Scaled Beta Models and Feature Dilution for Dynamic Ticket Pricing

Jonathan R. Landers

jonathan.robert.landers@gmail.com
GitHub Repository: github.com/jonland82/seatgeek-beta-modeling

Abstract

A novel approach is presented for identifying distinct signatures of performing acts in the secondary ticket resale market by analyzing dynamic pricing distributions. Using a newly curated, time series dataset from the SeatGeek API, we model ticket pricing distributions as scaled Beta distributions. This enables accurate parameter estimation from incomplete statistical data using a hybrid of quantile matching and the method of moments. Incorporating the estimated α and β parameters into Random Forest classifiers significantly improves pairwise artist classification accuracy, demonstrating the unique economic signatures in event pricing data. Additionally, we provide theoretical and empirical evidence that incorporating zero-variance (constant-value) features into Random Forest models acts as an implicit regularizer, enhancing feature variety and robustness. This regularization promotes deeper, more varied trees in the ensemble, improving the bias-variance tradeoff and mitigating overfitting to dominant features. These findings are validated on both the new ticket pricing dataset and the standard UCI ML handwritten digits dataset.

1. Introduction

In 2024, the secondary concert ticket resale market experienced significant growth, driven by factors such as marketplace convenience, ticket scarcity, better seating options, and advancements in mobile technologies. The global secondary tickets market, encompassing concert tickets, is projected to expand by USD 132.1 billion between 2024 and 2028, with a compound annual growth rate (CAGR) of 34.25% [1].

This study introduces a novel method to identify the unique economic signatures of performing artists using concert ticket resale data, focusing on dynamic pricing distributions. The research uses a newly curated time series dataset of ticket pricing details for entertainment events, collected via the SeatGeek API from May 2023 to May 2024 [2]. The dataset includes a wide range of events, from high-profile artists like Metallica and Taylor Swift to emerging local acts. SeatGeek's influence in the secondary ticket market has been bolstered by partnerships with major organizations, such as becoming the official secondary marketplace for Paciolan, a leader in ticketing solutions for college athletics and live entertainment venues [3].

We propose an innovative parameter estimation method that uses composite quantile and moment matching to reverse engineer and model ticket pricing distributions as scaled Beta distributions. This approach effectively handles incomplete statistical data, leveraging the flexibility of Beta distributions to capture varied pricing patterns. Random Forest models, using time series subsequence statistics and estimated distribution parameters, accurately classify performing artists. By supplementing limited subsequence distributional information with inferred parameters, we achieve a notable improvement in classifier accuracy, exhibiting the unique economic signatures present in event pricing data. Expanding on these results, we mathematically demonstrate that the observed improvement in Random Forest accuracy metrics indicates that the estimation algorithm yields a more precise and informative representation of the true underlying pricing distributions. Moreover, the ability to explicitly reverse engineer data distributions has numerous practical applications. These range from market forecasting and pricing optimization in the secondary ticket market to broader uses in financial modeling, demand forecasting, and supply chain analytics. In each case, understanding and reconstructing underlying distributional patterns is crucial for effective decision-making and risk management.

Another key contribution of this research is demonstrating how zero-variance (constant-value) features act as implicit regularizers in Random Forest ensembles. These features dilute the selection probability of dominant features during tree splits, promoting variety by increasing the likelihood of selecting less prominent but informative features, such as, potentially, the α_i and β_i parameters derived from pricing distributions. This effect enables trees to grow deeper and more varied, enhancing model robustness and accuracy by reducing the ensemble's reliance on a narrow subset of predictors. As explored later, this probabilistic rebalancing fosters greater tree variety, quantified through pairwise feature usage distances, and encourages the model to leverage a broader range of predictors, thereby improving its ability to capture nuanced patterns in the data. Mathematically, we derive selection probability approximations to show how constant features adjust feature rankings, alongside decomposing error terms to reveal variance reduction. Notably, the configuration yielding the highest accuracy on both the ticket pricing and UCI handwritten digits datasets would not be identified through standard grid searching of the parameter space; it is only through implicit regularization that this optimal setup emerges. This mechanism shifts the bias-variance tradeoff favorably, mitigating overfitting, especially in scenarios where a few strong features might otherwise overshadow subtler but valuable signals or where high correlation among tree predictions could degrade performance. Empirical results validate these findings using both the newly curated ticket pricing dataset from SeatGeek and the UCI ML handwritten digits dataset [4], demonstrating consistent improvements in generalization performance under specific conditions. This regularization approach complements the parameter estimation strategy, offering a practical enhancement to Random Forest models with applications extending beyond ticket pricing to domains requiring robust handling of complex, dynamic datasets.

Implicit regularization effects emerged during our investigation of parameter estimation, revealing a critical insight: engineered features with subtle predictive power may be systematically undervalued by conventional hyperparameter tuning approaches in Random Forests. This relationship is particularly consequential when ground truth distributions are unavailable, as our theoretical bounds demonstrate that improvements in classification accuracy correspond to quadratic improvements in distributional fidelity. Such enhancements are valuable when developing class-conditioned generative models of dynamic systems like ticket markets, where accurate reconstruction of underlying distributions directly impacts model quality. The event data from SeatGeek offers a rich resource for analyzing economic patterns in secondary ticket markets across varied artists and venues. By addressing both the challenge of parameter estimation with limited data and the implicit regularization mechanisms that affect feature utilization, we provide a unified approach to extracting and leveraging distributional information in classification tasks, while introducing a valuable dataset for research in dynamic pricing analysis.

The remainder of the paper is structured as follows: Section 2 reviews related work spanning distribution estimation under limited information, quantile- and moment-based inference, time series classification methods, and implicit regularization mechanisms in ensemble learning. Section 3 introduces the curated SeatGeek dataset, outlines the raw time series representation of event-level pricing data, and formalizes the transformations used to derive distributional and statistical feature sets for artist classification tasks. Section 4 presents the distribution-based modeling framework, using scaled Beta distributions to approximate pricing dynamics from limited moment statistics. The α_i and β_i parameters are estimated via composite quantile and moment matching, and their contribution to classification performance is analyzed through both kernel density estimation (KDE) comparisons across performing acts and Random Forest accuracy improvements in pairwise artist classification. Section 5 introduces the implicit regularization framework, demonstrating how zero-variance features modify the probabilistic structure of feature selection in Random Forests, enhancing tree depth, feature variety, and generalization by rebalancing dominance effects. Section 6 summarizes the combined contributions and discusses implications.

2. Related Work

Parameter estimation in the absence of full ground-truth distributions and time series classification have been widely studied across machine learning, statistics, and econometrics. Our work focuses on estimating the parameters of a scaled Beta distribution using composite quantile and moment matching, leveraging limited summary statistics (minimum, maximum, mean, median) to infer the underlying probability distribution. This approach connects to multiple research threads in quantile-based estimation, moment matching, statistical learning theory, time series classification, functional data analysis (FDA), and implicit regularization in machine learning. Classic methods for time series classification, such as Dynamic Time Warping (DTW) and 1-Nearest Neighbor (1-NN) [5, 6], provide foundational work relevant to our comparative analysis. A recent survey of time series classification methods [7] provides an extensive comparison of approaches, including ensemble-based methods like Time Series Forest (TSF) and Canonical Interval Forest (CIF), which are relevant to our work on feature-based classification. Ye and Keogh [8] introduced time series shapelets, which are discriminative subsequences that capture local patterns in time series data. Their approach computes a distance between each candidate shapelet and the time series to derive a classification rule. Discriminative subsequences play a part in this study whereby they are used for distribution recovery and ultimately act classification.

2.1. Distribution-Based Parameter Estimation

Classical references on Beta distributions provide essential foundations for parameter estimation. Krishnamoorthy [9] discusses traditional methods for fitting Beta parameters, typically assuming full sample data. Our approach extends these fundamentals by deriving estimators that work under limited summary statistics, namely min, max, mean, median. This enables our scaled Beta-based modeling to capture event-specific pricing dynamics with minimal data requirements.

Christ et al. [10] introduced TSFresh, a Python package that extracts a comprehensive set of 794 time series features spanning statistical, Fourier-based, and time-domain characteristics, with an automated hypothesis-testing-based feature selection process to retain only the most relevant features. Likewise, Lubba et al. [11] developed Catch22, a minimalistic feature extraction framework that selects 22 domain-agnostic time series features optimized for interpretability and efficiency while maintaining strong classification performance across various datasets. More recently, Zhang et al. [12] developed Quantile Flows for Generative Flow Networks (GFlowNets), illustrating how quantile-based estimation can replace point-estimate methods in generative modeling. This work reinforces the core principle that quantiles alone can capture meaningful distributional characteristics, paralleling our own quantile-driven approach to scaled Beta modeling for ticket prices.

Zhou et al. [13] proposed Beta Diffusion, showcasing the versatility of Beta distributions for range-bounded inference, such as our setting in event ticket pricing, by retaining both flexibility and interpretability in the learned representations. Dempster et al. [14] introduced QUANT, achieving state-of-the-art classification performance on time series benchmarks solely through quantiles extracted from dyadic intervals. Their findings mirror the strength of our sparse-data strategy, as we also focus on extracting pivotal distributional features rather than full empirical distributions. Similarly, Wei et al. [15] showed in Latent Quantile Matching (LQM) that even limited quantile information can preserve essential statistical properties for downstream tasks, reinforcing the viability of our quantile-oriented approach in data-scarce scenarios.

Beyond quantiles, our approach also builds on moment-matching techniques for probability distribution estimation. Salimans et al. [16] applied moment matching to diffusion models, demonstrating that aligning conditional expectations (first moments) along the sampling trajectory enables more efficient generative modeling without full-step sampling. While their work focuses on accelerating sampling, it underscores the broader utility of moment constraints in improving probabilistic inference. Similarly, Zhang et al. [17] integrated moment matching into denoising Gibbs sampling, enhancing sampling efficiency in energy-based models by leveraging moment constraints to guide the learning process. These works reinforce the principle that carefully chosen statistical moments can meaningfully constrain and improve inference in data-driven models. Drawing from this base, our framework combines quantile- and moment-based techniques to estimate distributional properties in ticket-pricing contexts, particularly when detailed observations are sparse. To our knowledge, no prior work has derived closed-form expressions for Beta distribution parameters using only the mean and median (scaled via min/max) in a low-data setting, nor applied this technique to real-time event classification tasks.

2.2. Statistical Learning Theory and Parameter Estimation Accuracy

Lin [18] introduced the Jensen–Shannon divergence as a symmetrical, bounded measure of distributional distance, demonstrating how classification accuracy can deteriorate significantly when estimated and true distributions diverge. Devroye et al. [19] then provided probabilistic bounds on classification risk, directly linking distribution-estimation error to predictive accuracy. Tsybakov [20] introduced margin conditions under which classification error rates converge optimally, establishing a deeper connection between parameter-estimation precision and classification performance.

Our findings follow these foundational insights: improved estimation of Beta parameters leads to more accurate classification of event types, while misestimation propagates into downstream classification error. By mapping the proposed parameter-estimation method to these theoretical frameworks, we demonstrate how precisely characterizing the underlying distribution supports robust predictive performance. Moreover, the observed classification accuracy itself provides indirect validation that the estimated distributions faithfully capture key aspects of the true underlying pricing dynamics. In this way, our scaled Beta approach echoes the broader principle in statistical learning that well-characterized data distributions are essential for achieving strong generalization, and conversely, strong generalization serves as empirical evidence of distributional fidelity.

2.3. Implicit Regularization and Entropy-Based Learning in Random Forests

Research on ensemble methods, particularly Random Forests, has long recognized implicit regularization as a key factor in robust generalization. Lin [18] initially established the Jensen–Shannon divergence as a theoretical tool for measuring distributional shifts, a perspective that later influenced ensemble learning paradigms. Breiman [21] introduced ”bagging” as a variance-reduction technique, demonstrating that bootstrapping weak learners stabilizes predictions and improves generalization. Ho [22] expanded on these ideas by pioneering the Random Subspace Method (RSM), training trees on randomly selected subsets of features to limit over-reliance on any single subset. Geurts et al. [23] later introduced ”Extremely Randomized Trees”, an approach that injects additional randomness into tree splits, enhancing variance reduction beyond standard Random Forests.

Breiman [24] formalized these viewpoints by decomposing Random Forest generalization error into “strength” (individual tree accuracy) and “correlation” (similarity among trees), illustrating how lower correlation improves overall performance. Bousquet and Elisseeff [25] further underscored the role of entropy-based stability in achieving robust generalization, offering insights closely tied to ensemble variety. Wyner et al. [26] later depicted Random Forests as “interpolating classifiers” that can perfectly fit training data yet still generalize effectively through ensemble “self-averaging.” This self-averaging perspective aligns with the findings of Wager and Athey [27], who demonstrated how causal forests provide implicit regularization by adaptively choosing neighborhoods, thus reducing estimation variance and improving robustness in causal inference settings. Mentch and Zhou [28] expanded on these insights by showing how random feature selection (`max_features` or m) reduces variance without explicit parameter constraints, functioning as an implicit regularizer, an idea with parallels in recent approaches to budget-constrained hyperparameter tuning [29].

Our work builds on these developments by highlighting an additional, often overlooked source of implicit regularization: constant-value (zero-variance) features. By altering the probability distribution over feature selection, these features act as an “entropy-based stabilizer,” preventing dominant predictors from monopolizing splits and encouraging more varied tree constructions. This dynamic alleviates over-concentration on a small subset of features, fosters deeper decision trees, and reduces correlation across the ensemble. Moreover, from our standpoint of scaled Beta modeling and limited summary statistics, such regularization can be crucial: it ensures that newly introduced distributional parameters (α and β) have a fair chance to inform each tree’s splits, rather than being overshadowed by more obvious predictors.

There are many related studies that employ regularization through penalty terms, starting with foundational algorithms such as ridge regression [30, 31]. Through the lens of KL-divergence-based smoothing, Random Forests, though not explicitly Bayesian, can exhibit entropy-driven regularization, a phenomenon paralleling the stability perspective outlined by Bousquet and Elisseeff [25]. This feature-probability redistribution resonates with function-smoothing techniques in functional data analysis (FDA); more recent studies continue to extend this line of inquiry, such as roughness penalization in free-knot spline estimation [32], where avoiding over-concentration preserves a balanced, informative representation of underlying structure. Moreover, the relationship between implicit regularization, feature selection probabilities, and classification accuracy established in our approach bears resemblance to recent developments by Dunbar et al. [33], who address hyperparameter optimization for randomized algorithms by formulating it as a stochastic inverse problem solved via Ensemble Kalman Inversion.

3. Preliminaries and Data

3.1. Raw Event Time Series Data

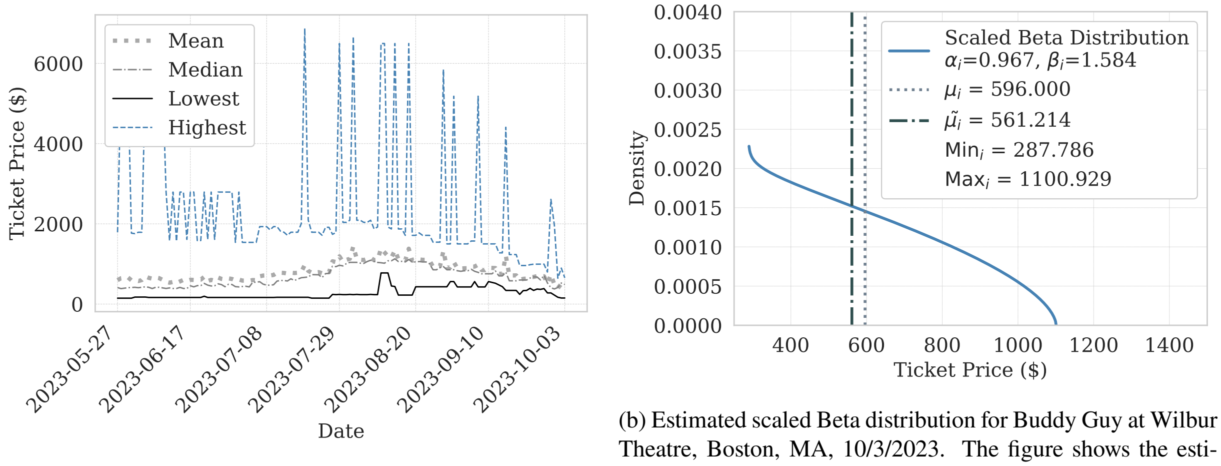
The data is comprised of daily snapshots of concert ticket prices from the SeatGeek API for approximately 130,000 events, 15,400 artists/acts, and 6,700 venues across the United States between May 2023 and May 2024. For each event, price information was recorded from the initial sale date (or first available date) through the event date, yielding a comprehensive view of the pricing lifecycle. We denote the raw time series data as

$$\mathcal{D}_{\text{raw}} = \{\mathbf{x}_t\}_{t=1}^{T},$$

where $\mathbf{x}_t = [x_t^{(1)}, x_t^{(2)}, \ldots, x_t^{(d)}]^\top$ is a vector of d observed variables at time t, and T is the total number of recorded time steps. Variables include artist, event date/time, venue, price collection date/time, mean price, median price, low price, high price, and listing count. This can equivalently be represented as a matrix:

$$\mathbf{X} \in \mathbb{R}^{T\times d},$$

with rows corresponding to time steps and columns to variables. Figure 1a illustrates this representation using ticket price data for blues guitarist Buddy Guy at the Wilbur Theatre in Boston on 10/3/2023.



(a) Ticket prices over time for Buddy Guy at Wilbur Theatre, Boston, MA, 10/3/2023, showing the Mean, Median, Lowest, and Highest prices.

SEAT GEEK

(b) Estimated scaled Beta distribution for Buddy Guy at Wilbur Theatre, Boston, MA, 10/3/2023. The figure shows the estimated α_i and β_i parameters, the mean price (μ_i), the median price ($\tilde{\mu}_i$), the lowest price (Min$_i$), and the highest price (Max$_i$). These quantities define the scaled Beta distribution at snapshot T', leading up to event i on 10/3/2023. They represent the economic signature and corresponding feature value set for this event in the Random Forest model.

Figure 1: Event Overview, Buddy Guy at Wilbur Theatre, Boston, MA, 10/3/2023

3.2. Derived Data Representations for Classification

To prepare the data for artist classification, we define a transformation f from the raw time series data into a structured feature space:

$$f : \mathbb{R}^{T\times d} \to \mathbb{R}^{E\times n}.$$

The resulting training dataset is given by:

$$\mathcal{D}_{\text{train}} = f(\mathcal{D}_{\text{raw}}) = \{(\mathbf{z}_i, y_i)\}_{i=1}^{E},$$

where $\mathbf{z}_i \in \mathbb{R}^n$ is the derived feature vector for the i-th event, and y_i is the corresponding artist label. This structured format is amenable to standard machine learning methods.

We frame artist classification as a series of pairwise binary classification tasks, where each model distinguishes between events for a given pair of artists (e.g., The Pixies vs. Billy Joel). These models are not merely for artist identification, they serve as tests of whether artist-specific pricing distributions are distinct enough to be reliably classified. High classification accuracy reflects a strong representation of underlying distributional dynamics, while lower accuracy suggests the need for further investigation.

Random Forest classifiers are well-suited to this task due to their ability to model complex feature interactions and their robustness to noise. Prior research in time series classification has shown the effectiveness of feature-based approaches, where summary statistics provide strong discriminative power [7]. Methods like TSFresh [10] and Catch22 [11] extract handcrafted statistical features, achieving high accuracy without explicitly modeling temporal dependencies. Similarly, ensemble-based approaches such as Canonical Interval Forest (CIF) enhance performance by incorporating interval-based statistics [7]. Our method builds on these insights by summarizing pricing time series into compact distributional representations, capturing artist-specific patterns in pricing dynamics.

Basic Statistical Features. Initially, we derive basic summary statistics for the subsequence T' leading up to each event:

$$x = \frac{1}{|T'|}\sum_{t\in T'} x_t, \quad x_t \in \{\mu_t, \tilde{\mu}_t, \text{Max}_t, \text{Min}_t\}, \tag{1}$$

where $\mu_t, \tilde{\mu}_t, \text{Max}_t, \text{Min}_t$ denote mean, median, maximum, and minimum prices respectively. This yields feature vectors:

$$\mathbf{z}_i = [\mu_i,\ \tilde{\mu}_i,\ \text{Max}_i,\ \text{Min}_i]^\top,$$

and dataset:

$$\mathcal{D}_{\text{basic}} = \{(\mathbf{z}_i, y_i)\}_{i=1}^{E}. \tag{2}$$

Distribution-Augmented Features. Basic statistics alone omit nuanced distributional shapes. To address this, we estimate scaled Beta distribution parameters α_i, β_i for each event-artist pair over $[\mathrm{Min}i, \mathrm{Max}_i]$ (Section 4). These parameters enrich the feature vector:

$$\mathbf{z}_i = [\mu_i,\ \tilde{\mu}_i,\ \mathrm{Max}_i,\ \mathrm{Min}_i,\ \alpha_i,\ \beta_i]^\top,$$

yielding dataset:

$$\mathcal{D}_{\alpha\beta} = \{(\mathbf{z}_i, y_i)\}_{i=1}^{E}. \tag{3}$$

3.3. Implicit Regularization via Constant-Value Features

To implicitly regularize our Random Forest models, we augment dataset $\mathcal{D}_{\alpha\beta}$ with n_{ZV} zero-variance (constant-value) features $\mathbf{c} \in \mathbb{R}^{n_{\mathrm{ZV}}}$:

$$\mathbf{z}_i = [\mathbf{z}_i,\ \mathbf{c}]^\top,$$

yielding:

$$\mathcal{D}_{\alpha\beta}^{(\mathrm{reg})} = \{(\mathbf{z}_i, y_i)\}_{i=1}^{E}. \tag{4}$$

Although counterintuitive, constant-value features subtly shift Random Forest feature-selection probabilities, implicitly promoting deeper, more robust trees and improved generalization, as explored in detail in Section 5.

Additional Validation (handwritten Digits). To verify that implicit regularization effects generalize beyond ticket pricing data, we replicate our approach using the standard UCI handwritten digits dataset [4]. Specifically, we form two analogous datasets: $\mathcal{D}_\delta$, containing the original digit features, and $\mathcal{D}_\delta^{(\mathrm{reg})}$, which includes additional constant-value features to mirror the ticket pricing methodology. This parallel validation confirms (Section 5) the consistency and generalizability of the observed regularization effects across distinct data domains.

4. Parameter Estimation in the Absence of Ground Truth from Limited Summary Statistics

Concert ticket price distributions are modeled for each event-artist pair using a scaled Beta distribution. The approach estimates key distribution parameters, α_i and β_i, by way of composite quantile and moment matching, from the limited statistical information provided by the SeatGeek API: the minimum, maximum, mean, and median ticket prices. The scaled Beta distribution is chosen for its flexibility in capturing varying distribution shapes, providing a more nuanced representation of event pricing dynamics at a given snapshot in time. In contrast to methods like those of Wei et al. [15] that match multiple quantiles in a latent space, or Zhang et al. [12] that parameterize entire flows via quantiles, our approach targets only one quantile (the median) and the mean to fit α and β. This limited yet effective summary-based strategy leverages minimal data. Dempster et al. [14] also highlight the power of quantile-based distribution features, but they rely on richer raw data coverage than what we have from SeatGeek. Meanwhile, classical references such as Krishnamoorthy [9] provide the foundational Beta moments and parameter-estimation formulas that underlie scaled Beta modeling.

4.1. Scaled Beta Distribution

Let the minimum, maximum, mean, and median ticket prices for each event-artist pair be denoted as Min_i, Max_i, μ_i, and $\tilde{\mu}_i$, respectively. The probability density function (PDF) of the scaled Beta distribution is given by:

$$f(x; \alpha_i, \beta_i, \mathrm{Min}_i, \mathrm{Max}_i) =$$

$$\frac{(x - \mathrm{Min}_i)^{\alpha_i - 1}\,(\mathrm{Max}_i - x)^{\beta_i - 1}}{(\mathrm{Max}_i - \mathrm{Min}_i)^{\alpha_i + \beta_i - 1}\,B(\alpha_i, \beta_i)}, \tag{5}$$

where x is the ticket price, and $B(\alpha_i, \beta_i)$ is the Beta function. This formulation transforms the standard Beta distribution from $[0, 1]$ to $[\mathrm{Min}_i, \mathrm{Max}_i]$. Such a scaled Beta framework is also seen in other contexts, like Zhou et al. [13], who exploit Beta distributions for bounded data in generative modeling, underscoring the flexibility of Beta-based parameterizations. Classical discussions in Krishnamoorthy [9] elaborate on these Beta formulations and offer general moment-based inference approaches that set the stage for our scaled version.

4.2. Parameter Estimation Using Median and Mean: Composite Quantile and Moment Matching

To estimate the parameters α_i and β_i, we reparameterize the scaled Beta distribution using the mean and median provided by the SeatGeek API. The mean μ_i and median $\tilde{\mu}_i$ for the Beta distribution on $[\mathrm{Min}_i, \mathrm{Max}_i]$ are given by:

$$\mu_i = \mathrm{Min}_i + \frac{\alpha_i}{\alpha_i + \beta_i}\,(\mathrm{Max}_i - \mathrm{Min}_i), \tag{6}$$

$$\tilde{\mu}_i \approx \mathrm{Min}_i + (\mathrm{Max}_i - \mathrm{Min}_i)\left(\frac{\alpha_i - \frac{1}{3}}{\alpha_i + \beta_i - \frac{2}{3}}\right). \tag{7}$$

We first scale the mean and median to $[0, 1]$:

$$s = \frac{\mu_i - \mathrm{Min}_i}{\mathrm{Max}_i - \mathrm{Min}_i}, \quad q = \frac{\tilde{\mu}_i - \mathrm{Min}_i}{\mathrm{Max}_i - \mathrm{Min}_i}. \tag{8}$$

From the mean equation, we express β_i in terms of α_i and s:

$$\beta_i = \alpha_i \Big(\frac{1-s}{s}\Big). \tag{9}$$

Substituting this into the median equation gives:

$$q = \frac{\alpha_i - \frac{1}{3}}{\frac{\alpha_i}{s} - \frac{2}{3}}, \tag{10}$$

which simplifies to:

$$\alpha_i = \frac{s\,(2q-1)}{3\,(q-s)}, \quad \beta_i = \frac{(1-s)\,(2q-1)}{3\,(q-s)}. \tag{11}$$

This method leverages limited statistical features to estimate the underlying price distribution, capturing both central tendencies and distributional shape, improving the predictive performance of machine learning models. Salimans et al. [16] show that matching specific moments can effectively distill or preserve generative behavior, supporting our stance that one well-chosen quantile (the median) plus the mean can significantly influence an inferred distribution. While Dempster et al. [14] exploit a richer set of raw-data quantiles, our composite-quantile-and-moment-matching approach operates with minimal summary data from the SeatGeek API. Canonical Beta-distribution identities from Krishnamoorthy's handbook [9] further validate the soundness of this parametric inference, reinforcing our tactic of fitting α and β with so few statistics.

Furthermore, Wei et al. [15] demonstrate that matching multiple quantiles in a latent space can further refine distribution alignment. Our approach is simpler: fitting scaled Beta parameters in the *observable* ticket-price space from only the mean and median, but it likewise underscores that a modest set of carefully selected statistics can unlock deeper distributional insights. Lubba et al. [11] highlight how even a small feature subset maintains strong classification performance, which supports our reliance on $\{\mu_i,\ \tilde{\mu}_i,\ \alpha_i,\ \beta_i\}$ plus Min_i and Max_i. An example of an estimated price distribution for a specific artist-event snapshot is shown in Figure 1b.

4.3. Kernel Density Estimation for Distributional Features

Given the derivations for α_i and β_i alongside the original statistical features, it is useful to compare these components across events for specific acts to determine where α_i and β_i contribute additional predictive power. Larger distances between the feature distributions of two acts indicate greater separability. For example, consider two acts $\{1, 2\}$ in a pairwise setting.

Formally, for a given act (artist), we define:

$$\mathbf{z}_i = \big[\mu_i,\ \tilde{\mu}_i,\ \text{Max}_i,\ \text{Min}_i,\ \alpha_i,\ \beta_i\big]^\top$$

for each event i. Let $x \in \{\mu, \tilde{\mu}, \text{Max}, \text{Min}, \alpha, \beta\}$. The kernel density estimate (KDE) for each feature is

$$\hat{f}_x(x) = \frac{1}{E\,h}\sum_{i=1}^{E} K\Big(\frac{x - x_i}{h}\Big), \tag{12}$$

where K is the kernel function, E is the number of data points (events), and h is the bandwidth.

Using the KDE for each feature and act, $\{\hat{f}_\mu^{act}, \hat{f}_{\tilde{\mu}}^{act}, \hat{f}_{\text{Max}}^{act}, \hat{f}_{\text{Min}}^{act}, \hat{f}_\alpha^{act}, \hat{f}_\beta^{act}\}$, we assess distributional similarity using the Hellinger Distance $H(\hat{f}_x^1, \hat{f}_x^2)$ and the Jensen-Shannon (JS) Divergence $JS(\hat{f}_x^1 \parallel \hat{f}_x^2)$:

1. Hellinger Distance:

$$H(\hat{f}_x^1, \hat{f}_x^2) =$$

$$\frac{1}{\sqrt{2}}\sqrt{\int \big(\sqrt{\hat{f}_x^1(t)} - \sqrt{\hat{f}_x^2(t)}\big)^2 dt}. \tag{13}$$

2. Jensen-Shannon Divergence:

$$JS(\hat{f}_x^1 \parallel \hat{f}_x^2) =$$

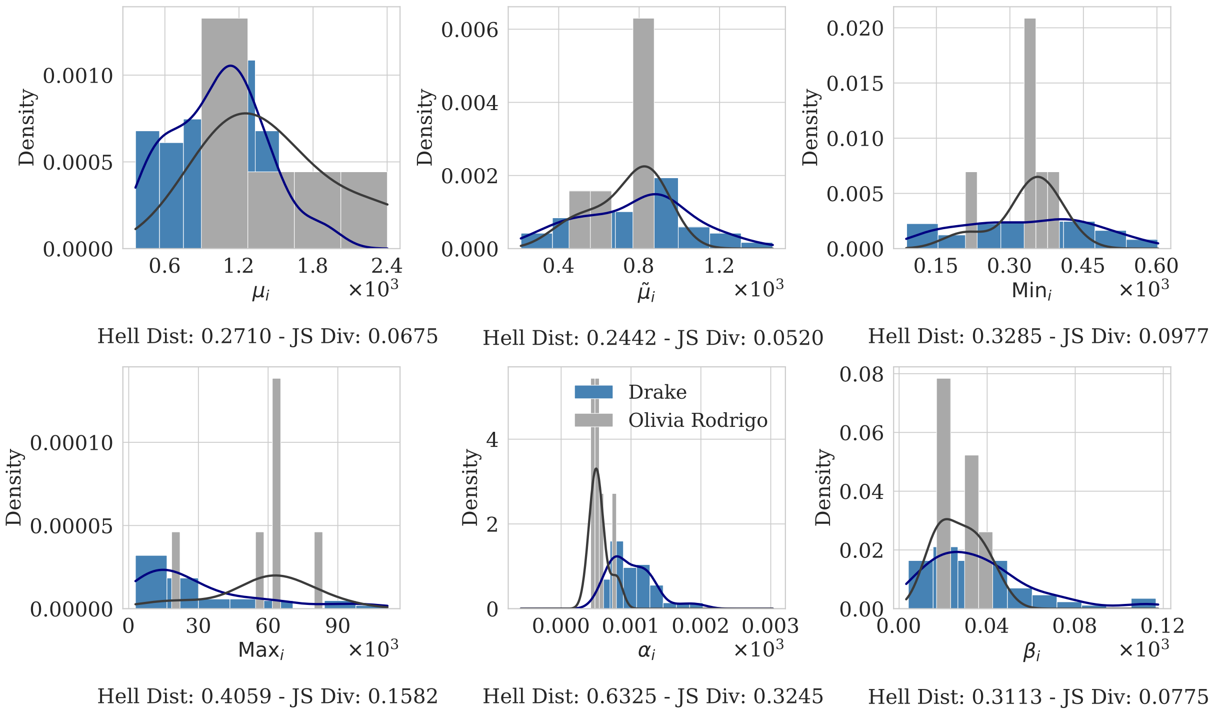
$$\tfrac{1}{2}\, D_{KL}(\hat{f}_x^1 \parallel M) + \tfrac{1}{2}\, D_{KL}(\hat{f}_x^2 \parallel M), \tag{14}$$

where $M = \frac{1}{2}(\hat{f}_x^1 + \hat{f}_x^2)$ and

$$D_{KL}(\hat{f}_x^1 \parallel \hat{f}_x^2) = \int \hat{f}_x^1(t)\ \log\Big(\frac{\hat{f}_x^1(t)}{\hat{f}_x^2(t)}\Big)\, dt. \tag{15}$$

These metrics evaluate each feature's ability to differentiate between acts, with larger distances indicating better separability and more distinct economic signatures. The estimated parameters, α_i and β_i can enhance predictive power, often improving Random Forest accuracy. For example, Drake and Olivia Rodrigo are well-known pop artists. In Figure 2, the KDEs for each feature and performer show that the calculated α_i exhibits a more distinct distribution than all of the original features, as reflected in both Hellinger and JS distances.

While Dempster et al. [14] leverage a rich set of quantile features extracted from raw data, our approach compensates for the lack of complete observations by using summary statistics to compute α_i and β_i. Krishnamoorthy's discussion [9] of Beta parameters similarly emphasizes the role of shape parameters in capturing subtle distributional differences, an insight that translates well here,



Hell Dist: 0.2710 - JS Div: 0.0675

Hell Dist: 0.2442 - JS Div: 0.0520

Hell Dist: 0.3285 - JS Div: 0.0977

Hell Dist: 0.4059 - JS Div: 0.1582

Hell Dist: 0.6325 - JS Div: 0.3245

Hell Dist: 0.3113 - JS Div: 0.0775

Figure 2: The plots show the distributions of each feature across all events for artists Drake and Olivia Rodrigo. The Hellinger Distance and Jensen-Shannon divergence are calculated between each distribution. In this particular comparison of artists, the α_i parameter offers the most distinctive density profile across all events, as indicated by the distribution distance metrics.

where these shape and skew measures both contribute to and are validated by classification.

Note that our use of the Jensen-Shannon distance in this section serves a different purpose than in Section 4.4, where it is employed to derive formal bounds on distributional convergence. Here, it is used empirically to compare density profiles across artists.

4.4. Validating Parameter Estimation via Classification Accuracy

To further justify classification as a validation technique for parameter estimation, we establish a theoretical link between classification accuracy and parameter estimation accuracy. Specifically, if the estimated parameters $\hat{\theta}$ closely approximate the true distribution parameters θ, then classification performance should improve. Conversely, if classification performance is high, this serves as empirical evidence that the estimated parameters faithfully capture the underlying distribution. This connection builds upon Tsybakov's margin assumption [20] and Devroye et al.'s probabilistic bounds on classification risk [19], both of which highlight the role of estimation precision in classification performance. We extend these foundational results by precisely characterizing the relationship between classifier accuracy and distributional similarity through both total variation distance and Jensen-Shannon divergence [18], demonstrating that improvements in classification accuracy yield increasingly substantial improvements in distributional fidelity, particularly with quadratic convergence properties in the information-theoretic context.

Proposition 1 (Parameter Estimation Consistency via Classification Accuracy)**.** *Let $\Theta \subset \mathbb{R}^d$ be the space of parameters, where each probability distribution P is parameterized by $\theta \in \Theta$. Define a feature map*

$$\phi(\theta) = \Big(f_1(P), f_2(P), \dots, f_{k-d}(P), \theta\Big), \tag{16}$$

where $f_i(P)$ represents summary statistics of P, such as minimum, maximum, median, and mean. A classifier $f : \mathbb{R}^k \to \{0, 1\}$ is trained to distinguish between two classes based on $\phi(\hat{\theta})$, where $\hat{\theta}$ is an estimated parameter obtained from observed data.

If f achieves a classification error rate ε, then there exists a function $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$ such that the estimation error satisfies

$$\|\hat{\theta} - \theta\| \;\leq\; \delta(\varepsilon). \tag{17}$$

Proof. **Propagation of Estimation Error to Feature**

Space. Define the "true" feature vector by $X^* = \phi(\theta)$ and the observed feature vector by $X = \phi(\hat{\theta})$. By the Lipschitz condition,

$$\| X - X^*\| = \|\phi(\hat{\theta}) - \phi(\theta)\| \ \leq \ L\,\|\hat{\theta} - \theta\|. \tag{18}$$

Relating Feature Perturbation to Classification Error. Under the margin separation assumption, the ideal feature vectors for two classes are separated by at least Δ. Suppose that a perturbation of size γ in feature space is tolerable without altering class assignment. Then,

$$R(f) \ \geq \ \mathbb{P}\big(\| X - X^*\| \ \geq \ \tfrac{\Delta}{2}\big). \tag{19}$$

Bounding the Probability of Large Feature Perturbation. Using Markov's inequality,

$$\mathbb{P}\big(\| X - X^*\| \ \geq \ \tfrac{\Delta}{2}\big) \ \leq \ \frac{2}{\Delta}\,\mathbb{E}\big[\| X - X^*\|\big]. \tag{20}$$

Combining with the Lipschitz bound,

$$\mathbb{P}(E) \ \leq \ \tfrac{2}{\Delta}\,L\,\|\hat{\theta} - \theta\|. \tag{21}$$

Then using the risk bound $R(f) \leq \varepsilon$, we obtain:

$$\varepsilon \ \geq \ \tfrac{2L}{\Delta}\,\|\hat{\theta} - \theta\|. \tag{22}$$

Rearranging,

$$\|\hat{\theta} - \theta\| \ \leq \ \tfrac{\Delta}{2L}\,\varepsilon. \tag{23}$$

Setting $\delta(\varepsilon) = \frac{\Delta}{2L}\,\varepsilon$, we see $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$, proving the proposition. □

Building on this foundation, we can more precisely characterize the relationship between classification accuracy and distributional similarity. The following theorems extend our theoretical analysis to establish rigorous bounds between classifier error and common measures of distributional difference.

Theorem 2 (Classification Accuracy and Total Variation Distance). *Let $P_{\hat{\theta}}(x)$ and $P_\theta(x)$ denote probability distributions parameterized by estimated parameters $\hat{\theta}$ and true parameters θ, respectively. Let ε be the misclassification error probability of a classifier built upon the distribution $P_{\hat{\theta}}(x)$. Then the total variation distance between the distributions is bounded by:*

$$TV(P_{\hat{\theta}}, P_\theta) = \frac{1}{2}\int |P_{\hat{\theta}}(x) - P_\theta(x)|\,dx \leq \eta(\varepsilon),$$

$$\textit{where} \quad \eta(\varepsilon) \to 0 \quad \textit{as} \quad \varepsilon \to 0. \tag{24}$$

Proof. **Misclassification and distributional differences.** Consider a binary classifier with decision regions $C_{\hat{\theta}}$ and C_θ corresponding to estimated and true parameters, respectively. The misclassification probability ε is given by:

$$\varepsilon = \frac{1}{2}\int \left[P_{\hat{\theta}}(x)I(x \in C_\theta) + P_\theta(x)I(x \in C_{\hat{\theta}})\right]\,dx, \tag{25}$$

where $I(\cdot)$ is the indicator function.

We compare ε to the Bayes-optimal classification error ε^*, given explicitly by the total variation distance [19]:

$$\varepsilon^* = \frac{1}{2}\left[1 - TV(P_{\hat{\theta}}, P_\theta)\right]$$

$$= \frac{1}{2}\left[1 - \frac{1}{2}\int |P_{\hat{\theta}}(x) - P_\theta(x)|\,dx\right]. \tag{26}$$

Since the achieved error ε must exceed the Bayes-optimal error ε^*, we have:

$$\varepsilon \geq \varepsilon^* = \frac{1}{2}\left[1 - TV(P_{\hat{\theta}}, P_\theta)\right]. \tag{27}$$

Rearranging terms explicitly isolates the total variation distance:

$$TV(P_{\hat{\theta}}, P_\theta) \leq 1 - 2\varepsilon. \tag{28}$$

This provides a fundamental lower bound linking classification error and distributional differences. However, we also seek a meaningful upper bound.

Upper bound via parameter continuity. From the previous proposition, we have a direct parameter-based bound:

$$\|\hat{\theta} - \theta\| \leq \delta(\varepsilon), \quad \text{with} \quad \delta(\varepsilon) \to 0 \quad \text{as} \quad \varepsilon \to 0.$$

Assume distributions $P_\theta(x)$ belong to a family that is Lipschitz-continuous in parameters, meaning there exists a constant $L > 0$ such that:

$$TV(P_{\hat{\theta}}, P_\theta) \leq \frac{L}{2}\|\hat{\theta} - \theta\|. \tag{29}$$

(This condition typically holds for parametric distributions like scaled Beta distributions considered in this work, where densities vary smoothly with respect to parameters.)

Substituting the result from the Proposition, we get a tighter, upper-bound formulation:

$$TV(P_{\hat{\theta}}, P_\theta) \leq \frac{L}{2}\delta(\varepsilon). \tag{30}$$

Define $\eta(\varepsilon) = \frac{L}{2}\delta(\varepsilon)$, which clearly approaches zero as $\varepsilon \to 0$.

Thus, we have established a rigorous upper bound directly relating classifier error to total variation distance:

$$TV(P_{\hat{\theta}}, P_\theta) \leq \eta(\varepsilon), \quad \eta(\varepsilon) \to 0 \quad \text{as} \quad \varepsilon \to 0.$$

□

While the Total Variation distance provides a natural measure of distributional difference, information-theoretic measures can offer additional insights with stronger convergence properties. The following theorem establishes an even more precise relationship using the Jensen-Shannon divergence.

Theorem 3 (Classification Accuracy and Jensen-Shannon Divergence)**.** *Under the same conditions as the previous theorem, the Jensen-Shannon divergence between the distributions can be bounded more tightly by:*

$$D_{JS}(P_{\hat{\theta}}||P_\theta) \leq \xi(\varepsilon),$$

$$\textit{where} \quad \xi(\varepsilon) \to 0 \quad \textit{as} \quad \varepsilon \to 0. \tag{31}$$

Furthermore, this bound exhibits a quadratic convergence rate as ε approaches zero.

Proof. **Relationship between JS divergence and Total Variation distance.** The Jensen-Shannon divergence between distributions $P_{\hat{\theta}}$ and P_θ is defined as:

$$D_{JS}(P_{\hat{\theta}}||P_\theta) = \frac{1}{2}D_{KL}(P_{\hat{\theta}}||M) + \frac{1}{2}D_{KL}(P_\theta||M) \tag{32}$$

where $M = \frac{1}{2}(P_{\hat{\theta}} + P_\theta)$ is the mixture distribution, and D_{KL} is the Kullback-Leibler divergence. We can leverage Lin's inequality [18], which states that:

$$D_{JS}(P_{\hat{\theta}}||P_\theta) \leq \frac{1}{2\ln(2)}TV^2(P_{\hat{\theta}}, P_\theta) \tag{33}$$

where $TV(P_{\hat{\theta}}, P_\theta)$ is the total variation distance as defined in the previous theorem.

Applying Total Variation bound from the previous theorem. From the previous theorem, we have established that:

$$TV(P_{\hat{\theta}}, P_\theta) \leq \eta(\varepsilon) = \frac{L}{2}\delta(\varepsilon)$$

Substituting this into Lin's inequality:

$$D_{JS}(P_{\hat{\theta}}||P_\theta) \leq \frac{1}{2\ln(2)}TV^2(P_{\hat{\theta}}, P_\theta)$$

$$\leq \frac{1}{2\ln(2)}\left(\frac{L}{2}\delta(\varepsilon)\right)^2 = \frac{L^2}{8\ln(2)}\delta^2(\varepsilon) \tag{34}$$

Establishing the quadratic convergence bound. Let $\xi(\varepsilon) = \frac{L^2}{8\ln(2)}\delta^2(\varepsilon)$. Since $\delta(\varepsilon) \to 0$ as $\varepsilon \to 0$, we also have $\xi(\varepsilon) \to 0$ as $\varepsilon \to 0$. However, the key insight is that $\xi(\varepsilon)$ approaches zero at a quadratic rate relative to $\delta(\varepsilon)$, which itself approaches zero as ε decreases.

Therefore:

$$D_{JS}(P_{\hat{\theta}}||P_\theta) \leq \xi(\varepsilon),$$

$$\text{where} \quad \xi(\varepsilon) \to 0 \quad \text{as} \quad \varepsilon \to 0$$

□

The progression from total variation distance to Jensen-Shannon divergence reveals a more nuanced relationship with quadratic convergence properties. This result has significant implications for our ticket pricing application. As classification accuracy improves (i.e., as ε decreases), the estimated Beta distribution parameters converge to the true parameters at an accelerating rate, not just linearly.

The Jensen-Shannon divergence provides several advantages over the total variation formulation: (1) tighter convergence rate through the quadratic relationship, (2) natural information-theoretic interpretation related to the distinguishability of distributions, (3) bounded measure between 0 and 1, and (4) symmetric property unlike KL divergence.

In our ticket pricing context, these results show that modest improvements in classifier accuracy yield increasingly substantial improvements in how well our estimated Beta distribution parameters reflect the true underlying pricing distributions. This accelerating improvement suggests that our parameter estimation approach using limited statistics can achieve high fidelity with relatively modest classification performance. Furthermore, these theorems strengthen our theoretical justification for using the estimated parameters α_i and β_i in downstream tasks.

By clearly relating classification errors to both traditional distance measures (total variation) and information-theoretic measures (Jensen-Shannon divergence), these results unify practical machine-learning classification performance with rigorous statistical inference, establishing a bridge between theoretical statistical learning theory and our practical modeling task.

Application to Ticket Pricing and Artist Classification. In our case, the probability distribution P represents a scaled Beta distribution modeling ticket prices, with parameters $\theta = (\alpha_i, \beta_i)$. The summary statistics $f_i(P)$ correspond to the minimum, maximum, mean, and median ticket prices, and the classifier f aims to distinguish between concerts of different artists based on pricing information.

According to the theorem, the accuracy of artist classification directly validates the accuracy of the estimated parameters $\hat{\alpha}_i$ and $\hat{\beta}_i$. When the classifier performs well, it implies that the estimated Beta distribution closely approximates the true underlying ticket-price distribution. Empirical results in the following section confirm this relationship, demonstrating that including α_i and β_i in feature representations improves classification performance. This is consistent with Tsybakov [20] and Devroye et al. [19], who emphasize that accurate estimation leads to better predictive outcomes. Notably, Krishnamoorthy [9] highlights how Beta inference precision often hinges on matching a few carefully chosen statistics, a principle that resonates here: tight approximation of α_i, β_i yields tangible boosts in classification accuracy.

4.5. Random Forest Results

Random Forests are ensemble learning methods that build multiple decision trees during training and combine their predictions to enhance generalization performance [22, 24]. Specifically, the prediction for an input x is

$$\hat{y} = \frac{1}{B}\sum_{b=1}^{B} h_b(x), \tag{35}$$

where $h_b(x)$ is the prediction of the b-th decision tree and B is the total number of trees (estimators). Each tree is trained on a bootstrap sample of the data, and a random subset of features is selected at each split. This approach reduces variance and helps prevent overfitting compared to individual decision trees. The standard scikit-learn [34] implementation is leveraged for this research.

In our classification task, we aim to identify the artist performing at an event based on its pricing information. Specifically, each Random Forest model is trained to distinguish between two artists. By framing the problem as dyadic classification, we avoid significant class imbalances that can occur when trying to classify one artist against all others.

Empirical Classification Improvements with α and β Estimates. To assess the effect of including α_i and β_i estimates in the feature set, we performed a pairwise comparison of Random Forest classifiers trained on two different feature sets. The first feature set, $\mathcal{D}_{\text{basic}}$, includes the mean, median, minimum, and maximum ticket prices as described in Section 3. The second feature set, $\mathcal{D}_{\alpha\beta}$, builds upon the basic set by adding the estimated α_i and β_i parameters derived from the scaled Beta distribution, as detailed in Section 4.

The classification task involves distinguishing between two artists, where the target y_i for each example corresponds to the performing act or artist. We evaluate performance improvements by comparing models trained with $\mathcal{D}_{\text{basic}}$ versus $\mathcal{D}_{\alpha\beta}$. We measure overall accuracy and also count how many models show improved performance when using $\mathcal{D}_{\alpha\beta}$.

A dataset of $N_{\text{pair}} = 20{,}000$ paired observations was used, each pair representing the same artist classification problem solved using both $\mathcal{D}_{\text{basic}}$ and $\mathcal{D}_{\alpha\beta}$. Among these pairs, there are $N_{\text{artist}} = 954$ unique artists. On average, each pair includes $N_{\text{event}}^{\text{train}} \approx 37$ training events and $N_{\text{event}}^{\text{test}} \approx 10$ testing events, resulting from an 80/20 train/test split. In total, this setup evaluates $N_{\text{models}} = 20{,}000$ Random Forest models. The selected dataset is statistically representative; experiments on additional subsets confirmed similar results. The Random Forest hyperparameters are shown in Table 1, with other hyperparameter choices yielding comparable conclusions.

The outcomes for each pair fall into three categories: $\mathcal{D}_{\alpha\beta}$ outperformed $\mathcal{D}_{\text{basic}}$, both performed equally, or $\mathcal{D}_{\text{basic}}$ outperformed $\mathcal{D}_{\alpha\beta}$. Among the $N_{\text{pair}} = 20{,}000$ total pairs, we observed 12,739 ties, 4,488 cases where $\mathcal{D}_{\alpha\beta}$ performed better, and 2,773 cases where $\mathcal{D}_{\text{basic}}$ performed better. Excluding tied cases, the effective sample size for statistical testing is

$$N' = 20{,}000 - 12{,}739 = 7{,}261.$$

Under the null hypothesis H_0, the $\mathcal{D}_{\alpha\beta}$ feature set is equally likely to improve or degrade classification performance compared to $\mathcal{D}_{\text{basic}}$. We therefore assume a probability $p = 0.5$. We test this by comparing the observed "better" outcomes ($n_{\text{better}} = 4{,}488$) to a binomial distribution with $p = 0.5$.

Using a normal approximation to the binomial, we have

$$\mu = N' \cdot 0.5 = 7{,}261 \times 0.5 = 3{,}630.5,$$

$$\sigma = \sqrt{N' \times 0.5 \times 0.5} \approx 42.61.$$

Applying a continuity correction, the Z-score becomes

$$Z = \frac{n_{\text{better}} - 0.5 - \mu}{\sigma} =$$

$$\frac{4{,}488 - 0.5 - 3{,}630.5}{42.61} \approx 20.13.$$

A Z-score of about 20.13 corresponds to an extremely small p-value ($p < 10^{-89}$). Table 2 summarizes these calculations.

We observe a statistically significant improvement in classification performance when using $\mathcal{D}_{\alpha\beta}$ (Figure 3a).

Table 1: Random Forest hyperparameters for $\mathcal{D}_{\text{basic}}$-$\mathcal{D}_{\alpha\beta}$ comparisons.

Hyperparameter	Value
`random_state`	42
`class_weight`	balanced
`max_features`	2
`max_depth`	100
`n_estimators`	100

Table 2: Summary of statistical results comparing $\mathcal{D}_{\alpha\beta}$ to $\mathcal{D}_{\text{basic}}$.

Statistic	Value
Effective sample size (N')	7,261
n_{better}	4,488
n_{worse}	2,773
Mean ($\mu = N'/2$)	3,630.5
Std. dev. (σ)	42.61
Z-score	20.13
p-value	$< 10^{-89}$

While the average accuracy improvement across all 20,000 models is modest, as illustrated in Figure 3b, the large volume of models that benefit from α_i, β_i features is noteworthy. The small but consistent accuracy gain, coupled with the extremely low p-value, suggests that the improvement is nontrivial. Incorporating α_i and β_i parameters improves the Random Forest classifier's ability to distinguish artists, highlighting the value of distributional features for capturing unique economic signatures in dynamic pricing data.

This strategy shows that, while Dempster et al. [14] fully exploit raw data for quantile features, our mathematical treatment of limited summary statistics offers a novel pathway for enhanced distributional characterization. Such moment-based insights also resonate with Zhang et al. [12] and Salimans et al. [16], wherein matching or estimating key distribution properties can significantly boost downstream tasks. Krishnamoorthy's foundational work [9] on Beta-based inference lends additional theoretical justification, illustrating that limited but carefully chosen statistics can often yield sufficiently accurate parameter estimates to drive improved classification performance.

Beyoncé or Ed Sheeran? Improving Concert Classification with Beta Distribution Parameters. To illustrate the value of incorporating estimated α_i and β_i parameters, we examine a concrete example: an Ed Sheeran concert held on 6/29/2023 at the Boch Center Wang Theatre in Boston, MA. Initially, the event's ticket price summary, characterized only by minimum, maximum, mean, and median prices, closely resembles a typical Beyoncé pricing profile. Consequently, the model incorrectly classifies this Ed Sheeran concert as a Beyoncé event (Figure 4a). However, upon incorporating estimated α_i and β_i parameters into the model, the detailed scaled beta distribution reveals subtle yet distinctive features. Specifically, the ticket prices for this Ed Sheeran event exhibit a noticeably sharper price drop compared to a typical Beyoncé concert, aligning clearly with Ed Sheeran's pricing profile (Figure 4b). This refined distributional insight corrects the initial misclassification, highlighting the enhanced accuracy and robustness gained from integrating estimated distribution parameters into the Random Forest classification framework, while also validating the underlying estimated distribution.

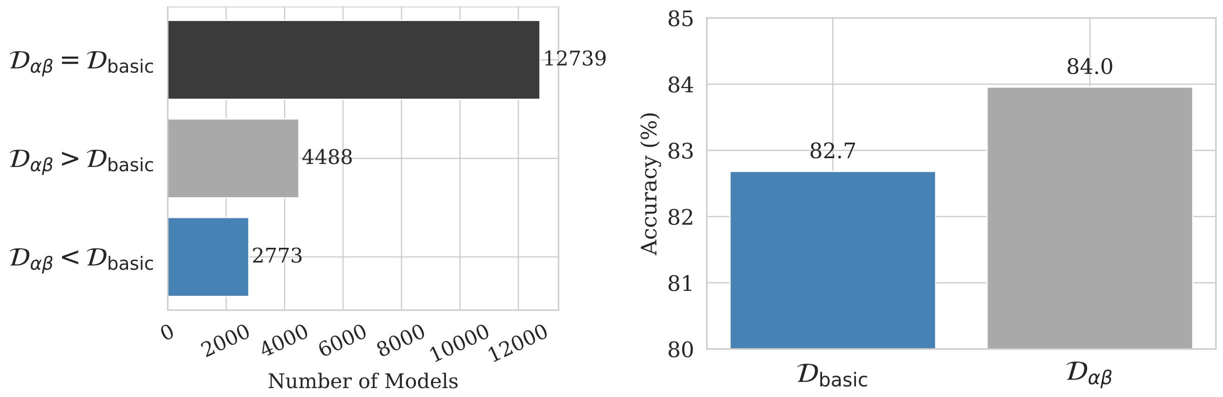
These findings align with publicly available ticket pricing information, where Ed Sheeran's 2023 North American tour featured general admission tickets priced around \$139, with lower-tier tickets starting as low as \$49 [35]. In contrast, Beyoncé's recent tours have exhibited significantly higher pricing structures, ranging widely for standard tickets and higher for extravagant VIP packages [36] [37]. Thus, the disparity in pricing patterns identified by our model reflects real-world pricing dynamics for these artists.

5. Implicit Regularization in Random Forests via Constant-Value Features

Regularization is pivotal for enhancing the generalization performance of machine learning models by mitigating overfitting to noisy or irrelevant features. In Random Forest classifiers, the ensemble nature, together with randomness in both data (bagging, as introduced by Breiman [21]) and feature sampling (random subspace method pioneered by Ho [22]), induces implicit regularization that bolsters robustness. Multiple bootstrapped decision trees reduce variance, while random feature selection at each split introduces decorrelation among trees, curbing model complexity and overfitting. Breiman [24] formalized these ideas by decomposing Random Forest generalization error into components of tree strength and correlation, providing a theoretical underpinning for ensemble variation.

Mathematically, the generalization error of a Random Forest on a binary classification problem can be expressed as

$$\mathbb{E}[(f(x) - \mathbb{E}[y \mid x])^2]$$



(a) Random Forest model performance comparisons for $N_{\text{models}} = 20{,}000$, using the typical default feature selection size of $m = \text{round}(\sqrt{n}) = 2$. The bars show the number of cases in which models trained on $\mathcal{D}_{\alpha\beta}$ performed the same, better, or worse than models trained on $\mathcal{D}_{\text{basic}}$.

(b) Percent Accuracy by Feature Subset for $N_{\text{models}} = 20{,}000$. Although the overall accuracy difference between $\mathcal{D}_{\alpha\beta}$ and $\mathcal{D}_{\text{basic}}$ appears relatively small, it is statistically significant given the large sample size and the substantial proportion of models showing improvement.

Figure 3: Random Forest performance comparison using $\mathcal{D}_{\text{basic}}$ vs. $\mathcal{D}_{\alpha\beta}$ features.

$$= \text{Var}(f(x)) + \text{Bias}^2(f(x)) + \sigma^2, \tag{36}$$

where $\mathbb{E}[y \mid x]$ denotes the true conditional probability of class y given x, $\text{Var}(f(x))$ is the variance of the model's predicted probabilities, $\text{Bias}^2(f(x))$ is the squared bias of the predictions relative to the true probabilities, and σ^2 is irreducible label noise. This variance reduction through randomness aligns with Geurts et al. [23], who enhanced decorrelation via ”Extremely Randomized Trees”, further mitigating variance and promoting stability. Additionally, the entropy-based stability described by Bousquet and Elisseeff [25] underscores the importance of varied ensembles for robust generalization. In this work, we investigate how additional implicit regularization mechanisms, such as introducing zero-variance (constant-value) features, can further enhance Random Forests by affecting feature selection and decision boundaries.

5.1. Analysis of Probabilistic Feature Selection

Notation. For notational clarity in what follows, we let m = `max_features` denote the number of features randomly selected at each split in the Random Forest, following the `scikit-learn` [34] implementation.

In a standard Random Forest construction with fixed-size feature selection, exactly m out of the n total features are chosen at each split node, uniformly over all $\binom{n}{m}$ subsets. Hence, the probability that a particular feature X_j is included in the candidate set at any node is

$$P(\text{Include } X_j) = \frac{m}{n}. \tag{37}$$

Across B trees, each containing an average of L split nodes, the expected total number of times X_j appears in candidate sets is then

$$\mathbb{E}[\text{Count}_{\text{in-candidate}}(X_j)] = B \cdot L \cdot \frac{m}{n}. \tag{38}$$

Feature Selection via Gini Impurity Reduction. For binary classification, the Gini impurity is

$$G = 2\,p\,(1-p), \tag{39}$$

where p is the proportion of one class. Splitting on X_j changes this impurity, reducing it by $\Delta G(X_j)$. We define the **rank** or **score** of feature X_j as

$$r(X_j) = \Delta G(X_j). \tag{40}$$

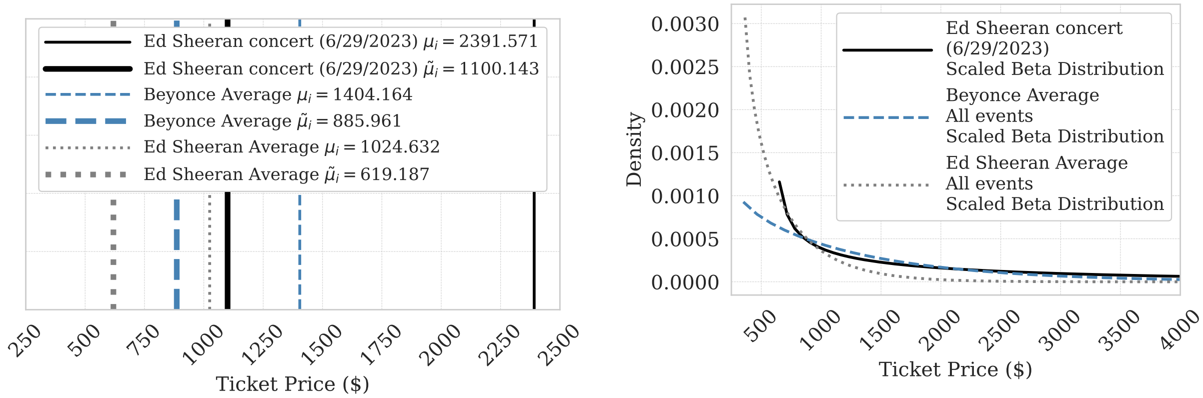
A higher $r(X_j)$ means a larger impurity reduction and thus a higher rank among the available features at a node. Moderately predictive parameters, such as α_i and β_i in the artist classification use-case, can still achieve some positive $r(X_j)$, even if not as large as top-ranked features.

Competitive Advantage of Highly Ranked Features. Although each feature X_j has a nominal $\frac{m}{n}$ chance of appearing in the size-m candidate set at a node, the final split is awarded to whichever feature yields the greatest score. If we assume a proportional “weighted by $r(X_j)$” selection among the m chosen, then for a subset

$$S \subseteq \{1, \ldots, n\}, \quad |S| = m, \tag{41}$$

we have

$$P(S) = \frac{1}{\binom{n}{m}}. \tag{42}$$



(a) Ticket pricing summary for the Ed Sheeran concert on 6/29/2023 at Boch Center Wang Theatre, Boston, MA, using basic statistics (μ_i, $\tilde{\mu}_i$, Min_i, Max_i). Without distribution parameters, the median and mean prices align closely with typical Beyoncé concert values, leading to misclassification.

(b) Comparison of scaled Beta distributions after estimating α_i and β_i parameters for the Ed Sheeran concert (6/29/2023). The estimated distribution shows a more pronounced price drop relative to the typical Beyoncé concert profile, accurately reflecting Ed Sheeran's pricing pattern and correcting the previous misclassification.

Figure 4: Statistical vs. distributional pricing representations for the Ed Sheeran concert on 6/29/2023 at Boch Center Wang Theatre.

Conditioned on S, the probability that X_j wins the split is

$$P(X_j \mid S) \;=\; \frac{r(X_j)}{\sum_{k\in S} r(X_k)}. \tag{43}$$

Hence, the unconditional probability of X_j being chosen for a split is

$$P(X_j) \;=\; \sum_{S:\, j\in S} P(S) \,\cdot\, P(X_j \mid S), \tag{44}$$

which expands to

$$P(X_j) \;=\; \frac{1}{\binom{n}{m}} \sum_{S:\, j\in S} \frac{r(X_j)}{\sum_{k\in S} r(X_k)}. \tag{45}$$

Closed-Form Approximation. When $n \gg m$, or simply for conceptual ease, one can approximate $\sum_{k\in S} r(X_k)$ by its expectation,

$$r(X_j) \;+\; (m-1)\,\mathbb{E}\big[\,r(X_k)\big], \tag{46}$$

yielding

$$P(X_j) \;\approx \tag{47}$$

$$\frac{1}{\binom{n}{m}} \sum_{S:\, j\in S} \frac{r(X_j)}{r(X_j) + (m-1)\,\mathbb{E}\big[\,r(X_k)\big].} \tag{48}$$

Since there are $\binom{n-1}{m-1}$ subsets that include X_j, and $\frac{\binom{n-1}{m-1}}{\binom{n}{m}} = \frac{m}{n}$, we obtain

$$P(X_j) \;\approx\; \frac{m}{n} \,\cdot\, \frac{r(X_j)}{r(X_j) + (m-1)\,\mathbb{E}\big[\,r(X_k)\big].} \tag{49}$$

Thus, even though every feature has the same nominal $\frac{m}{n}$ chance of entering the candidate set, those with consistently higher $r(X_j)$ can dominate, overshadowing less ranked predictors.

Probabilistic Effects of Zero-Variance Variables. Earlier (Section 3), the datasets $\mathcal{D}_{\alpha\beta}$ and $\mathcal{D}_{\delta}$ were extended into $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$ and $\mathcal{D}_{\delta}^{(\text{reg})}$ by including zero-variance (constant-value) features. The number of constant-value features is denoted by n_{ZV}, and these features have approximately zero Gini scores since they provide no impurity reduction. While such features may seem unhelpful, they alter the Random Forest's probabilistic dynamics by providing lower-score competition that can lessen the dominance of top-ranked features.

Let n be the total number of non-constant features (e.g., $\mu_i, \tilde{\mu}_i, \text{Max}_i, \text{Min}_i, \alpha_i, \beta_i$) and n_{ZV} the number of zero-variance features. Thus, the total feature set has $n + n_{\text{ZV}}$ features. If each zero-variance feature has rank score $r_{\text{ZV}} \approx 0$, it dilutes the sum of scores in the approximate denominator, thereby boosting the selection probability of mid-ranked features relative to the scenario without any zero-variance features.

Theorem 4 (Zero-Variance Dilution Effect)**.** *Suppose* n_{ZV}

zero-variance features with $r_{\mathrm{ZV}} \approx 0$ are added, enlarging the feature set from n to $n_{\mathrm{eff}} = n + n_{\mathrm{ZV}}$. Let

$$\bar{r}_{\mathrm{eff}} = \frac{\sum_{j=1}^{n} r(X_j) + n_{\mathrm{ZV}}\, r_{\mathrm{ZV}}}{n_{\mathrm{eff}}} \approx \frac{n\,\bar{r}}{n + n_{\mathrm{ZV}}} \tag{50}$$

$$\text{so that} \quad \bar{r}_{\mathrm{eff}} < \bar{r}. \tag{51}$$

For any two informative features X_h, X_ℓ with scores $a = r(X_h) > b = r(X_\ell) > r_{\mathrm{ZV}}$, the closed-form odds ratio between their selection probabilities satisfies

$$\frac{P_h^{(\mathrm{eff})}(m)}{P_\ell^{(\mathrm{eff})}(m)} < \frac{P_h(m)}{P_\ell(m)}, \tag{52}$$

where

$$P_j(m) = \frac{m}{n}\,\frac{r(X_j)}{r(X_j) + (m-1)\,\bar{r}}, \tag{53}$$

$$P_j^{(\mathrm{eff})}(m) = \frac{m}{n_{\mathrm{eff}}}\,\frac{r(X_j)}{r(X_j) + (m-1)\,\bar{r}_{\mathrm{eff}}}. \tag{54}$$

Thus adding zero-variance features compresses the relative dominance of higher-scoring over lower-scoring variables, giving mid-ranked features more splitting opportunities.

Proof. Write $K = (m-1)\,\bar{r}$ and $K_t = (m-1)\,\bar{r}_{\mathrm{eff}}$ with $K_t < K$. The prefactors $\frac{m}{n}$ and $\frac{m}{n_{\mathrm{eff}}}$ cancel in the ratio, giving

$$\frac{P_h^{(\mathrm{eff})}(m)}{P_\ell^{(\mathrm{eff})}(m)} = \frac{a}{b}\,\frac{b + K_t}{a + K_t}, \quad \frac{P_h(m)}{P_\ell(m)} = \frac{a}{b}\,\frac{b + K}{a + K}. \tag{55}$$

Define

$$R(K) = \frac{a}{b}\,\frac{b + K}{a + K}. \tag{56}$$

A direct derivative gives

$$\frac{dR}{dK} = \frac{a(a-b)}{b\,(a+K)^2} > 0 \tag{57}$$

because $a > b > 0$; hence $R(K)$ is strictly increasing in K. Since $K_t < K$, we have

$$R(K_t) < R(K), \tag{58}$$

establishing the claimed inequality. □

Geurts et al. [23] demonstrated that increasing the randomization of split selection in Extremely Randomized Trees leads to deeper decision trees by weakening the dependence of split choices on the target variable. This increased depth arises because random splits reduce the impurity reduction at each node, thus requiring additional splits to achieve sufficient purity. Formally, this can be expressed as: $\mathbb{E}[d_{\mathrm{random}}] > \mathbb{E}[d_{\mathrm{optimal}}]$, where $\mathbb{E}[d_{\mathrm{random}}]$ and $\mathbb{E}[d_{\mathrm{optimal}}]$ represent the expected tree depths for randomized and optimal splits, respectively ([23]). Our approach and experiments reveal comparable effects, with zero-variance features increasing tree depth and encouraging more variation among splits.

Corollary 5 (Increased Expected Tree Depth)**.** *Consider a Random Forest whose effective feature set is $n_{\mathrm{eff}} = n + n_{\mathrm{ZV}}$, with n informative features ($r(X_j) > 0$) and n_{ZV} zero-variance features ($r_{\mathrm{ZV}} \approx 0$). Let d denote the depth of a decision tree grown under a fixed impurity-based stopping rule. Then, holding all other training hyperparameters constant,*

$$\mathbb{E}\big[d(n_{\mathrm{eff}})\big] > \mathbb{E}\big[d(n)\big]. \tag{59}$$

Proof. Theorem 4 shows that adding zero-variance features compresses the odds of high- versus mid-ranked variables:

$$\frac{P_h^{(\mathrm{eff})}(m)}{P_\ell^{(\mathrm{eff})}(m)} < \frac{P_h(m)}{P_\ell(m)}. \tag{60}$$

Consequently, top-scoring features win fewer splits relative to before, and more mid-ranked features are selected. Because those mid-ranked features achieve smaller impurity reductions ($r_\ell < b < r_h$), the expected impurity drop per internal node is lower. A lower per-split reduction means the chosen impurity threshold is reached later in the recursive partitioning process, so additional levels are needed before termination. Hence the expected depth increases: $\mathbb{E}[d(n_{\mathrm{eff}})] > \mathbb{E}[d(n)]$. □

Increased randomness in split selection explicitly reduces the correlation among trees ([23]), expressed mathematically as:

$$\sigma^2 = \rho\,\frac{\mathrm{Var}(h(x))}{B}, \tag{61}$$

where lower correlation ρ directly reduces ensemble variance σ^2. Our theoretical analysis and empirical results confirm this assertion.

Corollary 6 (Reduced Ensemble Correlation)**.** *Let ρ denote the pair-wise correlation between base learners in a Random Forest. Adding n_{ZV} zero-variance features lowers* *that correlation and hence the variance term $\rho\,\mathrm{Var}(h)/B$ in Breiman's bias–variance decomposition.*

Proof. Without zero-variance features, the highest-ranked variables win a large fraction of candidate splits; many trees therefore grow similar decision paths, inflating ρ. Theorem 4 shows that after augmentation the odds ratio $P_h^{(\mathrm{eff})}(m)/P_\ell^{(\mathrm{eff})}(m)$ shrinks for every pair of scores $a > b > 0$. Consequently, top-ranked variables win fewer splits relative to mid-ranked ones, and different features now have a greater chance of initiating branches. This increased heterogeneity of split choices makes the predictions of individual trees less correlated, so ρ decreases; the factor $\rho\,\mathrm{Var}(h)/B$ is therefore reduced. □

5.2. Expanding the Regularization Search Space.

Prior work by Mentch and Zhou [28] studies how tuning the m (`max_features`) hyperparameter, the number of features considered at each split, regularizes Random Forests by changing the probability that any given feature is selected. Their approach relies on discrete steps (since m must be an integer), which partitions the feature-selection probabilities into a finite set. In contrast, our method of adding constant-value (zero-variance) features changes the total feature count continuously, producing a near-continuum of feature-selection probabilities. The theorem below formalizes how introducing n_{ZV} constant features can approximate any target probability in a broad interval, thereby "filling in the gaps" left by the discrete m adjustments alone.

Theorem 7 (Continuous Approximation via Constant-Feature Dilution)**.** *Let n denote the number of truly informative features, and fix an integer m such that $1 \le m \le n$. Let n_{ZV} be the number of constant (zero-variance) features added. In the absence of constant features, the effective probability of selecting an informative feature at a split is*

$$\gamma \;=\; \frac{m}{n}. \tag{62}$$

If we add $n_{\mathrm{ZV}} \ge 0$ constant (zero-variance) features, then the total number of features is $n + n_{\mathrm{ZV}}$, and the effective selection probability of an informative feature becomes

$$\gamma' \;=\; \frac{m}{n + n_{\mathrm{ZV}}}. \tag{63}$$

For any desired probability

$$0 \;<\; \gamma^* \;\le\; \frac{m}{n}, \tag{64}$$

there exists an integer n_{ZV} such that γ' can be made arbitrarily close to γ^*. Consequently, the set

$$S_L \;=\; \left\{ \tfrac{m}{n+n_{\mathrm{ZV}}} : n_{\mathrm{ZV}} \in \mathbb{N}_0 \right\} \tag{65}$$

is dense in the interval $\left(0, \tfrac{m}{n}\right]$. Equivalently, we can say that adding constant features expands the tuning space from a discrete set $\{1/n,\, 2/n,\, \ldots,\, 1\}$ to a near-continuum.

Proof. **Discrete Set Without Constant Features.** Following Mentch and Zhou [28], let n be the number of informative features and let m be the chosen subset size at each split. The probability that any one informative feature appears in a candidate set is then $\gamma = \frac{m}{n}$. Because m must be an integer with $1 \le m \le n$, the set of possible probabilities (as m varies) is

$$S_{MZ} \;=\; \left\{ \tfrac{1}{n}, \tfrac{2}{n}, \ldots, \tfrac{n}{n} \right\}. \tag{66}$$

This set is finite and discrete.

Continuous Dilution via Constant Features. Now, fix m. Instead of varying m itself, we *add* n_{ZV} constant (zero-variance) features to the existing n informative ones, so the total feature count is $n + n_{\mathrm{ZV}}$. As a result, the effective probability of picking an informative feature becomes

$$\gamma' \;=\; \frac{m}{n + n_{\mathrm{ZV}}}.$$

Hence, each nonnegative integer n_{ZV} in $\{0, 1, 2, \ldots\}$ produces one element of the set

$$S_L \;=\; \left\{ \tfrac{m}{n+n_{\mathrm{ZV}}} : n_{\mathrm{ZV}} \in \mathbb{N}_0 \right\}.$$

Because n_{ZV} can grow arbitrarily large, the values of γ' can get arbitrarily close to 0. Also note that when $n_{\mathrm{ZV}} = 0$, $\gamma' = \frac{m}{n} = \gamma$. Thus, S_L spans probabilities in $(0, \frac{m}{n}]$.

Approximating Any Target γ^*. We now show S_L is dense in the interval $\left(0, \frac{m}{n}\right]$. Take any target probability γ^* satisfying

$$0 < \gamma^* \le \frac{m}{n}.$$

We want to find n_{ZV} so that $\left|\frac{m}{n+n_{\mathrm{ZV}}} - \gamma^*\right|$ is arbitrarily small. Observe that

$$\frac{m}{n + n_{\mathrm{ZV}}} \;=\; \gamma^* \quad\Longleftrightarrow\quad n_{\mathrm{ZV}} \;=\; \frac{m}{\gamma^*} \;-\; n. \tag{67}$$

Since $n_{\mathrm{ZV}} \in \mathbb{N}_0$, we cannot in general solve this equation *exactly*, but we can pick

$$n_{\mathrm{ZV}} \;=\; \left\lfloor \tfrac{m}{\gamma^*} - n \right\rfloor \quad\text{or}\quad n_{\mathrm{ZV}} \;=\; \left\lceil \tfrac{m}{\gamma^*} - n \right\rceil, \tag{68}$$

whichever is nonnegative. Either choice differs from $\frac{m}{\gamma^*} - n$ by less than 1. This yields

$$\left|\frac{m}{n+n_{\mathrm{ZV}}} - \gamma^*\right| = \left|\frac{m}{n+\lfloor\frac{m}{\gamma^*} - n\rfloor} - \gamma^*\right| < \epsilon \tag{69}$$

for arbitrarily small ϵ, provided we choose n_{ZV} large enough. Thus $\gamma' \in S_L$ can be made arbitrarily close to γ^*.

Since any real number $\gamma^* \in (0, \frac{m}{n}]$ can be approximated arbitrarily well by a suitable choice of n_{ZV}, the set $\left\{\frac{m}{n+n_{\mathrm{ZV}}} : n_{\mathrm{ZV}} \in \mathbb{N}_0\right\}$ is dense in $(0, \frac{m}{n}]$. This shows that by adding zero-variance features, we transform the discrete selection probabilities $\{\frac{1}{n}, \frac{2}{n}, \ldots, 1\}$ into a near-continuum $\{\frac{m}{n+n_{\mathrm{ZV}}} : n_{\mathrm{ZV}} \in \mathbb{N}_0\}$, offering finer granularity in regularizing Random Forests than integer m adjustments alone. □

Corollary 8 (Continuous Accuracy Expansion via Selection Probability)**.** *Let*

$$\gamma' = \frac{m}{n + n_{\mathrm{ZV}}}$$

be the effective probability of selecting an informative feature when the original n features are augmented with n_{ZV} constant (zero-variance) features. Assume that the mapping from γ' to the classifier's accuracy ν is continuous, and let $(\nu_{\min}, \nu_{\max}]$ denote the interval of achievable accuracy values under the original discrete scheme. Then, for any target accuracy ν^* satisfying

$$\nu_{\min} < \nu^* \leq \nu_{\max}, \tag{70}$$

and for any $\epsilon > 0$, there exists an integer $n_{\mathrm{ZV}} \geq 0$ such that the accuracy ν' obtained with $\gamma' = \frac{m}{n+n_{\mathrm{ZV}}}$ satisfies

$$|\nu' - \nu^*| < \epsilon. \tag{71}$$

In other words, the set of achievable accuracies is dense in $(\nu_{\min}, \nu_{\max}]$, providing near-continuous control over the model's performance by fine-tuning the effective selection probability.

Proof. By Theorem 7, the set

$$S_L = \left\{\frac{m}{n + n_{\mathrm{ZV}}} : n_{\mathrm{ZV}} \in \mathbb{N}_0\right\}$$

is dense in $(0, \frac{m}{n}]$. Thus, for any effective selection probability γ^* with $0 < \gamma^* \leq \frac{m}{n}$ and any $\delta > 0$, there exists an $n_{\mathrm{ZV}} \in \mathbb{N}_0$ such that

$$\left|\frac{m}{n + n_{\mathrm{ZV}}} - \gamma^*\right| < \delta. \tag{72}$$

Since the mapping from γ' to the classifier's accuracy ν is assumed to be continuous, the corresponding accuracy value ν' can be made arbitrarily close to any target accuracy ν^* in $(\nu_{\min}, \nu_{\max}]$. Formally, for any $\epsilon > 0$ there exists an $n_{\mathrm{ZV}} \geq 0$ such that

$$|\nu' - \nu^*| < \epsilon,$$

which completes the proof. □

Adjusting m indeed influences a Random Forest's effective complexity, as demonstrated in [28], but the discrete nature of integer m limits the granularity with which one can tune the selection probability. By contrast, expanding the feature set from n to $n + n_{\mathrm{ZV}}$ via constant-value features achieves γ' values for any integer $n_{\mathrm{ZV}} \geq 0$. Since those values densely fill the interval $(0, \frac{m}{n}]$, the user gains near-continuous control over the effective regularization level. In essence, this approach allows one to mimic or surpass the effect of "small" m while refining probabilities in ways not possible through integer steps alone. Similar continuous approximations from discrete parameter spaces have been explored in related contexts such as hyperparameter tuning, where carefully refining discrete grids approximates continuous optimization [29]. These techniques collectively broaden the applicability and flexibility of discrete-choice methods in machine learning.

5.3. Relation to Penalty-Based Regularization, Functional Data Analysis, and other domains.

Our approach connects fundamentally to classic regularization methods, such as ridge regression [30, 31], where explicit quadratic penalties emerge naturally from Gaussian priors:

$$x_{MAP} = \arg\min_x\{\|Ax - b\|^2 + \lambda\|x\|^2\}. \tag{73}$$

Extending beyond foundational work, the feature probability reweighting structurally parallels recent advancements in regularization across various domains. For instance, in functional data analysis (FDA), recent methods such as roughness penalization in free-knot spline estimation [32] redistribute information to avoid over-concentration on specific knots, maintaining balanced representations. Similarly, our implicit regularization dynamically adjusts feature selection probabilities, preventing dominance by specific features:

$$p_i' = \frac{p_i}{1 + \lambda \sum_j p_j}, \tag{74}$$

This formulation resembles penalized optimization used in FDA:

$$C = \arg\min_{C} \|Y - \Phi^T C\|^2 + \lambda C^T R C, \tag{75}$$

which explicitly penalizes abrupt variations to enforce smoothness.

Furthermore, our implicit feature-selection regularization is related to another penalty-based approach, the inverse-problem hyperparameter optimization framework introduced by Dunbar et al. [33], whose formulation includes a log-determinant regularization:

$$L_M^{(EKI)}(u) = \|\Gamma(u)^{-1/2}(z - G(u))\|^2 - 2\log P(u). \tag{76}$$

The interplay between implicit and explicit regularization frameworks presents an intriguing intersection of theoretical and applied perspectives.

5.4. Analogous Effects in Ticket Pricing & Digit Classification

The same phenomena appear in both the (i) ticket pricing dataset with features derived from distribution summaries and (ii) handwritten digit classification data. In these applications, highly ranked features (e.g., strong distributional predictors or highly informative pixel locations) dominate the splits, often stifling the expression of less dominant but still important features. Introducing zero-variance variables can mitigate this dominance, allowing subtly useful features to be chosen more frequently, thereby enriching the model.

In essence, the zero-variance features act like an implicit regularization mechanism in a manner similar to setting $m = 1$ in Random Forests: they increase the probability of selecting secondary features by reducing the effective weight of extremely high-ranked ones. As a result, weaker signals gain more splitting opportunities, improving the overall variety of the ensemble and, potentially, leading to better performance.

5.5. Experimental Results

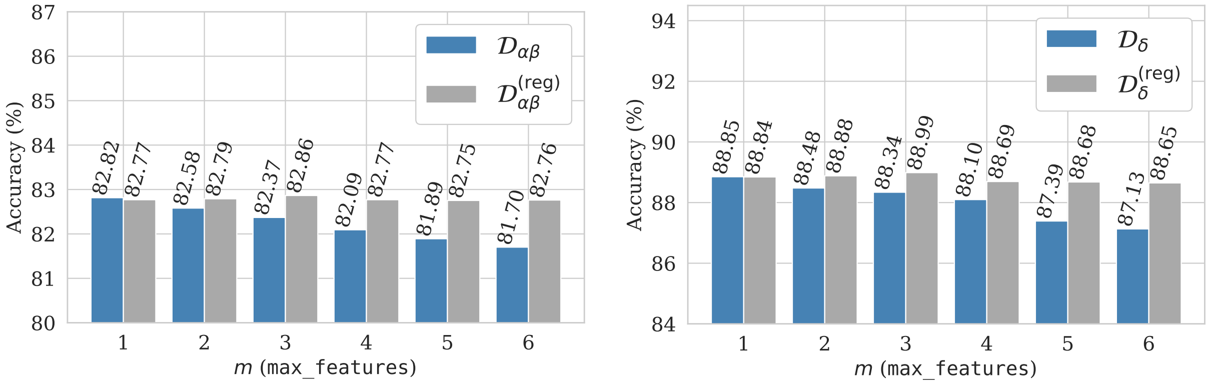
This section applies the same pairwise Random Forest classification methodology used in the $\mathcal{D}_{\alpha\beta}$ experiments to investigate how zero-variance features serve as an implicit regularizer in two distinct domains: the new concert ticket pricing dataset and the UCI handwritten digits dataset [4]. First, we compare $\mathcal{D}_{\alpha\beta}$ to $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$, wherein the new ticket pricing dataset contains 5,000 artist-pair classification models spanning 954 unique artists, each trained and tested on an 80/20 split that yields, on average, ≈ 40 training events and ≈ 10 test events per model. Second, we examine $\mathcal{D}_\delta$ versus $\mathcal{D}_\delta^{(\text{reg})}$ using the standard UCI handwritten digits dataset, constructing 90 digit-pair classifiers (10 unique digits) with a similar 80/20 split that provides around ≈ 287 training comparisons and ≈ 72 test comparisons per pair. For models that use regularization, $n_{\text{ZV}} = 20$. For $\mathcal{D}_\delta$, a random subset of $n = 6$ representative features is chosen for consistency with $\mathcal{D}_{\alpha\beta}$. These complementary experiments highlight the consistently beneficial effect of implicit regularization via zero-variance features, as observed both in real-world secondary ticket pricing data and a classic benchmark dataset.

Accuracy and the Selection Size, m. Figures 5a and 5b illustrate the accuracy trends for the concert pricing and handwritten digit datasets as m is varied. While iterating m in a standard Random Forest adjusts the probability of selecting an informative feature, our experiments reveal that the highest accuracy is achieved only when implicit regularization is applied. As shown in Theorem 7, adding constant (zero-variance) features expands the tuning space from discrete m steps to a near-continuum of effective selection probabilities, and corollary 8 confirms that this continuous parameter space enables fine-grained adjustments of the classifier's performance. The theoretical selection probability reaches extreme values at the boundaries of m (e.g., $m = 1$ yields uniform selection probabilities, while $m = n$ favors highly ranked features), and our empirical findings demonstrate that leveraging the continuous tuning enabled by implicit regularization provides an additional degree of flexibility. By incorporating constant features, the effective average rank of candidate features is diluted, balancing the influence of dominant predictors with that of less prominent ones, thereby achieving accuracy levels that would be difficult or impossible to reach by iterating m alone.

Scope of Model Improvements. We analyze the improvements when $m = 6$, the value with the largest discrepancy, for both ticket pricing and digit classification datasets. Figures 6a and 6b illustrate the performance outcomes, showing statistically significant improvements due to regularization via zero-variance features.

For the ticket pricing data, $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$ includes zero-variance features, while $\mathcal{D}_{\alpha\beta}$ incorporates distributional parameters α and β but no zero-variance features. A paired comparison reveals $n_{\text{better}} = 1084$, $n_{\text{worse}} = 675$, and $N' = 1759$. Under the null hypothesis H_0 that there is no difference (with $p = 0.5$), the mean and standard deviation are:

$$\mu = N' \cdot 0.5 = 879.5, \quad \sigma = \sqrt{N' \cdot 0.5 \cdot 0.5} \approx 20.98.$$



(a) Percent accuracy across feature selection size (m) iterations for artist event training datasets $\mathcal{D}_{\alpha\beta}$ and $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$. The impact of constant features is evident, with $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$ showing declining accuracy as m increases. The highest accuracy is achieved using zero-variance feature regularization, exceeding what standard hyperparameter tuning alone can reach.

(b) Percent accuracy across feature selection size (m) iterations for handwritten digit training datasets $\mathcal{D}_{\delta}$ and $\mathcal{D}_{\delta}^{(\text{reg})}$. The impact of constant-value features is again evident, with $\mathcal{D}_{\delta}^{(\text{reg})}$ showing declining accuracy as m increases. Zero-variance feature regularization achieves the highest accuracy, unattainable via standard tuning alone.

Figure 5: Accuracy trends across feature selection sizes (m) for artist and digit datasets, highlighting the implicit regularization effects of constant-value features.

The Z-score is:

$$Z = \frac{n_{\text{better}} - 0.5 - \mu}{\sigma} = \frac{1084 - 0.5 - 879.5}{20.98} \approx 9.72,$$

yielding $p < 10^{-21}$. Figure 6a confirms this statistically significant improvement.

For the digit classification data, $\mathcal{D}_{\delta}^{(\text{reg})}$ includes zero-variance features while $\mathcal{D}_{\delta}$ excludes them. A paired comparison reveals $n_{\text{better}} = 52$, $n_{\text{worse}} = 14$, and $N' = 66$. The mean and standard deviation are:

$$\mu = N' \cdot 0.5 = 33, \quad \sigma = \sqrt{N' \cdot 0.5 \cdot 0.5} \approx 4.06.$$

The Z-score is:

$$Z = \frac{n_{\text{better}} - 0.5 - \mu}{\sigma} = \frac{52 - 0.5 - 33}{4.06} \approx 4.56,$$

yielding $p < 10^{-5}$. Figure 6b highlights the improvements.

Table 3 summarizes the statistical significance of the improvements for both datasets.

Feature Re-ranking and Usage in the Models. In both $\mathcal{D}_{\alpha\beta}$ vs. $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$ and $\mathcal{D}_{\delta}$ vs. $\mathcal{D}_{\delta}^{(\text{reg})}$, introducing zero-variance features modifies the selection probabilities in the approximate formula

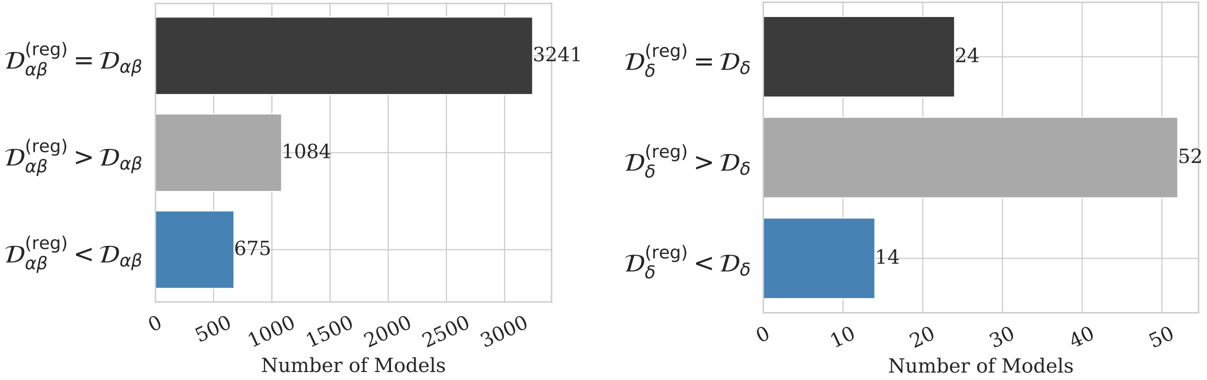
$$P(X_j) \approx \frac{m}{n_{\text{eff}}} \cdot \frac{r(X_j)}{r(X_j) + (m-1)\,\overline{r},}$$

where $n_{\text{eff}} = n + n_{\text{ZV}}$ when constant features are present and equals n otherwise. By injecting low-scoring features into the ensemble, these datasets effectively dilute the dominance of highly ranked predictors X_j. Consequently, the final usage distribution, aggregated across all base learners, becomes more balanced, giving subtle but informative features more opportunities at split nodes. This re-ranking serves as a form of implicit regularization, stabilizing the Random Forest. Figures 7a and 7b show the redistribution of feature usage, highlighting the increased prominence of moderately-ranked predictors. Notably, the shifts shown in these figures underscore how implicit regularization effectively promotes model robustness through enhanced feature breadth.

Increased Tree Depth with Implicit Regularization. We further explore the effects of zero-variance features on Random Forest models by calculating the average tree depth across all models. The analysis demonstrates that models with zero-variance features (from $\mathcal{D}_{\delta}^{(\text{reg})}$ and $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$) produce deeper trees than their non-regularized counterparts. This finding, as formalized in corollary 5, supports the idea that implicit regularization encourages the training process to exploit more nuanced feature representations. The experimental results validate our analysis of expected tree depth, building on and extending prior findings by [23] on how randomized split selection increases depth through reduced impurity gains. Figures 8a and 8b demonstrate the increase in tree depth observed when zero-variance features are added. The consistent and significant depth enhancements seen in these figures

Table 3: Summary of statistical results for improvements with $m = 6$.

Statistic	**Tickets**	**Digits**
Effective Sample Size (N')	1759	66
n_{better}	1084	52
n_{worse}	675	14
Mean ($\mu = N'/2$)	879.5	33
Std. Dev. (σ)	20.98	4.06
Z-score	9.72	4.56
p-value	$< 10^{-21}$	$< 10^{-5}$



(a) Performance comparison for $m = 6$ across $N_{\text{models}} = 5{,}000$ artist classification models. Bars indicate how often $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$ performed the same, better, or worse than $\mathcal{D}_{\alpha\beta}$ (see Figure 5a).

(b) Performance comparison for $m = 6$ across $N_{\text{models}} = 90$ digit classification models. Bars indicate how often $\mathcal{D}_{\delta}^{(\text{reg})}$ performed the same, better, or worse than $\mathcal{D}_{\delta}$ (see Figure 5b).

Figure 6: Effect of constant-value feature regularization at $m = 6$, showing the distribution of model performance changes for artist and digit classification tasks.

highlight the stabilizing effect of implicit regularization across distinct datasets. Specifically, for the ticket pricing models, the median depth increased from 3.0 to 4.0, and the average depth increased from 3.18 to 4.16. For the handwritten digits models, the median depth increased from 8.0 to 10.0, and the average depth increased from 8.35 to 10.42.

Tree Variety as Measured by Feature Count Distance. We quantify the ensemble “variety” by examining the pairwise Euclidean distance between trees’ feature usage vectors, $\mathbf{v}_i \in \mathbb{R}^d$. Defining the distance between trees i and j as $\|\mathbf{v}_i - \mathbf{v}_j\|_2$, we compute the sum over all pairs:

$$V(m) = \sum_{1 \le i < j \le n} \|\mathbf{v}_i - \mathbf{v}_j\|_2. \tag{77}$$

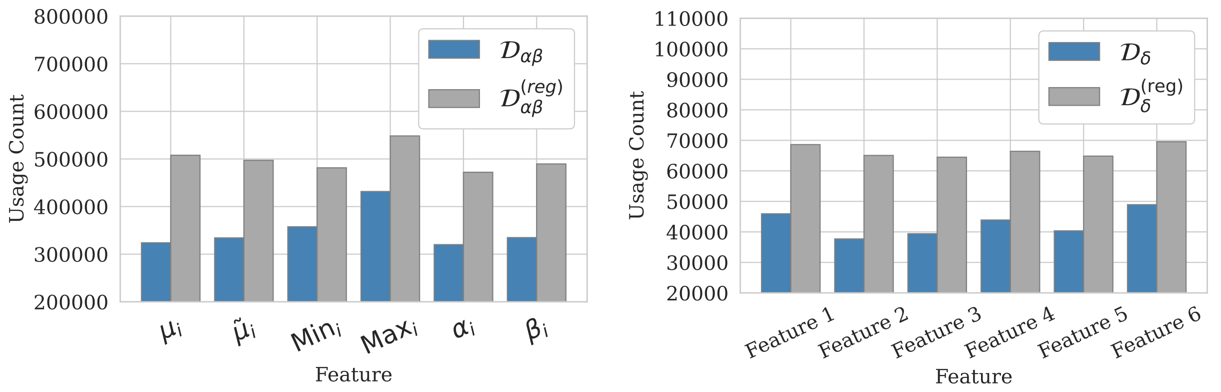
This measure is computed for each model in both the regularized and non-regularized datasets (see Figures 9a and 9b). A higher average $V(m)$ indicates more varied feature usage among trees, which is consistent with the effect of zero-variance features. Figures 9a and 9b illustrate the increased variety in feature usage induced by zero-variance features. This enhanced range drives the reduced correlation among ensemble members, thereby confirming the stabilizing impact of implicit regularization across distinct modeling scenarios. Specifically, for the ticket pricing models, the median variety increased from 2.00 to 2.83, and the average variety increased from 2.26 to 2.99. For the handwritten digits models, the median variety increased from 4.90 to 7.35, and the average variety increased from 5.27 to 7.67.

Geurts et al. [23] show that the fully randomized split selection in Extra-Trees reduces correlation among trees, which we quantify via the average cosine similarity of their normalized feature usage vectors. In particular, for any two trees with vectors $\mathbf{v}_i$ and $\mathbf{v}_j$ (with $\|\mathbf{v}_i\|_2 = \|\mathbf{v}_j\|_2 = 1$), we have

$$\mathbf{v}_i^\top \mathbf{v}_j = 1 - \frac{1}{2}\|\mathbf{v}_i - \mathbf{v}_j\|_2^2. \tag{78}$$

Defining the average correlation p as

$$p = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \mathbf{v}_i^\top \mathbf{v}_j, \tag{79}$$



(a) Average feature usage counts across $N_{\text{models}} = 5{,}000$ for $\mathcal{D}_{\alpha\beta}$ versus $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$.

(b) Average feature usage counts across $N_{\text{models}} = 90$ for $\mathcal{D}_{\delta}$ versus $\mathcal{D}_{\delta}^{(\text{reg})}$.

Figure 7: Comparison of average feature usage between unregularized and regularized models, across both artist and digit datasets. Regularization via constant-value features leads to more varied and balanced feature selection.

we obtain

$$p = 1 - \frac{1}{n(n-1)} \sum_{1 \le i < j \le n} \|\mathbf{v}_i - \mathbf{v}_j\|_2^2. \tag{80}$$

Thus, as our variety measure $V(m)$ increases, the average correlation p decreases, demonstrating that greater tree variety leads to reduced inter-tree correlation. This mathematical relationship aligns with our analysis, showing that implicit regularization via zero-variance features promotes a more decorrelated ensemble.

Error Decomposition and Implicit Regularization Effects. Recalling the bias–variance decomposition, the inclusion of zero-variance features modifies the ensemble variance by diluting the impact of top-ranked features and increasing tree variety. Formally, if ρ is the correlation among tree predictions, then

$$\mathrm{Var}(f(x; n_{\mathrm{ZV}})) = \frac{\rho(n_{\mathrm{eff}})\,\mathrm{Var}(h(x; n_{\mathrm{ZV}}))}{B}. \tag{81}$$

This effect is again consistent with earlier observations by Geurts et al. [23], who demonstrate that increased randomness directly reduces ensemble correlation. Zero-variance features decrease $\rho(n_{\mathrm{eff}})$ (see Corollary 6), thus reducing overall variance. While the bias may increase slightly when moderate features are used more often, the net effect is improved generalization, as supported by our experiments.

The regularization perspective is strengthened by the findings of Wyner et al. [26], who demonstrated that AdaBoost generalizes successfully due to averaging ensembles of interpolating classifiers, thus achieving ”spiked-smooth” decision boundaries. Their results suggest that regularization emerges implicitly through ensemble averaging rather than explicit complexity penalties, precisely as we observe in our Random Forests with zero-variance features. Mathematically, their derived variance reduction for AdaBoost,

$$\mathrm{Var}(F) = \frac{\mathrm{Var}_H}{J}, \tag{82}$$

directly parallels our Random Forest variance expression,

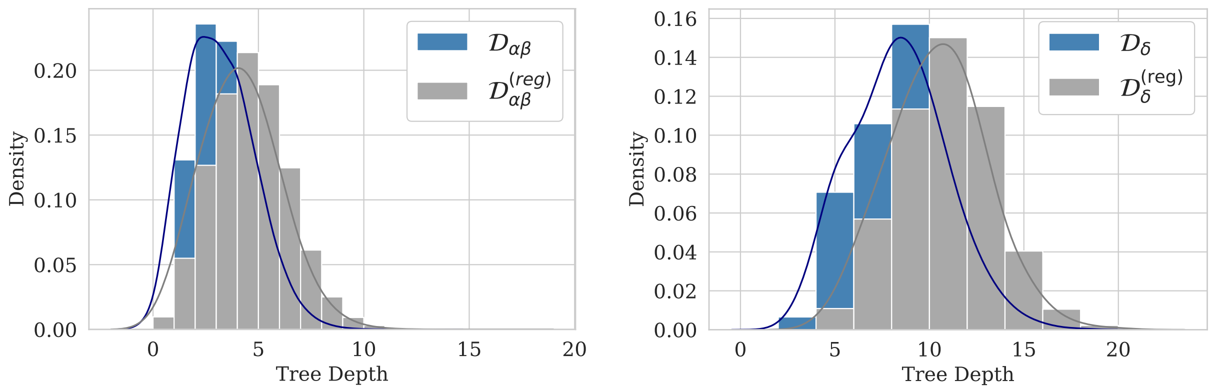
$$\mathrm{Var}(F) \approx \frac{\rho \mathrm{Var}_h}{B}. \tag{83}$$

Thus, both methodologies validate implicit regularization through variance reduction as the underlying principle for generalization performance.

Dropkick Murphys vs. The Avett Brothers - implicit regularization rescues a borderline case.

A Dropkick Murphys concert on 10/28/2023 is initially mislabelled by the baseline Random Forest as an Avett Brothers show. In the unregularized model the four location statistics - mean (μ_i), median ($\tilde{\mu}_i$), Min_i, and Max_i - hold the highest empirical ranks, so they dominate the candidate-split lotteries in almost every tree (Figure 11a, blue bars). The artists' feature distributions across all events, as shown in Figure 10, experience significant overlap. The ensemble, initially driven by location parameters (Figure 11a), overlooks the clear visual match between the event price density (black curve) and the Dropkick template (blue dashed) shown in Figure 11b.

The implicit regularization mechanism introduced in Section 5 adds n_{ZV} zero-variance columns, diluting the feature pool from n to $n + n_{\mathrm{ZV}}$ and lowering the expected rank of every genuine predictor. Theorem 4 (Zero-



(a) Average tree depth for $N_{\text{models}} = 5{,}000$ models: $\mathcal{D}_{\alpha\beta}$ vs. $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$. Median depth increased from 3.0 to 4.0, and average depth from 3.18 to 4.16.

(b) Average tree depth for $N_{\text{models}} = 90$ models: $\mathcal{D}_{\delta}$ vs. $\mathcal{D}_{\delta}^{(\text{reg})}$. Median depth increased from 8.0 to 10.0, and average depth from 8.35 to 10.42.

Figure 8: Effect of zero-variance feature regularization on tree depth. Both artist and digit models trained with regularized datasets grow deeper trees on average, suggesting increased robustness and feature utilization.

Variance Dilution Effect) shows that this strictly increases the sampling probability of all mid-ranked features, most notably the shape coefficient β_i. Empirically, β_i exhibits a marked increase in split selection frequency (grey bars in Figure 11a), accompanied by a corresponding decline across the location statistics. With a signature shape that typifies Dropkick concerts now amplified, a majority of trees flip their vote and the ensemble classifies the concert correctly.

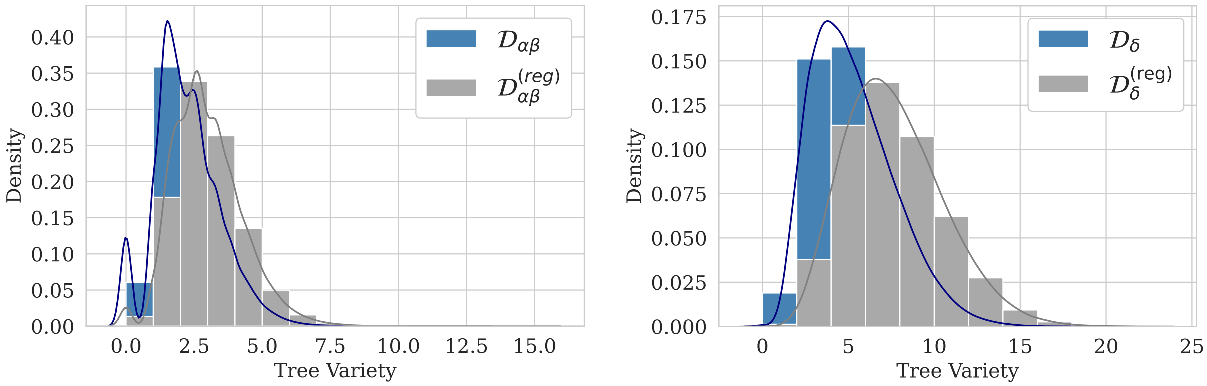
Every prediction that was already correct for either artist remains unchanged. Regularization does not redraw the decision boundary, nor does it suppress informative location cues; it simply grants the discriminative shape parameters enough opportunity to steer borderline cases toward the proper class while also trimming tree-to-tree correlation for a modest bias-variance bonus across the artist pair. Crucially, this success emerges only from the combination of scaled-Beta parameter estimation and zero-variance regularization; either component in isolation leaves the forest confined to a discrete subset of models that standard hyperparameter tuning cannot reach. The theoretical analysis in Section 5 anticipates this behavior and Figures 5a and 5b visualize it by showing how regularization unlocks a near continuous space of feature weightings inaccessible to both the baseline feature set and a standard hyperparameter search.

6. Conclusions

We present a novel approach for identifying distinct economic signatures in the secondary ticket resale market by modeling dynamic pricing data using scaled Beta distributions. Our method estimates the distribution parameters α_i and β_i from limited statistical features $(\text{Min}_i, \text{Max}_i, \mu_i, \tilde{\mu}_i)$, significantly improving artist classification accuracy and strengthening confidence in the fidelity of the modeled distributions. These results underscore the value of probabilistic modeling in uncovering unique patterns in dynamic systems.

The theoretical foundation for our parameter estimation builds upon classical work on Beta distributions [9], while extending these results to the challenging setting where only minimal summary statistics are available. Prior work has explored quantile-based methods for estimating distributions under limited information [12, 14, 15]. Our composite quantile-and-moment matching approach is distinguished by its ability to recover scaled Beta parameters from just mean and median estimates, quantities readily available in real-world data pipelines such as those provided by APIs like SeatGeek [2]. This capability is particularly valuable for large-scale market data where full empirical distributions are unavailable, allowing us to extract meaningful latent structure from highly compressed observations.

The ticket pricing use case is compelling, given the widespread popularity of live concerts and the ubiquity of secondary market platforms such as SeatGeek, StubHub, and Ticketmaster. These markets handle billions of dollars annually, making them both economically significant and rich sources of dynamic data. Our findings not only illuminate the unique pricing signatures of artists but also provide a foundation for further research into the economics of live entertainment and event dynamics. Recent empirical studies in time series classification [7, 11, 10] demonstrate the importance of feature-based methods in high-



(a) Average tree variety for $N_{\text{models}} = 5{,}000$ models: $\mathcal{D}_{\alpha\beta}$ vs. $\mathcal{D}_{\alpha\beta}^{(\text{reg})}$. Median variety increased from 2.00 to 2.83, and average variety from 2.26 to 2.99.

(b) Average tree variety for $N_{\text{models}} = 90$ models: $\mathcal{D}_{\delta}$ vs. $\mathcal{D}_{\delta}^{(\text{reg})}$. Median variety increased from 4.90 to 7.35, and average variety from 5.27 to 7.67.

Figure 9: Effect of zero-variance feature regularization on Random Forest tree variety. Regularization increases both the median and average distances of tree structures in the ensemble, improving generalization capacity.

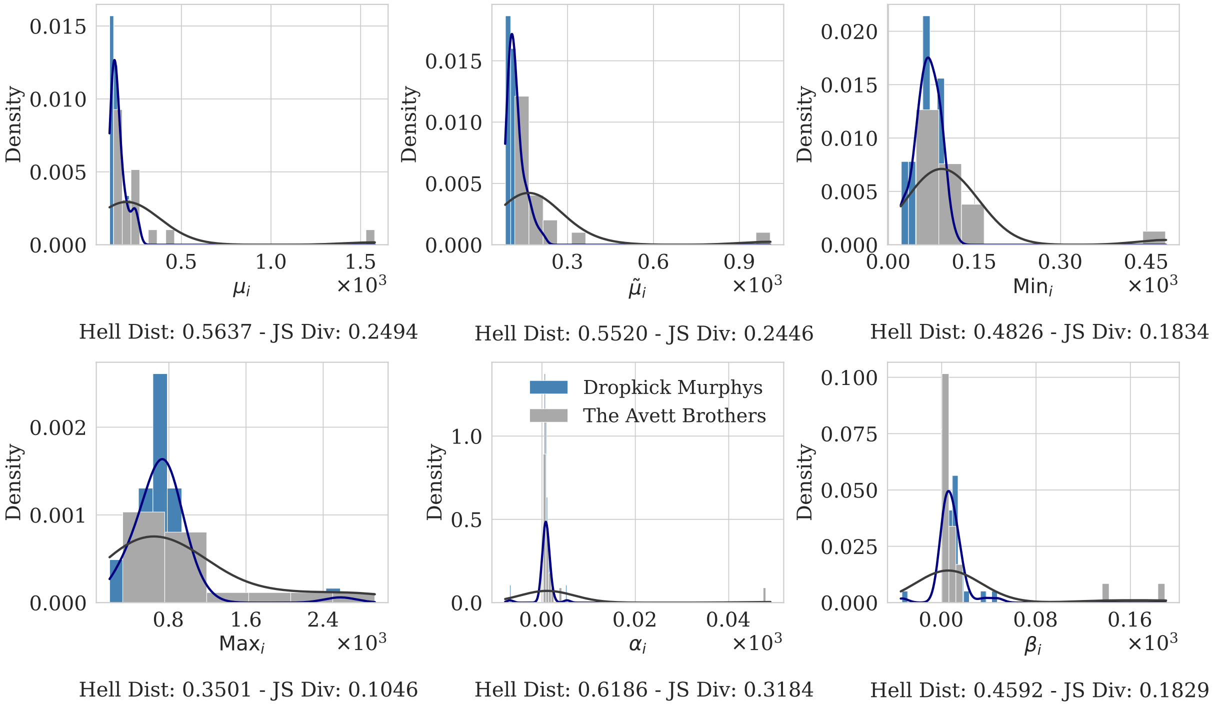
dimensional time series settings. Our work contributes to this literature by introducing a targeted, distribution-based feature representation that captures artist-specific pricing behavior.

In addition to the applied contribution, this work introduces a formal theoretical framework linking classification accuracy to distributional fidelity, grounded in information-theoretic measures such as Total Variation Distance and Jensen-Shannon divergence [18, 19, 20]. Building on these foundations, we show that improved classification accuracy corresponds to increasingly precise parameter estimates for the underlying pricing distributions. The quadratic convergence rate established through Jensen-Shannon divergence analysis offers strong guarantees on the stability and efficiency of our estimation process, even when applied to sparse and noisy real-world data.

Additionally, we demonstrate that incorporating zero-variance (constant-value) features into Random Forest models acts as an implicit regularizer. This mechanism reduces the dominance of highly ranked features, promoting feature variety, deeper trees, and improved generalization performance. The regularization effect balances the bias-variance tradeoff, mitigating overfitting and enhancing robustness. This work draws upon and extends a large body of literature on implicit regularization and entropy-based learning in ensemble methods [21, 24, 22, 23, 25, 26, 27, 28, 33]. Our findings illustrate that even deceptively simple interventions, such as introducing constant-value features, can strategically reshape the feature-selection dynamics of Random Forests by influencing the entropy structure of the model's splitting process. This rebalancing provides a form of stochastic regularization that complements existing hyperparameter-based approaches [28, 33], and in some cases opens previously inaccessible regions of the hyperparameter space.

Random Forests remain a widely used algorithm in industry due to their versatility, interpretability, and strong performance across various domains such as finance, healthcare, energy, and e-commerce. By enhancing these models with the proposed methods, including implicit regularization and probabilistic feature modeling, we verify their continued relevance and ability to address modern, dynamic datasets effectively. Our analysis echoes prior work that framed Random Forests as interpolating ensembles capable of achieving strong generalization even in high-capacity regimes [26], while offering a new perspective on how structural adjustments to the feature space can further improve ensemble variety and stability.

Our findings are validated across both a newly curated ticket pricing dataset and the UCI handwritten digits benchmark dataset [4]. This cross-domain validation confirms the generalizability of the proposed techniques and their applicability to a wide range of dynamic datasets.



Hell Dist: 0.5637 - JS Div: 0.2494

Hell Dist: 0.5520 - JS Div: 0.2446

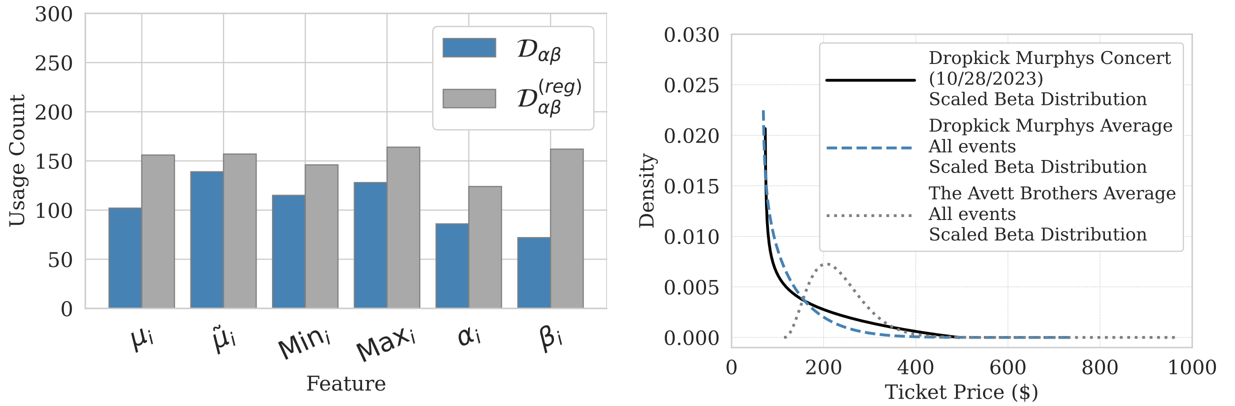
Hell Dist: 0.4826 - JS Div: 0.1834

Hell Dist: 0.3501 - JS Div: 0.1046

Hell Dist: 0.6186 - JS Div: 0.3184

Hell Dist: 0.4592 - JS Div: 0.1829

Figure 10: The plots show the distributions of each feature across all events for artists Dropkick Murphys and The Avett Brothers. In each case, there is considerable overlap, making classification as determined by location statistics alone difficult. With a satisfactory re-ranking of feature importances by the zero-variance implicit regularization mechanism, the Random Forest model can classify more effectively based on the estimated distributional shape as shown explicitly in Figure 11.



(a) Feature-usage counts across the Random Forest before (blue) and after (grey) the zero-variance regularization. Location statistics (Min$_i$, Max$_i$, μ_i, $\tilde{\mu}_i$) remain frequent but lose relative dominance, while the shape parameter β_i increases its appearances in split decisions significantly, an explicit illustration of the Zero-Variance Dilution Effect that corrects the misclassification and leaves earlier correct calls intact.

(b) Scaled-Beta ticket-price densities: the fitted distribution for the 10/28/2023 concert (solid) versus the mean Dropkick Murphys profile (dashed) and The Avett Brothers profile (dotted). The signature matches the Dropkick template, but the signal is muted when the model relies chiefly on statistics such as Min$_i$, Max$_i$, μ_i, $\tilde{\mu}_i$.

Figure 11: Illustration of how summary statistics can mislead model classification when the true distribution shape is more informative. Dropkick Murphys and Avett Brothers show overlapping location descriptors despite nuanced and differing density shapes. Regularization shifts the focus of the Random Forest to shape.

Taken together, this work makes both theoretical and applied contributions. Theoretically, we advance the understanding of distribution parameter estimation from minimal data, derive tight classification-based consistency guarantees, and introduce a novel form of implicit regularization grounded in feature-selection probability dynamics. Practically, we offer scalable algorithms that can be readily deployed in live-market contexts where incomplete distributional information is common, and where model robustness and generalization are critical for operational success. These results offer potential value not only in secondary ticket markets, but also in broader domains where dynamic pricing, limited data access, and ensemble learning intersect, including financial markets, healthcare resource allocation, demand forecasting, and energy systems.

Keywords

Scaled Beta Distribution, Quantile Matching, Moment Matching, Random Forest, Implicit Regularization, Time Series Classification, Ticket Pricing, Ticket Resale, Feature Dilution, Constant-Value Features, Zero-Variance Features, Distribution Parameter Estimation, Ensemble Learning, SeatGeek

Acknowledgment

The author used publicly available event data accessed via the SeatGeek API (SeatGeek, Inc.) in accordance with SeatGeek's API Terms of Use. SeatGeek is not affiliated with this research and does not endorse it. All trademarks and content remain the property of their respective owners. Proper attribution is provided at `https://seatgeek.com` as required. Raw API data is not redistributed per licensing requirements. Derived statistical features and methodology code are available at [GitHub repo] for reproducibility purposes.

SEAT
GEEK

The author also used standard computational tools and programming libraries (including Python packages and the large language model ChatGPT-4 by OpenAI) for coding support, symbolic manipulation, and editorial refinement during the preparation of this manuscript. All outputs were reviewed by the author for accuracy and appropriateness.

References

[1] PR Newswire. Secondary tickets market size is set to grow by usd 132.1 billion from 2024-2028 – rising popularity of sports events to boost the revenue. November 2024. Accessed: February 11, 2025. URL: `https://www.prnewswire.com/news-releases/secondary-tickets-market-size-is-set-to-grow-by-usd-132-1-billion-from-2024-2028--rising-popularity-of-sports-events-to-boost-the-revenue--technavio-302315097.html`.

[2] SeatGeek. SeatGeek API Documentation. `https://platform.seatgeek.com`. Accessed: 2025-07-02.

[3] Learfield. Seatgeek partners with paciolan, the largest ticketing company in college athletics, February 2023. Accessed: February 11, 2025. URL: `https://www.learfield.com/2023/02/seatgeek-partners-with-paciolan-the-largest-ticketing-company-in-college-athletics/`.

[4] E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits [dataset], 1998. Accessed: February 11, 2025. URL: `https://doi.org/10.24432/C50P49`.

[5] Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, pages 359–370. AAAI Press, 1994.

[6] Thanawin Rakthanmanon, Bilson Campana, Abdullah Mueen, Gustavo Batista, Brandon Westover, Qiang Zhu, Jesin Zakaria, and Eamonn Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 262–270, New York, NY, USA, 2012. Association for Computing Machinery. `doi:10.1145/2339530.2339576`.

[7] Matthew Middlehurst, Patrick Schäfer, and Anthony Bagnall. Bake off redux: a review and experimental evaluation of recent time series classification algorithms. *Data Mining and Knowledge Discovery*, 38:1958–2031, 2024. `doi:10.1007/s10618-024-01022-1`.

[8] Lexiang Ye and Eamonn Keogh. Time series shapelets: a new primitive for data mining. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 947–956, New York, NY,

USA, 2009. Association for Computing Machinery. doi:10.1145/1557019.1557122.

[9] K. Krishnamoorthy. *Handbook of Statistical Distributions with Applications*. Chapman and Hall/CRC, 2nd edition, 2016. doi:10.1201/b19191.

[10] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. Time series feature extraction on basis of scalable hypothesis tests (tsfresh – a python package). *Neurocomputing*, 307:72–77, 2018. URL: https://www.sciencedirect.com/science/article/pii/S0925231218304843, doi:10.1016/j.neucom.2018.03.067.

[11] Carl H. Lubba, Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher, and Nick S. Jones. catch22: Canonical time-series characteristics: Selected through highly comparative time-series analysis. *Data Mining and Knowledge Discovery*, 33(6):1821–1852, November 2019. doi:10.1007/s10618-019-00647-x.

[12] Dinghuai Zhang, Ling Pan, Ricky T. Q. Chen, Aaron Courville, and Yoshua Bengio. Distributional GFlownets with quantile flows. *Transactions on Machine Learning Research*, 2024. Expert Certification. URL: https://openreview.net/forum?id=vFSsRYGpjW.

[13] Mingyuan Zhou, Tianqi Chen, Zhendong Wang, and Huangjie Zheng. Beta diffusion. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL: https://openreview.net/forum?id=zTSlm4nmlH.

[14] A. Dempster, D. F. Schmidt, and G. I. Webb. quant: a minimalist interval method for time series classification. *Data Mining and Knowledge Discovery*, 38:2377–2402, 2024. doi:10.1007/s10618-024-01036-9.

[15] Wei Wei, Tom De Schepper, and Kevin Mets. Dataset condensation with latent quantile matching. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7703–7712, 2024. doi:10.1109/CVPRW63382.2024.00766.

[16] Tim Salimans, Thomas Mensink, Jonathan Heek, and Emiel Hoogeboom. Multistep distillation of diffusion models via moment matching. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL: https://openreview.net/forum?id=C62d2nS3KO.

[17] Mingtian Zhang, Alex Hawkins-Hooker, Brooks Paige, and David Barber. Moment matching denoising gibbs sampling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL: https://openreview.net/forum?id=NWrN6cMG2x.

[18] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi:10.1109/18.61115.

[19] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. 01 1996. doi:10.1007/978-1-4612-0711-5.

[20] Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004. doi:10.1214/aos/1079120131.

[21] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. doi:10.1023/A:1018054314350.

[22] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998. doi:10.1109/34.709601.

[23] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, April 2006. doi:10.1007/s10994-006-6226-1.

[24] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi:10.1023/A:1010933404324.

[25] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, March 2002. doi:10.1162/153244302760200704.

[26] Abraham J. Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, 18(48):1–33, 2017. URL: http://www.jmlr.org/papers/volume18/15-240/15-240.pdf.

[27] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi:10.1080/01621459.2017.1319839.

[28] Lucas Mentch and Siyu Zhou. Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171):1–36, 2020. URL: http://jmlr.org/papers/v21/19-905.html.

[29] Lukas Cironis, Jan Palczewski, and Georgios Aivaliotis. Automatic model training under restrictive time constraints. *Statistics and Computing*, 33(1), 2022. doi:10.1007/s11222-022-10166-3.

[30] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi:10.1080/00401706.1970.10488634.

[31] Andrey N. Tikhonov. On the stability of inverse problems. *Proceedings of the USSR Academy of Sciences*, 39:195–198, 1943. URL: https://api.semanticscholar.org/CorpusID:202866372.

[32] A. De Magistris, V. De Simone, E. Romano, and G. Toraldo. Roughness regularization for functional data analysis with free knots spline estimation. *Statistics and Computing*, 34(5), 2024. doi:10.1007/s11222-024-10474-w.

[33] Oliver R. A. Dunbar, Nicholas H. Nelsen, and Maya Mutic. Hyperparameter optimization for randomized algorithms: a case study on random features. *Statistics and Computing*, 35(3), February 2025. doi:10.1007/s11222-025-10587-w.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[35] The Economic Times. Ed sheeran's mathematics tour ticket price for north america announced, October 2023. Accessed: April 6, 2025. URL: https://m.economictimes.com/news/international/us/ed-sheerans-mathematics-tour-ticket-price-for-north-america-announced/articleshow/94820106.cms.

[36] Hochman David. Here's how much beyoncé's cowboy carter tour will really cost you, May 2025. Accessed: June 18, 2025. URL: https://www.forbes.com/sites/davidhochman/2025/05/02/heres-how-much-beyonces-cowboy-carter-tour-will-really-cost-you/.

[37] Stevens Matt. How much does it cost to see beyoncé? it depends., May 2025. Accessed: July 2, 2025. URL: https://www.nytimes.com/2025/05/15/arts/music/beyonce-cowboy-carter-tour-ticket-prices.html.