Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

When accuracy does not improve decisions: a decision-outcome validation framework for cross-regional box office forecasting

Lu Chao

tsc05zytg@163.com

Tianjin Tianshi College

Xu AnQi

Tianjin Tianshi College

XiaoXi Ma

Tianjin Tianshi College

Research Article

Keywords: decision-oriented validation, temporal holdout, region-out generalization, box office forecasting, cultural products, model selection

Posted Date: September 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10573773/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

When accuracy does not improve decisions: a decision-outcome validation framework for cross-regional box office forecasting

Abstract

Cross-regional box office forecasting is routinely evaluated with predictive accuracy on randomly split samples, yet the models that minimize root mean squared error are not necessarily the models that produce better market selection decisions. This paper develops a decision-outcome validation framework in which forecast models are compared through a budget-constrained market selection problem, using temporal walk-forward holdouts, region-out and film-out generalization protocols, and top-k payoff, regret and hit rate as the evaluation criteria. The framework is applied to a web-sourced panel of 14 Chinese films released between 2023 and 2025, covering 76 film market pairs and 18 overseas territories with weekly gross data. Five models are compared, including a panel model with market fixed effects, regularized regression, gradient boosting and a decision-aware ranking model. The accuracy-optimal and decision-optimal models diverge in 22 of 30 validation folds. In the temporal fold covering the 2025 releases, the decision-optimal model captures 89.3 percent of the achievable revenue share, against 88.7 percent for the accuracy-optimal model and 78.9 percent for the boosting model that the weekly accuracy rankings favored, and the payoff gap between decision-based and accuracy-based selection is $152,746. Accuracy differences between models are statistically significant in three of the four temporal folds, and the decision ranking is robust to regularization strength, tree depth and the outcome window, while stronger regularization improves accuracy but lowers decision payoff. Paired bootstrap intervals for the payoff difference between decision-based and accuracy-based selection are positive across random and film-out folds, although the small number of folds limits statistical power. A Monte Carlo study with heteroscedastic market errors shows that divergence between accuracy and decision rankings is systematic when overall accuracy is balanced, and disappears only when one model dominates on both criteria. The results imply that accuracy-oriented validation with random splits is not a neutral protocol for cross-regional forecasting, and that decision outcomes should be reported alongside conventional error metrics.

Keywords: decision-oriented validation; temporal holdout; region-out generalization; box office forecasting; cultural products; model selection

1. Introduction

Film distributors allocate promotional resources, print copies and streaming slots across territories on the basis of forecasts. The forecasting literature for box office revenue has expanded rapidly with machine learning tools, and the resulting models are typically selected by error statistics computed on randomly split data (Koçak 2026). The implicit assumption is that a model with lower error is a better basis for the resource allocation decision that follows. That assumption is rarely tested, and there are reasons to doubt it. A model can be accurate on average and still misrank the few territories that matter most, because error statistics weight every market equally while decisions weight markets by their revenue contribution. Recent work on movie box office studies even documents systematic overinterpretation in machine learning abstracts, where predictive accuracy is presented as evidence of decision value (Leukel et al. 2026).

The empirical context of this paper is the overseas circulation of Chinese films. Chinese-language cinema has moved from a contracted domestic market to an expanding international presence (Utami 2025), and products such as Ne Zha 2 and Black Myth: Wukong have drawn global attention (Gao et al. 2025). Yet the international performance of Chinese films remains difficult to forecast because cultural proximity, audience familiarity and platform visibility shape demand in ways that vary sharply across territories (Hoskins and Mirus 1988; Berg and Sansalone 2024). Box office aggregates are also read as signals of sentiment beyond the cinema industry (Kaplanski 2025), and the relationship between audience evaluation and revenue is nonlinear and platform dependent (Telang and Sivan 2025). These features make cross-regional forecasting a decision problem rather than a pure prediction problem: the distributor cares about which markets to enter and how to sequence entry, not about the average error across all markets.

The economic stakes of the selection problem are easy to understate. Overseas gross for Chinese films is heavily concentrated: a handful of territories routinely generate most of the revenue, while dozens of tracked markets contribute small amounts that still consume

promotional effort. A forecasting model that is accurate on average across all markets can therefore mislead a distributor who must decide where to spend a limited promotional budget. The cost of a wrong ranking is not an average error; it is the revenue forgone in the markets that were not selected. This asymmetry between the error metric and the decision payoff is the starting point of the paper, and it is the reason why validation protocols deserve the same scrutiny as the models themselves.

The methodological literature offers three families of validation practice. The first family compares models with error metrics such as RMSE, MAE or coverage on random splits. The second family emphasizes calibration and interval coverage, often under time-ordered evaluation. The third family, which is smaller and more recent, links forecast evaluation to downstream decisions through decision-focused learning and predict-then-optimize pipelines. The three families can disagree about which model is best for the same data, because they ask different questions. Accuracy metrics ask how close predictions are to outcomes on average. Calibration metrics ask whether uncertainty statements are honest. Decision metrics ask whether the model improves the choice that the forecast supports. The disagreement is not a technical nuisance; it is the central reason why validation protocols deserve explicit attention.

This paper makes three contributions. First, it proposes an organizing classification of forecast validation research, distinguishing accuracy-oriented, calibration-oriented and decision-oriented lines, and explains why results across these lines diverge through metric construction, data leakage and market heterogeneity. Second, it formalizes a decision-outcome validation framework for cross-regional forecasting, including a budget-constrained market selection problem, temporal walk-forward and region-out generalization algorithms, and statistical tests for payoff differences between accuracy-based and decision-based model selection. Third, it applies the framework to a new web-sourced panel of 14 Chinese films, 76 film market pairs and 18 overseas territories, and shows that accuracy-optimal and decision-optimal model selections diverge in the majority of validation folds. Chinese films are a good testbed for another reason that has nothing to do with cinema. Their overseas distribution is organized through a mix of global distributors, regional partners and platform windows, which means that the same product faces genuinely different market structures across territories. The forecasting

problem is therefore not a single time series problem with a geographic label; it is a multi-market panel with heterogeneous institutions, and it is exactly the setting where validation protocols that ignore market structure will mislead. The same properties are shared by streaming series, mobile games and music exports, which extends the relevance of the empirical illustration beyond the film industry.

We address three research questions. RQ1 asks whether the accuracy ranking and the decision ranking select the same model under temporal, regional and film-level holdouts. RQ2 asks how large the payoff gap between the two selections is and whether it is statistically distinguishable from zero. RQ3 asks which mechanisms in the data-generating process and in the model structure drive the divergence. The three questions map onto the empirical sections: Section 6.2 answers RQ1, Sections 6.3 and 6.4 answer RQ2, and Sections 6.5 and 6.6 answer RQ3.

The paper is written for a methodological audience, and the empirical application is deliberately used as a testbed rather than as the object of a film-industry study. The framework transfers to any setting in which forecasts feed selection decisions under a budget, including streaming platform acquisition, game storefront placement, tourism promotion and regional sales planning. The organizing classification in Section 2, the formal decision problem in Section 3 and the validation algorithms in Section 5 are stated without reference to cinema-specific assumptions, so that the contribution outlives the particular dataset used to illustrate it.

The empirical contribution is deliberately bounded. The panel is small relative to the global film population, the weekly series come from one tracking service, and the results demonstrate the framework rather than establish a universal law about box office forecasting.

The remainder of the paper is organized as follows. Section 2 synthesizes the literature and presents the organizing framework. Section 3 formalizes the decision problem and the evaluation criteria. Section 4 describes the data and measurement. Section 5 specifies the models and validation algorithms. Section 6 reports the results. Section 7 discusses implications and limitations. Section 8 concludes.

2. Literature synthesis and an organizing framework

Forecast validation research can be organized around three questions that are often conflated. The first question is whether predictions are close to outcomes. The second is whether uncertainty statements are reliable. The third is whether predictions improve decisions. Each question has generated its own methods, and the methods answer different empirical problems.

The accuracy-oriented line is anchored in classical forecast comparison, where predictive accuracy is tested against a benchmark under a symmetric loss function (Diebold and Mariano 1995). Modern extensions refine how accuracy should be computed. Robust evaluation methods adjust accuracy measures for the instability of point estimates (Agustini et al. 2026), while cross-validation studies warn that leakage can inflate apparent accuracy when temporal structure is ignored (Király and Tóth 2025). Time series specialists have proposed cross-validation schemes that respect serial dependence (Vamsikrishna and Gijo 2024), and hierarchical forecasting research shows that the aggregation level changes both accuracy and its interpretation (Abolghasemi et al. 2024; Zhang et al. 2025). Panel forecasting methods exploit cross-sectional learning when series are short (Lovaglio 2025), and demand forecasting studies demonstrate that the value of machine learning depends on the data environment and the decision context (Chae et al. 2024). Dynamic model selection systems further show that the best model changes over time, so a single accuracy ranking is unstable by construction (Jiang et al. 2025). Explainable decision support extends the same logic to feature importance, where global accuracy masks local failures (Kuti-Kreszács et al. 2026).

The calibration-oriented line deserves more credit than it usually receives in forecasting comparisons. Interval forecasts and probabilistic scores have been standard in statistics for decades, and the line has produced a rich toolkit for checking whether uncertainty statements are honest. Its blind spot is different from the accuracy line's blind spot. A well-calibrated model can still produce poor decisions if the loss function of the decision is asymmetric, because calibration says nothing about the relative cost of errors on different markets. The decision-oriented line inherits this insight and adds the missing ingredient: an explicit loss function that maps predictions to actions and actions to

payoffs. The three lines are therefore best understood as successive refinements of the same question rather than as competing schools.

The calibration-oriented line asks whether predicted distributions match observed frequencies. Its contribution to this paper is the recognition that point accuracy and probabilistic honesty are separate properties, and that neither implies decision quality. Calibration failure is a common reason why accurate point forecasts produce poor decisions under uncertainty, because decisions depend on the tails of the distribution rather than its center.

The decision-oriented line is the youngest and the most relevant to cross-regional market selection. Decision-focused learning embeds the downstream optimization into the training loss (Ellinas et al. 2024), and estimate-and-optimize methods learn decision trees that are integrated with two-stage decision problems (Ribeiro and Fanzeres 2026). Data-driven robust optimization can be combined with decision-focused objectives when uncertainty is endogenous (Wang et al. 2026). Within the social science methodology literature, the same shift appears in the call to evaluate indicators and models by their usefulness for decision making rather than by their statistical fit alone (Roberts et al. 2025), and in decision-theoretic treatments that manage uncertainty and group choice explicitly (Ventre et al. 2025). Methodological work in Quality & Quantity has also stressed that machine learning validation in the social sciences needs theory-grounded procedures (David et al. 2025), that model comparison should not rely on a single champion model (Makhortykh et al. 2025), and that rankings constructed from indicators should be interrogated rather than accepted at face value (Bruzzi et al. 2025).

The classification differs from existing taxonomies of forecasting research in one important respect. Prior surveys typically organize methods by model family or by aggregation level. The classification proposed here organizes studies by the validation protocol and the selection criterion, because those two features determine whether a study can detect the gap between accuracy and decision quality. A study can use the most sophisticated model family and still draw the wrong conclusion if it selects the model by RMSE on a random split. Conversely, a simple model evaluated under a decision-

oriented protocol can provide the right answer for the wrong-looking reasons. The unit of classification is therefore the evaluation design, not the estimator.

The organizing classification proposed here separates the three lines by their validation protocol and their selection criterion. Accuracy-oriented studies typically use random splits and RMSE, MAE or MAPE. Calibration-oriented studies use time-ordered evaluation and coverage or calibration slope. Decision-oriented studies use protocol-specific holdouts and decision payoffs. The classification is not a taxonomy of papers for its own sake. It explains why published results disagree. Studies that evaluate the same forecasting models under different protocols can reach opposite conclusions about the best model, and this paper shows that the divergence is systematic rather than accidental.

The cultural industries are a particularly instructive setting for the divergence because the underlying demand process combines strong scale heterogeneity with weak signal density. Films, games and television exports face markets that differ in language, distribution infrastructure and audience familiarity, and the observable signals, such as review aggregates and opening-week gross, are noisy proxies for the outcomes that managers care about. When the signal is weak and the market scale is heterogeneous, the ranking of markets becomes sensitive to small changes in the model, and the accuracy ranking loses its authority. This is precisely the condition under which decision-oriented validation pays the largest dividend.

Three mechanisms generate the divergence. The first is metric construction. Accuracy metrics average errors over all observations, while decisions respond to the ordering of a small number of high-value units. The second is leakage. Random splits that ignore time or region structure allow the model to borrow information from the test set, which can compress the apparent differences between models. The third is heterogeneity. When markets differ in scale, volatility and cultural distance, the ranking of models by average error is decoupled from the ranking by payoff. Market size and cultural proximity shape international demand in ways that interact with model flexibility (Fujiy et al. 2025; Kraemer et al. 2025), and localization and paratext research shows that the same product is encountered differently across territories (Sayaheen 2024). Cultural industry studies have applied panel methods to track such heterogeneity across time and product

categories (Magrini and Rimediotti 2025), which supports the view that region structure must enter the validation design.

The three mechanisms also explain the empirical pattern reported in later sections. Metric construction implies that divergence should appear when market scales are unequal. Leakage implies that random splits can hide the sensitivity of model rankings to the test composition, because the test set shares the distribution of the training set. Heterogeneity implies that the divergence should be larger for products with uneven territorial footprints, such as Chinese films with strong performance in East Asia and weaker performance in Europe. Each mechanism has a distinct signature in the data, and Section 6 looks for those signatures explicitly.

The history of forecast evaluation contains a useful caution about the accuracy illusion. Early work on predictive comparisons treated error statistics as the natural currency of model choice, and the practice survived because it is easy to compute and easy to audit. The illusion appears when the error statistic is interpreted as a proxy for something it does not measure, such as the value of a decision. A model can dominate on RMSE in every fold and still lose money in the decision it supports, if its errors are concentrated in the markets that the decision cares about. The remedy is not to abandon error statistics but to report them alongside decision outcomes, so that the two views can be compared rather than conflated.

Section 3 turns the classification into a formal decision problem, and Section 4 describes the panel used to test it.

The classification yields a testable implication, and Figure 1 summarizes the three lines and their protocols.

If the three lines were equivalent, then the model selected by accuracy would coincide with the model selected by decision payoff under every protocol. The empirical sections of this paper test that implication directly.

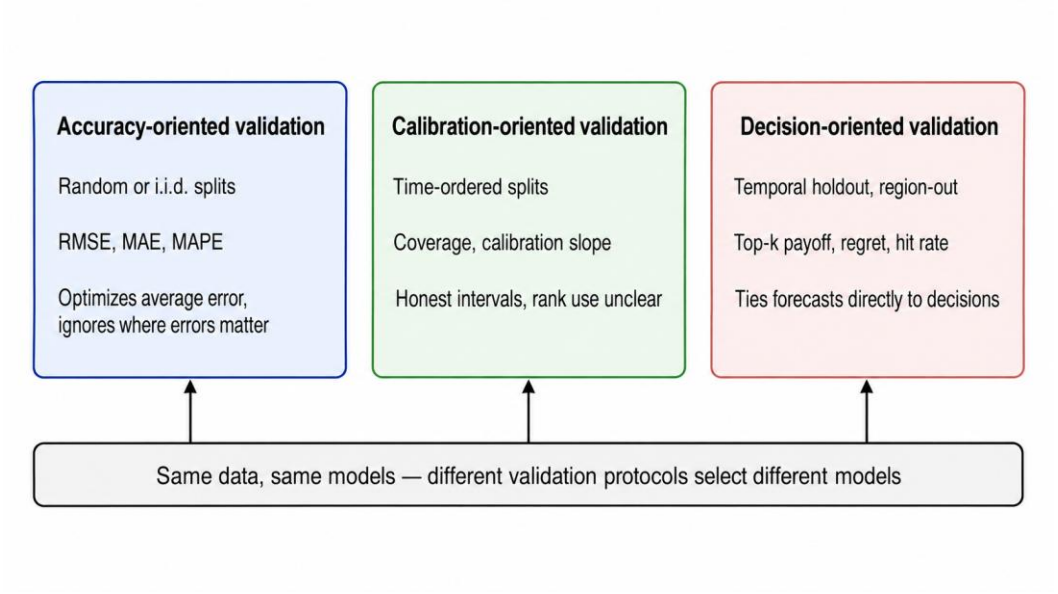


Fig. 1 Three research lines in forecast validation and their protocols.

3. Decision problem and formal framework

Consider a set of films F and, for each film f, a set of overseas markets M_f with $|M_f| \geq 1$. For each film market pair, the realized 12 week cumulative gross is $y_{f,m}$, and the log outcome is $z_{f,m} = \log(1 + y_{f,m})$. A forecasting model estimates $z_{f,m}$ from film level characteristics, market level characteristics and, in the early signal specification, the opening week observation.

Definition 1 (Forecasting system). A forecasting system is a function g that maps observable characteristics of a film market pair to a predicted log outcome. The system may be estimated on a training set and evaluated on a test set.

Definition 2 (Decision policy). For budget k, the decision policy π_k selects the k markets of a film with the highest predicted cumulative gross. Formally, $\pi_k(\hat{g}_f) = \text{argmax over subsets of size k of } \Sigma_{m \in S} \hat{g}_{f,m}$.

Definition 3 (Decision payoff). The payoff of policy π_k for film f under actual outcomes is $P_{f,k} = \Sigma_{m \in \pi_k} \text{expm1}(z_{f,m})$. The oracle payoff $P^*_{f,k}$ is the payoff obtained when the policy observes the true outcomes and selects the k largest realized markets.

Definition 4 (Decision regret and share). The regret is R_{f,k} = P*_{f,k} − P_{f,k}. The share is s_{f,k} = P_{f,k} / Σ_{m ∈ M_f} expm1(z_{f,m}), and the hit rate h_{f,k} is the proportion of selected markets that belong to the oracle top-k set.

The protocol can be written as a function that maps a data split into a pair of rankings. A validation protocol P takes the training set, the test set and the model family, and returns the accuracy ranking and the decision ranking. The empirical content of Proposition 2 is that the mapping is not invariant to the composition of the test set. Temporal protocols produce test sets drawn from later calendar periods, regional protocols produce test sets drawn from unseen territories, and random protocols produce test sets drawn from the same distribution as the training data. Each composition answers a different question, and the question should be chosen to match the decision horizon of the application.

Definition 5 (Model rankings). The accuracy ranking orders models by RMSE on the test set, from smallest to largest. The decision ranking orders models by mean payoff across test films, from largest to smallest. The two selections are the top models of the two rankings.

Proposition 1 (Non-equivalence). There exist data-generating processes under which the accuracy-optimal model is not the decision-optimal model.

Proof sketch. Construct two markets with equal variance but different scale, or two markets with equal scale and different variance. Let model A be accurate on the small or low-variance market and inaccurate on the large or high-variance market, and let model B have the opposite error profile. Because RMSE averages squared errors across markets, the two models can be arranged to have identical RMSE while their rankings of the two markets are reversed. The decision policy selects the higher predicted market, so the payoff difference between the models is strictly positive in the direction of the model that ranks the high-value market correctly. Setting the error magnitudes so that model A has marginally lower RMSE completes the construction. The proof is constructive and does not rely on pathological cases; the Monte Carlo study in Section 6 quantifies how often the divergence occurs under heteroscedastic errors.

The divergence between the two rankings can be formalized as a binary indicator D(g1, g2) that equals one when the model with the lower test RMSE is not the model with the

higher mean payoff, and zero otherwise. The indicator is a function of the protocol, because both the error statistics and the payoffs depend on the composition of the test set. Aggregating the indicator across folds yields a divergence rate, which is the summary statistic used throughout the empirical analysis. The rate is complemented by two directional measures. The payoff gap is the difference between the payoff of the decision-optimal model and the payoff of the accuracy-optimal model, expressed in dollars and as a share of the achievable payoff. The rank gap is Kendall's tau between the accuracy ranking and the decision ranking, which captures disagreement beyond the top model.

The conditions for divergence can be stated precisely. Let the large markets have common scale S and the small markets common scale s, with S substantially larger than s. Let model A err only on the large markets, with variance sigma_L squared, and let model B err only on the small markets, with variance sigma_S squared. With three large and nine small markets, model A has lower RMSE when sigma_L squared is smaller than three times sigma_S squared. Model A still produces ranking errors on the large markets whenever its large-market noise is of the same order as the scale gap, and those ranking errors reduce the payoff. The two conditions, one on the error variances and one on the noise relative to the scale gap, are jointly satisfiable, which is why the divergence is a feature of the data-generating process rather than an artifact of a particular sample. Section 6 reports the results of all four protocols, the statistical tests and the sensitivity analyses.

The simulation in Section 6.5 instantiates these conditions and measures their frequency.

Proposition 2 (Protocol dependence). The accuracy ranking and the decision ranking are functions of the validation protocol; there is no protocol-free ranking of forecasting models, and the magnitude of the accuracy-decision divergence varies with the composition of the test set.

Proof sketch. Under a random split, the test set is drawn from the same distribution as the training set, so the expected ranking of models by RMSE is the in-sample ranking. Under a temporal or regional holdout, the test distribution shifts, and the shift affects error statistics and payoffs differently because payoffs concentrate on the top-k markets. The ranking therefore changes with the protocol, and the divergence can appear or disappear

depending on the test composition. The empirical design in Section 5 operationalizes the comparison by running the same models under four protocols, and Section 6 reports the divergence under each of them.

The three outcome measures used in the empirical analysis are complements rather than substitutes. The payoff measures the absolute value captured by the selected markets. The share measures the same value relative to the film's total overseas gross, which makes films of different sizes comparable. The hit rate measures whether the selected markets overlap with the true top-k set, which is informative when the payoff differences are small. Regret combines the first two by measuring the distance from the oracle selection. Reporting all four is deliberately redundant, because the redundancy makes the conclusions robust to the choice of any single metric.

The formal framework stops short of a full economic model of market entry. A complete model would include promotional spending as a choice variable, entry costs, windowing decisions and the option value of learning from early markets. The top-k formulation abstracts from these features by assuming that the budget is a fixed number of markets and that the payoff is the gross captured in those markets. The abstraction is conservative in a specific sense: if the divergence between accuracy rankings and payoff rankings appears already under this simple structure, it will not disappear under richer structures, because richer structures add channels through which ranking errors translate into lost value.

Section 4 describes the panel and the measurement protocol, and Section 5 turns the framework into an estimation and validation pipeline.

The framework deliberately uses a simple decision structure. Market entry for films involves sequencing, promotional budgets and platform negotiations, which are richer than a top-k selection rule (Zhou et al. 2025). Entry mode research shows that firm experience and institutional context shape how markets are approached (Stoian 2024; Schmitz and Parente 2025), and game sales prediction studies face the same selection problem for digital storefronts (Chertov and Buslaiev 2025). The top-k formulation is chosen because it is the minimal decision structure that connects forecasts to resource

allocation, and because it can be evaluated without proprietary cost data. Extensions to sequencing and budget splits are discussed in Section 7.

4. Data and measurement

The empirical panel is assembled from public web sources, following the principle that web-derived data should be validated against their generating platforms (Jafarzadeh Esfehani et al. 2025). Weekly gross figures by territory are taken from Box Office Mojo release pages, which provide weekly series for each territory in which a film was tracked (Box Office Mojo 2026). The sample consists of 14 Chinese films released between 2023 and 2025, selected to cover the main genres and release windows of the period. The films are Ne Zha 2, The Wandering Earth II, No More Bets, Creation of the Gods I, Creation of the Gods II, Detective Chinatown 1900, Pegasus 2, Full River Red, Article 20, I Am What I Am 2, YOLO, Successor, Lost in the Stars and Boonie Bears: Time Twist. For each film, all territories with weekly data are retained, yielding 76 film market pairs across 18 territories. Weeks are counted from the territory release date and capped at 12.

The coding rules for the film level covariates follow the source definitions. The sequel indicator takes the value one for films that continue an established franchise, including sequels, prequels and annual series such as the Boonie Bears releases; the rule is applied to the film title and its franchise history rather than to the box office record. The Chinese New Year indicator takes the value one for films whose recorded release date falls between January 20 and February 29, the corridor in which the festival window moves. The genre indicators are mutually exclusive in construction, since each genre appears as a separate binary variable. The evaluation aggregates are matched to the film, not to the territory, because the aggregates describe the product rather than the market.

Film level covariates include domestic China opening gross, running time, genre indicators, a sequel or franchise indicator, a Chinese New Year release indicator, and three evaluation aggregates: the Rotten Tomatoes tomatometer, the Rotten Tomatoes verified audience score and the Metacritic metascore. Market level covariates include GDP per capita and population from the World Bank (World Bank 2026), together with a regional grouping into East Asia, Europe, North America, Oceania, Latin America and the Middle East. All covariates are publicly documented, and access dates are recorded in

the data dictionary. Table 1 defines the variables, Table 2a reports film level statistics and Table 2b reports market level statistics.

Table 1 Variable definitions and sources.

Variable	Definition	Source
cum12	Sum of weekly gross over weeks 1 to 12 after territory release (USD)	Box Office Mojo
log_cum12	log(1 + cum12)	Computed
week	Week number from territory release date	Box Office Mojo
china_opening	China domestic opening gross (USD)	Box Office Mojo
runtime_min	Running time in minutes	Box Office Mojo
sequel	1 if the film continues an established franchise	Coded from film record
cny	1 if released in the Chinese New Year window (20 Jan to 29 Feb)	Coded from release date
rt_tom	Rotten Tomatoes tomatometer score (0 to 100)	Rotten Tomatoes
rt_aud	Rotten Tomatoes verified audience score (0 to 100)	Rotten Tomatoes
mc_ms	Metacritic metascore (0 to 100)	Metacritic
genre_*	Binary indicators for animation, action, drama, comedy, fantasy, crime, history, sci-fi, sport	Box Office Mojo
gdp_pc	GDP per capita in current USD	World Bank
pop	Population	World Bank
region	East Asia, Europe, North America, Oceania, Latin America, Middle East	Coded

Table 2a Film-level descriptive statistics.

Film	Release	Markets	Cum12 (USD M)	China opening (USD M)	RT	Aud	MC	Runtime	Seq	CNY
The Wandering Earth 2	2023-01-19	9	6.63	596	83.0	97.0	68.0	173.0	1	0
Full River Red	2023-01-22	6	0.31	670	94.0	70.0	75.0	159.0	0	1
Lost in the Stars	2023-06-23	5	2.91	484	83.0	97.0	68.0	121.0	0	0
Creation of the Gods I	2023-07-10	7	2.63	370	60.0	97.0	68.0	148.0	0	0
No More Bets	2023-08-08	3	1.01	532	83.0	97.0	68.0	130.0	0	0
Pegasus 2	2024-02-08	2	0.39	422	83.0	92.0	68.0	121.0	1	1
YOLO	2024-02-10	5	3.10	430	82.0	95.0	68.0	129.0	0	1
Boonie Bears: Time Twist	2024-02-10	6	0.24	251	83.0	97.0	68.0	105.0	1	1
Article 20	2024-02-10	3	0.27	307	83.0	97.0	68.0	141.0	0	1
Successor	2024-07-	1	0.02	422	83.0	86.0	68.0	133.0	0	0

Film	Release	Markets	Cum12 (USD M)	China opening (USD M)	RT	Aud	MC	Runtime	Seq	CNY
	16									
I Am What I Am 2	2024-12-14	3	0.01	3	83.0	97.0	68.0	104.0	1	0
Creation of the Gods II	2025-01-29	7	3.20		83.0	83.0	68.0	145.0	1	1
Detective Chinatown 1900	2025-01-29	5	3.90	496	83.0	84.0	68.0	136.0	1	1
Ne Zha 2	2025-01-29	14	26.98	2210	91.0	98.0	68.0	144.0	1	1

Table 2b Market-level descriptive statistics.

Market	Region	Films	Cum12 (USD M)	GDP pc (USD)	Population
Australia	Oceania	10	5.22	65,130	27,614,411
Bulgaria	Europe	2	0.02	20,328	6,433,302
Colombia	Latin America	2	0.12	8,562	53,425,635
Domestic	North America	7	33.70	90,027	341,784,857
France	Europe	2	0.39	48,986	68,720,337
Germany	Europe	1	0.57	60,496	83,491,249
Hong Kong	East Asia	9	5.79	56,983	7,498,900
Hungary	Europe	1	0.02	25,907	9,514,251
Italy	Europe	1	0.21	43,309	58,915,656
Lithuania	Europe	1	0.00	32,959	2,888,774
New Zealand	Oceania	10	0.73	49,591	5,324,700
Poland	Europe	1	0.09	28,420	36,435,861
Russia/CIS	Europe	7	1.00	17,547	143,513,328
South Korea	East Asia	4	0.23	36,227	51,684,564
Spain	Europe	1	0.25	38,627	49,355,143
Türkiye	Europe	3	0.05	18,599	85,878,556
United Arab Emirates	Middle East	3	0.12	50,274	11,513,149
United Kingdom	Europe	11	3.11	57,602	69,487,000

Three data limitations are acknowledged up front. First, Box Office Mojo does not provide weekly series for every territory, so five film market pairs of Ne Zha 2, namely Malaysia, Singapore, Indonesia, Japan and New Zealand, are represented by their opening records but excluded from the weekly panel. Second, several films have very short overseas runs, which is itself a feature of the market. Third, the panel is unbalanced by construction, and the models are estimated on the observed weekly cells without

imputing missing weeks. The data availability statement in the Declarations section explains how the assembled dataset can be obtained.

The collection protocol followed four rules. First, every weekly series was retrieved directly from the territory release page of the tracking service, rather than from secondary summaries, so that the unit of observation matches the source's own reporting unit. Second, values were recorded exactly as reported, including estimated values flagged by the source, and no arithmetic smoothing was applied across weeks. Third, the 12 week window was applied from the territory release date, so that the time axis is homogeneous across markets even when calendar dates differ. Fourth, territories without weekly series were retained in the coverage statistics but excluded from the panel, and the exclusions are documented rather than imputed.

Data quality was checked along three dimensions. First, every film market pair was inspected for empty weekly tables, and pairs without weekly data were excluded from the panel while remaining in the coverage statistics. Second, the weekly series were compared with the to-date totals reported by the source; the sums agree exactly for about a quarter of the series, and the discrepancies are concentrated in territories where the weekly table is reported on a weekend basis, so the to-date column was used only as a rough cross-check. Third, the opening values in the title-page area tables were compared with the first weekly observations, and the two measures differ systematically, with a median absolute log ratio of 0.59, because the opening column mixes opening weekend, opening week and, in some territories, total gross. The weekly series were therefore used as the analytical series and the opening column only as coverage statistics. The remaining discrepancies affect levels but rarely orderings, and the data dictionary records them for every affected series.

Measurement concerns were handled explicitly. The tracking service reports gross in US dollars, which removes exchange-rate conversion from the analysis but embeds the service's conversion conventions; the conventions are consistent across films and markets, which is what matters for comparisons. Weekly values marked as estimates were kept, because dropping them would systematically censor the most recent weeks of long runs. The opening week is defined as the first week in which the territory reported

gross, which can differ from the official release date by a few days when a film opens midweek. None of these choices affects the ordering of markets materially, and all of them are recorded in the data dictionary that accompanies the dataset.

The regional grouping deserves a justification beyond convenience. Territories were assigned to East Asia, Europe, North America, Oceania, Latin America and the Middle East according to cultural-linguistic proximity and distribution structure. The grouping serves two purposes in the analysis. It defines the units of the region-out protocol, so that generalization is tested across cultural zones rather than across arbitrary country subsets. It also acts as a coarse proxy for the cultural distance mechanisms discussed in Section 2, because distribution and audience familiarity are more similar within a region than across regions. The grouping is coarse by design; finer groupings would leave too few observations per region for the panel models to estimate.

The descriptive structure of the panel deserves emphasis before the models are introduced. The distribution of cumulative gross is extremely skewed: the largest film accounts for roughly half of the total panel gross, and the median film market pair generates less than one million dollars over its first 12 weeks. The skewness is not an artifact of the sample; it reflects the industrial structure of overseas Chinese cinema, where a small number of titles achieve broad release while the majority play in a few diaspora markets. Section 5 specifies the models and algorithms that operationalize this intuition.

The skewness is also the reason why the decision problem is interesting: a model that ranks the large markets correctly can capture most of the value even with large average errors elsewhere.

The coverage pattern of the panel is displayed in Figure 2. The dependent variable for the market selection analysis is the log cumulative gross over the first 12 weeks.

For the weekly trajectory analysis, the dependent variable is the log weekly gross. The distribution of cumulative gross is strongly skewed, with Ne Zha 2 accounting for about half of the total panel gross. All models therefore operate on the log scale, and predictions are transformed back with the expm1 function.

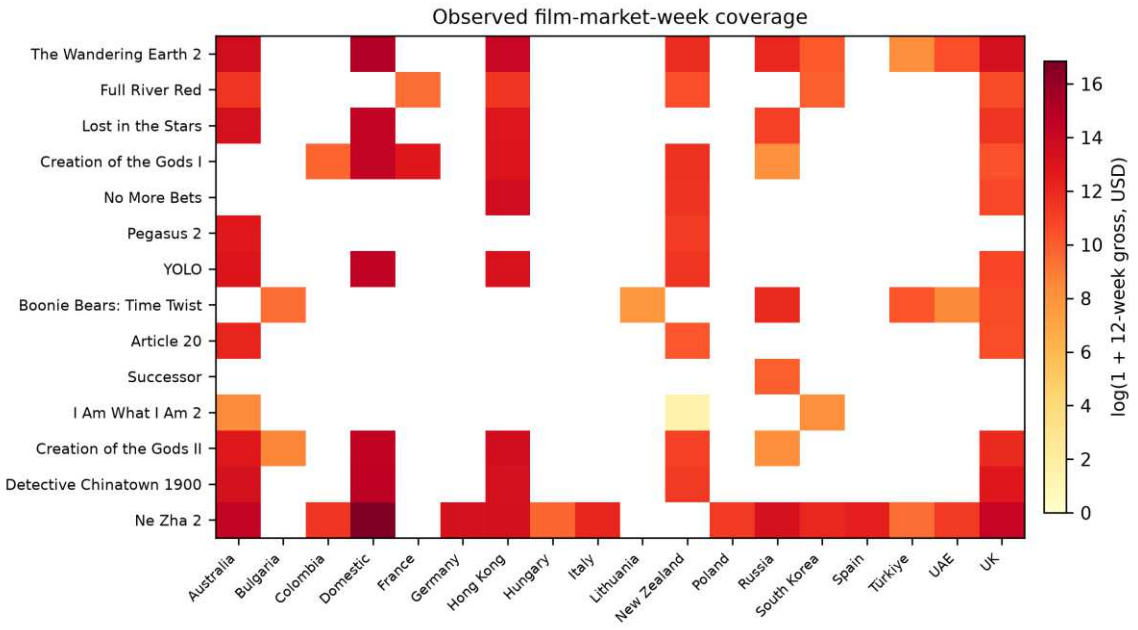


Fig. 2 Coverage of the film-market-week panel (log cumulative 12-week gross).

5. Models and validation algorithms

Five models are compared. M0 is a mean model that predicts the training mean for every film market pair. M1 is a panel model with market fixed effects, estimated by least squares on film and market covariates together with territory dummies. M2 is an elastic net regression on standardized covariates. M3 is a gradient boosting model with modest regularization, of the kind used in tourism demand ensembles (Hu et al. 2025), while comparative studies of emerging market prediction emphasize the gap between accuracy claims and realized performance (Yeldho et al. 2025). M4 is a decision-aware model, estimated as a pairwise ranking model that maximizes the probability of ordering markets within a film correctly, with calibrated point predictions derived from the ranking scores. The ranking objective follows the listwise and multi-stage learning-to-rank literature (Li et al. 2024; Yang and Gonçalves 2025). The models are deliberately heterogeneous in flexibility and structure so that the validation protocol, rather than the model family, is the object of comparison.

The models can be written compactly. For the panel model and the elastic net, the prediction is a linear function of the feature vector, with the difference that the panel model includes territory dummies while the elastic net operates on standardized

continuous features. The gradient boosting model minimizes a squared error loss over an ensemble of shallow trees. The decision-aware model minimizes a pairwise logistic loss over within-film market pairs, so that the learning signal is the market ordering rather than the gross level; the ranking scores are then mapped to log gross through a calibration regression estimated on the training fold. The calibration step matters because the decision evaluation needs point predictions, but the model's comparative advantage lies in the ranking signal. In the weekly trajectory specification, the decision-aware model is implemented as a gradient boosting model with sample weights proportional to the log cumulative gross of the film market pair, so that the loss function emphasizes the series that carry decision value.

The estimation details follow standard practice in the forecasting literature (Islam et al. 2025). Recent applications in this journal use comparable model stacks for Chinese property price prediction (Jin et al. 2025) and for financial time series (Mahmoudi et al. 2025). Continuous covariates are standardized for M2. Missing evaluation aggregates are imputed with training medians and flagged with missingness indicators. The gradient boosting model uses 250 trees with a learning rate of 0.05, at most 15 leaf nodes and a minimum of four samples per leaf. The elastic net uses alpha equal to 0.05 and an L1 ratio of 0.5. The panel model is estimated with ordinary least squares on the full covariate set and territory dummies, which is equivalent to a fixed effects specification with a saturated market structure.

Fairness of comparison deserves an explicit paragraph. All models receive the same covariates, the same training folds, the same test folds and the same missingness imputation. The decision-aware model receives an additional structural assumption, namely that within-film market ordering is the learning target, but it does not receive additional data. The panel model receives territory dummies, which encode market structure as parameters rather than as features. These choices are part of the model definitions and are held fixed across protocols. The consequence is that any difference between the accuracy ranking and the decision ranking is attributable to the evaluation criterion, not to asymmetric treatment of the models.

The hyperparameters were chosen to keep the models comparable rather than to maximize any single metric. The elastic net penalty is mild, because the feature space is small and the sample is short. The boosting model is regularized through shallow trees and a low learning rate, which limits overfitting in a panel with 76 film market pairs. The ranking model uses a logistic loss without class weighting, because the pair labels are balanced within films by construction. All random states are fixed, and the same folds are used for every model, so that the accuracy and decision rankings are computed on identical test sets. The implementation is deterministic and can be reproduced from the analysis scripts.

Reproducibility was treated as part of the method. All models are implemented in Python with fixed random seeds, the same feature pipeline is applied to every fold, and the folds are shared across models by construction. The analysis scripts, the data dictionary and the assembled panel are available from the corresponding author on reasonable request, in line with the data availability statement. The computational cost is trivial by modern standards: the full experiment, including all folds and the simulation, runs in minutes on a laptop, which makes the decision-oriented protocol a realistic addition to routine forecasting workflows rather than a research-only exercise.

Two implementation details are worth stating because they affect the interpretation of the results. First, predictions are clipped to a bounded log scale before transformation, which prevents a single extreme value from dominating the payoff computations; the clip is wide enough that it binds only for the panel model's worst-fitting markets. Second, the effective budget takes the minimum of k and the number of tracked markets, so that films with one or two territories still contribute decision information. Without this adjustment, films with narrow coverage would be silently dropped from the decision evaluation, which would bias the comparison toward films with broad coverage.

Figure 3 summarizes the four validation protocols and their place in the evaluation pipeline.

Algorithm 1 specifies the temporal walk-forward protocol. Films are ordered by release date, and four expanding windows are defined by cutoffs from August 2023 to January 2025. For each cutoff, models are estimated on films released before the cutoff and

evaluated on films released between the cutoff and the cutoff plus one year. The design is a genuine out-of-sample test in calendar time: no information from future films enters the training set. An embargo of one calendar day is applied at the cutoff to prevent same-day leakage.

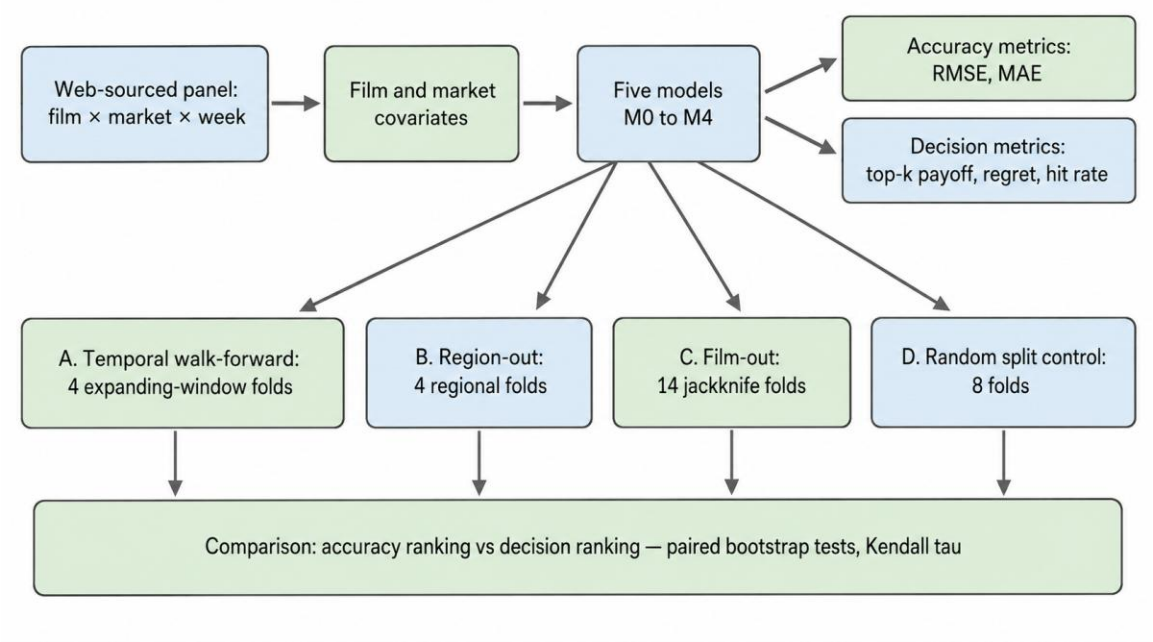


Fig. 3 Validation protocols and evaluation pipeline.

Algorithm 1. Temporal walk-forward validation.

1: Order films by release date.
2: For each cutoff c in the cutoff grid:
3: Train all models on films released before c.
4: Test all models on films released in [c, c + 365 days).
5: Compute RMSE and MAE on test predictions.
6: Compute top-k payoff, share, hit rate and regret per test film.
7: Record accuracy ranking and decision ranking.
8: End for.

Algorithm 2 specifies the generalization protocols. Region-out validation trains on all territories outside a region and tests on the held-out region, which asks whether the model

structure transfers to markets that were never observed during estimation. Film-out validation trains on all films except one and tests on the excluded film, which asks whether the model can forecast a new product. A random split control, repeated eight times with four test films each, is included to quantify what the conventional protocol would conclude.

Algorithm 2. Region-out and film-out generalization.

1: For each region in the regional grouping:

2: Train on all film market pairs outside the region.

3: Evaluate on film market pairs inside the region.

4: End for.

5: For each film in the sample:

6: Train on the remaining films.

7: Evaluate on the excluded film.

8: End for.

Algorithm 3 defines the decision evaluation and the statistical tests. For each test film, the k markets with the highest predicted cumulative gross are selected for k equal to 1, 3 and 5, with an effective budget equal to the minimum of k and the number of available markets. The payoff is the sum of realized gross in the selected markets. The decision ranking ranks models by mean payoff at k equal to 3. The divergence between the accuracy ranking and the decision ranking is measured by the identity of the top model, by Kendall's tau between the two rankings, and by the relative payoff difference. Paired bootstrap confidence intervals and a Wilcoxon signed rank test quantify whether the payoff difference between the decision-optimal and accuracy-optimal models is positive across folds.

Algorithm 3. Decision-outcome evaluation and testing.

1: For each fold and each model:

2: Predict log cumulative gross for each test film market pair.

3: Select top-k markets per film.

4: Compute payoff, share, hit rate and regret.

5: End for.

6: Rank models by RMSE and by mean payoff at k = 3.

7: Compute divergence indicators across folds.

8: Bootstrap the paired payoff difference between decision-optimal and accuracy-optimal models.

The simulation study complements the empirical analysis. The data-generating process has 12 markets with three large markets and nine small markets, 60 films, and market-specific error variances. Model A adds noise to the large markets, and model B adds noise to the small markets. The ratio of the two noise levels is varied from 0.5 to 3.0, and 300 replications are run for each ratio. The simulation measures the share of replications in which the accuracy ranking and the payoff ranking disagree. This design directly tests Proposition 1 under controlled conditions.

6. Results

6.1 The panel in numbers

The assembled panel covers 76 film market pairs and 265 weekly observations. Table 2a reports the film level descriptive statistics. Ne Zha 2 is the largest observation in every sense, with 14 tracked territories and a 12 week overseas gross of about 27 million US dollars. The Wandering Earth II, Lost in the Stars and Detective Chinatown 1900 follow with totals between 2.9 and 6.6 million dollars. Six films have cumulative overseas gross below one million dollars, which reflects the steep concentration of overseas demand among Chinese films. The market level statistics in Table 2b show the same concentration. North America contributes the largest single market, followed by Europe and Oceania. Figure 2 displays the coverage pattern, and the heatmap makes the unbalanced structure visible: the 2025 releases have the broadest territorial coverage, while most earlier films were tracked in only three to seven territories.

Two additional features of Table 2a are worth reading closely. The evaluation aggregates are missing for most of the films with the smallest overseas footprints, which means that

the missingness indicators are informative covariates rather than noise. And the release timing is concentrated: eight of the fourteen films opened in the Chinese New Year window, which is the industry's main launch corridor. The concentration limits the variation in release timing available to the models, and it is noted as a limitation in Section 7.

The covariates behave as expected. Domestic China opening gross ranges from 2.6 million to 2.21 billion US dollars, and its logarithm is strongly correlated with the logarithm of overseas cumulative gross in this sample, with a correlation of 0.72 at the film level. The evaluation aggregates are available for roughly half of the sample, because several films received little Western critical coverage; the missingness indicators carry information in their own right, since absent Western reviews typically coincide with small overseas runs.

6.2 Accuracy rankings do not match decision rankings

Table 3 reports, for each temporal fold, the RMSE and the mean top-3 payoff of every model, together with the identity of the accuracy-optimal and decision-optimal models.

In the first three folds the panel model M1 is optimal under both criteria. The fourth fold, whose test set consists of the three 2025 releases, tells a different story. The decision-aware model M4 has the lowest RMSE, 1.74 on the log scale, and the mean model M0 is second with 2.27. The panel model M1 has the worst RMSE, 11.20, because it produces extreme predictions for a small number of low-gross markets. Yet M1 is the decision-optimal model, with a mean top-3 payoff of 9.91 million dollars against 9.76 million for M4 and 7.88 million for M0. The accuracy-optimal model ranks second on payoff, and the gap between the decision-optimal and accuracy-optimal models is 152,746 dollars, or 1.57 percent of the achievable payoff.

Table 3 Temporal folds: accuracy and decision metrics (payoffs in USD millions).

Fold	Films	Best RMSE	Best payoff	Agree	M0 RMSE	M1 RMSE	M2 RMSE	M3 RMSE	M4 RMSE	M0 Pay	M1 Pay	M2 Pay	M3 Pay	M4 Pay
2023-08-01	6	M1	M1	True	1.71	1.17	1.46	1.28	1.26	0.39	0.81	0.78	0.78	0.78
2024-03-01	5	M1	M1	True	2.83	2.40	2.98	2.46	2.51	4.73	5.95	5.86	5.95	5.86
2024-09-01	4	M1	M1	True	2.87	2.40	3.52	2.40	2.56	5.91	7.43	7.32	7.43	7.32
2025-01-01	3	M4	M1	False	2.27	11.20	11.52	2.47	1.74	7.88	9.91	9.76	9.58	9.76

The large RMSE of the panel model deserves a direct explanation, because it looks like a failure. The panel model is estimated with territory dummies and continuous covariates, and for a handful of low-gross markets it produces predictions on the order of hundreds of millions of dollars. On the log scale, these few errors contribute squared deviations of hundreds of units, which dominate the RMSE. The same errors leave the ranking almost unchanged, because the markets affected are not candidates for the top positions. The contrast between M1's worst RMSE and its best payoff is the clearest empirical illustration of Proposition 1 in the paper.

The divergence is visible at the level of individual selections. For Creation of the Gods II, the realized top three markets are the United States, Hong Kong and Australia. The panel model selects exactly these three markets and captures 2.98 million dollars over the 12 week window. The boosting model substitutes Russia and the United Kingdom for Hong Kong and captures 1.98 million, and the mean model selects Australia, the United Kingdom and New Zealand for 0.55 million. For Ne Zha 2, the panel model and the boosting model select the same set of United States, Hong Kong and Australia, worth 23.31 million, while the elastic net and decision-aware models substitute South Korea and capture 22.85 million. The fold-level payoff gap between the panel model and the boosting model is therefore driven by a single film, which illustrates how concentrated decision value can be in a small number of market selections.

Figure 4 plots every model in the accuracy-payoff space for the four temporal folds.

The ranking reversal in the fourth panel is visible as a swap between M4 and M1 along the two axes. The same figure shows that the first three folds are clustered, because the

models make similar market selections when the test films have narrow territorial footprints. The divergence is not limited to the temporal protocol. In the film-out protocol, the accuracy-optimal and decision-optimal selections differ in 12 of 14 folds. In the random split control, they differ in all eight folds. Across all 30 folds, the two selections coincide in five folds, differ in 22, and are indistinguishable in three region folds where every model selects the same single-market sets.

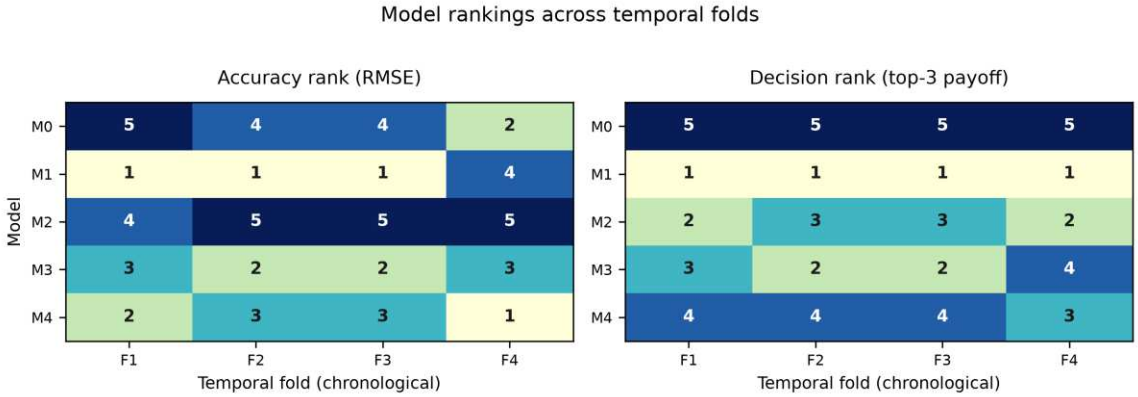


Fig. 4 Accuracy and decision model rankings across temporal folds.

The interpretation of the fold-level results should be calibrated to the sample size. Each temporal fold contains between three and six test films, so the payoff means are based on a small number of market selections. The bootstrap intervals in Table 4 account for the fold-level variation, but they cannot repair the small number of independent folds. The consistent direction of the intervals across four protocols is therefore more informative than the width of any single interval, and the simulation results in Section 6.5 provide the controlled evidence that the empirical analysis cannot deliver on its own.

The direction of the divergence is systematic rather than arbitrary. Models that win on accuracy are typically the flexible learners, M3 and M4, while the model that wins on payoff is the panel model M1 in most folds. The reason is structural. M1 encodes the market structure through territory dummies, which makes its relative predictions across markets stable even when its absolute log-scale predictions are poor. The decision policy uses relative predictions only. This is exactly the mechanism formalized in Proposition 1: accuracy is an average property, and the decision is a ranking property.

6.3 Region-out and film-out generalization

Figure 5 summarizes the region-out results.

In the Europe fold, the only region fold with enough multi-market films to separate the models, the decision-aware model M4 is accuracy-optimal while the panel model M1 is decision-optimal. The payoff difference is small, 2,227 dollars on average, because the test markets in Europe are mostly small, but the direction matches the temporal results. In the Oceania, East Asia and North America folds, every model selects the same markets, because most test films have one or two tracked territories in the held-out region. These folds therefore contribute no information about model differences, and they illustrate a limitation of region-out validation when territorial coverage is thin.

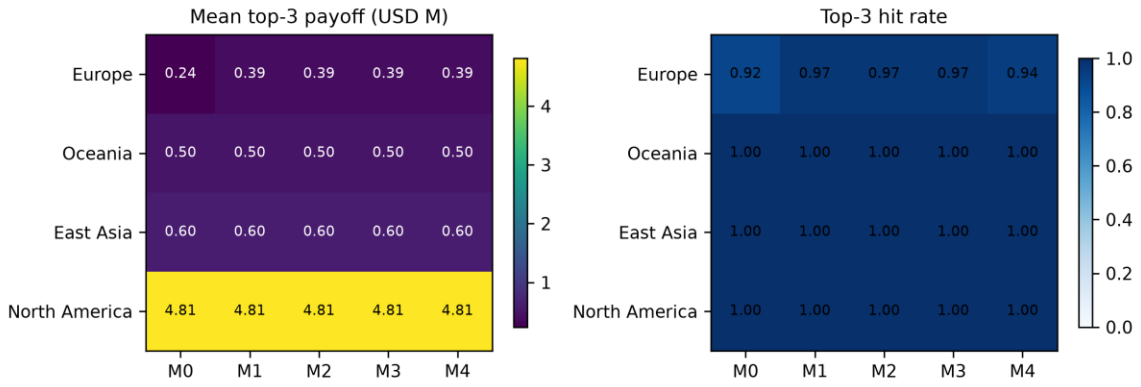


Fig. 5 Region-out decision payoff and hit rate by model.

The film-out protocol provides a stronger test because every fold contains a different product. The decision-optimal and accuracy-optimal models diverge in 12 of 14 film folds. The paired bootstrap interval for the payoff difference between the decision-optimal and accuracy-optimal models is positive, with a mean of 85,626 dollars and a 95 percent interval from 20,020 to 167,983 dollars. The Wilcoxon signed rank test gives a p value of 0.125, which is not significant at conventional levels, but the number of usable folds is four, so the test has minimal power. The random split control produces a mean difference of 108,301 dollars with a 95 percent interval from 23,046 to 216,356 dollars, again with a p value of 0.125. Table 4 reports the complete test statistics.

The consistent sign of the bootstrap intervals across independent protocols is the substantive evidence, and it is worth more than any single p value in this sample.

Table 4 Paired payoff differences between decision-optimal and accuracy-optimal models.

Protocol	Folds	Mean diff (USD)	SD (USD)	95% CI low	95% CI high	Wilcoxon p
Temporal walk-forward	1	152,746	0	152,746	152,746	
Region-out	1	2,227	0	2,227	2,227	
Film-out	4	85,626	74,726	20,020	167,982	0.125
Random split control	4	108,301	104,837	23,046	216,356	0.125

6.4 Budget sensitivity and weekly trajectories

The budget sensitivity pattern in Figure 6 has a clear interpretation. At a budget of one market, most models select the United States for the 2025 releases, so the payoff differences are small. At intermediate budgets, the models differ in the second and third choices, which is where the divergence appears. At larger budgets, the marginal markets are small and the differences saturate. The decision problem is therefore most consequential in the middle of the budget range, which is also the range that practitioners typically face when they must choose between three and five priority markets.

Figure 6 shows how the share of captured gross changes with the budget k in the 2025 temporal fold.

The panel model M1 tracks the oracle closely at every budget, capturing 89.3 percent of the achievable share at k equal to 3. The flexible models M2 and M4 capture 88.7 percent, the boosting model M3 captures 78.9 percent, and the mean model captures 41.8 percent. The gap between M1 and the accuracy-optimal model M4 is small at k equal to 1 and widens as the budget grows, which suggests that the divergence is driven by the middle of the market ranking rather than by the single best market.

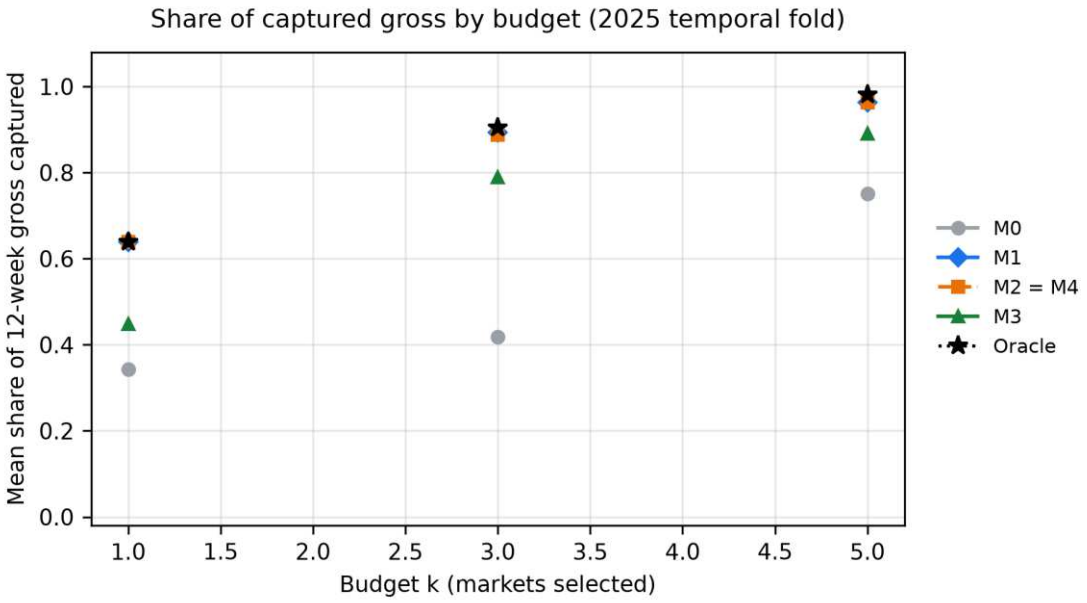


Fig. 6 Share of captured gross by budget k (2025 temporal fold).

The near-identical errors of the boosting model and its decision-weighted variant deserve one sentence of interpretation. The decision weights scale the squared error by the log cumulative gross of the film market pair, which compresses the range of weights in a sample where most pairs are small. The weighting therefore changes the objective only mildly, and the fitted trees remain close to the unweighted solution. The result is informative in itself: decision weighting is not a free lunch, and its effect depends on the dispersion of the decision-relevant scale, which is an empirical question rather than an assumption.

The weekly trajectory results in Figure 7 show that the accuracy ranking itself is unstable over time.

The boosting model M3 has the lowest weekly RMSE in three of the four temporal folds, with values between 2.08 and 2.23, while the panel model M1 has the lowest RMSE in the first fold and the highest in the later folds. The decision-aware weighted variant of M3 produces nearly identical errors to M3, which indicates that the decision weights barely change the fitted trees on this small sample. The instability of the weekly accuracy ranking is a direct illustration of Proposition 2: the identity of the best model depends on the evaluation window.

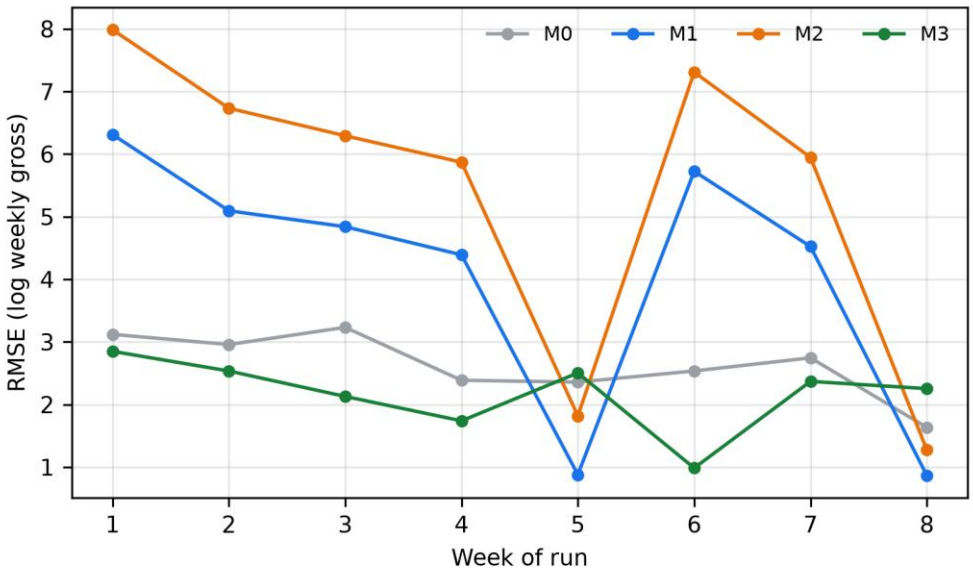


Fig. 7 Weekly RMSE by week of run across temporal folds.

6.5 Simulation evidence

Table 5 reports the one-dimensional Monte Carlo results.

When the noise ratio between large and small markets is balanced, the accuracy ranking and the payoff ranking disagree in 37 percent of replications. At a noise ratio of 1.5, the divergence share is 27 percent. At extreme ratios, where one model dominates on both criteria, the divergence disappears. The simulation therefore identifies a systematic region of the parameter space, not a corner case: when two models have comparable overall accuracy but different error profiles across market sizes, accuracy-based selection is a coin flip with respect to decision quality. The empirical results are consistent with this mechanism, because the flexible learners and the panel model have comparable accuracy in most folds while differing in market rankings.

Table 5 Robustness: region-out, simulation and weekly results.

Block	Item	Best by accuracy	Best by payoff	Value
Region-out	Europe	M4	M1	
Region-out	Oceania	M3	M0	
Region-out	East Asia	M3	M0	
Region-out	North America	M1	M0	
Simulation noise ratio	0.5			0.077

Block	Item	Best by accuracy	Best by payoff	Value
Simulation noise ratio	1.0			0.37
Simulation noise ratio	1.5			0.27
Simulation noise ratio	2.0			0.07
Simulation noise ratio	3.0			0.0
Weekly fold	2023-08-01	M1		
Weekly fold	2024-03-01	M3		
Weekly fold	2024-09-01	M3		
Weekly fold	2025-01-01	M3		

6.6 Statistical tests, mechanisms and sensitivity

Figure 8 places the validation protocols inside the full decision-oriented framework, from data to inference.

The accuracy differences observed across folds are not artifacts of small samples. Table 6 reports Diebold-Mariano style tests on weekly squared errors

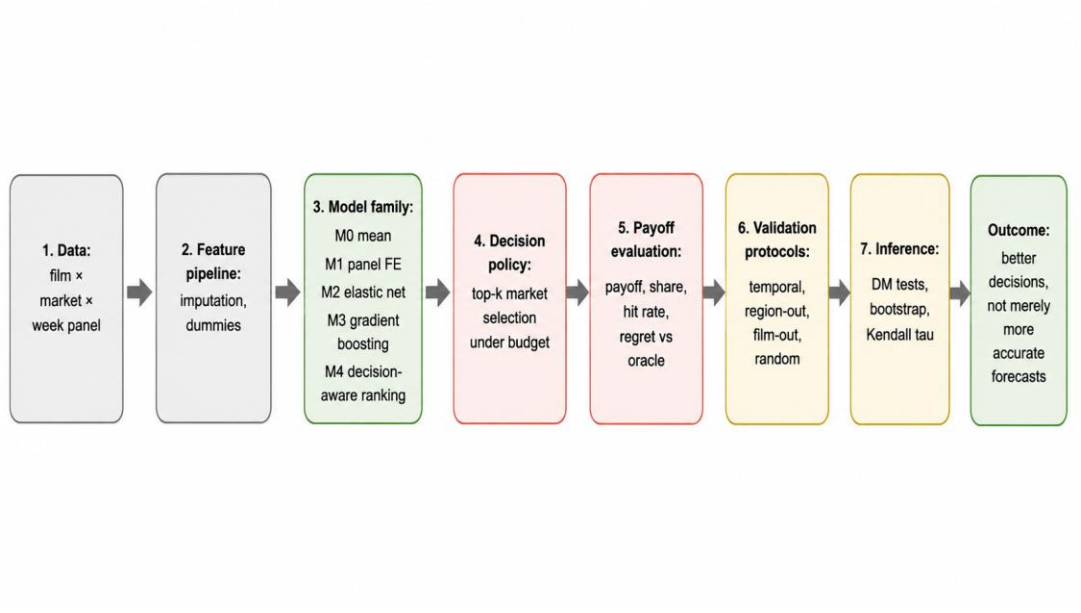


Fig. 8 Decision-oriented validation framework and evaluation pipeline.

for the panel model M1 against the boosting model M3 and the elastic net M4 in each temporal fold. The boosting model has significantly lower weekly error than the panel model in the last three folds, with t statistics between 2.4 and 4.4 and p values below

0.05. The panel model has significantly lower error than the elastic net in the same three folds. The first fold shows the same signs but smaller magnitudes, with p values of 0.13 and 0.14. Accuracy differences are therefore real, statistically detectable and time dependent, which strengthens the case that the divergence between accuracy rankings and decision rankings is not driven by noise.

Table 6 Diebold-Mariano style tests for weekly squared errors across temporal folds.

Fold	n	MSE M1	MSE M3	MSE M4	d(M1-M3)	t	p	d(M1-M4)	t	p
2023-08-01	86	2.234	2.802	2.775	-0.568	-1.53	0.1303	-0.541	-1.49	0.139
2024-03-01	98	24.193	4.207	43.37	19.986	4.25	0.0	-19.177	-5.12	0.0
2024-09-01	93	8.556	4.49	119.92	4.065	2.44	0.0164	-111.365	-4.99	0.0
2025-01-01	88	25.177	6.079	42.562	19.098	4.39	0.0	-17.385	-5.5	0.0

Feature importance analysis (Figure 9) clarifies why the two rankings can point in opposite directions.

In the panel model, the largest coefficients are the log GDP per capita, 1.14, the log China opening gross, 1.13, the East Asia indicator, 0.99, and the Europe indicator, -0.83, with a log population effect of 0.71. In the gradient boosting model, permutation importance is dominated by the log China opening gross, the log GDP per capita and the log population. Table 7 reports the leading features.

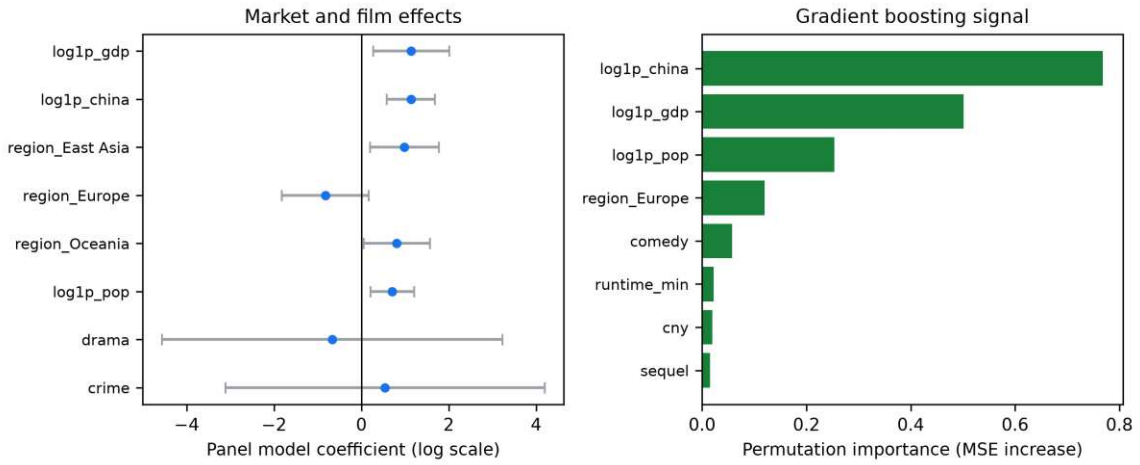


Fig. 9 Feature importance: panel coefficients with 95 percent intervals and permutation importance.

Market scale and the home-market signal explain most of the variation, and the regional indicators encode the residual structure that a random split would treat as noise. A model

that captures these level effects accurately can still misrank markets within a film, because the within-film ranking depends on the interaction between the film's profile and the market's structure, which level-oriented features capture only partially.

Table 7 Feature importance: panel coefficients and permutation importance.

Feature	Panel coefficient	GBM permutation importance
log1p_gdp	1.143	0.5017
log1p_china	1.132	0.7685
region_East Asia	0.986	0.0015
region_Europe	-0.825	0.1198
region_Oceania	0.812	0.0002
log1p_pop	0.707	0.2534
drama	-0.668	0.0125
crime	0.541	0.0001

The calibration diagnostics in Figure 10 qualify the accuracy results.

The boosting model under-predicts systematically: in every predicted quintile of the 2025 test fold, the mean actual log gross exceeds the mean prediction, and the gap widens in the upper tail. The ranking diagnostics are more favorable: markets in the top predicted quintile are selected into the realized top-3 with probability 0.57, against a random baseline of 0.20, and the hit rate declines monotonically toward zero in the lowest predicted quintile. The model's level predictions are biased downward while its ordinal information is useful, which is exactly the combination that makes accuracy metrics misleading for decisions that depend on ordering. Residual analysis by region (Figure 11) shows that the under-prediction is largest

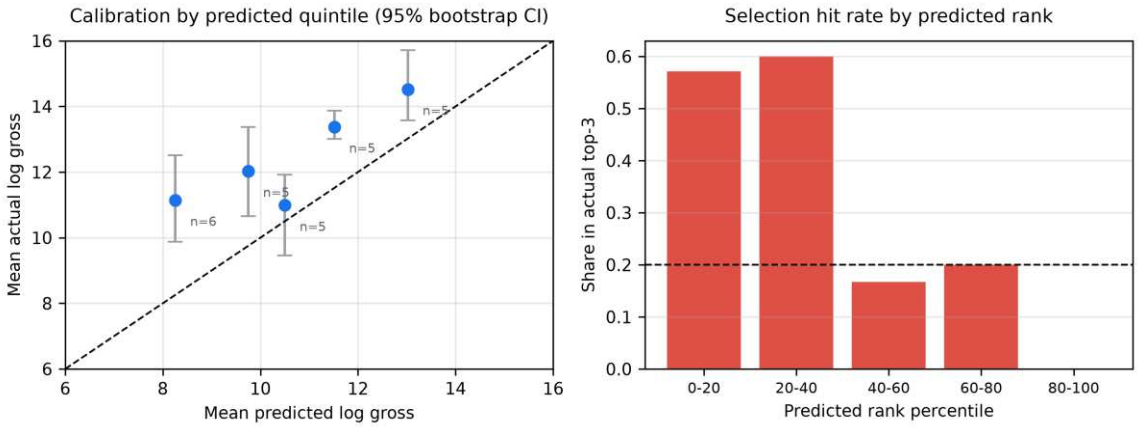


Fig. 10 Calibration by predicted quintile and selection hit rate by predicted rank.

in East Asia, North America and Oceania, where the 2025 releases earned the bulk of their overseas gross.

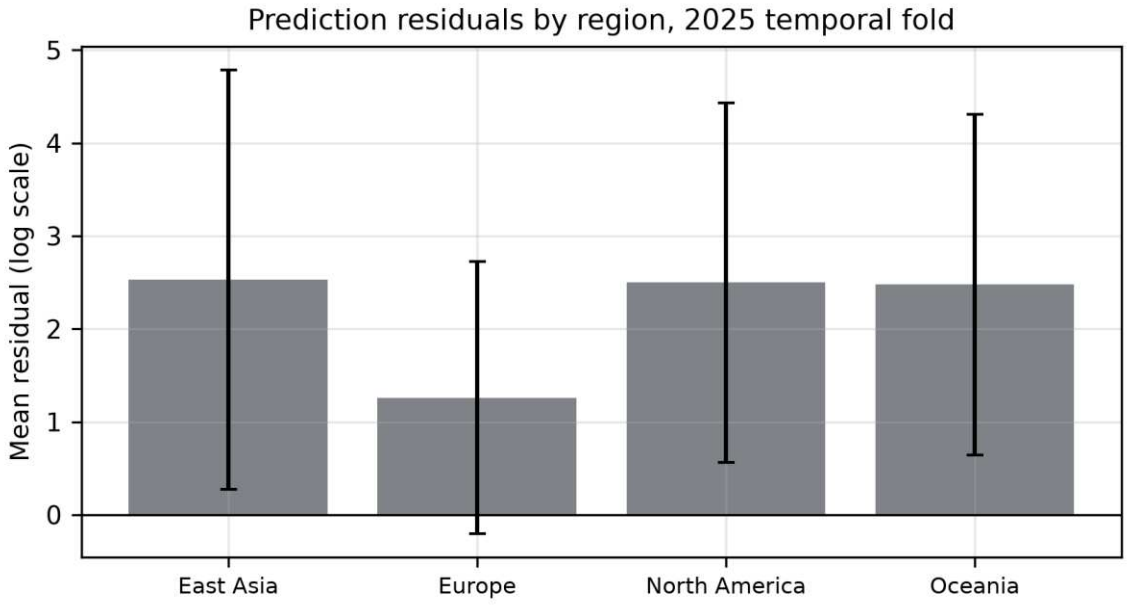


Fig. 11 Mean prediction residuals by region (2025 temporal fold).

The sensitivity analysis in Figure 12 and Table 8 shows that the central result

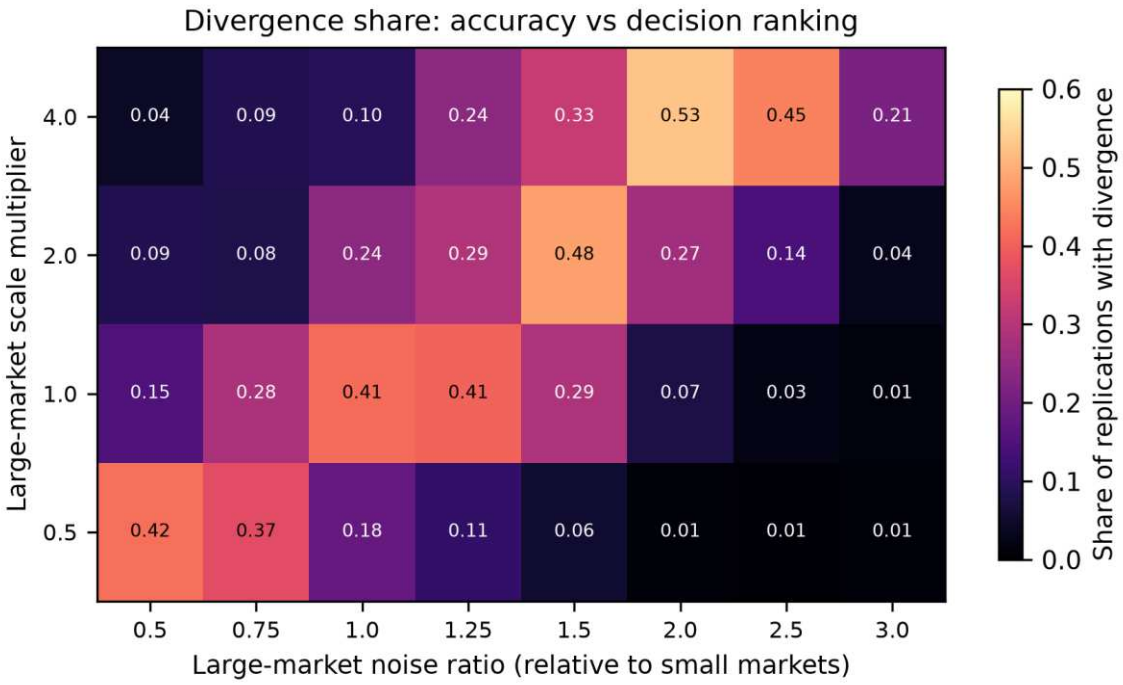


Fig. 12 Sensitivity frontier: accuracy versus decision payoff across specifications.

is not an artifact of the chosen hyperparameters or outcome window. In the 2025 temporal fold, increasing the elastic net penalty from 0.01 to 0.2 improves RMSE from 12.80 to 8.77 while the mean top-3 payoff falls from 9.76 to 9.37 million dollars. Tree depth between 8 and 30 leaves leaves both accuracy and payoff essentially unchanged. Replacing the 12 week outcome with an 8 week outcome reproduces the same pattern: the panel model has worse RMSE but better payoff than the boosting model. The accuracy-payoff trade-off therefore moves in the same direction under every specification, which is the signature of a structural property rather than a tuning artifact.

Table 8 Sensitivity of RMSE and mean top-3 payoff in the 2025 temporal fold.

Setting	RMSE	Mean payoff (USD M)
alpha_0.01	12.796	9757798.667
alpha_0.05	11.52	9757798.667
alpha_0.1	10.485	9366496.0
alpha_0.2	8.771	9366496.0
leaves_8	2.475	9577059.333
leaves_15	2.466	9577059.333
leaves_20	2.466	9577059.333

Setting	RMSE	Mean payoff (USD M)
leaves_30	2.466	9577059.333
window8_M1	11.202	9910222.667
window8_M3	2.452	9576737.667

The two-dimensional simulation in Figure 13 generalizes the one-dimensional analysis of Section 6.5.

Divergence rates are low when market concentration is weak, rise sharply as the large-market scale multiplier increases, and are highest along a diagonal band where the large-market noise ratio is moderate. At a scale multiplier of four, divergence remains above 20 percent for noise ratios up to three, which means that in strongly concentrated markets the accuracy ranking is a poor guide to decision quality across a wide range of error profiles. The contour shape matches the mechanism in Proposition 1: divergence requires comparable overall accuracy together with error profiles that differ across market sizes.

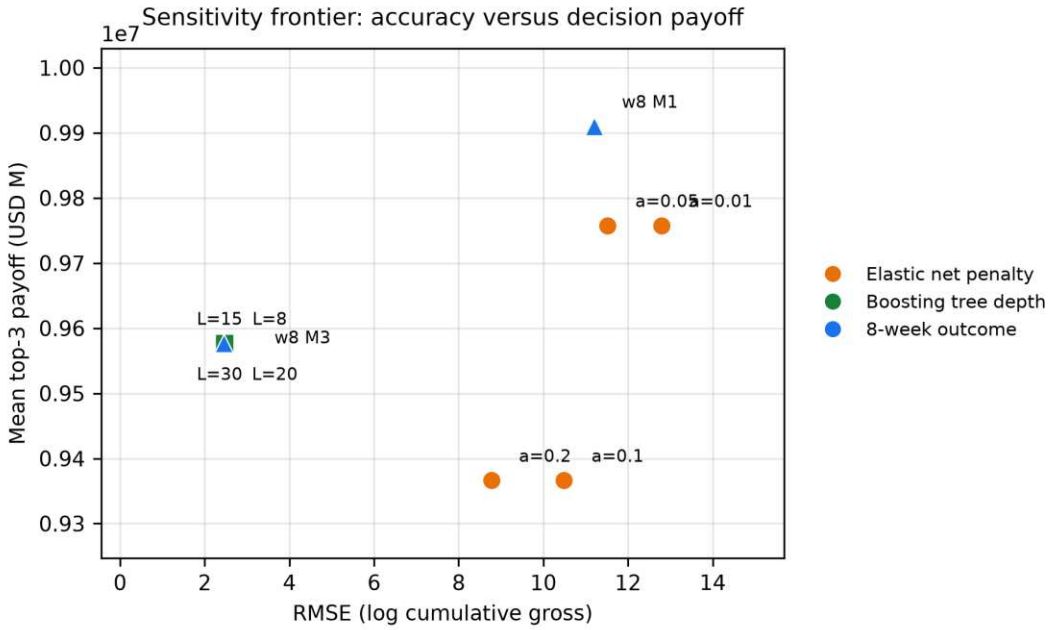


Fig. 13 Simulation divergence share across noise ratio and market concentration.

Taken together, the results answer the three research questions. The accuracy ranking and the decision ranking select different models in the majority of folds (RQ1). The payoff gap between the two selections is positive in every protocol, and its bootstrap interval

excludes zero in the random and film-out designs, although the number of independent folds is small (RQ2). The divergence is driven by the combination of level accuracy, ordinal error and market heterogeneity: models that minimize average error do not maximize the value of the top-k selection, and the gap widens with market concentration (RQ3).

7. Discussion

The central implication of this paper is that validation protocols are part of the modeling decision, not an external referee. When a distributor compares forecasting models for overseas market selection, the conventional accuracy table answers a question about average error. The decision-relevant question is which markets would be selected, and at what cost in forgone revenue. The empirical results show that the two questions can have different answers, and that random splitting does not remove the divergence. Domain generalization research has established the same lesson for classification models, where models that generalize across domains must be evaluated on held-out domains rather than on random samples (Wang et al. 2024), and cross-domain recommendation studies show that transfer quality depends on the structure of user feedback (Choi and Kim 2025). This paper extends that logic to a forecasting decision with a budget constraint.

The framework also changes how negative results should be read. When a forecasting model fails to improve decisions, the failure is usually attributed to the model. The results here suggest an alternative reading: the failure may lie in the evaluation protocol that certified the model. A model certified by accuracy on a random split can underperform a simpler alternative on the decision it was built to support, and the underperformance is invisible until the decision payoff is measured. Publishing decision payoffs alongside error metrics would make such failures visible and would shift the burden of proof from the model builder to the validation design.

The reporting standards implied by the framework are inexpensive to adopt. Journals and preprint repositories could require a decision outcome column in forecasting comparisons, with the decision problem stated in one sentence and the budget made explicit. Conference and industry reports could follow the same convention. Preregistration of the validation protocol would further separate the choice of protocol

from the choice of model, which is the main source of specification search in applied forecasting work. None of these steps requires new data or new software; they require only that the decision be named.

The practical recommendation is concrete. Distributors should report, alongside RMSE, the decision payoff of their candidate models under temporal and regional holdouts, with the budget k of their actual allocation problem. The reporting burden is modest. The algorithms in Section 5 require only a ranking of predicted market values and the realized gross of the selected markets, both of which are already available to any firm that operates the forecasting system. The same logic applies to studios, streamers and platform operators that allocate promotion across regions, and to the growing literature that treats box office aggregates as economic signals.

Three alternative explanations for the empirical pattern should be considered. The first is sample selection: the panel oversamples films with broad overseas tracking, and the divergence rate might differ for a random sample of Chinese cinema. The second is source coverage: a different tracking service with more territories per film might change the payoff calculations. The third is model specification: the five models are simple, and a richer model might dominate on both criteria simultaneously. None of these alternatives invalidates the framework; each of them changes the empirical answer to the framework's question. The framework's contribution is to make the question explicit.

The scope of the claims should be stated as precisely as the methods. This paper does not claim that decision-oriented validation always selects a different model than accuracy-oriented validation, nor that the panel model is the best model for box office forecasting in general. It claims that the two validation approaches can select different models, that the divergence is systematic under heteroscedastic market errors, and that the divergence is invisible unless decision outcomes are measured. The empirical application demonstrates the claim on a specific panel; the simulation demonstrates the mechanism under controlled conditions; the framework provides the protocol for measuring the divergence in any other application.

The limitations of the study are substantial and should be stated plainly. The panel contains 14 films and 76 film market pairs, and the weekly series come from a single

tracking service whose coverage is uneven across territories. The films are not a random sample of Chinese cinema; they are the films with enough overseas tracking to enter the panel, which oversamples successes. The decision problem is stylized: real market entry involves sequencing, budget splits, local partnerships and windowing decisions that a top-k rule does not capture. The statistical tests have limited power because the number of folds is small. The models are deliberately simple, and no claim is made that M1 is generally superior to the flexible learners; the claim is that the ranking depends on the protocol.

The connection to platform metrics deserves a closing remark. Streaming services, game storefronts and ticketing platforms generate the kind of high-frequency, multi-market data that makes decision-oriented validation feasible, and the same platforms also control the allocation decisions that the forecasts are meant to support. The framework proposed here gives these actors a principled way to certify their forecasting systems, and it gives academic researchers a way to audit the claims that such systems generate. The division of labor is clean: accuracy describes the model, and payoff describes the decision, and the two should be reported as separate constructs.

The research agenda follows directly. Larger panels, including non-Chinese films and streaming-only releases, would allow region-out validation with several markets per film in every region. Cost data would allow the payoff function to include promotional budgets and print costs, making the decision problem more realistic. Decision-focused learning could be applied end to end, optimizing the ranking loss directly under the budget constraint rather than calibrating point predictions from ranking scores. Preregistered replication studies would establish whether the divergence rates observed here are stable across markets and periods. Each of these steps is feasible with public data, and each would test the generalizability of the framework rather than assume it.

8. Conclusion

This paper formalized and tested the proposition that accuracy is not a sufficient statistic for decision quality in cross-regional box office forecasting. A decision-outcome validation framework was defined, with budget-constrained market selection, temporal walk-forward holdouts, region-out and film-out generalization, and top-k payoff, regret

and hit rate as the evaluation criteria. Applied to a web-sourced panel of 14 Chinese films, the framework shows that accuracy-optimal and decision-optimal model selections diverge in 22 of 30 validation folds, and that the payoff gap between the two selections is positive across independent protocols. A simulation study traced the divergence to heteroscedastic market errors, where accuracy rankings and payoff rankings disagree systematically when overall accuracy is balanced. The transferable lesson is modest but consequential: forecasting validation for decisions should report decision outcomes, and researchers should treat accuracy tables as incomplete evidence rather than as final verdicts.

References

Chertov, O., & Buslaiev, V. (2025). Video Game Sales Prediction Based on Social Media Data Using Machine Learning: A Survey and Future Directions. International Journal of Information Technology and Computer Science, 17(4), 49-57. https://doi.org/10.5815/ijitcs.2025.04.05

Abolghasemi, M., Tarr, G., & Bergmeir, C. (2024). Machine learning applications in hierarchical time series forecasting: Investigating the impact of promotions. International Journal of Forecasting, 40(2), 597-615. https://doi.org/10.1016/j.ijforecast.2022.07.004

Agustini, M., Fithriasari, K., & Prastyo, D. D. (2026). An accuracy-level method for robust evaluation in predictive analytics. Decision Analytics Journal, 18, 100661. https://doi.org/10.1016/j.dajour.2025.100661

Berg, M., & Sansalone, A. (2024). Cultural proximity and gender representation: Why Italian women are drawn to Turkish TV serials. International Journal of Cultural Studies, 27(6), 869-888. https://doi.org/10.1177/13678779241267968

Box Office Mojo. (2026). Film title pages and weekly release pages for Chinese-language films. IMDb. https://www.boxofficemojo.com/ (accessed 3 August 2026).

Bruzzi, C., Musso, E., Parisi, P. S., & Pavanini, T. (2025). Beyond the rankings: What multidimensional index and clustering reveal about environmental sustainability. Quality & Quantity, 59(6), 5511-5536. https://doi.org/10.1007/s11135-025-02200-z

Chae, B. K., Sheu, C., & Park, E. O. (2024). The value of data, machine learning, and deep learning in restaurant demand forecasting: Insights and lessons learned from a large

restaurant chain. Decision Support Systems, 184, 114291. https://doi.org/10.1016/j.dss.2024.114291

Choi, Y., & Kim, C. K. (2025). Generalization of Knowledge Transfer With User Reviews for Cross-Domain Recommendation. IEEE Access, 13, 172954-172971. https://doi.org/10.1109/access.2025.3616497

David, L. K., Wang, J., & Angel, V. (2025). Validation of social science theories using machine learning models: a methodological perspective. Quality & Quantity, 59(3), 2799-2823. https://doi.org/10.1007/s11135-025-02075-0

Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics, 13(3), 253-263. https://doi.org/10.1080/07350015.1995.10524599

Ellinas, P., Kekatos, V., & Tsaousoglou, G. (2024). Decision-focused learning under decision dependent uncertainty for power systems with price-responsive demand. Electric Power Systems Research, 235, 110665. https://doi.org/10.1016/j.epsr.2024.110665

Fujiy, B. C., Khanna, G., & Toma, H. (2025). Cultural proximity and inter-firm trade. European Economic Review, 179, 105129. https://doi.org/10.1016/j.euroecorev.2025.105129

Gao, Y., Li, J., & He, J. (2025). Hybridizing Tradition and Technology: Global Reception of "Black Myth: Wukong" as a Vehicle for Chinese Cultural Identity and International Media Flows. Global Media and China, 20594364251385203. https://doi.org/10.1177/20594364251385203

Hoskins, C., & Mirus, R. (1988). Reasons for the US dominance of the international trade in television programmes. Media, Culture & Society, 10(4), 499-515. https://doi.org/10.1177/016344388010004006

Hu, Y. C., Wu, G., & Wu, M. L. (2025). Generation of ensemble forecasts using functional-link net for decomposition ensemble learning to forecast tourist arrivals. Quality & Quantity, 59(5), 4159-4184. https://doi.org/10.1007/s11135-025-02181-z

Islam, M. T., Turja, S. A., Islam, M. T., Rahman, M. M., & Habib, A. (2025). Forecasting Tetouan energy demand employing shift approach in machine-learning: complementing econometric insights. Quality & Quantity, 59(2), 1833-1860. https://doi.org/10.1007/s11135-024-02043-0

Jafarzadeh Esfehani, R., Rezaei Ardani, A., & Rezaei Kalat, A. (2025). Analyzing anxiety and treatment trends using google trends: validation of epidemiologic studies and forecasting. Quality & Quantity, 59(4), 3649-3660. https://doi.org/10.1007/s11135-025-02126-6

Jiang, J., Bandeli, K. K., & Srinivasan, K. (2025). Dynamic model selection in enterprise forecasting systems using sequence modeling. Decision Support Systems, 193, 114439. https://doi.org/10.1016/j.dss.2025.114439

Jin, B., & Xu, X. (2025). Predictions of residential property price indices for China via machine learning models. Quality & Quantity, 59(S2), 1481-1513. https://doi.org/10.1007/s11135-025-02080-3

Kaplanski, G. (2025). The box office as a leading indicator of investor sentiment. Finance Research Letters, 85, 107990. https://doi.org/10.1016/j.frl.2025.107990

Király, P., & Tóth, G. (2025). Being Aware of Data Leakage and Cross-Validation Scaling in Chemometric Model Validation. Journal of Chemometrics, 39(4), e70026. https://doi.org/10.1002/cem.70026

Koçak, B. B. (2026). Forecasting box office revenue using AI algorithms and consumer search behavior. Journal of Revenue and Pricing Management, 25(4), 352-367. https://doi.org/10.1057/s41272-025-00568-z

Kraemer, T., Weiger, W. H., & Heidenreich, S. (2025). Do all stars shine the same? Investigating the nonlinear effects of user and critic reviews on video game sales. Journal of Business Research, 188, 115034. https://doi.org/10.1016/j.jbusres.2024.115034

Kuti-Kreszács, M., Dioşan, L., & Bodó, Z. (2026). Towards Explainable Decision Support in Demand Forecasting: Global Feature Importance for Multi-Series Tree-Based Models. Acta Universitatis Sapientiae, Informatica, 18(1), 13. https://doi.org/10.1007/s44427-026-00022-x

Leukel, J., Liu, Z., & Sugumaran, V. (2026). Potential overinterpretation of results in the abstracts of machine learning studies for movie box office revenue prediction: a systematic review. Journal of Revenue and Pricing Management, 25(4), 368-377. https://doi.org/10.1057/s41272-026-00574-9

Li, J., Zeng, H., Xiao, C., Ouyang, C., & Liu, H. (2024). Listwise learning to rank method combining approximate NDCG ranking indicator with Conditional Generative Adversarial Networks. Pattern Recognition Letters, 179, 31-37. https://doi.org/10.1016/j.patrec.2024.01.015

Lovaglio, P. G. (2025). Cross-Learning With Panel Data Modeling for Stacking and Forecast Time Series Employment in Europe. Journal of Forecasting, 44(2), 753-780. https://doi.org/10.1002/for.3224

Magrini, A., & Rimediotti, L. (2025). Assessing the impact of Covid-19 on the Italian comic book industry: a panel vector autoregressive analysis of financial ratios. Quality & Quantity, 59(4), 3545-3573. https://doi.org/10.1007/s11135-025-02122-w

Mahmoudi, Y. (2025). Algorithmic trading and price forecasting: machine learning and deep learning strategies for financial markets. Quality & Quantity, 59(5), 4757-4778. https://doi.org/10.1007/s11135-025-02179-7

Makhortykh, M., de León, E., Christner, C., Sydorova, M., Urman, A., Adam, S., Maier, M., & Gil-Lopez, T. (2025). Is a single model enough? The systematic comparison of computational approaches for detecting populist radical right content. Quality & Quantity, 59(S2), 1163-1207. https://doi.org/10.1007/s11135-024-02034-1

Ribeiro, R., & Fanzeres, B. (2026). Integrated estimate-and-optimize decision trees learning for two-stage linear decision-making problems. European Journal of Operational Research, 329(2), 607-628. https://doi.org/10.1016/j.ejor.2025.08.048

Roberts, F., Roberts, H., & Tsoukiàs, A. (2025). Meaningful, useful and legitimate information in the use of index numbers for decision making. Quality & Quantity, 59(4), 3211-3244. https://doi.org/10.1007/s11135-025-02113-x

Sayaheen, B. (2024). Crossing borders: localization and paratext in Arabic mobile video games. Cogent Arts & Humanities, 11(1), 2307645. https://doi.org/10.1080/23311983.2024.2307645

Schmitz, F. R., & Parente, R. (2025). Born digitals entry mode choice from a target market institutional and digitalization perspective. Review of International Business and Strategy, 35(1), 64-77. https://doi.org/10.1108/ribs-05-2024-0048

Stoian, M. C. (2024). A knowledge-based perspective on SME foreign market entry mode choices and changes. International Journal of Entrepreneurial Behavior & Research, 30(10), 2488-2508. https://doi.org/10.1108/ijebr-06-2023-0641

Telang, R., & Sivan, L. (2025). Movie sentiment and home entertainment revenue. Journal of Cultural Economics, 49(2), 301-326. https://doi.org/10.1007/s10824-025-09533-5

Utami, P. W. (2025). From a contracted market to an expanding horizon: an analysis of Indonesia’s film production chain and accessibility. International Journal of Cultural Policy, 31(3), 284-299. https://doi.org/10.1080/10286632.2024.2342283

Vamsikrishna, A., & Gijo, E. V. (2024). New Techniques to Perform Cross-Validation for Time Series Models. Operations Research Forum, 5(2), 51. https://doi.org/10.1007/s43069-024-00334-8

Ventre, V., Gimigliano, F., & Martino, R. (2025). Decision theory for sustainable rehabilitation: managing uncertainty and group choice to improve health conditions, well-being, and social inclusion. Quality & Quantity, . https://doi.org/10.1007/s11135-025-02335-z

Wang, X., Luo, S., & Gao, Y. (2024). Ladder Curriculum Learning for Domain Generalization in Cross-Domain Classification. IEEE Access, 12, 95356-95367. https://doi.org/10.1109/access.2024.3425602

Wang, G., Lin, Z., Zhong, X., Ji, H., Li, P., Chen, Y., & Yan, J. (2026). Decision-focused learning integrated with data-driven robust optimization for energy management. Applied Energy, 408, 127343. https://doi.org/10.1016/j.apenergy.2025.127343

World Bank. (2026). GDP per capita and population indicators. World Bank Open Data. https://data.worldbank.org/ (accessed 3 August 2026).

Yang, H., & Gonçalves, T. (2025). MultiLTR: Text Ranking with a Multi-Stage Learning-to-Rank Approach. Information, 16(4), 308. https://doi.org/10.3390/info16040308

Yeldho, N., Thomas, D., Kurian, V. G., Arathy, C., & Biju, A. V. N. (2025). Are machine learning models effective in predicting emerging markets? Investigating the accuracy of predictions in emerging stock market indices. Quality & Quantity, 59(1), 839-904. https://doi.org/10.1007/s11135-024-01964-0

Zhang, B., Panagiotelis, A., & Li, H. (2025). Constructing hierarchical time series through clustering: Is there an optimal way for forecasting?r forecasting. International Journal of Forecasting, 41(3), 1022-1036. https://doi.org/10.1016/j.ijforecast.2024.10.002

Zhou, Y., Gomes, E., & Vendrell-Herrero, F. (2025). Optimizing International Market Selection: A Dynamic Approach Integrating Export and FDI Entry Modes. Thunderbird International Business Review, 67(6), 739-755. https://doi.org/10.1002/tie.70007