**Peer response: Machine Learning Model Use in Risk Management**

Natali addresses the use of machine learning (ML) for financial fraud detection. She focuses first on supervised ML, namely by employing ensemble learning methods using gradient boosting machines (GBMs). GBMs are indeed effective at predicting instances of fraud (Botchey, Qin and Hughes-Lartey, 2020; Zhao and Bai, 2022). However, GBM training can be particularly demanding in terms of time and computing power. They are also prone to overfitting, and may suffer from poor interpretability (Natekin and Knoll, 2013; Russel and Norvig, 2021).

Some of the proposed challenges of supervised ML are the need for investments of money and time to develop specific infrastructure, which hinder swift implementation of new algorithms. One should add the complexity of integrating AI within legacy systems, and the difficulty in training existing staff to understand this technology and integrating it in their workflows (Davenport and Ronanki, 2018). Nonetheless, it should be noted that these are wider challenges of implementing AI in any scenario, not specifically of supervised ML.

Natali then describes unsupervised learning approaches, for example self-organising maps (a subtype of artificial neural networks) (Yadav, Ramu and Deb, 2023). In this case, unsupervised ML would group together transactions with similar features, without explicit legitimate/fraudulent labels, and outliers flagged as requiring investigation. This paradigm could help address the challenges mentioned with obtaining appropriately categorised data (as required for supervised ML). Another potentially useful approach is semi-supervised learning, where models are initially trained on a small subset of labelled data and refined on larger sets of unlabelled data (Russel and Norvig, 2021; Bergmann, 2023).

However, she raises doubts about the value of unsupervised ML given the need to identify the specific data points related to a fraud prediction (namely individual people or transactions). This seems to reflect a concern about model interpretability, or the need for data anonymisation, rather than its unsupervised nature. Self-organising maps are indeed limited by poor interpretability, but the same is true of some supervised models (such as GBMs). Moreover, both models have been the subject of recent developments to provide explainable outputs (Konstantinov and Utkin, 2020; Yadav, Ramu and Deb, 2023). Finally, pseudonymisation approaches can be employed so that the model is trained on pseudonymised identifiers, with identifiable data kept in a separate table (with restricted access). Unsupervised learning approaches may therefore prove valuable for fraud detection, possibly within wider risk-management platforms also leveraging supervised and semi-supervised models.

**References:**

Bergmann, D. (2023) *What is semi-supervised learning?*, *IBM*. Available from: https://www.ibm.com/topics/semi-supervised-learning#:~:text=Contributors%3A%20Dave%20Bergmann-,What%20is%20semi%2Dsupervised%20learning%3F,for%20classification%20and%20regression%20tasks. (Accessed: 12 June 2024).

Botchey, F.E., Qin, Z. and Hughes-Lartey, K. (2020) ‘Mobile Money Fraud Prediction—A Cross-Case Analysis on the Efficiency of Support Vector Machines, Gradient Boosted Decision Trees, and Naïve Bayes Algorithms’, *Information*, 11(8), p. 383. Available from: https://doi.org/10.3390/info11080383.

Davenport, T.H. and Ronanki, R. (2018) ‘Artificial Intelligence for the Real World’, *Harvard Business Review* [Preprint]. Available from: https://hbr.org/2018/01/artificial-intelligence-for-the-real-world (Accessed: 6 June 2024).

Konstantinov, A.V. and Utkin, L.V. (2020) ‘Interpretable Machine Learning with an Ensemble of Gradient Boosting Machines’. arXiv. Available from: https://doi.org/10.48550/ARXIV.2010.07388.

Natekin, A. and Knoll, A. (2013) ‘Gradient boosting machines, a tutorial’, *Frontiers in Neurorobotics*, 7. Available from: https://doi.org/10.3389/fnbot.2013.00021.

Russel, S. and Norvig, P. (2021) *Artificial intelligence: A modern approach*. Global Edition. Pearson Higher Education.

Yadav, D., Ramu, P. and Deb, K. (2023) ‘Interpretable self-organizing map assisted interactive multi-criteria decision-making following Pareto-Race’, *Applied Soft Computing*, 149, p. 111032. Available from: https://doi.org/10.1016/j.asoc.2023.111032.

Zhao, Z. and Bai, T. (2022) ‘Financial Fraud Detection and Prediction in Listed Companies Using SMOTE and Machine Learning Algorithms’, *Entropy*, 24(8), p. 1157. Available from: https://doi.org/10.3390/e24081157.