International Journal of Pure and Applied Mathematics
Volume 117 No. 1 2017, 185-201
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: 10.12732/ijpam.v117i1.15

ELECTRIC ENERGY PRICE FORECASTING: DESCRIPTIVE ANALYSIS AND FEATURES SELECTION

Maria Teresa Grifa

Department of Engeneering and Computer Science
University of L'Aquila
Via Vetoio, 67100, Coppito (AQ), ITALY

Abstract: The present paper is focused on the analysis of electricity market, after its recent liberalization. In particular we provide a detailed analysis of the latter exploiting descriptive analysis and the feature selection approach for a multivariate time series dataset. Moreover we will apply a pool of regression models on the features selection methodology focusing our study on the *2014-Global Energy Forecasting Competition* dataset.

AMS Subject Classification: 62H99, 62P20, 68U20
Key Words: time series,electricity price forecasting, descriptive analysis, feature selection, machine learning

1. Introduction

Since the past two decades, the electricity market has been liberalized. The previous monopolistic situation has been replaced by a competitive and deregulated market in which consumers have the freedom to decide the operator from whom to receive the service. Electricity energy has very different characteristics compared to other material product, for instance it is distinctive because of its limited storage capacity and transportability. The daily demand and prices are

Received: 2017-09-08
Revised: 2017-10-27
Published: December 1, 2017

determined the day before the physical delivery by means of hourly concurrent auctions. In the day-ahead market every transaction for each hour of the next day is programmed. This means that hourly prices for the next day delivery are fixed at the same time. As a result, electricity prices disclose volatility whereas the electricity demand has strong relation with the atmospheric conditions and effects prices dynamic. It is worth to mention that accurate forecasting techniques lead to substantial savings in operating and maintenance cost and correct decisions for future development. Electricity (demand/price) data exhibit peculiar features: daily, weekly and annual periodic patterns as well as dependency on calendar effects, price jumps and mean-reversion. Different types of seasonality can be detected in electricity prices: annual, related to the seasons during the year and to the economic and social activities during different months; weekly, related to working days and weekends; and intra-day cycles, related to variations among different hours of the day. Spikes in electricity prices are sudden upward movements of the dynamic and they can be attributed to the low level of flexibility in energy markets. A spike occurs when the price exceed a specific threshold for short time range. Spikes are not homogenous within the time period, but they are mainly common during on-peak hours, namely between 9:00 and 18:00 of working days. The electricity loads fluctuations affect peak prices, and this depends both on the marginal cost supply and the demand that has seasonal behaviors. If the demand is high, a small variation of usage will cause electricity prices changes. Mean-reversion is a phenomenon such that after a price spike, prices get back to a mean level. The current forecasting methods for the estimation of electricity price are mainly based on two approaches: statistical models and automated learning techniques. Statistical methods attempt to estimate the future price from past values and they are based on both statistics instruments and stochastic analysis calculus approaches, see, e.g., citeDPPerin,AEBK. However, due to the nonlinear relationship between price and other factors, as, e.g., loads, peaks of production, meteorological exogenous influences, etc., it is difficult to obtain accurate forecast within statistical models frameworks, see, e.g., [7, 8, 9, 10, 11] An important concern about electricity price forecast is the construction and selection of a set of variables on which we aim to built our future predictive methods.

In what follows we analyze a dataset provided from the Global Energy Forecasting Competition, in 2014, the original competition goal was to predict quartile electricity prices for every hour on a given day.

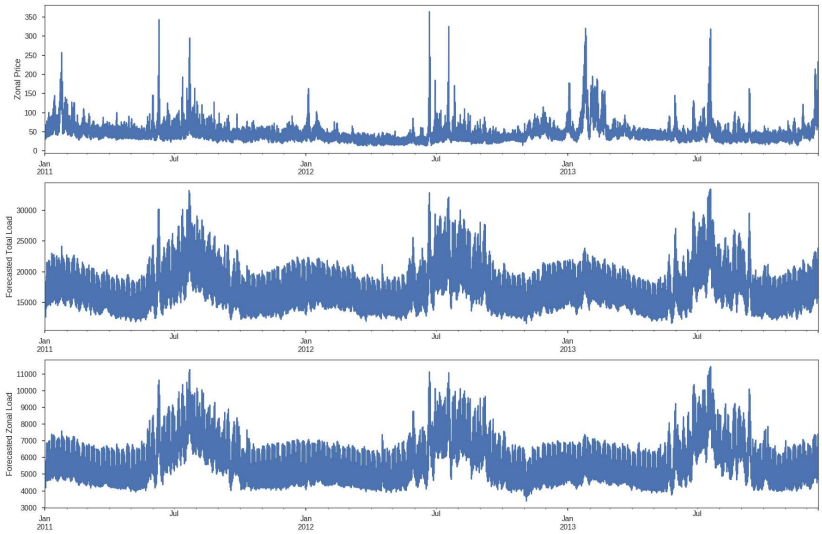

Figure 1: Time Series Plots

2. Dataset

The dataset contains three time series on hourly resolution from January 1st, 2011 to December 17th, 2013 for an unknown zone located in Australia. The target variable is the locational marginal price while two exogenous time series are: the forecasted zonal load and the forecasted total load (see Figure 1). The first variable is day-ahead predictions of zonal loads referred to the same zone of the price data. The second exogenous variable is day-ahead predictions of system loads, i.e. the forecasted total electricity load in the provider network. The unit of measurement for these variables are unknown. The dataset does not show any repetitions or missing values. Table 1 shows the descriptive statistic values for the target and the exogenous variables. We illustrate the distributions of the three time series, Figure 2. The histogram of Zonal Price (target) is bent to the left and has a havy tail on the right due to presence of spikes, in fact it reveals some unusual hight values. This suggests to take the natural logarithm value of Zonal Price variable for future modeling steps. The distribution of the other two exogenous variables are quite similar. We plot the correlation matrix

Table 1: Descriptive statistics of taget and exogenous variables

	Zonal Price	**Forecasted Zonal Load**	**Forecasted Total Load**
count	25944	25944	25944
mean	48.14603	6105.56618	18164.10330
variance	683.42024	1715538.219	11930368.109
kurtosis	20.98	0.5623	0.9568
std	26.52	1309.78	3454.036
min	12.52	3395	11544
max	363.8	11441	33449

between target and exogenous variables, Figure 3. The exogenous variables are highly correlated with a correlation value of ~ 0.97, but they are not so much correlated with the prices series itself ($\sim 0.5 - 0.58$).

3. Descriptive Analysis

In order to undestand the meaningful features of our times series, we conducted some data visualizations. We plotted the Zonal Price series locked into years. In 2011, the price series shows peak values by the second half of January, mean reverse behavior during Autumn and we can see that prices increase and report more peaks during Winter time. This behavior is almost the same whithin 2012 and 2013. We plotted autocorrelation function and partial autocorrelation function with lags = 60. The autocorrelation function allows to evaluate temporal dependency within the data, i.e. it takes the correlation of a series with is past (lagged values). The partial autocorrelation function is an extension of autocorrelation function, where the dependence on the intermediate elements (those within the lag) is removed. In this framework, Zonal Price time series shows a sine-wave shape pattern because it has a strong time persistence. It smoothly decreases over the lags but it does not cut-off completely. Moreover, it shows cyclic patter of period 24. This fact allows us to take in account hourly seasonal components.

The partial autocorrelation function plot shows spikes at lag = 2. The next few lags are at the borderline of statistical significance. The ACF plots of Forecasted Total Load and Forecasted Zonal Load show sine-wave shape with repeating patter at lag = 24 while the PACF plots exhibit significant correlation at lag = 2 and after gradually tails-off, see Figure 5. Another way to detect seasonality is via looking at the scatter plots of a time series and its lagged

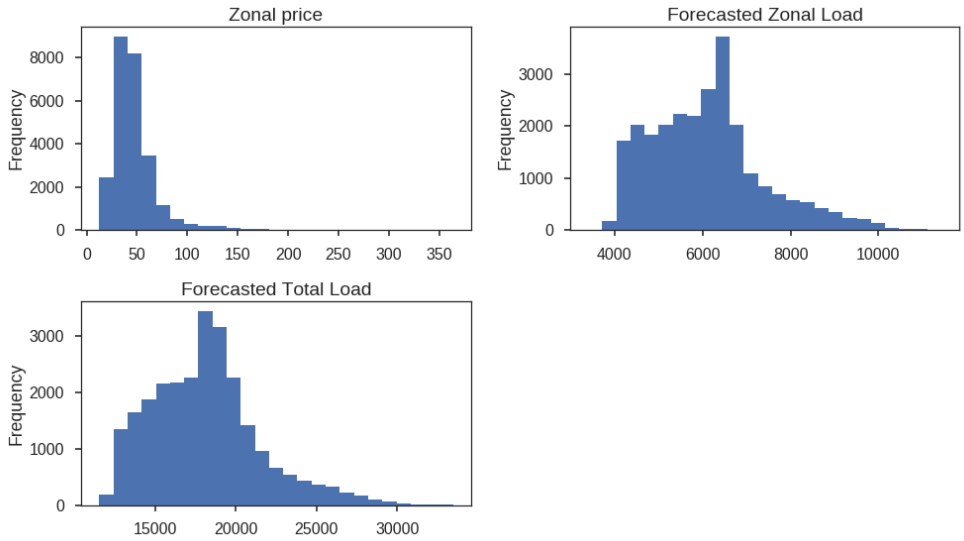

Figure 2: Time series distribution

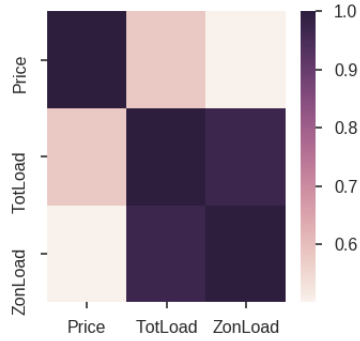

Figure 3: Correlation matrix

values. Table 2 shows the correlation values of the provided time series and some of their lags. We plotted the prices time series dynamic for every months within the three different years, we did not ended up with any significant behavior. It is possible to depict weekly, hourly and monthly seasonality using box-plots, Figure 7. The weekly box-plot highlight several outliers. Working days have similar graphs, Monday and Friday have a mean level slightly lower than mid weekdays because their nearness with weekend. Saturday box-plot is moved

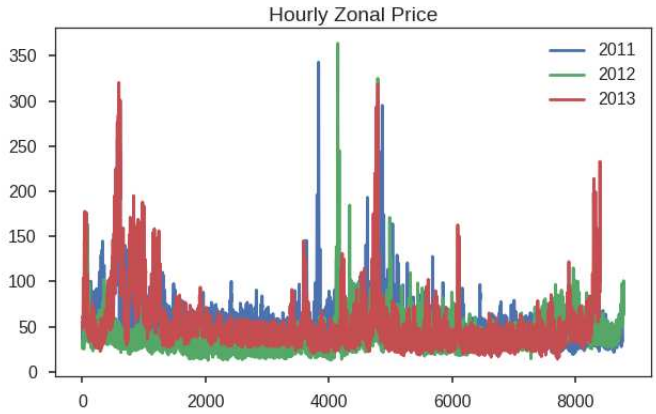

Figure 4: Zonal Price grouped by years

	Zonal Price	**Forecasted Zonal Load**	**Forecasted Total Load**
t-1	0.96	0.98	0.97
t-2	0.90	0.93	0.90
t-12	0.40	0.19	0.08
t-24	0.85	0.91	0.92

Table 2: Correlation: variable vs lagged variable

downward with respect to other days, and this features is more clear on Sunday. The variability during the week is linked to working day and weekend electricity usage. From the monthly box-plot we observed that on January, February and July the mean level is higher that other months. The box-plot of Zonal Prices grouped by hours depicts and increasing mean level during working hours. From hourly time series electricity prices, we built daily prices by taking the mean of 24 hourly data in order to pre-processing spikes phenomenon Figure 8 reports daily prices dynamic its autocorrelation function (for 30 days) and displays the hourly mean prices throughout the week and the hourly mean prices separately for working days and weekend days. These plots show:

1. strong seasonal componets within the day and the month

2. presence of multiple spikes

3. different price levels during the weekdays

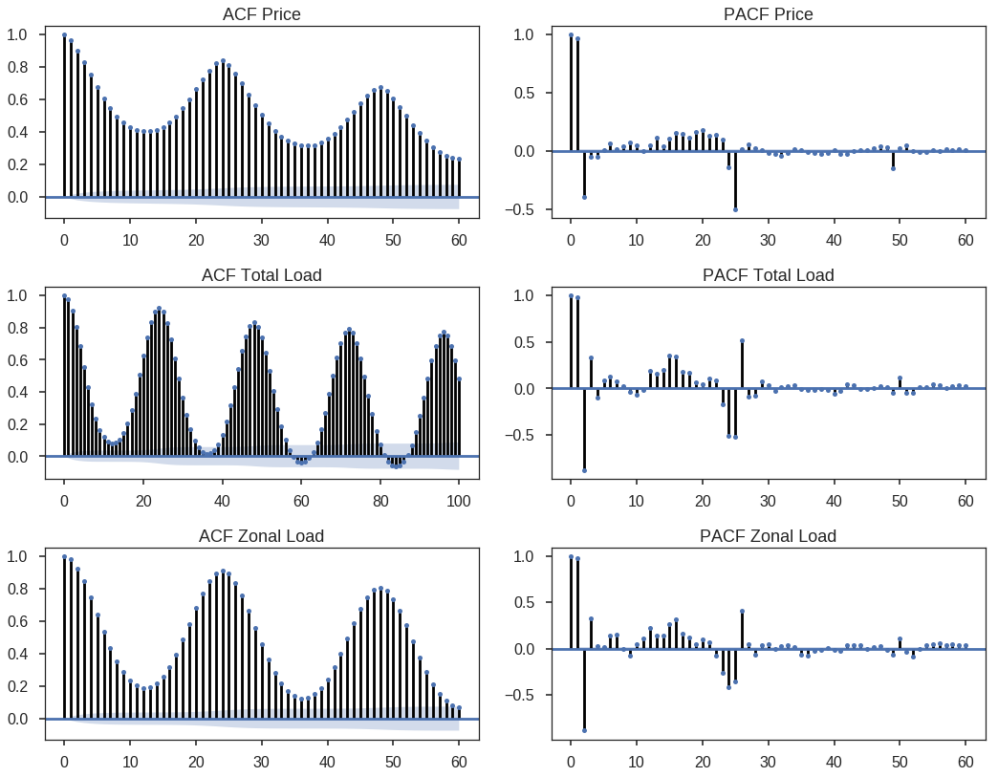

Figure 5: Autocorrelation and Partial Autocorrelation plots

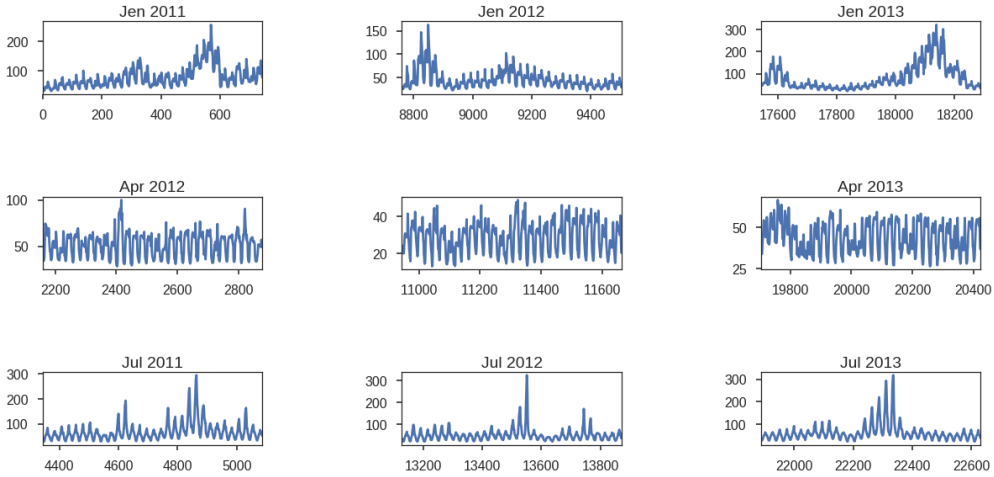

Figure 6: Monthly dynamic

4. presence of volatility cluster

We tried to understand the relationship between Zonal Price and its lagged-seasonal values. Before enter into details concerning this approach, it is worth to mention that, as outlined in [4, 5] see also references therein, seasonality/cyclic components can be also analyse taking into consideration the *regime switching* approach. In Figure 9, we plotted the autocorrelation price values in specic hours and they are shifted in days (24 hours). The autocorrelation values of night time (21:00-23:00) and early morning (4:00-5:00) are higher than afternoon hours (12:00-15:00). It seems reasonable to us to conduct the future experimental methodology using 24 time series belonging to 24 hours of the day and include in the model shifted variables.

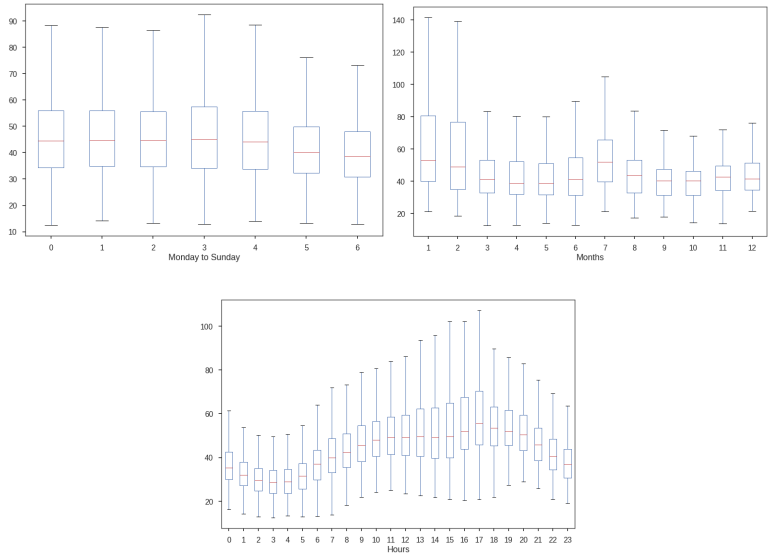

Figure 7: Box Plot-Zonal Price

4. Features Selection

In order to select the best modeling approach, we considered additional explanatory variables using the date-time range and the original times series. Figure 10 shows the correlation between explanatory variables. The main task when dealing with several variables is to select the optimal set of features. In this regard, we used the methodology reported in [3, 2]:

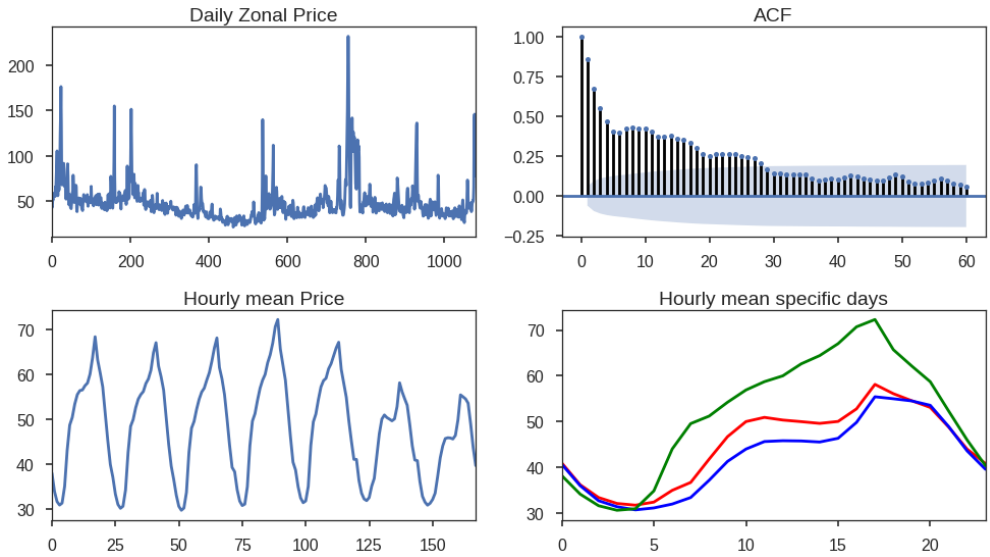

Figure 8: top left: daily prices, top right: daily prices autocorrelation plot, bottom left: Hourly average patterns for each day of the week, bottom right: hourly average patterns for working days and weekend days: red=Saturday, blue = Sunday, green= working day.

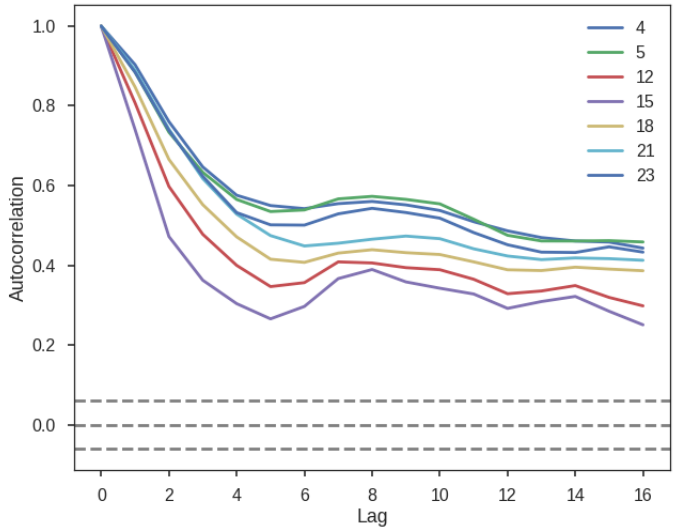

Figure 9: Autocorrelation of price in specific hours shifted within a day.

	Variable description
day	Day of the week, integer, between 0 and 6
month	Month of the year, integer between 0 and 12
yeard	Day of the year, integer, between 0 and 365
yearw	Week of the year, integer, between 1 and 52
work	Weekday and weekend, boolean values
tot24back	Total Load from 24 hours earlier
tot48back	Total Load from 48 hours earlier
zon24back	Zonal Load from 24 hours earlier
zon48back	Zonal Load from 48 hours earlier
zontotDiff	Difference between Zonal and Total Load
zon24zonDiff	Difference between Zonal and zon24back
tot24totDiff	Difference between Total and tot24back
maxP	Maximal Zonal Price within a day
minP	Minimal Zonal Price within a day
meanP	Mean Zonal Price within a day
maxZ	Maximal Zonal Load within a day
minZ	Minimal Zonal Load within a day
meanZ	Mean Zonal Load within a day

- find the correlation matrix between features,
- generate four dataset with respect to fixed correlation coefficients ($Corr < 0.5, Corr < 0.75, Corr < 0.95, 0 \leq Corr \leq 1$),
- fit the same model on the aforementioned groups and plot the ranking of variables that are considered important by the model
- take the variables in importance ranking order, fit the same method on every sub-group of the four dataset and perform the root mean square error on every sub-groups.

We use a pool of regression models: linear, ridge, lasso, random forest regression. We compared the explained methodology [2], see also [3] ,with the well known recursive feature elimination using linear regression. We applied each of the above listed methods on the four data sets and normalize the scores so that thay are between 0 (lowest rank) and 1 (highest rank). For recursive feature elimination, the top eight feature will all get score 1, with the rest of the ranks spaced equally between 0 and 1 according to their rank.

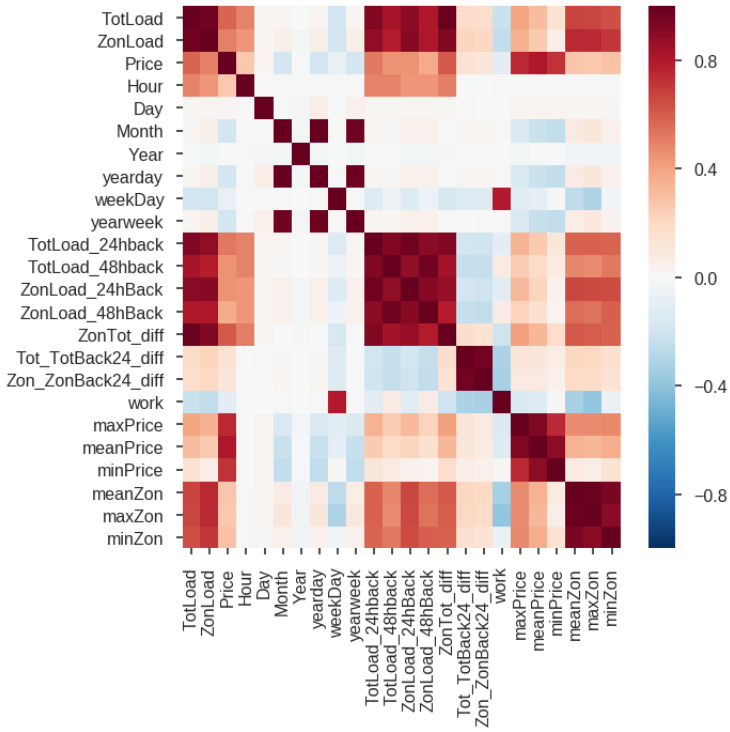

Figure 10: Features Correlation Matrix

Random Forest Regression

The ranking is base on the impurity, also known as mean decrease impurity or Gini impurity. This ranking is defined as the total decrease in node impurity averaged over all trees of the ensemble. Figure 11 shows the features scores according to each groups. When the dataset has correlated features, we can use any of these correlated features as the predictor. Once one of them is used, the importance of others is significantly reduced since effectively the impurity they can remove is already removed by the first feature. As a consequence, they will have a lower reported importance. This is an issue when we want to interpret the data. This can cause the incorrect conclusion that one of the variables is a strong predictor while the others in the same group are unimportant, while actually they are very connected with the response variable. The next figure reports the bootstrapped root mean square error for every groups

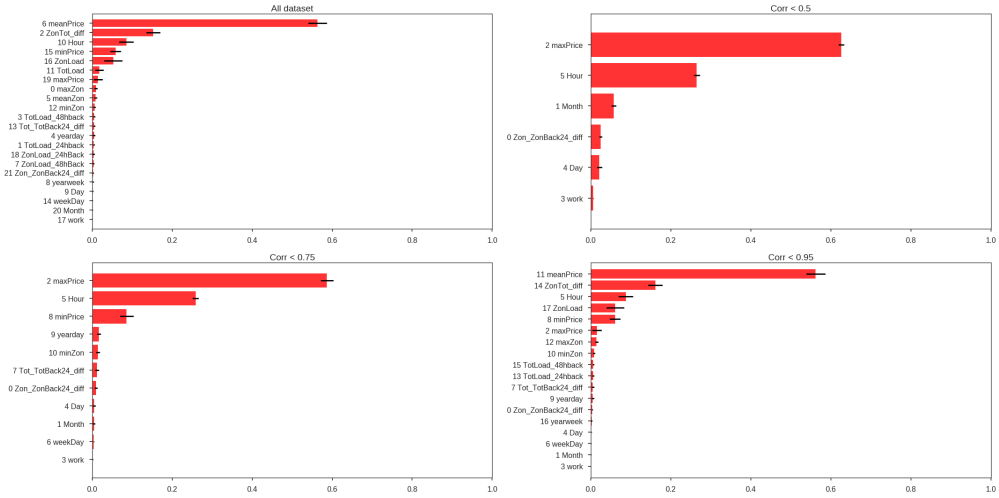

Figure 11: Features Importance Random Forest

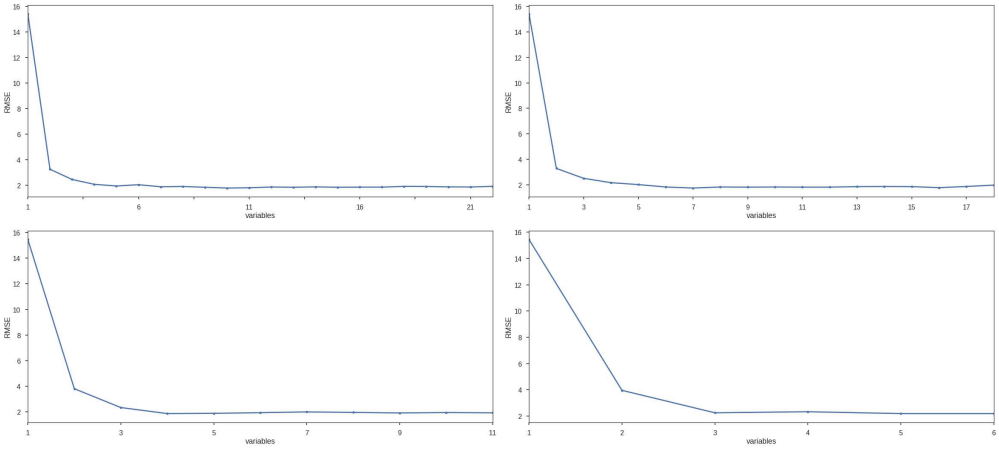

Figure 12: Accuracy of subgroups of features

Linear Regression

We used the coefficients of regression model for features ranking. When there are multiple correlated features, the model becomes unstable, i.e. small changes in the data can cause large changes in the coefficient values, making model interpretation very difficult.

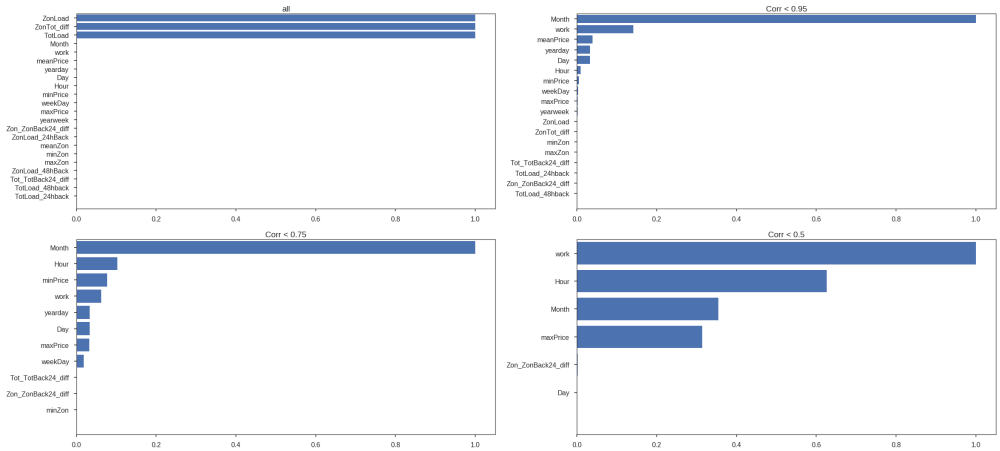

Figure 13: Features Importance Linear Regression

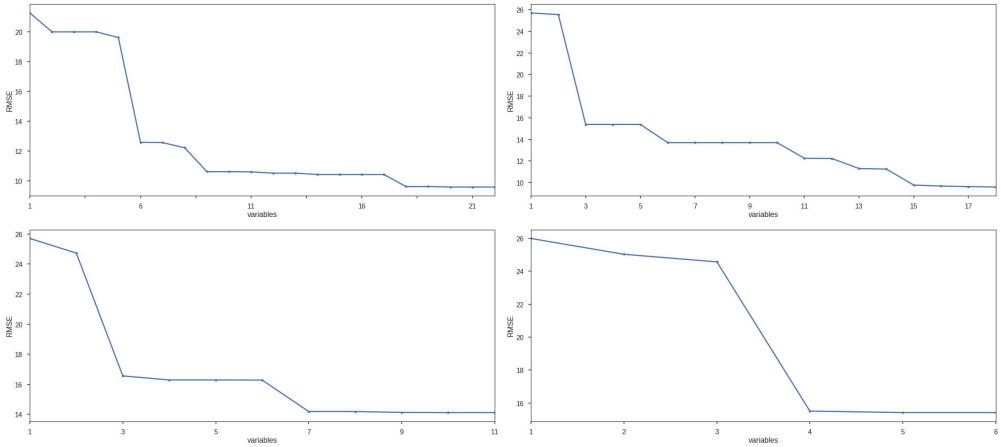

Figure 14: Accuracy of subgroups of features

Lasso

Lasso produces sparse solutions and as such is very useful selecting a strong subset of features for improving model performance. Lasso picks out the top performing features, while forcing other features to be close to zero. We used hyperparameters optimization to obtain Lasso parameter, we done it using grid search and cross-validation.

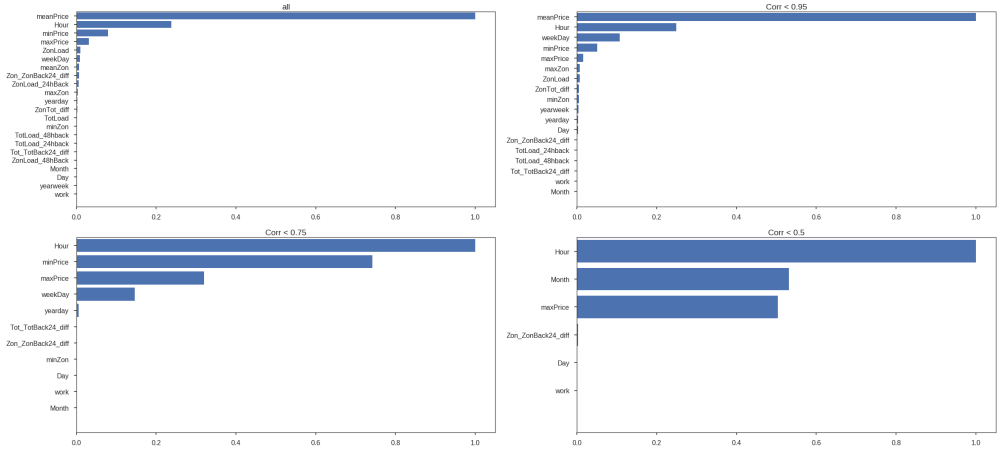

Figure 15: Variable Importance Lasso Regression

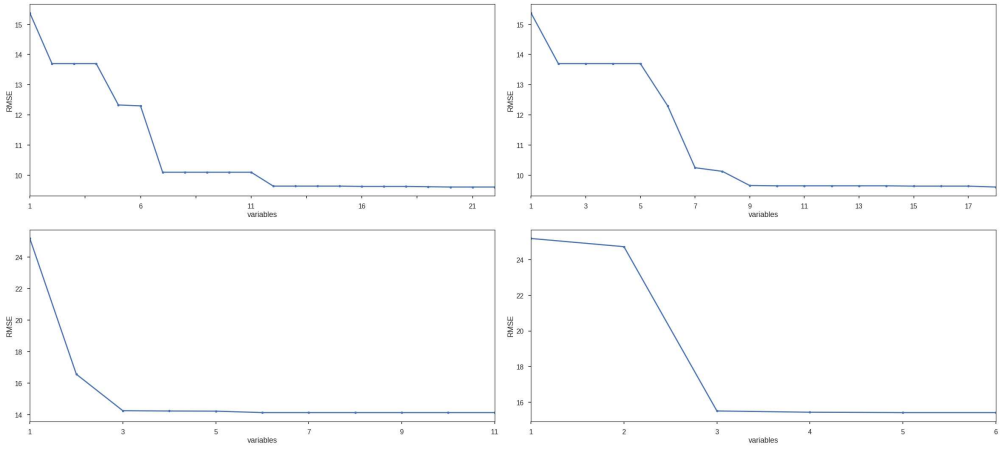

Figure 16: Accuracy of subgroups of features

Ridge Regression

In Ridge regression, useful features tend to have non-zero coefficients.Ridge regression forces regressions coefficients to spread out similarly between correlated variables. We used hyperparameter optimization to obtain Ridge parameter, we done it using grid search and cross-validation.

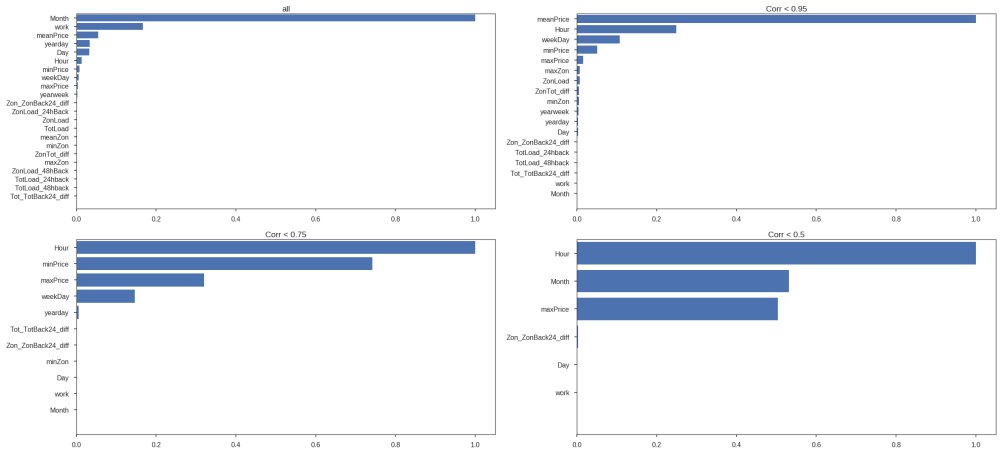

Figure 17: Features Importance Ridge Regression

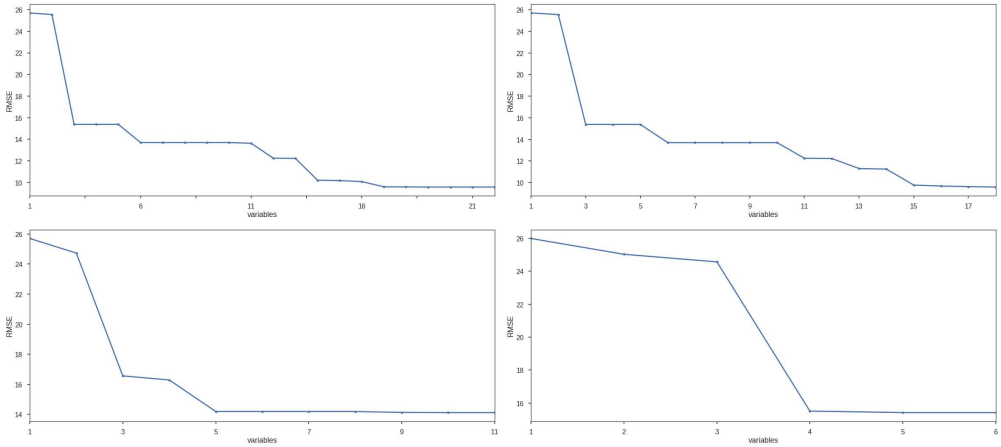

Figure 18: Accuracy of subgroups of features

Recursive Features Elimination Feat. Linear Regression

Recursive features elimination with linear regression is based on the idea to repeatedly construct a model and choose the best/worst performing feature based on coefficients, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. Features are then ranked according to when they were eliminated. The top selected features have 0 ranking.

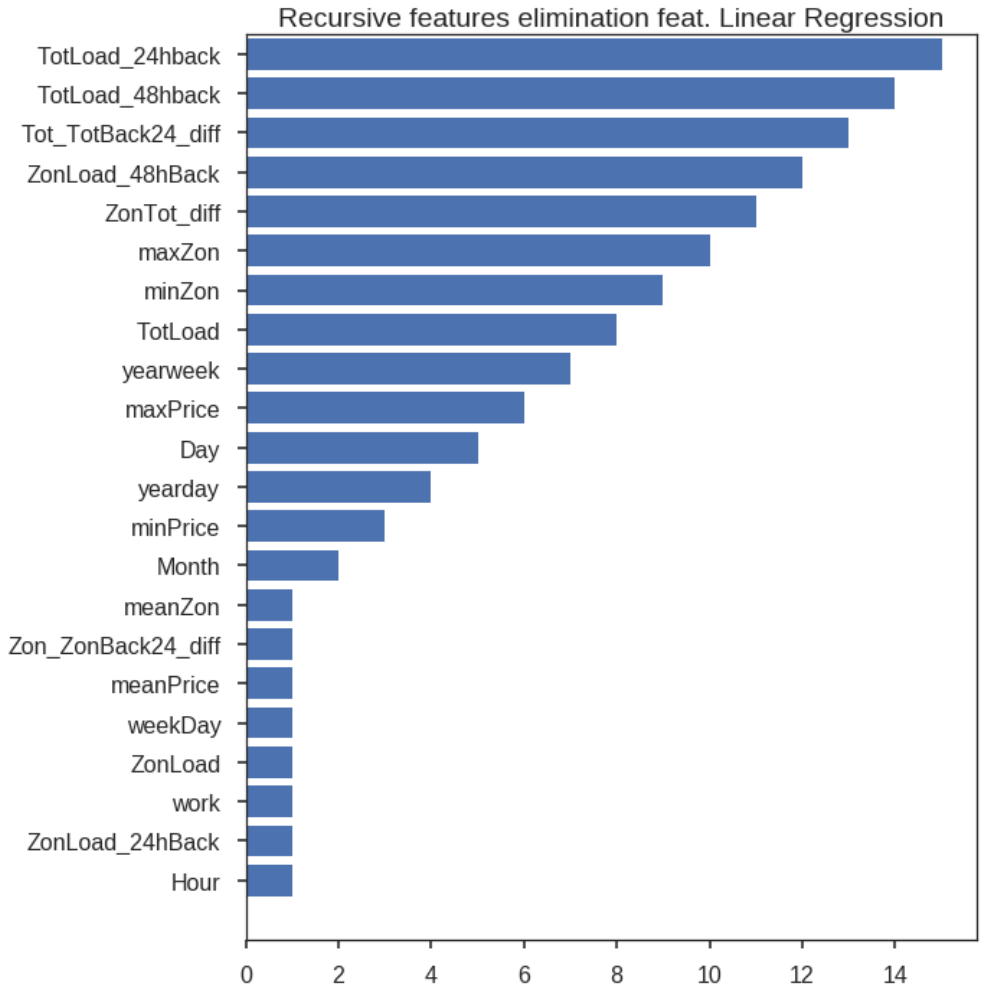

Figure 19: Last eight features are considerated important from the model

5. Conclusion

We deeply analyzed time series dataset, taken from 2014-Global Energy Forecasting Competition, with exogenous variables regarding electricity prices. We built a set of additional variables by means of the hidden structure of the dataset itself. We conducted a research of an optimal set of explanatory variables for our future forecasting models. Then, we compared different regression methods with the well-known feature elimination methods, highlighting pros and cons of analyzed solutions.

References

[1] René Aïd (auth.), Fred Espen Benth, Valery A. Kholodnyi, Peter Laurence (eds.)- *Quantitative Energy Finance: Modeling, Pricing, and Hedging in Energy and Commodity Markets*, Springer-Verlag New York (2014).

[2] F. Cordoni, *On variables selection in energy markets forecasting*, HPA (2016).

[3] L. Di Persio, A. Cecchin, F. Cordoni, Novel approaches to the energy load unbalance forecasting in the Italian electricity market, *Journal of Mathematics in Industry*, **7**, No. 1, 1–5 (2017).

[4] L. Di Persio, M. Frigo, Gibbs sampling approach to regime switching analysis of financial time series, *Journal of Computational and Applied Mathematics*, 300, pp. 43–55 (2016).

[5] L. Di Persio, M. Frigo, Maximum likelihood approach to markov switching models, *WSEAS Transactions on Business and Economics*, 12, pp. 239–242 (2015).

[6] L. Di Persio, I. Perin, An ambit stochastic approach to pricing electricity forward contracts: The case of the German Energy Market, *Journal of Probability and Statistics*,626020 (2015).

[7] V. Kaminski, *Energy markets*, Incisive Media, Risk Books Series (2014).

[8] G. Swindle *Valuation and Risk Management in Energy Markets* CUP (2014).

[9] A. Swishchuk, *Modeling and Pricing Of Swaps For Financial and Energy Markets with Stochastic Volatilities*, World Scientific Publishing Company (2013).

[10] R. Weron,*Modeling and forecasting electricity loads and prices: A statistical approach*, John Wiley & Sons, Vol. 403 (2007).

[11] R. Weron, Electricity price forecasting: A review of the state-of-the-art with a look into the future, *International journal of forecasting* , **30**, No. 4 (2014), 1030–1081.