International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 4 Volume No.7 July - August – 2023
DOI:10.46647/ijetms.2023.v07i04.081 ISSN: 2581-4621

The Importance of Data Mining & Predictive Analysis

Sreejit Ramakrishnan
Saintgits College of Applied Sciences, Pathamuttom, Kottayam, Kerala
https://orcid.org/0009-0004-6154-2724

ABSTRACT
Data mining is the process of analyzing enormous amounts of information and datasets, extracting (or “mining”) useful intelligence to help organizations solve problems, predict trends, mitigate risks, and find new opportunities. Data mining is like actual mining because, in both cases, the miners are sifting through mountains of material to find valuable resources and elements. Data mining also includes establishing relationships and finding patterns, anomalies, and correlations to tackle issues, creating actionable information in the process. Data mining is a wide-ranging and varied process that includes many different components, some of which are even confused for data mining itself.
Keywords— Knowledge Discovery in Data, or KDD, knowledge extraction, data pattern analysis, data archaeology, data dredging, information harvesting, business intelligence.

1. Introduction

Data Mining History:

Data warehousing, BI and analytics technologies began to emerge in the late 1980s and early 1990s, providing an increased ability to analyze the growing amounts of data that organizations were creating and collecting. The term data mining was in use by 1995, when the First International Conference on Knowledge Discovery and Data Mining was held in Montreal.

The event was sponsored by the Association for the Advancement of Artificial Intelligence, or AARI, which also held the conference annually for the next three years. Since 1999, the conference -- popularly known as KDD 2021 and so on -- has been organized primarily by SIGKDD, the special interest group on knowledge discovery and data mining within the Association for Computing Machinery.

A technical journal, Data Mining and Knowledge Discovery, published its first issue in 1997. Initially a quarterly, it's now published bimonthly and contains peer-reviewed articles on data mining and knowledge discovery theories, techniques and practices. Another publication, the American Journal of Data Mining and Knowledge Discovery, was launched in 2016.

For millennia, people have excavated places to find hidden mysteries. "Knowledge discovery in databases" refers to the act of shifting through data to uncover hidden relationships and forecast future trends. In the 1990s, the phrase "data mining" was invented. Data mining emerged from the convergence of three scientific disciplines: artificial intelligence, machine learning, and statistics.

Artificial intelligence is the human-like intelligence demonstrated by software and machines, machine learning is the term used to describe algorithms that can learn from data to create predictions, and statistics is the numerical study of data correlations.

Data mining takes advantage of big data's infinite possibilities and inexpensive processing power. Processing power and speed have grown significantly in the recent decade, allowing the globe to undertake rapid, easy, and automated data analysis.

Why is data mining important?

Data mining is a crucial component of successful analytics initiatives in organizations. The information it generates can be used in business intelligence (BI) and advanced analytics applications that involve analysis of historical data, as well as real-time analytics applications that examine streaming data as it's created or collected.

Effective data mining aids in various aspects of planning business strategies and managing operations. That includes customer-facing functions such as marketing, advertising, sales and customer support, plus manufacturing, supply chain management, finance and HR. Data mining supports fraud detection, risk management, cyber security planning and many other critical business use cases. It also plays an important role in healthcare, government, scientific research, mathematics, sports and more.

Data mining process: How does it work?

Data mining is typically done by data scientists and other skilled BI and analytics professionals. But it can also be performed by data-savvy business analysts, executives and workers who function as citizen data scientists in an organization.

Its core elements include machine learning and statistical analysis, along with data management tasks done to prepare data for analysis. The use of machine learning algorithms and artificial intelligence (AI) tools has automated more of the process and made it easier to mine massive data sets, such as customer databases, transaction records and log files from web servers, mobile apps and sensors.

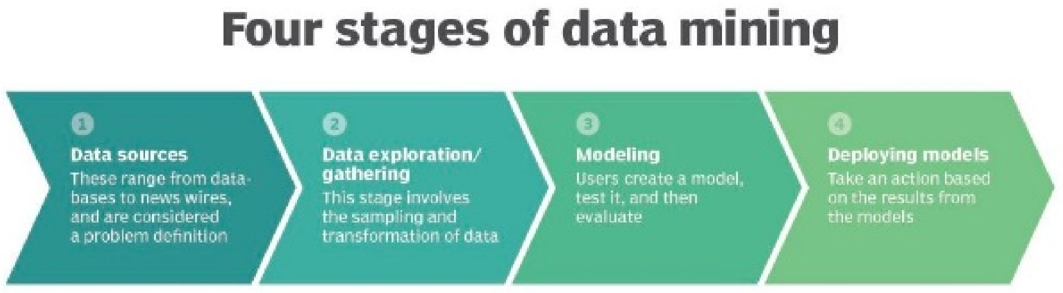
The data mining process can be broken down into these four primary stages:

1) **Data gathering**: Relevant data for an analytics application is identified and assembled. The data may be located in different source systems, a data warehouse or a data lake, an increasingly common repository in big data environments that contain a mix of structured and unstructured data. External data sources may also be used. Wherever the data comes from, a data scientist often moves it to a data lake for the remaining steps in the process.

2) **Data preparation**: This stage includes a set of steps to get the data ready to be mined. It starts with data exploration, profiling and pre-processing, followed by data cleansing work to fix errors and other data quality issues. Data transformation is also done to make data sets consistent, unless a data scientist is looking to analyze unfiltered raw data for a particular application.

3)**Mining the data**: Once the data is prepared, a data scientist chooses the appropriate data mining technique and then implements one or more algorithms to do the mining. In machine learning applications, the algorithms typically must be trained on sample data sets to look for the information being sought before they're run against the full set of data.

4) **Data analysis and interpretation**: The data mining results are used to create analytical models that can help drive decision-making and other business actions. The data scientist or another member of a data science team also must communicate the findings to business executives and users, often through data visualization and the use of data storytelling techniques.


Four stages of data mining
1
Data sources
These range from data-bases to news wires, and are considered a problem definition
2
Data exploration/ gathering
This stage involves the sampling and transformation of data
3
Modeling
Users create a model, test it, and then evaluate
4
Deploying models
Take an action based on the results from the models

Types of data mining techniques

Various techniques can be used to mine data for different data science applications. Pattern recognition is a common data mining use case that's enabled by multiple techniques, as is anomaly detection, which aims to identify outlier values in data sets. Popular data mining techniques include the following types:

1) **Association rule mining**: In data mining, association rules are if-then statements that identify relationships between data elements. Support and confidence criteria are used to assess the

relationships -- support measures how frequently the related elements appear in a data set, while confidence reflects the number of times an if-then statement is accurate.

2) **Classification**: This approach assigns the elements in data sets to different categories defined as part of the data mining process. Decision trees, Naive Bayes classifiers, k-nearest neighbor and logistic regression are some examples of classification methods.

3) **Clustering**: In this case, data elements that share particular characteristics are grouped together into clusters as part of data mining applications. Examples include k-means clustering, hierarchical clustering and Gaussian mixture models.

4)**Regression**: This is another way to find relationships in data sets, by calculating predicted data values based on a set of variables. Linear regression and multivariate regression are examples. Decision trees and some other classification methods can be used to do regressions, too.

5) **Sequence and path analysis**: Data can also be mined to look for patterns in which a particular set of events or values leads to later ones.

6) **Neural networks**: A neural network is a set of algorithms that simulates the activity of the human brain. Neural networks are particularly useful in complex pattern recognition applications involving deep learning, a more advanced offshoot of machine learning.

Data mining software and tools

Data mining tools are available from a large number of vendors, typically as part of software platforms that also include other types of data science and advanced analytics tools. Key features provided by data mining software include data preparation capabilities, built-in algorithms, predictive modeling support, a GUI-based development environment, and tools for deploying models and scoring how they perform.

Vendors that offer tools for data mining include Alteryx, AWS, Databricks, Dataiku, DataRobot, Google, H2O.ai, IBM, Knime, Microsoft, Oracle, RapidMiner, SAP, SAS Institute and Tibco Software, among others.

A variety of free open source technologies can also be used to mine data, including DataMelt, Elki, Orange, Rattle, scikit-learn and Weka. Some software vendors provide open source options, too. For example, Knime combines an open source analytics platform with commercial software for managing data science applications, while companies such as Dataiku and H2O.ai offer free versions of their tools.

Benefits of data mining

In general, the business benefits of data mining come from the increased ability to uncover hidden patterns, trends, correlations and anomalies in data sets. That information can be used to improve business decision-making and strategic planning through a combination of conventional data analysis and predictive analytics.

Specific data mining benefits include the following:

1. More effective marketing and sales: Data mining helps marketers better understand customer behavior and preferences, which enables them to create targeted marketing and advertising campaigns. Similarly, sales teams can use data mining results to improve lead conversion rates and sell additional products and services to existing customers.
2. Better customer service: Thanks to data mining, companies can identify potential customer service issues more promptly and give contact center agents up-to-date information to use in calls and online chats with customers.
3. Improved supply chain management: Organizations can spot market trends and forecast product demand more accurately, enabling them to better manage inventories of goods and supplies. Supply chain managers can also use information from data mining to optimize warehousing, distribution and other logistics operations.
4. Increased production uptime: Mining operational data from sensors on manufacturing machines and other industrial equipment supports predictive maintenance applications to identify potential problems before they occur, helping to avoid unscheduled downtime.

5. Stronger risk management: Risk managers and business executives can better assess financial, legal, cybersecurity and other risks to a company and develop plans for managing them.
6. Lower costs: Data mining helps drive cost savings through operational efficiencies in business processes and reduced redundancy and waste in corporate spending.

Ultimately, data mining initiatives can lead to higher revenue and profits, as well as competitive advantages that set companies apart from their business rivals.

Industry examples of data mining

Here's how organizations in some industries use data mining as part of analytics applications:

Retail: Online retailers mine customer data and internet clickstream records to help them target marketing campaigns, ads and promotional offers to individual shoppers. Data mining and predictive modeling also power the recommendation engines that suggest possible purchases to website visitors, as well as inventory and supply chain management activities.

Financial services: Banks and credit card companies use data mining tools to build financial risk models, detect fraudulent transactions and vet loan and credit applications. Data mining also plays a key role in marketing and in identifying potential upselling opportunities with existing customers.

Insurance: Insurers rely on data mining to aid in pricing insurance policies and deciding whether to approve policy applications, including risk modeling and management for prospective customers.

Manufacturing: Data mining applications for manufacturers include efforts to improve uptime and operational efficiency in production plants, supply chain performance and product safety.

Entertainment. Streaming services do data mining to analyze what users are watching or listening to and to make personalized recommendations based on people's viewing and listening habits.

Healthcare. Data mining helps doctors diagnose medical conditions, treat patients and analyze X-rays and other medical imaging results. Medical research also depends heavily on data mining, machine learning and other forms of analytics.

Predictive Analytics:

Predictive analytics is extracting information from large datasets to predict and estimate future outcomes.

Predict on top of data mining results by applying domain knowledge –

- What customer will buy next?
- What will be the customer churn rate?
- How many new subscriptions will be started if this offer is given?
- What is the amount of stock of a product needed for the coming month?

Apply business knowledge on data-mine patterns with any additional data needed to get business-valid predictions. Predictive analytics tries to find answers to the pattern by applying business knowledge and thus making it a more actionable piece of information. Business-specific knowledge and a clear business objective are a must here. Business analysts and other domain experts can analyze and interpret the patterns discovered by the machines, making useful meaning out of the data patterns and deriving actionable insights.

Predictive analytics is the process by which information is extracted from existing data sets for determining patterns and predicting the forthcoming trends or outcomes. It uses data, statistical algorithms, and machine learning techniques to identify the likelihood of future outcomes based on historical data. In other words, the aim of predictive analytics is to forecast what

As the name implies, it is an analytic process used to explore a large amount of data in regards to consistent patterns and systematic relationships between variables. Businesses prefer data mining because it aims to predict. Predictive analyses, on the other hand, refine data resources, in particular, to extract hidden value from those newly discovered patterns.

“Data mining + Domain knowledge => predictive analytics => Business Value”

Overall, predictive analysis and data mining, both make use of algorithms to discover knowledge and find the best possible solutions around.

Predictive analytics aims to identify the likelihood of future events based on historical data. By using data, mathematical algorithms and machine learning technology, predictive analytics has the potential to provide the best evaluation of what will happen. Moreover, it offers a perfect view of what's going on and what needs to be done to succeed.

As a result, predictive analytics can offer:

1) Valuable insight
2) Increase competitive edge
3) Predict trends
4) Identify new business opportunities in time

Make the Most of Data Mining and Predictive Analytics

Knowing what your customers are most likely to do or what they want or how much they are likely to spend to get it, are one of the best possible ways to hit your target audience. For example – think of Netflix binge recommending sci-fi shows, this is a pure example of predictive analytics results.

Furthermore, both the procedures data mining as well as predictive analytics deal with discovering secrets within big data but people often get confused with these methodologies. Data mining uses software to search for patterns, while predictive analytics uses those patterns to make predictions and direct decisions. So it is safe to say that data mining turns out to be a stepping stone for predictive analysis. Apart from this, data mining is passive while predictive analytics is active and has the potential to offer a clear picture.

Being a marketer or business owner, it is imperative for you to navigate the whole world of big data. No matter how intimidating the world of information seems, you need to keep embracing it at regular intervals.

CONCLUSION

Predictive analysis is a progressive branch of data engineering that usually does the prediction of any existence or probability of data. Predictive analytics makes use of data- mining methods for making predictions about the events in the future and then yields recommendations by these predictions. The procedure consists of a historic data analysis, and depending on that evaluation, the prediction of the future events is done.

Classification and regression hail to be the two chief goals of predictive analytics. It comprises different statistical and analytical methods that are employed for evolving the models which will do the prediction of future occurrence, events, or chances. Predictive analytics is capable of dealing with continuous and discontinuous changes. Classification, prediction and, to the particular extent, affinity analysis comprise the analytical techniques used in predictive analytics. The role taken by these predictive models differs based on the data which are used by them. In this survey paper, the list of reviews are provided and discussed, how researchers already used predictive analytics for business and medical industry and also mentioned the techniques and algorithms with the issues while applying on big data. Thus, a novel approach or model could be generated to predict making use of predictive analytics modelling methods suitable for big data.

References

1. Aneeshkumar, A.S. and Venkateswaran, C.J. (2012) 'Estimating the surveillance of liver disorder using classification algorithms', International Journal of Computer Applications, Vol. 57, pp.39–42.

2. Babu, P. and Sastry, S.H. (2014) 'Big data and predictive analytics in ERP systems for automating decision making process', 5th IEEE International Conference on Software Engineering and Service Science (ICSESS), 27–29 June, Beijing, China, pp.259–262.

3. Banjade, R and Maharjan, S (2011) 'Product recommendations using linear predictive modeling', Second Asian Himalayas International Conference on Internet (AH-ICI), Kathmandu, Nepal, 46 Nov, pp.1–4.

4. Bellaachia, A. and Guven, E. (2005) Predicting breast cancer survivability using data mining techniques, Department of Computer Science, the George Washington University, Washington.
5. Bhat, V.H., Rao, P.G., Krishna, S. and Shenoy, P.D. (2011) An Efficient Framework for Prediction in Healthcare, Springer-Verlag, Berlin Heidelberg, p.522–532.
6. Bhat, V.H., Rao, P.G., Shenoy, D., Venugopal, K.R. and Patnaik, L.M. (2009) 'An efficient prediction model for diabetic database using soft computing techniques', Proceedings of the 12th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Springer-Verlag, Berlin Heidelberg, p.328–335.
7. Chaitrali, S., Sulabha, D. and Apte, S. (2012) 'Improved study of heart disease prediction system using data mining classification techniques', International Journal of Computer Applications, Vol. 47, No. 10, pp.44–48.
8. Chandra Shekar, K., Ravi Kanth, K. and SreeKanth. K. (2012) 'Improved algorithm for prediction of heart disease using case based reasoning technique on non-binary datasets', International Journal of Research in Computer and Communication Technology, Vol. 1, No. 7, pp.420–424.
9. Chiang, H-J., Tseng, C-C. and Torng, C-C. (2013) 'A retrospective analysis of prognostic indicators in dental implant therapy using the C5.0 decision tree algorithm', Journal of Dental Sciences, Vol. 8, No. 3, pp.248–255.
10. Chinchor, N., Thomas, J. and Wong, P. (2010) 'Multimedia Analysis + Visual Analytics = Multimedia Analytics', IEEE Computer Graphics, Vol. 30, No. 5, pp.52–60.
11. Hall, L., Chawla, N. and Bowyer, K. (1998) 'Decision tree learning on very large data sets', International Conference on Systems, Man and Cybernetics, San Diego, CA, USA, 14 Oct, pp.2579–2584.
12. Huang, Z., Wong, P.C., Mackey, P., Chen, Y., Ma, J., Schneider, K. and Greitzer, L. (2009) 'Managing complex network operation with predictive analytics', Proceedings of the AAAI Spring Symposium on Techno social Predictive Analytics, California, USA, 23–25 March 2009, pp.59–65.
13. Huang, C.H., Yang, K.C. and Kao, H.Y. (2014) 'Analyzing big data with the hybrid interval regression methods', The Scientific World Journal, Vol. 2014, pp.1–9.
14. Jakrarin, T. and Piromsopa, K. (2013) 'An analysis of suitable parameters for efficiently applying K-means clustering to large TCP dump data set using Hadoop framework', Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 10th International Conference, Krabi, Thailand, 15–17 May, pp.1–6.
15. Jun, S., Lee, S.J. and Ryu, J.B. (2015) 'A divided regression analysis for big data', International Journal of Software Engineering and its Applications, Vol. 9, No. 5, pp.21–32.