Putting Big Data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests

Author(s): Ludwig, Nicole; Feuerriegel, Stefan; Neumann, Dirk

Publication Date: 2015-01

Permanent Link: https://doi.org/10.3929/ethz-b-000182404


Rights / License: In Copyright - Non-Commercial Use Permitted
Putting Big Data Analytics to Work:  
Feature Selection for Forecasting Electricity Prices using the LASSO and Random Forests

Nicole Ludwig, Stefan Feuerriegel, Dirk Neumann

Chair for Information Systems Research, University of Freiburg, Platz der Alten Synagoge, 79098 Freiburg, Germany

Abstract
Successful companies are increasingly those companies that excel in the task of extracting knowledge from data. Tapping the source of Big Data requires powerful algorithms combined with a strong understanding of the data used. One of the key challenges in predictive analytics is the identification of relevant factors that may explain the variables of interest. In this paper, we present a case study in predictive analytics in which we focus on the selection of relevant exogenous variables. More specifically, we attempt to predict the EPEX electricity spot prices with reference to historical prices and a deep set of weather variables. In order to choose the relevant weather stations, we use the LASSO and random forests to implicitly execute a variable selection. Overall, in our case study of German weather data we manage to improve forecasting accuracy by up to 16.9% in terms of mean average error.

Keywords: Predictive analytics, decision support, exogenous predictors, feature selection, electricity prices, weather data

*Corresponding author. Mail: stefan.feuerriegel@is.uni-freiburg.de; Tel: +49 761 2032395; Fax: +49 761 2416.
Email addresses: nludwig21@t-online.de (Nicole Ludwig), stefan.feuerriegel@is.uni-freiburg.de (Stefan Feuerriegel), dirk.neumann@is.uni-freiburg.de (Dirk Neumann)
1. Introduction

In today’s world, companies face fierce competition that obliges them to store data about their customers, their processes and the underlying business environment. Only those companies that are successful in extracting useful knowledge from such data have a chance of surviving this fierce competition. While data requirements in the past ten years were restricted to company internal data, companies are now pushed into the Big Data era (Halper, 2011; Power, 2014), where literally all available information has to be stored and subsequently processed.

From Business Intelligence development (Turban, 2011), it is known that the majority of stored data is waste; the real challenge lies in how to identify useful information. This is contrary to the conventional perception that “large data sets offer a higher form of intelligence and knowledge” and possess an “aura of truth, objectivity, and accuracy” (Boyd & Crawford, 2012). Big Data analytics requires the use of appropriate algorithms coupled with an intuitive feeling for the data being processed. Objectivity and accuracy do not come for free but need to be developed (e.g. Bendler et al., 2014) within a well-conceived process.

One crucial facet in today’s Big Data efforts revolves around prediction. A good prediction of environmental factors avoids overproduction, false production, mispricing, etc. and is thus a key activity in all companies (Boylan & Syntetos, 2012; Chen et al., 2012; Davenport & Harris, 2007). The term predictive analytics (Shmueli & Koppius, 2011) denotes the sub-area of data mining that is concerned with extracting knowledge from data in order to predict behavioral patterns or long-term trends. Predictive analytics attempts to identify the relevant relationships between explanatory variables and the predicted variables from past occurrences. Subsequently, these relationships are exploited to predict the future outcome of the variable of interest.

With the advent of Big Data (Davenport et al., 2012; IBM, 2013), which is tantamount to the availability of thousands of potential variables that may offer explanatory power, predictive analytics faces the challenge of identifying useful exogenous variables (Varian, 2014) that may offer a causal link to the variables of interest. Since the variety of potential explanatory variables is vast, manual selection is inappropriate. Taking the many challenges, such as spurious correlations, omitted variables and missing data into consideration, it is not surprising that a fully-fledged automated selection procedure does not exist; it is even questionable if such a process will ever exist. Nevertheless,
it seems that the selection of exogenous variables is still an open field in research and in practice.

This paper attempts to contribute to the selection of adequate explanatory variables by presenting a predictive analytics case study. In this study, we attempt to predict electricity spot prices by using weather data from the preceding days. As more electricity from regenerative sources is fed into the power system, we expect electricity prices to react to weather changes. Unlike previous studies, we abstain from using aggregated weather data but instead use data from many distributed weather stations. Since the local weather proxies local electricity generation from wind-mills or photovoltaic installations, we expect the weather to act as a good predictor for future prices. Nonetheless, as not all areas in a country generate renewable power, our predictive analytics approach needs to identify those stations that are relevant and discard all other stations.

![Figure 1: Predictive analytics incorporating exogenous parameters to improve electricity price forecasts.](image)

In our analysis (see Figure 1), we employ the LASSO (least absolute shrinkage selection operation) and random forests to execute the feature selection process. While the LASSO (Hastie et al., 2013; Tibshirani, 1996) adds a regularization term to the OLS estimator in order to keep the coefficients of the regression small, it implicitly conducts a feature selection, since variables are inherently driven to zero. Random forests (Breiman, 2001; Hastie et al., 2013) are also very useful in feature selection: they randomly create
many small and thus generalizable decision trees to explain the data. Variables that are frequently used in different trees may exhibit a measurable effect on the variable we would like to predict. This relevance information can be condensed by a so-called variable importance metric. The case study shows that we can greatly improve our electricity price forecast by employing weather data from several weather stations. Despite the fact that our data sample may not be large enough to be dubbed Big Data, it offers an intriguing insight into the selection of variables for Big Data analytics.

Interestingly, a recent MISQ review claims that “predictive analytics are rare in mainstream IS literature, and even when predictive goals or statements about predictive power are made, they incorrectly use explanatory models and metrics” (Shmueli & Koppius, 2011). In the domain of electricity price forecasts, the existing literature (cf. Aggarwal et al., 2009; Weron, 2006) focuses predominantly on classical time series analysis and statistical learning. What these works all have in common is the fact that they frequently ignore the predictive power of external parameters, such as weather conditions. As shown in Figure 1, making use of more accurate price forecasts and implementing an effective strategy can provide a competitive advantage (Davenport & Harris, 2007; Lett, 2004; Keen, 1981; Turban, 2011).

The remainder of this paper is structured as follows. Section 2 provides a literature overview of publications that forecast electricity prices, in which the majority of models ignore external impacts. To close this research gap, Section 3 utilizes both time series analysis and statistical learning to present models that incorporate exogenous predictors. These models are then evaluated in Section 4, which reveals that the implemented external inputs improve forecasting accuracy at a statistically significant level.

2. Related Work on Predictive Analytics for Electricity Prices

Electricity prices reveal several (unique) characteristics, for example, they exhibit hourly specific mean-reversion (Huisman et al., 2007). Further features (Cartea & Villaplana, 2008; Weron, 2006) include high volatility, strong seasonal patterns and price spikes. Altogether, these characteristics excite particular requirements that models of electricity price forecasts must meet.

In this paper, we restrict our analysis to short-term forecasts that cover time spans of a few hours to a few days, which are common in day-to-day market operations, for example, adjusting the production schedules of a power plant (Weron, 2006). Electricity exchanges typically work in the form of
hourly day-ahead auctions and, thus, the short-term forecasting of next-day prices is of major importance (Misiorek et al., 2006). To support day-to-day market operations, this paper focuses on short-term forecasts of day-ahead prices. This provides “companies with valuable information when facing their decision making and risk-management process.” (Bello & Reneses, 2013).

In addition to forecasting horizons, one can also classify price forecasting according to methodology. In the literature, countless forecasting strategies are proposed in order to estimate the future development of spot prices. This section does not attempt to make a thorough presentation of the known techniques but rather provides a general classification. Existing approaches can be broadly divided into six groups (Weron, 2006): cost-based models, game theoretic approaches, fundamental (or structural) methods, econometric models, statistical approaches and artificial intelligence-based techniques.
<table>
<thead>
<tr>
<th>Reference</th>
<th>Market</th>
<th>Time</th>
<th>Price</th>
<th>Model</th>
<th>Forecast Comp.</th>
<th>Re-Estimation</th>
<th>Out-of-Sample Predicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contreras et al. (2003)</td>
<td>California, Spain</td>
<td>2000</td>
<td>Hourly day-ahead</td>
<td>ARIMA, ARIMAX</td>
<td>✔ ✔ ✔</td>
<td></td>
<td>Load, daily production of hydro units</td>
</tr>
<tr>
<td>Erni (2012)</td>
<td>EPEX</td>
<td>2010–2012</td>
<td>Hourly day-ahead</td>
<td>GARCH, threshold regression, time-varying parameter regression</td>
<td>✔ ✔ ✔</td>
<td></td>
<td>Load, wind speed, weather (e.g., sunshine duration, cloud cover), prices (CO₂, etc.), power plant availability</td>
</tr>
<tr>
<td><strong>This paper</strong></td>
<td>EPEX</td>
<td>2013/14</td>
<td>Hourly day-ahead</td>
<td>ARMA, ARMAX, LASSO, random forest</td>
<td>✔ ✔ ✔</td>
<td></td>
<td>Temperature &amp; wind speed across multiple stations</td>
</tr>
</tbody>
</table>
Recently, researchers have introduced spot price models that include *exogenous variables*. This approach is motivated by the assumption that price patterns are the characteristic result of underlying fundamentals (Erni, 2012). Examples of these fundamentals are, for instance, demand and supply (Cartea & Villaplana, 2008). In addition, “variables to consider in forming expectations of prices are ... the weather” (Bello & Reneses, 2013). Consequently, this motivates the incorporation of external variables in electricity price models.

Previous research (Kosater, 2006) notes that “surprisingly, weather, as one important driving force of electricity demand and hence electricity prices, has been neglected so far”. In fact, we are aware of only a few studies that actually deal with models that employ weather as an external predictor. All publications found are listed in Table 1. When looking at these references in detail, we observe that weather-related time series only appear in an aggregated form, mostly averaged across several weather stations. To improve on that, we gather local weather information from several weather stations instead of using a single, aggregated value.

Furthermore, Contreras et al. (2003), as well as Weron & Misiorek (2008), measure the beneficial effect of re-estimating models for each trading day, though abstain from a detailed comparison. In addition, none of the papers analyze to a sufficient extent how the addition of external parameters can result in improved accuracy. However, day-ahead spot markets are highly volatile and exhibit a substantial seasonal pattern, while also being strongly dependent on external weather conditions. In summary, none of the aforementioned references study the positive influences on forecasting accuracy when weather information from individual stations is included in the model.

Hence, this paper – based on our earlier work (Feuerriegel et al., 2014) – proposes and compares models to forecast hourly day-ahead electricity prices with actual weather data as an exogenous predictor. Thus, we outline a forecast framework that includes time series describing both temperature and wind speed as measured at multiple weather stations. In addition to that, we implement a rigorous evaluation of out-of-sample errors and perform a feature selection. Based on this, we identify and measure the beneficial effects of incorporating weather as an exogenous predictor. To allow for time-dependent parameters, we also use a moving window of training data in order to study the rolling re-estimation of coefficients.
3. Predictive Analytics Models for Electricity Prices

Once we have obtained the parameters from our training set, we can plug these parameters into the final models, which are confronted with the out-of-sample data from the testing data set. The resulting forecasts are compared in terms of the mean average error. Statistical tests are subsequently performed to show that differences in forecasting quality are not subject to randomness.

In addition, this section provides a mathematical introduction to the models used for electricity prices. Following the existing literature, we start our analysis with two benchmark models: a naïve approach and an ARMA model. Following this, we add exogenous predictors to our forecasting models, resulting in an ARMAX model from time series analysis and, furthermore, random forests. In the following, let \( p_t \) with \( t \in \{1, \ldots, N\} \) denote the time series of electricity prices. In order to include external variables, let the \( k \) exogenous time series be given by \( W_t^{(1)}, \ldots, W_t^{(k)} \), which measures the wind speed at \( k \) weather stations. Furthermore, the inputs in the form of measured temperature across \( k \) stations are denoted by \( T_t^{(1)}, \ldots, T_t^{(k)} \).
3.1. Benchmark: Naïve Approach

Following Conejo et al. (2005) and Misiorek et al. (2006), we employ a naïve but challenging test to verify that our proposed models are better than random guessing. According to this test, we set the forecast $\hat{p}_t$ to $p_{t-168}$, which is the historic price from 7 days before time $t$. Any proposed model passes the naïve test if the error is smaller than the error of the naïve approach. Misiorek et al. (2006) report that in some time periods, almost all of their models have trouble passing this test. The reason why the naïve model provides such a good estimate is due to the strong repetitive pattern of intra-day and intra-week seasonality of electricity prices.

3.2. Time Series Analysis: ARMA/ARMAX Model

In contrast to the above linear regression models, time series analysis focuses on approaches that also take into account correlations within the series under study. Thus, in time series analysis, electricity prices are not merely modeled as a linear combination of exogenous predictors but as a linear combination of past observations of the time series and, possibly, past errors. Extensions can also incorporate exogenous variables.

Autoregressive-moving-average (ARMA) models describe a stationary stochastic process in terms of two summands: one for auto-regression and the second for the moving average. More precisely, the process is modeled as a linear combination of previous values and past errors. In that case, both the mean and autocovariance of the series are independent of time. The notation ARMA($p,q$) refers to a model with $p$ autoregressive terms and $q$ moving-average terms given by

$$p_t = \vartheta + \sum_{i=1}^{p} \varphi_i p_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \beta D_t^{\text{Weekend}} + \varepsilon_t$$  \hspace{1cm} (1)

with suitable coefficients $\vartheta$, $\varphi_1$, ..., $\varphi_p$, $\theta_1$, ..., $\theta_q$, $\beta$ and white noise error terms $\varepsilon_1, \ldots, \varepsilon_t$. These coefficients are estimated by maximum likelihood estimation (MLE). Let $L$ denote the lag (or backshift) operator $L^h p_t = p_{t-h}$. Thus, the ARMA($p,q$) process can be rewritten more concisely as

$$\varphi_p(L) p_t = \theta_q(L) \varepsilon_t + \beta D_t^{\text{Weekend}}$$  \hspace{1cm} (2)

with

$$\varphi_p(L) = 1 - \sum_{i=1}^{p} \varphi_i L^{(i)} \quad \text{and} \quad \theta_q(L) = 1 + \sum_{i=1}^{q} \theta_i L^{(i)}. \hspace{1cm} (3)$$

9
When it comes to determining the ARMA orders $p$ and $q$, one frequently relies on the minimization of an information criterion, such as AIC (Akaike Information Criterion) or BIC (Bayesian Information Criterion).

The autoregressive-moving-average model with exogenous inputs (ARMAX) extends the classical ARMA model by an additional linear combination of exogenous variables. The notation $\text{ARMAX}(p,q)$ with weather variables as external influences refers to the process

$$p_t = \vartheta + \sum_{i=1}^{p} \varphi_i p_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \beta D_{\text{Weekend}} + \sum_{i=1}^{k} \gamma_i W_{t}^{(i)} + \sum_{i=1}^{k} \delta_i T_{t}^{(i)} + \varepsilon_t$$

with suitable coefficients $\vartheta$, $\varphi_i$, $\theta_i$, $\beta$, $\gamma_i$ and $\delta_i$, as well as white noise error terms $\varepsilon_1, \ldots, \varepsilon_t$. Obviously, ARMAX models can be helpful tools for modeling electricity prices when prices are dependent not only on past values but also on exogenous inputs, such as weather variables.

3.3. LASSO

As a result of e.g. multicollineary, the ordinary least squares (OLS) technique can result in a low predictive power. To overcome this problem, regularization techniques shrink some coefficient estimates towards zero; this increases the bias but improves the forecasting accuracy. A common choice (Hastie et al., 2013; Tibshirani, 1996) of regularized least squares versions is the so-called LASSO (least absolute shrinkage and selection operator) defined by

$$\hat{\beta}_{\text{LASSO}} = \arg\min_{\beta} \sum_{i=1}^{N} \left[ y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right]^2 \quad \text{s. t.} \quad \sum_{j=1}^{p} |\beta_j| \leq t, \quad (5)$$

or, similarly,

$$\hat{\beta}_{\text{LASSO}} = \arg\max_{\beta} \left( \frac{1}{2} \sum_{i=1}^{N} \left[ y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right]^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right). \quad (6)$$

Because of the $L_1$ penalty term, the LASSO typically produces estimates in which some of the coefficients are set exactly to zero (Hastie et al., 2013). Thereby, the LASSO performs a feature selection. Selecting a good value of $\lambda$ for the LASSO is critical; in practice, the parameter $\lambda$ is tuned using cross-validation. In fact, the LASSO technique has been applied to electricity markets in which, for example, drivers of electricity consumption (Wang et al., 2007) and electricity prices (Suard et al., 2010) are identified.
3.4. Random Forests

In addition to the classical time series models introduced above, we include random forests (Breiman, 2001; Hastie et al., 2013) from statistical learning. These are robust to errors resulting from the inclusion of predictors that do not contribute to the model by performing feature selection (Kuhn & Johnson, 2013). Random forests are an ensemble learning method that construct a large collection of de-correlated decision trees during training. As only continuous values are relevant in this case, the output is the average value over all individual trees. The training algorithm for random forests applies bagging (also named bootstrap aggregating) to the single tree learner. Here, bagging repeatedly selects a total of $B$ bootstrap samples of the training set and fits trees $t_b$ – using the Gini impurity – to these samples. As notation, we introduce inputs $x = [x_1, \ldots, x_N]^T \in X$ and responses $y \in Y$ for training to estimate the model parameters. Then, after training, predicted values for unseen examples $x$ are calculated by averaging the predictors from all individual regression trees via

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} t_b(x).$$

(7)

In the above algorithm, the number of bootstrap samples $B$ is a free parameter. Interestingly, increasing the number of trees\footnote{In all subsequent evaluations, we use $B = 500$ trees.} tends to decrease the variance of the model without increasing the bias.

3.5. Forecasting Accuracy

To assess the predictive power of the proposed models, we utilize different statistical metrics to measure the forecasting accuracy. Let $\tilde{p}_t$ with $t \in \{1, \ldots, N\}$ denote the predicted time series. We can then check the out-of-sample forecasting accuracy afterwards, once the true market prices $p_t$ are available. For the time horizon under study, we calculate the following prediction errors, typically used in the literature on electricity price forecasting (see e.g. Conejo et al., 2005; Keles et al., 2013; Misiorek et al., 2006). Common measures are the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE), which are defined as:

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^{N} |\tilde{p}_t - p_t|$$

and

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (\tilde{p}_t - p_t)^2}.$$
Squared Error (RMSE) given by

$$\text{MAE} = \frac{1}{N} \sum_{t=1}^{N} |p_t - \tilde{p}_t| \quad \text{and} \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (p_t - \tilde{p}_t)^2}. \quad (8)$$

A common scale-independent measure is the Mean Absolute Percentage Error (MAPE) defined as

$$\text{MAPE} = \frac{1}{N} \sum_{t=1}^{N} \left| \frac{p_t - \tilde{p}_t}{p_t} \right|. \quad (9)$$

Although this measure has the benefit of giving the proportional difference between forecasted and true values, it is not useful for our data because values approach infinity when prices become close to zero.

4. Evaluation: Measuring the Influence of External Variables

This section evaluates the previously specified models for forecasting electricity prices: we present the data and the forecasting framework. Subsequently, we compare the performance of different prediction models and demonstrate that the inclusion of exogenous predictors in the form of weather data improves forecasting accuracy.

4.1. Data and Descriptive Statistics

We integrate data sources\(^2\) as follows. The electricity prices used for our analysis are the German and Austrian day-ahead spot prices from January 01, 2013 to April 30, 2014, retrieved from the European Power Exchange (EPEX SPOT). Since market operations are carried out 24 hours a day, 7 days a week, this time span results in a total number of 11,640 observations. During this period, the prices range from $100.03/\text{MWh}$ to $130.27/\text{MWh}$, with a mean price of $36.60/\text{MWh}$. As one can see in Figure 3, this time series is characterized by high volatility and price spikes, as well as strong seasonal patterns.

\(^2\)Missing spot prices, e.g., due to time conversion from Central European Time to Coordinated Universal Time, are handled by duplicating the previous value. Missing weather values are set to zero.
In addition to electricity spot prices, we use wind and temperature data as exogenous predictors. The necessary data is obtained from the Global Data Set of the German Meteorological Service, DWD (Deutscher Wetterdienst) for short. The DWD, a public German institution, is legally obliged to provide information about the German weather. Since approximately 8% of the 2011 German gross electricity production\(^3\) originates from wind turbines, we choose to include wind speed as one of our exogenous predictors. We use the data from 69 weather stations providing hourly averages in m s\(^{-1}\). As listed in Table 2, the hourly wind speed ranges from calm air (0 m s\(^{-1}\)) to hurricane force (>32 m s\(^{-1}\)), with an average over all weather stations of 4.04 m s\(^{-1}\).

Additionally, we include the data of 75 weather stations providing hourly temperature averages in °C. Hourly temperatures range from −23.70 °C measured at Zugspitze on October 2, 2013 up to 37.60 °C measured in Rheinstet-

---

ten on July 27, 2013. The average temperature over all stations is 7.73°C. Figure 4 provides an overview of the weather stations used in the analysis; note that all stations are fairly evenly distributed across Germany.

Figure 4: Weather stations in Germany used in our analysis, where red points mark stations providing either only temperature or only wind data, whereas black labels represent stations providing both.

Table 2: Descriptive statistics of day-ahead electricity prices and exogenous predictors in the form of weather data. Values for the latter are given for the joint combination of all weather stations.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Day-Ahead Price</td>
<td>€/MWh</td>
<td>36.60</td>
<td>15.98</td>
<td>−100.03</td>
<td>130.27</td>
<td>0.13</td>
<td>2.59</td>
</tr>
<tr>
<td>Hourly Temperature</td>
<td>°C</td>
<td>7.73</td>
<td>7.49</td>
<td>−23.70</td>
<td>37.60</td>
<td>0.29</td>
<td>−0.33</td>
</tr>
<tr>
<td>Hourly Wind Speed</td>
<td>m s⁻¹</td>
<td>4.04</td>
<td>2.19</td>
<td>0.00</td>
<td>34.70</td>
<td>0.98</td>
<td>1.35</td>
</tr>
</tbody>
</table>

A prerequisite of the ARMA/ARMAX model is that the underlying time series must be stationary. As a statistical test, one commonly relies on the augmented Dickey-Fuller test (ADF), with the null hypothesis of the time series not being stationary. For electricity spot prices, this hypothesis test
returns a test statistic of $-6.97$ with a $P$-value $< 0.01$. Thus, we can assume stationarity at common significance levels. We also consider the weather to be stationary since the climate does not change in such a short time frame.

4.2. Predictive Analytics Framework

In this section, we introduce the framework on which our predictive analysis is based. We compare the aforementioned models according to their accuracy on out-of-sample data. Therefore, it is necessary to specify a set of training data to estimate our model and a set of test data to check the predictive accuracy of our models. We use a training set of 365 d, such that our initial training set is the year 2013, leaving us with a test set of 2880 h, corresponding to the first 120 d in the year 2014 (i.e. January 1 to April 30). Following our previous work (Feuerriegel et al., 2014), we iteratively forecast 24 h ahead due to the auction design at electricity exchange.

When estimating the parameters of our models, we can choose between two options: (i) coefficients can be determined based on a training set and then remain fixed throughout the whole testing set. Alternatively, (ii) coefficients can be re-estimated daily for each forecast using a moving window of training data. Each re-estimation takes a new training set into account, shifted by one day. This idea is referred to as a rolling re-estimation of parameters (c.f. Feuerriegel et al., 2014; Hu et al., 1999). Figure 5 compares both approaches graphically, where the upper section presents the idea of fixed parameters and the bottom shows how parameters are re-estimated based on a shifted training set. Though the re-estimation increases computational requirements, one can claim that this approach improves accuracy since the model is updated according to new price information.
4.3. Variable Importance

Since we have a large number of possible predictors, we need to determine those predictors that are important for our forecasting accuracy. In order to do so, we use random forests as a variable importance measure. The increase in the mean squared error (MSE) over all external inputs ranges from 6.5 to 197.14, with the 20 best variables having a value of above 26. As we can see in Figure 6, the inclusion of the lagged prices $p_{t-168}$ has the greatest impact on the MSE. Interestingly, the best weather stations that are included are those of the wind stations. The first temperature station is ranked 22nd. A possible explanation might be that far more wind power is fed into the grid than solar power (Erni, 2012). Altogether, the clear variation in the importance of variables motivates and justifies the development of models incorporating feature selection mechanisms, detailed in the following section.
4.4. Results: Model Comparison

In this section, we compare our models according to their performance in terms of forecasting accuracy. The models vary in the length of the training periods, the usage of re-estimation and the set of included externals. We denote the lagged prices with $p_{t-24}$, $p_{t-48}$, $p_{t-168}$ and a dummy for weekends with $D_{Weekend}$. The set of external weather stations used differs implicitly according to the method of variable selection. All results are reported in Table 3.

As a first striking fact, we see that rolling re-estimation clearly improves forecasting accuracy. In fact, re-estimation decreases the mean absolute error for all models. For example, the MAE for the random forest in model (6) drops by 1.78% from 5.779 down to 5.676. Thus, we report only errors originating from daily re-estimation in the following discussion.

As a second observation, we notice that the best-performing model is given by a random forest. This model (6) achieves a mean average error of
5.676, followed by the LASSO with 5.712. On the one hand, both models behave almost equally well. The random forest results in a MAE that is around 1% lower, possibly because random forest can model non-linearities. On the other hand, the LASSO shrinks the search space and, thus, features an overall estimation time\textsuperscript{4} of a staggering 3082.08 s compared to 7210.66 s – an increase of 134.0%. Altogether, the two best-performing models are those that implicitly incorporate feature selection.

Furthermore, we can see that exogenous predictors in the form of weather data considerably improve forecasting accuracy. The best benchmark model, i.e. the random forest in model (3), yields a MAE of 6.834. When we compare this benchmark with the two aforementioned best-performing models, we receive the following results: the MAE reduces from 6.834 to 5.676 or 5.712, respectively – a drop of up to 16.9%. When looking at e.g. the root mean squared error (RMSE), the results show the same direction. Here, errors decrease by up to 7.89%. This supports the hypothesis that electricity prices nowadays are highly dependent on weather characteristics. Overall, we see that all models benefit substantially from the inclusion of exogenous predictors.

\textsuperscript{4}Estimation times are measured on an Intel Core i5-2400 CPU at 3.10 GHz with 4 GB RAM and a 64-bit Windows 7.
<table>
<thead>
<tr>
<th>Model</th>
<th>Lagged Prices</th>
<th>Dummies</th>
<th>Externals</th>
<th>Without Re-Estimation</th>
<th>With Daily Re-Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>(2) ARMA(1,1)</td>
<td>$p_t - 24$, $p_t - 48$, $p_t - 168$</td>
<td>$D_{Weekend}$</td>
<td>—</td>
<td>6.151</td>
<td>220.989</td>
</tr>
<tr>
<td>(4) ARMAX(1,1)</td>
<td>$p_t - 24$, $p_t - 48$, $p_t - 168$</td>
<td>$D_{Weekend}$</td>
<td>All</td>
<td>5.809</td>
<td>156.259</td>
</tr>
<tr>
<td>(5) LASSO</td>
<td>$p_t - 24$, $p_t - 48$, $p_t - 168$</td>
<td>$D_{Weekend}$</td>
<td>Implicit Subset</td>
<td>5.810</td>
<td>160.347</td>
</tr>
<tr>
<td>(6) Random Forest</td>
<td>$p_t - 24$, $p_t - 48$, $p_t - 168$</td>
<td>$D_{Weekend}$</td>
<td>Implicit Subset</td>
<td>5.779</td>
<td>164.420</td>
</tr>
</tbody>
</table>

Table 3: Comparison of forecasting accuracy between January 01, 2014 and April 30, 2014 across various models.
Finally, in order to provide statistical proof that the models which include external weather data more accurately predict electricity prices, we apply the Diebold-Mariano (DM) test (Diebold & Mariano, 1995). The null-hypothesis tests whether the models with weather data are at least as precise in their forecasts as those without weather data. When comparing models (4) to (6) with all benchmark models, all $P$-values remain below 0.001. Hence, we can reject the null hypothesis in all cases at all significance levels. In short, we conclude that those models that include exogenous predictors in the form of weather data achieve a superior predictive accuracy than models without.

4.5. Managerial Implications

The way in which companies master the task of Big Data analytics will become the most important edge in the context of fierce competition. Only those companies that succeed in the efficient extraction of knowledge will remain competitive over a longer time period. This particularly holds true for prediction. The companies that initiate better predictions will enjoy a rare advantage (Davenport & Harris, 2007; Lett, 2004; Keen, 1981; Turban, 2011).

As Niels Bohr once conceived: “predictions are difficult, especially for the future”, this is not an easy task. One remedy is to employ as many variables as possible in the estimation that may potentially have a causal link to the variables of interest. However, this requires good algorithms (cf. Varian, 2014) that automatically support the selection of variables. On the other hand, the computational requirements of the algorithm should not be too extensive.

In this paper, we contribute to the quest for algorithms that are suitable for Big Data analytics. More precisely, we attempt to improve the predictive quality of electricity spot prices by feeding in not only historical prices, but also detailed weather information harvested from many locally dispersed weather stations. The intuitiveness of the inclusion of weather data is straightforward, as the weather directly affects electricity generation from regenerative sources (Erni, 2012). Conversely, the weather may also impact the electricity consumption patterns of the users. While the general strategy is undisputed, it is difficult to determine which weather data is of interest for the prediction.

Our case study of the German electricity market ultimately unveiled the following four managerial findings that are noteworthy to any data analyst interested in predictions:
Finding 1. The inclusion of external variables considerably increased the prediction quality (almost 17% in terms of mean average error). This finding is not as astonishing when our constructed causal link is taken into account. Nevertheless, we report this finding as it bolsters the strength of Big Data analytics. Critics of the Big Data analytics approach, such as the renowned economist Bruno Frey, argue against the inclusion of extensive external variables (Cueni & Frey, 2014). Their critique pointed to the observation that the prediction of physical processes will be improved when more information (i.e. variables) is included – but this does not (necessarily) hold for social processes. Our example contradicts this proposal, although we share their caution in the exaggeration of Big Data on the general impact of predictions.

Finding 2. A rolling re-estimation procedure is necessary for all algorithms. Otherwise, all algorithms – albeit very sophisticated – fail to improve on their one-time estimation counterparts. While the necessity of rolling re-estimations is obvious, it is not so clear that static predictions perform poorly. The implications of this finding is that the applicability of computationally intensive algorithms, such as neural networks, is diminished since they require the training of a new neural network for each day. To ensure robustness using cross-validation, the training effort consequently becomes extremely high. Even more importantly, rolling re-estimation is a remedy against the time-varying influence of weather on electricity prices and, to some extent, the reaction issue (Cueni & Frey, 2014) described by Bruno Frey is solved.

Finding 3. Algorithms that perform an automated variable selection are superior, in terms of mean average error, to other approaches. Our results suggest that the LASSO and random forests are adequate for use in Big Data analytics. Interestingly, such methods could also be used to identify important variables – so-called features – that are subsequently fed into support vector machines or in connected restricted Boltzmann machines to perform deep learning. As feature selection establishes itself as a crucial task in Big Data analytics, we strongly promote the use of these algorithms in predictive analytics.

Finding 4. Feature selection reduces the computational burden (see Guyon & Elisseeff, 2003) of the algorithms, as the problem size is pruned. Since
computational tractability is the key to analyzing big data, this observation supports Finding 3, favoring methods which embody automated feature selection.

While our algorithm suggestions are not comprehensive, the quest for adequate algorithms will be continued. Nonetheless, we provide the reader with the recommendation of two different methods: the LASSO and random forests.

5. Conclusion

Today’s competition coerces companies to invest in Big Data analytics technologies, in order to master knowledge extraction from stored data. While Big Data promises sustaining success, it is in many cases deceptive. The reason for this stems from the vast amount of waste data, which algorithms have to discard in their analysis. Effectively, Big Data analytics promises a lot but, in most cases, will never be achieved due to missing best practice processes combined with sophisticated algorithms. One of the crucial aspects an algorithm needs to support is the automated identification of important variables. In this paper, we contribute to the development of such algorithms by testing the LASSO and random forests against econometric algorithms from time series analysis.

In our case study, we attempt to forecast electricity prices using local weather data, such as wind and temperature, at various weather stations. This paper analyzes how LASSO and random forests (out-)perform in their forecasting of electricity prices compared to alternative econometric approaches. More specifically, we employ the autoregressive-moving-average model with exogenous predictors (ARMAX) from time series analysis. Furthermore, we use feature selection to improve our ARMAX model. The findings demonstrate that by including exogenous parameters, we can achieve more accurate forecasts. In addition, we statistically verify via the use of a Diebold-Mariano Test that the exogenous predictors yield better performance. More precisely, the mean average error (MAE), as a measure for forecasting accuracy, drops by up to 16.9%. The LASSO and random forests work particularly well in their forecasts due to their ability to select the right variables as a byproduct of their algorithm. Accordingly, we strongly advocate the use of the LASSO and random forests for use in Big Data analytics.
In future work, we will advance the above methods in three directions. First, our rigorous predictive analysis would benefit greatly from the comparison of additional models. Examples of other advantageous approaches are jump-diffusion and regime-switching models, Artificial Neural Networks, Support Vector Regression or autoregressive models with conditionally heteroskedastic errors. Second, extending the selection of external predictors would open an avenue for further gains in accuracy. Thus, it might be beneficial to include additional exogenous predictors, such as electricity load and detailed weather data like wind direction or precipitation. Third, the models could also be tested for the prediction of intraday electricity prices.

Acknowledgment

The valuable contributions of Laura Cuthbertson and Simon Riedlinger are gratefully acknowledged.

References


25


