
Forecasting the carbon intensity of the Portuguese electric power system

António Leandro Pacheco Campos

Dissertation

Master in Modelling, Data Analysis and Decision Support Systems

Supervised by:

PhD Ricardo Jorge Bessa

Co-supervised by:

PhD Bruno Miguel Delindro Veloso

2024

Acknowledgments

I would like to express my genuine gratitude to Professors Ricardo Bessa and Bruno Veloso for their support and guidance throughout my research. Their expertise and insights have been decisive to the success of this study.

I'm grateful for the rich insights and camaraderie shared with my fellow students and colleagues throughout the course. A special thank you to Bruno Teixeira and Paulo Lima for their companionship, invaluable brainstorming sessions, and the many laughs we shared, all of which greatly inspired me.

Lastly, I am also immensely thankful to my family and friends for their patience and support during this challenging period. I eagerly look forward to making up for the time spent apart.

Abstract

This dissertation advances the methodology for forecasting carbon intensity in the Portuguese electric power system, integrating machine learning techniques with an emphasis on advanced feature engineering and the use of Evolutionary Forest algorithms. The primary intention was to enhance the precision of day-ahead carbon intensity predictions by developing and refining predictive models that can effectively integrate complex patterns derived from electricity market variables, energy consumption data, and renewable energy projections. In addressing these objectives, the study employed a variety of machine learning models, both deterministic and probabilistic, and introduced an application of Evolutionary Forest for feature enrichment. This method allowed for the extraction of non-linear interactions within the data, improving forecasting accuracy. The models were assessed using rigorous statistical methods and compared through a comprehensive benchmark analysis, primarily using the mean absolute error metric (MAE), to show the benefits of integrating advanced feature engineering techniques. Key insights revealed that the use of Evolutionary Forest elucidated the importance of features, enabling more interpretable modelling and insightful predictions. CatBoost model emerged as a top performer, representative of its effectiveness in handling complex feature interactions. This research aims to contribute to global sustainability efforts, focusing on enhancing the operational efficiency and environmental compliance of the energy sector.

Keywords: carbon intensity forecasting, machine learning, feature selection, evolutionary forest

Contents

1	Introduction	1
1.1	Motivation	2
1.2	Objectives	2
1.3	Contributions	3
1.4	Thesis Structure	3
2	State-of-the-art	5
2.1	Advancements in Carbon Intensity Forecasting	5
2.2	Methodologies in Carbon Intensity Forecasting	6
2.2.1	Innovative Machine Learning Approaches for Short-term Forecasting of Carbon Intensity in Power Grids	7
2.3	Carbon Intensity Forecasting in Europe and Similar Contexts	8
2.4	Applications of Carbon Emissions Forecasts	9
2.5	Data Sources and Challenges	10
2.6	Gaps and Opportunities	11
2.6.1	Opportunities directly related to the field under study	11
2.6.2	Opportunities related with ML evolution	11
2.6.3	Portuguese context	11
2.7	Concluding Remarks	12
3	Forecasting Methodology	13
3.1	Forecasting Problem Formulation	13
3.1.1	Factors Influencing Carbon Intensity	13
3.1.2	Methodological Approach to Forecasting Carbon Intensity	14
3.2	Feature engineering with Evolutionary Forest	16
3.2.1	Advanced Feature Engineering with Evolutionary Forest	17
3.3	ML Models	19
3.3.1	Random Forest and Extra Trees	19
3.3.2	AdaBoost and Gradient Boosting	19
3.3.3	XGBoost, LightGBM, and CatBoost	19
3.3.4	Linear Regression and LASSO	20
3.3.5	Artificial Neural Networks (ANN)	20
3.3.6	Support Vector Machine (SVM)	20
3.3.7	Dropouts meet Multiple Additive Regression Trees (DART)	20
3.4	Hyperparameter Tuning	21
3.5	Evaluation measures, MAE, RMSE, R^2 and Autorank Analysis	21

3.5.1	Mean Absolute Error (MAE)	22
3.5.2	Root Mean Squared Error (RMSE)	22
3.5.3	R-squared (R^2)	22
3.5.4	Methodological Framework of Autorank Analysis	22
3.6	Quantile Loss Function	23
3.6.1	Theoretical Foundation	23
3.7	Ensemble Models for Different Time Frames	24
3.7.1	Theoretical Foundations	25
3.7.2	Application to Different Time Frames	25
4	Case Study	27
4.1	Data Description and Exploratory Data Analysis	27
4.1.1	Datasets	27
4.1.2	Trend, Seasonal and Cycling Analysis	28
4.1.3	Correlation	33
4.2	Feature Engineering	35
4.2.1	Pre-processed Dataset	35
4.2.2	Further Feature Engineering with Evolutionary Forest	38
4.3	Models' evaluation	43
4.3.1	Autorank Analysis	44
4.3.2	Quantile Loss Function	46
4.3.3	Ensemble Models for Different Time Periods	48
5	Conclusion	51
5.1	Final Remarks	51
5.1.1	Key Insights and Implications	51
5.2	Limitations and Future Work	53
	Bibliography	55

Glossary

AdaBoost - Adaptive Boosting
AI - Artificial Intelligence
ARIMA - Autoregressive Integrated Moving Average
ARIMAX - Autoregressive Integrated Moving Average with Exogenous Inputs
ANN - Artificial Neural Network
CNN - Convolutional Neural Network
CO₂ - Carbon dioxide
DACF - Day-ahead carbon intensity forecasting of power grids using machine learning
DART - Dropouts meet Multiple Additive Regression Trees
DK2 - Danish bidding zone
DL - Deep learning
DPSF - Differenced Pattern Sequence-based Forecasting method
DR - Demand response
EDA - Exploratory Data Analysis
EEMD - Ensemble Empirical Mode Decomposition
EF - Extra trees
ENTSO-E - European Network of Transmission System Operators for Electricity
EV - Electric vehicles
FE - Future engineering
FFNN - Feedforward Neural Network
GBDT - Gradient boosting decision trees
GP - Genetic programming
HVAC - Heating, Ventilation, and Air Conditioning
IMFs - Intrinsic Mode Functions
INESC TEC - *Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência*
IoT - Internet of Things
LASSO - Least Absolute Shrinkage and Selection Operator
LightGBM - Light Gradient Boosting Machine
LIME - Local Interpretable Model-agnostic Explanations
LSTM - Long Short-Term Memory
MAE - Mean Absolute Error
MAPE - Mean Absolute Percentage Error
MDA - Mean Decrease in Accuracy
MDI - Mean Decrease in Impurity
ML - Machine learning
R² - R-squared
RF - Random forest

REN - *Redes Energéticas Nacionales*
REE - *Red Eléctrica de España*
RMSE - Root Mean Squared Error
SDG - Sustainable Development Goals
SHAP - SHapley Additive exPlanation
SVM - Support Vector Machine
XGBoost - Extreme Gradient Boosting

List of Figures

1.1	Thesis structure and its sections	4
2.1	Structure of literature review chapter and its sections	6
2.2	The Electricity Maps calculates and visually presents the source of the electricity, along with its corresponding global and real-time carbon footprint . .	9
3.1	Average Carbon Intensity in California. Source: Maji, Shenoy, and Sitaraman (2022)	14
3.2	Implemented methodological approach scheme	14
4.1	Energy production from hard coal in Spain, nuclear in Spain, gas in Portugal, and hydro run-of-river in Portugal	30
4.2	Hydro and solar energy production in Portugal	31
4.3	Hourly energy production for different sources in Portugal and Spain	32
4.4	Correlation Heatmap for different energy sources in Portugal and Spain . . .	33
4.5	Comparative performance analysis of MAE across various models for Input A - Raw Variables and Input B - Manual FE prior to the application of the Evolutionary Forest	37
4.6	Comparison of MAE performance across various ML models before and after application of the Evolutionary Forest	41
4.7	Top 10 features for the top 3 performing models	43
4.8	95% confidence intervals for MAE across different ML models, ranked to show variability in model performance	45
4.9	Actual values and predicted quantiles (10%, 50%, 90%) over one week, illustrating the quantile range and its alignment with actual outcomes	47
4.10	Sum of absolute deviations for the 10%, 50%, and 90% quantiles, comparing their deviation from actual observed values	47

Chapter 1

Introduction

Nowadays, there is a concerted effort to reduce carbon emissions as part of a global strategy to meet sustainability goals and mitigate the environmental impacts arising from the climate crisis. The challenge of lowering carbon footprints spans several economic sectors worldwide, with an increasing focus on developing innovative methods to precisely forecast carbon intensity. The drive towards sustainable development is expressed by initiatives and studies aimed at understanding and controlling emissions at multiple levels, including local, regional, and global scales.

The urgency to address these environmental emphasize the need for nations to adopt cleaner energy practices and improve the precision of environmental impact assessments. As part of this universal endeavour, many innovative technologies and methodologies are being explored in order to improve the predictability and manageability of energy production and consumption impacts.

In response to these needs of the energy sector, the application of machine learning (ML) techniques in environmental science represents a promising boundary. These techniques offer noteworthy advancements in processing and analysing complex datasets, enabling more accurate predictions and deeper understandings into the dynamics influencing carbon emissions. This dissertation explores the integration of such knowledges, focusing on how they can be leveraged to polish carbon intensity forecasting, consequently contributing to more the overall global endeavour.

Sustainable Development Goals

The Sustainable Development Goals (SDG) set by the United Nations assist as a blueprint for reaching a more sustainable future for all. This dissertation, specifically aligns with SDG 7, SDG 11, and SDG 13, crucial in guiding global efforts toward cleaner energy solutions, sustainable urban development, and climate action, respectively. SDG 7 encourages universal access to affordable, reliable, sustainable, and modern energy. SDG 11 promotes for making cities and human settlements inclusive, safe, resilient, and sustainable. SDG 13 underlines urgent actions to combat climate change and its impacts.

Considering and leveraging these goals offers a valuable framework for this research. By

integrating advanced forecasting techniques to enhance the predictability of carbon emissions, this study aims to contribute and addressing these objectives. It advocates for energy efficiency, the adoption of renewable resources, and the development of sustainable communities.

1.1 Motivation

As societies worldwide wrestle with the escalating impacts of climate change, the imperative to reduce carbon emissions has become a central concern across several sectors, particularly the energy sector. Accurately forecasting carbon intensity—the measure of carbon dioxide emissions per unit of electricity generated—is critical for active environmental management and policy-making. This forecasting not only informs about strategies to decrease the carbon footprint of energy production but also supports operational and financial planning in the energy sector.

The significance of this research is emphasised by the transition towards more sustainable energy systems, where precise predictions of carbon intensity are vital for incorporating renewable energy sources effectively and minimizing environmental impact. In that respect, ML offers advanced capabilities in handling the complex dynamics of carbon emissions data, which traditional statistical methods might not capture proficiently.

1.2 Objectives

Carbon intensity is a measure of the amount of carbon dioxide emissions produced per unit of electricity generated, typically measured in grams of CO₂ per kilowatt-hour. This metric is crucial for understanding the environmental impact of electricity production and for developing strategies and policies to reduce emissions.

The primary goal of this research is to develop a robust methodology for forecasting carbon intensity in the Portuguese electric power system for the day ahead, leveraging front-line ML techniques. Distinct to conventional methods that generate separate forecasts for different generation technologies, this approach aims to produce a comprehensive forecast for system-wide carbon intensity. This move necessitates a detailed examination of electricity market variables, energy consumption tendencies, and renewable energy predictions through advanced feature engineering (FE). This process comprises transforming existing data features and creating new ones to capture essential dynamics and interactions more effectively.

The challenge lies in designing, implementing, and comparing different ML configurations, combined with feature engineering, in order to improve the predictive performance and accuracy of the models. This is driven by the objective of refining the forecasting capabilities, thereby yielding a more informed and accurate tool for forecasting carbon intensity. This improved forecasting ability aims, at the end, to provide insights into the management and planning of energy resources, reflecting a try to step forward towards operational efficiency and sustainable practices.

1.3 Contributions

This work advances the field of carbon intensity forecasting in Portugal by leveraging feature engineering, with a clear emphasis on interpretability and improving the models' abilities to handle complex data dynamics. Enthused by the paper from Zhang, Zhou, and Zhang (2021), the integration of Evolutionary Forests (EF) is central in this study, which significantly expands the models' ability to discern and utilize non-linear relationships within the dataset. This method is able to capture essential temporal patterns through sophisticated engineering strategies such as time-based aggregations, lagged variables, and mathematical transformations expecting to encapsulate cyclical behaviour inherent in energy data.

The research observes twelve diverse models, including ensemble methods, linear models, neural networks, and support vector machine (SVM), enabling a benchmark analysis to discern the most effective techniques and respective impact on the models. The findings shown in the conclusions chapter reveal that the refined feature engineering processes are contributory in boosting the predictive performance across the models, particularly in respect to the incorporation of advanced feature engineering techniques. Moreover, the use of the Autorank tool enriches this comparative analysis by providing a statistical validation on the performance differences among the trained model.

The best-performing model was enhanced by applying quantile loss function analysis and ensemble techniques across different time periods. This integrated approach, which combines probabilistic forecasting with temporal multiplicity, suggestively improves the model's accuracy. It expertly balances short-term fluctuations and long-term trends, bringing possibilities to improve decision-making under uncertainty.

1.4 Thesis Structure

The structure of this thesis, as shown in the Figure 1.1, is designed to methodically explore and address the forecasting of carbon intensity within the Portuguese electric power system using advanced ML techniques. The thesis is organized into several chapters, each focusing on a distinct aspect of the research. This structure ensures a logical progression from the introduction of the problem and its research context to the detailed methodology, analysis, and conclusions.

The introduction of the thesis sets the stage by highlighting the importance of forecasting carbon intensity within the context of global sustainability initiatives and the specific demands of the energy sector. It enunciates the motivations driving the research, framing the discussion within the broader efforts to improve environmental conservation and operational efficiency through improved predictive models. Following the introduction, the State-of-the-Art chapter analyses existing methodologies and advances in carbon intensity forecasting, recognising gaps and opportunities for further research, in that way establishing a foundation for the study. The methodology chapter exposes the forecasting techniques used, highlighting the combination of advanced ML methods and feature engineering through Evolutionary Forest to refine forecasting accuracy. A practical application of these methods is verified in the

Chapter 1 Introduction					
	Motivation	Objectives	Contributions	Thesis Structure	
Chapter 2 State-of-the-art			Chapter 3 Forecasting Methodology		
Advancements in Carbon Intensity Forecasting	Main Methodologies	Carbon Intensity Forecasting in Europe and Similar	Formulation of the Forecasting Problem	Feature Engineering (Evol. Forest)	ML Models and Hyperparameter Tuning
Applications	Data Sources and Challenges	Gaps and Opportunities	Evaluation Measures	Quantile Loss Function	Ensemble Models for Different Time Frames
Chapter 4 Case-study			Chapter 5 Conclusions		
Exploratory Data Analysis	Feature Engineering	Models' Evaluations with Autorank Tool	Key Insights and Implications Limitations and Future Work		
	Best Performing Model (Quantile Loss Function and Ensemble Models for Different Time Periods)				

Figure 1.1: Thesis structure and its sections

Case Study chapter, which focuses on the carbon intensity of the Portuguese electric power landscape, assessing several ML models' effectiveness. The dissertation concludes with a chapter that recapitulates the key insights, discusses the wider implications of the findings, and suggests future research directions, reflecting on the study's contributions to the field and possible policy impacts.

Chapter 2

State-of-the-art

The evolution of carbon intensity forecasting has witnessed a notable shift towards advanced methodologies, especially ML, for enhanced predictive accuracy. Initial studies focused on understanding the relationship between energy consumption and carbon emissions, employing basic statistical models like autoregressive methods based on historical data. Recent advancements in data processing capabilities have enabled the adoption of more sophisticated techniques.

2.1 Advancements in Carbon Intensity Forecasting

Influential studies have played a crucial role in this evolution. Leerbeck et al. (2020) introduced a comprehensive ML algorithm that forecasts carbon intensity while distinguishing between average and marginal emissions, essential for applications like load scheduling. The *Day-ahead carbon intensity forecasting of power grids using machine learning* (DACF) system by Maji, Sitaraman, and Shenoy (2022) emphasized the significance of day-ahead forecasts, optimizing energy generation and consumption by incorporating weather forecasts and individual source production.

The *CarbonCast* methodology proposed by Maji, Shenoy, and Sitaraman (2022), adopted a two-tier hierarchical approach, integrating Artificial Neural Networks (ANN), Convolutional Neural Network (CNN), and Long Short-Term Memory Network (LSTM) to enhance forecast accuracy over longer periods (up to 96 hours) for grid carbon intensity.

The literature on forecasting in electric power systems covers load forecasting, renewable energy forecasting, and carbon intensity forecasting. Studies like Weron (2006), which discusses "Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach," explore power grid dynamics, considering factors such as demand, production, and environmental impacts. ML techniques play a vital role in successful forecasting methods, addressing non-linear associations and dynamic relationships between features.

Despite these advancements, challenges persist, including the need for high-quality datasets, uncertainties in factors like weather conditions, and ensuring that models are flexible enough to accommodate continuous changes in the energy sector.

Chapter 2 State-of-the-art						
Sec. 2.1 Advancements in Carbon Intensity Forecasting			Sec. 2.2 Methodologies		Sec. 2.3 Forecasting in Europe and Similar Contexts	
Fundamentals of Carbon Intensity Forecast	Machine Learning	Day-ahead Forecast	Feature Selection (e.g. LASSO)	Autoregressive Models Limitations	Deep Learning (ANN, CNN, LSTM)	Regional Differences
			SoftMax Weighted Average	Time Series Decomposition (EEMD, IMFs)	DACF Model	Custom-made Models
						Incorporating Renewable Energy Sources
Sec. 2.4 Applications			Sec. 2.5 Data Sources and Challenges		Sec. 2.6 Gaps and Opportunities	
Demand Response	Electricity Maps	Web-based Dashboard	Limited/Noisy Datasets	Non-linear Relationships Within the Data	Thesis Dataset: 100% Publicly Data	Local Studies to Capture Region Dynamics
						Negative Impacts from ML Models
						Policies and Regulatory Frameworks
						Integration of Emerging Technologies
						Interpretability of ML Models
						Holistic View on Energy Consumption
Sec. 2.7 Inferences						

Figure 2.1: Structure of literature review chapter and its sections

Common metrics used to evaluate forecasting accuracy include RMSE, Mean Absolute Percentage Error (MAPE), and Forecast Bias, providing quantitative measures to evaluate the effectiveness of forecasting models.

As represented in the Figure 2.1 this chapter highlights the journey from traditional methods to sophisticated ML approaches, contributing to more precise and detailed carbon intensity forecasting in electric power systems.

2.2 Methodologies in Carbon Intensity Forecasting

Many methodologies are utilized in carbon intensity forecasting, oscillating from traditional statistical models like autoregressive models to more advanced ML algorithms. Linear regression is usually employed to model the relationship between factors such as power production, demand, import/export data, weather conditions, and carbon emissions, as demonstrated by Tuaimah and Abass (2014). Feature selection techniques such as the Least Absolute Shrinkage and Selection Operator (LASSO), as shown by Tibshirani (1996), are applied to improve linear regression models by selecting relevant features and minimizing overfitting. Furthermore, Autoregressive Integrated Moving Average (ARIMA) models, as detailed by ?, are helpful in correcting correlated residuals, in that way improving forecast accuracy. Another predominant approach, as outlined by Hastie (2009), involves combining multiple linear regression models through a weighted average to benefit from the strengths of each model and enhance overall prediction accuracy.

Traditional methods, such as ARIMA models, have limitations in seizing complex non-linear relationships present in carbon intensity data. These traditional methods may face difficulties with diverse sets of explanatory variables and often rely on historical patterns, potentially overlooking unexpected (sudden) changes or external factors. Linear regression models, while useful, assume a linear relationship between variables, which may not accurately represent the particulars of carbon intensity dynamics influenced by numerous factors. Z.-X. Wang and Ye (2017) explores traditional methods in China, by pointing out findings which reveal the superiority of the non-linear grey multivariable model over traditional grey and autoregressive integrated moving average model

An emerging tendency in carbon intensity forecasting is the incorporation of ML, particularly deep learning (DL) techniques like ANN (as proposed by Acheampong and Boateng (2019)), CNN, and LSTM (as proposed by Maji, Shenoy, and Sitaraman (2022)). These approaches outperform in capturing non-linear relationships, delivering more accurate predictions in dynamic and interconnected energy systems. In Qader, Khan, Kamal, Usman, and Haseeb (2021) where various forecasting methods are employed and evaluated, in Bahrain, the neural network model outperforms other methods. In fact ANN has been a very popular and have been applied in innumerable studies for several countries around the world as exemplified in the study carried by Acheampong and Boateng (2019), that employs ANN to predict carbon emissions intensity in high-emitting countries (Australia, Brazil, China, India, USA). Innovations, such as hierarchical forecasting models like CarbonCast (Maji, Shenoy, and Sitaraman (2022)) showcase the combination of diverse ML architectures to improve accuracy over longer time periods. ML models offer flexibility to adapt to moving conditions and include a diverse range of variables, contributing to more robust and accurate carbon intensity forecasts.

Being more specific in respect to the main methodologies, this section will now present in more detail three articles as being the groundbreaking studies in the field of carbon intensity forecasting. I will use them as the state-of-the-art and reference for my own work.

2.2.1 Innovative Machine Learning Approaches for Short-term Forecasting of Carbon Intensity in Power Grids

An important innovation in the study by Leerbeck et al. (2020) on short-term forecasting of CO₂ emission intensity in power grids through ML is its ability to distinguish between average and marginal emissions, indispensable for applications like load scheduling.

The research evaluates 473 explanatory variables such as power production, demand, imports, and weather conditions, which assist as inputs for the Autoregressive Integrated Moving Average with Exogenous Inputs (ARIMAX) model. To increase efficiency and avoid overfitting, techniques like LASSO and Forward Feature Selection reduce these variables to less than 30. The model's predictive accuracy is further increased by employing the Soft-Max Weighted Average to integrate outputs from three linear regression models and applying cross-validation to ensure model consistency.

Short-term CO₂ emissions forecasting based on decomposition approaches and its impact on electricity market scheduling, by Bokde, Tranberg, and Andresen (2021), improves short-term CO₂ emissions forecasting through time series decomposition. It compares a four-step decomposition process—identifying trends, extracting seasonal patterns, and isolating random noise—with Ensemble Empirical Mode Decomposition (EEMD). The simpler, four-step method generally outperforms, enabling better forecasting by modeling individual components, each optimized with specific techniques such as Feedforward Neural Networks (FFNN) and ARIMA.

Lastly, in the study by Maji, Sitaraman, and Shenoy (2022) on day-ahead carbon intensity forecasting using ML (DACF), the attention shifts from CO₂ intensity to energy produc-

tion time series, utilizing both historical data and future insights such as weather forecasts to enhance forecast accuracy. The DACF model incorporates DL to refine predictions, successfully addressing both long-term and short-term patterns in energy production.

2.3 Carbon Intensity Forecasting in Europe and Similar Contexts

It is important to understand the geographical context of those methodologies, and carefully analyse if they are applicable to Portuguese context. In exploring articles and papers, it becomes obvious that different European regions or countries may show unique characteristics in terms of energy production, consumption patterns, and environmental policies. Bhattacharya, Inekwe, and Sadorsky (2020) examines carbon emissions intensity convergence across 70 countries, differentiating between consumption-based and territory-based metrics. This research addresses gaps in understanding convergence determinants and forecasts future emission patterns. Therefore, forecasting models need to be custom-made to these regional graduations to ensure accuracy, which means that the transferability of forecasting models to Portugal is not direct and should be carefully evaluated.

By looking at the studies that concentrate on carbon intensity forecasting in the European energy landscape reveals common characteristics across different regions. It is possible to observe the interconnected energy grids, the diversity on energy sources, and it is possible to admit that there is a general and genuine concern of various countries on environmental subject and how to mitigate climate change impacts in the future. For example, study from Leerbeck et al. (2020) focus on Danish bidding zones (DK2) offer insights into challenges and opportunities specific to each region. DK2 refer to specific geographic areas in Denmark that are elected for the trading of electricity on the energy market. These zones are part of the larger European electricity market and are related with specific challenges and opportunities related to energy production, consumption, and trading.

The interconnected nature of European energy grids is a topic of crucial consideration. Countries are often part of interconnected grids, requiring forecasting models that account for cross-border exchanges and dependencies. One important key message is that the models must be drawn to accommodate the dynamic nature of energy flows between different regions and countries.

Europe's leadership in incorporating renewable energy sources presents challenges related to the irregular nature of renewables sources and their impact on carbon intensity. Another important key message is that forecasting models must be robust enough to address these challenges and deliver good predictions.

In conclusion, concerning to the transferability of forecasting models between regions, it is vital to evaluate the adaptableness of a model to the unique characteristics of each region or country, because factors like as energy mix, grid infrastructure, and environmental policies play a significant role in determining the model's success across the diverse European

frameworks.

2.4 Applications of Carbon Emissions Forecasts

In the context of real-world practical applications, existing solutions are proving to be essential in the reduction of carbon emissions, offering valuable insights for informed decision-making. In that regards, Lowry (2018) discusses the importance use of demand response (DR) in building HVAC systems to reduce carbon emissions in the context of the United Kingdom electricity grid. To promote the informed decision-making by the economic agents, certain solutions are already in the market today, playing a crucial role and position themselves to become integral components in our efforts to mitigate environmental impact.

One noteworthy example is the utilization of carbon intensity forecasts by tech giant Google. Google uses Electricity Maps (Figure 2.2) forecasts to schedule their Artificial Intelligence (AI) models, ensuring that computational tasks are executed when renewable energy sources are abundant—essentially, when the wind blows or the sun shines. This approach aligns with the broader objective of optimizing energy consumption and reducing reliance on non-renewable sources.

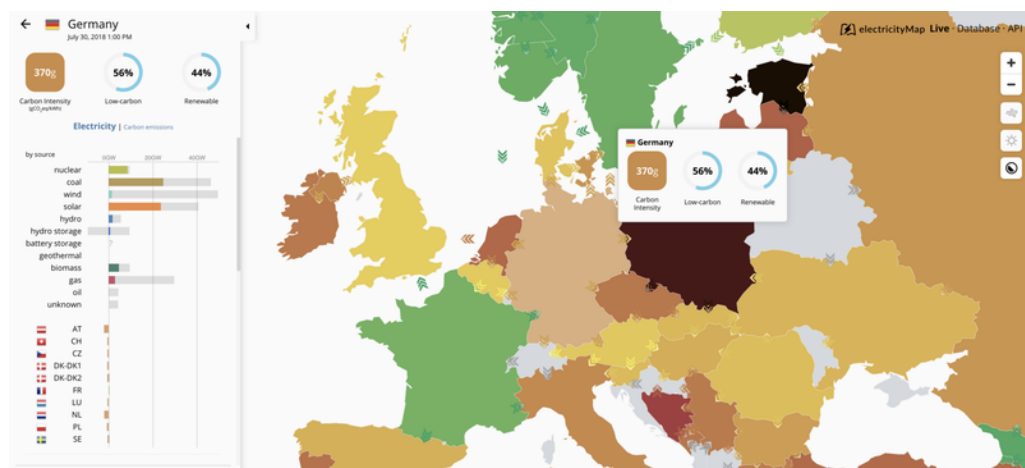


Figure 2.2: The Electricity Maps calculates and visually presents the source of the electricity, along with its corresponding global and real-time carbon footprint

The common use case involves a web-based dashboard designed for the analysis of CO₂ emissions. Ashok (2020) emphasizes the significant role of such web-based dashboards as invaluable tools for monitoring electricity exchange patterns among countries and regions, alongside tracking CO₂ emissions.

While Electricity Maps stands out as a prominent application in this domain, it's worth remarking that the service is not public and comes with a substantial cost. Despite its popularity, accessibility remains a challenge (specially for researchers), motivating the exploration of alternative solutions that can offer similar benefits on a costly manner.

The significance of these applications extends beyond individual businesses, influencing broader strategies for sustainability and carbon footprint reduction. By incorporating car-

bon intensity forecasts into operational planning, companies can make data-driven decisions that contribute to a more sustainable and environmentally conscious approach to energy consumption.

2.5 Data Sources and Challenges

Carbon intensity forecasting depend on on diverse data sources to capture the complex dynamics of energy systems. Frequently used data include historical records of energy consumption (time series), carbon emissions, weather conditions, and production patterns. These sources provide essential information for building models that can predict carbon intensity with required accuracy.

Nevertheless, there are challenges and limitations that are intrinsic in dealing with such data. One main challenge (perhaps the major one) is the availability of complete, consistent and high-quality datasets. In some cases, historical data may be limited or not sufficiently detailed or even incomplete, making it challenging to capture the complete range of factors influencing carbon intensity. Furthermore, the consistency of data across different sources can be an issue, leading to potential imprecisions in forecasting models.

The community of researchers have actively addressed these challenges to improve the reliability of carbon intensity forecasting. The growing integration of ML techniques has allowed for better handling of complex, non-linear relationships within the data. Additionally, advancements in data processing capabilities have also contributed and enabled researchers to extract expressive insights even from limited or noisy datasets. Besides that, in order to mitigate the challenges related with data availability, researchers have also explored innovative solutions. For instance, research such as that conducted by Duras, Javed, Månsson, Sjölander, and Söderberg (2023) has effectively utilized feature selection techniques like LASSO to identify the most relevant variables, thus minimizing the impacts of data limitations. The use of ARIMA models for residual correction in linear regression models is another strategy commonly applied to increase forecast accuracy, as explored by Y. Wang, Wang, Zhao, and Dong (2012).

Transitioning to the specific case of Portugal (and the specific case of my thesis), it's noted that data availability issues persist, requiring proactive measures. Despite challenges, exclusively relying on 100% publicly available data extracted from REN (*Redes Energéticas Nacionais*), REE (*Red Eléctrica de España*), and ENTSO-E (European Network of Transmission System Operators for Electricity). The dataset spans one year (from 2022-09-02 00:00 to 2023-09-27 23:00), comprising 9,384 entries (one entry per hour) and an extensive set of features, totalling 114. These features include the target variable (CO₂ intensity), and other exogenous variables like electricity price (as greatly explored by García-Martos, Rodríguez, and Sánchez (2013)), solar energy production, wind energy production, and others.

In summary, while data challenges are inherent in carbon intensity forecasting, researchers have demonstrated a proactive approach to addressing these questions. By leveraging advanced practices and exploring groundbreaking techniques.

2.6 Gaps and Opportunities

We can identify two types of opportunities. The first type is directly related to the field under study: carbon intensity forecasting models. The second type is more general and not specifically tied to the field of study but pertains to the broader evolution of ML, being these opportunities applicable to other fields of study besides carbon intensity forecasting.

2.6.1 Opportunities directly related to the field under study

The existing literature on carbon intensity forecasting in energy systems has made significant steps, but still some gaps continue to exist. One noteworthy gap is the need for more comprehensive studies concentrating on region-specific dynamics, particularly within the context of the Portuguese electric power system. While wider European studies provide valuable insights, a more local approach that reflects the unique characteristics of the Portuguese energy scenery would contribute to more precise and actionable forecasts.

It is essential to acknowledge the paradoxical nature of ML and AI models designed to reduce CO₂ emissions. Delanoë, Tchuente, and Colin (2023) introduced a method to quantify both positive impacts (emission reduction during model use) and negative impacts (emissions from model training and use). This methodology holds potential for specific adoption within this field of study.

Exploring the potential impact of policy changes and regulatory frameworks on carbon intensity forecasting is another area where more research could provide valuable insights, especially considering the lively nature of environmental policies, general political agendas and geo-political contexts.

2.6.2 Opportunities related with ML evolution

Additional research opportunities lie in exploring the integration of emerging technologies, such as blockchain and Internet of Things (IoT), to improve data accuracy and transparency. Opportunities for forward-moving the field might also be extended to the improving the interpretability of ML models. Developing models that not only deliver accurate predictions but also offer understandings into the factors influencing carbon intensity could improve the decision-making process for energy stakeholders.

Furthermore, there is a chance to create forecasting frameworks that explicitly consider the socio-economic aspects of energy consumption, contributing to a more holistic understanding of carbon intensity subtleties.

2.6.3 Portuguese context

Particular attention to the challenges and opportunities within the national energy infrastructure is also important. This includes addressing the unique mix of energy sources, the grid structure, and respective policy framework. By tailoring research efforts to these particulars, scholars may directly contribute to the development of more accurate and applicable forecasting models for the Portuguese electric power system.

2.7 Concluding Remarks

In a nutshell, the reviewed literature emphasises the evolution of carbon intensity forecasting in energy systems. The historical move from simplistic models to advanced ML methodologies reproduces the growing emphasis on predictive accuracy. Important milestones include the introduction of ML algorithms, day-ahead forecasting systems, and hierarchical forecasting models.

Influential works by researchers like K Leerbeck, D Maji, P Shenoy, and RK Sitaraman have significantly influenced the field, introducing innovative approaches and frameworks. Traditional methods, while still pertinent, face limitations in seizing complex non-linear relationships, leading to the growing adoption of ML techniques. Numerous methodologies, such as linear regression (ARIMA), LASSO, and weighted average models, validate diverse approaches to carbon intensity forecasting. ML, particularly DL techniques, emerges as a trend, presenting improved accuracy and robustness to the forecasting models.

The literature underlines challenges like the need for comprehensive datasets and the adaptability of models to the evolving energy landscapes. Exploring European contexts discloses the importance of considering regional specifics, interconnected grids, and challenges in renewable energy integration. Transferability of models between regions is not that straightforward, depending on adaptability to unique characteristics. Looking at the existing case studies and applications, real-world implementations offer valuable lessons. Assessing the effectiveness of forecasting methodologies across various scenarios showcases the practical implications of research. Common metrics used by researchers for evaluating forecasting accuracy include RMSE, MAPE, and Forecast Bias.

In respect to the data sources, the literature highlights the implications of not having accurate and consistent data. Very often, there are challenges in that respect, including data availability, quality, and consistency, with researchers employing strategies to address these limitations.

The literature also suggests a need for more localized studies, considering the identified gaps, particularly within the Portuguese electric power system. Opportunities for future research involve integrating emerging technologies, considering policy influences, and being able to enhance model interpretability.

In light of these conclusions, this review sets the stage for further research aiming to contribute directly to the Portuguese electric power system's forecasting accurateness. By addressing recognised gaps and leveraging emergent opportunities, the forthcoming research should aim to improve the precision and applicability of carbon intensity forecasting models.

Chapter 3

Forecasting Methodology

This chapter delves into the development of a forecasting methodology that attaches the capabilities of ML to predict carbon intensity in Portugal’s electric power sector. Through a thorough exploration of methodological innovations, such as the application of Evolutionary Forest in feature engineering and the employment of varied ML models, this study purposes to elevate the accuracy of carbon intensity forecasts.

3.1 Forecasting Problem Formulation

The focus of this dissertation is on day-ahead forecasting to predict the overall system carbon intensity of the Portuguese electric power system. This approach marks a departure from traditional methodologies that typically forecast carbon intensity for individual generation technologies. By using a comprehensive set of variables from the electricity market, as well as data on energy consumption and renewable energy forecasts, this research aims to improve the accuracy and relevance of its predictions. Each day, by noon, the task is to accurately predict the carbon intensity for the subsequent 24 hours. Addressing this requires to tackle a complex time series problem and strategically applying feature engineering techniques to improve forecasting accuracy.

3.1.1 Factors Influencing Carbon Intensity

The main challenge in this forecasting problem is to capture the complex interactions between various factors that influence carbon intensity. These include:

- **Electricity Market Variables:** Prices, demand, and generation schedules.
- **Energy Consumption Patterns:** Hourly, daily, and seasonal variations in electricity consumption.
- **Renewable Energy Forecasts:** Predictions of wind and solar power generation, which are inherently variable and uncertain.

The following Figure 3.1 is taken from Maji, Shenoy, and Sitaraman (2022), which is representative of the factors influencing carbon intensity in California (similar latitude as Portugal,

with both known as sunny destinations in the northern hemisphere of the globe). It exemplifies how carbon intensity varies throughout the day, showing lower values during the day when solar production is high.

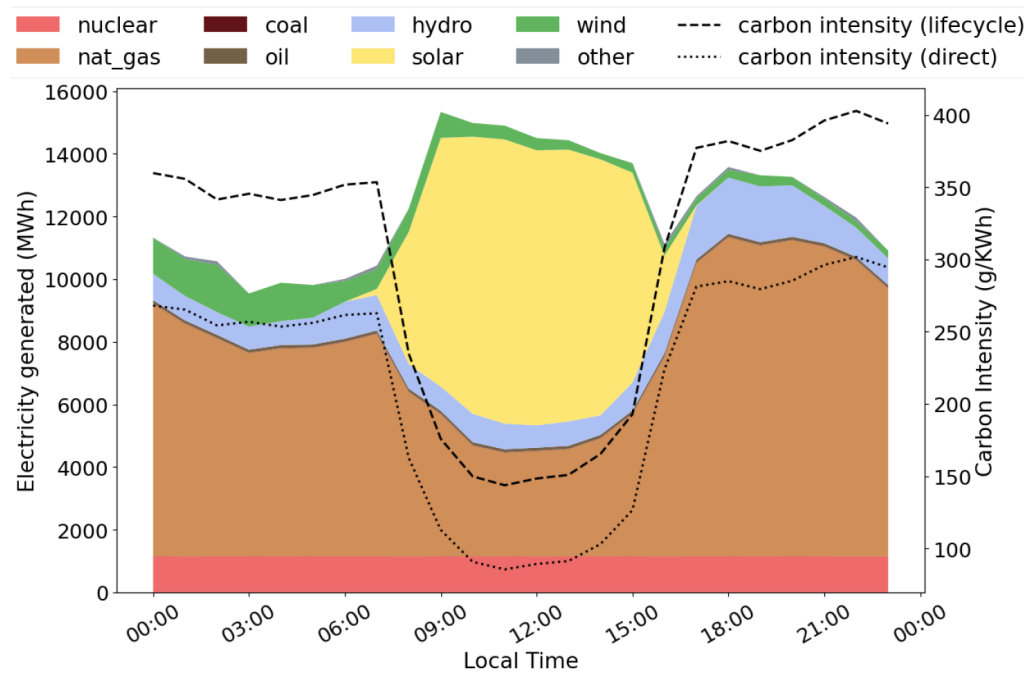


Figure 3.1: Average Carbon Intensity in California. Source: Maji, Shenoy, and Sitaraman (2022)

3.1.2 Methodological Approach to Forecasting Carbon Intensity

The subsequent sections outline the schematic representation from Figure 3.2 and provide a short explanation for each step. Further theoretical foundations will be elaborated upon in subsequent sections, particularly focusing on the highlighted blue boxes, which are the fundamental elements carrying significant value to the case study presented in chapter 5.

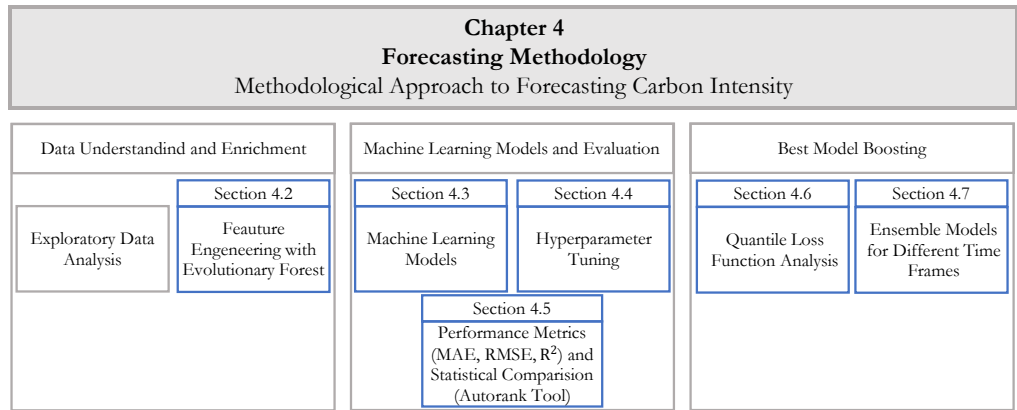


Figure 3.2: Implemented methodological approach scheme

The Figure 3.2 represents the approach taken, beginning with extensive Exploratory Data Analysis (EDA) to identify trends and patterns in the data. Feature engineering is executed using the EF method to generate complex features, which are then appended to the original dataset. Firstly, various ML models are evaluated in their basic forms. Hyperparameter tuning is conducted using Optuna to optimize the models by minimizing MAE. Optuna is an open source hyperparameter optimization framework to automate hyperparameter search created by (Akiba, Sano, Yanase, Ohta, and Koyama (2019)). The performance of these optimized models is compared using metrics such as MAE, RMSE, and R-squared (R^2), supported by Autorank analysis and statistical tests. The best-performing model is further analyzed through probabilistic forecasting using quantile loss functions and ensemble models for different time frames to measure stability and consistency.

Data Understanding and Enrichment

Exploratory Data Analysis: The initial phase of the methodology involved extensive EDA to understand the underlying patterns in the data. This included a special focus on trend, seasonal, and cyclic analysis to capture the temporal dynamics of the dataset. EDA helps in identifying significant patterns, which are crucial for building robust predictive models.

Feature Engineering with Evolutionary Forest: Feature engineering was conducted using the EF method, which involves genetic algorithms with decision trees to construct features capable of capturing more complex interactions within the data. Hyper-parameters were tuned so that the evolutionary process produces the most informative features.

Appending New Features: The new features generated by the EF were appended to the original dataset. An optimal number of features were selected based on feature importance scores and model performance, balancing the trade-off between model complexity and predictive accuracy.

ML Models and Evaluation

Hyper-parameter Tuning: Hyper-parameter tuning was conducted using Optuna, an open-source framework for automated hyper-parameter optimization. Optuna employs efficient search algorithms to find the optimal set of hyper-parameters that minimize the MAE on the validation set, thereby enhancing model performance.

Evaluation of Models: Initially, a wide range of ML models were evaluated in their basic forms without hyper-parameter tuning. This step provided a baseline performance for each model and helped in understanding their inherent capabilities and limitations. The models considered included ensemble methods, linear models, neural networks, and SVM.

Model Performance Comparison: The performance of the optimized models was compared using several metrics such as MAE, RMSE, and R^2 . The Autorank tool was employed for statistical comparison, providing detailed insights into the ranks, confidence intervals, and significance of differences between models. This comparison was reinforced by significance testing through repeated measures ANOVA and post-hoc Tukey HSD tests.

Best Model Boosting

Quantile Loss Function Analysis: Probabilistic forecasting was performed using quantile loss functions to provide a more complete view of the predictive distribution, especially under uncertainty. This approach involves predicting the 0.1, 0.5, and 0.9 quantiles to create prediction intervals that incorporate the central 80% of the potential outcomes, offering a clear view of the uncertainty in the forecasts.

Ensemble Models for Different Time Frames: The best-performing model was further analysed to understand its strong points and limitations. This analysis included the ensemble of the model for different time dimensions (6 months, 3 months, and 1 month), to assess their stability and consistency over varying time frames of the dataset.

3.2 Feature engineering with Evolutionary Forest

Feature engineering is a crucial step in the ML pipeline, involving the transformation of raw data into informative inputs that significantly enhance a model's predictive performance, as discussed in Kuhn and Johnson (2019). One important aspect of feature engineering is featurization, where new features are derived to better capture underlying patterns in the data. The EF technique exemplifies this approach by applying principles of genetic programming (GP). It begins with an initial population of decision trees and evaluates their performance using a fitness function. The following text intends to provide the fundamentals of GP and EF, and understand how they correlate.

GP is a kind of evolutionary algorithm that mimics the process of natural selection to find optimal solutions to complex problems. It is an iterative process where, at each stage, the algorithm selects the fittest individuals (offspring) to reproduce in the next generation, directed by a fitness function. This fitness function evaluates how well each solution meets the desired objectives. Comparable to biological evolution, GP algorithms can include random mutations in offspring, but only those with the highest fitness are selected for reproduction, leading to progressive improvement over generations. The algorithm typically ends once it reaches a predefined fitness measure. That is deeply described in the article Koza (1994).

The concept of GP was initially proposed by Alan Turing in 1950, but it wasn't until the 1980s that practical implementations emerged. John Koza patented the first genetic algorithm in 1988 and has significantly contributed to the field since then. In GP, the process begins with the initialization of a population of decision trees. Each tree is evaluated using a fitness

function, $f(x)$, which is often defined as

$$f(x) = \sum_{i=1}^n w_i \cdot x_i$$

where w_i are weights, and x_i are features of the solution. Based on their fitness scores, the top-performing trees are selected for reproduction. The crossover operation combines parts of two parent solutions, x_1 and x_2 , to create a new solution, x' , represented mathematically as

$$x' = \text{crossover}(x_1, x_2)$$

Following crossover, mutation introduces random changes to the new solution, x' , to produce x'' , expressed as

$$x'' = \text{mutate}(x')$$

This process forms a new generation of decision trees with improved fitness. The selection, crossover, and mutation steps are repeated iteratively, substituting the old population with the new one. The iterations last until a predefined fitness measure is reached or a set number of generations is completed. By following these steps, GP iteratively improves the solutions to reach an optimal or near-optimal result.

3.2.1 Advanced Feature Engineering with Evolutionary Forest

There has been extensive research utilizing GP techniques for feature construction as shown in Smith and Bull (2005). The EF method combines GP and decision trees to create a strong and effective ensemble learning framework for regression problems.

The algorithm framework of EF includes several main steps (pseudocode steps represented in the Table 3.1. As explained in Zhang et al. (2021) the process begins with the random initialization of a population of multitree GP individuals. Each GP individual contains multiple GP trees, each representing a potential feature. New GP individuals are then created through genetic operations such as crossover and mutation applied to parent GP individuals.

The fitness of each GP individual is evaluated using a fivefold cross-validation scheme on the training data, with the fitness function typically based on the Mean Squared Error (MSE). A lexicase selection operator is used to select the fittest GP individuals based on their performance across different data subsets. The best-performing GP individuals are stored in an archive, which forms the final EF model at the end of the evolutionary process. This iterative process lasts until a predefined number of generations or a acceptable fitness level is reached.

In EF, each GP individual is represented as a set of GP trees, where each tree corresponds to a feature. This multitree approach allows EF to construct multiple features simultaneously. The genetic operations include subtree mutation, which randomly replaces a subtree with a newly generated tree, and subtree crossover, which exchanges subtrees between two parent trees to create offspring.

Fitness is estimated using a fivefold cross-validation scheme. The evaluation involves training a DT on the features generated by each GP individual and measuring the predictive performance using the absolute deviation between the predicted and actual values. The lexibase selection operator maintains diversity in the population by selecting individuals that perform well across different subsets of the data, using multiple fitness values to ensure diverse predictive behaviour.

An archive is then used to store the best GP individuals. New GP individuals are added to the archive if they perform better than the worst individual currently in the archive. The time complexity of EF depends on the number of decision trees constructed during the fitness evaluation process. Even though EF has a higher time complexity than traditional boosting methods, it has a lower time complexity compared to previous wrapped-based feature construction methods.

Step	Description
1	Randomly initialize a population of s GP individuals $P = \{\phi_1, \phi_2, \dots, \phi_s\}$.
2	Initialize the archive A to be empty.
3	Set the generation counter gen to 0.
4	While $gen \leq max_gen$ do
5	For each individual i in the population
6	If $i \% 2 = 1$ and a random number is less than the crossover probability P_{cr} then
7	Perform crossover between ϕ_i and ϕ_{i+1} .
8	If a random number is less than the mutation probability P_{mu} then
9	Perform mutation on ϕ_i .
10	Evaluate the fitness of each GP individual using cross-validation.
11	Select the best individuals based on their fitness.
12	Update the archive with the selected individuals.
13	Replace the old population with the selected individuals.
14	Increment the generation counter.
15	Return the archive A containing the best GP individuals.

Table 3.1: Pseudocode of Evolutionary Forest algorithm

The EF method leverages GP to evolve a population of decision trees, creating high-quality, non-linear features that enhance the performance of regression models. This approach, through its iterative process of selection, crossover, and mutation, ensures the generation of robust and diverse features, making it a powerful tool for regression problems. This detailed explanation aligns with the descriptions and illustrations provided in the document, effectively combining the strengths of GP and decision trees to offer a sophisticated and efficient solution for feature construction and regression modelling.

3.3 ML Models

Employing a diverse range of ML models offers several advantages. It permits us to leverage the unique strengths of each model, ensuring that various aspects of the data are captured effectively. This comprehensive approach improves the overall predictive performance and robustness of our forecasts. By comparing the performance of different models, we can identify the most suitable model for our specific task and gain deeper insights into the dataset's characteristics. This diversity also facilitates the application of ensemble methods, such as blending or stacking, which might further enhance prediction accuracy. This section provides a theoretical framework and rationale for using these models, emphasizing their strengths and contributions to the forecasting problem.

3.3.1 Random Forest and Extra Trees

Random forest (RF) and extra trees (ET) are ensemble learning methods that utilize multiple decision trees to improve predictive performance and robustness. As explained by Breiman (2001) these models are particularly effective in capturing non-linear interactions between features, making them suitable for complex datasets like the one we are using in our case-study. By averaging the predictions of multiple trees, RF and ET reduce overfitting and improve generalization. Their ability to handle large datasets with high dimensionality and to provide feature importance metrics makes them valuable tools for feature selection and interpretation.

3.3.2 AdaBoost and Gradient Boosting

Adaptive Boosting (AdaBoost) and gradient boosting decision trees (GBDT) are boosting algorithms explored by Freund and Schapire (1997) that combine weak learners to form a strong predictive model. AdaBoost focuses on improving the performance of weak classifiers by adjusting their weights based on their errors, while GBDT builds models sequentially to minimize a specified loss function. These models are effective in handling various types of data and capturing intricate patterns, making them ideal for improving prediction accuracy in our carbon intensity forecasting task. Their flexibility in optimizing different loss functions, including quantile loss for probabilistic forecasting, further enhances their applicability.

3.3.3 XGBoost, LightGBM, and CatBoost

Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Category Boosting (CatBoost) are advanced gradient boosting frameworks designed for efficiency and scalability. XGBoost implements optimized gradient boosting techniques with regularization to prevent overfitting. LightGBM leverages histogram-based algorithms to speed up training and handle large datasets efficiently. CatBoost excels in handling categorical features (as demonstrated by Dorogush, Ershov, and Gulin (2018)) and reducing the impact of data order, making it highly suitable for our dataset, which includes categorical variables. These models are known for their high performance in predictive tasks, providing

robust and accurate forecasts.

3.3.4 Linear Regression and LASSO

Linear Regression and LASSO are fundamental linear models used for regression tasks. Linear Regression as explained by Seber and Lee (2012) delivers a simple yet effective approach to modelling relationships between features and the target variable. LASSO extends this by performing feature selection and regularization, reducing the risk of overfitting and improving model interpretability. These models are useful for understanding linear relationships in the data and for providing a baseline comparison against more complex models.

3.3.5 Artificial Neural Networks (ANN)

ANN are powerful models capable of capturing complex non-linear relationships in data. Goodfellow, Bengio, and Courville (2016) provides an exhaustive overview of ANN and DL, including the theoretical foundations and practical applications of ANN. ANN are particularly suitable for time-series forecasting tasks due to their ability to model intricate patterns and dependencies. By utilizing multiple hidden layers and neurons, ANN can learn from historical data to make accurate predictions.

3.3.6 Support Vector Machine (SVM)

Cortes and Vapnik (1995) introduces SVM and their use in classification and regression tasks. SVM is a solid ML algorithm used for classification and regression tasks. SVM constructs a hyperplane that maximizes the margin between data points of different classes, making it effective in high-dimensional spaces. For regression, SVM aims to minimize the prediction error within a pre-defined threshold, making it apt for our forecasting problem where precision is crucial. Its ability to handle non-linear relationships through kernel functions improves its predictive power.

3.3.7 Dropouts meet Multiple Additive Regression Trees (DART)

Dropouts meet Multiple Additive Regression Trees (DART) is an extension of gradient boosting that incorporates dropout techniques from neural networks to improve model generalization and robustness as detailed by Vinayak and Gilad-Bachrach (2015). By randomly dropping trees during the training process, DART reduces overfitting and increases the model's ability to generalize to unseen data. This approach is particularly helpful for our forecasting task, where the model needs to adapt to varying patterns in energy consumption and production.

3.4 Hyperparameter Tuning

Hyperparameter tuning is a key step in optimizing ML models to accomplish the best performance. In this study, Optuna, an open-source hyperparameter optimization framework, was considered to automate and enhance the process of hyperparameter tuning for the *Evolutionary Forest Regressor*. Optuna efficiently searches for the optimal set of hyper-parameters by exploring the hyperparameter space and identifying the combinations that minimize the MAE on the validation dataset.

The process was applied over all ML models, leveraging the new features generated by the EF technique to find their best configurations. The process involves the following phases:

1. Initialization: Optuna initializes the search by randomly selecting a set of hyperparameters within the defined ranges.
2. Evaluation: The selected hyperparameters are used to train the respective models, and their performance is evaluated using the MAE metric.
3. Optimization: Optuna employs sophisticated algorithms such as Tree-structured Parzen Estimator (TPE), as explored by Watanabe (2023), to iteratively refine the search space, concentrating on the most promising hyperparameter combinations.
4. Selection: The hyperparameter combination that yields the lowest MAE is selected as the optimal set.

It is decisive to ensure that ML models are always executed with tuned parameters to achieve optimal performance. In the context of our case study, hyperparameter tuning must be conducted both before and after applying the EF procedure, as the feature transformations may disturb the models' parameter space and optimal configurations. In addition, determining the optimal number of features is crucial to balancing model complexity and predictive performance. Therefore, the tuning process should interactively consider different feature set sizes to identify the best threshold that maximizes accuracy while avoiding overfitting. Incorporating this approach certifies that the hyperparameter optimization process fully aligns with the evolved feature landscape, enabling the most efficient utilization of computational resources and yielding the best possible outcomes in MAE minimization.

3.5 Evaluation measures, MAE, RMSE, R^2 and Autorank Analysis

In this dissertation, several evaluation measures were employed to comprehensively assess the performance of the ML models presented. These measures include MAE, RMSE, and R^2 . Additionally, the Autorank tool was utilized to systematically compare the performance of these models. Autorank is a Python package designed to simplify statistical comparisons between multiple paired populations Herbold (2020). The principal reason for employing these evaluation metrics and the Autorank analysis is to statistically determine the best-performing models while considering the variability and consistency of their performance across different subsets of the data. This step is crucial in identifying models that not only perform well on

average but also maintain stability and reliability in their predictions.

3.5.1 Mean Absolute Error (MAE)

MAE is a measure of errors between paired observations expressing the same phenomenon. It is defined as the average of the absolute differences between the predicted and actual values. The formula for MAE is:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where y_i represents the actual values, $\hat{y}_i$ represents the predicted values, and n is the number of observations. MAE provides a simple interpretation of the average error degree, making it a useful metric for understanding model performance in practical terms.

3.5.2 Root Mean Squared Error (RMSE)

RMSE is another commonly used measure of prediction error. It is defined as the square root of the average of squared differences between predicted and actual values. The formula for RMSE is:

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

RMSE is mostly useful because it penalizes larger errors more significantly than smaller ones, providing a measure of the model's ability to predict with precision.

3.5.3 R-squared (R^2)

R^2 , or the coefficient of determination, measures the proportion of variance in the dependent variable that is predictable from the independent variables. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of the actual values. R^2 values range from 0 to 1, with higher values representing better model performance. An R^2 value of 1 implies that the model explains all the variability in the response variable, whereas a value of 0 implies no explanatory power.

3.5.4 Methodological Framework of Autorank Analysis

The Autorank tool enables the comparison of multiple models through statistical tests that account for the distributions of their performance metrics, as explored in Maxwell, Delaney, and Kelley (2017). Here is a breakdown of the process:

1. **Normality Testing:** Each model's performance metrics are tested for normality to ensure that the data meets the premisses required for parametric tests. The Shapiro-Wilk test is regularly used for this purpose, where a non-significant result indicates that the data is normally distributed.
2. **Homogeneity of Variance:** Bartlett's test is employed to check for homogeneity of variance (homoscedasticity) among the models. A non-significant result specifies that the variances are equal across groups, satisfying one of the assumptions for ANOVA.
3. **Repeated Measures ANOVA:** If the assumptions of normality and homogeneity are met, a repeated measures ANOVA is executed to determine if there are significant differences between the mean performances of the models. This test is appropriate for comparing the means of three or more groups when the same subjects are used in each group, providing insights into whether any of the models perform significantly better than others.
4. **Post-hoc Tukey HSD Test:** When significant differences are found through ANOVA, the Tukey HSD (Honestly Significant Difference) test is activated as a post-hoc examination to identify which specific models differ from each other. This step certifies that we can pinpoint the models that consistently outperform others by providing pairwise comparisons among the groups.

3.6 Quantile Loss Function

Quantile loss function analysis is a statistical approach used in predictive modelling to provide interval predictions. This method is particularly useful for generating probabilistic forecasts, which offer a more comprehensive understanding of potential results by estimating the conditional quantiles of the response variable. This method can successfully capture the uncertainty and variability in the predictions, making it a valuable tool for assessing and improving the reliability of predictive models.

3.6.1 Theoretical Foundation

The quantile loss function, also recognised as the pinball loss function, is used to optimize a model for a explicit quantile. Given a quantile level $q \in (0, 1)$, the quantile loss for a prediction $\hat{y}$ relative to the true value y is defined as:

$$L_q(y, \hat{y}) = \begin{cases} q \cdot (y - \hat{y}) & \text{if } y \geq \hat{y} \\ (q - 1) \cdot (y - \hat{y}) & \text{if } y < \hat{y} \end{cases}$$

This loss function punishes overestimation and underestimation differently, depending on the quantile being predicted. For instance, the 0.1 quantile (10th percentile) loss will penalize overestimation more deeply, while the 0.9 quantile (90th percentile) loss will penalize underestimation more heavily. By minimizing the quantile loss, the model's predictions might be tuned to capture the specified quantiles accurately. The median or the 0.5 quantile prediction minimizes the MAE. It is represented as:

$$L_{0.5}(y, \hat{y}) = |y - \hat{y}|$$

whereas the general quantile loss for quantile q is minimized when:

$$\hat{y}_q = \arg \min_{\hat{y}} \mathbb{E}[L_q(y, \hat{y})]$$

In the context of this thesis, we focus on predicting the 0.1, 0.5, and 0.9 quantiles to create prediction intervals that incorporate the central 80% of the potential outcomes, thus offering a clear view of the uncertainty in the forecasts. In practice, the quantile loss function analysis considers the following steps:

1. *Model Training:* The selected model, such as CatBoost or RF, is trained separately for each quantile (0.1, 0.5, and 0.9). Each model aims to minimize the corresponding quantile loss function, resulting in three distinct models.
2. *Prediction Intervals:* The trained models are used to predict the respective quantiles for the test set. For example, the model trained for the 0.1 quantile delivers the lower bound of the prediction interval, the model for the 0.5 quantile gives the median prediction, and the model for the 0.9 quantile delivers the upper bound.
3. *Evaluation:* The prediction intervals are evaluated by calculating the coverage, which is the proportion of actual values that fall within the predicted intervals. The formula for coverage is the following:

$$\text{Coverage} = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(y_i \in [\hat{y}_{0.1}, \hat{y}_{0.9}])$$

where $\mathbf{1}$ is an indicator function that equals 1 if the condition inside is true, and 0 otherwise.

4. *Calibration and Sharpness:* Besides coverage, it is indispensable to evaluate the calibration and sharpness of the prediction intervals. Calibration refers to the alignment of predicted quantiles with the observed frequencies. Sharpness, or amplitude, refers to the concentration of the predictive distribution and can be quantified by measuring the average width of the prediction intervals. Narrow intervals may indicate high confidence, while wider intervals suggest greater uncertainty. The formula for sharpness is:

$$\text{Sharpness} = \frac{1}{n} \sum_{i=1}^n (\hat{y}_{0.9,i} - \hat{y}_{0.1,i})$$

3.7 Ensemble Models for Different Time Frames

Ensemble models combine the predictions of numerous base models to improve the overall performance, robustness, and generalizability of the predictive system. In the exact context of carbon intensity forecasting, using ensemble models across different time frames (6 months, 3 months, and 1 month) can meaningfully increase the accuracy and reliability of predictions. This section explains the theoretical foundations of ensemble models, the advantages of using them over varying time frames, and their application to the best-performing

model identified in the case-study.

This method offers us several benefits. Firstly, they offer better generalization by falling the risk of overfitting through leveraging multiple models trained on different subsets of data, as demonstrated in Breiman (1996) work on ensemble methods. Secondly, they offer robustness to noise. By combining predictions from different time frames, the ensemble model can smooth out noise and anomalies, leading to more consistent forecasts as demonstrated by Dietterich (2000). Thirdly, they are adaptive to changes. Including models trained on shorter time frames permits the ensemble to rapidly adapt to recent changes in the data, while longer time frames provide stable trend analysis as shown by Zhou (2012). Lastly, they improve accuracy. Empirical results, such as those by Caruana, Niculescu-Mizil, Crew, and Ksikes (2004), show that ensemble models repeatedly outperform individual models in terms of predictive accuracy and reliability.

3.7.1 Theoretical Foundations

Ensemble learning methods, such as bagging, boosting, and stacking, leverage the collective power of multiple models to produce better predictions. At this point, we elaborate on the theoretical aspects of these methods.

Bagging (Bootstrap Aggregating) encompasses training multiple models on different subsets of the training data, created through random sampling with replacement. Simply, the predictions of these models are then averaged (for regression tasks) to produce a final prediction. Mathematically, if $f_1, f_2, \dots, f_m$ are the individual models, the bagging prediction is given by:

$$\hat{y} = \frac{1}{m} \sum_{i=1}^m f_i(x)$$

Boosting successively trains models, where each model tries to correct the errors of its predecessor. Popular boosting algorithms include AdaBoost, GBDT, and XGBoost. For GBDT, the model prediction at iteration t is:

$$\hat{y}^{(t)} = \hat{y}^{(t-1)} + \eta \cdot h_t(x)$$

where η is the learning rate and $h_t(x)$ is the base learner at iteration t .

Stacking includes training multiple base models and a meta-model that learns to combine the predictions of these base models. The base models' predictions serve as inputs to the meta-model. If $f_1, f_2, \dots, f_m$ are the base models and g is the meta-model, the final prediction is:

$$\hat{y} = g(f_1(x), f_2(x), \dots, f_m(x))$$

3.7.2 Application to Different Time Frames

Ensemble models might be applied to different time frames to capture the varying dynamics of the data. For this study, the best-performing model from previous evaluations was used to construct ensemble models for 6 months, 3 months, and 1 month time frames. The justification behind this approach includes:

Capturing Temporal Variations: Different time frames capture varying seasonal, trend, and cyclic patterns in the data. Shorter time frames (e.g., 1 month) can better adapt to recent changes, while longer time frames (e.g., 6 months) provide a wider perspective.

Improving Robustness: By averaging predictions over multiple time frames, the ensemble model decreases the impact of anomalies or noise that might be present in any single period.

Improved Stability: Ensemble models across different time frames safeguard that the predictions are stable and consistent, mitigating the risk of overfitting to short-term fluctuations.

Mathematical Formulation

Let's assume we have an ensemble of models trained on data subsets from different time frames: 6 months, 3 months, and 1 month. Let f_6, f_3, f_1 be the models trained on these respective time frames. The combined prediction $\hat{y}$ can be formulated as:

$$\hat{y} = \alpha_6 f_6(x) + \alpha_3 f_3(x) + \alpha_1 f_1(x)$$

where $\alpha_6, \alpha_3, \alpha_1$ are the weights assigned to the predictions from each time frame model. These weights might be optimized based on historical performance or might be set equally if no past information is available.

Different Types of Ensemble Methods

In addition to averaging predictions, there are other ensemble methods that could be tested in our CO2 forecasting problem:

Averaging Predictions: Averaging predictions considers taking the mean of predictions from multiple models. This method helps to reduce the variance of the predictions and usually improves the model's robustness.

Stacking: Stacking involves training a meta-model to combine the predictions of different base models. The base models are first trained on the training data, and their predictions are used as input features to train the meta-model. This method often produces better predictive performance because the meta-model learns how to best combine the strengths of each base model.

Blending: Blending is analogous to stacking but typically involves a simpler combination method. Instead of using a complex meta-model, blending often combines the predictions of base models using weighted averages or other simple techniques. This might be quicker to implement and still improve performance by leveraging multiple models.

Chapter 4

Case Study

In this chapter, a case study focused on forecasting carbon intensity in the Portuguese electric power system is explored leveraged by publicly available datasets. The exercise details the data exploration process and underlines the challenge of feature engineering. Initial data processing sets the stage, while advanced techniques like the EF method meaningfully enhance the dataset by creating insightful new features.

4.1 Data Description and Exploratory Data Analysis

4.1.1 Datasets

This dissertation leverages from publicly available data sources, making them highly accessible for replication and further study. Specifically, data is extracted from REN, REE, and ENTSO-E. The data sources are as follows:

- REN data is extracted from the Mercados platform (<https://mercado.ren.pt/>).
- REE data is sourced from e-sios (<https://www.esios.ree.es>).
- ENTSO-E data comes from the Transparency Platform (<https://transparency.entsoe.eu>).

The dataset utilized in this research, although originating from a single source, has been segmented into multiple versions through varied methods of processing and feature engineering. This results in four primary datasets each designed for distinct analysis and application of ML techniques.

The original dataset, labeled as **Input A - Raw Variables**, comprises the raw data collected. It includes 9,384 entries and 28 variables, capturing complete hourly records of various energy metrics from Spain (ES) and Portugal (PT), covering over a year from September 2, 2022, to September 27, 2023. Beyond the target variable CO2 intensity (labeled as CO2_PT), this dataset includes information on energy load, cross-border physical flow, next-day forecasts for energy load and price, and projections for wind and solar generation, providing a comprehensive view of the energy generation and consumption dynamics within the Iberian

market.

Input B - Manual FE reflects the manual feature engineering performed by an expert from the Instituto de Engenharia de Sistemas e Computadores, Tecnologia e Ciência (INESC TEC). This process has expanded the original dataset from 28 to 114 variables. It includes time-based aggregations such as 24-hour averages and 7-day summaries, lagged variables to capture temporal dependencies, and transformations like gradients and power terms, which are crucial for identifying nonlinear relationships in the data.

Input C - Automatic FE represents the dataset enhanced through automated feature engineering techniques. From 28 raw variables, dataset was extended to 223 features in total. This version introduces variables transformed through EF designed to optimize variable selection and transformation autonomously.

Finally, **Input D – Combined FE** merges both manual and automated feature engineering methods, aiming to leverage the strengths of both approaches to enhance the accuracy and robustness of the forecasting models. From 114 raw variables, dataset was extended to 223 features in total.

For clarity of reading, the new added variables obtained from **Input A - Raw Variables** created the new extended dataset labelled as **Input C - Automatic FE**. Similarly, the new added variables obtained from **Input B - Manual FE**, created the new extended dataset labelled as **Input D – Combined FE**.

In order to enable a complete understanding of the dataset, it is essential to describe the key variables which illustrate the energy generation and consumption profiles within Spain and Portugal. Below, in the following Table 4.1 we see a complete description of the main variables present into the dataset.

4.1.2 Trend, Seasonal and Cycling Analysis

In this section, we illustrate the trends of key variables, specifically different energy sources, over time. The following Figure 4.1 provide a comprehensive representation of the main production sources in Portugal and Spain throughout the period covered by the dataset from September 2022, to September 2023.

Fossil Hard Coal production from Spain shows fluctuations, potentially reflecting operational changes or policy shifts. Nuclear Energy production from Spain appears quite stable, which is typical for nuclear energy given its role as a base load power source. Fossil Gas production in Portugal exhibits variations which might be influenced by market demands or supply issues. Hydroelectric Run-of-River production in Portugal also shows some variability, potentially revealing of seasonal patterns in water availability.

Fossil Hard Coal production from Spain shows fluctuations, potentially reflecting operational changes or policy shifts. Nuclear Energy production from Spain appears quite stable, which is typical for nuclear energy given its role as a base load power source. Fossil Gas

Category	Variables	Description
Time-based Features	Date, Hour, Week_day	Record the timestamp of each data entry, providing a chronological framework for the dataset. They are crucial for time series analysis, allowing segmentation and analysis over different time frames to identify patterns or anomalies. Week_day is a binary (0 or 1) categorization that simplifies analysis, showing different patterns in energy consumption or production during weekdays versus weekends.
Fossil Fuel Energy Sources	Fossil_hard_coal_ES, Fossil_gas_PT	These variables quantify the energy generated from hard coal in Spain and gas in Portugal, respectively, highlighting each country's dependence on fossil fuels within their energy mix.
Nuclear Energy	Nuclear_ES	Measures the nuclear energy produced in Spain, reflecting its role in the national energy strategy.
Renewable Energy Sources	Biomass_PT, Hydro_run_river_PT, Hydro_run_river_ES	Detail the biomass and hydroelectric power generation in Portugal and Spain, illustrating the commitment to sustainable energy practices.
Other Energy Sources	Other_ES, Other_renewable_ES, Waste_ES	Encompass various other forms of energy generation in Spain, including both renewable sources and waste energy, underlining the diversity of the energy portfolio.
Interdependencies and Exchanges	Cross_Border_Physical_Flow_PT_to_ES	Provides insights into the dynamics of scheduled commercial exchanges between Portugal and Spain, critical for understanding the interdependencies and operational efficiencies within the Iberian energy market.

Table 4.1: Overview and corresponding description of the dataset features

production in Portugal exhibits variations which might be influenced by market demands or supply issues. Hydroelectric Run-of-River production in Portugal also shows some variability, potentially revealing of seasonal patterns in water availability.

In respect to seasonality, particularly looking at the hydroelectric and solar production in Portugal, it is possible to identify patterns related to seasonal fluctuations. In the following Figure 4.2 we may observe the seasonal trends more evidently.

The charts illustrate seasonal trends in energy production. Hydroelectric Energy production shows significant fluctuations, likely corresponding to seasonal water availability; this includes higher production during wetter months and lower output during drier periods. In the same way, Solar Energy production shows variability, with peaks likely happening during sunnier months (or days), reflecting the higher solar irradiance characteristic of summer time.

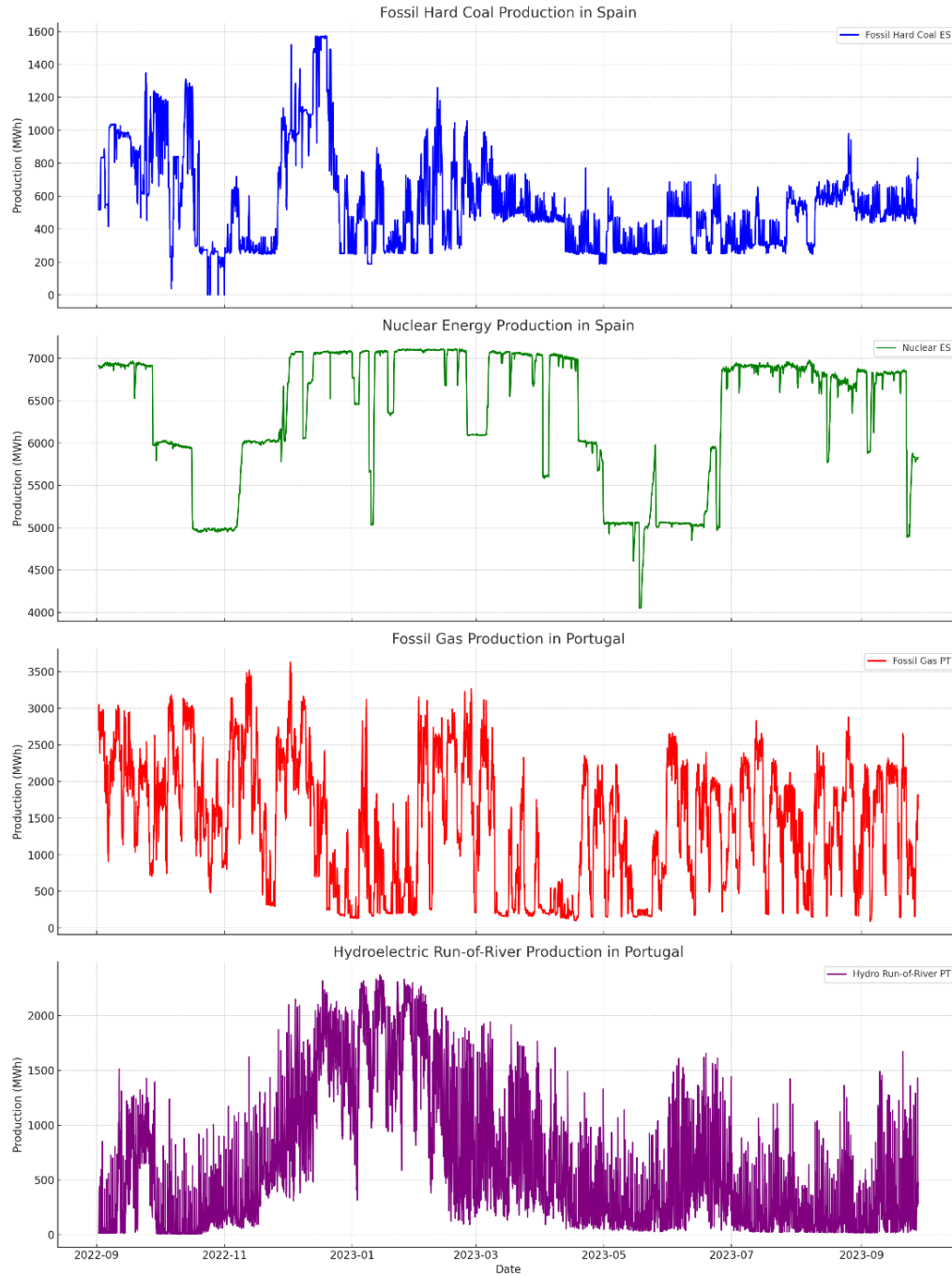


Figure 4.1: Energy production from hard coal in Spain, nuclear in Spain, gas in Portugal, and hydro run-of-river in Portugal

In addition to trend and seasonality analyses, exploring hourly cycles in energy production is crucial for forecasting CO₂ intensity. The cyclic nature of daily energy usage patterns, represented by hourly variations, can significantly influence the environmental impact of energy production. By analysing hourly data, we can identify predictable daily peaks and troughs in energy production. This regular pattern is important for understanding how different types of energy sources contribute to overall CO₂ emissions at different times of the day. The use

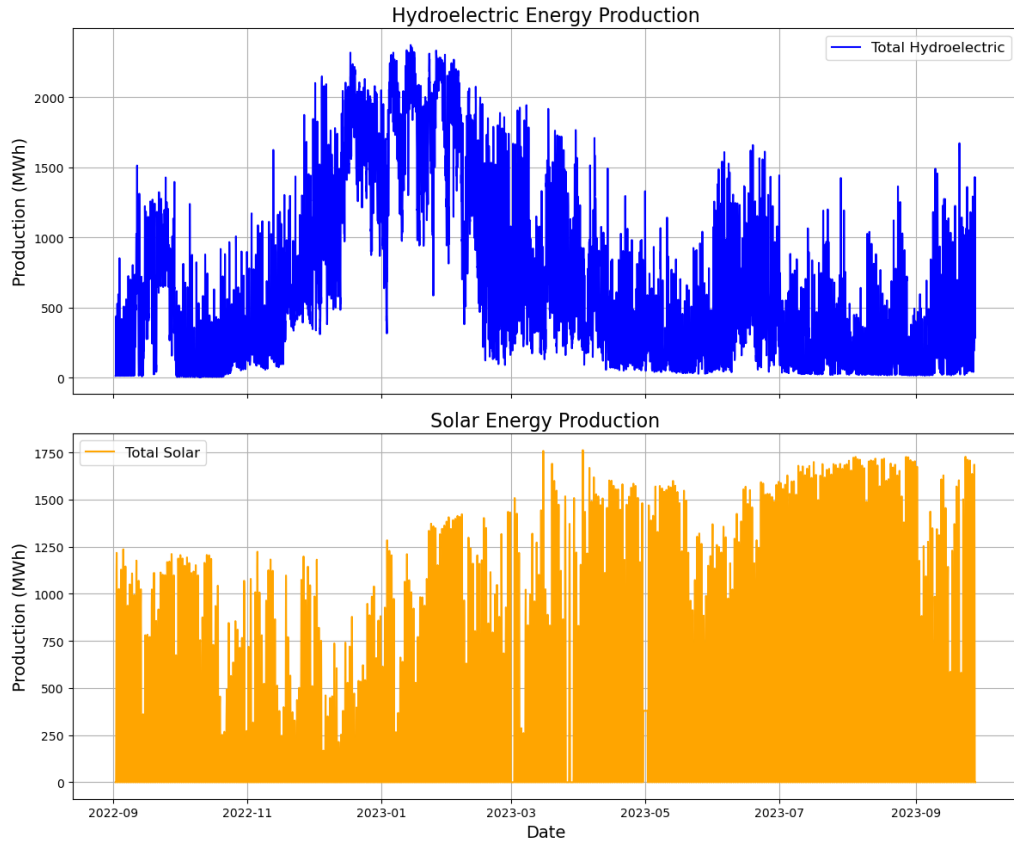


Figure 4.2: Hydro and solar energy production in Portugal

of hourly cycles to represent the time of day helps in capturing these cycles very effectively, for each type of energy production, as we may observe in the following Figure 4.3.

The daily cyclical patterns of various energy production types in Portugal and Spain, as illustrated by the box plots below, provide significant insights into the nature of energy generation.

Solar energy production displays a pronounced bell-shaped curve in both countries, reflecting the availability of sunlight. Peak production aligns with the highest sun elevation around midday, evidenced by the central peaks in the plots. The absence of solar energy production during the early morning and late evening aligns with the predictable lack of sunlight during these periods.

Hydroelectric production shows variability throughout the day, with noticeable peaks and troughs that reflect the natural flow of water and demand for energy.

Wind energy production displays a more consistent output throughout the day but still shows some variability. Unlike solar energy, wind generation does not cease at any specific time of the day, as wind conditions can vary widely, regardless of solar patterns. The plots indicate fluctuations in wind energy output, which could be influenced by daily changes in wind patterns or geographical factors that affect wind speed and steadiness.

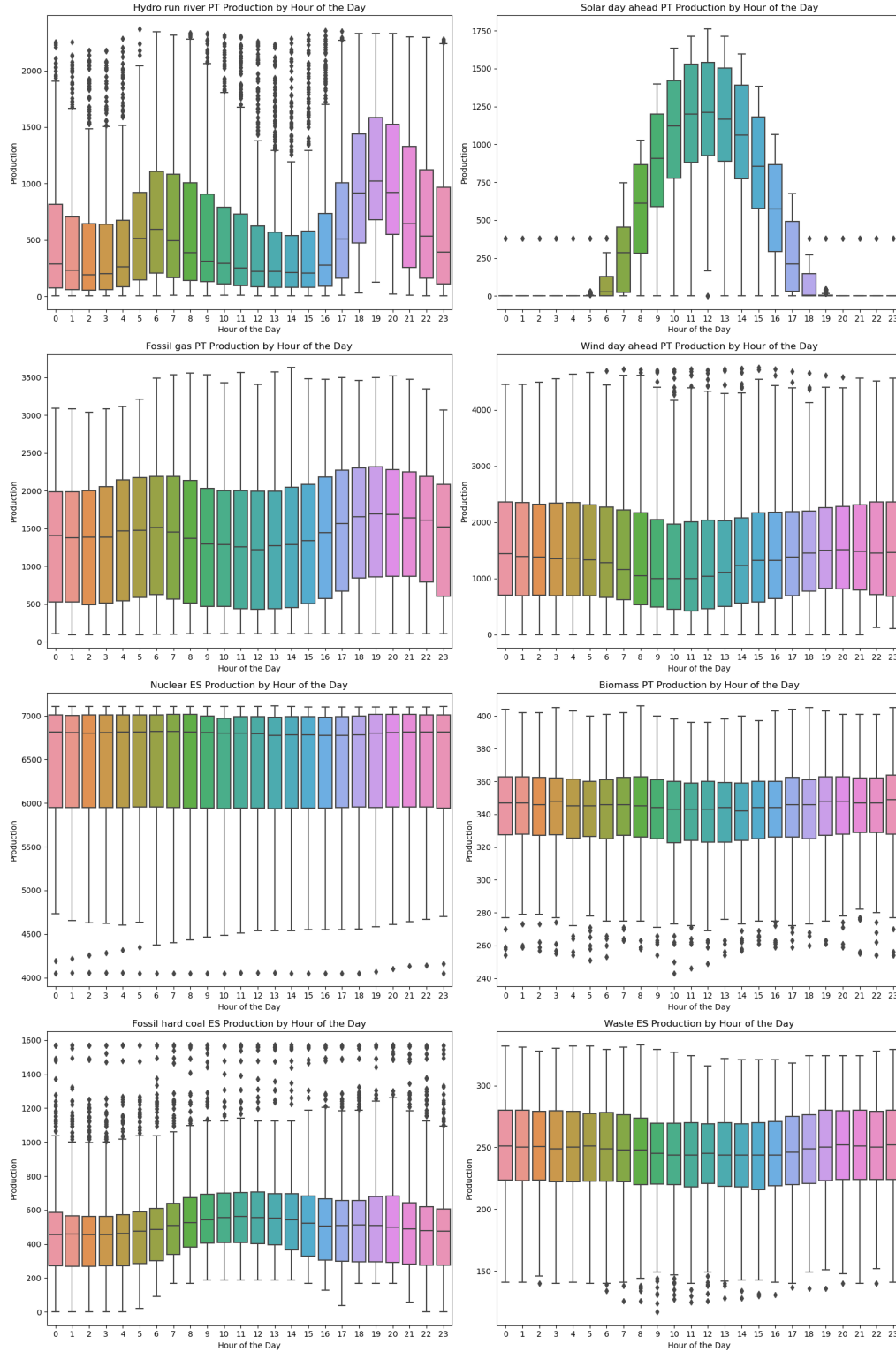


Figure 4.3: Hourly energy production for different sources in Portugal and Spain

Fossil fuel-based energy production (coal and gas) tends to show less variability but still follows a daily cycle with higher production during peak demand hours.

Nuclear energy production remains relatively stable throughout the day, reflecting its role as a base-load power source.

Biomass energy production shows some variability but generally follows a consistent pattern, highlighting its role in providing steady renewable energy.

4.1.3 Correlation

By examining the correlations among these variables, we can comprehend how different sources of energy generation might be related to CO2 emissions and to each other. This analysis helps in identifying dependencies or effects between various energy sources and environmental impacts. The heatmap in the provided image (Figure 4.4) is based on the calculation of the Pearson correlation coefficient between pairs of features in the dataset.

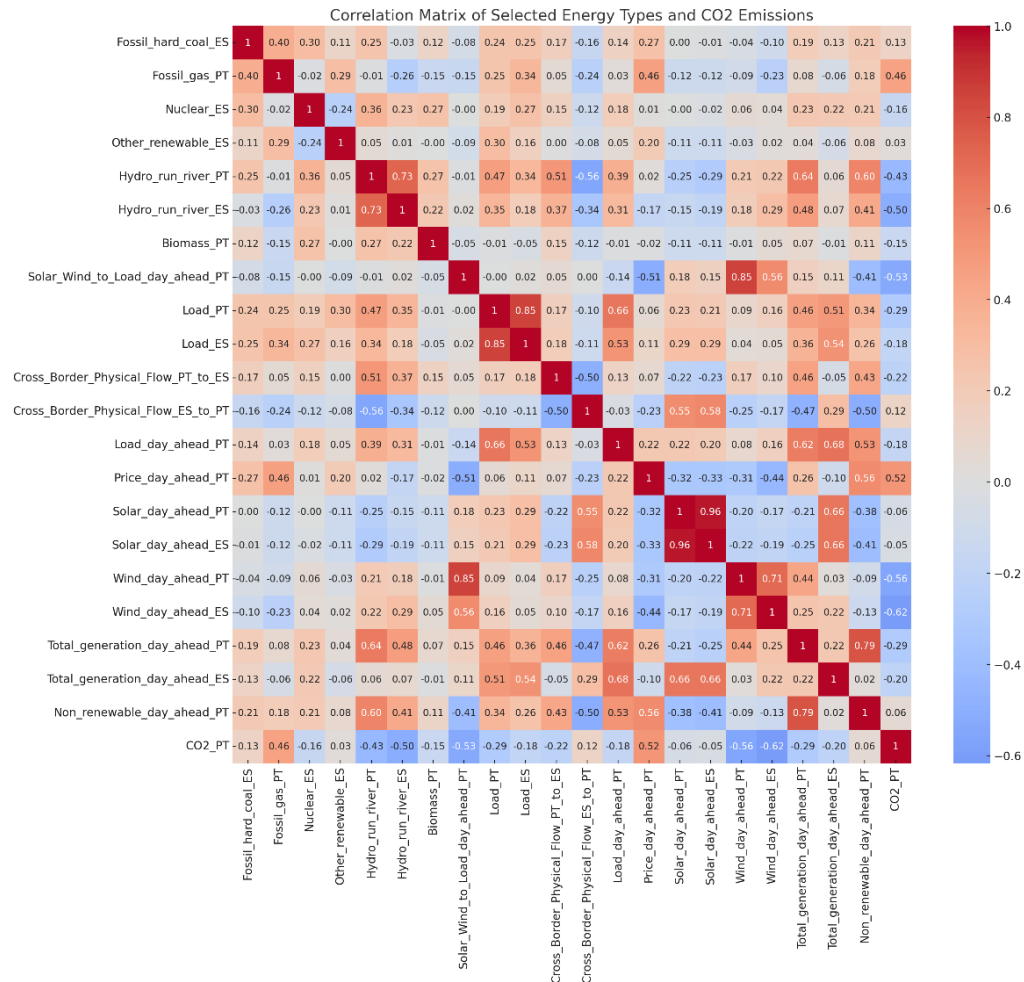


Figure 4.4: Correlation Heatmap for different energy sources in Portugal and Spain

The heatmap displayed above elucidates the relationships between various energy generation methods and CO2 emissions. Notably, fossil fuel generation sources exhibit positive

correlations with CO2 emissions, indicated by the variable 'CO2_PT'. Fossil gas demonstrates a particularly strong correlation of 0.46, underscoring its significant impact on carbon emissions. Similarly, fossil hard coal also shows a positive correlation of 0.13, reflecting its contribution to CO2 levels.

Renewable and hydro power (both 'Hydro_run_river_PT' and 'Hydro_run_river_ES') demonstrate negative correlations with CO2 emissions. For instance, 'Hydro_run_river_PT' has a correlation of -0.43 , indicating that increased utilization of hydroelectric power is associated with significant reductions in CO2 emissions. 'Other_renewable_ES' exhibits a weak positive correlation of 0.03, suggesting a minimal impact on carbon emissions reduction.

Nuclear energy ('Nuclear_ES') displays a low negative correlation of -0.16 with CO2 emissions, aligning with the expectation that nuclear power, which does not directly emit CO2, would have a minimal direct impact on carbon emission levels. This is consistent with the general perception of nuclear energy as a cleaner alternative to fossil fuels, despite other environmental and safety concerns.

Interestingly, the solar and wind energy variables show varied correlations. For example, 'Solar_day_ahead_PT' has a weak negative correlation of -0.06 , suggesting that higher solar energy production is only slightly linked with lower CO2 emissions. Similarly, 'Solar_day_ahead_ES' shows a weak negative correlation of -0.05 . This weak correlation can be attributed to the intermittency and variability of solar energy, which relies deeply on weather conditions and time of day, making it an inconsistent energy source. Moreover, the current energy mix and the need for fossil fuel-based backup systems during periods of low solar output smooth out the impact of solar energy on overall CO2 emissions. Furthermore, limitations in grid integration and energy storage mean that excess solar energy produced during peak times cannot be effectively stored and used during periods of low production, thereby reducing its potential to displace fossil fuel-generated electricity and lower CO2 emissions. In contrast, 'Wind_day_ahead_PT' shows a strong negative correlation of -0.56 , indicating that wind energy significantly contributes to reducing CO2 emissions. 'Wind_day_ahead_ES' also shows a strong negative correlation of -0.62 . In this context, the stronger correlation might indicate that wind energy production is generally more consistent and less dependent on daily and seasonal variations compared to solar energy. Additionally, the infrastructure and grid integration for wind energy may be more developed, allowing for a more effective displacement of fossil fuel-generated electricity..

The feature 'Solar_Wind_to_Load_day_ahead_PT', calculated as $(\text{'Solar_day_ahead_PT'} + \text{'Wind_day_ahead_PT'}) / \text{'Load_day_ahead_PT'}$, represents the proportion of solar and wind energy relative to the total load. This variable shows a negative correlation of -0.53 with CO2 emissions, indicating that a higher proportion of renewable energy in the energy mix is associated with lower CO2 emissions.

These insights underline the variable impacts of different energy sources on environmental outcomes, particularly in terms of carbon emissions. Naturally, fossil fuels remain major contributors to CO2 emissions, while increased adoption of renewable and hydroelectric energy sources presents a viable lane for reducing carbon footprints.

4.2 Feature Engineering

4.2.1 Pre-processed Dataset

The pre-processed dataset (**Input B - Manual FE**) included numerous derived features designed to improve the predictive performance of the ML models. These features encapsulate various aspects of the energy production, consumption, and market dynamics, presenting a rich and detailed basis for further analysis.

The feature engineering done by INESC TEC focused on creating a comprehensive set of variables that reflect both the temporal and categorical characteristics of the data. Some of the key features provided summarized in the following Table 4.2. This additional feature set provides a solid basis for predictive modelling, summarising a wide range of information from the dataset.

The MAE comparisons between **Input A - Raw Variables** and **Input B - Manual FE** across all models are summarized in Table 4.3. This table highlights key patterns and draws conclusions about the efficacy of the models and their sensitivity to the application of feature engineering.

The first conclusion is very evident in the results. All models except ANN exhibit significant improvements in performance when shifting from **Input A - Raw Variables** to **Input B - Manual FE**, as demonstrated by the decrease in MAE values and the negative percentage changes. This displays the efficacy of the feature engineering applied in **Input B - Manual FE**. As mentioned, the ANN model is the only one that shows a slight increase in MAE, indicative of a potential mismatch or overfitting due to the additional features in **Input B - Manual FE**. The slight increase in MAE for ANN (1.94%) might be attributed to a few factors. ANN needs a large amount of data to generalize well, and the feature engineering might have introduced complexity that led to overfitting or struggle in learning effective patterns from the data. It is also likely that the additional features were not sufficiently informative for ANN compared to other models that might better seize non-linear relationships with smaller datasets.

Tree-based models such as RF, ET, and GBDT reveal substantial improvements, with percentage decreases in MAE ranging from approximately 9% to 12%. This advocates that tree-based models benefit significantly from the developed feature set, likely due to their capability to capture complex interactions between features. The ensemble nature of these models permits them to successfully capture a wide range of feature interactions, making them a appropriate choice for datasets with high feature range.

Boosting models like CatBoost and XGBoost also show noticeable improvements. CatBoost shows an improvement of 6.75%, while XGBoost displays a reduction of 8.69% in MAE. The capacity of these models to handle categorical features and high-dimensional data effectively might explain their greater performance. Datetime-derived features provide valu-

Category	Variables	Description
Daily Averages	Fossil_hard_coal_ES_24_hour_mean, Nuclear_ES_24_hour_mean, Hydro_run_river_PT_24_hour_mean, Renewable_percentage_day_ahead_PT, Load_day_ahead_PT_24_hour_mean, Solar_day_ahead_PT_24_hour_mean	Represent the 24-hour mean values of various energy sources, capturing daily trends in energy generation.
Lagged Variables	Load_day_ahead_PT_1h_lag, Wind_day_ahead_PT_2h_lag, Total_generation_day_ahead_ES_3h_lag	Represent lagged values, capturing temporal dependencies and the influence of past values on future predictions.
Higher-Order Terms	Load_day_ahead_PT_power_2, Solar_day_ahead_PT_power_3	Introduce polynomial terms to capture non-linear relationships.
Extreme Values	Load_day_ahead_PT_daily_max, Wind_day_ahead_PT_daily_min	Capture the daily maximum and minimum values, highlighting extreme behaviours.
Gradients	Load_day_ahead_PT_grad, Solar_day_ahead_PT_grad	Represent the gradients or rate of change, providing insights into the dynamics of changes in the data.
Temporal Sine and Cosine Transformations	sin_hour, cos_hour	Capture the cyclical nature of hourly data, reflecting daily patterns in energy consumption and production.

Table 4.2: Key features developed through feature engineering by INESC TEC

able information to the boosting models, likely leading to better performance in capturing time-dependent patterns.

LightGBM and DART, which are both hybrid tree-based boosting models, exhibit clear improvements in MAE when shifting from **Input A - Raw Variables** to **Input B - Manual FE**. LightGBM shows a significant decrease of 14.6% in MAE, while DART displays a decrease of 7.92%. This improvement aligns with the performance of other tree-based and boosting models.

Linear models such as Linear Regression and LASSO show the most significant improvements, with LASSO showing a decrease of over 20%. This specifies that linear models might benefit from well-engineered features that capture linear relationships more effectively. The significant 23.12% improvement in Linear Regression emphasises the positive impact of fea-

Model	Input A - Raw Variables	Input B - Manual FE	Abs. Change	% Change
RF	30.29	26.57	-3.72	-12.28%
ET	29.51	26.39	-3.12	-10.57%
AdaBoost	34.87	33.34	-1.53	-4.39%
GBDT	28.89	26.37	-2.52	-8.73%
XGBoost	28.66	26.17	-2.49	-8.69%
CatBoost	27.42	25.57	-1.85	-6.75%
Linear Regression	33.03	25.39	-7.64	-23.12%
ANN	30.33	30.92	+0.59	+1.94%
SVM	29.30	26.11	-3.19	-10.89%
LASSO	32.17	25.61	-6.56	-20.40%
LightGBM	30.01	25.63	-4.38	-14.60%
DART	28.30	26.06	-2.24	-7.92%

Table 4.3: Comparison of MAE results for **Input A - Raw Variables** and **Input B - Manual FE**, before application of Evolutionary Forest

ture engineering that enhances linearity in the dataset.

SVM also benefits from the feature engineering as well, showing an improvement around 10%, suggesting that the supplementary features help in better defining the decision boundaries in the feature space, allowing SVM to improve its performance meaningfully.

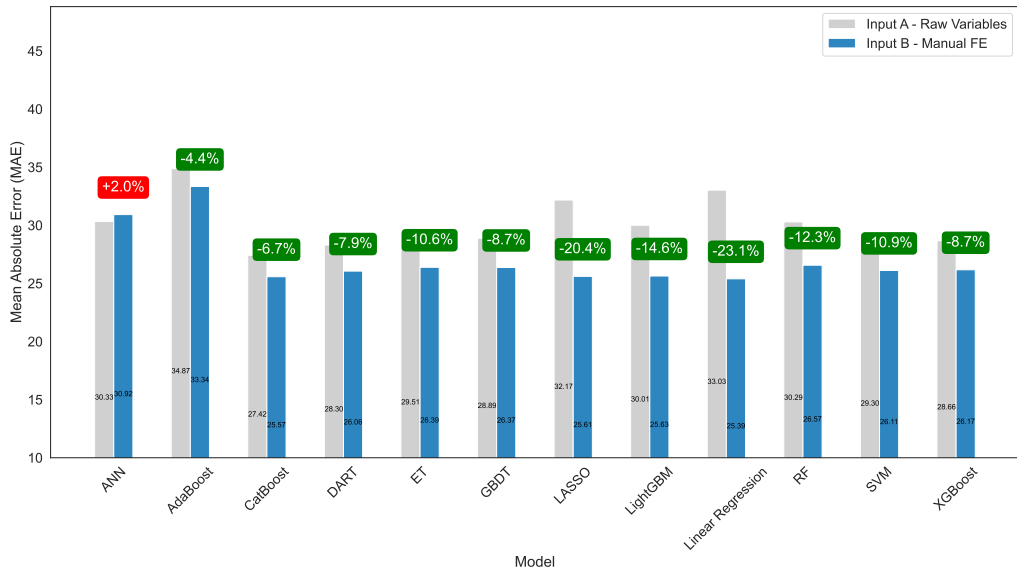


Figure 4.5: Comparative performance analysis of MAE across various models for **Input A - Raw Variables** and **Input B - Manual FE** prior to the application of the Evolutionary Forest

The Figure 4.5 provides a visual interpretation of how different models respond to feature engineering. Tree-based and boosting models generally display substantial developments. Linear models benefit the most from improved linear relationships, while ANN, interestingly, experiences a small degradation in performance.

4.2.2 Further Feature Engineering with Evolutionary Forest

Building on the pre-processed dataset, additional feature engineering will be conducted to enhance the model's predictive capabilities. This additional phase involves the application of EF technique to uncover complex patterns and interactions within the data, that were not captured in the initial feature set. In order to evaluate the value added and effectiveness, this procedure will be applied over **Input A - Raw Variables** and **Input B - Manual FE**.

In the first step, Optuna was used to perform hyperparameter optimization for an *EvolutionaryForestRegressor*. It was defined a range of possible values for several hyperparameters and searched for the combination that minimizes the MAE on the test dataset. The best combination of hyperparameters was selected for both **Input A - Raw Variables** and **Input B - Manual FE**.

Parameter	Description	Input A	Input B
basic_primitives	Basic operations used in the model: Addition (Add), Subtraction (Sub), Multiplication (Mul), Analytic Quotient(AQ)	Add,Sub,Mul,AQ	Add,Sub,Mul,AQ
boost_size	Number of boosts to perform in the evolutionary process	142	140
cross_pb	Crossover probability in the genetic algorithm	0.67297	0.91911
ensemble_size	Number of trees in the ensemble	142	140
gene_num	Number of genes in each individual tree	20	20
mab_parameter	Multi-Armed Bandit parameter settings (default: standard exploration-exploitation settings)	{}	{}
map_elite_parameter	Parameters for the Map-Elite evolutionary strategy (default: basic grid discretization). Options include map_size, elite_fraction, niches	{}	{}
max_height	Maximum height of the trees	10	4
n_gen	Number of generations for the evolutionary process	9	6
n_pop	Population size for the genetic algorithm	150	123
random_state	Seed for random number generation for reproducibility	42	42
select	Selection method used in the evolutionary process	Automatic Lexicase	Automatic Lexicase
verbose	Verbosity of the training process	True	True

Table 4.4: Hyperparameter configurations for *Evolutionary Forest Regressor*

The *Evolutionary Forest Regressor* was configured with specific parameters, as represented in the Table 4.4, such as basic primitives (addition represented as 'Add', subtraction as 'Sub',

multiplication as 'Mul', and a 'AQ'), boost size, cross probability, ensemble size, number of generations, population size, and selection method. During the training process, the *Evolutionary Forest Regressor* generated an initial population of decision trees, evaluated their performance, and iteratively evolved the population through selection, crossover, and mutation over several generations. The fitness of the trees was assessed using the R^2 metric, and the process aimed to optimize this fitness score.

After the final iteration, the trained *Evolutionary Forest Regressor* was used to transform the original features and generate new features. Detailed feature importance was calculated, and the top features were selected based on their importance scores.

Finally, the ML models were retrained using the new feature set generated by the *Evolutionary Forest Regressor*. The performance metrics of these models were compared to the baseline metrics to evaluate the improvement.

Evolutionary Forest new features

The transform method of the *Evolutionary Forest Regressor* is used to generate new features from the training and testing sets. The importance of these new features is then calculated, using a custom function 'get_feature_importance'. The top features are selected based on their importance scores. The elected top features absolute number are different between **Input A - Raw Variables** and **Input B - Manual FE**, even though, the optimal criteria (threshold defined in percentage) for selection is similar. In both cases (**Input A** and **Input B**) the top 50% of features were selected. In total, out of 391 created new features, 195 were selected over **Input A - Raw Variables**, while over **Input B - Manual FE**, 257 new features were selected out of 514 newly created features. Therefore, that means, for perfect understanding, the new variables derived from **Input A - Raw Variables** resulted in the formation of **Input C - Automatic FE**. Likewise, **Input D - Combined FE**, is an extended dataset originated from **Input B - Manual FE**.

The optimal threshold for selecting top features was determined based on their importance scores. An iterative process was employed to identify the optimal percentage of top features to retain, ranging from 25% to 75% in increments of 5%. For each iteration, the MAE was evaluated across all models.

The Table 4.5 shows the MAE comparison between **Input B - Manual FE** and **Input D - Combined FE** across all the trained and tested models. This table discloses several insightful patterns and conclusions regarding model performance and their response to the feature engineering techniques employed by EF.

The outcomes indicate a broad improvement in performance for most models post-EF application. A vast majority of the models, such as RF, ET, GBDT, and various boosting algorithms like AdaBoost, CatBoost, and XGBoost, demonstrate a decrease in MAE values, demonstrating the effectiveness of EF in increasing the predictive capabilities of these models.

Tree-based models, including RF, ET, and GBDT display good improvements, with percentage decreases in MAE of approximately 1.39%, 4.81%, and 4.25%, respectively. These

Model	Input B - Manual FE	Input D - Combined FE	Abs. Change	% Change
RF	26.57	26.20	-0.37	-1.39%
ET	26.39	25.12	-1.27	-4.81%
AdaBoost	33.34	31.91	-1.43	-4.29%
GBDT	26.37	25.25	-1.12	-4.25%
XGBoost	26.17	25.76	-0.41	-1.57%
CatBoost	25.57	24.50	-1.07	-4.19%
Linear Regression	25.39	28.58	+3.19	+12.56%
ANN	30.92	28.96	-1.96	-6.34%
SVM	26.11	27.73	+1.61	+6.17%
LASSO	25.61	26.50	+0.89	+3.47%
LightGBM	25.63	26.43	+0.80	+3.12%
DART	26.06	26.28	+0.22	+0.84%

Table 4.5: Detailed comparison of MAE results for the selected ML models before and after implementation of Evolutionary Forest

outcomes suggest that the complex feature interactions taken by EF provided an additional layer of refinement to these ensemble models. ET and GBDT experienced better improvements compared to RF due to their ability to get advantage more from the greater feature diversity and interaction depth. ET uses randomness in feature selection, permitting it to exploit on the enriched feature space, while GBDT incrementally refines weak learners, effectively leveraging new feature connections introduced by EF. RF, on the other hand, is more limited in its aptitude to exploit these enhancements since it relies on averaging over many decision trees, which might smooth out subtle feature interactions. This averaging process may make RF less sensitive to the supplementary complexities captured by the refined features, consequently resulting in relatively modest gains compared to its counterparts.

Boosting models, including CatBoost, XGBoost, and AdaBoost, show also good improvements (overall similar improvements as tree-based models), with CatBoost realizing a 4.19% reduction, XGBoost a 1.57% decline, and AdaBoost a 4.29% diminution in MAE. CatBoost and AdaBoost benefit more from the enriched feature set introduced by EF due to their unique algorithmic strengths. CatBoost excels at handling categorical features and employs an efficient application of gradient boosting with ordered boosting, which decreases overfitting and captures elaborate feature interactions effectively. This makes CatBoost particularly well-suited to leverage the upgraded feature space. AdaBoost, on the other hand, improves the accuracy of weak learners by tuning the weights of misclassified instances, letting it to adapt rapidly to the new patterns identified by EF. In contrary, XGBoost, while extremely effective in general, benefits less due to its dependence on traditional gradient boosting without the specialized handling of categorical data and the same degree of adaptation to misclassification as AdaBoost, therefore, the EF presents transformations that align more closely with the specialized capabilities of CatBoost and AdaBoost.

Interestingly, while hybrid tree-based boosting models like LightGBM and DART are typically strong performers across various contexts, they experienced a slight increase in MAE following the application of EF. This could be attributed to their inherent ability to handle complex features effectively, which may limit the incremental benefits gained from the addi-

tional features introduced by EF.

Linear models (Linear Regression and LASSO), present an interesting outcome. While Linear Regression displays a 12.56% increase in MAE, reflecting a deterioration in performance, LASSO shows a moderate increase of 3.47%. The observed performance deterioration suggests that EF's feature transformations may have accentuated non-linear relationships that these linear models could not leverage. In place of benefiting from the enriched feature space, these models face challenges due to their intrinsic linearity assumptions, incapacity to manage multicollinearity, and sensitivity to noise.

ANN reveal a notable decrease in MAE of 6.34%. This reduction may reflect ANN capacity to adapt to new patterns carried by EF, provided the network has sufficient capacity (it includes both the network size and the ability to generalize from the data, network architecture, depth, and width) and data to train on.

SVM shows an increase in MAE by 6.17%, pointing towards a probable misalignment between the new features and the decision boundaries required for optimal performance in SVM. The increase in error suggests that while SVM may take advantages from refined features, the transformations introduced may not have effectively captured the linear relationships or boundaries required for optimal regression. This misalignment may lead to sub-optimal hyperplane positioning, particularly if the data's true complexity is not adequately addressed by the SVM's kernel.

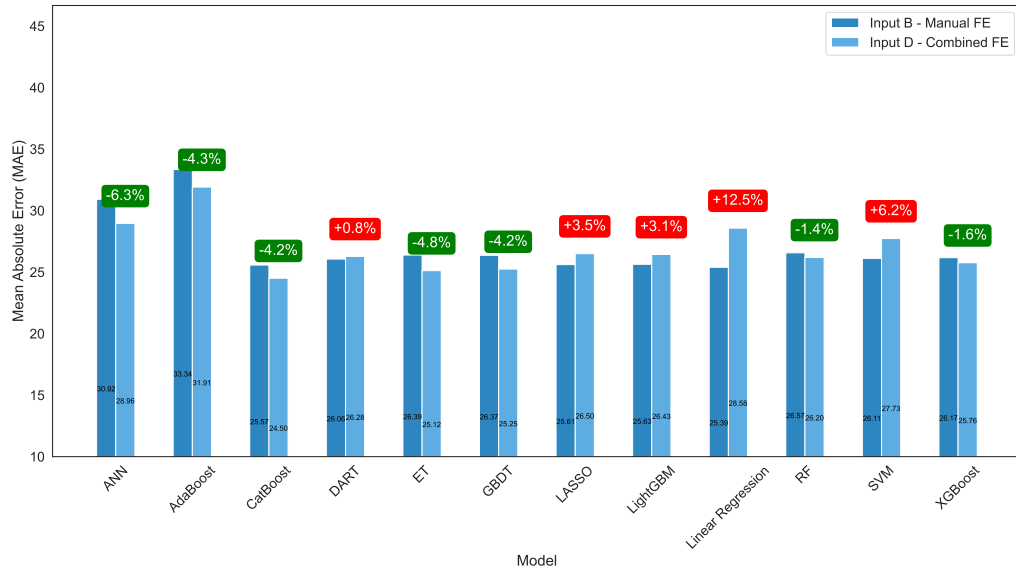


Figure 4.6: Comparison of MAE performance across various ML models before and after application of the Evolutionary Forest

In perspective, as visually shown in the Figure 4.6, EF added significant value to most models by enhancing feature complexity and interaction, which benefited models that exploit on such features, particularly boosting and tree-based algorithms. Nevertheless, the results also highlight that not all models equally benefit from this method, as seen in the less positive results for Linear regression models, tree-based boosting models (DART and LightGBM),

and SVM. Accordingly, while EF is a powerful tool for feature transformation, its success in improving model performance is dependent upon the alignment of the engineered features with the specific strengths and weaknesses of each model. Obviously, it is relevant to consider the nature of the dataset and model when implementing such advanced feature engineering techniques.

Feature Importance

Feature importance is a crucial concept in ML and data science, referring to methods that determine each feature's impact on predicting the target variable. By calculating feature importance, we identify the most powerful features, enhancing model interpretability and performance. This process helps in feature selection, simplifying models, and reducing overfitting, while also diagnosing model performance by revealing which features the model relies on. Besides, it also might contribute to uncover biases or irrelevant features.

In this study, the feature importance of various models was assessed both with and without the use of EF. By comparing the top 10 features identified through feature importance for models with and without EF, we can evaluate the added value of this advanced feature engineering technique.

When evaluating feature importance for models like RF and GBDT, which inherently provide importance scores through their structure, we can directly interpret the impact of each feature. These models calculate feature importance based on metrics like Mean Decrease in Impurity (MDI) or Mean Decrease in Accuracy (MDA), which measure how much each feature contributes to reducing the loss function or improving prediction accuracy.

Nevertheless, for models such as ANN and SVM, feature importance isn't inherently available due to their nature. To overcome this, *Permutation Importance* is employed. *Permutation Importance* works by arbitrarily shuffling the values of each feature and measuring the change in model performance, typically using metrics like MAE or accuracy. The rank of a feature is quantified based on the increase in the error when its values are shuffled; a greater increase in error specifies higher feature importance.

By evaluating the top 10 features, it is possible to identify the most significant contributors to the model's predictive power, and we may observe the effectiveness with and without the application of the application of EF. Looking carefully at the ranked scores there is a clear evidence about the dominance of EF-generated features in the top 10 lists across the ML models. In the case of the DART and AdaBoost models, a striking result was observed: all 10 of the top features were generated by the EF. Then, we have 7 models (RF, ET, GBDT, XGBoost, Linear Regression, LightGBM) in which we observe that 9 out of the top 10 features were created by the EF. The following image (Figure 4.7) illustrates the substantial impact of EF on the feature importance for top-performing models. Curiously, the feature 'Price_day_ahead_PT_24_hour_mean' stands out as the only feature (engineered by INESC TEC) that remained among the top ten most important features across the models after applying the EF technique.

Before Evolutionary Forest			After Evolutionary Forest		
Feature	Score	Feature Source	Feature	Score	Feature Source
Catboost					
Price_day_ahead_PT_24_hour_mean	13.613	Feature Engineering INESC	AQ(Hydro_water_reservoir_PT_24_hour_mean, Non_renewable_day_ahead_PT_daily_min)	4.973	Evolutionary Forest
Non_renewable_day_ahead_PT_daily_min	8.576	Feature Engineering INESC	Mul(month, Non_renewable_day_ahead_PT_daily_min)	4.744	Evolutionary Forest
Hydro_run_river_ES_24_hour_mean	5.404	Feature Engineering INESC	AQ(Price_day_ahead_PT_24_hour_mean, Wind_day_ahead_ES_24_hour_mean)	4.596	Evolutionary Forest
Hydro_run_river_ES	4.575	Raw Feature	Sub(Wind_day_ahead_PT_daily_min, Fossil_gas_PT)	4.449	Evolutionary Forest
Wind_day_ahead_PT_daily_max	4.213	Feature Engineering INESC	AQ(Non_renewable_day_ahead_PT_daily_max, Hydro_run_river_PT_24_hour_mean)	4.114	Evolutionary Forest
Wind_day_ahead_ES	4.051	Raw Feature	Price_day_ahead_PT_24_hour_mean	3.446	Feature Engineering INESC
Price_day_ahead_PT	3.551	Raw Feature	AQ(Price_day_ahead_PT_24_hour_mean, Hydro_water_reservoir_PT)	2.589	Evolutionary Forest
month	2.857	Raw Feature	AQ(Non_renewable_day_ahead_PT_daily_min, Hydro_run_river_ES_24_hour_mean)	2.554	Evolutionary Forest
Hydro_water_reservoir_PT_24_hour_mean	2.621	Feature Engineering INESC	AQ(Fossil_gas_PT, Wind_day_ahead_PT_3h_lag)	2.321	Evolutionary Forest
Renewable_percentage_PT	2.247	Feature Engineering INESC	AQ(AQ(Price_day_ahead_PT, Renewable_percentage_day_ahead_PT_grad), Mul(Total_generation_day_ahead_ES, Load_PT))	1.860	Evolutionary Forest
ET - Extra Trees Regressor					
Price_day_ahead_PT_24_hour_mean	0.134	Feature Engineering INESC	Sub(Wind_day_ahead_PT_daily_min, Fossil_gas_PT)	0.085	Evolutionary Forest
Wind_day_ahead_ES	0.079	Raw Feature	Price_day_ahead_PT_24_hour_mean	0.056	Feature Engineering INESC
year	0.057	Raw Feature	AQ(Price_day_ahead_PT_24_hour_mean, Wind_day_ahead_ES_24_hour_mean)	0.042	Evolutionary Forest
Non_renewable_day_ahead_PT_daily_min	0.039	Feature Engineering INESC	AQ(Hydro_water_reservoir_PT_24_hour_mean, Non_renewable_day_ahead_PT_daily_min)	0.041	Evolutionary Forest
Price_day_ahead_PT	0.038	Raw Feature	Sub(Hydro_run_river_PT_24_hour_mean, Fossil_gas_PT)	0.035	Evolutionary Forest
Wind_day_ahead_PT_daily_max	0.037	Feature Engineering INESC	Add(Wind_day_ahead_ES, Hydro_run_river_ES)	0.035	Evolutionary Forest
Wind_day_ahead_PT	0.036	Raw Feature	Mul(month, Non_renewable_day_ahead_PT_daily_min)	0.029	Evolutionary Forest
Wind_day_ahead_ES_24_hour_mean	0.034	Feature Engineering INESC	Add(Price_day_ahead_PT_24_hour_mean, solar_day_ahead_PT_daily_min)	0.023	Evolutionary Forest
Hydro_run_river_ES_24_hour_mean	0.030	Feature Engineering INESC	AQ(Non_renewable_day_ahead_PT_daily_min, Hydro_run_river_ES_24_hour_mean)	0.021	Evolutionary Forest
Wind_day_ahead_PT_1h_lag	0.026	Feature Engineering INESC	Add(Hydro_water_reservoir_PT, Wind_day_ahead_PT_3h_lag)	0.021	Evolutionary Forest
GBDT - Gradient Boosting Decision Trees					
Price_day_ahead_PT_24_hour_mean	0.319	Feature Engineering INESC	Sub(Wind_day_ahead_PT_daily_min, Fossil_gas_PT)	0.143	Evolutionary Forest
Wind_day_ahead_ES	0.091	Raw Feature	AQ(Fossil_gas_PT, Wind_day_ahead_PT_3h_lag)	0.115	Evolutionary Forest
Hydro_run_river_ES_24_hour_mean	0.061	Feature Engineering INESC	AQ(Price_day_ahead_PT_24_hour_mean, Wind_day_ahead_ES_24_hour_mean)	0.110	Evolutionary Forest
Wind_day_ahead_PT_daily_max	0.053	Feature Engineering INESC	AQ(Hydro_water_reservoir_PT_24_hour_mean, Non_renewable_day_ahead_PT_daily_min)	0.093	Evolutionary Forest
Non_renewable_day_ahead_PT_daily_min	0.047	Feature Engineering INESC	Mul(month, Non_renewable_day_ahead_PT_daily_min)	0.051	Evolutionary Forest
Hydro_water_reservoir_PT_24_hour_mean	0.042	Feature Engineering INESC	Add(Hydro_water_reservoir_PT, Wind_day_ahead_PT_3h_lag)	0.045	Evolutionary Forest
Renewable_percentage_day_ahead_PT	0.036	Feature Engineering INESC	AQ(Non_renewable_day_ahead_PT_daily_max, Hydro_run_river_PT_24_hour_mean)	0.044	Evolutionary Forest
Renewable_percentage_PT	0.034	Feature Engineering INESC	Price_day_ahead_PT_24_hour_mean	0.042	Feature Engineering INESC
Price_day_ahead_PT	0.030	Raw Feature	AQ(Non_renewable_day_ahead_PT_daily_min, Hydro_run_river_ES_24_hour_mean)	0.037	Evolutionary Forest
Nuclear_ES	0.023	Raw Feature	Add(Other_ES_24_hour_mean, Price_day_ahead_PT)	0.029	Evolutionary Forest

Figure 4.7: Top 10 features for the top 3 performing models

Complementary Feature Engineering

The feature sets created by INESC TEC and the EF are complementary rather than redundant. The features engineered by INESC TEC provide a solid foundational dataset by capturing essential linear and non-linear relationships, time dependencies, and cyclic behaviours. This complete feature set lays the basis that enables further enrichment. In the meantime, the EF builds upon this foundation by exploring even more complex feature relationships and transformations, which traditional feature engineering methods may overlook. By combining INESC TEC's features using advanced mathematical operations, the EF increases the existing feature set, efficiently expanding the information available to models.

As demonstrated in the earlier subsection the top features identified by the EF recurrently utilize features originally created by INESC TEC, as exposed in the feature importance analysis. For instance, the EF generates features like 'AQ(Price_day_ahead_PT_24_hour_mean, Wind_day_ahead_ES_24_hour_mean)' by combining INESC TEC's already-informative features. This indicates a kind of dependency, where the success of EF's feature creation is reliant on the quality of INESC TEC's initial features. In conclusion, INESC TEC's feature engineering provides a robust foundation by capturing essential patterns through statistical and mathematical transformations, offering the raw materials for EF's advanced operations. In our dataset, the improvements are particularly pronounced in boosting and tree-based models.

4.3 Models' evaluation

As we have seen in the feature engineering section, several ML models are implicated, in order to better understand the different impacts across the different models (always with tuned hyperparameters for models to exploit optimal performance). Thereafter by using *Autorank Tool*, we can objectively (statistically) rank the models based on performance metrics to identify the best-performing one. Finally, to refine predictive accuracy, we further explored the best-performing model by applying the quantile loss function and implementing an en-

semble approach across different time frames.

4.3.1 Autorank Analysis

Based on the MAE values presented in the Table 4.6 for all tested models, CatBoost emerges as the top-performing model for CO2 forecasting in Portuguese context. However, it is important to statistically validate this superiority to ensure that the observed performance differences are significant and not due to random variations in the data. This validation can be done using *Autorank Tool*, which offers a robust statistical framework to rank the models objectively based on their performance metrics.

Model	Input A - Raw Variables	Input B - Manual FE	Input D – Combined FE
RF	30.29	26.57	26.20
ET	29.51	26.39	25.12
AdaBoost	34.87	33.34	31.91
GBDT	28.89	26.37	25.25
XGBoost	28.66	26.17	25.76
CatBoost	27.42	25.57	24.50
Linear Regression	33.03	25.39	28.58
ANN	30.33	30.92	28.96
SVM	29.30	26.11	27.73
LASSO	32.17	25.61	26.50
LightGBM	30.01	25.63	26.43
DART	28.30	26.06	26.28

Table 4.6: Comparative analysis of MAE across ML models

The statistical analysis was conducted on 12 different models, each with 6 paired samples. The tests were completed with a family-wise significance level of $\alpha = 0.05$. I started by checking the normality of all populations using Shapiro-Wilk test, where that fails to reject the null hypothesis, confirming that the populations are normal (minimum observed p-value = 0.111). Bartlett's test for homogeneity was applied subsequently, and again failed to reject the null hypothesis ($p = 0.076$), indicating that our data is homoscedastic.

The mean values, standard deviations for each model were calculated and d 95% confidence intervals are presented in the Table 4.7 below:

The Tukey HSD test revealed that the models CatBoost, DART, ET, LASSO, Linear Regression, GBDT, LightGBM, XGBoost, and SVM, did not have statistically significant differences in their mean performance scores. This shows that these models perform similarly well on our dataset.

CatBoost's mean rank was the highest among all models, with insignificant effect size differences compared to DART and ET. The effect sizes for models such as XGBoost, SVM, ANN, RF, and AdaBoost were large, indicating substantial differences from CatBoost, highlighting that these models significantly diverge in performance from CatBoost.

Model	Mean Rank	Mean (M)	Std (SD)	CI Lower	CI Upper
CatBoost	10.67	27.91	2.27	25.06	30.77
DART	9.00	28.47	1.31	25.62	31.33
ET	8.83	28.47	2.72	25.62	31.33
LASSO	8.83	28.54	2.42	25.68	31.39
Linear Regression	7.67	29.02	2.44	26.16	31.87
GBDT	7.17	29.35	2.82	26.49	32.20
LightGBM	6.50	29.79	3.84	26.94	32.65
XGBoost	5.00	31.42	4.70	28.56	34.27
SVM	5.00	30.51	2.02	27.65	33.36
ANN	4.17	30.96	1.01	28.10	33.81
RF	4.00	31.61	4.48	28.76	34.47
AdaBoost	1.17	35.58	2.34	32.73	38.44

Table 4.7: Mean ranks, mean (M), standard deviation (SD), and 95% confidence intervals (CI) for different models

Remarkably, as shown in the Figure 4.8, CatBoost consistently showed the lowest MAE, making it the top-performing model. Nevertheless, the statistical analysis confirms that other models in this group also deliver reliable and robust performance.

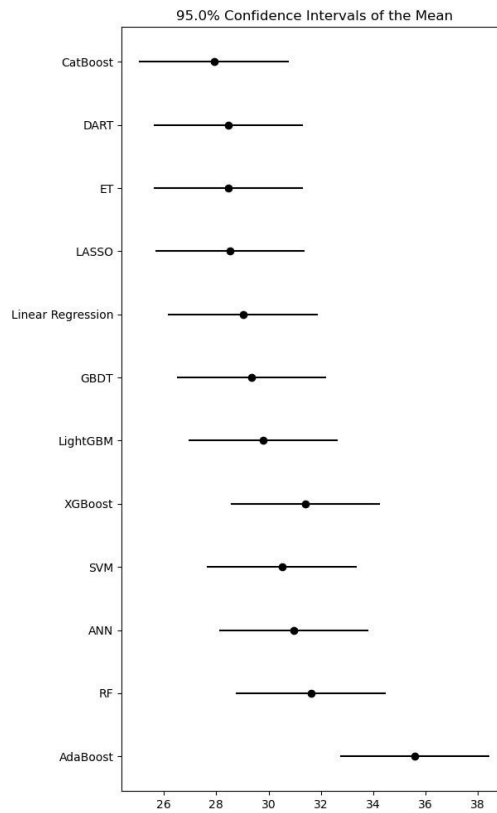


Figure 4.8: 95% confidence intervals for MAE across different ML models, ranked to show variability in model performance

The *Autorank* analysis statistically validates CatBoost as the top-performing model for CO2 forecasting, with relevant performance improvements attributed to its advanced handling of feature interactions and robustness against overfitting. This rigorous approach safeguards that the model selection is grounded on solid statistical evidence, confirming CatBoost’s suitability for our specific problem context.

4.3.2 Quantile Loss Function

After identifying CatBoost as the best performing model through extensive assessment, it was analysed its performance using the quantile loss function. The quantile loss function, important for quantile regression, estimates conditional quantiles of a response variable, contributing to a more complete view than traditional regression techniques that estimate only the conditional mean. By predicting different quantiles (the exact quantiles chosen were 0.1, 0.5, and 0.9, representing the lower bound, median, and upper bound of the predicted values, respectively), this method allows us to understand the distribution and variability of the target variable. That is especially suitable in the context of CO2 forecasting, where it can be beneficial to predict not just a single point estimate but a range of possible outcomes. This approach can well capture the uncertainties and fluctuations intrinsic to energy production and consumption data.

To optimize the CatBoost model for quantile regression, Optuna was employed to find the best hyperparameters. Optuna’s optimization procedure involved running multiple trials (300) to identify the optimal set of parameters that produced the lowest quantile loss, thereby enhancing the model’s predictive performance for the selected quantiles. The procedure also required to define an objective function using the pinball loss, which measures the accuracy of quantile predictions by tuning parameters such as the number of estimators, learning rate, maximum depth, and minimum samples split, aiming to minimize the quantile loss.

The Figure 4.9 depicts the behaviour of predicted quantiles (0.1, 0.5, 0.9) over a one-week period compared to actual values (based on **Input D – Combined FE**). The average width of the prediction intervals is 70.00, closely mirroring the standard deviation of the target variable at 64.49. This closer alignment suggests that the model’s prediction intervals are appropriately scaled to the variability observed in the data, indicating a refined balance between accommodating uncertainty and maintaining precision. The coverage of actual values between the 10% and 90% quantiles, was 82.50% during the training phase, indicating strong model performance. However, in the test set, as it is visible in the Figure 4.9, this coverage slightly decreased to 68.78%, suggesting a reduction in the model’s ability to predict the range accurately under unseen conditions.

The calibration and amplitude analysis shed light on the capability of the model’s predictive intervals. The Figure 4.10 shows the sum of absolute deviations for each quantile against actual observed values. The visible deviations for the 10th and 90th quantiles, represented by the blue and red bars respectively, suggest a potential misalignment in their calibration. This makes parallel with the referenced study on probabilistic electricity demand forecasts from Chen, Mohammad, Totsuka, Laliashvili, and Navarro (2024), which explores the trade-off between calibration and sharpness when using the pinball loss function. The pinball loss

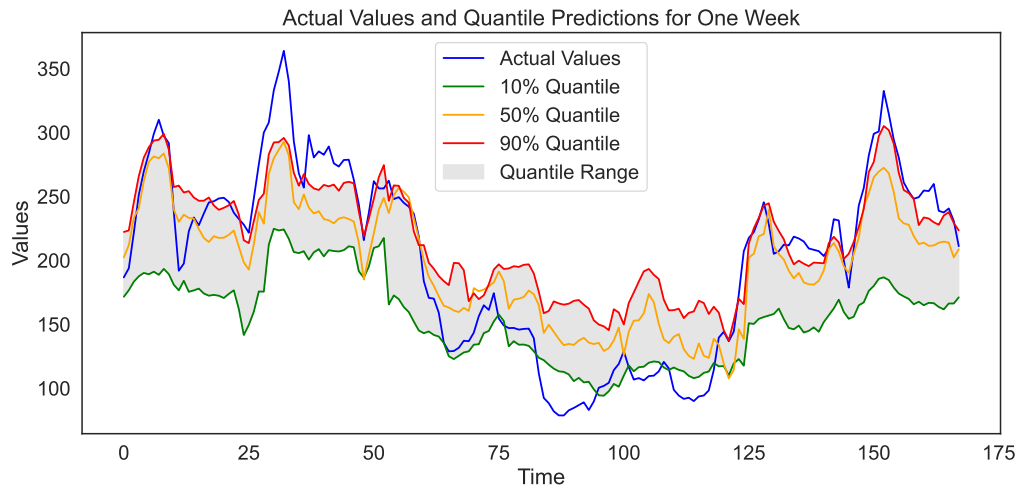


Figure 4.9: Actual values and predicted quantiles (10%, 50%, 90%) over one week, illustrating the quantile range and its alignment with actual outcomes

function, while effective for minimizing prediction errors, can sometimes prioritize sharpness over calibration.

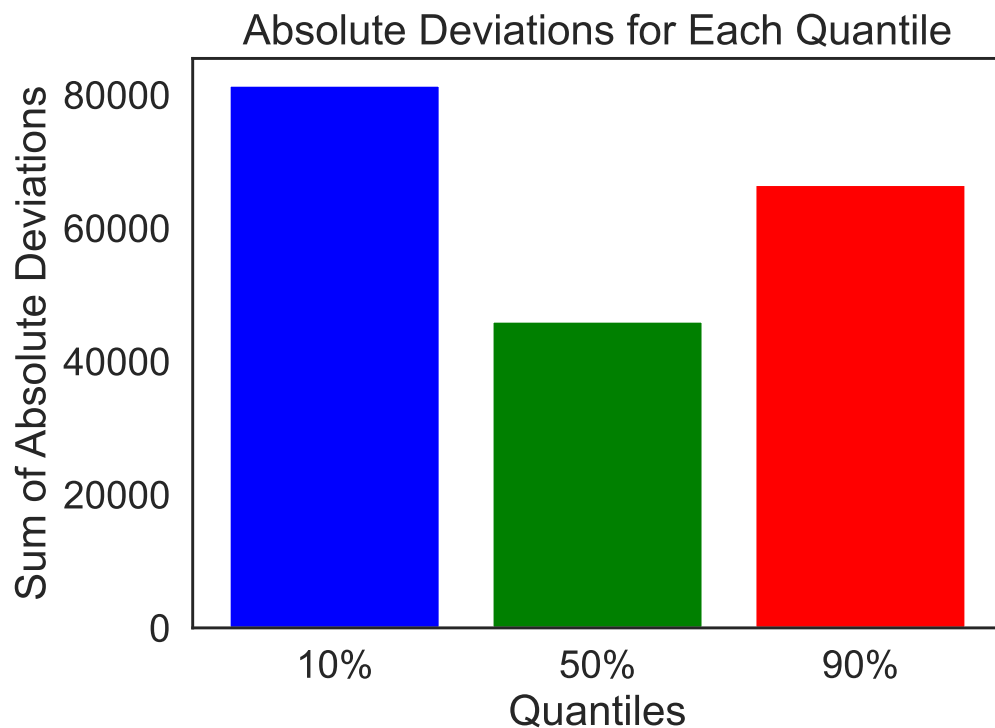


Figure 4.10: Sum of absolute deviations for the 10%, 50%, and 90% quantiles, comparing their deviation from actual observed values

4.3.3 Ensemble Models for Different Time Periods

In this segment, it was explored the performance of ensemble models constructed using CatBoost (best performing model) with evolutionary features (**Input D – Combined FE**) over different time periods. The goal is to assess whether incorporating data from varying temporal scopes (such as 6 months, 3 months, and 1 month) can improve the predictive accuracy of our model. The ensemble approach combines predictions from these different models to produce a stronger and reliable forecast.

Methodology

Initially, and as already explored and explained before, it was applied the EF technique to generate new features from the training dataset, selecting the top 50% most important features based on their importance scores. The CatBoost model was then trained on the complete dataset (1 year and 1 month) with an 80/20 train-test split. This attended as our baseline model for comparison. To examine the impact of using different temporal data scopes, we created subsets of the dataset for the last 6 months, 3 months, and 1 month. Each subset was split into an 80/20 train-test set, and the CatBoost model was trained using the evolutionary features resulting from the full dataset. Predictions from each of the time-period-specific models were averaged (simple average) to form an ensemble prediction. The logic behind this is that dissimilar time periods might capture various patterns and seasonal effects that a single time period might overlook.

Results and Discussion

The outcomes are shown in the Table 4.8 indicate that models trained on shorter, more recent data (such as the last 3 months) tend to perform better, with the lowest MAE and highest R^2 values. This indicates that the most recent data captures the most relevant patterns and trends for forecasting. However, the ensemble model, which combines predictions from models trained on different time periods, achieves a balance between capturing long-term trends and short-term variations.

Model	MAE	RMSE	R^2
Full Dataset (1 year 1 month)	24.50	30.87	0.77
Last 6 Months	15.47	23.99	0.86
Last 3 Months	9.72	17.90	0.92
Last 1 Month	21.71	31.38	0.76
Ensemble (Simple Average)	15.84	20.77	0.90
Blending Ensemble	15.59	20.50	0.90

Table 4.8: Performance metrics for the considered different time periods

The ensemble model's performance validates that integrating multiple temporal perspectives can lead to more accurate and reliable forecasts. The lower MAE and RMSE values, coupled with a high R^2 , indicate that the ensemble approach effectively mitigates the limitations of individual models trained on specific time periods. This ensemble method is especially useful in scenarios with volatile or seasonal data, such as energy consumption and production forecasting, where different time frames might reveal different aspects of the underlying

patterns.

The 3-month window strikes an optimal balance, capturing the essential seasonal and trend components long enough to capture short-term seasonal while maintaining the most relevant and up-to-date information for forecasting CO₂ levels. A 3-month period seems to be long enough to seize short-term seasonal variations and trends but short enough to avoid the noise and potential changes in underlying patterns that may arise over longer periods. This balance results in improved performance metrics, making it the most suitable time period for this dataset.

The comparison between the simple average ensemble and the blending ensemble determines the superiority of the latter. The blending ensemble achieved a lower MAE (15.59 vs. 15.84) and RMSE (20.50 vs. 20.77), while maintaining a comparable R^2 (0.90). The blending ensemble's better performance can be credited to its ability to leverage the strengths of individual models more effectively by using a meta-model to learn the optimal combination of predictions. This method captures the nuances and changes existing in different time periods.

Chapter 5

Conclusion

5.1 Final Remarks

The application of the EF technique in this study has generated some insights into improving carbon intensity forecasting for the Portuguese electric power system. This study's principal objective was to evaluate the effectiveness of the EF in seizing complex, non-linear relationships in order to improve traditional ML models. The presented results provide a positive indication that the EF is particularly effective in producing sophisticated feature transformations that might contribute to improvements in predictive accuracy.

5.1.1 Key Insights and Implications

Boosted Feature Engineering: In respect to our case study, the EF demonstrated its capacity to create complex associations and high-level features that were not captured by traditional methods. These advanced features were particularly beneficial for ensemble models such as RF, GBDT, and CatBoost. This improvement is attributed to EF's use of GP to simulate evolution, where potential solutions (features) are iteratively improved through operations analogous to natural selection, mutation, and crossover. This approach permits the algorithm to effectively explore complex feature interactions, resulting in a more informative feature set that significantly improves model learning by uncovering patterns that would otherwise be difficult to detect with traditional feature engineering methods.

Complementarity with INESC TEC Features: There is a synergy between INESC TEC's engineered features and those generated by the EF, which plays an important role in improving model performance. INESC TEC's features provide a solid basis by capturing essential linear, non-linear, and temporal dependencies within the dataset, applying mathematical transformations such as polynomial expansions, lagged variables, and sinusoidal functions to encapsulate cyclical patterns and historical trends. In contrast, the richness of the EF is in its ability to explore the feature space beyond human intuition, using GP to iteratively refine and combine these initial features, applying operations such as addition, subtraction, multiplication, and division to expose higher-order interactions and new patterns. Ultimately, the choice between these approaches depends on the specific problem context, data characteristics, and the balance between interpretability and exploratory power needed for the modeling task. Combining both methods, as seen in our case study, seems to leverage the strengths of

each.

Model-Specific Performance Variability: The application of evolutionary feature engineering techniques, such as those obtained by the EF, produced varied performance enhancements across different model clusters. Tree-based models and boosting models exhibited substantial improvements due to their inherent effective ability to leverage enriched feature sets. Hybrid tree-based boosting models, like LightGBM and DART, however, showed a slight increase in their performance. These models are intrinsically designed to handle complex features, potentially decreasing the incremental benefits from additional features introduced by the EF. Linear models, such as Linear Regression and LASSO, faced a deterioration in performance that suggests that the non-linear feature transformations accentuated by the EF were not compatible with the linear assumptions of these models. ANN displayed 6.34% decrease in MAE, demonstrating their adaptability to complex feature transformations due to their non-linear learning competencies. SVM encountered an increase in MAE by 6.17%, indicating a misalignment between SVM's linear boundary requirements and the non-linear transformations introduced by the EF. A fundamental element in enhancing model performance was the strategic use of hyperparameter tuning with Optuna. Employing techniques like the Tree-structured Parzen Estimator (TPE), Optuna resourcefully explored the hyperparameter space and adapted to the evolving feature landscape, optimizing configurations based on performance metrics (MAE). Optuna identified optimal settings that aligned with the enriched feature sets from the EF, allowing models to reach their peak potential.

Feature Importance and Interpretability: The analysis of feature importance exposed the importance of advanced features in improving model interpretability and predictive power. The EF-generated features consistently ranked high in importance across various models, demonstrating their impact on forecasting accuracy. Picking the top three performing models (CatBoost, ET, and GBDT) the feature importance analysis showed that nine out of the top ten features were created by the EF. This is representative of the algorithm's capability to contribute to a deeper understanding of carbon emissions dynamics.

CatBoost as the Top Performer: CatBoost emerged as the top-performing model for CO₂ forecasting in the Portuguese power system, consistently achieving the lowest MAE across various datasets. Statistical validation confirmed CatBoost's ranking as the best model, indicative of its adaptability to enriched feature sets like those introduced by the EF. Its superior performance is attributed to its effective handling of time-related features such as date, month, day, and hour. CatBoost excels at seizing the temporal patterns and dependencies inherent in these features, allowing it to accurately model cyclical behaviors and temporal swings that are decisive for precise CO₂ forecasting. In addition to its ordinary performance, CatBoost's implementation with the quantile loss function further improved its fitness to predict a range of outcomes. This probabilistic approach delivers a more inclusive view of potential variations in CO₂ levels, which is particularly important in the context of fluctuating energy production and consumption. Moreover, the ensemble models' analysis discovered that training CatBoost on a 3-month data range was the most effective, striking a balance between catching short-term trends and avoiding long-term noise. Ensemble approaches, such as averaging or blending, may offer even greater improvements by combining insights from CatBoost models trained on diverse time frames, such as 1 month, 3 months, and 6 months.

This approach leverages the temporal diversity of the training data, allowing the ensemble to capture a wider range of trends and seasonal variations for greater predictive performance.

In a nutshell, the EF technique is a powerful tool for enhancing feature engineering in carbon intensity forecasting, as it generates complex, high-level features that meaningfully boost model performance. This study determines the importance of combining traditional and advanced feature engineering methods to cover the full spectrum of data dynamics.

5.2 Limitations and Future Work

The case study on the EF technique for carbon intensity forecasting in the Portuguese power system discourses some limitations and areas for improvement. Firstly, the data availability and quality posed for potential important challenges, with existing datasets often missing sufficient granularity and consistency. This could potentially influence the accuracy of the models developed. The feature engineering procedure, while sophisticated, risked increasing model complexity and overfitting due to the introduction of complex features that might capture noise rather than meaningful patterns. The models used, such as CatBoost and EF, balance between performance and interpretability. While EF offers improved interpretability through GP, the complexity of relationships it captures can pose difficulties in fully understanding every feature transformation. CatBoost, on the other hand, with its sophisticated feature treatment, frequently functions as a black-box, limiting interpretability compared to simpler models.

As well, the choice of temporal scope, although offering some insights, may not adequately capture long-term tendencies due to the focus on specific time windows. This could lead to the omission of pertinent seasonal patterns that occur over extended periods. The computational cost related with techniques like GP and Optuna for hyperparameter tuning is also a concerning topic, as it requires significant resources and time, creating a barrier to real-time applications.

To address these limitations, numerous potential improvements and areas for future work are proposed. Improving data collection through diverse sources like IoT sensors and smart grid data could improve data quality and deliver deeper insights into energy patterns. Advanced feature selection techniques should be executed to ensure that only the most informative features are retained, in that way reducing overfitting risks. Likewise, integrating interpretable ML methods such as SHapley Additive exPlanation (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) can improve model transparency, making the predictions further accessible to stakeholders.

A dynamic threshold optimization strategy could also be implemented, allowing for more polished feature selection tailored to each model's needs. This could be accomplished by integrating threshold selection with hyperparameter tuning using tools like Optuna. The study of DL models, such as Recurrent Neural Networks or transformers, presents an exciting chance to capture long-term dependencies and temporal dynamics that traditional models may fail to. These models could be combined with EF-generated features to measure their impact

on performance.

Interdisciplinary collaboration with experts in energy systems and energy strategic policy can enhance the practical applicability of the EF technique. By ensuring that features correspond with real-world contexts and conditions, the models may become more relevant and effective in addressing carbon intensity challenges. Furthermore, testing the proposed methodology across different regions (such as various countries or specific smaller areas in Portugal) could evaluate its adaptability and robustness, potentially bringing new insights that could improve predictions.

In conclusion, while the EF technique has exposed promise in enhancing feature engineering for carbon intensity forecasting, it is indispensable to address the identified limitations and explore the suggested enhancements and future work. This will safeguard the development of more accurate, interpretable, and generalizable models, ultimately contributing to more sustainable energy system in Portugal.

Bibliography

- Acheampong, A. O., & Boateng, E. B. (2019). Modelling carbon emission intensity: Application of artificial neural network. *Journal of Cleaner Production*, 225, 833–856.
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*.
- Ashok, S. K. (2020). *Development of a web-based application for analysis and visualization of co2 emissions of european power systems* (Unpublished master's thesis).
- Bhattacharya, M., Inekwe, J. N., & Sadorsky, P. (2020). Consumption-based and territory-based carbon emissions intensity: determinants and forecasting using club convergence across countries. *Energy Economics*, 86, 104632.
- Bokde, N. D., Tranberg, B., & Andresen, G. B. (2021). Short-term co2 emissions forecasting based on decomposition approaches and its impact on electricity market scheduling. *Applied Energy*, 281, 116061.
- Breiman, L. (1996). Bagging predictors. *Machine learning*, 24, 123–140.
- Breiman, L. (2001). Random forests. *Machine learning*, 45, 5–32.
- Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on machine learning* (p. 18).
- Chen, H. H., Mohammad, A., Totsuka, T., Laliashvili, A., & Navarro, M. (2024, 06). Evaluation of probabilistic electricity demand forecasts.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20, 273–297.
- Delanoë, P., Tchuente, D., & Colin, G. (2023). Method and evaluations of the effective gain of artificial intelligence models for reducing co2 emissions. *Journal of Environmental Management*, 331, 117261.
- Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems* (pp. 1–15).
- Dorogush, A. V., Ershov, V., & Gulin, A. (2018). Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*.
- Duras, T., Javed, F., Månsson, K., Sjölander, P., & Söderberg, M. (2023). Using machine learning to select variables in data envelopment analysis: Simulations and application using electricity distribution data. *Energy Economics*, 120, 106621.
- Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119–139.
- García-Martos, C., Rodríguez, J., & Sánchez, M. J. (2013). Modelling and forecasting fossil fuels, co2 and electricity prices and their volatilities. *Applied Energy*, 101, 363–375.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- Hastie, T. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

- Herbold, S. (2020). Autorank: A python package for automated ranking of classifiers. *Journal of Open Source Software*, 5(48), 2173. Retrieved from <https://doi.org/10.21105/joss.02173> doi: 10.21105/joss.02173
- Koza, J. R. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4, 87–112.
- Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC.
- Leerbeck, K., Bacher, P., Junker, R. G., Goranović, G., Corradi, O., Ebrahimi, R., ... Madsen, H. (2020). Short-term forecasting of co2 emission intensity in power grids by machine learning. *Applied Energy*, 277, 115527.
- Lowry, G. (2018). Day-ahead forecasting of grid carbon intensity in support of heating, ventilation and air-conditioning plant demand response decision-making to reduce carbon emissions. *Building Services Engineering Research and Technology*, 39(6), 749–760.
- Maji, D., Shenoy, P., & Sitaraman, R. K. (2022). Carboncast: multi-day forecasting of grid carbon intensity. In *Proceedings of the 9th acm international conference on systems for energy-efficient buildings, cities, and transportation* (pp. 198–207).
- Maji, D., Sitaraman, R. K., & Shenoy, P. (2022). Dacf: day-ahead carbon intensity forecasting of power grids using machine learning. In *Proceedings of the thirteenth acm international conference on future energy systems* (pp. 188–192).
- Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective*. Routledge.
- Qader, M. R., Khan, S., Kamal, M., Usman, M., & Haseeb, M. (2021). Forecasting carbon emissions due to electricity power generation in bahrain. *Environmental Science and Pollution Research*, 1–12.
- Seber, G. A., & Lee, A. J. (2012). *Linear regression analysis*. John Wiley & Sons.
- Smith, M. G., & Bull, L. (2005). Genetic programming with a genetic algorithm for feature construction and selection. *Genetic Programming and Evolvable Machines*, 6, 265–281.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288.
- Tuaimah, F. M., & Abass, H. M. A. (2014). Short-term electrical load forecasting for iraqi power system based on multiple linear regression method. *International Journal of Computer Applications*, 100(1).
- Vinayak, R. K., & Gilad-Bachrach, R. (2015). Dart: Dropouts meet multiple additive regression trees. In *Artificial intelligence and statistics* (pp. 489–497).
- Wang, Y., Wang, J., Zhao, G., & Dong, Y. (2012). Application of residual modification approach in seasonal arima for electricity demand forecasting: A case study of china. *Energy Policy*, 48, 284–294.
- Wang, Z.-X., & Ye, D.-J. (2017). Forecasting chinese carbon emissions from fossil energy consumption using non-linear grey multivariable models. *Journal of Cleaner Production*, 142, 600–612.
- Watanabe, S. (2023). Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*.
- Weron, R. (2006). *Modeling and forecasting electricity loads and prices: A statistical approach* (Vol. 396). John Wiley & Sons.
- Zhang, H., Zhou, A., & Zhang, H. (2021). An evolutionary forest for regression. *IEEE Transactions on Evolutionary Computation*, 26(4), 735–749.

Zhou, Z.-H. (2012). *Ensemble methods: foundations and algorithms*. CRC press.

