# Data quality issues leading to sub optimal machine learning for money laundering models

Abhishek Gupta
*Department of Management, Bharathidasan Institute of Management, Tiruchirappalli, India*

Dwijendra Nath Dwivedi
*Department of Economics and Finance, UEK, Krakow, Poland and Department of Development, IGIDR, Mumbai, India*

Jigar Shah
*Department of Management, Narsee Monjee Institute of Management and Higher Studies, Mumbai, India, and*

Ashish Jain
*Indian Institute of Management Lucknow, Lucknow, India*

## Abstract

**Purpose** – Good quality input data is critical to developing a robust machine learning model for identifying possible money laundering transactions. McKinsey, during one of the conferences of ACAMS, attributed data quality as one of the reasons for struggling artificial intelligence use cases in compliance to data. There were often use concerns raised on data quality of predictors such as wrong transaction codes, industry classification, etc. However, there has not been much discussion on the most critical variable of machine learning, the definition of an event, i.e. the date on which the suspicious activity reports (SAR) is filed.

**Design/methodology/approach** – The team analyzed the transaction behavior of four major banks spread across Asia and Europe. Based on the findings, the team created a synthetic database comprising 2,000 SAR customers mimicking the time of investigation and case closure. In this paper, the authors focused on one very specific area of data quality, the definition of an event, i.e. the SAR/suspicious transaction report.

**Findings** – The analysis of few of the banks in Asia and Europe suggests that this itself can improve the effectiveness of model and reduce the prediction span, i.e. the time lag between money laundering transaction done and prediction of money laundering as an alert for investigation

**Research limitations/implications** – The analysis was done with existing experience of all situations where the time duration between alert and case closure is high (anywhere between 15 days till 10 months). Team could not quantify the impact of this finding due to lack of such actual case observed so far.

**Originality/value** – The key finding from paper suggests that the money launderers typically either increase their level of activity or reduce their activity in the recent quarter. This is not true in terms of real behavior. They typically show a spike in activity through various means during money laundering. This in turn impacts the quality of insights that the model should be trained on. The authors believe that once the

Journal of Money Laundering Control
Vol. 25 No. 3, 2022
pp. 551-555
Emerald Publishing Limited
1368-5201
DOI 10.1108/JMLC-05-2021-0049

financial institutions start speeding up investigations on high risk cases, the scatter plot of SAR behavior will change significantly and will lead to better capture of money laundering behavior and a faster and more precise "catch" rate.

**Keywords** Data quality, Anti money laundering, Machine learning, Model efficiency

**Paper type** Research paper

## 1. Introduction

The application of machine learning in money laundering has been an important topic of discussion. Jullum *et al.* (2020) have documented their approach to money laundering detection through machine learning. Starting from simple predictive algorithms to sequence matching for anomalic transaction identification, Liu *et al.* (2008). During the review of various machine learning mechanisms for money laundering, Chen *et al.* (2018) also mentioned various other techniques often used including link analysis, behavioral analysis and others.

However, techniques can only take you as far as your data can provide insights. The impact of data quality on model quality has not been a well-documented topic. However, in our experience, the most value is generated from quality data. Hence, the majority of the advisory firms, McKinsey, Deloitte or others, stress on the data quality in the context of artificial intelligence in anti-money laundering (AML) operations.

In this paper, we focus on one very specific area of data quality, the definition of an event, i.e. the suspicious activity reports (SAR)/ suspicious transaction report. For developing any model, there are three important characteristics that a modeler needs to freeze:

1. definition of event;
2. definition of observation window and performance window; and
3. definition of predictors.

In our discussions, depending on the number of SAR cases available within a bank, the definition of events has generally been a filed SAR case or filing of an internal SAR. In few cases, the definition has stretched to include created cases for investigation as a proxy for defining an event in the machine learning model context.

In these cases, the definition of the observation window becomes critical. Typically, a modeler will rely on the case close date, i.e. the date on which the case completed the investigation and the SAR was filed to track the transaction behavior prior to that of the model predictions. This is where we identified the specific problem on data quality.

## 2. Methodology and approach

The team analyzed the transaction behavior of four major banks spread across Asia and Europe. Based on the findings, the team created a synthetic database comprising 2,000 SAR customers mimicking the time of investigation and case closure.

The team also made it a point to understand the SAR behavior of only those customers who are not filed due to adverse media or at the central bank's request. The reason for eliminating them is simple; a customer might be a money launderer based on the other bank's transaction. The customer might be a completely normal customer in another bank. Analyzing these kinds of customers can bias the analysis and, hence, can help in cherry-picking the right customer profiles for analyzing the customers.

For machine learning models, the team typically choses 3–6 months for an observation window. This is ideal as it provides enough time frame to observe the "normal" behavior of a customer. Any abnormality in behavior can also be easily observed in an ideal scenario. It fulfills the criteria of recency and sufficient time frame to observe relevant customer transaction behavior. However, some of the glaring facts the team have observed during model development are as follows:

- There are 15%–20% inactive SAR customers (depending on the segment of customers) during the observation window, i.e. the sum total of credits and debits across transaction modes for 6 months period is 0. It suggests that the investigation continued for over 6 months after the customer completed suspicious transactions.
- The cash and wire trends of the customers are shown in Figure 1.

Figure 2 shows two spikes for segment 1 – one spike at the beginning of observation window (26 weeks prior) and another one at around week 15–18. The spike in activity for segment 2 is around week 13–15 and then another spike in week 17–20. Either way – the immediate past weeks transaction behavior is a downward slope/flat line, suggesting that the model would never be trained properly 4–6 weeks prior to the case close date.

Similar observations on the above exhibit shows spikes at different points in time going up to 26 weeks prior to case closure.

Authors expected a typical higher transaction spike in the last 4–6 weeks because as per the bank's compliance teams, the investigation on alerts starts within a week of the alert generation. Also, there are typically turn around times for case closures. However, from the chart, it is clear that the investigations are spread over a much larger time frame. Second, this is not the case for a negligible number of SAR customers. The phenomenon is observed for a fairly large number of SAR customers (Figure 3).

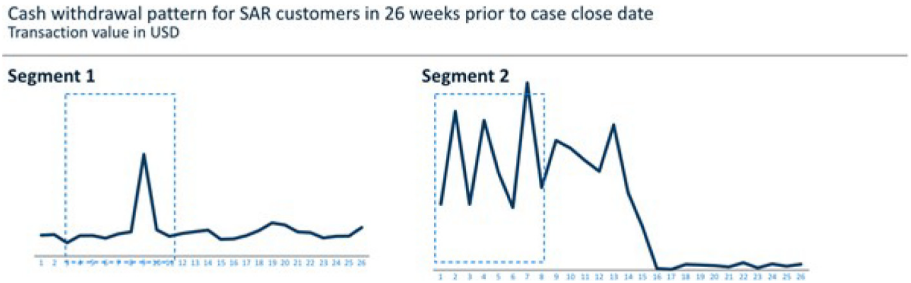

**Figure 1.** Cash withdrawal patterns for SAR customers showing differing behavior – long back in the history

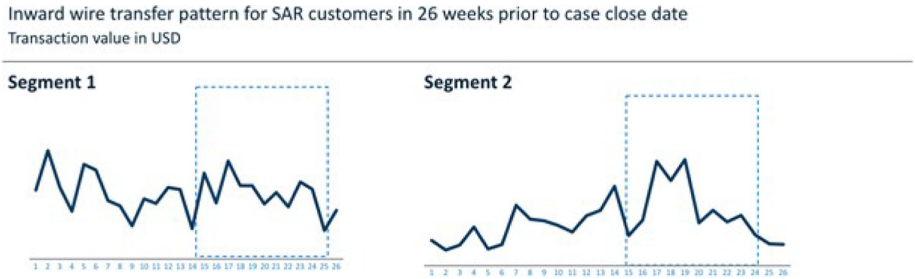

**Figure 2.** Wire-in transactions of the SAR customers

## 3. Conclusion

Given the large variation in transaction behaviors, driven by delayed case closures, the training of the models is also biased. Hence, in few cases, it would be giving a delayed warning signal – much later than the time when money laundering activity happened. Typically, the percentage change in activity between historical and recent past is intuitively shown in Figure 4.

The above graph suggests that the money launderers typically either increase their level of activity or reduce their activity in the recent quarter. This is not true in terms of real behavior. They typically show a spike in activity through various means during money laundering. We expect the velocity of a spike to be 2–3 weeks maximum, depending on the segment. However, the delayed filing and consequent delayed case close date changes the time horizon. This in turn impacts the quality of insights that the model should be trained on.

One of the ways which it can be handled is through dynamic analysis of time when the transaction spike ends and logically define that at the beginning of the performance

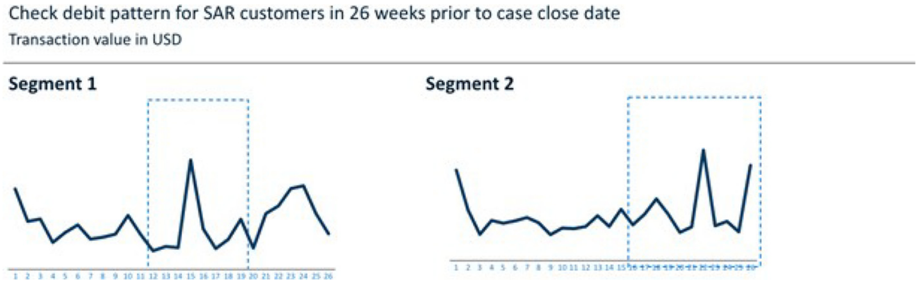

**Figure 3.**
Check debit patterns for two segments

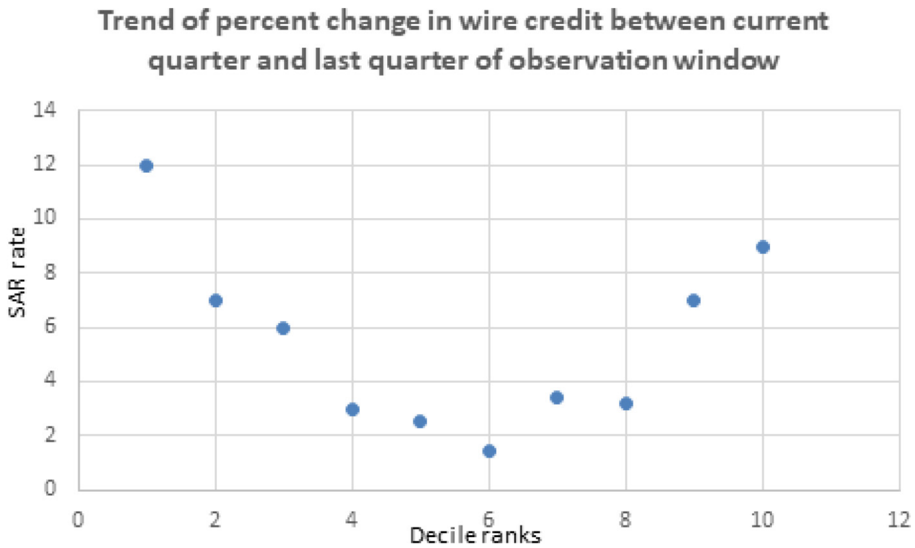

**Figure 4.**
Percent change inactivity between historical and recent past

window. However, this runs the risk of biasing the model and could also result in overfitting (self-fulfilling prophecy of observed behavior).

We believe that once the financial institutions start speeding up investigations on high risk cases, the scatter plot of SAR behavior will change significantly and will lead to better capture of money laundering behavior and a faster and more precise "catch" rate.

## References

Chen, Z., Van Khoa, L.D., Teoh, E.N., Nazir, A., Karuppiah, E.K. and Lam, K.S. (2018), "Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review", *Knowledge and Information Systems*, Vol. 57 No. 2, pp. 245-285, doi: 10.1007/s10115-017-1144-z.

Jullum, M., Løland, A., Huseby, R.B., Ånonsen, G. and Lorentzen, J. (2020), "Detecting money laundering transactions with machine learning", *Journal of Money Laundering Control*, Vol. 23 No. 1, pp. 173-186, doi: 10.1108/JMLC-07-2019-0055.

Liu, X., Zhang, P. and Zeng, D. (2008), "Sequence matching for suspicious activity detection in anti-money laundering", in Yang, C.C., Mao, W., Zheng, X. and Wang, H. (Eds), *Intelligence and Security Informatics*, ISI 2008, Lecture Notes in Computer Science, Vol 5075, Springer, Berlin, Heidelberg, doi: 10.1007/978-3-540-69304-8_6.

## Further reading

available at: www.mckinsey.com/industries/financial-services/our-insights/banking-matters/the-aml-industry-in-2019

Gupta, A., Dwivedi, D.N. and Jain, A. (2021), "Threshold fine-tuning of money laundering scenarios through multi-dimensional optimization techniques", *Journal of Money Laundering Control*, doi: 10.1108/JMLC-12-2020-0138.

## Corresponding author

Dwijendra Nath Dwivedi can be contacted at: dwivedy@gmail.com