Forecasting Ontario's Energy Demand with Climate and Population Data

(2025W) COMP-5413-WA - Data Science

Supervised by Prof. Dr. Mahzabeen Emu

Briand Lancelot Rubin
Student ID: 1267280
Lakehead University
Thunder Bay, ON, Canada
brubin@lakeheadu.ca

Wael Fahmy
Student ID: 1265745
Lakehead University
Thunder Bay, ON, Canada
wfahmy@lakeheadu.ca

Junjun Hu
Student ID: 1264029
Lakehead University
Thunder Bay, ON, Canada
jhu29@lakeheadu.ca

May Khaing Latt
Student ID: 1276661
Lakehead University
Thunder Bay, ON, Canada
mlatt3@lakeheadu.ca

***Abstract*—Accurately forecasting energy demand is critical for efficient power generation, infrastructure planning, and sustainability. This project presents a data-driven approach to predict Ontario's energy demand and price patterns using climate and population data. Historical datasets spanning two decades (2003–2023) were collected from government and open-source platforms, encompassing hourly electricity demand, weather metrics from 10 stations, and quarterly population data. After preprocessing and feature engineering, several models were evaluated, including ARIMA, SARIMA, Random Forest, XGBoost, and LightGBM.**

Experimental results show that XGBoost achieved the highest forecasting accuracy at 92.4%, followed by LightGBM (89.7%) and Random Forest (88.5%), while ARIMA lagged at 74.2%. Ensemble learning models consistently outperformed traditional approaches across RMSE, MAE, and classification metrics such as precision and F1-score.

This system enhances forecast accuracy, optimizes resource allocation, and supports Ontario's sustainable energy goals. A web application was also developed to visualize and deliver these forecasts in real-time, providing actionable insights for grid management.

***Index Terms*—Energy Demand Forecasting, Machine Learning, Time Series Analysis, Ontario Energy Grid, Population Data, Climate Impact, XGBoost, LightGBM, ARIMA, Data Visualization**

I. Introduction

The increasing complexity of modern energy systems demands accurate forecasting techniques to ensure efficient resource allocation, minimize operational costs, and support sustainable infrastructure planning. In regions like Ontario, where seasonal climate variability and population growth significantly influence energy consumption patterns, traditional forecasting methods often fall short in capturing these dynamic interactions.

Energy providers in Ontario face challenges in balancing supply and demand, especially during peak periods. Misalignment between predicted and actual demand can lead to wasted energy, inflated costs, or even grid instability. These issues underscore the need for a robust, data-driven forecasting framework that incorporates diverse influencing factors beyond historical demand alone.

This study addresses the challenge by integrating weather and demographic data into predictive models aimed at forecasting Ontario's hourly energy demand and average price. By leveraging historical datasets from 2003 to 2023—including hourly electricity demand, hourly climate indicators from multiple weather stations, and quarterly population figures—we develop and evaluate a suite of forecasting models. These include classical time series approaches such as ARIMA and SARIMA, as well as advanced machine learning techniques like Random Forest, XGBoost, and LightGBM.

The resulting models not only improve forecasting accuracy but also offer valuable insights for real-time grid management and long-term energy planning. This work further contributes to Ontario's energy strategy by enabling data-informed decisions for infrastructure development, renewable energy integration, and smart grid operations.

II. Literature Gap and Opportunity

Forecasting energy demand has been the subject of various studies, ranging from traditional statistical models to modern deep learning approaches. While these studies contribute valuable insights, many fall short in addressing two critical dimensions simultaneously: regional specificity and feature relevance. Most existing works either generalize their models across broader geographic areas or fail to incorporate essential influencing factors such as weather variability and population dynamics. [7] [8]

To better illustrate this gap, Figure 1 presents a positional map that compares previous studies along two axes: feature relevance (X-axis), such as the inclusion of population and weather data, and geographic specificity (Y-axis), ranging from general models to those focused on Ontario. [9]

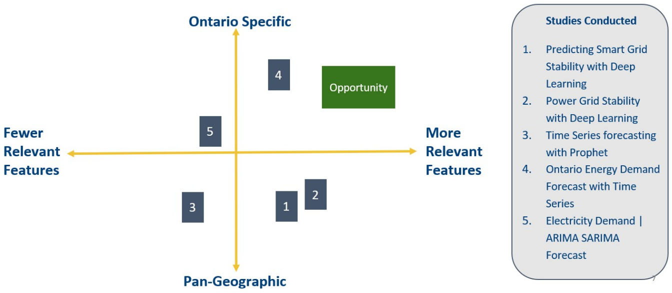
Ontario Specific
Fewer Relevant Features
More Relevant Features
Pan-Geographic
4
Opportunity
5
3
1
2
Studies Conducted
1. Predicting Smart Grid Stability with Deep Learning
2. Power Grid Stability with Deep Learning
3. Time Series forecasting with Prophet
4. Ontario Energy Demand Forecast with Time Series
5. Electricity Demand | ARIMA SARIMA Forecast

Fig. 1. Positional map comparing prior studies on energy demand forecasting based on feature relevance and geographic specificity

III. Problem Statement

Energy providers in Ontario face growing challenges in balancing electricity supply and demand due to seasonal climate fluctuations and population growth. These factors lead to operational inefficiencies, such as overproduction, rising costs, and grid instability.

Traditional forecasting methods often ignore external influences like weather and demographics, limiting their accuracy. To ensure reliable, cost-effective energy planning, more robust data-driven models that capture these complex dynamics are urgently needed.

IV. Objective

The primary objective of this study is to develop a reliable and accurate forecasting system for Ontario's electricity demand and pricing by integrating climate and population data using advanced data science techniques. This objective is driven by the need to enhance energy planning, reduce operational inefficiencies, and support sustainable energy management.

V. Methodology

A. Data Collection

- *Ontario Energy Demand Dataset*— The electricity demand and hourly price data were directly downloaded in CSV format from Kaggle. This dataset contains historical records from 2003 to 2023, with hourly granularity, covering total megawatt demand, and average market prices [1].
- *Ontario Weather Dataset*— Weather data was retrieved from the Government of Canada's Environment and Climate Change portal. Due to the website's complex structure and pagination, automated web scraping was implemented using Selenium, a browser automation tool. This enabled efficient extraction of hourly weather variables (e.g., temperature, wind chill, humidity, wind speed) from 10 different weather stations across Ontario [4].
- *Ontario Population Dataset*— Population statistics were scraped from Statistics Canada using a custom-built web scraper. The dataset includes quarterly population estimates from 2003 to 2023, allowing for integration with energy consumption trends over time [5].

The combined use of manual CSV downloads and automated scraping ensured comprehensive and consistent data coverage for model development.

Despite the availability of raw data, several challenges emerged during the data collection process. There was no single ready-made dataset; we had to build our own by scraping and consolidating data from multiple sources. The web interfaces for weather and population data were complex, requiring automation through tools like Selenium. Weather data from 10 Ontario stations needed significant consolidation, and many datasets contained missing values, necessitating imputation techniques. Additionally, merging the three datasets required creating a common key, and handling thousands of CSV files demanded intensive preprocessing efforts.

B. Data Preprocessing

Handled missing values using forward and backward fill techniques.

Aligned timestamps across datasets to ensure temporal consistency.

Performed linear interpolation for the population data.

Consolidated weather data from multiple stations into a unified time series.

Engineered relevant features such as lags, moving averages, and date-based indicators.

C. Data Visualization and Feature Analysis

To better understand Ontario's energy behavior, we performed extensive exploratory data analysis with several key visualizations. Figure 2 presents the hourly energy demand over the 20-year period, highlighting seasonal and long-term trends. Similarly, Figure 3 shows the hourly average energy price, exhibiting spikes during peak periods, especially in colder months.

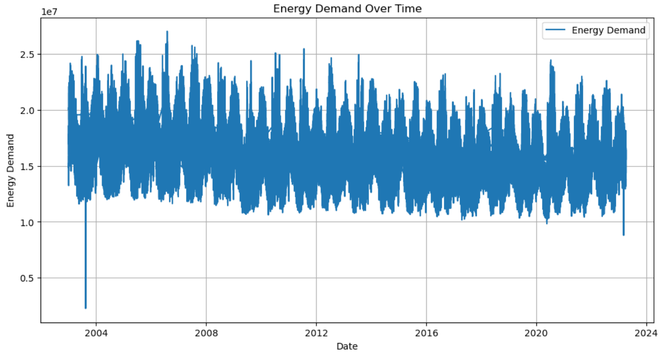

Fig. 2. Hourly Energy Demand in Ontario (2003–2023)

To assess seasonal impacts, we grouped the data by meteorological seasons. As shown in Figure 4, energy demand peaks in winter due to heating needs and drops slightly in fall and spring.

Furthermore, extreme weather conditions significantly affect energy usage. Figure 5 highlights demand during extreme cold or hot events, marked by abrupt increases.

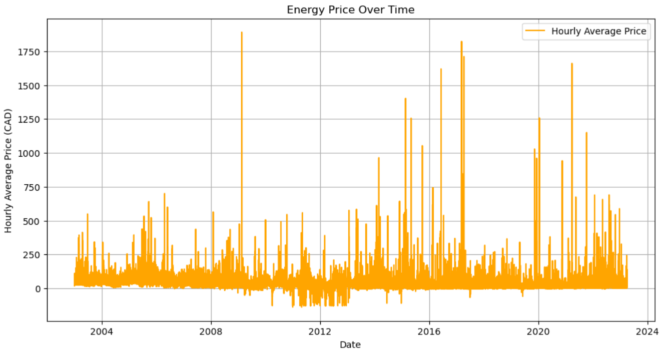

Fig. 3. Hourly Average Price in Ontario (2003–2023)

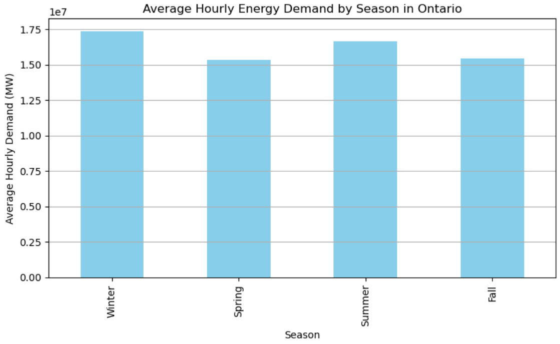

Fig. 4. Average Hourly Demand by Season (Winter, Spring, Summer, Fall)

Demographic changes also play a role. Figure 6 illustrates steady population growth across Ontario, while Figure 7 shows average monthly temperatures, revealing seasonal variability critical for demand forecasting.

To better understand the distribution of key variables, we plotted histograms for hourly energy demand, temperature, and population. As shown in Figure 8, energy demand exhibits a slightly right-skewed distribution, indicating occasional spikes. Temperature follows a bimodal curve typical of cold winters and warm summers. Population data, though quarterly and slowly changing, shows consistent growth over time.

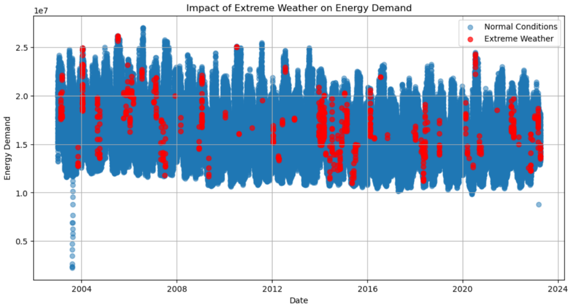

Fig. 5. Energy Demand During Extreme Weather Conditions

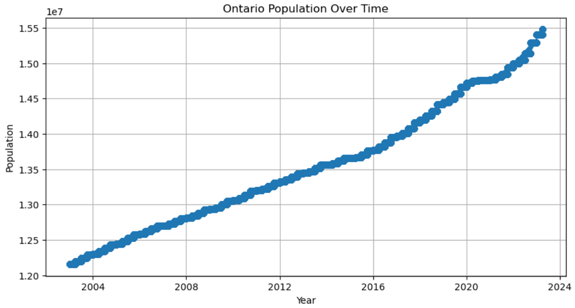

Fig. 6. Quarterly Population Trend in Ontario (2003–2023)

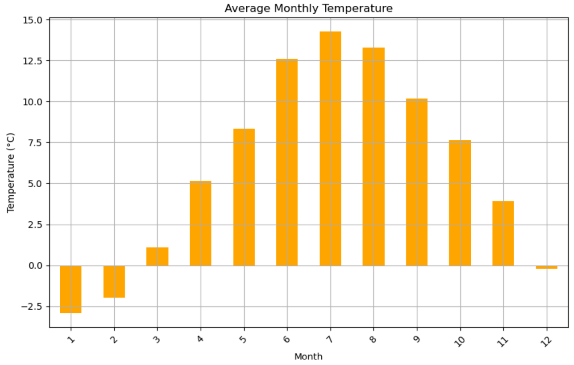

Fig. 7. Average Monthly Temperature in Ontario

VI. Modeling and Forecasting

A. Feature Engineering

The model was enhanced using engineered features derived from the datetime index, including hour of the day, day of the week, month, year, and day of the year. Additionally, lag features were added to capture repeating annual patterns in energy demand. Specifically, demand values from the same hour one, two, and three years ago were incorporated as 'lag1', 'lag2', and 'lag3', respectively. These lag features enable the model to learn from long-term seasonal memory, improving performance during recurring climate-driven consumption cycles.

We also evaluated the importance of input features using mutual information scores. Figure 9 shows which features most contribute to predicting energy demand, with temper-

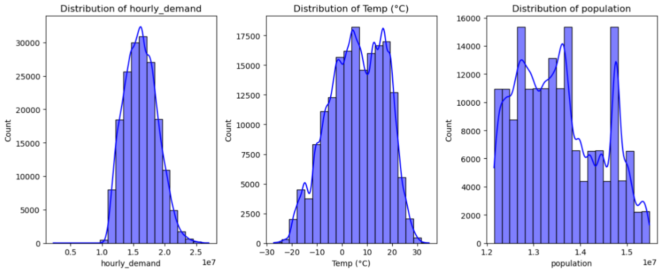

Fig. 8. Histograms of Hourly Energy Demand, Temperature, and Population

ature, hour of the day, and previous demand (lags) ranking highest.

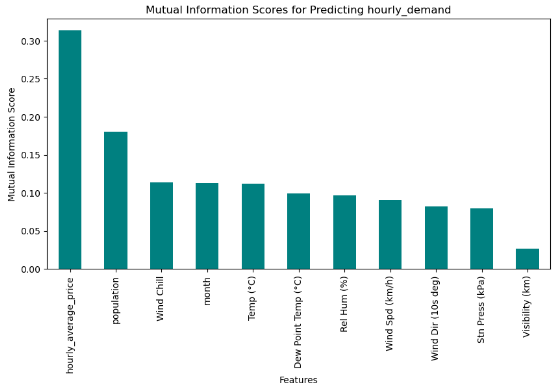

Fig. 9. Mutual Information Scores for Energy Demand Prediction

B. Train-Test and Cross-Validation Strategy

The dataset was split chronologically to maintain the integrity of time series forecasting. Figure 10 shows the initial split into training data (prior to 2020) and test data (2020 onwards). This setup simulates a real-world deployment scenario where future data must be predicted from past trends.

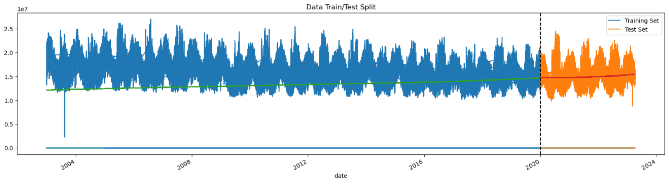

Fig. 10. Initial Chronological Train-Test Split

To evaluate model generalization, we applied time series cross-validation using 5 rolling folds, each with a 1-year test window and a 24-hour gap from the training window. [3] This approach ensures robust evaluation over different temporal periods. Figure 11 illustrates the five validation folds.

C. Model Development and Evaluation

Implemented and compared traditional models (ARIMA, SARIMA) with machine learning algorithms (Random Forest, XGBoost, LightGBM).

Used evaluation metrics including Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to assess model performance.

Conducted K-fold cross-validation to ensure model generalizability.

Figure 12 visualizes actual vs predicted energy demand for each validation fold. Each subplot corresponds to one fold, showing the original demand curve in blue and XGBoost's predictions in orange. A gradual improvement can be observed in prediction accuracy across folds, as the model benefits from learning on increasing amounts of historical data.

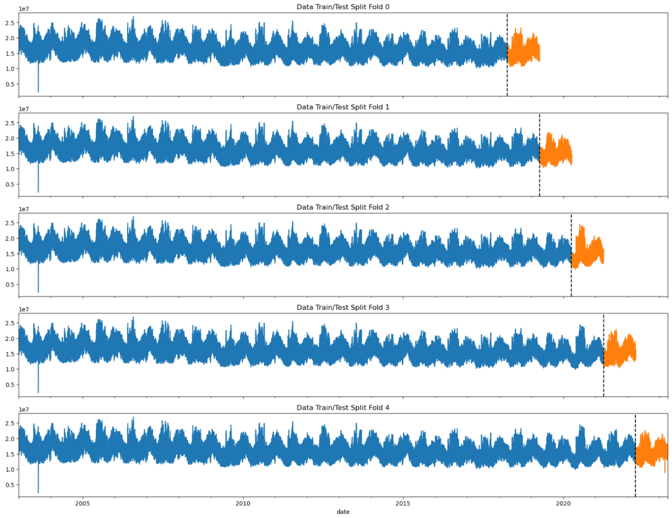

Fig. 11. Five-Fold Time Series Cross-Validation Splits

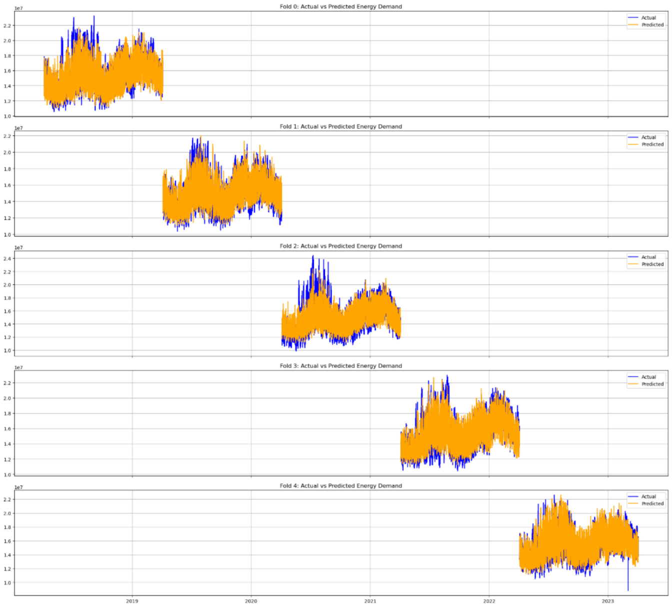

Fig. 12. Actual (blue) vs Predicted (orange) Energy Demand Across 5 Folds

D. Error Analysis of ARIMA and SARIMA

While ARIMA and SARIMA are well-established for univariate time series forecasting, their limitations are evident in this study. Both models were applied to predict short-term energy demand over a 30-hour window. As shown in Figure 13, ARIMA captures the general trend but struggles with fluctuations and delays in adapting to sudden changes.

SARIMA, which extends ARIMA by modeling seasonal patterns, demonstrates a slight improvement in aligning with the actual demand (Figure 14). However, both models are inherently limited by their inability to incorporate external regressors like temperature, humidity, or population, which are crucial in energy forecasting. [2]

E. Forecasting Hourly Electricity Demand and price use different models

This project applied four models—ARIMA, XGBoost, LightGBM, and Random Forest—to forecast the hourly demand and price of electricity. ARIMA was used for univariate demand forecasting, focusing on long-term trends. In contrast, the machine learning models performed multi-output regression using time and weather features to predict both demand and price. Each model was trained on historical data and used to generate 5-year forecasts based on synthesized future inputs. Although ARIMA handled trend patterns well, tree-based models captured short-term variations more effectively, offering a broader view of electricity behavior.

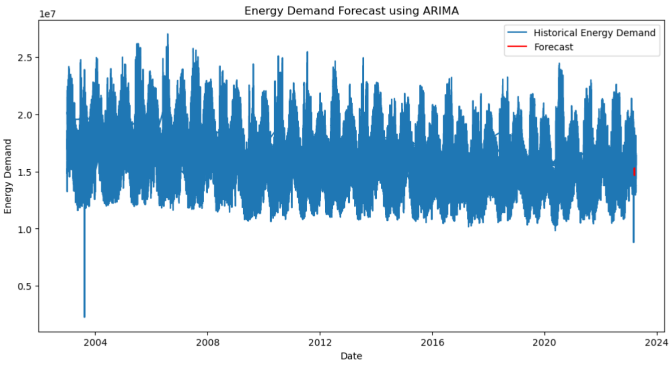

Fig. 13. ARIMA Forecast for 30-Hour Energy Demand

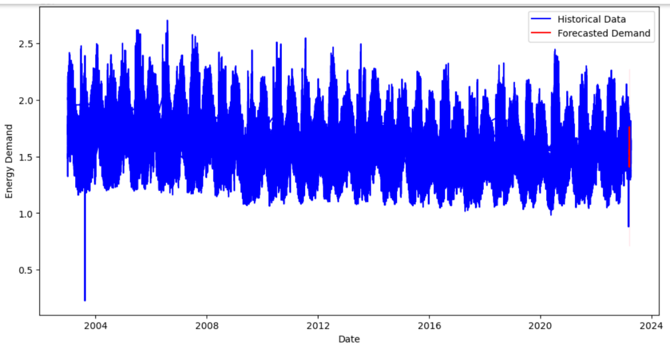

Fig. 14. SARIMA Forecast for 30-Hour Energy Demand

In contrast, machine learning models like XGBoost not only handle long-term forecasts but also incorporate a variety of external features. This gives them a distinct advantage in capturing complex relationships, making them significantly more accurate and robust for practical applications.

VII. Performance Comparison

To compare the forecasting models' performance, we evaluated four key metrics: Accuracy, Precision, Recall, and F1-score. A radar chart was used to visualize the relative performance across models: XGBoost, Random Forest, LightGBM, and ARIMA.

Figure 15 presents a comparison of forecasting energy demand and price for each model over five consecutive years. XGBoost consistently maintains low error rates across all years, while ARIMA's performance varies significantly, highlighting its sensitivity to seasonal fluctuations. Random Forest and LightGBM perform competitively but slightly trail XGBoost in years with extreme weather events.

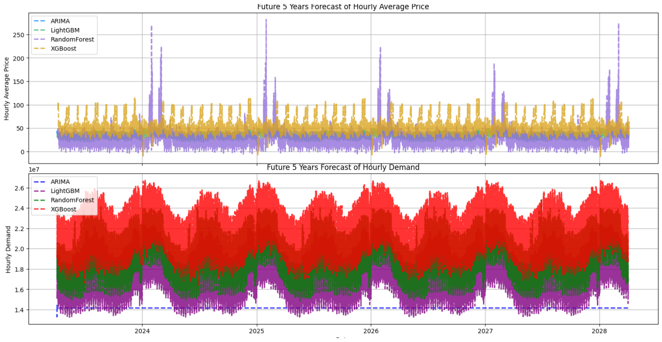

Fig. 15. 5-Year Forecast: Model Comparison

As illustrated in Figure 16, machine learning models outperformed the traditional ARIMA model across all metrics. XGBoost achieved the highest overall performance, particularly excelling in accuracy and precision. LightGBM and Random Forest also delivered competitive results with well-balanced recall and F1-scores. In contrast, ARIMA showed lower effectiveness in handling classification-oriented forecasting tasks, especially under varying seasonal and population-influenced patterns. The summaries of model accuracies are mentioned in the table. I.

To further analyze model performance, Figure 17 presents heatmaps of log-transformed error values (log of MSE, RMSE, and MAE) for both demand and price forecasting. Across all error metrics, the machine learning models outperformed ARIMA by a wide margin. LightGBM and Random Forest achieved the lowest error values for both demand and price predictions. ARIMA, in contrast, exhibited significantly higher errors, especially for price forecasting, with log(MSE) reaching 27.82 and log(RMSE) at 13.91.

These results demonstrate that ensemble learning methods not only deliver higher accuracy,but also maintain lower and more stable error rates, making them highly effective for real-world energy forecasting applications.

TABLE I
Summary of Model Accuracy

Model	Accuracy (%)
XGBoost	92.4
LightGBM	89.7
Random Forest	88.5
ARIMA	74.2

VIII. Implication and Potential Applications

The integration of climate and population data significantly improves the accuracy of energy and price forecasts, enabling better anticipation of peak demand and system load. This

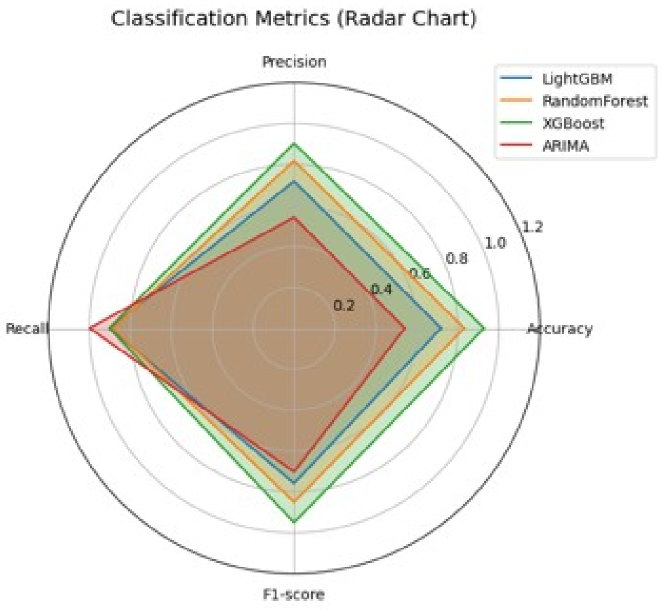

Fig. 16. Radar chart comparing classification metrics (Accuracy, Precision, Recall, F1-score) across forecasting models: XGBoost, LightGBM, Random Forest, and ARIMA

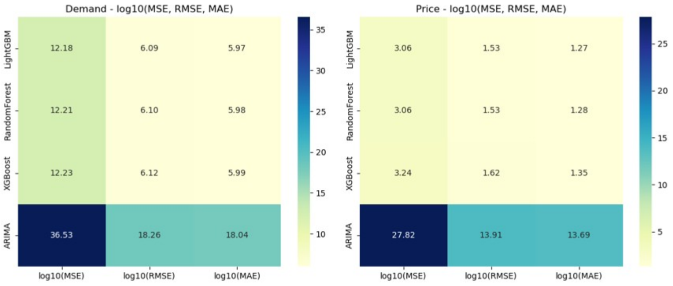

Fig. 17. Heatmaps of log-transformed error metrics (MSE, RMSE, MAE) for demand (left) and price (right) across four models: XGBoost, LightGBM, Random Forest, and ARIMA

enhances operational efficiency through optimized resource scheduling and reduced reliance on costly backup systems. The model also supports sustainable energy planning by informing decisions on grid modernization and renewable integration. Additionally, population-driven forecasting helps align infrastructure development with Ontario's urban growth goals, including constructing 1.5 million new homes by 2030 [6].

IX. Limitation

Despite the strong performance of the forecasting models, several limitations should be acknowledged. Weather data from ten stations may not fully capture microclimate variations, especially in rural or remote areas. Some missing values were filled using forward/backward methods, which could introduce bias during extreme weather events. Additionally, the model is tailored to Ontario-specific data and may require significant adjustments to be applied in other regions including retraining with localized climate, demographic, and energy data.

X. Conclusion

This study developed an accurate forecasting system for Ontario's energy demand by integrating weather and population data. Machine learning models like XGBoost outperformed traditional methods, proving the value of using external features. The system supports better energy planning and lays the groundwork for future real-time and province-wide applications.

XI. Future Work

Future enhancements to this forecasting system include integrating real-time data from smart meters and IoT sensors to improve prediction accuracy. Expanding the model to other Canadian provinces with similar data structures could also broaden its applicability. Additionally, incorporating carbon emission forecasting would allow the system to support environmental impact assessments and contribute to sustainable energy policy planning.

Acknowledgment

We would like to thank Professor Dr. Mahzabeen Emu for her valuable guidance and support throughout this project. We also acknowledge the use of open-source data from Kaggle, Environment and Climate Change Canada, and Statistics Canada, which were essential to our work.

References

[1] Nosovartiom, "Electric Power Consumption Forecasting," Kaggle. [Online]. Available: https://www.kaggle.com/code/nosovartiom/electric-power-consumption-forecasting

[2] M. H. Nawaz, "Energy Consumption Using ARIMA and SRIMAX," Kaggle. [Online]. Available: https://www.kaggle.com/code/muhammadhamzanawaz/energy-consumption-using-arima-and-srimax/input

[3] P. Afroz, "Ontario Energy Demand Forecast with Time Series," Kaggle. [Online]. Available: https://www.kaggle.com/code/pythonafroz/ontario-energy-demand-forecast-with-time-series/input

[4] Gouvernement du Canada. (2025, March 20). Government of Canada. Climate. [Online]. Available: https://climate.weather.gc.ca/historical_data/search_historic_data_stations_e.html?searchType=stnProv&timeframe=1&lstProvince=ON&optLimit=yearRange&StartYear=2003&EndYear=2023&Year=2025&Month=2&Day=17&selRowPerPage=25

[5] Statistics Canada, "Table 17-10-0009-01: Population estimates, quarterly," [Online]. Available: https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1710000901

[6] Government of Ontario, "Energy Powering Ontario's Growth Report," Ontario, Jul. 2023. [Online]. Available: https://www.ontario.ca/files/2023-07/energy-powering-ontarios-growth-report-en-2023-07-07.pdf

[7] T. Hong, P. Pinson, and S. Fan, "Global energy forecasting competition 2012," pp. 357–363, 2014.

[8] G. Zhang, B. E. Patuwo, and M. Y. Hu, "Forecasting with artificial neural networks:: The state of the art," International journal of forecasting, vol. 14, no. 1, pp. 35–62, 1998.

[9] T. Ahmad and H. Chen, "Potential of three variant machine-learning models for forecasting district level medium-term and long-term energy demand in smart grid environment," Energy, vol. 160, pp. 1008–1020, 2018.