Enhancing Reliability in Recommendation Systems: Beyond point estimations to monitor population stability

Yingshi Chen, Mohit Jain, Vaibhav Sawhney and Liyasi Wu

Indeed Inc.

Abstract

Ensuring reliable recommendations is essential for a company's success and user trust. Indeed has traditionally used point estimation to maintain consistent model predictions during model refinement and retraining. However, despite extensive research on robustness, there has been less focus on reliability and population monitoring. This study introduces the Cumulative Probability Stability Index (CPSI), which is derived from the Probability Stability Index (PSI), to monitor distribution stability. CPSI assesses the stability of a model's population and allows for targeted adjustments. Our implementation of CPSI proved effective in identifying significant instabilities during model transitions, demonstrating its versatility across various model types and calibration methods.

Keywords

Recommender Systems, Model Stability, Production Monitoring

1. Introduction

The #1 job site in the world, Indeed is committed to offering job seekers with high-quality opportunities through advanced recommendation systems. To maintain the accuracy of recommendations, we regularly retrain and enhance our models to effectively accommodate shifts in the job market and job seeker behavior. However, the process of retraining might produce diverse outcomes, occasionally resulting in unforeseen variations in scores, thereby compromising user confidence and product excellence [1].

Most research on recommender systems has generally concentrated on accuracy, business metrics, and diversity, often overlooking the crucial aspect of stability. Stability measures how recommendations change with updates and their consistency over time [2]. Traditional definitions emphasize strict alignment with prior predictions [1, 3], whereas more recent studies acknowledge the possibility of some deviations to accommodate new information [4]. We define stability as the ability to provide reliable and consistent recommendations while effectively adapting to changes without significant interruptions.

Point estimation methods, such as the mean or median score of prediction scores, are insufficient to assess the stability of recommender systems [5]. We utilize the Population Stability Index (PSI), a risk modeling metric that measures consistency between two probability distributions based on the Kullback-Leibler divergence [6, 7, 8, 9].

Limited research was conducted to understand the properties of PSI. There is a general rule of thumb for interpreting PSI values[8]: if PSI is less than 10%, there is no change in the population; if PSI is between 10% and 25%, the population has changed slightly, and investigation is needed; and if PSI exceeds 25%, there are significant changes in the population, and the models should be retrained[6, 7, 10]. Later research has discussed the arbitrary nature of the general 'rule of thumb' and explored the statistical properties of PSI[10].

This paper presents the Cumulative Population Stability Index (CPSI), an improved version of PSI. CPSI efficiently identifies alterations in distribution patterns and maintains robustness against noise. We demonstrate the efficacy of CPSI through simulations and real-life examples.

RecSys in HR'24: The 4th Workshop on Recommender Systems for Human Resources, in conjunction with the 18th ACM Conference on Recommender Systems, October 14–18, 2024, Bari, Italy.

yolandac@indeed.com (Y. Chen); mjain@indeed.com (M. Jain); vsawhney@indeed.com (V. Sawhney); lwu@indeed.com (L. Wu)

2. Related Work

2.1. Measurement of Stability: PSI

PSI is a metric for assessing population stability between two samples. It classifies scores into predefined bins or categories, evaluating the difference between a given probability distribution and a reference distribution.

Let N be the sample size for the reference population and M be the sample size for the target population, each being divided into B bins. Then PSI can be defined as:

$$PSI = \sum_{i=1}^{B} (\hat{p}_i - \hat{q}_i) \times (\ln \hat{p}_i - \ln \hat{q}_i) \tag{1}$$

where n_i and m_i are counts in the i-th bin, $\sum n_i = N$, $\sum m_i = M$, $\hat{p}_i = \frac{n_i}{N}$, and $\hat{q}_i = \frac{m_i}{M}$. ln denotes the natural logarithm.

2.2. Limitation of PSI

PSI focuses on local bin proportions, ignoring cumulative distribution patterns, which can result in false positives during cumulative score shifts. Additionally, fixed bin boundaries based on percentiles may not accurately capture skewed distributions.

2.2.1. Local Comparisons

PSI focuses on local comparisons of bin proportions. PSI does not account for cumulative or global distribution patterns.

According to Figure 1, the PSI value is 28.4%. Based on the general rule of thumb, if the PSI exceeds 25%, it strongly suggests that the model needs to be recalibrated. However, the Kolmogorov-Smirnov (KS) goodness-of-fit test [11] yielded a p-value of 0.173, indicating that there is no significant drift in the predictions. Additionally, the Global Comparisons (CDF) plots demonstrate that the cumulative distribution functions (CDFs) of the two distributions remain closely aligned.

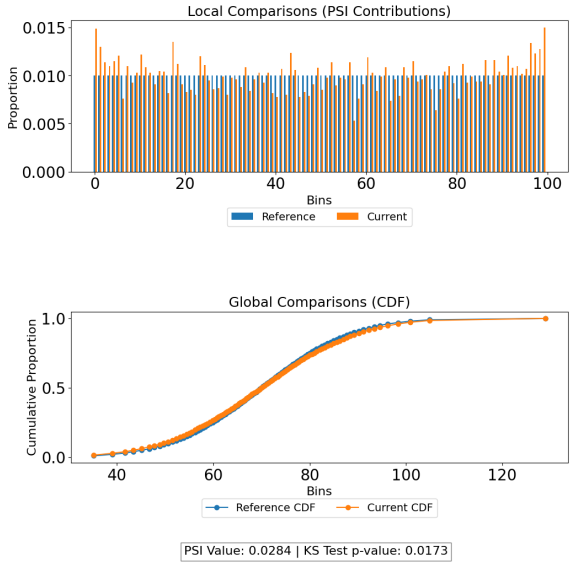

Figure 1: Comparative Analysis of Score Distributions with PSI.

2.2.2. Fixed Bin Boundaries

PSI uses predetermined bin limits determined on percentiles of the anticipated distribution. Although this method can effectively partition the data into equal-sized groups, it may not accurately capture the actual distribution's structure, especially for distributions that are skewed or have several modes.

The fixed bin boundaries intersect many modes of the distribution, which may result in an inaccurate representation of the discrepancies. This can result in bins containing many peaks or valleys, leading to less accurate stability assessments.

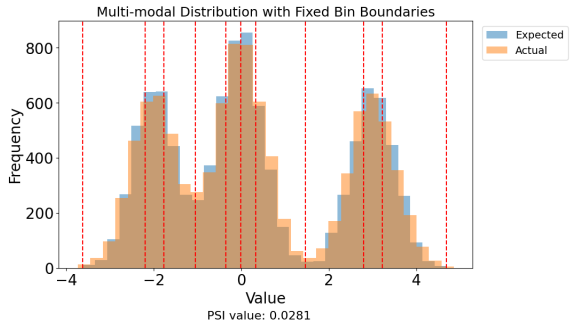

Figure 2: Multi-modal Distribution Comparison with PSI.

3. Proposed Method

In this section, we propose the Cumulative Population Stability Index (CPSI) to provide a comprehensive view of distribution changes. CPSI provides a detailed assessment of distributional changes by computing localized cumulative sums, allowing tailored analysis and evaluation within sliding windows of bins.

3.1. Definition

The CPSI is defined as:

$$\text{CPSI} = \sum_{i=1}^{B} (\tilde{P}_{i-k,i+k} - \tilde{Q}_{i-k,i+k}) \times \ln\left(\frac{\tilde{P}_{i-k,i+k}}{\tilde{Q}_{i-k,i+k}}\right) \tag{2}$$

where:

- $\tilde{P}_{i-k,i+k}$ and $\tilde{Q}_{i-k,i+k}$ are the cumulative proportions of the initial and new distributions, respectively, from bin $\max(1, i-k)$ to bin $\min(B, i+k)$, defined as:

$$\tilde{P}_{i-k,i+k} = \sum_{j=\max(1,i-k)}^{\min(B,i+k)} P_j \quad \text{and} \quad \tilde{Q}_{i-k,i+k} = \sum_{j=\max(1,i-k)}^{\min(B,i+k)} Q_j$$

- P_i and Q_i represent the proportions in the reference and current (prediction) distributions, respectively, for bin i.
- B is the total number of bins.
- k is the number of bins included in the cumulative sum on either side of bin i.
- N and M are the sample sizes of the reference and current (prediction) distributions, respectively.
- ln denotes the natural logarithm.

CPSI can be viewed as a variation of PSI. Recall the definition of PSI:

$$PSI = \sum_{i=1}^{B} (\hat{p}_i - \hat{q}_i) \times (\ln \hat{p}_i - \ln \hat{q}_i) \tag{3}$$

When comparing these definitions, we see that CPSI is a transformation of PSI, where $\hat{p}_i$ in PSI corresponds to $\tilde{P}_{i-k,i+k}$ in CPSI.

3.2. Statistical Properties of CPSI

3.2.1. Expectation of CPSI

As proved by Yurdakul and Naranjo [10], the expectation of PSI is:

$$E(\text{PSI}) = \sum_{i=1}^{B} (p_i - q_i)(\ln p_i - \ln q_i) + \frac{B-1}{N} + \frac{B-1}{M} + \frac{B - \sum_{i=1}^{B} \frac{q_i}{p_i}}{2N} + \frac{B - \sum_{i=1}^{B} \frac{p_i}{q_i}}{2M} \tag{4}$$

Since $\hat{p}_i$ in PSI corresponds to $\tilde{P}_{i-k,i+k}$ in CPSI, the expected value $E(\text{CPSI})$ can be expressed analogously to $E(\text{PSI})$, with $\tilde{P}_{i-k,i+k}$ and $\tilde{Q}_{i-k,i+k}$ substituting for $\hat{p}_i$ and $\hat{q}_i$, respectively.

$$E(\text{CPSI}) = \sum_{i=1}^{B} (\tilde{P}_{i-k,i+k} - \tilde{Q}_{i-k,i+k})(\ln \tilde{P}_{i-k,i+k} - \ln \tilde{Q}_{i-k,i+k}) + \frac{B-1}{N} + \frac{B-1}{M} + \frac{B - \sum_{i=1}^{B} \frac{\tilde{Q}_{i-k,i+k}}{\tilde{P}_{i-k,i+k}}}{2N} + \frac{B - \sum_{i=1}^{B} \frac{\tilde{P}_{i-k,i+k}}{\tilde{Q}_{i-k,i+k}}}{2M} \tag{5}$$

Under the null hypothesis $H_0 : p_i = q_i,\ i = 1, \ldots, B$, we have:

$$\tilde{p}_{i-k,i+k} = \tilde{q}_{i-k,i+k}$$

Where:

$$\tilde{P}_{i-k,i+k} = \sum_{j=\max(1,i-k)}^{\min(B,i+k)} P_j \quad \text{and} \quad \tilde{Q}_{i-k,i+k} = \sum_{j=\max(1,i-k)}^{\min(B,i+k)} Q_j$$

The expectation of CPSI is:

$$E(\text{CPSI}) = (B-1)\left(\frac{1}{N}+\frac{1}{M}\right) \tag{6}$$

Proof. Under H_0, the first term is 0. Also $B-\sum_{i=1}^{B}\frac{\tilde{Q}_{i-k,i+k}}{\tilde{P}_{i-k,i+k}} = 0$ and $B-\sum_{i=1}^{B}\frac{\tilde{P}_{i-k,i+k}}{\tilde{Q}_{i-k,i+k}} = 0$.

3.2.2. Theorem and Variance Calculation

Yurdakul and Naranjo[10] use the following theorem to identify the variance of PSI. **Theorem 1** Let $E(Y) = \mu$ and $\text{Cov}(Y) = \Sigma$. The proof of this theorem can be found in Searle's [12].

Then:

$$\text{Var}(Y'AY) = 2\text{Tr}(A\Sigma A\Sigma) + 4\mu' A\Sigma A\mu. \tag{7}$$

Given that $\mu = E(Y) = 0$, Theorem 1 implies:

$$\text{Var}(Y'AY) = 2\text{Tr}(A\Sigma A\Sigma).$$

For PSI, assuming that the null hypothesis $H_0 : p_i = q_i$, $i = 1, \dots, B$ is true, they prove that $\mu = 0$ and

$$Tr(A\Sigma A\Sigma) = \left(\frac{1}{N}+\frac{1}{M}\right)^2 \times \sum_{i=1}^{B}(1-p_i)$$

finally leading to

$$\text{Var}(\text{PSI}) = 2\left(\frac{1}{N}+\frac{1}{M}\right)^2 \times (B-1)$$

Recall that CPSI can be viewed as a transformation of PSI, where the elements of the matrix A are replaced by cumulative sums over k bins:

$$\text{CPSI}^* = \sum_{i=1}^{B}\left(\tilde{P}_{i-k,i+k} - \tilde{Q}_{i-k,i+k}\right)^2 \frac{1}{\tilde{P}_{i-k,i+k}} = Y^T\tilde{A}Y$$

Under H_0, we have

$$\tilde{A} = \begin{bmatrix} w\frac{1}{\tilde{P}_{1-k,1+k}} & 0 & \cdots & 0 \\ 0 & \frac{1}{\tilde{P}_{2-k,2+k}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\tilde{P}_{B-k,B+k}} \end{bmatrix}$$

Thus, the variance of CPSI is:

$$\text{Var}(\text{CPSI}) = 2\text{Tr}(\tilde{A}\Sigma\tilde{A}\Sigma).$$

Under the null hypothesis $H_0 : \tilde{p}_{i-k,i+k} = \tilde{q}_{i-k,i+k}$, $i = 1, \dots, B$, for CPSI we can say that $\mu = 0$ and

$$Tr(\tilde{A}\Sigma\tilde{A}\Sigma) = \left(\frac{1}{N}+\frac{1}{M}\right)^2 \times \sum_{i=1}^{B}(1-\tilde{P}_{i-k,i+k})$$

$$= \left(\frac{1}{N}+\frac{1}{M}\right)^2 (B-1)$$

Then:

$$\text{Var}(\text{CPSI}) = 2\left(\frac{1}{N}+\frac{1}{M}\right)^2 (B-1)$$

3.2.3. Robustness and Invariance of CPSI

The distribution of CPSI, controlled by B, N, and M (total number of bins, and sample sizes of the reference and target populations, respectively), is unaffected by the underlying variable distributions, ensuring it remains a reliable and robust measure of divergence between model predictions.

3.3. Parameter Selection

3.3.1. Determining Sample Sizes N and M

The robustness and dependability of the Cumulative Population Stability Index (CPSI) depend critically on the sample sizes of the target population M and the reference population N. The sensitivity and stability of the index can be greatly affected by the choice of N and M.

To make sure that N and M are big enough to find significant differences, we have performed a power analysis. Furthermore, the selection of these parameters can be guided by preliminary exploratory data analysis, and model performance can be optimized through iterative refinement. A final decision between N and M should take practical limits, stability, and sensitivity into account.

3.3.2. Determining the Optimal Number of Bins

Determine the number of bins is crucial to accurately capture the distributional characteristics of the data. An optimal value for the number of bins must balance both bias and variance. There are several studies that have attempted to determine an optimal number of bins, each offering different advantages based on the size and distribution of the data:

Square-Root Choice:

$$B = \sqrt{N}$$

The square-root choice method recommends using the square root of the data points to determine the number of bins, providing a balanced approach for moderate-sized datasets[13].

Sturges' Formula:

$$B = \lceil \log_2 N + 1 \rceil$$

Sturges' formula, commonly used for smaller datasets, assumes that the data follow an approximate normal distribution. The objective is to ascertain an appropriate number of bins that properly reflect the distribution of the data points, while avoiding unnecessary complexity in the model.[14].

Rice Rule:

$$B = \lceil 2 \times n^{1/3} \rceil$$

The Rice Rule proposes determining the appropriate number of bins by achieving a balance between granularity and simplicity. This method is particularly effective for larger datasets.[15] [16].

Each of these methods provides a pragmatic way to select bins, depending on the specific attributes of the data set and the objectives of the research. We looked at a few different approaches to figure out how many bins there should be, and we used those approaches to figure out the values to work from. We then analyzed the trade-off between bias and variance as a function of the parameter B, with N and M kept constant. The parameter B was selected to minimize the overall error.

3.3.3. Determining the number of k

A key factor in balancing the sensitivity and robustness of the Cumulative Population Stability Index (CPSI) is the selection of k, which regulates the number of bins included in the cumulative sum on either side of a bin i.

We determined the optimal value of k, using a combination of domain-specific knowledge and exploratory data analysis. We have deliberately chosen to maximize the noticeable variability in CPSI values during calibration while minimizing penalty for small shifts between adjoining bins. This approach aligns with our objective of reducing penalization for modest distribution shifts caused by calibration, so that predictions remain reasonably close to the true likelihood.

3.4. Rule of Thumb

By comparing subsequent model version distributions using the Cumulative Population Stability Index (CPSI), we can quantify changes and establish a benchmark for stability between iterations. To effectively apply CPSI, we require two population samples: the base sample, which represents the score distribution from a previous model version, and the test sample, representing the predicted score distribution from the current model version. We propose the following 'rule of thumb' values derived from empirical data, specifically using the 90th and 99th percentiles. (details in Appendix 8.1) to monitor system stability across model retraining versions by assessing the CPSI measure over different historical time frames.

4. Results: evaluating the Effectiveness of CPSI

We conducted a simulation study to assess CPSI performance using the normal approximation for critical values. Based on the statistical properties of CPSI, we can construct the following test:

$$\text{CPSI} > \left(\frac{1}{N}+\frac{1}{M}\right)(B-1) + z_{0.95}\left(\frac{1}{N}+\frac{1}{M}\right)\times\sqrt{2(B-1)}$$

where the right-hand side (RHS) is the critical value, defined as the 95th percentile of the CPSI normal approximation.

We created a right-skewed baseline using a Beta distribution and introduced small shifts and noise to simulate real-world conditions. PSI and CPSI values were calculated for the expected and new distributions. We sampled 10,000 values from the baseline and challenger distributions, conducting 30 simulations. The CPSI results were computed with the number of bins (B) set to 1,000 and K set to 1. We compared these results with the critical value of 0.0215, as suggested by normal approximation.

The results table (Table 1) along with the plot (Fig.3), shows that PSI identified small shifts as unstable, indicating a high sensitivity to local changes. In contrast, CPSI smooths out local variations and focuses on cumulative proportions, proving robust against noise while effectively detecting distribution shifts.

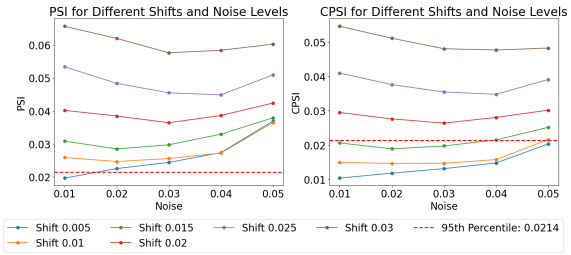

Figure 3: PSI and CPSI plots for different shifts and noise levels

shift	noise	PSI_rejection_rate	CPSI_rejection_rate
0.005	0.01	23.33	0.00
0.005	0.02	36.67	0.00
0.005	0.03	100.00	0.00
0.005	0.04	100.00	0.00
0.005	0.05	100.00	0.00
0.010	0.01	100.00	0.00
0.010	0.02	100.00	0.00
0.010	0.03	100.00	0.00
0.010	0.04	100.00	0.00
0.010	0.05	100.00	46.67
0.015	0.01	100.00	3.33
0.015	0.02	100.00	0.00
0.015	0.03	100.00	0.00
0.015	0.04	100.00	33.33
0.015	0.05	100.00	100.00
0.020	0.01	100.00	100.00
0.020	0.02	100.00	100.00
0.020	0.03	100.00	100.00
0.020	0.04	100.00	100.00
0.020	0.05	100.00	100.00

Table 1
Rejection Rates for different shifts and noise levels

5. Discussion: Real-world applications

Our system uses various algorithms for Apply Rate prediction, focusing on Deep&Cross V2 [17] models in our ad recommender systems. Despite DNNs' strong prediction performance, identical DNN models trained on the same data can yield different results [18]. Detecting significant distribution shifts and raising accurate alarms is challenging.

We conducted a re-evaluation of a stability experiment for offline assessment, which consisted of millions of emails sent to job seekers. This experiment aimed to test a treatment designed to mitigate instability between different retrained versions of predictive models. Treatments included altering the calibration method and transitioning the training dataset splitting from timestamps to job IDs. The accompanying figure (Fig.4) illustrates the effectiveness of this treatment in reducing instability, measured by the mean apply rate. Observations showed that the control models experienced spikes and sudden drops, whereas the version transitions in the test model were markedly smoother. Although point estimates alone are insufficient to confirm stability, they do offer some directional confidence.

This experiment provides a valuable case study for evaluating the effectiveness of the Cumulative Population Stability Index (CPSI) and comparing it with other population stability metrics. These metrics include the Population Stability Index (PSI), the recently introduced Population Ac-

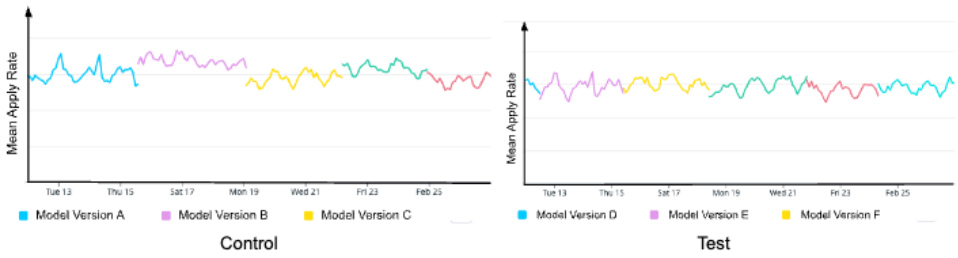

Figure 4: Comparison of Mean Apply Rates: Control vs. Test

curacy Index (PAI) [19], and the Kolmogorov-Smirnov (KS) goodness-of-fit test [11].

The Population Accuracy Index (PAI) offers an alternative approach to scorecard stability testing by measuring the change in the variance of the estimated mean response since development. PAI Interpretation: $0 \leq \text{PAI} < 1.1$ indicates no substantial change, $1.1 \leq \text{PAI} < 1.5$ suggests a small change, and $\text{PAI} \geq 1.5$ indicates a substantial change. The Kolmogorov-Smirnov (KS) test statistic quantifies the maximum difference between two empirical distribution functions, providing a measure of the discrepancy between observed and expected distributions.

Versions	CPSI	PAI	PSI	KS_statistic	KS_p_value	Treatment
Model B compared with A	0.04	0.971	0.054	0.07	0	Control
Model C compared with B	0.071	1.145	0.085	0.113	0	Control
Model D compared with C	0.011	0.996	0.025	0.026	0	Control
Model E compared with D	0.019	0.987	0.030	0.04	0	Control
Model B' compared with A'	0.012	0.959	0.028	0.021	0	Test
Model C' compared with B'	0.011	1.025	0.027	0.018	0	Test
Model D' compared with C'	0.01	1.036	0.023	0.011	0	Test
Model E' compared with D'	0.011	0.966	0.025	0.025	0	Test

Table 2
Comparison of Models with Various Metrics

KS test is overly sensitive to small changes when the sample size is large, often labeling any model changes as instability [20]. Even minor distributional changes can lead to the rejection of population stability at nominal significance levels, potentially misrepresenting true instability. Similarly, PSI is prone to detecting small local changes, frequently marking all movements as unstable.

However, the experimental observations are at odds with the PAI results, which suggest little change between model transitions in both the test and control groups.

This study demonstrates how CPSI outperforms other well-established techniques in determining population stability due to its resilience and comprehensiveness.

6. Monitoring System Implementation with CPSI

In this section, we present the implementation of an online recommender monitoring system that incorporates CPSI Metrics.

We designed a testing infrastructure to leverage the requests coming to the incumbent model in production to test against the challenger model which was trained using a different set of data. The 'incumbent' model refers to the machine learning model that is currently deployed in production and actively handling real-world requests or tasks. It is the established model that new or 'challenger' models are compared against to determine if an upgrade or replacement is warranted. The architecture of the said infrastructure contains two modules: one to poll for a newly trained challenger model called Model Score Verification Initiator and another to test it against the incumbent model called Model Score Verifier.

Step by step depiction of how the infrastructure is laid out to perform testing.

- **Gathering data:** Collect a set of sampled requests from the past 14 days. These requests should include either a list of multiple jobs being matched to one job seeker or a list of multiple job seekers being matched to one job. They were originally inferred using the model being tested in the past.
- **Preparing testing infrastructure:** Set up the necessary testing environment, including Model Score Verification Initiator, databases, and any required software or tools.
- **Triggering test:** Initiate the test by triggering an instance of the Model Score Verifier for the models being tested.
- **Loading the right models:** Model Score Verifier ensures that the correct model are loaded and active in the application responsible for inferring the requests.
- **Sending and inferring requests:** Forward the gathered requests to the loaded models for inference.
- **Logging responses:** Record the responses generated by the models for later analysis, each containing one score attached to multiple unique job-job seeker pairs, respectively, that were part of the request.
- **Deciding to promote or drop:** On gathering 100k unique pairs with their relevance score we calculate CPSI. The process of gathering 100k unique pairs with their scores is repeated 30 times, and a mean CPSI score is calculated. Evaluate the results to decide whether the tested model should be promoted to production or discarded using the mean CPSI score. By algorihtms such as Jackknife resampling, we can find the standard error of CPSI through the 30-time calculation. This can be used to calculate the confidence interval of CPSI. Here the number 30 is to insure we can have a statistically sound conclusion.
- **Triggering alerts:** If the test identifies any critical issues or anomalies, automatically trigger alerts to notify the relevant stakeholders.

Figure 5 provides a simplified overview of the monitoring system. This system enables proactive monitoring, investigation, and improvement of recommendation system performance in production. By integrating this system into our

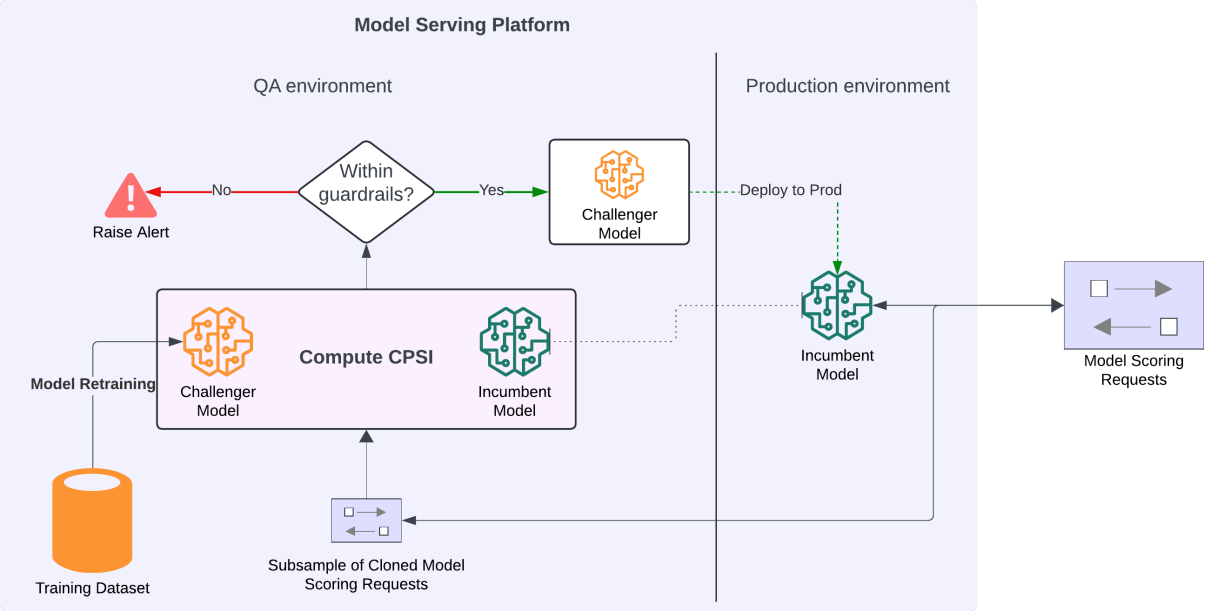

Figure 5: Using CPSI for Model deployment monitoring

workflow, we can detect major issues early without depending on manual monitoring, preventing negative impacts on customer trust and reducing churn.

7. Conclusion and Future Work

In this paper, we introduced the Cumulative Probability Stability Index (CPSI) as a tool for monitoring large-scale recommender systems. We demonstrated CPSI's effectiveness in detecting significant instabilities during model transitions and its robustness against prediction variations through simulations, real-world implementations, and monitoring systems. CPSI has proven to be a reliable metric for evaluating the stability of recommender systems, both offline and online, especially for DNN-based recommendation systems. We believe CPSI has potential applications in other domains as well.

In the future, we aim to extend the application of the proposed stability monitoring methods to a broader range of scenarios, including various model types such as reinforcement learning models. In addition, we plan to evaluate the effectiveness of these methods in different domains, explore their scalability in large-scale systems, and assess their adaptability to real-time monitoring environments.

References

[1] Adomavicius, G., and Zhang, J. 2010. On the stability of recommendation algorithms. In Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10, 47–54. New York, NY, USA: ACM.

[2] O'Mahony, M., Hurley, N., Kushmerick, N., and Silvestre, G. 2004. Collaborative recommendation: A robustness analysis. ACM Trans. Internet Technol. 4(4):344–377.

[3] Adomavicius, Gediminas and Zhang, Jingjing. 2012. Stability of recommendation algorithms. ACM Transactions on Information Systems 30, 4 (Nov. 2012).

[4] Shriver, D., Elbaum, S., Dwyer, M., and Rosenblum, D. 2019. Evaluating Recommender System Stability with Influence-Guided Fuzzing. Proceedings of the AAAI Conference on Artificial Intelligence 33:4934-4942. https://doi.org/10.1609/aaai.v33i01.33014934.

[5] Ekstrand, M., Carterette, B., and Diaz, F. 2023. Distributionally-Informed Recommender System Evaluation. ACM Transactions on Recommender Systems 2, 1:1–27. Online publication date: 31-Mar-2024.

[6] Thomas, L. C., Edelman, D. B., and Crook, J. N. 2002. Credit Scoring and its Applications. SIAM monographs on mathematical modeling and computation. Philadelphia: SIAM.

[7] Siddiqi, N. 2017. Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards. John Wiley & Sons.

[8] Lewis, E. M. 1994. Introduction to Credit Scoring. The Athena Press.

[9] Kullback, S., and Leibler, R. A. 1951. On Information and Sufficiency. The Annals of Mathematical Statistics 22, 1:79–86.

[10] Yurdakul, B., and Naranjo, J. 2020. Statistical Properties of the Population Stability Index. Journal of Risk Model Validation 14, 3:89–100.

[11] Kolmogoroff, A. 1933. Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari*, 4:83–91.

[12] Searle, S.R. 1971. *Linear Models.* Wiley, New York. 560 pages.

[13] Lohaka, H.O. 2007. Making a grouped-data frequency table: development and examination of the iteration algorithm. Doctoral dissertation, Ohio University. p. 87.

[14] Sturges, H.A. 1926. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66. doi:10.1080/01621459.1926.10502161.

[15] Lane, D.M. 2015. Guidelines for Making Graphs Easy to Perceive, Easy to Understand, and Information Rich. In: McCrudden, M.T., Schraw, G., and Buckendahl, C. (Eds.), *Use of Visual Displays in Research and Testing:*

Coding, Interpreting, and Reporting Data. Information Age Publishing, Charlotte, pp. 47–81.

[16] Lane, D.M. 2015. Histograms. Rice University, Houston.

[17] Wang, R., Shivanna, R., Cheng, D., Jain, S., Lin, D., Hong, L., and Chi, E. 2021. DCN V2: Improved Deep & Cross Network and Practical Lessons for Web-Scale Learning to Rank Systems. In Proceedings of the Web Conference 2021, 1785–1797.

[18] Chen, Z., Wang, Y., Lin, D., Cheng, D. Z., Hong, L., Chi, E. H., and Cui, C. 2021. Beyond Point Estimate: Inferring Ensemble Prediction Variation from Neuron Activation Strength in Recommender Systems. In Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21), 76–84. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3437963.3441770.

[19] Taplin, R., and Hunt, C. 2019. The Population Accuracy Index: A New Measure of Population Stability for Model Monitoring. *Risks*, 7(2).

[20] du Pisanie, J. 2023. A Critical Review of Existing and New Population Stability Testing Procedures in Credit Risk Scoring.

8. Appendix

8.1. Determining Critical Values for CPSI

There are two approaches for determining the critical values for CPSI. The first and most straightforward approach involves utilizing the normal approximation. Instead of relying on predetermined critical values, it is more advantageous to utilize the theoretical percentiles of the normal approximation distribution [10]. As seen in the preceding section, the distribution of CPSI is affected by the parameters B, N, and M. Using the normal approximation, we can determine the desired percentiles to establish the critical values.

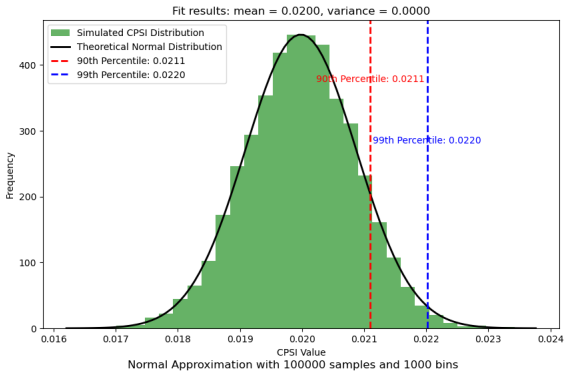

Figure 6: Normal Approximation of CPSI.

The second method involves using the empirical distribution of CPSI values collected during production. This method does not depend on hypothetical estimations, but rather utilizes actual distribution data. Through the implementation of offline simulations using historical prediction data, we can collect results to determine the critical values of CPSI. We found that using the empirical distribution of CPSI values to set critical values is more promising. The critical value should reflect the system's tolerance for instability. We have identified inherent instability resulting from variability in training neural network models with medium-sized datasets. Sorely depending on the normal approximation may lead to false alerts, as it might misinterpret natural score fluctuations as significant deviations. Hence, using empirical critical values from actual system performance data provides a more accurate and reliable stability assessment.