Politecnico di Torino

Department of Electronics and Telecommunications
Master’ s degree in ICT For Smart Societies

Master Thesis

Market power analysis and detection in the Italian electricity market based on big data analysis and deep learning

Supervisor:

Prof. Tao Huang

Author:

Shuo Huang

August 2020

Acknowledgments

First of all, I would like to express my gratitude to all those who helped me during the writing of this thesis. I gratefully acknowledge the help of my supervisor, Prof.Tao Huang, who has offered me valuable suggestions in the academic studies. In the preparation of this thesis, he has spent much time reading through each draft and provided me with inspiring advice. Without his patient instruction, insightful criticism and expert guidance, the completion of this thesis would not have been possible.

Then, I would like to thank all my friends and classmates who supported me and gave me valuable advice. During the writing, every time I met problem and bottlenecks, they are always with me and help me to think out the solution.

Last, I should finally like to express my gratitude to my beloved parents and girlfriend who have always been helping me out of difficulties and supporting me without a word of complaint.

Author :
Shuo Huang

Abstract

In today's society, electricity is undoubtedly the largest energy source. The vast majority of production activities and people's daily lives are based on a stable supply of electricity. Therefore, the power industry itself has produced great commercial value. With the development of liberalized markets, free electricity markets based on competition also arises. Electricity is a necessity for production and life. If the electricity market is maliciously manipulated or monopolized, it will greatly affect people's lives, social welfare and economic development. Therefore, as an important indicator for evaluating the level of market monopoly, market power has also become an important research object in the electricity market industry in recent years. This thesis focuses on the Italian electricity spot day-ahead market and aims to use big data technology and deep learning to study its market power level and supervision.

In the first chapter, the research background of this thesis is introduced, including the electricity market and its development, the generation and composition of the Italian electricity market. In the second chapter, in order to better understand this thesis, some basic economics knowledge such as the competitive relationship of the electricity market is introduced. In addition, the current research status of market power in the electricity market is analyzed. Also explains the reality which is lacking unified standard and inefficiency of market power behavior identification from the aspects of power particularity, market power definition, market power precedents and EU regulations,

Chapter 3 introduces the technologies and methods used in this thesis, including big data technologies and deep learning. Besides, the detailed derivation of the variational auto-encoder algorithm used in our anomaly detection model is introduced. In Chapter 4, the used data is first introduced. Then, based on big data technologies, the traditional long-term static indicators and the short-term dynamic indicators that proposed in this thesis are used to analyze the potential market power of the Italian electricity market. On this basis, the anomaly detection model is established, and its composition and method selection are briefly introduced. In Chapter 5, in order to verify the model performance, we manually evaluate some of the abnormal points detected by the model and do the final judgment from many aspects such as economic and empirical methods. The results show that the abnormal points do have abuses of market power, indicating that the model has achieved the expected goal, which can effectively reduce the detection range and significantly improve the efficiency of market power detection. The thesis lists the detailed analysis process of two abnormal points as the case studies. Finally, Chapter 6 shows the summary, reflections and prospects of the thesis.

Keywords: market power, big data, deep learning, anomaly detection

Table of Contents

Table of Figures

Table of Tables

Chapter 1

Background

This thesis focuses on the research of the market power of the Italian electricity market. Compared with the ordinary markets, the electricity market has some particularities, and its characteristics are not well known by people outside the electricity industry. To help better understand the entire thesis, this chapter will briefly describe some important knowledge and the background about the electricity industry, including electricity trading, particularities of electricity, the Italian electricity industry development and its composition.

1.1 Electricity trading

Since the second industrial revolution, electricity has become the most critical energy source in the world today, whether in industrial production or daily life, and the following information technology revolution is also based on the electricity industry. At the same time, the ways of electricity production, transmission, and sales have also taken a significant change: from the early government Manipulation model evolved into a free trading model that follows the laws of the market economy, and the electricity started to be traded as a kind of actual commodity.

With the development of electricity trading, some electricity markets started to appear in some countries, which also led that the forms of power trading started to evolve into different directions, including the bidding model and long-term contract model.

The bidding model means that market participants (producers and customers) need to submit their orders in particular markets, containing power quantity and unit price. The electricity market maintainer will calculate the clearing price based on the strict power balance and supply-demand relationship. Finally, this clearing price will be used to clear all the electricity power accepted by the market. On the other hand, the long-term contract model mainly refer to some bilateral transaction contracts, which are not affected by the clearing system but will also affect the balance of power supply and demand.

At the same time, the forms of the electricity market was also changing; from the end of 20 century, some countries started the privatization process, like Italy, the United Kingdom and New Zealand. Privatization can effectively increase competition level, promote the healthy and orderly development of the power industry, and ensure the interests of users to a greater extent. On the other hand, countries like Japan and France still put the nation force as the major promotion for the development of their electricity markets. This kind of model can better take advantage of the power of the electricity industry and can be more effectively adapt to the social conditions and concentrate limited resources for development. However, no matter which model is selected, the electricity market should ensure fair competition and adapt to the development condition of the country and society.

1.2 The particularity of electricity

Different from conventional commodities, electricity as a commodity has some particularities; this also distinguished the electricity market from the others. The most important feature is its higher natural monopoly, which mainly stems from the following points:

1. **Difficult for storage**

   With current power storage technology, it is difficult to stably store electricity on a large scale. Therefore, common storage models are difficult to apply to the electricity market.

2. **Grid real-time balance**

   Besides, the power injection and power consumption in the grid must keep balance in real-time. Otherwise, frequency fluctuations and voltage changes will damage the electrical equipment. If it exceeds the power system's regulation capability, it will endanger the entire power system and cause unpredictable losses. Therefore, the production and consumption of electricity must be carried out at the same time.

3. **Irreplaceability of electricity**

   Electricity is the most important energy source in modern society. It is one of the bases for production and life. Therefore, the demand is enormous and cannot be replaced by other sources.

All the reasons above make electricity market more easily manipulated than ordinary ones. Therefore, Monopoly oligarchs often with a large market share can affect the transaction and price of electricity. We will discuss this in more detail in the following chapters.

1.3 History of the Italian electricity industry

1.3.1 Early period

The earliest power plant in Italy was established in Milan in 1883. It is mainly used for lighting and power supply of important buildings[1]. Limited by direct current and low-voltage transmission, the location of power plants and loads were relatively close. With the development of high-voltage transmission technology, Italy began to use hydroelectric power generation and firstly develop geothermal power generation in 1904[2]. For a long time after this, Italy had been the global leader in renewable energy power generation, and renewable energy can even meet all the electricity demand in Italy, including industrial needs. This state was not broken by until the rapid development of the economy and the surge in population in the 1960s.

1.3.2 State-owned period

Before 1962, the entire electricity industry in Italy was privately owned. However, due to the inherent monopolistic nature of electricity trading and request for fair competition, France and the United Kingdom took the lead in nationalizing the electricity industry in 1946 and 1957 respectively, and significant results had been achieved soon. Since then, the nationalization of the power industry has become a development trend in Europe.

Soon after, Italy established the state-owned enterprise ENEL S.P.A in 1962 and integrated more than 1,000 private power generation companies in the market, monopolizing all links in the production, transmission, and distribution of electricity. At this point, the Italian power industry has entered a state-owned period. With the substantial increase in power demand, ENEL has also begun try to use other energy sources such as fossil and nuclear energy [1] for power supply and has begun to import electrical energy actively.Now, Italy is still an electricity net importing country.

1.3.3 Privatization

With the free economy's global development, people have gradually lost confidence in public enterprises, so the power industry started its privatization process.

Since the promulgation of the Italian legislative decree 79/1999 ("Decreto Bersani") of 1999, the privatization of the electricity industry in Italy was dramatically accelerated, and the electricity production capacity that controlled by ENEL started transferred to private enterprises. Soon after, the dispatch and transmission institution Terna was established in 2002. Italy finally has its national electricity market, which is managed by Gestore dei Mercati Energetici S.P.A (GME). Since then, the privatization of the Italian power industry has continued to deepen. Now, it has already been opened to general users.

[1] Nuclear generation was banned by the referendum in 1978, because of the impact of the Chernobyl nuclear power plant accident.

1.4 The Italian electricity markets and their clearing mechanism

The Italian electricity market IPEX (Italian Power Exchange) is one of the European electricity market representatives, based on the EU Directive (EU, 2009). It is operated and managed by GME and be divided into a day ahead auction market(MGP), Intra-Day auction market(MI), and auxiliary service market (MSD). Each sub-market takes different responsibilities and cooperates to keep the electricity system a stable, healthy, and safe operating environment.

1.4.1 Day Ahead Market and clearing process

In the day-ahead market, the next day's electricity transaction is coupled in advance and is divided into 24 energy blocks by hour. Therefore, producers and consumers submit their produced and needed power quantity for each block, with acceptable prices, respectively. The entire MGP market opens at 8:00 am and closes at 12:00 am. The bidding results will be released at 12:55 pm. The market will eventually match each time block's supply and demand according to the price sorting to determine the clearing price. In the end, all accepted orders will be settled at this price.

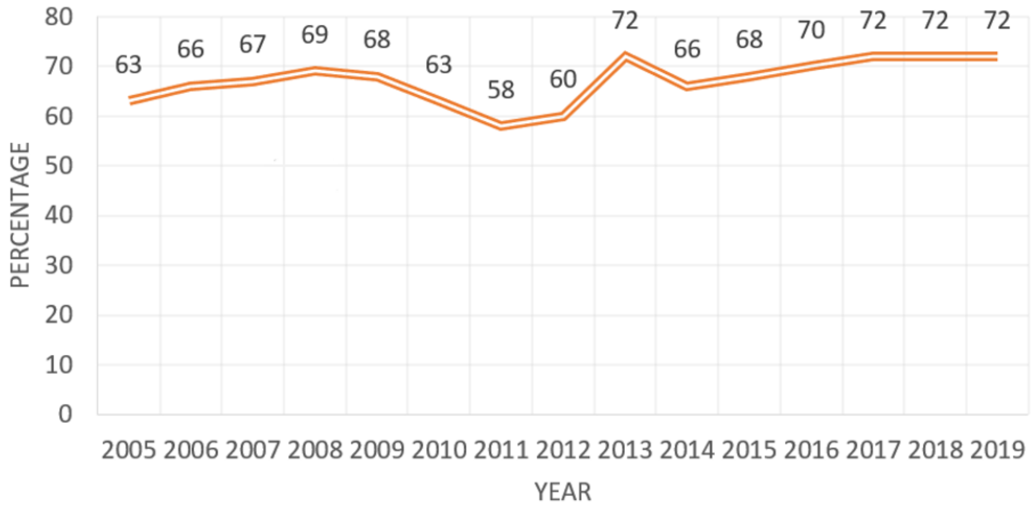


Figure 1-1: Electricity Liquidity of MGP (2005-2019)

The MGP market occupies a major position in the entire Italian electricity market. Its total electricity liquidity has been steadily maintained at more than 70 percent in recent years and has kept an upward trend, as shown in figure 1-1. Besides, the MGP market follows competitive rules and is the basis of MI, MSD. So it can best reflect the economic laws that exist in the entire Italian electricity market. Therefore, this thesis mainly focuses on the MGP market.

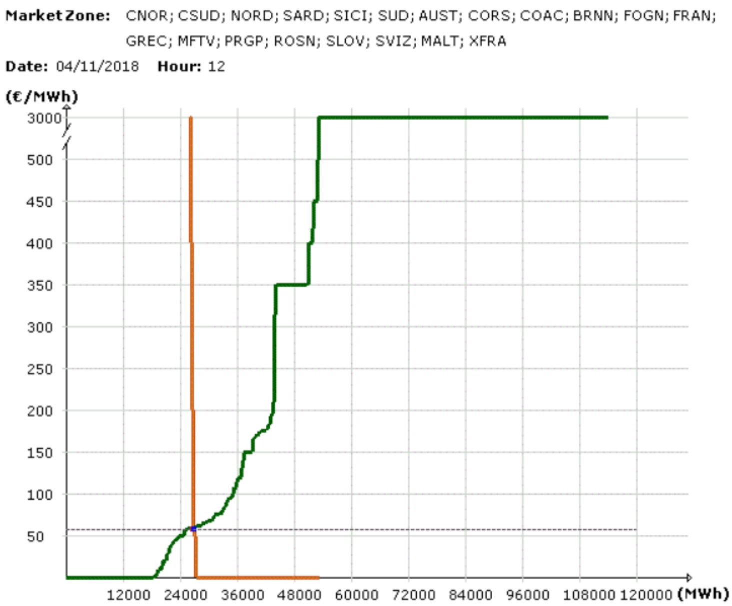


Figure 1-2: Clearing process in 4/11/2018

Figure 1-2[3] contains the supply-demand curve and the clearing process. The intersection of the two curves in the figure is the clearing point. The price at this point is the market clearing price, also known as the marginal clearing price. GME sorts the quotations of producers and consumers by ascending and descending order, respectively, and accumulates the corresponding power quantity. Usually, the marginal generator' s bid price will decide the market. At the equilibrium point, the demand and supply just entirely equivalent. For the generator side, all the bids lower(left side of the clearing point) than this price will be accepted, and the others will be rejected(right the clearing point).

1.4.2 Intra-Day Market

The Intra-Day Market (MI) market follows the same clearing mechanism as the MGP market. It starts after the results of the MGP market is released and is divided into seven segments according to time, from 12:55 pm on the previous day to 3:45 pm on the same day. Producers and consumers can adjust their output and consumption in the MI market in order to deal with some unexpected incidents to avoid affecting the balance of the power system.

1.4.3 Ancillary Services Market

The Ancillary Services Market (MSD) is the venue where Terna S.p.A.[2] procures the resources that it requires for managing and monitoring the system, relief of intra-zonal congestions, creation of energy reserve, real-time balancing. In the MSD, Terna acts as a central counterparty and accepted bids are remunerated by pay-as-bid, which makes the price often much higher than normal price[3].

[2]Terna is the owner of the Italian national transmission grid (NTG) for high and extra high voltage electricity and is the largest independent electricity transmission system operator (TSO) in Europe.Its tasks are to monitor and maintain the power grid uninterruptedly, and solve unexpected situations that occur during the operation through control and power dispatch, and ensure the orderly transportation and load balance.

[3]The offered price includes the basic cost and ”Generator startup fee” and other extra fees

Chapter 2

Competition and Market power

Competition and market power are the sources of the problem that need to be solved in this thesis, and it is also a very abstract concept. Unlike the data and information fields we are familiar with, the economics content is complex and changeable. Sometimes the factors considered in the different research of the same issue are often extraordinarily divergent, and many of them are based on empirical science; thus the definitions and interpretations are often not unified.

This is also one of the motivations of this thesis. We hope to combine machine thinking with economic problems to form a data-based problem model that minimizes human subjective factors.

This chapter will provide a brief overview of the core economic theories and their development involved in this thesis, including competition and market power, the relationship between market power behavior and abnormal behavior, and the introduction of problems and solutions.

2.1 Competition in electricity markets

Competition is generally accepted as a decisive factor of markets, and results from scarcity—there is never enough to satisfy all conceivable human wants—and occurs "when people strive to meet the criteria that are being used to determine who gets what."[4].

In economic activities, competition is a way that different economic firms seek to obtain a share of a limited good by varying the elements of the marketing mix: price, product, promotion and place. In classical economic theories, competition causes commercial firms to develop new products, services and technologies, which would give consumers greater selection and better products. The greater selection typically causes lower prices for the products. On the other hand, the unfair competition will have the opposite effect, such as dumping, where powerful companies grab market share by lower prices, crowd out small companies, and ultimately lead to industry monopoly[5].

As we mentioned above, electricity as a commodity has a higher monopolistic. Also, from the Italian electricity market development, we can find that the electricity industry has been a natural monopoly since its birth. Especially during the period of nationalization, the production, transmission, and sales of electricity are controlled by ENEL, and the price of electricity depends entirely on national policies and conditions. Even though the privatization process in the 21st century has broken this monopoly, it is still difficult to avoid unfair competition.

2.2 Market power in electricity markets

2.2.1 Definition of market power

Market power is one of the most important factors which cause unfair competition. It defines market participants' ability to affect the market and a necessary condition for unfair competition. In almost all economic textbooks, market power is defined as "the ability to profitably alter prices from competitive levels" (Krugman and Wells [2009]; Mankiw[2008]). However, although there are similar definitions in market power researches, the identification criteria are often different.

2.2.2 Market power behaviors and Abnormal behaviors

Because of electricity's particularity, there are many more constraints on many aspects of electricity trading like transaction mode and supply-demand relationship. So do the behaviors of participants, making it far more complicated than other commodity markets.

On the other hand, the fundamental purpose of business activities and competition is profit. That is, in general speaking, no market participant will harm their interests without purpose. Simultaneously, considering the clearing mechanism of the electricity market and the general electricity production costs, for most generating units, get the right to be clear in the market means guarantees the acceptable profit or minimize the loss in some particular cases[1]. Combining the above three points, the commercial behavior in the electricity market can be roughly divided into two types, normal behavior and abnormal behavior. The normal behavior is for obtaining a stable income, and another one is used to obtain a higher income.

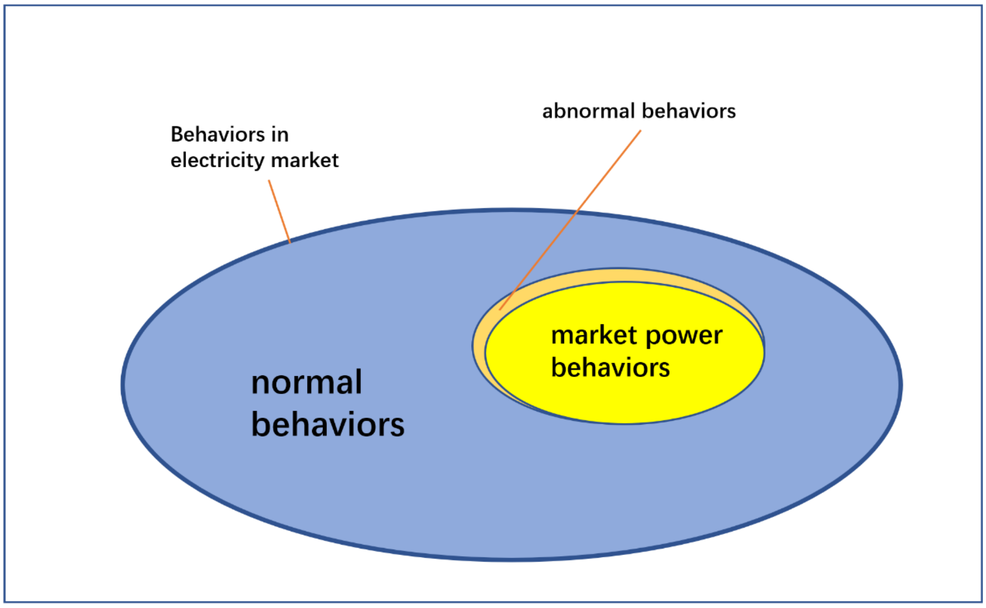


Figure 2-1: Behavioral Relation in the Electricity Market

Figure 2-1 shows the ideal relationships between all kinds of behaviors in electricity market to help better understand abnormal behaviors and market power behaviors.

[1]For example, some plants would prefer to keep the market access qualification even though submit a lower price rather than let the generators idle. Because the startup cost is often considerable. In this case, power plants can reduce the loss

2.2.3 Issues of Market power behaviors

As we mentioned, market power has various similar but different definitions, which also caused the difficulty and complexity of the identification of market power. The definations can be summarised from the following points:

1. **As an ability**

Market power is often defined as an ability, that is, market power can have two states: have and exercising. This means that participants who have market power may not certainly use it. Therefore, it is impossible to distinguish market power behavior by only considering the ability of participants and market share.

2. **Manipulate prices**

Participants with market power can use low prices to repel weaker competitors, but in this situation, the both sides will suffer losses at the same time. However, in fact, the side who exercising market power often has a stronger stress resistance. Therefore, Weak competitors will exit the market sooner. Companies that regain their monopoly position will make up for these losses in various ways in the future, which means that market price fluctuations are unpredictable; besides, they are often long-term operations. This makes it difficult to Judge market power behavior on single and short-term price changes.

3. **Different from competitive levels**

This is the most important part of the definition of market power, and also the most confusing part. Because "above the competitive level" is often reflected in price changes, but at the same time, the prices will also be affected by various factors. Even in the power spot market with more restrictive conditions, some large price fluctuations exist due to simple supply and demand relations, short-term environmental factors, regulatory policies, system operational status, players' bidding strategies, and other factors. For example, the California power crisis [2], although in this case, the imbalance of supply

[2] A demand-supply gap was created by energy companies, mainly Enron, to create an

and demand is mainly a human-made operation of gas supply, the subsequent sharp increase in electricity prices is common behavior on the power generation side. Therefore, judging whether the price is at a competitive level is itself a problem that needs to be solved first.

In summary, identifying market power behavior is still in a relatively vague state, and there is no specific and clear criterion for judgment. Without a unified standard, it is undoubtedly a great challenge to market regulation, which means it is difficult for regulators to judge market behavior effectively, also dramatically affects market efficiency and social welfare. This is the starting point of this thesis. We try to build a model that uses machine learning thought to replace human judgment to find out the deeper potential connections hidden in the data and make preliminary judgments on market power behavior in a short time.

2.2.4 Identification standard of market power behavior in EU

As mentioned above, the exercise of market power can take many different forms, such as controlling power generation resources, controlling short-term supply and demand, and often be completed in multiple steps, links,market participants, and long-term operation. This section will analyze two typical cases in Spain and Italy and their judgment process and relevant EU regulations to assess the current stage identification standards for market power behavior in EU countries in the current stage[6].

Case 1 :Spanish temporary congestion case

The Spanish competition authorities charged that four electricity producers had abused the monopolistic positions during three consecutive days with congestion problems. Each vertically integrated producer submitted unusually

artificial shortage. Energy traders took power plants offline for maintenance in days of peak demand to increase the price. Traders were thus able to sell power at premium prices, sometimes up to a factor of 20 times its normal value.

high prices during that time, far higher than the normal level.

Finally, the Tribunal sentenced each company to a penalty of 900,000 euros. However, what is interesting is that the court said they abused their dominant position because ”the producers should have priced as if they would have done in the absence of that problem.” This shows the absence of market power behavior judgment standards.

Case 2 : ENEL price manipulation

The Italian Energy regulator and competition accused ENEL of price manipulation in 2005 and 2006. However, publicly documents did not describe this case's details, just said ENEL used its leadership to repel other competitors. This company eventually offered to auction 700 MW per year in order to increase liquidity in markets. The warning signs of this case is obviously greater than the practical meaning. Finally, this case ended with ENEL auctions part of the production capacity.

Besides, in «Treaty on the Functioning of the European Union, Article 102», there are no very clear regulations on market power behavior and malicious competition, but only mentioned the vague market position abuse concept.

In summary, the current EU electricity market's regulatory system is not perfect and even lacks judgment standards, especially on market power abuse. The review time is often months or even years, which is almost no timeliness. It also makes judgments very difficult and far-fetched, makes it lack warning effect. Because of the power industry's importance, these problems have caused serious waste of resources, increased the burden on citizens, reduced social welfare, and hindered economic development.

2.2.5 Indicators of potential market power

In the current market power research field, the potential market power is mainly evaluated based on market share. Common classic indicators are as follows:

Concentration ratio

$$R_m = \sum_{i=1}^{m} a_i \quad (m \leq n) \tag{2.1}$$

In equation 2.1, n means the total number of participants, m is the number of first m largest ones. a_i is the market share of participant i.

	High	Median	Low	Perfect competition
R_4	> 71%	14% ~ 71%	0.01% ~ 14%	< 0.01%

Table 2-1: Evaluation ranges of Concentration ratio

Concentration ratio represents the sum of the market shares of the most influential companies in the market. The larger the value, the greater the influence of a few large companies, and the easier the market is to be manipulated.But it does not consider the influence of other competitors in the market.In some cases, these competitors should not be ruled out from considering, competitors with a smaller market share may jointly implement significant market power. For example, in the two markets of the same value of concentration ratio, if their total number of competitors is different, then their potential market power should be different either.

Herfindhal—Hirschman index (HHI)

$$\zeta = 10000 \cdot \sum_{i=1}^{n} a_i^2 \tag{2.2}$$

In equation 2.2, n means the total number of participants, a_i is the market share of participant i.

	No	Median	High
ζ	≤ 1000	$1000 \sim 1800$	> 1800

Table 2-2: Evaluation ranges of HHI

HHI can reflect the distribution of the market share of competitors in the market, that is, the degree of dispersion of the manufacturer's scale. Compared with concentration ratio, it considers the influence of the total number of competitors.

Entropy coefficient (EC)

$$E = -\sum_{i=1}^{n} a_i \cdot \ln(a_i) \tag{2.3}$$

In equation 2.3, n means the total number of participants, a_i is the market share of participant i.

Evaluation of CE :

$$\text{Monopoly } \Rightarrow 0 \leq E \leq \ln(n) \Leftarrow \text{ Equal share}$$

This indicator is derived from the information entropy formula. The larger the value, the greater the uncertainty and more uniform of the unit market share distribution.

Lerner index

The above indicators are mainly used to analyze the potential market power of the entire market and cannot evaluate the behavior of individual participants. Lerner index often be used for evaluate individual market power. Lerner index is also called the monopoly index, which reflects the potential meaning of the behavior by measuring the degree of deviation between the price and the marginal cost. The higher it is, the more unreasonable its behavior. Therefore, the Lerner Index can be used to evaluate the abnormal behavior in markets.

$$L = \frac{P - MC}{P} \tag{2.4}$$

Where P is the market price set by the firm and MC is the firm's marginal cost. The index ranges from 0 to 1.

However, in the electricity market, marginal cost is often one of the core secrets of power producers, which is difficult to obtain and predict. So we use approximate thinking to propose a conjectured cost curve.

Conjectured cost curve

From the publicly available data, we have no way of knowing the marginal cost curve of each power plant, so it is difficult to evaluate its Lerner index. Therefore, in this thesis we tried to use big data framework to extract the historical lowest bid price corresponding to the different capacities of each generator from the massive historical data. And use this to form a curve that approximates the marginal cost curve.

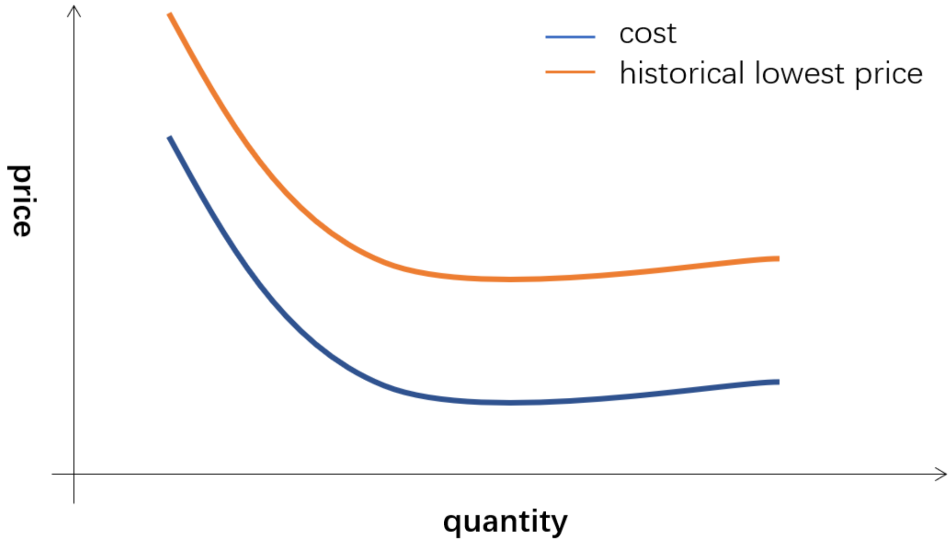


Figure 2-2: Ideal cost and lowest price curves

Based on this curve, we can get information from two aspects:

1. Approximate Lerner Index

   Generally speaking, cost is the lower bound of the lowest price, so the curve composed by the lowest prices can conjecture its cost curve. Therefore, in practice, we can observe the deviation of the specific bid price and the correspond lowest price, that is, approximate Lerner Index, to analyze the abnormality level of a specific bidding. And use this as one of the basis for verifying the results of our deep learning anomaly detection model.

2. Curve fluctuation

   Observing the changes of this curve of a specific generator can help us estimate the stability of its bidding. In theory, small fluctuations mean that the generator tends to make stable bids, and the probability of exercising market power is relatively small; otherwise, the opposite is true.

2.3 Motivation and goals

After the discussion of market power behavior and analysis of several more typical cases, it is not difficult to find that, compared with other goods, electricity as a commodity itself with its monopoly and irreplaceable significantly reduced the difficulty of market power exercising in electricity markets. Moreover, the damage and loss caused by excessive market power abuse are immeasurable; even every citizen is now unwittingly paying for all kinds of abuse behaviors now.

Determining market power behavior through pure economic theory is itself a difficult task at the current stage. This has led to inaccuracies and disagreements in the judgment of market power. Also, empirical statistical methods or expert systems currently do not have uniform standards and good performance; on the other hand, the data explosion brought by the vast daily trading volume also brings great difficulties.

Besides, the methods are also limited by human's current subjective, that is, most studies are limited to the surface or shallow connections of the data, unable to dig more in-depth information. Therefore, this thesis attempts to study from the Italian power spot market, mainly the MGP market's historical data, through deep learning to dig out the deep correlation of the data, in order to establish a preliminary judgment model of the power market abnormal data (bids). To establish a set of standards in the premise of considering as little artificial subjective factors as possible, and just depends on the data and machines' thought.

Firstly, according to historical data and the market power theories, the Italian electricity spot market's overall market power needs to be assessed. We decided to adopt a big data processing framework for this task in terms of data processing and calculation because of the large-scale data. On this basis, we then use deep learning algorithms to model and evaluate data and establish a preliminary detection model for abnormal data. For result verification, we will use Multi-aspect verification based on different economic and empirical theories.

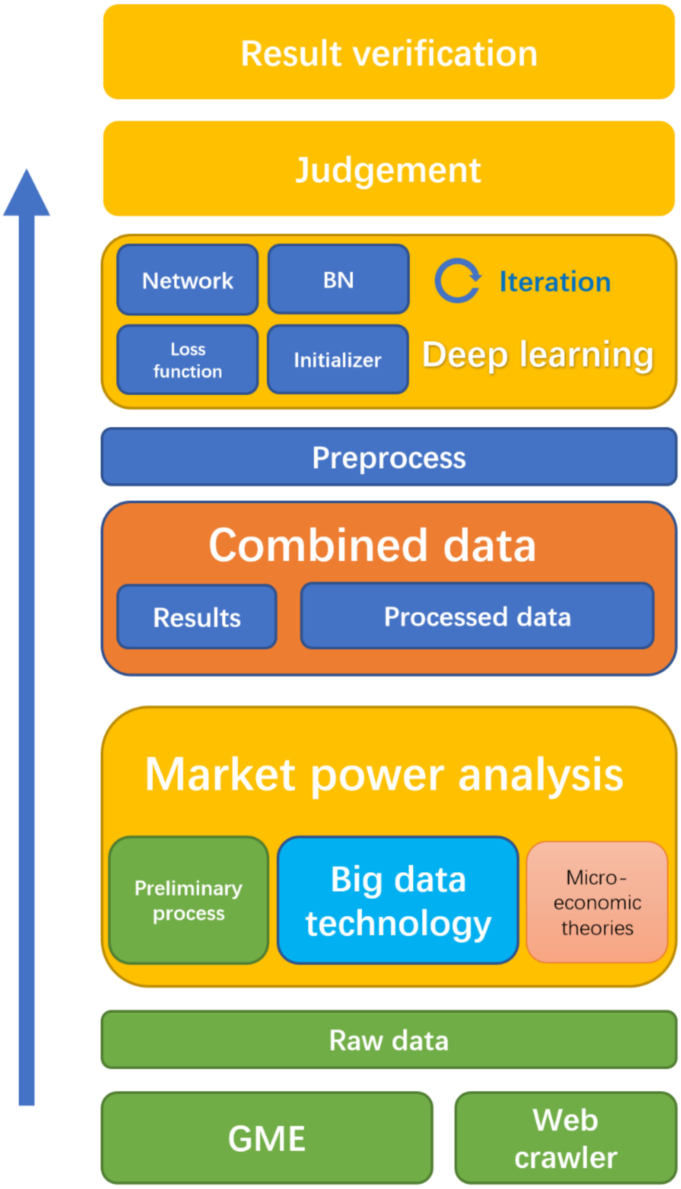


Figure 2-3: flow chart

Chapter 3

Technologies and tools

This chapter provides a brief introduction and description of the technologies and models that be used in this thesis, including big data processing frameworks, machine learning and deep learning, and specific algorithms. Moreover,this chapter will also give a brief description of data acquisition and preliminary processing.

3.1 Big data processing framework

Big data usually refers to data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time[7]. To process and analyze this kind of data, Big data processing technologies are invented, which has four characteristics: High volume, High Variety, High Velocity, High Veracity [8].Next, we are going to introduce the common tools and development of big data technology.

3.1.1 Hadoop

Apache Hadoop is a collection of open-source software utilities that facilitate using a network of many computers to solve problems involving massive amounts of data and computation. It provides a software framework for distributed storage (HDFS) and processing of big data using the MapReduce programming model[9].

3.1.2 HDFS

The Hadoop Distributed File System (HDFS) is the primary data storage system used by Hadoop applications. It provides a reliable method for managing pools of big data and supporting related big data analytics applications. It employs a NameNode and DataNode architecture to implement a distributed file system that provides high-performance access to data across highly scalable Hadoop clusters[10].

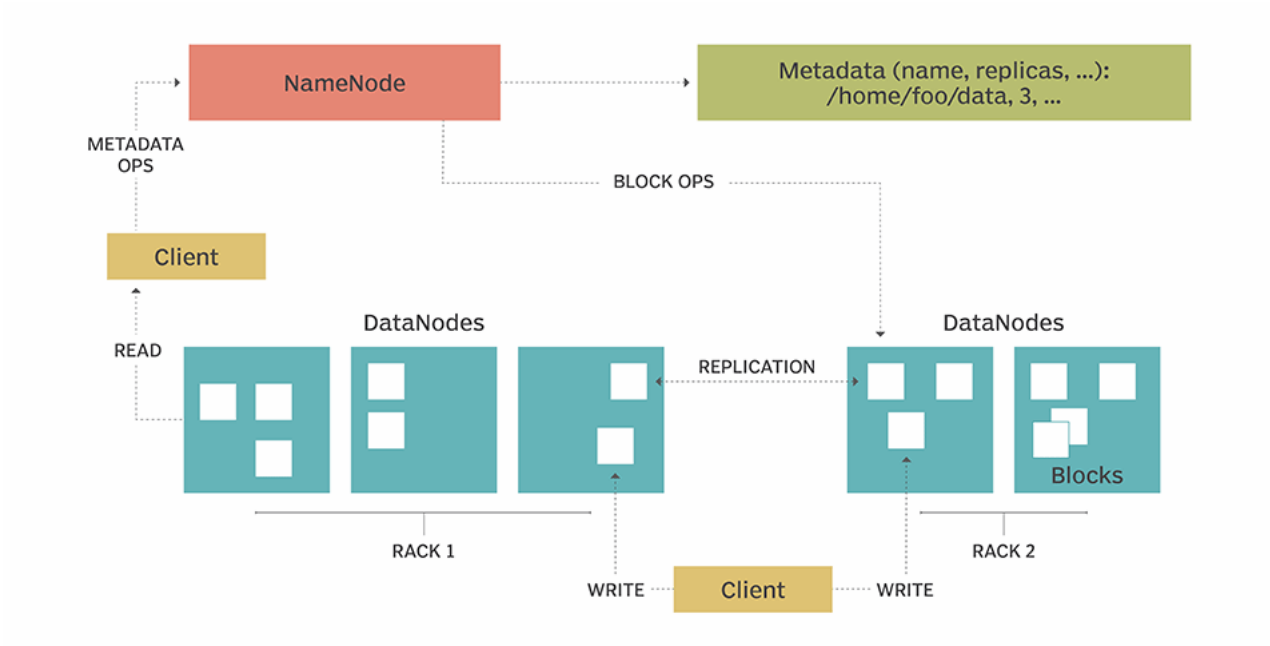


Figure 3-1: Architecture of HDFS

Figure 3-1 is the architecture of HDFS[10]. In HDFS, data is sliced into multiple blocks. All blocks in a file except the last block are the same size. Each block will generate multiple replicas and be stored in different DataNodes as backups. Namenode periodically receives a Heartbeat and a Blockreport from each of the DataNodes in the cluster. Receipt of a Heartbeat implies that the DataNode is functioning properly. A Blockreport contains a list of all blocks on a DataNode, and NameNode will gather and maintain all the information together. By this structure, HDFS mostly solved the problem of data failure due to hardware failures and other reasons, and significantly improves data security, system fault tolerance and reliability.

3.1.3 MapReduce

MapReduce is a framework proprietary software and introduced by Google to support distributed computing on large amounts of data in clusters of computers. Combined with HDFS, the same data set can be processed in parallel, which can significantly reduce the time required.

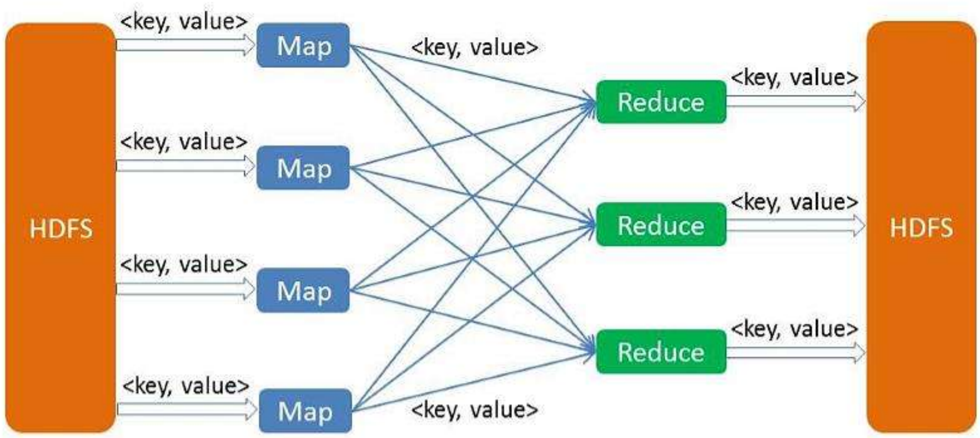


Figure 3-2: The process of Map-reduce

From figure 3-2, we can see the working process of MapReduce: In a MapReduce job, usually splits the input data-set into independent chunks, which are processed by the Map tasks in a completely parallel manner. Then, the framework sorts the outputs of the maps, which are then inputted to the reduce tasks. Typically both the input and the output of the job are stored in HDFS. The framework takes care of scheduling tasks, monitoring them.

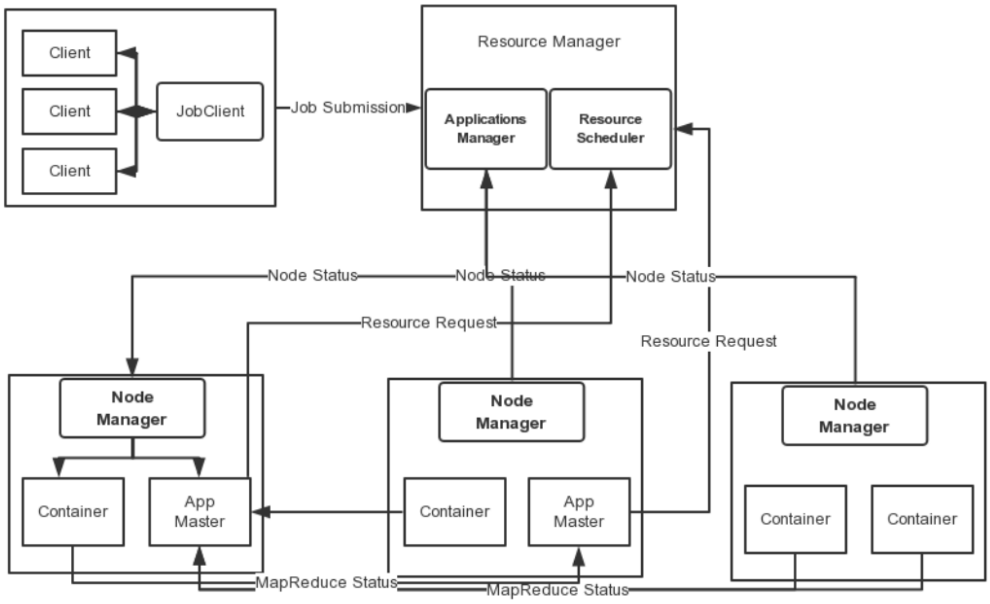


Figure 3-3: Architecture of Map-reduce

Figure 3-3 is the architecture of MapReduce. It consists of a single master ResourceManager, one worker NodeManager per cluster-node, and MRAppMaster per application. Applications specify the input and output locations and supply map and reduce functions by implementations of appropriate interfaces. Then, the Hadoop job client submits the job with configurations to the ResourceManager, which will distribute the software and configurations to the workers, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client[11].But MapReduce still has some limitation,and next, we will introduce the Apache Spark which is based on MapReduce and actually used in this thesis.

3.1.4 Spark

Apache Spark is a fast and versatile computing engine designed for large-scale data processing. Spark is an open source Hadoop MapReduce-like general parallel framework developed by UC Berkeley AMP lab (University of California, Berkeley). Compare to the Hadoop MapReduce, Spark not only has the all the advantages of Hadoop MapReduce, but also allows the intermediate output results can be stored in memory, so there is no need to read and write DFS (Distributed File System) when computing[12]. Thanks to the memory computing, Spark is potentially 100 times faster than Hadoop. Therefore, Spark is better suited for MapReduce algorithms that require many iterations, such as data mining and machine learning.

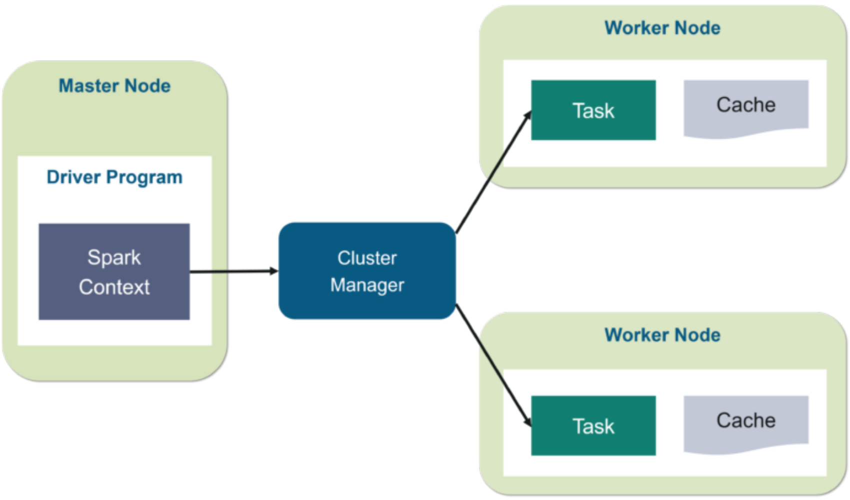


Figure 3-4: Spark working process

Figure 3-4 is the Spark work process.The The entire Spark cluster is composed of Master nodes and worker nodes. Among them, the Master daemon and Driver processes reside on the Master node. The Master is responsible for turning serial tasks into a set of tasks that can be executed in parallel, and also responsible for error handling. On the other hand, the Worker daemon resides on the Worker node. The Master manages all the Worker nodes, while the Worker node is responsible for executing tasks. The function of the Driver is to create a SparkContext and is responsible for executing the MAIN function of the Application written by the user. Process, Application is a program written by the user. There are one or more Executor processes with a thread pool on each Worker, and each thread is responsible for the execution of a Task task. According to the number of CPU-cores on the Executor, each of them can parallel multiple tasks at the same time.

Spark SQL

Spark SQL is a Spark module for structured data processing. It provides a programming abstraction called DataFrames and can also act as a distributed SQL query engine. Spark SQL brings native support for SQL to Spark and streamlines the process of querying data stored both in RDDs (Spark's distributed datasets) and in external sources. Spark SQL conveniently blurs the lines between RDDs and relational tables. Unifying these powerful abstractions makes it easy for developers to intermix SQL commands querying external data with complex analytics, all within a single application[13].

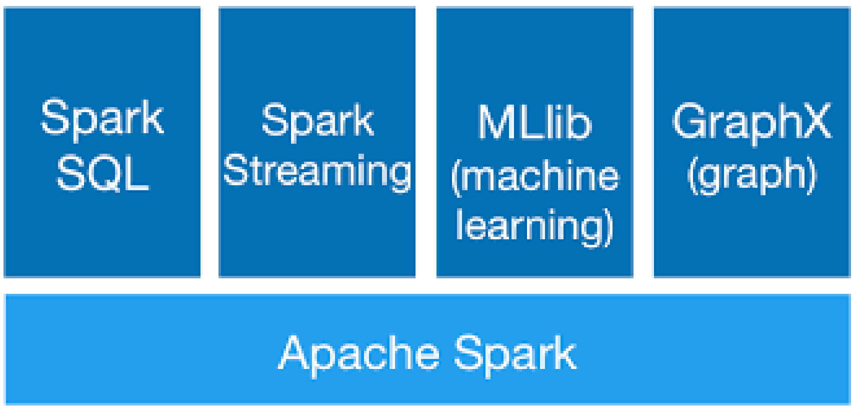


Figure 3-5: Spark framework

In this thesis, the historical data of market transactions is huge, and with the development of the market and accumulation year by year, it is destined to become larger in the future. Therefore, we use Spark SQL to process, integrate data and do preliminary analysis, which brought a significant improvement in work efficiency.

3.2 Machine learning

Machine learning (ML) is a branch of artificial intelligence, a study of computer algorithms that improve automatically through experience [14]. Machine learning algorithms build a mathematical model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so[Arthur Samuel,1959].

Machine learning has been used in many fields, such as architecture, medical treatment, communications, energy, and its mathematical models have provided help to solve many problems that are difficult to conventionally model, and greatly improved the efficiency. The main idea of machine learning is to learn autonomously from data, artificially propose solutions and algorithms to problems and program them, and the computer is continuously adjusted and optimized according to the goals, make the results are more and more accurate.

Machine learning approaches are traditionally divided into three broad categories, depending on the nature of the "signal" or "feedback" available to the learning system:

- **Supervised learning**

  The data contains various features and example output, which is called 'label'. What the computer needs to do is to try to find the relationship between each feature and the label, and the way of mapping from the feature data to the target output. It is often used as a classification and prediction problems.

- **Unsupervised learning**

  The data does not contain 'label'. Computers need to learn the structure and relationships in the data independently. It is commonly used in clustering problems and various problems with unclear definitions and data that are difficult to label, such as the problem in this thesis.

- **Reinforcement learning**

  Reinforcement learning is a computer program that interacts with a dynamic environment in which it must perform a certain goal. Reinforcement learning problems involve learning what to do—how to map situations to actions—so as to maximize a numerical reward signal. The learner is not told which actions to take, as in many forms of machine learning, but instead must discover which actions yield the most reward by trying them out[15].

3.3 Deep learning

Deep learning is an important branch of machine learning, known as deep structured learning, based on artificial neural networks, also can be supervised, semi-supervised, or unsupervised. The types of deep learning mainly include deep neural networks, deep belief networks, recurrent neural networks, and convolutional neural networks, which have been widely used in various industries such as medical services, natural language processing, computer vision, and biotechnology. The results are comparable to even surpassing human expert performance in some cases.

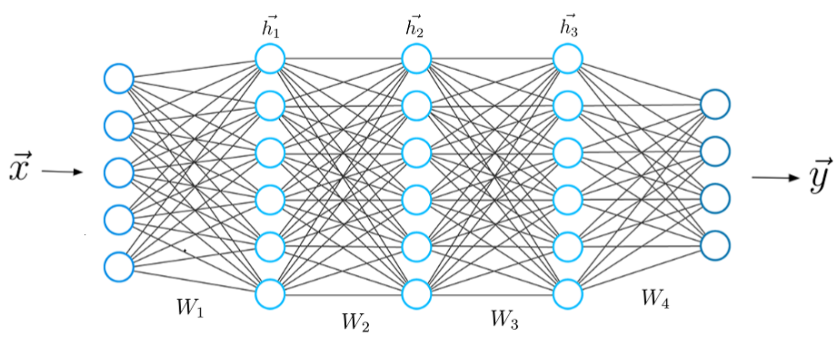


Figure 3-6: Structure of a dense neural network with 3 hidden layers

Like what is shown in Figure 3-6. The word 'deep' means the structure of deep learning, which is multiple layers. Compared with traditional machine learning, deep learning is good at using nonlinear expressions to express and build model complex problems, rather than linear models. The limitation of the linear model lies in the inability to fit well complex problems or the internal connection of data that humans cannot observe now. The nonlinear model can break through this limitation, the deep learning models can describe the relationship between input and output almost infinitely "finely" through the combination of multiple hidden layers and multiple neurons, which even can exceed the scope of human cognition. Therefore, the interpretability of deep learning is not as good as machine learning; that is why it is often referred to as a "black box".

3.4 Variational Auto-Encoder

In this thesis, we chose Variational Auto-Encoder as the algorithm of the anomaly detection model. It is a variant of Auto-Encoder, but the internal principle is quite different. In this section, we will introduce the idea of Auto-encoder and the inference of Variational Auto-Encoder in detail.

3.4.1 Auto-Encoder

Auto-encoder is a neural network algorithm commonly used for semi-supervised learning and unsupervised learning. The structure can be divided into two parts: encoder and decoder. In applications, it is often used as a data dimensionality reduction tool, and can also be used for anomaly detection.

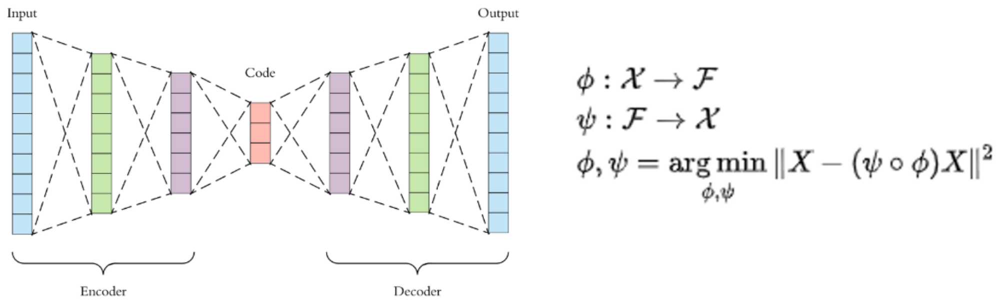


Figure 3-7: Structure and objective of Auto-Encoder

Firstly, the first half of the neural network (encoder) converts the input data into an intermediate layer (code), which can be regarded as the reduced-dimensional expression of the input data. The second half (decoder) maps the code to a reconstruction of the original input. The algorithm self-update iteratively to reduce the error between input and output.

3.4.2 Auto-Encoder based anomaly detection

Anomaly detection using dimensionality reduction is based on the following assumption: data has variables correlated with each other and can be embedded into a lower-dimensional subspace in which normal samples and anomalous samples appear significantly different[16].

In 2015, Mayu Sakurada and Takehisa Yairi proposed using deep auto-encoders to identify anomalies based on reconstruction errors[17]. They used normal data as a training set and deep auto-encoders as a model to simulate its internal connection. In the test phase, they judged whether it is abnormal by observing the reconstruction error of the data.

Algorithm 3.1: Autoencoder based anomaly detection

Input: Normal dataset X, Anomalous dataset $x^{(i)}$ $i = 1, \cdots, N$, threshold α

Output: reconstruction error $\|x - \hat{x}\|$

$\phi, \theta \leftarrow$ train a autoencoder using the normal dataset X

$\phi \leftarrow$ encoder, $\theta \leftarrow$ decoder

for i = 1 *to* N **do**

  reconstruction error$(i) = \left\| x^{(i)} - g_\theta\left(f_\theta\left(x^{(i)}\right)\right)\right\|$

  if *reconstruction error* $(i) > \alpha$ **then**

    $x^{(i)}$ is an anomaly

  else

    $x^{(i)}$ is not an anomaly

  end

end

Algorithm 3.1 shows how the Auto-Encoder based anomaly detection model works. However, auto-encoder also has its limitation. It is too dependent on the training set to represent the general law of the whole data set accurately. But the historical data of electricity market is changeable. Therefore, we decide to use variational auto-encoder to solve this problem, which will be introduced in the next section.

3.4.3 Variational Auto-Encoder

Variational auto-encoder is a generative model, which is firstly proposed by Diederik P. Kingma and Max Welling in 2013[18], and was published as tutorial[19] by Carl Doersch in 2016. Instead of using a compressed representation to represent the data like Auto-encoder, variational Auto-encoder was designed to learn the parameters of a probability distribution that can represent the data. So it has a better Robustness and less reliance on original data. In a word, variational auto-encoder can make the model applicable to not only the current training set data but a type of data. Next we are going to introduce the inference of variational auto-encoder and some relevant knowledge.

Kullback—Leibler divergence

In mathematical statistics, the Kullback—Leibler divergence, also called relative entropy, is a measure of how one probability distribution is different from a second, reference probability distribution[20]. Kullback—Leibler divergence of zero indicates that the two distributions in question are identical. It is a very useful statistics tool with diverse applications such as applied statistics, fluid mechanics, neuroscience and machine learning[21].

$$D_{\mathrm{KL}}(P \| Q) = \sum_{x \in \mathcal{X}} P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{3.1}$$

Where $\mathcal{P}$ and $\boldsymbol{Q}$ are defined discrete probability distributions on the same probability space $\mathcal{X}$.

Inference of Variational auto-encoder

For a generative model, the goal is to find the law of observed data and generate data that conforms to it. We can assume this law as an existed latent distribution $Q(Z)$, that is, the observation data set X can be mapped from $Q(Z)$. But in practice, we can't get $Q(Z)$ directly, so we try to use the observation data set X to infer $Q(Z)$, which is the posterior probability $P(Z \mid X)$.

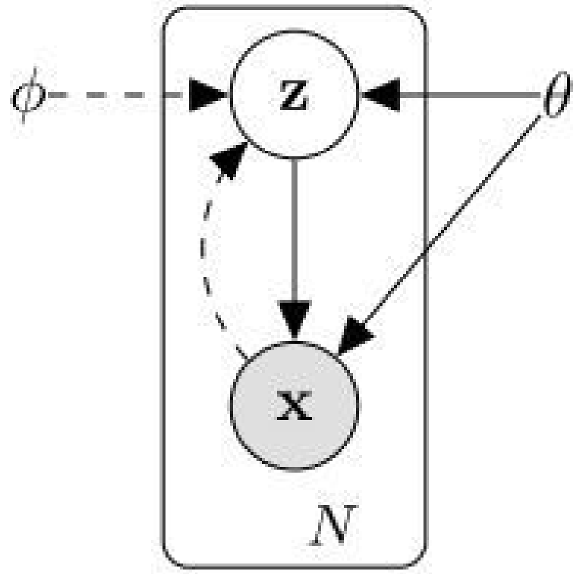


Figure 3-8: The generative model

However, the calculation of posterior probability is often very complicated and cannot be solved in polynomial time, which means it is an intractable problem. Therefore, variational inference uses the idea of problem transformation to approximate $P(Z \mid X)$ with tractable $Q(Z \mid X)$. That is, $Q(z \mid x) \approx P(z \mid x)$, like figure 3-9.

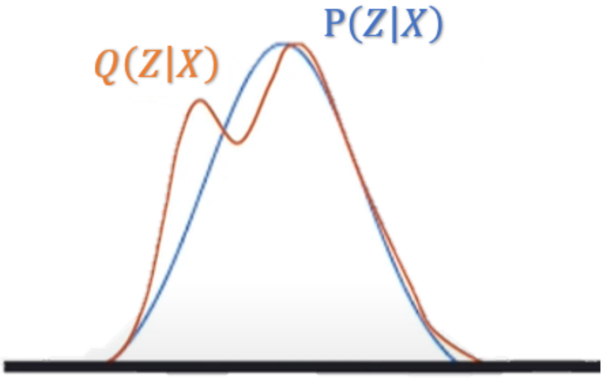


Figure 3-9: Idea of approximation

To make the approximation, we first need a standard to measure the difference between $Q(Z \mid X)$ and $P(Z \mid X)$, which is the KL divergence we mentioned earlier. For making $Q(Z \mid X)$ a tractable distribution, We only need to manually define it and make it close to $P(Z \mid X)$. The process is as follows:

The KL divergence between $Q(Z \mid X)$ and $P(Z \mid X)$ is:

$$
\begin{aligned}
D_{KL}(Q(Z \mid X) \| P(Z \mid X)) &= -\sum_Z Q(Z \mid X) \log \frac{P(Z \mid X)}{Q(Z \mid X)} \\
&= -\sum_Z Q(Z \mid X) \log \frac{P(X, Z)}{Q(Z \mid X) P(X)} \\
&= -\sum_Z Q(Z \mid X) \log (\frac{P(X, Z)}{Q(Z \mid X)} \frac{1}{P(X)}) \\
&= -\sum_Z Q(Z \mid X) [\log \frac{P(X, Z)}{Q(Z \mid X)} - \log P(X)] \\
&= -\sum_Z Q(Z \mid X) \log \frac{P(X, Z)}{Q(Z \mid X)} + \sum_Z Q(Z \mid X) \log P(X)
\end{aligned} \tag{3.2}
$$

From inference 3.2, we can get a equation 3.3:

$$
logP(X) = KLD + \sum_Z Q(Z \mid X) \log \frac{P(X, Z)}{Q(Z \mid X)} \tag{3.3}
$$

We know that $KLD \geqslant 0$ and $logP(X)$ is fixed, which means if we want to minimize the KLD, just need to maximize the second term on the right side, which is called variational lower bound of $logP(X)$. Like figure 3-10.

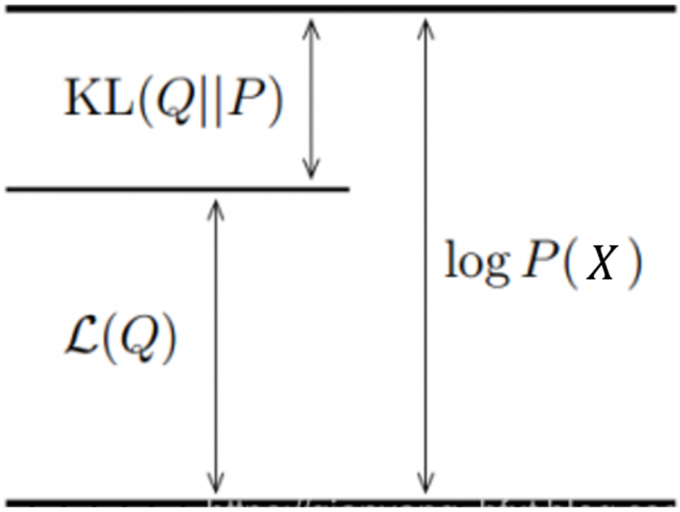


Figure 3-10: Lower bound

Then, the maximum process is:

$$
\begin{aligned}
L[D_{KL}(Q\|P)] &= \sum_Z Q(Z \mid X) \log \frac{P(X,Z)}{Q(Z \mid X)} \\
&= \sum_Z Q(Z \mid X) \log \frac{P(X \mid Z)P(Z)}{Q(Z \mid X)} \\
&= \sum_Z Q(Z \mid X)[\log P(X \mid Z) + \log \frac{P(Z)}{Q(Z \mid X)}] \\
&= \sum_Z Q(Z \mid X) \log P(X \mid Z) + \sum_Z Q(Z \mid X) \log \frac{P(Z)}{Q(Z \mid X)}
\end{aligned}
\tag{3.4}
$$

Here, equation 3.4 can be rewrite as equation 3.5

$$
\begin{aligned}
L[D_{KL}(Q\|P)] &= E_{Q(Z|X)} \log P(X \mid Z) + D_{KL}(Q(Z \mid X)\|P(Z)) \\
&= \|x_i - \hat{x}_i\|^2 + D_{KL}(Q(Z \mid X)\|P(Z))
\end{aligned} \tag{3.5}
$$

So the Minimum process can be turned as:

$$
\min L[D_{KL}(Q\|P)] = \min \|x_i - \hat{x}_i\|^2 + \min D_{KL}(Q(Z \mid X)\|P(Z)) \tag{3.6}
$$

Equation 3.6 is the objective function of variational auto-encoder. It tells us we need to minimize reconstruction error and make $Q(Z \mid X)$ close to a manually selected $P(Z)$ as much as possible. To do that, we can use neural networks to simulate the processes of $Q(Z \mid X)$ and $P(X \mid Z)$ like figure 3-11

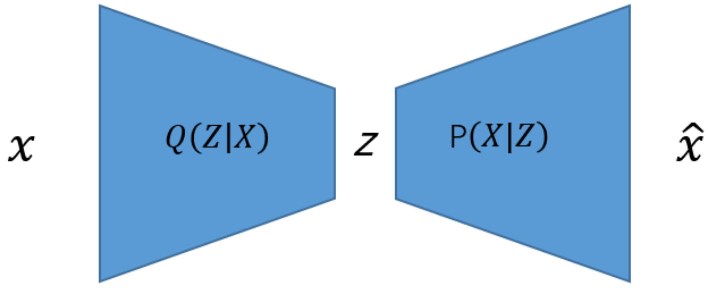


Figure 3-11: VAE without re-parameter trick

Besides, we need to use the re-parameter trick to fulfill the requirements second term of objective function $\min D_{KL}(Q(Z \mid X) \| P(Z))$. Making the encoder part composed of two neural networks to generate parameters of distribution mean μ and variance σ.Then,randomly sample a code ϵ from $P(X)$, and combine them by $\epsilon * \sigma + \mu$ as the intermediate latent code Z. The final structure is in figure 3-12. For anomaly detection, the process is similar to algorithm 3.1; the only difference is the objective function.

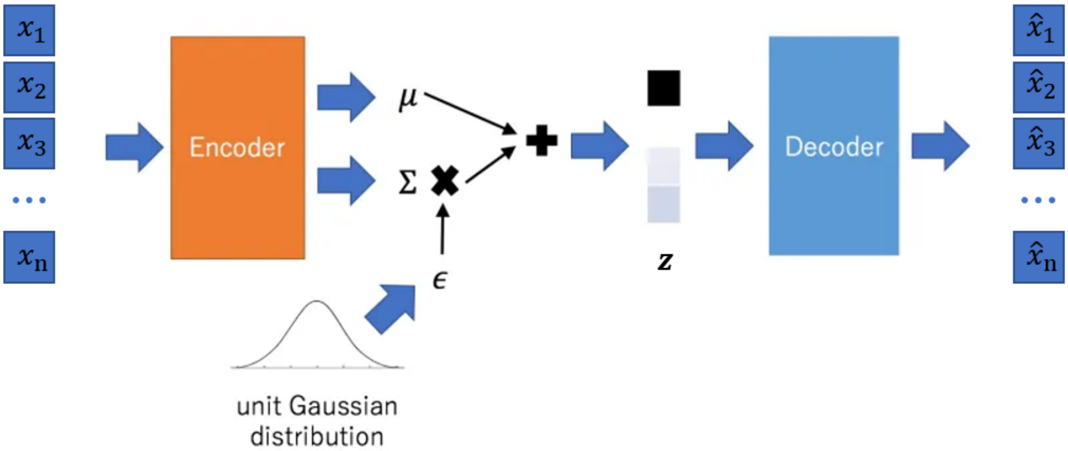


Figure 3-12: structure of variational auto-encoder

Chapter 4

Market power analysis and anomaly based market power behavior detection model

In this chapter, we will introduce the actual experimental processes, including data selection and initial processing, the overall analysis of the potential market power of the Italian electricity market based on big data technologies, and the establishment of abnormal model based on Variational Auto-Encode.

4.1 Data introduction

4.1.1 Data acquisition

As we mentioned before, GME is the manager of the Italian electricity market, including the day-ahead market, intraday market, auxiliary market, and even gas and other markets. All the real-time transaction data are also recorded and stored by GME in real-time.

The research objective of this thesis is the market power and abnormal behavior analysis of the day-ahead market, so the initial data comes from the official data set published by GME.

In terms of data selection, considering the hardware conditions and the timeliness of the data, we selected two years of real-time transaction data of the day-ahead market. Due to the impact of the Covid-19 epidemic in 2020, the electricity market has been dramatically affected. According to the GME report published in July 2020, the average transaction price had a drop of more than 40% compared to 2019, shown in figure4-1. Therefore, we have removed the data of this period.

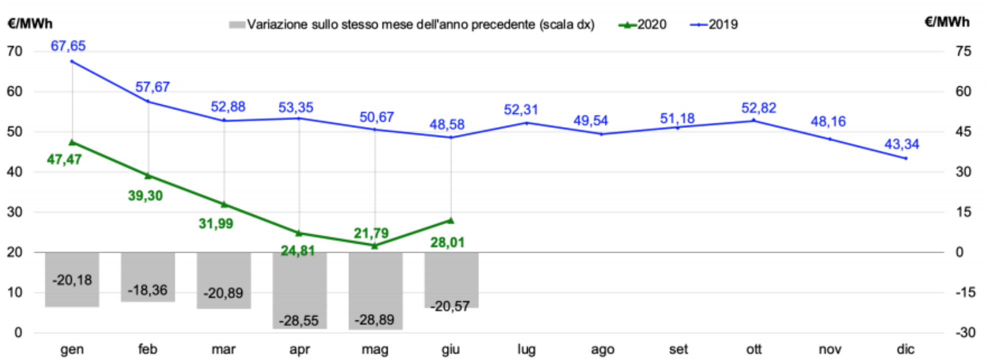


Figure 4-1: Epidemic impact on electricity wholesale price

4.1.2 Data composition

MGP public transaction data includes power transactions in 24 time blocks each day, contains the supply bid price of each generator on the power generation side and the purchase price of users on the power consumption side, as well as specific participant information, transaction power and other information. The table 4-1 shows the features and their brief description of the public data set.

Feature	Description
ADJ_ENERGY_PRICE_NO	Price possibly adjusted by the system
ADJ_QUANTITY_NO	Submitted volume, possibly adjusted by the system
AWARDED_PRICE_NO	Price awarded by the market
AWARDED_QUANTITY_NO	Volume awarded by the market
BID_OFFER_DATE_DT	Flow date in the YYYY/MM/DD format
BILATERAL_IN	It indicates whether the bid/offer comes from the PCE platform
ENERGY_PRICE_NO	Price submitted by the participant
GRID_SUPPLY_POINT_NO	Grid supply point with which the unit is associated
INTERVAL_NO	Relevant period to which the bid/offer refers
MARKET_CD	Market code
MERIT_ORDER_NO	Merit order of the bid/offer as calculated by the market solution algorithm
OPERATORE	Registered name of the participant
PARTIAL_QTY_ACCEPTED_IN	Indicator of partially accepted bid/offer
PRODOTTO	For the MPEG only: specify the identification code of the product being offered
PURPOSE_CD	Purpose of bid/offer
QUANTITY_NO	Volume submitted by the participant
STATUS_CD	Status of bid/offer after market execution
SUBMITTED_DT	Time of submission
TRANSACTION_REFERENCE_NO	GME's identifier of the bid/offer
TYPE_CD	It indicates whether the bid/offer is predefined or current
UNIT_REFERENCE_NO	Unit code
ZONE_CD	Zone to which the unit belongs

Table 4-1: features of data

4.2 Analysis of the potential market power of the Italian electricity market

In this section, we will introduce the analysis of potential market power in the Italian electricity market. Based on the theories and analysis mentioned in chapter 2, we use Spark framework to do this job, including the overall trend, traditional long-term market power indicator analysis. Besides, we also calculated the short-term(per hour) market power indicators and simulate the conjectured cost curve of all market participants using spark's strong computation ability.

4.2.1 Overall trend

The Italian electricity market came into being earlier and developed for a long time. It has been continuously optimizing related laws, regulations and trading rules. In the GME's annual reports of recent ten years, there have always been a small space describing the market concentration. The report said that the market concentration has been decreasing year by year after 2010. Although the change is small, it still means that potential market competition is on the rise.But there is a lack of detailed description.To have a better overview of current situation of the Italian electricity market,based on recent years' market transaction data, we calculated several indicators to analyze potential market power in the Italian electricity market.

4.2.2 Market share

We accumulated the awarded bidding power quantity of all market participants to obtain the total annual capacity of the market and calculate the capacity of each operator to obtain the corresponding market share. Table 4-2 and table 4-3shows the top operators and generators in the MGP market respectively.

OPERATORE	**Market share %**
ENEL PRODUZIONE S.P.A.	31.66367
GSE SPA	9.624354
EDISON SPA	7.86633
A2A SPA	7.559669
EP PRODUZIONE SPA	4.536951
DXT COMMODITIES SA	4.096402
AXPO ITALIA SPA	3.164841
IREN ENERGIA SPA	2.855435
EGO TRADE SPA	2.77098
ENI SPA	2.482258
EGO ENERGY S.r.l.	2.434658
ERG POWER GENERATION S.P.A.	2.214254
SORGENIA S.P.A.	2.045767
ENGIE ITALIA SPA	1.899757
DOLOMITI ENERGIA TRADING SPA	1.89582
TIRRENO POWER S.P.A.	1.759566
ALPERIA TRADING SRL	1.632136
DANSKE COMMODITIES A.S	1.310612
EDELWEISS ENERGIA S.P.A.	1.193733
C.V.A. TRADING S.R.L. A S.U.	1.160818
Total	91.8

Table 4-2: Top 20 operators in market share

UNIT_REFERENCE_NO	OPERATORE	market_share(%)
UP_DI8888_NORD_Y	GSE SPA	4.239477
UP_LCSELLACLE_5	Enel Produzione S.p.A.	1.537101
UP_DI8888_SUD_Y	GSE SPA	1.258541
UP_ETQCHIOTAS_1	UP_ETQCHIOTAS_1	1.234594
UP_PRESENZAN_1	Enel Produzione S.p.A.	1.081897
UP_DI8888_CSUD_Y	GSE SPA	0.9681
UP_DI8888_CNOR_Y	GSE SPA	0.941045
UP_SMRICRICHI_1	EDISON SPA	0.935099
UP_OSTIGLIA_12	Enel Produzione S.p.A.	0.909076
UP_NCTLVRNFRR_1	EP PRODUZIONE SPA	0.904449
UP_TORVISCOSA_1	EDISON SPA	0.893346
total		14.9

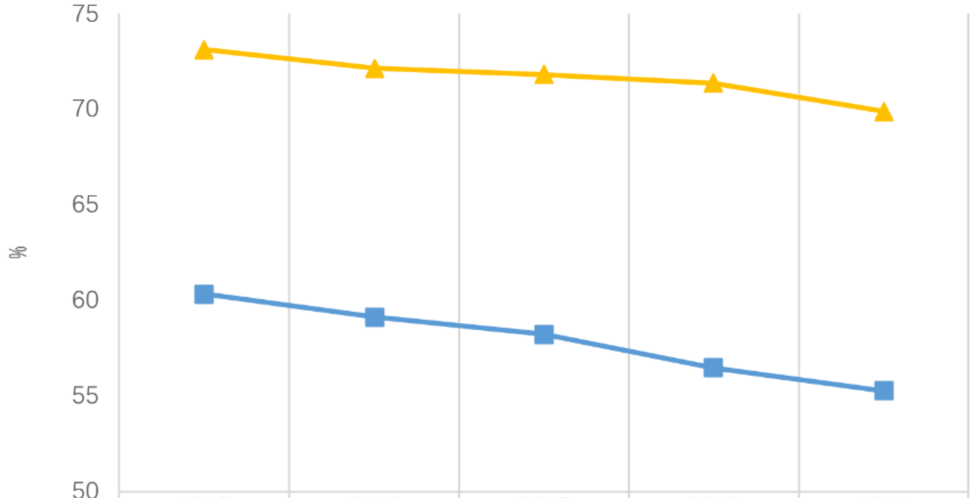
Table 4-3: Top generators in market share

It can be seen from table 4-2 that the current Italian electricity market share is unevenly distributed. A small number of large competitors control most of the market share. The top 20 operators occupy 90% of the market, and the largest suppliers even control Occupies about 30% of the market capacity. In addition, we calculate the market share of each single generator, as shown in Table 4-3 .Combining the two tables, it can be seen that the most of generators with larger capacity are controlled by large companies, which further illustrates the higher monopoly risk.

4.2.3 Long-term indicators

Concentration ratio

Based on the market share, we calculated the concentration ratios of the electricity market with the top 4 and 8 largest operators respectively.



	2015	2016	2017	2018	2019
R4	60.34	59.12	58.22	56.45	55.26
R8	73.11	72.15	71.81	71.37	69.89

Figure 4-2: Concentration ratio in recent 5 years

As seen from the figure 4-2, the concentration ratio has slowly declined in recent years, but still stay at the medium level, which means there still exists a considerable space to exercise market power .

Herfindhal—Hirschman index (HHI)

In order to take into account the contribution of all market participants to potential market power, we calculated the HHI indicator.

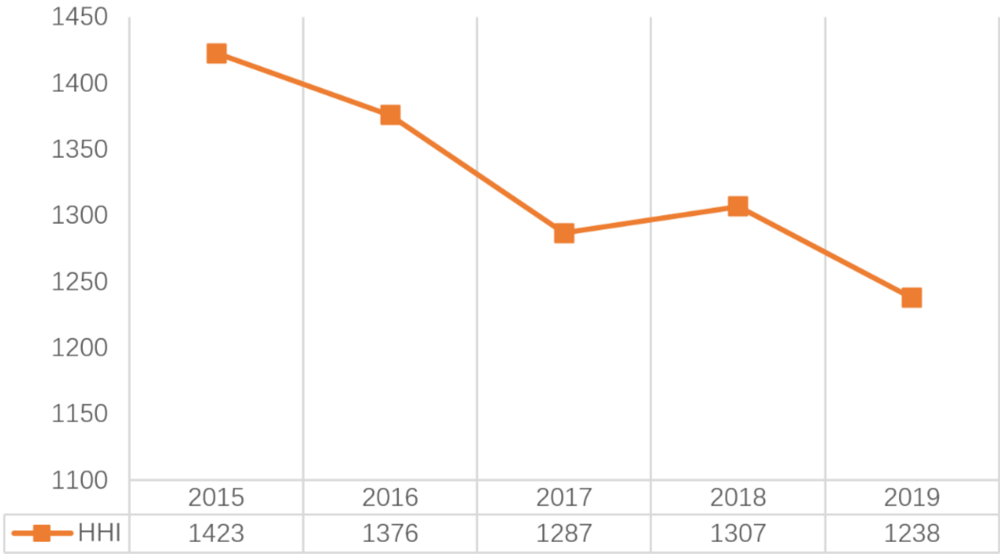


Figure 4-3: Herfindhal—Hirschman index in recent 5 years

The change of HHI is similar to CR, indicating a medium level of potential market power in the market. It also reflects that although participants have increased year after year, small competitors only contribute little for the improvement of competition level in the market.

Entropy coefficient (EC)

Year	EC	Ln(n)
2018	2.67	4.36
2019	2.89	4.49

Table 4-4: Entropy coefficient of 2018 and 2019

Similarly, taking 2018 as an example, the results of the Entropy coefficient indicate that the market has considerable potential market power.

4.2.4 Short-term indicators

The above indicators can reflect the potential market power in the long-term. However, in reality, the abuse of market power often occurs in a short period. Therefore, we took advantage of big data calculations, and regarded each time interval as a complete market and analyzed various market power indicators in each of them:

Risk	Date	Interval	R4	R8	HHI	EC	Ln_n
High	20190602	14	60.31025	71.70554	1871.375	2.452034	4.204693
	20190602	15	59.1852	70.83533	1859.349	2.458346	4.174387
	20190602	13	57.45413	70.46831	1713.023	2.544793	4.248495
	20190421	15	65.31603	76.14421	1701.98	2.513027	4.127134
Medium	20190421	13	57.64303	71.51679	1377.609	2.729933	4.234107
	20190330	13	59.01871	73.05293	1377.536	2.737831	4.317488
	20190623	12	50.47228	67.07312	1373.093	2.756992	4.204693
	20190616	12	53.85427	71.43303	1369.712	2.70033	4.26268
Low	20190205	22	35.8008	56.98114	532.3672	3.228324	4.219508
	20190211	21	33.98701	58.37648	531.2387	3.186753	4.204693
	20190112	22	35.5401	57.36011	524.4473	3.137496	4.219508
	20190509	24	36.6434	54.4378	522.0256	3.210685	4.248495

Table 4-5: Indicators of several short-term markets

From these two tables, We can find, that at some times, different indicators show different information, for example, in 2/06/2019, the HHI increased over the high-level threshold, but can not find significant abnormal from other indicators at the same time.This may be caused by the specific operation of some competitors, also prove that the market power behaviors is very complex which is hard to be identified from single or only a few aspects.

4.2.5 Summary

All the results show that there is a medium level potential market power in the Italian electricity market. Moreover, from the comparison, we can find that long-term indicators only can give an overall static analysis but can not help with short-term identification. In the current market power analysis field, they are still accustomed to using long-term static indicators, which caused the weakness of regulation. Market participants may exercise market power behavior in the short term to obtain high profits and use long-term operations to evade supervision. Therefore, it is necessary to have a real-time, effective anomaly detection methods.

4.3 Anomaly based market power behavior detection model

4.3.1 Data preprocessing

Raw data processing

In the preprocessing stage, we need to perform some operations on raw data, such as filtering, transforming, and merging. Some operations require complicated computation, such as grouping computation on every single generator and integrating the results. To have the best efficiency, we used some operation methods of spark SQL, including Filter, Group, Join, and other methods.

Besides, the data format conversion is another necessary process. For example, some features are string formats like ”ZONE_CD” and the IDs of the operators and generators. These kinds of data cannot be directly used in the neural network, so we use stringIndexer in Spark ML to quantify them into indexes.

UNIT_REFERENCE_NO	CorrespondingIndex
UP_DI8888_NORD_Y	1
UP_SARLUX_1	32
UP_SMRICRICHI_1	143
UP_TORVISCOSA_1	562
UP_ALTOMONTE_1	9
UP_VADOTERM_5	56

Table 4-6: Examples of string data conversion

Standardization

Standardization is a data preprocessing method for neural networks. If the values of different features in the data set are quite different, it will affect the size of the updated weights in each iteration. Our data set exists this problem, so we used standardization to reduce the impact of the feature magnitude. At the same time, it will also make the data follow distribution $X \sim N(0,1)$, in order to improve the model convergence speed and reduce the probability of overfitting.

$$z = \frac{x_i - \mu}{\sigma} \tag{4.1}$$

4.3.2 Activation function

The activation function is one of the cores of neural networks; it can make the network able to express or simulate specific functions in a non-linear way.

Without the activation function, the expression of each neuron is linear: $y = wx + b$. Therefore, no matter how many hidden layers and neurons the neural network has, the output will always be a combination of linear expressions. This kind of network is hard to simulate complex functions. However, using activation functions can make the expression of the network non-linear and give full play to the role of hidden layers and neurons to maximize the expression ability of the network. Figure 4-4 shows the comparison between linear and non-linear classification.

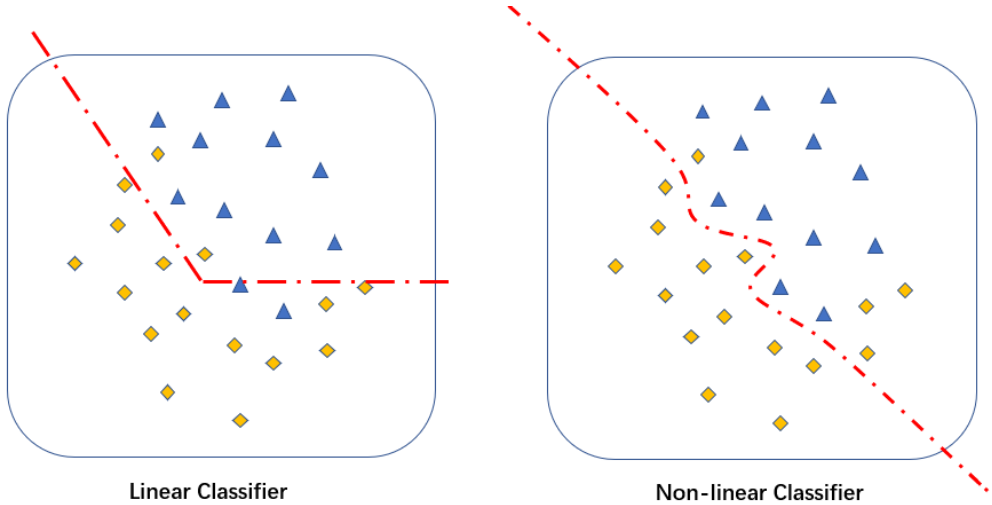


Figure 4-4: linear and non-linear classification

Commonly used activation functions include sigmoid, tanh, relu. In this thesis, we selected tanh function due to its better convergence efficiency. However, this may cause gradient vanishing problem. Make tanh as an example, if the network contains many hidden layers, the continuous multiplications of the weights during the back-propagation process will make the results of tanh close to the edge of the function diagram(figure 4-5), which will make the gradient close to 0. In this case, the hidden layers at the beginning of the network will not be updated effectively. To solve this problem, we will use initialization and batch normalization.

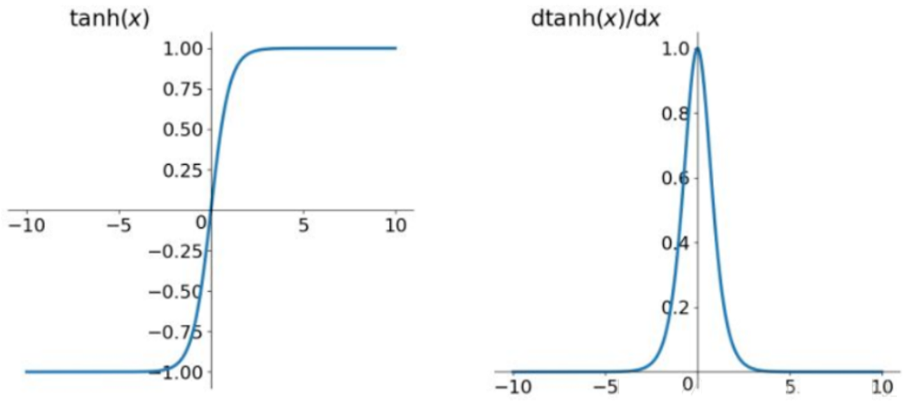


Figure 4-5: tanh(x) and derivative tanh(x)

4.3.3 Initialization

Before training, we need to initialize the weights in the network. A good initialization can significantly reduce the convergence time and find the global lowest point faster. On the contrary, an inappropriate method will take more time or even make it fall into a local lowest point.

Common initialization methods include random initialization, Xavier and He methods. Among them, random initialization can only ensure the running of the network, but may occur the problem of forward and backward propagation attenuation in deep networks.

Because we used tanh function in this thesis, so the Xavier[22] is the best method that fits it. The main idea of the Xavier method is to try to make the variance of the each hidden layer's input and output equal, by initialing weight within the range shown in equation 4.2, to optimize the transmission efficiency. In addition, it can reduce the possibility of gradient vanishing caused by the tanh function.

$$W \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}, \frac{\sqrt{6}}{\sqrt{n_j + n_{j+1}}}\right] \tag{4.2}$$

4.3.4 Batch normalization

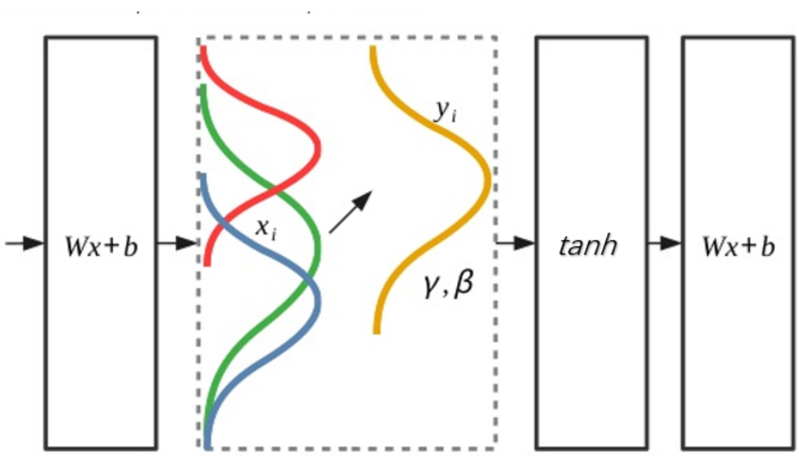


Figure 4-6: Batch normalization process

The idea of Batch normalization[23] is to normalize the output of each layer make it follows the $X \sim N(0,1)$ distribution. At the same time, two parameters are introduced so that the network can adjust the final output by itself during training. Batch normalization can make the output value always within the sensitive part of the activation function, as shown in Figure 1; which can improve the update efficiency and reduce the possibility of gradient vanishing.

Algorithm 4.1: Batch Normalization

Input: $\{x_1, \cdots, x_m\}$

Output: $\{y_1, \cdots, y_m\}$

1 Calculate batch mean: $\mu_B \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$

2 Calculate batch variance: $\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{\mathrm{m}}\sum_{\mathrm{i=1}}^{\mathrm{m}} (\mathrm{x_i} - \mu_{\mathcal{B}})^2$

3 Normalization: $(\hat{x}_i) \leftarrow \frac{x_i - \mu_B}{\sqrt{\sigma_B^2 + \epsilon}}$

4 Pan and zoom: $y_i \leftarrow \gamma \hat{x}_i + \beta$

4.3.5 Optimizer

Neural networks usually use gradient descent to optimize the loss, and the commonly used ones are SGD (Stochastic gradient descent), Adagrad, Momentum, Adam.

The advantage of SGD is that it is simple and easy to implement. But it is inefficient because sometimes the direction of the gradient does not point to the lowest point, which causes repeated oscillations in the inefficient direction(figure 4-7) and easily falls into the local lowest point.

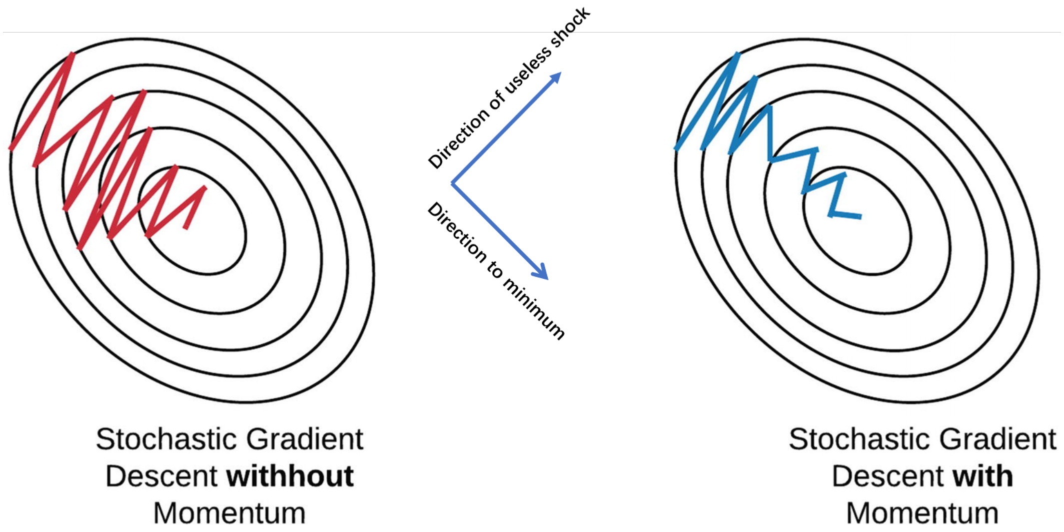


Figure 4-7: SGD vs Momentum

The improvement of Momentum is that it will accumulate the updated value. It slows down the gradient update in the direction of the oscillation, while the step size pointing to the lowest point keep increasing, thereby speeding up the convergence. Simultaneously, the larger step size will also increase its ability of leaving the local lowest point.

By combining SGD and Momentum, the Adam method has all their advantages. So we chose to use it in this thesis.

4.3.6 Training

Figure 4-8 is the final structure of the model.

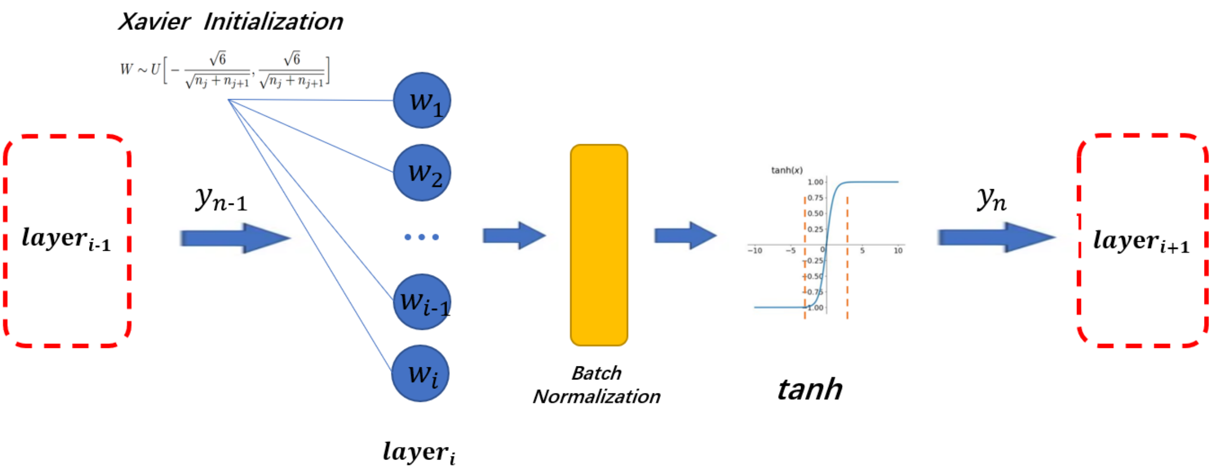


Figure 4-8: structure of anomaly detection model (1 layer)

Figure 4-9 shows the loss change during training.

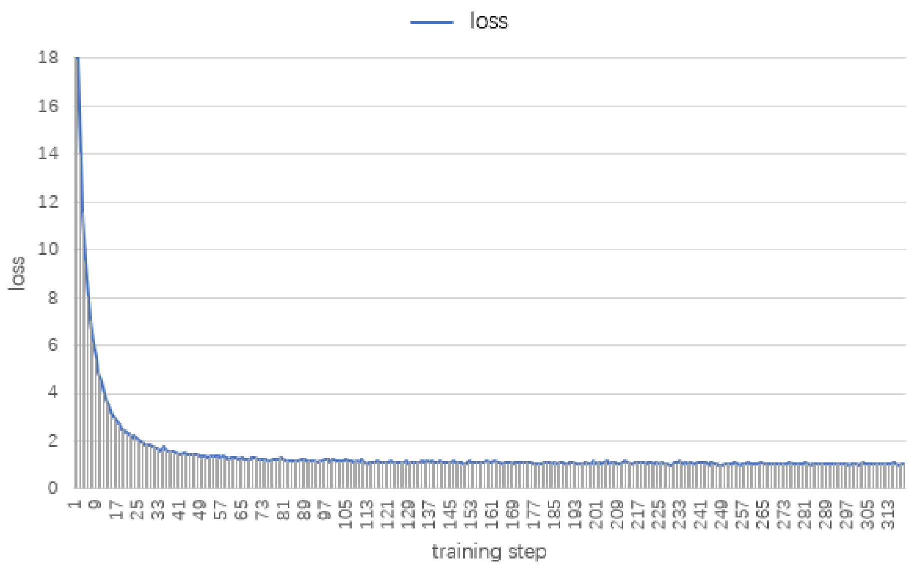


Figure 4-9: loss change during training (intercepted)

4.3.7 Abnormal judgment

For abnormal judgment, we use the loss value of each input data as its anomaly score. Then, randomly select a part of the data as input to the model and analyze the results. The output anomaly score distribution as shown in figure 4-10.

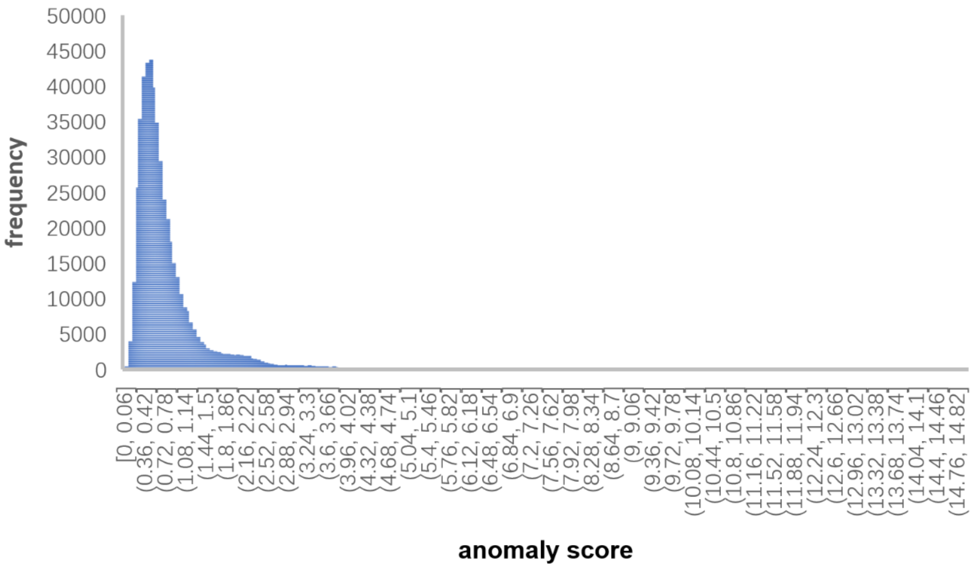


Figure 4-10: distribution of anomaly score

It can be seen that the distribution of anomaly scores is similar to the Gaussian distribution, most of which are concentrated in the range $[0, 5]$, which means that a small number of data with anomaly scores higher than this 5 has a relatively higher probability of anomaly. Therefore, we set the threshold to 5 according to the distribution, that is, the data with an anomaly score higher than 5 will be defined as abnormal data, and the opposite is normal data.

Chapter 5

Case study

In this chapter, we will evaluate the judgment performance of the model.We randomly sampled part of the data in the training set, about 450,000 records, and used it as input and analyzed the data with higher anomaly scores.

The results show that the model screened out 7288 abnormal data, the abnormality rate was about 1.6%. Among them, most of the abnormal data were continuous-time and belonged to the same generator, which is basically consistent with the fact that market power behaviors account for a small proportion and are mostly consistent operations. We analyzed some of the data based on the economic knowledge and experience introduced earlier, and randomly selected 2 cases for illustration.

5.1 Case 1

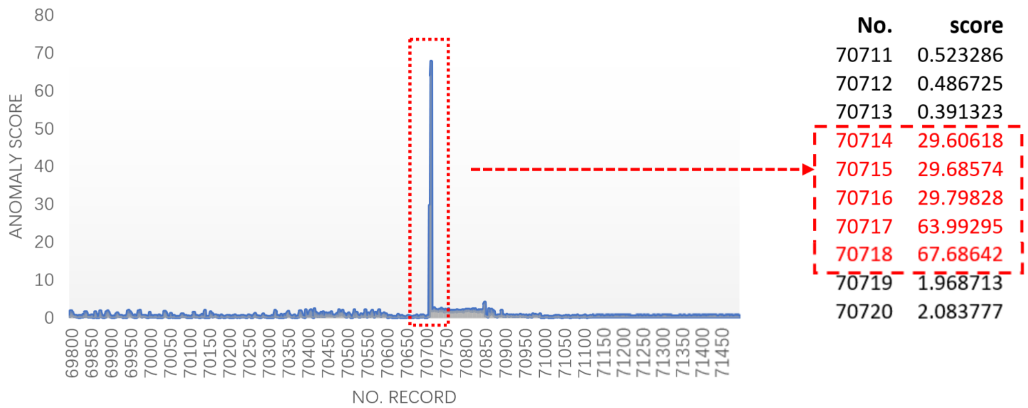


Figure 5-1: anomaly scores around case 1

As can be seen from the figure 5-1, when 70,714 transaction records are detected, the anomaly score is much larger than normal, which means that the model judges it to be an anomaly. According to the location of the anomaly, we checked its original data, which is concentrated on October 5, 2018. And its basic quotation information is shown in table 5-1

UP	OP	DATE	INTERVAL	CAPACITY	QUANTITY	BID PRICE	AWARDED PRICE
UP_TOR**_1	ED** SPA	20181005	18	530	10	75.33	77
UP_TOR**_1	ED** SPA	20181005	19	530	10	75.33	84.5
UP_TOR**_1	ED** SPA	20181005	20	530	10	75.33	93.31
UP_TOR**_1	ED**SPA	20181005	21	530	85	75.33	81.83
UP_TOR**_1	ED** SPA	20181005	22	530	85	75.33	76.2

Table 5-1: basic bidding information of case 1

It can be seen that the abnormal points are all from the same unit, which has a capacity of up to 530, which is a rare large-scale generator and belongs to the same operator. It is one of the top 3 largest companies and has considerable potential market power.

Secondly, its bid price is high, very close to the final clearing price. On the other hand, observing its bid quantity, it can be found that its reported electricity only accounts for less than 2% of its total production capacity, which is a very suspicious phenomenon.

On this basis, we inquired about other quotation information at the same time and found that there is another quotation order for this generator as shown in the table 5-2

UP	OP	DATE	INTERVAL	CAPACITY	QUANTITY	BID PRICE
UP_ TOR**_1	ED** SPA	20181005	18	530	290	0
UP_ TOR**_1	ED** SPA	20181005	19	530	290	0
UP_ TOR**_1	ED** SPA	20181005	20	530	290	0
UP_ TOR**_1	ED** SPA	20181005	21	530	215	0
UP_ TOR**_1	ED** SPA	20181005	22	530	215	0

Table 5-2: other bids of UP_TOR** at the same time

Obviously, it can be seen that a large amount of its production capacity is released in the market with 0 quotations, which means that this part of the electricity will definitely win the bid and be used to fill part of the market demand. And the small amount of electricity in table 5-1 will be used to raise the electricity price before the clear point deliberately. Shown as figure 5-2

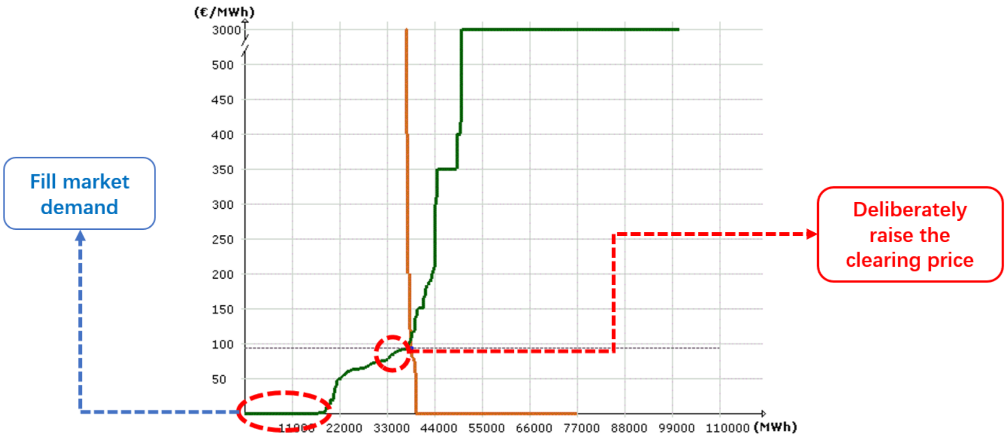


Figure 5-2: clearing process in 5/10/2018

Besides, we have conducted a more in-depth investigation over the same operators' overall quotation information during this period. The results are shown in table 5-3, its quotations below 10 euros exceed 60% of the total capacity, and the power quantity whose price below 60 euros is close to 75% . The rest is all invested around the clearing point to drive up electricity prices.

<10 euro	60%
<60 euro	75%
around clear price	25%

Table 5-3: overall bid of ED** SPA

It can be seen that not only the generator is likely to exercise market power, it is also a collective behavior of the entire group.

5.2 Case 2

Similarly, the model detected an abnormality score higher than usual at position 116805, as shown in the figure 5-3

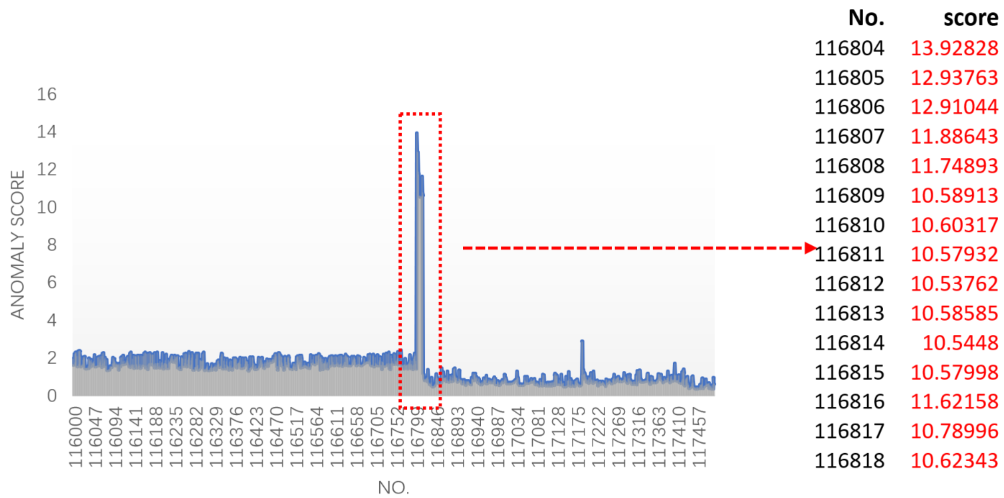


Figure 5-3: anomaly scores around case 2

According to the location of the abnormal point, we checked the original data as shown in table 5-4.

UP	OP	DATE	INTERVAL	CAPACITY	QUANTITY	BID PRICE	AWARDED PRICE
UP_CI**_1	DOL**	20181122	18	180	20	99	141.25
UP_CI**_1	DOL**	20181122	19	180	20	99	140.02
UP_CI**_1	DOL**	20181122	20	180	20	99	101.99
UP_CI**_1	DOL**	20181123	10	180	20	99	99
UP_CI**_1	DOL**	20181123	11	180	10	90	96.07

Table 5-4: basic bidding information of case 2

From the data, it can be seen that the generator's capacity is large, and the operator it belongs is also a large operator in the market, both have high potential market power.In terms of quotations and reported volume, it is also similar to case 1; that is, the submitted power quantity is far less than its production capacity. It shows that there is a suspicion that the segment quotation deliberately pulled the clear price. This has also been confirmed in other quotations in the same period. As shown in table 5-5.

UP	OP	DATE	INTERVAL	CAPACITY	QUANTITY	BID PRICE	AWARDED PRICE
UP_CI**_1	DOL**	20181122	18	180	160	0	141.25
UP_CI**_1	DOL**	20181122	19	180	169	0	140.02
UP_CI**_1	DOL**	20181122	20	180	169	0	101.99
UP_CI**_1	DOL**	20181123	10	180	160	0	99
UP_CI**_1	DOL**	20181123	11	180	140	0	96.07

Table 5-5: other bids of UP_CI**_1 at the same time

Similar to case 1, almost all of its remaining capacity is quoted at 0 euro to fill the market demand, and a small amount of high-quoted electric energy is used to raise the electricity price, which is a very obvious market power behavior. In addition, according to the conjectured cost curve of the UP previously obtained by big data technology, its quotation seriously deviates from the conjectured cost , which is very suspicious.

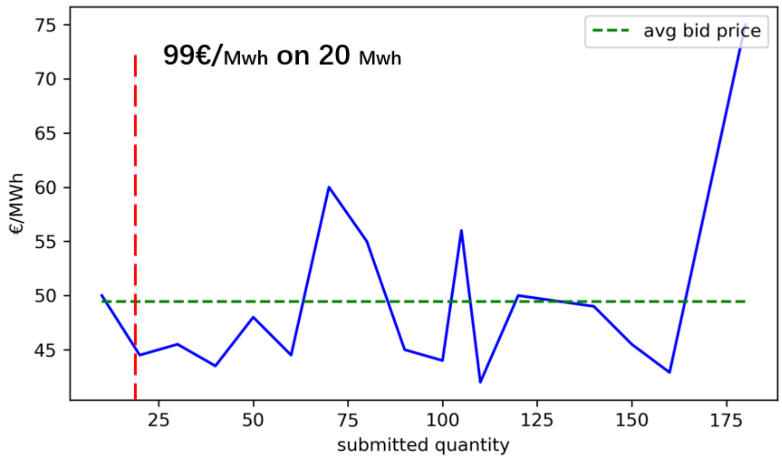


Figure 5-4: compare with conjectured cost curve of UP_C**_1

In addition, another major problem of the data here is that the clearing price was very high, about twice the average price of that month, so did the bid price of the generator. Among them, the fourth data successfully makes the unit become the marginal unit of the clearing process, which is the final price setter of the clearing price (quotation equals to clearing price). Accordingly, we inquired about the MGP clearing coupling diagram released by GME in 22/11/2018. As shown in figure 5-5.

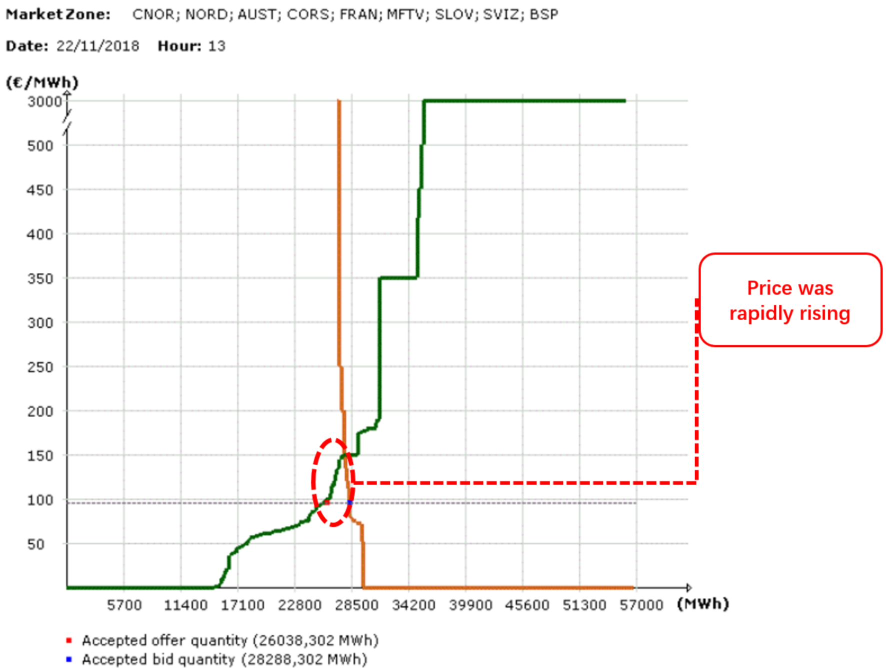


Figure 5-5: MGP clearing coupling diagram in 22/11/2018

It can be found that before the clearing, the price rises rapidly, and this phenomenon is often difficult to complete by a single operator, so it is most likely a collective behavior within the market. After checking the weekly report[24] released by GME at that time (figure 5-6), we found that the price of this week was much higher than the price of the same period of the previous week. This supports our previous hypothesis.

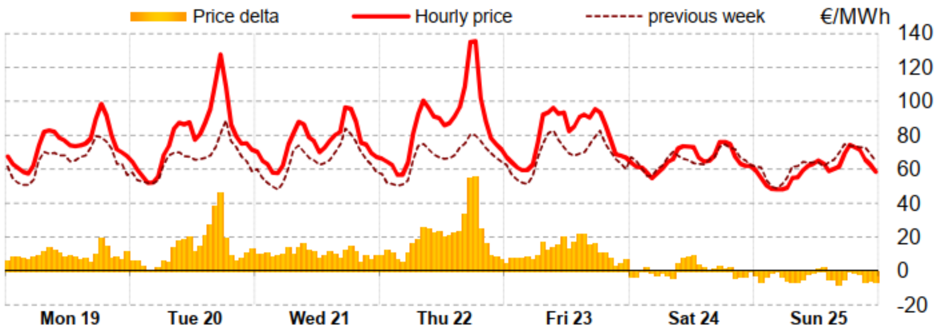


Figure 5-6: clearing prices on week no.46 and no.week 45 November 2018

Chapter 6

Conclusions, limitations, and prospects

6.1 Conclusions

For the analysis of the potential market power of the Italian electricity market, the indicators obtained through the big data framework all proved that the Italian electricity market still has medium-scale potential market power, including the imbalance of market sharing, the existence of monopolistic power manufacturers; the weak competitiveness of small-scale power plants and other issues.

In practice, the analysis and determination of market power are very complicated, and the lack of unified evaluation standards and huge data makes it more difficult. The current power market supervision system is relatively weak compared to other industries. If the power industry, as a pillar industry of the country and society, cannot be effectively supervised, unfair competition and monopolistic behavior in the market will increase the burden on residents and significantly affect economic development and social welfare.

For the anomaly detection model based on the variational auto-encoder, the model is good at filtering massive data and initially locate the abnormal points, so the regulator can conduct targeted investigations, which significantly saves time and greatly improves efficiency. Although there are still many factors (environment, physical obstruction, etc.) that have not been taken into consideration and the final determination requires human participation, it still has the ability to detect market anomalies initially. From the examples introduced in chapter 5, we can find that it not only can discover the individual market power behaviors of generators but also help to mine the collective behavior or market abnormal state related to it.

6.2 Limitations

The first problem is about data. Limited by trade secrets and the high confidentiality of the power industry, many important data are not available. The lack of important data is the main problem of the current model.

In addition, due to the lack of uniform judgment standards in the industry, the final judgment of abnormal behavior still requires human participation. Therefore, it is very difficult to conduct large-scale analysis and accurately evaluate model performance.

6.3 Prospects

First of all, a unified standard for determining market power is very important. If a set of practical unified standards can be developed in the power market field, it will greatly reduce the restrictions on model selection, data screening, and result verification. On this basis, supervised learning or large-scale semi-supervised learning can be carried out, which can be believed to improve the model’s performance significantly.

On the other hand, for the power industry, as a pillar industry of the state, should also regulate the supervision system of the electricity market at the national level in laws, regulations and policies, so as to achieve the synchronization and balance between market liberalization and supervision Strengthen. To really give full play to the advantages of the power industry, reduce the people's additional expenditure, and help economic development.

References

[1] G. L. Lapini, “The power station in via santa radegonda.” [Online]. Available: http://www.storiadimilano.it/citta/milanotecnica/elettricita/radegonda0.htm

[2] “Geothermal explained: Geothermal power plants.” [Online]. Available: https://www.eia.gov/energyexplained/geothermal/geothermal-power-plants.php#:~:text=The%20first%20geothermal%20power%20plant,steam%20to%20drive%20generator%20turbines.

[3] “Gme: results and statistics.” [Online]. Available: https://www.mercatoelettrico.org/En/Esiti/MGP/EsitiMGP.aspx

[4] Heyne, Paul, Boettke, P. J, Prychitko, and D. L, *The Economic Way of Thinking(13th ed.).* Pearson, 2014.

[5] A. T. Tyler Cowen, *Modern Principles: Microeconomics.* Worth Publishers, 2009.

[6] M. van der Woude, “Unfair and excessive prices in the energy sector,” *European Review of Energy Markets*, vol. 2, no. 3, 2008.

[7] C. Snijders, U. Matzat, and U. Reips, “”big data” : big gaps of knowledge in the field of internet science,” *International Journal of Internet Science*, vol. 7, no. 1, pp. 1–5, 2012.

[8] R. Kitchin and G. McArdle, “What makes big data, big data? exploring the ontological characteristics of 26 datasets,” *Big Data and Society*, vol. 3, no. 1, pp. 1–10, 2016.

[9] A. Woodie, “Why hadoop on ibm power.” [Online]. Available: https://www.datanami.com/2014/05/12/hadoop-ibm-power/

[10] M. Rouse, "Hadoop distributed file system (hdfs)." [Online]. Available: https://searchdatamanagement.techtarget.com/definition/Hadoop-Distributed-File-System-HDFS

[11] "Mapreduce tutorial." [Online]. Available: https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

[12] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," *NSDI'12: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, p. 2, 2012.

[13] "What is apache spark sql." [Online]. Available: https://databricks.com/glossary/what-is-spark-sql

[14] C. M. Bishop, *Pattern Recognition and Machine Learning.* New York: Springer-Verlag, 2006.

[15] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* Cambridge, Massachusetts ; London, England: The MIT Press, 2014, 2015.

[16] V. K. Varun Chandola, Arindam Banerjee, "Anomaly detection: A survey," *ACM Computing Surveys*, vol. 41, no. 3, 2009.

[17] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis - MLSDA*, no. 14, 2014.

[18] M. W. Diederik P Kingma, "Auto-encoding variational bayes." [Online]. Available: https://arxiv.org/pdf/1312.6114.pdf?source=post_page

[19] C. DOERSCH, "Tutorial on variational autoencoders." [Online]. Available: https://arxiv.org/pdf/1606.05908.pdf%20http://arxiv.org/abs/1606.05908.pdf

[20] Cowen, Tyler, Tabarrok, and Alex, *Information Theory and Statistics.* Courier Corporation, 1959.

[21] R. A. L. S. Kullback, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 2007.

[22] Y. B. Xavier Glorot, "Understanding the difficulty of training deep feedforward neural networks," *Journal of Machine Learning Research*, 2010.

[23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv*, 2015.

[24] GME, "The report of : Week no. 47 from 19 to 25 november 2018." [Online]. Available: https://www.mercatoelettrico.org/En/MenuBiblioteca/Documenti/20181126MGPreportsettimanale47En.pdf

[25] C. S. S. Ioffe, "Understanding the difficulty of training deep feedforward neural networks," *ArXiv*, vol. abs/1502.03167, 2015.

[26] T. Pham, "Market power in power markets in europe : the cases in french and german woholesale electricity markets. economics and finance," *Université Paris Dauphine - Paris IX*, 2015.

[27] E. Union, "Treaty on the functioning of the european union (tfeu)." [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A12012E%2FTXT