

Forecasting Daily Online Attention to K-Pop Artists: A Retrospective Comparison of Baselines, Ridge Regression and Random Forests

Frankie Gao

perspectiveopen.com

Correspondence: zcfbfga@ucl.ac.uk

Abstract

Daily public attention measures offer a low-cost basis for audience monitoring, but forecast improvements must be assessed against strong simple methods and with chronological information boundaries. This study compares persistence, seasonal naive and moving-average baselines with ridge regression and random forests using English Wikipedia user-classified pageviews for BTS, BLACKPINK, TWICE, Stray Kids, SEVENTEEN and aespa. The primary dataset contains 10,404 artist-day observations from January 2022 to September 2026. Parameters were selected using eight expanding quarterly validation blocks in 2024–2025. Retrospective evaluation used 1,638 artist-days in January–September 2026, with models fixed before the first forecast origin and lagged inputs updated as observations became available. Log-target ridge regression achieved a mean absolute error of 512.50 pageviews, compared with 603.18 for persistence, a 15.03% reduction (paired 14-day block-bootstrap 95% interval: 6.14%–21.51%). The development-selected log-target forest achieved 551.13 pageviews, an 8.63% reduction (–2.95%–18.20%). Both improved the point-estimate error for all six primary artists, although monthly gains were uneven. Forest error was stable across five random seeds; removing the latest count raised MAE by 33.50%. Only 26 large-traffic artist-days accounted for 38.33% of forest absolute error. Seven-day forecasts were less accurate, and transfer to two artists excluded from fitting did not establish a reliable advantage over persistence. These findings support transparent, lightweight forecasting for aggregate attention monitoring while showing that nonlinear models need not outperform a well-specified linear comparator. Pageviews measure article visits, not unique fans, music consumption or campaign effects.

Keywords: K-pop; online attention; Wikipedia pageviews; time-series forecasting; random forest; ridge regression; chronological evaluation

1. Introduction

Online publicity teams encounter a practical planning problem: attention changes from day to day, while decisions about monitoring, reporting and content preparation must be made before the next day’s activity is observed. A useful forecast need not explain every cultural event. It can provide a transparent estimate of the attention level that would be expected from recent history, allowing a team to recognize departures from that expectation and organize its workload. The relevant research question is therefore concrete: how much information about tomorrow’s aggregate article traffic is contained in the preceding days, and does a relatively simple machine-learning model improve on repeating the latest observed count?

English Wikipedia provides a consistent public measure for examining this problem. The Wikimedia Analytics API reports article-level pageviews with explicit choices of project, access method, agent category and temporal granularity [1]. Its numerical data are released under CC0 [2]. This combination makes the measurement and computations inspectable without access to private advertising accounts or fan-platform records. The measure nevertheless has a specific meaning. A pageview is a recorded visit to an article; the same person can contribute multiple visits, and reading may reflect curiosity, news exposure or information seeking. The variable is consequently described here as article attention rather than fandom size or commercial demand.

The focus on K-pop artists supplies a recognizable applied setting while keeping the data requirement modest. The six primary artist series were taken from an existing pilot: BTS, BLACKPINK, TWICE, Stray Kids, SEVENTEEN and aespa. This is a purposive sample of group-level articles. It is not a census of Korean entertainment and does not include solo-member articles, local-language projects or social-platform interactions. The sample is useful because it contains different traffic levels and enough daily history for repeated chronological evaluation. A forecasting result for these series can inform a small monitoring exercise; it cannot, by itself, establish a general account of how the entire music industry responds to publicity.

Three difficulties make an apparently straightforward prediction task scientifically meaningful. First, a slowly changing series can make a complex model look effective even when persistence performs nearly as well. Second, large attention surges can dominate aggregate error and obscure performance on more ordinary days. Third, mixing future observations into model fitting or feature construction can produce a comparison that could not have been carried out at the stated forecast origin. A credible study therefore needs competitive baselines, results in the original units, and a precise account of when every input and fitted target becomes available. These requirements determine the design used here.

The analysis evaluates more than a single random-forest fit. It includes three simple one-day baselines, a regularized linear model, two target representations for each learned model, eight validation quarters, and a final nine-month comparison. The main forest is selected using development performance rather than final-period error. Repeated seeds test whether the forest result depends on a favorable random draw. Feature ablations examine which parts of the predictor set carry useful information for that fitted system, while shorter training windows test sensitivity to historical coverage. A separate seven-day specification examines how the task changes when the most recent observable count is farther from the target.

A second extension concerns artist transfer. A pooled model may learn relationships that work for artists already present in training but fail for another artist with a different traffic level. IVE and LE SSERAFIM are therefore evaluated without contributing any targets to model fitting or parameter selection. Their own

past observations remain available as inputs, making this a test of transfer of the fitted relationship rather than forecasting an artist with no history. This distinction matters for practical use: adding a new monitored article is easier than learning from no observations at all, but it still exposes weaknesses in the treatment of artist identity and scale.

The contribution is an empirical comparison and an auditable research workflow, rather than a new forecasting algorithm. The study asks four connected questions: whether learned models reduce one-day error against persistence; whether any improvement is consistent across artists and months; whether the selected forest is stable under seed, feature and training-window changes; and whether its behavior extends to a longer horizon and artists excluded from fitting. All results arise from executed local computations on public aggregate data. The analysis deliberately reports weak and negative findings alongside favorable ones, because deciding when a model adds little value is part of evaluating a monitoring method.

2. Related Work and Analytical Rationale

Wikipedia activity has previously been studied as a signal of public attention. Mestyán, Yasseri and Kertész used Wikipedia activity in a model of movie box-office outcomes [3]. That work supports the broader idea that publicly recorded information-seeking behavior can contain predictive information, but its outcome, features and setting differ from this study. Predicting daily visits to an artist article is not equivalent to predicting ticket sales, and the relationship between article traffic and music consumption is not tested here. The present study keeps the dependent variable within the same measurement system instead of treating pageviews as a validated proxy for another outcome.

Ridge regression and random forests represent two accessible modeling choices. Ridge regression penalizes coefficient magnitude, which is useful when recent lags and rolling summaries contain overlapping information [4]. Random forests average predictions from multiple fitted trees and can represent nonlinear interactions without specifying them manually [5]. These capabilities motivate a comparison; they do not establish that either model must win. The data may mostly favor a smooth relationship between recent and subsequent counts, in which case a regularized linear predictor can be competitive. Conversely, a forest may adapt to recurrent combinations of level and recent variability while failing to extrapolate to an abrupt new peak.

Forecast evaluation depends on the intended information set. Cerqueira, Torgo and Mozetič studied performance-estimation methods for time series and found that the suitability of validation schemes depends on characteristics such as nonstationarity [6]. Hewamalage, Ackermann and Bergmeir likewise emphasize matching evaluation design, baselines and error measures to the forecasting problem [7]. The present study uses successive calendar blocks because the intended task is prediction of later days from earlier information. It does not claim that randomized validation is invalid for every autoregressive setting; it chooses chronological evaluation because changing traffic levels and an operational next-day interpretation are material to this application.

The target representation is also part of the modeling question. Fitting squared error to original counts and fitting squared error to log-transformed counts are different objectives. Exponentiating a predicted mean on the log scale does not automatically produce an unbiased mean count, nor does it guarantee a conditional median. Gneiting’s discussion of point forecasts and scoring functions provides the relevant distinction between a predicted functional and the score used to assess it [8]. Here all methods are evaluated on the

original count scale, and both MAE and RMSE are reported. The transformations are treated as empirical specifications rather than mathematical guarantees about the best possible forecast.

3. Data and Methods

3.1. Data source, measurement and sample

Daily extracts were requested from the Wikimedia Analytics API using `project en.wikipedia.org`, access all-access, agent user and granularity daily [1]. The user category is the API classification of traffic, not a unique-person identifier. All dates are UTC calendar dates. Six fixed article titles define the primary sample: BTS, Blackpink, Twice, Stray_Kids, Seventeen_(South_Korean_band) and Aespa. Artist display names are standardized separately from these source identifiers. The numerical API extracts are CC0 data [2]. The study uses counts only; it does not reproduce article text, images, biographies or identifiable visitor records.

The primary observation window runs from 1 January 2022 through 30 September 2026 inclusive. Each artist contributes 1,734 daily observations, yielding 10,404 artist-days. The earlier pilot supplied 6,024 records from January 2024 onward, and 4,380 additional records extended training history through January 2022. The two parts were combined on artist and date without averaging or filling records. Raw JSON responses, the preserved pilot CSV, the combined CSV and source metadata are retained in the accompanying research folder. Request URLs, source titles, record counts and SHA-256 digests permit a reader to distinguish the exact extract from a later API retrieval.

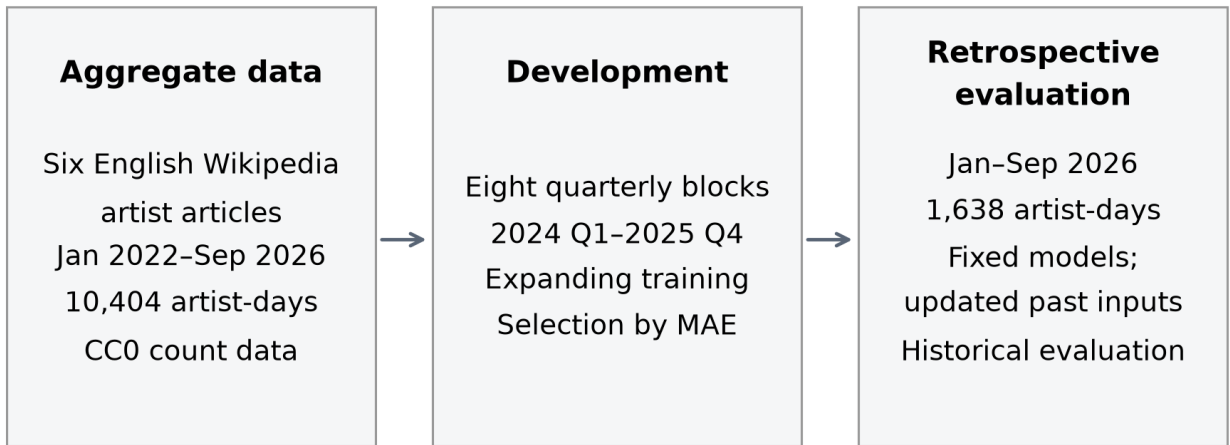
Data checks required one record for every artist-date combination in the stated range, no duplicate artist-date keys, no missing counts, and nonnegative integer values. These conditions passed for the primary series; all observed primary counts were positive. No imputation, winsorization, removal of high-traffic dates or synthetic augmentation was applied. The additional IVE and LE SSERAFIM extracts cover 1 January 2025 through 30 September 2026, producing 1,276 records in total. Their 2025 observations provide lag history and an error-scale reference, while the 2026 comparison contains 546 artist-days. These artists are absent from all model-fitting and parameter-selection rows.

This is a retrospective study. The primary 2026 records had been examined in an earlier pilot, and the additional artist extracts were accessed during pipeline verification. The chronological separation described below prevents future labels from entering a fitted forecast, but the evaluation is not presented as a prospectively registered or previously untouched experiment. Model parameters for the reported pipeline are selected from 2024–2025 validation performance, and neither the additional artists nor 2026 target errors determine the reported parameter choices. This distinction concerns the interpretation of the evidence, not the availability of the data or the execution of the analyses.

Table 1. Primary sample and traffic levels

Artist	Mean daily views	Median daily views	Maximum daily views	Training 95th percentile	2026 test views
BTS	10,512.3	9,433.5	140,144	19,104	2,604,815
BLACKPINK	7,413.3	6,159.0	50,338	14,965	1,304,094
TWICE	2,848.8	2,698.0	17,371	5,061	574,687
Stray Kids	3,137.2	3,009.0	20,174	4,768	631,104
SEVENTEEN	2,303.3	2,083.0	16,238	4,359	343,783
aespa	1,764.5	1,599.5	15,261	2,962	433,032

Each artist has 1,734 observations. Mean, median and maximum use January 2022–September 2026; the threshold uses January 2022–December 2025; test totals use January–September 2026. All values count article visits.



Forecast information boundary

For target day t at horizon h , count inputs end at $t - h$.
Fitted targets end no later than the first forecast origin.

Robustness and transfer

Five forest seeds; feature ablations; training-window and seven-day checks
Paired date-block bootstrap; IVE and LE SSERAFIM excluded from fitting

Figure 1. Study design and information boundaries. Primary evaluation uses historical 2026 observations. Two additional artists are excluded from model fitting and parameter selection; their past counts are available as forecast inputs. Source: Wikimedia Analytics API; local calculations.

3.2. Forecast origins and predictor construction

Let a identify an artist, t identify a target UTC day, and h denote the forecast horizon in days. The available count history ends at t minus h . The one-day task predicts the count for day t using observations no later than day t minus one; the seven-day task ends the available count history at day t minus seven. The target weekday is known from the calendar in both cases. The fitted model is kept fixed throughout an evaluation block, while the historical inputs are updated at successive origins. Consequently, the experiment is a sequence of fixed-horizon daily forecasts, not a single nine-month forecast issued at the end of 2025.

$$\hat{y}_{a,t|t-h} = f_h(\mathbf{x}_{a,t-h}, a, \text{weekday}(t)) \quad (1)$$

Four individual lag predictors use counts at target offsets h , $h + 6$, $h + 13$ and $h + 27$ days. For the one-day task these are lags 1, 7, 14 and 28. Two rolling means use the last 7 and 30 available observations, and a rolling standard deviation uses the last 7 available observations with the sample standard-deviation convention. Two calendar predictors encode the target weekday with sine and cosine terms. A one-hot artist identifier completes the predictor set. Count-based and rolling predictors are transformed with $\log(1 + x)$ before fitting every learned specification; weekday terms remain unchanged. Appendix A states the feature windows explicitly.

Feature construction sorts each artist’s records by date before shifting and rolling. Rolling windows are applied to the shifted count series, so their final observation is always the latest count available at the stated origin. Warm-up rows lacking a full 30-observation window are excluded: 30 days per artist for the one-day task and 36 for the seven-day task. These exclusions affect fitting eligibility but do not remove any date from the 2026 evaluation. A direct source-window check recomputes the latest count and both rolling means. A future-perturbation check also replaces later observations with very large values and verifies that earlier predictor values remain unchanged at both horizons.

The fitting boundary follows the first forecast origin, not merely the first target date. For the primary one-day evaluation, fitted target observations end on 31 December 2025. For the seven-day evaluation, they end on 25 December 2025, seven days before the first target of 1 January 2026. This prevents the first seven-day forecasts from using fitted labels that would still be in the future at their origin. The same rule is applied at every validation-quarter boundary. A saved grid records the last fitted target and first validation origin, and an independent check verifies their ordering for every candidate and fold.

3.3. Baselines and learned specifications

Persistence repeats the most recent count available at the forecast origin. Seasonal naive repeats the count from the same target weekday in the preceding week. The seven-day-mean baseline averages the last seven observations available at the origin. For $h = 1$, these supply distinct comparisons: yesterday’s count, the count seven days earlier and a smoothed recent level. For $h = 7$, persistence and seasonal naive both equal the count at t minus seven, so the duplicate seasonal-naive result is omitted. Baselines receive the same historical availability boundary as the learned methods and require no fitted hyperparameters.

Ridge regression is fitted in two target forms: original pageview counts and $\log(1 + \text{pageviews})$. Numeric predictors are standardized using training rows only, and artist indicators are one-hot encoded. Candidate penalty strengths are 1, 10, 100 and 1,000. The unpenalized intercept is retained. Log-target predictions are returned to the count scale using $\exp(\text{prediction})$ minus one, and all learned predictions are bounded below

by zero. No residual-based retransformation adjustment is applied. Predictions remain real-valued count estimates rather than being rounded to integers before scoring. The two ridge specifications share transformed predictors, making their target representation an explicit part of the comparison.

Random forests also use original-count and log-transformed targets. Each fit has 200 trees, bootstrap sampling, squared-error splitting and all available transformed features considered at each split. Three depth-and-leaf configurations are evaluated: maximum depth 6 with at least 10 samples per leaf; depth 12 with at least 5; and unrestricted depth with at least 20. Numeric predictors are not standardized for forests, while the same artist encoding is retained. Seed 42 defines the primary model fit. The grid is intentionally small enough to inspect and reproduce locally; conclusions concern these stated configurations, not every possible forest architecture or optimization objective.

3.4. Development selection and retrospective evaluation

Eight expanding validation blocks span 2024 Q1 through 2025 Q4. Each candidate is fitted to eligible targets ending no later than the first origin of that quarter and evaluated on every target date within the quarter. Models are refitted between quarters but remain fixed within a quarter; available historical predictors continue to update. Each quarter contains 90–92 target days per artist. The selection score is the arithmetic mean of the eight quarterly MAEs, giving every quarter equal weight. The minimum score selects one parameter configuration within each model family and horizon. This rule is independently recomputed from the saved validation grid.

The one-day log forest has the lower development MAE of the two forest target forms and is designated the primary forest. Its selected parameters are unrestricted depth and a minimum of 20 samples per leaf. The count-target forest selects the same configuration. One-day ridge selects penalty 1,000 for counts and 10 for the log target; seven-day ridge selects 1,000 in both target forms. Seven-day forests retain unrestricted depth and leaf size 20. The configuration file records these choices, the data and analysis-code digests, horizon-specific fitting cutoffs, the fixed primary seed and the validation quarters. All competing families remain in the final results rather than reporting only the designated forest.

Final primary evaluation covers 273 target days from 1 January through 30 September 2026 for all six artists, yielding 1,638 forecasts per method and horizon. The one-day learned models use 8,586 fitted rows; the seven-day models use 8,514. The data split is by time across the full artist panel, avoiding training on a later date for one artist while testing an earlier date for another. Calendar-based predictors are permitted because their values are already determined at the origin. No future count, contemporaneous target traffic, release-day label or later social-media measurement enters the predictor matrix. Hyperparameter selection does not use final-period errors.

3.5. Error measures and paired uncertainty estimates

MAE is the main selection and reporting metric, expressed as pageviews per artist-day. RMSE supplies a second perspective that gives larger errors more influence. WAPE divides total absolute error by total observed traffic, and mean bias is defined as prediction minus observation so that negative values indicate underprediction. These measures are calculated from unrounded predictions in the original count units. WAPE is a traffic-normalized aggregate error ratio; it is not a classification accuracy or a percentage of fans correctly predicted. Reporting more than one metric helps reveal when a model behaves differently on ordinary observations and large peaks.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|, \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (2)$$

$$\text{WAPE} = \frac{\sum_{i=1}^N |y_i - \hat{y}_i|}{\sum_{i=1}^N y_i}, \quad \text{Bias} = \frac{1}{N} \sum_{i=1}^N (\hat{y}_i - y_i) \quad (3)$$

MASE normalizes each absolute error by that artist’s mean absolute consecutive-day change in the pre-2026 reference history [9]. The primary reference spans 2022–2025; the additional artists use their own 2025 histories. This denominator is held fixed when comparing training windows and horizons, so changes in MASE reflect forecast errors rather than a changing normalization. A value below one indicates an error below that historical scale, not necessarily superiority to persistence on the 2026 test period. Artist-specific MAE and WAPE are also reported to retain the original measurement meaning and prevent a pooled score from hiding a weaker result for a smaller series.

$$q_a = \frac{1}{T_a - 1} \sum_{s=2}^{T_a} |y_{a,s} - y_{a,s-1}|, \quad \text{MASE} = \frac{1}{N} \sum_{i=1}^N \frac{|y_i - \hat{y}_i|}{q_{a_i}} \quad (4)$$

Relative MAE reduction is calculated against the matching persistence forecast on the same observations. Artist-level and monthly reductions use their own corresponding persistence denominators. The pooled MAE averages equally over artist-days because every artist contributes the same number of evaluated dates. This gives each artist the same observation count but does not equalize the size of errors measured in counts. A separate mean of the six artist-specific percentage reductions is therefore reported as a complementary view of scale heterogeneity. It answers a different question from the pooled reduction and should not be substituted for it.

$$\text{MAE reduction}(m) = 100 \left(1 - \frac{\text{MAE}_m}{\text{MAE}_{\text{persistence}}} \right) \% \quad (5)$$

Uncertainty is estimated with a paired circular moving-block bootstrap of evaluation dates, building on block-resampling methods for dependent observations [10]. For each date, absolute errors are first averaged over the six primary artists. Blocks of 14 successive dates are sampled with replacement, wrapping at the end of the evaluation period, until 273 dates are obtained. The same sampled dates are used for all models, preserving within-date artist dependence and model pairing. Two thousand replicates generate percentile 95% intervals for average error and relative reduction versus persistence. Block lengths 7 and 28 provide sensitivity checks; the fixed random seed for a given length is 20261002 plus that length.

The intervals describe variability across the observed historical date sequence conditional on this artist sample and fitted pipeline. They are not per-day prediction intervals and do not include model-selection uncertainty, new-artist sampling uncertainty or a guarantee of coverage under a future regime shift. Resampling whole date blocks addresses short-range temporal dependence more plausibly than treating all artist-days as independent, but it does not make a changing historical series stationary. A further paired comparison calculates the log forest minus log ridge MAE difference using the same 14-day date blocks. Additional-artist pooled intervals are computed in the same way over two artists and are interpreted as an exploratory transfer comparison.

3.6. Robustness, transfer and computational checks

Five forest seeds—7, 19, 42, 73 and 101—are evaluated under the selected parameters for both target representations. Every seed is reported; the smallest final-period error is not used to redefine the primary fit. Feature ablations refit the selected one-day forest after separately removing the latest count, earlier individual lags, rolling summaries or artist identifier. The grid and seed remain unchanged, making this a controlled specification comparison rather than additional parameter tuning. Ablations can reveal dependence of this model on a predictor group, but correlated inputs mean they are not a unique attribution of information or a causal effect of a feature.

Training-window checks restrict the eligible fitted target dates to 48, 24 or 12 months ending in December 2025. The underlying historical series remains available to form lagged inputs at the beginning of each window; the change concerns which target rows are fitted. The seven-day models are constructed directly at that horizon, rather than recursively feeding one-day predictions back into the predictor matrix. For traffic-regime evaluation, a large-traffic artist-day is one whose observed count exceeds its artist’s 95th percentile from 2022–2025. The threshold is fixed from earlier data, but regime assignment uses the realized test outcome and is performed only after forecasts are generated.

Artist transfer applies the six-artist fitted one-day models to IVE and LE SSERAFIM. Their recent counts and target weekdays are provided, but their targets never enter fitting or parameter selection. Unknown artist categories are handled by zeroing all fitted one-hot artist columns. This implementation is explicit because its behavior can matter: an unseen artist does not receive a newly estimated intercept, and it is not assigned a known artist’s indicator. The additional-artist comparison therefore tests both transfer of the count relationship and the existing unknown-category handling. It does not test a separately adapted or fine-tuned model for these two groups.

The computations used Python 3.12.14, pandas 2.2.3, NumPy 2.3.5 and scikit-learn 1.9.1 on the local macOS machine [12]. Saved predictions support independent recomputation of every headline MAE. Further checks verify complete date grids, unchanged past features after future-count perturbation, direct rolling-window identities, the fitting boundary at every forecast horizon, and agreement between saved validation scores and selected parameters. These checks address concrete leakage and reproducibility risks rather than treating successful model execution as proof of validity; leakage research provides the broader methodological motivation [11]. The package preserves the source data, configurations, executed outputs, figure code and a notebook that refits and reconciles the primary comparison.

4. Results

4.1. Historical traffic and evaluation composition

The six articles differ substantially in traffic scale. Across the full window, BTS averages 10,512.3 daily pageviews and BLACKPINK 7,413.3, compared with 1,764.5 for aespa and 2,303.3 for SEVENTEEN. The largest full-window daily count is 140,144 for BTS. Monthly medians in Figure 2 reveal changing levels within the historical period, while the shared logarithmic axis retains comparability across artists. These descriptive summaries use all observations to characterize the sample; none is introduced as a full-window predictor. The differences justify reporting both original-count errors and artist-normalized comparisons instead of drawing conclusions from a single pooled percentage.

The primary evaluation contains 5,891,515 observed pageviews. BTS contributes 44.21% of this traffic and 60.75% of persistence absolute error, despite representing one sixth of the evaluated artist-days. This concentration shows why equal numbers of observations do not imply equal influence on a count-scale loss. A model can produce a visible pooled improvement by handling BTS better while providing only a modest change for the other series. The artist-level results below therefore distinguish the pooled error reduction from the mean artist-specific percentage reduction. Neither weighting is intrinsically universal; each corresponds to a different practical emphasis.

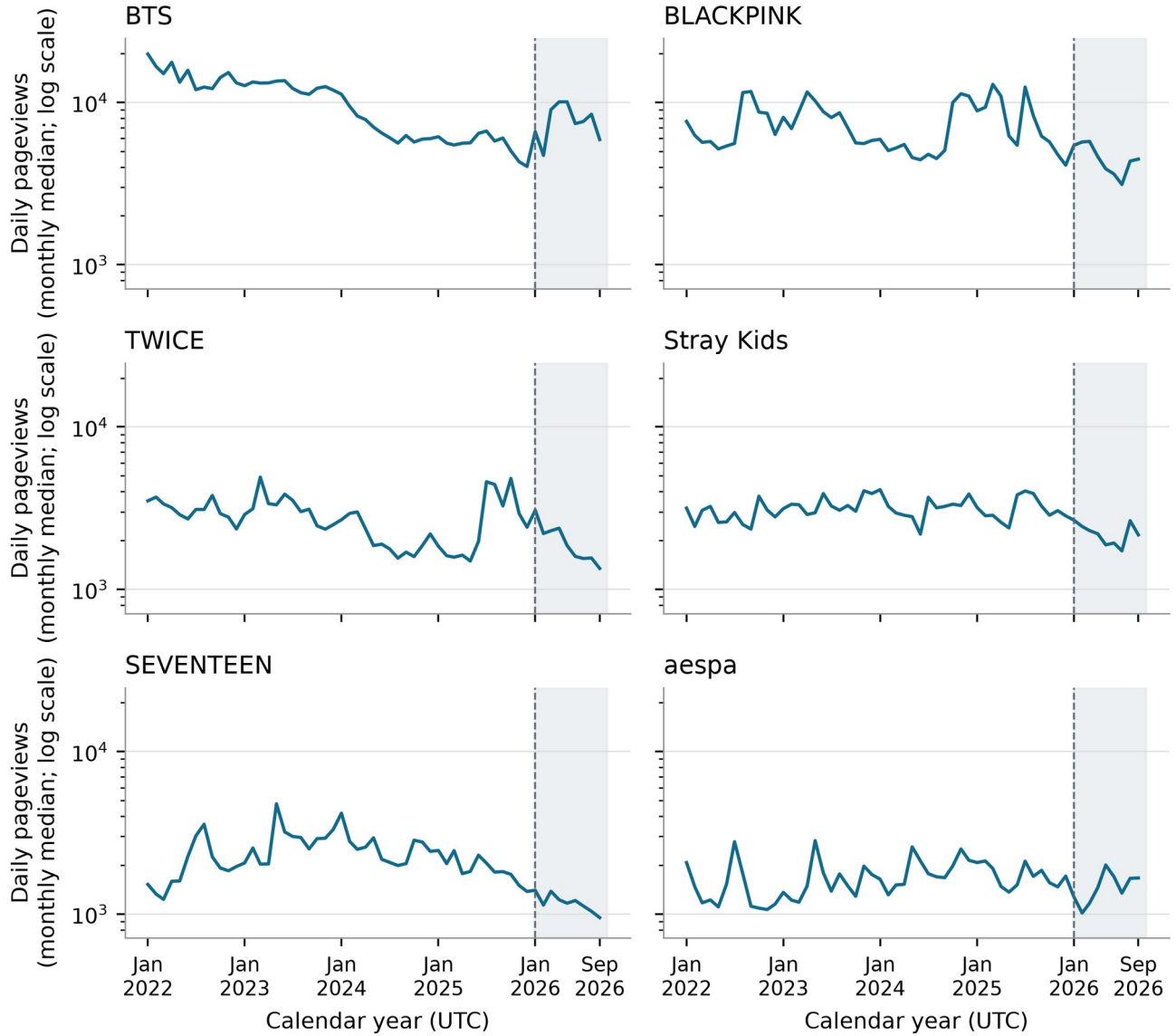


Figure 2. Monthly medians of daily user-classified pageviews, January 2022–September 2026. All panels share the same logarithmic vertical scale. Shading marks the 2026 evaluation period; monthly aggregation is used only for display. Source: Wikimedia Analytics API; local calculations.

4.2. Overall one-day comparison

Persistence achieves an MAE of 603.18 pageviews per artist-day. Seasonal naive is substantially worse at 1,118.15, and the seven-day mean reaches 789.14. Thus, recent article traffic contains useful short-term

persistence, while simply moving one week back or smoothing the recent history does not improve the primary comparison. This result is an essential benchmark for the learned methods: a model must improve on the latest observable count, not merely on a less suitable seasonal comparator. The large gap between persistence and seasonal naive also warns against assuming that a weekly calendar rhythm is the dominant source of predictive information for these series.

Log-target ridge regression has the lowest evaluated MAE, 512.50, and the lowest RMSE, 2,909.40. Its WAPE is 14.25% and MASE is 0.778. Relative to persistence, it reduces MAE by 90.68 pageviews or 15.03%. The development-selected log forest reaches MAE 551.13, WAPE 15.32% and MASE 0.836, improving MAE by 52.05 pageviews or 8.63%. The count-target forest reaches MAE 592.04, only 1.85% below persistence. The count-target ridge specification performs poorly, with MAE 1,293.87. These values describe the stated feature and target specifications, not a general failure of linear regression or a general advantage of forests.

Table 2. One-day forecast performance in January–September 2026

Method	MAE	RMSE	WAPE	MASE	MAE reduction
Persistence	603.18	3,310.24	16.77%	0.889	+0.00%
Seasonal naive	1,118.15	4,720.60	31.09%	1.568	-85.38%
7-day mean	789.14	3,441.98	21.94%	1.110	-30.83%
Ridge (counts)	1,293.87	3,370.64	35.97%	2.601	-114.51%
Ridge (log1p)	512.50	2,909.40	14.25%	0.778	+15.03%
RF (counts)	592.04	3,036.57	16.46%	0.906	+1.85%
RF (log1p)	551.13	3,084.27	15.32%	0.836	+8.63%

Each method evaluates the same 1,638 artist-days. MAE and RMSE are in pageviews; reduction is relative to persistence. MASE uses the fixed artist-specific pre-2026 scale. RF denotes random forest; log1p denotes $\log(1 + \text{count})$ target fitting.

The 14-day paired bootstrap interval for ridge’s relative improvement is 6.14%–21.51%, while the forest interval is –2.95%–18.20%. The historical comparison therefore supplies clearer evidence of an improvement for log ridge than for the forest. The point-estimate forest improvement is positive, but the paired interval includes no improvement. This interpretation does not follow from overlap between the MAE intervals drawn in Figure 3; it follows from the bootstrap distribution of paired model-versus-persistence differences. The distinction is relevant because every method faces the same difficult dates, creating substantial correlation between its errors and the baseline errors.

The further paired log-forest-minus-log-ridge MAE difference is 38.63 pageviews, with a 14-day bootstrap interval of 3.47–95.79. This supports log ridge’s lower historical error within that comparison, subject to the interval’s conditional and exploratory interpretation. Results are stable to changing the date-block length: ridge’s improvement intervals remain positive at 7, 14 and 28 days, while the forest intervals include zero at each length. No multiplicity-adjusted claim is made across every model, subgroup and robustness comparison. Appendix C provides the block-length sensitivity values so that the main statement can be assessed without relying on a single resampling setting.

The selected log forest had a slightly lower development score than log ridge: 458.78 versus 462.20 pageviews, based on the mean of eight quarterly MAEs. Their ranking reverses in the later evaluation period. This change is a useful finding rather than an inconsistency. The validation difference is small, traffic changes over time, and the models use different inductive assumptions. A model that wins by a few pageviews in development should not be treated as a guaranteed future winner. Reporting the full comparator set prevents the development-selected forest from becoming the only visible result when the later evidence favors the linear alternative.

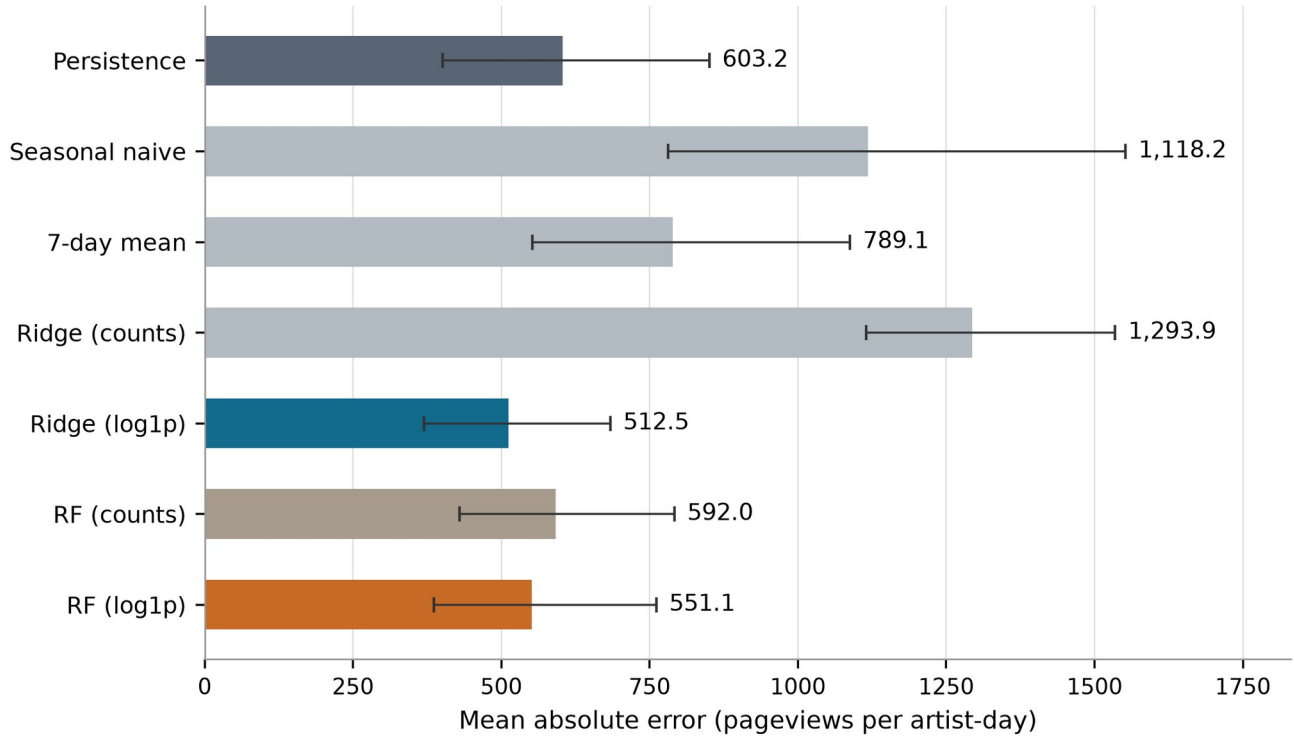


Figure 3. One-day mean absolute error on 1,638 artist-days in January–September 2026. Whiskers are percentile 95% intervals from 2,000 paired circular date-block resamples with block length 14 days. These are intervals for average error, not prediction intervals. Source: Wikimedia Analytics API; local calculations.

4.3. Artist and month heterogeneity

Both log-target methods improve the MAE point estimate for each of the six primary artists. Ridge’s reductions range from 5.14% for BLACKPINK to 19.48% for BTS, with intermediate gains of 13.29% for SEVENTEEN, 6.03% for Stray Kids, 9.54% for TWICE and 11.07% for aespa. The log forest’s reductions are smaller: 4.32%, 11.96%, 2.29%, 2.32%, 3.49% and 3.91%, respectively, in the same artist order. The mean artist-specific reduction is 10.76% for ridge and 4.72% for the forest, compared with their pooled reductions of 15.03% and 8.63%. The gap between these summaries reflects the greater original-count influence of BTS.

Table 3. Artist-specific one-day errors and relative improvements

Artist	Persistence MAE	Log ridge MAE	Log forest MAE	Ridge reduction	Forest reduction
BTS	2,198.77	1,770.51	1,935.72	19.48%	11.96%
BLACKPINK	514.33	487.89	492.10	5.14%	4.32%
TWICE	254.31	230.05	245.43	9.54%	3.49%
Stray Kids	235.94	221.71	230.47	6.03%	2.32%
SEVENTEEN	219.74	190.53	214.72	13.29%	2.29%
aespa	196.00	174.30	188.33	11.07%	3.91%

Each row evaluates 273 dates. Reductions are computed from that artist’s persistence MAE; they are descriptive point estimates, not artist-specific significance tests.

The count forest illustrates how a pooled result can hide divergent behavior. It improves MAE for only one primary artist despite a pooled reduction of 1.85%, and its mean artist-specific percentage reduction is negative. Seasonal naive and the seven-day mean fail to improve on persistence for any of the six artists. Figure 4 displays all these contrasts using artist-specific denominators. The numerical labels are important because the poorly performing count ridge creates a broad range of percentage changes; a cell’s color alone should not be used to judge a small favorable difference. The comparison preserves weak results instead of dropping specifications after observing their final scores.

Monthly comparisons are also uneven. Log ridge improves on persistence in seven of nine months; the log forest improves in six. Ridge is slightly worse in January and June, while the forest is worse in January, March and June. In March the forest MAE is 1,202.77 compared with 1,184.23 for persistence, whereas ridge reaches 938.70. In July both improve on the much higher persistence error: 858.46 for the forest and 889.83 for ridge, against 1,189.80. The aggregate nine-month ranking consequently combines months with different relative strengths. These monthly summaries are diagnostic descriptions of the fixed models, not a rule that selects a different method after observing each month.

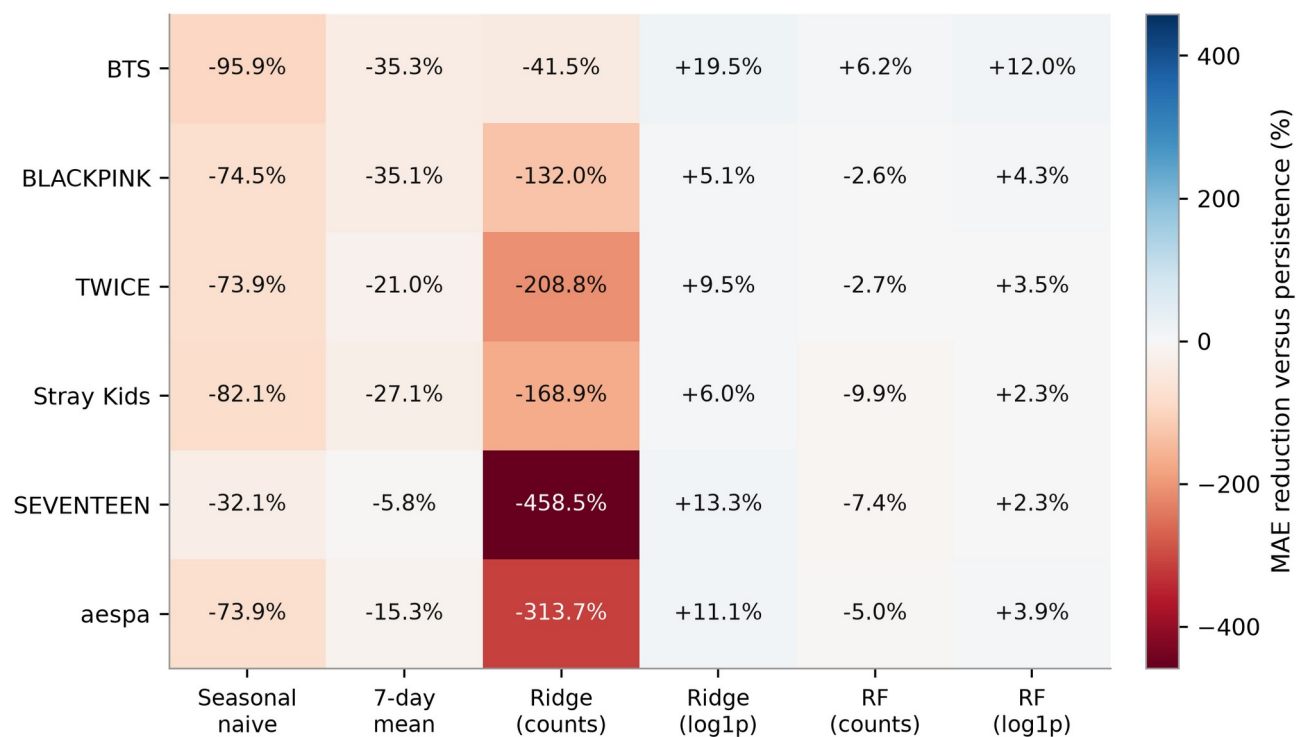


Figure 4. Artist-specific one-day MAE reductions relative to persistence. Positive percentages indicate lower error. Each cell uses 273 artist-days and its own artist-specific persistence denominator. The diverging color scale is symmetric around zero. Source: Wikimedia Analytics API; local calculations.

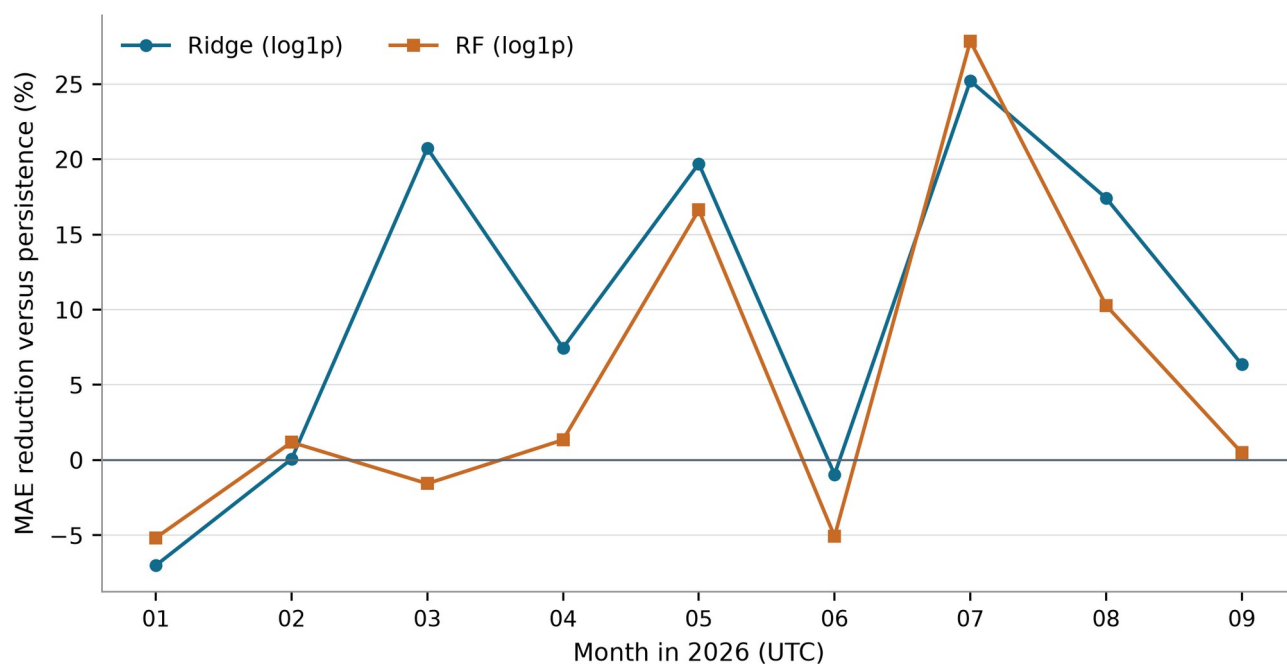


Figure 5. Monthly one-day MAE reductions relative to persistence, January–September 2026. Monthly samples contain 168–186 artist-days. The zero line indicates equal MAE; negative values indicate worse performance. Source: Wikimedia Analytics API; local calculations.

4.4. Large traffic days, forecast bias and visible failures

Only 26 of the 1,638 evaluation artist-days exceed their artist-specific training 95th percentile. These observations make up 1.59% of evaluated rows and 10.51% of observed pageviews, but they account for 38.33% of log-forest absolute error. The forest's MAE is 345.38 on the other 1,612 observations and 13,307.51 on the large-traffic subset. Ridge reaches 332.90 and 11,647.39, respectively, while persistence reaches 372.83 and 14,885.23. All three methods improve their average behavior away from the rare largest traffic values, but the higher-traffic tail remains a substantial source of error. The threshold does not force exactly 5% of the later period to be large because it is estimated from an earlier traffic distribution.

The largest observed primary test count is 96,360 pageviews for BTS on 19 July 2026. The log forest predicts 8,822.14, leaving an absolute error of 87,537.86. This single forecast is a concrete example of a failure that an average score can conceal. The fixed July–September display in Figure 6 shows the subsequent adjustment along with the other artists' trajectories. The chart window is a complete calendar quarter selected for consistent presentation, not a set of individually chosen successful forecasts. The study does not assign a release, news item or campaign as the cause of this spike because no event-level explanatory variable or causal design is included.

Bias reveals another distinction. Log ridge underpredicts by 114.33 pageviews per artist-day on average, and the log forest by 167.69. Persistence has a small positive mean bias of 2.88, despite its greater absolute error. A nearly zero mean bias does not imply accurate individual forecasts because positive and negative errors can cancel. Conversely, a lower MAE can coexist with underprediction if large peaks are systematically underestimated. The forest's mean bias for BTS is −994.59, while the other artist biases are much smaller. These patterns support an operational interpretation in which point forecasts are useful reference levels but insufficient estimates of peak-day workload.

The count forest has lower RMSE than the log forest, 3,036.57 versus 3,084.27, although its MAE is higher. This is consistent with different weighting of larger deviations under the two scores and target objectives. It does not contradict the MAE-based choice of the primary forest, and it shows why the selected objective must be stated. Log ridge still has lower RMSE than either forest in the final one-day comparison. Figure 7 reports error concentration and regime-specific MAE without treating the observed large-traffic category as something known before prediction. A future operational alert would need a separate rule based entirely on information available at the origin.

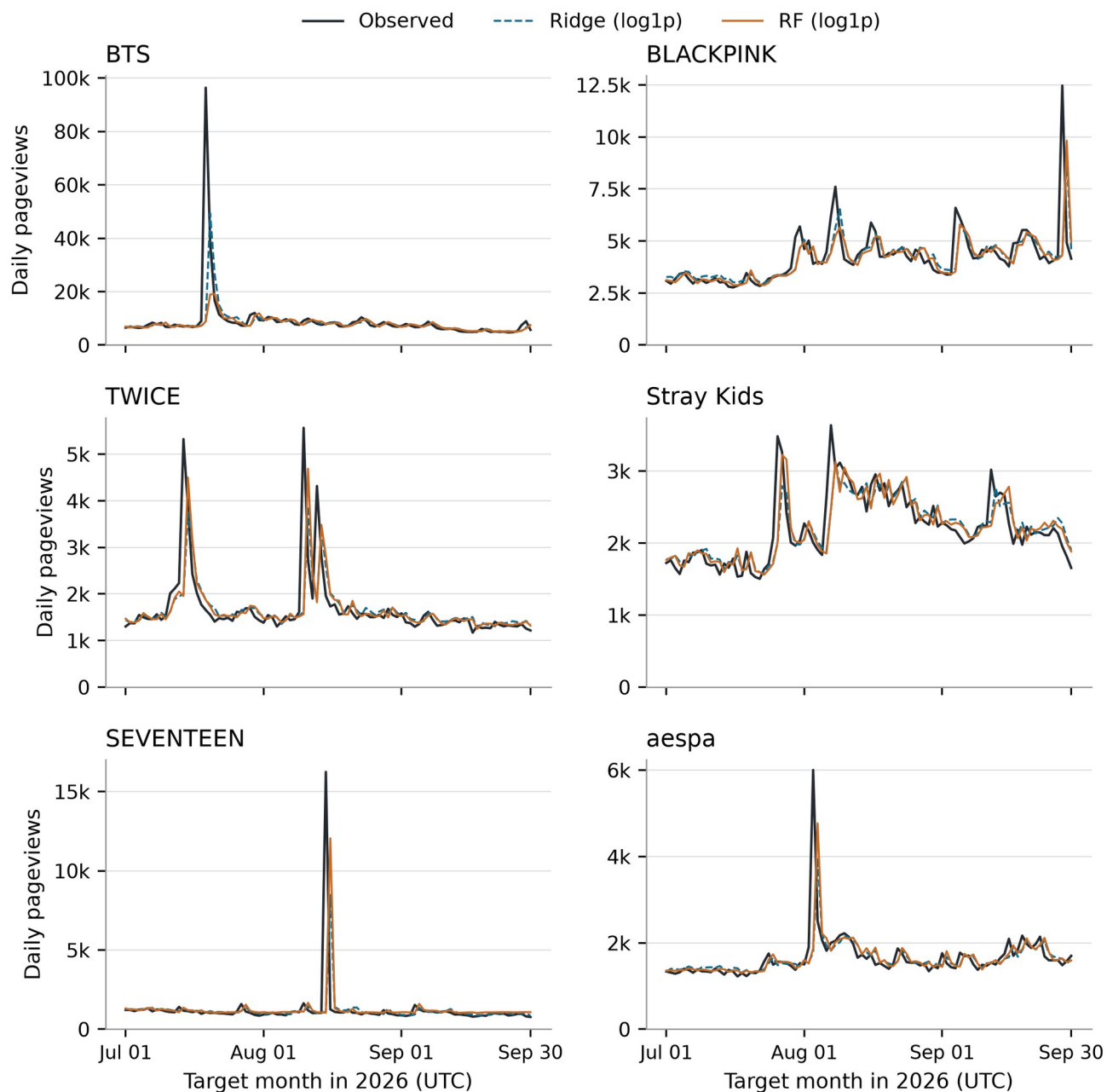


Figure 6. Observed and one-day predicted counts during the fixed July–September 2026 window (92 days per artist). Panels use separate vertical ranges starting at zero to reveal each trajectory; levels should not be compared across panels without reading the axes. No forecast uncertainty bands are claimed. Source: Wikimedia Analytics API; local calculations.

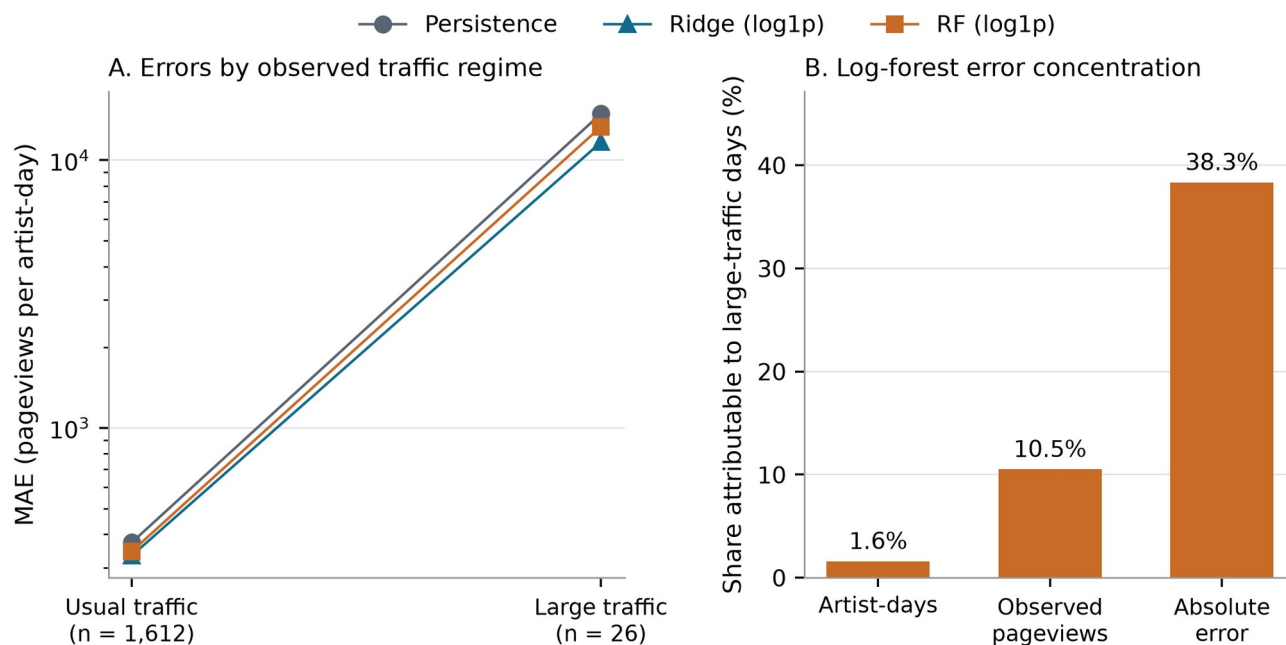


Figure 7. Large-traffic artist-days exceed that artist's 95th percentile calculated from 2022–2025 counts. This observed-outcome stratification is used only to evaluate errors after forecasting. Panel A has a logarithmic error axis; panel B reports shares of artist-days, actual counts and log-forest absolute error. Source: Wikimedia Analytics API; local calculations.

4.5. Seeds, features and training history

The log forest is stable across the five reported random seeds. Its MAE ranges from 550.65 to 551.77, a span of only 1.12 pageviews. The count forest ranges from 588.74 to 594.32, a span of 5.57. These differences are small relative to the difference between the log forest and log ridge, and much smaller than the variation associated with difficult dates. Seed 42 remains the primary run even though another seed has a slightly lower final score. Repeating seeds therefore addresses stochastic fitting variability, while the paired date-block analysis addresses a separate source of uncertainty across the historical evaluation sequence. The two should not be interpreted as interchangeable intervals.

Feature ablation identifies the latest observable count as the dominant contributor within the selected forest. Removing it raises MAE from 551.13 to 735.76, an increase of 184.63 pageviews or 33.50%. Removing the earlier individual lags yields 552.58; removing rolling summaries yields 551.46; removing the artist identifier yields 550.43. The last result is slightly better than the full model, suggesting that artist identity is not necessary to obtain this forest's observed primary accuracy. The very small differences among these latter variants should not be promoted as precise feature rankings. The remaining predictors overlap strongly, and no variant receives its own new tuning search.

Table 4. One-day log-forest feature ablations

Variant	MAE	Change in MAE	Relative change
Full feature set	551.13	+0.00	+0.00%
Without latest count	735.76	+184.63	+33.50%
Without earlier lags	552.58	+1.45	+0.26%
Without rolling summaries	551.46	+0.33	+0.06%
Without artist identifier	550.43	-0.70	-0.13%

Every variant is refitted with the same selected parameters and seed 42. Changes use the full feature set as the denominator; positive values indicate worse error.

Longer fitted target history helps the selected forest in this retrospective comparison. Restricting fitting to 24 months increases MAE to 592.03, and restricting it to 12 months increases MAE to 596.90, compared with 551.13 under the 48-month window. The changes are approximately 7.42% and 8.31%, respectively. This is evidence against assuming that the latest year alone is automatically the best training set for these specified models. It does not establish that more distant observations will always help, because changes in audience behavior or measurement can make older records less relevant. The horizon and parameters are held constant, so these results isolate the stated fitting-window choice more directly than comparing independently optimized pipelines.

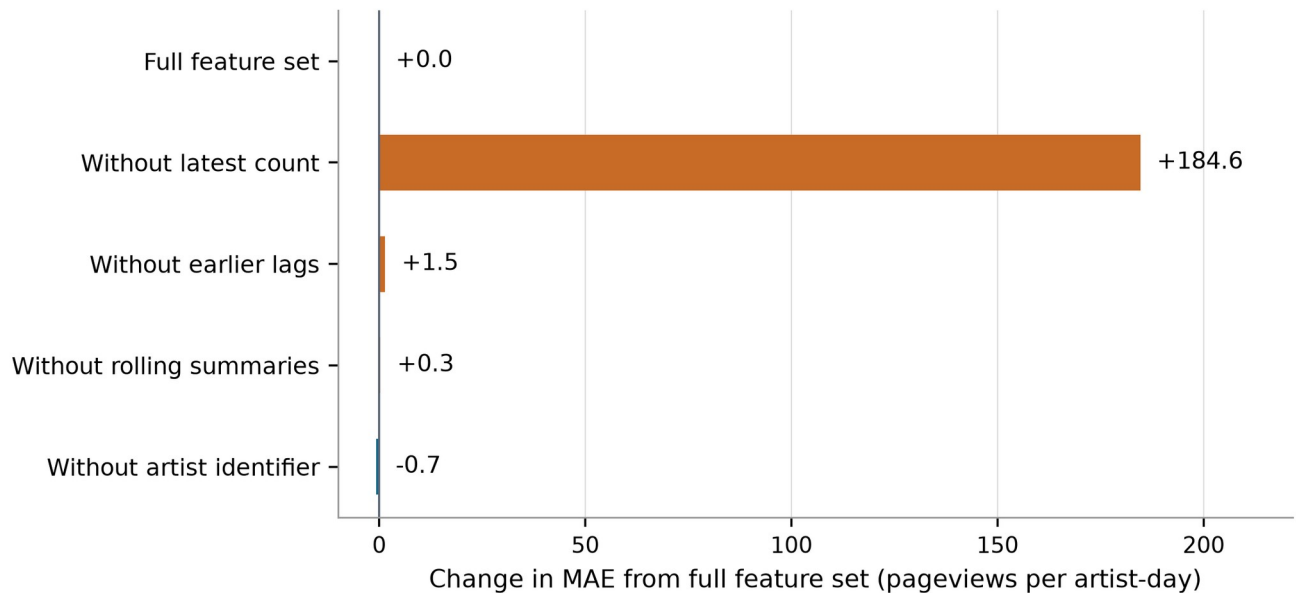


Figure 8. Feature ablations for the development-selected log forest, with parameters fixed and every variant refitted. Positive values indicate higher one-day MAE than the full feature set. The comparison diagnoses this fitted system and does not identify causal effects of features. Source: Wikimedia Analytics API; local calculations.

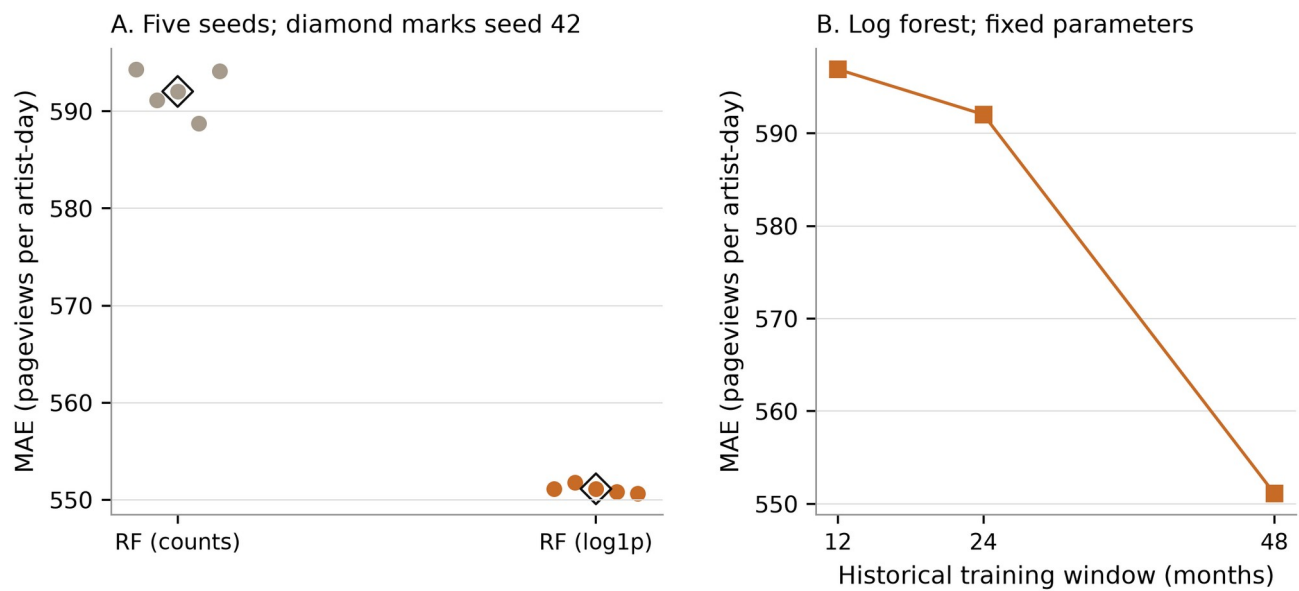


Figure 9. Random-seed and training-window checks on the same 1,638 one-day evaluation observations. Panel A reports all five seeds (7, 19, 42, 73, 101); seed 42 remains the primary run. Panel B retains the selected forest parameters while changing only the available training targets. Vertical axes are intentionally not zero-based because these panels show small differences between repeated fits. Source: Wikimedia Analytics API; local calculations.

4.6. Seven-day forecasting and artist transfer

At the seven-day horizon, errors increase substantially. Persistence reaches MAE 1,118.15 and the seven-day mean reaches 1,052.90. Log ridge and the log forest reach 896.39 and 898.00, respectively, representing reductions of approximately 19.83% and 19.69% relative to seven-day persistence. Their errors are nearly equal in this comparison, while the count forest reaches 970.65. These are horizon-specific direct models with horizon-specific development selection and a fitting cutoff of 25 December 2025. The larger relative reduction does not mean seven-day forecasts are more accurate than one-day forecasts: the seven-day persistence comparator is itself much less accurate, and the absolute errors of both learned log models are higher.

Table 5. Seven-day forecast performance

Method	MAE	RMSE	WAPE	MAE reduction
Persistence	1,118.15	4,720.60	31.09%	+0.00%
7-day mean	1,052.90	4,023.34	29.27%	+5.84%
Ridge (counts)	1,377.66	3,658.04	38.30%	-23.21%
Ridge (log1p)	896.39	3,648.22	24.92%	+19.83%
RF (counts)	970.65	3,604.95	26.99%	+13.19%
RF (log1p)	898.00	3,626.81	24.97%	+19.69%

All methods evaluate 1,638 artist-days. At this horizon seasonal naive is identical to persistence and is omitted. Reduction uses the matching seven-day persistence error; all count inputs end at least seven days before the target.

Artist transfer exposes a more consequential limitation. On IVE and LE SSERAFIM pooled, persistence has MAE 235.30 and log ridge 224.84, a descriptive reduction of 4.45%. The log forest instead reaches 263.90, 12.15% worse than persistence. For IVE alone, the log forest’s MAE is 190.64 compared with persistence 121.67, a deterioration of 56.68%; log ridge is also slightly worse at 125.20. For LE SSERAFIM, both log methods improve their point estimates, reaching 337.15 for the forest and 324.48 for ridge against persistence 348.93. The different outcomes show that a positive result on all fitted artist categories does not automatically transfer to an artist category absent from training.

Table 6. Transfer to artists excluded from fitting

Artist	Persistence MAE	Log ridge MAE	Log forest MAE	Ridge reduction	Forest reduction
IVE	121.67	125.20	190.64	-2.89%	-56.68%
LE SSERAFIM	348.93	324.48	337.15	+7.01%	+3.38%
Pooled external artists	235.30	224.84	263.90	+4.45%	-12.15%

Each individual artist contributes 273 dates; the pool contains 546 artist-days. Models retain their six-artist fit and selected parameters. Each additional artist’s own past counts are supplied as inputs; no adaptation is performed.

The additional-artist 14-day bootstrap intervals are -10.97% – 14.49% for the ridge reduction and -36.02% – 0.57% for the forest reduction. Neither establishes a dependable improvement over persistence in this two-artist historical comparison. The point estimate and interval agree that the forest transfer result is weak, but its interval still includes a very small improvement. It would be too strong to declare a universal transfer failure from two articles, just as it would be too strong to call the small pooled ridge gain reliable. Figure 10 places these transfer results beside the horizon check, making both limits visible within the same evaluation framework.

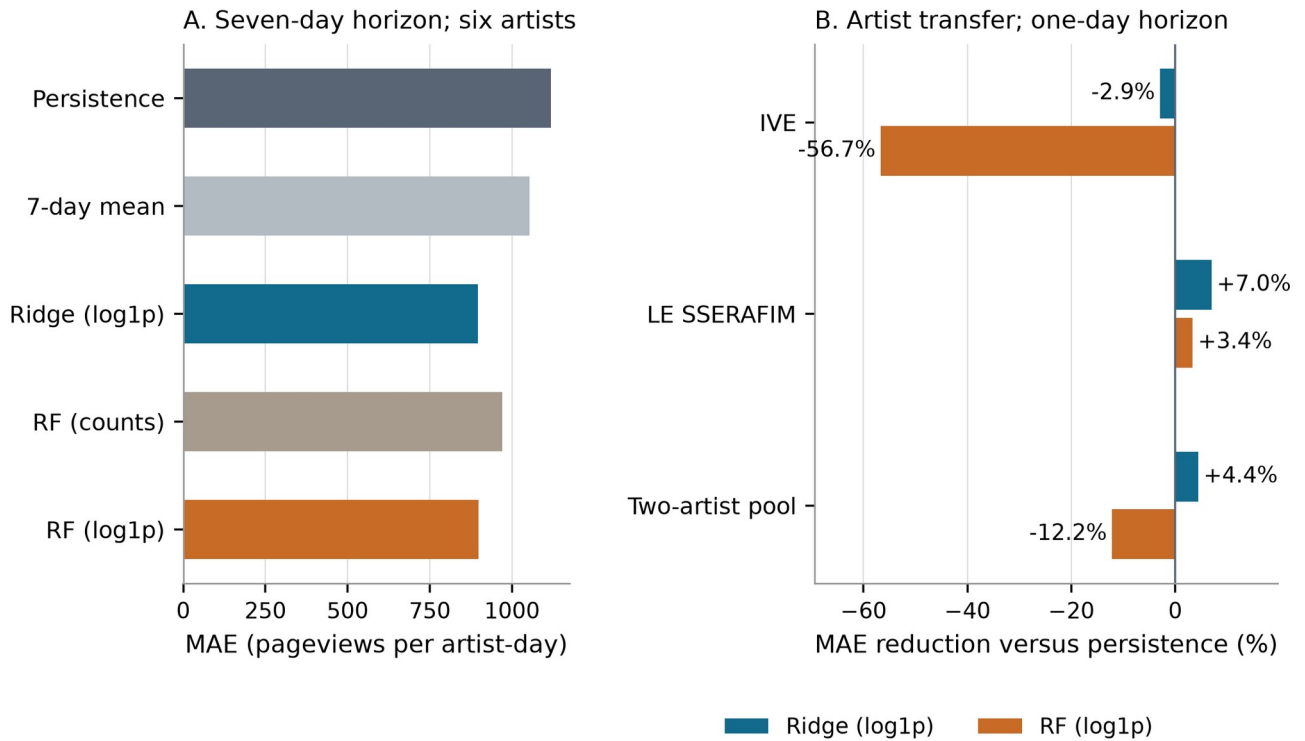


Figure 10. Seven-day and artist-transfer checks. Panel A evaluates 1,638 primary artist-days with horizon-specific training cutoffs and parameters; seasonal naive equals persistence at this horizon. Panel B evaluates 273 days per additional artist (546 pooled), with both artists excluded from fitting and parameter selection. Artist transfer still uses each artist’s own past counts. Percentages are descriptive point estimates. Source: Wikimedia Analytics API; local calculations.

5. Discussion

5.1. What the model comparison establishes

The main finding is that model complexity alone does not explain useful forecasting performance. A regularized linear model fitted to log counts outperforms the evaluated forests in the one-day primary comparison. At the same time, the log forest improves the point-estimate MAE over persistence and over its count-target counterpart. These results suggest that the transformation, treatment of scale and strength of the recent-history baseline deserve as much attention as the choice between a linear and nonlinear estimator. The experiment does not establish that the learned data-generating relationship is truly linear; it establishes which stated forecasting procedures performed better on these recorded artist-days.

The latest observable count carries substantial predictive information, as shown by the strong persistence baseline and the large forest deterioration when that feature is removed. Earlier lags, rolling summaries and artist indicators contribute little additional error reduction within the ablation design. This pattern is consistent with a monitoring setting in which much of tomorrow’s traffic level follows the current level, while sudden changes are less inferable from a month of count history alone. Because the inputs are correlated and only one forest configuration is ablated, the results support a practical simplification argument rather than a claim that those other variables contain no information in any model.

The development and final rankings also illustrate the uncertainty of model choice. Log forest wins narrowly over log ridge in the eight-quarter development mean, yet ridge wins in the later period. A model-

selection result should therefore be accompanied by the size of the development difference, later errors and suitable uncertainty estimates. A single “best model” label can conceal the fact that two methods are close during selection and react differently to later peaks. For an attention-monitoring implementation, retaining persistence and a simple regularized comparator alongside the forest would make performance degradation observable instead of assuming that a chosen algorithm remains best indefinitely.

5.2. Practical use for publicity and audience monitoring

A modest operational use follows from the measurement being forecast. An analyst could record each article’s recent counts, produce a next-day reference estimate and compare later observed traffic with that estimate. The result could help prioritize article monitoring and explain whether a day’s volume is close to its recent pattern. It could also support regular reports that distinguish a high count from an unexpectedly high count. These are proposed uses of a forecast of article visits. They are not demonstrated effects on campaign efficiency, media placement, stream growth or internship performance, and the present study contains no intervention that would identify such benefits.

The accuracy numbers give this proposal boundaries. A reduction of 90.68 pageviews in pooled one-day MAE for ridge may be useful for a recurring aggregate report, but it is not a promise to anticipate every publicity surge. The extreme July example shows a much larger single-day miss. Point forecasts should consequently be presented as expected reference levels with a visible record of recent forecast errors, and people using them should retain the ability to inspect the actual counts. The bootstrap intervals in this paper should not be displayed as uncertainty bands around tomorrow’s forecast, because they quantify the average historical error comparison rather than the distribution of a particular future count.

The transfer results recommend a separate check when a new article is added to monitoring. Past counts for that article are necessary but not sufficient to ensure that the six-artist fitted relationship works well. The zeroed unknown-category indicators are a concrete modeling assumption, and IVE’s weaker results show its practical relevance. An implementation could compare the transferred model with persistence before relying on it, then evaluate whether removing identifiers or fitting an article-specific calibration changes performance. Such possibilities are implications of the completed transfer experiment; no adapted model or prospective performance value is claimed in the current results.

5.3. Sources of large errors and scope of inference

The concentration of error on a small set of high-traffic dates is one of the most informative findings for this setting. More than one third of forest absolute error is associated with 1.59% of the evaluated observations. When attention is driven by an event not represented in recent counts, an autoregressive model may have little basis for predicting the initial jump. The evaluated forest also produces predictions by averaging training-leaf target values, limiting how it can express an unprecedented scale change. These mechanisms are consistent with the observed peak errors, but the study does not decompose their causes or establish which external event produced a particular count.

Adding release schedules, independently collected news volume or other online covariates could change this forecasting problem, but those inputs would need publication timestamps that establish availability at the forecast origin. A contemporaneous article about an already observed spike would not be a legitimate predictor of the spike’s earlier onset. The present feature set has the advantage that its information boundary is straightforward to inspect. Its lack of event context is also a substantive limitation. A richer

model should be judged against the same persistence and log-ridge comparators, rather than assuming that extra data or a deeper estimator will necessarily improve the retrospective scores.

The largest original-count series influences pooled results disproportionately. For a team interested in total monitoring volume, that weighting may be appropriate. For a team assigning equal attention to each artist, the mean artist-specific percentage improvement is more relevant and smaller. Neither figure can be described simply as the percentage by which “K-pop forecasting” improves. Stating the sample, denominator and aggregation rule is therefore part of the substantive interpretation. This is also why the apparent seven-day relative gain should not be compared directly with the one-day percentage without considering the much larger seven-day baseline error.

5.4. Limitations

The study has four main limits. First, it measures attention to fixed English Wikipedia article titles, without combining alternative titles, member articles, other language editions or other platforms. Geographic origin is not inferred from article language. Visits cannot identify unique fans or establish commercial engagement. Second, the six primary artists and two additional artists form a purposive and small series sample. The results do not represent every artist, audience or publicity cycle. Each artist supplies many days, but repeated observations do not turn six artist series into a large random sample of the entertainment industry.

Third, the evaluation reconstructs daily forecast information from a historical extract. It is retrospective, and it does not measure live API publication delay, late revisions or the exact operational hour at which the completed previous UTC day’s count becomes available. The daily-origin interpretation assumes that the count history ending at that origin has been observed. A real deployment must align issuance time with actual data availability. The historical split, source hashes and leakage checks make the stated reconstruction inspectable, but they cannot turn it into evidence of live forecasting performance. The earlier pilot and pipeline-verification access are disclosed in the methods for the same reason.

Fourth, the statistical comparison is conditional on the stated models, limited grid, artist sample and evaluation period. The bootstrap preserves paired dates and short blocks but does not integrate the uncertainty of the original artist choice, parameter selection or all future regimes. Ablations have not been separately retuned, the count-target ridge uses log-transformed predictors, and the learned estimators are fitted using squared error even though MAE is the primary evaluation criterion. These choices bound the conclusion: log ridge performs best among these executed one-day specifications. They do not rule out alternative transformations, objectives, seasonal statistical models or more extensive tuning that could perform differently.

6. Conclusion

Public Wikipedia pageviews support a reproducible, lightweight study of daily aggregate attention to K-pop artist articles. On 1,638 later primary artist-days, log-target ridge regression reduces MAE by 15.03% relative to persistence, while the development-selected log forest produces an 8.63% point-estimate reduction with a paired interval that includes no improvement. The latest available count is the most consequential forest input, and five-seed repetition shows little stochastic variation. Rare high-traffic days account for a substantial share of forecast error, seven-day absolute errors are higher, and transfer to two artists excluded from fitting does not establish a reliable advantage over persistence.

The practical implication is to begin attention forecasting with a strong recent-count baseline and a transparent regularized comparator, then evaluate whether a forest adds value for the actual artist and period being monitored. Forecasts should retain their measurement meaning as expected article visits and be assessed across artists, months and unusual traffic conditions. Preserved data, prediction tables, figures, a results workbook and an executed notebook support local inspection and reproducibility. The study demonstrates both the potential of accessible online data and the importance of reporting where a modest predictive method fails.

Data and Code Availability

The numerical source data are openly available through the Wikimedia Analytics API under CC0 [2], using the article titles, date ranges and request settings reported in this paper. The locally executed research package contains preserved raw API responses, CSV extracts, provenance records, scripts, configuration files, prediction tables, figures, a workbook, an executed notebook and pinned dependency records. These working materials are retained privately by the author and publisher and are not deposited as public supplementary files. Readers may direct methodological and reproducibility enquiries to the corresponding author.

Authorship and Computational Assistance

Frankie Gao is the sole named author. The study uses publicly available aggregate data, and the computations were executed locally. AI tools assisted with code development, analysis execution, figure preparation and language drafting.

References

- [1] Wikimedia Foundation. Page view analytics. Wikimedia Analytics API documentation. Accessed 2 October 2026. <https://doc.wikimedia.org/generated-data-platform/aqs/analytics-api/reference/page-views.html>
- [2] Wikimedia Foundation. Access policy: data licensing and downloads. Wikimedia Analytics API documentation. Accessed 2 October 2026. <https://doc.wikimedia.org/generated-data-platform/aqs/analytics-api/documentation/access-policy.html>
- [3] Mestyán M, Yasseri T, Kertész J. Early prediction of movie box office success based on Wikipedia activity big data. PLOS ONE. 2013;8(8):e71226. <https://doi.org/10.1371/journal.pone.0071226>
- [4] Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12(1):55–67. <https://doi.org/10.1080/00401706.1970.10488634>
- [5] Breiman L. Random forests. Machine Learning. 2001;45:5–32. <https://doi.org/10.1023/A:1010933404324>
- [6] Cerqueira V, Torgo L, Mozetič I. Evaluating time series forecasting models: an empirical study on performance estimation methods. Machine Learning. 2020;109:1997–2028. <https://doi.org/10.1007/s10994-020-05910-7>
- [7] Hewamalage H, Ackermann K, Bergmeir C. Forecast evaluation for data scientists: common pitfalls and best practices. Data Mining and Knowledge Discovery. 2023;37:788–832. <https://doi.org/10.1007/s10618-022-00894-5>
- [8] Gneiting T. Making and evaluating point forecasts. Journal of the American Statistical Association. 2011;106(494):746–762. <https://doi.org/10.1198/jasa.2011.r10138>
- [9] Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. International Journal of Forecasting. 2006;22(4):679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>

- [10] Künsch HR. The jackknife and the bootstrap for general stationary observations. The Annals of Statistics. 1989;17(3):1217–1241. <https://doi.org/10.1214/aos/1176347265>
- [11] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. <https://doi.org/10.1016/j.patter.2023.100804>
- [12] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>

Appendix A. Feature Definitions and Evaluation Boundaries

Table 7. Count-history features for target day t at horizon h

Predictor	Available source observations	Transformation
Latest count	Count at $t - h$	$\log(1 + \text{count})$
Earlier individual lags	Counts at $t - h - 6$, $t - h - 13$ and $t - h - 27$	$\log(1 + \text{count})$
7-day mean	Mean of counts from $t - h - 6$ through $t - h$, inclusive	$\log(1 + \text{mean})$
30-day mean	Mean of counts from $t - h - 29$ through $t - h$, inclusive	$\log(1 + \text{mean})$
7-day standard deviation	Sample standard deviation over $t - h - 6$ through $t - h$	$\log(1 + \text{standard deviation})$
Target weekday	Sine and cosine of 2π times weekday/7; Monday = 0	Unchanged
Artist	Category learned from training rows only	One-hot; unknown categories ignored

The two weekday terms are known calendar information. Count and rolling windows never use observations after the origin. The seven-day specification is fitted directly with $h = 7$.

The first eligible target date is 31 January 2022 for $h = 1$ and 6 February 2022 for $h = 7$. The primary test always begins on 1 January 2026 and ends on 30 September 2026. Final fitted target dates end on 31 December 2025 for $h = 1$ and 25 December 2025 for $h = 7$. The additional-artist one-day series have 30 warm-up days from their January 2025 start, and all 2026 dates remain evaluable. The 24- and 12-month fitting-window variants use 4,386 and 2,190 fitted target rows, respectively, with prior count history retained for boundary lookbacks.

Appendix B. Development Grid and Selection

Each horizon evaluates four ridge penalty candidates in each target representation and three forest configurations in each target representation, for 14 learned candidates across eight validation quarters. The saved validation grid consequently contains 224 candidate-fold records over two horizons. Baselines and selected learned specifications are also summarized by fold in a separate table. Hyperparameters are selected by the mean of the eight fold MAEs. The primary forest is the lower-development-MAE forest representation for the one-day task. Fitting ends at the first origin of each quarter; no coefficient or tree is updated within that quarter.

Table 8. Mean quarterly development MAE

Method	One-day horizon	Seven-day horizon
Persistence	507.19	853.73
Seasonal naive	853.73	Same as persistence
7-day mean	622.48	852.74
Ridge (counts)	1,231.73	1,374.76
Ridge (log1p)	462.20	771.64
RF (counts)	495.02	919.28
RF (log1p)	458.78	817.59

Development quarters span 2024 Q1–2025 Q4. Values are equally weighted means of quarterly MAEs. Seasonal naive is omitted as a duplicate at the seven-day horizon.

Appendix C. Interval and Repeat-Run Sensitivity

Table 9. Date-block length sensitivity of one-day MAE reductions

Method	Block length (days)	Point reduction	95% lower	95% upper
RF (log1p)	7	8.63%	-2.72%	18.98%
RF (log1p)	14	8.63%	-2.95%	18.20%
RF (log1p)	28	8.63%	-1.49%	18.35%
Ridge (log1p)	7	15.03%	5.13%	22.47%
Ridge (log1p)	14	15.03%	6.14%	21.51%
Ridge (log1p)	28	15.03%	6.75%	20.70%

Intervals use 2,000 paired circular date-block replicates. They quantify the historical average-error comparison, not future individual-count uncertainty.

Table 10. All forest random-seed results

Seed	Count forest MAE	Log forest MAE
7	594.32	551.13
19	591.12	551.77
42	592.04	551.13
73	588.74	550.83
101	594.12	550.65

Selected parameters are held fixed. Seed 42 defines the primary run and is not replaced by the best final-period seed.

Appendix D. Reproduction and Data Identity

Run the data-preparation, development, evaluation, report-diagnostic and figure scripts in that order, using the commands in README.md. Preserved input files make historical data preparation repeatable without requiring the original pilot workspace or repeating the API requests. The executed notebook independently refits the primary one-day comparison and reconciles its saved predictions and metrics. Raw source responses remain separate from derived tables. The complete prediction files contain target date, artist, source article, horizon, observed count, model, prediction and absolute error, enabling alternative aggregation without refitting the models.

The primary CSV SHA-256 digest is 053cc6e06a71a4871c36c3bcfab558359542a3acd70e093276ba12ad2842f325. The configuration file stores this digest alongside the analysis-code digest. Execution and report-diagnostic records confirm the direct rolling-window checks, both-horizon future-perturbation checks, fitting-origin boundaries, validation-based parameter selection and recomputed headline MAEs. Provenance records distinguish initial retrieval timestamps from later reads of cached responses. Artist transfer is explicitly recorded as retrospective, with the two additional artists excluded from fitting and parameter selection throughout.