5-29-2024

Detecting Drifts in Data Streams Using Kullback-Leibler (KL) Divergence Measure for Data Engineering Applications

Jeomoan Francis Kurian
Chapman University

Mohamed Allali
Chapman University, allali@chapman.edu

Recommended Citation

Kurian, J.F., Allali, M. Detecting drifts in data streams using Kullback-Leibler (KL) divergence measure for data engineering applications. *J. of Data, Inf. and Manag.* (2024). https://doi.org/10.1007/s42488-024-00119-y

Detecting Drifts in Data Streams Using Kullback-Leibler (KL) Divergence Measure for Data Engineering Applications

Journal of Data, Information and Management
https://doi.org/10.1007/s42488-024-00119-y

ORIGINAL ARTICLE

Detecting drifts in data streams using Kullback-Leibler (KL) divergence measure for data engineering applications

Jeomoan Francis Kurian[1] · **Mohamed Allali**[2]

Received: 25 December 2023 / Accepted: 20 March 2024

Abstract
The exponential growth of data coupled with the widespread application of artificial intelligence(AI) presents organizations with challenges in upholding data accuracy, especially within data engineering functions. While the Extraction, Transformation, and Loading process addresses error-free data ingestion, validating the content within data streams remains a challenge. Prompt detection and remediation of data issues are crucial, especially in automated analytical environments driven by AI. To address these issues, this study focuses on detecting drifts in data distributions and divergence within data fields processed from different sample populations. Using a hypothetical banking scenario, we illustrate the impact of data drift on automated decision-making processes. We propose a scalable method leveraging the Kullback-Leibler (KL) divergence measure, specifically the Population Stability Index (PSI), to detect and quantify data drift. Through comprehensive simulations, we demonstrate the effectiveness of PSI in identifying and mitigating data drift issues. This study contributes to enhancing data engineering functions in organizations by offering a scalable solution for early drift detection in data ingestion pipelines. We discuss related research works, identify gaps, and present the methodology and experiment results, underscoring the importance of robust data governance practices in mitigating risks associated with data drift and improving data observability.

Keywords Kullback-Leibler divergence(KL) · Data drift · Population Stability Index(PSI) · Real-time data validation · Explainable AI · Concept drift · Data observability

1 Introduction

Organizations worldwide are experiencing exponential data growth and effective use of such data in analytical workflows presents unprecedented challenges for data engineering functions (Abedjan et al. 2016). Extraction, Transformation and Loading (ETL) process within data management functions is tasked with validation of incoming data input files, file structure verification and audit of source formats (Kimball and Ross 2013). While the process ensures the data files are ingested without error, it does not validate the content within a data field. For instance, when multiple sources feed a data field, scale, unit, or plus-minus signs can be different for a newly added data source, but they often go undetected during the ETL process. Derived data fields like a score, output of a mathematical model or a formula, could easily hide issues with data inputs from standard ETL validations. Such data input issues are commonly observed by the downstream users of the model and are denoted as manifestation of "concept drift".

Concept drift refers to the changes in distributions and statistical properties within data over time (Gama et al. 2014; Riess 2022) This makes it challenging for machine learning models to accurately project previously learned patterns to new circumstances. This results in a degradation of model performance, and depending on the application, consequences can be severe. However this is a gradual process, with a primary emphasis on adapting to evolving data patterns and implementing corrective actions to the model

Remedying data issues promptly in a production environment is expensive and any delay in such intervention risks the automated analytical functions making decisions based on erroneous data prior to issue detection. This becomes increas-

✉ Jeomoan Francis Kurian
kurian@chapman.edu

Mohamed Allali
allali@chapman.edu

[1] Researcher, Computational and Data Sciences, Chapman University, 1 University Dr, Orange 92866, CA, USA

[2] Associate Professor, Fowler School of Engineering, Chapman University, 1 University Dr, Orange 92866, CA, USA

Published online: 29 May 2024

ingly critical in the context of recent artificial intelligence (AI) applications, where data-driven decision-making occurs with minimal supervision (Polyzotis et al. 2017). Many of these data observability challenges caught the interest of the data management research community only recently. A major issue is that the behavior of AI systems depends on the data ingested, which can change due to errors in upstream data pipelines. As a consequence, algorithmic and system-specific challenges can often not be disentangled in complex AI applications (Polyzotis et al. 2018).

1.1 Problem statement

This study addresses the problem of detecting drifts in data distributions and divergence within the same data fields (input variables) processed from two different sample populations. We will elaborate on this problem using a hypothetical bank example. Bank A obtains a monthly performance data file from a credit bureau for all its credit card holders. One of the fields is a customer behavior model(CBM) score, which is useful for the bank as it helps predict the future payment delinquency of the credit card holders. The bank automates the credit card renewal processes, and the automated policy prevents auto renewals when CBM score is less than 620. In November 2022, the bank observed a decline in auto-renewal rates, falling from 95 percent to 90 percent. With a million customers renewing every month, this translated to 50,000 credit cards requiring manual renewal. Upon reviewing the input files, analysts noticed an issue with the data file, which is shown in the graphs below.

The below graph shows that the November 2021 CBM score distribution was centered around a mean value of 675, but as the November 2022 file was processed, there was a drift in this distribution, with the average score shifting to 600, even though there was no known change in the profile of the credit card holders. Later, analysts discovered that the CBM scoring model at the bureau did not accurately process one of the inputs, resulting in more customers falling into less than 600 CBM buckets. Such hard-to-detect ETL data issues are expensive for a bank that relies on automation. In this instance, 50,000 customers experienced automatic renewal denials, necessitating manual review efforts, and adversely impacting the overall customer experience. It's not just rolling back the incorrect data in production, but the downstream impact of reversing a decision poses operational and reputational risks to the Bank. As more organizations embrace AI for automating and decisioning processes, the severity of challenges related to input data problems becomes more pronounced.

The efficiency of ETL processes lies in their ability to handle input files of diverse frequencies and sizes. However, these processes lack a built-in mechanism to assess the variance of content within data fields. The presence of inconsistent data can significantly distort the results of models, often negating the benefits of AI approaches (Hellerstein 2008).As data continues to grow exponentially, and the adoption of black-box machine learning models rises, it becomes crucial to monitor less obvious data issues such as drift, as manual front-end validations prove impractical.

The concept of data drift can be traced back to early studies in information theory and statistics, laying the groundwork for subsequent advancements in research related to drift detection, adaptation strategies, and their integration into machine learning frameworks. A seminal paper published in 1951 (Kullback and Leibler 1951), which discussed data drift and later became widely acknowledged as Kullback-Leibler (KL) divergence in data distributions, has played a foundational role in many such studies. This paper builds upon these research and proposes a novel method for early drift detection in data ingestion pipelines.

1.2 Objectives of the sudy

The objective of this study is two-fold: firstly, to present Kullback-Leibler (KL) divergence as a method for detecting drifts early in data distributions, and secondly, to propose a solution addressing the identified drift problem, with the specific aim of enhancing data engineering functions in organizations that have adopted AI in automation and decision-making.

The rest of the paper is organized as follows. In Section 2, we discuss related research works and how we identified the gap and formulated the objective. In Section 3, we present the methodology in two components: first, the derivation of Population Stability Index(PSI) as a variant of Kullback-Leibler divergence, and second, the description of the simulation approach employed to generate data for the application of the PSI technique developed in the preceding section. In Sections 4, we present the experiment results and their implications, followed by a concise summary and concluding remarks in Section 5, outlining potential avenues for future research.

2 Related work

The literature we have reviewed in this context can be categorized into three groups. The first set of studies addresses

various ETL and non ETL data issues, including incorrect or inconsistent data, outliers, duplicates, missing values, integrity constraint violations, data validity in model quality, schema evolution, training-serving skew, and overall data management challenges in the context of machine learning model management (Fig. 1).

The common theme among the second set of studies is the exploration of concept drift in machine learning models, particularly in online supervised learning scenarios. These studies delve into adaptive learning processes and strategies to handle evolving data distributions. Additionally, they highlight the utilization of various techniques such as evolutionary algorithms, metaheuristics, and ensemble methods to effectively detect and adapt to concept drift in non-stationary data streams.

The last set of studies specifically focus on the application of Kullback-Leibler (KL) divergence, or its variants, in various domains to address issues related to data distribution shifts, concept drift detection, and classifier evaluation. These studies utilize KL divergence as a statistical measure to quantify dissimilarity between probability distributions, enabling the detection of anomalies, monitoring of system behavior, and identification of distributional shifts. he following paragraphs will summarize these three groups of studies.

Hellerstein (2008) examines data quality challenges in large organizations, particularly focusing on incorrect or inconsistent data. They emphasize data cleaning techniques like outlier detection and exploratory data analysis to effectively address these issues. Abedjan et al. (2016) explore data cleaning for enterprise applications, addressing errors such as outliers, duplicates, missing values, and integrity constraint violations. They stress the importance of using a combination of tools and strategies for comprehensive error coverage. Gudivada et al. (2017) discuss data quality considerations in the context of big data and machine learning, suggesting a reevaluation of traditional approaches. They introduce a data governance-driven framework and highlight tools for managing data quality beyond traditional cleaning and transformations. Polyzotis et al. (2017) tackle data management challenges within machine learning pipelines, focusing on tasks such as comprehending, validating, cleaning, and enriching training data. They emphasize the significance of data validity in model quality and address challenges like schema evolution and training-serving skew. Polyzotis et al. (2018) address data management issues in the context of machine learning model management, covering various aspects from training to deployment and monitoring. They underline the complexity of managing ML models and call for further research on data management challenges specific to ML systems.

Gama et al. (2014) provide a comprehensive examination of concept drift in online supervised learning, detailing adaptive learning processes, categorizing strategies for handling concept drift, and surveying techniques and algorithms. Their review serves as a valuable resource for understanding concept drift adaptation. In contrast, Ghomeshi et al. (2019) focus on addressing concept drifts in non-stationary data stream classification by introducing the Evolutionary Algorithm-based Concept Drift (EACD) ensemble method. This approach dynamically adjusts its ensemble to detect and resize types, offering superior performance in diverse non-stationary environments compared to existing algorithms. Riess (2022) explores automated adaptation to concept drift in machine learning models, highlighting population-based methods like Genetic Algorithm and Particle-Swarm Optimization. The study identifies challenges in evaluating minority class performance and transparency in real-world

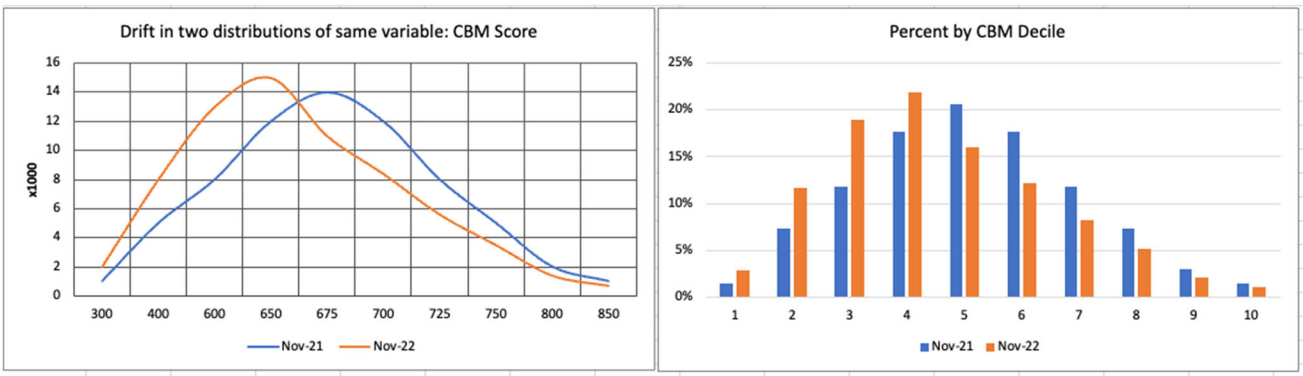

Fig. 1 The first graph shows the drift in most recent month compared to same month previous year. The second histogram arranges the CBM score by deciles and shows the percentage difference in each bucket. In a data intensive environment, where data files are processed daily and every file contains hundreds of fields, front-end validations like this is not practical

data drift characteristics, suggesting future research directions for improved concept drift detection and correction.

Zeng et al. (2014) develop statistics based on KL divergence for monitoring large-scale technical systems. Their study focuses on detecting anomalous system behavior by comparing estimated density functions with reference density functions, particularly for Gaussian distributed process variables. Basterrech and Wozniak (2022) address concept drift in continual learning, introducing Kullback-Leibler divergence for ongoing monitoring of changes in probability distributions in multi-dimensional data streams. Their method, KL-divergence-based concept drift detector (KLD), offers a fast and robust decision rule to predict and understand concept drift occurrences. Ponti et al. (2017) introduces the decision cognizant Kullback-Leibler divergence (DC-KL) as a measure for evaluating classifier agreement in decision-making systems with multiple classifiers. This research contributes to discerning between classifier congruence and incongruence in pattern recognition systems. Lin (2017) applies a variant of KL divergence called population stability index (PSI) in financial model validation, aiming to measure distributional shifts between two samples over time. Yurdakul (2018) explores KL divergence and specifically PSI properties in scorecard monitoring, providing statistical properties of PSI. This study provided a valuable reference to PSI as a distinct case of KL divergence offering deeper insights into the interpretability of PSI statistics.

The existing literature extensively investigates data management challenges, offering valuable insights into data quality, cleaning, and management. However, there is a noticeable gap in integrating scalable techniques like KL divergence or its variants for drift detection in data ingestion pipelines. While KL divergence and similar algorithms are employed for concept drift detection or front-end model validations, they primarily focus on adjusting to evolving data patterns and are slow to detect data issues. As organizations increasingly adopt AI technologies, there is a pressing need for robust data governance practices to mitigate this risk. This paper aims to address this gap by proposing a scalable drift detection algorithm, within data ingestion pipelines, utilizing a variant of KL divergence.

3 Methodology

The selection of Kullback-Leibler (KL) divergence as the evaluation metric in this study is based on the comprehensive review of existing literature, which highlights its significance in addressing data distribution shifts and concept drift detection. Unlike other algorithms that primarily focus on adjusting to evolving data patterns, KL divergence offers a statistical measure to quantify dissimilarity between data distributions, enabling the early detection of anomalies and automated intervention.

3.1 PSI as a variant of KL divergence

Given two probability distributions P (actual), and Q (expected) of a discrete random variable $x, x = x_1, x_2, ..., x_B$, KL divergence is defined as:

$$D_{KL}(P(x) \,||\, Q(x)) = \sum_{i=1}^{B} P(x_i) \cdot \ln\left(\frac{P(x_i)}{Q(x_i)}\right) \tag{1}$$

An interpretation of KL divergence is that it measures the expected excess surprise in using the actual distribution versus the expected distribution as a divergence of the actual from the expected. B is the number of buckets (discrete) of the distribution.

D_{KL} measures divergence however, researchers note that it's not a true distance measure as its definition is not symmetric. That is:

$$D_{KL}(Q(x) \,||\, P(x)) \neq D_{KL}(P(x) \,||\, Q(x))$$

A symmetric measure is obtained by defining:

$$\begin{aligned} D(P, Q) &= D_{KL}(Q \,||\, P) = D_{KL}(P \,||\, Q) \\ &= \sum_{i=1}^{B} P(x_i) \ln\left(\frac{P(x_i)}{Q(x_i)}\right) + \sum_{i=1}^{B} Q(x_i) \ln\left(\frac{Q(x_i)}{P(x_i)}\right) \\ &= \sum_{i=1}^{B} P(x_i) \ln\left(\frac{P(x_i)}{Q(x_i)}\right) - \sum_{i=1}^{B} Q(x_i) \ln\left(\frac{P(x_i)}{Q(x_i)}\right) \\ &= \sum_{i=1}^{B} (P(x_i) - Q(x_i)) \ln\left(\frac{P(x_i)}{Q(x_i)}\right) \end{aligned}$$

This variant of KL divergence is known as Population Stability Index (PSI) and is widely used in machine learning and model validations as a divergence measure. The following steps will show how to compute PSI using the CBM score data we discussed in the problem statement.

From the derivation above,

$$PSI = \sum_{i=1}^{B} (P(x_i) - Q(x_i)) \ln\left(\frac{P(x_i)}{Q(x_i)}\right) \tag{2}$$

In the context of CBM score data distribution, B is the number of bins CBM accounts data was grouped into. For example, bin 1 contains the number of accounts with CBM score between 300 and 400. $P(x_i)$ is the percent of accounts in bin i, in November 2022. This is the actual data. $Q(x_i)$ is the percent of accounts in bin i in November 2021. This is the baseline or expected data distribution in that bin. PSI is then calculated as shown in the Table 1 below.

PSI calculated in this example is 0.1106. PSI thresholds are used to determine similarity between the baseline and new samples. PSI less than 0.1 is considered similar or no significant drift. PSI between 0.1 and 0.2 is considered substantial

Table 1 Calculation of PSI

Bin B	%Base-Nov 21 Q	%TTD Nov-22 P	(1) P-Q	(2) ln(P/Q)	(1)*(2) Part PSI
1	1%	3%	0.0144	0.6844	0.0099
2	7%	12%	0.0431	0.4612	0.0000
3	12%	19%	0.0719	0.4767	0.0343
4	18%	22%	0.0422	0.2144	0.0090
5	21%	16%	−0.0455	−0.2499	0.0114
6	18%	12%	−0.0540	−0.3655	0.0197
7	12%	8%	−0.0360	−0.3655	0.0132
8	7%	5%	−0.0225	−0.3655	0.0082
9	3%	2%	−0.0090	−0.3655	0.0033
10	1%	1%	−0.0045	−0.3655	0.0016
	100%	100%	PSI(Sum of part PSI)=		0.1106

divergence and $PSI > 0.2$ is considered significant shift. However, these are only guidelines and confidence intervals can be different for different distributions.

In data validation applications, the PSI threshold can be adjusted to capture even minor changes, depending on the risk appetite of the business. Additionally, the computation of PSI can be extended to encompass all data fields that impact downstream AI models. By adopting this approach, comprehensive real-time data validation is ensured before critical decisions are made by these systems.

3.2 Simulation approach

Simulation of the data to reflect the real-life data scenarios is an important step in this study. The advantage of simulation is that we could reproduce all known data issues without having to wait to experience them in the production environment. Also, we could experiment and document how the proposed technique solves the issues.

To test the similarity of base(Q) and target(P) distributions we created the following four scenarios. Base file had four data fields, Advertisement Response, Sales Volume, Deposits, and CBM Score. Sample size, Mean and standard deviation used for each field are summarized in the Table 2 below. Assumption of normality is not necessary for PSI calculations, but these data fields tend to be normal in real life around the specified mean.

Next step is to simulate the target sample of the same data fields by introducing data issues from the real world. The error scenarios applied to the above four data series are summarized in the Table 3 below.

Ad response Advertisement response is a data series that reports the number of responses to various advertisement campaigns from the online advertisements delivered through advertisement platforms like Google, Facebook, Twitter, etc. At times incomplete files may be delivered from these platforms. 10 percent of the values selected at random were set to missing to mimic data missing from one of the major platforms.

Sales volume The field represents the sales transactions of an international luxury car dealer. It's quite unlikely that prices fluctuate significantly in this segment, so a significant price increase suggests some double counting or accounting

Table 2 Base sample(Q) - simulation criteria

Data field	Sample size	Mean	Standard deviation
Ad response	100,000	8,000	1,000
Sales volume	100,000	350,000	13,000
Deposits	100,000	75,000	20,000
CBM score	100,000	610	50

SAS: Mersenne-Twister pseudo-random number generator was used

Table 3 Through the door sample (P) simulation

Data field	TTD Sample simulation criteria
Ad response	10% observations have missing value
Sales volume	50% records in Q1 had sales value increased by 10%
Deposits	20% of random observations are reported in $ 1000s
CBM score	10% Q4 customers had 50 points drop in CBM

mistakes during a system migration. Sales transaction price was increased by 10 percent in the quartile one for 50 percent of the random records. This would shift many of them to the 2nd quartile of the base sample.

Deposits Deposit distribution of a major bank for millions of their customers. A newly added branch banking system reports the numbers in 1000s instead of actual numbers for 20 percent of random cases.

CBM score CBM is a monthly behavior score of the credit card holders of a bank, refreshed monthly to monitor the health of the portfolio. 10 percent of the quartile four customers of a bank show a score drop of 50 points due to some input error in the model, sending them to lower buckets.

Simulated base and target (TTD) distributions are plotted below (Figs. 2, 3, 4 and 5) to visualize the divergence in samples.

4 Experiment results

Tables 4 and 5 exhibit the summary of PSI calculations for each data distribution at decile and demi-decile levels respectively. For Ad Response, the PSI component values range from 0.00043 (at decile 5) to 0.10097 (at decile 1), with an overall PSI of 0.06724. While the overall PSI suggests no significant drift, the high PSI value at decile 1 indicates a potential anomaly in the data. This observation is supported by the graph for Ad Response. The PSI's capability to detect such drifts at the component level offers valuable flexibility in implementing a configurable rule to pause the ETL process for an investigation. Furthermore, Table 5 for Ad Response demonstrates that when we expanded the number of bins to 20, the issue was magnified, with the total PSI value now reaching 0.1. For Sales Volume, Deposits, and CBM Score, the total PSI values indicate a moderate to significant level of data drift, aligning with the graphical representations. Additionally, in all cases, expanding the bins led to increased PSI values, highlighting the sensitivity of PSI values to bin sizes. We provide a comprehensive breakdown of calculations at both decile and demi-decile levels in the Appendix.

5 Summary and conclusion

As detailed in Section 4, in order to simulate the data divergence issue within data streams, we chose four baseline data fields: Ad Response, Sales Volume, Deposits, and CBM Score. To introduce realistic variations, we deliberately introduced real-life errors, causing distortions in the distributions. Subsequently, we computed the Population Stability Index (PSI) with various bin sizes, and the summarized results are presented in Table 6 below.

The guidelines used are as follows: when PSI is less than 0.1, the distributions are considered similar or show 'little drift.' PSI values falling between 0.1 and 0.25 indicate a 'moderate drift,' which warrants a review. On the other hand, PSI greater than 0.25 suggests significant divergence or 'significant drift' from the baseline distribution, requiring immediate attention.

As expected, PSI effectively detected the distortions introduced into the data fields during the simulations. Ad

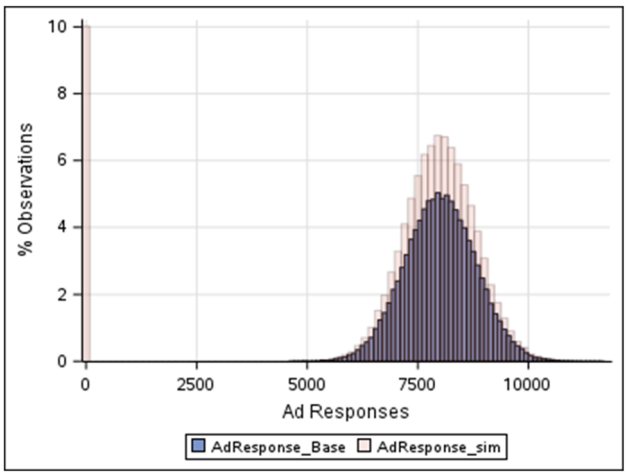

Fig. 2 Ad response

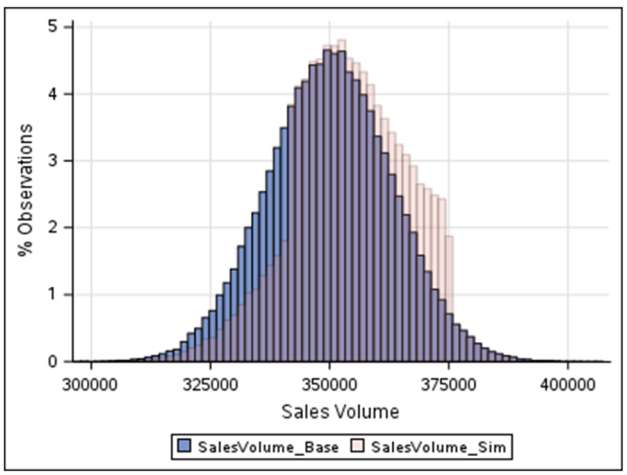

Fig. 3 Sales volume

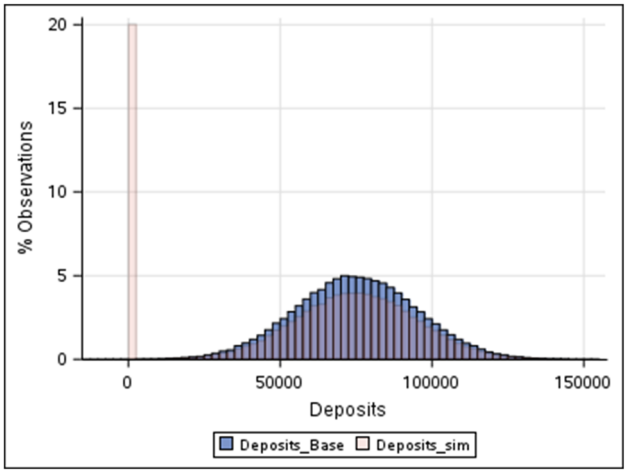

Fig. 4 Bank deposits

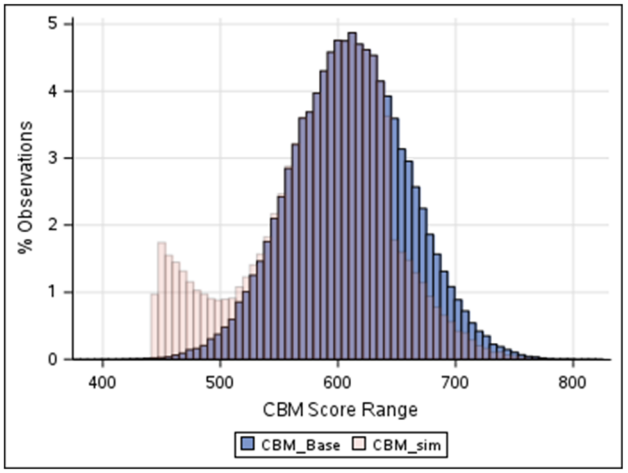

Fig. 5 CBM score

Response is the only data field that showed a below threshold number when PSI was measured using deciles. However, when binned into twenty buckets, PSI was significant with 0.11. Demi-decile binning in general produced a higher PSI value.

To conclude, PSI provides a straightforward and interpretable metric of distributional shifts between two samples over time, making it easy to understand and implement in practical scenarios. Unlike many complex drift detection algorithms, PSI calculation involves simple computations, which is easy to implement using SQL while processing data. Additionally, PSI is robust to changes in data volume and frequency, allowing for effective monitoring of data drift in dynamic environments where data streams may vary in size and frequency of updates. Moreover, PSI can detect subtle shifts in data distributions by setting appropriate thresholds to detect issues early and intervene timely to mitigate potential issues arising from data drift. Overall, the simplicity, robustness, and sensitivity of PSI make it a valuable tool for detecting data drift and maintaining the integrity of analytical workflows in data-driven organizations.

Future research PSI thresholds followed currently are from the industry best practices borrowed from engineering and modeling applications. The properties of PSI need to be studied in the context of large volume data engineering applications. The cost of false positives and false negatives differ with type of data fields so determination of PSI thresholds should be based on the cost benefit analysis. Optimal discretization (binning) is another area left to explore in a future study.

Table 4 Summary of PSI results - PSI at deciles

Deciles	Ad response	Sales volume	Deposits	CBM score
1	0.057695	0.03445	0.18514	0.09691
2	0.001071	0.03493	0.0043	0.00002
3	0.000995	0.00638	0.00465	0.00000
4	0.001062	0.00001	0.00453	0.00000
5	0.000976	0.00002	0.00477	0.00000
6	0.001008	0.00009	0.00416	0.00000
7	0.001123	0.00034	0.00471	0.00000
8	0.001108	0.00124	0.00422	0.00632
9	0.00104	0.00644	0.00434	0.03397
10	0.001162	0.04062	0.00445	0.03385
PSI	0.06724	0.12452	0.22527	0.17107

Table 5 Summary of PSI Results - PSI at Demi-deciles

Demi-deciles	Ad response	Sales volume	Deposits	CBM score
1	0.10097	0.01759	0.29832	0.14313
2	0.00053	0.01687	0.00234	0.00016
3	0.0005	0.01724	0.00197	0.00002
4	0.00058	0.01769	0.00234	0.00000
5	0.00043	0.01562	0.00235	0.00000
6	0.00057	0.00000	0.0023	0.00000
7	0.00053	0.00000	0.0022	0.00000
8	0.00053	0.00000	0.00233	0.00000
9	0.00053	0.00001	0.00254	0.00000
10	0.00045	0.00001	0.00223	0.00000

Table 5 continued

Demi-deciles	Ad response	Sales volume	Deposits	CBM score
11	0.00045	0.00003	0.00207	0.00000
12	0.00055	0.00007	0.00209	0.00000
13	0.0006	0.00013	0.00235	0.00000
14	0.00052	0.00021	0.00237	0.00000
15	0.00054	0.00039	0.0022	0.00000
16	0.00056	0.00089	0.00203	0.01648
17	0.00048	0.00169	0.00213	0.01726
18	0.00057	0.00518	0.00222	0.01672
19	0.00059	0.01792	0.00217	0.01658
20	0.00057	0.02276	0.00228	0.01727
PSI	0.11105	0.1343	0.34083	0.22762

Table 6 Summary of PSI results - PSI at demi-deciles

	Decile	Demi-Decile
Ad response	0.067	0.111
Sales volume	0.125	0.134
Deposits	0.225	0.341
CBM score	0.171	0.228

Appendix

Detailed PSI Calculation at decile and demi-decile levels

Table 7 Ad Response Table

Bin B	Ad response decile range	%Base Q	%TTD P	(1) P-Q	(2) ln(P/Q)	(1)*(2) Part PSI
1	4630−6970	10%	19%	0.0902	0.6396	0.057695
2	6980−7320	10%	9%	−0.01	−0.1069	0.001071
3	7330−7570	10%	9%	−0.0097	−0.1029	0.000995
4	7580−7790	10%	9%	−0.0101	−0.1048	0.001062
5	7800−7990	10%	9%	−0.0096	−0.102	0.000976
6	8000−8200	10%	9%	−0.0099	−0.1014	0.001008
7	8210−8410	10%	9%	−0.0101	−0.1108	0.001123
8	8420−8670	10%	9%	−0.0104	−0.1065	0.001108
9	8680−9010	10%	9%	−0.0098	−0.106	0.00104
10	9020−11670	10%	9%	−0.0105	−0.1104	0.001162
					PSI=	0.067239

Table 8 Sales volume table

Bin B	Sales volume Decile range	%Base Q	%TTD P	(1) P-Q	(2) ln(P/Q)	(1)*(2) Part PSI
1	296260−333270	10%	5%	−0.0499	−0.6908	0.03445
2	333280−339080	10%	5%	−0.0502	−0.6963	0.03493
3	339090−343200	10%	8%	−0.0236	−0.2702	0.00638
4	343210−346760	10%	10%	0.0009	0.0089	0.00001
5	346770−350070	10%	10%	0.0016	0.0155	0.00002
6	350080−353310	10%	10%	0.003	0.0298	0.00009
7	353320−356890	10%	11%	0.006	0.0579	0.00034
8	356900−361020	10%	11%	0.0114	0.1081	0.00124
9	361030−366720	10%	13%	0.027	0.2388	0.00644
10	366730−407000	10%	17%	0.0736	0.5517	0.04062
					PSI=	0.12453

Table 9 Deposits table

Bin B	Deposits Decile Range	%Base Q	%TTD P	(1) P-Q	(2) ln(P/Q)	(1)*(2) Part PSI
1	−63520	10%	28%	0.1799	1.0292	0.18514
2	49320−58140	10%	8%	−0.0197	−0.2188	0.0043
3	58150−64460	10%	8%	−0.0204	−0.2281	0.00465
4	64470−69860	10%	8%	−0.0201	−0.2248	0.00453
5	69870−74880	10%	8%	−0.0206	−0.2311	0.00477
6	74890−80020	10%	8%	−0.0194	−0.2151	0.00416
7	80030−85460	10%	8%	−0.0205	−0.2295	0.00471
8	85470−91730	10%	8%	−0.0195	−0.2168	0.00422
9	91740−100590	10%	8%	−0.0198	−0.2199	0.00434
10	100600−153830	10%	8%	−0.02	−0.2229	0.00445
				PSI=	0.22527	

Table 10 CBM score table

Bin B	CBM score Decile range	%Base Q	%TTD P	(1) P-Q	(2) ln(P/Q)	(1)*(2) Part PSI
1	378−545	10%	22%	0.1211	0.8002	0.09691
2	546−567	10%	10%	0.0013	0.0131	0.00002
3	568−583	10%	10%	0.0003	0.0026	0.00000
4	584−597	10%	10%	0.0001	0.0007	0.00000
5	598−609	10%	10%	0.0000	0.0002	0.00000
6	610−622	10%	10%	0.0000	0.0001	0.00000
7	623−635	10%	10%	0.0000	0 .0000	0.00000
8	636−651	10%	8%	−0.0237	−0.2674	0.00632
9	652−673	10%	5%	−0.0499	−0.6815	0.03397
10	674−823	10%	5%	−0.0493	−0.6865	0.03385
					PSI=	0.17108

Table 11 Ad response table

Bin B	Ad response Decile range	%Base Q	%TTD P	(1) P-Q	(2) ln(P/Q)	(1)*(2) Part PSI
1	4630−6680	5%	15%	0.0952	1.0604	0.10097
2	6690−6970	5%	5%	−0.005	−0.105	0.00053
3	6980−7170	5%	5%	−0.0049	−0.1017	0.0005
4	7180−7320	5%	4%	−0.0051	−0.1123	0.00058
5	7330−7450	5%	4%	−0.0045	−0.0962	0.00043
6	7460−7570	5%	5%	−0.0052	−0.1095	0.00057
7	7580−7680	5%	4%	−0.005	−0.1062	0.00053
8	7690−7790	5%	5%	−0.0051	−0.1035	0.00053
9	7800−7890	5%	4%	−0.0049	−0.1078	0.00053
10	7900−7990	5%	5%	−0.0046	−0.0965	0.00045
11	8000−8090	5%	4%	−0.0046	−0.099	0.00045
12	8100−8200	5%	5%	−0.0054	−0.1035	0.00055
13	8210−8300	5%	4%	−0.0052	−0.1159	0.0006
14	8310−8410	5%	4%	−0.0049	−0.1058	0.00052
15	8420−8530	5%	5%	−0.0051	−0.1063	0.00054
16	8540−8670	5%	5%	−0.0053	−0.1067	0.00056
17	8680−8820	5%	4%	−0.0047	−0.1016	0.00048
18	8830−9010	5%	4%	−0.0051	−0.1104	0.00057
19	9020−9310	5%	5%	−0.0054	−0.1103	0.00059
20	9320−11670	5%	4%	−0.0052	−0.1104	0.00057
					PSI=	0.11105

Table 12 Sales volume table

Bin B	Sales volume Decile range	%Base Q	%TTD P	(1) P-Q	(2) ln(P/Q)	(1)*(2) Part PSI
1	296260−328560	5%	2%	−0.0252	−0.6996	0.01759
2	328570−333270	5%	3%	−0.0247	−0.682	0.01687
3	333280−336510	5%	2%	−0.0249	−0.6919	0.01724
4	336520−339080	5%	2%	−0.0253	−0.7006	0.01769
5	339090−341270	5%	3%	−0.024	−0.652	0.01562
6	341280−343200	5%	5%	0.0003	0.0066	0.00000
7	343210−345020	5%	5%	0.0004	0.0085	0.00000
8	345030−346760	5%	5%	0.0005	0.0093	0.00000
9	346770−348440	5%	5%	0.0008	0.0153	0.00001
10	348450−350070	5%	5%	0.0008	0.0157	0.00001
11	350080−351690	5%	5%	0.0012	0.0235	0.00003
12	351700−353310	5%	5%	0.0018	0.0359	0.00007
13	353320−355050	5%	5%	0.0026	0.0508	0.00013
14	355060−356890	5%	5%	0.0033	0.0645	0.00021
15	356900−358800	5%	5%	0.0045	0.086	0.00039
16	358810−361020	5%	6%	0.0069	0.1291	0.00089
17	361030−363540	5%	6%	0.0096	0.1757	0.00169
18	363550−366720	5%	7%	0.0174	0.298	0.00518
19	366730−371400	5%	8%	0.0343	0.5222	0.01792
20	371410−407000	5%	9%	0.0393	0.5796	0.02276
					PSI=	0.13429

Funding Open access funding provided by SCELC, Statewide California Electronic Library Consortium.

Open Science Compliance/data availability statement SAS codes developed for simulation, data generation and PSI computation are archived in a public github repository.Reproducible package can be downloaded using the link provided in the reference section. Kurian (2023)

Declarations

Conflict of Interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Abedjan Z, Chu X, Deng D, Fernandez R, Ilyas I, Ouzzani M, Papotti P, Stonebraker M, Tang N (2016) Detecting data errors: Where are we and what needs to be done? Proceedings of the VLDB Endowment 9(12):993–1004. https://doi.org/10.14778/2994509.2994518

Basterrech S, Wozniak M (2022) Tracking changes using kullback-leibler divergence for the continual learning, pp 3279–3285. https://doi.org/10.1109/SMC53654.2022.9945547

Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia H (2014) A survey on concept drift adaptation. ACM Computing Surveys (CSUR) 46. https://doi.org/10.1145/2523813

Ghomeshi H, Gaber MM, Kovalchuk Y (2019) Eacd: evolutionary adaptation to concept drifts in data streams. Data Mining and Knowledge Discovery 33(3):663–694

Gudivada V, Apon A, Ding J (2017) Data quality considerations for big data and machine learning: going beyond data cleaning and transformations. Int J Adv Softw 10(1):1–20

Hellerstein JM (2008) Quantitative data cleaning for large databases. United Nations Economic Commission for Europe (UNECE) 25:1–42

Kimball R, Ross M (2013) The Data Warehouse Toolkit. Wiley, Boston

Kullback S, Leibler RA (1951) On information and sufficiency. Ann Math Stat 22(1):79–86

Kurian JF (2023) Open science compliance: Archived SAS codes to reproduce the results. GitHub https://github.com/FrancisKurian/Kullback

Lin AZ (2017) Examining Distributional Shifts by Using Population Stability Index (PSI) for Model Validation and Diagnosis. Paper presented at the WUSS, SAS Conference Proceedings,September 2017

Polyzotis A, Zinkevich MA, Whang S, Roy S (2017) Data management challenges in production machine

learning. In: Proceedings of the 2017 ACM international conference on management of data, New York, USA, pp 1723–1726
Ponti M, Kittler J, Riva M, Campos T, Zor C (2017) A decision cognizant kullback–leibler divergence. Pattern Recognition 61:470–478
Riess M (2022) Automating model management: a survey on meta-heuristics for concept-drift adaptation. J Data Inf Manag 4(3):211–229
Schelter S, Biessmann F, Januschowski T, Salinas D, Seufert S, Szarvas G (2018) On challenges in machine learning model management. IEEE Data Eng Bull 41:5–15
Yurdakul B (2018) Statistical Properties of Population Stability Index. Western Michigan University, Ph.D Dissertation
Zeng J, Kruger U, Geluk J, Xun W, Xie L (2014) Detecting abnormal situations using the kullback–leibler divergence. Automatica 50(11):2777–2786