Detection and ~~Overcome the~~ Challenges in Predictive Analytics in Data Mining using Machine Learning

By

Group 99

SAIKUMAR REDDY RAJU

NELANUTHALA KARTHIK

YELLURU VENKATA KRISHNA

BOMMINENI SHRAVAN REDDY

AVINASH GANISETTI

# ABSTRACT

Predictive Analytics are the key application of data mining algorithm. It can be used for getting insight into the facts of daat and to predict future incidents. In this context, the data can be collected and analysed using the algorithms of data mining. However, there are certain limitations of predictive analytics that can be recovered by the application of machine learning with the help of statistical computation. In this paper, the research will be conducted to validate the fact and by collecting the research papers by using the search string., the final paper shave been selected for the review using the inclusion and exclusion criteria and exclusion criteria. Those selected and includes papers have been reviewed and the result has been discussed.

# 1. Introduction

Prediction is one of the key issues in today’s technology to transform the systematic approach to be automated. Nowadays, automation has been captured in most industries where predictive analytics facilities them to make future decisions. Decision making is the process that can be done by the implication of predictive learning and to assume the future interpretants of the profit and loss scenario (Deng & Gao, 2015). All those scenarios are based on the present and historical data that plays an important role in making the future decision and strategies for the business. Thus, predictive analytics, decision making and business intelligence are correlated and the main methodology that plays behind the scenario is machine learning. Predictive analytics is a subdomain of advanced data analytics that is widely used to predict unknown and unseen future events. Predictive analytics applied different techniques for making the prediction of the future based on the present and past data and is extensively used in the domain like data mining, statistical learning, data aggregation, data modelling, machine learning and artificial intelligence etc (Subbalakshmi & Prabhu, 2018). Predictive analytics is essentially used in the business environment where the decision support system is an important division through which future incidents can be predicted. However, for an organizational issue, not a single method is applied to predict the future data using the data mining algorithms, rather a combination of algorithms can be used to predict the data more accurately and sophistically.

## 1.1 Context

Predictive analytics emphasize especially the historical data of the organization that they use to keep in their database. Historical data helps to facilitate the predictive models in data mining with statistical incorporation so that the production of the future value can be made possible. Predictive analytics models use to take the relationships of several factors of the data to assess the statistics and thereby helps to determine the future values and the associated risks. By utilizing such data, the management of the organization can get helpful and meaningful insight into the data statistics and get benefitted using predictive and machine learning (Ge, et al., 2017). Sometimes, the predictive analytics that is used in data mining is misinterpreted with the algorithms in machine learning. At its core, predictive analysis encompasses a variety of statistical techniques. These outcomes might be behaviours a customer is likely to exhibit of the possible change in the market, for example, Predictive analysis helps us to understand possible future occurrences by analysing the past. Machine Learning on other hand is a subfield of computer science that evolved from the study of pattern recognition and explores the notion that algorithms can learn from and make predictions on data and as they begin to become more intelligent (Smirnov & Levashova, 2020). These algorithms can overcome program instructions to make highly accurate, data-driven decisions. By combining both predictive analysis and machine learning, a good and comprehensive outcome can be achieved for predicting future values.

## 1.2 Problem & General Research Question

Predictive analytics facilitates the organizations to become forward-looking, proactive, anticipating outcomes and behaviours that are based upon the historical data and not on false or vague assumptions. Several ways are present there to analyze the data and to present the statistics that will be helpful and focused on making the organization profitable. In the context of business, the organization ~~use to~~ sell the product and services and customers use to consume those. In this context, the data should contain the total volume of transactions that have been done within a certain period, the product selling price and the profit or loss (Ramesh, 2017). On the other hand, customers use to provide their feedback through comments in the digital platform like social media or others that reflect the views and sentiments towards the products and services that the organization provides. So, to vet the insight of the business, the financial data and the textual data both are required to decide on the future strategy. In such contexts, if the predictive analytics models will contain flaws in generating the data statistics, the future prediction may be wrong (Pechenizkiy, 2015). If it will happen, it will place a negative impact on the business for which the organization may face loss in future. So, the determination of the strength of the predictive models is the essential and key issue to detect and overcome the challenges. So, a practical determination is essential to find that fact whether the challenges will be faced at the time of predictive analytics in data mining operations using machine learning or not.

To conduct the research, the review of the previous researchers will be done based on the research question that is mentioned below:

*How the Machine Learning algorithms can overcome the drawbacks of predictive Analytics in Data Mining?*

## 

## 1.3 Methodology and Approach to Research

The approach for the research will be oriented to the determination of the fact whether the challenges in predictive analytics can be faced in the data mining operations using machine learning (Vargas, et al., 2020). To conduct the research, the following steps of approaches will be taken into considerations:

1. To design the search string through which the research articles can be obtained from different journals and other resources.
2. To select the papers with the application of search string.
3. To design the policy for inclusion and exclusion criteria.
4. To map the policy to filter the collected research papers.
5. To accept the papers that will satisfy the inclusion criteria.
6. To review the selected research papers to get the insight and information regarding the taken research question
7. To discuss the important factors of the researches.
8. To conclude the research

## 1.4 Results

The result of the present research will aim to reflect the outcome that will be obtained through the reviews of the previous papers concerning the taken research question. The result will be obtained by maintaining the sequential operations that have been mentioned in the last section.

## 1.5 Implications/conclusions

The research will imply the final outcome of the reviews of the previous researches. In this context, those researches will be reviewed that will be taken into consideration through the inclusion criteria. From the review of the previous researches, the insight of predictive analytics can be obtained along with the application area for it. In the previous researches, it can be found that the researchers have used different predictive models in data mining with the inclusion of machine learning techniques to predict future data (Smirnov & Levashova, 2020). From those researches, the methodology and the application scenario will be obtained which will be essential to fulfil the aim and objective of the research. So, by founding the information and model overviews that have been applied by previous researchers, the research will be concluded by emphasizing the factors of the research.

# 2. Background

Predictive analytics emphasizes the implications of statistics on the data to create effective modelling for the outcomes of the predictions. Predictive Analytics exclusively applied the statistics on the historical data to predict the future value of profit, loss and volume transactions of the products (Asniar & Surendro, 2019). Thus, it has a direct impact and effect on the business organization where the predictive context is essential to determine the future growth and thereby determining the future strategy. Many researchers are involved in proposing new models and approaches to predict the data more accurately and precisely. ~~Every research is done to overcome certain challenges~~. So, the researchers also emphasize the possible challenges and propose the solution through which the challenges of predictive learning in data mining can be overcome when those will be used along with machine learning (Shanthakumari & Jayakarthik, 2021).

Predictive analytics models use a number of data mining algorithms to test the historical data and to create an effective model for future purposes so that the prediction can be done by utilizing that model (Deng & Gao, 2015). However, the strength and the perfection of the predictive models are one of the key issues for which those require proper observation and judgements. Not all algorithms in data mining can play a good role in predicting future values (Shanthakumari & Jayakarthik, 2021).

The researchers of the paper Geng (2017) have emphasized the application of machine learning to predict historical data. As the future prediction exclusively depends upon what was happened in the past, the database is one of the important factors in this context. Thue has argued that the Knowledge discovery Database is the most fruitful in the case of prediction where the challenges can be overcome regarding the predictive models using machine learning. As the database is involved in the predictive model, thus data mining models play important roles in storing, retrieving and analyzing the data for prediction and to make future strategies. The researchers of the paper Ge et al. (2017) have proposed the model to predict the data with effectiveness using predictive analytics and machine learning. According to their research, machine learning provides the basic tool for data analytics. It facilitates the analyst to extract the information, features and to identify the patterns of the data. the identification of data patterns facilitates the analyst and engineer to overcome the challenge that they may face at the time of prediction. In most cases, organizations use to store and retrieve the data in their database and the analytics on that data can be done using data mining algorithms. So, by applying the data mining model in view of predictive analytics, machine learning can help to detect and recover errors using the scoring method that determines the efficiency of machine learning models. The researchers of the paper Asniar & Surendro (2019), Vargas et al. (2020) and Smirnov & Levashova (2020) have focused on the predictive learning method for the decision support system. In the practical view of customer segmentation, the decision support system is one of the essential facts that helps to determine the demand of the scustomers. Using the realization of demands, the busienss organization can thek their future decision about the sales and matketting strategy. The most important part of this decision making is predictive analytics which can be done by fetching and analyzing the data from the database. In this context, the analyst uses to apply the predictive model by identifying and rectifying the errors using machine learning algorithms.

The researchers of the papers Kumar et al. (2020) and Al-Tameemi et al. (2020) have emphasied the beneficial factors of data mining models using predictive learning in the educational sector. The growth of the students in a certain institute depends upon the curriculum structure and the placement statistics of the students. If it will be found that structire and orientation of study in that institute are good enough, more students will gatehr there. Hiwever, to attract the student in the educational institutes, the administrators use to take the prediction of the future aspects concerning the past data that has been stored in the database. Using machine learning and a predictive model, the data mining strategy can be developed through which the future prediction cab eb doen for the educational sector.

The researchers of the papers Subbalakshmi & Prabhu (2018) and Ramesh (2017) have emphaseid the fact of the application of predictive models using machine learning in the financial sector. The financial sector is one of the most important and sensitive sectors in the world. Most of the financial transactions are done using the network. So, the growth of financial institutions like banks depebds upon the number of customers they have presently. So, the predictive learning and model can be applied to predict the possible number of customers that they can seen in the next financial years. Based upon the prediction, the financial institition use to draw the strategy. Based on the strategy, they use to place certain offers to attract more customers where they also invest huge amount for advertising. If the productive model will be erroneous, the financial institiute may face a huge loss. This can be avoided using the appl;icatiomn of machine learning along with predictive analytics in data mining.

# 3. Research Method

This section will be containing the detailed method or research along with the design & application of the protocol to select the research papers.

## 3.1 Protocol for Searching Research Papers

The topic that is based on predictive analytics, data mining and machine learning are the hot topic of research and a huge number of researchers are engaged in it to propose their new models. Thus, a huge number of researches are done. However, the research papers will be selected systematically by incorporating the inclusion and exclusion criteria. The process of selection of research papers by employing the selection policy is shown below:

## 3.2 Search String

Search String is the essential party of the research through which the research papers can be selected and reviewed. The research vaults use to generate the result by predicting the list of research papers by taking the proper search string as input. In this research, the search string has been designed as follows:

*Predictive Analytics AND(machine learning OR Data Mining)*

To create the search string, the following process has been followed:

The search string will be applied to collect the research papers from IEEE Xplore and IEEE Transactions. By collecting the research papers, those will be reviewed as seen in the background section and the information will be extracted from those research papers which will be emphasized in the discussion section.

## 3.3 Inclusion and Exclusion Criteria

The inclusion and exclusion criteria for selection and rejection of the obtained papers are shown below:

|  |  |  |
| --- | --- | --- |
| **Policy of Selection** | **Inclusion Criteria** | **Exclusion Criteria** |
| Language | The English language will be preferred for the selection of the research articles and manuscripts | The research papers in other languages except English will be excluded. |
| Quality of research | The research paper must belong to well-known journals and conferences. | The research papers will be excluded if will not belong to well-known journals or conferences |
| Domain of Research | The domain of the research should fall under the category of Predictive analytics, data mining and machine learning | The papers from other domains but related to Predictive analytics, data mining and machine learning will be excluded. |
| Year of Publication | The research papers will e included if the year of publication will be within 5 years of the current date | The research papers will be excluded if they have been published earlier than 5 years. |
| Digital Object Identifier | The research papers will be included if it will have a valid DOI | The research papers will be excluded if they do not have any valid DOI |
| Method and approach | The research paper should contain lucid and interesting methods for the execution of the objectives | The research papes will be excluded if they do not have any clear and lucid idea and approach. |

## 3.4 Extraction of Data

The research papers will be included using the inclusion criteria. The information of the papers regarding the approaches and the methods that the previous researchers have used will be extracted based on the following criteria:

1. Type of approaches and methods
2. Applied algorithms for predictive analytics
3. Applied algorithms for data mining and machine learning
4. Challenges in the researches
5. The proposed solution of the researches

# 4. Results

The research papers have been searched out using the search string. In this context, by applying the search string, 1152 Conferences papers, 177 Journals, 31 Magazines and 8 books have been found. The research papers have been found to e published from 2000 to 2021. So, by applying the inclusion and exclusion criteria, 70 papers have been selected from which 17 papers have been taken finally emphasizing the aim and research question. The research papers are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **Title** | **Author** | **DOI** |
| 2017 | Data Analytics and Predictive Analytics in the Era of Big Data | Hwaiyu Geng | 10.1002/9781119173601.ch19 |
| 2017 | Data Mining and Analytics in the Process Industry: The Role of Machine Learning | Zhiqiang Ge; Zhihuan Song; Steven X. Ding; Biao Huang | 10.1109/ACCESS.2017.2756872 |
| 2006 | Data Mining and Predictive Analytics in Public Safety and Security | C. McCue | 10.1109/MITP.2006.84 |
| 2016 | Predictive Policing and Civilian Oversight: What Will It Take to Get It Right? | Jesse Hirsh | 10.1109/MPOT.2016.2569723 |
| 2007 | Logistics Management in a Mobile Environment: A Decision Support System Based on Trajectory Mining | Gianni Costa; Giuseppe Manco; Riccardo Ortale; Domenico Sacca; Alessandro D'Atri; Stefano Za | 10.1109/ICONS.2007.33 |
| 2013 | Predictive analytics and proactive service | Detlef Nauck | 10.1049/ic.2013.0234 |
| 2019 | Predictive Analytics for Predicting Customer Behavior | Asniar; Kridanto Surendro | 10.1109/ICAIIT.2019.8834571 |
| 2020 | Intelligent Decision Support Based on User Digital Life Model: Principles and Conceptual Framework | Alexander Smirnov; Tatiana Levashova | 10.23919/FRUCT49677.2020.9211055 |
| 2020 | Prediction of financial product acquisition for Peruvian savings and credit associations | Emmanuel Roque Vargas; Ricardo Cadillo Montesinos; David Mauricio | 10.1109/CONIITI51147.2020.9240413 |
| 2015 | Predictive analytics on evolving data streams anticipating and adapting to changes in known and unknown contexts | Mykoa Pechenizkiy | 10.1109/HPCSim.2015.7237112 |
| 2020 | Predictive Learning Analytics in Higher Education: Factors, Methods and Challenges | Ghaith Al-Tameemi; James Xue; Suraj Ajit; Triantafyllos Kanakis; Israa Hadi | 10.1109/ICACCE49060.2020.9154946 |
| 2018 | Protagonist of Big Data and Predictive Analytics using data analytics | Sakineti Subbalakshmi; CSR Prabhu | 10.1109/CTEMS.2018.8769141 |
| 2021 | An Enhanced Predictive Health Care Analytics in Data Mining Techniques for Chronic Kidney Diseases | A.S. Shanthakumari; R. Jayakarthik | 10.1109/ICOEI51242.2021.9452903 |
| 2017 | Predictive analytics for banking user data using AWS Machine Learning cloud service | Ranjith Ramesh | 10.1109/ICCCT2.2017.7972282 |
| 2020 | Campus Placement Predictive Analysis using Machine Learning | Nikhil Kumar; Ajay Shanker Singh; Thirunavukkarasu K; E. Rajesh | 10.1109/ICACCCN51052.2020.9362836 |
| 2012 | Distributed Big Advertiser Data Mining | Ashish Bindra;Srinivasulu Pokuri;Krishna Uppala;Ankur Teredesai | 10.1109/ICDMW.2012.73 |
| 2015 | An advertising analytics framework using social network big data | Lei Deng; Jerry Gao | 10.1109/ICIST.2015.7289018 |

# 5. Discussion

The research aimed to detect and identify the challenges in the application of the predictive analytics model in data mining with the implication of machine learning algorithms. To fulfil the research, the review of the previous researches have been done and from that review, the following factors have been found:

1. Predictive Learning: The authors of the research papers (Geng, 2017) and (Ge, et al., 2017) have emphasized the application fioeld of predictive learning and have shown that there is a wide range of applications ar there. Using predictive learning with the implicatiomn of data mining models, future prediction can be done and the errors can be detected and rectified using machine learning models.
2. Decision Support System: The authors of the research papers (Asniar & Surendro, 2019), (Shanthakumari & Jayakarthik, 2021), (Asniar & Surendro, 2019), (Vargas, et al., 2020) and (Smirnov & Levashova, 2020) have focused on the decision support system that can be designed and created using predictive models and machine learning algorithms. In this context, the main challenge that the organization can face is the erroneous prediction that can be rectified using machine learning models

# 

# 6. Conclusions

In this paper, the review of the literature has been done based on the research questions. The research papers have not been collected from any random search and rather those have been collected through the implication of search string. The search string has been formulated to cover all possible research papers in the journals and conferrences. During the search by applying the search string, 1152 Conferences papers, 177 Journals, 31 Magazines and 8 books have been found. The research papers have been found to e published from 2000 to 2021. So, by applying the journal selection policy, 70 papers have been selected. Finally, 17 papers have been taken by emphasizing the aim and research question. The research papers have been reviewed and evaluated concerning the fulfilment of the research questions and the approaches & methods have been emphasized. Finally, the important factors of the research paper have been discussed in the discussion section.

# 7. References

Al-Tameemi, G. et al., 2020. Predictive Learning Analytics in Higher Education: Factors, Methods and Challenges. *International Conference on Advances in Computing and Communication Engineering (ICACCE),* 6(2), pp. 1-5.

Asniar & Surendro, K., 2019. Predictive Analytics for Predicting Customer Behavior. *International Conference of Artificial Intelligence and Information Technology (ICAIIT),* 5(2), pp. 1-4.

Bindra, A., Pokuri, S., Uppala, K. & Teredesai, A., 2012. Distributed Big Advertiser Data Mining. *IEEE 12th International Conference on Data Mining Workshops,* 15(8), pp. 1-5.

Costa, G. et al., 2007. Logistics Management in a Mobile Environment: A Decision Support System Based on Trajectory Mining. *Second International Conference on Systems (ICONS'07),* 12(8), p. 34.

Deng, L. & Gao, J., 2015. An advertising analytics framework using social network big data. *Information Science and Technology (ICIST) 2015 5th International Conference,* 8(2), pp. 470-475.

Geng, H., 2017. Data Analytics and Predictive Analytics in the Era of Big Data. *Internet of Things and Data Analytics Handbook,* 3(2), pp. 329 - 345.

Ge, Z., Song, Z., Ding, S. X. & Huang, B., 2017. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access,* 23(5), pp. 20590 - 20616.

Hirsh, J., 2016. Predictive Policing and Civilian Oversight: What Will It Take to Get It Right?. *Potentials IEEE,* 36(5), pp. 19-22.

Kumar, N., Singh, A. S., K, T. & Rajesh, E., 2020. Campus Placement Predictive Analysis using Machine Learning. *2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN),* 17(12), pp. 1-6.

McCue, C., 2006. Data Mining and Predictive Analytics in Public Safety and Security. *IT Professional,* 8(4), pp. 12-18.

Nauck, D., 2013. Predictive analytics and proactive service. *IET Seminar on Data Analytics 2013: Deriving Intelligence and Value from Big Data,* 72(6), pp. 1-4.

Pechenizkiy, M., 2015. Predictive analytics on evolving data streams anticipating and adapting to changes in known and unknown contexts. *International Conference on High Performance Computing & Simulation (HPCS),* 1(1), pp. 1-4.

Ramesh, R., 2017. Predictive analytics for banking user data using AWS Machine Learning cloud service. *2nd International Conference on Computing and Communications Technologies (ICCCT),* 1(1), pp. 1-4.

Shanthakumari, A. & Jayakarthik, R., 2021. An Enhanced Predictive Health Care Analytics in Data Mining Techniques for Chronic Kidney Diseases. *5th International Conference on Trends in Electronics and Informatics (ICOEI),* 14(5), pp. 1-7.

Smirnov, A. & Levashova, T., 2020. Intelligent Decision Support Based on User Digital Life Model: Principles and Conceptual Framework. 1(1), pp. 247-252.

Subbalakshmi, S. & Prabhu, C., 2018. Protagonist of Big Data and Predictive Analytics using data analytics. *International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS),* 1(4), pp. 1-5.

Vargas, E. R., Montesinos, R. C. & Mauricio, D., 2020. Prediction of financial product acquisition for Peruvian savings and credit associations. *Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI),* 12(2), pp. 1-5.