Irving Fisher Committee on
Central Bank Statistics

BIS

IFC-Bank of Italy Workshop on "Machine learning in central banking"

19-22 October 2021, Rome / virtual event

A novel machine learning-based validation workflow for financial market time series [1]

Magdalena Erdem and Taejin Park,
Bank for International Settlements

[1] This presentation was prepared for the Workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the central banks and other institutions represented at the event.

A novel machine learning-based validation workflow for financial market time series[1]

Magdalena Erdem and Taejin Park[2]

Abstract

The size and complexity of data managed by central banks have been increasing providing them with opportunities to make evidence-based policy decisions. As a result, data validation has become a more challenging task. In this paper, we propose a highly automated validation workflow that outperforms traditional approaches and is suitable for a large volume of financial market time series, based on machine learning algorithms. Using some real-life examples, we illustrate how machine learning (ML) algorithms can help address key challenges, such as understanding the context of various financial instruments and dynamically coping with constantly evolving market environments.

Keywords: machine learning, data science, data validation; financial market time series

JEL classification: C55, C80, G17, O33

[1] The views expressed in this article are those of the authors and do not necessarily reflect those of the Bank for International Settlements.

[2] Respectively, Head of Departmental Research Support, BIS (Magdalena.Erdem@bis.org); and Head of Financial Markets and Research Support, BIS (Taejin.Park@bis.org)

Introduction

Data validation is a key accountability of statistics teams in central banks. With more data available for central banks to make evidence-based policy decisions, the amount of data managed by central banks have been increasing accordingly. As a result, data validation has become a more challenging task, especially with large volumes of data, such as financial market series that are usually available at high frequencies – daily or intraday. Validating such large financial market data can requires significant resources so it is often outsourced to data providers which requires trust in the data they provide. Accepting third-party data without due verification could lead to high operational risks.

While the need to guarantee high quality data for policy makers is being given more importance, there has been little research focused to central banks' (or financial authorities') toolkits for checking financial market time series. Eurostat publishes a comprehensive data validation framework focusing on official statistical agencies (Eurostat, 2018). Among central banks, the European Central Bank (ECB) provides a framework, mainly covering governance and data quality dimensions and metrics, and an additional list of data quality checks on supervisory reporting data (ECB; Hogan, 2017). The US Board of Governors of the Federal Reserve System (Federal Reserve Board) has developed Information Quality Guidelines for data quality, objectivity, utility and integrity as required by the US Office of Management and Budget. The Bank of England offers its data quality framework to enable data users to be informed about the quality of the data they use (Bank of England, 2014). Such frameworks can be very useful to understand the various aspects of data validation principles. However, complementing them with more detailed practices and examples can further help statisticians address their day-to-day challenges.

Meanwhile, in the machine learning community, research has been conducted in detecting anomalies and outliers in financial market data (Au Yeung et al, 2020; Ahmed et al, 2016; Golmohammadi and Zaiane, 2015; Ferdousi and Maeda, 2006). Such research aims mostly at detecting fraud or informing investment decisions. Whilst this can provide useful insights for developing data validation techniques specific to financial market data, it is not directly applicable to *error* detection or confirming *true outliers* (ie seemingly suspicious but actually correct data points).

To fill the research gap in practical guidance for validating large volumes of high-frequency financial market series, this paper proposes a solid data validation workflow that would allow for full automation requiring low maintenance costs.

Recent developments in data science provide great opportunities for central banks. Among many, machine learning techniques have been developing quickly in recent years. Thanks to its popularity, machine learning is also now much more accessible. There are free software packages for machine learning analysis (eg Python and R), active online forums (eg Stack Overflow) and open source off-the-shelf code libraries (eg Scikit-learn, TensorFlow and Keras). In addition, advancing computing capacity enables statisticians and data scientists to solve complex algorithms. In particular, big data platforms, GPU units and cloud computing can significantly boost efficiency and broaden the scope of machine learning analysis.

Leveraging on these opportunities, this paper proposes a highly *automated* validation workflow that *outperforms* traditional approaches and is suitable for a large volume of *financial market time series*. The main objective of our analysis is to develop an end-to-end workflow for data validation, with examples of step-by-step machine learning applications. We do not intend to propose any single specific machine learning model as individual circumstances will predominantly determine model selection.

Challenges with traditional validation approaches

To better understand the requirements of a good validation tool, we first reviewed the most commonly used traditional validation methods – graphical method, conditional controls, threshold-based warnings and cross-referencing. These traditional methods enable some degree of automation but still require frequent human intervention, making them unsuitable for large scale validation processes (Table 1). Validation of even a small number of financial market series, taking account of the context of various financial instruments in different market segments, can be time-consuming. Furthermore, financial market environments are continuously evolving and often entail structural changes in the market due to central bank policy actions (eg policy rate changes or quantitative easing), financial conditions and new or outdated instruments. Keeping up with such changes in diverse financial market segments can be very labour-intensive.

Overview of common traditional validation approaches

Table 1

	Description	Limitations
Graphical method	Visual inspection of time series graphs to detect any anomalies	Time consuming; vulnerable to oversight; difficult to validate plausible but erroneous data points or true outliers
Conditional controls	Implementing pre-defined "If-Then" controls	Requires a good understanding of market contexts; setting appropriate parameters for controls/thresholds to many heterogeneous series is challenging; becomes ineffective in case of any structural changes or turmoil in the market
Threshold-based warnings	Setting certain thresholds to give warnings, for example, based on percentage changes or z-scores. Usually used together with conditional controls.	
Cross-referencing	Cross checking with the same series from another source or other related series to see there's any divergence in their relationship	Requires a good understanding of market contexts to determine appropriate reference series to validate the target series; alternative source might not be available; especially challenging when validating many series in different market segments.

Source: Authors' elaboration.

Graph 1 illustrates some of the challenges in validating financial market time series in practice. The daily time series of the preliminary euro short-term rate (Pre-€STR) exhibits various anomalies at a glance (left panel). If the series is reviewed by any graphical methods and/or threshold-based controls, it is likely giving warnings for the spike each quarter and the sudden break in September 2019. However, the real error in this series is the repeating values at the end of the series when the instrument is discontinued on the official launching of €STR on 2 Oct 2019. Since the legacy preliminary series was treated as a separate instrument by its data provider, instead of being replaced by the official one, the last quoted value kept appearing in the following days. That repeated values is not unusual in some illiquid financial markets can make a statistician consider the data plausible, potentially leading to a false positive error (ie Type I error).

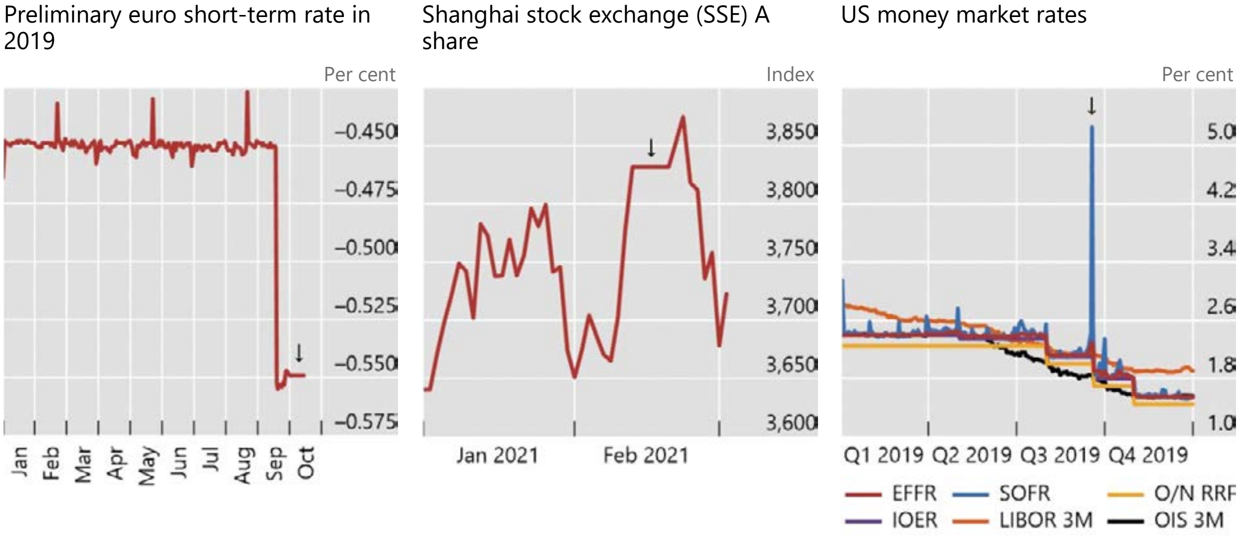
A similar pattern is observed in the Shanghai stock exchange (SSE) A share index where a value is repeated for eight straight days in mid-February 2021 (middle panel). As opposed to the previous case, however, the long repetition is what happens in the market during Chinese (Lunar) New Year holidays, which fall on different dates in January or February of the Gregorian calendar each year.

These two examples show similar data patterns in appearance but require different actions, indicating that data validation necessitates a good understanding of the market context. Those who have sufficient knowledge about euro money markets would know that such quarterly spikes and a break in series due to a policy rate change are common. Similarly, those who are familiar with Chinese markets would not regard the long repetition as an error. However, keeping up with many financial instruments in diverse market segments is challenging.

The right panel of Graph 1 shows the US money market rates in 2019. Immediately noticeable is the striking outlier of the Secured Overnight Financing Rate (SOFR) in Q3 2019. Again, any outlier detection controls in place will likely give warnings given the serious magnitude of change in a day (ie Type II error, (false negative error)). Even for an expert in money market instruments, such a spike in a secured interbank overnight interest rate so much above other unsecured rates would be perceived as an extremely rare event. To validate this outlier, one can compare it with other US secured interbank rates that usually follow similar trends, such as the Broad General Collateral Rate (BGCR) or the Tri-Party General Collateral Rate (TGCR). Since the other secured rates appear to have similar shocks on the same date, the outlier can eventually be confirmed as a true value.

Traditional validation methods

Graph 1


Preliminary euro short-term rate in 2019

Shanghai stock exchange (SSE) A share

US money market rates

Sources: Bloomberg; Refinitiv.

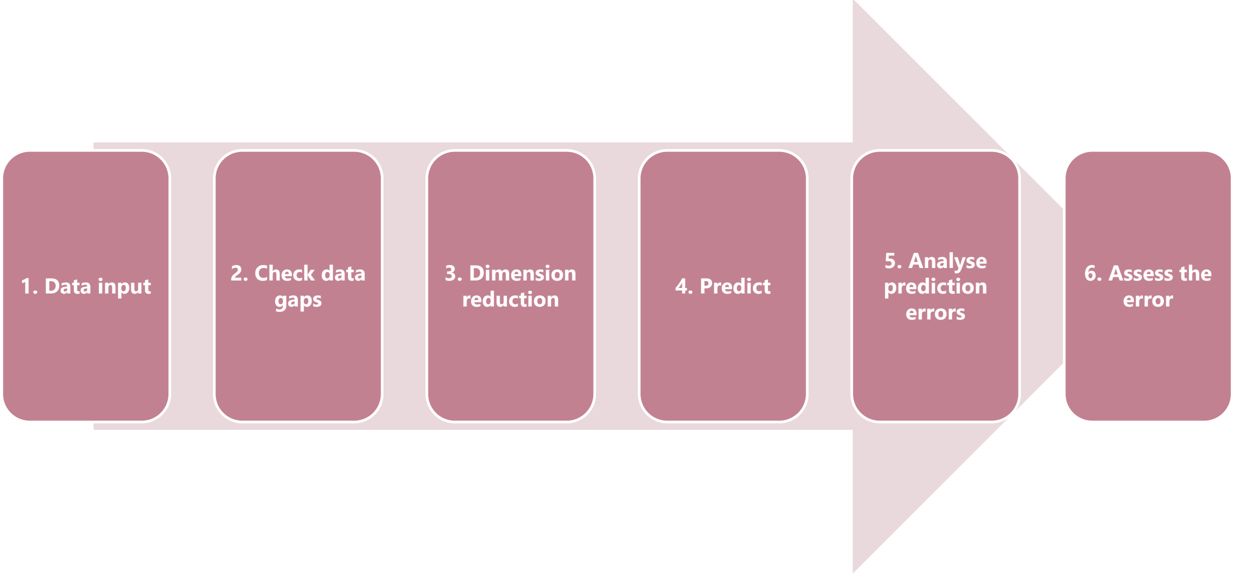
One might argue that such false negative errors are still useful to maintain alertness and in fact, receiving false negative errors for suspicious data points is better than missing any true errors. However, in the long term frequent false negative errors can make statisticians become accustomed to the warning messages and thus ignore them without investigation. Reducing false negative errors is also therefore highly important to manage operational risks. To summarise, a good data validation system should set a precise threshold that is free from both false positive and false negative errors.

Proposed data validation workflow using machine learning models

The challenges illustrated in the previous section highlight the main requirements of a good data validation model for financial market series. Without a good understanding of the context of each instrument, those common traditional methods can produce frequent Type I or II errors. An effective data validation model should imitate the behaviours of a subject matter expert who is familiar with the unique characteristics of the instrument and keeps up with recent market development. Therefore, analysing the underlying processes of a human expert validating a financial market series can provide insight for developing an optimal validation model.

A human expert can access all readily available information and identify only the relevant information for a specific context. They can then make a best judgement to assess if a given value is within a reasonable range according to the key information collected. In statistical terms, the process can be translated into: input data collection; filter useful explanatory data (or reduce dimensions); and then assess the target value if it is within the range predicted by the explanatory data. Based on this intuitive approach to the validation, we propose a model workflow of financial market data validation (Graph 2). We predict any data point to validate as if it is invisible to us for the date. The prediction fully makes use of many other financial market time series including the date where the target series is assumed invisible. This workflow uses machine learning models in various steps to minimise human intervention. In other words, the process is highly automated for scalability and dynamic adaptation to evolving market environments. Once implemented, the workflow can continuously run every day to detect anomalies. In the remainder of this section, we will provide more details about each of the six steps with example machine learning applications for the anomaly cases shown in the previous section.

Stylised workflow of data validation process using machine learning models Graph 2



Source: Authors' elaboration.

1. Data input

For the purpose of the illustration, we use about 3,000 daily incoming financial market time series, covering a variety of market segments as input data. The aim is to develop a workflow that is suitable for validating any or all of them. Additionally, the 3,000 series can be used as input to validate any specific series within the sample.

These input data have the typical financial market time series characteristics that can pose challenges to effective data validation. First, the size of dataset is so big that manual validation processes can't be applied. Second, as previously illustrated, anomalies are not always easily detectable when analysing them in isolation. Third, they show frequent structural changes within a data series, which makes it difficult to fit any specific model that can reflect such changes dynamically. Lastly, the dataset covers diverse market segments (eg equity, interest rates, FX, credit, commodities, etc) for around 70 geographical markets, making it resource intensive to track and monitor all market developments.

However, having such a large financial market dataset can also provide an important opportunity for data validation. A financial market series usually has strong explanatory series. For example, there can be several equity indices that are tracking the same market. Yields of instruments with similar maturities in the same market show similar trends. Exchange rates that are pegged to the same currency also move together. This is a valuable feature of financial market time series that can enable precise predictions.

2. Checking data gaps

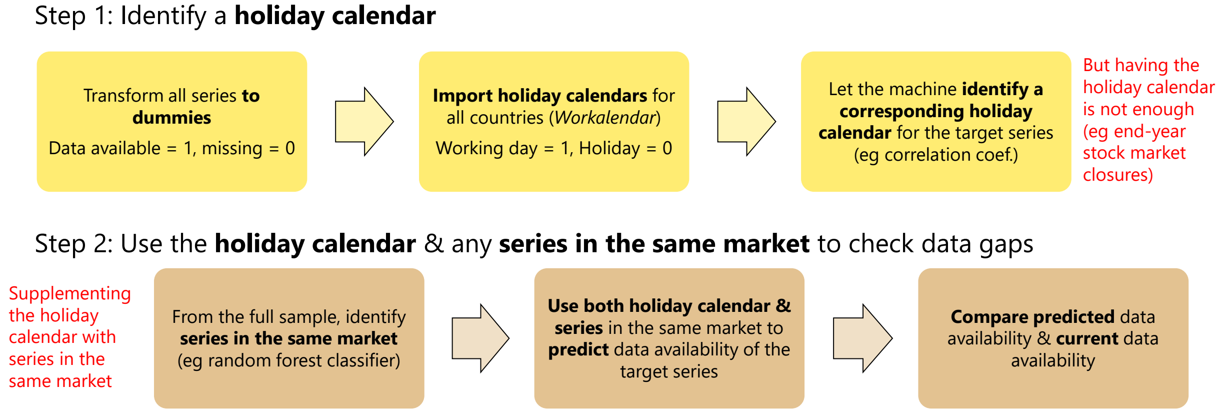
Data input should be followed with a check of data availability in the target series for a recent date (or any other date of one's interest). Missing data in the financial market series is usually due to market closure. Most financial markets follow the holiday calendar of the country in operation. Therefore, detecting any erroneously missing data can start with cross-checking the data with its corresponding holiday calendar. However, checking against holiday calendars alone is insufficient because of exceptions. For example, some FX markets are open every day including weekends and holidays. Some stock exchanges are exceptionally closed on a few non-holidays. For more accurate validation, supplementing holiday calendars with additional information is essential. Financial instruments traded in the same exchange usually follow the same opening and closing schedule. Hence data availability of a financial market series can also be assessed by cross-checking with another financial market series traded in the same market (eg Nasdaq & S&P500). If two financial market series historically show similar availability patterns, data availability of one series can be predicted by reference to the other series. Based on the two datasets, holiday calendars and series in the same market, this paper proposes a two step approach for checking data gaps in the target series.

The first step automatically identifies a holiday calendar for the target time series. As summarised in Graph 3, all input series first need to be re-coded as binomial variables to only indicate data availability. Holiday calendars are imported also as binomial variables. The holiday calendars for most countries are retrieved from a Python module (*workcalenda*[3]). Then, a simple algorithm (eg correlation coefficients) can identify the most relevant holiday calendar for the target series.

The second step is to identify any series in the same market. From the 3,000 financial market series in binomial terms, a machine learning algorithm (eg random forest classifier) can identify any financial market series that are likely traded in the same market as the target series. Once they are identified, together with the holiday calendar, they can collectively be used to predict today's data availability of the target series. For the prediction, a similar algorithm can be applied as in the identification of the series in the same market. The predicted data availability then can be used to evaluate the recorded data availability.

Two step approaches to check data gaps

Graph 3


Step 1: Identify a **holiday calendar**
Transform all series **to dummies**
Data available = 1, missing = 0
Import holiday calendars for all countries (*Workalendar*)
Working day = 1, Holiday = 0
Let the machine **identify a corresponding holiday calendar** for the target series (eg correlation coef.)
But having the holiday calendar is not enough (eg end-year stock market closures)
Step 2: Use the **holiday calendar** & any **series in the same market** to check data gaps
Supplementing the holiday calendar with series in the same market
From the full sample, identify **series in the same market** (eg random forest classifier)
Use both holiday calendar & series in the same market to **predict** data availability of the target series
Compare predicted data availability & **current** data availability

Source: Authors' elaboration.

[3] Workalendar Maintainers (2021), https://github.com/workalendar/workalendar

3. Dimension reduction

Dimension reduction is considered to be a common step in machine learning analysis to better fit a model to improve prediction power and to estimate the model more efficiently. Since our analysis used around 3,000 input variables, filtering irrelevant variables and focusing only on a small set of series that best explain the target series is imperative.

In this paper, dimension reduction is done in two steps. First, largely irrelevant variables are filtered following a simple traditional algorithm, that is, correlation coefficients. After this filtering, only dozens of series that are somewhat related to the target series can be taken forward for further dimension reduction (or feature selection). In the second step, we apply a random forest regression model to leave few key features that will become explanatory variables for a prediction model in the next step of the validation workflow.

Graph 4 shows results of an application of the dimension reduction processes to the challenging cases discussed at the beginning of the paper. For the SSE A share index, the model returned the SSE composite index as a dominant feature to explain the target series (left panel). The SSE composite index consists of both A shares and B shares that are traded in the SSE. Since market capitalisation of B shares takes only 0.2%[4] of that of A shares, the composite index should move together with the SSE A share index in a highly synchronised way. Therefore, one can precisely predict movement of the SSE A share index if the SSE composite index is known. For the US SOFR, the most important features turned out to be other US secured overnight money market rates, namely the TGCR and the BGCR, followed by unsecured overnight money market rates (right panel). For both the Chinese and US cases, it is interesting to note that the model automatically learns market contexts from the data and returns useful series. If a subject matter expert manually conducted the same exercise, the results would be very similar to what the model selected.

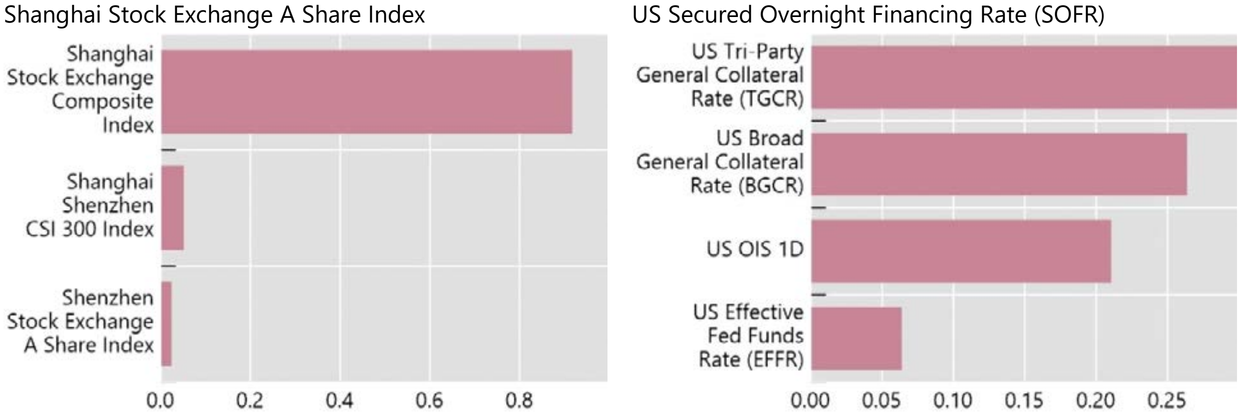
The number of features to be used for prediction can be determined based on the prediction model specifications and specific use cases. A minimum Gini importance cut-off value can be applied to individual series or to top-N series combined. Another approach is to develop any feature selection algorithm, such as backward/forward/recursive eliminations or exhaustive selection. The algorithm tries to find the best combination of features that can maximise the prediction performance of the exact model that will be used for prediction. In this approach, computing capacity is an important consideration.

[4] As of 11 October 2021.

Key features that can best explain the target series

Gini importance

Graph 4


Shanghai Stock Exchange A Share Index

US Secured Overnight Financing Rate (SOFR)

Sources: Bloomberg; Refinitiv; authors' calculations.

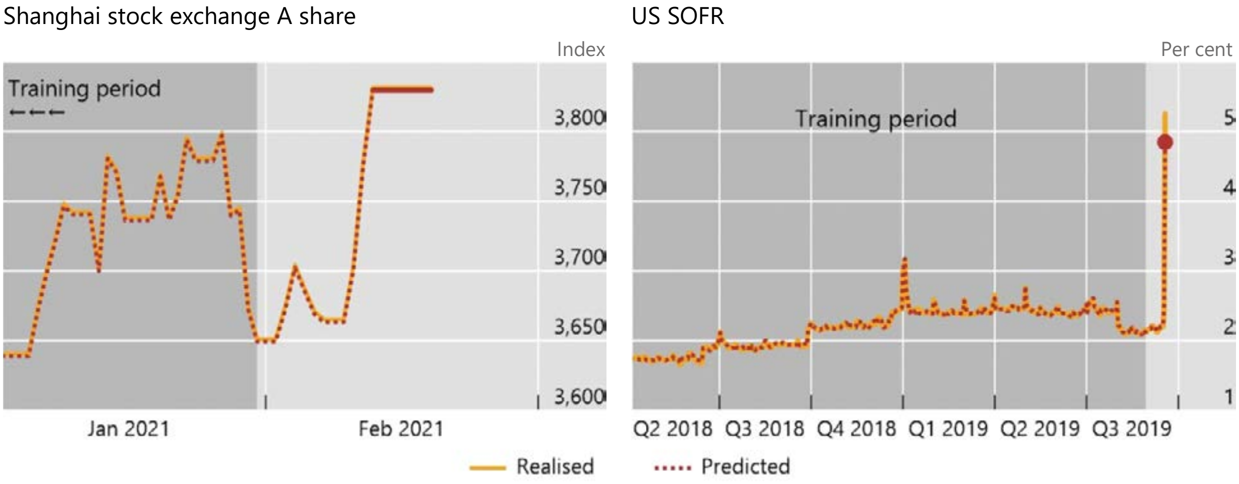
4. Prediction

Using the small set of features selected above, we can fit a machine learning model to predict a value for today or any date of interest for the target series as if the value is unknown. For model selection, we opted for a recurrent neural network (RNN), a deep learning model suitable for sequential data such as time series. More specifically, we used the Long Short-Term Memory (LSTM) model to capture long-term contexts without vanishing gradients and exploding gradients problems (Hochreiter and Schmidhuber, 1997). Once again, the main purpose of this paper is not to propose any specific model that can fit all different cases but rather to illustrate the proposed workflow.

Based on the two anomaly examples of financial market time series – the SSE A share index and US SOFR – we fit the LSTM model to predict values for the suspicious data points that were discussed in the previous section. The out-of-sample prediction results appear to be precise in both cases (Graph 5). Both predictions are based on the same model specifications, except for input data. This is one of the main advantages of using a machine learning algorithm as the machine itself can find a model to best fit input data with minimum human intervention.

Comparison of predicted and realised values

Graph 5



The prediction is based on the long short-term memory (LSTM) model.

Sources: Bloomberg; Refinitiv; authors' calculations.

5. Analyse prediction errors

Once a predicted value is available, the realised value in question needs to be compared to the predicted value. The difference between the predicted and the realised value (ie prediction error) would indicate how much the realised value in question deviates from a reasonable expectation. When the difference is *significantly large* it would be worth investigating further.

An important question is then how to determine whether a prediction error is large enough to suspect erroneous data. One possibility is to set certain thresholds to evaluate a prediction error based on percentage deviations or standard deviations. However, if a series is highly volatile by nature or difficult to predict due to limited feature availability, such a static threshold-based approach can signal frequent Type I or II errors. Therefore, this paper applies an unsupervised machine learning algorithm to decide whether a prediction error is acceptable or not, based on historical patterns of prediction errors in the target series.

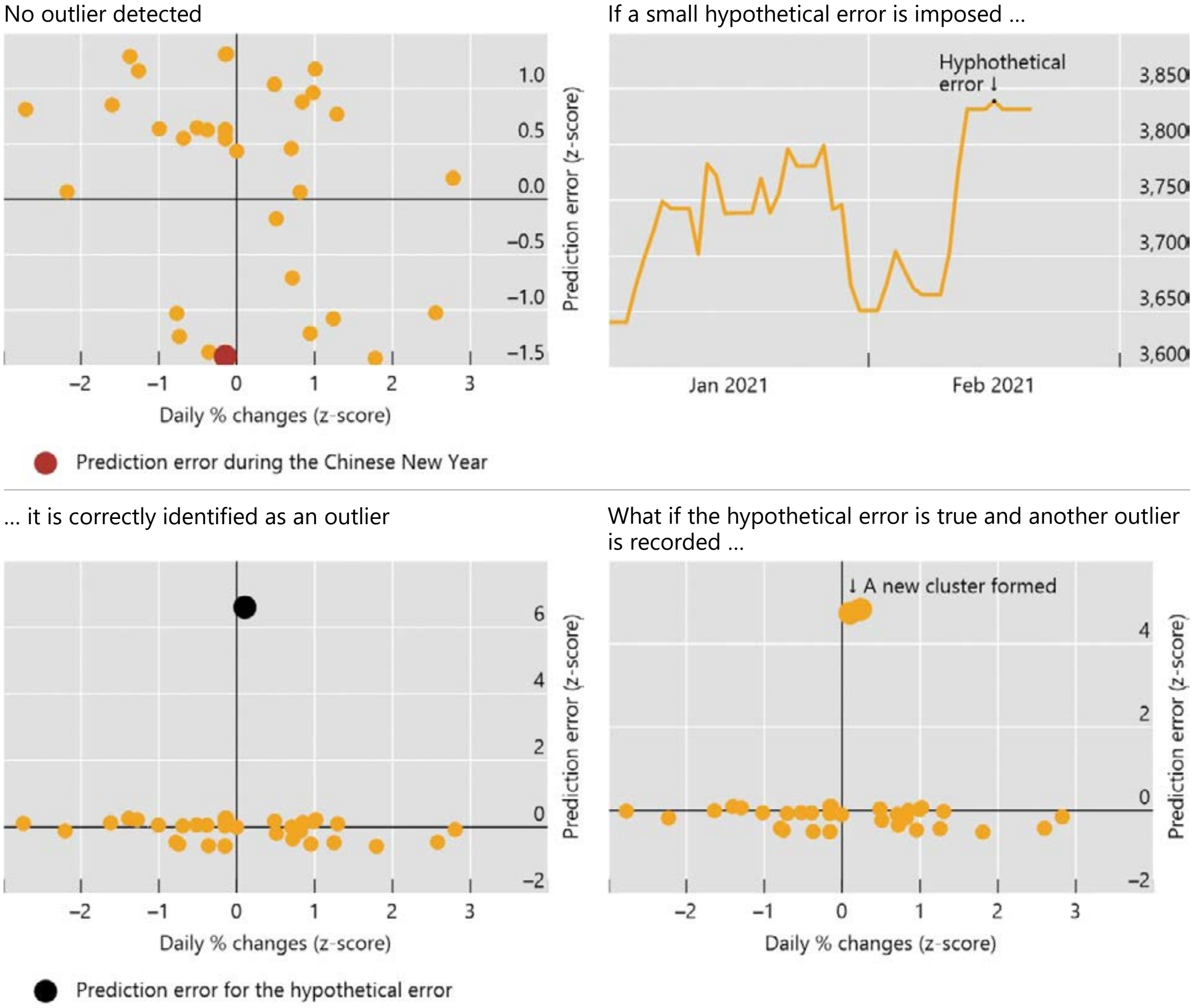
6. Assess the errors

To assess the prediction errors, we applied an unsupervised clustering algorithm, Density-Based Spatial Clustering of Applications with Noise (DBSCAN). DBSCAN requires minimum input parameters and discovers clusters with arbitrary shape, which makes it suitable for a large dataset (Ester et al, 1996).

When DBSCAN is applied to assess prediction errors for the SSE A share index case, the repeating values during the Chinese New Year are identified as non-outliers (Graph 6, top left). To check the robustness of the algorithm, we intentionally recorded a small hypothetical error equal to the average daily percentage change of the original series (top right). Since the time series data were precisely predicted in the previous section, such a small error that looks insignificant in the graph is identified as an obvious error in this algorithm (bottom left). To highlight how this algorithm can cope with a structural change in the market, we assume that the small error turns out to be true and a similar outlier is recorded again on the next date. In this case, the second outlier is not considered as an error anymore because the minimum number of samples to form a new cluster is set to 2 (bottom right). This implies that the model returns a false negative error message when it first encounters a significant structural change in the market, but it can correctly validate any similar types of outlier for the coming days. This is an advantage of machine learning models as they can dynamically learn contexts from data and reflect such structural changes in the market.

Outlier detection based on an unsupervised machine learning algorithm

Graph 6



No outlier detected

If a small hypothetical error is imposed …

… it is correctly identified as an outlier

What if the hypothetical error is true and another outlier is recorded …

The analysis is based on density-based spatial clustering of applications with noise (DBSCAN). The maximum distance between two samples for one to be considered as in the neighbourhood of the other is set to 2 Euclidian distance. The minimum number of samples to form a new cluster is set to 2.

Sources: Bloomberg; Refinitiv; authors' calculations.

Conclusion

While financial market time series data are key inputs to important policy decisions by central banks, little research focuses specifically on data validation processes for them. This paper first reviewed common practices used for time series data validation and their limitations in this emerging data-intensive environment. It then proposed an end-to-end workflow of daily data validation routines to overcome key challenges in ensuring high quality for large financial market time series datasets. While describing each step, we illustrated how machine learning algorithms can help address the key challenges, such as understanding the context of many financial instruments and dynamically coping with constantly evolving market environments. During our analysis, we intentionally focused on a few carefully selected examples to best illustrate key challenges and solutions in each step from central bank practitioners' perspectives.

We would like to reiterate that machine learning techniques are now more accessible than ever, even to non-experts. This provides a great opportunity for data validation work. At the same time, the abundance of models, code libraries and references available can create additional challenges, especially without a clear overview of one's unique business requirements. For this reason, we did not recommend which machine learning model would work best in each situation as such discussion would be less meaningful without a more detailed context of characteristics and materiality of datasets, infrastructure, business environments, etc.

Future work and investigation should focus precisely on describing the suggested methods and their application to specific cases. We hope this paper will provide a stimulating basis for emerging research on this topic to the benefit of the central banking community. The ultimate aim is to build knowledge blocks for more efficient quality assurance of data.

References

Au Yeung, J.F.K., Wei, Zk., Chan, K.Y. et al (2020): "Jump detection in financial time series using machine learning algorithms", *Soft Computation*, no 24, pp 1789–1801

Bank of England Statistics and Regulatory Data Division (2014): *Data Quality Framework*

Board of Governors of the Federal Reserve System: *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by the Federal Reserve Board*, https://www.federalreserve.gov/iq_guidelines.htm

Chollet, F. et al (2015): "Keras", https://github.com/fchollet/keras

European Central Bank: *Additional supervisory data quality checks*, https://www.bankingsupervision.europa.eu/banking/approach/dataqualitychecks/html/index.en.html

Eurostat (2018), *Methodology for data validation 2.0*

K. Golmohammadi and O. R. Zaiane (2015): "Time series contextual anomaly detection for detecting market manipulation in stock market," *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp 1–10

Mohiuddin Ahmed, Abdun Naser Mahmood, Md. Rafiqul Islam (2016): "A survey of anomaly detection techniques in financial domain", *Future Generation Computer Systems*, Volume 55, pp 278–288

Pedregosa et al (2011): "Scikit-learn: Machine Learning in Python", *JMLR,* no 12, pp. 2825–2830

P Hogan (2017): "ECB Supervisory Data Quality Framework, Tools and Products", presented at the Supervisory Reporting Conference, ttps://www.bankingsupervision.europa.eu/press/conferences/shared/pdf/sup_rep_conf/2017/Data_quality_framework_tools_and_products.pdf

Workalendar Maintainers (2021): https://github.com/workalendar/workalendar

Z. Ferdousi and A. Maeda (2006): "Unsupervised Outlier Detection in Time Series Data", *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, pp x121–x121

Irving Fisher Committee on
Central Bank Statistics

BIS

Deep learning as a novel validation tool for financial market time series

IFC and Bank of Italy Workshop on "Data science in central banking", part 1, 19-22 Oct 2021

Taejin Park, Head of Financial Markets and Research Support (FMRS), BIS (presenter)

Magdalena Erdem, Head of Departmental Research Support (DRS), BIS

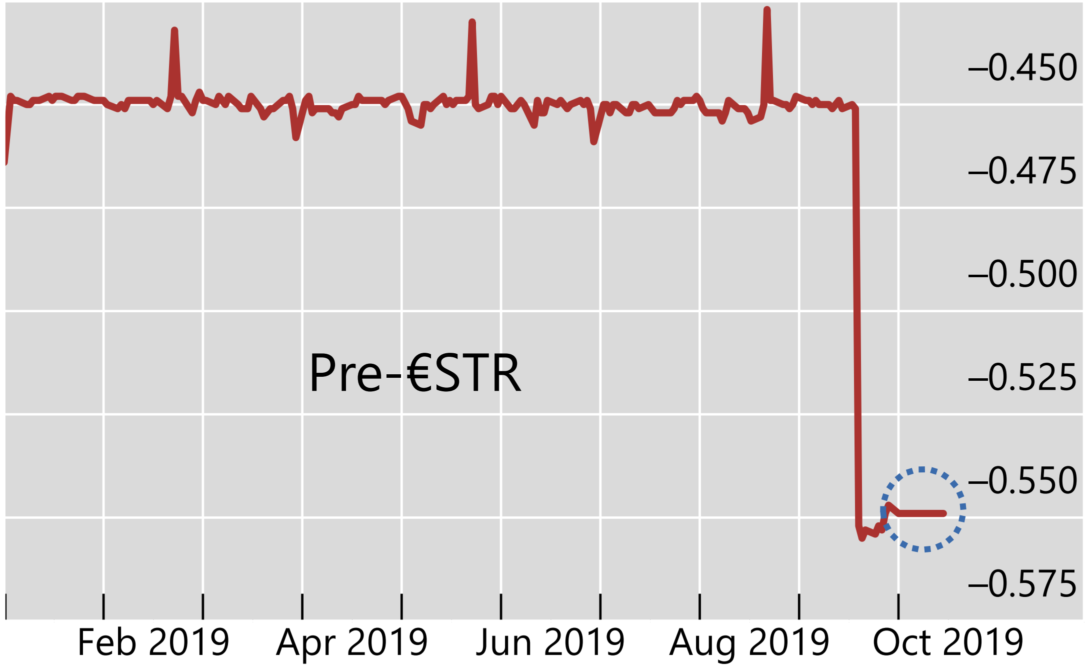
Introduction

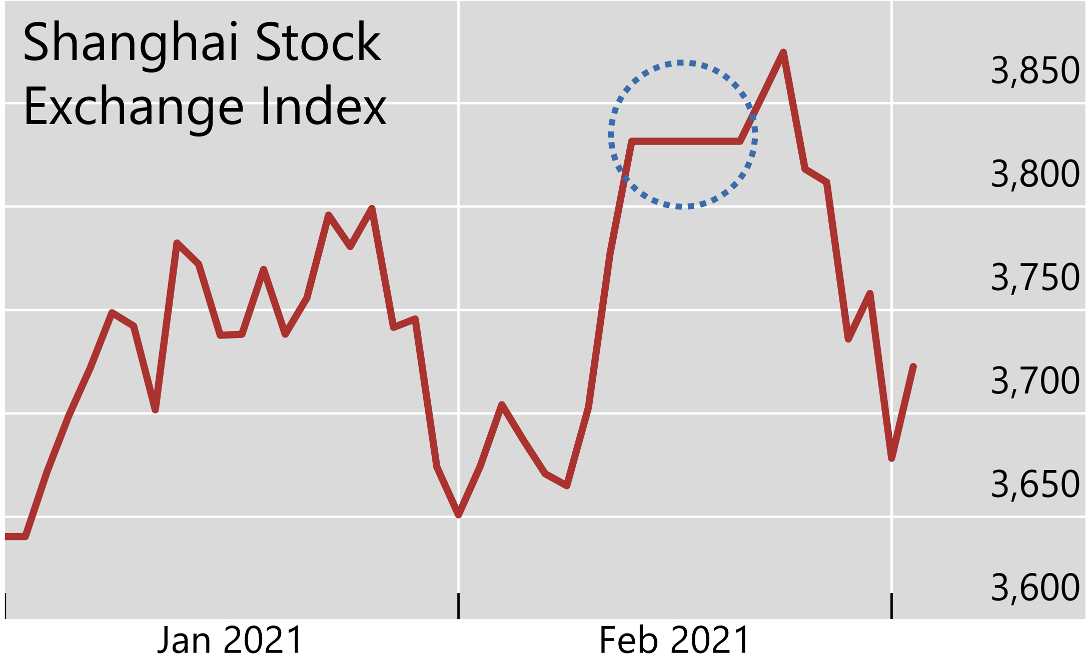
- Data validation has been becoming **challenging**

- With machine learning & enhanced computing capacity, we propose:
  - A highly **automated** validation **work flow**
  - that **outperforms** traditional approaches
  - suitable for **a large volume of financial market** data

Challenges with the traditional approaches

- Are there any issues with these series?



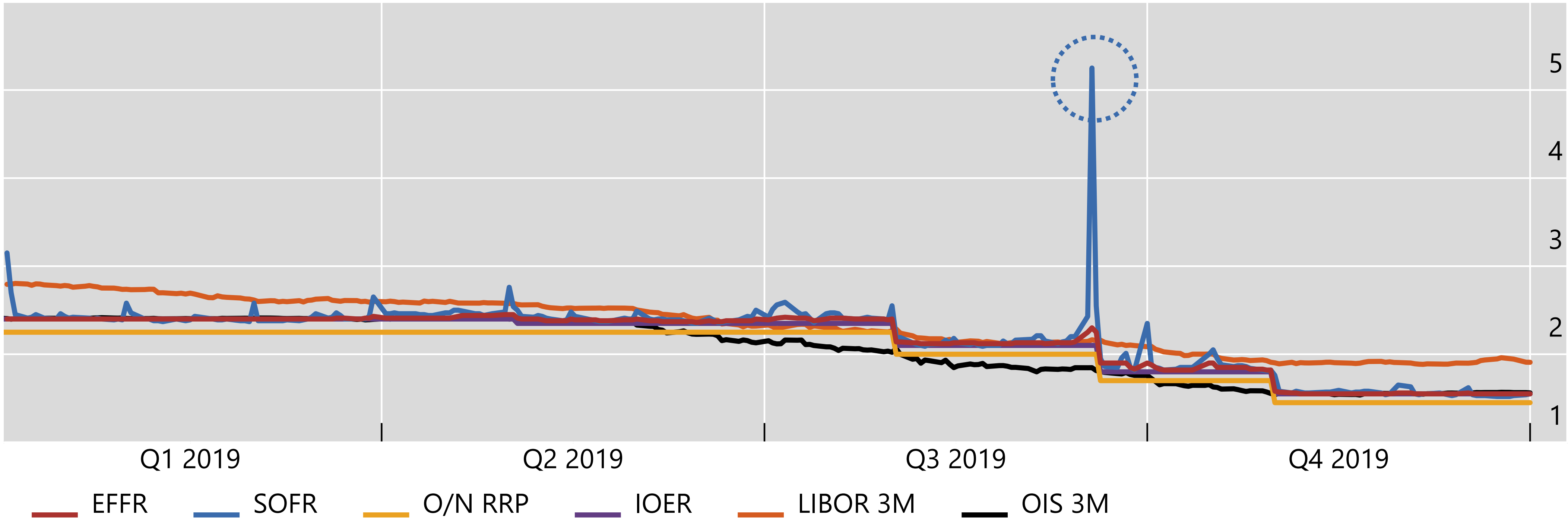


Sources: ECB; Refinitiv.

Plausible but wrong data (eg missing values, repeating values, ticker changes) are difficult to detect (False positive error (ie Type I error))

Challenges with the traditional approaches

- Is there any issue with these series?

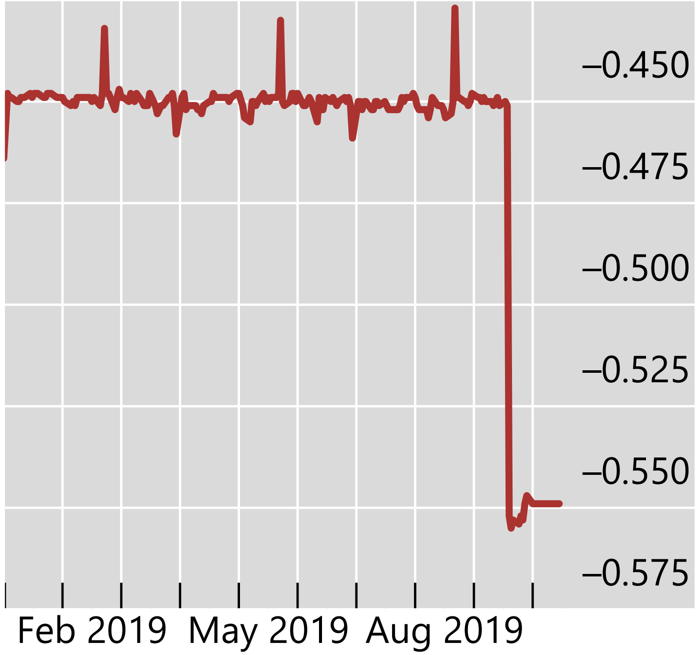


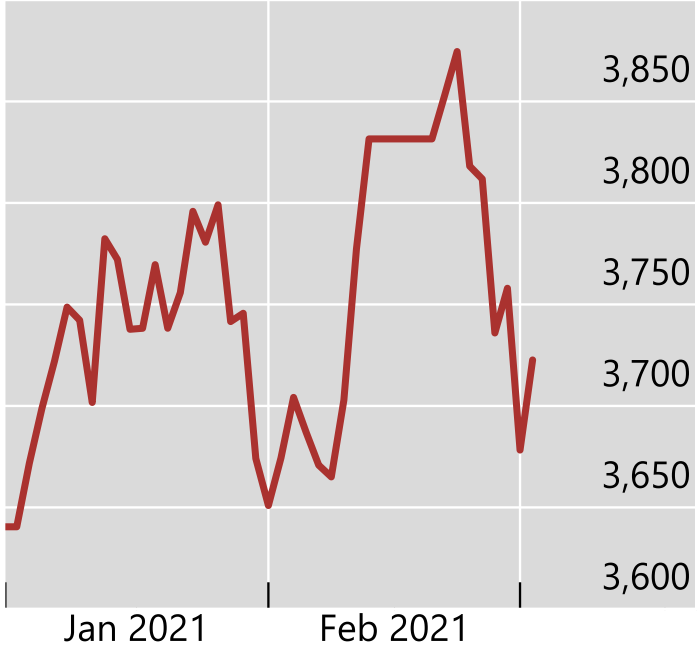
Sources: Bloomberg; Refinitiv.

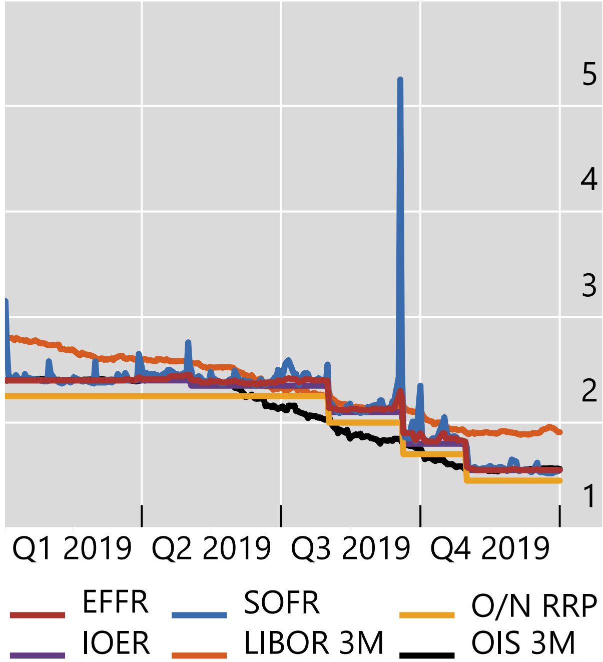
- Suspicious but correct data are also difficult to detect (False negative error (ie Type II error))
- What can be done to minimise such false alarms?

Overview of common traditional data validation techniques

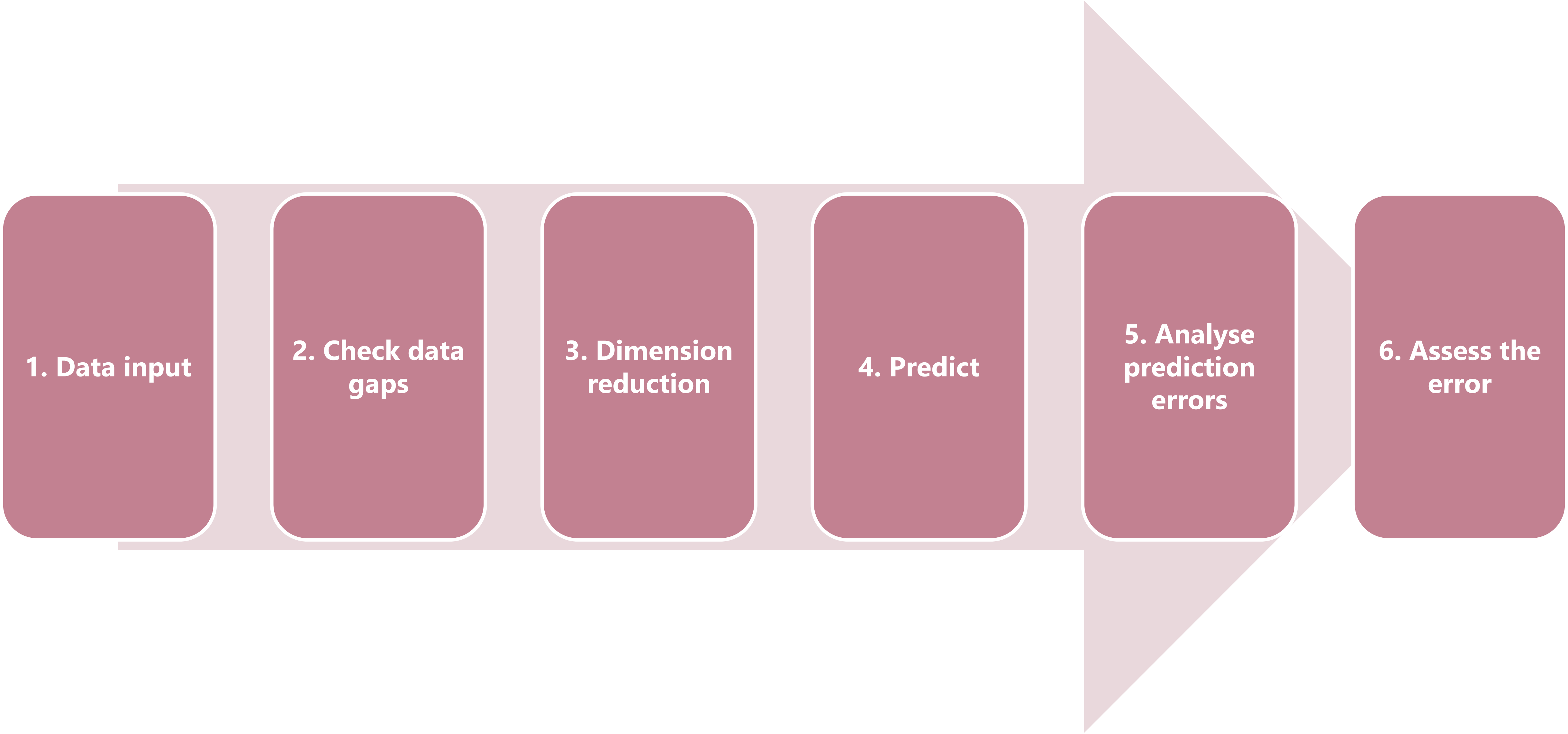
- Alone or combination of
  - Graphical method
  - Conditional controls (eg if … then …)
  - Threshold-based warnings
  - Cross-referencing
- Are they enough to validate the previous cases?







Stylized work flow of data validation process using machine learning models



1. Data input

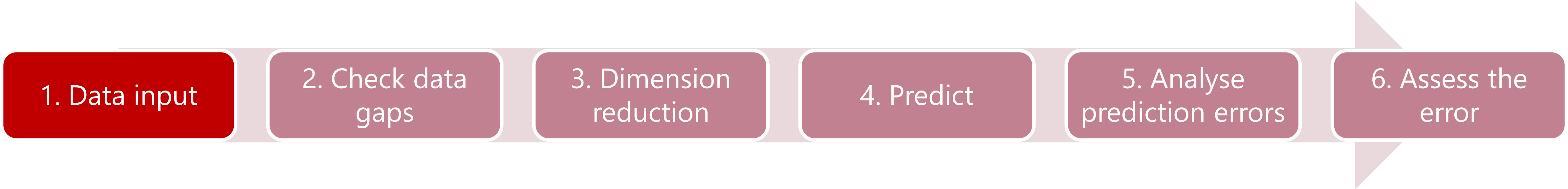
2. Check data gaps

3. Dimension reduction

4. Predict

5. Analyse prediction errors

6. Assess the error



1. Data input

2. Check data gaps

3. Dimension reduction

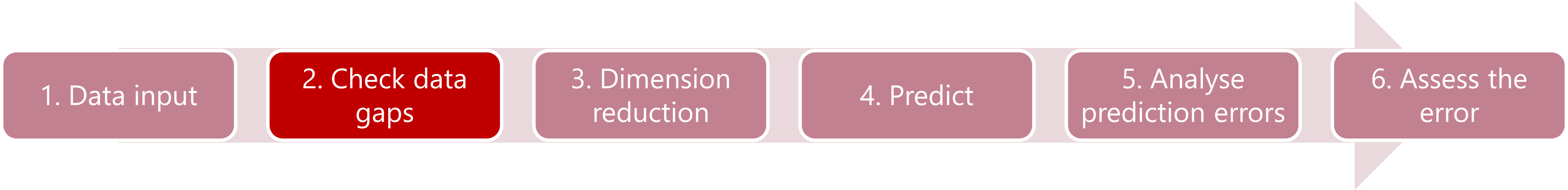
4. Predict

5. Analyse prediction errors

6. Assess the error

- About 3,000 daily incoming FM data from Bloomberg
- Characteristics of the financial market time series
  - **High-frequency big** data
  - Anomalies are **not easily visible** to human eyes
  - Frequent market **structural changes** due to market conditions, policy changes, new & outdated instruments
  - In the context of BIS – global coverage makes it difficult to understand the contexts

  - **Most series have highly correlated and/or good explanatory series, for example:**
    - NASDAQ Index, NASDAQ 100, MSCI US, S&P500, ..
    - Yields of similar maturities
    - Pegged FX rates



1. Data input

2. Check data gaps

3. Dimension reduction

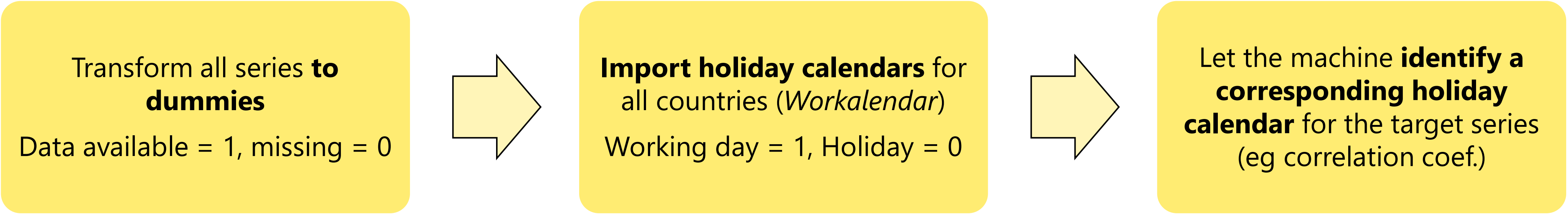
4. Predict

5. Analyse prediction errors

6. Assess the error

- How to verify if a data gap of a target series is due to market closure?

Step 1: Identify a **holiday calendar**



Transform all series **to dummies**

Data available = 1, missing = 0

Import holiday calendars for all countries (*Workalendar*)

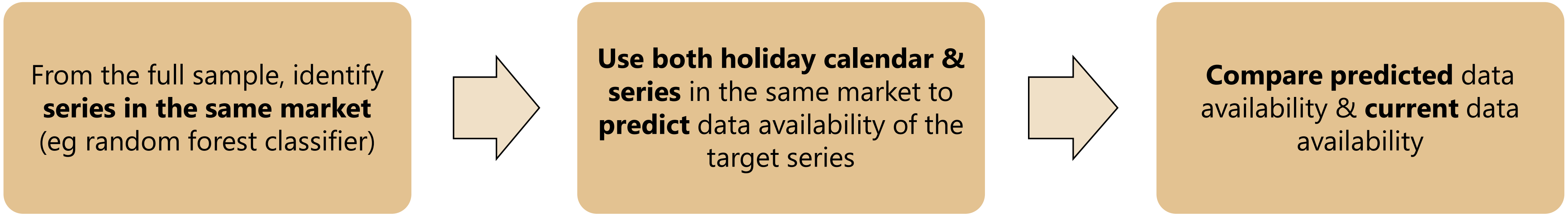
Working day = 1, Holiday = 0

Let the machine **identify a corresponding holiday calendar** for the target series (eg correlation coef.)

But having the holiday calendar is not enough (eg end-year stock market closures)

Step 2: Use the **holiday calendar** & any **series in the same market** to check data gaps

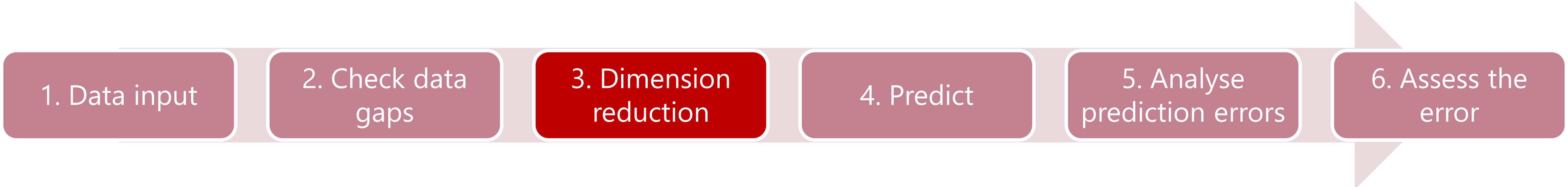
Supplementing the holiday calendar with series in the same market



From the full sample, identify **series in the same market** (eg random forest classifier)

Use both holiday calendar & series in the same market to **predict** data availability of the target series

Compare predicted data availability & **current** data availability


1. Data input

2. Check data gaps

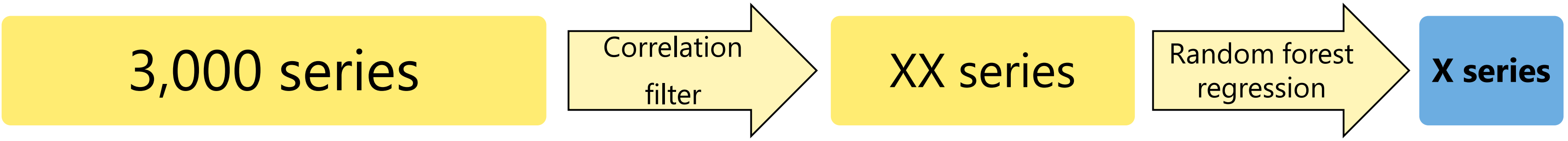
3. Dimension reduction

4. Predict

5. Analyse prediction errors

6. Assess the error

- From the full sample, **reduce dimensions** to identify a small set of series that best explain the target series (ie the most important features)


3,000 series

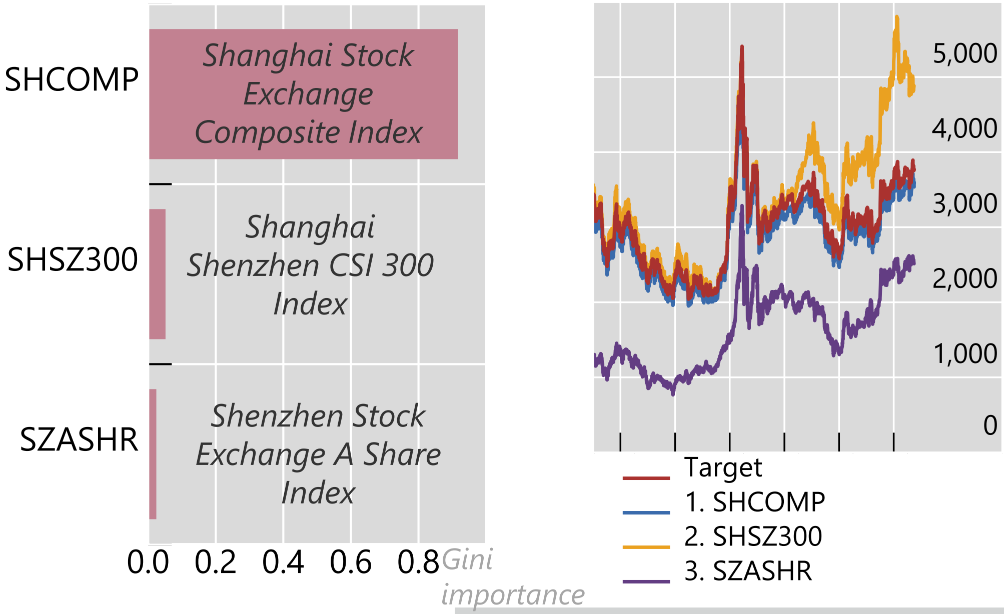
Correlation filter

XX series

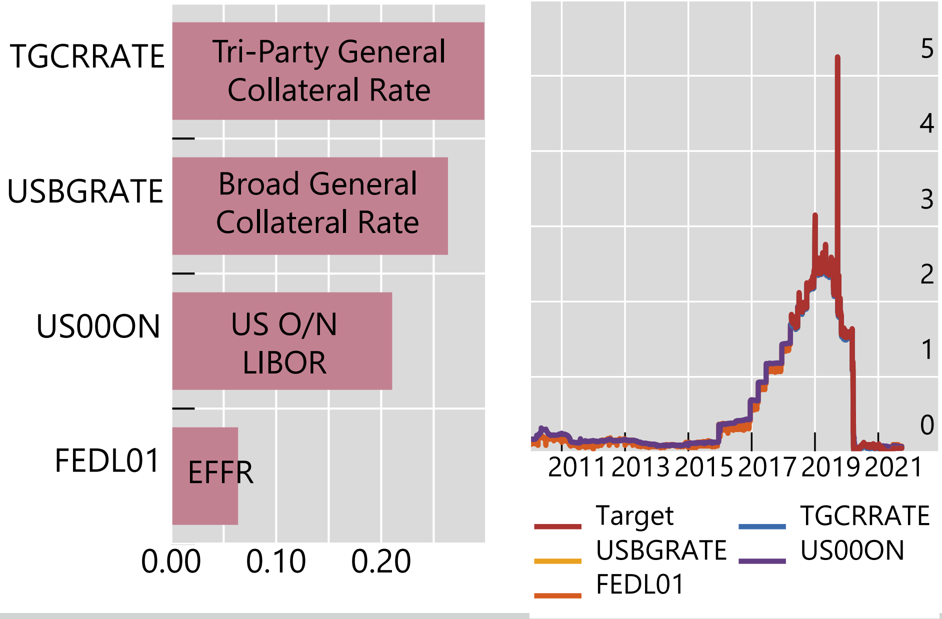
Random forest regression

X series

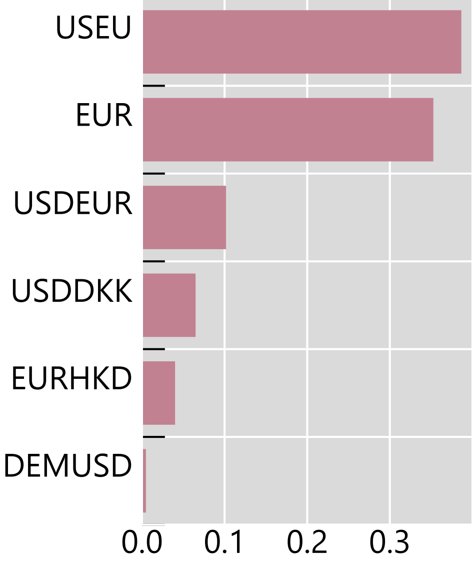
Shanghai Stock Exchange A Share Index

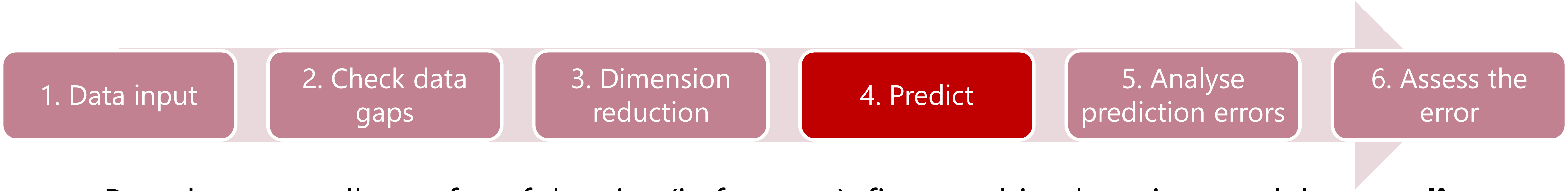


US Secured Overnight Financing Rate (SOFR)

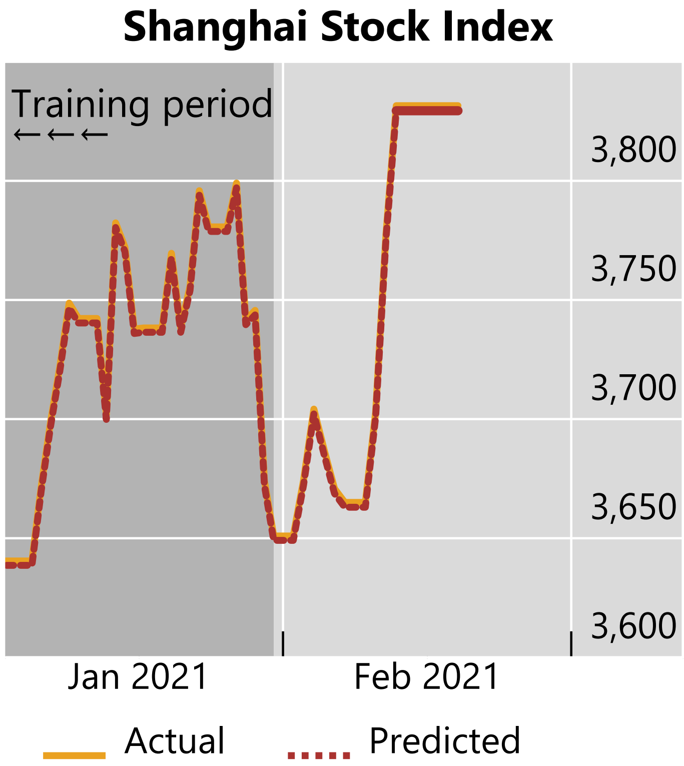


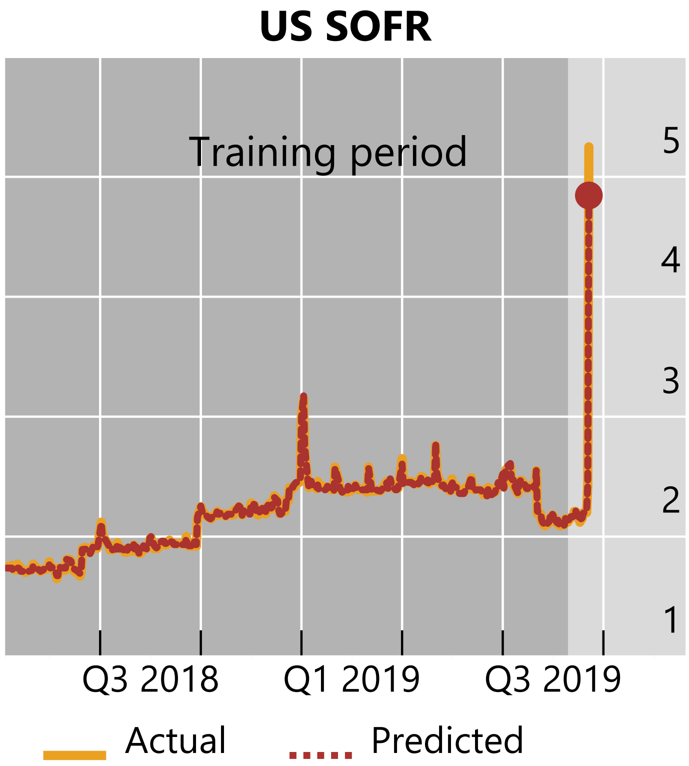
EUR/USD



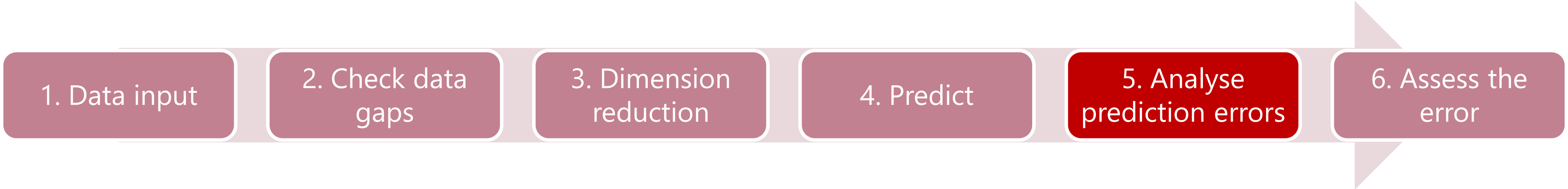


- Based on a small set of useful series (ie features), fit a machine learning model to **predict today's value** of the target series as if today's value is unavailable.





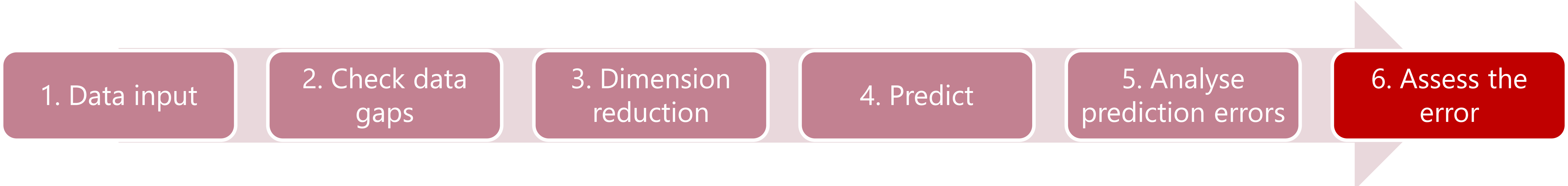
Both were predicted based on the **exactly same LSTM specifications**.



1. Data input
2. Check data gaps
3. Dimension reduction
4. Predict
5. Analyse prediction errors
6. Assess the error

- Prediction error = Predicted value – actual value
- If a prediction error is significantly *larger than usual (ie outlier)*, it is worth investigating
- Then, how can we decide whether a prediction error is an outlier?

➢ **Unsupervised learning algorithm** can help



1. Data input

2. Check data gaps

3. Dimension reduction

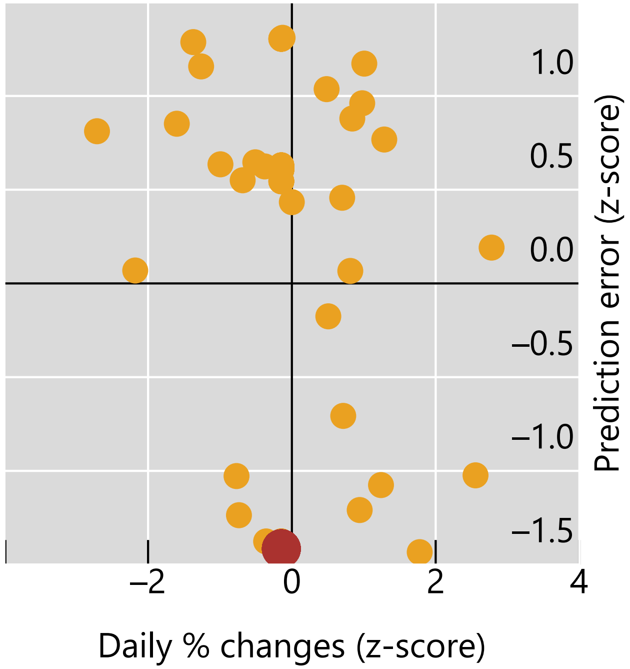
4. Predict

5. Analyse prediction errors

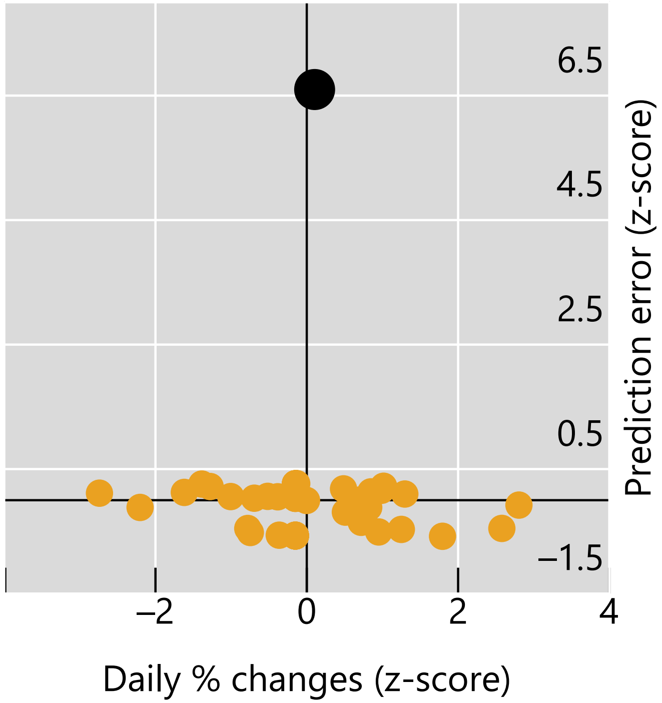
6. Assess the error

- A prediction error can be assessed based on unsupervised clustering algorithm (eg DBSCAN)
- With Shanghai Stock Index example:

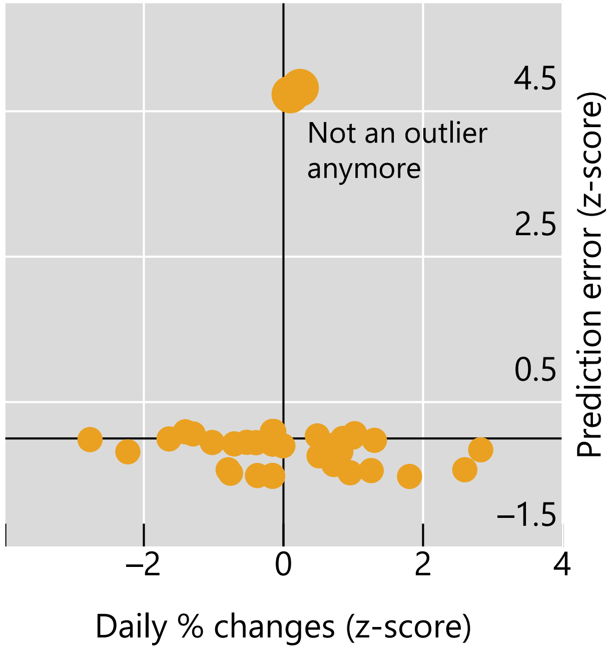
No outlier identified (eps=2, min sample=2)



What if we impose a small error in the data, will it be identified?



What if that was <u>not</u> an error but a *structural change* in the market



Thank you

- Select tools and models used in our analysis:
  - **Workalendar**: https://github.com/workalendar/workalendar
  - **Random forest classifier/regressor**: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html; https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
  - **Long short-term memory (LSTM):** https://keras.io/api/layers/recurrent_layers/lstm/
  - **Density-based spatial clustering of applications with noise (DBSCAN)**: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html