International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 12 | Dec 2024 www.irjet.net p-ISSN: 2395-0072

Optimizing Machine Learning Algorithms for Enhanced Data Quality and Integrity in Real-Time Processing Environments

Purvaja Biche[1], Aditya Utpat[2]

[1]*Department of Computer Science SP College, Pune, India*
[2]*Department of Computer Science JSPM's Rajarshi Shahu College Of Engineering, Pune, India*

***

***Abstract*—Continual advancement in real-time data processing has brought to light the demand for high-quality and high- integrity data to support judgment in a dynamic atmosphere. The current study ambitions to upscale data quality and in- tegrity through optimization of the machine learning model as illustrated by the context of JPMorgan transactional data. This environment often brings together high-frequency trading and incredible volumes of real-time transactions. The proposed study relies on a robust methodology to evaluate the impact of hyper parameter tuning on three predictive models, i.e., CNN, SVMs, and Random Forests. By mandating to a duteous data pre-processing process and enacting measured hyper parameter optimization, the study finds that model performance notably im- proved. Discovery highlights the SVM and Random Forest models that demonstrated refined predictive capability as measured by a substation reduction in RMSE and a notable enhancement in accuracy. By contrast, while performance remained stabilized, the CNN model showcased a trade-off between RMSE and persistence, suggesting adaptable output in dynamic settings. This finding demonstrates the fine balance amidst precision and adaptiveness critical to real-time usage. Outcome indicate that upgraded, optimized model exhibit potential transformational abilities when utilized in real-world use cases including fraud detection, predicting stock market shifts, and image identification. The study augments the existent literature regarding algorithmic harmony while also enabling a particular course of action to make maximal utilization of machine learning models in real-life, fast-paced data ecosystems. The study contributes to promoting data integrity as a crucial aspect that underpins efficacy, consistency, and judgment making in contemporary finance.**

***Index Terms*—Machine Learning, Real-Time Processing, Data Quality, Data Integrity, Hyperparameter Tuning, Financial Data Analytics, Convolutional Neural Networks (CNN),Support Vector Machines (SVM),Random Forest Models**

I. INTRODUCTION

In the crucible of contemporary data-driven landscapes, the optimization of machine learning algorithms emerges as a decisive factor in the pursuit of impeccable data quality and integrity (Allioui et al., 2023). Our focus narrows onto a spe- cific arena, leveraging JPMorgan's transaction data, unraveling the intricate tapestry of challenges embedded in real-time data processing environments.

A. *Contextualizing the Challenge*

Institutions like JPMorgan navigate the complexities of the financial world by managing vast volumes of transactional data, a stark contrast to the traditional batch processing sys- tems (George, 2024). This real-time data ecosystem, especially evident in high-frequency trading, demands instantaneous decision-making, introducing a set of unique challenges. For example, during a particularly volatile trading day, JPMorgan's systems must accurately process over 100,000 transactions per second, each needing validation and execution within mil- liseconds to capitalise on fleeting market opportunities. This scenario underscores the critical need for ultra-low latency and high reliability in their trading infrastructure to ensure competitive advantage and operational integrity in a landscape where every fraction of a second counts (Bi et al., 2024). Additionally, maintaining data accuracy and security amidst this high-speed transactional flow poses significant challenges, requiring sophisticated algorithms and robust cybersecurity measures to navigate the dynamic, high-stakes environment of real-time financial markets.

B. *Implications of Inaccurate Data*

The implications of a misstep in this real-time dance with data are profound, especially when tethered to the intricate web of JPMorgan's operations (Hoffman, 2022). Imagine a momentary glitch distorting transactional records a ripple effect emerges. From inaccurate financial reporting to operational inefficiencies and potential regulatory non-compliance, the fallout is tangible. Consider a scenario where a minor data inconsistency in a high-frequency trading algorithm results in erroneous buy/sell decisions, potentially translating into significant financial losses.

C. *Associated Costs for Large Organizations*

The financial repercussions of managing inaccurate data in the realm of real-time processing are profound, extending beyond mere monetary losses to impact operational effi- ciency and reputational standing for institutions like JPMorgan (haloumis, n.d). This financial giant allocates millions annually to data cleansing initiatives, endeavoring to correct errors, purify datasets, and uphold order amidst the relentless influx of real-time information. Such efforts are not only resource-intensive but also critically time-sensitive. For instance, a report highlighted that JPMorgan spent approximately $600 million in one year on data management and cleansing op- erations alone, aiming to mitigate the repercussions of data inaccuracies. This investment reflects a strategic necessity rather than a discretionary choice, underlining the importance of accuracy for maintaining competitive edge and operational smoothness (Abdulsalam, n.d). Each instance of inaccuracy not only dents the firm's financial health by potentially mil- lions but also challenges the integrity of its transactional processes, showcasing the high stakes involved in the real- time data ecosystem.

D. *Defining Key Concepts*

To navigate this labyrinth, it's pivotal to grasp our key con- cepts. Optimization here signifies more than mere efficiency; it's about crafting machine learning algorithms finely tuned to the nuances of real-time financial data at JPMorgan. Data quality transcends beyond correctness; it embodies precision, consistency, and completeness, ensuring that each transactional record is an accurate representation of the financial reality it mirrors (Rambe et al., 2020). Simultaneously, data integrity underscores the reliability and trustworthiness of this data journey, from the initiation of a transaction to its assimilation into decision-making processes.

E. *Research Question, Aims, and Objectives*

As we set the stage, the focus sharpens on our guiding question: **How can machine learning algorithms be cali- brated to not just process but enhance the quality and in- tegrity of JPMorgan's real-time transaction data?** Our aims encompass unraveling the intricacies of data optimization, forging methodologies that transcend traditional paradigms, and proposing solutions that echo across the sprawling cor- ridors of large organisations (Allioui, 2023). The objectives unfurl as follows: delving into existing literature to fortify our understanding, devising a systematic methodology attuned to the nuances of real-time data, and presenting results that not only address the outlined challenges but pave the way for pragmatic recommendations. The journey unfolds in the subsequent sections, grounded in the tangible complexities of real-world financial data intricacies.

II. LITERATURE BACKGROUND

A comprehensive survey titled "A survey of machine learn- ing for big data processing" was published in the EURASIP Journal on Advances in Signal Processing. This survey is where our adventure begins as we investigate the landscape of machine learning techniques for real-time data processing (Roshan et al., 2024). The purpose of this in-depth review is to investigate various advanced learning approaches, including transfer learning, deep learning, distributed and parallel learn- ing, and representation learning. The issues that are presented by the processing of large amounts of data in contexts that are dynamic are addressed by these methods, which are particularly relevant to real-time scenarios and offer insights into how they do so (Boppiniti, 2021). In spite of the fact that the sources do not contain any clear talks on the difficulties associated with preserving the quality of real-time data, the article titled "A Review on Machine Learning Strategies for Real-World Engineering Applications" written by Hindawi offers a more comprehensive viewpoint. Although the paper is primarily concerned with applications of machine learning in a variety of engineering fields, it does, in a roundabout way, shed light on potential difficulties associated with the management of real-time data (Karnati et al., 2024). When we are trying to optimise algorithms for real-time data processing, it is absolutely necessary for us to have a solid understanding of the current state of the art in machine learning. For the purpose of our inquiry, the literature on optimisation strategies for machine learning algorithms, which is described in the article on massive data processing, becomes an essential compo- nent. This source dives into more advanced machine learning techniques, with a particular focus on the effectiveness of computational and statistical methods (Karnati et al., 2024). It investigates methods such as deep learning and distributed learning, both of which are essential in the process of optimis- ing machine learning algorithms for use in real-time contexts. The use of these optimisation tactics is necessary in order to improve the efficiency and speed with which data is processed. In the process of moving from the realm of academia to the sphere of practical applications, the "Contract Intelligence" (COiN) platform developed by JPMorgan Chase emerges as an intriguing case study. COiN demonstrates the revolutionary power of machine learning in practice by utilising Natural Language Processing (NLP) to automate the extraction of crucial data from legal documents. This demonstrates how machine learning can be used to real-world circumstances (Karnati et al., 2024). The dramatic reduction in time, which went from performing human review, which required 360,000 labour hours, to using machine learning, which only requires a few hours, not only shows enhanced efficiency but also demonstrates significant cost savings. The enormous budget allocation of $15.3 billion in 2023 that JPMorgan has made for technology, which is evidence of their strategic commitment to technology, highlights how important it is to remain at the forefront of artificial intelligence and machine learning tech- nologies. JADE, which stands for JPMorgan Chase Advanced Data Ecosystem, and Infinite AI are two of the bank's internal platforms that demonstrate the institution's commitment to developing robust infrastructures for data management and advanced analytics (Karnati et al., 2024). Our objective is to optimise machine learning algorithms for improved data quality and integrity in real-time processing contexts, and this strategic approach is in line with that objective. A concrete illustration of the application of machine learning may be seen in JPMorgan's effort to automate the processing of legal papers through the implementation of COiN there. The bank was able to accomplish a dramatic reduction in the amount of time spent by increasing the number of jobs that were automated, such as the interpretation of commercial-loan agreements, from 360,000 hours annually to mere seconds (Karnati et al., 2024). This case not only highlights the costs- effectiveness of such applications in large organisations, but it also highlights the efficiency benefits that may be achieved through the implementation of machine learning.

III. METHODOLOGY

The research question at hand focuses on developing ma- chine learning models that can effectively address a specific problem. The methodology adopted for this research encom- passes a systematic and multi-faceted approach, integrating data preprocessing techniques, hyperparameter tuning, and the utilization of diverse machine learning algorithms (Wilson et al., 2024). This section will provide a detailed account of the steps involved, including data gathering, preprocessing, and the design and improvement of machine learning models.

A. *Data Preprocessing*

The first crucial step in the methodology is data prepro- cessing, where raw data is refined to enhance its quality and prepare it for modeling (Mishra et al., 2020). Python libraries, such as numpy, pandas, and scikit-learn, are employed for efficient data handling. The dataset undergoes scaling using the StandardScaler to standardise feature values. A careful train- test split is conducted to ensure that the models are evaluated on unseen data, promoting generalisation.

B. *Hyperparameter Tuning with Keras Tuner*

To optimise the performance of neural network models, a robust hyperparameter tuning approach using Keras Tuner is integrated into the methodology. Keras Tuner facilitates an efficient search for the optimal hyperparameters of a neural network (Shawki et al., 2021). The *model_builder* function is defined to construct the neural network architecture, and a Hy- perband tuner is configured. The search is guided by an objec- tive metric, in this case, the *val_root_mean_squared_error*. This process not only enhances the model's predictive capa- bilities but also prevents overfitting through early stopping.

C. *Support Vector Machine (SVM) Model*

In addition to neural networks, a Support Vector Machine (SVM) model is incorporated into the methodology. The SVM model offers a different approach to the problem and diversifies the machine learning techniques employed. Grid search is conducted over a defined parameter grid to identify the optimal combination of hyperparameters (Alibrahim et al., 2021). The SVM model is trained using the best parameters obtained from the search, and its performance is evaluated based on Root Mean Squared Error (RMSE).

D. *Random Forest Model*

Further diversifying the machine learning ensemble, a Ran- dom Forest model is included in the methodology. Similar to the SVM model, grid search is employed to find the best hyperparameters for the Random Forest model (Reddy et al., 2020). The model is then trained using the optimal parameters, and its performance is evaluated based on RMSE.

E. *Justification of the Methodology*

The chosen methodology is justified based on its com- prehensive and systematic approach. By integrating different machine learning algorithms and techniques, the research ensures a robust evaluation of the problem (Azevedo et al., 2024). The use of Keras Tuner for hyperparameter tuning enhances the efficiency of neural network models, while the inclusion of SVM and Random Forest models provides a holistic perspective. The methodology not only addresses the research question but