Original Article

A REVIEW ON THE ROLE OF STATISTICAL TOOLS IN EFFECTIVE FUNCTIONALITY OF DATA SCIENCE

Ravi Kiran.G[1], Geetha Manoharan[2], Rajyalakshmi K[3], Sunitha Purushottam Ashtikar[4]

[1]School of Sciences, SR University, Warangal, India.
g.ravikiran@sru.edu.in,
[2]School of Business, SR University, Warangal, India.
geethamanoharan1988@gmail.com
[3]School of Sciences, SR University, Warangal, India.
rajyalakshmi.raji4@gmail.com
[4]School of Business, SR University, Warangal, India.
sunithaashtikar@gmail.com
Corresponding author: Ravi Kiran.G

DOI: 10.47750/pnr.2023.14.02.374

Abstract

For any corporate and financial organization, data science is the most reliable tool for assisting their leaders in making choices based on current events and facts. It was created using a variety of technologies and academic fields. The most significant field that contributes to the deep and basic workings of data science is statistics. In order to conduct essential mechanisms, functions, processes, and procedures to analyze and quantify uncertainty, statistics is assisting the most common tools and approaches in data science. The importance of statistics in delivering essential capabilities such as data gathering and refinement, data processing, data analysis, modeling, model selection, and model validation, as well as visualization for decision-making, is presented in this work. This research article elucidated the essential features of data science using various statistical tools and techniques.

INTRODUCTION

Data science uses a statistical technique combined with machine learning to turn the information source into a high-dimensional display Ethem [1]. One of the most popular fields in data science to analyze and quantify uncertainty is statistics. It is also used in the collection, enrichment, exploration, modeling, validation, analysis, presentation, and reporting of data. Data science is powered by statistical visualization to provide analytical results in visual components like charts, graphs, and maps (Pallavi and Nitin [2]). The visualization tools provide an approachable manner to exhibit and comprehend informational trends, flows, and patterns. The purpose of these visualization tools is to make information accessible to end users in an intelligible manner (Adanma [3]).

It is common practice in the field of data science to use the statistical methods developed within the realm of mathematics in order to gather and analyze numerical information. Knowledge is extracted from datasets compiled from many sources, including social media, via the use of scientific methodologies, processes, procedures, and systems that make up the interdisciplinary area of data science (Jin et al. [4]). When it comes to gaining a more in-depth understanding of data sets, the tools and techniques provided by data science—which include a heavy emphasis on statistics—are invaluable. Data sets are analyzed, uncertainty is quantified, and the most insightful and accurate findings are reported to decision makers in the form of visualization reports (Tracey et al. [5]).

Normal distribution, central tendency, variability, variance, standard deviation, modal, skewness, and kurtosis are some of the most often used statistical operations in data science (He and Huang [6]). In statistics, a bell curve represents the normal distribution. Statistics relies heavily on central tendency procedures such as the mean, median, and mode. In data science, the statistical variability with 25%, 50%, and 5% quartiles are significant when analyzing data sets (He [7]). Statistical inferences are mathematical calculations used to extrapolate conclusions from data. Data analysis in data science makes use of descriptive statistics and probability theory. Major contributions of statistics to data science outside of probability distributions, hypothesis testing, and regression, are presented in Strunk and White [8].

RELATED INFORMATION

The significance of statistics in data science was highlighted by Weihs and Ickstadt [9]. This document outlines the processes and procedures necessary to extract meaningful patterns from data sets and conduct statistical analysis. This study provides a summary of the many potential architectures of data science and shows how statistics affects data collection, enrichment, exploration, analysis, modeling, validation, and reporting via visualization and representation Kozak et al. [10].

In their study, Galeano and Pena [11] details seven ways in which statistical methodologies have evolved. Big Data is thought to have a significant impact on these seven domains. High-dimensional visualization; multiple-testing challenges; heterogeneity analysis; automated model selection; estimating techniques for sparse models; and the integration of network data into statistical models are all examples of this. In this study, we draw parallels between the statistical method and the work done by machine learning and computer science. Importance of statistical analysis and reporting in data science was discussed (Date and Tanaka [12]).

Chen et al. [4] discussed the importance of statistics in the context of Big Data analytics within the discipline of Data Science. Decisions rely heavily on accurate and timely data, and new developments in data science are helping to address these issues. The examples of corporate statistics and biostatistics have been shown to illustrate the role of statistics in the area of data sciences and the latest developments to address the significant issues. New concepts and practical applications in statistics and data science have been presented in Ahmed et al. [13], Jalajakshi and Myna [14].

METHODOLOGY

The explanation of what statistics is and how it fits into the area of data science can be found in the Cross Industry Process for Data Mining document. Himani and Sunil [15], and Liu [16] list a number of the most important definitions, including data storage and processing, data quality improvement, data modeling and interpretations, deep data analytics, acquiring knowledge and exploration, high-performance processing, computation, experiment design, communication, networking, and data-to-inference and action. These are just some of the main definitions.

DATA ACQUISITION AND ENRICHMENT

Data collection and enhancement are the cornerstones of the statistical method. This may be thought of as the underlying process by which statistical approaches produce new data from existing data sets that are already noisy and unstructured. Optimizing algorithms may be implemented into the data collecting and enrichment process. Imputation of missing data is another area where this procedure aids data enrichment. Science is mostly utilized in data science for refining data from unstructured data sets across databases and filling the gaps with data flood (Bischl et al. [17]).

DATA EXPLORATION

Another important task that may be accomplished with the aid of statistics is data exploration. It's a must-have for any pre-processing of data. This statistical method is the backbone of exploratory data analysis, which is used to better comprehend and shape the data warehouse's contents. Users are given the tools they need to examine the data via statistical and visual analysis thanks to this process. This is very helpful in identifying issues and trends in the dataset, and in making a choice on which model or algorithm to employ next. It is mostly on the basis of summary statistics and techniques of visualization in data science (Bottou et al. [18]).

STATISTICAL DATA ANALYSIS

Analysis of statistical data is a pivotal step in the statistical method used in data science. Hypothesis testing, classification, regression, and time series analysis are just some of the most common applications. Testing hypotheses is a crucial element of statistical research. Hypotheses are developed in response to the questions posed by data-driven concerns. Statistical tests, inquiries, and theory all depend on, and even need the formulation of, assumptions. In light of the evidence at hand, we can test these. When the same information is used for many tests, significance thresholds must be adjusted. One answer to the major issues in applied statistics is correct multiple testing (Bühlmann, P., Van De Geer, S [19]).

The ability to identify and predict subpopulati1919ons from data is crucial to many fields, and classification algorithms are a key mechanism for doing so. Finding instances of subpopulations inside a dataset is a challenge that calls for no previous information about their existence. Clustering may make this feasible. Classification rules make up the supervised case. The results from the training data are then used to implement classification. Classification and segmentation of the target object or element will be accomplished using this output. This method of pattern recognition is by far the most common. Both descriptive and inferential statistics form the foundation of the categorization procedures that are put into effect. There are 7 main classification ML algorithms to use for statistical categorization (Cao [20]).

A major use of statistical regression is making predictions based on the best linear relationship between a dependent variable and an independent variable. Incorporating the best fit is as simple as making sure the total of all distances between the form and the actual observations at every place is minimal. This method is used to disentangle the impact of several independent factors on a single dependent variable. Microsoft Excel's built-in statistical features may be used to conduct a linear regression study. Estimating the strength of a link between variables is the cornerstone of statistical regression. Regression methods in various forms are used extensively in machine learning. Linear regression, logistic regression, Ridge regression, Elastic Net regression, Polynomial regression, Lasso regression, and step wise regression are some of the most common ones. Regression analysis is used in data science to differentiate between predictors' strengths, predict an impact, and estimate a trend.

The process of dealing with time series data or doing trend analysis is known as time series analysis. Data organized in a sequence across a range of times or intervals might benefit from its incorporation. Cross-sectional data, data of one or more variables, and data gathered at the same moment in time may all be analyzed using time series analysis. It's the order in which numerical values are recorded. Secular trend, seasonality, cyclicity, and irregularity are the four components of time series analysis. In order to make accurate forecasts, time series data is primarily collected. In data science, this capability is greatly sought after so that visualization reports may be generated for use in making decisions. Time series utilizes the autoregressive integrated moving average components of the data set functionality model to deal with trends, seasonality, cycles, mistakes, and non-stationary characteristics of a data set (Fahrmeir et al. [21]).

STATISTICAL MODELLING

Graphs and networks may be used for statistical modeling. Stochastic differential and difference equations are useful tools for implementing statistical modeling. Using both local and global models might be helpful when

doing statistical modeling. Models in the natural and engineering sciences are modeled using stochastic differential and difference equations. The equations may be solved using approximate statistical models, yielding insights into scientific and engineering ideas. Validity is restricted to subregions of the domain of the relevant variables in statistical models that use local models and globalization (Claeskens and Hjort [22]). When analyzing structural breakdowns to locate areas in a time series, local models are used. The extension of local to global models is another common use of mixture models. If we characterize real-world connections using a mixture model, we may then combine relevant models in the process. Standard mathematical models may be used for analysis of diverse data or wider areas of interest (Cooper et al. [23]).

MODEL VALIDATION AND MODEL SELECTION

In the field of data science, one of the most important tasks is validating and selecting models to use. Model validation is most often used to verify the accuracy of a statistical model's predictions in light of the actual procedure by which the corresponding data was generated. Within regulatory guidelines, model validation may be determined as the process and actions used to ensure that a model is performing as intended in light of its intended business usage and design goals. The major use of this tool is to verify the effect of any presumptions or restrictions that may be present. Selecting the best statistical model from a pool of possible ones is known as model selection (Fahrmeir et al [21]). This term is also used to describe the difficulty inherent in making decisions based on a small subset of available computational models. Optimization under uncertainty is another common use. The nature of the associations between the dependent and explanatory variables should be taken into account when deciding on a statistical model (Dyk et al. [24]).

REPRESENTATION AND REPORTING

Visualization is used for representation and reporting. Using statistics for this purpose is crucial in the field of data science. This crucial part of statistical analysis is used to convey the findings of statistical analysis performed on complicated data sets in the most easily digestible graphical representation style. Knowing your target demographic is essential to creating an effective data visualization strategy, as is determining the best graph or chart style, report format, and ordering, layout, and hierarchy to prioritize your large amounts of data. Throughout the four phases of research—exploration, analysis, synthesis, and presentation—visualization may be used. R-Studio, Tableau, Zoho analytics, IBM Cognos analytics, Qlik Sence, Microsoft Power BI, Python, and many more may all be used as part of a software development environment to create visualizations for use in data science. Python has a plethora of libraries that provide access to interactive data visualization tools.

EVOLUTION

Utilizing a Statistical Method for Data science is on the rise as a viable field for extracting actionable insights from complex data sources. Data gathering, data exploration, data analysis, reporting, and representation for decision making have all been covered in the study report's presentation of the core statistical procedures (Dyk et al. [24]). The field of statistics plays an essential part in the data validation process. The primary value added by statistics is in the realm of distribution. Exploring and modeling data may be valuable for reporting when statistical analysis is applied to the gathered information. Data science is mostly based on statistical analysis. The first step in data exploration is collecting and cleaning up raw data. The foundation for statistical modeling is being laid by this method of analyzing statistical data. Data science may execute the representation with visualization with the use of statistical modeling, including model validation and model selection (Donoho [25]).

CONCLUSION

Because of the statistical approaches' ability to examine and verify data thoroughly, statistics is having an increasingly noticeable effect on data science. Statistics have been shown to be superior than computer science in terms of both accuracy and ease of application. It is established that data science has produced the most wanted reports for difficult decision making in a variety of business solution domains. To do this, a statistical method might be used to several enormous data sets culled from different sectors, such as the financial industry, social media, and online retail.

Statistical methods in data science using artificial intelligence machine learning algorithms have provided useful commercial solutions in several fields. Simple computer algorithms, statistical reasoning, and straightforward comparisons based on appropriate methodologies and models enhance the statistical procedures. Businesses, governments, and nonprofits may use these statistical tools, together with data science, to better foresee their future actions.

REFERENCES

1. A. Ethem, Machine Learning, 35-69, MIT Press, (2021).
2. G. Pallavi and V. T. Nitin, International Journal of Computer Applications, **176**(24), 10 – 14, (2020).
3. E. Adanma, International Journal of Computer Trends and Technology, **38**, 46 – 50, (2016).
4. X. Jin, B. W. Wah, X. Cheng and Y. Wang, Big Data Research, **2(2)**, 59 – 64, (2015).
5. W. Tracey, M. Natasa, W. Stacey, G. Vesna, PLoS Biology **13**, e1002128, (2015).
6. L. He and Z. Huang, 2020 Chinese Control and Decision Conference (CCDC), 4710–4714, (2020).
7. X. He, 2021 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), 747–751, (2021).
8. W Strunk Jr. and E. B. White, The Elements of Style, 3rd ed, Macmillan, New York, (1979).
9. C. Weihs, K. Ickstadt, International Journal of Data Science and Analytics, **6**, 189–194, (2019).
10. M. Kozak, J. Hartley, A. Wnuk, M. Tartanus, Journal of Scholarly Publishing, **46**, 282–289, (2015).
11. P. Galeano, and D. Peña, Data science, big data and statistics, TEST, (2019).
12. K. Date, Y. Tanaka, 2020 International Symposium on Semiconductor Manufacturing (ISSM), 1 – 4, (2020).
13. R. Ahmed, M. Faizan and A.I. Burney, 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS), 1–9, (2019).
14. V. Jalajakshi and A.N. Myna, Global Transition Proceedings, **3**, 326 – 331, (2022).
15. S. Himani, K. Sunil, International Journal of Science and Research, **5**, 2094 – 2097, (2019).
16. J. Liu, 2020 International Conference on Computing and Data Science, 343–346, (2020).
17. B. Bischl, O. Mersmann, H. Trautmann and C. Weihs, Evolutionary Computation, **20(2)**, 249–275, (2012).
18. L. Bottou, F. E. Curtis, J. Nocedal, SIAM Review, **60(2)**, 223 – 311, (2018).
19. P. Bühlmann, S. VanDeGeer, Statistics for High-Dimensional Data: Methods, Theory and Applications, Springer, Berlin, (2011).
20. L. Cao, ACM Computing Surveys, **50(3)**, 1 – 42, (2018).
21. L. Fahrmeir, T. Kneib, S. Lang and B. Marx, Regression: Models, Methods and Applications. Springer, Berlin, (2013).
22. G. Claeskens, and N.L. Hjort, Model Selection and Model Averaging. Cambridge University Press, Cambridge, (2008).
23. H. Cooper, L. V. Hedges, J. C. Valentine, The Handbook of Research Synthesis and Meta-analysis. Russell Sage Foundation, New York City, (2009)
24. D.V. Dyk, M. Fuentes, M. I. Jordan, M. Newton, B. K. Ray, D. T. Lang, H. Wickham, ASA Statement on the Role of Statistics in Data Science, (2015).
25. D. Donoho, Journal of Computational and Graphical Statistics, **26(4)**, 745 – 766, (2017).