St. Petersburg University

Graduate School of Management

Master in Information Technologies and Innovation Management

Prediction and decision making support in Dutch electricity market using big data techniques: a comparative analysis of different approaches

Master's Thesis by the 2nd year student

Concentration — MITIM

Zhao Yuchen

Research advisor:

Associate Professor, Maria M. Smirnova

St. Petersburg

[2016]

Content

Abstract

Big data analytics in energy industry is a relatively new topic, but a number of cases in different countries have already been studies for it can create real value for the producers or the consumers of energy. The prediction of energy market situation in order to provide decision making support is a widely studies area, however for Dutch electricity market, related studies are not many. The Dutch electricity market is quite different from the situation of other countries, first of all it has a power exchange where the energy price is dominated by supply and demand, and an electricity operator that regulates the price of electricity based on the market output -imbalance. Meanwhile, renewable energy such as solar energy, wind energy and greenhouse is on a rise in the liberalized Dutch energy market, and those green energy output can be freely traded in the market. Our study aims to look into the situation in Dutch electricity market by means of big data analytics, to examine and compare different data mining methods of market situation prediction and to examine if weather can affect the price of electricity in Dutch market, and hence provide decision support for energy purchase for those small electricity operators in the Dutch market who have access to two different market prices. As a result, we have found that big data mining techniques outperforms traditional data analytics method, and exist certain data mining technique that can give us the best decision making support in selecting bidding strategy of electricity. Meteorological data, seemingly irrelevant to the market price fluctuation, indeed has power of predicting market condition in the Netherlands. Current study is practice-oriented and has strong managerial implication, a conclusion can be drawn by our research is that the models we built in our study can actually help operators from private sectors to save energy purchasing cost thus have the value to be further developed.

Keywords: Big data, data mining, electricity price prediction, Dutch power market

Аннотация

Аналитика Больших данных в отрасли энергии является относительно новой темой, но исследования уже были проведены для разных случаев в некоторых странах потому, что такие исследования имеют большие значения для производителей и потребителей энергии. Прогнозирование ситуации на рынке энергии в целях поддерживать принятия решений для бизнеса – тема широко изученая, однако для голландского рынка электроэнергии, связанных с этим исследований немного. Голландский рынок электроэнергии отличается от того в других странах, в первую очередь, тем что в Голландии есть биржа электроэнергии, где цена энергии определяется спросом и предложением, и в Голландии главный оператор электроэнергии регулирует рынок с помощью цены-несбалансированность, что определяется несбалансированностью между спросом и предложением электроэнергии на рынке. В то же время, доля возобновляемой энергии, такой как солнечная энергия, ветровая энергия и гидроэнергия, увеличивается в этом рынке, и такие энергии можно торговать свободно в Голландии. В данной работе мы анализируем ситуации в голландском рынке с помощью аналитики больших данных, изучаем и сравняем различные методы аналитики данных для прогнозирования рыночных цен в этом рынке, определяем влияют ли метеорологические факторы на цены электроэнергии. Следовательно, мы сможем поддержать операторам, имеющим доступ в две разных цены, в принятии бизнес-решении о покупке электроэнергии. В результате нашей работы мы обнаружили что методы интеллектуального анализа данных превосходят традиционый метод аналитики данных, и существует определенный метод интеллектуального анализа данных, который дает нам лучшее результат в прогнозировании рыночного состояния. Метеорологические данные, оказалось бы не имеют отношение к цене на рынке, на самом деле влияют на колебания рыночной цены в Голландии. Данная работа ориентированна на практику и имеет значения для менеджеров в принятии решении в покупке электроэнергии на голландском рынке, модель прогнозирования, созданные в данной работе, могут помогать операторам сократить затраты покупки электроэнергии и зато имеют значения для дальнейшего исследования.

Ключевые слова: Большие данные, интеллектуальный анализ данных, прогнозирование цены электроэнергии, голландский рынок электроэнергии.

Introduction

Research background

With the wave of information technology innovation and the impact of computer technology evolution, the information and the size of the data people can take advantage of is growing exponentially, that information are gathered based on cloud computing, the internet of things, and the development of social media, and are rapidly converging into an enormous amount of raw data. Data mining, data analysis and the utilization of big data is not something new in the field of computer science, in fact, the IT industry is never lack of fact and precedents of the processing of massive data and information, however the concept "big data" came into people's awareness only in the most recent years, when the study of big data has finally been connected to the progress of human society.

The recipient of Turing Award Jim Gray pointed out already in 2007 in his talk to the Computer Science and Telecommunications that data capture, curation and analysis, or the data-intensive paradigm is going to be the fourth paradigm of scientific research(Gray, 2012); Chris Anderson(2008), chief editor of the Wired magazine, asserted that Big Data is going to make the scientific method "obsolete" and this is the "End of Theory". Viktor Mayor-Schonberger(2013) in his book Big Data: A Revolution That Will Transform How We Live, Work, and Think pointed out how big data have changed our life in business and state of mind, especially he mentioned that some value can be unleashed from big data analysis by a shift from causation to correlation.

Although big data is being praised for changing our lives, there are also some doubtful voices in the field. The Havard business Review has concluded that good data sets don't always result in brilliant decision support. In fact, for most population, they either put too much faiths in data science or simply don't pay enough attention to the power of data, only few understand how to use data scientifically (Liang et al, 2013). Meanwhile the big data approaches are also criticized for the incapability to explain the essential laws of natural and social phenomena,

which is the core debate of the big data approaches: whether we can address questions based on correlation relationship without understanding the causation relationship behind things.

Because of the necessity for big data analysis in addressing different unfamiliar data and their unknown impacts to the task at hand as a whole, it is usually unclear about the correlation relationship between variables, not mentioning the causation relationship which is sometimes pure experiences and assumptions. This situation has made data mining the most applicable research method for big data analytics, for its multiple functions such as classification, association analysis and more. And with varied data mining techniques under each function, a great range of relationships can be discovered by big data mining techniques.

Despite all those uncertainties, the concept of big data analysis and data mining approaches still attracts public awareness, and there have been tons of attempts of utilizing these methods for value creation in different fields and industries. For instance, Amazon has been collecting data of the purchase record and the bowering habits of their clients, by analyzing those data, the company was able to provide personal recommendation for purchase to their customers, thus increase the company's operating income. DataSift company, by means of sentimental analysis of posts on Twitter, was able to provide possible trend prediction of Facebook stock price. However, Data mining, although is capable of discovering correlations, still requires analysts to understand the full picture of the situation, for the correlation discovered can mean nothing at all if there is not a reasonable assumption based on causation relationship to back the case up.

Energy industry, as a specific industry whose condition is influenced by a big variety of factors, requires great effort of data analysis, in particular big data analytic methods, considering that many of this industry's influencing factors and the impact of those factors are still waiting to be discovered. So when dealing with big data in energy industry, is not only about what data to use, but also about why we use those data.

How to process more and more available data is the next question. According to Mansi (2014), the increasing availability of big data in energy industry has add towards the importance of applying big data techniques in this field. In fact, big data techniques have already been studied and applied for decision support in companies and institutes in many energy sectors including oil and gas, electricity and other. Data mining techniques is most widely applied

method, many of its functions including classification, clustering, neuron network and genetic algorithm have already been applied for providing feasible solutions for mining and exploiting information crucial for management decision making and achieving productivity gains in the energy industry (Mansi, 2014).

Our main question to be addressed in current study is about electricity price pattern prediction, which is a very direct approach of extracting value from database, and is of many practices and cases of big data mining studies in energy industry, since under the restructuring of electric power industry, different participants namely generation companies and consumers of electricity need to meet in a marketplace to decide on the electricity price (Schweppe, 1988) and with some natural characteristics of the electricity market such as non-storability, high vitality and inelasticity of demand in short term, price-forecasting tools are essential for all market participants for their survival under new deregulated environment (Sanjeev et al, 2009). Moreover, the electricity market is believed to be subject to many different factors, including market structure, market uncertainties, behavioral indices and temporal effects, the true impact of many of those factors on the market price fluctuation is yet to be studied (). Dutch electricity market, for its typical market structure and the fact that green energy is being directly traded in this market, draws our special attention.

Research scope and research questions

There are not many countries in the world with deregulated electricity market that has different electricity prices to charge, most of those countries are European countries. The Netherlands, in particular, has a power exchange where the energy price is dominated by supply and demand, and an electricity operator that regulates the price of electricity based on the market output imbalance. Meanwhile, renewable energy such as solar energy, wind energy and greenhouse is on a rise in the liberalized Dutch energy market, although the Netherlands is not even on the list of top renewable energy producing countries (Bal, 2013) in Europe, it is still presumable that the uprising of the quantity of greenhouses still has an impact on the electricity output, thus influence the market price. (Hong T., Chang W. K., Lin H. W, 2013)

However, there is still no thorough study about the detailed impact of the continuous rising

in renewable energy output, for example private greenhouse operators, on the electricity market price. Although there are already some researches indicating the influence of renewable energy to the Dutch electricity market, this is still a relatively new topic especially considering the particular electricity market structure in the Netherlands, moreover, previous researches did not imply decision making support in business in Dutch energy industry, thus no practical implications and suggestions can be made such as should company utilize traditional statistical methods or big data analytic methods to support its decision making process, as well as when data mining techniques are being used, which specific functions and related techniques of big data mining should be chosen.

Current study has a focus in the energy market, in specific, the prediction and decision making support for those players of electricity market in the Netherlands. The study also pays attention to a particular market phenomenon in the Dutch electricity market, that is the growing number of greenhouse operators in the Netherlands in selecting whether and when they should trade the electricity produced by their greenhouses in the imbalance market directly or should they trade on Amsterdam energy exchange which takes day ahead bidding and has a more volatile APX price.

Power market is under the influence of many different factors, many of which are not directed related to the price fluctuation, for this reason, in this research it is desirable to study this particular situation by applying data mining techniques in this particular situation incorporating big data, in our case we propose time-series and meteorological big data, for the analysis of market price, trying to find out whether meteorological data has influence on the market price, and to identify best analytical method and data mining prediction models for such situation for decision making support. In this research we also try to compare data mining methods with traditional statistical method that has already been out into utilization by some companies in this particular market.

The general research question to be answered in this paper is: Is there a proper method of price prediction analysis to be applied for business decision support in Dutch electricity market.

To answer this question, some sub questions need to be answered as well:

- *Does big data methods outperform traditional method in this particular case.*
- *Is there are specific big data model that outperforms other models under this particular circumstance.*
- *Does big data analytics result meet our previous assumption that meteorological data as well as time-serial data impact market price in the Netherlands.*

Chapter 1. Big data analytics: application and research methods

1. Big data analytics: an overview

1.1 Definition and characteristics of big data

The term "Big Data" was officially appeared in the 1998 in the article “A Handle for Big Data” in magazine "Science", However, according to scholars, by that time the concept "Big Data" is still not the Big Data as we see today (Liu et al., 2013). The real beginning of the academic awareness of Big Data is the proposition of data-intensive paradigm in 2007 (Gray, 2012), followed by a special issue of Nature magazine about Big Data (Nature, 2008), in 2011 Science also launched a special issue “Dealing with Data”(Science, 2011), that is when begins the era of big data. Nowadays, the academic circle, the business world and the government have all begun to understand and to get involved in big data, academic institutes, enterprises and states have all launched their Big Data development strategy (Zhao et al, 2013; Ostp, 2012). The figure below illustrates the rapid growth of public awareness of "big data" on Google(Google, 2016).

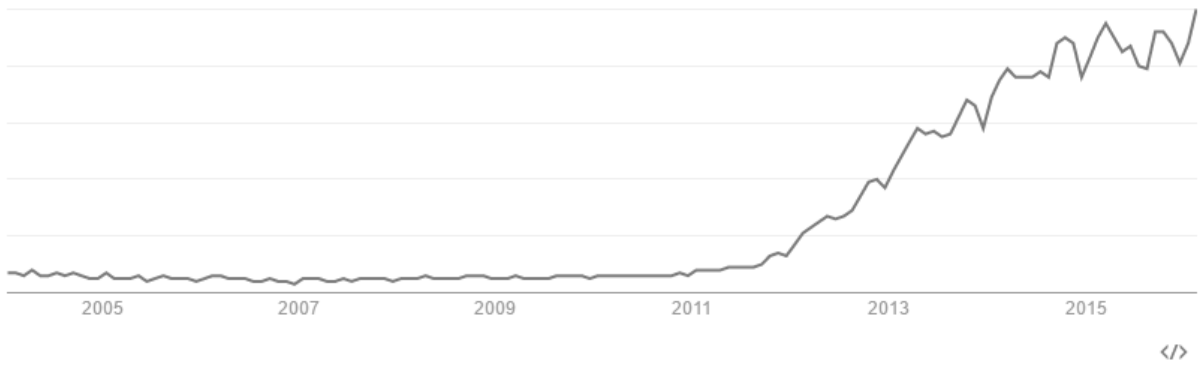


Figure 1. Searching record of keyword “Big data” on Google, retrieved in 2016.

Although the concept has already attracted lots of attention from the academia, from industries and from the government, there still lack a perfect definition of it.

In a report made by McKensey&Company Big Data: The next frontier for innovation, competition and productivity, which spurred the trend of big data analysis, the definition was made as: datasets which exceed the capability of, this definition contains the following

1. The size of the data sets is increasing over time and will increase as a result of technology revolution.

2. Different organizations and institutions have different view on how big should the dataset be. (Mckinsey&Company, 2011)

American IT consulting firm Gartner holds the view that, Big data is a kind of asset that has big volume, high velocity of generation and diversity of data type, it requires new cost-efficient data processing techniques in order to better enhance insight and automate decision making. (Lapkin, 2012)

As the first institution to pay attention to the uprising of big data at the government level, the Office of Science and Technology Policy in its governmental report <Big Data research and development Initiative> made the point that Big data is the capability of extracting valuable information from huge and diversified data. The key topic of Big data should be how to make advantage of the data asset to serve the interests of individual life, enterprise decision-making and state governance. (Ostp, 2012)

Chinese scientist Zhu Yangyong (2009) made his definition on data science, claiming that data science is “The theories, methods and techniques of researching and exploring the data environment.” He noted data environment as the research objective of the data science, which aims to “understand the types, conditions, properties and regular patterns of the data, thus to understand the patterns of natural environment as well as human behavior”. Big data, accordingly, is a kind of data science.

“Fourth paradigm: data-intensive scientific discovery” attributes special attention to big

data analytics, and refers to big data problems as “Data-Intensive Research”. Data-Intensive research usually consists of three basic activities, the collection of data, the management of data and the analysis of data. In data-intensive research, according to Jim Gray, data is no more the result of scientific researches, it has become its basis as well as the objective and tool of it. In a scientific point of view, the big data phenomena doesn’t mean only a mass of data, but also indicates the change in role of data in academic studies.

To sum up, in defining big data, it is most voiced that the data should be in big volume and velocity (Lapkin, 2012), and the big diversity is also one key element of big data (Ostp, 2012). Many researches have attributed big data analysis as kind of data science, and have identified the basic requirements of solving big data questions as mentioned by Lapkin (2012). At present the most widely accepted theory of describing big data is the 4 “V” theory, which we will address below.

Volume, namely the volume of data, is the first, and in certain occasions, the only factor that leaps out at the mention of big data. According to International data corporation (IDC), in every single month Google will process more than 400PB of data, Facebook can create more than 300TB of post data in a single day. Chinese E-business giant Taobao generates 20TB of transaction data in a single day and still increasing. Science laboratories also generates millions of data. Pan-STARRS(Panoramic survey telescope and Rapid Response System) project creates over 2.5PB of data every year, and LHC has 50-100PB data produced annually.(Hey et al, 2012)

Variety, simply the diversification of data. Big data comes in different types, including structured, semi-structured and unstructured data, in modern world the generating speed of unstructured data is surpassing that of the structured data, which caused the ascending need of new data processing and cleaning methods.

Velocity of big data has two dimensions. Firstly the speed of data generating; secondly the generated data requires just-in-time process, otherwise the data can lose its value. For instance, e-business data requires immediate analysis or no up-to-data suggestions can be provided because the customer changes mind very quickly.

Value, another dimension of big data. From one side, there can be hidden value behind huge

amount of data that can be extracted by data mining techniques; From the other side, the density of value is relatively low in big data, for only data in great scale has the capability of revealing hidden patterns of the object of study.

Those characteristics of big data questions has actually differentiated big data questions from traditional statistics in many different ways, according to (Davenport et al, 2012), the organizations that capitalize on big data differ from traditional data analysis environment in that they pay attention to data flows as opposed to stocks, and they rely on professional data scientists and product and process developers rather than data analysts who majors in statistics, most importantly, in those who capitalize in data science, big data analytics have already been moved towards a core position, providing decision support for core function in business like operation and production.

And this is actually the most difficult part of big data analytics: data don't talk for themselves, big data analytics requires self-updating data, professional skills from data scientists and the brains that can implement the findings in real life. Moreover, it is believed that the biggest difference between big data methods and traditional data science approaches lies in that in traditional approaches people incorporates only data that is directly related to the question to be solved, while in big data analytics people not only include data that is directly relevant to the case, they also put data that is not directly related to the task at hand. This is the so-called the debate of correlation and causation. In other word, the data dealt with in big data analysis holds more correlation relations other than causation relations.

1.2 Previous research on Big data questions

Big data is a relatively new topic, therefore researches and studies about big data is still in an early stage, and many big data related questions are still being explored. However, a number of important researches have already been done. Existing some important studies in this field that give us many insightful conclusions to provide guidance in our own study.

In western countries researches about big data began earlier, and more detailed and in-depth. But the time when big data questions really draw attention of the public is after the article

by Anderson(2008) “The End of Theory” in which the author claimed that the era of petabyte has come. He has three main ideas: First, with significant amount of data, data can “talk for themselves”. Second, all data science models are wrong and fail to describe the real world, yet some of them are useful and can provide insight from particular aspect. The most debated point in his article is that he claims correlation to be more important than causation relation, which has drawn lots of deputes afterwards by the academia.

The research carried out by Microsoft, The Fourth Paradigm: Dara-Intensive Scientific Discovery (Hey et al, 2012) is the first work to depict how Big data has revolutionized our life from the view of researchers, in this book the author illustrated how data flow has changed the way of scientific researches. The book begins with Jim Gray’s claim that data-intensive research is the fourth paradigm of scientific discovery, and provided an insight of some changes that is taking places in the scientific world due to the emergence of big data questions.

Right after Gray’s claim, Werner Callebaut (2012) discussed in his work the future of data-driven researches. He made his point that the viewpoint made by Anderson is untenable since theories cannot be explained and replaced by data models, correlation relationship and causation relationship cannot be comprehended as one. Werner also asserted that Scientific methods should be multidimensional, traditional methods and new born big data approaches can coexist, and in scientific researches, we should incorporate every research method that fits our purpose.

But when we accept big data mining approaches as a methods of data science research, some issues occur. Wolfgang Pietsch (2013) in his study discussed some points of attention of data mining as the key research method of big data analytics, he concludes that big data mining methods is different from computer simulation methods and traditional statistics. By comparing the extracted knowledge from data mining with the knowledge we perceived to be true through causation relationship of things, Pietsch refuted the idea that causation relationship is of minor importance in big data analytics, and that big data approaches lacks explanation power due to its nature that it focuses more on digging for correlation relationship. The author also points out that modelling in data-intensive researches tend to be “horizontal” — lacking of hierarchical, nested structure that is familiar in conventional research methods. The most inspiring finding of Pietsch is that he concluded that prediction models using data mining techniques is capable of noticing

patterns which are not easily explained by existing theories, hence sometimes building a prediction model without theoretical support can lead to useful, unexpected findings.

In terms of prediction models in data-intensive researches as a methods of data mining, Vasant Dhar (2013) discussed the topic in several ways. He has concluded that the prediction model is being more and more applied in data driven industries and the internet for the abundancy of data, and the big data mining approaches can not only be used to find solutions, but also be used to find problems. And it is important to know that when evaluating new knowledge extracted from big data mining approaches that is to be applied in decision making support, emphasis should be made on the predictive power of the prediction model other than the descriptive power of the model to historical data.

Interestingly enough, although big data approach is practice oriented, many previous researches focus in the debate of the rationality of the big data approach, more precisely whether correlation relationship weighs more compared to causation relationship in big data analysis. After Anderson's (2008) assertion, there have been many debates around this topic. Viktor Mayor-Schonberger in his book asserted that the correlation gives us more insight into the questions, but causation is relationship is necessary in analyzing the question. And many have been criticizing Anderson's point by pointing out that it is wrong to just ignore the causation relationship when tackling questions (Callebaut, 2012; Pietch, 2013). Some other researches have questioned the rationality of big data mining models for they incorporate too much variables without solid theory background (Pietch, 2013). Nowadays, after years of discovery, people have agreed on that big data approach is a useful tool of discovering knowledge, however, it lacks explanation power when we focus only on digging correlation relationships. In fact, Pietch (2013) and Todri (2015) have all reached to the conclusion that classical techniques, which incorporate only relevant data for analysis, in collaboration with big data analysis techniques which focus on the discovery of hidden relations, usually can come up with more fruitful and accurate results. In practice, this mean we should understand the whole picture before we run the analysis, and we need to incorporate not only directly relevant data.

2. Data mining as method of big data analysis

With the booming of information technology, the volume of data to be stored, the speed all types of data are generated and the speed of computers that are used to process data has grown dramatically, and traditional statistical methods already don't meet our requirement of data analysis. Data mining techniques, which is developed for big data research, is becoming the most widely applied approach of analyzing big data in the modern world.

The nature of data mining is the process of extracting extract hidden unknown, yet useful and reliable knowledge from massive, incomplete raw data. The raw data can be both structured, like tables and charts from the database, or semi-structured even unstructured, like texts and pictures. And the way of extracting knowledge can be with or without statistical support, can be both inductive or deductive. The extracted knowledge can be helpful for information management, optimization of procedure, decision-making support and process control. Data mining is a broad interdisciplinary tool which has been applied in many areas.

2.1 Definition of data-mining and related concepts

Although history of data mining is still short, many breakthroughs have been made based on its application. The concept of data mining has many different definitions thanks to its interdisciplinary application (Han J et al, 2001), SAS institute, leader of business analytics, business intelligence and data management, concludes data mining as "the model building and data exploring method based on big amount of relevant data" Bhavani (1998) puts more emphasis on tools and techniques: "The process of discovering new pattern, mode and trend in massive data by applying pattern identifying, statistical and mathematical techniques." Hand et al(2000) make the definition simpler by saying "Data mining is the process of exploring valuable information in data base of big scale".

All previous stated definitions define the nature of data mining as pattern discovery within datasets, which is the most commonly accepted description of data mining: "A complex process of extracting unknown, valuable knowledge such as pattern and mode from massive data". Besides, some refer to data mining process as knowledge discovery in databases (KDD). There is

an important distinction between data mining and KDD, that is data mining is usually focusing in the discovery part within KDD, which aims to find untouched information that holds potential value and build models for predictions. The results of data mining can be used in KDD in order to extract knowledge of the real case. Considering that the pattern discovered in data should serve the purpose of solving real questions, the use of data mining results should always influence and inform the data mining process itself. (Provost, Foster, and Tom F, 2013)

Besides the key definition of data mining about how it works and the main goal data mining is to achieve, there are also some other concepts based on previous researches which address data mining questions from the detail, those concepts are generated as results of previous researches and can be instructive for us to deal with real big data questions: (Provost, Foster, and Tom. F, 2013). For example, the identification of informative attributes requires the user to be able to select data that is correlated with or informative about an unknown quantity of interest. The complexity of the data mining model should be restricted and it is necessary to find a trade-off between generalization and over fitting, which are typical mistakes in data science.

2.2 Standard process of data mining

Data mining is being applied in different areas, and the detailed process differs in accordance with the task at hand. The data mining process is interactive and iterative, involving numerous steps with many decisions made by the user (Fayyad, 1996). Brachman and Anand (1996) made a practical view of the data mining process, exploring the interactive and iterative nature of the process.

The knowledge discovery in databases is commonly defined in several stages, we broadly outline the basic steps of the knowledge discovery in database. (Fayyad, 1996).

1. Developing an understanding of the application domain, relevant prior knowledge and the goals of the end-user.
2. Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed. The integration of data is in this step.

3. Data cleaning and preprocessing. This may include the removal of noise or outliers, collecting necessary information to model or account for noise and developing strategies for handling missing data fields or accounting for time sequence information and known changes.
4. Data reduction and projection. This step is to reduce the effective number of variables so as to find valid representation of the data. This step should be under the guidance of the KDD goal we aim to achieve.
5. Choosing the data mining function. Classification, clustering, etc.
6. Choosing the data mining method. Algorithms applied are defined here.
7. Mining the data to find valuable information
8. Interpreting mined patterns.

Then we can move to the last step of data mining, that is to act on the extracted knowledge, whether by incorporating the knowledge directly or to carry out further research on it. We may need to double check if the knowledge extracted is contradictory to previous gained knowledge. Worth noticing is that in the very first data mining process definition, Fayyad (1996), pictured the knowledge extraction process for the first time, in 2000, Shearer (Shearer C., 2000) proposed a data mining process model based on commonly used approaches for experts to deal with data analytic problems. A great progress in Shearer's research is that his model employs the deployment of the extracted knowledge into real situation, and his point is accepted by latter researches. In the textbook for studying data mining "Data Science for Business" by Provost, Foster and Tom Fawcett it is mentioned that the whole data mining process should be supervised by the target that the KDD process wish to achieve, meaning that the deployment of knowledge is always in consideration throughout the whole process. (Provost, Foster, and Tom. Fawcett. 2013)

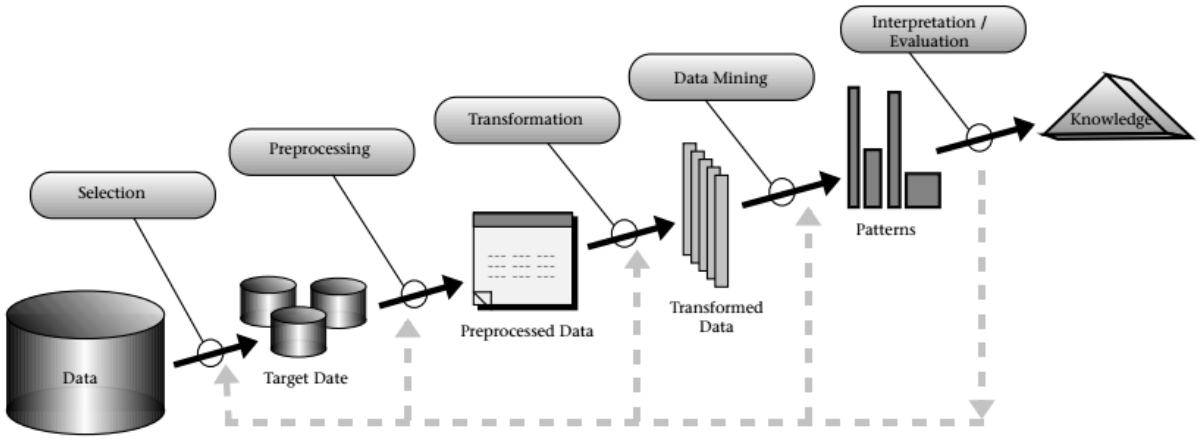


Figure 2 illustrates the general process of data mining by Fayyad (1995).

2.3 Functions of data mining techniques

Put forward by Fayyad (1995), this data mining approach contains two basic functions: descriptive function and predictive function. According to Han (2012), the functions of data mining techniques can be sorted into four classes: classification, association analysis, cluster analysis, and anomaly detection. While cluster analysis and anomaly detection functions are more descriptive function, many techniques under classification and association function are for pattern discovering and predictions.

Classification:

Classification techniques aims to produce certain classification model through data learning process, this model is able to identify correlation between independent variables and dependent variables thus can attribute classes to those variables. Among classification techniques, decision tree model, rule-based classifier model, Artificial neuron network model, support vector machine leaning model and simple Bayesian model are most applied models.

Association analysis:

Association analysis is one big data mining function which aims to find the hidden

relationship that provides useful insights among existing datasets, for example frequent appeared item and dataset in the data base can provide valuable information for decision making support under certain circumstances.

The fundamental part of association analysis lies in the identification and establishment of appropriate association rules, representing a method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. One famous example of applying association rules for the interest of marketing decision making is the discovery that that beer and diapers frequently appear together in a shopping basket using data mining techniques. Knowledge can be gained after the decision rule defined and decision support can be built. In a study done by Huang Ling and Xu Jianmin (2008), a prediction model of traffic jam was built based on knowledge discovered by association analysis of floating cars data.

Clustering analysis:

Cluster analysis is to divide the data set into different clusters, and every individual cluster differs significantly from other clusters while data in same cluster are highly similar in certain dimension. Basic algorithms in clustering analysis include partitioning, hierarchical method, density-based method and grid-based method. The major difference of cluster analysis from classification techniques is the clustering is an unsupervised data mining method. Knowledge gained from cluster analysis can be helpful in grouping consumers thus is of great insight in business research. For instance, by studying the performance of diffusion of the Internet based on Bass model indicators, (diffusion) Chinese scholar concluded four different patterns of the diffusion of telecom technologies in the global economy by clustering the innovator coefficient and imitator coefficient of technology growth in different countries.

Anomaly detection:

Anomaly detection is to find out-of-pattern subjects, sometimes call an outlier. This technique is majorly applied for fraud detection, medical treatment, public security, damage detection in industry, digit monitoring and so on. The application of anomaly detection is sometimes conducted in the frame of traditional statistical analysis, for example outlier detection

in numeric values. Anomaly detection can also be done in the process of other data mining techniques, k-mean clustering is often used for the detection of outliers.

Usually for different data mining functions there are different techniques and algorithms for settling the problem. A variety of data mining techniques exist, an important distinction made is between supervised and not-supervised methods. (Provost, Foster, and Tom. F. 2013). Unlike unsupervised methods, supervised methods always involve a target, a target is quantity to predict, such as transaction value of incident of switching, and there are different requirements for the value of the target in different algorithms. When extend this to different functions, whether a method is supervised or non-supervised is what differs clustering function from classification. Some of the most widely applied algorithms of each function are listed below:

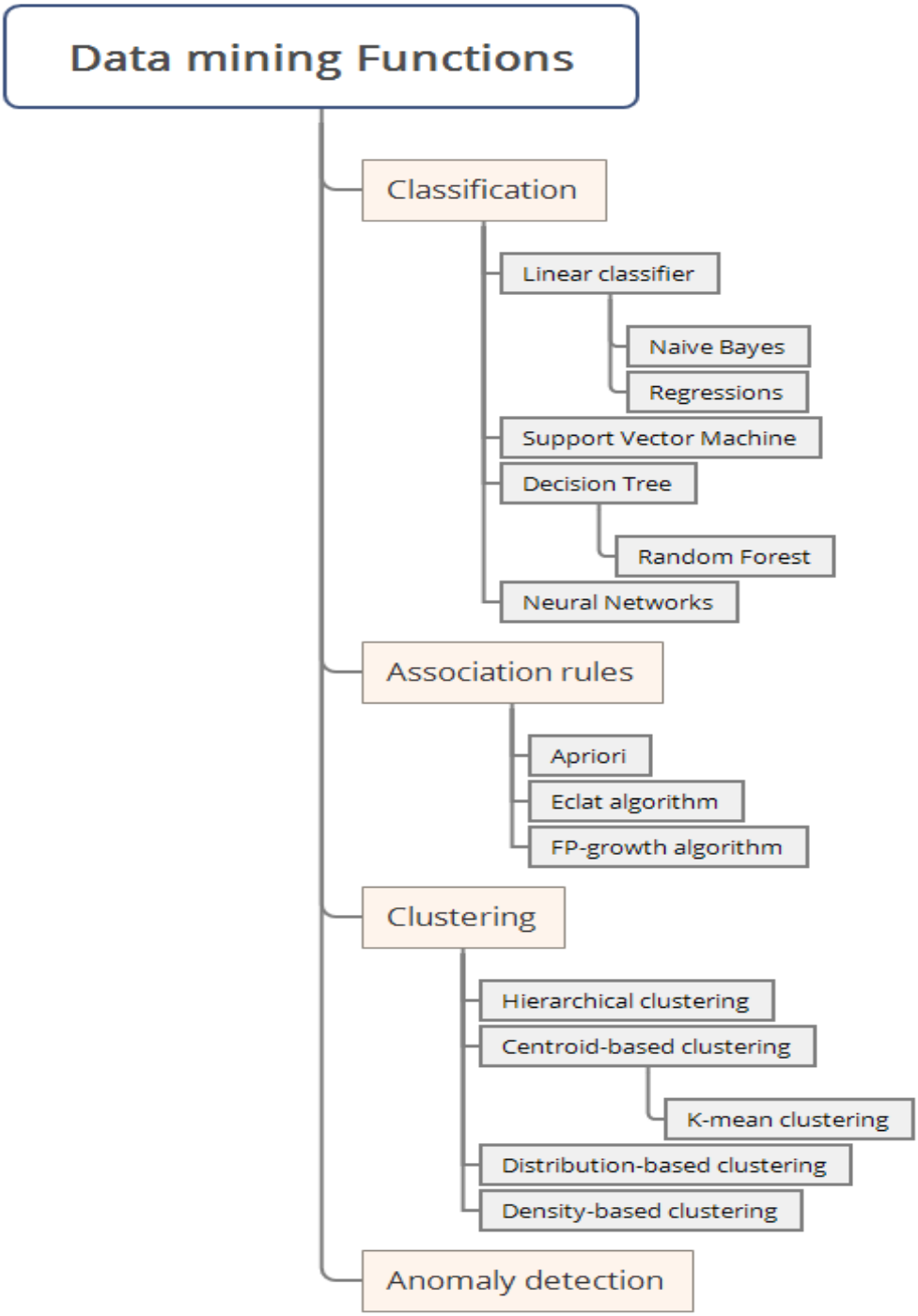


Figure 3. Different data mining functions and the data mining techniques under each function.

2.4 Classification techniques in Big data mining questions

As of classification function is most widely used for prediction for it's a supervised data

mining function with pre-defined target value, it has a wide range of algorithms to solve prediction problems. Since the study we are now carrying is categorized as classification problem, from here we review some of the most widely applied algorithms, familiarizing with their mechanism and pros and cons.

Support Vector Machine (SVM)

Support Vector Machines (SVMs) are supervised learning methods used for classification and regression tasks that originated from statistical learning theory (Vapnik, 1998). It is a classification technique that employs non-overlapping partitions for classifying and that is not picky for input variables. The entity space is partitioned in a single pass, so that flat and linear partitions are generated. SVMs are based on maximum margin linear discriminants, and are similar to probabilistic approaches, but do not consider the dependencies among attributes (Zaki and Meira, 2010)

The picture below illustrates the general idea of how a number of linear classifiers can be used to classify the data. The optimal classification is usually granted by a linear classifier which maximizes the distance between itself and the nearest data samples (Vapnik, 1998). It is intuitively expected that this classifier generalizes better than the other options (Gunn, 1998.).

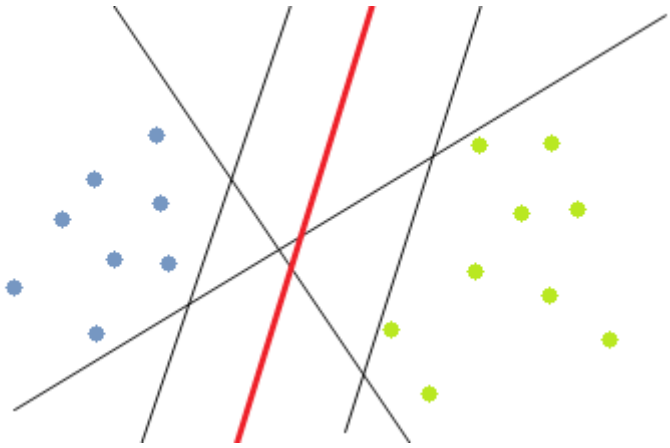


Figure 4: General illustration of SVM classifier

Neural Network (NN)

Artificial neural networks are commonly used for classification in data science. They group

feature vectors into classes, allowing you to input new data and find out which label fits best. This can be used to label anything, like customer types or music genres. The idea was developed in accordance with the elementary principle of human central nervous system. This network consists of three or more neuron layers: one input layer, one output layer and at least one hidden layer. In most cases, a network with only one hidden layer is used to restrict calculation time, especially when the results obtained are satisfactory. (David R., Sovan L., Ioannis D., Jean J., Jacques L., Stéphane A., 1997).

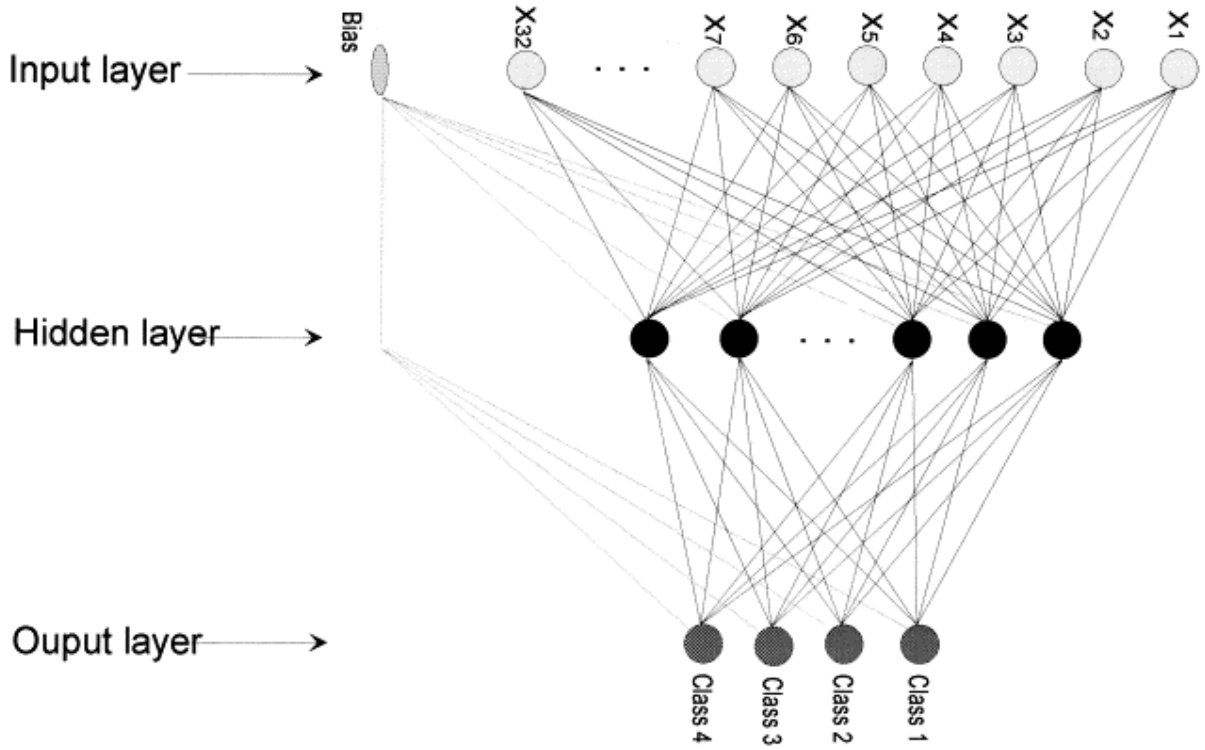


Figure 5: General illustration of Neural network classifier

Decision tree model

Decision tree learning is a method commonly used in data mining (Rokach, Lior; Maimon, O., 2008) It is a supervised function aims to predict pre-defined outcomes by analyzing a bunch of input variables. The fundamental algorithm in this classification technique is a greedy algorithm defined by Rose Quinlan (Quinlan, J. R. 1986), and is a top-down, greedy search type of algorithm with no back tracking. The running principle of decision tree is based on entropy and information gain, by comparing the entropy outcomes in every step, a decision is made for each branch.

Random forest

Random forest classifier is a more complex version of decision tree classifier, in which the forest is grown by a bunch of decision trees. Assume that the user already knows about the structure of single decision tree, and each tree in the forest makes decision separately. To classify a new object from an input vector, the model will put the input vector down each of the trees in the forest. Each tree delivers a classification result, and we say the trees "vote" for that every class. The forest chooses the classification having the most votes (over all the trees in the forest). The mechanism usually leads to the result that when having same input and target variables, random forest usually produces a better result than a decision tree.

Logistic regression

Logistic regression is a classifier for fitting a regression curve, y = f(x), when y consists of proportions or probabilities, or binary coded (0,1--failure, success) data. When the output is a binary variable, and inputs are numeric, logistic regression fits a logistic curve to the relationship between x and y. The logistic curve is an S-shaped or sigmoid curve, often used to model population growth, survival from a disease, the diffusion process of disease and so on. The basic assumption for logistic regression classifier is that the predictors are all numeric variables, and the outcome should be binary or variable of proportions, other type of input will not be recognized. Logistic regression is usually not a good model for big data problems because of the diversificity of data.

Each of the upper stated data mining methods has its own assumptions, when carrying out a prediction using those models, we should follow the rules when evaluating questions to be addressed and selecting input data.

3. Data mining in energy industry

3.1 Introduction of data mining application in energy market

Data mining in energy industry and its application in this field are more common and hold more importance than any previous time (*Ghodsi, M, 2014)*. First of all, as the energy market

today is growing more and more volatile and unpredictable, it is of great meaning to have ensured risk management process in order to guarantee the survival of the organization, maximizing opportunities whilst minimizing potential threats. However, data minining techniques have the potential of enabling the organization to capture previously unnoticed trends and patterns in the market, thus ensuring risk management capabilities of the organization. Moreover, as a result of recent technology development and revolution in other fields, the quantity of available information to be analyzed in the energy industry surged enormously, giving space for data analytic techniques to uncover useful information of the energy market. (Hassani, H., Saporta, G. and Silva, E. S, 2014) Last but not the least, classical techniques such as times-series and statistical prediction methods are out of date because their incapability of process such great quantity and variety of big data emerging in the energy industry.

However, even though we figured out the broad direction of analyzing energy market, in specific the energy price prediction for decision support, there are still many obstacles lying in our way. Designing a price-forecasting model for electricity market is a very complex task, according to Sanjeev and Ashwani (2009), identification of market scope, forecast horizon, selection of input variables, selection of data mining models, data preprocessing, parameter estimation and accuracy estimation can all be complex and requires lots of efforts. We believe that the review of some previous researches in this sector, the analysis of their techniques and research methods and applications, there design of data mining models, can be instructive for us to deal with our task at hand.

3.2 Data mining techniques and their application in energy industry

Up until now a variety of data mining techniques have been applied for different purposes in the energy industry, this includes and is not restricted to classification approaches, association analysis approach and clustering approach. Among techniques in data mining analyses, the most applied methods are decision tree model, neuron network, k-mean clustering, genetic algorithm, dynamic regression model, Bayesian model and support vector machine leaning. The large variety of analysis techniques come as a result of the diversified situation to be analyzed and

different aims of studies in this field of research. For instance, in several researches (Zhao, J. H., Member, S., Dong, Z. Y., Member, S., Li, X., & Wong, K. P, 2007) the support vector machine leaning algorithm is applied and proved to be a preferable method for the prediction of the occurrence of electricity price spike in the market. And in a study of Spanish power system carried in 2006, (Mandal, P., & Senjyu, T, 2006), decision tree model is applied for the prediction of the value of the reactors and the capacitors in the Spanish power system. The following table shows the data mining techniques applied in the energy industry in recent years and the purpose of their applications.

Data mining function	Data mining techniques	Application	References
Clustering	Visual DM techniques	Decision making support based on ARM (Automatic meter reading) data	(Liu, H., Yao, Z., Eklund, T. and Back, B. 2012)
	K-mean clustering	Fault detection analysis based on electricity consumption data in an Italian office building	(Khan, I., Capozzoli, A., Corgnati, S. P. and Cerquitelli, T, 2013)
	K-mean Clustering	Profiling energy usage among residencies in the UK	(Dent, I., Aickelin, U. and Rodden, T, 2011).
	K-mean together with SOM (Self-organizing map) technique)	Identifying electricity consumption patterns in different countries	(Figueiredo, V., Rodrigues, F. and Vale, Z, 2005).
	SOM clustering	Cluster the	(Valero, S.,

	and visualization techniques	customer policies in new electricity markets	Ortiz, M., Senabre, C., Alvarez, C., Franco, F. J. G. and Gabald´on, A, 2007).
Classification	Classification tree	classify electricity prices by simulating electricity market prices in New York, Ontario and Alberta.	(Huang, D., Zareipour, H., Rosehart, W. D. and Amjady, N, 2012).
		Optimize energy consumption management in a building.	(Gao, Y., Tumwesigye, E., Cahill, B. and Menzel, K, 2010).
	Neural network	Forecast system imbalance price of electricity in competitive markets	(Garcia, M. P. and Kirschen, D. S, 2006).
	Random forest	Wind power forecasting, using different methods.	(Fugon, L., Juban, J. and Kariniotakis, G, 2008).
		Predict the pattern of status of wind turbines in electricity generation	(Kusiak, A. and Verma, A, 2011).
	Bayesian model	Forecast electricity price spike in Queensland	(Lu, X., Dong, Y. Z. and Li, X, 2005).

		Forecast electricity price spike in reginal Chinese market, compared different techniques.	(Wu, W., Zhou, J., Mo, L. and Zhu, C, 2006).
	Support Vector Machine-Learning		
		Electricity spike forecasting and decision making, compare different techniques	(Zhao, J. H., Dong, Z. Y. and Li, X, 2007).
	K-nearest neighbors		
		Forecast electricity price based on hourly data	(Lora, A. T., Santos, J. R., Santos, J. R., Exposito, A. G. and Ramos, J. L. M, 2002).
		Point forecast of future electricity price	(Lora, A., Santos, J., Exposito, A., Ramos, J. and Santos, J, 2007).

Table 1. Review of data mining functions and techniques in energy industry

Ghodsi (2014), by reviewing previous researches about Big data mining in this particular industry, has concluded that the mostly applied data mining technique for researches in this field is decision tree model (Hassani, H., Saporta, G. and Silva, E. S, 2014), followed by support vector machine leaning. Clustering analysis is also a popular method of research for energy industry. The selection of methods could be based on the task at hand that is to be addressed, and in many researches the author chose to define a research scope and chose a data mining function before selecting a technique or algorithms to address the task. As a result of reviewing previous findings, many practical and value creating suggestions in the energy industry were made, for instance, decision tree model outperforms other data mining techniques in the prediction of

clearing price of electricity of New England market (Li, E. and Luh, P. B., 2001) and the random forest model is proved to have outperformed other algorithms in the prediction of wind power output (Fugon, L., Juban, J. and Kariniotakis, G. 2008)

Nevertheless, there is no benchmark for checking the continued out-performance of a single model over other models. Most of the results reported by different researchers cannot be put in a single framework because of the diversity in their presentations, as stated Sanjeev (2009), no single available model has been applied across data from larger number of markets. There is little systematic evidence yet that one model may explain the behavior of price signal in different electricity markets, which is an indicator of participants' collective response to uncertainties, on a consistent basis. Which means there can be no single optimal choice of data mining models for different study background. That is to say, although existing many researches on across different situations in different markets, to figure out the mining technique that fits our task the best, we still need to design our own model and exam it.

3.3 Selection of variables that may affect electricity price in data mining

The collection of informative data is an important part of KDD, (Provost, 2013). Variable selection in data mining is a complex task that requires a though understanding of the task we are addressing. Usually when selecting variables for prediction models, we choose only variables that is proved or believed to be predictive for the output of the model. For instance, in a study aims to build a prediction model for the prediction of global solar radiation, Mori took global solar radiation as output variable and for input variable he chose mostly meteorological variables at current time interval, including sunshine duration, temperature, wind speed, humidity, altitude of sun and so on. (Mori, H, and A Takahashi. 2012). Yet the task of selecting input variables is still challenging in many ways (Guyon, I, and A Elisseeff. 2003) First of all, too many predictive variables may cause statistical overwhelm, thus influence the accuracy of prediction, secondly, there are different nature of variables, they can be continuous numerical or categorical, they should be chosen in accordance with the requirements of statistical algorithms of the prediction models.

As of energy industry, some factors may have or have already been proved to have influence on the price of electricity, thus it is reasonable to incorporate them in the data base for analysis. According to (Sanjeev et al, 2009), the factors influencing electricity price can be classified into five classes:

- Market characteristics: such as the structure of the market, import/export, renewable energy share and so on.
- Nonstrategic uncertainties: such as fossil fuel energy prices, weather, temperature.
- Other nonstrategic uncertainties such as linear contingency, generation outages
- Behavior indices such as demand and spike existence and historical price
- Temporal effect: Price may under the influence of season, day type, time of the day and so on.

In model building it is desirable to use only the variables that hold the most predictive power. Since there is no record of literature which explores the predictive power ranking of different factors and it is believed to be very challenging to do so, (Sanjeev, 2009) in real life cases we usually select the predictors based on our assumptions that the predictor has an impact on the outcome. In this case, a thorough understanding of the situation is a must. Take our case for instance, meteorological variables are used for energy price prediction mainly because of the assumption that weather may affect renewable energy output hence change the supply of energy, leading to the fluctuation of energy prices.

Chapter 2 Decision support with big data analytics for Dutch energy market

In the first chapter, the literature review part, we have developed a theoretical framework for big data mining questions, about how big data questions are different from traditional statistics, how should we conduct data mining process step by step in a scientific way, and the issues in data mining applied for energy industry.

In general, those information is of useful guidance for our study. In reviewing ideas and view about big data replacing traditional data analytics, we reached the conclusion that although big data is an insightful tool for digging information, we cannot simply ignore causation relationship, and it is often the case that when traditional statistics works together with big data methods lead to a better result. In the data mining part we reviewed different functions and methods for problem solving, and we noticed that the development of prior knowledge and to let it supervise the mining process is of great importance to the whole problem solving.

Then we reached the part of data mining in energy industry, we have reviewed the methods and targets in previous data mining research in energy industry, leading to a conclusion that our input variables in the data mining process can be based on our assumptions and we should examine different methods since there is not optimal one for every question type in every market.

Our research is practice oriented and we incorporate real life situation, by analyze real situation we are able to give practical results.

1. Research background and overview

The application of big data mining techniques in the field of energy price prediction has been discussed and studied in different countries and in different cases, and many researches are done by far, mostly for business purpose (Weron, Rafa, 2014) Despite the complexity of market prediction in energy industry, energy price prediction, however, is not a less complicated a case in terms of research scope and the quantity of cases to be studied. This complexity is caused mainly by two factors. Firstly, energy industry is a highly volatile market that is under the impact of many different factors, some of which are even not studied, and it requires a lot of efforts to explain even a little pattern in the dynamic of electricity price. Secondly, exists certain

restrictions in the realization of speculation, for instance, in order to realize a profitable transaction, the company should be able to store the energy that is bought, and wait for the time to come when it is profitable to be sold or utilized.

From the side of big data mining, it is also required to have access to a great amount of relevant data, this is not restricted only to adequate data of different market prices at different points of a period of time, but also require the data of independent variables that can possibly affect the fluctuation of prices, in our case, those variables can be anything that affects the price of energy, this can be supply and demand, price elasticity, generation output and even predictors that seems to be irrelevant at first sight, such as the sun altitude or some meteorological variables (Mori,Takahashi, 2012). As a matter of fact, data mining is already becoming one of the most attractive topic in energy market forecast thanks to the nature of the market and the predictive power of data mining techniques (Weron, 2014).

Dutch electricity market, as a particular unregulated market, has many specificities that is interesting to study. The Netherlands, in particular, has a power exchange where the energy price is dominated by supply and demand, and an electricity operator that regulates the price of electricity based on the market output imbalance. Meanwhile, renewable energy output such as solar energy, wind energy and greenhouse is on a rise in the liberalized Dutch energy market, and private energy producers are in close connection with particular energy exchange market and are actively participating in it. In 2010, the world's first biomass energy exchange, the APX Biomass Exchange was opened in Rotterdam, the Netherlands, it is fully controlled by the Dutch government and allows private biomass power producers of the country to participate freely in the Dutch power market, few years later in 2013, the APX Biomass Exchange was combined into APX power exchange. (APX Group, 2016)

Broadly speaking, the price of electricity in the Netherlands is determined by two mechanisms: the mechanism of supply and demand and the price incentives that arise to counter mismatches between supply and demand. To specify, there are two different prices in Dutch electricity market, each is determined by one of the previous mechanism:

- Supply and demand

As in all markets, the price of energy is determined through the interaction of supply and demand. For electricity in the Netherlands, this happens in the APX market. Producers can indicate per hour of the day how much power they can produce for which price. Consumers can similarly indicate, for each hour of the day, how much they wish to consume and at which price. These bids combine to generate the price of electricity for the next day, which varies by hour, this is how the APX system works.

- Imbalance

  Besides the APX price, there is a second mechanism that influences the price of electricity: the costs of imbalances. For the electric grid to operate properly, supply and demand must always be in balance. TenneT is the grid operator who is responsible for maintaining this balance in the Netherlands. When imbalances occur, foe example because of incorrectly predicted renewable power generation, this party determines a so called imbalance price. This imbalance price indicates how much it is worth to engage or disengage a certain amount of capacity. This system is more complex, and only power suppliers are able to enter this market. (APX, 2015)

Dutch electricity market is not the only one with imbalance measures, the UK electricity market is using the same system, however, the Netherlands has the world's first biomass exchange for private energy producers to join the market, which means that there is a significant day to day activities of renewable energy producers in this market. The appearance of Amsterdam Power exchange(APX), which is an independent fully electronic exchange for anonymous trading on the spot market offering distributors, producers, traders, brokers and industrial end-users a spot market trading platform for day ahead transactions as well as Intraday transactions for on-the-day trading, provides customers with the possibilities and the option to purchase energy on APX prices or Imbalance prices. To struggle between prices is interesting, yet we still need further information of the detailed market situation to carry out our study, that is to come up with a data mining model for the market prediction.

To sum up, existing certain particular characteristics in the Dutch electricity market which made the whole business problem interesting to look into. That uniqueness in this market are the key factors we need to address through every step of data mining model building, especially

when we are selecting informative input variables and set target of the whole knowledge discovery process.

- The existence of different prices of electricity that is accessible to majority market players in the market, namely the APX price and the imbalance price.

- The increasing number of renewable energy producers in the market whose output is assumed to be under the influence of many additional factors, and their participation can possibly influence APX price and the imbalance price.

- The existence of an energy exchange where a number of green energy producers is participating in.

2. Market situation details

To better merge into the situation, we incorporate a Dutch company which is operating in the field of energy exchange and is offering contracts to end-customers. This company is currently trying to address the problem we mentioned above. Let's call this company Greenfield energy. The main business of this company is to provide consultation on the bidding behavior to end-customers which purchases energy in the stated market. Currently the company serves mostly for small greenhouse operators, as they are relatively small players in this market and inexperienced in comparison with other energy producers.

As market grows and the renewable energy output surges, many greenhouses appeared in the Dutch market. However, considering that they are relatively inexperienced in operating their own greenhouses, many of them consult energy consultancy firms for bidding strategy suggestions. Many of Greenfield-Energy's clients are those Dutch greenhouse operators. They are big electricity consumers and generators at the same time. After noticing that there was a chance of participating in imbalance market and purchasing energy other than market price, the greenhouse would like to optimize the energy cost by reducing purchase prices. In order to answer to this client's problem, Greenfield-Energy turned its eye to big data analysis and data

mining techniques.

Greenfield Energy has a newly developed software product that helps greenhouse customers to maximize their profit. Greenhouses require a lot of warmth and during this process of warmth production a lot of electricity is generated as well. This energy is sold on the Amsterdam Power Exchange (APX). In other word, greenhouses in the Netherlands are buying and selling electricity at the same time, while they always sell their produced energy in APX exchange, they as power producers can gain access to the imbalance market and can purchase electricity at the imbalance price. Choosing where to buy electricity is fun, interesting and lucrative but also consumes a lot of time of the (inexperienced) greenhouse operator. A well designed prediction model based on the market situation can be especially helpful under this circumstances, data mining methods can be applied here, to build a market prediction model that probably can provide decision making suggestions for power purchase, maximize the profit on APX by reducing energy purchase cost through a proprietary algorithm that incorporates traditional statistical methods and big data mining techniques.

The market of participation is regulated by the Amsterdam power exchange and players have to follow particular rules. When greenhouse operators doing biddings in the market, several restrictions exist for the decision to be made, customers are able to take positions on the APX a day in advance of the actual delivery. For example, a client states on Tuesday 10:00 AM how much electricity he will buy on the APX market for every hour the next APX-day (Wednesday 10:00 – Thursday 10:00). Meanwhile, the imbalance price is set by the operator of Amsterdam Power Exchange, the Tennet company, and is open to most energy producers to balance the supply hence to control the supply-demand balance in the market to keep the APX price stable.

In terms of volatility, the two different market price are show in the graph below, the APX prices are relatively stable compared to the Imbalance prices as can be seen in the image below.

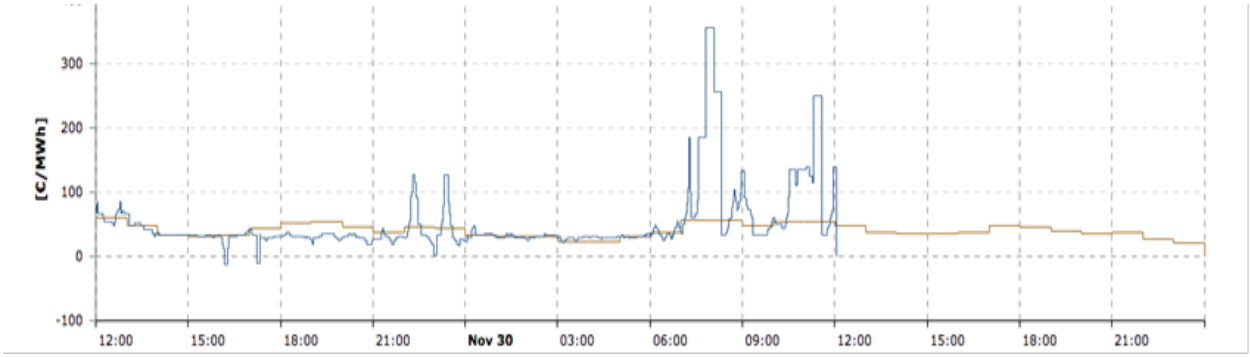


Figure 6. The APX prices (brown, or the more stable one) and Imbalance prices (blue, or the more volatile one) on the 30th of November, 2015; source: https://services.tenergy.nl/public.aspx/actualimbalanceprices

It is desirable for Greenfield-Energy to expand its current traditional statistical prediction models with a forecasting ability to address both the imbalance and the market price at the same time. As can be seen from the figure above, the imbalance prices are more volatile because this is the mechanism that ensures equilibrium of supply and demand. If the supply of electricity exceeds demand of electricity for a certain quarter, the Imbalance prices fall (see for example at 22:00 in figure 1), making it more attractive to buy energy on Imbalance, resulting in that the demand will rise up to equilibrium and that the APX and Imbalance prices become on par again (approximately at 22:30).

According to the information we have gained from the company, currently Greenfield-Energy is running a pilot with a small group of customers in which they have a basic forecasting model run every morning that sends out a bidding strategy. The forecasting model is a traditional statistical model based on fixed weighted averages of historical prices of the past 10 days (i.e. the weight each day carries do not change over time). As competitors of Greenfield-Energy are expected to already have a more sophisticated model to tackle the problem of dynamics in difference between Imbalance and APX prices, it is crucial for Greenfield-Energy to also have such a model in place. Otherwise, it might risk to lose current customer to competitors that have this sophisticated model in place, being able to provide more valuable business suggestions. The lack of a more sophisticated model creates a results gap, which is depicted in figure 6.

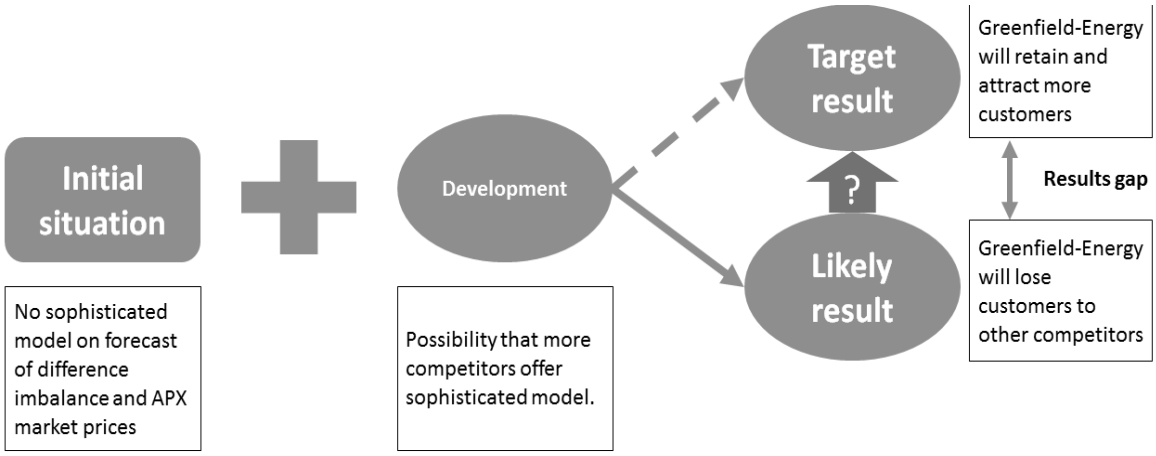


Figure 7. Results gap of the Dutch consulting firm in dealing with current problem.

Now that we have a clear visual of the market detail, we also got to understand the task that need to be addressed here in this case, that is the decision of when to buy electricity on APX price and when on the market price. According to our knowledge, the decision should be made when any of the two prices is significantly lower than the other. As Greenfield-Energy is only operating in the Dutch market, we will look at Dutch sources to obtain the data. During an interview with the Director of Sales and the Head of Product Development of Greenfield-Energy, it was indicated by the director that weather presumably plays an important role in forecasting the Imbalance prices, due the fact that electricity from solar cells and wind energy is rising (Feijen & Van Dijk, 2015). According to the company, this typical phenomenon in the Dutch market has increased the volatility of the Imbalance prices in the last few years. Ideally, Greenfield-Energy would like to receive a model that successfully advises on the decision whether electricity should be bought on the Imbalance prices or the APX prices, based on different variables input. Since our study is strongly business oriented, it has specific managerial implication which aims to provide solution for the following questions when managerial decisions are being made:

Can we advise clients of Greenfield-Energy whether to buy electricity on APX or to buy electricity on imbalance based on a variety of selected inputs including weather variables and time series variables such as the hour of the day, or whether it is weekday or weekend.

Splitting the question that is to be answered into several sub questions can help us to

understand the task in front of us more precisely, thus help us to build the data mining model for prediction.

What is the correlation between the weather variables and the difference of the Imbalance and APX price?

What is the correlation between the time of the day and the difference of the Imbalance and APX price?

How does the developed model compare to the current model of Greenfield-Energy?

What are the limitations of the developed model?

The business problem should be in par with the aim of our research that is to incorporate big data analysis methods and data mining techniques to provide insight to Dutch energy market, and identify an optimal method to be useful for business decision making support under specific market condition in the Netherlands.

3. Research design

To answer the research questions, an overview of the research design is a must. In general, we are trying to provide a suggestion of which analytical method to choose and which model to use so as to be able to provide best results in giving purchasing suggestions in the Dutch power market.

The backbone of current research is a KDD process that aims to provide optimal decision making suggestions for power purchase of greenhouse operators in the Netherlands, the main part of the research will follow the basic data mining process proposed by Fayyad (1996) and Shearer (2000), The main part of current research is designed following basic data mining process, to specify, the data mining part of this research is designed as follow:

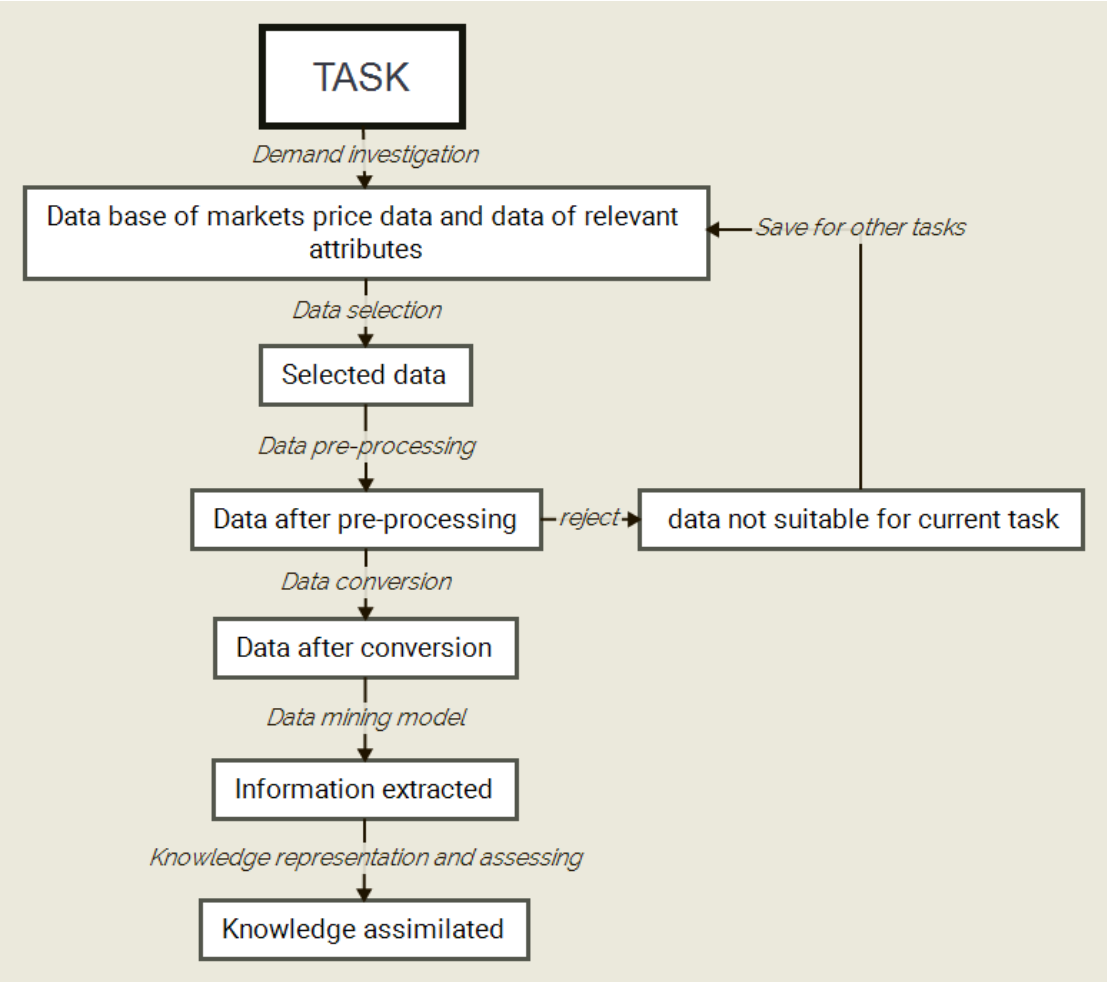


Figure 8. Framework of current study.

4. Data Preparation

In order to find a solution to the above stated business problem, target variables and explanatory variables have to be defined accordingly. Thus identifying the data base is a must for preliminary processing. In our case, the historical price data is a necessary element without question. Another data base that is required in our study is the meteorological data, for that we need to examine the impact of green energy output to the market price fluctuation and to evaluate the big data method's prediction power applying different data mining techniques.

4.1 Collection of the data

The collection of the data involved collecting data from the Imbalance prices, APX prices, meteorological data specified in weather variables and the exact time point of happening of

different attributes. The Imbalance price data is publicly available on the website of Tennet (an electricity provider in the Netherlands) and can be downloaded as a .csv file. The weather data is publicly available as well on the KNMI (Royal Dutch Weather Institute) website. In order to get the data from the APX prices, it is needed to request this data to the APX Group. This Group will determine whether the data is being used for the right purposes. As Greenfield-Energy already obtained this data, we managed to get collect the dataset directly from the company.

The imbalance price and the weather data is available until current date, however, the APX price data is acquired on request, and the Amsterdam Power Exchange is able to provide price data till December 2014, which limited the integrity of the data and thus limited the integrity of the input of the prediction model.

In addition, to serve the purpose of comparing the big data mining approach with the traditional statistical approach, it is desirable to have the perdition result of the basic model as well. For that the Greenfield-Energy company have already implemented a prediction model based on fixed weighed averages of the prices of the past 10 days, we decided to collect the prediction result of that model in a short period of time. After informing the company of what is necessary for the research, the company provided us with the suggested buying price data calculated by their model, as of the whole year of 2014. Directly comparing the prediction results of the year 2014 generated by both models is a more convenient way of comparing models' prediction power.

Data	Link
Imbalance prices	Tennet (Dutch energy operator) http://www.tennet.org/bedrijfsvoering/ExporteerData.aspx
Weather data	Royal Dutch Weather institute. http://www.knmi.nl/klimatologie/uurgegevens/selectie.cgi
APX prices	Provided by Greenfield Energy company on request.
Suggested bidding decision by Greenfield	Provided by Greenfield Energy company on request.

Energy, as of year 2014	

Table 2. Collected data and sources.

4.2 Target variable

After analyzing the market background and current situation, it is believed a model based on the interrelation between prices and external factors can be built for market performance prediction, since we are about to predict whether one price in the Dutch market is significant lower than another price at certain time point, we decided not to focus on the prices themselves but to focus on the price difference of the two prices.

The target variable is determined by the difference of the Imbalance price and the APX market price (Imbalance-APX). From there, the company has to decide whether to buy electricity on the Imbalance price or the APX price. As determined by Greenfield-Energy, a threshold of €5/MWh has to be taken into account when switching from buying electricity from APX to Imbalance. This is because, as opposed to buying on APX, buying from Imbalance is not risk-free as it requires accurate forecasting. As it is not risk free, a margin of €5 is set by Greenfield-Energy which is based on experience with 50+ greenhouse operators in the last years (Feijen and Van Dijk, 2015). In our classification problem we take two options for clients into account: “Buy on Imbalance price” and “Buy on APX price”.

Buy on APX price: Greenfield-Energy will advise its customers to buy electricity on APX prices if the difference between (Imbalance-APX) is higher than -€5/MWh (e.g. if Imbalance price is €70/MWh and APX price is €60/MWh). In our datasets, we have classified this option as 0.

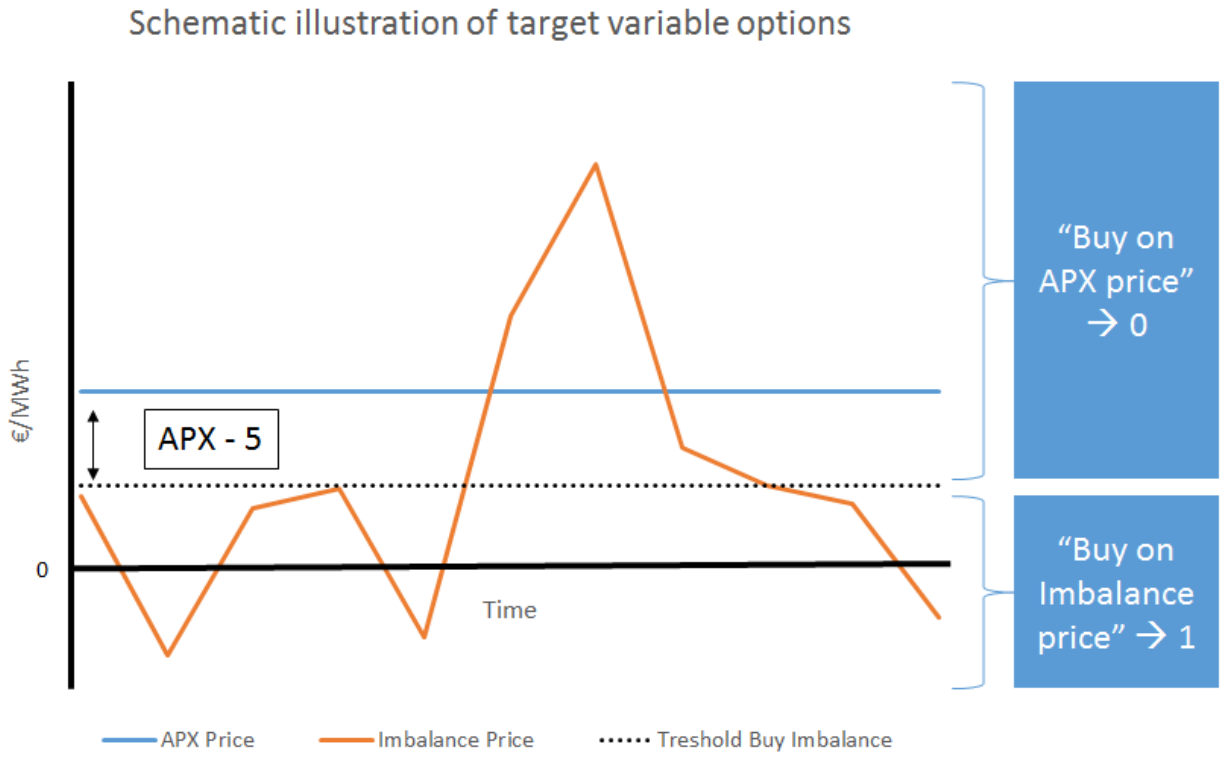
Buy on Imbalance price: Greenfield-Energy will advise its customers to buy electricity on Imbalance prices if the difference between the Imbalance and APX price (Imbalance-APX) is lower than -€5/MWh (e.g. if Imbalance price is €50/MWh and APX price is €60/MWh). In our datasets, we have classified this option as 1.

In the figure 8 which is shown below, the two different options are depicted in a schematical way. Information collected based on historical data of prices. Important to note is that no data

was used to create this diagram, it is merely for the quick understanding of the task. The fact that the APX market price is relatively stable and the Imbalance price is volatile (and possible negative) is derived from the data (using standard deviation and the ranges) and can be seem from figure No.8 and this is confirmed during the interview with Greenfield-Energy.

Variable	Standard deviation	Minimum value
APX.price	13.6	0
Imbalance.price	58.8	-435.10

Table 3. Variable characteristics that confirm that the Imbalance prices are more volatile than APX prices and can have a negative value.



igure 9. Schematic illustration of target variable options.

4.3 Exploratory variables

As stated in the problem definition, our model will analyze the relation between our target variable (the difference between the Imbalance price and APX market price) and selected input

variables. In a traditional statistical model, it is incorporated only proven related data, yet for big data analysis all supposed relevant data should be incorporated so as to discover and extract hidden knowledge from database, and our real task is to examine if there is correlation. The input variables we selected include meteorological variables and the time point of each event in the dataset. A description of the different weather variables is given in the table below.

Name	Description
HH	time (HH uur/hour, UT. 12 UT=13 MET, 14 MEZT. Hourly division 05 runs from 04.00 UT to 5.00 UT
DD	Mean wind direction (in degrees) during the 10-minute period preceding the time of observation (360=north, 90=east, 180=south, 270=west, 0=calm 990=variable)
FH	Hourly mean wind speed (in 0.1 m/s)
FF	Mean wind speed (in 0.1 m/s) during the 10-minute period preceding the time of observation
FX	Maximum wind gust (in 0.1 m/s) during the hourly division
T	Temperature (in 0.1 degrees Celsius) at 1.50 m at the time of observation
TD	Dew point temperature (in 0.1 degrees Celsius) at 1.50 m at the time of observation
SQ	Sunshine duration (in 0.1 hour) during the hourly division, calculated from global radiation (-1 for <0.05 hour)
Q	Global radiation (in J/cm2) during the hourly division
P	Air pressure (in 0.1 hPa) reduced to mean sea level, at the time of observation
N	Cloud cover (in octants), at the time of observation (9=sky invisible)
U	Relative atmospheric humidity (in percents) at 1.50 m at the time of observation
M	Fog 0=no occurrence, 1=occurred during the preceding hour and/or at the

	time of observation
R	Rainfall 0=no occurrence, 1=occurred during the preceding hour and/or at the time of observation
S	Snow 0=no occurrence, 1=occurred during the preceding hour and/or at the time of observation
O	Thunder 0=no occurrence, 1=occurred during the preceding hour and/or at the time of observation
Week days	Week days 0=working days (Mo-Fr), 1=weekend days (Sa-Su)

Table 4. Description of explainer variables and predictors.

4.4 Data cleaning

The data cleaning process in our study include the removal of noises, the reduction of data volume and the alignment of data structure.

All data processing and modeling processes are done in Rstudio with R language as the basic statistical language.

From the sources described in table 2, the datasets that are collected are as follows:

1.Imbalance price data from 2010-2015, march 30th, collected per 15 minutes; The data set has 183837 rows and 11 columns, the attributes to those columns are related to the time period of data collection, the imbalance market taking prices and infeed prices, and imbalance market situations such as if there were emergency power during certain period of time.

2.APX price data from 2010-2014, collected per hour; The dataset has 25 columns representing the date id and the hour of the day, the number of rows is the number of the days during the period. Data sets are collected by years, so we have managed to merge the datasets of different years into one data set.

3.Weather data from 2010-2014, collected per hour. Data sets are collected per year, so we have managed to merge the datasets of different years into one data set. The data sets are in text

form, after transforming the data into a csv file, it has 22 columns indicating the weather variables, the number of rows is the number of hours within the time period.

Here is what we did to the datasets step by step.

1. Remove all missing values listwise;

2. Extract all data from 01-01-2010 till 31-12-2014 in the imbalance price dataset, this is because for the APX price dataset and meteorological dataset we were not able to collect data of the year 2015, we have to ensure the consistency of the data;

3. Select and save every 4th row in the Imbalance price dataset. This is because Imbalance price dataset is four times larger than other datasets, which is due to the fact that the Imbalance prices were collected every 15 minutes while other data are collected hourly;

4. The APX price dataset is not properly structured: the prices are listed in horizontal rows by each day, with each column representing an hour. We have therefore extracted all values as a vector, and transformed the vector into a data frame with only one vertical column. This column has been named: “APX.price”;

5. The weather data and the processed Imbalance price data, together with the column “APX.price” are merged into one dataset;

6. We created additional columns:

Difference: The difference between Imbalance price and the APX price, calculated as (Imbalance price - APX price)

BUY: A vector of 0 and 1, whereby 0 indicates that the difference between prices is lower than 5, 0 indicates elsewise.

Weekdays: as we are interested in whether weekends have an influence on the price difference, this new variable is created. The column is a vector of 0 and 1, whereby 0 represents Monday to Friday and 1 represents Saturday and Sunday. The reason we are interested in this, is because electricity consumption is expected to be lower in the weekends (e.g. due to less company activity). It is interesting to see whether this has a correlation with the difference of Imbalance and APX price.

5. Modelling and testing

5.1 Selection of input variables

A first step in modelling and testing is to test the coefficient between independent and dependent variables, and assess whether significant coefficients exist, this is a preliminary step of input variable selection. In this way, it is possible to determine which independent variables are significant, and as well if they correlate in a positive or negative way. A linear regression model is built to do that:

Difference ~ FH + FF + FX + T + TD + SQ + Q + P + N + U + M + R + S + O + Weekdays

Note that wind direction (DD) and hour of the day (HH) are not included. Although we assume these are important factors, they are structured in a way that they are not suitable for linear regression test. This is because both variables can be seen as 'circular', i.e. the higher the value, the closer it gets to the lowest value. An example is that HH=23 (eleven o' clock in the evening) is very close to HH=1 (one o' clock in the night). This 'circular' characteristic excludes these variables to be included in the regression model but are later on incorporated in several other models which do work characteristic. Input variables are majorly numeric that meets the basic assumptions of a linear regression.

After running the linear regression model, we can see the coefficients in the table below:

	Coefficient	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	102.034198	35.474328	2.876	0.00403	**
FH	-0.249154	0.054918	-4.537	5.73e-06	***
FF	0.071398	0.044923	1.589	0.1119	
FX	-0.003245	0.025617	-0.127	0.8992	
T	0.088726	0.055469	1.600	0.10970	
TD	-0.080502	0.057339	-1.404	0.16034	

SQ	1.070367	0.136826	7.823	5.28e-15	***
Q	-0.004915	0.007608	-0.646	0.5182	
P	-0.013027	0.003230	-4.033	5.52e-05	***
N	0.827455	0.101732	8.134	4.27e-16	***
U	0.337055	0.119830	2.813	0.00491	**
M	-3.082416	1.428370	-2.158	0.03093	*
R	3.479327	0.813675	4.276	1.91e-05	***
S	11.690550	2.238389	5.223	1.77e-07	***
O	6.092429	3.021407	2.016	0.04376	*
Weekdays	1.168232	0.602788	1.938	0.05262	.

*Table 5. Coefficient table. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

As displayed above, 10 out of 15 variables are significant and we believe these coefficients meet the situations in the real world. For instance, when wind speed (FH) is higher, we expect more energy output from wind turbines, which leads to a lower market (APX) price. The same can be said for solar radiation, because we expect more solar energy output when solar radiation is strong, which leads to an increase of energy supply, which in turns results into a lower market price.

As for the insignificant variables, they can still be useful in our models because the target value in our models is the "buy on price" decision instead of the difference between prices (on which the regression analysis is based). It is possible that a variable can cast comparable influence on both the imbalance price and the APX price, which would lead to an insignificant result in the regression analysis, in which case the selection of predictors can be based on our understanding of the situation and the fitness of the data to the prediction model. (Guyon, I, and A Elisseeff. 2003. "An Introduction to Variable and Feature Selection." Journal of Machine Learning Research 3: 1157–82.)

For example, some variables are widely believed to have influence on the energy output, hence on our decision making, such as wind and temperature. We incorporate them into our

model tests, despite the insignificance according to the regression analysis, for the target of research should guide the data mining process and variable selection is supervised by our assumption of the real life. We incorporated most of the variables into our decision tree model and random forest, because these models make decisions based on information gain, and incorporating new variables will not reduce information gain we can get from other variables.

Another issue need to be noticed is that it is not the case the more predictors, the better the result. Our input variables are structured in different ways, and our prediction models have different requirement to the input data. Sometimes we need to try many times with different predictor combination to get a satisfactory result for a single prediction model.

5.2 Selection of prediction models

Our business problem has been structured as a classification problem. A classification task usually involves training datasets and test sets which consist of data instances. Each instance in the training set contains one target value (class label) and several attributes (features). The goal of a classifier is to produce a model able to predict target values of data instances in the testing set, for which only the attributes are known. Current classification problem can be viewed as a two-class problem in which our objective is to separate the two classes with the help of different attributes by a function induced from available examples. We've set a target value with two levels, these being "Buy on Imbalance price" and "Buy on APX price". In previous steps we have classified these two decisions in the column named "Buy", with respectively value 1 and 0 standing for the decisions levels. Considering that our goal is to produce a classifier that generalizes well, or to say that works well on unseen examples, we should incorporate different models for testing and comparison.

By developing various models, we are able to compare the preliminary results and focus our remaining efforts predominantly on 1 model. As a starting point, as we have categorized our problem as a classification, we decide to examine some of the most widely used classification models, which we have already reviewed in previous chapter of current paper. To specify, we are going to use the Artificial Neural Network classifier, the Support Vector Machine, the

classification tree model, the random forest model and logit regression model.

We did not include Naive Bayes, as we assume it does not fit our problem. This is because Naive Bayes modelling is based on probability learning and has strong feature independence assumptions. Considering the nature of the data, we believe logistic regression is also not a good prediction model for this question because of the diversified data types and the different nature of the data. For example, we expect more energy output when there is high solar radiation rate, high wind speed and high temperature, but it is possible that an increasing temperature and a decreasing solar radiation happen at the same time, causing the prediction to be less reliable. Another limitation is that when there are wrong independent variables, the prediction model shows little or no value at all. Judging from the aspect of big data analysis, the value of many of our input variables are not certain, thus is very likely that to predict the outcome with logit regression is not a good idea. However, we decide to examine the logit model anyway and see if this assumption is correct.

Considering that we do everything in R language environment, we incorporate several algorithm packages to support upper stated data mining models, these include:

"plyr", "C50", "nnet", "e1071", "rpart", "rpart.plot", "rattle", "randomForest".

5.3 Methods of model comparison

The next step is to test the acceptability of different models, usually this is done by evaluating prediction accuracy and data availability. Usually when we run and evaluate prediction models, there should be two datasets, one for model training and one for data testing. In our case we have one single data frame, to prevent the data from affecting the accuracy of our prediction, one proper method is to apply k-fold cross validation.

When carrying out k-fold cross-validation, we randomly divide original dataset into k equal sized subsets. then a single subsample is retained as the test data for testing the model, and the remaining k − 1 subsets are used as training data. After that we run the cross-validation k times, so that each of the k subsamples is used exactly once as the testing data. There will be k results from the folds, we calculate the mean accuracy to produce final accuracy evaluation of the

model. Such method perfectly addresses the one-dataset problem without influencing the prediction result. The advantage of this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, but in general k remains an unfixed parameter.

In our study, the number of folds applied is mainly decided by the time needed for computing. For instance, the computing of SVM model can take up to hours, which is why only 2 fold were assigned in the cross validation for SVM model. Because our dataset is big enough and have a big number of different variables, the number of folds won't affect the accuracy too much.

When the problems of dividing train data and test data is solved, our next question is how should compare the outcome of different prediction models. In general, the evaluation of the prediction models' predicting power is conducted based on three indicators:

- Sensitivity or True Positive Rate: measures the proportion of actual positives which are correctly identified. The formula is displayed below:

$$TPR = TP/P = TP/(TP + FN)$$

- Specificity or True Negative Rate: measures the proportion of negatives which are correctly identified. The formula is displayed below:

$$SPC = TN/N = TN/(TN + FP)$$

- Accuracy: measures of proposition of positives and negatives which are correctly identified. The formula is displayed below:

$$ACC = (TP + TN)/(P + N)$$

After the collection, selection and processing of data, the selection of input variables and the selection big data models, as well as the identification of model comparison methods, we could finally move to test our models. Although the initial coding requires some time, and there

have been errors keep occurring during our modeling process, it is repeat works after the first stage of the coding.

The method of carrying our study is not without problems, one major shortcoming of the methodology development lies in that we try to evaluate decision making support models merely by the accuracies of them, not considering other restrictions that may have during the data mining process. For example, big data models require a large amount of data, and all data should be well structured for analysis, this is one major difficulties in carrying big data analysis; Big data techniques are based on mathematical functions and the application of which requires thorough understanding of the task and the situation, as well delicate work on the data, which is a skill that is impossible for non-professionals. Developing new big data solutions requires financial input, so is hiring professional data analysts, and such cost may offset the benefit that big data model can bring to the company in the short run. Considering that the ease of application is also a factor to take into account in order to evaluate a decision making method under certain circumstances, judging a method based on accuracy is not a good idea for all situations.

Chapter 3. Prediction evaluation and comparison of methods

1. Examining the prediction power of meteorological attributes

One of our task is to identify if the input meteorological data have prediction power, in order to do that, we can simply conduct a prediction without meteorological input, then we incorporate meteorological data, repeat the prediction and compare the prediction results. If the prediction which incorporate meteorological data outperforms that without weather data, then we can say that one outperforms the other. While conducting the research, one important issue is, since the meteorological attributes are many, some of them may cause negative effect other than positive prediction power adding effect because of too much different input variables. A data pre-selection will help to reduce such effect. Despite the linear regression test we have done in the methodology chapter, some other methods can be taken to narrow the range of selection of input variables, this includes filtering out data that doesn't fit the assumptions of selected data mining techniques and carrying out tree model classification to rank information gain of different predictors.

Knowing that we have a big number of weather factors waiting to be input for prediction, tree model is an idea model for data selection in our case because classification tree model can effectively ignore wrong or inefficient predictors and label those factors that has more contribution to the final prediction result than other factors. Following this logic, we plot a classification tree with all input variables including weather data, time data, and dates. The visualized result can show us the ranking of prediction power of different variables. The plot is as follow, when tested for prediction with 10-folds cross validation, and find out that the most powerful predictor is HH (hour of the day), followed by Q (Solar radiation) and T (Temperature). When we increase the number of levels of the classification tree, we come to know that FH (Hourly mean wind speed) is the next significant predictor.

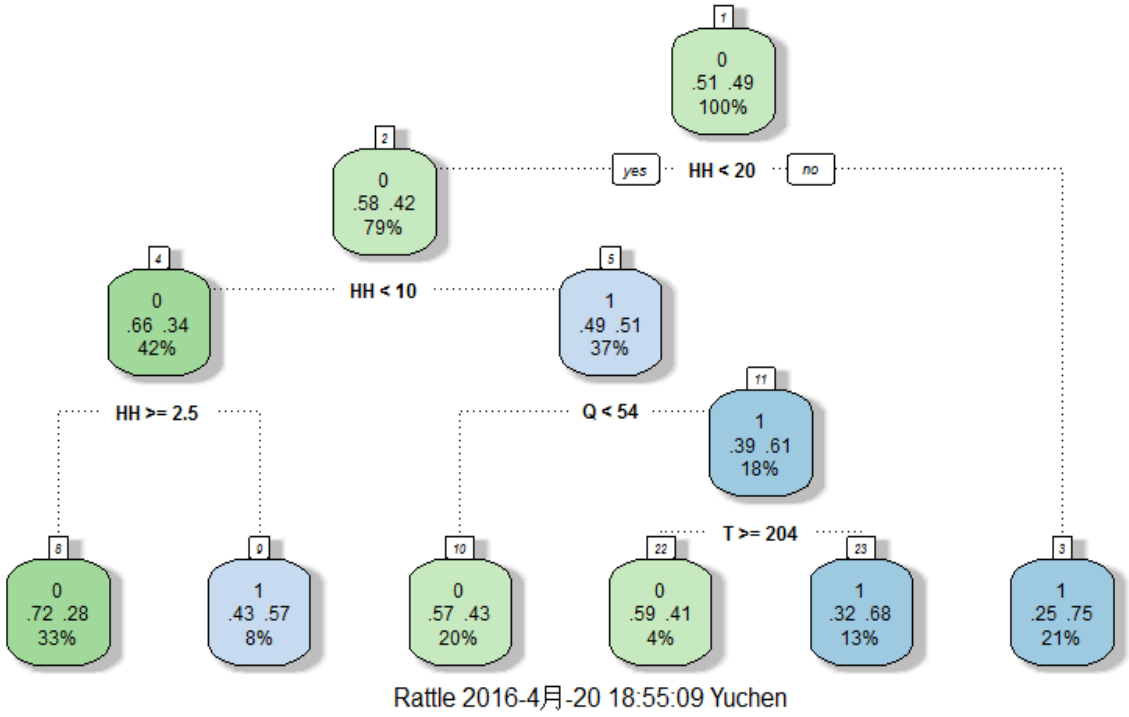


Figure 10: Classification tree to check entropy of different factors.

This tree built has a mean accuracy of 0.6724, if we perform prediction with random forest model, the final mean accuracy is around 0.7099.

However, when we restrict the predictor to only HH (Hour of the day) and day type (the weekdays), the accuracy reduces to 0.6555 and 0.6356 relatively.

Accuracy	**With all predictors**	**Without weather data**
Tree model	0.6724	0.6555
Random forest	0.7099	0.6356

The results show that the accuracies of the classification tree and random forest model with weather predictors is higher than the prediction accuracies of the same model without weather inputs. Based on the results so far, we can conclude that although weather data is not the major contributor to the prediction of correct bidding strategy in the Dutch electricity market, the meteorological can contribute to the accuracy of the prediction and holds prediction power. For testing different data mining models, we give priorities to hour of the day, solar radiation,

temperature, and hourly mean wind speed as input predictors. When we bring this logic into testing different models, it was discovered that for other big data mining models, the same situation happened that the incorporation of weather data, especially those three attributes we have identified above, can improve the result of the prediction.

2. Comparison of data mining techniques

Based on our pre-research assumption, there should exist a big data mining model which suits our question best and gives an optimal prediction result in giving decision making support for which-price-to-buy decision. Now we perform prediction with each selected models, namely the artificial neural network, Support vector machine, logic regression, tree model and random forest. For each model, a number of equally sized folds is created. The number of folds applied is mainly decided by the time needed for computing. For instance, the computing of SVM model can take up to hours, which is why only 2 fold were assigned in the cross validation for SVM model.

For each model, at the very beginning most predictors are incorporated, then some predictors are removed step by step and the calculation of prediction accuracy is conducted repeatedly to make sure that most combination of predictors are considered and tested. During this procedure, priorities are given to attributes such as hour of the day, solar radiation, temperature and mean hourly wind speed. As a result, for each data mining model in our case, we are able to pick up one final combination of predictors which has the best prediction outcome for different big data mining techniques.

The most accurate SVM is structured as:

Buy ~ HH + FH + T + Q

The most accurate ANN model is structured:

Buy ~ HH + T + Q + weekdays, size=10

The tree model with best performance that was built:

Buy ~ HH + DD + FH + FF + FX + T + TD + SQ + Q + P + N + U + M + R + S + O + Weekdays

The most accurate random forest model that was built:

Buy ~ HH + DD + FH + FF + FX + T + TD + SQ + Q + P + N + U + M + R + S + O + Weekdays, ntree=100

We created a confusion matrix for each model to check the performance of each model. Based on this confusion matrix, the accuracy, specificity and sensitivity of each model are calculated.

classification_NN	1	0
0	3665	7818
1	7162	3264

Table 6. Artificial neural network prediction result.

classification_logit	1	0
0	6	28
1	4294	4436

Table 7. Logit regression prediction result

classification_tree	1	0
1	1293	588
0	842	1659

Table 8. Classification tree model prediction result.

classification_forest	1	0
1	1455	620
0	660	1647

Table 9. Random forest prediction result.

classification_SVM	1	0
1	7023	3092

0	3678	8116

Table 10. Support Vector Machine prediction result.

We calculate the mean accuracy of prediction from the prediction result of each fold in cross-validation approach, the specificity and sensitivity are calculated separately for the comparison of different big data techniques.

Model name	# of folds (K)	Mean accuracy	Specificity	Sensitivity
SVM	2	0.6902415	0.6562938	**0.7241256**
NN	10	0.6825277	0,6614944	0.7054683
Classification tree	10	0.6724178	**0.7445964**	0.6018913
Random forest	**10**	**0.7098683**	0,7096283	0,7117463
Logit model	5	0.4916472	0,508132875	0,1764705

Table 10. Comparison of big data prediction results

Among all models, random forest has the best performance in term of overall accuracy, the negative decision given by classification tree model is more reliable than the negative decision support of other models, the support vector machine which incorporates HH (hour of the day), FH (Wind speed), T (temperature) and Q (solar radiation) as predictor has the best performance in term of sensitivity. Worth mentioning is that for SVM and ANN models, the restriction of number of input predictors can significantly increase the overall accuracy of the prediction, when all predictors were applied, the accuracy were merely over 50%, when the predictors were reduced to only hour of the day and several important weather factors, the accuracy increased to over 65%.

Models other than random forest model, although the overall prediction accuracy is slightly lower, still hold quite remarkable prediction power, Neural networks and SVM model all have accuracies over 68%, outperforming the result of single tree classifier. SVM model has a sensitivity of over 72%, meaning that it works best in predicting true positive results, the positive suggestions given by the SVM model is worth following. The classification tree model has the highest specificity value, which means that the negative decision suggestion given by the tree models is more trustworthy compared to other models. The Logit model is as expected a not suitable model - just like we stated before-due to the diverse data type.

Therefore, it was decided to use the random forest model as the final model to be compared with previous installed model of the Greenfield company, the logic of which is, if even the best big data model is not able to outperform currently implemented prediction model, then it is considered that big data model is not a better approach in current business scenario than traditional method.

3. Benchmarking the big data model with previous applied model in the company

In order to compare the big data model with the traditional model, and to assess the usability of our big data model, we have benchmarked our model against the model from Greenfield-Energy. As currently the company model is a work-in-progress, little resources have been devoted to the development of it and statistics on the output are non-existent. However, because of the relative simple structure of the model, and as we have all required data accessible, we reverse engineered the bidding strategies the company model would have generated for 2010-2014. The mechanism of the current model Greenfield-Energy uses is to look at the historical Imbalance prices to determine whether the client has to buy on the APX or buy on Imbalance. For every hour that needs to be determined, the model looks at the Imbalance price on the same hour for the past 10 days. Each day has received a particular weighting, primarily based on instinct of the client and previous market experience. An example calculation is depicted below:

	12:00 Day-2	**12:00 Day-3**	**12:00 Day-4**	**12:00 Day-5**	**12:00 Day-6**	**12:00 Day-7**	**12:00 Day-8**	**12:00 Day-9**	**12:00 Day-10**	**12:00 Day-11**
Weights	40%	30%	5%	5%	5%	5%	3%	3%	3%	1%
Imbalance price	38,55	46	33,01	37,32	45	24,17	-2,19	9,55	19,55	-3
Contribution	15,42	13,8	1,6505	1,866	2,25	1,2085	-0,0657	0,2865	0,5865	-0,03
Expected Imbalance price	€36,97									

Table 11. The current model of Greenfield-Energy is based on weighted averages of historical

Imbalance prices

We have calculated what the accuracy of the current company model would have been if it had been deployed for the past 5 years. Same as ours, we give the company model two possible output-variables: buy on APX or buy on Imbalance. As a result, the current company model has an overall accuracy of 66.56%, meaning that during the last 5 years in 66.56% of the total cases the model made a correct advice. Our random forest model has an overall accuracy of 71%, an improvement of almost 5% compared to the current model Greenfield-Energy uses. In terms of accuracy, big data model outperforms the traditional one.

4. Cost-Benefit Analysis

To perform a close to real-life cost-benefit analysis, we have tested both models on a generic client's transactions. The difference between this analysis and the comparison of accuracy is the timeslots of when buy on APX or Imbalance come into play. The exchange market does not trade energy 24 hours a day, i.e. is not always able to make buying decisions. We have used a general operation setup of the Amsterdam power exchange which is displayed below. This general operation setup is given by Greenfield-Energy based on market information.

	Weekdays		Weekend	
	Start	**End**	**Start**	**End**
Winter	**07:00**	**23:00**	**12:00**	**23:00**
Spring	**09:00**	**20:00**	**12:00**	**20:00**
Summer	**10:00**	**17:00**	**x**	**x**
Autumn	**09:00**	**20:00**	**12:00**	**20:00**

Table 12. The general times of operations for an average customer. These timeslots display the windows of opportunities in which decisions to either buy on APX or Imbalance can be made.

We have deployed the old model, our own model, and a perfect model using the above setup, for the year 2014. The perfect model always takes the best price, either APX or Imbalance, and hence results in the minimum price possible of energy purchase during last years. Both the cost and the benefit of our decisions can be expressed in monetary values. In other words,

making a *wrong* decision (advising to buy imbalance, when APX would've been better) can similarly to a *right* decision be expressed as a monetary value. We collected the data of decision record suggested by the company model, in the year of 2014, the comparison below assumes 1MWh to be sold per transaction and has been run over the whole year of 2014 to take into account any seasonal differences.

	Costs	
Greenfield-Energy Model	€ 225.100,51	100%
Big Data Model	€ 189.714,79	84,82%
Cost Saving	**€ 35.385,72**	**15,72%**
Perfect Model	€ 163.857,89	72,79%

Table 13. Cost-benefit analysis based on the comparison of prediction methods.

The result shows that compared to current model implemented by Greenfield Energy, the best performing big data model can actually reduce the energy purchasing cost of that greenhouse operator by 15.72%, in the year 2014, with significant cost-saving effect, we believe the model is worth to be developed further. Although this result is based on the assumption that a fixed quantity of electricity is purchased per transaction, and the real situation is sure to be different, but we believe that people tend to purchase more when the prediction suggest them to buy, if the price is not influenced by the purchase amount change, it is possible that this model can perform even better than our calculation.

5. Discussion of results

Our main target of current study is to find out if there is a proper method of price prediction analysis for business decision making in the Dutch power market. There are currently two basic approaches on the table, one traditional and one big data approach which is a relatively new method, our task includes finding out which one is a better business tool, and if big data out performs, which exact technique should be applied and how to the decision making tool should be built, what kind of data can be applied and how may this model improve business performance in the real world. Current study simply addressed above questions by trying to build

our own big data models, following instructions and suggestions of the Green-field Energy company which is an experienced player in Dutch power industry. As a result, a big data classifier which uses weather data, time seral data and price dynamic data to predict cost-saving buying strategy was built. This result is the outcome of a serial of comparative analysis between big data approach and traditional price dynamic approach, and amongst different big data classifiers. As is shown in our result, it is suggested to use our random forest model for most accurate decision support, while the support vector machine leaning and artificial neural networks is of specific value in predicting special results. Although far from perfect, according to our current result, big data methods can help operators in the Dutch power industry to save energy purchase costs, for managers in the company of this industry, the practical value of the model should be analyzed and probably further developed, and in accordance with such, data analysts are the people to recruit.

However, our study is carried follow a bunch of strict assumptions and thus have a number of limitations, making it not being able to be put into practice right away.

The whole job done is based on historical data. We use past weather and pricing data in order to determine if the future APX price will be higher or lower than the Imbalance one. Based on that classification, we give advice on whether clients should purchase energy on the APX or Imbalance price. Is not always a good approach to predict future based on history, the result can be informative and valuable, but it has shortcomings. Weather and pricing values are known to be volatile and hard to predict. Basing our model on the historical data does not take into consideration uncertainty to a proper extent. It thus cannot be an automatic tool deciding of the actions to be taken throughout the day. The model is meant to serve as a decision support system. Similarly to what Greenfield-Energy is currently doing, it could provide a report every morning with a suggestion for actions to be taken throughout the day, leaving the greenhouses with the final decision.

In order to have an automatic tool, a few additional elements should be added. First of all, the predictions on weather data, Imbalance and APX prices should be incorporated into the data set. However, this data is either expensive to buy, weather data licenses range from € 14,000 to € 140,000 (ECMWF.int), or not available, as predictions on APX and Imbalance prices are not

available. Moreover, when using this predictive data, the accuracy of the model depends on the way predictions are made. For example, we can try to generate a bidding strategy for tomorrow, the weather input is tomorrow's weather forecast, the weather predictions are not always correct thus may influence our prediction result.

Another important limitation relies on the dynamics of the energy market. The changes are driven by three main effects. Firstly, concerning the climate change, some drastic changes have occurred. The average temperature in the Northern-Hemisphere increased by nearly 1°C on land, and there is a clear snow deficit compared to previous periods. It can be seen that precipitations in Northern Europe, including the Netherlands, have increased (World Meteorological Organization, 2013). These changes may impact the demand for energy. As an example: snow influences energy demand, as lower amounts of snow may decrease the demand.

The second driver relies on stakeholder behavior. In the Netherlands, when the biomass power exchange first established in 2010, renewables were having a 3% share in the energy mix. The government's target for 2020 is 12% (Rabobank International, 2012). Renewable energy is influenced by the climate, especially wind and solar radiation. The supply of renewables is thus not constant. By increasing their share, the government may increase volatility of energy supply, which decreases future predictability. Moreover, economic growth impacts the demand for energy. In period of prosperity there is a growth of energy demand, whereas in recessions like the one in 2008 it drops (International Energy Agency, 2012). Finally, household energy use is slowly changing. The use of smart meters and improved building isolation reduce the consumption and demand of energy (Rijksdienst voor Ondernemend Nederland, 2014).

The third limitation is caused by technologic improvements. We mentioned the used of smart meters, there are a few other trends that will probably impact the energy market. The implementation of electric cars is one of them. It has been discussed that a dynamic electricity pricing could be offered. Drivers would recharge their vehicles off peak hours, and thus reduce their recharging costs. Ultimately it might create demand during hours where it was less important and reduce fluctuations in demand for energy distribution across the day (Erasmus Energy Forum, 2014). This could have a significant impact on our model since hour of the day as a significant impact on the prices. Moreover, technologies drive down the price of renewable

energy generators, enabling more and more people to produce their own electricity. Finally, companies such as Tesla are researching on batteries that would enable storage of electricity (Lindsay, 2015). This might have a significant impact in the future, as it will enable people to store energy and release it in the market whenever they want. People would for instance be able to store solar energy on a sunny day and release on the market on a rainy one, markets will be then less influenced by weather and time variables.

Forth, we did not include in our model the supply of energy from the foreign markets such as Germany, the Netherlands imports more or less 2400 GWh per month (TenneT, 2014). This supply has an impact on prices and thus may influence our model.

Furthermore, we used data only from 2010 onwards, as we were advised by the Head of Product Development in Greenfield Energy because data prior to 2010 is considered irrelevant due to the changed dynamics in the energy market caused by the APX biomass exchange. Prior to 2010 the energy market was more stable due to the lesser influence of above stated environmental drivers. As these drivers (e.g. renewable energy) are becoming more of an influence, our model will become less accurate over time as these are not taking into consideration. It will be necessary to keep updating the data of input every after a certain period. Our testing data is until the December 31. 2014, for we have got no data of 2015 from the company, making our test a not up-to-dated one.

Finally, we give a recommendation to Greenfield's clients on the action to take on the energy market. We suggest them to buy from APX or Imbalance, but there are no indications on the optimal quantity that should be placed in the market. The model will give recommendation to clients, meaning that they will not make gut feeling decisions anymore but educated decisions. This will increase or decrease the energy supply in the market and also the prices. Our model will then influence the market and has the potential to change the patterns of energy pricing, which in the long term will affect the validity of the tool.

Conclusions and further research implications

The result of the modelling and testing have given us the answers to our research questions, our assumption that exists a most accurate data mining technique for current situation was met. Given the fact that Greenfield Energy is now using a statistic model that employs only price data and time serial data, we have found that big data mining techniques outperforms traditional data analytics method, and exist certain data mining technique that can give us the best decision making support in selecting bidding. And according to our result, our random forest model has the best predictive power in dealing with dataset among all prediction models, whether is based on big data mining techniques or traditional approach.

By controlling the predictors in the models we have built, we have also reached the conclusion that, meteorological data, seemingly irrelevant to the market price fluctuation, indeed has power of predicting market condition in the Netherlands. Among all prediction weather variables, the wind speed, temperature and solar radiation has the most prediction power. This actually coincide with our knowledge that those factors can influence green energy output.

To sum up, for the Dutch electricity market, we can suggest Greenfield Energy to look more into big data models like random forest model in making the buy-or-not buy decision. The model can still be further improved by adding more data. Although we think it is not always a good decision to evaluate the prediction models only by the its accuracy, and big data models, though hold highest predictive power, may be difficult to use and costly to develop. Finally, though the outcomes are delightful, lots of limitations exist in our study, our model is far from perfect, it requires lots of work before it can be finally put into practice.

This study addresses the research gap we have mentioned in previous chapters, focusing on specific problem in the Dutch energy market and provided solutions, but for researchers there are still space for further studies. From the perspective of researchers, we have proved that big data models and data mining techniques provides more accurate predictions to support business decision. Our model is far from perfect, but it can shed light upon further studies.

First of all, although we have proved that in general weather data can improve the

prediction power of our big data models, it is not identified the actual power of each specific weather variable. Although in big data researches more attention is placed on the outcomes, it is wise to examine if the outcomes fit our assumption based on causation relationship to make big data researches more credible.

Another recommendation for researches is that the model has further development potential. In the Netherlands green energy is not only generated by wind turbine and greenhouses, the hydraulic system in the Netherlands is one of the most developed in the world and is being used to create power. So it is suggested that hydrology data is also a predictor that should be examined in further studies. And there can be predictors other than weather and hydrology data waiting for us to discover.

In the end, although every country is a unique market, the basic idea of using big data for business decision support in energy industry can be applied for other countries, to use a series of predictors for cost-saving analysis can be used for every situation similar to ours, and it will be interesting to discover cases in countries other than the Netherlands.

As for the companies, especially for energy consulting firms like Greenfield energy, it is recommended to adopt our model in a test environment, continue with the development of it, and eventually deploy it in their workflow processes.

Results from the benchmark demonstrate that our model in its current existence already outperforms the current model of Greenfield-Energy. The benchmark however has been based on a generic customer and we would recommend to benchmark both models on a wide set of customers. Not only in terms of operations but also in terms of a diverse set of strategies, e.g. risk-averse versus risk-seeking strategies. This would better demonstrate the robustness of both models making comparisons more accurate and informational.

As has been described in the limitations it’s paramount to keep the model up to date with current data. As the electricity market is influenced by countless factors, and the drivers of energy prices are forever changing the model needs to be kept current. We recommend to update the model on a monthly basis at a minimum, preferably more often. The best way is that the application which implements this model can update itself automatically.

Furthermore, it is recommended that Greenfield-Energy incorporates the model with other models. We know that Greenfield-Energy possesses various model that might be beneficial. For example, we know that Greenfield-Energy owns an APX-forecasting model and we deem it valuable to look into potential synergies if this is combined with our model. A specific recommendation for combining our model would be a model that can identify which hours over the year are most volatile. Having more accurate predictions of variables allows for a better overall prediction.

As Greenfield-Retail is pursuing a diversifying strategy we furthermore recommend to look into companies that are familiar with modelling and in specific prediction modelling. This is of course a more long-term strategy and would require more in-depth analysis of market potential for this model but the preliminary results are promising.

Literature

1. Grey J. (2009). “Jim Gray on eScience: A Transformed Scientific Method.” The Fourth Paradigm, 17–31. doi:10.1038/embor.2010.47.
2. Anderson C. (2008), The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine, 16(7).
3. Mayer-Schönberger, Viktor, and Kenneth Cukier. (2013), Big Data: A Revolution That Will Transform How We Live, Work and Think. London: John Murray, Print.
4. Liang N, Yan Z, (2013), Promote the discovery of data-intensive science: New model, methods and challenge. Bulletin of the Chinese Academy of Sciences:1:115-121.
5. Ghodsi, Mansi. (2014). “A Brief Review of Recent Data Mining Applications in the Energy Industry” 2 (1): 49–57. doi:10.1142/S2335680414500045.
6. Fred C. Schweppe, Michael C. Caramanis, Richard D. Tabors, Roger E. Bohn. (1988), “Spot Pricing of Electricity”. The Kluwer International Series in Engineering and Computer Science, ISBN: 978-1-4612-8950-0
7. Sanjeev.K.A. (2009). “Electricity Price Forecasting in Deregulated Markets : A Review and Evaluation.” INTERNATIONAL JOURNAL OF ELECTRICAL POWER & ENERGY SYSTEMS, no. February. doi: 10.1016/j.ijepes.
8. Bal, Remco. (2013). “Development of the Imbalance of the Dutch Electricity Grid,” no. Utrecht University, Department of Innovation, Environmental and Energy Sciences.
9. Hong L, Xinhe H, (2012), Preliminary research in philosophy in data science. Philosophical Trends, 12: 82-88.
10. Nature. (2008), Big Data. http://www.nature.com/news/specials/bigdata/index.html.
11. Pietsch W. Big Data – The New Science of Complexity. 6th Munich-Sydney-Tilburg Conference on models and Decision, 2013.
12. Office of Science and Technology Policy, (2012), Obama Administration Unveils" Big

Data" Initiative: Announces $200 Million in New R&D Investments. http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.

13. McKinsey & Company. (2011). “Big Data: The next Frontier for Innovation, Competition, and Productivity.” McKinsey Global Institute, no. June: 156. doi:10.1080/01443610903114527.
14. Lapkin A. (2012), Hype Cycle for Big Data, Gartner, G00235042.
15. Yangyong Z, Zan X. (2009), The science of data, Fudan University Press.
16. Hey, T., Tansley, S. and Tolle, K. (2009). The Fourth Paradigm: Data-intensive Scientific Discovery, Redmond: Microsoft Research.
17. Davenport, Thomas H, Paul Barth, and Randy Bean. (2012). “How ‘ Big Data ’ Is Different How ‘ Big Data ’ Is Different.” MIT Sloan Management Review 54 (1): 43–46.
18. Callebaut W. (2012), Scientific perspectivism: A philosopher of science’s response to the challenge of big data biology. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Science, 43(1):69-80.
19. Dhar V.(2013) Data science and prediction. Communications of the ACM, 56(12):64-73.
20. Fayyad U M, Uthurusamy R,(1995). Proceedings. First International Conference on Knowledge Discovery and Data Mining: August 20-21, 1995, Montréal, Québec, Canada. AAAI Press.
21. Todri, Vilma. (2015). “Big Data : From Correlation to Causation.” Department of Information, Operations and Management Sciences, no. Stern School of Business, New York University: 1–3.
22. Han J, (2001), Data Mining: Concepts and techniques. Morgan Kaufmann Publisher, 2001
23. Provost, Foster, and Tom Fawcett. (2013). “Data Science for Business”, O'Reilly Media

24. Fayyad, Usama; Piatetsky-Shapiro, Gregory; Smyth, Padhraic (1996). "From Data Mining to Knowledge Discovery in Databases" (PDF).

25. Pyle, D., (1999). Data Preparation for Data Mining. Morgan Kaufmann Publishers, Los Altos, California.

26. Shearer C., (2000), The CRISP-DM model: the new blueprint for data mining, J Data Warehousing; 5:13—22.

27. Mori, H, and A Takahashi. (2012). "A Data Mining Method for Selecting Input Variables for Forecasting Model of Global Solar Radiation." Transmission and {Distribution} {Conference} and {Exposition} ({T} {D}), 2012

28. Friedman, J. H. (1997). Data Mining and Statistics: Whats the Connection? In: 29th Symposium on the Interface: Computing Science and Statistics, 14–17 May 1997,Houston, TX, 3–9.

29. Hand, D. J. (1998). Data Mining: Statistics and More? The American Statistician,52(2), 112–118.

30. Guyon, I, and A Elisseeff. (2003). "An Introduction to Variable and Feature Selection." Journal of Machine Learning Research 3: 1157–82.

31. Hassani, H., Saporta, G. and Silva, E. S. (2014). Data Mining and Official Statistics: The Past, the Present and the Future. Big Data, 2(1), 1–10.

32. Zhao, J. H., Member, S., Dong, Z. Y., Member, S., Li, X., & Wong, K. P. (2007). A Framework for Electricity Price Spike Analysis, 22(1), 376–385.

33. Mandal, P., & Senjyu, T. (2006). Neural networks approach to forecast several hour ahead electricity prices and loads in deregulated market, 47, 2128–2142.

34. Liu, H., Yao, Z., Eklund, T. and Back, B. (2012). A Data Mining Application in Energy Industry. In: Proceedings of the 12th Industrial Conference, ICDM 2012, July 13-20, Berlin, Germany.

35. Khan, I., Capozzoli, A., Corgnati, S. P. and Cerquitelli, T. (2013). Fault Detection Analysis of Building Energy Consumption Using Data Mining Techniques. Energy

Procedia, 42, 557–566.

36. Dent, I., Aickelin, U. and Rodden, T. (2011). The Application of a Data Mining Framework to Energy Usage Profiling in Domestic Residencies using UK Data. In: Proceedings of the Research Students Conference on “Buildings Dont Use Energy, People Do?” Domestic Energy Use and CO2 Emissions in Existing Dwellings, 28 June 2011, Bath, UK.

37. Figueiredo, V., Rodrigues, F. and Vale, Z. (2005). An Electric Energy Consumer Characterization Framework Based on Data Mining Techniques. IEEE Transactions on Power Systems, 20(2), 596–602.

38. Valero, S., Ortiz, M., Senabre, C., Alvarez, C., Franco, F. J. G. and Gabald´on, A. (2007). Methods for Customer and Demand Response Policies Selection in New Electricity Markets. IET Generation, Transmission & Distribution, 1(1), 104–110.

39. Huang, D., Zareipour, H., Rosehart, W. D. and Amjady, N. (2012). Data Mining for Electricity Price Classification and the Application to Demand-Side Management. IEEE Transactions on Smart Grid, 3(2), 808–817.

40. Gao, Y., Tumwesigye, E., Cahill, B. and Menzel, K. (2010). Using Data Mining in Optimisation of Building Energy Consumption and Thermal Comfort Management. In: 2nd International Conference on Software Engineering and Data Mining (SEDM), 23-25 June, Chengdu, 434–439.

41. Garcia, M. P. and Kirschen, D. S. (2006). Forecasting System Imbalance Volumes in Competitive Electricity Markets. IEEE Transactions on Power Systems, 21(1), 240–248.

42. Fugon, L., Juban, J. and Kariniotakis, G. (2008). Data Mining for Wind Power Forecasting. In: European Wind Energy Conference & Exhibition EWEC 2008, Brussels, Belgium, 1–6.

43. Kusiak, A. and Verma, A. (2011). Prediction of Status Patterns of Wind Turbines: A Data-Mining Approach. Journal of Solar Energy Engineering, 133(1), 1–10.

44. Lu, X., Dong, Y. Z. and Li, X. (2005). Electricity Market Price Spike Forecast with Data

Mining Techniques. Electric Power Systems Research, 73(1), 19–29.

45. Wu, W., Zhou, J., Mo, L. and Zhu, C. (2006). Forecasting electricity market price spikes based on Bayesian expert with support vector machine. Advanced Data Mining and Applications Lecture Notes in Computer Science, 4093, 205–212.

46. Zhao, J. H., Dong, Z. Y. and Li, X. (2007). Electricity Market Price Spike Forecasting and Decision Making. IET Generation, Transmission & Distribution, 1(4), 647–654.

47. Lora, A. T., Santos, J. R., Santos, J. R., Exposito, A. G. and Ramos, J. L. M. (2002). A comparison of two techniques for next-day electricity price forecasting. Intelligent Data Engineering and Automated Learning IDEAL 2002 Lecture Notes in Computer Science, 2412, 384–390.

48. Lora, A., Santos, J., Exposito, A., Ramos, J. and Santos, J. (2007). Electricity Market Price Forecasting Based on Weighted Nearest Neighbors Techniques, IEEE Transactions on Power Systems, 22(3), 1294–1301.

49. Li, E. and Luh, P. B. (2001). Forecasting power market clearing price and its discrete using a Bayesian-based classification method. In: Proceedings of the IEEE power engineering society winter meeting, 28 January–01 February, Columbus OH, 1518–1523.

50. Weron, Rafa, (2014). “Electricity Price Forecasting: A Review of the State-of-the-Art with a Look into the Future.” International Journal of Forecasting 30 (4). Elsevier B.V.

51. Feijen, C. and Van Dijk, T. (2015). Problem statement interview Agro-Energy, Personal Interview, Delft, 31 April 2015.

52. International Energy Agency, (2012), Oil & Gas Security Emergency Response Of IEA Countries. Paris: OECD/IEA, Print.

53. Lindsay, Rowena. (2015), 'How Tesla's New Battery May Revolutionize Energy Consumption'. The Christian science monitor 2015.

54. Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. Tourism management, 29(3), 458-468.

55. Rabobank International, (2012), An Outlook For Renewable Energy In The Netherlands. Utrecht: cooperatieve centrale Raiffeisen-Boerenleenbank B.A., Print. Rabobank Industry Note #320.

56. Rijksdienst voor Ondernemend Nederland, (2014), Dutch Energy Savings Monitor For The Smart Meter. Print.

57. TenneT,. Market Review (2014), H1 Electricity Market Insights First Half 2014. Amsterdam: TenneT, 2014. Print.

58. Thuraisingham B. (1998), Data mining: technologies, techniques, tools, and trends. – CRC press.

59. V. Vapnik, (1998), Statistical learning theory. Wiley, New Tork.

60. M. J. Zaki and W. Meira Jr. (2010), Fundamentals of Data Mining Algorithms. Cambridge University Press.

61. Gunn, (1998), Support vector machines for classification and regression. Tech. rep., University of Sothampton, UK.

62. David Reby, Sovan Lek, Ioannis Dimopoulos, Jean Joachim, Jacques Lauga, Stéphane Aulagnier, (1997), Artificial neural networks as a classification method in the behavioural sciences, Behavioural Processes, Volume 40, Issue 1, April 1997, Pages 35-43, ISSN 0376-6357.

63. Rokach, Lior; Maimon, O. (2008). Data mining with decision trees: theory and applications. World Scientific Pub Co Inc. ISBN 978-9812771711.

64. Quinlan, J. R. (1986). Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106

65. Hong T., Chang W. K., Lin H. W. (2013), A fresh look at weather impact on peak electricity demand and energy use of buildings using 30-year actual weather data //Applied Energy. – T. 111. – C. 333-350.

Appendix:

Appendix 1. Codes for data processing:

```
# Read file that contains the Imbalance prices for every 15 minutes (in Dutch)
price <-read.csv("export.csv")
str(price) # View structure of Imbalance

price.2010<-price[1:35040,]
price.2011<-price[35041:70080,]
price.2012<-price[70081:105216,]
price.2013<-price[105217:140256,]
price.2014<-price[140257:175296,]
price.2015<-price[175297:183836,]

#select every 4th row in the dataset(as hourly imbalance price)
hourly.price.2012<-price.2012[(seq(4,to=nrow(price.2012),by=4)),]
# We do this to every years' data from 2010 through 2014

# Read file that contains the 2012 APX prices for every hour (the prices are in a sheet named 'prices_2')
APX2012<-read.xlsx("DAM - Historical data - 2012.xls",sheetName="prices_2",header=TRUE)
# Delete irrelevant columns
APX2012.2<-APX2012[-c(1,26:39)] # Delete irrelevant columns
# Rotate the dataset so as to extract all the values
APX2012.3<-as.data.frame(t(APX2012.2)) # Rotate the dataset so as to extract all the values
#Extract all the value in the dataset into a vector
APX.price2012<-unlist(APX2012.3[,c(1:366)], use.names = FALSE) #Extract all the value in the dataset into a vector
# Convert the extracted vector into a data frame
APX.price.2012<-as.data.frame(APX.price2012) # Convert the extracted vector into a data frame
# Combine the imbalance price dataset and the APX price dataset
total.2012<-data.frame(APX.price.2012,hourly.price.2012) # Combine the imbalance price dataset and the APX price dataset
# We do this to every year from 2010 through 2014
# Read file that contains the 2010 weather data for every hour
weather2010<-read.csv("KNMI_2010_hourly.txt")
# We do this with weather data from 2010 through 2014

# Now merge the weather data into the price dataset we got earlier
Total.2010<-data.frame(total.2010,weather2010) # Add 2010 APX prices and weather data together
# We do this with weather data from 2010 through 2014
```

```
Total.2010<-Total.2010[ -c(3:9,11:13) ] # Remove irrelevant columns
# We do this with weather data from 2010 through 2014

# Rename APX.[year] columns, so that rbind is possible. Rbind requires similar columns
colnames(Total.2010)[1] <- "APX.price"
# We do this with weather data from 2010 through 2014

# Combine 2010:2014 data frames
TOTAL<-rbind(Total.2010,Total.2011,Total.2012,Total.2013,Total.2014) # Combine 2010:2014 data frames

colnames(TOTAL)[3] <- "Imbalance.price"  # Change 'Afnemen' (Dutch) column to Imbalance.prices

TOTAL$Sell<-as.vector(0)
TOTAL$Buy<-as.vector(0)

# Function to change comma to dot, which is case with Imbalance.price.
# By this the difference can be calculated of Imbalance and APX price, as Imbalance.price will be detected as numeric
myfun <- function(x) {sub(",",".",x)}

# Apply the function to Imbalance.price variable
var<-as.numeric(sapply(TOTAL[,3], FUN=myfun))
TOTAL$Imbalance.price<-var
str(TOTAL) # Check that Imbalance.price is indeed numeric right now

#Create new variable "sell", which means the difference of imbalance price and the APX price.
TOTAL$difference<-TOTAL$Imbalance.price - TOTAL$APX.price

#Transform into new variable, 'Sell' which indicates if the difference exceeds the treshold of 5 euros/MWh.
TOTAL$sell <- ifelse(TOTAL$difference > 5, c(-1), c(0))
TOTAL$Sell <- ifelse(TOTAL$difference > 5, c(1), c(0))

# Transform into new variable, 'Buy' which indicates if the difference is lower than -5 euros/MWh.
TOTAL$Buy <- ifelse(TOTAL$difference < -5, c(1), c(0))

TOTAL$Decision<-TOTAL$sell+TOTAL$Buy

#New column which indicates whether it's a weekend
TOTAL$Weekdays<-as.vector(0)

n <- length(TOTAL$Weekdays)
TOTAL$Weekdays[seq(2, n, 7)] = 1
```

```
TOTAL$Weekdays[seq(3,n,7)]=1

#Check for missing values
colSums (is.na( TOTAL ))
# Listwise deletion of all rows with any missings
TOTAL_listwise<-TOTAL[ complete.cases (TOTAL$APX.price),]
TOTAL_listwise<-TOTAL_listwise[ complete.cases (TOTAL_listwise$N),]
```

Appendix 2. Codes for modelling and testing:

```
# Change the value into factors for classification
TOTAL_listwise$Decision <- factor (TOTAL_listwise$Decision,
                    levels = c(1,0,-1))
# Change the value into factors for classification
TOTAL_listwise$Sell <- factor (TOTAL_listwise$Sell,
                  levels = c(1,0))
# Change the value into factors for classification
TOTAL_listwise$Buy <- factor (TOTAL_listwise$Buy,
                 levels = c(1,0))
# Listwise deletion of all rows with any missings
TOTAL_listwise<-TOTAL[ complete.cases (TOTAL$APX.price),]
TOTAL_listwise<-TOTAL_listwise[ complete.cases (TOTAL_listwise$N),]

#Linear regression, target - difference
modelA<-difference ~ FH + FF +
  FX + T + TD + SQ + Q + P + N + U + M + R + S + O + Weekdays
class ( modelA )
rsltModelA <- lm(modelA , TOTAL_listwise)
#To see the coefficience
summary ( rsltModelA )

#Linear regression, target - Decision
modelB<- Buy ~ FH + FF +
  FX + T + TD + SQ + Q + P + N + U + M + R + S + O + Weekdays
rsltModelB <- lm(modelB , TOTAL_listwise )
summary ( rsltModelB )

# Loading the required libraries
library ("plyr")
library ("C50")
library ("nnet")
library ("e1071")
library ("rpart")
```

```
library ("rpart.plot")
library ("rattle")
library ("randomForest")

#Cross validation for testing Neuron network model
# Randomize the data
data<-TOTAL_listwise[sample(1:nrow(TOTAL_listwise)) ,]
# Create 10 equally sized folds
number_folds<-10
folds<-cut(seq(1,nrow( data )),
           breaks=number_folds ,
           labels=FALSE)
# Vectors to store the results initialized with 0s
accuracy_NN<-rep(0,number_folds)

# Perform 10 fold cross validation
for(i in 1:number_folds ){
 #Segement the data by fold using which - function
 testIndexes <- which ( folds==i,arr.ind =TRUE )
 testData <- data [testIndexes, ]
 trainData <- data [-testIndexes, ]
 # Fit the models
  NN_model <-nnet(Buy ~ HH + FH + FF + T + TD + SQ + Q + P + N + U + R + S,
data=trainData, size=10)

 # Make predictions on the test set
 classification_NN<-predict(NN_model,
                            testData,
                            type="class")

 # Measuring accuracy
 accuracy_NN[i]<-sum(classification_NN==
                       testData$Buy)/nrow(testData)
} # end of the for loop

#Create confusion matrix
table(classification_NN,testData$Buy)

#Cross validation for testing SVM model
# Randomize the data
data<-TOTAL_listwise[sample(1:nrow(TOTAL_listwise)) ,]
# Create 2 equally sized folds
number_folds<-2
folds<-cut(seq(1,nrow( data )),
        breaks=number_folds ,
```

```
        labels=FALSE)
# Vectors to store the results initialized with 0s
accuracy_SVM<-rep(0,number_folds)

# Perform 2 fold cross validation
for(i in 1:number_folds ){
  #Segement the data by fold using which - function
  testIndexes <- which ( folds==i,arr.ind =TRUE )
  testData <- data [testIndexes, ]
  trainData <- data [-testIndexes, ]
  # Fit the models
  SVM_model <-svm(Buy ~ FH + FF + T + TD + SQ + Q + P + N + U + R + S, data=trainData)

  # Make predictions on the test set
  classification_SVM<-predict(SVM_model,
                        testData,
                        type="response")

  # Measuring accuracy
  accuracy_SVM[i]<-sum(classification_SVM==
                    testData$Buy)/nrow(testData)
} # end of the for loop

#Cross validation for testing the tree model
# Randomize the data
data<-TOTAL_listwise[sample(1:nrow(TOTAL_listwise)) ,]
# Create 10 equally sized folds
number_folds<-10
folds<-cut(seq(1,nrow( data )),
        breaks=number_folds ,
        labels=FALSE)
# Vectors to store the results initialized with 0s
accuracy_tree<-rep(0,number_folds)

# Perform 10 fold cross validation
for(i in 1:number_folds ){
  #Segement the data by fold using which - function
  testIndexes <- which ( folds==i,arr.ind =TRUE )
  testData <- data [testIndexes, ]
  trainData <- data [-testIndexes, ]
  # Fit the models
  tree<-rpart(Buy ~ HH + DD + FH + FF +
               FX + T + TD + SQ + Q + P + N + U + M + R + S + O + Weekdays,
             data=trainData,
             method="class",
```

```
            parms=list(split="information"))
  # Make predictions on the test set
  classification_tree<-predict(tree,
                          testData,
                          type="class")

  # Measuring accuracy
  accuracy_tree[i]<-sum(classification_tree==
                      testData$Buy)/nrow(testData)
} # end of the for loop

#Create confusion matrix
table(classification_tree,testData$Buy)

#Cross validation for testing random forest model
# Randomize the data
data<-TOTAL_listwise[sample(1:nrow(TOTAL_listwise)) ,]
# Create 10 equally sized folds
number_folds<-10
folds<-cut(seq(1,nrow( data )),
           breaks=number_folds ,
           labels=FALSE)
# Vectors to store the results initialized with 0s
accuracy_forest<-rep(0,number_folds)

# Perform 10 fold cross validation
for(i in 1:number_folds ){
  #Segement the data by fold using which - function
  testIndexes <- which ( folds==i,arr.ind =TRUE )
  testData <- data [testIndexes, ]
  trainData <- data [-testIndexes, ]
  # Fit the models
  forest <- randomForest(Decision ~ HH + DD + FH + FF + FX + T + TD + SQ + Q + P + N + U
+ M + R + S + O + Weekdays,data=trainData, ntree=100)
  # Make predictions on the test set
  classification_forest<-predict(forest,
                           testData,
                           type="class")

  # Measuring accuracy
  accuracy_forest[i]<-sum(classification_forest==
                        testData$Decision)/nrow(testData)
} # end of the for loop

#Cross validation for testing random forest model
```

```
# Randomize the data
data<-TOTAL_listwise[sample(1:nrow(TOTAL_listwise)) ,]
# Create 10 equally sized folds
number_folds<-10
folds<-cut(seq(1,nrow( data )),
        breaks=number_folds ,
        labels=FALSE)
# Vectors to store the results initialized with 0s
accuracy_forest<-rep(0,number_folds)


# Perform 10 fold cross validation
for(i in 1:number_folds ){
  #Segement the data by fold using which - function
  testIndexes <- which ( folds==i,arr.ind =TRUE )
  testData <- data [testIndexes, ]
  trainData <- data [-testIndexes, ]
  # Fit the models
  forest <- randomForest(Buy ~ HH + DD + FH + FF + FX + T + TD + SQ + Q + P + N + U + M
+ R + S + O + Weekdays,data=trainData, ntree=100)
  # Make predictions on the test set
  classification_forest<-predict(forest,
                          testData,
                          type="class")

  # Measuring accuracy
  accuracy_forest[i]<-sum(classification_forest==
                      testData$Buy)/nrow(testData)
} # end of the for loop
```