Towards Assessing the Electricity Demand in Brazil: Data-Driven Analysis and Ensemble Learning Models

João Vitor Leme [1], Wallace Casaca [1,2,*] and **Marilaine Colnago [1]** and **Maurício Araújo Dias [3]**

1   Department of Energy Engineering, São Paulo State University (UNESP), Rosana, SP 19274-000, Brazil; joao.leme@unesp.br (J.V.L.); marilaine.colnago@unesp.br (M.C.)
2   Center of Mathematical Sciences Applied to Industry (CeMEAI), São Carlos, SP 13566-590, Brazil
3   Faculty of Science and Technology (FCT), São Paulo State University (UNESP), Presidente Prudente, SP 19060-900, Brazil; ma.dias@unesp.br
*   Correspondence: wallace.casaca@unesp.br

Abstract: The prediction of electricity generation is one of the most important tasks in the management of modern energy systems. Improving the assertiveness of this prediction can support government agencies, electric companies, and power suppliers in minimizing the electricity cost to the end consumer. In this study, the problem of forecasting the energy demand in the Brazilian Interconnected Power Grid was addressed, by gathering different energy-related datasets taken from public Brazilian agencies into a unified and open database, used to tune three machine learning models. In contrast to several works in the Brazilian context, which provide only annual/monthly load estimations, the learning approaches Random Forest, Gradient Boosting, and Support Vector Machines were trained and optimized as new ensemble-based predictors with parameter tuning to reach accurate daily/monthly forecasts. Moreover, a detailed and in-depth exploration of energy-related data as obtained from the Brazilian power grid is also given. As shown in the validation study, the tuned predictors were effective in producing very small forecasting errors under different evaluation scenarios.

Keywords: energy forecasting; data-driven analysis; machine learning; Brazilian power grid

1. Introduction

Electricity is part of a composite market that involves generation, transmission, and consumption agents. Such a free market has become highly competitive in recent years, leveraging the participation of several investors, electric companies, and public agencies [1]. Once these stakeholders seek to maximize their profits while minimizing their expenses, a suitable prediction of energy generation has been mandatory for those who interact in this business, especially because of the competitive electricity market, influenced by the supply and demand conditions. Moreover, as most system operational decisions occur as a response to data gathered and processed at the control center, the use of data-driven platforms are crucial to get useful information and make intelligent choices [2]. These data-guided frameworks are particularly important in the Brazilian context—as the goal of our work—since the national power grid is currently operated by a general grid operator that arbitrates when and how much each power plant will produce from official computer models [3]. As the Brazilian electricity matrix is mainly composed by renewable power sources, which vary in nature [2], the electricity market prices may reflect the stochastic behavior of these sources. This means that, in most cases, the wholesale market prices in Brazil are determined by the opportunity costs of these renewable power plants, based on the acquired data and supply and demand tendencies [4].

Techniques devoted to forecasting electricity demand aim at estimating the amount of energy needed, over a historical time series, for transmission and later consumption by others. Despite the adaptation of several machine learning models to properly address the problem [5–9], tracking the progressive use of the electricity is not a straightforward task in practice. In fact, the electricity dispatch is intrinsically related to the internal operations of the power systems such as the periodic scheduling of power generation in hydroelectric plants, the preventive maintenance of the generators, the reliability evaluation of power systems, etc. [10]. Moreover, the problem becomes even more challenging and especially interesting when one has to deal with the highly nonlinear tendency of power data, as it is mathematically modeled by highly-oscillating time series whose parameters can be affected by exogenous variables such as weather/ambient conditions [11] and economy-related factors [12].

Formally, the problem of predicting the power demand on time series can be described as follows: given a time series of electric load $X_1, X_2, \ldots, X_t$, in which X_i accounts for the historical energy load at the instant i, $i = 1, 2, ..., t$, the goal is to predict the quantity $X_{(t+h)}$, where h establishes the forecast horizon [13,14]. Taxonomically speaking, this kind of prediction usually comprises three categories of planning horizons: (i) long-term (years/months); (ii) regular-term (days/weeks); and (iii) short-term (minutes/hours). Since estimating the electricity demand becomes harder as the planning horizon increases, the predictions can be strongly influenced by several nonuniform variables such as electric consumption, temperature, air humidity, and socioeconomic aspects. Moreover, long- and regular-term time series make the problem more difficult to be technically managed and solved, as obtaining a computationally robust solution to act in real scenarios requires the integration of customized tuning approaches and non-linear models as a unified framework to properly work [15–19]. Therefore, in this paper, our main interest lies in designing well-behaved forecasters to assess and predict the electricity demand in Brazil for both long- and regular-term time series.

Considering the recent advances in Machine Learning (ML) for electricity load forecasting, the literature offers a variety of approaches, most of them specifically designed to solve a particular case study of energy consumption. For instance, Qiu [20] proposed a Support Vector Regression (SVR) variant based on Particle Swarm Optimization (PSO) to forecast the energy trade of the Taiwan electricity market. Hybrid methods based on well-established ML models such as random forest, neural networks, and fuzzy logic adjusted to gauge electric consumption were also given in [21], where the authors monitored the amount of energy consumed in buildings located at the Polytechnic University of Catalonia, in Spain. Hernandez et al. [22] described a hybrid decision-making tool to analyze and inspect the energy consumption in an industrial park, also in Spain. They combined self-organizing maps and k-means clustering into a cascade-based application to supervise the power consumption flow in the evaluated industrial park.

In contrast to the above study cases, covering larger territories takes a lot of data from different resources to produce meaningful results. As a consequence, studies dedicated to investigating larger areas such as full countries only have contributed to the minority of literature in electricity demand. This was the case faced by Zhao et al. [23], as their method integrates a *grey* forecasting model with parameters optimization to assess the electricity consumption in Mongolia. Zhao's method was later modified by Liang and Liang [24] to cope with the electricity demand of China between 2016 and 2020. Their method divides the forecasting task into two steps: two prior single predictions, and the combination of both forecasts as a new, more accurate, one to produce the definitive estimation. A rich discussion comprising the energy demand for different sectors in U.S. was presented by Ameyaw and Yao [25], through a Recurrent Neural Network (RNN) designed as an assumption-free-based predictive model for regular-term predictions. RNN was also the learning strategy used by Bouktif et al. [26], who faced the forecasting problem by applying an RNN architecture, namely Long Short-Term Memory (LSTM), to predict the electric load in France.

Finally, concerning the core works devoted to covering the electricity demand in Brazil—as the goal of our work—contributions were made towards assessing the energy consumption, but assuming

as data source only global indicators such as Gross Domestic Product (GDP) and Population Growth Rate (PGR). Maçaira et al. [27] presented different projections for the Brazilian energy consumption taking the GDP—as an independent variable—and a dynamic regression approach. Despite predicting consumption until 2050, their forecasts were only performed annually, giving no details for the demanded energy during the months. Another long-term forecasting study was conducted by Torrini et al. [28]. The authors used a fuzzy logic-based methodology, which was calibrated with GDP and PGR indices, and compared their results with official projections for the sector, as provided by the Brazilian Energy Research Office (EPE - Empresa de Pesquisa Energética, in Portuguese: http://www.epe.gov.br/en). Similarly, a year-by-year estimation for the electric demand in Brazil was also carried out by Trotter et al. [29], by modeling uncertainty in the estimates of weather variables. Their approach relies on basic features such as population size and national income together with the electricity demand so that a multiple linear regression model is obtained to yield annual forecasts.

Other studies have also been published concerning the electric demand in Brazil, most of them focused on annual estimations [30–32], wherein no conclusive forecasts can be delivered if one intends to predict the daily/monthly electricity load in a more detailed level of resolution. Moreover, the literature lacks more comprehensive studies which consider a wider range of data resources to go deeper into what kind of information is really a good match for energy demand predictability in Brazil. As motivation, to cite a few works that can perform robust data exploration while improving the predictability of ML approaches for energy demands in other countries, one may consider Dai et al.'s work [33], which exploited the problem under the perspective of how to apply ensemble models to improve Support Vector Machines (SVM) to get the energy consumption in China. Another interesting study was conducted by Utterbäck [34], whose goal focused on answering how weather data geographically vary in Scandinavia, and whether geographical properties are useful to give relevant information to the predictors. Assessing the impact of weather conditions in the forecasting task was also the goal of Zhang et al. [35], but in the sense of how photovoltaic power systems can be drastically affected due to abrupt weather changes that occur during a whole day. Facing a similar problem, Ceci et al. [36] applied entropy-based metrics for online training of artificial neural networks to better exploit the non-linear dependencies between the feature space (weather conditions) and the target space (observed power production). Finally, feature space analysis was the key point handled by Sarhani and Afia [37] for load forecast, based on the combination of PSO and feature selection to improve their SVM variant, and by Liang et al. [38], by means of a hybrid model which integrates several tuning strategies such as empirical mode decomposition and minimal redundancy maximal relevance into a regression neural network to produce the forecasts.

Contributions

In this study, the problem of predicting the electricity demand in Brazil is addressed for both long- and regular-term time series. Different strategies to optimize the performance and accuracy of the presented ML approaches are discussed in details so as to promote a comprehensive analysis of the explored data while still elucidating how electricity load predictions can be achieved and driven by data exploration and ensemble-based learning models. In contrast to other works that only provide annual forecasting assessments for the Brazilian electricity demand, this paper establishes a solid methodological pipeline for daily/monthly forecasts, by introducing several variables related to the national electrical system instead of employing macro-indices, as previously discussed. As accurateness in predicting the electric demand depends on the amount of available data and how to properly handle the data to build well-behaved predictive models, a new database composed by two national (and official) data repositories in Brazil is also given and discussed, thus filling the gap with respect to the absence of a comprehensive and reliable dataset in the Brazilian context.

This paper is organized as follows. Section 2 introduces the data analysis apparatus, learning pipelines, and the datasets utilized in our investigation, while Section 3 gives the results, main findings, and their discussion. Finally, Section 4 summaries the conclusion of our research.

2. Materials and Methods

2.1. Data Resources and Feature Description

Aiming at exploiting the energy-related data publicly available by Brazilian government agencies, two specific data repositories were analyzed and then combined into a unique database, named here as *Brazilian Interconnected Power Grid Dataset* (BRAIPG). The first data resource was obtained from the National Electrical System Operator (ONS) [39], which is the core body responsible for coordinating and controlling the operations of electricity generation and transmission in the Brazilian Interconnected Power Grid (SIN). The ONS agency is under the supervision and regulation of the National Electric Energy Agency (ANEEL) [40]. The interconnected grid of the Brazilian electrical system allows the achievement of synergistic gains, as it takes into account the diversity of the hydrological regimes of the country's basins [41]. For a better readability, Table 1 lists the abbreviations used in this section.

Table 1. List of abbreviations for Brazilian agencies.

Acronyms	Description (Portuguese)	Description (English)
ONS	Operador Nacional do Sistema Elétrico	National Electrical System Operator
SIN	Sistema Interligado Nacional	Brazilian Interconnected Power Grid
ANEEL	Agência Nacional de Energia Elétrica	National Electric Energy Agency
INMET	Instituto Nacional de Meteorologia	National Institute of Meteorology
MAPA	Ministério da Agricultura, Pecuária e Abastecimento	Ministry of Agriculture, Livestock and Supply

ONS has developed a new way of making available historical results stemmed from the SIN, based on the principle of transparency and information reliability. The data are tabulated in daily and hourly time, by subsystems (Brazilian's regions), or total demanded. In particular, eleven of fifteen features available by the ONS for the years ranging from 2005 to 2018 were collected at national level, stored, and then analyzed in a daily format (see Table 2). Notice that our choice in selecting variables concerning water generation is due to the fact that Brazil has an electrical system in which hydroelectric plants predominate, i.e., more than 60% of the national electricity generation comes from water bodies [41].

Table 2. Main set of collected data (energy-related features).

Feature	Description	Unity
Energy Load	Energy charge consumption	GWh
Max Demand	Electric load demand peak	GWh
Stored Energy	Amount of stored energy	GWh
Generated Energy	Total of all energy resources	GWh
Border Power	Power flow transmission at Brazil's borders	GWh
Date	The day, month, and year	-
Influent Flow	Influent flow in all water reservoirs	m^3/s
Water Flow	Water flow in all hydroelectric	m^3/s
Total Volume	Total volume available in the water reservoirs	%
Hydroelectric Gen.	Total generated by hydroelectric	GWh
Hydroelectric Gen. SE/CO	Total generated by hydroelectric plants in SE/CO regions	GWh

As mentioned above, a secondary data repository was also considered in our analysis: the weather-related information, taken from the National Institute of Meteorology (INMET) [42]. The INMET agency is a federal branch under the supervision of Ministry of Agriculture, Livestock and Supply (MAPA), which provides reliable information regarding the Brazilian climate and meteorology conditions. Although the constituted database intends to cover the whole country whose data can significantly range due to the large territorial extension and regional diversity of Brazil, only the most representative states in terms of energy dispatch were selected in this study, more precisely: *Bahia*, *Goiás*, *Mato Grosso do Sul*, *Minas Gerais*, *Paraná*, *Rio Grande do Sul*, *São Paulo*, and *Tocantins*. Indeed, these

states concentrate high levels of agricultural activities, industrial production, and high capacity of electric generation and transmission (see Figure 1).
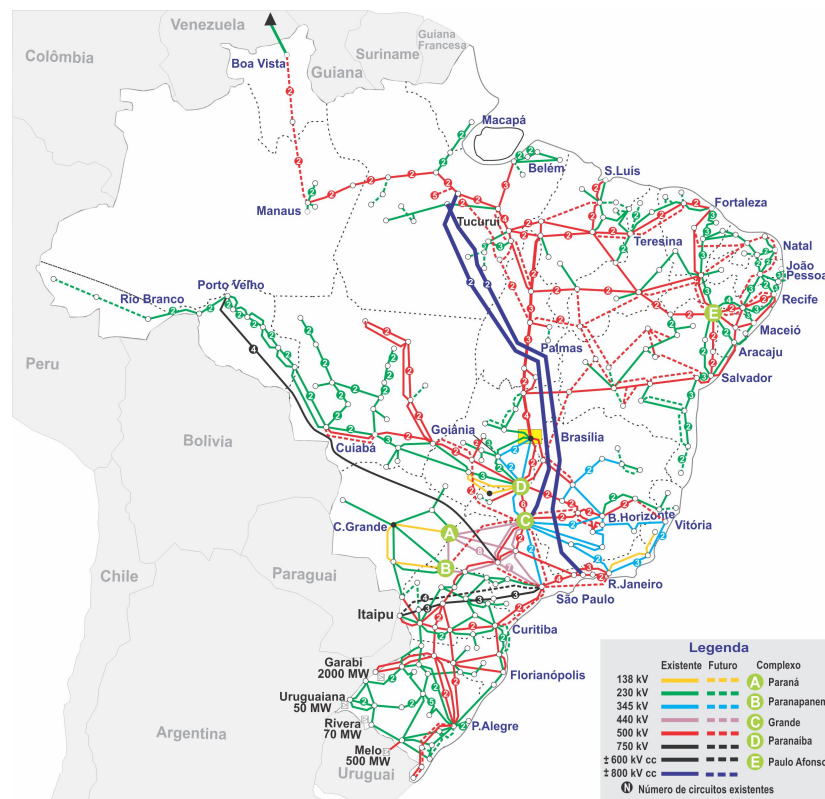

Figure 1. Map of the National Interconnected System with main transmission trunks [43].

Therefore, to enrich our data collection with the features from INMET while keeping the number of variables feasible to work with, the following weather measurements were acquired for each state capital according to Table 3.

Table 3. Additional set of collected data (other features).

Feature	Description	Unity
Average Temperature	Average temperature of all states considered	°C
Average Relative Humidity	Average relative humidity of all states considered	%
Average Wind Speed	Average wind speed of all states considered	m/s

2.2. Predictive Models for Energy Demand Forecasting in Brazil

This section describes the ML models, parameters tuning, and the learning strategies used to forecast the electricity load in the Brazilian Interconnected Power Grid. Three nonlinear learning models were customized and then applied to yield the predictions: *Random Forest*, *Gradient Boosting*, and *Support Vector Machines*. All of them were adjusted to perform regression. Notice that the use of three distinct learning approaches to cope with the investigated problem is endorsed by Zhou's conjecture [44]. In fact, he postulated that, in real problems, there is no definitive ML model which optimally solves the majority of the problems in a given solution space. As a result, each model may reach different performances and accuracies when applied on a particular dataset, thus leading to the use of various ML-based pipelines, especially in the context of time series, as addressed in this work.

2.2.1. A Random Forest-Based Ensemble Model

Random Forest (RF) can be seen as an ensemble method, that is, a set of estimators which induces the creation of its own learners and decision rules, wherein the primary learners are all classification/regression trees (CARTs) [45]. In general, RF enables the use of two tuning strategies: *Bagging* and the *Random Subspace Technique*. The main difference between the random subspace method and bagging is that, on a given node, instead of using all variables, RF takes only a random subset to pick up the variables in the division criteria. Moreover, it was observed that such a randomization procedure allows for reducing the correlation between the regression trees so that the forecasting performance can be substantially improved [46,47].

The RF-inspired pipeline as implemented in this work relies on three core steps:

- Generate *n bootstrap* sample sets from the training set (as discussed below).
- For each *bootstrap* sample, compute an unadjusted regression tree with the following modification: on each node, generate a random sample *p* for the input variables from the training set, and choose the best division for these variables, with $p < m$, where *m* represents the number of variables as listed in our dataset.
- Predict the new output, by averaging the outputs of *n* regression trees when new variables are entered into the model.

Mathematically, the obtained RF-based model is defined as a collection of random predictor trees: $H_\gamma = \{h(x, \theta_n, \gamma) : n = 1, ..., N\}$, where $h(x, \theta_n, \gamma)$ denotes the *n*th random predictor tree, i.e., it forms the basis for the general forecaster $h(x)$. It is important to highlight that the predictors are independent and randomly distributed in quantities, allowing to embed the randomization process into the decision trees, thus improving the end predictor. Notice also that θ_n is selected before tree growth, and it is independent of the learning data, γ.

Finally, predictor trees are then combined so as to generate the following definitive forest-based estimator of $h(x)$:

$$h(x, \theta_1, ..., \theta_N, \gamma) = \frac{1}{N} \sum_{n=1}^{N} h(x, \theta_n, \gamma). \tag{1}$$

2.2.2. A Gradient Boosting-Based Ensemble Model

As RF, the Gradient Boosting (GB) can also be formulated as a tree-based ensemble regression method. Basically, it employs the *boosting strategy* instead of bagging. This learning strategy is an improvement of the bagging, which consists in training several submodels with random samples and combining them for a less "individualized" performance, hence diminishing overfitting. Moreover, in the GB ensemble model, the constituent trees do not have the same weights in the voting, i.e., the goal of GB is to find the optimal tree combination constrained to customized weights [48,49].

The GB ensemble model creates a prior regression tree and applies the stochastic variant of the descending gradient method to optimize the trees during the iterations, according to a cost function. Formally, the output of the GB is given in terms of the sum of tree estimates [50,51]. The classic principle of this kind of formulation involves the analysis of two key factors [52]:

- Weak predictive models for making weak learner predictions (e.g., decision trees).
- An additional model of weak learners to minimize the cost function.

Considering its mathematical formulation [53], the GB takes the following additive form:

$$f(x) = \sum_{n=1}^{N} \gamma_n h_n(x), \tag{2}$$

where $h_n(x)$ represent the basis functions, i.e., they are commonly viewed as weak learners. Notice that the GB takes decision trees of a fixed size as weak learners. Decision trees bear several attractive

properties which make them quite effective for boosting, as the capability of modeling complex functions and handling data of mixed types. In practice, an additive model is built in a greedy fashion:

$$f_n(x) = f_{n-1}(x) + \gamma_n \, h_n(x), \tag{3}$$

where the newly added tree $h_n(x)$ aims to minimize the loss L, given the previous ensemble $f_{n-1}(x)$ and the actual value y_i of the time series:

$$h_n = \min \sum_{i=1}^{n} L(y_i, f_{n-1}(x_i) + h(x_i)). \tag{4}$$

Finally, Equation (4) is numerically solved by the method of steepest descent, given a differentiable loss function L (for numerical details, see [53]).

2.2.3. Support Vector Regression Model

Support Vector Machine (SVM) is an ML approach that relies on decision surfaces to segregate instances of different classes. The method generates an optimal hyperplane which maximizes the margin, i.e., the distance between support vectors of different classes. These vectors received such a denomination because of their proximity to the decision surfaces, contributing in a decisive way to the definition of the surfaces. It is worth mentioning that, to create a decision surface for problems that cannot be separated linearly, SVM makes use of kernel functions, which enable the method to perform linear cuts in the attribute space in order to create larger dimensions wherein the classes are separated by a hyperplane [54,55].

The SVM model can be easily reformulated to tackle regression problems, by managing a loss function, minimized by means of a regularizer. Indeed, when SVM assumes its regression form (e.g., to perform predictions in time series), it is commonly called SVR (Support Vector Regression). In our study, SVR was adapted and applied on the Brazilian electricity demand context, wherein the forecasting issue is modeled so as to determine a nonlinear function f which minimizes the prediction error from the training set. First, the input x is mapped onto a n-dimensional feature space using some fixed (nonlinear) mapping, and then a linear model is built in this space. Following SVR [56], function $f(x, y)$, which controls the prediction error in the feature space, can be written as:

$$f(x, y) = \sum_{i=1}^{n} y_i g_i(x) + \varepsilon, \tag{5}$$

where $g_i(x)$ denotes a set of nonlinear transformations, while parameter ε regulates the error [47].

2.3. Tuning Hyperparameters with Random Search Strategy

One of the main drawbacks of ML methods is that they have several hyperparameters to be adjusted so that the definitive predictor can reach better data fitting and adherence. Another complication is that the number of valid hyperparameter settings can grow exponentially in practice [57,58].

To address these issues while still making our RF and GB models more assertive, optimal hyperparameters were computed by applying the Random Search strategy, whose the selected values and the tuning universe are shown in Table 4. In contrast to the standard ML approaches that usually apply grid search-based schemes to test all the possibilities to find out the best learning pipeline, the Random Search verifies a much smaller number of cases to get the definitive pipeline [59], resulting in a more effective scheme for parameter exploration. In the Random Search scheme, a fixed number of iterations is defined, together with the candidates for the pipelines (tuning universe) with or without a pre-specified selection criterion. As a result, the best pipeline constrained to the tuning universe is computed more quickly than other search schemes.

Table 4. Optimal hyperparameters (last two columns) and their search spaces for the RF and GB ensemble models.

Hyperparameter	Description	Tuning Universe	RF	GB
max_features	Number of feats. when searching for the best division	auto, sqrt, log2, none	none	none
max_depth	Maximum depth of the tree	2, 3, 5, 10, 15	15	5
min_samples_leaf	Minimum number of samples for a leaf node	1, 2, 4, 6, 8	2	8
min_samples_split	Minimum number of samples to split an internal node	2, 4, 6, 10	2	8
N_estimators	Number of trees to be generated	100, 500, 900, 1100, 1500	900	500
loss (only for GB)	Loss function to be optimized	ls, lad, huber	-	huber

As the implemented SVR-based forecaster is not precisely an ensemble method in essence, its hyperparameters were chosen empirically from the following space of parameters (Table 5), similar to the one given in [60].

Table 5. Optimal hyperparameters (last column) and their search spaces for the SVR model.

Hyperparameter	Description	Tuning Universe	SVR
kernel	Kernel function to be applied	rbf, linear, poly	rbf
degree	Kernel degree (for kernel = poly)	1, 2, 3, 4, 5	1
C	Error penalty parameter	1000, 2000, 5000, 10,000, 15,000	10,000

Considering the optimal settings as computed by the Random Search strategy for the trained models (Tables 4 and 5), it can be observed that, for both RF and GB, *none* has been considered optimal for the *max_features* attribute, i.e., without any kind of maximum restriction. Concerning the loss function, the Random Search found *huber* as the best choice, which is basically a combination of "ls" (Least Squares) and "lad" (Least Absolute Deviation). For the SVR model, the kernel selected was *rbf*, which is the same as the default kernel. The only parameter that changed considerably was C (10,000), i.e., the penalty criterion, since by default it assumes the unit constant.

2.4. Improving the Learning Task by Resource Engineering

To improve the predictability of the learning models, new variables were created from the existing ones and then incorporated into our data collection. More specifically, three new categories of variables were generated and introduced as new features to be handled by our forecasting models: the *log returns* [61] on the daily horizon; the *simple* and *exponential moving averages* [62] for both the daily and weekly horizons; and the *Dummy Coding* [63], to transform the day and week into vector features—as part of a vectorization process—totalizing three additional *continuous features* and seventeen *categorical features*, used by our learning approaches to increase accurateness.

The rationale behind the utilization of new variables is that if there are no sufficient discriminative variables to train the models, they may not accomplish the forecast task satisfactorily. On the other hand, if there are too many variables, or if most of them carry irrelevant/duplicate information, the models tend to be more computationally expensive, as well as more complex to be trained. Therefore, the new discriminative variables allow the deployed pipelines to significantly improve predictability in time series as the ones handled in this work.

3. Results and Discussion

3.1. Knowledge Data Discovery (KDD Analysis)

To better understand the relationships among the collected data, a descriptive statistical analysis was carried out, by computing several summarization metrics such as basic statistics, the coefficient of variation [64], Pearson correlation [65], and energy load histogram Q-Q plot [66].

The first calculated statistic was the Coefficient of Variation (CV), which gave us 13.39%. Such a percentage indicates that the variation of electricity demand in Brazil is homogeneous, i.e., it basically follows a stable distribution, as the CV establishes a measure of dispersion (the ratio) between the standard deviation and the average. Maxima and minima of the target variable—the *Energy Load*—were also collected, as exposed in Table 6. It was found that a variation between these scores (supposedly) ranges 3–4 times the value of the standard deviation above or below the average. Furthermore, one can verify that the average and the median (quartile 50%) are very close to each other, suggesting that the explored data hold a symmetric distribution.

Table 6. Descriptive statistical analysis of the Brazilian energy demand.

Statistics	Energy Load (GWh)
Average	1330.52
Standard deviation	178.25
Min value	856.56
Max value	1804.52
25%	1198.64
50% (median)	1320.01
75%	1467.63

The Pearson correlation was also computed to check how strongly the input variables are linearly related (see Figure 2). From the tabulated scores, one may conclude that the features with the highest positive (directly proportional) correlations for the *Energy Load* are those which delivery the highest scores, such as the maximum consumption demand on a day (*Max Demand*), the electricity generated from demand (*Generated Energy*), and the hydroelectric generation (*Hydroelectric Gen.*) Not surprisingly, the last feature, *Hydroelectric Gen.*, plays a central role in Brazil, as it is the main source of energy production, corresponding around 60% of the country's electricity. Among the variables with the highest negative correlations (inversely proportional) with respect to *Energy Load*, *Stored Energy* was revealed, which makes sense, as if there is too much potential energy stored in water reservoirs, then the electricity consumption is low. In summary, all those observations can be visualized in Figure 3.
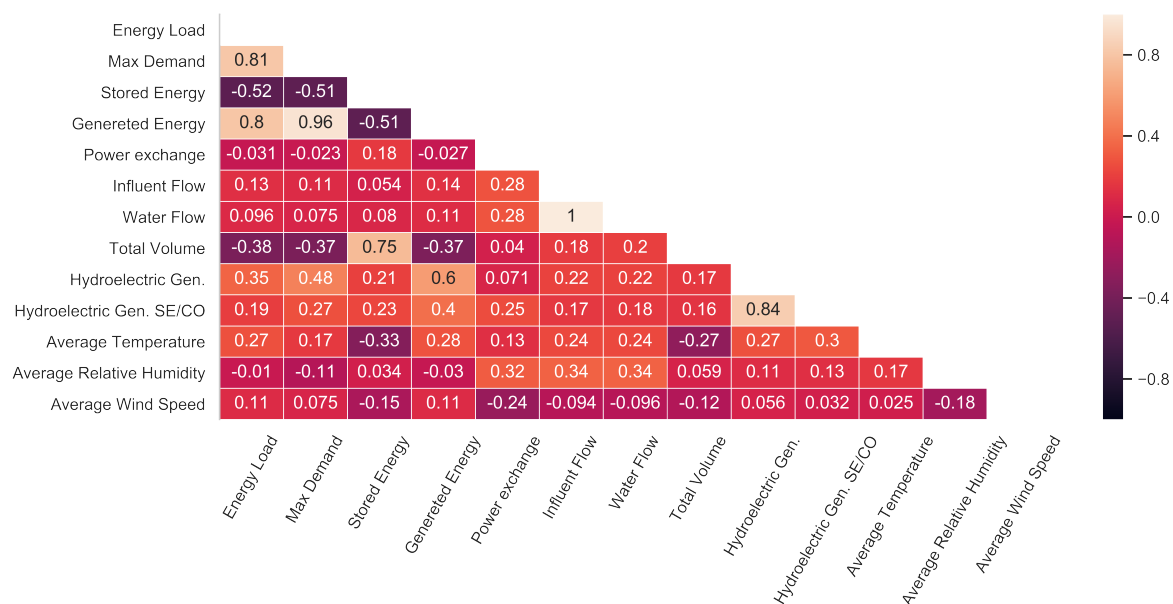

Figure 2. Pearson correlation, represented as a heat map for the original input variables.
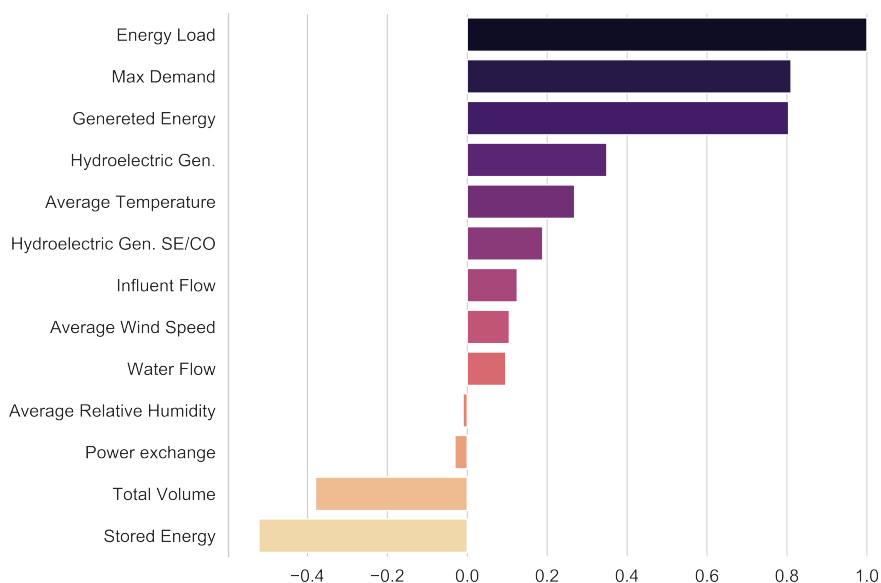
Figure 3. Pearson correlation for the *Energy Load*, only.

Figure 4 shows the normalized frequency distribution of the *Energy Load*, as a histogram during the period from 2005 to 2018. Notice that this normalization is determined so that the total area under the histogram is equal to 1 (see [67] for details). The Kernel Density Estimation (KDE), a continuous version of the histogram obtained by summing the individual Kernel contributions (i.e., Gaussians) at every data point [67], is also displayed. The plotted histogram indicates that the electricity consumption in Brazil is more often found in two peaks of almost the same density, but with different dispersions. The first peak is located close to the average and median values, as reported in Table 6, while the second peak is close to the average value plus the standard deviation, which comprises the 75% quartile. Between these peaks, there is also the presence of a gap, which corresponds to the region with the lowest occurrence in consumption, i.e., a range of 120 GWh. This valley can be explained by the expansion of the Brazilian industry during the period considered (2005–2018), as well as the climatic seasonality between the regions of the country, which are co-integrated by the National Interconnected Power System (SIN).
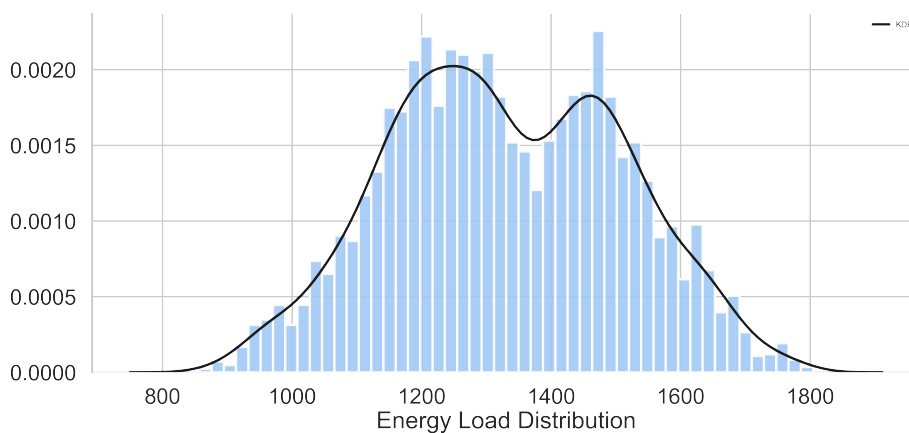

Figure 4. *Energy Load* histogram and the KDE curve. The x-axis groups the energy charge (in GWh) into bins, while the y-axis quantifies the resulting probability density values, i.e., a normalized histogram.

Statistical analysis to verify the existence of outliers in the BRAIPG dataset was also performed, by inspecting the probability distribution plot with the QQ-Plot (see Figure 5). From the plotted curve, one can see that there is a high linear relationship between the quartiles distributions, both visually identified when one checks the values above the line, and through the coefficient of determination, which is 0.9918. Statistically speaking, this means that the predictable variable (*Energy Load*) has no discrepant values, not pointing to outliers that could lead to noise in the data distribution.
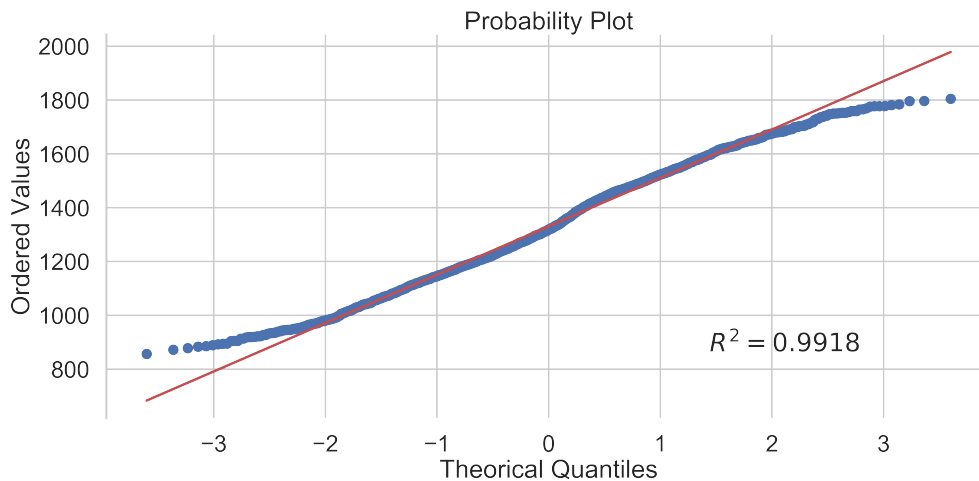

Figure 5. QQ-Plot computed for the target variable.

Finally, the graphs in Figures 6–8 show the average of the target variable over the weekly, monthly and annual horizons. The resulting distributions confirm that working days have the highest electrical consumption, while, on Sundays, it is smaller. The months of highest consumption are: February, March, and April. Notice that these months correspond to the summer season, as well as the transition to fall in Brazil, while the lowest consumptions are: June, July, and August (i.e., during the winter). It is also possible to see in Figure 8 that some sheet-like bars are flatter and longer, while others are thicker and shorter. This behavior comes from the data distribution in a given month so that the months of highest consumption are those with the largest variations in their distributions.
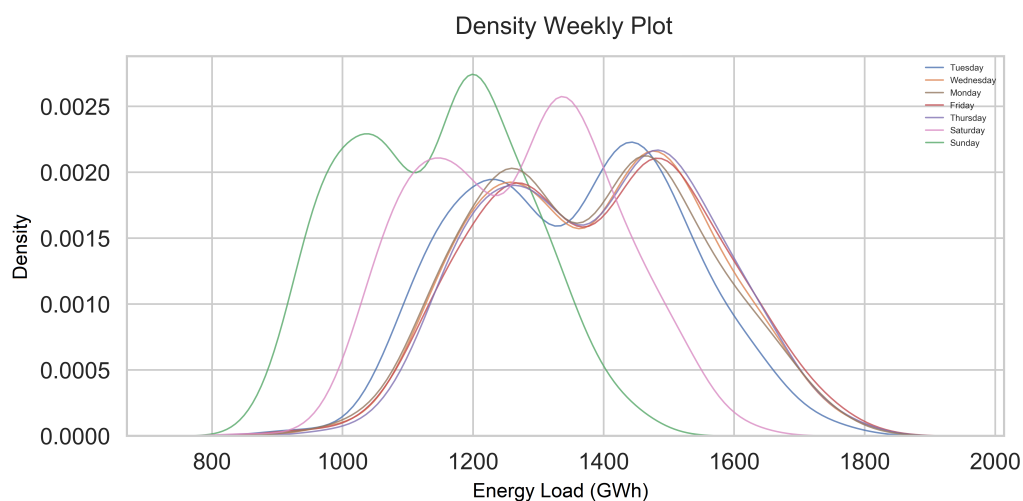

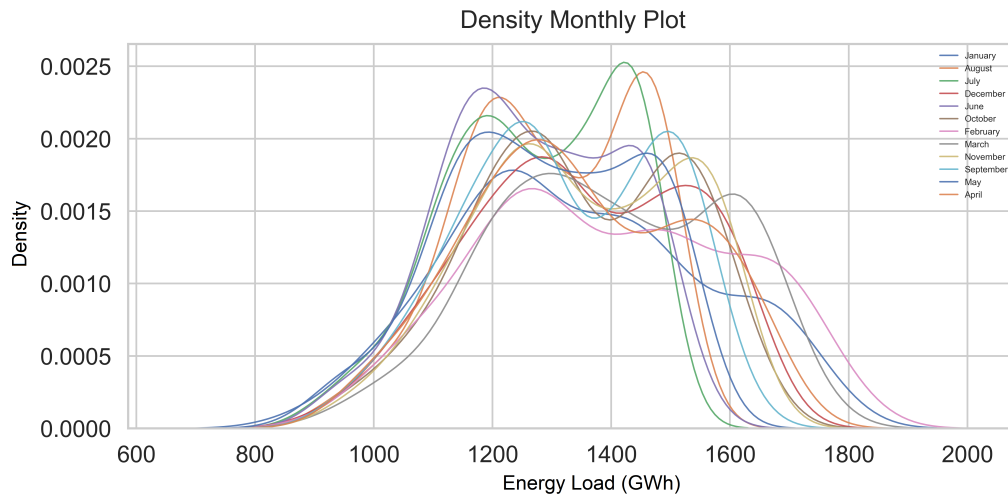**Figure 6.** Density plot of weekly average distribution for the *Energy Load*.

Figure 7. Density plot of monthly average distribution for the *Energy Load*.
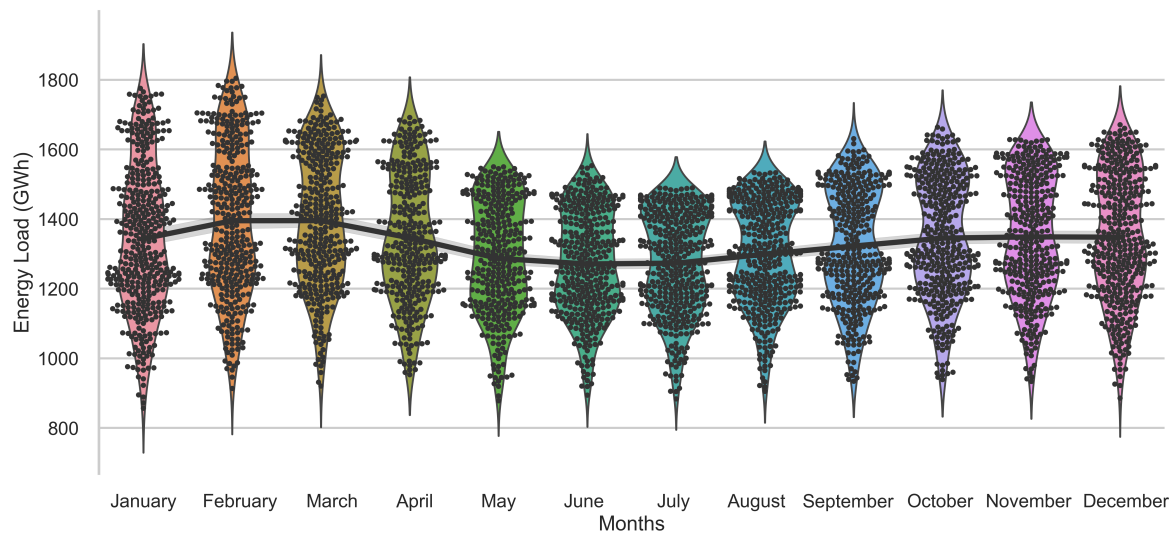

Figure 8. Average distribution of the *Energy Load* during a whole year (2018).

3.2. Data Preparation and Standardization

After performing the data exploration of our integrated dataset, around 90% of the collected data were taken for training the leaning models (period: 2005–2017), while the remaining portion of the data, around 10%, was used in the validation study (testing samples covering the whole year of 2018). Following the well-established protocol of separating the data into training and testing subsets, the normalization step was then computed for both subsets. More explicitly, this procedure consisted in normalizing all variables on a common scale of 0–1 to decrease the effects of different units between the variables, hence reducing the scalability bias naturally imposed by the variables.

The vectorization process of the categorical variables in the examined dataset was also carried out, which comprises: the days of the week and the months of the year (for implementation details, see [68]). Finally, the machine learning models were implemented using the *Scikit-learn* Python library, where the default parameters setting for each model was initially taken. Next, the optimization of the hyperparameters for each learning approach was performed to improve the forecasters in obtaining the best possible precision for the predicted load values.

3.3. Application of the Trained Models for Electricity Load Forecasting in Brazil

To assess the accuracy of the trained models in predicting the energy demand in the Brazilian Interconnected Power Grid, two evaluation scenarios have been investigated: the use of standard versions of the forecasting methods, as they are commonly applied in other related applications, and the utilization of our optimized pipelines with parameter tuning. Additionally, the well-established time-series based methods *Autoregressive Integrated Moving Average* (ARIMA) and *Long Short-Term Memory* (LSTM) were taken as baseline methods in our comparative analysis. ARIMA assumes three parameters: ARIMA(p,d,q), where p, d, and q are the orders of the model for the number of autoregressive, differences, and moving average parameters, respectively [69]. Its default parameters are given by $(2, 1, 0)$, and after tuning, by $(7, 1, 0)$, where $p \in [1, 15]$, $d \in [0, 5]$ and $q \in [0, 2]$. LSTM depends on two parameters: LSTM(w,n), where w is window used to split the training division in the database, and n is the number of neurons in the input layer. MSE was taken as a loss function, while ADAM as the optimizer and activation function. Finally, default parameters of LSTM are given by $(4, 50)$, and after tuning, by $(10, 500)$, where $w \in [1, 15]$ and $n \in [50, 1000]$. The codes were built based on the *Keras* implementations, a high-level neural networks API, and *TensorFlow*, a popular and robust open source deep learning tool designed by Google [70]. As the designed approaches, two parameter settings were considered for the time-series methods: by default and after parameter tuning with *Random Search*, similarly to what was done for our learning approaches.

In our experiments, the assessments were obtained from quality validation metrics as usually employed in the Machine Learning field. More specifically, the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE) [71,72] were used to assess the results:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \dot{y}_i|, \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \dot{y}_i)^2}, \tag{7}$$

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \dot{y}_i}{y_i} \right| \times 100, \tag{8}$$

where y_i and $\dot{y}_i$ account for the actual and forecasted values of energy load, respectively.

The predictive performance of both standard and optimized models was first verified in monthly horizons for the whole year of 2018, as shown in the left-side columns in Table 7. From the tabulated scores, one can compare the implemented approaches under different perspectives. First, when checking the individual performance of the standard models, one can observe that GB delivers the smallest prediction errors, while the SVR the biggest ones. After optimizing the models, SVR becomes very competitive, giving the best results for February and May-August. Moreover, both RF and GB increase their performance after hyperparameter optimization in most of the months, allowing the forecasters to be more effective in predicting the load demand in the Brazilian Interconnected Power Grid. By analyzing all the results together, both implemented approaches, RF and GB, clearly outperform the baseline time-series methods ARIMA and LSTM with/without parameter optimization. The only exception occurs with SVR without a suitable hyperparameter treatment, which is surpassed by the time-series models; however, the opposite holds when the SVR is properly optimized.

Now, if one checks the best and the worst months of predictions for each model individually (as highlighted in orange and green in Table 7, respectively), July and August are the ones with the lowest errors, which include the MAPEs close to zero for the tuned RF, GB, and SVR methods, while February and March return the highest errors (in overall). Notice, however, there are nuances for the worst results such as the months of January (optimized SVR), and May and December (ARIMA). Finally, one can also observe that the SVR is notably improved after a more appropriate parameter tuning, as the highest MAPE decreases drastically.

Table 7. MAPE (%) monthly over the test data (2018) for all the evaluated models with default (non-optimized) parameters against the optimized (tuned) pipelines.

| | Implemented Models | | | | | | Time-Series Based Models | | | |
| | Standard | | | Optimized | | | Standard | | Optimized | |
Months	**RF**	**GB**	**SVR**	**RF**	**GB**	**SVR**	**ARIMA**	**LSTM**	**ARIMA**	**LSTM**
1	1.208	0.862	15.753	0.924	0.931	1.687	6.643	8.272	3.632	6.037
2	1.197	0.877	18.318	0.681	0.769	0.597	5.864	10.878	2.603	7.903
3	2.212	1.481	15.947	1.699	1.242	1.331	5.664	8.208	2.796	8.148
4	1.587	1.423	12.487	1.204	1.024	1.239	5.851	7.575	3.331	4.813
5	1.441	1.129	11.114	1.367	1.072	0.796	7.233	8.301	3.371	5.447
6	1.257	1.055	9.634	1.004	1.013	0.961	6.305	7.468	3.541	6.146
7	1.070	0.880	8.206	0.571	0.507	0.418	6.537	7.017	4.384	5.757
8	0.760	0.646	9.365	0.566	0.682	0.559	5.555	6.636	2.438	4.114
9	1.068	0.776	10.888	0.862	0.746	0.859	6.524	7.726	4.023	5.422
10	0.977	1.017	11.875	1.058	0.911	0.932	6.280	8.409	2.496	6.913
11	1.640	1.374	11.128	1.255	1.324	1.410	6.556	7.367	2.891	6.345
12	0.850	0.849	14.435	0.825	0.753	0.762	6.587	7.427	4.608	4.472

Aiming at assessing which model delivers the best predictions, Table 8 summarizes the mean values of MAPE and MAE over the whole year of 2018 (test data). In all evaluation scenarios, GB gives the best results, with a MAPE, MAE, and RMSE of 0.918%, 13.832 GWh, and 19.798 GWh, respectively, resulting in an accuracy of 99.082%, i.e., an improvement of 0.116% with respect to the predicted values without parameter optimization (accuracy of 98.966%). Notice also that the SVR comes from a MAPE of 12.567% to an error less than 1 with parameter tuning, leading to prediction errors very close to the one generated by the GB. Concerning the time-series based approaches, ARIMA and LSTM, there was a considerable gain after tuning their parameters, but still they were less effective when compared against the implemented learning pipelines, especially the RF and GB ones, with/without tuning. In summary, the three improved forecasters RF, GB, and SVR were able to produce the lowest prediction errors. Such a high assertiveness can be explained due to the nature of the target variable (*Energy Load*), as the QQ-Plot (Figure 5) has returned a coefficient of determination of 0.9928% between the quartiles, demonstrating that the demand of electricity charge behaves well with respect to the external factors, besides having a good correlation to the fundamental variables present in our dataset.

Table 8. Mean of MAPE (%), MAE (GWh), and RMSE (GWh) over the test data (2018) for all the evaluated models with default and optimized parameters.

| | Implemented Models | | | | | | Time-Series Based Models | | | |
| | Standard | | | Optimized | | | Standard | | Optimized | |
Metrics	**RF**	**GB**	**SVR**	**RF**	**GB**	**SVR**	**ARIMA**	**LSTM**	**ARIMA**	**LSTM**
MAPE	1.279	1.034	12.567	1.006	0.918	0.962	6.255	7.137	3.317	4.707
MAE	19.279	15.576	189.313	15.160	13.832	14.497	94.196	107.316	49.957	70.821
RMSE	24.543	21.186	221.172	23.890	19.798	21.522	118.013	116.412	75.203	90.947

Since the GB has reached the best scores, its level of importance was calculated to rank the most relevant variables in the regression task. Figure 9 brings such a feature analysis, where the artificially created variable *MME Weekly Energy Load* was classified as the one with the highest weight in the GB predictability (75%), followed by *Energy Load log_return*, another side variable generated via resource engineering. This demonstrates the importance of setting new composite variables in the training step.

The prediction of *Energy Load*, as given by the trained models with parameter tuning, for the first quarter of 2018 is shown in Figure 10. The difference (residue) between the real and predicted values is also displayed (see Figure 11). Despite the high volatility usually found in energy load time series, the predictions follow the actual data very closely, capturing the cyclical nature of the ground-truth curve such as undulations and local extrema. The only exception occurs at two ill-behaved points,

located after the 80th day (see the gap in Figure 11). Among the trained models, the best one in terms of fitting data capability was the GB, since it produces residues closer to zero when compared to the others.
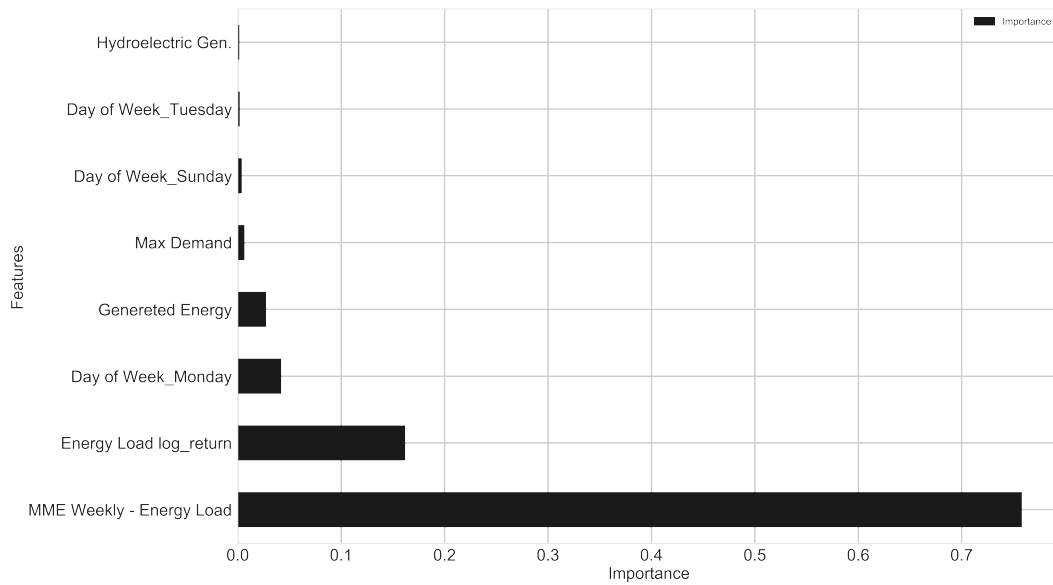

Figure 9. Importance of variables when predicting the *Energy Load* by the tuned GB.
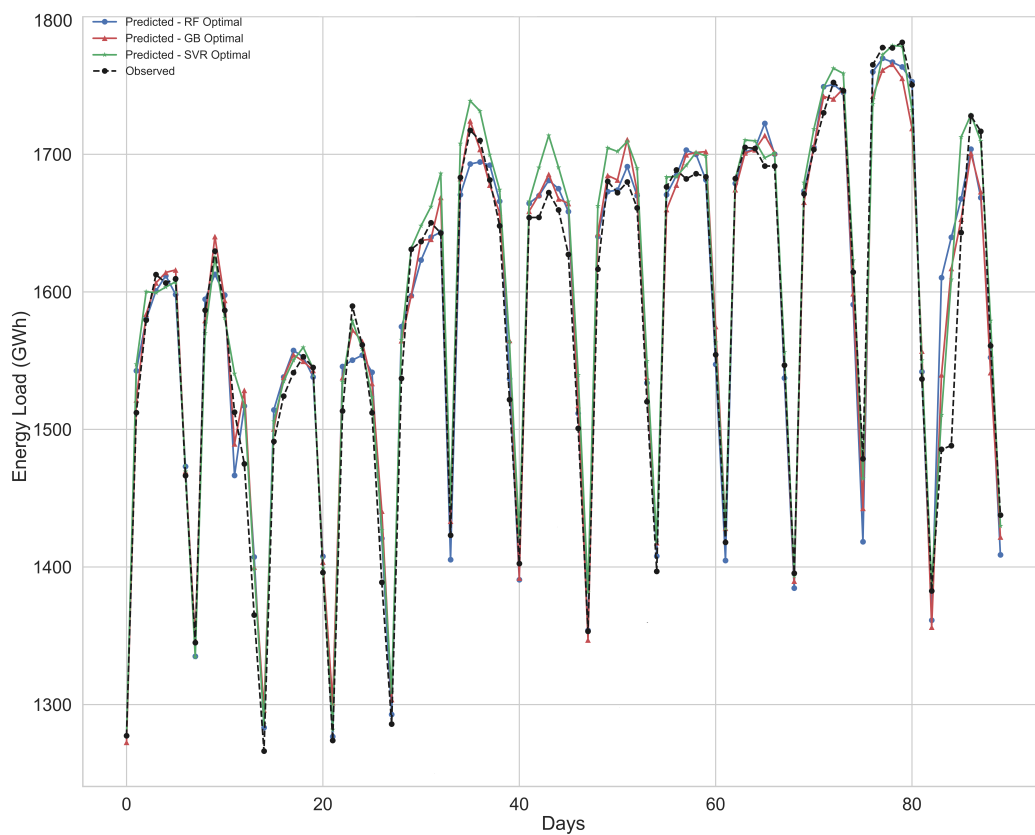

Figure 10. *Energy Load* predictions (obtained by the optimized forecasting models) × real values for the first quarter of 2018.
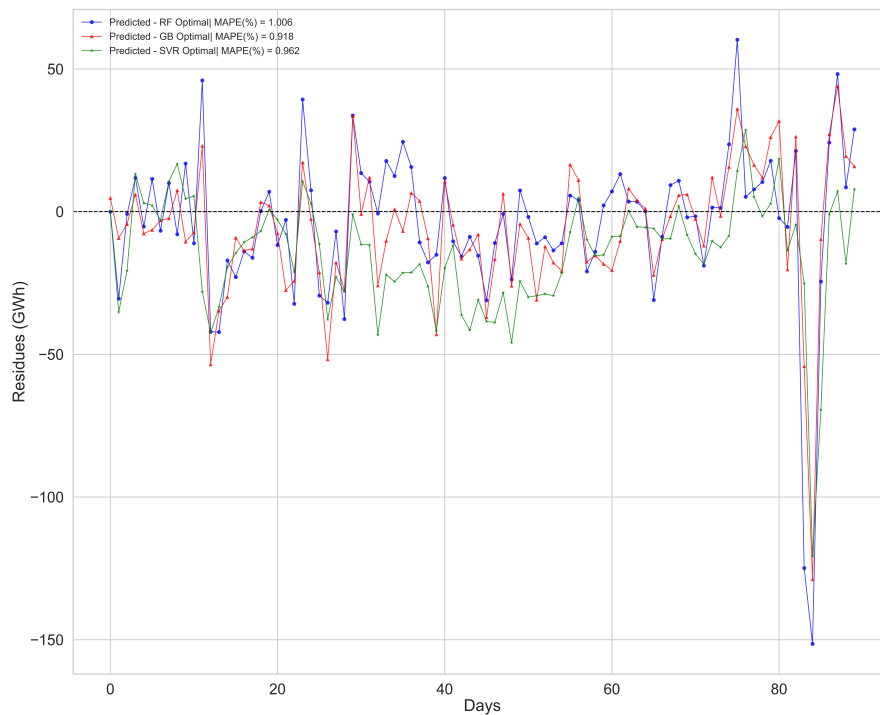
Figure 11. *Energy Load* residues, as obtained by the optimized models, for the first quarter of 2018.

The distributions of the prediction results, given by the optimized learning models, and the true data are shown in Figure 12. One can observe that, in the valley of the learning curves (1300–1400 GWh), the SVR underestimates the actual load, and, at the peak between 1400 and 1500 GWh, the opposite holds, for all the trained models. In contrast, for load values less than 1300 GWh, both RF and GB are able to produce highly accurate distribution curves with respect to real data, while SVR sightly overestimates the original distribution.
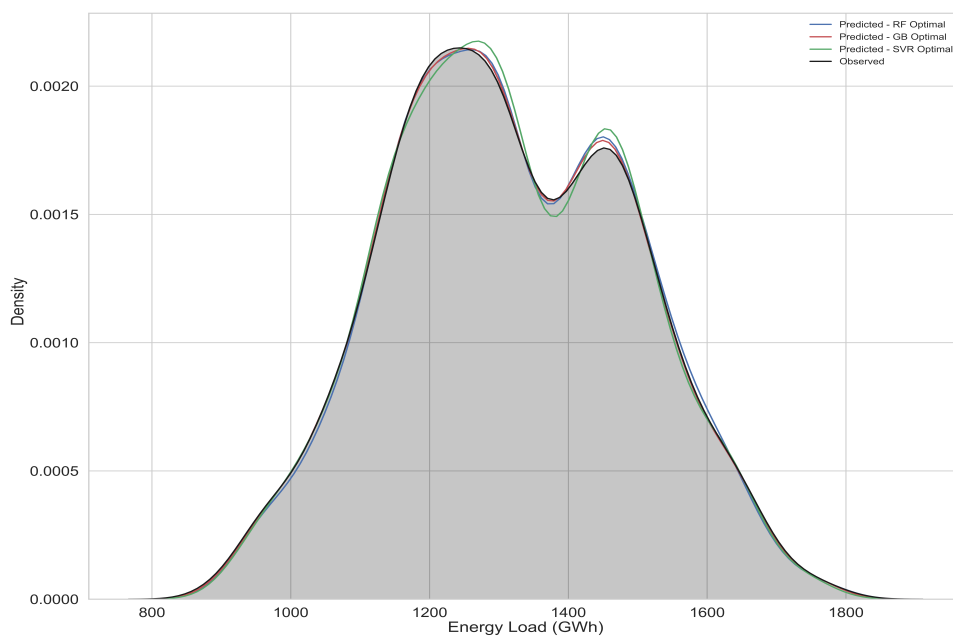

Figure 12. Density plot for the true observed data and the predictions generated by the trained models.

4. Conclusions

This study focused on the study of electricity load prediction in the Brazilian Interconnected Power Grid by means of different machine learning strategies and data exploration tools. In contrast to most existing works, which give only annual/monthly estimations for the electricity demand in Brazil, here, three ML models were applied and then optimized as new ensemble-based predictors with optimal hyperparameters to provide accurate daily/monthly forecasts. As verified in the evaluation study, the predictive model with the best performance was the GB, surpassing the other methods in terms of accuracy (tuned model: 99.082%) and MAPE/MAE (tuned model: 0.918% and 13.832%, respectively), therefore attesting the efficacy of GB in the predictability of electricity load demand in the Brazilian context.

The Knowledge Data Discovery (KDD), as conducted via the data analysis tools presented in Section 3, was also of paramount importance to reveal the statistical behavior and other intrinsic relationships of the collected data. Moreover, there was a substantial gain due to the creation of new artificial variables, as the ones delivered by the resource engineering scheme, which was crucial for weighing the ensemble-based models, as well as improving the SVR, since it did not achieve a satisfactory performance without a proper adjustment of parameters.

Finally, in addition to establishing new methodological pipelines to forecast the energy demand in Brazil and to go deeper into the acquired data, this work provides a full data collection of data taken from official Brazilian agencies to the industry and those who are interested in studying load demand, especially in the Brazilian context.

Author Contributions: Conceptualization, J.V.L., W.C., M.C. and M.A.D.; methodology, J.V.L., W.C. and M.C.; software, J.V.L. and M.A.D.; validation, J.V.L. and M.A.D.; formal analysis, W.C. and M.C.; resources, W.C. and M.C.; data curation, J.V.L.; writing–original draft preparation, J.V.L.; writing–review and editing, W.C. and M.C.; funding acquisition, W.C. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Martinez-Alvarez, F.; Troncoso, A.; Asencio-Cortes, G.; Riquelme, J.C. A survey on data mining techniques applied to electricity-related time series forecasting. *Energies* **2015**, *8*, 13162–13193. [CrossRef]
2. Kassakian, J.G.; Schmalensee, R.; Desgroseilliers, G.; Heidel, T.; Afridi, K.; Farid, A.; Grochow, J.; Hogan, W.; Jacoby, H.; Kirtley, J.; et al. *The Future of the Electric Grid: An Interdisciplinary MIT Study*; Massachusetts Institute of Technology: Cambridge, MA, USA, 2011; p. 280.
3. Hochstetler, R.L.; Cho, J.D. Assessing Competition in Brazil's Electricity Market If Bid-based Dispatch were Adopted. *Rev. Econ. Contemporanea* **2019**, *23*. [CrossRef]
4. Vieira, V.; Hochstetler, R.; Mello, J.C.O.; Barroso, L.A.N. Aligning Regulatory Incentives and Price Signals in the Brazilian Wholesale and Retail Electricity Markets. In Proceedings of the CIGRE Session, Paris, France, 22–26 August 2016; pp. 1–10.
5. Suganthi, L.; Samuel, A.A. Energy models for demand forecasting—A review. *Renew. Sustain. Energy Rev.* **2012**, *16*, 1223–1240. [CrossRef]
6. Ghalehkhondabi, I.; Ardjmand, E.; Weckman, G.R.; Young, W.A. An overview of energy demand forecasting methods published in 2005–2015. *Energy Syst.* **2017**, *8*, 411–447. [CrossRef]
7. Huang, J.; Tang, Y.; Chen, S. Energy Demand Forecasting: Combining Cointegration Analysis and Artificial Intelligence Algorithm. *Math. Probl. Eng.* **2018**, *2018*, 1–13. [CrossRef]
8. Seyedzadeh, S.; Rahimian, F.P.; Glesk, I.; Roper, M. Machine Learning for Estimation of Building Energy Consumption and Performance: A Review. *Vis. Eng.* **2018**, *6*, 1–20. [CrossRef]
9. Bouktif, S.; Fiaz, A.; Ouni, A.; Serhani, M. Multi-Sequence LSTM-RNN Deep Learning and Metaheuristics for Electric Load Forecasting. *Energies* **2020**, *13*, 391. [CrossRef]

10. Abdoos, A.; Hemmati, M.; Abdoos, A.A. Short term load forecasting using a hybrid intelligent method. *Knowl.-Based Syst.* **2015**, *76*, 139–147. [CrossRef]

11. Fidalgo, J.N.; Matos, M.A. Forecasting Portugal Global Load with Artificial Neural Networks. In Proceedings of the International Conference on Artificial Neural Networks (ICANN), Porto, Portugal, 9–13 September 2007; pp. 728–737.

12. Li, K.; Zhang, T. Forecasting Electricity Consumption Using an Improved Grey Prediction Model. *Information* **2018**, *9*, 18. [CrossRef]

13. Rana, M.; Koprinska, I. Forecasting electricity load with advanced wavelet neural networks. *Neurocomputing* **2016**, *182*, 118–132. [CrossRef]

14. Graff, M.; Pena, R.; Medina, A.; Escalante, H.J. Wind speed forecasting using a portfolio of forecasters. *Renew. Energy* **2014**, *68*, 550–559. [CrossRef]

15. Koprinska, I.; Rana, M.; Agelidis, V.G. Correlation and instance based feature selection for electricity load forecasting. *Knowl.-Based Syst.* **2015**, *82*, 29–40. [CrossRef]

16. Ferraro, P.; Crisostomi, E.; Tucci, M.; Raugi, M. Comparison and clustering analysis of the daily electrical load in eight european countries. *Electr. Power Syst. Res.* **2016**, *141*, 114–123. [CrossRef]

17. Dong, Y.; Wang, J.; Wang, C.; Guo, Z. Research and application of hybrid forecasting model based on an optimal feature selection system—A case study on electrical load forecasting. *Energies* **2016**, *10*, 490. [CrossRef]

18. Paulos, J.P.; Fidalgo, J.N. Load and Electricity Prices Forecasting using Generalized Regression Neural Networks. In Proceedings of the International Conference on Smart Energy Systems and Technologies (SEST), Sevilla, Spain, 10–12 September 2018; pp. 1–6.

19. Sarhani, M.; Afia, A.E. Generalization Enhancement of Support Vector Regression in Electric Load Forecasting with Model Selection. In Proceedings of the International Conference on Learning and Optimization Algorithms: Theory and Applications, Rabat, Morocco, 2–5 May 2018.

20. Qiu, Z. Electricity consumption prediction based on data mining techniques with particle swarm optimization. *Int. J. Database Theory Appl.* **2013**, *6*, 153–164. [CrossRef]

21. Juradoa, S.; Nebot, A.; Mugica, F.; Avellana, N. Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques. *Energy* **2015**, *86*, 276–291. [CrossRef]

22. Hernández, L.; Baladrón, C.; Aguiar, J.; Carro, B.; Sánchez-Esguevillas, A. Classification and Clustering of Electricity Demand Patterns in Industrial Parks. *Energies* **2012**, *5*, 5215–5228. [CrossRef]

23. Zhao, H.; Zhao, H.; Guo, S. Using GM (1,1) Optimized by MFO with Rolling Mechanism to Forecast the Electricity Consumption of Inner Mongolia. *Appl. Sci.* **2016**, *6*, 20. [CrossRef]

24. Liang, J.; Liang, Y. Analysis and Modeling for China's Electricity Demand Forecasting Based on a New Mathematical Hybrid Method. *Information* **2017**, *8*, 33. [CrossRef]

25. Ameyaw, B.; Yao, L. Sectoral Energy Demand Forecasting under an Assumption-Free Data-Driven Technique. *Sustainability* **2018**, *10*, 2348. [CrossRef]

26. Bouktif, S.; Fiaz, A.; Ouni, A.; Serhani, M.A. Optimal deep learning lstm model for electric load forecasting using feature selection and genetic algorithm: Comparison with machine learning approaches. *Energies* **2018**, *11*, 1636. [CrossRef]

27. Maçaira, P.M.; Silva, F.L.C.; Oliveira, F.L.C.; Calili, R.F.; Lourenço, P.M. Statistical analysis of the Brazilian electricity sector: A top-down long range energy consumption and supply forecast model. In Proceedings of the XLVI Brazilian Symposium of Operational Research, Salvador, Brazil, 16–19 September 2014; pp. 1182–1193.

28. Torrini, F.C.; Souza, R.C.; Oliveira, F.L.C.; Pessanha, J.F.M. Long term electricity consumption forecast in Brazil: A fuzzy logic approach. *Socio-Econ. Plan. Sci.* **2016**, *54*, 18–27. [CrossRef]

29. Trotter, I.M.; Bolkesjø, T.F.; Féres, J.G.; Hollanda, L. Climate change and electricity demand in Brazil: A stochastic approach. *Energy* **2016**, *102*, 596–604. [CrossRef]

30. International Energy Agency. *Technology Roadmap: Hydropower*; International Energy Agency: Paris, France, 2012; p. 68.

31. da Silva, F.L.; Souza, R.C.; Oliveira, F.L.C.; Lourenço, P.M.; de C. Fagundes, W. Forecast of Long-term Electricity Consumption of the Industrial Sub-sector of Pulp and Paper in Brazil Using a Bottom-up Approach. *Procedia Comput. Sci.* **2015**, *55*, 514–522. [CrossRef]

32. Resendea, L.; Soaresb, M.; Ferreirac, P. Electric Power Load in Brazil: View on the Long-Term Forecasting Models. *Production* **2018**, *28*, 1–12. [CrossRef]

33. Dai, S.; Niu, D.; Li, Y. Forecasting of energy consumption in China based on ensemble empirical mode decomposition and least squares support vector machine optimized by improved shuffled frog leaping algorithm. *Appl. Sci.* **2018**, *8*, 678. [CrossRef]

34. Utterback, O. Feature Selection Methods with Applications in Electrical Load Forecasting. Master's Thesis, Lund University, Lund, Sweden, 2017.

35. Zhang, X.; Fang, F.; Liu, J. Weather-Classification-MARS-Based Photovoltaic Power Forecasting for Energy Imbalance Market. *IEEE Trans. Ind. Electron.* **2019**, *66*, 8692–8702. [CrossRef]

36. Ceci, M.; Corizzo, R.; Malerba, D.; Rashkovska, A. Spatial Autocorrelation and Entropy for Renewable Energy Forecasting. *Data Min. Knowl. Discov.* **2019**, *33*, 698–729. [CrossRef]

37. Sarhani, M.; Afia, A.E. Electric load forecasting using hybrid machine learning approach incorporating feature selection. In Proceedings of the International Conference on Big Data Cloud and Applications, Tetuan, Morocco, 25–26 May 2015; pp. 1–7.

38. Liang, Y.; Niu, D.; Hong, W.C. Short term load forecasting based on feature extraction and improved general regression neural network model. *Energy* **2019**, *166*, 653–663. [CrossRef]

39. National Electrical System Operator (ONS Brazil). Available online: http://ons.org.br (accessed on 1 February 2019).

40. Brazilian Electricity Regulatory Agency (ANEEL Brazil). Available online: http://www2.aneel.gov.br/aplicacoes/capacidadebrasil/capacidadebrasil.cfm (accessed on 1 February 2019).

41. Soccol, F.J.; Pereira, A.L.; Celeste, W.C.; Coura, D.J.C.; de Lorena Diniz Chaves, G. Challenges for Implementation of Distributed Energy Generation in Brazil: An Integrative Literature Review. *Braz. J. Prod. Eng.* **2016**, *2*, 31–43.

42. National Institute of Meteorology (INMET Brazil). Available online: http://www.inmet.gov.br/portal/index.php?r=home2/index (accessed on 2 February 2019).

43. Brazilian Interconnected Power Grid Map. Available online: http://ons.org.br/paginas/sobre-o-sin/mapas (accessed on 2 February 2019).

44. Zhou, Z.H. *Ensemble Methods: Foundations and Algorithms*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2012.

45. Lin, L.; Wang, F.; Xie, X.; Zhong, S. Random forests-based extreme learning machine ensemble for multi-regime time series prediction. *Expert Syst. Appl.* **2017**, *83*, 164–176. [CrossRef]

46. Qiu, X.; Zhang, L.; Suganthan, P.N.; Amaratunga, G.A. Oblique random forest ensemble via least square estimation for time series forecasting. *Inf. Sci.* **2017**, *420*, 249–262. [CrossRef]

47. Wang, X.; Yu, F.; Pedrycz, W.; Wang, J. Hierarchical clustering of unequal-length time series with area-based shape distance. *Soft Comput.* **2019**, *23*, 6331–6343. [CrossRef]

48. Wang, P.; Li, Y.; Reddy, C.K. Machine learning for survival analysis: A survey. *ACM Comput. Surv.* **2019**, *51*, 110:1–110:36. [CrossRef]

49. Keprate, A.; Ratnayake, R.C. Using gradient boosting regressor to predict stress intensity factor of a crack propagating in small bore piping. In Proceedings of the IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), Singapore, 10–13 December 2017; pp. 1331–1336.

50. Kazemi, V.; Sullivan, J. One millisecond face alignment with an ensemble of regression trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014; pp. 1867–1874.

51. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

52. Brownlee, J. A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning. Available online: http://machinelearningmastery.com/gentleintroductiongradient-boosting-algorithm-machine-learning (accessed on 9 July 2018).

53. Scikit-learn: Gradient Boosting. Available online: https://scikit-learn.org/stable/modules/ensemble.html#gradient-boosting (accessed on 1 February 2019).

54. Izmailov, R.; Vapnik, V.; Vashist, A. Multidimensional splines with infinite number of knots as svm kernels. In Proceedings of the International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 4–9 August 2013; pp. 1–7.

55. Marmaras, C.; Javed, A.; Cipcigan, L.; Rana, O. Predicting the energy demand of buildings during triad peaks in gb. *Energy Build.* **2017**, *141*, 262–273. [CrossRef]

56. Kernel SVM: Support Vector Machine Regressor. Available online: http://kernelsvm.tripod.com/ (accessed on 1 October 2019).

57. Scikit-learn: Random Forest Regressor Documentation. Available online: http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html (accessed on 2 April 2019).

58. Scikit-learn: Gradiente Boosting Regressor Documentation. Available online: http://scikitlearn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (accessed on 22 May 2019).

59. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.

60. Scikit-learn: Support Vector Regressor Documentation. Available online: http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html (accessed on 22 May 2019).

61. Hudson, R.S.; Gregoriou, A. Calculating and comparing security returns is harder than you think: A comparison between logarithmic and simple returns. *Int. Rev. Financ. Anal.* **2015**, *38*, 151–162. [CrossRef]

62. Chen, B.; Choi, J.; Escanciano, J.C. Testing for Fundamental Vector Moving Average Representations. *Quant. Econ.* **2017**, *8*, 149–180. [CrossRef]

63. Morgan, D.L.; Zilvinskis, J.; Dugan, B. Opening the Activism and Postsecondary Education Black Box: Relating High-Impact Practices and Student Identity With Activist Behaviors. *J. Political Sci. Educ.* **2019**. [CrossRef]

64. Cheng, Z.; Liu, D.; Zhang, L. Random Two-Frame Phase-Shifting Interferometry via Minimization of Coefficient of Variation. *Appl. Phys. Lett.* **2019**, *115*, 121107. [CrossRef]

65. Ly, A.; Marsman, M.; Wagenmakers, E.J. Analytic posteriors for Pearson's Correlation Coefficient. *Stat. Neerl.* **2018**, *72*, 4–13. [CrossRef] [PubMed]

66. Pandit, R.; Infield, D. QQ plot for Assessment of Gaussian Process wind Turbine Power Curve Error Distribution Function. In Proceedings of the 9th European Workshop on Structural Health Monitoring. British Institute of Non-Destructive Testing, Manchester, UK, 10–13 July 2018; pp. 1–11.

67. VanderPlas, J. *Python Data Science Handbook: Essential Tools for Working with Data*; O Reilly Media: Sebastopol, CA, USA, 2016; p. 541.

68. Potdar, K.; Pardawala, T.S.; Pai, C.D. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *Int. J. Comput. Appl.* **2017**, *175*, 7–9. [CrossRef]

69. ARIMA Implementation. Available online: https://www.statsmodels.org/stable/generated/statsmodels.tsa.arima_model.ARIMA.html (accessed on 27 February 2020).

70. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Zhifeng, C.; Citro, C.; Corrado, G.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available available online: tensorflow.org (accessed on 26 February 2020).

71. Keitsch, K.A.; Bruckner, T. Input data analysis for optimized short term load forecasts. In Proceedings of the IEEE Innovative Smart Grid Technologies—Asia (ISGT-Asia), Melbourne, Australia, 28 November–1 December 2016; pp. 1–6.

72. Adhikari, R.; Agrawal, R. *An Introductory Study on Time Series Modeling and Forecasting*; LAP Lambert Academic Publishing: Saarbrücken, Germany, 2013; p. 76.