand dynamics. Precise energy production forecasts are vital for navigating these price volatilities and ensuring market stability.
- Many countries are committed to reducing their carbon footprint and transitioning towards a more sustainable energy future. Accurate forecasting of energy production from renewable and non-renewable sources is essential for monitoring progress towards these targets and making informed policy decisions.

2.2 List the specific problem which your project is solving:

This project specifically addresses the problem of short-term energy production forecasting for Turkey and Romania. The primary focus is on predicting daily energy production for both countries with high accuracy and reliability by leveraging historical data and relevant features. The specific problems being addressed include:

- Stable short-term predictions are vital to ensure that supply meets demand which tends to vary at certain times of the day or year. When the energy generation is estimated with a high level of certainty, grid operators will be in a position to control energy distribution, storage and demand in

a more effective manner which will help in increasing the stability of the grid to the extent that blackouts are less frequent.

- The forecast of the energy production is useful in minimizing the cost of the energy production in the production process. This includes reducing costs of starting up or shutting down energy production units, buying energy from other sources, and avoiding imbalance costs.
- To the energy producers and policymakers, the forecasted production levels are useful in the planning process due to resource allocation. This involves coordination in maintenance of power plants, investment in the infrastructure, and planning for the future energy demand.
- Energy markets are well known for their volatility and their price is determined by the basic laws of supply and demand. Such forecasts are useful in stabilizing these markets as they offer the required data to conduct the markets and make necessary strategic plans. This includes issues to do with trading plans, price setting systems, and handling of risks.

2.3 Provide a detailed explanation of how this project solves the problem(s).

This project tackles the challenge of short-term energy production forecasting by developing and comparing the performance of three distinct machine learning models: LSTM: Long Short Term Memory, XGBoost: Extreme Gradient Boosting, Random Forest: Random Forest Algorithm. We apply these models to two distinct datasets: One of the datasets is covering the energy consumption and energy price for Turkey while the other is presenting the hourly electricity generation of Romania.

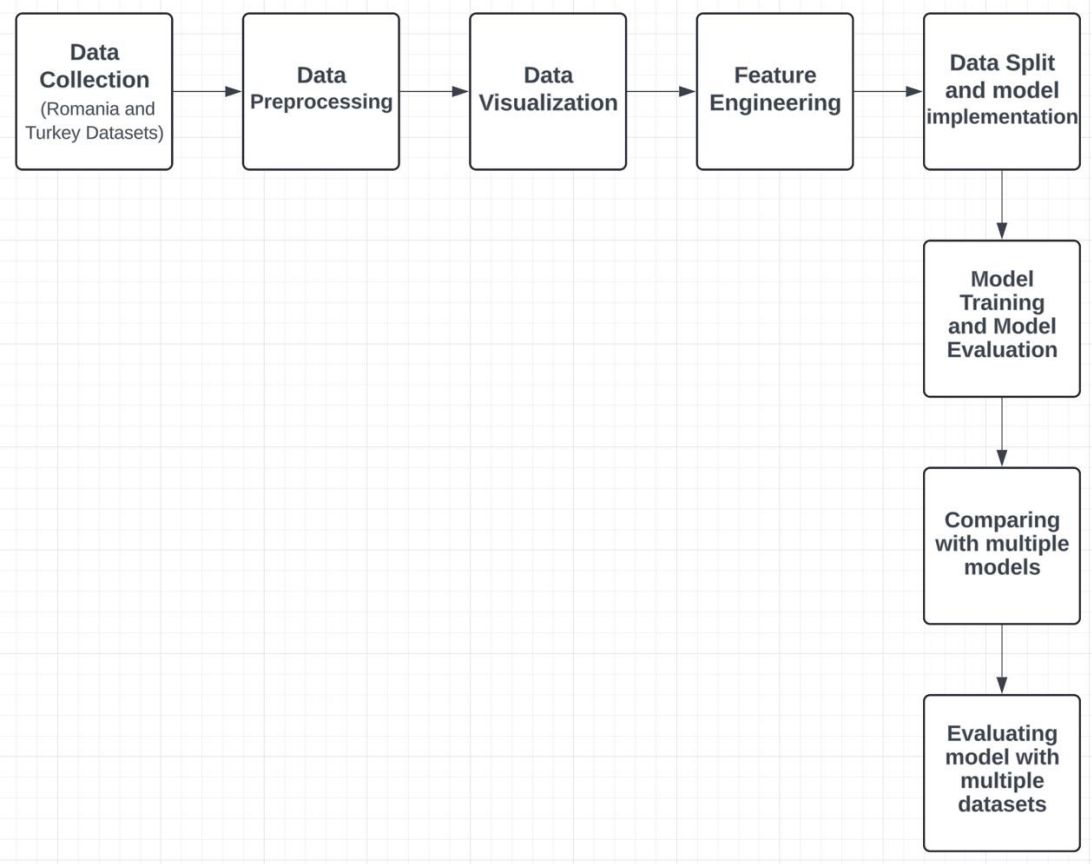

Figure 1: Framework for Energy Production Forecasting

By employing this rigorous framework, our project aims to deliver accurate and reliable short-term energy production forecasts for Turkey and Romania. The insights gleaned from this project will be valuable for various stakeholders, contributing to efficient energy management, grid stability, and informed decision-making in the energy sector.

3. LITERATURE REVIEW

3.1 Research Papers

The performance of the short-term consumption forecast was examined by the authors in [1] using the LSTM deep learning model. An energy consumption dataset from the Benin City regional 132/33KV transmission station of the Transmission Company of Nigeria (TCN) was used for this. The dataset consisted of half-hourly daily load readings that were recorded between August and December of 2021. The model was utilized to show that, even with the odd and inadequate energy consumption readings, it was still possible to produce a reliable short-term load forecast for the case study. The statistical evaluation metrics that are employed are Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE). With a 100 time-step, the method yields remarkably high levels of accuracy (MAPE of 0.010 and RMSE of 19.759).

A study [2] created a new method for time series forecasting to be applied to solar energy forecasting involving the use of the nonlinear autoregressive network with an exogenous input network. The model was tested in the context of absolute active power and in this case the proposed model was used with 744 samples collected at one-hour time intervals through a smart meter. It was trained using techniques such as Levenberg-Marquard LSG, Scaled conjugate gradient, and Bayesian Regularization. The best model was a nonlinear autoregressive network with an exogenous inputs model incorporating the Bayesian regularization algorithm was the best since it provided the least MSE value of 0.0031 and 0.0029 in the training and testing stages, respectively.

In another study [3], a new TSF-CGANs algorithm was derived, which incorporated the use of CGANs coupled with CNN and Bi-LSTM to enhance hourly PV power prediction error. Specifically, the generator in the TSF-CGANs network was adopted with the general structure of the regression forecast model, which took the features of history and random noise vectors as inputs. The actual value and the predicted value generated by the Bi-LSTM model were passed to the discriminator, which decided the authenticity of the values. The effectiveness of the proposed method was many folds substantiated with a real-life data set out compared LSTM, recurrent neural network, back-propagation neural network, support vector machine, and persistence models in terms of prediction accuracy. The TSF-CGANs model brought the improvement with RMSE reduced by 32% and the forecast skill (FS) was improved by 0. 4863 compared to Persistence.

In paper [4] was published that proposed a feedback-based forecasting method. It also improves the Fourier Series expansion of the hourly prediction by applying the current hour's error to the next hour. Using only historical demand data, the identified methodology, which is used in this paper with the example of the Turkish power market for 2012–2017, allows providing an effective instrument for the demand forecast on an hourly, daily, and annual basis. To obtain slightly improved results, the Fourier series expansion predictions are also passed through an autoregressive (AR) model. Thus, relative to the Mean Absolute Percentage Error (MAPE) benchmark, the hourly forecasting error of the demand is 2. 80% for day-ahead, 3. 53 for hour-ahead and 0 for day-ahead. 87% for hour-ahead horizons.

The authors in [5] utilized time series and nonlinear regression issues pertaining to the forecasting of building energy consumption. To address this issue, the differential evolution (DE) algorithm is used in this study. The performance of SVR is highly dependent on the choice of its parameters. Nu-SVR and epsilon-SVR weighted SVR models are used to develop the forecasting model. The weights assigned to each model are once more determined using the DE algorithm. The effectiveness of the suggested model is demonstrated through a case study involving time series data on energy consumption from an institutional building in Singapore. For the same building, the suggested model can be used to forecast daily and half-hourly electricity consumption time series data. The data on daily energy consumption has a mean absolute percentage error (MAPE) of 5.843, while the data on half-hourly energy consumption has a MAPE of 3.767.

A novel approach to long-term load forecasting with hourly resolution is put forth in paper [6]. Recurrent neural networks with Long-Short-Term-Memory (LSTM-RNN) cells form the core of the model. With LSTM-RNN, the long-term relationships in a time series data of electricity load demand are considered, leading to more precise forecasts. The suggested model is applied to ISO New England electricity market real-time data. To be exact, twelve years' worth of publicly accessible data, spanning from 2004 to 2015, were gathered to train and validate the model. Forecasts regarding the demand for electricity have been prepared on a rolling five-year basis, spanning from 2011 to 2015. The proposed model is found to be highly accurate with a Mean Absolute Percentage Error (MAPE) of 6.54 within a confidence interval of 2.25%.

The CRISP-DM data mining methodology was employed in the research [7] as a typical problem solver for both business and research. Two normalization scenarios and five attribute selection scenarios based on correlation values based on target attributes were the scenarios tested in the study. Then, Bi-LSTM with hyperparameter tuning grid search is the deep learning model that is employed. R2, RMSE, and MAPE are

used to evaluate performance measurements. The results of the tests indicated that the Bi-LSTM model yielded the best MAPE, 7.7256%. R2 of 0.6151 at min-max normalization and an RMSE of 0.1234. By contrast, the z-score normalization yields lower results, with the best MAPE value being produced at 10.5525%. R2 of 0.4186 and RMSE of 0.7627.

BigPSF, the new algorithm for prediction of big data time series is described in the work [8]. This new method is built upon the Pattern Sequence-based Forecasting algorithm, which has gained much prominence in the literature. First, prediction accuracy was increased in the original algorithm; secondly, the algorithm was transferred to big data context when significant scalability results were reached. The algorithm under consideration is a ready to use application with little tunable parameters that come bundled with the Apache Spark distribution for distributed computing. It involves the computation of the algorithm on real data correlating to Uruguay's electricity demand using physical and cloud clusters. The bigPSF has the RMSE values of 61.23, MAE value of 57.15 and MAPE value of 4.70.

To generate a short-term forecast of electricity demand, the paper [9] examined the advantages of merging data features. The nature of electricity typically exhibits a complex characteristic as well as a clear seasonal trend. Adaptive Fourier decomposition's benefit is first utilized in this paper to extract the fluctuation characteristics. Once the linear and stationary sequence conditions are met, the seasonal pattern is measured and eliminated using the sub-series approach. A quantitative identification of the average periodicity length is made during the seasonal adjustment process. Additionally, the sine cosine optimization algorithm is used to choose the support vector machine's kernel and penalty parameters to achieve the generalization performance on actual electricity demand data.

The study [10] used 1-hour unit electricity consumption data from 136 households to predict household AMI under a smart grid environment. The study used Euclidean and DTW distance calculation methods to analyze the performance of various models. The results showed that univariate seasonal time-series models (DSHW and TBATS) had more accurate predictive power without clustering. However, clustering using the Euclidean or DTW distance improved the prediction results significantly. The performance was also improved after clustering using the ARIMA, ARIMAX, DSHW, TBATS, NN-AR, and NARX models. The ETS-based DSHW and TBATS showed smaller improvements, but both methods with clustering showed better performance. The best models for Euclidean distance calculation were the ARIMA, ARIMAX, and TBATS models, with their MAPE values improving by 26.2% to 22.4%, 22.4% to 17.6%, 8.9% to 8.8%, and 9.1% to 8.8%, respectively. The optimal models for DTW distance calculation were the DSHW, NN-

AR, and NARX models, with their MAPE values improving by 9.1% to 8.8%, 19.5% to 9.4%, and 12.7% to 7.8%, respectively. NARX, which performed best among several models, the RMSE was 6.348 for the Euclidean distance calculation, but for DTW, it was 5.939, revealing a significant difference depending on the distance calculation method.

3.2 Patents

S. No.	Existing state of art	Drawbacks in existing state of art	Overcome
1	US Patent No. 9,882,023 - "System and Method for Short-Term Load Forecasting Using Machine Learning"	Limited to specific data sources (e.g., weather data)	Our project is adapted to incorporate diverse data streams (e.g., historical production data, grid topology).
2	US Patent No. 10,230,432 - "Method and Apparatus for Load Forecasting Using Time Series Data and Deep Learning"	Focuses on a particular deep learning architecture	Our project is extended to explore other deep learning architectures (e.g., convolutional neural networks) for potential improvements.
3	EP 3 213 007 A1 - "Method for Short-Term Load Forecasting Based on a Hybrid Neural Network and Fuzzy Inference System"	Relies on a specific hybrid model	Our project investigates in conjunction with more recent ensemble learning techniques.
4	US Patent Application No. 2020/0380242 - "Electricity Load Forecasting Method and System Using Attention-Based Long Short-Term Memory Networks"	Lacks exploration of transfer learning techniques	Our project investigates by pre-training the LSTM model on a broader energy production dataset for improved generalizability.
5	WO 2019/022532 A1 - "Method and System for Short-Term Photovoltaic Power Forecasting Based on Echo State Networks"	Restricted to a specific renewable energy source (photovoltaic)	Our project is adapted to handle predictions for various generation sources (e.g., wind, hydro) by incorporating relevant features.

6	CN 111932122 B - "A Short-Term Load Forecasting Method Based on Multi-Objective Grey Wolf Optimization and Extreme Learning Machine"	Employs a less common optimization technique (Grey Wolf Optimization)	Our project is compared with more established optimization algorithms (e.g., Adam optimizer) to assess effectiveness for hyperparameter tuning.
7	US Patent No. 11,331,923 - "Short-Term Load Forecasting Method and System Using a Convolutional Neural Network"	Lacks incorporation of temporal dependencies	Our project is hybridized with LSTM networks to leverage the strengths of recurrent learning for capturing spatial and temporal patterns.
8	EP 3 456 123 A1 - "Method for Short-Term Load Forecasting Based on Probabilistic Ensemble of Random Forests"	Relies solely on Random Forests	Our project is extended to explore a more diverse ensemble approach, combining Random Forests with other models like XGBoost for potentially enhanced robustness.

4. METHODOLOGY

4.1 Exploratory Data Analysis

For our study, we used Kaggle datasets to forecast hourly electricity output in Romania and Turkey.

Hourly Electricity Consumption and Production for Romania:

Date range: Five years.

Details: Hourly data on energy consumption and production were included, with types such as nuclear, wind, hydroelectric, oil and gas, coal, solar, and biomass.

Importance: This dataset is valuable for study due to Romania's diverse energy sources, which include considerable contributions from renewables such as solar and wind, as well as nuclear power.

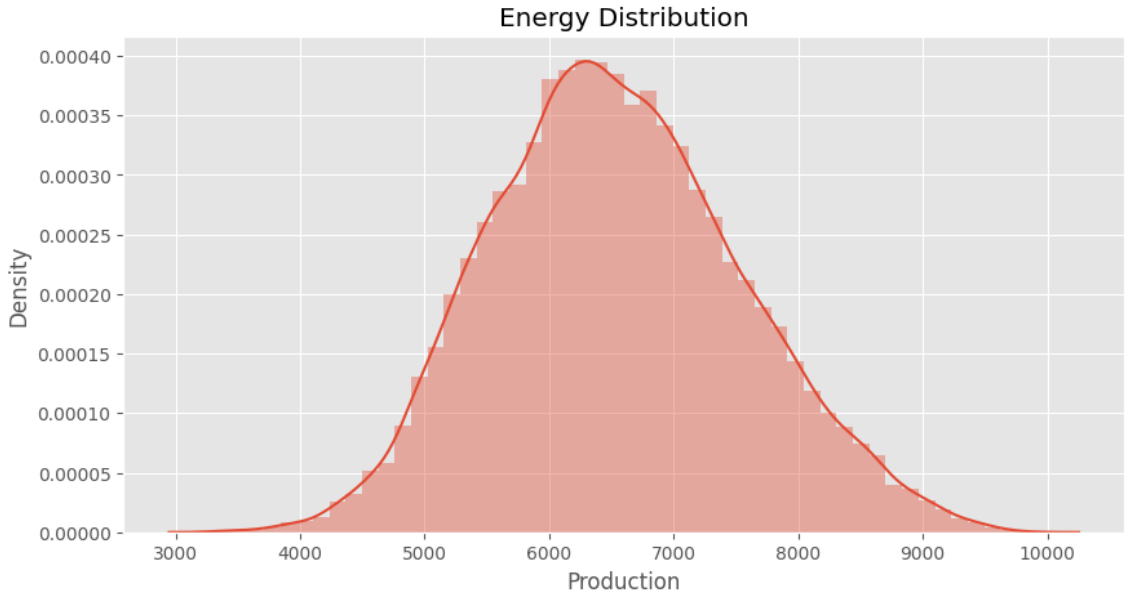

Figure 2: Energy distribution for the Romania Dataset

Hourly Energy Data for Turkey:

Data range: January 1, 2018 to December 31, 2023.

Details: Includes hourly time series data on consumption of energy and generation by type.

Importance: Provides a full overview of energy production and consumption patterns in Turkey, allowing for a thorough examination of the country's energy sector.

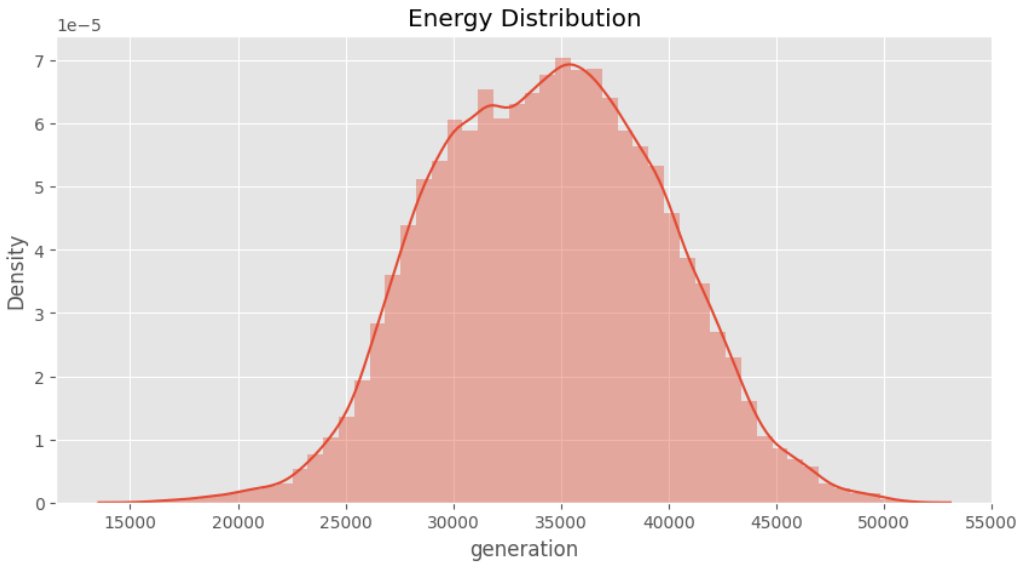

Figure 3: Energy distribution for the Turkey Dataset

We performed an Exploratory Data Analysis (EDA) to gather insights and better understand the patterns in the dataset.

The EDA procedure involved the following steps:

- **Data Visualization:**

  Time Series Plots: Plotted hourly electricity production is plotted to show trends and patterns across time.

  Box Plots: we used box plots to look at the distribution of energy production by day of the week, highlighting daily production changes.

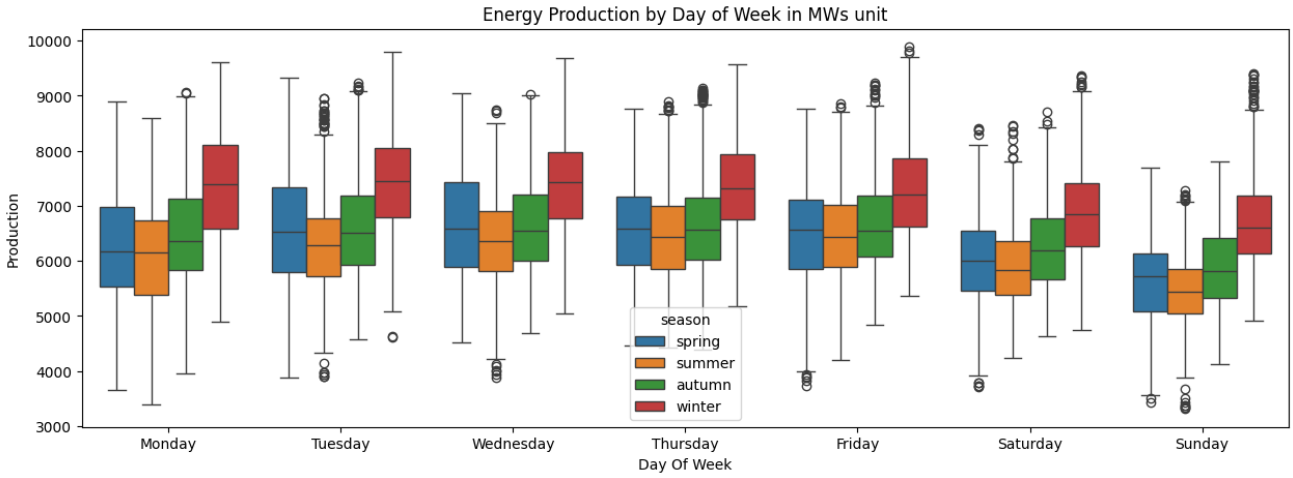

Figure 4: Box Plot for the Energy Production by Day of Week in MWs unit

- **Energy Data Profiling Report:** Generated an in-depth report summarizing key statistics, identifying abnormalities, and identifying missing and zero values in the dataset.
- **Data Cleaning:**

  Dropping Columns: Removed irrelevant and redundant columns like types and consumption.

  Handling Missing Values: Missing and zero values were addressed to ensure data quality.
- **Descriptive Statistics:** Means, medians, and standard deviations are calculated statistics that summarize the central tendencies and variances in data.
- **Feature Engineering:** To allow for more detailed analysis, we extracted additional time-based data such as month, year, date, time, week, and day of the week from the “datetime” column.
- **Data Splitting:** The dataset was divided into training and testing sets to prepare for model training and evaluation.

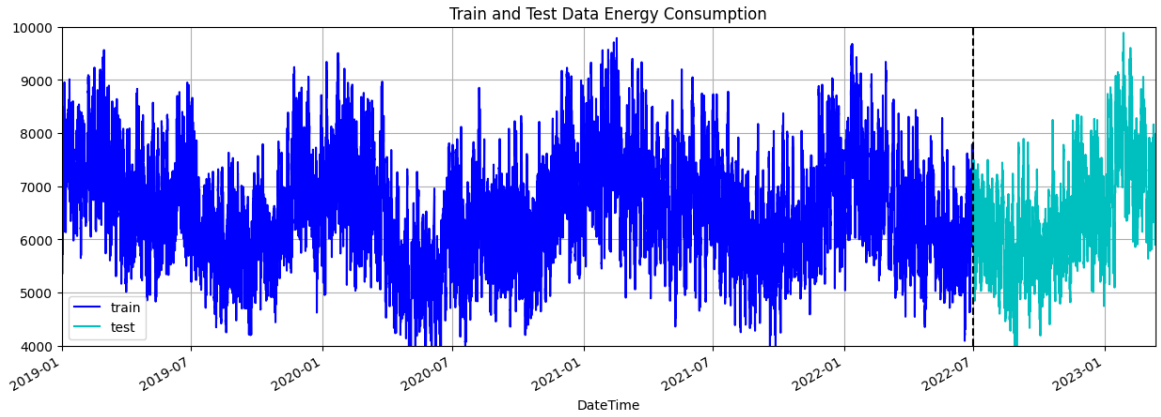

Figure 5: Train and Test Data Energy Consumption in Prophet model

The EDA process provided us with useful insights about the dataset's structures and properties, allowing us to efficiently prepare the data for the model prediction.

4.2 Languages and Packages including Libraries

Python Programming Language:

As for the language of the project, the major language used is Python which is an effective programming language. In the recent years Python programming language has rapidly grown in uses in industries such as machine learning, web development, and data analysis given the benefits that accompany it such as being easy to use, flexible and enjoys the backing of a large library support. Its extensive ecosystem of libraries

and frameworks, specifically designed for data manipulation, analysis, and model building, makes it an ideal choice for this project.

- **Pandas:** This library provides data structures like DataFrames, enabling efficient data loading, cleaning, transformation, and analysis. With the help of it we easily handled the time series data, performed data aggregation, and prepared the data for model input.
- **NumPy:** NumPy was essential for numerical computations, providing functionalities for array manipulation, mathematical operations, and linear algebra. We used it to scale and normalize the data, creating training and test datasets, to perform evaluations within our models.
- **Scikit-learn:** Scikit-learn offers a comprehensive suite of tools for data preprocessing, model training, evaluation, and traditional machine learning algorithms. We used Scikit-learn for tasks such as data splitting, feature scaling, model fitting, hyperparameter tuning, and performance metric calculation.
- **XGBoost:** This specialized library implements the XGBoost gradient boosting algorithm, known for its high performance and efficiency in predictive modeling tasks. We used it to train and evaluate the XGBoost model, optimizing its parameters for best performance on our energy production data.
- **TensorFlow:** TensorFlow, a robust open-source framework for deep learning, enabled us to build and train the LSTM model. It provided the building blocks for neural network construction, automatic differentiation capabilities for gradient-based optimization, and efficient computation on GPUs.
- **Keras:** Keras, a user-friendly high-level API running on top of TensorFlow, simplified the process of designing, training, and evaluating deep learning models. We utilized its intuitive interface to define the LSTM architecture, compile the model with appropriate optimizers and loss functions, and monitor its training progress.
- **Matplotlib & Seaborn:** These libraries provided the tools for data visualization, allowing us to create informative plots and charts to represent the data, model predictions, and performance metrics. We used them to generate plots showcasing the actual vs. predicted energy production, forecasting trends, and model comparisons.

- **Prophet:** This specialized library, developed by Facebook, implements the Prophet time series forecasting model. It provided a convenient framework for fitting the model to our data, generating forecasts, and analyzing its components.

4.3 List the Technical Features and Elements of the Project

This project incorporates a range of technical features and elements that contribute to its comprehensive approach to energy production forecasting:

- **Time Series Forecasting:** At its core, the project focuses on predicting future energy production values based on historical patterns and trends observed in the data. This involves understanding the different time dependencies and seasonality present in energy production.
- **Deep Learning:** The LSTM model leverages the power of deep learning, a subfield of machine learning, to capture complex non-linear relationships and long-term dependencies within the time series data. The model learns intricate patterns from historical data to make accurate predictions.
- **Ensemble Learning:** The Random Forest model employs ensemble learning, a technique where multiple decision trees are combined to improve prediction accuracy and generalization ability. It aggregates predictions from individual trees to form a more robust and reliable forecast.
- **Gradient Boosting:** The XGBoost model utilizes gradient boosting, a powerful ensemble learning method that sequentially trains weak learners (decision trees) and combines their predictions with weights to create a strong predictive model. It iteratively learns from the errors of previous models to enhance performance.
- **Model Evaluation:** We used evaluation metrics like Test and Train Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared to assess the accuracy and reliability of our forecasting models. These metrics quantify the model's performance by comparing its predictions to the actual energy production values.

4.4 Model Architecture:

Prophet Model:

- The Prophet model, designed for time series forecasting, decomposes data into three main components: Trend, seasonality, and holidays are other factors that are commonly encountered in most business environments. The trend component examines the long-term behavior of the series and different growth paths can be fitted. The seasonality sub-models seasonality at various levels, for instance, annual or weekly, by applying Fourier series. It also takes into consideration impacts

that are associated with holidays and other occasions. The use of Prophet is efficient in dealing with missing data, shift in trends, and the presence of outliers. However, it is relatively basic and may not pick up very intricate patterns from the data, as indicated by an RMSE test score of 1287. 015.

LSTM Model

- LSTMs are a type of recurrent neural networks that can be used to learn dependencies in time series. Each LSTM cell maintains a cell state controlled by input, forget, and output gates, allowing the network to selectively remember or forget information. This helps the LSTMs to capture complex time patterns. The steps taken included data pre-processing where data was normalized and data split into training and testing sets, LSTM layers configured, and the model trained. The LSTM model provided the minimum Train RMSE of 81. A Test RMSE was obtained as 6567 and a Test RMSE of 80. 2113, which moreover illustrates its ability to capture time patterns well.

Random Forest Model

- Random Forest is an ensemble learning method that constructs multiple decision trees and merges their results to improve accuracy. Each tree is trained only on a randomly sampled subset of the data and features, which brings the element of diversity into the model and improves its capacity to generalize. The final predictions are made by taking the average or by voting which is done individually in each tree. This model is especially suitable for data that are high dimensional and have non-linear dependencies. It involved using feature selection to divide data into training and testing sets and training several trees (n_estimators=100). While comparing the results of the Random Forest model, the RMSE scores were used.

XGBoost Model

- XGBoost is a type of boosting algorithm that constructs decision trees in a stepwise manner, and each tree in the sequence minimizes the errors of the previous tree. It avoids overfitting through regularization and performs well when handling large datasets. XGBoost models can learn about the feature interactions and can give high accurate energy forecasting. Similar to the Random Forest model, the data preparation included feature selection and normalization, model training was conducted with gradient boosting, and the resulting RMSE scores were comparable to the previous model.

- These models were used to forecast hourly electricity production, and the results of the models were assessed using RMSE. It helped to understand the strengths and limitations of each model and select the model that showed the best result on the Turkish dataset for implementation.

4.5 Unique Features of our Project

This project distinguishes itself by focusing on the comparative analysis of multiple forecasting models, on two different energy datasets from Romania and Turkey. This approach thus focuses on the model generalization capabilities in the real-world energy forecasting domains. Other works could dedicate their analysis to one or several models, or focus on one or several datasets, while this work offers a larger view of the topic by examining how various models behave in different contexts of energy production.

One unique aspect is that we used three various machine learning algorithms for the prediction of hourly energy production: LSTM, Random Forest, and XGBoost. By comparing the model's performance across different time stamps, namely seasonal and yearly, we gained deeper insights into the strengths and limitations of each model. This kind of approach allows us to identify the most suitable model for specific forecasting needs, ensuring more accurate and reliable predictions.

One of the major unique features is detailed year-wise and season-wise analysis, where we split the data and apply the models to compare trends across different periods. This type of analysis helps in identifying different time dependencies and seasonal patterns, providing valuable insights for improving forecasting accuracy.

Overall, the key strengths of our work are grounded in the comparative analysis, the multiple-model approach, and the detailed different time dependencies, which together help to progress the area of energy production prediction.

4.6 Alternative ways of implementing our Project

Alternative approaches for implementing energy production forecasting exist, each with its own strengths and limitations:

- **Statistical Time Series Models:** Models like Autoregressive Integrated Moving Average (ARIMA) and Seasonal ARIMA models are statistical models that capture dependencies and seasonality and are effective in forecasting. These models are especially useful when the variables on which the data is based exhibit seasonality and autocorrelation.

- **Other Machine Learning Algorithms:** Support Vector Regression (SVR), Artificial Neural Networks (ANN's) with different architectures and other machine learning algorithms. Every algorithm has its advantages thus, SVR is efficient in working with high-dimensional data, and at the same time, ANNs are suitable for modeling non-linear dependencies in the data.
- **Ensemble Learning Techniques:** Techniques like Bagging, Boosting, and Stacking can be used to enhance the prediction accuracy by combining the strengths of multiple models. For instance, the Random Forest and XGBoost can be used together to build an improved predictive model.
- **Hybrid Models:** Using statistical and machine learning strategies, it is possible to improve the quality of forecasting. For example, combining the linear models and seasonality to forecast the large-scale trends, while using machine learning models for the residuals.
- **Transfer Learning and Fine-Tuning:** Utilizing pre-trained models on related tasks and fine-tuning them on specific energy production data could improve forecasting performance. This approach leverages the knowledge gained from large datasets in related domains to enhance prediction accuracy with limited training data.

However, it is important to note that while the general solution provided in this project has been discussed, it is not assured that there will be a successful implementation of the same through the proposed alternatives. The training algorithm, the choice of the hyperparameters, the feature selection methods, and data pre-processing strategies significantly affect the accuracy and robustness of the models. To replicate the outcomes of a similar project, one would need to have access to the given datasets, the concrete instantiations of the models, and the justification for the decisions made.

4.7 Status of our project

The project has been successfully implemented using Google Colaboratory. Initial testing of individual components has shown promising results, and further refinement and optimization are ongoing before full-scale implementation. The first successful implementation occurred in May, 2024, at BML Munjal University, by Anish Borkar, Harsha Vardhan and Sai Nikhil.

5. RESULTS AND DISCUSSION

The training for the LSTM model on the Romania dataset is plotted in Figure 6. The train and validation loss is plotted and the decrease in both the parameters indicates no overfitting in the LSTM model. Being a regression task, the train RMSE and test RMSE obtained were 81.65667766792257 and 80.2113467138468

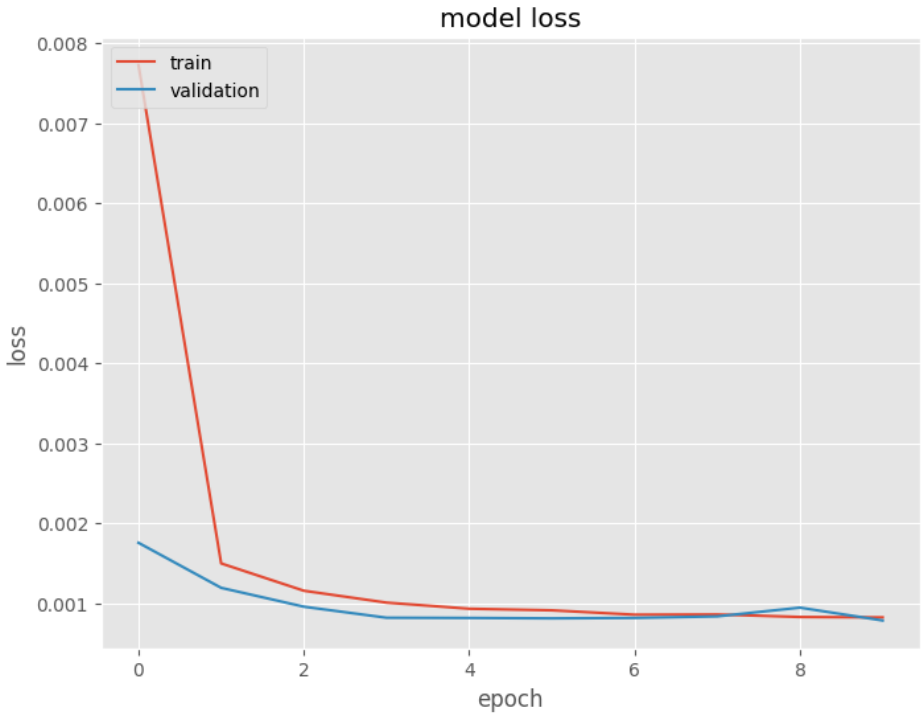

Figure 6: Loss Graph for LSTM on the Romania dataset

Figures 7 and 8 illustrate the actual vs train predictions and actual vs test predictions, respectively.

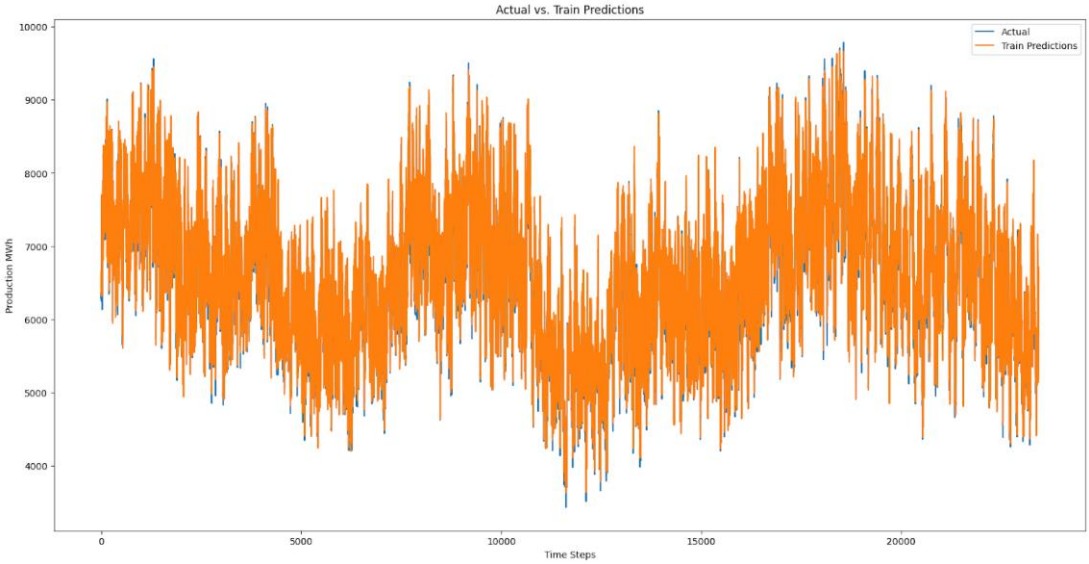

Figure 7: Actual vs train predictions

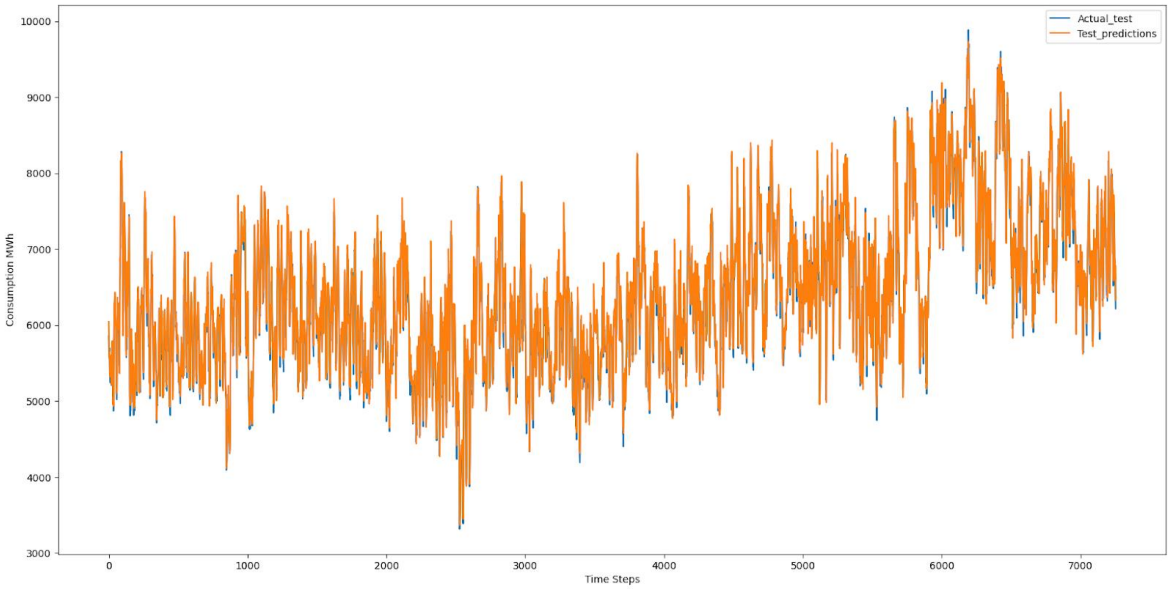

Figure 8: Actual vs test predictions

Figure 9 shows the production forecast for the next 30 days using LSTM model on the Romania dataset.

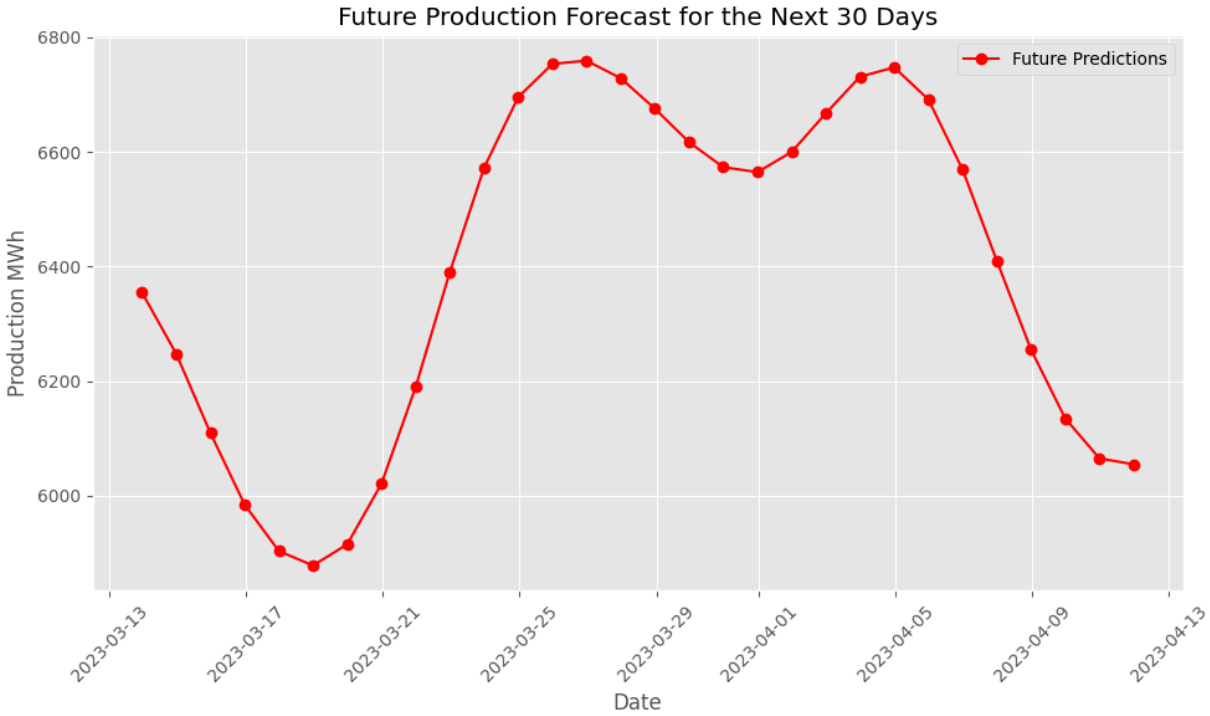

Figure 9: Production forecasting for next 30 days

Figure 10 shows the data forecasting done splitting year wise.

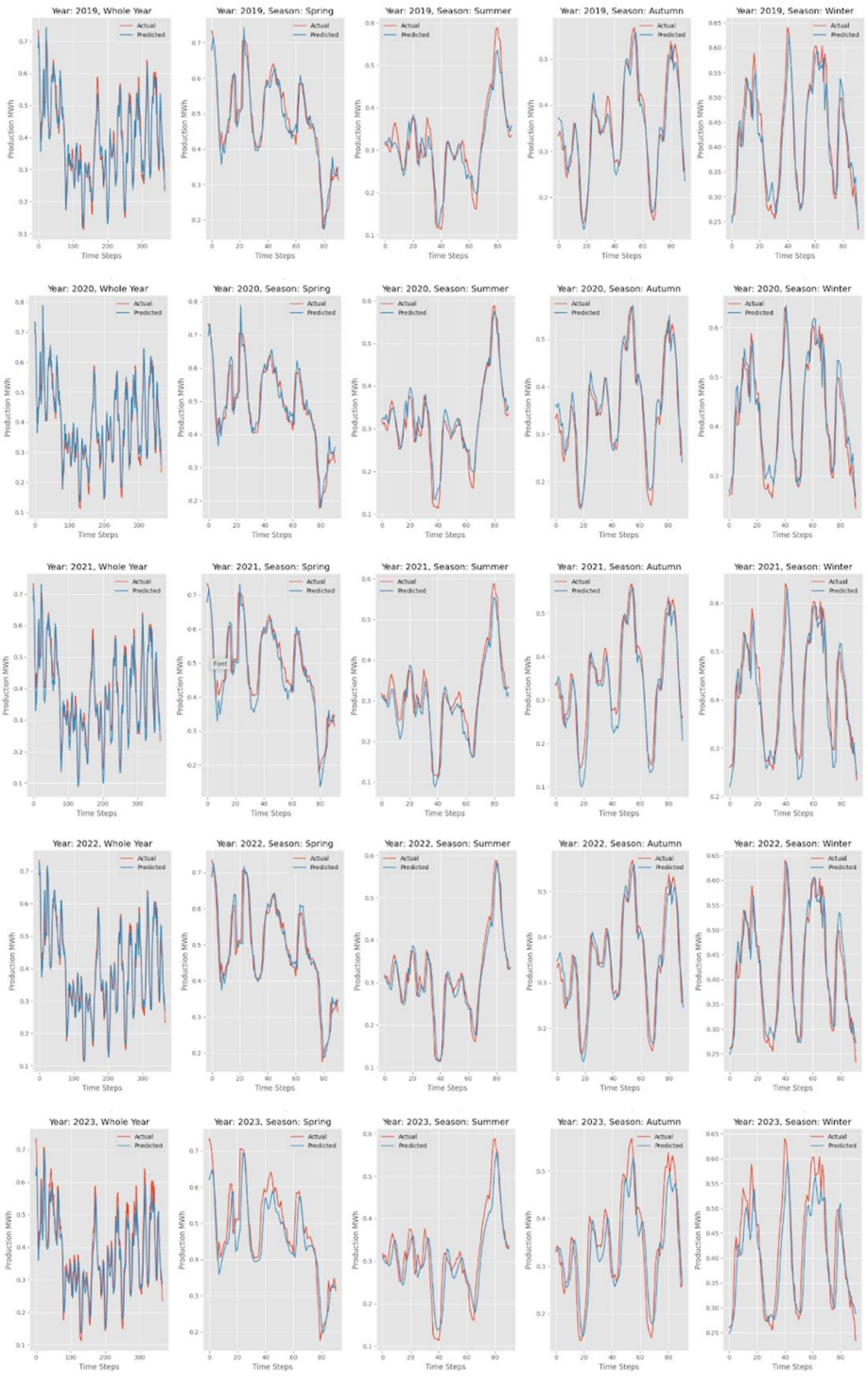

Figure 10: Forecasting done splitting year-wise

Figure 11 shows the data forecasting done splitting season wise.

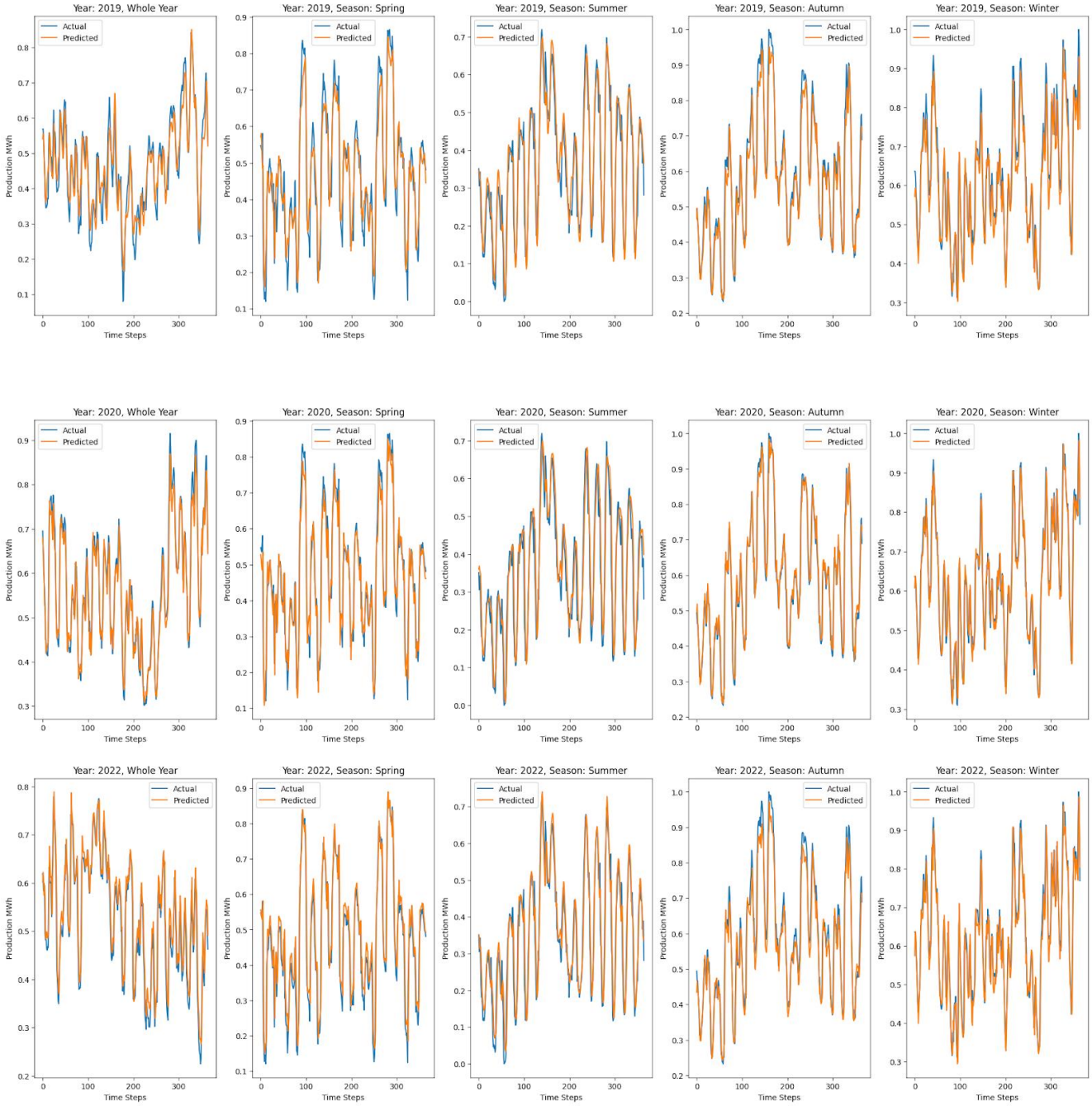

Figure 11: Forecasting done splitting season-wise

Figure 12 shows the comparison of LSTM predictions with Random Forest and XGBoost for the different year splits. Different time stamps are considered across the test data.

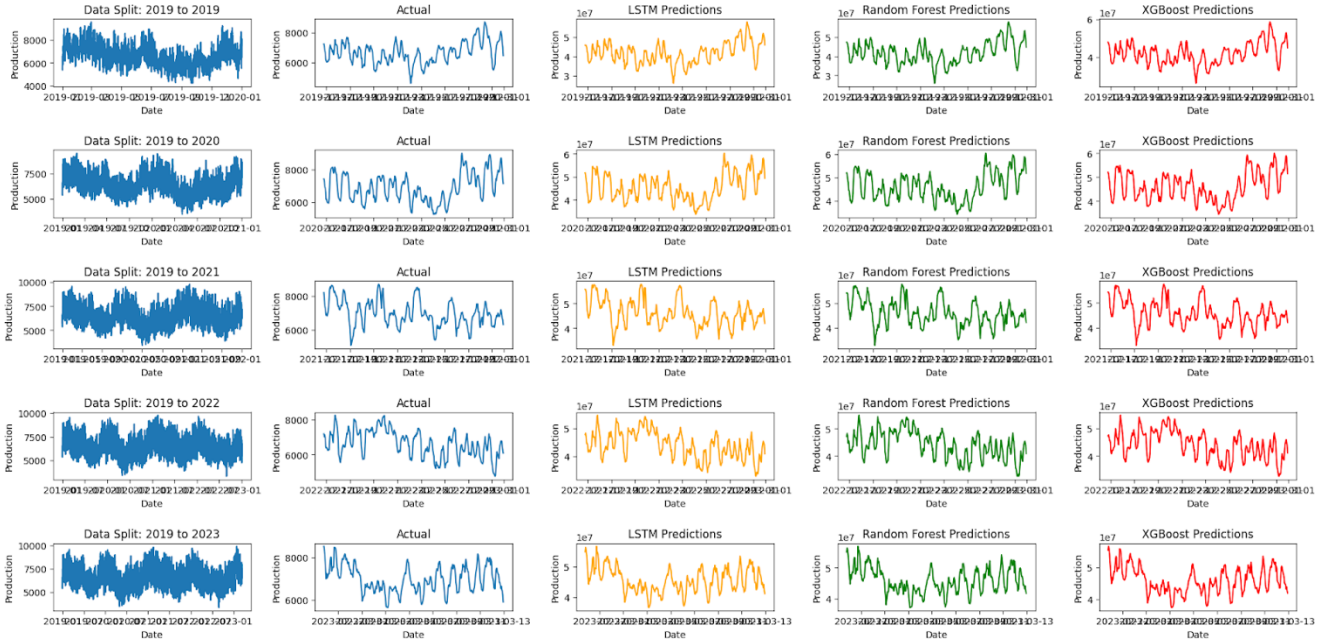

Figure 12: Comparison of LSTM with Multiple Models across the different time stamps by taking only test data

In the evaluation of LSTM and RMSE on the Turkey dataset, the Train and Test RMSE in the LSTM model came out as 187.46261792886517 and 191.46144719790988, respectively. In case of Random Forest, the RMSE came out as 4154.831780617651. All these results indicated the overfitting of the models on the Turkey dataset.

6. Conclusion and Future Scope

6.1 Conclusion

This project set out to develop robust and accurate short-term energy production forecasting models for two distinct European countries: Romania and Turkey. To achieve this, we implemented and compared three powerful machine learning algorithms: Long Short-Term Memory (LSTM) networks, XGBoost, and Random Forest. Each model was carefully trained and optimized using a detailed Romanian electricity production dataset that captures hourly generation patterns.

The evaluation results on the Romanian dataset revealed promising outcomes. All three models demonstrated a strong ability to forecast electricity production accurately.

Here's a summary of their performance:

- **LSTM:** Showed excellent capability in capturing temporal dependencies, resulting in highly accurate forecasts.
- **XGBoost:** Performed well with a good balance between speed and accuracy, leveraging its ability to handle non-linear relationships in the data.
- **Random Forest:** Provided robust predictions with a relatively lower risk of overfitting compared to more complex models like LSTM.

However, a critical observation emerged when we applied these trained models to the Turkish energy dataset. This dataset is broader, including not only production data but also consumption, pricing, and production details across various energy sources from 2018 to 2023.

The models displayed a clear tendency toward overfitting with the Turkish data, indicating a lack of generalization ability across different energy contexts. This crucial finding underscores a fundamental principle in machine learning: a model’s success in one context does not guarantee its effectiveness in another, even if the tasks appear similar.

Factors Contributing to Lack of Generalization

1. **Distinct Data Distributions:** The Romanian and Turkish datasets likely have different underlying data distributions. These reflect unique energy mixes, consumption patterns, and influencing factors within each country. The broader scope of the Turkish dataset may further contribute to these differences.

2. **Feature Relevance Variance:** Features that are crucial for accurate electricity production forecasting in Romania might not hold the same significance in Turkey. This difference can lead to models overemphasizing irrelevant information, hindering their ability to generalize.
3. **Model Complexity and Overfitting:** The inherent complexity of certain models, particularly LSTMs, can make them prone to overfitting, especially when trained on data with limited generalizability. In some cases, simpler models might offer better generalization.

By concluding this the advanced machine learning models such as LSTM, XGBoost, and Random Forest can achieve high accuracy in specific settings, their effectiveness can vary significantly across different datasets and energy landscapes. This emphasizes the necessity for careful consideration of data characteristics and feature relevance when developing and applying forecasting models.

6.2 Future Scope:

This project highlights the critical importance of understanding the relationship between model performance and data characteristics. While machine learning provides powerful tools for energy forecasting, it's crucial to recognize that models are not one-size-fits-all. We need to carefully consider data distributions, feature relevance, and model complexity to develop models that generalize well and provide accurate, reliable energy production forecasts across various energy landscapes.

Future research could explore techniques to improve model generalization, such as:

- **Transfer learning**: Using knowledge gained from the Romanian dataset to improve performance on the Turkish dataset.
- **Domain adaptation:** Adjusting models to perform well on data from different but related domains, like electricity production and overall energy production.
- **Ensemble methods:** Combining predictions from multiple models to enhance robustness and generalization.

By embracing these strategies, we can aim to develop energy production forecasting models that work well in various contexts, offering valuable insights and reliable predictions to help navigate the complexities of the ever-changing global energy landscape.

References

[1] Edoka, E. O., Abanihi, V. K., Amhenrior, H. E., Evbogbai, E. M. J., Bello, L. O., & Oisamoje, V. (2023). Time series forecasting of electrical energy consumption using deep learning algorithm. *Nigerian Journal of Technological Development*, *20*(3), 163-175.

[2] Gopi, R. R., & Annamalai, C. (2023). Day Ahead Energy Consumption Forecasting Through Time-Series Neural Network. *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, *11*(1), 164-179.

[3] Huang, X., Li, Q., Tai, Y., Chen, Z., Liu, J., Shi, J., & Liu, W. (2022). Time series forecasting for hourly photovoltaic power using conditional generative adversarial network and Bi-LSTM. *Energy*, *246*, 123403.

[4] Yukseltan, E., Yucekaya, A., & Bilge, A. H. (2020). Hourly electricity demand forecasting using Fourier analysis with feedback. *Energy Strategy Reviews*, *31*, 100524.

[5] Zhang, F., Deb, C., Lee, S. E., Yang, J., & Shah, K. W. (2016). Time series forecasting for building energy consumption using weighted Support Vector Regression with differential evolution optimization technique. *Energy and Buildings*, *126*, 94-103.

[6] Agrawal, R. K., Muchahary, F., & Tripathi, M. M. (2018, February). Long term load forecasting with hourly predictions based on long-short-term-memory networks. In *2018 IEEE Texas Power and Energy Conference (TPEC)* (pp. 1-6). IEEE.

[7] Wibawa, A. P., Fadhilla, A. F., Paramarta, A. K. A. I., Triono, A. P. P., Setyaputri, F. U., Akbari, A. K. G., & Utama, A. B. P. (2024). Bidirectional Long Short-Term Memory (Bi-LSTM) Hourly Energy Forecasting. In *E3S Web of Conferences* (Vol. 501, p. 01023). EDP Sciences.

[8] Pérez-Chacón, R., Asencio-Cortés, G., Martínez-Álvarez, F., & Troncoso, A. (2020). Big data time series forecasting based on pattern sequence similarity and its application to the electricity demand. *Information Sciences*, *540*, 160-174.

[9] Li, R., Jiang, P., Yang, H., & Li, C. (2020). A novel hybrid forecasting scheme for electricity demand time series. *Sustainable Cities and Society*, *55*, 102036.

[10] Kim, H., Park, S., & Kim, S. (2023). Time-series clustering and forecasting household electricity demand using smart meter data. *Energy Reports*, *9*, 4111-4121.