Indian Institute of Information Technology Allahabad
Prayagraj Uttar Pradesh, India

Real-Time Electricity Consumption Forecasting in Office Buildings using Ensemble Learning with Spark and Apache Kafka

Shubham Kumar(IIT2020007)[#1], Raj Chhari (IIT2020010)[#2], Shashikant Thakur(IIT2020024)[#3], Nilesh Singh (IIB2020038)[#4], Ankit Kumar (IIT2020011)[#5]

Table of Contents:

Abstract

Reducing energy costs and achieving sustainability goals depends on the effective management of electricity consumption in office buildings. To achieve this goal, ensemble learning techniques, specifically the random forest algorithm, offer an approach to predict electricity consumption in real time. In the face of increasing demand for energy use and sustainable business practices, accurate forecasting models can significantly improve resource utilization and minimize impact. The joint approach uses a number of learning algorithms instead of a single one to boost predictions. By combining their capabilities, it achieves a better overall prediction result and reduces the disadvantages of each individual model, leading to optimal performance and greater reliability.

The proposed solution uses Apache Spark for data processing and Apache Kafka for real-time data streaming. By combining this technology with ensemble learning algorithms, the approach aims to provide accurate and timely forecasts of electricity consumption in office buildings, leading to energy efficiency and cost-effectiveness results.

Introduction

There has been an increasing demand for effective prediction of power usage in office structures because of increased pressure on saving energy and environmental friendliness. With highly accurate forecasts it is possible for a building manager to take right and timely actions in their planning efforts, which could improve efficiency of an operation and lower cost. However, real-time forecasting is especially important in situations where one can adjust and respond quickly to any disproportions between actual demand and current predictions. This report presents in real-time the use of ensemble learning coupled with Apache Spark and Apache Kafka for predicting electricity consumption in an office building.

Proper control over electricity utilization in offices is one of the critical aspects of eco-efficient approaches and resources use rationalization. As more people seek effective use of power and a need to minimize environmental degradation through pollution, prediction of electricity usage holds much importance. However, many traditional forecasting approaches have limitations when trying to capture all the complicated and mutually exclusive factors that influence energy use patterns.

To address the above-mentioned challenges, a new approach of electrical consumption prediction in office buildings based on envelope learning using Random Forest algorithm is introduced in this article.

Ensemble learning is about combining various models together aimed at improving performance capabilities mainly in cases of predicting or handling diverse patterns in a dataset. In addition, the dynamic in character and multidimensional building-related data demands an approach that can respond to complex patterns or interactions during electrical predictions. Additionally

Efficient management in terms of energy efficiency and the overall sustainability effort involves electricity consumption management in office buildings. Correct prediction of electricity demand is essential due to the necessity to improve energy performance and reduce the burden on nature. In office buildings electricity prediction has its many determinants. For example,socio-economic change that other prediction methods cannot handle. Therefore this paper presents an innovative method of electricity forecasting in office buildings via ensemble learning particularly,Random Forest algorithm.

Literature Review

Electricity consumption is an issue that has caught the eye of researchers recently. Therefore, researchers have used different methods in order to solve this problem. In this section, we briefly present essential studies for Electricity Consumption Forecasting in an Office Building.

S. No	***Authors***	***Paper title***	***Description***	***Methodology***	***Result***
1.	Tiago Pinto , Isabel Praca , Zita Vale , Jose Silva	Ensemble learning for electricity consumption forecasting in office buildings	Importance of precise energy forecasting, existing model constraints, ensemble learning proposed as solution, comparative study with real office	The study creates and tests three ensemble models for short-term load forecasting in power systems, favoring the adapted Adaboost model for superior hour-ahead electricity consumption	The study proves ensemble learning's efficacy in improving short-term load forecasting for power systems. The adapted Adaboost consistently outperforms gradient boosted regression trees and random forests in office building data analysis.

S. No	*Authors*	*Paper title*	*Description*	*Methodology*	*Result*
			data evaluating various forecasting methods and influential factors.	predictions using real office building data.	
2.	Jeevantika Lingalwar	Improvising Processing of Huge Real Time Data Combining Apache Kafka and Spark Streaming	The paper emphasizes Cloud Computing's transformative impact on computing technologies, highlighting its role in data processing, storage, and internet security. It underscores the importance of real-time data processing, advocating Apache Spark Streaming and Apache Kafka integration for efficient handling of vast real-time data, showing that while Spark excels with large datasets, the combined framework's efficacy varies with dataset size.	This paper highlights efficient real-time data processing needs in Cloud Computing, addressing challenges of data overload on the World Wide Web. It proposes Apache Spark Streaming integrated with Apache Kafka, showing Spark's swift processing of large datasets and the Kafka-Spark framework's variable execution time based on dataset size, offering insights for optimizing real-time data processing. .	This study shows how Cloud Computing changes technology by handling data and internet security. It says processing data quickly is crucial and suggests using Apache Spark Streaming and Apache Kafka together for managing lots of real-time data. It found that Spark works great with big datasets, but the Spark-Kafka mix's success varies with dataset size.

S. No	*Authors*	*Paper title*	*Description*	*Methodology*	*Result*
3.	HOUDA DAKI, Asmaa El Hannani, Hassane OUAHMANE	Forecasting Electricity Consumption in a Moroccan Educational Institution	This research shows how predictive analytics helps ensure steady electricity in schools. It uses six years of data from a Moroccan school, including class schedules and weather info, to find the best way to predict energy use and save electricity.	This study assesses electrical energy consumption prediction models in educational institutions, focusing on El Jadida's National School of Applied Sciences, Morocco. Using a six-year dataset including schedules and weather data, it benchmarks forecasting models to find the best fit for managing electricity usage efficiently.	The research found that using predictive analytics is crucial for ensuring a stable electricity supply in schools. By examining six years of data from a Moroccan school, including schedules and weather details, effective methods for accurately predicting energy use were identified, aiming to improve energy management and reduce electricity consumption in educational settings.

Methodology

1. Data Collection and Streaming

In the first instance, electric power consumption data will be gathered by making use of different meters as well as sensors in the office block. This information will be sent through in real time, and it will use a software called Apache kakaa (distributed event streaming platform). Accordingly, Kafka supports efficient mechanisms of handling streams in distributed systems.

2. Data Preprocessing

When data is sent to Kafka, Apache Spark processes the data since it is a powerful data processing system. Preprocessing of data involves cleaning, normalization, and feature engineering. This is what spark does.4 This involves working on handling missing data, disposing outliers, and developing suitable features including meteorological conditions, morning/evening, and occupancy state that are proven to be major factors influencing power consumption.

3. Ensemble Learning Models

The ensemble learning enables one to come up with multiple machine-learning models that are used together to enhance the prediction capability and performance of the forecasting model. In this approach, we employ the following ensemble learning algorithms:

1.Bagging (Bootstrap Aggregating):

Example: Random Forest

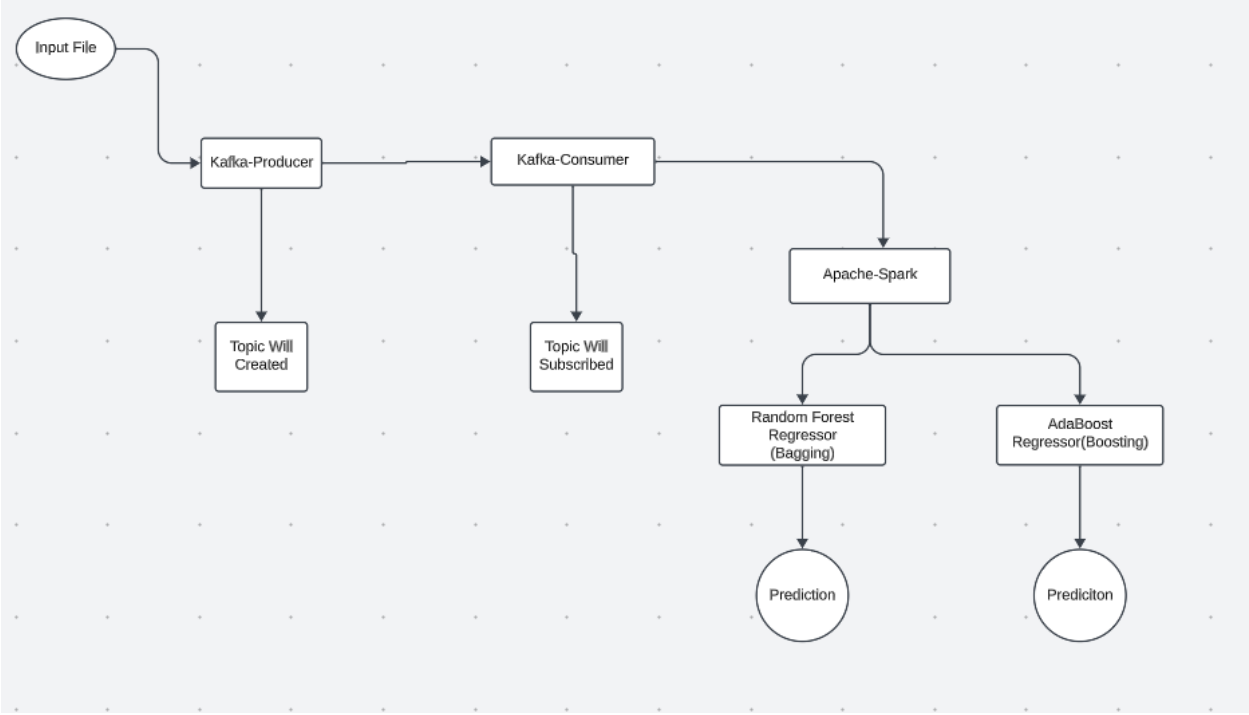
```
rf_regressor = RandomForestRegressor(n_estimators=100, random_state=42)
model = rf_regressor.fit(X_train, y_train)
y_pred = rf_regressor.predict(X_test)
```

2.Boosting Example:

AdaBoost(Adaptive Boosting)

```
# # Initializing the base regressor (Decision Tree in this case)
base_regressor = DecisionTreeRegressor()
adaboost_model = AdaBoostRegressor(base_regressor, n_estimators=100, learning_rate=0.1, random_state=42)
```

FlowChart



4. Model Training and Evaluation

The preprocessed data was split into training and test sets, utilizing historical data for training and recent data for testing. Subsequently, the training dataset is used for training the ensemble learning models that are later applied in the prediction of the testing dataset. Performance

evaluation measures like MAE, MSE, or RMSE are employed here with the aim to evaluate the effectiveness of the employed predictions.

5. Real-time Forecasting

Each time new data flows into Kafka, the ensemble learning models produce real-time predictions of electricity consumption. User-friendly dashboard provides building managers with these forecasts, which allow them to manage and adjust energy consumption.

Results and Benefits

The real-time approach for electricity consumption forecasting in office buildings using ensemble learning and Apache Spark with Apache Kafka integration offers several key benefits:

1. Accuracy: The ensemble learning models provide accurate forecasts, which enable precise control of electricity consumption.

2. Timeliness: Real-time streaming ensures that building managers have access to the most up-to-date consumption forecasts, enabling rapid response to any deviations from expected patterns.

3. Cost Reduction: Accurate forecasts allow for better load balancing and scheduling of energy-intensive operations, resulting in cost savings.

4. Sustainability: Improved energy management helps reduce the carbon footprint of the office building.

5. Scalability: The solution can be easily scaled to accommodate larger buildings or multiple sites, making it suitable for various office building configurations.

	Date	Time	Global_active_power	Global_reactive_power	\
count	10000.000000	10000.000000	10000.000000	10000.000000	
mean	13501.698800	43033.680000	1.744016	0.118801	
std	2.036392	24995.239152	1.339772	0.111437	
min	13498.000000	0.000000	0.194000	0.000000	
25%	13500.000000	21420.000000	0.388000	0.000000	
50%	13502.000000	42840.000000	1.478000	0.100000	
75%	13503.000000	64920.000000	2.560500	0.178000	
max	13505.000000	86340.000000	7.884000	0.724000	

	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	\
count	10000.000000	10000.000000	10000.000000	9998.000000	
mean	241.085660	7.376535	0.885277	2.079816	
std	3.673307	5.644298	5.479666	7.675109	
min	228.910000	0.800000	0.000000	0.000000	
25%	238.500000	1.800000	0.000000	0.000000	
50%	241.550000	6.200000	0.000000	0.000000	
75%	243.920000	10.800000	0.000000	1.000000	
max	249.480000	34.200000	40.000000	73.000000	

	Sub_metering_3
count	9998.000000
mean	8.242849
std	8.735647
min	0.000000
25%	0.000000
50%	0.000000
75%	17.000000
max	20.000000

Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3	features
13498	62640	4.216	0.418	234.84	18.4	0.0	1.0	17.0	[13498.0,62640.0]
13498	62700	5.36	0.436	233.63	23.0	0.0	1.0	16.0	[13498.0,62700.0]
13498	62760	5.374	0.498	233.29	23.0	0.0	2.0	17.0	[13498.0,62760.0]
13498	62820	5.388	0.502	233.74	23.0	0.0	1.0	17.0	[13498.0,62820.0]
13498	62880	3.666	0.528	235.68	15.8	0.0	1.0	17.0	[13498.0,62880.0]
13498	62940	3.52	0.522	235.02	15.0	0.0	2.0	17.0	[13498.0,62940.0]
13498	63000	3.702	0.52	235.09	15.8	0.0	1.0	17.0	[13498.0,63000.0]
13498	63060	3.7	0.52	235.22	15.8	0.0	1.0	17.0	[13498.0,63060.0]
13498	63120	3.668	0.51	233.99	15.8	0.0	1.0	17.0	[13498.0,63120.0]
13498	63180	3.662	0.51	233.86	15.8	0.0	2.0	16.0	[13498.0,63180.0]
13498	63240	4.448	0.498	232.86	19.6	0.0	1.0	17.0	[13498.0,63240.0]
13498	63300	5.412	0.47	232.78	23.2	0.0	1.0	17.0	[13498.0,63300.0]
13498	63360	5.224	0.478	232.99	22.4	0.0	1.0	16.0	[13498.0,63360.0]
13498	63420	5.268	0.398	232.91	22.6	0.0	2.0	17.0	[13498.0,63420.0]
13498	63480	4.054	0.422	235.24	17.6	0.0	1.0	17.0	[13498.0,63480.0]
13498	63540	3.384	0.282	237.14	14.2	0.0	0.0	17.0	[13498.0,63540.0]
13498	63600	3.27	0.152	236.73	13.8	0.0	0.0	17.0	[13498.0,63600.0]
13498	63660	3.43	0.156	237.06	14.4	0.0	0.0	17.0	[13498.0,63660.0]
13498	63720	3.266	0.0	237.13	13.8	0.0	0.0	18.0	[13498.0,63720.0]
13498	63780	3.728	0.0	235.84	16.4	0.0	0.0	17.0	[13498.0,63780.0]

only showing top 20 rows

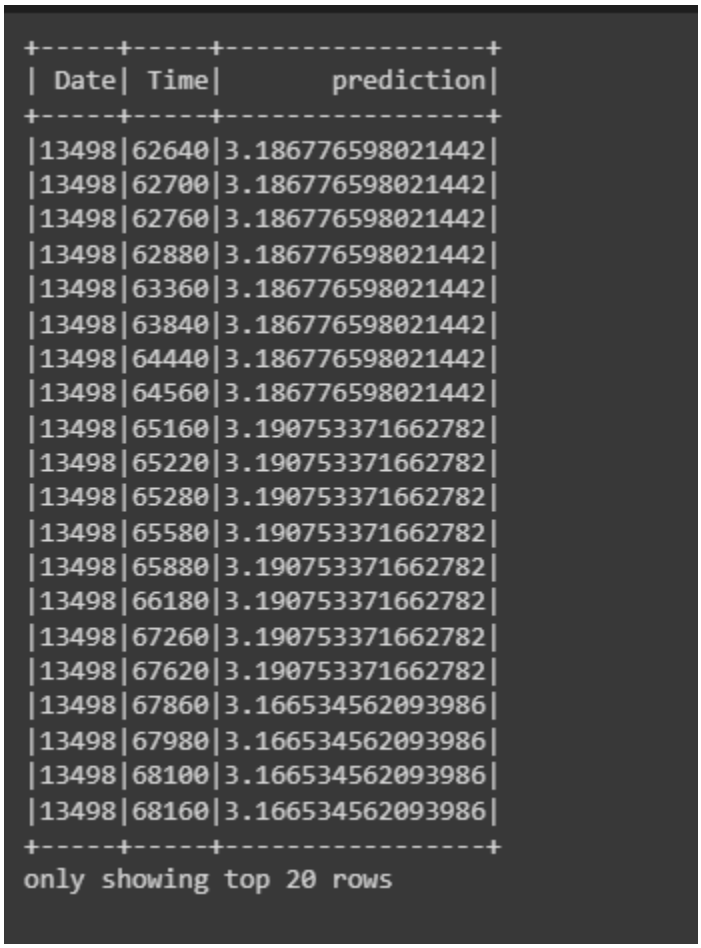
```
print(trainingData.show())
```

```
+-----+-----+------------------+--------------------+-------+---------------+-------------+-------------+-------------+--------------------+
| Date| Time|Global_active_power|Global_reactive_power|Voltage|Global_intensity|Sub_metering_1|Sub_metering_2|Sub_metering_3|            features|
+-----+-----+------------------+--------------------+-------+---------------+-------------+-------------+-------------+--------------------+
|13498|62820|             5.388|               0.502| 233.74|           23.0|          0.0|          1.0|         17.0|[13498.0,62820.0]|
|13498|62940|              3.52|               0.522| 235.02|           15.0|          0.0|          2.0|         17.0|[13498.0,62940.0]|
|13498|63000|             3.702|                0.52| 235.09|           15.8|          0.0|          1.0|         17.0|[13498.0,63000.0]|
|13498|63060|               3.7|                0.52| 235.22|           15.8|          0.0|          1.0|         17.0|[13498.0,63060.0]|
|13498|63120|             3.668|                0.51| 233.99|           15.8|          0.0|          1.0|         17.0|[13498.0,63120.0]|
|13498|63180|             3.662|                0.51| 233.86|           15.8|          0.0|          2.0|         16.0|[13498.0,63180.0]|
|13498|63240|             4.448|               0.498| 232.86|           19.6|          0.0|          1.0|         17.0|[13498.0,63240.0]|
|13498|63300|             5.412|                0.47| 232.78|           23.2|          0.0|          1.0|         17.0|[13498.0,63300.0]|
|13498|63420|             5.268|               0.398| 232.91|           22.6|          0.0|          2.0|         17.0|[13498.0,63420.0]|
|13498|63480|             4.054|               0.422| 235.24|           17.6|          0.0|          1.0|         17.0|[13498.0,63480.0]|
|13498|63540|             3.384|               0.282| 237.14|           14.2|          0.0|          0.0|         17.0|[13498.0,63540.0]|
|13498|63600|              3.27|               0.152| 236.73|           13.8|          0.0|          0.0|         17.0|[13498.0,63600.0]|
|13498|63660|              3.43|               0.156| 237.06|           14.4|          0.0|          0.0|         17.0|[13498.0,63660.0]|
|13498|63720|             3.266|                 0.0| 237.13|           13.8|          0.0|          0.0|         18.0|[13498.0,63720.0]|
|13498|63780|             3.728|                 0.0| 235.84|           16.4|          0.0|          0.0|         17.0|[13498.0,63780.0]|
|13498|63900|             7.706|                 0.0| 230.98|           33.2|          0.0|          0.0|         17.0|[13498.0,63900.0]|
|13498|63960|             7.026|                 0.0| 232.21|           30.6|          0.0|          0.0|         16.0|[13498.0,63960.0]|
|13498|64020|             5.174|                 0.0| 234.19|           22.0|          0.0|          0.0|         17.0|[13498.0,64020.0]|
|13498|64080|             4.474|                 0.0| 234.96|           19.4|          0.0|          0.0|         17.0|[13498.0,64080.0]|
|13498|64140|             3.248|                 0.0| 236.66|           13.6|          0.0|          0.0|         17.0|[13498.0,64140.0]|
+-----+-----+------------------+--------------------+-------+---------------+-------------+-------------+-------------+--------------------+
only showing top 20 rows
```

AdaBoost

	Code	n_estimators	learning_rate	MSE	R-squared
0	Code 1	100	0.1	0.247019	0.892165
1	Code 2	50	Default	0.258944	0.838683
2	Code 3	150	0.01	0.245315	0.899805
3	Code 4	50	Default	0.258944	0.838683

Final Prediction



```
+-----+-----+-----------------+
| Date| Time|       prediction|
+-----+-----+-----------------+
|13498|62640|3.186776598021442|
|13498|62700|3.186776598021442|
|13498|62760|3.186776598021442|
|13498|62880|3.186776598021442|
|13498|63360|3.186776598021442|
|13498|63840|3.186776598021442|
|13498|64440|3.186776598021442|
|13498|64560|3.186776598021442|
|13498|65160|3.190753371662782|
|13498|65220|3.190753371662782|
|13498|65280|3.190753371662782|
|13498|65580|3.190753371662782|
|13498|65880|3.190753371662782|
|13498|66180|3.190753371662782|
|13498|67260|3.190753371662782|
|13498|67620|3.190753371662782|
|13498|67860|3.166534562093986|
|13498|67980|3.166534562093986|
|13498|68100|3.166534562093986|
|13498|68160|3.166534562093986|
+-----+-----+-----------------+
only showing top 20 rows
```

Conclusion

In conclusion, the method of predicting electricity usage in office buildings in time using a combination of learning, Apache Spark and Apache Kafka is a highly effective and adaptable solution. It empowers building managers with timely insights to optimize energy consumption. By utilizing learning techniques alongside the real time capabilities of Apache Kafka and Spark this approach offers a solution for enhancing energy efficiency, reducing costs and contributing to a more sustainable future. I would recommend it to any organization that wishes to improve their electricity consumption forecasting, in office buildings.

References

[1]https://www.researchgate.net/publication/341251832_Ensemble_Learning_for_Electricity_Consumption_Forecasting_in_Office_Buildings
[2] https://norma.ncirl.ie/4249/1/jeevantikalingalwar.pdf
[3] https://www.researchsquare.com/article/rs-248534/v1
[4]https://site.ieee.org/pes-iss/data-sets/
[5] https://www.sciencedirect.com/science/article/pii/S0925231220307372