1. INTRODUCTION
The electricity sector has a strategic and vital role in the economy of any country, upholding the primary material for moving industrial processes, commerce and everyday life. Alas still a commodity, and as in any other similar sector, the development of strategies that can mitigate market risks and uncertainties, and generate earnings opportunities, is critical to energy companies.

Load forecasting strategies have been widely applied to minimize operational costs and maximize the economic efficiency of power utilities, still remaining an active field of research. To this end, the use of Machine Learning (ML) algorithms has been regularly proposed, and frequently compared, as very effective approaches.1-3

This paper is not intended to be a survey or review in the area of load forecasting (LF), as there are comprehensive books and related references covering this aspects3-5; it will also not further debate as to the importance of LF and their contribution to the planning and operation of power system, as they are also well versed and conventionally discussed in literature. Instead, our goal to present an analysis of a general application of LF, implemented through the process of Data Mining (DM); in this line, from the perspective of a ML practitioner, highlight some of the traits that differentiates this field and discuss practices for LF. In this focus, we expect to provide new scholars in this area with a succinct and objective exposition to the DM application for LF, and also bridge some viewpoints of its main actors, the domain and the technology experts.

2. CONNECTING/MATCHING THE DOMAIN WITH METHODS
2.1 Data mining and the introduction to a new domain
From the perspective of a data analyst or ML expert, joining on a new application domain can be an exciting endeavor, but one whose trail toward the desired results is not without a few chasms to transpose. Those, in turn, can be bridged by either: a deepening into the domain, through researched correlated works and literature reviews in the new area; or by continued assistance (or tutoring) by the domain expert.

While it is ideal to bring both approaches together, as they are complementary, the eagerness in obtaining results faster sometimes lead newcomers to preferably take the direction of learning predominantly from the domain expert. Just as often, however, differences in familiar notations, background and sheer unexpressed tacit knowledge from both sides can lead to misinterpretations; hence the thought speed up might turn into massive reworking to what can already be a laborious task.

Generally speaking, in the area of pattern analysis and knowledge discovery, whichever the application case, a standard process should be applied to better direct the research. Whereas different processes or roadmaps are presented, they are arguably representing slight variations or integrations (e.g. text mining, web mining, CRISP-DM) of...
the Knowledge Discovery in Database (KDD) process, nowadays synonymously referred simply as DM process. Further still, though idealized for application when vast amount of data is available, its framework is still successfully applied on various domains without much distinction to the actual amount of data or dimensionality. The KDD, or DM, primarily encompasses the steps of data selection, pre-processing and transformation (sometimes jointly referred as data treatment), and the pattern analysis by applying ML algorithms.

2.2 Load forecasting on power systems

The task of power LF is a key element in decision making processes in electrical systems, being an essential part to the efficient planning and operation of power utilities, involving decisions on the purchase and generation of electricity, load switching and development of infrastructures\(^7\). The problem of LF falls into the DM task of prediction/regression, where methods are used to identify existing patterns on historical data and thus estimate the series’ future values based on the defined forecasting horizon; whether: very-short term, for systems with data sampling rate smaller than an hour; short-term forecasts, which are usually from one hour to one week; medium-term forecasts, one week to one year; and long-term forecasts, for periods longer than one year.

For ML practitioners, as a good practice, the definition of data processing methods, or ML algorithm to apply, must not be overly biased on the success of previous related researches in the area, as many levels of differentiation can exist on each case.

Aspects that usually distinguish LF case studies mainly comprehend the class of the power load (e.g. residential, rural, industrial, total) to be estimated, and the origination region of the power supplier and its samples; the latter, in turn, can also be as flexible as its discerning groups or subgroups. Based on their specific regulatory systems and laws, socioeconomic and climatic conditions, area characteristics, population distributions, among others, different Countries, even within them, their Regions, States and Cities, can display very unique consumption behaviors\(^7\). As a case in point, taking the example of Brazil, it is a Country that is only now beginning to implement a system that consider the use of different tariffs for energy consumption based on time of day (or differentiating pricing for specific days/months in general); it is also one that almost entirely relies on hydropower systems for energy generation, consequentially quite susceptible to variations of climatic and hydrological nature. Brazil’s large territorial extension and uneven population distribution, specially in the northern (Amazon) region, and existing access difficulties on several cities makes it hard for the power suppliers to provide a stable energy distribution, of even to timely support and repair problems that (often) occur in the distribution system; depending on the period of year and/or weather conditions, some locations can only be reached by boat or helicopter. With such diverse characteristics, and economical contrasts found throughout the Country’s territorial clusters, to consider a single model (whether National, Regional or even State wise) to be generally used as rule of thumb for managing LF would be unsound.

The good news is that, even without much expert experience in the area of power systems, following a DM process and good practices for applying ML algorithms can be enough to achieve effective results. The reason is that, disregarding the domain’s particularities, it is still falls as a prediction task over a time series (usually structured) dataset.

3. MACHINE LEARNING AND DATA MINING, FOLLOWING THE STANDARDIZED ROADMAP

3.1 Data selection and treatment

The training data represents load consumption values, with samples collected over a period of time. With a dependent consumption variable \(Y\), and a historical dataset, the objective is to find the most suitable model to estimate future values of \(Y\) based on a set of \(X\) independent variables. When dimensioning samples and variables, the domain data available can be facing one of the two orders of problems for DM analysis: i) whether too much data is available, making it hard for analysis, or unfeasible for a good performance of ML algorithms; ii) or few samples are available, a much bigger problem as, even if it is a representative sample, it might not be sufficient for effectively training the ML algorithm. While the former can be more easily treated by means of sampling or aggregation methods, a solution for the latter requires knowledge and insight from experts for proposing alternatives, or even substantial financial cost, for purchasing data or realizing new experiments.

3.2.1 Dataset samples

Sample size reduction can sometimes be a matter of sim-
ple aggregation (e.g. by average) to a higher discrete interval, i.e. hourly sampled values into days, or daily values into weeks or months and so on; as long as it does not conflict with the constraints of the forecasting goal, and is a level of detail that all existing \( X \) variables can abide as well. Aggregating to a higher level can also work as a way to minimize the effect of missing values in the samples, if they occur. Since the data follows a time series, the option of simply removing missing data samples could impair the analysis, and so it is prudent to treat them prior to the ML application.

When considering the samples component, aggregating and rearranging the data can also provide a different view of the series trends; as one of the main strategies for treating time series is finding and separately treating its periodicities (e.g. whether daily, weekly, monthly, seasonal). Figure 1 provides such a transposed view of a daily sampled data: by aggregating hourly or daily to a higher (monthly) level, periodicities are easier to visually distinguish (Figure 1L); and by partitioning the single monthly series into 12 yearly series (Figure 1R), a simplified view for forecasting was found, as an almost linear trend. In this case, less complexity in the algorithmic solution, and possibly more accuracy in results, can be obtained for the price of creating a higher number of models for individually treating each component.

Additionally, as for outlier treatment, values with higher variation from the series’ trends would have to be estimated directly (e.g. based on the average, regression or data imputation methods), or corrected based on other features, as highlighted in the next section. To this end, filtering techniques have also been successfully applied, not solely for smoothing samples, but also providing more information on the dataset’s behavior, allowing for segmenting the series into different levels of seasonalities and noise; hence they can be separately evaluated for observing the foremost trend of the series and the occasional fluctuations.

In the case when only a small amount of samples is available, approaches usually settle for using less complex algorithms, with fewer parameters and a less demanding training procedure. In contrast, if the available samples are representative enough, a data densification process can be implemented for producing virtual training samples based on the probability distribution of the original dataset; the virtual samples can then be used for training, while the real historical data is used for validating the model.

3.1.2 Dataset variables

In a case where the order of \( X \) variables is substantial, or their quality for estimating \( Y \) is uncertain, dimensionality reduction can be obtained by applying a factor analysis or a feature selection method over \( X \). On the opposite (and habitual) side, as primarily only one variable is available (i.e. the energy consumption – \( Y \)), the selection or expansion of \( X \) with new variables depend on the domain aspects or forecasting goals. Common considerations for additional variables for constituting the \( X \) matrix set are:

A. Past observations

By far, the most regarded feature in explaining the future trend of a dependent variable \( Y \) in a time series is its own past behavior variables – \( X = [Y_{-1}; Y_{-2}; \ldots; Y_{-s}] \); a sensible and also comfortable approach, since \( Y \) is available from the start. The challenge here then becomes to find the best lagged window size \( s \) that can more accurately represent the trend \( [Y_{n}, Y_{n+1}, \ldots] \).

Applying feature selection approaches or known metrics (e.g. correlation, analysis, regression explaining factor \( R^2 \)) can provide an evaluation on the best number of lagged points to estimate \( Y \) (i.e. \( \{Y_{l}|Y_{-l}\}, \{Y_{j}|Y_{-j}\}, \{Y_{k}|Y_{-k}\}, \ldots \} \); or better yet, the most significant individual points (i.e. \( \{Y_{l}|Y_{-l}\}, \{Y_{j}|Y_{-j}\}, \{Y_{k}|Y_{-k}\}, \ldots \) as they can co-
incidently lead to insight on future approaches to pursue in the DM process. A realization on X’s most relevant lagged points can also motivate and/or better fundament considering the approach to partition the time series into complementing models; that is, building a different model for each of the seven days of the week or, as in Figure 1R, one for every month. For instance, consistently higher values found for lagged points that are a multiple of 7 on a daily sampled dataset can indicate a weekly periodicity; and similarly for multiples of 12 on a monthly sampled.

B. Exogenous and artificial variables

Besides “time”, energy consumption variation and growth also suffers impact of variables exogenous to the power systems area, mainly climatic and socioeconomic factors. However, as they are also subjected to dynamic variations, and not so easy to predict themselves, additional layers of errors might gradually increase the deviation and errors for longer forecasting points.

When consistent, however, this expense pays off. Generally, as socioeconomic variables (e.g. growing rate of industry sectors, variations in currencies, directions of governmental policies) and knowledge of the power supplier’s market (e.g. when a new major consumer would enter the system) can improve long term forecasting for the majority of consumption classes; and more specifically adjusted, as: commercial consumption varies depending on time of day, notable events, and day of the week; industrial consumption can be additionally affected on the location’s (Country, State) economical situation and policies; whereas residential consumption can furthermore adhere to hour/cost constraints if a system of different tariffs is employed by the suppliers.

These considerations can have their values incorporated in the model directly as artificial variables $P = [P_1; P_2; \ldots; P_n]$ with values for each $P_i$ assigned based on the types of each $P_i$ possible time profiles (e.g. day of week, workday or non-work day, period of day).

Adding climatic variables and other specific weather related events can equally improve the predictive model’s accuracy. Similarly, their addition to the model could be as $P$ discrete variables with classes symbolizing seasonal weather events, or as continuous variables (e.g. values for maximum and minimum temperature, precipitation rates for rain and snow).

C. Outliers and special days as artificial variables

Additional $P$ variables can be used to fine-tune the impact of irregular events, with values duly assigned for asserting the corresponding samples as an abnormal point or period (as the marked event O in Figure 1).

Also habitually accounted as additional artificial variables, specially for short term forecasting, the impact of holidays receives a special attention in their discrimination. Categorical $P$ variables and approaches used for describing special days can be as varied as: two categories, one for normal work days and one for weekends and holidays; three categories, one for normal work days, one clustering Saturdays and days before and after a holiday, and one for Sundays and holidays; four categories, normal work days, holidays, days before holiday and days after holiday; seven categories for holidays, expressing the impact of a holiday event when occurring on the different days of the week; different categories set for specific holidays, particularly Mother’s Day, Christmas and New Year; and so on, with even finer levels of calendar discriminations defined, in reflection to the case study location and consumption class considered. Time profiles on shorter forecasting horizons can be similarly found, e.g. the variation in hourly consumption data, as well as its periodicity, is inherently affected by the period of day and day of the week it is collected.

3.2 Data Mining

After a good time spent preparing the data to serve as train/test sets for the ML algorithms, the DM becomes much swifter, almost a process of triad and error involving: i) experimentation of different ML algorithms; ii) adjustment of parameters for the ML algorithms; iii) iterate back to a previous selection or treatment step, toward adding/removing variables, evaluation of outlier points, etc.

In fact, with a broad collection of ML tools readily available for use, most of which simplify the DM process to a connected sequence of “Read input dataset file”, “Select variables”, “Choose algorithm” and “Run”, and as the prediction error is usually the only measure considered to evaluate the quality of different approaches and algorithms used, DM can be mistaken as an all too straightforward process. It is here, however, that some cautions should be taken, specially regarding the domain expert, as to assure the reliability of results.

The easiness of use derived from interfaces and toolboxes, along with an array of implemented, ready-to-use (and sometimes optimized) ML algorithms, should not be neglected, as they can drastically decrease the simulation time.
and expand the progress of the overall investigative research; still, they should not be blindly trusted, and applied preferably with a minimum knowledge of the tool’s inside procedures and the workings of the selected ML algorithms. With this in mind, some important aspects to consider are as follows:

- Familiarize with how the tool handles the data prior and during training/testing of the ML algorithm is paramount. As illustration, the data separation into training and testing must be guaranteed not to be at random, as sometimes is the case for general DM tools and default configurations, but sequential, in conform to the requisites of a forecasting task. A similar attention is needed when forecasting several points ahead, as to assure that the testing process recurrently uses estimated values to predict the succeeding ones.

- Using more complex ML algorithms, of which there is already a familiarity and know-how is often preferred, moreover if correlated works support their application to the task, but testing simpler algorithms should also be encouraged and attempted. Applying an Artificial Neural Networks or novel hybrid algorithms over the data is prominently reinforced in literature, but might not be so justifiable if a simpler regression based model, over a well prepared training dataset, can suffice.

- ML algorithms are known to presents several parameters for adjusting and thus improving its adherence to the data’s patterns and likewise quality of results; a common fault, however, would be to take the tool’s default parameters settings as universally optimal and, consequentially, regard the algorithm as ineffective if results are not satisfactory on their trial. It is sometimes the case that taking more time fine-tuning parameters of ML algorithms can be more effective than prematurely jumping to a different one.

- Besides predictive accuracy, other criteria can also be considered for choosing (or ranking) the forecasting algorithms; still, it is the case where the weight of their evaluation is commonly overlooked in the trade-off with accuracy. Examples are: complexity of the algorithm, from linear multivariate models to non-linear and kernel based approaches, it relates to the number of parameters and calculations involved; occasionally included herein is also the necessary memory for storing the model representation, examples can range from a small vector of coefficients, or set of rules, to large weights or probability matrices; computational performance, though ordinarily not approached as a constraint, unless for very-short or real-time forecasting horizons, computational cost is a quite distinctive feature among ML algorithm, with some taking several minutes to appropriately train, and others but a few seconds or less.

4. FINAL REMARKS

Surmising, we see that the data’s behavior can be as diverse and distinctive as their application scenarios, sharing complexity features (e.g. non-linearity, high frequencies, mean shifts) that directly difficult the LF task. In consequence, application approaches, renewed and novel, largely focus their efforts to consider the many possible avenues to take for improving the quality of samples, evaluating different types of smoothing and levels of decomposition, and realize the existing types of seasonalties, as to enrich the available data with new features. It is also readily understood that there is not a standard or one-to-one relation as to ML algorithms; with a varied range being regularly applied and compared (e.g. neural networks, RBF networks, neuro-fuzzy systems, support vector machines) and also novel hybrid models steadily emerging.

This paper presented a succinct reference as the the process of DM, highlight commonly found and handled aspects in the task of LF; practices of which can be similarly applied to other problems in the area of power systems, such as energy price and wind power forecast.

Amidst the abundant literature in the area, we fundamentally encourage the newcomer to this topic, as much room is still available for contribution; as it is so generally to the energy sector. In a time marked by conscious investments toward the use of renewable energy, power deregulation, and the technology prospects found in smart grids to provide large volumes of data at a faster rate, DM and ML algorithms will continue to be well sought approaches is the pursuit for faster and more reliable results.

REFERENCES