A Data Science Approach to Short Term Electricity Forecasting in New South Wales

Michael Kingston (z5372750),
William Stephan (z3404800),
Noel Singh (z5267015).
Reuben Bowell (z5382909)
19th March 2024

Contents

Introduction and Motivation

Electric power is the lifeblood of modern society, powering everything from homes and businesses to critical infrastructure. To ensure the smooth and reliable operation of the power grid, accurately forecasting electricity demand is essential. This is where Short-Term Load Forecasting (STLF) comes in.

STLF is a technique of predicting energy needs over a short horizon, typically ranging for one hour to one day ahead. Precise STLF predictions reduces blackout risks, ensure grid stability and reliability and lowers operating costs for energy suppliers. This project investigates the suitability of forecasting techniques for predicting energy demand over the next seven days, with a specific focus on New South Wales (NSW). The project aims to develop a viable competitor to the current prediction method and hopes to gain valuable insights to the field of STLF.

Primary Motivation: The ever-increasing demand for reliable electricity necessitates continuous improvement in forecasting techniques. Accurate short-term forecasts are crucial for grid stability, as they enable utilities to optimise generation and minimise the risk of blackouts.

Secondary Motivation: This project provides a valuable learning experience. By exploring different forecasting models and analysing real-world data, we aim to gain a deeper understanding of the factors influencing electricity demand and the challenges associated with short-term prediction.

Brief Literature Review

Preliminary research into STLF methods reveal that machine learning techniques are preferred over tradition smoothing and regression methods due to their high accuracy, ability to model complex relationships and capacity for exogenous variables (Ahmad et al., 2014; Metaxiotis et al., 2003; Rodrigues et al., 2023). Techniques of interest from articles that achieved satisfactory accuracy caught in the initial search net are summarised.

For artificial neural networks (ANN) trained for STLF Houimli et al. (2019) identifies 3 important categories of input namely, previous load demand, climate conditions and calendar event variables. Khwaja et al. (2017) provides a convincing case for using ensemble learning methods that outperform single ANN models. Problems typical of ANN such as convergence rates and overfitting were addressed by selective data practices and using momentum during training (Gabaldón et al., 2021; Rodrigues et al., 2023).

Guo et al. (2021) reviews support vector machine, random forest and long short-term memory neural networks. They present a statistical improvement using a fusion of the three models. Mor et al. (2021) present their findings on using a hidden Markov model. They remark that the method is dynamic enough for load forecasting but can suffer when the data

has short term fluctuations, making it potentially better suited for longer term load forecasting. Andriopoulos et al. (2020) trained a convolution neural network on stationarity pre-processed data that outperformed long short-term memory models.

Some common themes for project consideration:

- Impact of residents who store and produce power
- Impact that response programs have on energy consumer demand behaviours (Andriopoulos et al., 2020)
- Improving STLF accuracy with hybrid models and multistep methodologies (some suggested included similar pattern, variable selection, hierarchical forecasting and weather station selection) (Gabaldón et al., 2021)
- Assessment of case-by-case comparison metrics and evaluation methods (Guo et al., 2021; Rodrigues et al., 2023)

This research informs project direction together with insight from AEMO modelling practices. Methods that demonstrate STLF accuracy improvement will be given priority. The project will be able to compare some of the proposed models but won't be able to completely bridge the gap in knowledge of how to effectively compare techniques for separate cases.

Methods, Software and Data Description

Methods

The project method follows the 'cross-industry standard process for data mining' guidelines. Specific choices are informed by current research in literature and industry practices. Project flow is visualised in Appendix B.

Data collection and pre-processing

The National Electricity Market (NEM) and Bureau of Meteorology (BOM) websites are the primary data sources of the project (Bureau of Meteorology, 2023; *Nemweb Market Data*, 2024). They offer clean, complex data that is readily accessible. Along with the Git repository data, the project has access to important predictors for the STLF task. They are summaries in the table below. Data is then pre-processed via cleansing, time series feature engineering and standardizing ready for exploratory analysis in Python.

Data	Source	Format
Total Electricity Demand (NSW)	NEM	DATETIME (5min intervals), TOTAL DEMAND (MW), REGIONID
Total Electricity Forecast (NSW)	NEM	DATETIME (30min intervals), FORECASTDEMAND (MW), REGIONID, PREDISTPATCHSEQNO, PERIODID, LASTCHANGE
Tempreture (NSW)	BOM	DATETIME, TEMPRETURE (air temperature °C), LOCATION (region (i.e Bankstown)

Data	Source	Format
Forecast Demand (NSW)	NEM	REGIONID, INTERVAL_DATETIME, LOAD_DATE, OPERATIONAL_DEMAND_POE10, OPERATIONAL_DEMAND_POE50, OPERATIONAL_DEMAND_POE90, LASTCHANGED
Rooftop_PV	NEM	ROOFTOP_PV_ACTUAL, INTERVAL_DATETIME, REGION_ID, POWER, QI, TYPE, LASTCHANGED
Electricity Demand	NEM	Measured by metering supply to the network, not consumption. 'Operational' = total electricity used by consumers and businesses.
Heating Degree Days (HDD)	NEM	Degrees below a critical temperature indicating deviation from normal weather.
Cooling Degree Days (CDD)	NEM	Degrees above a critical temperature indicating deviation from normal weather.
Dummy for shock effect	NEM	Dummy variable, represent changes in economic activity due to external shocks affecting electricity consumption.

Table 1: Data sources & formats

Exploratory Data analysis

Exploratory techniques for time series data will be used for initial data understanding and pattern identification. Python will be used to perform:

- Data visualizations such as heat maps, variable plots and time series plots
- Decomposition to check for seasonality, trends, and anomalies in the data.
- Correlation analysis
- Outlier detection
- Descriptive statistics

Feature engineering

For STLF we are advised to derive additional time features from the data. These features are based on 24 hour, weekly and yearly timeframes that capture clear 'seasonal' impacts that time has on electrical load. Indicator features based on calendar events will also be generated.

Data splitting & CV strategy

For model evaluation, data will be divided into training, validation and testing subsets. Due to the temporal nature of the data, specific time series cross validation strategies will be used. The python library scikit-learn has appropriate methods for this namely `TimeSeriesKFold( )`.

Selecting Models

Priority models used in the project will include neural networks ensemble, a hybrid model and a regression model (which is used by AEMO for both residential and business load forecasting). In line with our project goal and due to the team's strength in modelling and programming, other valid techniques will be considered for comparison in this project.

- Smoothing techniques, ranging from exponential smoothing, Holt-Winters etc.

- Autoregressive models (ARIMA and SARIMA)
- Tree based models such as random forest and gradient boosted decision trees
- Deep learning approaches are promising, but time constraints may limit our ability to implement models like LSTMs and Transformers.

Model evaluation

Evaluation of the performance for each trained model will be conducting using the test dataset. A comparison of predicted values against actuals will be determine the accuracy of the specific model.

Tracking key metrics to be used in comparing models includes: MSE (mean square error), RMSE (Root Mean Square Error), MAPE (mean absolute percentage error) will help to optimise model performance.

Visualisation of predicted values against actuals will also help identify areas for improvement.

Software

The forthcoming research project will utilize the following data toolkit. Please note that this is not an exhaustive list of applications and may change through the course of the project.

- **Python:** Python to be used by the team for data cleansing, analysis and modelling due to the significant amount of experience all team members have with Python.
- **Pandas, NumPy, scikit-learn, matplotlib, seaborn, tqdm, StatsModels, XGBoost, LGBM, Prophet, Pytorch**: Part of the primary libraries and packages to be used for data manipulation, data modelling and visualization.
- **Poetry**: Python Poetry simplifies dependency management and packaging by using a single `pyproject.toml` file for configuration, offering deterministic builds and resolving dependency conflicts more efficiently than pip or conda.
- **Jupyter Notebook**: Project execution and report generation and allows the team to lay out the steps taken through the course of analysis.
- **GitHub**: Serving as the project repository, VCS (version control system), data storage (archived / zipped data), backlog management, Kanban board for task tracking and Code Review.

- **Teams**: The official communication channel to be used for the project including team meetings, staff catch-ups, sharing documents not committed to GitHub and collaboration.
- **Microsoft Office Suite**: Specifically, Excel, Word and PowerPoint, for supporting data analysis activities and capturing additional project documentation and final presentation.

Project Plan

Team Structure

Group O comprises of four members with diverse backgrounds, skill sets and levels of experience. This diversity enables each member to assume multiple roles and make unique contributions to the project's ultimate outcome.

There are several roles with a research project team which need to be filled to produce a successful project, these include: (Indeed Editorial Team, 2022)

Project Role	Abb.	Responsibilities
ML engineer	PM	Responsible for developing models. Responsible for the day-to-day management of the project specifically in planning, risks and mitigation, maintaining quality standards, running the project on time and budget and handling change requests. (What does a project manager do?, 2024)
Data Engineer	DE	Deal with structured and unstructured data. Provide data in a usable format to data scientists to run queries on. Need to be good with understanding ETL tools and API's. (Lutkevich, 2021)
Data Scientist	DS	Coordinates data analysis and interprets results from data. Proficient in statistical analysis, machine learning and data visualisation.
Business Analyst	BA	Ensuring that data analysis efforts are closely aligned with the strategic objectives of the organization, and the delivery of tangible business value.

Table 2: Data research project roles and responsibilities

The table below captures the candidate and project role based on the mind-map (Appendix A).

Candidate	Project Role	Rationale
Michael Kingston	ML Engineer	Experienced with Git and GitHub, some experience with ML projects.
William Stephan	Data Scientist	Experienced programmer with a background in mathematics and statistics. Experienced with data visualization, regression and machine learning packages.
Reuben Bowell	Data Scientist	Mathematics background and working as a quantitative analyst

Candidate	Project Role	Rationale
Noel Singh	Business Analyst	Experience as a business analysis working on Actuarial data platforms and other complex projects.

Table 3: Group 'O' roles

Key Activities

The schedule of activities is captured in the table below and have been identified based on analysis of the team's skills, interests, and experience.

#	Task	Team Role	Rationale
1	**Business Problem Understanding**		
1.1	Research electricity market	All	Build domain knowledge for team
1.2	Initial review of project statement	All	Initial project setup and problem statement definition.
2	**Data Collection**		
2.1	Understand data requirements	All	Define data gaps and formulate data ingestion plan
2.2	Collate data	DE	Needs to understand methods of accessing data for ingestion pipeline.
3	**Data Preparation**		
3.1	Ingestion jobs	DE / DS	Develop ingestion jobs into platform.
3.2	Data cleansing, scrubbing	DE / DS	Tools and techniques for data cleansing and scrubbing employed.
3.3	Outlier remediation	DE / DS / BA	
3.4	Standardize data format	DE / DS	Conformed data required for modelling.
4	**Modelling**		
4.1	Explanatory Analysis	All	Different perspectives brought by experienced team members.
4.2	Data modelling	PM/DE / DS /BA	ML models and tools employed based on experience and project parameters.
5	**Report**		
5.1	Final Version	All	Input from all users to produce report.
6	**Video Presentation**		
6.1	Final Version	All	Input from all users for final presentation.

Table 4: Key Activities & Roles

Gantt Chart

The project plan is represented in the below Gantt chart:

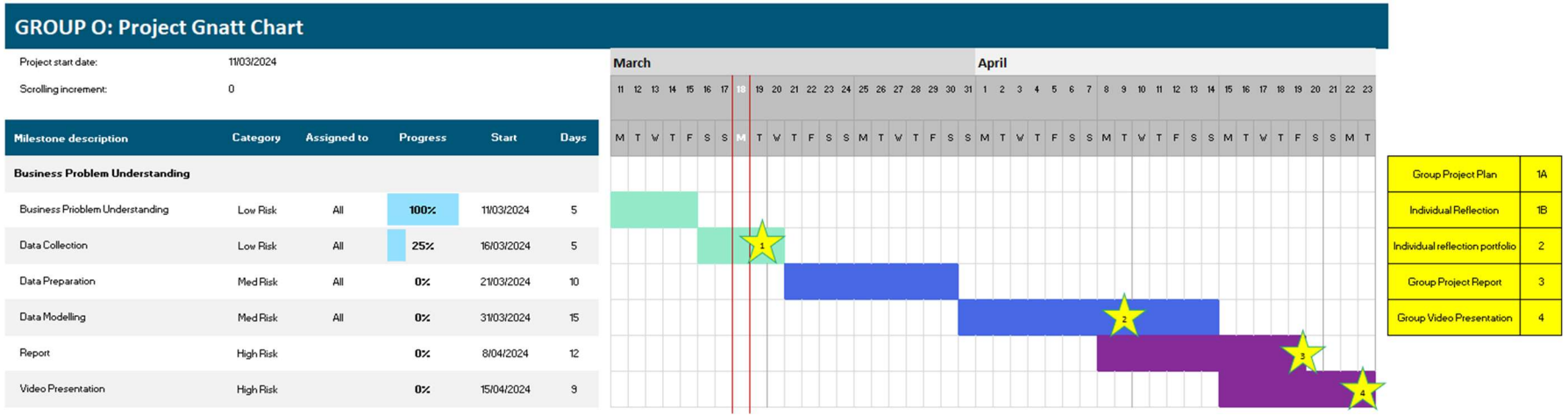

Figure 1: Project Gantt Chart

Bibliography

Andriopoulos, N., Magklaras, A., Birbas, A., Papalexopoulos, A., Valouxis, C., Daskalaki, S., Birbas, M., Housos, E., & Papaioannou, G. P. (2020). Short Term Electric Load Forecasting Based on Data Transformation and Statistical Machine Learning. *Applied Sciences*, *11*(1), 158. https://doi.org/10.3390/app11010158

Bureau of Meteorology. (2023). *Climate Data Online*. Bom.gov.au; Bureau of Meteorology. http://www.bom.gov.au/climate/data/

Data Manager. (2022, 12). Retrieved 03 2024, from APSC: https://www.techtarget.com/searchdatamanagement/definition/data-engineer

Department of Electrical Engineering, Iowa State University. (2005). Modeling electricity markets with hidden Markov model. _Archives of Computational Methods in Engineering,_ Available online 15 December 2005.

GabaldónA., María Carmen Ruiz-Abellón, & Alfredo, L. (2021). *Short-term load forecasting 2019*. Mdpi.

Guo, W., Che, L., Shahidehpour, M., & Wan, X. (2021). Machine-Learning based methods in short-term load forecasting. *The Electricity Journal*, *34*(1), 106884. https://doi.org/10.1016/j.tej.2020.106884

Houimli, R., Zmami, M., & Ben-Salha, O. (2019). Short-term electric load forecasting in Tunisia using artificial neural networks. *Energy Systems*. https://doi.org/10.1007/s12667-019-00324-4
Indeed Editorial Team. (2022, 10). *Responsibilities of Research Teams (With Key Roles)*. Retrieved 03 2024, from Indeed: https://www.indeed.com/career-advice/career-development/responsibilities-of-research-teams

Khwaja, A. S., Zhang, X., Anpalagan, A., & Venkatesh, B. (2017). Boosted neural networks for improved short-term electric load forecasting. *Electric Power Systems Research*, *143*, 431–437. https://doi.org/10.1016/j.epsr.2016.10.067

Lewis, C. (n.d.). Chapter 6 Project Roles and Responsibilities | Data Management in Large-Scale Education Research. In *datamgmtinedresearch.com*. Retrieved March 19, 2024, from https://datamgmtinedresearch.com/project-roles-and-responsibilities

Lutkevich, B. (2021, 03). *Definition Data Engineer*. Retrieved 03 2024, from TechTarget Data Management: https://www.techtarget.com/searchdatamanagement/definition/data-engineer

Malki, H. A., Karayiannis, N. B., & Balasubramanian, M. (2004). Short-term electric power load forecasting using feedforward neural networks. *Expert Systems*, *21*(3), 157–167. https://doi.org/10.1111/j.1468-0394.2004.00272.x

Metaxiotis, K., Kagiannas, A., Askounis, D., & Psarras, J. (2003). Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher. *Energy Conversion and Management*, *44*(9), 1525–1534. https://doi.org/10.1016/s0196-8904(02)00148-6

Mor, et al. (2020). A Systematic Review of Hidden Markov Models and Their Applications. Archives of Computational Methods in Engineering, 28(2).

Nasr, G. E., Badr, E. A., & Younes, M. R. (2001). Neural networks in forecasting electrical energy consumption: univariate and multivariate approaches. *International Journal of Energy Research*, *26*(1), 67–78. https://doi.org/10.1002/er.766

Nemweb market data. (2024). Aemo.com.au. Retrieved March 19, 2024, from https://www.aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem/market-data-nemweb#mms-data-model

Research Team Structure. (2021, November 9). Elsevier Author Services - Articles. https://scientific-publishing.webshop.elsevier.com/research-process/research-team-structure/

Rodrigues, F., Cardeira, C., Calado, J. M. F., & Melicio, R. (2023). Short-Term Load Forecasting of Electricity Demand for the Residential Sector Based on Modelling Techniques: A Systematic Review. *Energies*, *16*(10), 4098. https://doi.org/10.3390/en16104098

Veeramsetty, V., Chandra, D. R., & Salkuti, S. R. (2020). Short-term electric power load forecasting using factor analysis and long short-term memory for smart cities. *International Journal of Circuit Theory and Applications*, *49*(6), 1678–1703. https://doi.org/10.1002/cta.2928

What does a project manager do? (2024). Retrieved 03 2024, from Association of Project Management: https://www.apm.org.uk/jobs-and-careers/career-path/what-does-a-project-manager-do/

Appendix

Appendix A: Individual Skills Mind-map

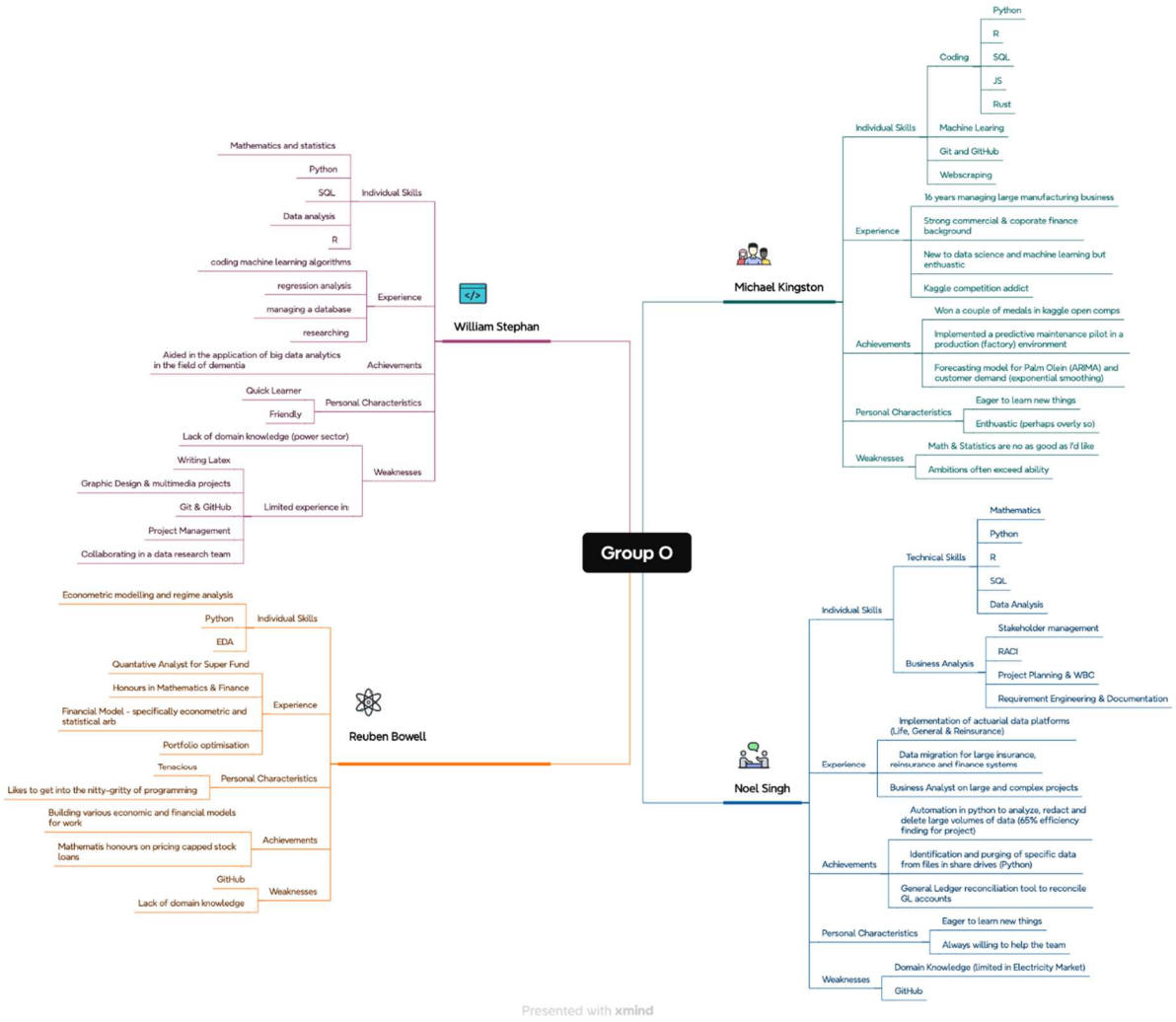

Figure 2: Skills Mind-map

Appendix B: Proposed Methodology Chart

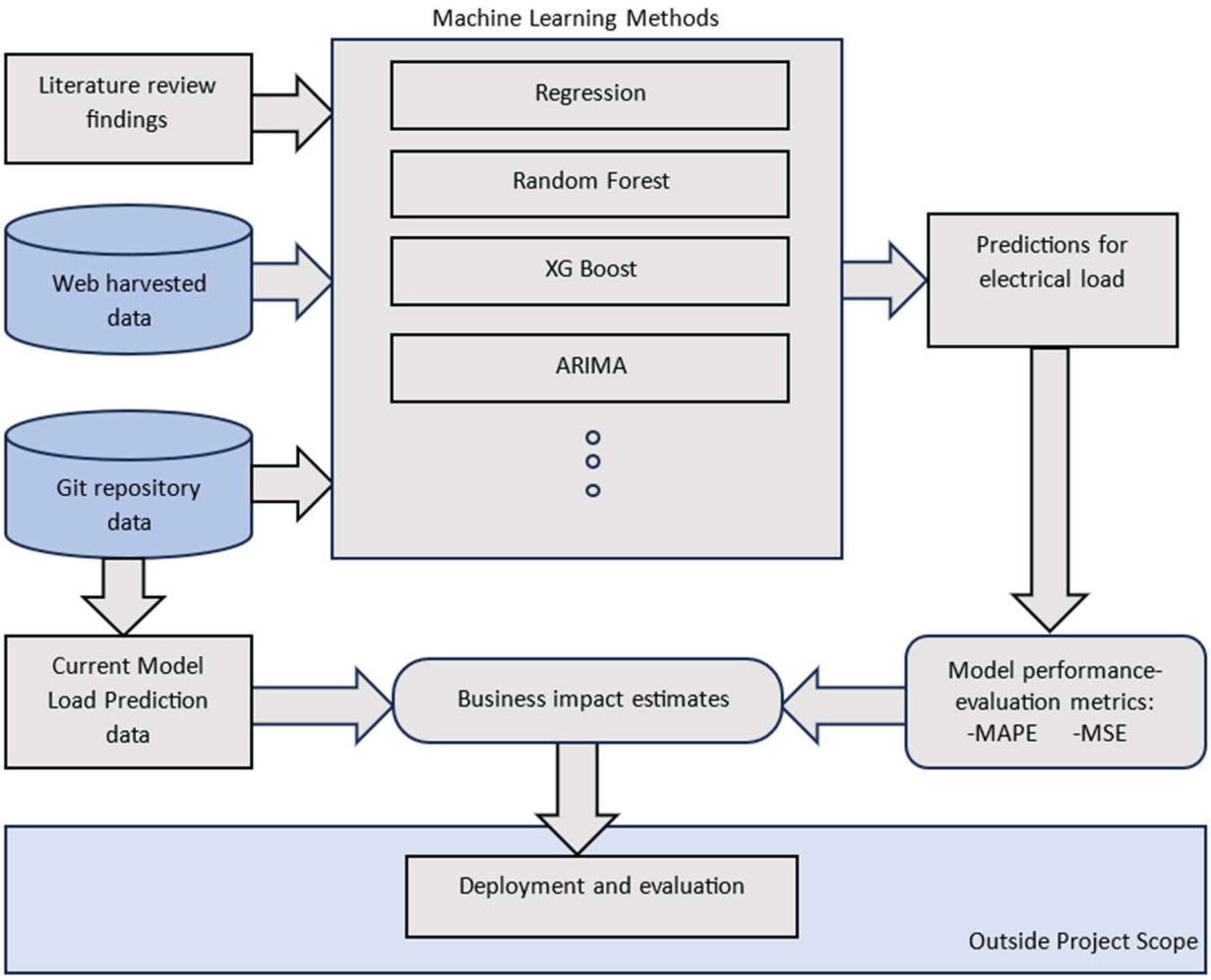

Figure 3: Methodology Process Flow