

## Introduction to Data Science and Machine Learning to Support Business Decisions Mini-track

**Dursun Delen**

Oklahoma State University

**Behrooz Davazdahemami**

University of Wisconsin-  
Whitewater

**Hamed M Zolbanin**

University of Dayton

In this mini-track, we are thrilled to present a diverse selection of ten exceptional papers, each contributing valuable insights to the ever-evolving landscape of business analytics research. These contributions span a wide array of high-level topics, including innovative approaches to healthcare data analysis, novel methodologies in recommender systems, robust optimization techniques in data analysis, and advanced machine learning methods for classification tasks. These papers offer a glimpse into cutting-edge research and practices across various domains, demonstrating the HICSS' commitment to fostering knowledge exchange and exploration in the rapidly changing world of academia and technology.

The study by *Amini et al.* addresses the challenges within Electronic Health Record (EHR) systems. These systems contain vast patient data, and the authors leverage explainable AI techniques and predictive analytics to enhance Clinical Decision Support Systems (CDSS). The primary obstacle lies in managing the extensive EHR data, with millions of patient records and various features often plagued by missing values. Their groundbreaking framework offers a solution by efficiently handling the issue of incomplete EHR data. This tool allows researchers to identify crucial variables while maintaining an acceptable level of missing data. The authors exemplify its effectiveness by applying the framework to develop a CDSS for Parkinson's disease diagnosis. Since Parkinson's disease can be elusive to diagnose, the framework's adaptability is noteworthy. It can be integrated into EHR systems or serve as a standalone tool, making it accessible even to non-specialist healthcare practitioners in remote areas. Their results demonstrate improved predictive model accuracy, ultimately identifying undiagnosed patients, a crucial step in improving healthcare outcomes.

The paper authored by *West and Deuse*, delves into the realm of Machine Learning (ML) for anomaly detection in time series data originating from screw-driving operations, a pivotal component of the manufacturing process. The authors leverage an innovative, open-access real-world dataset to explore the effectiveness of various unsupervised and

supervised ML models. Notably, within the realm of unsupervised models, DBSCAN emerges as the frontrunner, boasting an accuracy of 96.68% and a Macro F1 score of 90.70%. In the supervised arena, the Random Forest classifier takes the lead, achieving an impressive accuracy of 99.02% and a Macro F1 score of 98.36%. These results not only underscore the potential of ML in elevating manufacturing quality and efficiency but also shed light on the practical challenges involved in their implementation. This research serves as a catalyst for further exploration and refinement of ML techniques in the domain of industrial anomaly detection, ultimately contributing to the advancement of resilient, efficient, and sustainable manufacturing processes. Notably, the authors demonstrate their commitment to open science by making the entire analysis, complete dataset, and Python-based scripts available through a dedicated repository, a move aimed at supporting practical applications and future adaptations of their work in facilitating business decisions within quality management and the manufacturing industry.

*Lin et al.*, in their study, delve into the dynamic and intricate landscape of livestreaming e-commerce, which continually introduces new products, resulting in a unique and evolving context. In order to effectively navigate the balance between exploration and exploitation within this rapidly changing recommendation environment, the authors put forth a reinforcement learning-based solution that places a strong emphasis on the interplay between customers, streamers, and products. The authors employ Recurrent Neural Networks (RNN) to model the shifting preferences of users for streamers and products, maintaining a focus on long-term engagement. Their novel recommendation system termed the Livestreaming E-commerce Recommendation System (LERS), is designed to enhance the exploration of new products by integrating uncertainty into neural networks through Variational Autoencoders (VAE) for user modeling and Bayesian Neural Networks (BNN) for product recommendations. Their findings underscore the promising practical applications of their algorithm in the context of live-streaming e-commerce, which is

characterized by its ever-evolving product landscape and the intricate relationships between users, streamers, and products.

In the paper authored by *Sun et al.*, the authors address a critical aspect of modern recommender systems, focusing on sequential recommendation methods. These methods are essential for their ability to grasp the evolving context of user interactions, primarily based on their recent behaviors. Despite the success of these approaches, the authors argue that they often overlook a crucial dimension: the time intervals between user interactions. Neglecting this temporal information can hinder the ability of recommender systems to learn high-quality user representations. To overcome this limitation, the authors introduce a time interval-sensitive mechanism for sequential recommendation. They seamlessly incorporate this mechanism into the Gated Recurrent Unit (GRU) and aptly named it "2Gated-TimeGRU". Their innovative approach involves utilizing the time intervals between consecutive user interactions as an additional model feature. Through comprehensive experiments on real-world datasets, the authors demonstrate that their proposed method outperforms state-of-the-art baseline models. This enhancement translates to a more adept capacity for capturing sequential user preferences and improving the overall recommendation accuracy, making it a significant contribution to the field of recommender systems.

In a rapidly evolving construction industry, accidents, particularly earthwork foundation pit collapses, pose significant risks, causing casualties and economic losses. *Shi et al.*, in their study, aim to identify the causes of these accidents and their relationships to enhance safety and prevention in construction. Their approach involves text-mining historical construction safety reports and extracting keywords related to causative factors. They then analyze risk factors to construct a fault tree for earthwork foundation pit collapse accidents, identifying the structural importance of these factors. In the final stage, they transform the fault tree into a Bayesian network, enabling them to predict the probability of top events and analyze node sensitivity. This comprehensive study offers a scientific reference for construction accident prevention and prediction, crucial for improving safety in the construction industry. Shi et al.'s approach provides valuable insights and strategies for effective supervision and prevention during the construction process, ultimately reducing the occurrence and impact of accidents.

In the paper authored by *Mesana et al.*, an innovative approach is introduced for assessing the risk of re-

identification of individuals within data release strategies. These strategies encompass data redaction, data anonymization, and data synthesis. The core of this approach involves simulating an attacker engaged in singling-out attacks, aligning with data protection regulations, and evaluating attacks based on record linkability and the information gain achieved by the attacker. Notably, the approach is further enhanced by modeling attacks as a cooperative game, wherein the value of attackers' information resources is determined using the Shapley value borrowed from game theory. The authors substantiate the effectiveness of their approach by applying it to the Adult Income Census (AIC) dataset. Subsequently, they delve into the economic implications associated with privacy breaches. This research addresses the critical need for a better understanding of the inherent trade-off between preserving privacy and maximizing data utility.

In the study conducted by *Goldberg et al.*, the critical role of financial disclosures in comprehending a firm's status and future performance is explored. This study delves into the predictive potential of communications that occur during investor relations calls. These calls capture unscripted narratives exchanged between a firm's senior leadership and industry analysts. The primary objective is to investigate the extent to which the interplay between the tone of public questions and senior leadership's responses can anticipate a firm's future performance. The findings reveal that the average sentiment of questions posed has a persistent positive association with the average stock price in the subsequent quarter. In contrast, the sentiment of the answers provided does not significantly predict future performance. This study offers a novel perspective on financial disclosures, underscoring the value of oral communications and their tones as essential tools for gaining insights into a firm's prospects.

In the research conducted by *Schechter and Li*, the growing practice of using supervised machine learning (SML) algorithms to extract structured information from unstructured data, such as text or images, is examined. These derived variables from SML are often incorporated into regression models for making inferences and testing theories. However, these SML-generated variables typically introduce measurement errors, compared to the underlying constructs they represent. To mitigate the negative impact of these errors and produce less biased coefficient estimates while enhancing the accuracy of hypothesis testing, the authors propose a novel approach involving robust optimization. This method is particularly relevant in the generalized research context where SML algorithms measure a flexible number of dependent

and independent variables. The authors combine recent robust optimization techniques to fit a linear regression model in the presence of uncertain measurement errors, theoretically demonstrating the consistency and efficiency of this robust approach. Their findings are further validated through simulations, highlighting the method's effectiveness.

*Grote et al.*, in their study, challenge the traditional machine classification approach, which assumes complete knowledge of all classes during training. This assumption often does not hold, particularly in fast-changing environments and safety-critical applications like self-driving cars or tumor detection. Instead, they introduce the concept of open set recognition, which acknowledges the presence of incomplete knowledge about classes during training and the potential emergence of unknown classes during testing. Crucially, the study simulates an open-set scenario on four well-established datasets and demonstrates how Open Set Nearest Neighbor classification can be enhanced through metric learning. Their findings suggest that the prior application of the Large Margin Nearest Neighbor algorithm consistently improves classification results and strengthens the ability to reject unknown instances—a vital aspect in scenarios with numerous unknown classes. These results underscore the significance of metric learning and provide a benchmark for further exploration at the intersection of metric learning and open set recognition, with valuable implications for applications in rapidly changing and safety-critical domains.

In the study by *Mahdavi and Carvalho*, the focus is on the potential of machine learning-based techniques to enhance the extraction of valuable insights from data, aiding businesses in making informed decisions. The challenge lies in the conventional closed-set scenario, where most techniques assume identical label spaces for training and test sets, which often do not reflect the complexities of real-world scenarios. To bridge this gap, the study delves into open set recognition (OSR), aiming to bring classification closer to reality. OSR involves classifying known classes while effectively handling unknown classes, which is vital for addressing the practical constraints of natural dynamic environments. Training a model to account for all possible examples of unknown items is often prohibitively expensive, leading to potential failures when the model is tested in real-world scenarios. The study introduces an algorithm that explores a novel representation of the feature space to enhance classification in OSR tasks. By integrating OSR, businesses can improve the efficacy and efficiency of their processes and decision-making, leading to more

precise and insightful predictions of outcomes. The study's performance on three established datasets demonstrates that the proposed model outperforms baseline methods, highlighting its potential for delivering more accurate and valuable insights to support business decisions.