Exploring the Limits of Machine Learning in Near-Random Systems:

A Study of Signal Detection in the Kerala State Lottery

Niteesh Krishna M

Independent Researcher (Grade 12) | IAMNK

info@iamnk.com

Date: May 2026

Abstract

This paper investigates whether machine learning methods can detect and exploit weak signals in a system designed to produce uniformly random outcomes. Using the Kerala State Lottery—one of India's largest government-operated lottery systems—as a controlled test environment, six algorithmically distinct ranking families are evaluated across six temporal windows, producing 36 independent rankers. Each is assessed over a 91-day strictly out-of-sample evaluation period encompassing more than 28,000 book-level evaluations and over 33,000 individual winner observations. The evaluation framework combines hit-rate metrics, prize-tier-weighted scoring, return-on-investment simulation, and formal binomial hypothesis testing with Bonferroni correction for 144 simultaneous tests. Results indicate that several algorithms achieve statistically significant lift over the random baseline ($p < 10^{-6}$), with the strongest exhibiting a lift factor of 1.154. However, the detected signals are insufficient to produce positive economic returns under any realistic scenario, with ROI ranging from −54% to −85%. The central finding is that a meaningful gap exists between statistical detectability and economic exploitability—a boundary with implications for signal detection methodology in fraud detection, anomaly scoring, and other domains where the null hypothesis is randomness.

1. Introduction

Can machine learning outperform randomness in a system designed to be random? This question is foundational to applied statistics. Every machine learning system implicitly claims to identify structure in data. When the underlying data-generating process is, by design, structureless—uniformly distributed and independent across observations—any detected "pattern" must be scrutinized as a potential artifact of finite sampling, overfitting, or methodological error.

The Kerala State Lottery provides a near-ideal controlled environment for studying this question. Established in 1967 by the Government of Kerala, India, it conducts daily draws across seven lottery series with prizes distributed across nine tiers, from ₹100 to ₹10,000,000. Prize determination for most tiers is based on the last four digits of ticket numbers, creating a well-defined outcome space of 10,000 possible endings (0000–9999). The expected distribution under the null hypothesis is uniform. The data is publicly available, government-audited, and spans years of consistent operation.

This paper does not attempt lottery prediction. Rather, it investigates whether the historical distribution of winning endings exhibits any detectable departure from uniformity, and if so, whether such departures can be captured algorithmically and translated into economic advantage. The study is positioned as an empirical investigation of signal detection limits in near-random systems—a research area with direct relevance to fraud detection, financial anomaly scoring, medical screening, and any domain where distinguishing genuine signal from statistical noise is the central methodological challenge.

Existing research on lottery analysis suffers from three gaps. First, most studies report hit rates without formal hypothesis testing, without multiple comparison correction, and without economic utility analysis. Second, there is no systematic comparison of diverse algorithmic approaches on the same dataset under identical evaluation conditions. Third, papers either overstate weak signals or dismiss the domain entirely, failing to characterize the boundary between detection and exploitation. This paper addresses all three gaps through a rigorous, multi-dimensional evaluation framework applied to six distinct algorithm families.

2. Problem Definition

Given: A time series of lottery draw results $D = \{d_1, d_2, ..., d_T\}$, where each draw d_t contains a set of winning ticket endings $W_t \subset \{0000, ..., 9999\}$, with each ending associated with a prize tier and corresponding prize amount.

Produce: For each future draw d_{T+1}, a ranking function $f: \{0000, ..., 9999\} \rightarrow \mathbb{R}$ that assigns a score to each possible ending, such that the set of top-K endings contains more actual winners than expected under uniform random ranking.

Evaluate: Whether the ranking function achieves statistically significant improvement over the random baseline at multiple K thresholds, and whether such improvement translates to positive expected value at a ticket cost of ₹50.

The evaluation design enforces strict temporal causality: predictions generated at time t are evaluated exclusively against results from time t+1. No information from future draws may influence feature computation or ranking generation. This constraint eliminates data leakage—the most common source of inflated performance in prediction system evaluations.

Success is defined through a hierarchy of metrics. Statistical significance (one-sided binomial test, Bonferroni-corrected) establishes whether the signal is real. Lift factor quantifies its magnitude. ROI simulation determines whether the signal has practical value. This hierarchy prevents the common error of conflating statistical significance with practical utility.

3. Methodology

3.1 Algorithm Families

Six algorithmically distinct ranking families are employed, each capturing a different hypothesized signal source:

- **Positional Digit Analysis:** Ranks endings by combining positional digit frequencies at the hundreds, tens, and ones positions independently, then aggregating the three marginal distributions into a composite score.

- **Frequency-Based Scoring:** Scores endings by their observed frequency within a rolling temporal window, operationalizing the "hot number" hypothesis with information-theoretic normalization.
- **Gaussian Mixture Model Classification:** Applies Gaussian Mixture Models trained on high-tier winners to score endings by their digit-level feature similarity to recent prize-winning patterns.
- **Recency-Based Ranking:** Assigns scores based on the time elapsed since each ending last appeared as a winner, formalizing the "overdue number" hypothesis associated with the Gambler's Fallacy.
- **Composite Scoring:** Combines frequency, structural pattern, and positional digit signals into a single composite score.
- **Ensemble (Rank Fusion):** Aggregates rankings from multiple source algorithms using weighted Reciprocal Rank Fusion with adaptive weight learning.

Each family is evaluated across six temporal windows (15, 30, 60, 90, 180, and 365 draws), producing 36 independent rankers. The use of multiple windows tests whether detected signals are transient or persistent.

3.2 Evaluation Framework

The evaluation framework constitutes the primary methodological contribution. It measures performance across five complementary dimensions:

Hit rate at K: The fraction of actual winners appearing in the top-K ranked endings. Random baseline: K/10,000 for full-pool evaluation, or k/10 for book-level evaluation (where a "book" is a contiguous block of 10 endings sharing the first three digits).

Prize-tier-weighted scoring: Weights hit rates by prize tier importance, reflecting the extreme concentration of economic value in upper tiers (1st Prize: ₹10,000,000; 9th Prize: ₹100).

ROI simulation: Computes return on investment at ₹50 per ticket, forcing the analysis to confront whether detected signals have practical value. Both theoretical (unrestricted ticket access) and realistic (book-level, single-ticket) scenarios are evaluated.

Book-level evaluation: Measures whether the algorithm's top-ranked ending within each 10-ending book matches an actual winner. This models the realistic decision scenario: a buyer has already chosen a number range and seeks guidance on the final digit.

Statistical testing: One-sided binomial tests against the random baseline, with Bonferroni correction for 144 simultaneous tests (36 algorithms × 4 K thresholds). The corrected significance threshold is $\alpha = 0.05/144 \approx 0.00035$.

4. Experimental Setup

Parameter	Value
Data source	Official Kerala State Lottery result publications
Evaluation period	91 days, strictly out-of-sample
Total winner observations	33,332
Book-level evaluations per algorithm	28,305 (full pool) / 26,751 (reduced pool)
Algorithms evaluated	36 (6 families × 6 windows)
Book K thresholds	1, 2, 3, 5 (within 10-ending book)
Ticket price (ROI)	₹50
Statistical test	One-sided binomial, Bonferroni-corrected
Significance threshold	$p < 0.05/144 \approx 0.00035$
Average winners per draw	~367

Table 1. Experimental parameters.

The evaluation follows a strict rolling temporal design. Historical predictions are archived with timestamps, and each prediction set is evaluated against the subsequent draw's results. This folder-based archival structure makes the temporal boundary explicit and independently auditable.

The 91-day period encompasses draws from all seven Kerala lottery series, providing exposure to the full diversity of draw mechanisms. Statistical power analysis confirms that with 28,305

book-level evaluations per algorithm and a baseline rate of 10%, a one-percentage-point improvement is detectable at $\alpha = 0.001$ with power exceeding 0.99.

5. Results

5.1 Hit Rate Analysis

Book-level evaluation results are summarized in Table 2. The random baseline for top-1-in-book selection is 10.0% (1 of 10 endings).

Algorithm Family	Best Window	Hit Rate (%)	Lift	p-value
Composite Scoring	15	11.55	1.154	$< 10^{-6}$
Positional Digit Analysis	180	11.34	1.134	$< 10^{-6}$
Frequency-Based	15	10.80	1.080	1×10^{-6}
Ensemble (Rank Fusion)	15	10.77	1.077	2×10^{-6}
GMM Classification	15	10.72	1.072	7×10^{-6}
Recency-Based	15	9.93	0.993	0.661 (NS)

Table 2. Best book-level performance per algorithm family. NS = not significant.

Three findings are notable. First, positional digit analysis achieves statistical significance across all six temporal windows—the only family to do so—indicating a stable, window-independent signal in positional digit frequencies. Second, composite scoring achieves the highest single-window lift (1.154 at the 15-day window), suggesting that combining frequency, structural pattern, and positional signals captures complementary information. Third, the recency-based approach performs consistently worse than random (lift < 1.0 at all windows ≥ 30 days), providing direct empirical evidence against the Gambler's Fallacy.

5.2 Window Size Effects

A consistent pattern emerges: shorter temporal windows (15 and 30 draws) outperform longer windows (180 and 365 draws) for most algorithm families. This indicates that whatever non-uniformity exists is transient, emerging over short periods and dissipating over longer horizons. The exception is positional digit analysis, which performs stably across all windows, suggesting that positional frequencies represent a more fundamental distributional feature.

5.3 Statistical Significance

After Bonferroni correction for 144 tests:

- 17 of 36 algorithm-window combinations achieve statistical significance.
- All 6 positional digit analysis windows are significant (most robust finding).
- 4 of 6 composite scoring windows are significant.
- GMM, frequency-based, and ensemble methods are significant only at the 15-day window.
- Zero recency-based variants achieve significance.

5.4 ROI Analysis

Under the theoretical model (unrestricted ticket access), ROI is deeply negative at all K thresholds, ranging from −60% to −85%. Under the realistic book-level model (single ticket, ₹50), the best algorithm improves expected return from ₹20.00 (random) to ₹23.08 (11.54% × ₹200 average lower-tier prize)—an improvement of ₹3.08 per ticket, yielding ROI of approximately −54% versus −60% for random selection.

Neither scenario produces positive expected value. The detected signal is insufficient to overcome the structural negative expected value inherent in lottery ticket pricing.

6. Discussion

6.1 The "Detectable but Not Exploitable" Boundary

The central finding is that a meaningful gap exists between statistical detectability and economic exploitability. Multiple algorithms detect signals that survive rigorous hypothesis testing ($p < 10^{-6}$, Bonferroni-corrected), yet these signals produce deeply negative returns under any realistic

scenario. This boundary—where a signal is strong enough to be statistically real but too weak to be practically useful—is the most important result of this study.

The mathematics are instructive. A lift of 1.154 means 11.54% accuracy where 10.0% is expected. With 28,000+ evaluations, this 1.54 percentage-point difference is overwhelmingly significant in statistical terms. In economic terms, it corresponds to approximately ₹3 additional expected value on a ₹50 ticket—an improvement that is both real and useless.

6.2 Why Statistical Significance Does Not Imply Utility

Three factors compound to prevent the translation of detected signals into economic returns:

- **Prize concentration:** Prize value is concentrated in the top three tiers (1st Prize: ₹10,000,000), which collectively produce only 2–3 unique winning endings per draw. The algorithmic signal is strongest in lower tiers (7th–9th prizes, ₹100–500), which are numerically abundant but economically negligible.
- **Physical constraints:** The assumption that a buyer can purchase any arbitrary set of K tickets is unrealistic. Lottery tickets are distributed across thousands of retail outlets, and no buyer can assemble a targeted ending portfolio. The book-level evaluation models the realistic scenario more faithfully.
- **Structural negative EV:** The lottery is structurally designed as a negative expected-value proposition. No small-magnitude signal can overcome the gap between ticket cost (₹50) and expected return (~₹20 per ticket under random selection).

6.3 Implications for the Gambler's Fallacy

The recency-based algorithm provides a clean empirical test of the Gambler's Fallacy—the belief that outcomes which have not occurred recently are "overdue" and therefore more likely. This approach performs consistently and significantly worse than random, indicating that recently absent endings are, if anything, slightly less likely to appear. This finding is consistent with weak positive autocorrelation in ending frequencies and directly contradicts one of the most common heuristics used in lottery ticket selection.

6.4 Implications for Ensemble Methods

The ensemble approach (rank fusion with adaptive weight learning) does not consistently outperform its best individual component. Inclusion of the anti-predictive recency signal dilutes

ensemble quality. This finding has implications for ensemble design in low-signal environments: component quality dominates component quantity. An anti-predictive member actively degrades the ensemble, and adaptive weighting cannot fully compensate.

6.5 Broader Implications

The boundary between detectable and exploitable signals is not unique to lotteries. In fraud detection, medical diagnostics, and financial markets, practitioners routinely encounter statistical signals that are real but too weak to act upon profitably. The evaluation framework presented here—combining statistical testing, economic simulation, and multiple comparison correction—provides a template for rigorously characterizing this boundary in any domain where the null hypothesis is randomness.

7. Conclusion

This study demonstrates that machine learning methods can detect weak but statistically genuine signals in a system designed to produce uniformly random outcomes. Multiple independent algorithms, using fundamentally different signal sources, achieve significance at $p < 10^{-6}$ after Bonferroni correction—a finding that is difficult to attribute to chance or methodological artifact.

However, the detected signals are too small to overcome the structural constraints of the lottery system. Statistical significance does not imply economic utility. A p-value of 10^{-6} is impressive as a measure of confidence, but when the underlying effect is a 1.5 percentage-point improvement on a ₹50 ticket, it is scientifically interesting and economically irrelevant. The entire literature on algorithmic performance in near-random systems must reckon with this distinction.

Three specific findings merit emphasis:

- Positional digit frequency analysis detects a stable, window-independent signal in the Kerala lottery ending distribution, suggesting subtle structural non-uniformity in the prize allocation process.
- The Gambler's Fallacy is empirically refuted: recency-based selection performs worse than random, indicating that "overdue" outcomes are not more likely.

- Ensemble methods do not automatically improve upon their best component in low-signal environments; anti-predictive members actively degrade ensemble quality.

The primary contribution of this work is not the detection of weak signals in a lottery system. It is the rigorous, multi-dimensional evaluation methodology for characterizing the boundary between detectable and exploitable signals in near-random systems. This methodology—combining hit-rate analysis, prize-tier weighting, economic simulation under realistic constraints, consistency measurement, and formal hypothesis testing with multiple comparison correction—is applicable to any domain where distinguishing genuine signal from noise is the central challenge. In a landscape of machine learning systems that routinely claim to find signal in noise, having a rigorous method to determine whether that signal actually matters is more valuable than the signal itself.

References

Clotfelter, C. T., & Cook, P. J. (1993). The "gambler's fallacy" in lottery play. Management Science, 39(12), 1521–1525.

Cormack, G. V., Clarke, C. L. A., & Büttcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. Proceedings of the 32nd International ACM SIGIR Conference, 758–759.

Haigh, J. (1997). The statistics of the National Lottery. Journal of the Royal Statistical Society: Series A, 160(2), 187–206.

Hand, D. J. (2006). Classifier technology and the illusion of progress. Statistical Science, 21(1), 1–14.

Joe, H. (1993). Testing for uniformity in multi-dimensional data. Canadian Journal of Statistics, 21(3), 269–279.

Reynolds, D. A. (2009). Gaussian mixture models. Encyclopedia of Biometrics, 741–749.

Stern, H. S., & Cover, T. M. (1989). Maximum entropy and the lottery. Journal of the American Statistical Association, 84(408), 980–985.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. Psychological Bulletin, 76(2), 105–110.