Final Year Project

Optimizing Short-Term Energy Demand Forecasts Through Machine Learning: A Case Study at UCD's School of Computer Science

Zairui Zhang

Student ID: 19209905

A thesis submitted in part fulfilment of the degree of

BSc. (Hons.) in Computer Science

Supervisor: Prof. Eleni Mangina

UCD School of Computer Science

University College Dublin

April 26, 2024

Table of Contents

Abstract

Short-term energy demand forecasting continues to be at the frontier of grid stability enhancement, efficient resource allocation, and effective participation in the energy markets. It works towards making sure that there is a balance in supply versus demand. These, in turn, help not to face any scope of blackouts and, further, tune their strategies toward bettering energy generation, distribution, and trading. They become important also in emergency times, namely when the most relevant weather events or other disturbances take place. Load forecasting represents an essential part of the power system operational framework. It impacts, with great weight, on the dispatch of the system, and in a derived way, on diverse operational aspects.

The approach of this project has shown from traditional electricity load simulations to domain-preferring, machine learning-based black-box techniques requiring less domain-specific knowledge and enabling faster development cycles. We make use of actual consumption data arising from the School of Computer Science for the year 2023. The only findings we use and test for the application and assessment of the proficiency of five different machine learning models lie in a short-term predictive scenario. A sliding window is used with the active feature selection paradigm to ensure that our machine-learning models always reflect the respective underlying data trends.

Regarding prediction accuracy, in relation to the performance in these models, we used two standard measures: mean bias error (MBE) and root mean square error (RMSE). The result of this research is a feasible strategy that offers academic institutions machine learning in their short-term forecasting of energy demand. This strategy will ensure that such institutions have a new way in which to revolutionize how they manage the energy within their campus.

Chapter 1: Project Specification

This philosophy underpins the energy unit at the University College Dublin (UCD) that the most sustainable energy is the one not used. Therefore, the Management Energy System (EnMS) dominantly focuses on efficiency improvement and waste reduction[1]. This report presents a case study focusing on short-term prediction of electricity consumption alongside machine learning models.

Moreover, the report integrates cutting-edge predictive analysis using data extracted from the Cylon platform[2]. Leveraging Cylon's data extraction capabilities, pertinent energy-related datasets are gathered, encompassing consumption, generation, weather, and pertinent parameters essential for predictive modeling and machine learning. This amalgamation of data-driven insights and modeling techniques amplifies the project's analytical prowess.

Utilizing the Scikit-learn[3] and the Statsmodels library in Python, the report predicted electricity consumption of the Computer Science building on the Belfield UCD campus, allowing diverse scenario exploration and system optimization. Additionally, aims are such as minimizing energy costs, reducing energy-related carbon emissions, and assessing the cost-effectiveness of emission reductions, thereby shaping the UCD campus's sustainable energy future.

1.1 Core Objectives

1. Considering the dynamic nature of consumption patterns within university campus buildings.
2. To integrate a dynamic feature selection strategy that helps dimensionality reduction. Then use a sliding window approach, ensuring the adaptability and timeliness of the models in response to changing data patterns.
3. To evaluate the efficacy of various machine learning models in different predictive scenarios (day-ahead, hour-ahead, and step-ahead) using key performance metrics like MBE and RMSE.
4. To explore the potential of extending the methodologies and insights gained from this study to broader applications in energy demand forecasting across different sectors.
5. To use the insights from model performance to guide energy management strategies that can optimize consumption, reduce costs, and contribute to UCD's sustainability goals.

Chapter 2: **Introduction**

Against the challenging backdrop of ambitious climate targets in Ireland, aiming to reduce emissions by 51% from all sectors by the year 2030, it further re-emphasizes that our UCD school is committed to sustainability through outstanding research and cutting-edge initiatives in the built environment and retrofitting practices. Because of the national mandate for retrofitting 500,000 homes by 2030, the overriding importance of getting the retrofits right and done right in the first place cannot be overstated. However passionate about retrofitting as a way of energy efficiency and reducing carbon footprints, evidence-based research on the efficacy of such retrofits leaves a critical gap. To bridge this gap, the BIACE Lab at UCD dedicates its activities to inquiring into the real performance of the world's leading energy-saving measures, like super-insulated fabrics or even heat pumps, otherwise found to underperform due to a series of problems from installation to the very misunderstanding of their operation dynamics. Recent studies have highlighted these inefficiencies, emphasizing the need for accurate post-occupancy evaluations. UCD is leading research on the optimization of retrofit strategies with ongoing SEAI-funded projects, among which is the MacAirH project that includes but is not limited to, monitoring and improvement of performance for air-to-water heat pumps. This commitment to sustainable development and careful research was one of the cornerstones in making our report, reflecting the need and immediacy of data-driven, informed approaches to climate actions of the built environment [4].

However, their effort toward the carbon-reduction target was proceeding with difficulty within the broader environmental landscape at the UCD Computer Science Building. Some statistics, as represented in (Figure2.1), do represent the trend that carbon emissions have been decreasing from 2014 to 2023; and indeed, after the results of COP26, we may have some room for optimism. Still, no single drop could reach—let alone over-achieve—our very ambitious goals, especially in the light of the huge, pandemic-induced consequences on Europe's energy balance. From a peak consumption in 2014 of 281,616 kWh, related to emissions of 162,233 kg CO2e, the trend goes downward, with a notable year of 2020, where consumption and emission are respectively remarkable, down to 182,307 kWh and 105,009 kg CO2e. This drop is corresponding to the global drop due to activity worldwide, because of COVID-19. At first sight, this outcome may look very promising and positive. However, in the further years after consumption, one can observe some kind of rebound effect, as both consumption and emissions tend to level.

The data portrays a temporal shift not only in energy use but also its environmental toll. They point out the cost implications, that the highest expenditure on energy was in 2014, standing at €52,268.02, and the lowest in 2020, standing at €33,836.13. Notwithstanding the steady decline in emissions year after year, the impact of the very recent events shows a big question mark on the way to the net-zero objective. This highlights that strict scrutiny of energy strategies and sustainability initiatives are now inevitable in order to respond anew to the resurging challenges post-pandemic and the way forward towards adjustment in a changed energy landscape in Europe.

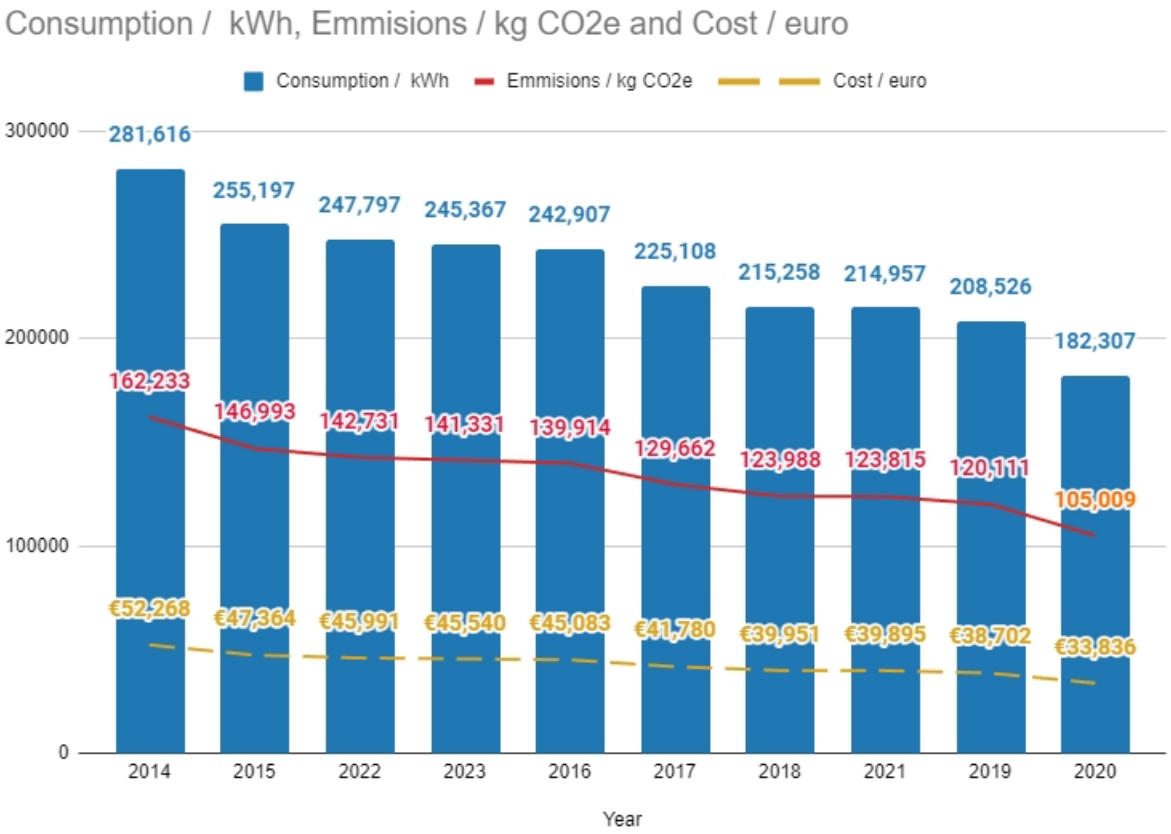

Figure 2.1: UCD Computer Science Building

Faced with spiraling worldwide energy needs that will be experienced in the post-pandemic period and more pressure due to the Russian invasion of Ukraine, which has critically affected the European energy industry and resulted in an explosion of oil and natural gas prices, the Irish government has set up the Energy Security Emergency Group (ESEG) [5]. The ESEG is a division of the Department of the Environment, Climate, and Communications. Its main aim is to ensure Ireland with secure energy, meaning that there would be continuous supply and availability of energy at prices afforded by the economy. This means that representatives in the realm will be expected to coordinate a comprehensive national strategy according to the outlined national energy security framework. This framework focuses on assistance in relation to energy poverty and supply security for the most exposed consumers and businesses, helping to reduce dependence on imported fossil fuels. The government has also rolled out the "Reduce Your Use" campaign, which will help promote energy efficiency, at the same time supporting household and business concerns with the upward-spiraling costs of energy and calling for responsible usage of energy in these challenging times for the globe[6].

It is generally believed that universities are the bastion of social change, heralding inclusion, truthfulness, sustainability, and bringing ideas and critical dialogue to many free-flowing rivers. They keep the burden of responsibility in sustainability cases and, if not an example for the bigger community, then at least one through vigilant monitoring of the environmental footprint[7]. Therefore, it behooves all students to be inculcated with the principles of sustainable development and be motivated to adopt and propound behaviors that are friendly to the environment. Thus, sustainability-focused education institutions have a twin responsibility of inculcating positive practices in their student body and reducing campus activities that are responsible for causing ecological damage. This should be clearly evident at the very crux of the curriculum and operational strategies and policies underpinning the university's role as an educator and practitioner of sustainable development[8].

This project tries to apply advanced techniques in machine learning to forecast future loads of electrical power with the required level of accuracy, while at the same time handling the inherent uncertainties. This comprises the Cylon platform for our primary data sources, which gives specialized datasets on electricity usage in the UCD Computer Science Building, paralleled by concurrent weather data. The present study focuses on a holistic time series analysis, with various time-dependent features engineered for the light to come out on the underlying consumption patterns, among them being operational hours and seasonal variances observed during winter and summer months.

In this work, the sliding window technique is applied with the objective of being able to forecast from very short-term forecasting (15 minutes ahead), up to longer ranges (one week ahead). This will be key in capturing the dynamic temporal relationship in data in accordance with the insights realized at the level of autocorrelation functions and partial autocorrelation plots. Such an augmented and solid analytical framework, if offered above, is a contribution to be made toward improving our understanding of electricity demand dynamics and, in this way, providing valuable inputs for efficient energy management and practice within academic institutions.

Chapter 3: Related Work and Ideas

This chapter will focus on the related work. The first section will discuss the scoping review process used to identify relevant related studies. The subsequent sections will include the literature review of the primarily related studies.

3.1 Scoping Review Process

A systematic literature review (SLR) was undertaken to identify pertinent literature for the scoping review. This rigorous and transparent search, as outlined by Snyder (2019)[9], spanned across multiple databases and was designed for reproducibility. SLR is a methodical approach used to identify and assess relevant research that will be conducive to the collection and analysis of relevant data. The systematic review tries to source all empirical evidence corresponding to pre-defined inclusion criteria in relation to a specific research question or hypothesis. The review process, being systematic and explicit in the methods used, ensures there is a reduction in the level of bias and therefore derives reliable findings for one to be able to conclude and make an informed decision. These are the main activities in SLR: planning the review, searching, literature analysis, and writing up the results report[10].

3.1.1 Search String Development:

A meticulously designed search string was employed to effectively retrieve relevant literature aligned with the research question and objectives. The search string utilized a combination of key terms, including but not limited to "Short-term prediction", "energy prediction", "Sustainable", "Green", "Energy Efficiency", "Campus," "sliding window" and "time series analysis" This strategy aimed to cast a wide net across databases and sources, ensuring the inclusion of diverse perspectives on the intersection of sustainability, energy efficiency, campus, and energy control system.

3.1.2 Platform Utilization

Subsequently, the search string was executed on reputable academic databases, namely Scopus [11] and Science Direct[12]. These platforms are recognized for their extensive coverage of academic literature, ensuring a broad scope in capturing pertinent papers related to the chosen topic.

3.1.3 Paper Management with Zotero[13]

Zotero reference management tool was used in sorting, managing, and organizing loads of papers that were exported from Scopus and Science Direct. This made it easy to collect, organize, cite, and share the research material, thus controlling the accumulated literature effectively and in an organized manner.

3.1.4 Results Consolidation

The next task was merging the results obtained from the Scopus and Science Direct searches within the Zotero platform. This process refers to the merging and synchronization of the two databases to have one database of papers all in a bid to get organized, systematic, and coherent analysis in preparation for the later stages of the literature review.

3.1.5 ASReview [14] for Efficient Filtering

Active Learning tool ASReview was used to assist in the screening and make use of pre-defined inclusion and exclusion criteria for filtering out papers. ASReview was created for systematic reviews and suggested an active learning approach in which the program could propose whether to include or exclude papers based on reviewers' judgment and screening. As the reviewer moved on to make the decision, the tool adapted and continued to refine its recommendations for good screening relevance and efficiency in the process.

3.1.6 Screening Process Execution

This step was to perform the screening using ASReview, for both inclusions and exclusions. As the review progressed, the tool itself improved its suggestions by learning from the decisions of the reviewer in real-time. This was through an iterative process within ASReview that ensured a productive and intensive screening of the papers was done in line with the criteria.

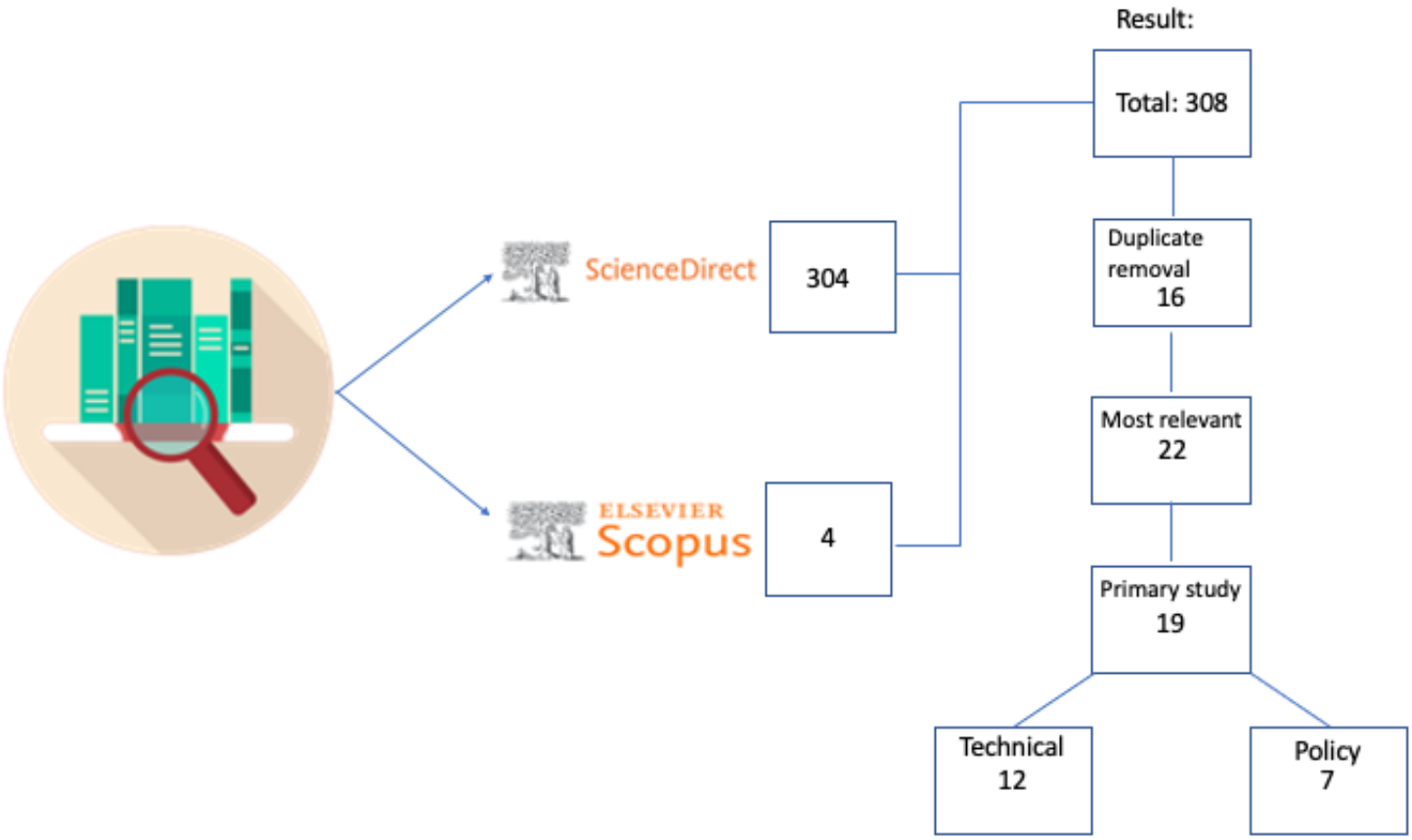

Figure 3.1: SRP

3.1.7 Discussion and Conclusion

Following the completion of the screening process, 19 papers were selected for primary studies which include 13 Technical papers and 6 Policy papers. The next step involves the extraction of pertinent data from the selected papers. This phase is crucial for acquiring detailed insights into each study. Information to be extracted encompasses various aspects, including study purpose, methodology, Dataset source, software mentioned and any other relevant details that contribute to a comprehensive understanding of the research.

Using these characteristics summarized by primary studies to summarize the literature review.

3.2 Methodology for short-term prediction

The technical methodologies employed across various studies demonstrate a diverse spectrum of approaches aimed at quantifying, modeling, and assessing energy consumption, and potential technological interventions.

In a more enhanced assessment of the energy flexibility potential in buildings, one may participate in the framework's methodology with the ensemble of machine learning frameworks, sliding window methods, and flexibility indicators. This new approach allows not only the quantification but also the characterization of flexibility in residential buildings at the level of the individual, considering a variety of energy systems and prediction horizons. Through ensemble learning techniques, data-driven algorithms are optimized, making it possible to train for the generation of optimal forecasts that, in practice, will mean a more precise assessment of the potential in energy flexibility. The paper has analyzed the existing occupant response models and highlighted their weak points through a critical analysis. In this sense, the approach selects features referring to occupancy profiles and prediction horizons, with the shorter prediction horizons which will be an essential element in carrying out accurate assessments of the potential impacts of demand response actions on occupants' comfort. This also takes into consideration the weather conditions and occupancy behaviors in the predictability of the flexibility potential in buildings; hence, it stresses the importance of accurate forecasting models in energy systems[15].

An approach to optimizing the short-term energy demand forecasts with regard to the long-term behavior in building performance is supposed to be built further. There is the aspect of detecting and quantifying the deviations between the modeled and observed data sets over different time frames. The submetering of the building allowed for the monitoring of the building's energy consumption against the predicted energy usage at periods of daily, weekly, monthly, and quarterly use of a deviation percentage formula. Further, it defines the limits in time and magnitude of deviation, for example, in the quarter, month, day, and week, whereby more dynamic and adaptive thresholds can be defined to set the limits that are related to different patterns of behavior of deviation. The scaling effects will be considered in the energy data analysis, whereby it would have the way in which the maximum usage peaks is to give maximum weight for it to represent the given quantity of energy performance. The model should also be recalibrated within an accurate energy forecast where discrepancies between the model and actual building performance are probably arising from different timescales. This requires the validation of the input data used in the building energy model and periodic recalibration to give accurate predictions and represent operating building characteristics at any time. This would support software tools like OpenStudio[16] and EnergyPlus[17], therefore, in sensitivity analysis, uncertainty quantification, and decision support for further optimization of short-term energy demand forecasts through machine learning. They offer an environment for rapid building energy models and permit fine-grained modeling with no requirement to create complicated, detailed models that are more and more inapplicable for

operational insight[18].

The Predictive system is a reliable and very accurate forecasting model for energy consumption patterns in residential buildings, based on the metaheuristic optimization of the sliding window and combining the MetaFA-LSSVR model with the SARIMA model. In this way, the hybrid accurately captures the compliances of the linearity and nonlinearity components in energy consumption data, hence enhancing the capability for prediction. In line with that, the system uses real-time data emanating from a smart grid metering infrastructure to provide useful information for the building owners. It is in line with the monitoring and prediction of energy consumption to take energy-saving action proactively from the prediction result. The system also metaheuristic optimizes to machine learning, making the system an exceedingly powerful tool for real-time energy management. The fact that the system also presents viable opportunities in its scalability to solve big-data problems related to smart grid energy consumption points to future research and development[19].

The application of the sliding window technique is crucial for forecasting the wind velocity time series. This is due to the uncertain impact caused by the discontinuity in the wind energy injected into the power grid. Accurate forecasting is needed on the part of the energy manager and electricity trader. It is proposed to use multi-layer perceptron (MLP) and radial basis function networks of artificial neural networks in the application for energy demand forecasting, and the network structure 4-7-13-1 architecture of the MLP proved best among others in predicting the wind velocity time series in Tehran, Iran. The most typical statistical indices used in the performance evaluation of forecasted data include RMSE, MSE, and R2, which bring forth insights in regard to the efficiency and accuracy of the network. The best MLP network structure of 4-7-13-1 gave acceptable values of RMSE, MSE, and R2, which made it clear that such a model was very effective in predicting wind velocity data using the sliding window technique[20].

A machine learning Sliding Window Regression (SWR) approach methodology for short-term energy demand forecasting optimization. SWR is a novel adaptive prediction algorithm in which it is data-driven using a sliding window of data for model training while new data is incorporated so that error propagation is avoided and model performance is maintained, even under dynamic load data conditions. The SWR methodology trains the model with historical data using the sliding window method, making a periodic update to the model so that it is able to trace errors and remain accurate. The ideal window size of the training data is determined automatically to make it into the model for better performance. The models can be evaluated through several metrics, such as mean absolute error (MAE), root mean square error (RMSE), mean absolute percentage error (MAPE), root mean square percentage error (RMSEP), and almost correct prediction error rate (ACPER). The MAPE is the relative accuracy metric in this paper that deals with the regression problem of relative accuracy between the values predicted and the actual load values within the training window. The SWR forecasting methodology used in the short term for energy demand foretells that, with such measures, researchers and practitioners would be in a position to improve the accuracy and reliability of predictions—particularly in conditions that are dynamic about load data[21].

3.3 Summary

The paper review mentioned the application of sliding window techniques, machine learning, and time series analysis as an approach to improve short-term energy demand forecasting. This approach lays special emphasis on the sliding window, viewed as a power in managing and analyzing sequentially updating data in the training set by disposal of the oldest and taking in the new data, hence ensuring well-calibrated models in even dynamic load conditions.

In the context of machine learning, these could be a type of energy consumption and wind velocity prediction. Such models perfectly fit linear as well as non-linear forms in energy data, which further enhances the predictive power of these models. The authors note that these systems are improved with ensemble learning techniques that generate the most accurate and reliable predictions.

Within such a context, the analysis of time series data is inevitable, with the use of machine learning models in handling and making predictions for energy consumption patterns from historical data. This is a robust integration underlining the importance of highly accurate, real-time forecasting in the management of energy systems, focusing on the leverage of continuous data streams from smart grid infrastructures.

These strategies work together in giving a stronger forecasting model for the prediction of energy requirements and achieving better energy-managed building adjustment to the variations in energy consumption patterns.

Chapter 4: Methodological Implementation

This study was conducted at the University College Dublin (UCD) campus, focusing on Short-term energy demand predictions. The methodology encompassed a multi-stage approach to comprehensively analyze, and optimize the campus's energy systems. The first chapter will introduce the implementation plan. Then the second section demonstrates applications of the Cylon platform. The strategy for implementing the code is described in the third part. The entire codebase and dataset are available on GitLab[22], an online platform leveraging Git for version control and collaboration.

4.0.1 Implementation Plan

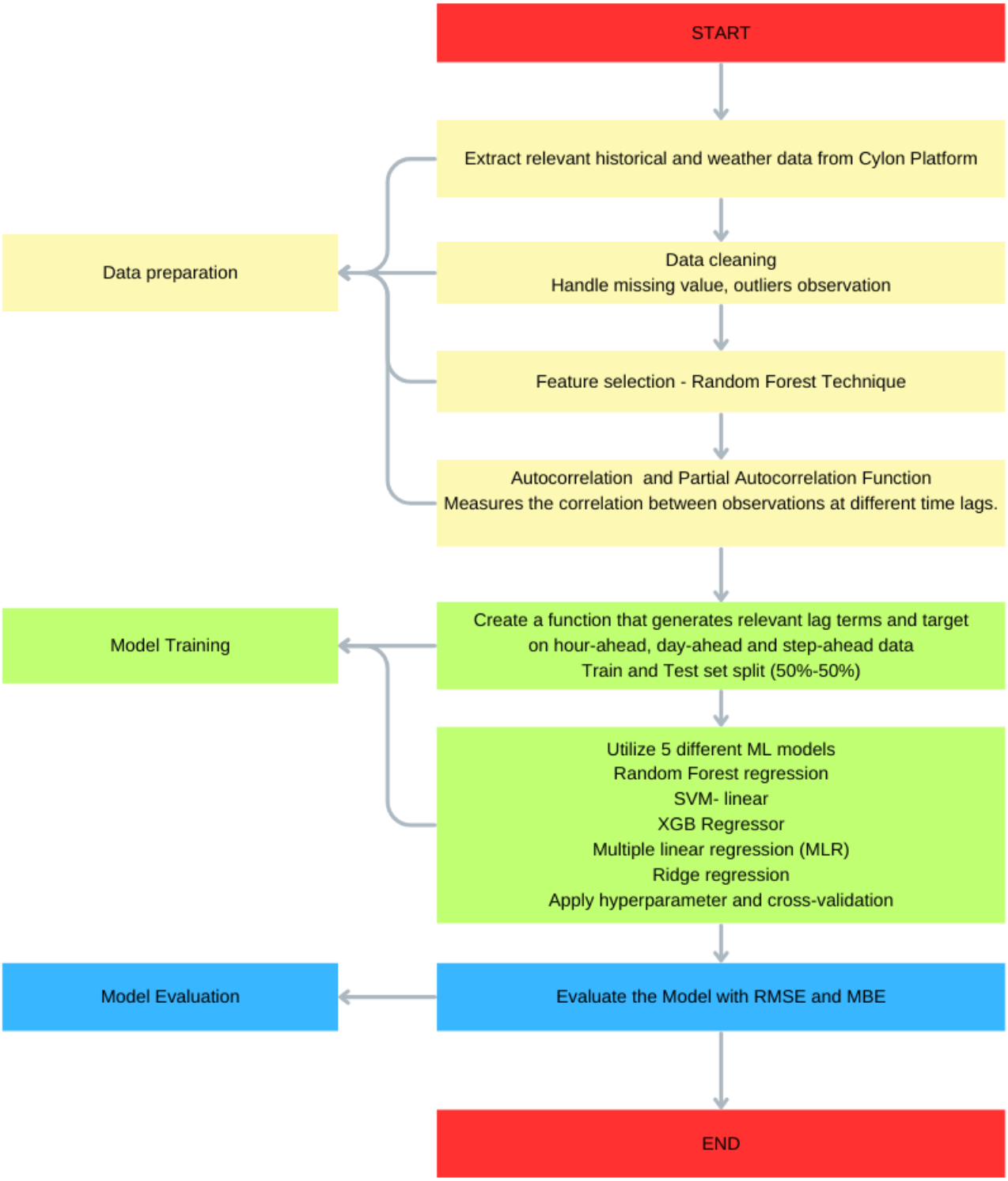

Figure 4.1: Implementation Flow chart

4.1 Data Preparation

The data preparation stage is vital to the modeling process, involving a series of methodical steps. We begin by extracting key historical and weather data from the Cylon Platform, ensuring a solid foundation for analysis. The next step is data cleaning, where we meticulously handle missing values and outliers to maintain data quality. Feature selection is then conducted using the Random Forest technique to identify influential predictors. Additionally, we utilize Autocorrelation and Partial Autocorrelation Functions to measure the correlation of data points across various time lags, offering insights into temporal dependencies. Each of these steps is vital for building a robust predictive model and will be discussed in greater detail in the following subsections.

4.1.1 Data resource

The project's energy analysis dataset is based on the ABB Cylon® solution integrated into the unit within the Unitron Building Management System (BMS). ABB Cylon® System represents state-of-the-art centralized control technology tailored to assist in the smooth running and efficient functions of a building for facility service. It demonstrates that this system can handle large complex environments; in fact, UCD's campus across approximately 300,000 m² of treated floor area.

The interface of ABB[23] with Unitron BMS provides a strong framework for energy management throughout the campus. This synergistic blend offers not only increased operational productivity but also indeed the most essential element for environmental sustainability. These state-of-the-art systems have allowed UCD to monitor, control, and analyze in Fig4.2 with great detail the consumption patterns of energy in the buildings they have ownership of.

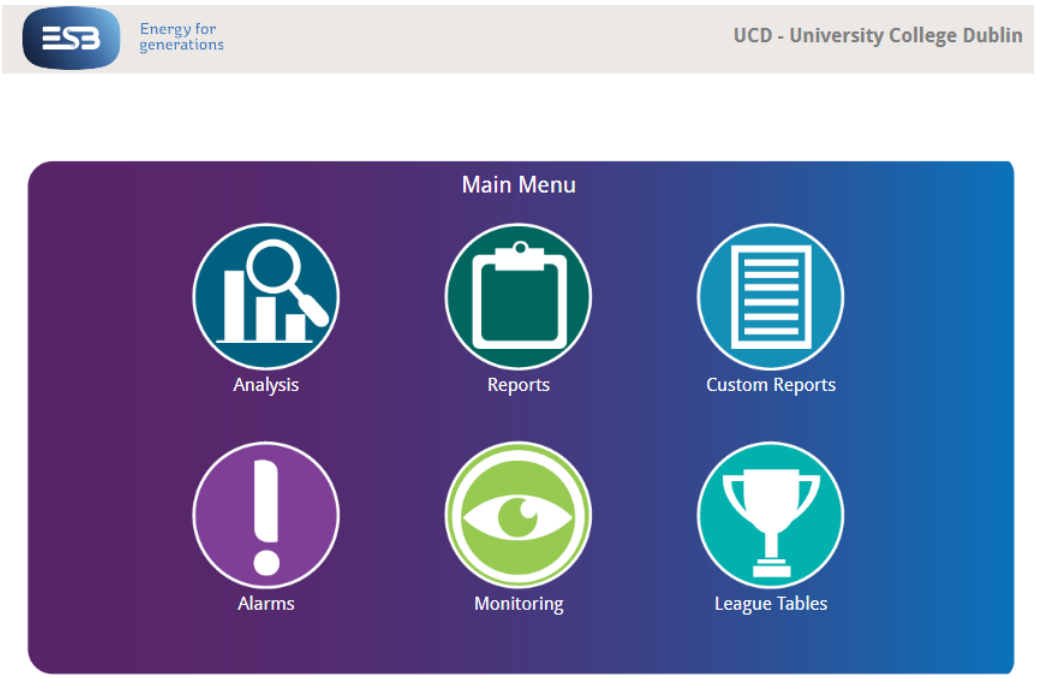

Figure 4.2: Cylon Home Page

This dataset delves into the collaborative efficiency of ABB Cylon® and Unitron BMS, attaining a granular view of the energy dynamics within the UCD campus. This includes a wide variety of datasets but is not limited to real-time energy consumption (gas, water, electricity), weather statistics, HVAC system operations, lighting controls, and occupancy trends. It further proves the potential of an integrated building management system in promoting a campus environment

that is sustainable and energy-efficient. This supports the relationship between ABB Cylon® with Unitron BMS and UCD.

Through employing these sophisticated control systems, UCD was able to prove its commitment to the cause of not only reducing the carbon footprint but also setting a benchmark of energy management for educational institutions. The project will utilize this broad set of data to assess efficiency in energy, unearth avenues of optimization, and propose pragmatic, sustainable energy solutions for the UCD campus[24].

The insights developed in this analysis will contribute critically toward the help of the much broader goals of enhancing energy sustainability and operational excellence across the campus.

Through access to the Computer Science building, the dataset unveils nuanced insights into 15-minute interval energy consumption patterns, the effectiveness of energy-saving initiatives, and the potential areas for optimization.

4.1.2 Data collection

In this section, our primary focus is on gathering essential energy consumption data and pertinent weather details, both of which have a substantial impact on the accuracy of our model predictions. Additionally, considering the variability in campus usage due to weekends, closing hours, bank holidays, special events and summer and winter vacations. Also, variations in prices can influence strategies for electricity use and conservation efforts. we will collect and generate related information to further refine our models. Following the data collection, we will meticulously process and prepare these datasets to ensure they are optimized for the training and testing phases of our models. This thorough preparation is designed to significantly improve the performance, accuracy, and reliability of our predictive models.

Identification of Relevant Variables

Our approach to discerning the variables that significantly affect energy usage and environmental patterns includes :

- **Historical Energy Usage Data:** Collection and analysis of historical energy consumption records across various campus buildings and facilities. In this step, we extract the Computer Science building relevant data as sample data to make predictions on electricity consumption.
- **Weather Patterns:** Integration of meteorological data to understand seasonal variations and their impact on energy demand.
- **Local Environmental Data:** Incorporation of local environmental factors influencing energy consumption trends.

Data Quality Assessment

Ensuring the quality and reliability of the collected data:

- **Data Accuracy:** Verification and validation processes to confirm the accuracy of recorded data.
- **Completeness and Consistency:** Assessment of data completeness and consistency across all variables.

- **Expert Validation:** Collaboration with domain-specific stakeholders or experts for data validation and verification.

Feature Engineering Techniques

Transformation of raw data into informative features:

- **Aggregation and Derivation:** Employing techniques to aggregate data or derive new features to capture complex relationships.
- **Pattern Recognition:** Creation of features to identify and capture intricate patterns affecting energy consumption or environmental trends.

4.1.3 Data Preprocessiong

This section describes the employment of quadratic polynomial interpolation for processing missing data and exhibits it to be more sensitive than linear interpolation if the data possesses characteristics of a non-linear set. For the fixed degree equal to 2, a quadratic curve will be fit to the known data points, and this model will l be used to estimate the missing values confidently. Generate key temporal features with the data timestamp to capture the intricate seasonal variations when the energy consumption that the present study focuses on is occurring within the campus buildings. The initial features after development are shown in Table4.1, which include the calendar information, Weather variables, historical data.

Feature Category	**Candidate Variables**
Calendar Information	Timestamp index, hour, day, weekend, month, is working hour(Binary), is working month(Binary)
Weather Information	Outside Temperature, Humidity, Rain rate, Wind Speed, Wind Direction, Light Lux
Historical Data	Electricity consumption

Table 4.1: Candidate variables for predicting electricity consumption.

It should be recognized that the observed trends are: energy use falls very considerably at weekends and on holidays and outside operational hours in parallel to the decreased activities at those times when the campus facilities are not being utilized. During the summer months, there will be much less electricity and gas use, since the demand for heating is so much, much lower than in winter time. These temporal dynamics are very critical in sharpening our accuracy in the data analysis toward making our interpolation temporal, accounting for not only the temporal dynamics of missing values but also dovetailing with the cyclical patterns typical of a campus environment. Figure4.3 shows the example of UCD Computer Science Building Electricity consumption over time in 2023.

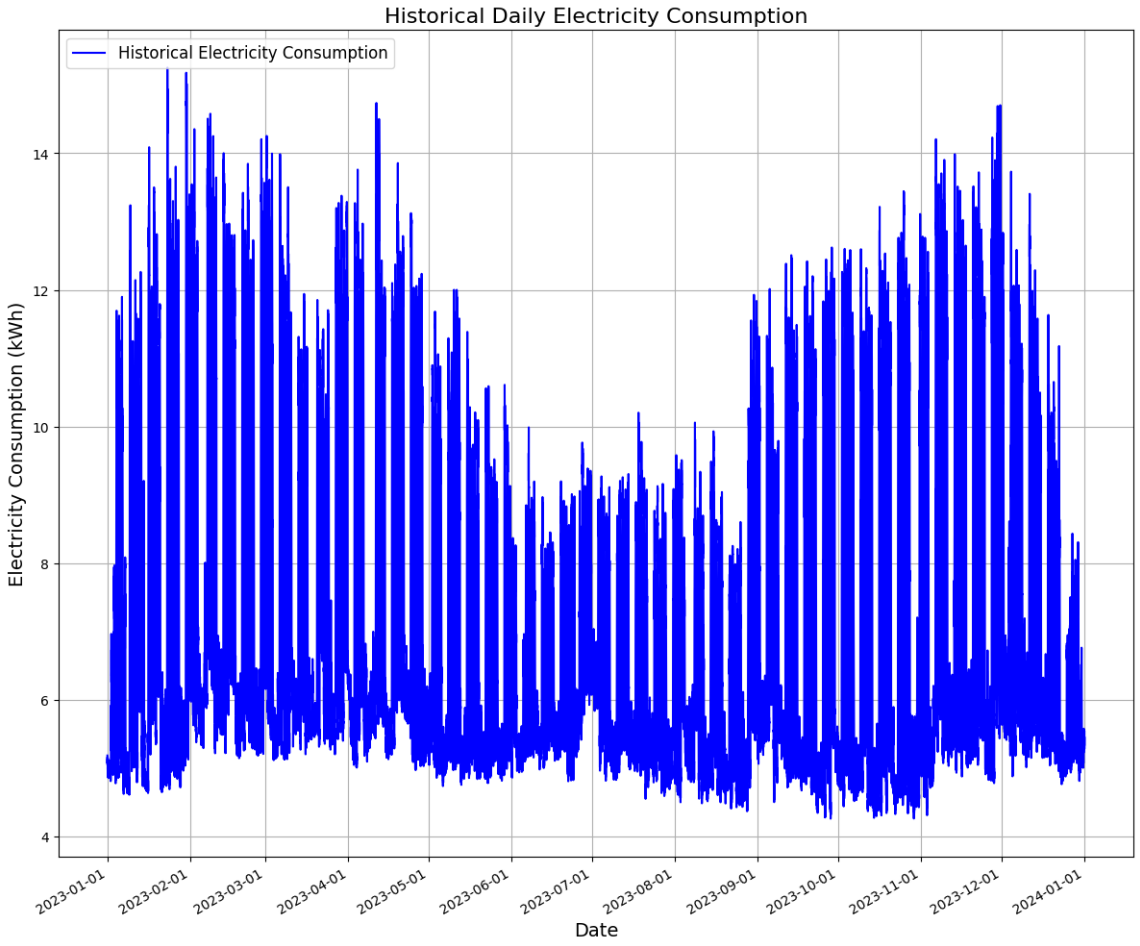

Figure 4.3: Electricity Consumption over 2023 for CS buiding in UCD campus

4.1.4 Feature Selection Methodology

Feature selection incorporates various steps in the identification and evaluation of relevant variables both in energy prediction and environmental patterns within the campus. These steps will be described in the following subsections that describe the methodology used.

Upon analysis of the dataset on the Cylon platform through bar plot visualizations, we observed that certain weather variables, such as rain rate, wind direction, and wind speed, exhibit negligible fluctuations. Consequently, these variables appear to have a minimal impact on model training and will be omitted to streamline the dataset and reduce complexity.

Feature selection comes across as one of the critical parts of the machine learning workflow, focusing specifically on pinpointing the most important features for a predictive model. This effort, on the other hand, tries to improve the performance of the model on the interpretability and highly computational demanding requirement by simply ignoring the irrelevant, less useful features. Recursive Feature Elimination (RFE) stands out among many feature selection methods—more so, it has come out to be the best when it is combined with ensemble methods like the RandomForestRegressor method.

We can note that RFE is a backward selection methodology: it iteratively builds the models by dropping the best- or worst-performing feature in each iteration. This process goes on, building the next model with one feature down in rank from the previous subset of features until all features are evaluated. The procedure focuses on the subset of features that contribute most significantly, though not collinearly, to the response/outcome variable of interest.

The approach uses RandomForestRegressor from sklearn.ensemble as the base model for RFE. RandomForestRegressor has a robust nature and, at the same time, it's very flexible. It develops through the formation of many decision trees in training and outputs the average prediction by these trees. This, therefore, deems it an excellent option for feature selection through RFE.

In this method, we use a RandomForestRegressor with 100 estimators. After training the model on the dataset, it is possible to extract feature importances and rank them. According to such ranking, each feature would make a relative contribution to predictions in this model. This would help provide a solid insight into which feature is the most powerful one for predicting the outcome.

The most valuable result of this process in Figure4.4 is a graphical representation of feature importances that clearly demonstrates the visual hierarchy of significance in the features. It helps in the choice of the most impactful features and, at the same time, contributes to ensuring transparency in the decision process of the model.

That thus leaves the RandomForestRegressor as more useful than just its normal, traditional role of a prediction model, but rather as an indispensable tool in meaningful feature analysis. Both usabilities of this lead us to a more informed and effective strategy for feature selection, with their inherent capabilities.[25]

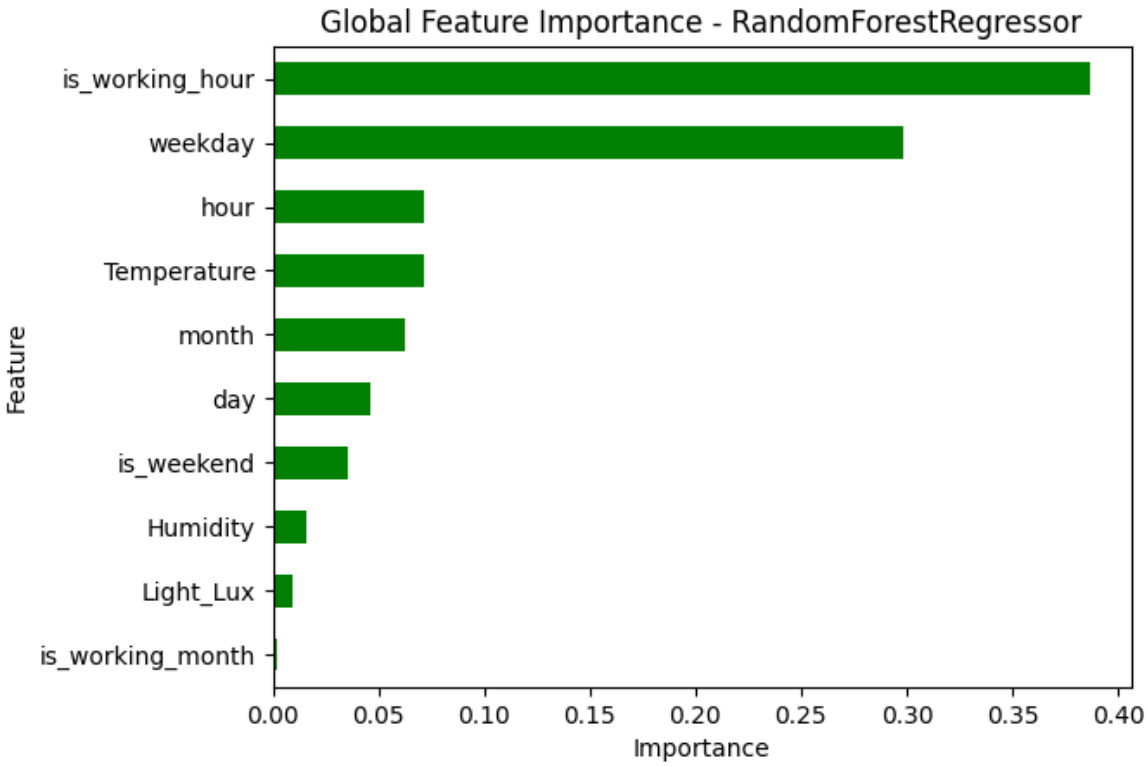

Figure 4.4: Feature importance sorting in descending

Through the RandomForestRegressor's feature importance analysis, we have ascertained the relative importance of various predictors in forecasting our target variable. The bar chart delineates that the features 'hour' and 'weekday' manifest as the most predictive factors, with 'is working hour' following closely. In light of these insights, and to streamline our model, we have elected to proceed with 'is working hours' and 'weekday' exclusively for our model training. This decision is founded on the objective to utilize the most influential features while optimizing the model's complexity and computational efficiency.

4.1.5 Autocorrelation Function (ACF))

Autocorrelation is a fundamental concept in time series analysis that measures the linear relationship between lagged values of the same time series. Specifically, if we have a time series represented by y_t, then the autocorrelation function (ACF) is defined as the set of correlations of y_t with its own lagged values y_{t-k}, for lags $k = 1, 2, \ldots$. The value of k signifies the lag, which could be in various time units such as days, months, or years. The ACF is particularly useful in identifying repeating patterns like trends or seasonality, as well as the degree to which past values influence future values.

Mathematically, the ACF at lag k is calculated as:

$$ACF(k) = \frac{\sum_{t=k+1}^{T}(y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}$$

where $\bar{y}$ is the mean of the time series, T is the total number of observations, and y_t is the value at time t[26].

The Autocorrelation Function (ACF) plot4.5 presented here exhibits the correlation between the

time series data and its past values over various lags, up to two weeks (1344 lags given the 15-minute intervals). Each peak in the plot represents a point where the data correlates with itself from an earlier time—a "lag." A lag of 96 corresponds to a full 24-hour cycle, signifying daily patterns, while the entire span of 1344 captures weekly cyclicality as well.

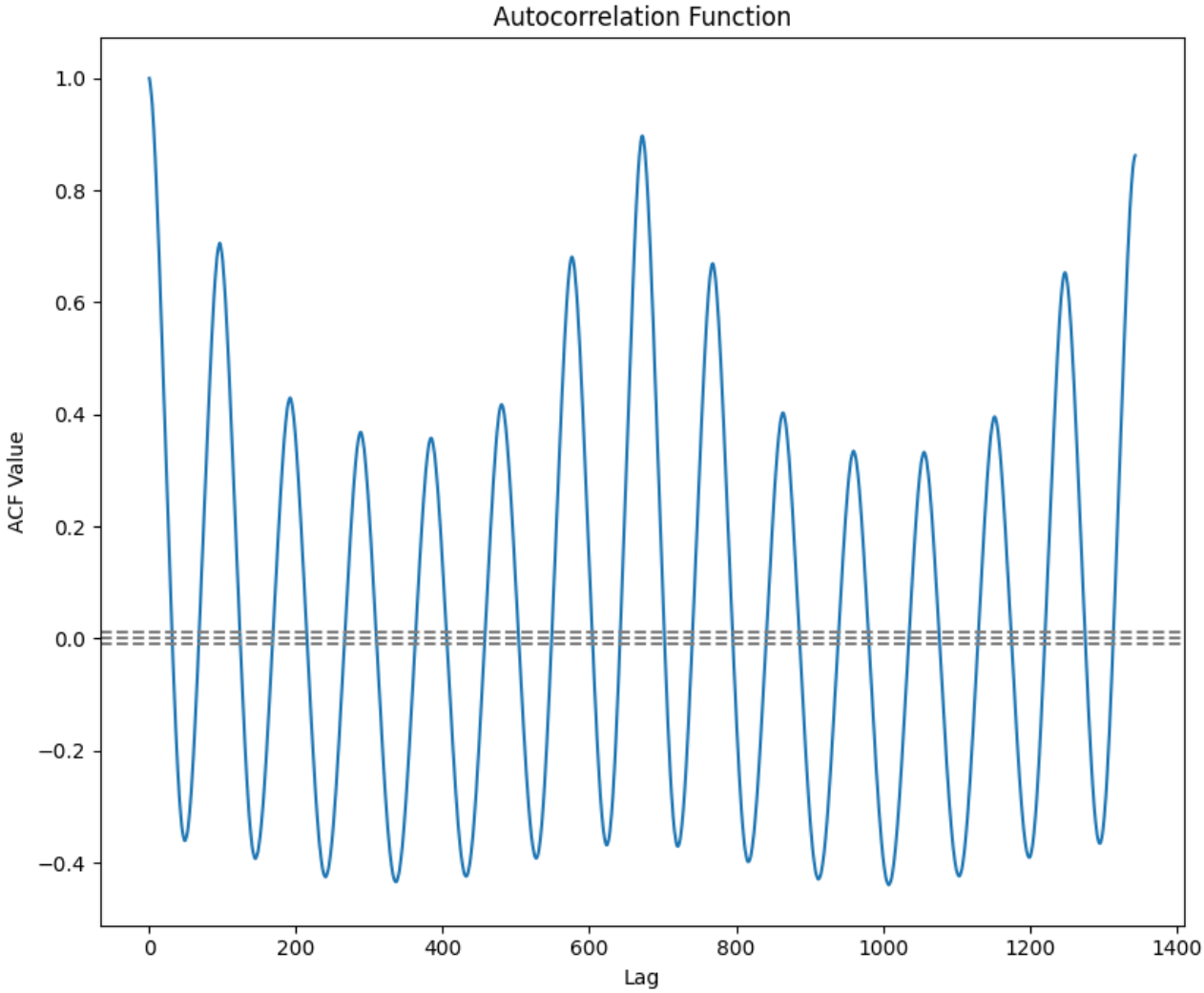

Figure 4.5: Autocorrelation Function ACF plot

The noticeable and recurring peaks that are shown at every 96th lag point to a significant daily pattern in the amount of power used, which is probably caused by regular operations in the campus building. These peaks show strong autocorrelation, which indicates that, after a day's break, the historical and present consumption patterns are remarkably comparable. In addition, the observation of peaks at multiples of 672 lags, or one week, would imply a weekly pattern, which is also critical for predicting.

Lag 4 is relevant for "hour-ahead" forecasting of the short term in such a way that since data is 15 minutes apart, it bears the effect of the most recent hour inside the future values. Temporal proximity would represent transient variations of power usage that do matter for one-hour-ahead forecasts.

With the collection of quarter-hourly data, the lag at 96 (one day before) should be accommodated for "day-ahead" forecasting. Most importantly, this lag plays a special role as it contains all activities and practices that are routine in nature and are most likely repeating themselves in the same time span the next day. Peaks at multiples of the lag 96 could be indicative of the diurnal cycle but would be taken into account in order to include the whole day rhythm in the model for

prediction.

Multiples of 672 lags will be at the weekly level, reflecting the weekly calendar of lectures, and operation cycles that could influence consumption patterns when looking for longer-term patterns related to forecasting.

This would have taken into account the repeating trend on an hourly, daily, and weekly basis by adding these lags specifically to the time series forecast model. In theory, this could make the algorithm much more precise as it makes predictions for future electricity usage.

In the context of sliding windows, which are used to create features for predictive models, the ACF helps determine the optimal "width" of the window, or in other words, how many past observations should be included to predict the current value. By identifying significant lags—those with correlations that cross the significance threshold on the ACF plot—an analyst can judiciously select relevant lag terms for the model. These significant lags are indicative of the intervals that hold predictive power due to their historical correlation with the target variable.

In summary, autocorrelation and its graphical representation through the ACF plot serve as essential tools in time series analysis for recognizing which historical data points (lags) have a statistically significant correlation with current data points. This significance is instrumental in setting up the sliding window for model input features, ensuring that the most relevant temporal dependencies are captured for forecasting purposes.

4.2 Machine Learning model strategy

First, in the model training stage, we are to build a function with the sliding window idea that would systematically generate features lagging the relevant forecasting interval. This is going to be very useful in being able to predict future values with great accuracy. Split of the dataset will be done equally for training and testing so that both stages are covered equally with the representation of data.

Moving on to model training, we leverage a variety of machine-learning algorithms such as Random Forest Regression, Linear SVM, XGBRegressor, Multiple Linear Regression, and Ridge Regression. This eclectic mix was chosen in order to compare a breadth of approaches, all looking at their effectiveness in using our dataset. Each model is fine-tuned through the optimization of hyperparameters and thoroughly cross-validated in its predicting power.

4.2.1 Sliding Window

The sliding window method is a systematic approach utilized in various domains such as machine learning and signal processing. It segments a dataset into overlapping segments, or "windows," of fixed size and sequentially processes each segment. By partitioning the data in this manner, the sliding window simplifies the analysis and application of algorithms by concentrating on smaller, more manageable portions of data at any given time[27].

When working with time series data, where the most recent observations may have a significant impact on forecasts, this strategy is especially helpful. Because the data maintains its temporal sequence in each window, features reflecting the dynamics of that particular era can be extracted. A thorough understanding of the time-dependent patterns and changes is provided by the window's ability to collect various subsets of the data as it moves across it. See example Figure4.6

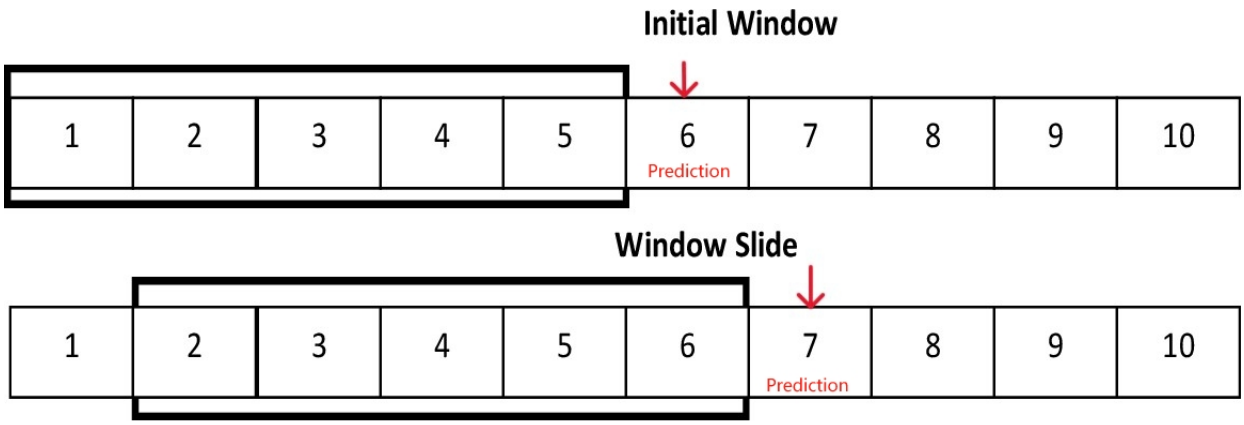

Figure 4.6: Sliding Window Example Source-[28]

Using the lagged data within each window to create feature sets that are used to predict future observations is how the sliding window is implemented in predictive modeling. By focusing on specific correlations within the time series, this adapted approach makes it easier to construct and develop reliable predictive models.

As illustrated in Figure 4.7, the sliding window method is applied to the dataset with hour-ahead prediction, using a selected range of lag terms from the past hour up to one day (t-4 to t-96) to predict the next value, as the target (t). This approach, while computationally intensive, captures the essential temporal dynamics for accurate forecasting. The next tasks may focus on reducing computational complexity by optimizing the selection of lag terms combined with the last step Autocorrelation function plot.

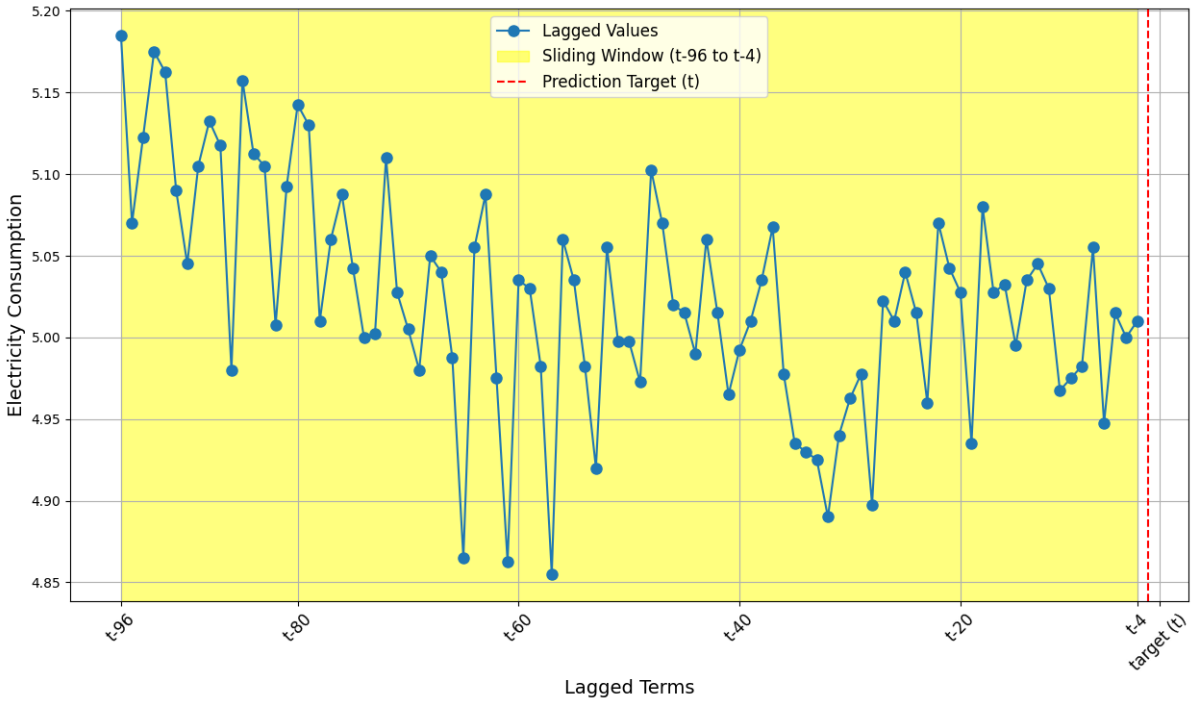

Figure 4.7: Example for Sliding Window apply in electricity consumption prediction

4.2.2 Transformation of Time Series Data for Supervised Learning

The sliding window technique, as implemented in this function, involves systematically creating lagged observations—windows of past data points—that serve as input features for the forecasting model. See the following Pseudocode for this function.

```
Function: series_to_supervised
Purpose: Transforms time series data into a structured format suitable for
         supervised learning, implementing a sliding window approach for
         feature and target generation.

Input:
    data - Time series data, either a list or a DataFrame.
    n_in -  lag terms
    n_out - number of future steps to forecast.

Process:
    1. Initialize:
       - Determine the number of variables (n_vars) based on
         the input type (list or DataFrame).
       - Convert input data to a DataFrame if not already.

    2. Generate Features:
       - For each lag in n_in:
           a. Shift the DataFrame down by 'i' positions to create lagged features.
           b. Name these new columns as 'var{j}(t-i)', where 'j' is the variable index.

    3. Generate Targets:
       - For each step in n_out:
           a. Shift the DataFrame up by 'i' positions to set future targets.
           b. Name these new columns as 'var{j}(t+i)', adjusting for forward-looking
           indices.

    4. Compile DataFrame:
       - Concatenate all generated columns into a single DataFrame.
Output:
    - Returns a DataFrame which should include all features and targets
      with (t-i) lag terms. which is ready to split into train and test sets.
```

This process transforms the time series into a format where each row in the dataset represents a separate instance of a sliding window, with the following characteristics:

Window Formation:

For each instance, the window comprises a set number of past observations (lags) defined by the n_in parameter. These observations represent the input features, capturing the essential dynamics up to that point in the series. Combined with the results of the Autocorrelation function in the previous stage, we obtained the optimal window size and the most relevant lag terms.

Target Alignment:

Correspondingly, the future values to be predicted, determined by the n_out parameter, are aligned

with each window. These values represent the outputs or targets for each set of inputs, facilitating the training of supervised learning models to predict future time series values based on past data.

Sequential Integrity:

Sequential integrity means each window slides forward by one timestep, making sure that each segment of the data is used for training the model. That process helps maintain the intrinsic temporal connection of the time series and is very important for learning from sequential data.

Scientific Justification and Utility:

The sliding window technique is foundational in time series analysis, particularly when the objective is to predict future values based on observed historical data. It aligns with the principle that 'the past informs the future, enabling models to uncover and utilize temporal patterns and relationships within the data. By reformatting the time series into a supervised learning framework, the technique allows the use of standard machine learning algorithms, which typically require fixed-length input and output vectors.

4.2.3 Training and Test set Split

In preparation for the creation of training and test sets, superfluous columns—particularly extraneous lag terms not aligned with the target variable—were excluded from the dataset. This pruning strategy, grounded in the redundancy of certain variables such as 'working hours' and 'day of the week', which directly coincide with our target test values, effectively simplifies the dataset's complexity.

After this dimensionality reduction, the data was split into training and test sets, allocated equally at a 50% / 50% distribution. This partition ratio was finalized after several experimental iterations aimed at optimizing the model's validation process. The X_train and X_test were normalized by the Min-Max scaler which is a normalization technique that adjusts the scale of input features or variables to a designated range, typically [0,1]. This scaling method ensures that for each feature, the smallest value is transformed to 0 (the minimum of the new scale), and the largest value is transformed to 1 (the maximum of the new scale). All other values are scaled to fall between 0 and 1 in a way that is directly proportional to their value with the minimum and maximum. This preserves the relationships in the data while standardizing the range of feature values[29].

4.2.4 Machine Learning short-term prediction

The proposed ML strategy [30] integrates diverse methodologies for optimizing energy efficiency on the UCD campus. Leveraging ML for energy consumption forecasting involves using historical data to predict usage patterns, aiding in identifying inefficiencies and optimizing resource allocation.

In this study, we employ five distinct machine learning models—Random Forest Regression, Support Vector Machine with a linear kernel (Linear SVM), XGBoost Regressor, Multiple Linear Regression (MLR), and Ridge Regression—to evaluate their efficacy in short-term electricity consumption forecasting. These models are tested across various predictive scenarios, including immediate step-ahead, hour-ahead, and day-ahead forecasts, utilizing appropriately selected lag terms to capture the temporal dynamics unique to each prediction horizon.

4.2.5 Hyperparameter and Cross-validation

Cross-validation is a very vital and robust method of modeling performance evaluation by dividing model data into several training and validation sets. It helps in conducting a detailed assessment through repeated training and measuring performance across these sets. Hyperparameter tuning includes strategies such as grid search, random search, and Bayesian optimization. These techniques assist in optimizing model performance by searching for the best values of the hyperparameters in a systematic manner. With knowledge of such hyperparameter tuning methods, practitioners are able to navigate the hyperparameter space in a way that enhances the model's accuracy and reliability [31].

In this study, we implement the Randomized Grid Search technique to systematically identify the optimal hyperparameters for an array of predictive models. Given the distinct dynamics presented by datasets curated for step-ahead, hour-ahead, and day-ahead forecasts, individualized Randomized Grid Searches are executed to tailor the model parameters to the specific temporal context.

Of great consideration in our hyperparameter tuning method is the use of TimeSeriesSplit in cross-validation. This method is best applicable to time series datasets, considering the natural temporal order presented by the dataset. TimeSeriesSplit splits the dataset in a forward chaining manner; each fold will contain all its preceding data, and hence the model training is done with no time-point discontinuity in the history[32].

The utility of the Randomized Grid Search, in this analytic framework, is twofold. Avoiding the computational intensity of the exhaustive grid search, thus making the process efficient, is the first merit of a random sampling of parameter combinations. Secondly, it injects stochasticity in the optimization of parameters and thus explores a wider potential solution space, which in many cases could yield more robust and generalizable models. TimeSeriesSplit combined with Randomized Grid Search provides a methodically sound approach to optimizing model performance for time series analysis. With respect to the sequential nature of the data, it still offers an efficient path across the hyperparameter landscape.

4.3 Performance evaluation strategy

For energy prediction, especially when dealing with data at 15-minute intervals, it's crucial to understand not just the magnitude of the forecast errors (as provided by RMSE), but also the direction and bias of these errors. Mean Bias Error (MBE)4.1 can be particularly informative as it helps identify systematic overestimations or underestimations made by the forecasting model[33].

$$\text{MBE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{4.1}$$

- n is the number of observations.
- y_i is the actual energy consumption at the i-th interval.
- $\hat{y}_i$ is the predicted energy consumption at the i-th interval.
- The differences between the predicted and actual values are summed and then averaged over all observations.

From the whole process of forecasting, forecast accuracy bears a lot of importance, for it provides a tool for the evaluation of how close a forecast is to the actual results. It takes the difference between the expected values and those actually observed, hence provides a very useful tool in the evaluation of how close a forecast is to the actual results. This will be important in industries like demand planning, finance, and weather prediction, energy consumption forecasting. Accurate prediction with high accuracy is the key to minimum risk and effective allocation of resources in making valid decisions. Among other ways, root mean square error (RMSE) gauges the forecast's accuracy. It may be used in many statistics to represent a measure of the strength of the forecast. The root mean square error (RMSE)4.2 is the square root of the average value of the sum of squares of forecasting errors, giving equal importance to both positive and negative deviations from the actual value[34].

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4.2}$$

- The sum of the squared differences between the predicted values and actual values is computed, then averaged, and finally, the square root of that average is taken to produce the RMSE.

Chapter 5: Testing and Evaluation

Day-Ahead Prediction(Table5.1)

For day-ahead predictions, aimed at forecasting values 24 hours in advance, the selected lag terms are determined based on the Autocorrelation Function (ACF) plot. These lag terms represent key cyclic patterns observed in the data:

- **Daily Lags:** We include lag terms for 1 to 7 days behind the target time (*t-96, t-192, t-288, t-384, t-480, t-576, t-672*). Each term represents a day's shift, calculated as 24 hours $\times$ 4 quarters/hour = 96 lags/day.
- **Biweekly Lag:** We also incorporate a biweekly lag (*t-1344*), capturing the pattern exactly two weeks prior to the prediction date. This term helps in understanding fortnightly cyclical trends that may affect the forecast.

In the Day Ahead Forecasting context, the RandomForestRegressor exhibited a respectable balance between error magnitude and bias, reflected by an RMSE of 0.7167 and an MBE of 0.0668, indicating a moderate overestimation. The SVR model demonstrated a higher RMSE, thus less accuracy, but also a small positive bias. The MLPRegressor showed a competitive RMSE of 0.7115 with the least bias among models at 0.0255, making it a strong candidate when considering both accuracy and bias.

Table 5.1: Day Ahead Forecasting Model Performance

Model type	**RMSE**	**MBE**
RandomForestRegressor: *max_depth=15, min_samples_leaf=4, min_samples_split=5*	0.7167	0.0668
SVR: *C=0.1, epsilon=1, kernel='linear'*	0.8288	0.0602
XGBoost: *colsample_bytree=0.9, learning_rate=0.05, max_depth=5, n_estimators=100*	0.7179	0.0592
MLPRegressor: *alpha=0.01, hidden_layer_sizes=(50,), max_iter=1000*	0.7115	0.0255
Ridge Regression: *alpha=1291.5497*	0.8099	0.0095

Hour-Ahead Prediction(Table5.2)

In the case of hour-ahead predictions, which forecast the electricity demand for the next hour, the lag terms include shorter intervals to capture more immediate trends:

- **Hourly Lags:** We use the lags corresponding to each 15-minute interval within the second hour prior to the target hour (*t-4, t-5, t-6, t-7*). These are critical for capturing the immediate past conditions influencing the next hour.
- **Daily Lag:** The lag at *t-96* represents the same quarter-hour of the previous day, providing a daily cyclical perspective.
- **Weekly Lag:** The lag at *t-672* (7 days prior at the same quarter-hour) helps to incorporate weekly seasonality into the forecast.

For the Hour-Ahead Forecasting. The MLPRegressor has the lowest RMSE of 0.4799, which means it has the best accuracy for short-term forecasts and a moderate overestimation bias (MBE of 0.0368). Both XGBoost and RandomForestRegressor had strong performance; however, the former had a little higher bias. In particular, the SVR model's negative MBE indicated a propensity toward underestimating.

Table 5.2: Hour Ahead Forecasting Model Performance

Model type	RMSE	MBE
RandomForestRegressor: *max_depth=10, min_samples_leaf=4, min_samples_split=10, n_estimators=200*	0.5103	0.0662
SVR: *C=10, kernel='linear'*	0.5758	-0.0244
XGBoost: *subsample=1, n_estimators=100, max_depth=5, learning_rate=0.05, colsample_bytree=0.9*	0.4953	0.0484
MLPRegressor: *hidden_layer_sizes=(100,), alpha=0.0001, activation='tanh'*	0.4799	0.0368
Ridge Regression: *alpha=0.0464*	0.5662	0.0226

Step-Ahead Prediction(Table5.3)

For step-ahead prediction, focusing on the very short term, likely predicting the next 15 minutes to an hour, we select lags that give a high-resolution view of the immediate past:

- **Immediate Past:** We use lags *t-1, t-2, t-3, t-4*, which represent the past four 15-minute intervals leading up to the prediction point. These lags are crucial for capturing the most recent trends that heavily influence the immediate future.
- **Daily Lag:** Similar to the hour-ahead model, the *t-96* lag offers a daily repeat cycle insight, helping to understand daily patterns at the exact time.
- **Weekly Lag:** The *t-672* lag provides insights from the same time a week ago, which is important for recognizing weekly patterns and anomalies.

In the step-ahead forecasting evaluation, the RandomForestRegressor emerged as the most accurate model, achieving the lowest RMSE of 0.2806, indicating superior precision in its predictions with

a small positive bias (MBE of 0.0142). This model strikes a balance between reliability and a marginal tendency to overestimate. The MLP Regressor, while presenting the lowest bias, does so at the expense of a slight underestimation. Overall, the RandomForestRegressor stands out as the optimal choice for scenarios where precision is paramount and a slight overestimation is acceptable within the operational context.

Table 5.3: Step Ahead Forecasting Model Performance

Model type	**RMSE**	**MBE**
RandomForestRegressor: *max_depth=10, min_samples_leaf=4, min_samples_split=10*	0.2806	0.0142
SVR: *C=10, kernel='linear'*	0.2942	-0.0082
XGBoost: *colsample_bytree=1, learning_rate=0.1, max_depth=3, n_estimators=50*	0.2852	0.0243
MLPRegressor: *activation='tanh', alpha=0.001, hidden_layer_sizes=(50,), max_iter=1000*	0.2833	-0.0184
Ridge Regression: *alpha=0.0001*	0.2909	0.0043

5.1 Summary

In our analysis of building energy consumption within a university campus, we evaluated the forecasting performance of five different models across three prediction intervals: day-ahead, hour-ahead, and step-ahead. The overall performance by RMSE in Plot5.1.

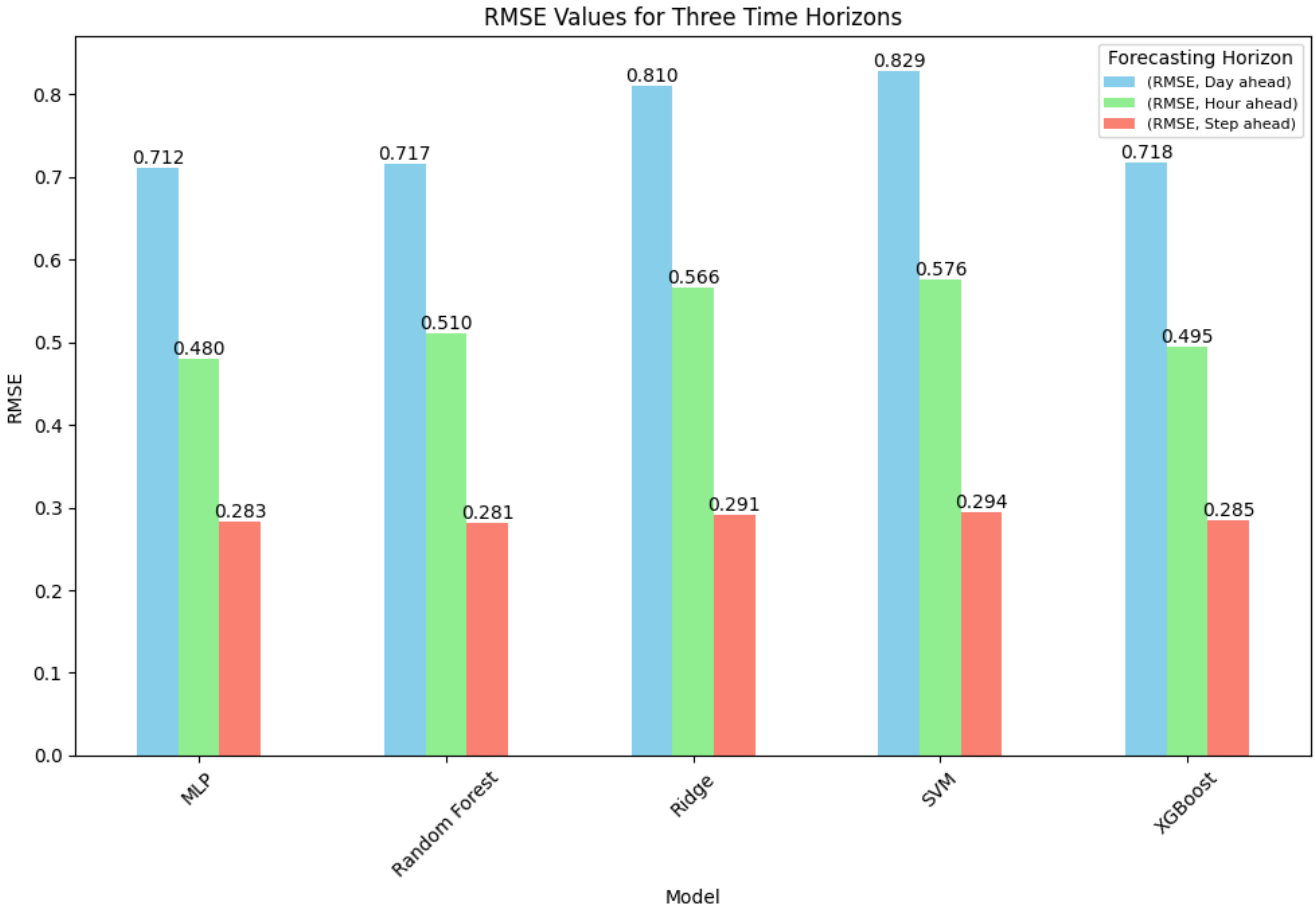

Figure 5.1: Models Performance by RMSE

- **Day-Ahead Prediction:**The best-performing model is the MLP Regressor, possessing the lowest RMSE- 0.712, for the day-ahead scenarios. Thus, it would be reasonable to select this model as the appropriate one for daily energy use in the campus building since it will help to forecast energy consumption accurately and choose relevant strategies to have enough energy in case of fluctuations, thus saving the acquisition cost.
- **Hour-Ahead Prediction:** The MLP Regressor model is superior for our hour-ahead prediction- 0.480. This model will work best when immediate decisions are required such as UCD's need to decide whether to adjust specific settings of HVAC. The model helps to increase energy efficiency and guarantee the maximum comfort of the users.
- **Step-Ahead Prediction:** For our step-ahead prediction, the most accurate model is RandomForest Regressor since the corresponding RMSE equals 0.281. This model also helps to increase the efficiency of the system using real data, which is particularly useful for sudden changes in UCD's activities, and events, or if the staff and students actually need more comfortable conditions.

5.1.1 Actual values and Prediction comparison

A Plot5.2 shows the actual and predicted electricity consumption for the CS building, the data selected from the last day with 96 data points segmented into 15-minute intervals.

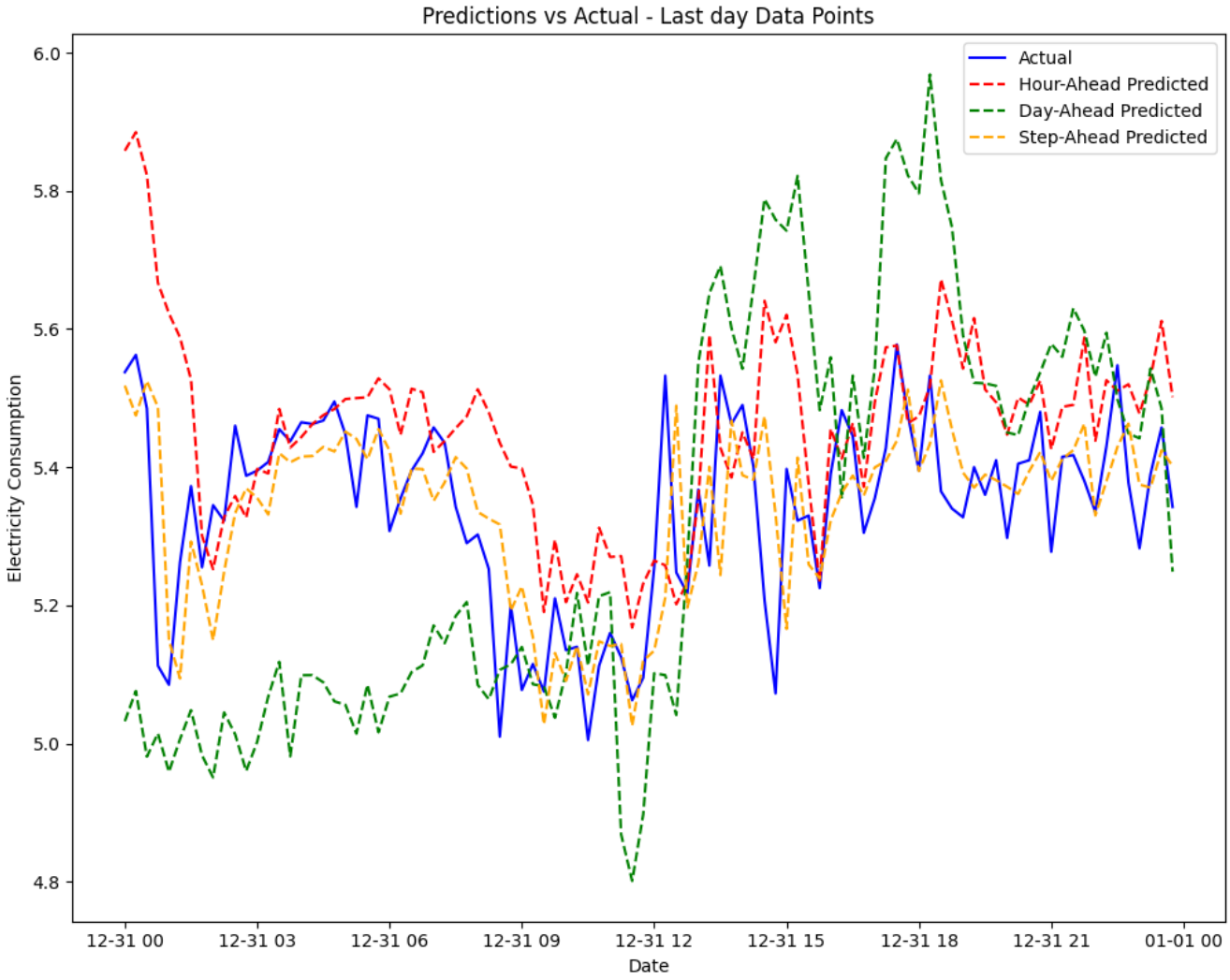

Figure 5.2: Actual VS Prediction on Computer Science building

- **Actual Consumption (Blue Solid Line):** This line represents the actual data recorded for electricity usage. It serves as an excellent reference for evaluating the performance of all predictive models applied to the data.
- **Hour-Ahead Predicted (Red Dashed Line):** The predictions made an hour in advance generally follow the actual consumption trend closely but include some deviations. These deviations suggest optimal performance, although there may be slight delays in capturing spikes or drops in consumption.
- **Day-Ahead Predicted (Green Dashed Line):** The day-ahead predictions show significant deviations from the actual data, which is expected given the complexity of forecasting further out. The deviations are particularly pronounced during peak and trough periods, highlighting the challenges of accurately predicting daily consumption patterns.
- **Step-Ahead Predicted (Orange Dashed Line):** This line tracks rapid changes in consumption more closely than the day-ahead predictions, focusing on the immediate future.

However, it sometimes overshoots or undershoots, indicating that the model is making rapid adjustments to real-time fluctuations.

From the graph, we observe that all predictive models capture the general trend of consumption, yet they each exhibit distinct variances from the actual data. These variances highlight the individual forecasting strengths and weaknesses over different horizons. The hour-ahead and step-ahead predictions are more volatile, adapting quickly to changes, while the day-ahead predictions offer a smoother outlook on expected consumption.

In practical terms, for a university building, the step-ahead and hour-ahead forecasts are critical for managing real-time energy systems, such as HVAC, to respond to occupancy patterns. Day-ahead predictions are useful for strategic decisions regarding energy procurement and daily operational planning, despite being less sensitive to immediate fluctuations.

The overall assessment indicates a sophisticated energy management system at the university, capable of leveraging various predictive models to optimize efficiency and reduce operational costs in both day-to-day and real-time energy use.

On-campus operations and forecasts move ahead in days; for UCD operations, the MLP Regressor model will work the best given its ability to forecast for the upcoming day, hour, or even next step. Therefore, the models that have been suggested will not only predict energy levels but manage energy on campus effectively. Thus, UCD's actual energy will be used in a more smart way in the future.

Chapter 6: Conclusions and Future Work

6.1 Conclusions

In the quest for energy optimization, short-term energy demand forecasting serves as a linchpin for efficient grid operation and resource management. The advancement of machine learning offers promising solutions to capture the non-linearity of energy patterns, propelling beyond traditional white-box modeling. This study harnesses the robustness of random forest algorithms and Multi-layer perceptron, capitalizing on their strong fitting capabilities to address the fluctuating nature of energy demand.

Employing a dynamic feature selection strategy through a sliding window approach, we developed a model utilizing 2023's electricity consumption data from UCD's School of Computer Science. The model's effectiveness was ascertained by its ability to closely predict actual loads, as indicated by low mean bias error (MBE) and root mean square error (RMSE). These metrics not only reflect the model's precision but also its practical relevance in real-world applications.

Our findings reveal that random forest and MLP models, updated periodically to adapt to new data, can significantly enhance short-term energy forecasting. The implication of such a development is manifold, impacting not only grid stability and pricing strategies but also contributing to informed decision-making in energy trading and emergency response planning.

The implementation of machine learning, in the realm of short-term energy demand forecasting, presents a transformative tool for power systems. The accuracy and adaptability demonstrated by the model in this study signal a shift towards more agile and data-driven energy management within university campuses and beyond. It establishes a foundation for future research to refine forecasting models further, potentially extending their applicability across different sectors and scales.

6.2 Future Work

6.2.1 Data Expansion:

The scope of this data has to be widened so as to bring over one year of energy consumption data into the scope to be used. Such widened scope shall allow capturing many more patterns, not only normal variants but also seasonal tendencies and anomalous trends.

6.2.2 Advanced Modeling Techniques:

Key success factors with the potential for further exploration in advanced approaches of machine learning and deep learning include Convolutional Neural Networks (CNN) and advanced configurations of Multi-Layer Perceptrons (MLP). This, for sure, will be the prominent approach of feature

extraction and pattern recognition, so foretelling can attain way more precision levels.

6.2.3 Special Day Analysis:

With bank holidays and special days like campus events, which have a lot of effect sizes on changes in energy usage, we plan to introduce a calendar-aware component into our models to cater to such changes in consumption driven by events and account for these irregular, non-working days. This will make the models respond better on such atypical days.

6.2.4 Scenario-Specific Forecasting:

We diversify the modeling approach towards estimating forecasts for more than one scenario—from the typical academic days period to heavy periods of events. This would mean further widening the dataset and fine-tuning the models to be both contextual and adaptive.

6.2.5 Long-term Forecasting Experimentation:

If at all the enriched-for-the-experiment data will be long-term in nature, then, of course, we too will experiment in a long-term forecast, past the horizons of prediction as envisaged at present, to a little further into the future. This could prove invaluable for strategic planning and long-term resource allocation.

The project aims to provide forecasts that are informed not only in their accuracy but in the context so that the derived management strategies are cost-effective and environmentally sustainable. The evolution of this project will continue to support UCD's transition to a more data-driven, energy-efficient campus.

Bibliography

1. *UCD ENERGY MANAGEMENT* https://ucdestates.ie/about/sustainability/energy-management/.
2. group, A. *Cylon* https://new.abb.com/low-voltage/products/building-automation/product-range/abb-cylon.
3. *scikit-learn* https://scikit-learn.org/stable/.
4. UCD, C. E. R. G. *Retrofit Strategy and Policy* https://www.ucd.ie/biace/projects/retrofitstrategyandpolicy/.
5. From Department of the Environment, C. & Communications. *Energy Security Emergency Group (ESEG)* https://www.gov.ie/en/publication/de3cf-energy-security-emergency-group-eseg/.
6. Of the Taoiseach, F. D. *Energy challenges* https://www.gov.ie/en/publication/96c7a-energy-challenges/#.
7. Yañez, P., Sinha, A. & Vásquez, M. Carbon Footprint Estimation in a University Campus: Evaluation and Insights. *Sustainability* **12,** 181. https://doi.org/10.3390/su12010181 (2020).
8. Adenle, Y. A. & Alshuwaikhat, H. M. Spatial Estimation and Visualization of CO2 Emissions for Campus Sustainability: The Case of King Abdullah University of Science and Technology (KAUST), Saudi Arabia. *Sustainability* **9,** 2124. https://doi.org/10.3390/su9112124 (2017).
9. Snyder, H. Literature review as a research methodology: An overview and guidelines. https://doi.org/10.1016/j.jbusres.2019.07.039 (2019).
10. *What are the different types of review?* https://support.covidence.org/help/types-of-review-explained.
11. Scopus. *Scopus Preview* https://www.scopus.com/.
12. Direct, S. *Science Direct* https://www.sciencedirect.com/.
13. Zotero. *Zotero* https://www.zotero.org/.
14. ASReview. *ASReview* https://asreview.nl/.
15. Bampoulas, A., Pallonetto, F., Mangina, E. & Finn, D. P. An Ensemble Learning-Based Framework for Assessing the Energy Flexibility of Residential Buildings with Multicomponent Energy Systems. https://doi.org/10.1016/j.apenergy.2022.118947 (2022).
16. OpenStudio®. https://openstudio.net/.
17. *Energy Plus* https://energyplus.net.
18. Lincoln C. Harmer, G. P. H. Using calibrated energy models for building commissioning and load prediction. https://www.sciencedirect.com/science/article/pii/S0378778814009244?via%3Dihub (2022).
19. Jui-Sheng Chou, N.-T. N. Time series analytics using sliding window metaheuristic optimization-based machine learning system for identifying building energy consumption patterns. https://www.sciencedirect.com/science/article/pii/S0306261916306717 (2016).
20. Vafaeipour, M., Rahbari, O., Rosen, M. A., *et al.* Application of Sliding Window Technique for Prediction of Wind Velocity Time Series. *International Journal of Energy and Environmental Engineering* **5,** 333–342. https://link.springer.com/article/10.1007/s40095-014-0105-5 (2014).

21. Irfan Ahmad Khan Adnan Akbar, Y. X. Sliding Window Regression based Short-Term Load Forecasting of a Multi-Area Power System. https://arxiv.org/ftp/arxiv/papers/1905/1905.08111.pdf.

22. *Zairui$_{Z}hang_{G}itLab$* https://gitlab.com/ucd_fyp24/short-term-prediction-on-ucd-cs-buiding.

23. ABB. https://global.abb/group/en.

24. *University College Dublin | UCD Centralised control of 15,500 points across 300,000 m2 campus* https://library.e.abb.com/public/3d7305b4488b49d2b2fd827ad26e7b7b/ABB_Cylon_A4_UniversityCollege_CaseStudy_global.pdf.

25. Seçkin, K. Recognition Model for Solar Radiation Time Series based on Random Forest with Feature Selection Approach. https://ieeexplore.ieee.org/document/8990664 (2024).

26. JimFrost. *Autocorrelation and Partial Autocorrelation in Time Series Data* https://statisticsbyjim.com/time-series/autocorrelation-partial-autocorrelation/.

27. DataOverload. *Sliding Window Technique — reduce the complexity of your algorithm* https://medium.com/@data-overload/sliding-window-technique-reduce-the-complexity-of-your-algorithm-5badb2cf432f.

28. Hota, H. S., Handa, R. & Shrivas, A. K. *Time Series Data Prediction Using Sliding Window Based RBF Neural Network* in (2017). https://api.semanticscholar.org/CorpusID:172129481.

29. Loukas, S. *Everything you need to know about Min-Max normalization: A Python tutorial* https://towardsdatascience.com/everything-you-need-to-know-about-min-max-normalization-in-python-b79592732b79.

30. Schoenfeld, J. *Using Machine Learning to Improve Building Energy Efficiency* https://www.buildingsiot.com/blog/using-machine-learning-to-improve-building-energy-efficiency-bd.

31. Singh, S. *Cross Validation and Hyperparameter Tuning: A Beginner's Guide* https://medium.com/@sandeepmaths04/cross-validation-and-hyperparameter-tuning-a-beginners-guide-96d258eedee7.

32. Staff, P. E. *Cross-Validation strategies for Time Series forecasting* https://hub.packtpub.com/cross-validation-strategies-for-time-series-forecasting-tutorial/.

33. Handbook of Energy Efficiency in Buildings, 2. *Mean Bias Error* https://www.sciencedirect.com/topics/engineering/mean-bias-error.

34. FasterCapital. *Root mean square error: Understanding Forecast Accuracy using Root Mean Square Error* https://fastercapital.com/content/Root-mean-square-error--Understanding-Forecast-Accuracy-using-Root-Mean-Square-Error.html.