Bad Data – does it really kill off AI and machine learning?

Z Pokrajcic[1], *P Stewart*[2]

1. Technical Director, Petra Data Science, Brisbane, Queensland, zpokrajcic@petradatascience.com
2. CEO, Petra Data Science, Brisbane, Queensland, pstewart@petradatascience.com

Keywords: data quality, digital twin, digital mine to mill, bad data, big data

ABSTRACT

With the availability of large volumes of data and extremely fast computing capability together with easy access to latest digital techniques using open source libraries, digital approaches to plant optimisation and orebody understanding using AI and machine learning are becoming common place. However, there is a persistent perception in industry that mining data is lacking and of low quality therefore will generate ineffective and bad prediction models. The concerns include insufficient data (missing data), data of poor quality (erroneous fields and strings) or data that is just wrong (values are not accurate). This is compounded by the fact that data is located and stored in many disparate systems and databases.

Fortunately, purpose built automated mathematical algorithms have been deployed across numerous applications to address these issues and include:

- Software to automatically clean data by automatically removing erroneous entries such as #NA or blank fields
- Calibration and autocorrection of key online sensors such as those coming from on stream analyses (OSAs) and particle size indicators (PSIs) with truth data, typical from control laboratory samples
- Monitoring online sensor for anomalies to trigger real-time alerts and sensor health for model inputs
- Automated real-time model switching contingent upon sensor health for model inputs
- Model confidence including real-time display of model output in DCS for continuous visibility of model performance
- Data valorisation using multi-sensor cross validation

Most importantly, it is pointless changing input data or fixing it for the purpose of building a model if when the model is deployed to operations it is exposed to poor quality data. Data cleaning and correction functions, such as those outlined here, need to be built into the software and must be able to handle raw data feeds in real-time.

It is true that raw mining data can be of poor quality but there are proven methods currently operational where these data issues are corrected and addressed in near-real-time to ensure AI and machine learning integrity.