Journal of Advanced Zoology

ISSN: 0253-7214

*Volume **45** Issue **S-4** **Year 2024** Page **107-112***

Impact Of Data Visualization In Data Analysis To Improve The Efficiency Of Machine Learning Models

Ms. Bhakti Govind Shinde[1*], Dr. Sunayana Shivthare[2]

[1]Assistant Professor, Indira College of Commerce and Science, Pune, Maharashtra, India. Email: krishnabhakti.shinde@gmail.com*
[2]Assistant Professor, MAEER's MIT Arts, Commerce and Science College, Alandi, Pune, Maharashtra, India. Email: sunayanashivthare@gmail.com

***Corresponding Author:** Ms. Bhakti Govind Shinde*
Email: krishnabhakti.shinde@gmail.com

Abstract

An essential component of machine learning is data visualization, which helps analysts comprehend and interpret patterns, connections, and trends in data. Data visualization is a crucial aspect of machine learning that enables analysts to understand and make sense of data patterns, relationships, and trends. Through data visualization, insights and patterns in data can be easily interpreted. This research paper explores the significant impact of data visualization on the efficiency of machine learning (ML) models during the data analysis phase. Data visualization serves as a powerful tool for data scientists and ML practitioners by offering intuitive insights into complex datasets, facilitating a deeper understanding of the underlying data characteristics, and guiding the decision-making process in model development. The visual techniques enhance various aspects of the data analysis phase, including exploratory data analysis (EDA), feature selection and engineering, anomaly detection, and assumption validation, ultimately leading to the development of more accurate and efficient machine learning models.

Keywords: Machine Learning, Data Visualization, Data Analysis.

1. Introduction

Since a picture truly is worth a thousand words, visualization techniques are probably the most appealing and useful forms of knowledge extraction techniques. With the increasing availability of big data, it has become more important to use data visualization techniques to explore and understand the data which can be used to improve decision-making process. The analysis phase in machine learning is critical for understanding the data, uncovering patterns, and preparing the dataset for modeling. Visualization techniques are indispensable during this phase, offering intuitive insights into the data those numerical summaries. Data visualization helps machine learning analysts to better understand and analyze complex data sets by presenting them in an easily understandable format. Data visualization is an essential step in data preparation and analysis as it helps to identify outliers, trends, and patterns in the data that may be missed by other forms of analysis.

With the increasing availability of big data, data visualization will continue to be an important part of the machine learning process, helping analysts to develop more accurate and reliable machine learning models. The use of data visualization in the data analysis phase of machine learning significantly contributes to the efficiency of ML models. It emphasizes the role of visualization in enhancing understanding, guiding preprocessing, feature engineering, and model selection, thereby leading to the development of more effective and efficient machine learning systems. Machine learning algorithms work best when they have high-quality and clean data, and data visualization can help to identify and remove any inconsistencies or anomalies in the data

2. Related Work

This section looks at recent experiments that used data visualization throughout the machine learning data analysis phase.

The author of [1] provides some clues to the current state and the near future of visualization methods within the framework of Machine Learning. Huge amount of information, coded in many different features, justifies the research on new methods of knowledge extraction: the great challenge is the translation of the raw data into useful information that can be used to improve decision making processes, detect relevant profiles, find out relationships among features, etc.

The author of [2] gives the data visualization skills possessed by data-centric practitioners reflect recent advances in computing systems and visual design principles, and their role in generating relevant and insightful output to support decision-making. Data visualization as a pillar for data analytics and business intelligence in organizations and businesses. Using data visualization to analyze data and present valuable insights is not only vital to engineering and computer science disciplines, but also to business disciplines. For instance, disciplines such as finance and accounting utilize visualization techniques to analyze and present financial and accounting data in ways that are easily digestible to stakeholders

According to author of [3] Data visualization involves presenting data in graphical or pictorial form which makes the information easy to take in. It helps to explain facts and determine courses of action. Criminology is an interesting application where data visualization plays an important role in terms of prediction and analysis. The primary purpose for data visualization is to assist people with processing large amounts of information. Data volumes are large and human cognitive capacities to remember and understand data are limited. Data visualization should be made to simplify visualization as much as possible to help people make more effective decisions. This study presents a contribution in the field of data visualization, with a focus on crime data. This involves designing and creating a data set, which will be used in data visualization and machine learning using various visualization techniques and machine learning algorithms. This is about doing data visualization for those who want to have knowledge of the data, interpret it and take decisions from the data.

The author of[4] Explored various visualization tools provide an insight for business schools across the world to be able to determine which software packages are worth investing the time and capital in to best prepare their students for employment opportunities

According to [5] representing data in a graphical or pictorial form, making it easier to understand and interpret. With the increasing availability of data in various domains, such as business, social sciences, humanities, sports, environmental sciences, and healthcare, the importance of data visualization has never been greater. This research paper provides a comprehensive overview of data visualization tools and techniques and their applications in various domains. This research aims to highlight the importance of data visualization in effectively communicating and analyzing data to provide insights into the various types of data visualization tools and techniques available.

The author of [6] provides information on the field of data visualization. It has appeared as a very powerful, widely acceptable and applicable tool to analyze and understand huge and complex data

The author of[7] explored the practice of data iteration in production machine learning model. The importance of designing data as equally important as designing models in the ML process, inspiring future research and tooling around evolving data.

According to [8] Machine learning algorithms and traditional data mining process usually require a large volume of data to train the algorithm-specific models, with little or no user feedback during the model building process

3. Discussions

3.1 Life cycle of machine learning

Machine learning life cycle involves seven major steps, which are given below.

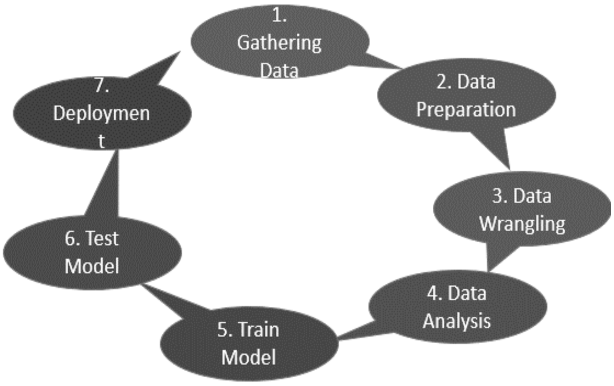

Figure 1: Life cycle of machine learning

The most important thing in the complete process is to understand the problem and to know the purpose of the problem. Therefore, before starting the life cycle, we need to understand the problem because the good result depends on the better understanding of the problem.

Gathering Data

The data gathering phase in the machine learning life cycle focuses on collecting relevant and diverse data from various sources to address the defined problem.

Data Preparation

Data preparation involves cleansing, structuring, and transforming raw data into a format ready for analysis and modeling. It addresses issues like missing values, outliers, and normalization to ensure data quality and consistency for effective model training.

Data Wrangling

Data wrangling involves organizing raw data into a structured, usable format for analysis. This phase addresses inconsistencies, missing values, and data integration, setting a clean foundation for modeling.

Data Analysis

Data analysis in the machine learning life cycle involves exploring and examining the prepared data to identify patterns, insights, and relationships. This phase uses statistical methods and visualization tools to inform the subsequent steps of feature engineering and model selection.

Train Model

The train model phase involves feeding prepared data into machine learning algorithms to learn from it, adjusting parameters to minimize error and improve accuracy. This critical step produces a predictive model capable of making decisions or forecasts based on new data.

Test Model

The test model phase evaluates the trained model's performance on a separate, unseen dataset to assess its generalization ability and accuracy. This step involves using metrics like precision, recall, and F1-score to ensure the model's reliability in real-world applications.

Deployment

The deployment phase integrates the trained and tested model into a production environment, enabling it to make predictions on new data. This step requires ensuring the model's scalability, performance monitoring, and maintenance for continuous operation.

3.2 Working Strategy of Machine Learning

The working strategy of machine learning encompasses defining a clear objective, collecting and preprocessing relevant data, dividing it into training, validation, and test sets, and selecting an appropriate algorithm for the task. The model is then trained on the training set, allowing it to learn from the data. Its performance is evaluated and fine-tuned using the validation set to optimize accuracy and reduce overfitting. After tuning, the model is tested on unseen data to assess its real-world applicability. Upon satisfactory performance, it is deployed into production where it makes predictions or decisions with new data.

Continuous monitoring and updating are essential to ensure the model adapts to new data or changes in the underlying problem space, maintaining its effectiveness and accuracy over time. This strategy encapsulates a cyclical process of development, evaluation, deployment, and maintenance, highlighting the iterative nature of machine learning projects.

The working strategy for machine learning involves certain crucial components, such as historical data, machinelearning algorithms, training programmes to hone those algorithms on historical data, and analysing a logical model based on the outcome. The model can be fed test data and can provide the necessary outcomes dependingon them after training on historical data and fine-tuning the logical model.

3.3 Working of data preparation and data analysis phase of machine learning

The data preparation and analysis phases in machine learning involve transforming raw data into a clean, organized format suitable for building robust models, and then examining this data to uncover patterns and insights. In data preparation, the raw data undergoes cleaning to remove inaccuracies and inconsistencies, imputation of missing values, normalization or scaling to ensure data is on a similar scale, and feature engineering to create new features that could enhance model performance. This stage is crucial for improving the quality and effectiveness of the data, ensuring that the machine learning models can learn from the most relevant and accurate information possible. Following preparation, data analysis involves statistical analysis and visualization techniques to explore the data, identify trends, outliers, and relationships among variables. This exploratory phase helps in understanding the underlying structure of the data, informing the choice of machine learning models, and guiding the subsequent steps of model training and evaluation with the goal of achieving accurate predictions or insights from the data.

3.4 Data visualization

Data visualization is a powerful technique that involves the graphical representation of data to enable easier understanding, interpretation, and insight extraction from complex datasets. By converting large volumes of data into visual formats such as charts, graphs, maps, and infographics, data visualization helps to uncover patterns, trends, outliers, and relationships that might not be apparent from raw data alone. It serves as an essential tool in data analysis and decision-making processes, allowing both technical and non-technical audiences to grasp the significance of data by highlighting key themes and summarizing vast amounts of information in a clear, concise, and engaging manner. Effective data visualization facilitates the communication of insights and findings in a visually appealing way, making it an indispensable element in data science, business intelligence, and any domain where data-driven decisions are crucial.

3.5 Data visualization role in data preparation and analysis phase of machine learning

Data visualization plays a pivotal role in the data preparation and analysis phases of machine learning by providing an intuitive means to explore, understand, and communicate the underlying patterns and insights within the data. During data preparation, visualization aids in identifying issues such as outliers, missing values, and distribution skewness, enabling more informed cleaning and preprocessing decisions. In the analysis phase, it becomes indispensable for exploring relationships between variables, understanding the data's structure, and uncovering potential features that could improve model performance. By transforming complex datasets into graphical representations, data visualization facilitates a deeper insight into the data, guiding the selection of appropriate modeling techniques, and enhancing the overall effectiveness of the machine learning process. It bridges the gap between raw data and actionable insights, making it easier for stakeholders to grasp complex concepts and make data-driven decisions.

3.6 Data visualization tools and techniques used in data preparation and analysis phase of machine learning

Data visualization tools and techniques are crucial in the data preparation and analysis phases of machine learning for uncovering insights, identifying trends and anomalies, and communicating findings effectively. Here's a list of commonly used tools and techniques:

- **Data Visualization tools:**

1. **Matplotlib:** A Python library that is widely used for creating static, interactive, and animated visualizations in Python.
2. **Seaborn:** Built on top of Matplotlib, Seaborn simplifies the creation of beautiful and informative statistical graphics.
3. **Pandas Visualization:** Offers built-in capabilities for data visualization in Python, leveraging Matplotlib

under the hood for plotting.

4. **Plotly:** A versatile tool that enables the creation of complex interactive plots. It supports various programming languages, including Python.
5. **Tableau:** A powerful and fast-growing data visualization tool used in the Business Intelligence industry. It helps in creating highly interactive and user-friendly visualizations.
6. **Power BI:** A business analytics service by Microsoft that provides interactive visualizations and business intelligence capabilities with an interface simple enough for end users to create their own reports and dashboards.
7. **D3.js:** A JavaScript library for producing dynamic, interactive data visualizations in web browsers. It's powerful for creating complex and rich data visualizations.
8. **GGplot2:** A data visualization package for the statistical programming language R, focusing on the principles of "The Grammar of Graphics" to create complex plots from data in a data frame.

- **Data Visualization Techniques**

1. **Histograms:** Useful for visualizing the distribution of numerical data.
2. **Scatter plots:** Great for observing relationships between two numeric variables and detecting potential correlations.
3. **Box plots:** Helpful in visualizing the distribution of data in terms of quartiles and detecting outliers.
4. **Heatmaps:** Useful for representing matrix data where colors represent values, excellent for spotting patterns and correlations in feature sets.
5. **Pair plots:** Provide a way to see both distribution of single variables and relationships between two variables, ideal for initial exploration of data.
6. **Line charts:** Essential for visualizing data trends over a period.
7. **Bar charts:** Effective for comparing quantities among different groups.
8. **Correlation matrices:** Visually explore the correlation between multiple variables at once, often represented as heatmaps.

These tools and techniques are foundational for data scientists and analysts to extract, analyze, and communicate insights from data, facilitating data-driven decision-making in the machine learning lifecycle.

4. Conclusion

Data visualization is an essential tool for machine learning analysts to analyze and understand complex data sets. By using data visualization techniques, analysts can identify trends, patterns, and anomalies in the data and communicate these insights to stakeholders in a format that is easily understandable. The strategic use of data visualization in the data analysis phase not only streamlines the development of machine learning models but also significantly improves their efficiency and effectiveness. By providing a visual understanding of the data, visualization acts as a catalyst for more informed and strategic decisions in the model development process, leading to the creation of robust, accurate, and efficient machine learning solutions. Data visualization serves as a powerful tool for data scientists and machine learning practitioners, enabling them to make informed decisions about feature selection, model choice, and tuning. By providing intuitive insights into the underlying data, visualization techniques facilitate a more effective preprocessing and cleaning process, ensuring that the data fed into machine learning models is of high quality and relevance. This directly impacts the efficiency of the models, as better-prepared data leads to more accurate predictions and classifications.

5. References

1. Vellido Alcacena, A., Martín, J. D., Rossi, F., & Lisboa, P. J. (2011). Seeing is believing: The importance of visualization in real-world machine learning applications. In *Proceedings: 19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN 2011: Bruges, Belgium, April 27-28-29, 2011* (pp. 219-226).
2. Asamoah, D. (2022). Improving Data Visualization Skills: A Curriculum Design. *International Journal of Education and Development Using Information and Communication Technology*, *18*(1), 213-235.
3. Llaha, O., & Aliu, A. (2023). Application of Data Visualization and Machine Learning Algorithms for Better Decision Making.
4. Diamond, M., & Mattia, A. (2017). Data visualization: An exploratory study into the software tools used

by businesses. *Journal of Instructional Pedagogies*, *18*.

5. Srivastava, D. An Introduction to Data Visualization Tools and Techniques in Various Domains.
6. Siddiqui, A. T. (2021). Data visualization: A study of tools and challenges. *Asian Journal of Technology & Management Research (AJTMR) ISSN*, *2249*(0892).
7. Hohman, F., Wongsuphasawat, K., Kery, M. B., & Patel, K. (2020, April). Understanding and visualizing data iteration in machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-13).
8. Li, H., Fang, S., Mukhopadhyay, S., Saykin, A. J., & Shen, L. (2018, December). Interactive machine learning by visualization: A small data solution. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 3513-3521). IEEE.
9. Beh, J. Z. (2023). Interactive machine learning by visualization: a small data solution.