IJCSMA
Rohan Kumar Mishra *et al*, International Journal of Computer Science and Mobile Applications,
Vol.10 Issue. 9, September- 2022, pg. 13-19 ISSN: 2321-8363
Impact Factor: 6.308
(An Open Accessible, Fully Refereed and Peer Reviewed Journal)

A Predictive Approach of Data Science to Transform Unstructured Web Data

Rohan Kumar Mishra[1]; Komal Sarraf[2]; Anirban Bhar[3]; Moumita Ghosh[4]
[1,2]B. Tech student, Department of Information Technology, Narula Institute of Technology, Kolkata, India
[3,4]Assistant Professor, Department of Information Technology, Narula Institute of Technology, Kolkata, India
[1] rohanmisraa2002@gmail.com
[2] komalsarraf19@gmail.com
[3] anirban.bhar@nit.ac.in
[4] moumita.ghosh@nit.ac.in

DOI: 10.5281/zenodo.7088931

Abstract

In the Fourth Industrial Revolution era, there is a wealth of data available in the digital world, including internet of things (IoT) data, corporate data, health data, mobile data, urban data, security data, and many more. Making wise decisions can be achieved in many application sectors by gaining knowledge or insightful information from these facts. In the field of data science, advanced analytics techniques like machine learning modelling can offer more in-depth understanding of the data or actionable insights, which makes computing more intelligent and autonomous.

It has always been a focus of study to apply predictive analytics to structured time-series data. Researchers have begun merging pertinent structured and unstructured data due to the abundance of textual content being generated across various sources on the web.

An active field of research has long been predictive analytics over structured time-series data. Due to the abundance of textual data being produced by many online sources, researchers have begun merging pertinent structured and unstructured data to enhance predictions. In this article, we provide a data science paradigm for predictive analytics that makes use of unstructured data.

We will examine data science and its relationship to artificial intelligence, machine learning, and deep learning in this essay. In this research, we attempted to exhibit the data science operations like data cleaning, data processing, data modelling, data visualisation, and data presenting approaches. The inclusion of these intellectual sciences in data science is valuable for perming many operations. Knowing your customers' demands and exceeding their future expectations through wise decision-making are essential for any firm looking to expand. The analytical algorithms or data operations used in data science make the data more useful for making decisions and enforcing decisions. We also emphasise the integration of mathematical and statistical techniques, logical thinking, and applications of artificial intelligence techniques in data science.

Keywords: Data Science, Machine Learning, Deep Learning, Predictive Analysis, Decision Making.

1. Introduction

Data science is the activity of analysing vast amounts of unstructured and organised raw data to find patterns and draw conclusions that can be put to use. The foundations of data science, an interdisciplinary topic, include statistics, inference, computer science, the development of machine learning algorithms, and new methods to extract knowledge from huge data.

Numerous methods and resources are employed in predictive data analytics. It makes the most precise future predictions using data, algorithms, and more recently, machine learning approaches.

Statisticians employed Predictive Analytics in the form of decision trees and linear/logistic regression to corroborate, classify, and predict business data. Predictive Data Analytics, however, has gained popularity for two reasons: easy access to technology that can gather and analyse enormous volumes of data, as well as technologies like Machine Learning, a branch of Artificial Intelligence.

Because data can forecast the future, marketers are already using it to identify new opportunities and evaluate future customer behaviour in addition to past trends.

Predictive analytics is being used by businesses more and more to learn how to interact with their customers more effectively by gathering data from the enormous volumes of data at their disposal, forecasting behaviour patterns, and spotting emerging trends.

To produce predictions, predictive data analytics use a range of approaches and tools shown in Figure 1.

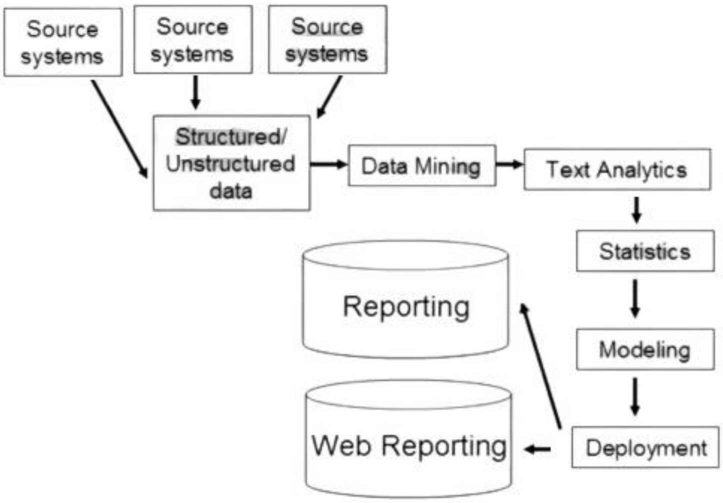

Figure 1: Predictive data analytics

To be clear, predictive data analytics has been around for a while, but it has only lately started to be used by a lot more companies. These are the explanations for this:

- Modern computers are quicker, less expensive, and simpler to use, all of which make the deployment of predictive data analytics simpler.
- Discovering new customer behaviour trends and business expansion opportunities is another benefit of predictive data analytics. It can be used by marketers to learn more about the market and uncover emerging trends, enabling them to model their goods or services to satisfy the wants of their target market.

Data analysts can construct predictive data models if they have gathered enough data. Using predictive analytics, a prediction score can be given to each customer. A prediction model that has been trained using your data is used to do this. Predictive modelling is a technique for forecasting outcomes using data models by fusing statistics and data.

The goal of predictive modelling is to predict the outcome of future occurrences by applying algorithms to data obtained from previous episodes. In a business model, this is most usually explained as the analysis of past sales data to forecast future sales results, followed by the application of those forecasts to determine which marketing initiatives should be made.

2. Related Work

Predictive analysis in business has primarily focused on structured data up until very recently. Numerous studies are looking into how consumer-generated unstructured text, such as complaints, service logs, social media data, etc., may impact business outcomes as more firms are embracing this approach. The influence of internet reviews on the purchase of goods and services have been a subject of much investigation. Effects of online evaluations on a few industries, including fashion and film, have been thoroughly researched. Few of these works offer mechanisms for adding the elements in a predictive model, even though the majority of these works examine the impact of the reviews.

The writers of [1] have presented a thorough overview of the numerous forecasting methodologies used to forecast the fashion sector. Numerous elements, such as shifting weather patterns, facilities for production across continents, holidays, public events, and economic conditions, among others, have an impact on the fashion market. In addition to these factors, the absence of historical data for novel types of fashion goods prompted analysts working in this field to explore especially for techniques that might include unstructured data into the predictive process in order to produce more accurate predictions.

Given that [2] and [3] have demonstrated that alterations to the statistical models based on the prior experience of experts lead to more accurate forecasts, one stream of study in this field concentrates on the integration of expert opinion and integrating it with statistical predictions.

A different line of research, however, focuses on the application of machine learning-based models that may easily incorporate unusual features like the ones mentioned above. It was discovered that for this domain, machine learning-based models outperformed conventional regression techniques in terms of results.

Extreme learning methods (ELM) were suggested by Sun et al. [4] as a tool for estimating sales at the item level. Thomassey and Happiette suggest using fuzzy inference systems and neural networks, among other soft computing techniques, to forecast sales in [5]. Teucke et al. [6] suggested combining decision trees with support vector machines for the actual forecasts in order to get more precise results. The decision trees would be used to assess which articles were likely to be reordered. The issue to be highlighted is that the majority of these works have only used data from the retail business to develop their models, even if numerous other variations of these models have been provided by other academics, as given in [1].

According to Yu et analysis. is of a sizable number of online movie reviews in [8], both the sentiments conveyed in the reviews and the quality of the reviews have a considerable impact on the performance of the film's future sales. Using probabilistic latent semantic analysis, sentiments are discovered (PLSA). They provide an autoregressive sentiment-aware model for sales prediction based on the sentiments. The quality of a review is taken into account in this approach, which enhances it further.

News for predicting stock market data is one area where text inputs have been used for prediction extensively. To examine correlations between stock prices and public emotion in response to social events and news, researchers in a variety of studies have examined texts from blogs, social network services (SNS), and the news [9–13]. In [14], Luss and Aspremont demonstrate how support vector machines can be used to predict intraday price fluctuations of financial instruments using data derived from news articles. Verma et al. provided findings in [15] to demonstrate that stock trends can be better anticipated by taking news events as well as actual stock values into consideration.

From the discussion above, it is clear that while machine-learning-based models are becoming more popular among business analysts, the use of unstructured data for prediction has not yet taken off. We can also see that deep-learning based models have not attracted the attention of many researchers in this sector, despite the fact that they have the ability to handle massive volumes of data. We provide a deep-learning based model that can do this in the parts that follow.

3. Methodology and Approach of Data Science

The Data Science Methodology is encountered by those who work in data science and are constantly seeking solutions to new problems. The process for locating answers to a particular problem is described by data science methodology. This is a cycle that experiences critic behaviour, which directs business analysts and data scientists to take the appropriate action. Figure 2 describes the method.

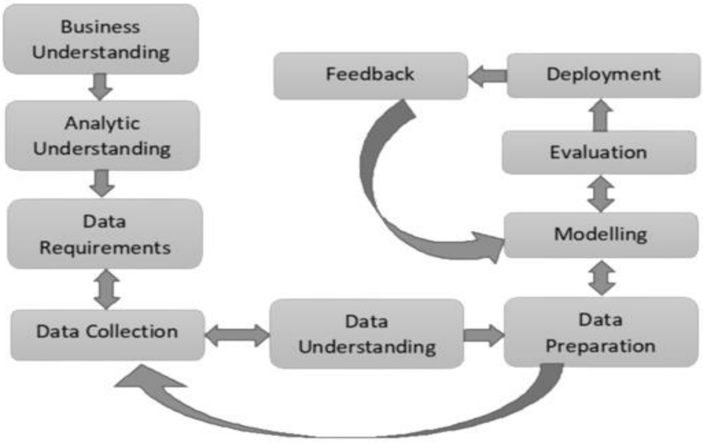

Figure 2: Data science approach

3.1 Business Understanding

Any problem in the business area needs to be thoroughly understood before it can be solved. A solid foundation created by business knowledge makes it easier to answer questions. We need to be clear about the precise issue we intend to address.

3.2 Analytic Understanding

One should choose the analytical strategy to use based on the business understanding discussed above. There are four different kinds of approaches: descriptive (current state and information presented), diagnostic (also known as statistical analysis, what is happening and why it is happening), predictive (it predicts trends or the likelihood of future events), and prescriptive (how the problem should be solved actually).

3.3 Data Requirements

The above-mentioned analytical approach identifies the pertinent data's required content, formats, and sources. Finding the answers to the following questions during the data needs process is necessary: "What," "Where," "When," "Why," "How," and "Who."

3.4 Data Collection

Any random format can be used to acquire the collected data. Therefore, the data gathered should be checked in accordance with the methodology used and the results expected. Therefore, if further information is needed, it can be collected or it can be discarded.

3.5 Data Understanding

The question "Is the data obtained representative of the problem to be solved?" is answered through data understanding. The measurements that are applied to the data in descriptive statistics are calculated in order to access the matter's quality and content. This step could result in going back to the previous stage to make adjustments.

3.6 Data Preparation

Let's connect this idea with two analogies to better grasp it. Washing just-picked veggies and only picking what you want from the buffet to put on your plate are the first two things to remember. Vegetable washing represents the elimination of impurities, or undesired items, from the data. Noise cancellation is done here. If we are only considering edible items on the plate and we don't require particular information, we shouldn't proceed with the process. Included in this entire process are transformation, normalisation, etc.

3.7 Modelling

Modelling determines whether the data that has been prepared for processing is suitable or needs extra seasoning and finishing. The development of predictive and descriptive models is the main goal of this stage.

3.8 Evaluation

Model development includes model evaluation. It examines the model's quality and determines whether it satisfies the business needs. It goes through a diagnostic measure phase (which determines whether the model functions as planned and where adjustments are needed) and a statistical significance testing phase (ensures about proper data handling and interpretation).

3.9 Deployment

The model is prepared for implementation in the business market as it is successfully evaluated. The deployment phase determines how well the model performs in comparison to competitors and how much external stress it can withstand.

3.10 Feedback

In order to improve the model and assess its performance and impact, feedback is a crucial goal. The steps in providing feedback include defining the review procedure, maintaining a record, assessing effectiveness, and reviewing with improvement.

After these ten steps have been successfully completed, the model shouldn't be left untreated; instead, an update should be performed based on user feedback and deployment. To ensure that the model continues to add value to the solutions, new trends should be assessed as new technologies are developed.

4. Evolution of Data Science and Data Analytics

The Statistical Analysis System (SAS), which served as the basis for a study at North Carolina State University's (NCSU) agricultural department, helped usher in the digital era of data science. The primary goal of data science when it was first introduced to the industrial level was to discover more precise and trustworthy solutions than those attained via business analysis. An analyst needs to possess the sought-after skill sets of data processing, predictive modelling, and visualisation to operate in the data science industry. Currently, the two most popular technologies used for data processing are Python and R [16].

But in the future, processing and analysing data might be done using Google's Go programming language. Because there are so many tools, technologies, and resources available, data science is developing at an exponential rate. It offers upbeat solutions to numerous actual word problems [17] [18] [19].

4.1 Machine Learning in Data science

Since the early 1950s, there has been machine learning technology. Machine learning is a result of a data-driven method that was developed in 1990. There was a change in emphasis toward information retrieval and natural language search from 1995 to 2005. The neural network, which was first used in 1957 for the first neural network computers, made a resurgence in 2005. Machine learning is one of those technologies that has both successful and unsuccessful applications, but there is a chance that it may become widely used in the near future. There should be increase in some of the aspects affecting the machine learning business, such as infrastructure and technical competence, to maintain its growth.

4.2 Deep Learning in Data Science

Alexey Grigoryevich Ivakhnenko and Valentin Grigor' evich Lapa introduced deep learning in 1965 [19]. They employed a few models with complex equations and polynomial functions that were statistically examined. A technique for identifying and mapping related or similar data was created in 1995. For recurrent neural networks, long short-term memory is established in 1997. With the advent of processors with quick computing speeds in the late 1990s, the efficiency of GPUs for image processing increased, eventually multiplying computation speed by a factor of about 1000. In the early 2000s, different pre-training layers were employed to enhance long short-term memory, and by 2011, as GPU speeds increased, computers could work with convolutional neural networks without needing layer-by-layer pre-training. Deep Learning is currently needed to process Big Data. Deep learning and AI are currently expanding, and more cutting-edge concepts are emerging.

4.3 Artificial Intelligence in Data Science

On a Ferranti mark 1 system, AI programmes are created and executed in 1951. The first workshop on artificial intelligence was held in 1956 at Dartmouth College [20]. The lack of computer hardware resources for computations was an issue back then. After the government insisted, business and the government contributed billions of dollars to the development of AI during the 1980s. After funds and interest were provided to advance the subject of artificial intelligence, it saw a boom from 2000 to 2010. Machine learning was an effective solution to a variety of issues in business and society after the creation of robust computer technology.

5. Conclusion

Because there were numerous combinations of viable solutions during the late 1980s and early 1990s, the algorithm employed to solve complex reasoning issues was insufficient. As the problems got more and bigger, this caused the computing rates to drop down dramatically. The notions of probability and economics were the remedies created by AI researchers to deal with insufficient or ambiguous information.

In the modern world, artificial intelligence is becoming more prevalent and we are making progress in that direction. It's also critical to recognise the technology' versatility in minimising hazards while maintaining safety. Analytics prediction is currently being done using machine learning techniques. When it comes to providing reliable results, deep learning is thought to offer more advantages than traditional machine learning methods. The intricacy and difficulties involved with this field grow as we advance in it.

References

[1] Samaneh Beheshti-Kashi, Hamid Reza Karimi, Klaus-Dieter Thoben, Michael Lütjen & Michael Teucke: A survey on retail sales forecasting and prediction in fashion markets, Systems Science & Control Engineering, Dec. 2014, Pages 154-161.

[2] Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to SKU-level demand forecasts. International Journal of Forecasting, 29(3), 510–522.

[3] Franses, P. H., & Legerstee, R. (2013). Do statistical forecasting models for SKU-level data benefit from including past expert knowledge? International Journal of Forecasting, 29(1), 80–87.

[4] Sun, Z.-L., Choi, T. M., Au, K-F., & Yu, Y.(2008). Sales forecasting using extreme learning machine with applications in fashion retailing. Decision Support Systems, 46(1), 411–419.

[5] Thomassey, S. (2010). Sales forecasts in clothing industry: The key success factor of the supply chain management. International Journal of Production Economics, 128(2), 470–483.

[6] Teucke, M., et al. (2014). Forecasting of seasonal apparel products. In H. Kotzab, J. Pannek, & K.-D. Thoben (eds.), Dynamics in Logistics. Fourth International Conference, LDIC 2014 Bremen, Germany, February 2014 Proceedings. Springer

[7] Hinton, G. E.; Osindero, S.; Teh, Y. W. (2006). "A Fast Learning Algorithm for Deep Belief Nets" (PDF). Neural Computation. 18 (7): 1527–1554.

[8] Xiaohui Yu, Yang Liu, Xiangji Huang and Aijun An: Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain, IEEE Transactions on Knowledge and Data Engineering, Volume: 24, Issue: 4, April 2012

[9] J. R. Nofsinger, "Social Mood and Financial Economics," J. Behav. Finance, vol. 6, no. 3, pp. 144–160, Sep. 2005.

[10] W. S. Chan, "Stock price reaction to news and no-news: drift and reversal after headlines," J. Financ. Econ., vol. 70, no. 2, pp. 223–260, Nov. 2003.

[11] N. Strauß, R. Vliegenthart, and P. Verhoeven, "Lagging behind? Emotions in newspaper articles and stock market prices in the Netherlands," Public Relat. Rev.

[12] G. Ranco, D. Aleksovski, G. Caldarelli, M. Grčar, and I. Mozetič, "The Effects of Twitter Sentiment on Stock Price Returns," PLOS ONE, vol. 10, no. 9, p. e0138441, Sep. 2015.

[13] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," J. Comput. Sci., vol. 2, no. 1, pp. 1–8, Mar. 2011.

[14] Luss, R., & d'Aspremont, A. (2015). Predicting abnormal returns from news using text classification. Quantitative Finance, 15(6), 999-1012.

[15] Ishan Verma, Lipika Dey, and Hardik Meisheri. 2017. Detecting, quantifying and accessing impact of news events on Indian stock indices. In Proceedings of the International Conference on Web Intelligence (WI '17). ACM, New York, NY, USA, 550-557.

[16] Data technology Landscape and evolution of data science. https://www.digitalvidya.com.

[17] Evolution of data analytics: then, now and later-Affineblog.

[18] The evolution of machine learning synectics for management decisions. http://www.smdi.com.

[19] A brief history of deep learning- Data versity.

[20] J Vincent, "Nvidia launches AI computer to give autonomous robots better brains" the verge, 2018.