Recency: Prediction with a Single Data Point

Florian M. Artinger, Gerd Gigerenzer, Nikita Kozodoi, Florian von Wangenheim*

Recency, that is, the time since an event occurred last, is commonly used to predict future events by experts. However, relying solely on recency and thereby ignoring frequency and other information is commonly regarded as a fallacy. We develop four conditions under which recency alone can predict future events at least as well as integrating all available information. In an extensive empirical study on predicting customer choice in 27 retail businesses and 31 further prediction problems including banking, crime, sports, and others, we analyze these conditions and show that they are central to the performance of the recency-based hiatus heuristic as compared with standard prediction models. We conclude that relying on recency can be ecologically rational.

Keywords: Bounded Rationality, Ecological Rationality, Heuristics, Prediction, Recency.
JEL: D01, D81, D84.

* Artinger: Simply Rational – The Decision Institute (florian.artinger@simplyrational.de); Gigerenzer: Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany (gigerenzer@mpib-berlin.mpg.de); Kozodoi: Amazon Web Services, Germany (this work was done prior to the author joining Amazon Web Services); Wangenheim: ETH Zurich, Department of Management, Technology and Economics, Weinbergstr. 56/58, 8092 Zürich, Switzerland (fwangenheim@ethz.ch).

1 Introduction

Consider a thought experiment by Nisbett and Ross from 1980 (1980, 15), which we have shortened and slightly edited here:

> *You wish to buy a new car. Today you choose between two alternatives: to purchase either a Volvo or a Saab. You use only one criterion for that choice, the car's life expectancy. You have information from reports that among many hundreds of Volvos and Saabs, only one Volvo broke down, whereas a dozen Saabs broke down. Just yesterday a neighbor told you that his new Volvo broke down. Which car do you buy?*

Following the principle of Bayesian updating, Nisbett and Ross argue that the neighbor's recent account changes the Volvo's record by only an iota and that you should buy the Volvo. The moral of their thought experiment is that rational reasoning entails updating the frequencies of events, and one should not be swayed by a single new case. In the case of buying a new car, they may well be right. Consider now a second thought experiment (Gigerenzer 2000, 263):

> *You visit a hotel in a jungle. Today you choose between two alternatives: to let your child swim in the river or let it climb trees instead. You use only one criterion for that choice, your child's life expectancy. You have information from reports that among many hundreds of children who visited, there was only one accident in the river, whereas a dozen children had accidents falling from trees. Just yesterday a neighbor told you that her child was eaten by a crocodile in the river. Where do you send your child?*

If applying the same reasoning as in the first thought experiment, parents should send the child to the river. The river's record has only changed from one to two incidents, so is still much safer than a dozen accidents by falling from trees. Yet many parents will not update past frequencies, and rely instead on the most recent information only. In this case, they may be right: a crocodile may now inhabit the river.

The two thought experiments illustrate the argument of this paper. Relying on recency only and ignoring past frequencies is not necessarily a bad decision. Frequency

and recency each have their own ecological rationality: frequency is the basis of rationality in a stable world where the future is like the past, whereas recency is the basis of rationality in a quickly changing world. In a stable world, the most recent event has no special status; in an unstable world, it does. Here, the past record may be obsolete.

This paper deals with the prediction of future events using recency. It makes three novel contributions. First, it provides a set of conditions under which relying on recency alone leads to equally or more accurate predictions than when relying on frequency and other valid predictors. These conditions were not known before. Second, we show empirically that a specific heuristic that relies on recency only – the hiatus heuristic – can predict customer purchases in 27 retail data sets with an accuracy that is on par with logistic regression, stochastic models capturing non-linear relations, and random forests, a machine learning technique. Third, we show empirically that the hiatus heuristic achieves similar levels of accuracy when predicting events in a total of 31 further data sets from domains such as banking, sports, crime, and others. Our key message is not that recency is always good, nor that frequency is always better, as suggested in earlier research. Rather, we show that there are identifiable conditions in which relying only on recency is ecologically rational.

We begin with a brief literature review of the role of recency and then develop four conditions under which relying on it can be a better predictor of future events than would integrating all available information, including recency, frequency, and other variables. This is followed by a description of the data used here. The empirical analysis follows.

2 Recency and Frequency

Research on recency dates back to the 19th century, when the "law of recency" was discovered (Brown 1838). This law states that memories of recent experiences come to mind more easily than memories from the distant past and are often the sole information that guides human decisions. Since this pioneering work, memory research has uncovered the "recency effect," the tendency to recall the last item in a series of words or numbers more readily than items in the middle Research on probability learning showed that when confronted with a nonstationary Bernoulli process, people quickly

detect substantial changes and adapt their estimates of the underlying Bernoulli parameter to the most recent trials rather than to the long run (Gallistel et al. 2014). People do not blindly apply recency but adapt their reliance on recency to the structure of the task, such as sequential dependencies between events (Jones and Winston R. Sieck 2003).

Recency is a feature of time. Yet neither the Kolmogorov axioms of probability nor Savage's axioms of rational choice incorporate time. Nor does Bayes' rule, as the thought experiment illustrates. Consequently, relying on recency when making predictions has been considered a fallacy in research that takes abstract choice axioms as a general standard of rationality. For instance, when predicting whether an event occurs, people reportedly rely on the ease with which an instance comes to mind, which has generally been considered an error and attributed to the availability heuristic (Tversky and Kahneman 1973). Other authors suggest that reliance on recency leads to maladaptive decisions. For instance, Kunreuther (1976) concluded that people overreact to the occurrence of natural disasters such as tornados and attributed this to reliance on recency. In areas hit by a natural disaster, the uptake for insurance policies against future damages spikes directly afterward (Slovic, Kunreuther, and White 1974). This holds even when records of the frequency of such disasters are publicly available, and it also holds for communities that have been hit more than once and where people actually have personal experience regarding the frequency of an event (Gallagher 2014). At the same time, there is evidence showing that if frequency is a reliable predictor and if one can learn from experience, then people do make choices that closely approximate expected value (Hertwig et al. 2004; Wulff, Mergenthaler-Canseco, and Hertwig 2017; Lejarraga and Hertwig 2021; but see Malmendier and Nagel 2016; for a review see Schulze and Hertwig 2021). A possible solution to this apparent contradiction is that under some conditions, frequency is a powerful predictor and under others it is recency.

Recency is used by many decision-makers across a broad spectrum of domains. Since the 1910s, marketing practitioners have been using the time since the most recent purchase as the sole indicator to predict whether a customer will buy from their company (Petrison, Blattberg, and Wang 1997). This practice remains in wide use today, despite the proliferation of data and modern prediction tools (Wübben and von

Wangenheim 2008). Reliance on recency has been observed in many other domains such as banking and investment (Malmendier and Nagel 2011; Haselhuhn et al. 2012; Agarwal et al. 2013), health (Davis 2004), crime (Gaissmaier and Gigerenzer 2012), and sports (Green and Zwiebel 2018). The general denominator of these phenomena is that people sometimes ignore the historical data except for the last data point. Such behavior seems to yield suboptimal results and therefore has frequently been interpreted as irrational. Yet Dosi et al. (2020) reported that in markets with unforeseeable shocks and other sources of uncertainty, a simple recency heuristic was superior to complex macroeconomic models in predicting demand. Given the widespread reliance on recency, the question is, do conditions exist where its use is rational in the sense that recency generates equally good or better out-of-sample predictions than do models that use additional information such as the frequency of events? These conditions have been unknown so far.

To derive these conditions, we first need to define the task and the rule for using recency. The task is to predict future events, such as whether a previous customer will make purchases in the next six months, or whether a sports team will win the cup again next year. In general, it is a prediction whether an event that happened in the past will happen again. The event is binary, such as purchase or no purchase, win or lose. To make this binary prediction, the hiatus heuristic can be used, which relies only on recency (Wübben and von Wangenheim 2008). We first formulate it for the concrete case of predicting customer purchases:

> *If a customer has purchased within the last* T *months, predict that the customer will make a purchase in the future. Otherwise, predict no purchase.*

Note that the heuristic relies on recency only – the time of the last purchase – and ignores information about how frequent purchases were in the past and other valid predictors that are available. It has a single free parameter, the hiatus T. Note that the time interval referred to here as "future" needs to be defined to measure the accuracy of both the heuristic and other prediction strategies.

The general version of the hiatus heuristic, which we will apply for predicting events in banking, sports, and other areas, can be formulated as follows:

If an event has occurred within the last T *time units, predict that it will occur again in the next* F *time units. Otherwise, predict that the event will not occur again in the next* F *time units.*

The hiatus heuristic is a special case of the class of one-reason heuristics (Gigerenzer and Gaissmaier 2011). It is used by both managers and frequent flyer programs to make binary predictions (Wübben and von Wangenheim 2008). Other uses of recency to make quantitative predictions include predicting the proportion of flu-related doctor visits (Katsikopoulos et al. 2022). Here, we deal with binary predictions only.

It is easy to imagine situations where a logistic regression that includes recency, frequency, and other variables will predict more accurately than the hiatus heuristic, which relies only on recency. In the next section, we pose the question, can we identify conditions under which the hiatus heuristic will predict as or more accurately than standard models of predictions, such as logistic regression?

3 Ecological Rationality of the Hiatus Heuristic

The study of the conditions under which heuristics perform well or fail is known as the study of ecological rationality. The term was used by Vernon Smith (2002) in his Nobel lecture as an alternative to constructivist or axiomatic rationality. Axiomatic rationality relates to consistency, while ecological rationality relates to success in attaining a goal, such as good prediction. A heuristic is ecologically rational to the degree that it is adapted to the structure of the environment (Gigerenzer, Todd, and the ABC Research Group 1999). The two thought experiments in the introduction provide a first intuitive answer to the question posed above: relying on the most recent event can be ecologically rational in environments prone to unexpected change, whereas relying on long-term frequency is rational in a stable environment, where probability updating is to be preferred.

In what follows, we present the bias-variance decomposition as a general framework for thinking about ecological rationality, and then look at specific conditions that decrease both the error due to bias and variance of the hiatus heuristic.

3.1 The Role of Bias and Variance in Prediction

The development of prediction models has been addressed particularly in the field of machine learning. The central question is, how one can infer a target variable y from a set of variables denoted as x given a sample S? The predictions of y are made out-of-sample. In other words, a prediction function $y = f(x)$ needs to be estimated based on a training sample $S = (x_1, y_1), \ldots, (x_N, y_N)$ that allows for determining y on some future observations of an attribute x. In contrast, many economic applications focus on parameter estimation, that is, producing good estimates of parameters β that govern the relationship between y and x, thereby seeking to uncover part of the data-generating process (Starmer 2005; Friedman et al. 2014; for an introduction to machine learning for an economic audience see Mullainathan and Spiess 2017). It is important to note that prediction algorithms are *not* built to uncover the underlying function of the data generating process but rather to maximize prediction accuracy. To illustrate, Mullainathan and Spiess (2017) analyze a data set from the American Housing Survey to predict house prices from 150 variables, such as the number of rooms, the base area, or the census region within the United States. From their total sample, they randomly select 10 training samples with about 5,000 units each and estimate a LASSO regression making predictions on a hold-out-sample.[1] Like most machine learning algorithms, LASSO uses regularization to limit overfitting. A perfect fit in-sample usually implies out-of-sample overfitting and a less-than-perfect prediction quality. Regularization in LASSO yields a sparse prediction function such that many of the coefficients are zero and are "not used" in the prediction. Mullainathan and Spiess (2017) find that less than half of the variables are used in any one of the training samples. Importantly, the variables that are used vary greatly between the different training samples. Nevertheless, the prediction quality remains constant across the different samples because the variables are commonly correlated with each other and act as substitutes in predicting house prices (but see in the housing market on the predictive power of a single variable Caplin et al. 2008). Such correlation also implies that often simple models using only a few or even one variable can perform as well as more complex

[1] LASSO uses a quadratic loss function, corresponds to a class of linear functions, and uses a regularizer that is the sum of the absolute values of the coefficients.

models that use a large set of variables (Rudin 2019; Semenova, Rudin, and Parr 2019; Rudin et al. 2022).

The results by Mullainathan and Spiess (2017) illustrate the typical methodology of machine learning. A data set is divided into a training sample and a hold-out sample, the parameters of the various models are fitted to the training sample, and the models with the fixed parameters are then tested on the hold-out sample. This procedure is repeated multiple times with different partitions to obtain reliable estimates of the out-of-sample accuracy of the competing models.

The error in the out-of-sample prediction consists of three components, bias, variance, and irreducible noise. Bias refers to the degree a model is systematically "wrong," deviating from the underlying function. Reducing bias is the focus of parameter fitting. However, there is also a second source that systematically affects prediction error, variance. An unbiased model can lead to large prediction error if it has too many free parameters estimated from small training samples, which together can increase the error due to variance. This relationship is studied in terms of the bias-variance decomposition of the prediction error (Geman, Bienenstock, and Doursat 1992).

In machine learning, the goal is typically to minimize the error on the unseen hold-out sample. To stress the dependence of the estimated prediction function *f* on the training sample, we write $f(x;S)$. One of the frequently used measures of the effectiveness of the estimated prediction function $f(x;S)$ for binary outcomes is the Brier score or mean squared error (MSE) between the binary target variable and the predicted probabilities. In the following, we use MSE to illustrate the role of bias and variance because it can easily be decomposed into the corresponding components (MSE; Brier 1950; Domingos 2000; Manning, Raghavan, and Schütze 2009):

$$(1) \qquad MSE = E_S[(f(x;S) - P(y|x))^2]$$

where E_S is the expectation with respect to the training sample S, that is, the mean of the ensemble of training samples *S*, and $P(y|x)$ is the true conditional probability from the data generating function. Given some training sample *S*, the estimated prediction function $f(x;S)$ can yield a good approximation of $P(y|x)$ and predict future events

well, although an estimated prediction function $f(x;S)$ may also vary substantially depending on the specific training sample S. Note that calculating the MSE requires the prediction function $f(x;S)$ to produce class probabilities.

To understand what drives the prediction error, consider its decomposition into bias, variance, and irreducible noise ε: [2]

$$(2) \qquad E_S\left[\left(f(x;S) - P(y|x)\right)^2\right] = bias + variance + \varepsilon$$
$$bias = (P(y|x) - E_S[f(x;S)])^2$$
$$variance = E_S[(f(x;S) - E_S[f(x;S)])^2]$$

If predictions of $f(x;S)$ differ on average from $P(y|x)$, then the estimated prediction function $f(x;S)$ is biased, as captured by the first term. However, an unbiased estimator may still have a large error due to the variance component, as captured by the second term. Even if $E_S[f(x;S)] = P[y|x]$, the estimated prediction function $f(x;S)$ can be highly sensitive to the training sample S and far removed from the generating function $P(y|x)$.

Error due to bias reflects the inability of a model to capture the underlying function. This can be the case, for instance, when the model is linear but the underlying function is quadratic. Error due to variance reflects the effect of the variation inherent in different samples on the estimated parameters. The effect of error due to variance has been commonly neglected when evaluating heuristic strategies (Brighton and Gigerenzer 2015). However, the variance component of the prediction error is a critical feature. For instance, Van Der Putten and Van Someren (2004) conduct a bias-variance analysis of eight models for predicting the sales of insurance policies: the differences between the models' performance overwhelmingly arise from error due to variance. They find the best performer to be the naïve Bayes classifier, a heuristic strategy that makes the "wrong" assumption that variables are conditionally independent. The naïve Bayes classifier tends to have a higher error due to bias but achieves lower overall prediction error due to low error from variance (Domingos and Pazzani 1997; Hand and

[2] In expectation, irreducible noise ε has a mean of zero. Because it bears relevance for empirically testing, we list it here.

Yu 2001). Naïve Bayes has been ranked in the top ten of the most widely used models in data mining (Wu et al. 2008).

In the next sections, we apply the bias-variance framework to the hiatus heuristic and define conditions that allow the heuristic to reduce error due to bias and variance: *attribute dominance*, *predictive validity*, *satisficing*, and *unpredictable change over time*. These environmental conditions define the ecological rationality of the hiatus heuristic, as well as that of similar one-reason heuristics.

3.2 Attribute dominance

The attribute dominance condition involves reducing bias. The general intuition about the role of the attribute dominance condition can be expressed in the framework of a linear model: if the weight of recency alone is equal to or larger than the sum of the other variables' additional contributions, then any linear model will lead to the same binary prediction as that made by the hiatus heuristic. This means that both models have the same error from bias. In that case, prediction implicitly depends only on recency and cannot be changed by the values of other attributes. To evaluate attribute dominance, we use three different measures. These allow for a robust evaluation of attribute importance from different angles.

3.2.1 Weight-Based Dominance

Linear models, particularly in the form of logistic regressions, are standard approaches to the prediction of binary outcomes. Assume that the probability of a positive outcome of entity i is distributed according to the sigmoid function, which is a standard assumption in the logistic regression framework:

$$(3) \qquad p(y_i = z_i) = \frac{1}{1+e^{-z_i}}$$

where z_i is the latent variable, which is a linear function of multiple entity attributes $x_1, \ldots, x_m$:

$$(4) \qquad z_i = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m$$

where $\beta_1, \ldots, \beta_m$ are the estimated attribute weights. Plugging in the probability distribution, one can formulate the prediction rule as:

$$(5) \qquad \hat{y}_i = \begin{cases} 1, & \text{if} \quad \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m > 0 \\ 0, & \text{if} \quad \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_m x_m < 0 \end{cases}$$

The prediction is determined by the sign of the latent variable z. Assume that attributes $x_1, \ldots, x_m$ are binary, with two possible values: -1 and 1. This can be achieved by scaling the continuous variables relative to their median value. Without loss of generality, we can now sort the attributes in the order of descending estimated absolute weights. *Weight-based dominance* is given if:

$$(6) \qquad \left|\hat{\beta}_1\right| \geq \sum_{j=2}^{m} \left|\hat{\beta}_j\right|$$

If this relationship holds, the sign of z is fully determined by the value of β_1. The weighted sum of attributes $|\beta_2 x_2 + \cdots + \beta_m x_m|$ cannot exceed $|\beta_1 x_1|$ and change the sign of z given any possible combination of the attribute values. This implies that attribute x_1 dominates all other attributes (Artinger et al. 2018). If recency is this dominant attribute, $x_1 = x_r$, then no sum of other variables can generate a better prediction. Not only does this allow the hiatus heuristic to reduce error from bias relative to linear models, but adding any further variables potentially incurs error due to variance that can lead to a performance loss of more complex models.

Note that the weight-based dominance condition formulated in equation (6) is conservative. It implies that there exists no combination of attribute values that allows overriding the prediction based on x_1. In practice, certain attribute value combinations that violate the weight-based condition may be observed rarely or be completely absent from the considered sample, implying that x_1 may still be considered as dominant. To overcome this potential limitation and avoid using the binary encoding of the attributes, we introduce two further measures of dominance.

3.2.2 Residual-Based Dominance

Running a regression with all variables combined will not illuminate the full power of recency. This is because the weight of recency usually shrinks when more variables are added to a regression. But what is the explanatory power of recency in isolation, and can other variables add any explanatory power above and beyond that already captured by recency? An answer to this question can be estimated by comparing the variance of the target variable explained by a certain attribute and the residual variance explained by all other attributes. We formalize this concept using a residual-based dominance metric.

The derivation of the residual-based dominance includes two steps. First, we estimate the percentage of the variance of the target variable explained by the potentially dominant attribute. We start out by using only recency, denoted as x_1, and fit a simple linear regression equation. The R-squared of this regression, denoted as R_1, indicates the percentage of the target variance explained by x_1.

$$(7) \qquad y_i = \hat{\alpha}_0 + \hat{\alpha}_1 x_1$$

Second, we calculate the remaining percentage of the target variance explained by all further attributes $x_2, \dots, x_m$. To estimate this figure, we calculate the residuals from the first-stage model to obtain the remaining part of the target not explained by x_1. We denote this residual as $r_i = y_i - \hat{y}_i$. Next, we regress the residual on the remaining attributes:

$$(8) \qquad r_i = \hat{\delta}_0 + \hat{\delta}_2 x_2 + \cdots + \hat{\delta}_m x_m$$

where $\delta_1, \dots, \delta_m$ are the estimated attribute weights. The R-squared of the second-stage regression denoted as $R_{2,\dots,m}$ indicates the percentage of the residual variance explained by $x_2, \dots, x_m$. Multiplying this value by $1 - R_1$ yields the percentage of the variance of the initial target variable explained by the other variables after accounting for recency.

We formalize *residual-based dominance* as follows:

$$(9) \qquad R_1 \geq R_{2,\dots,m} \, (1 - R_1)$$

If this relationship holds, the attribute x_1 explains a higher percentage of the target variable variance compared with $x_2, \dots, x_m$, and we consider x_1 to be a dominant attribute.

Note that residual-based dominance uses linear regression rather than logistic regression on the first stage for predicting the binary events. There are two reasons for this. First, using linear regression estimated with the Ordinary Least Squares procedure allows us to compute the R-squared coefficient that can directly be compared with the R-squared of the second-stage model. Second, we require continuous predictions of the linear model to acquire the residuals for the second-stage model.

3.2.3 Agreement-Based Dominance

Weight-based dominance and residual-based dominance use a linear framework. However, in environments that are governed by non-linear relationships, a decision-maker should employ other prediction strategies such as random forests aiming to identify the best prediction method. When using such models, calculating the contribution of each individual attribute is non-trivial and differs substantially from the linear model framework. It implies that we cannot rely on the above formalization of weight-based dominance, which motivates us to consider a third measure for capturing whether dominance is given in a certain environment. We refer to this measure as agreement-based dominance.

Instead of estimating the weight of recency, agreement-based dominance focuses on the predictions of future events produced by a given prediction method. Consider two variants of the same algorithm estimated from the training sample: (i) a model that uses a single attribute x_r, recency, to predict the future; (ii) a model that uses all available variables $x_1, \dots, x_m$, including recency. Let us compare binary predictions of the two variants for the hold-out sample denoted as $\hat{y}^{x_r}$ and $\hat{y}^{all}$. To compute agreement-based dominance, we simply calculate the number of entities for which $\hat{y}_i^{x_r} = \hat{y}_i^{all}$, where $i \in (1, \dots, K)$, K is the number of observations in the hold-out sample. The resulting measure lies between 0 and 1. Higher values of agreement-based dominance indicate that using all variables changes only a smaller number of model predictions, indicating that the relative importance of recency in the considered

environment is higher. This provides us with a model-agnostic and more refined measure of dominance in an environment.

In our empirical study, we compute all three variants of the dominance measure. If recency is a dominant variable, we expect that the relative performance of the heuristic improves in comparison to that of the more complex models.

3.3 Predictive Validity

Dominance indicates the relative predictive power of recency in comparison to other variables. Validity is a measure of the absolute predictive power of recency in an environment irrespective of other attributes available complementing the dominance analyses. High validity of recency allows achieving a smaller error due to bias.

Specifically, validity of an attribute can be defined as the correlation between the attribute and the target variable (Şimşek 2013; Martignon and Hoffrage 2002). We measure the validity of recency with Pearson's correlation between x_r and y. In our setup, recency of the last event is computed as the number of time units that have passed since the last occurred event in the training sample, whereas y is a binary variable indicating whether an event has occurred at least once in the hold-out sample. The higher the correlation between x_r and y, the larger the predictive power of recency. As such, the validity of recency potentially captures other elements that are not captured by the dominance measures we introduced above.

Measuring the validity of recency in our empirical study, one finds that a negative relationship is more widespread (i.e., an event is more likely to occur again if its recency is low). To account for both possible directions of the recency-target relationship and simplify the interpretation of results, we measure the predictive validity of recency by simply taking the absolute value of the Pearson correlation coefficient.

We expect to see a positive correlation between the performance of the hiatus heuristic and the predictive validity of recency.

3.4 Satisficing

Error due to variance results from the instability of the parameter estimates due to differences between samples. This error can be reduced by increasing the size of the training sample, which, however, is not always possible. It can be more efficiently

reduced by reducing the number of free parameters to be estimated in the first place. If a model has no free parameters, error due to variance is zero, and increasing the sample size is obsolete. Introducing fixed parameters instead of estimating "optimal" parameters from a training sample is a method in case. Simon (1955) refers to such a procedure as satisficing, in contrast to optimization (for a review see Artinger, Gigerenzer, and Jacobs 2022). Note that in the framework of out-of-sample prediction, optimization always refers to the training sample. In fact, the parameters with the optimal fit may not lead to the best out-of-sample predictions because "optimal" refers to bias only and neglects variance. Hence, optimizing the fit to a sample may lead to overfitting; this is why machine learning employs methods such as regularization and cross-validation.

Ignoring this insight, much of the literature on decision-making assumes that satisficing is equivalent to suboptimal results and that reducing biases to zero suffices (Tversky and Kahneman 1974; DellaVigna 2009). Satisficing, however, is a central property of heuristics that can constitute effective solutions to complex and dynamic problems (e.g., Caplin, Dean, and Martin 2011; Caplin and Dean 2011; Gabaix and Laibson 2006; Lant 1992; Manzini and Mariotti 2007; 2012; Artinger and Gigerenzer 2016). Moreover, satisficing is frequently employed by managers and other experts (Artinger et al. 2015).

The hiatus heuristic has one parameter: the duration of the hiatus, a threshold or aspiration level determining how many past time units T to consider. Although one can determine the optimal duration of the hiatus from the training sample and use this for prediction, that, as mentioned before, likely leads to overfitting. The satisficing alternative is to use a fixed hiatus instead. For instance, managers in retail companies reported using a 9-month hiatus (Wübben and von Wangenheim 2008), which can be used as a fixed threshold to test the heuristic. Alternatively, one can use a simple satisficing threshold that generalizes to other problems. The choice of the hiatus influences the bias. Whatever the method, a satisficing hiatus implies that error due to variance is zero for the hiatus heuristic, seeing as there is no parameter to estimate. This zero variance can provide an advantage in terms of the predictive accuracy of the hiatus heuristic in comparison to more complex models and even in comparison to an optimized version of the hiatus heuristic, which estimates the optimal duration of the

hiatus from the training sample. But blindly applying satisficing can also yield bad performance. For instance, if only a few time units are available as observations, using T smaller than the length of the training sample as the hiatus duration may be insufficient to generate sensible predictions.

Statistical models also entail the possibility of reducing error from variance by introducing systematic bias. The machine learning literature suggests that this can be vital for good performance (Geman, Bienenstock, and Doursat 1992). For instance, given the prediction of a binary outcome, such as whether an event will occur, statistical models provide a probability estimate. Given such a probability, one needs to determine a threshold above which the model predicts an event to occur. Alternatively, one may simply use $\tau = 0.5$ as a satisficing threshold, which may yield good performance by reducing error due to variance. Note that the standard practice in classification tasks is to start with default (i.e., satisficing) probability thresholds and then further improve the performance by switching to optimized thresholds (Freeman and Moisen 2008; Zou et al. 2016).

Geman et al. (1992), who introduce the bias-variance trade-off in machine learning, assert that "learning complex tasks is essentially impossible without the a priori introduction of carefully designed biases into the machine's architecture. (…) Despite a long preoccupation with learning per se, the identification and exploitation of the "right" biases are the more fundamental and difficult research issues." (Geman, Bienenstock, and Doursat 1992, 3). The effect of satisficing can be assessed by comparing a model's performance with a satisficing and an optimized threshold.

We expect that using a satisficing threshold to define what is considered a recent event need not be detrimental to performance in comparison to an optimally derived threshold and, at times, can be even advantageous, seeing as this strategy protects from error due to variance. We also evaluate the performance of the more complex models using a fixed threshold above which a model predicts an event to occur. Again, we expect that using such a fixed threshold need not be detrimental to the performance in comparison to an optimally derived threshold.

3.5 Unpredictable Changes over Time

The bias-variance dilemma is formulated for a situation of repeated drawings of random samples from a known population. The population from which the samples are drawn is assumed to be stable and not change over time. Over time, however, an environment might change in a way that is not captured in the learning sample, which can increase error from variance. For instance, Google Flu Trends' big data algorithm predicted the flu well in static out-of-sample tests but failed to predict the spread of flu over time; here, it was outperformed by a recency heuristic that better predicted the flu over a period of eight years by using a single data point (Katsikopoulos et al. 2022). Similarly, deep neural networks learned to make accurate predictions from chest X-rays in out-of-sample tests as to whether patients from one hospital were at high risk of pneumonia but failed in other hospitals (Zech et al. 2018). These failures illustrate the importance of testing models in environments in which the data distribution in the historical data differs from the distribution observed in the data to which the model is applied. The accuracy of a model may deteriorate when making predictions for an uncertain future because of the data set shift. Heuristics are meant to work under uncertainty and change, such as when making predictions about the future. As illustrated by the thought experiment in the introduction, relying on recency can be successful if the environment changes unexpectedly.

When making predictions about an unknown future that may be systematically different from the present, the conditions for the bias-variance decomposition may not be in place. Nonetheless, the conclusions in the previous sections still hold: one can reduce the variance of the prediction error to zero by introducing a satisficing threshold and reduce error due to bias by using a dominant attribute. An additional source of error will be introduced if the future systematically differs from the past. Such a change will likely increase the total prediction error. Specifically, models that fine-tune their free parameters on the training sample, that is, on the past, will likely incur additional error due to overfitting the past, that is, they will be more exposed to error due to variance than will simpler models.

Changes in the data between the training sample S and the hold-out sample H expose models whose parameters need to be estimated from S to error from variance. Such changes can, for instance, result from a non-stationary data-generating process or

too few time units in the training data such that it does not reveal all systematic changes over time. To train models in machine learning, one conventionally uses *k*-fold cross-validation. Within this framework, a data set with observations of N entities is partitioned into k folds of subsamples, where one of the folds is iteratively used as a hold-out sample to evaluate the performance of a given prediction model while training it on the remaining $k - 1$ folds.

In our empirical study, we assess the potential effect of changes over time by comparing traditional cross-validation with time-based validation (see Figure 1). In the time-based validation, we train the models on data from the starting time interval $\left[T = 0, T = \frac{1}{2}\right)$ and evaluate their performance on the following time interval $\left[T = \frac{1}{2}, T = 1\right)$. This ensures that data from future time units are not included at the training stage. In contrast, the regular cross-validation randomly splits the data into folds, ignoring the time dimension.

Quantifying the difference between the performance of different models in the time-based validation and regular cross-validation allows measuring the degree to which a model is susceptible to changes over time. At the same time, comparing the performance gaps across the data sets indicates which data set exhibits more substantial changes over time, which is likely to influence the accuracy of all prediction methods. Given the fact that the heuristic is less exposed to error from variance, we expect the relative performance of the hiatus heuristic in comparison to that of the more complex models to improve with stronger differences between time-based and cross-sectional validation.

Note that using the performance gap between the time-based validation and regular cross-validation allows us to quantify the impact of changes over time with respect to the specific time point $T = \frac{1}{2}$, which serves as a prediction point in our experiments. This gives this measure an advantage over other alternative statistics, such as clumpiness or autocorrelation, that do not focus on a specific time point.

4 Data and Methods

4.1 Data

Marketing executives are faced with the challenge of how to distinguish current customers who are likely to buy in the future from those who will not. The answer to this question helps in identifying profitable customers who should be targeted with marketing activities, such as mailings and catalogs, and in removing unprofitable customers from the customer base. The hiatus heuristic can be used to identify such active customers. Our analysis extends the previous study of Wübben and Wangenheim (2008), who compare the hiatus heuristic and the domain-specific Pareto/NBD and BG/NBD models on three customer purchase data sets. We collect 27 sets of publicly available customer purchase data from different retail companies and extend the number of models by also comparing the heuristic to logistic regression and random forest (Breiman 2001), a widely used forecasting model in machine learning.

Moving beyond the marketing context, we collect 31 further data sets that come from a range of other domains where decision-makers routinely predict future events, such as weather, sports, and crime. The data acquisition process aims at gathering a large and heterogeneous data library with relevant prediction tasks from different environments.

The data sets provide the context within which we empirically analyze the conditions laid out above, that is, dominant attributes, satisficing, and unpredictable changes over time.

To gather the data, we surveyed various data sources with publicly available data, including the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.php), data science competition platforms such as Kaggle (https://www.kaggle.com/datasets) and Data Mining Cup (https://www.data-mining-cup.com), and sociological surveys such as the Russian Longitudinal Monitoring Survey (https://rlms-hse.cpc.unc.edu). Acknowledging that gathering a representative data library is difficult, we develop an interactive website that allows a user to upload a custom data set and check whether the results for that set align with the results obtained for our data library. The web application is available at https://hiatus.shinyapps.io/shinyHiatus/. We believe that the constructed data library

and the accompanying website help demonstrate the external validity of our empirical findings.

In non-retail data sets, the challenge is the same: whether an event that happened in the past will occur again in the future, that is, within a specified time interval. The nature of the event to be predicted differs across domains. For instance, an event in health data sets may refer to whether a patient will be hospitalized again or recover from a particular disease; in sports data sets, an event may indicate whether a sports team will win the cup again; in weather data, an event may refer to whether a location will be hit again by a severe weather condition. The heterogeneous data library facilitates analyzing the role of recency across very different environments. Table 1 provides an overview of the used data sets, indicating the data domain and examples of the observed event.

Our data library consists of 58 natural data sets. Each data set contains information on a set of N individual entities $i \in (1, \dots, N)$ with the sequences of binary events observed for time T time units. The time units are measured on either a weekly, monthly, or yearly basis, depending on the domain. For each time unit $t \in T$, we observe whether an event $y \in \{0,1\}$ occurred. In retail data, the largest group, an entity is a customer with the observed purchased history; in weather data, an entity is a specific location with binary outcomes describing whether it was hit by a tornado or experienced a severe drought.

Table 1. Data Sets Overview

Environment	Example of the observed event	No. data sets
Retail	Customer makes purchase	27
Banking and investment	Client makes money transfer	6
Crime	City is hit by a terrorist attack	4
Sports	Basketball team wins NBA league	3
Corporate	Company purchases office supplies	3
Weather	Region experiences a severe drought	2
Other	Paper of scientist accepted at conference	13

The data sets vary in terms of the statistical properties, which are summarized in Table 2. An average data set contains 75,129 events observed for 2,757 entities over 57 time units. To ensure that we have enough data for model training, we use only those entities in a data set for which at least one year of historical data is available. The

number of observations ranges from 342 to above 1.3 million. The share of repeated events fluctuates between 1% and 99%. Note that the lower the share of repeated events, the rarer the observation of an event.

Each data set contains between three and ten variables that can be used for prediction, three of which are contained in all data sets: recency, frequency, and clumpiness. Recency refers to the time since the last event occurrence, frequency measures the total number of times an event has occurred, and clumpiness indicates how evenly spaced the temporal distance between events is (see Appendix C for more details). A number of studies show that accounting for clumpiness can improve the predictive power of models (Platzer and Reutterer 2016; Zhang, Bradlow, and Small 2015) and that it is a common feature in many decision environments (Anderson and Schooler 1991; Gabaix 2016; 1999). The other variables used for prediction differ between data sets. For instance, customer purchase data sets frequently contain purchase-specific data such as the total purchase amount and the number of bought items per customer.

Tables A1 and A2 in Appendix A provide a more detailed description of each of the data sets, including the list of variables used for prediction, the nature of the observed event type and the data dimensionality.

Table 2. Summary Statistics Across Data Sets

Characteristic	Mean	St. Dev.	Min	Max
Number of events	75,129	190,476	79	1,320,372
Number of entities	2,757	4,007	18	16,760
Number of time units	57	35	8	156
Number of variables	5	2	3	10
Share of repeated events	59%	23%	1%	99%

To make predictions about the future, we use the data partitioning scheme depicted in Figure 1. Each data set is split into two equal consecutive time intervals. We train the models on the training sample S that contains data from the first interval $\left[T = 0, T = \frac{1}{2}\right)$ and evaluate their performance on the hold-out sample H coming from the interval $\left[T = \frac{1}{2}, T = 1\right)$. The hold-out sample contains future observations of x and y that are unknown to the prediction function at the time that its parameters are

estimated. Testing models in terms of future occurrences is equivalent to the task that people solve when facing the uncertainties of an unknown future.

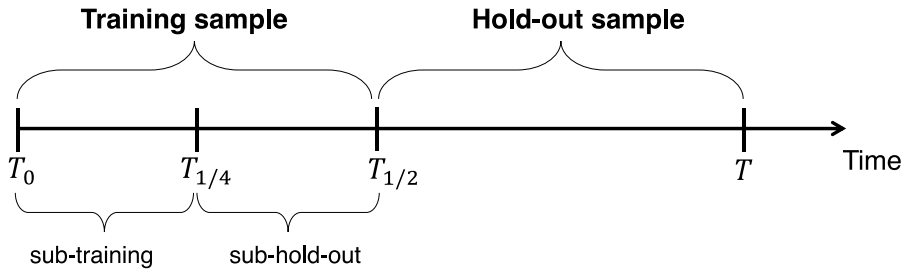

Figure 1. Data Partitioning Scheme

Note. The figure depicts the data partitioning scheme used for each data set in the study. The total time period is divided into two equal intervals: a training sample and a hold-out sample. We observe a set of entities (e.g., customers) during the training sample, tracking the occurrence of events and calculating the entity attributes such as recency, frequency, and clumpiness. The prediction methods are fitted to the training sample. Next, we use fitted models to predict the future event occurrence for each entity in the hold-out sample. To find the optimal thresholds, we further split the training sample into a sub-training and a sub-validation sample, where we tune the threshold before applying it to the model.

4.2 Prediction Strategies

We compare the predictive accuracy of the hiatus heuristic to multiple domain-specific and domain-general prediction algorithms. This includes logistic regression, random forest, and non-linear stochastic models designed specifically for predicting customer behavior. The models were selected to include a broad spectrum of strategies for evaluating the performance of the hiatus heuristic. In total, the empirical study uses prediction methods from the four families of models listed below.

(1) **Hiatus heuristic.** The hiatus heuristic predicts that an event will occur if the time since the event occurred last, that is, the recency of an event, is below a given time threshold. The heuristic has one parameter: a time threshold *t*. We apply a satisficing threshold of $t = \frac{1}{2} \times k$, where k is the length of the observed time interval. For a training sample of $T = \frac{1}{2}$, this yields a satisficing threshold of $\frac{1}{4}$ of the time units as depicted in Figure 1. In addition, we estimate an optimal threshold from the training sample.

(2) **Logistic regression** (denoted as LR). Logistic regression is a commonly used linear model in the context of a binary dependent variable and is frequently applied in economics, marketing, and medicine, among others. We use non-regularized logistic regression, which can be a further indication of overfitting in a given dataset.

(3) **Random forest** (RF). Random forest is an ensemble method based on a set of decision trees (Breiman 2001). In contrast to logistic regression, RF is able to capture complex, non-linear relationships in the data to improve prediction accuracy (Mullainathan and Spiess 2017). Furthermore, RF does not require strict assumptions with respect to the data structure, which makes it a universal and widely used method applicable across different areas.

(4) **Pareto/NBD** (P/NBD). Both logistic regression and random forest are general-purpose models. In the marketing domain, which provides the most data sets for our study, practitioners widely use domain-specific stochastic models to predict customer behavior. One such established model is Pareto/NBD, which can capture non-linear relationships (Fader, Hardie, and Lee 2005; Schmittlein, Morrison, and Colombo 1987). Although Pareto/NBD is designed for marketing settings, it can also be applied to predict future events in other data sets.

In contrast to the hiatus heuristic, which produces binary predictions of future event occurrence, logistic regression, random forest, and Pareto/NBD estimate the probability of occurrence. To derive a binary statement of whether an event will occur, one requires a probability threshold. Similar to the hiatus heuristic, we use two types of thresholds: satisficing and optimizing. As a satisficing threshold, we use $\tau = 0.5$. The optimized threshold is derived from the training sample, as it is for the hiatus heuristic.

Logistic regression and random forest models can incorporate an arbitrary number of variables describing an entity to predict future event occurrence. Keeping the model family constant and varying the input variables enables a direct comparison of the predictive power of recency and other variables. Therefore, we employ the following three variants of logistic regression and random forest:

(1) **Model (R).** The model uses only recency.

(2) **Model (F).** The model uses only frequency.

(3) **Model (All).** The model uses all available variables of a data set.

The performance of each prediction method is assessed in terms of the balanced accuracy (BACC), which is a metric commonly used in classification tasks (Fawcett 2006). It takes into account the correctly classified repeated events, that is, the total true positives *TP* given all positive events *P*, and correctly classified non-repeated events, that is, the true negatives *TN* given all negative events *N*:

$$BACC = \frac{\frac{TP}{P} + \frac{TN}{N}}{2} \tag{10}$$

BACC accommodates type I (false-positive) and type II (false-negative) errors and is not dependent on the share of repeated events in a data set, which fosters fairness when comparing performance across balanced and imbalanced environments. An imbalanced environment is one in which the frequency that an event occurs differs from the frequency that it does not occur. Furthermore, unlike some other established performance metrics such as MSE, BACC does not require predicted class probabilities and operates with binary class predictions. This makes BACC a more suitable metric for evaluating the performance of the hiatus heuristic.

5 Results

This section presents the empirical results. In order to facilitate a good understanding of process that leads to the overall results, we start by illustrating the predictive performance of the hiatus heuristic and other prediction methods on two selected data sets from the data library. These two data sets represent a case where recency and the hiatus predict well and an instance where this is not the case. Next, we compare the aggregate predictive performance across all data sets, distinguishing two environments: customer purchase data and other domains. Last, we focus on the environmental properties of the data sets, investigating the impact of different conditions on the predictive performance.

5.1 Individual Examples

Table 3 compares the performance of the hiatus heuristic and statical models that use all available information on two data sets coming from different domains and provides information about the conditions introduced in Section 3. The aggregated results for all

variants of the considered prediction models across the 58 data sets are provided in further sections.

The first data set *CDnow* represents the retail domain, where the hiatus heuristic was initially tested (Wübben and von Wangenheim 2008). The data set contains customer purchase data from the online shopping company CDnow, Inc and covers purchases of compact discs and music-related products made by 2,357 customers within 78 weeks in 1997 – 1998. The prediction task considered in this data set is to predict returning customers. That is, given the purchase data from the first 39 weeks, the task is to predict which customers will make at least one additional purchase in the following 39 weeks. The observed data contains five variables describing the customer behavior in the first 39 weeks: recency of the last purchase, purchase frequency, clumpiness of each customer, average order amount and the number of ordered items.

The second data set *Comcast* reports complaints by Comcast customers, a US-based telecommunications company. The data covers 1,477 complaints made by 630 Comcast customers in 2015. Here, the prediction task is to predict which customers will submit at least one service complaint in the last 26 weeks of 2015, given the complaints data from the first 26 weeks in 2015. The variables used for prediction are customer-level recency, frequency, clumpiness, as well as the average rating of the previous reviews of the customer.

As shown in Table 3, the recency-based hiatus heuristic demonstrates a good performance on *CDnow*. With a balanced accuracy of 70.62, the heuristic with the satisficing threshold outperforms LR (All) and RF (All) that use five variables for prediction. It is interesting to note that using an optimized threshold harms the heuristic's accuracy, which may indicate that optimization introduces too much variance on this data set. The good performance of the heuristic can be attributed to the strong ecological rationality of recency in this environment: the residual-based dominance condition is fulfilled, 91% of the recency-based predictions are not changed by introducing further variables into a model, and the correlation between recency and the target variable is 0.48. Moreover, the data set demonstrates strong changes over time: the difference in the performance of LR (All) between the time-based partitioning and regular cross-validation is 5 percentage points.

Table 3. Two Example Prediction Tasks

Results group	Characteristic	CDnow	Comcast
Balanced accuracy	Hiatus (satisficing threshold)	70.62	61.55
	Hiatus (optimized threshold)	66.10	58.11
	Logistic regression (LR All)	69.76	68.85
	Random Forest (RF All)	70.48	66.46
Data characteristics	Weight-based dominance	No	No
	Residual-based dominance	Yes	No
	Agreement-based dominance	0.91	0.69
	Recency validity	0.48	0.20
	Satisficing index	Yes	Yes
	Changes over time	5	1
Sample properties	Number of events	6,919	1,477
	Number of entities	2,357	630
	Number of time units	78	52
	Number of variables	5	4

In contrast, the heuristic performs relatively poorly on *Comcast*: with a balanced accuracy of 61.50, it falls behind both LR (All) and RF (All) that rely on four variables for prediction. This corresponds to the lower ecological rationality of recency in this environment, which is reflected in the absence of both weight-based and residual-based dominance, and lower values of the agreement-based dominance and the validity of recency at 0.69 and 0.20, correspondingly. In addition, *Comcast* exhibits more stability over time, indicated by a substantially smaller performance gap of 1 percentage point between the time-based and regular cross-validation.

The two individual examples show the importance of understanding the conditions in which the heuristic is ecologically rationality, that is, performs well. In the next section, we look at the aggregated predictive performance across 58 retail and non-retail data sets in the data library, and then move on to the ecological analysis of the heuristic.

5.2 Predicting Customer Purchases

Figure 2 compares the predictive performance of eight prediction methods for 27 retail data sets. Panel (a) depicts the mean BACC for each prediction method with both threshold types: optimizing and satisficing.

The recency-based hiatus heuristic achieves a mean BACC of 68% with a satisficing threshold and 69% with an optimized threshold. Generally, most models show a very similar performance regardless of whether they rely only on recency, frequency or integrate recency, frequency, and all other variables or whether they use a satisficing threshold or an optimized threshold. The error bars in Figure 2 indicate that there is no statistically significant difference between the performance of the recency-based prediction methods and methods that are more complex and require substantially more information, with the exception of P/NBD, which, when it relies on a satisficing threshold, demonstrates a substantially lower BACC.

This result suggests that when predicting customer purchases, much of the information relevant for prediction is already included in recency. A satisficing threshold does not need to be a disadvantage in comparison to an optimized threshold. Using the hiatus heuristic can be the preferred decision strategy for predicting future customer purchases, especially when further consumer attributes may be difficult or costly to collect or if transparency is important. Note that for random forests and logistic regression, using additional variables besides recency leads to only small and non-significant gains in predictive accuracy. It is of note that P/NBD, which was developed originally for making predictions regarding purchasing behavior, does not perform very well.

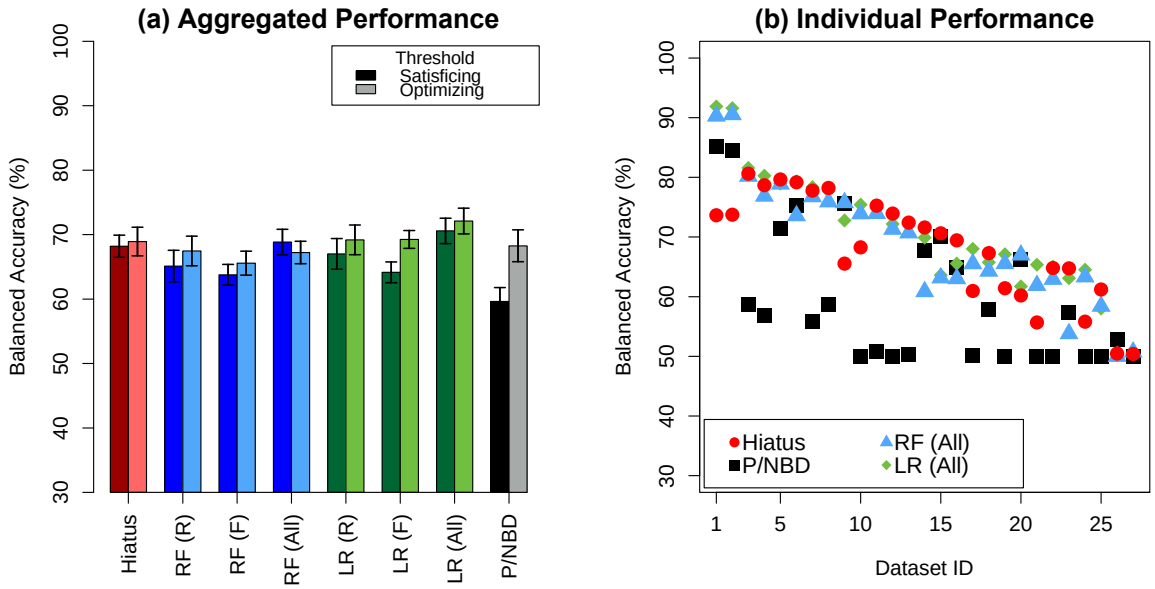

Figure 2. Comparing Performance Across 27 Retail Data Sets

Note. The figure depicts empirical results for 27 retail data sets. Panel (a) depicts the mean BACC ± one standard error for each model. Both satisficing and optimized thresholds are displayed. Panel (b) plots the BACC of selected models with the satisficing thresholds for each of the 27 data sets. The data sets are sorted by the maximum observed accuracy. Abbreviations: RF = random forest, LR = logistic regression, P/NBD = Pareto/NBD; (R) = model that uses only recency, (F) = model that uses only frequency, (All) = model that uses all variables.

Panel (b) of Figure 2 depicts the performance of the prediction methods for each data set separately and displays the best models from each model family threshold: the hiatus heuristic, LR (All), RF (All), and P/NBD. To ensure a fair comparison, we compare models using the same type of thresholds. We focus on satisficing thresholds in this section and present the results for optimized thresholds in Appendix E.

Counting how often a given model performs best reveals that LR (All) with a satisficing threshold is best in 44% of the data sets. The hiatus heuristic achieves the highest BACC in 41% of the data sets, whereas RF (All) is the best in just 11% of the data sets. Appendix E compares the performance of the model variants with optimized thresholds. In that case, LR (All) is best in 63% of the data sets, whereas the hiatus heuristic achieves the best accuracy in 15% of the data sets. This result indicates that LR (All) benefits more strongly from threshold optimization than does the heuristic. But as Figure 2a makes evident, the improvements in percentage points gains in BACC are rather small for LR (All) from optimization. In the next section, we assess whether these conclusions hold outside of the customer purchase domain.

5.3 Predicting Sports, Crime, Weather, and Other Events

Figure 3 compares the performance of the prediction methods across 31 data sets from different environments. As in Figure 2, panel (a) displays the mean BACC for each prediction method, whereas panel (b) depicts the individual performance of prediction methods for each of the data sets.

Compared with the results in the retail environment, the absolute accuracy values of all prediction strategies decrease by about two percentage points. At the same time, as in the retail environment, most models show a very similar average performance regardless of whether they rely only on recency or integrate recency, frequency, and all

other variables or whether they use an optimized threshold or a satisficing threshold (Figure 3a). As in the retail environment, logistic regression with all variables and an optimized threshold demonstrates the highest mean BACC with 69%. However, logistic regression relying on just recency achieves an accuracy of 67%. Close behind that performance, the hiatus heuristic achieves an accuracy of 66%, both with an optimized threshold and a satisficing threshold. As in the retail environment, when one compares the heuristic, the logistic regression, and the random forest, the differences in predictive performance are not large enough to be statistically significant.

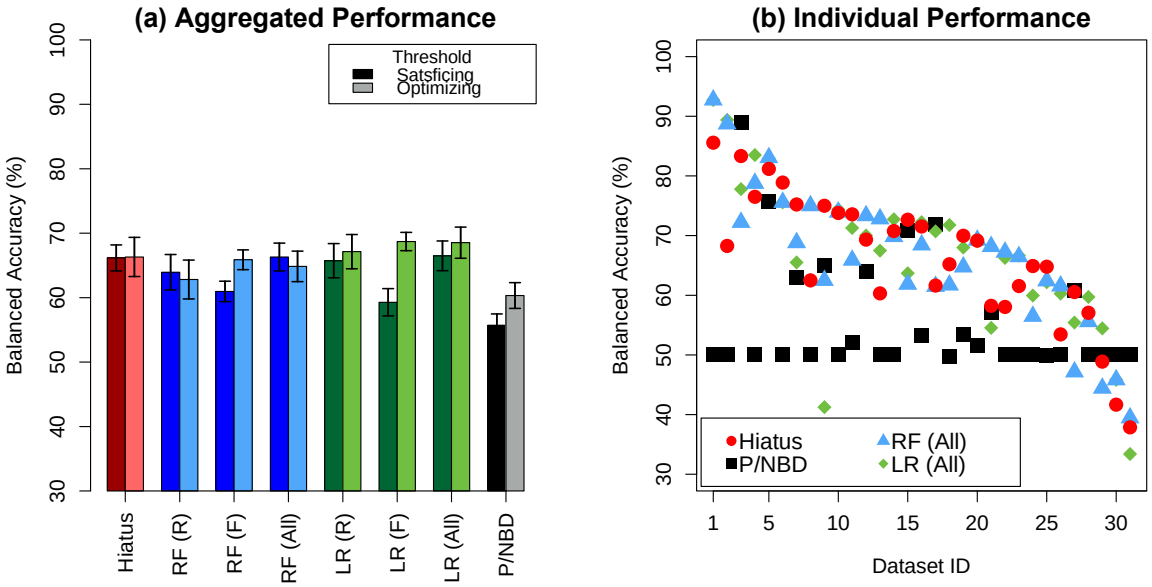

Figure 3. Comparing Performance Across 31 Non-Retail Data Sets

Note. The figure shows the empirical results for 31 non-retail data sets. Panel (a) depicts the mean BACC ± one standard error for each model. Both satisficing and optimized thresholds are displayed. Panel (b) plots the BACC of selected models with the satisficing thresholds for each of the 31 data sets. The data sets are sorted by the maximum observed accuracy. Abbreviations: RF = random forest, LR = logistic regression, P/NBD = Pareto/NBD; (R) = model that uses only recency, (F) = model that uses only frequency, (All) = model that uses all variables.

Panel (b) of Figure 3 compares the performance of the models for each data set, including the heuristic, LR (All), RF (All), and P/NBD with satisficing thresholds. The logistic regression performs best in 32% of the data sets, and the hiatus heuristic in 26% of the cases. When using optimized thresholds, LR (All) is best in 45% of the data sets,

and the hiatus heuristic demonstrates the best BACC in 16% of the cases (see Appendix E).

Overall, the results for retail data sets and non-retail data show the same patterns. If one combines these two categories in a larger analysis of 58 data sets, LR (All) with an optimized threshold achieves the highest mean BACC of 70%. The hiatus heuristic reaches a mean accuracy of 67% with a satisficing threshold and 68% with an optimized threshold. Appendix D compares the results in terms of the median performance, which further corroborates our observation of the small performance gap between the heuristic and other prediction methods. Overall, this result indicates that neither using more than one powerful variable nor optimizing a threshold necessarily yields a substantial improvement in predictive performance.

5.4 Ecological Rationality

5.4.1 Importance of Recency

Dominance and validity are two environmental characteristics that directly indicate the importance of recency for prediction. Figure 4 illustrates the relationship between the predictive performance of the hiatus heuristic and these indicators. The dominance of recency is estimated using three dominance measures suggested in Section 3.3.

Panel (a) of Figure 4 focuses on weight-based dominance. According to this measure, recency is a dominant attribute in 19% of the data sets. This indicates not only that weight-based dominance provides a necessary and sufficient condition for the hiatus heuristic to perform as well as a logistic regression that integrates recency, frequency and other variables, but also that the condition is by no means uncommon. The hiatus heuristic performs substantially better in environments that exhibit the dominance of recency, with a difference in the mean BACC of more than seven percentage points between the two data groups.

Weight-based dominance is a conservative estimate, given that the weights stem from a single regression. This usually implies that the weight of recency is reduced by the other variables contained in the regression. Residual-based dominance addresses this limitation and is depicted in panel (b). Computing the percentage of variance explained by recency alone in comparison to the *residual* percentage of variance explained by all other variables combined shows that recency explains a higher

percentage of the target variable variance in 76% of the data sets. The difference in the predictive performance of the hiatus heuristic depends on whether weight-based dominance is fulfilled and even higher compared to the weight-based measure: the mean BACC in environments where recency exhibits residual-based dominance is 74%, which is almost 15 percentage points higher than in environments in which residual-based dominance is not fulfilled.

Panel (c) depicts the correlation between the continuous agreement-based measure of dominance and the performance of the hiatus heuristic. The scatterplot indicates a positive relationship between the two metrics, suggesting that the accuracy of the heuristic is higher in environments with higher dominance. This supports the previous conclusions for the weight-based and residual-based dominance and provides another argument in favor of the importance of dominance.

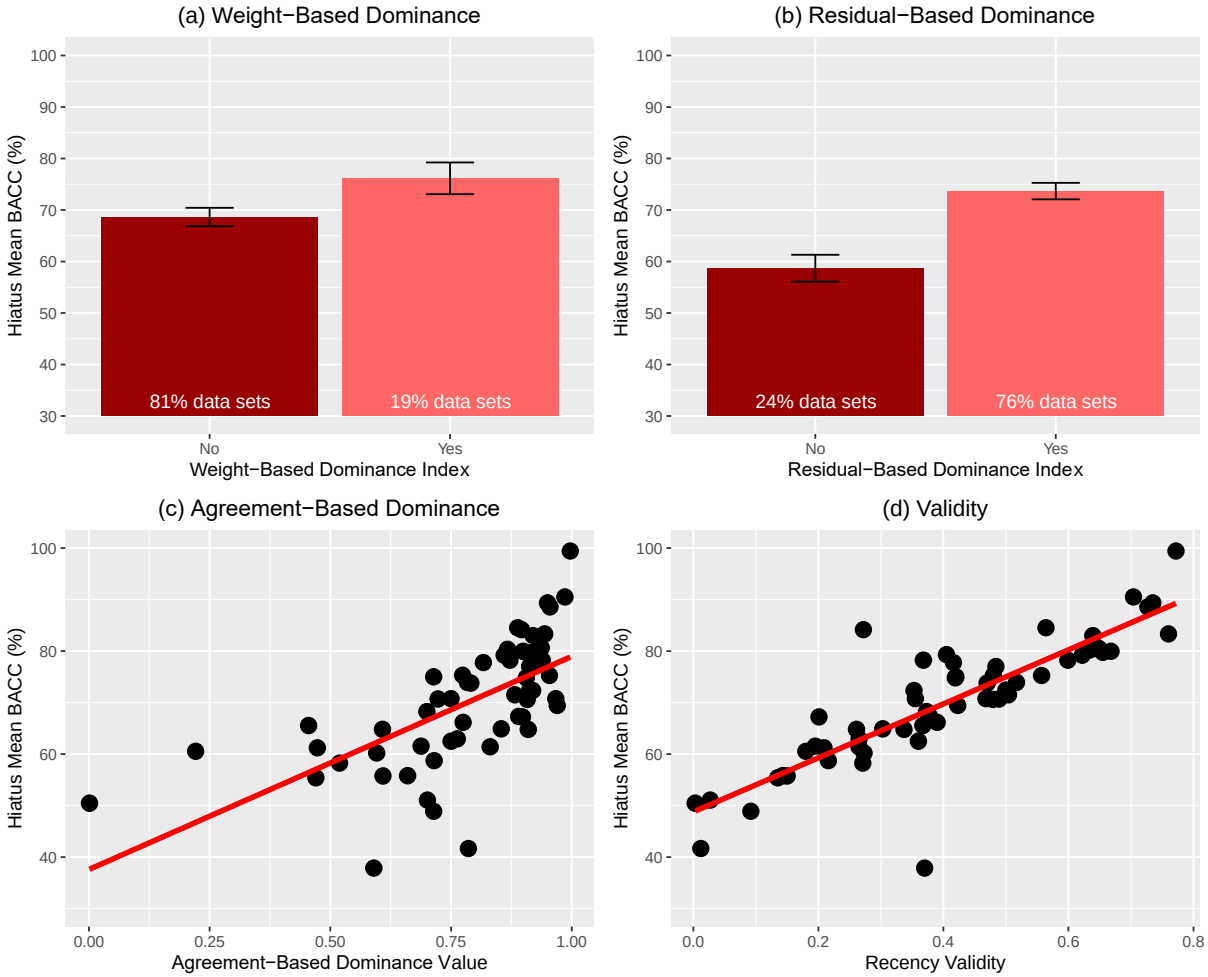

Figure 4. Impact of Dominance and Validity

Note. The figure depicts the impact of dominance and validity of recency on the hiatus heuristic's performance. Panels (a) and (b) compare the mean BACC of the hiatus heuristic ± one standard error for data sets that exhibit weight-based dominance (a) and residual-based dominance (b). Panel (c) illustrates the relationship between agreement-based dominance and BACC of the hiatus heuristic. The dots indicate individual data sets, and the red line illustrates the slope of the linear pairwise regression. Similarly, panel (d) plots the relationship between the predictive validity of recency and the performance of the hiatus heuristic.

The results are further corroborated by the observed relationship between the predictive validity of recency and the performance of the hiatus heuristic depicted in panel (d) of Figure 4. The panel demonstrates a strong positive correlation of $r = .86$ between the two metrics, indicating that the absolute importance of recency in the data environment is a requirement for a good performance of the recency-based heuristic.

Table 4. Comparing the Performance of Variables

	Recency ≥ Frequency		**Recency ≥ Frequency - 5%**		**Recency ≥ All variables - 5%**	
	LR	RF	LR	RF	LR	RF
Satisficing threshold	50%	47%	72%	69%	66%	71%
Optimized threshold	45%	41%	74%	72%	67%	72%

In studies asserting that people unduly rely on recency, the suggestion tends to be that people should instead always rely on the frequency of events. To compare the effectiveness of recency with that of frequency, one can evaluate each variable's performance within the same model family. Table 4 compares the performance of models across all 58 data sets.

As Table 4 indicates, logistic regression with only recency provides predictions that are as good as or better than a logistic regression with only frequency in 45% to 50% of the data sets, depending on the threshold type. For random forest, this share is between 41% and 45%. Note also that the average correlation between recency and frequency across the data sets is 0.52 and reaches up to 0.90 in some data sets, which suggests that these two variables may often be used interchangeably.

Together, these results suggest that recency, like frequency, is a powerful variable for prediction. This observation contrasts with the prior belief that relying on recency rather than on frequency is a fallacy. On the contrary, recency demonstrates a good performance on many of the data sets. In cases where relying on frequency is more accurate, the difference in the predictive performance tends to be insubstantial. As indicated in the second pair of columns of Table 3, recency-based models do not fall behind the frequency-based models by more than five percentage points for about 72% of the data sets. Furthermore, depending on the method, a model with all variables substantially outperforms the recency-based model in less than one third of the data sets. In other words, combining recency and frequency and using more than one powerful variable does not per se imply an improvement in predictive performance.

5.4.2 Satisficing

The bias-variance decomposition reveals error due to variance as one source of prediction error. Figure 5 compares the performance of models with different threshold types. Dark bars depict the share of data sets where optimizing predicts better than satisficing. Panel (a) indicates the data sets for which optimizing is strictly better (i.e., obtains a higher BACC). Because some differences in predictive performance appear minor, we also depict in panel (b) the share of data sets for which optimizing is better than satisficing by at least 5%. This allows us to understand the benefit of threshold optimization as opposed to satisficing.

Our analysis suggests that satisficing can yield effective performance without substantial performance losses. The satisficing hiatus heuristic, a model that decision-makers can readily implement, can often provide equal or even better performance than an optimized variant. In 52% of all data sets, the satisficing version predicts as well as or better than the optimized variant. Furthermore, the difference between optimizing and satisficing exceeds 5% in only 21% of the data sets. This suggests that simplicity, here in the form of a satisficing threshold, need not result in suboptimal performance. Even more complex models frequently benefit from the introduction of a satisficing threshold. For the random forest models, on average, the satisficing threshold performs at least as well as the optimized variant in 65% of the data sets. The logistic regression

benefits from the threshold optimization in most cases, but the accuracy gains are above 5% in less than 28% of the data sets.

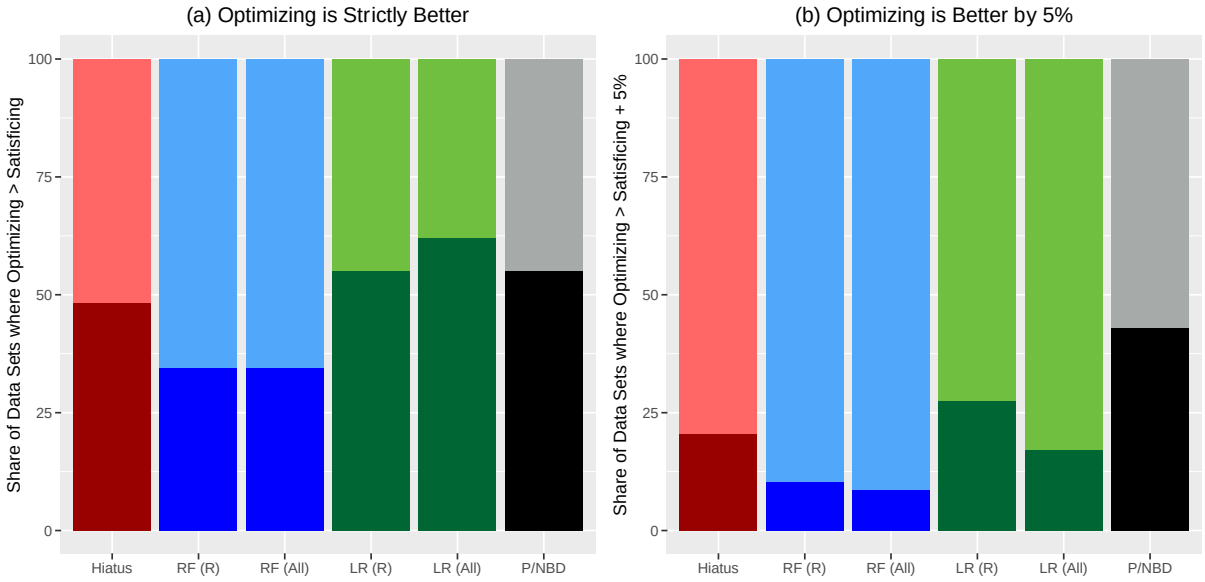

Figure 5[MO1][NK2]: Comparing Satisficing and Optimizing

Note. Dark bars in panel (a) depict the share of data sets for which a model with an optimized threshold is strictly better than the model with a satisficing threshold in terms of BAAC. The light bar indicates the remaining share of data sets, where optimizing is not strictly better than satisficing. Similarly, dark bars in panel (b) indicate the share of data sets for which a model with an optimized threshold outperforms the model with a satisficing threshold by at least 5%. Abbreviations: RF = random forest, LR = logistic regression, P/NBD = Pareto/NBD; (R) = model that uses only recency, (All) = model that uses all variables.

Focusing on the hiatus heuristic, it is interesting to compare its respective performance on data sets where satisficing or optimizing is more accurate. This helps us to better understand the nature of environments in which satisficing is ecologically rational. Figure 6 plots the mean BACC of the hiatus heuristic according to the binary satisficing index. The index is set to 'Yes' in data sets where satisficing performs better than or as well as optimizing, and 'No' in data sets where the reverse holds. This splits the entire data library into two groups.

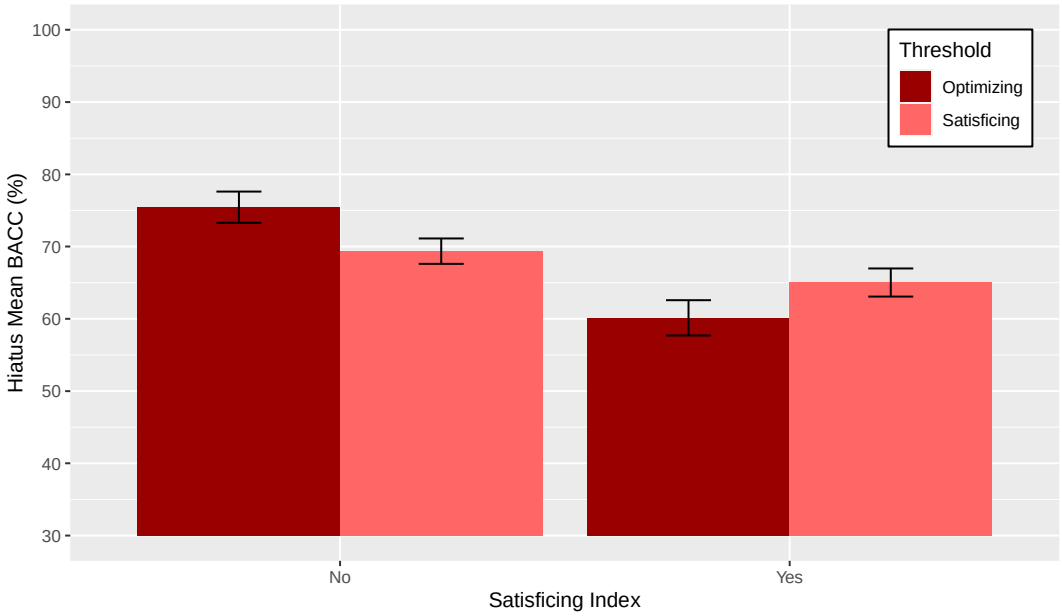

Figure 6: Impact of Satisficing

Note. The figure compares the mean BACC ± one standard error of the hiatus heuristic across the two groups of data sets: (1) data sets where the satisficing variant of the hiatus heuristic performs better than the optimizing variant (satisficing index = 'Yes'); (2) data sets where the optimizing variant is better (satisficing index = 'No').

As indicated in Figure 6, the accuracy of both heuristic variants is higher in environments in which the optimized variant of the heuristic performs better, that is, where prediction is easier. In contrast, satisficing performs better in data sets in which the accuracy of both heuristic variants is lower, that is, where prediction is more difficult. This result emphasizes that satisficing is expected to work better in more complex environments, where an additional optimization step may harm predictive performance due to the risk of overfitting.

5.4.3 Unpredictable Changes over Time

Another environmental characteristic we consider in the study is unpredictable changes over time. Figure 7 compares the predictive performance of the best prediction methods for two data partitioning schemes: a time-based partitioning used in the previous sections and a regular cross-validation framework that ignores the time dimension.

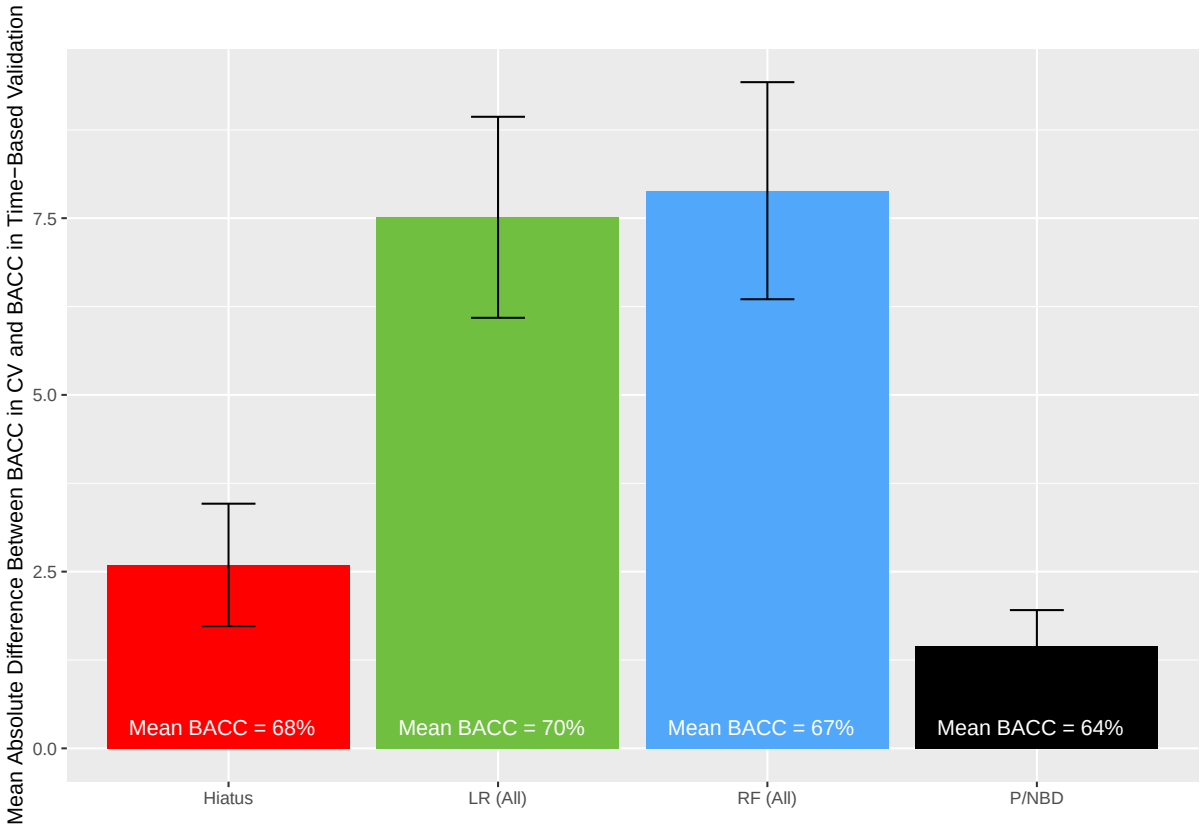

Figure 7: Impact of Unpredictable Changes over Time

Note. The figure depicts the mean absolute differences in the BACC ± one standard error for two data partitioning schemes: time-based validation and regular cross-validation. Lower differences indicate higher robustness of the prediction method to unpredictable changes over time. The differences are computed for 41 data sets, where cross-validation results are available for all prediction methods. Abbreviations: RF = random forest, LR = logistic regression, P/NBD = Pareto/NBD; (All) = model that uses all variables.

In Figure 7, we use only 41 out of 58 data sets because the sample size in some data sets makes it impossible to train all prediction methods using the cross-validation data split. The figure depicts the mean absolute differences between BACC achieved by each method in two data partitioning schemes. A low discrepancy between the two accuracy values indicates that a prediction method is stable toward the changes in the environment that occur in time, whereas a high discrepancy shows that a method is sensitive to changes in the environment. According to the results, LR (All) and RF (All) demonstrate the highest differences in accuracy, with above seven percentage points, indicating that these models using all variables may overfit the historical training data

and fail to maintain stable performance when predicting future events. The hiatus heuristic, in contrast, has a substantially lower discrepancy in accuracy between the time-based and traditional cross-validation, exhibiting a much lower sensitivity to the type of data partitioning and differences in mean accuracy below five percentage points on 82% of the data sets. This implies that the heuristic based on a single variable is more robust to unpredictable changes over time. Note that the Pareto/NBD model also demonstrates a stable performance, with a mean accuracy difference below two percentage points. However, this robustness comes at the cost of a lower predictive accuracy compared with all prediction method methods used in the study.

6 Summary

This paper investigates the role of recency in predicting future events by testing the hiatus heuristic as a simple heuristic that relies only on recency. On the basis of the bias-variance decomposition, we lay out four conditions under which relying only on recency can be ecologically rational: dominant attributes, validity, satisficing, and unpredictable changes over time. Dominance and validity relate to prediction error due to bias; satisficing and unpredictable changes over time relate to prediction error due to variance.

The empirical results for 58 real-world problems from different domains suggest that recency in general and the recency-based hiatus heuristic in particular yield predictive performance on par with more complex statistical and machine-learning methods that use recency and other valid predictors. These findings are observed in both the marketing domain, where the hiatus heuristic was initially observed when managers predicted future customer purchases, and other environments beyond the retail context, where we use the heuristic to predict future events in sports, weather, banking, and other areas.

The bias-variance decomposition helps us understand the conditions under which the hiatus heuristic can predict as accurately as or better than other models. Using a satisficing threshold for the hiatus reduces error from variance to zero because there is no free parameter, although it increases the risk of error due to bias. If recency is dominant and its validity high, moreover, error due to bias is reduced. If weight-based

dominance holds, the bias can be reduced to that of a linear model. Systematic differences between the future and the past can further increase the prediction error that is due to variance, particularly affecting those models where more free parameters need to be estimated. This combination explains the observation that a simple recency-based heuristic can make equally good or better predictions than models that use additional information, such as the frequency of events.

7 Conclusion

Relying solely on a single variable, and relying on recency in particular, has been generally dismissed as an inferior method of making predictions. For instance, Kahneman and Tversky (1973) identify the availability heuristic and the reliance on recency as a central strategy in human decision-making. Recency is used by people in such diverse settings as predicting whether a customer will buy again at a shop or whether a hurricane will strike at a certain location again. Kahneman and Tversky (1973) conclude that decision-making where people ignore much of the available information, most notably that of frequency, violates the laws of probability theory, which in turn yields suboptimal results.

The present research is not alone in calling this conclusion into question. Instead of uniformly dismissing violations of probability theory as a fallacy in decision-making, Simon (1956) noted that it is important to conceive of decision-making as constituting a pair of scissors, one blade being the environment, the other the decision strategy. A strategy performs well only if it is adapted to the environment. Savage (1954), the founding father of modern Bayesian updating, makes a distinction between two broad categories of environments (see also Keynes 1921; Knight 1921): risk and uncertainty. Under risk, the exhaustive and mutually exclusive set of all future states of the world, the probability with which a state occurs, and the set of consequences associated with each alternative are known. Under uncertainty, some of the information is lacking such that the exhaustive set of states of the world and their consequences is not knowable or foreseeable at the point of decision-making. Savage (1954, 16) cites as an example planning a picnic, where events can occur that one cannot know ahead of time. He points out that Bayesian updating cannot and should not be applied under uncertainty.

Under uncertainty, there is no predefined benchmark of rationality or a prediction method that predicts best under any circumstances. Instead, each model performs well under particular conditions, meaning that simple heuristics need not constitute an inferior method of making predictions per se. To explicate these conditions is the task of the study of ecological rationality. This research demonstrates that in face of an uncertain future, one needs to take into account both error due to bias and error due to variance to understand whether or not a decision strategy is appropriate for the given conditions. As Geman et al. (1992) observe, it is important to introduce the right "bias" into models to limit exposure of error due to variance. Note that the average data set that we use contains 75,129 binary event observations on 2,757 entities over 57 time units. Simply making a large amount of data available to prediction models does not suffice to introduce the right bias in our analysis. At the same time, the average decision-maker, who probably has far fewer observations in their memory than the average model in this study, readily utilizes recency, thereby incurring just the right bias.

Despite the fact that our data covers a wide range of different characteristics and environments, we have presumably not addressed all conditions that facilitate or hinder the performance of recency and the hiatus heuristic in comparison with the other models, nor is the list of models that we study by any means exhaustive. This points to further research opportunities for which this paper provides a steppingstone in understanding under what conditions the heuristics that people use can be powerful tools.

As Box and Draper famously observed, "all models are wrong, but some are useful" (1987, 424). Models, by their very nature, do not represent one-to-one copies of the real world but instead abstract and simplify. When people make decisions, a core challenge is to make good predictions. Simplicity in the form of a heuristic can protect against overfitting and, in tandem with reliance on recency, can yield a powerful decision tool.

Bibliography

Agarwal, Sumit, John C. Driscoll, Xavier Gabaix, and David I. Laibson. 2013. "Learning in the Credit Card Market." *National Bureau of Economic Research (NBER) Working Paper 13822*.

Anderson, John R., and Lael J. Schooler. 1991. "Reflections of the Environment in Memory." *Psychological Science* 2 (6): 396–408.

Artinger, Florian M., and Gerd Gigerenzer. 2016. "Heuristic Pricing in an Uncertain Market: Ecological and Constructivist Rationality." *SSRN Working Paper*.

Artinger, Florian M., Gerd Gigerenzer, and Perke Jacobs. 2022. "Satisficing: Integrating Two Traditions." *Journal of Economic Literature* 60 (2): 598–635.

Artinger, Florian M., Nikita Kozodoi, Florian Wangenheim, and Gerd Gigerenzer. 2018. "Recency: Prediction with Smart Data." In *American Marketing Association Winter Conference Proceedings*, edited by Jacob Goldenberg, Juliano Laran, and Andrew Stephen, 29:L–2. Chicago, IL: American Marketing Association.

Artinger, Florian M., Malte Petersen, Gerd Gigerenzer, and Jürgen Weibler. 2015. "Heuristics as Adaptive Decision Strategies in Management." *Journal of Organizational Behavior* 36 (S1): S33–52.

Box, George EP, and Norman Richard Draper. 1987. *Empirical Model-Building and Response Surfaces*. New York: Wiley.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78 (1): 1–3.

Brighton, Henry, and Gerd Gigerenzer. 2015. "The Bias Bias." *Journal of Business Research* 68 (8): 1772–84.

Brown, Thomas. 1838. *Lectures on the Philosophy of the Human Mind*. William Tait.

Caplin, Andrew, Sumit Chopra, John V. Leahy, Yann LeCun, and Trivikraman Thampy. 2008. "Machine Learning and the Spatial Structure of House Prices and Housing Returns." *SSRN Electronic Journal*, December.

Caplin, Andrew, and Mark Dean. 2011. "Search, Choice, and Revealed Preference." *Theoretical Economics* 6 (1): 19–48.

Caplin, Andrew, Mark Dean, and Daniel Martin. 2011. “Search and Satisficing.” *American Economic Review* 101 (7): 2899–2922.

Davis, Lucas W. 2004. “The Effect of Health Risk on Housing Values: Evidence from a Cancer Cluster.” *American Economic Review* 94 (5): 1693–1704.

DellaVigna, Stefano. 2009. “Psychology and Economics: Evidence from the Field.” *Journal of Economic Literature* 47 (2): 315–72.

Domingos, Pedro. 2000. “A Unified Bias-Variance Decomposition for Zero-One and Squared Loss.” In *Proceedings of the AAAI*, 564–69.

Domingos, Pedro, and Michael Pazzani. 1997. “On the Optimality of the Simple Bayesian Classifier under Zero-One Loss.” *Machine Learning* 29 (2–3): 103–30.

Dosi, Giovanni, Mauro Napoletano, Andrea Roventini, Joseph E. Stiglitz, and Tania Treibich. 2020. “Rational Heuristics? Expectations and Behaviors in Evolving Economies with Heterogeneous Interacting Agents.” *Economic Inquiry* 58 (3): 1487–1516.

Fader, Peter S., Bruce G. S. Hardie, and Ka Lok Lee. 2005. “‘Counting Your Customers’ the Easy Way: An Alternative to the Pareto/NBD Model.” *Marketing Science* 24 (2): 275–84.

Fawcett, Tom. 2006. “An Introduction to ROC Analysis.” *Pattern Recognition Letters* 27 (8): 861–74.

Freeman, Elizabeth A., and Gretchen G. Moisen. 2008. “A Comparison of the Performance of Threshold Criteria for Binary Classification in Terms of Predicted Prevalence and Kappa.” *Ecological Modelling* 217 (1–2): 48–58.

Friedman, Daniel, Mark R. Isaac, Duncan James, and Shyam Sunder. 2014. *Risky Curves: On the Empirical Failure of Expected Utility*. New York, NY: Routledge.

Gabaix, Xavier. 1999. “Zipf’s Law for Cities: An Explanation.” *Quarterly Journal of Economics* 114 (3): 739–67.

———. 2016. “Power Laws in Economics: An Introduction.” *Journal of Economic Perspectives* 30 (1): 185–206.

Gabaix, Xavier, and David Laibson. 2006. “Shrouded Attributes, Consumer Myopia, and Information Surpression in Competetive Markets.” *Quarterly Journal of Economics* 121 (2): 505–41.

Gaissmaier, Wolfgang, and Gerd Gigerenzer. 2012. “9/11, Act II: A Fine-Grained Analysis of Regional Variations in Traffic Fatalities in the Aftermath of the Terrorist Attacks.” *Psychological Science* 23 (12): 1449–54.

Gallagher, Justin. 2014. “Learning about an Infrequent Event: Evidence from Flood Insurance Take-Up in the United States.” *American Economic Journal: Applied Economics* 6: 206–33.

Gallistel, C R, Monika Krishan, Ye Liu, Reilly Miller, and Peter E Latham. 2014. “The Perception of Probability.” *Psychological Review* 121 (1): 96–123.

Geman, Stuart, Elie Bienenstock, and René Doursat. 1992. “Neural Networks and the Bias-Variance Dilemma.” *Neural Computation* 4 (1): 1–58.

Gigerenzer, Gerd. 2000. *Adaptive Thinking: Rationality in the Real World*. New York, NY: Oxford University Press.

Gigerenzer, Gerd, and Wolfgang Gaissmaier. 2011. “Heuristic Decision Making.” *Annual Review of Psychology* 62 (January): 451–82.

Gigerenzer, Gerd, Peter M. Todd, and the ABC Research Group. 1999. *Simple Heuristics That Make Us Smart*. New York: Oxford University Press.

Green, Brett, and Jeffrey Zwiebel. 2018. “The Hot-Hand Fallacy: Cognitive Mistakes or Equilibrium Adjustments? Evidence from Major League Baseball.” *Management Science* 64 (11): 5315–48.

Hand, David J., and Keming Yu. 2001. “Idiot’s Bayes: Not So Stupid after All?” *International Statistical Review* 69 (3): 385–98.

Haselhuhn, Michael P., Devin G. Pope, Maurice E. Schweitzer, and Peter Fishman. 2012. “The Impact of Personal Experience on Behavior: Evidence from Video-Rental Fines.” *Management Science* 58 (1): 52–61.

Hertwig, Ralph, Greg Barron, Elke U Weber, and Ido Erev. 2004. “Decisions from Experience and the Effect of Rare Events in Risky Choice.” *Psychological Science* 15 (8): 534–39.

Jones, Matt and Winston R. Sieck. 2003. “Learning Myopia: An Adaptive Recency Effect in Category Learning.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29 (4): 626–40.

Katsikopoulos, Konstantinos V., Özgür Şimşek, Marcus Buckmann, and Gerd Gigerenzer. 2022. “Transparent Modeling of Influenza Incidence: Big Data or a Single Data Point from Psychological Theory?” *International Journal of Forecasting* 38 (2): 613–19.

Keynes, John Maynard. 1921. *Treatise on Probability*. London: Macmillan.

Knight, Frank H. 1921. *Risk, Uncertainty and Profit*. Boston: Houghton-Mifflin.

Kunreuther, Howard. 1976. “Limited Knowledge and Insurance Protection.” *Public Policy* 24 (2): 227–61.

Lant, Theresa K. 1992. “Aspiration Level Adaptation: An Empirical Exploration.” *Management Science* 38 (5): 623–44.

Lejarraga, Tomás, and Ralph Hertwig. 2021. “How Experimental Methods Shaped Views on Human Competence and Rationality.” *Psychological Bulletin* 147 (6): 535–64.

Malmendier, Ulrike, and Stefan Nagel. 2011. “Depression Babies: Do Macroeconomic Experiences Affect Risk Taking?” *Quarterly Journal of Economics* 126 (1): 373–416.

———. 2016. “Learning from Inflation Experiences.” *The Quarterly Journal of Economics* 131 (1): 53–87.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2009. *An Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Manzini, Paola, and Marco Mariotti. 2007. “Sequentially Rationalizable Choice.” *American Economic Review* 97 (5): 1824–39.

———. 2012. “Choice by Lexicographic Semiorders.” *Theoretical Economics* 7 (1): 1–23.

Martignon, Laura, and Ulrich Hoffrage. 2002. “Fast, Frugal, and Fit: Simple Heuristics for Paired Comparison.” *Theory and Decision* 52 (1): 29–71.

Mullainathan, Sendhil, and Jann Spiess. 2017. “Machine Learning: An Applied Econometric Approach.” *Journal of Economic Perspectives* 31 (2): 87–106.

Nisbett, Richard, and Lee Ross. 1980. *Human Inference: Strategies and Shortcomings of Social Judgment*. *Prentice Hall*. JSTOR.

Petrison, Lisa A., Robert C. Blattberg, and Paul Wang. 1997. "Database Marketing. Past, Present, and Future." *Journal of Direct Marketing* 11 (4): 109–25.

Platzer, Michael, and Thomas Reutterer. 2016. "Ticking Away the Moments: Timing Regularity Helps to Better Predict Customer Activity." *Marketing Science* 35 (5): 779–99.

Rudin, Cynthia. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence* 1 (5): 206–15.

Rudin, Cynthia, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. 2022. "Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges." *Statistics Surveys* 16 (1): 1–85.

Savage, Leonard J. 1954. *The Foundations of Statistics*. New York: Wiley.

Schmittlein, David C., Donald G. Morrison, and Richard Colombo. 1987. "Counting Your Customers: Who-Are They and What Will They Do Next?" *Management Science* 33 (1): 1–24.

Schulze, Christin, and Ralph Hertwig. 2021. "A Description–Experience Gap in Statistical Intuitions: Of Smart Babies, Risk-Savvy Chimps, Intuitive Statisticians, and Stupid Grown-Ups." *Cognition* 210 (5): 104580.

Semenova, Lesia, Cynthia Rudin, and Ronald Parr. 2019. "A Study in Rashomon Curves and Volumes: A New Perspective on Generalization and Model Simplicity in Machine Learning." *Working Paper*.

Simon, Herbert A. 1955. "A Behavioral Model of Rational Choice." *Quarterly Journal of Economics* 69 (1): 99–118.

———. 1956. "Rational Choice and the Structure of the Environment." *Psychological Review* 63 (2): 129–38.

Şimşek, Özgür. 2013. "Linear Decision Rule as Aspiration for Decision Heuristics." *Advances in Neural Information Processing Systems* 26 (1): 2904–12.

Slovic, Paul, Howard Kunreuther, and Gilbert F. White. 1974. "Decision Processes, Rationality, and Adjustments to Natural Hazards: A Review of Some

Hypotheses." In *Natural Hazards: Local, National and Global*, edited by Gilbert White, 187–205. New York: Oxford University Press.

Smith, Vernon L. 2002. "Constructivist and Ecological Rationality in Economics." *American Economic Review* 93 (3): 465–508.

Starmer, Chris. 2005. "Normative Notions in Descriptive Dialogues." *Journal of Economic Methodology* 12 (2): 277–89.

Tversky, Amos, and Daniel Kahneman. 1973. "Availability: A Heuristic for Judging Frequency and Probability." *Cognitive Psychology* 5: 207–32.

———. 1974. "Judgment under Uncertainty: Heuristics and Biases." *Science* 185 (4157): 1124–31. https://doi.org/10.1126/science.185.4157.1124.

Van Der Putten, Peter, and Maarten Van Someren. 2004. "A Bias-Variance Analysis of a Real World Learning Problem." *Machine Learning* 57 (1): 177–95.

Wu, Xindong, Vipin Kumar, Quinlan J. Ross, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2008. "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems* 14 (1): 1–37.

Wübben, Markus, and Florian von Wangenheim. 2008. "Instant Customer Base Analysis: Managerial Heuristics Often 'Get It Right.'" *Journal of Marketing* 72 (3): 82–93.

Wulff, Dirk U., Max Mergenthaler-Canseco, and Ralph Hertwig. 2017. "A Meta-Analytic Review of Two Modes of Learning and the Description-Experience Gap." *Psychological Bulletin* 144 (2): 140–76.

Zech, John R., Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. 2018. "Variable Generalization Performance of a Deep Learning Model to Detect Pneumonia in Chest Radiographs: A Cross-Sectional Study." *PLOS Medicine* 15 (11): e1002683.

Zhang, Yao, Eric T Bradlow, and Dylan S Small. 2013. "New Measures of Clumpiness for Incidence Data." *Journal of Applied Statistics* 40 (11): 2533–48.

Zhang, Yao, Eric T. Bradlow, and Dylan S. Small. 2015. "Predicting Customer Value Using Clumpiness: From RFM to RFMC." *Marketing Science* 34 (2): 195–208.

Zou, Quan, Sifa Xie, Ziyu Lin, Meihong Wu, and Ying Ju. 2016. "Finding the Best Classification Threshold in Imbalanced Classification." *Big Data Research* 5 (September): 2–8. https://doi.org/10.1016/J.BDR.2015.12.001.

For Online Publication

Online Appendix

Appendix A: Data Description

This Appendix provides a description of each of the data sets used in the paper. The library contains 58 data sets, including 27 retail and 31 non-retail data sets. Some data sources provide several data sets, either coming from different places (e.g., transactional data from separate stores) or using different dependent variables. The maximum number of data sets from a single source is limited to five.

For each data set, we select a cohort of entities that satisfices a certain criterion. The cohort criterion is used to ensure that for all entities observed at least one includes a year of data. For instance, in the data set *data_bn1* we only retain entities (in this case, bank customers) that have their first transaction in January 2015. This ensures that there are at least one year of transactions for the selected customers.

Below, we summarize each of the data sets, indicating the data domain and the dependent variable (i.e., the event type), the list of variables used for prediction by statistical models, the time frame, the number of entities, and the number of the observed events. Table C1 overviews the retail data sets, whereas Table C2 describes data sets from the non-retail environments.

Table A1. Customer Purchase Data Description

#	Name	Nature	Predicted event	Variables used for prediction	No. entities	No. events	Time frame
1	data_air	Transactional data from an airline company	Purchasing a flight ticket	Recency, frequency, clumpiness, number of flights	1,430 customers	6,093	14 quarters (2000 –2004)
2	data_app	Transactional data from an apparel retailer	Making a purchase	Recency, frequency, clumpiness, order amount, number of items, item type, customer category	2,709 customers	24,609	87 weeks (2002 –2003)
3	data_cdn	Transactional data from CDnow, Inc. online retailer	Making a purchase	Recency, frequency, clumpiness, order amount, number of items	2,357 customers	6,919	78 weeks (1997 – 1998)
4	data_cl1	Carbohydrate rich food sales data in store 1	Making a purchase	Recency, frequency, clumpiness, order amount, number of items, whether a discount coupon is used	620 customers	14,876	104 weeks (2000 – 2001)
5	data_cl2	Carbohydrate rich food sales data in store 2	Making a purchase	Recency, frequency, clumpiness, order amount, number of items, whether a discount coupon is used	592 customers	15,126	104 weeks (2000 – 2001)
6	data_cl3	Carbohydrate rich food sales data in store 3	Making a purchase	Recency, frequency, clumpiness, order amount, number of items, whether a discount coupon is used	672 customers	10,647	104 weeks (2000 – 2001)
7	data_cl4	Carbohydrate rich food sales data in store 4	Making a purchase	Recency, frequency, clumpiness, order amount, number of items	509 customers	14,538	104 weeks (2000 – 2001)
8	data_cl5	Carbohydrate rich food sales data in store 5	Making a purchase	Recency, frequency, clumpiness, order amount, number of items, whether a discount coupon is used	666 customers	13,140	104 weeks (2000 – 2001)
9	data_cpn	Transactional data from an online discount coupons service	Purchasing a discount coupon	Recency, frequency, clumpiness	6,763 customers	82,613	51 weeks (2011 –2012)
10	data_dmc	Transactional data from an online retailer	Making a purchase	Recency, frequency, clumpiness, order amount, premium status	1,713 customers	4,911	10 weeks (2015)

Table A1. Customer Purchase Data Description (cont.)

#	Name	Nature	Predicted event	Variables used for prediction	No. entities	No. events	Time frame
11	data_fa1	Transactional data from a fashion store	Making a purchase	Recency, frequency, clumpiness, order amount, number of items, voucher amount, article size, payment method, WOAL	4,813 customers	68,643	52 weeks (2014)
12	data_fa2	Transactional data from a fashion store	Making a purchase	Recency, frequency, clumpiness, order amount, number of items, voucher amount, article size, payment type, WOAL	5,024 customers	5,024	52 weeks (2015)
13	data_fd1	Transactional data from a food chain	Making a purchase	Recency, frequency, clumpiness, order amount, number of units	2,981 customers	64,999	52 weeks (1997)
14	data_fd2	Transactional data from a food chain	Making a purchase	Recency, frequency, clumpiness, order amount, number of units	4,554 customers	134,651s	78 weeks (1998)
15	data_hs1	Household-level transactional data from an online retailer	Making a purchase	Recency, frequency, clumpiness, order amount, number of units, discount size	2,017 households	24,609	49 weeks (2001)
16	data_hs2	Household-level transactional data from an online retailer	Making a purchase	Recency, frequency, clumpiness, order amount, number of units, discount size	540 households	1,320,372	52 weeks (2000)
17	data_in1	Transactional data from an online retailer	Making a purchase	Recency, frequency, clumpiness	10,000 customers	165,676	52 weeks
18	data_in2	Transactional data from an online retailer	Making a purchase	Recency, frequency, clumpiness	10,000 customers	164,513	52 weeks
19	data_kw1	Juice drink sales data from a retailer store 1	Making a purchase	Recency, frequency, clumpiness, order amount	205 customers	551	52 weeks
20	data_kw2	Juice drink sales data from a retailer store 2	Making a purchase	Recency, frequency, clumpiness, order amount	139 customers	306	52 weeks

Table A1. Customer Purchase Data Description (cont.)

#	Name	Nature	Predicted event	Variables used for prediction	No. entities	No. events	Time frame
21	data_red	Transactional data from Red Hat company	Making a purchase-related activity	Recency, frequency, clumpiness, activity type	7,683 customers	30,026	35 weeks
22	data_ret	Transactional data from a retailer	Making a purchase	Recency, frequency, clumpiness, order amount, number of items	4,372 customers	406,829	53 weeks (2010 – 2011)
23	data_rt1	Transactional data from a retailer	Making a purchase	Recency, frequency, clumpiness, order amount, customer gender, state	8,962 customers	132,365	52 weeks (2012 – 2013)
24	data_sps	Transactional data from a retailer	Making a purchase	Recency, frequency, clumpiness, order amount, number of units, item category, order priority, shipment mode	642 customers	5,537	156 weeks (2010 – 2012)
25	data_tf1	Transactional data from a Chinese grocery store	Making a purchase	Recency, frequency, clumpiness, order amount, number of units, assets, customer age	16,578 customers	345,336	8 weeks (2001)
26	data_tf2	Transactional data from a Chinese grocery store	Making a purchase	Recency, frequency, clumpiness, order amount, number of units, assets, customer age	16,760 customers	336,413	9 weeks (2000)
27	data_tsh	Transactional data from a store selling T-shirts	Making a purchase	Recency, frequency, clumpiness, order amount, number of items	13,280 customers	13,591	96 weeks (2013 – 2014)

Table A2. Non-Retail Data Description

#	Name	Nature	Predicted event	Variables used for prediction	No. entities	No. events	Time frame
1	data_bn1	Transactional data from a bank	Withdrawing money via ATM	Recency, frequency, clumpiness, transaction amount	1,731 customers	51,346	66 weeks (2015 – 2016)
2	data_bn2	Transactional data from a bank	Making a purchase in a physical store	Recency, frequency, clumpiness, transaction amount	615 customers	15,048	66 weeks (2015 – 2016)
3	data_bn3	Transactional data from a bank	Transferring money to another person	Recency, frequency, clumpiness, transaction amount	1,250 customers	9,301	66 weeks (2015 – 2016)
4	data_com	Customer complaints data about Comcast Internet	Making an online complaint	Recency, frequency, clumpiness, rating	630 customers	1,477	52 weeks (2015)
5	data_ctr	Company-level transactional data from UK corporations	Making a purchase	Recency, frequency, clumpiness, order amount	237 companies	2,842	48 weeks (2014 –2015)
6	data_cz1	Transactional data from a Czech bank	Deposing money to the bank account	Recency, frequency, clumpiness, transaction amount	384 customers	2,426	52 weeks (1993)
7	data_cz2	Transactional data from a Czech bank	Transferring money to another person	Recency, frequency, clumpiness, transaction amount	538 customers	9,030	93 weeks (1995 – 1996)
8	data_don	Donation data from a charity	Making a donation	Recency, frequency, clumpiness	5,652 donors	29,639	11 yeas (1996 – 206)
9	data_drt	County-level weather data from the US	An extreme drought is observed	Recency, frequency, clumpiness	1,208 counties	62,623	66 quarters (2000 – 2016)
10	data_eur	Eurovision song contest data	Placing in top-2 in the contest finals	Recency, frequency, clumpiness, number of artists, gender, previous Eurovision experience	29 countries	83	41 years (1975 – 2015)

Table A2. Non-Retail Data Description (cont.)

#	Name	Nature	Predicted event	Variables used for prediction	No. entities	No. events	Time frame
11	data_jkl	Country-level data on journalists' killings	A journalist is killed	Recency, frequency, clumpiness, gender, whether journalist is local	62 countries	1,456	99 quarters (1992 – 2016)
12	data_nba	Data on annual NBA winners	Winning an NBA conference	Recency, frequency, clumpiness	29 teams	134	67 years (1950 – 2016)
13	data_nhl	Data on Annual NHL winners	Winning a Stanley Cup	Recency, frequency, clumpiness	21 teams	89	89 years (1927 – 2016)
14	data_pap	Data on papers presented at the annual NeurIPS	Presenting a paper on NeurIPS	Recency, frequency, clumpiness	2349 authors	6,555	29 years (1987 – 2016)
15	data_pod	County-level data on police officers' fatalities in the US	An officer dies on duty	Recency, frequency, clumpiness	2,459 counties	8,561	50 years (1967 – 2016)
16	data_prc	Transactional data for corporate procurement	Making a purchase	Recency, frequency, clumpiness, order amount	28 companies	422	154 weeks (2010 – 2012)
17	data_rl1	Survey data from Russian Longitudinal Monitoring Survey	Working at two jobs	Recency, frequency, clumpiness, age, gender, marriage status, education level, health level	80 individuals	322	16 years (2000 – 2015)
18	data_rl2	Survey data from Russian Longitudinal Monitoring	Growing fruits or vegetables for sale	Recency, frequency, clumpiness, age	710 individuals	3,312	16 years (2000 – 2015)
19	data_rl3	Survey data from Russian Longitudinal Monitoring	Breeding animals or birds for sale	Recency, frequency, clumpiness, age	575 individuals	3,471	16 years (2000 – 2015)
20	data_rl4	Survey data from Russian Longitudinal Monitoring Survey	Renting out a room or an apartment	Recency, frequency, clumpiness, age, gender, marriage status, education level, health level	63 individuals	164	16 years (2000 – 2015)

Table A2. Non-Retail Data Description (cont.)

#	Name	Nature	Predicted event	Variables used for prediction	No. entities	No. events	Time frame
21	data_rl5	Survey data from Russian Longitudinal Monitoring Survey	Having a savings account with interest	Recency, frequency, clumpiness, age, gender, marriage status, education level, health level	145 individuals	311	16 years (2000 – 2015)
22	data_rt2	Item return data from a fashion store	Returning a purchased item	Recency, frequency, clumpiness, order amount, customer gender, state	6,093 customers	54,416	52 weeks (2012 – 2013)
23	data_sal	Corporate-level transactional data	Making a purchase	Recency, frequency, clumpiness, order amount, number of units, item type	92 companies	2,823	125 weeks (2003 – 2005)
24	data_sfr	Municipality-level data on school fires in Sweden	A school fire is observed	Recency, frequency, clumpiness, number of cases, population	197 locations	1,161	16 years (1998 – 2014)
25	data_ter	City-level data on worldwide terrorist attacks	A terrorist attack is observed	Recency, frequency, clumpiness	2310 cities	3,749	45 years (1970 – 2015)
26	data_tor	State-level US weather data on tornadoes with fatalities	A fatal tornado has occurred	Recency, frequency, clumpiness, number of injuries, width, length	43 states	1,605	66 years (1950 – 2015)
27	data_trn	Arrival times data of regional trains in Philadelphia	Arriving at least 10 minutes late	Recency, frequency, clumpiness, delay in minutes, train direction	773 train routes	166,090	32 weeks (2016)s
29	data_ucl	Results data on UEFA Champions League	Playing in the finals	Recency, frequency, clumpiness	39 football teams	122	61 years (1956 – 2016)
30	data_ufo	City-level reports data on worldwide UFO sightings	An UFO is reported	Recency, frequency, clumpiness, observation duration, UFO shape	4,766 cities	24,391	65 years (1950 – 2014)
31	data_uni	Data on worldwide University rankings from Shanghai agency	Placing in the top-200	Recency, frequency, clumpiness, scores in categories: alumni, awards, HCI, PUB, PCP	225 universities	2,045	11 years (2005 – 2015)

Appendix B: Empirical Results for Each Data Set and Model

This Appendix provides individual performance values for each of the prediction methods on each data set. Table B1 depicts performance of methods with a satisficing threshold, whereas Table B2 displays performance of methods with an optimized threshold. The values in the tables refer to the BACC.

Note that some data sets exhibit a BACC of 50% for some statistical models. For some data sets, it is difficult to estimate the parameters of Pareto/NBD model because the underlying optimization algorithm may not converge. The convergence-related issues appear in data sets with specific structures. First, when a large proportion of the observed entities remain active throughout the entire observation period and do not stop being active at any point, the probabilities predicted by P/NBD are too close to one for all entities. Second, when the proportion of active entities is substantially larger than the share of passive entities, the probabilities cannot be calculated at all. Overall, the data sets with a small number of time units and/or a small number of entities are more likely to exhibit the data structure described above, which leads to the estimation issues. For data sets with such issues, if fitting a P/NBD model is not successful, we use a naïve prediction method instead (all events are predicted to be occurring or non-occurring depending on what is the majority class in the observed sample).

In the remaining cases, a BACC of 50% corresponds to a setting where a model predicts all events to be either occurring or non-occurring because of the learned patterns from the training sample. In that case, $BACC = \frac{\frac{TP}{P}+\frac{TN}{N}}{2} = \frac{100+0}{2} = 50$.

Table B1. Balanced Accuracy with Satisficing Thresholds

#	Data set	Hiatus	LR (All)	LR (R)	RF (All)	RF (R)	P/NBD
1	data_air	73.92	72.18	66.02	71.30	71.26	50.00
2	data_app	68.26	75.43	66.32	73.89	61.62	50.01
3	data_bn1	73.57	71.28	73.87	65.88	76.17	52.02
4	data_bn2	69.15	69.06	70.11	69.23	69.49	51.62
5	data_bn3	65.20	71.76	66.61	61.71	49.68	49.77
6	data_cdn	70.62	63.61	56.75	63.15	58.31	70.08
7	data_cl1	80.61	81.58	82.38	80.17	77.50	58.74
8	data_cl2	78.67	80.26	80.36	76.85	79.99	56.81
9	data_cl3	78.23	77.73	76.26	75.90	73.71	58.67
10	data_cl4	79.64	78.96	80.91	78.87	80.07	71.35
11	data_cl5	77.76	78.36	79.45	76.75	77.01	55.84
12	data_com	61.55	66.65	50.00	66.46	53.02	50.00
13	data_cpn	79.16	79.32	79.55	73.56	50.11	75.24
14	data_ctr	71.52	72.25	67.60	68.44	67.54	53.24
15	data_cz1	68.27	89.37	79.71	88.72	84.61	50.00
16	data_cz2	85.57	92.72	92.72	92.72	92.72	50.00
17	data_dmc	61.41	67.12	56.45	65.53	57.41	50.00
18	data_don	78.88	75.57	81.79	75.57	81.79	50.00
19	data_drt	60.54	55.43	50.00	47.16	55.77	60.87
20	data_eur	48.89	54.44	44.44	44.44	48.89	50.00
21	data_fa1	72.42	72.27	69.01	70.69	69.70	50.38
22	data_fa2	75.26	74.88	72.45	73.95	70.88	50.76
23	data_fat	41.67	45.83	37.50	45.83	37.50	50.00
24	data_fd1	61.21	58.01	60.82	58.41	57.08	50.00
25	data_fd2	64.81	65.00	66.72	62.91	61.84	50.00
26	data_hs1	64.74	63.10	62.87	53.80	49.80	57.32
27	data_hs2	71.59	69.89	71.59	60.85	77.29	67.70
28	data_in1	73.74	91.58	90.65	90.51	92.00	84.56
29	data_in2	73.65	91.88	91.51	90.26	92.68	85.25
30	data_jkl	61.63	70.71	69.79	61.49	71.91	71.91
31	data_kw1	60.21	61.71	50.00	66.98	50.41	66.24
32	data_kw2	65.54	72.77	50.00	75.77	58.40	75.59
33	data_nba	83.33	77.78	72.22	72.22	72.22	88.89
34	data_nhl	62.50	62.50	50.00	75.00	50.00	50.00
35	data_pap	72.66	63.69	50.00	61.85	66.59	70.82
36	data_pod	58.05	66.26	50.00	67.23	45.48	49.99
37	data_prc	75.00	41.25	50.00	62.50	32.50	65.00
38	data_red	69.43	65.55	62.29	63.02	58.20	64.87
39	data_ret	60.96	68.02	63.61	65.52	62.50	50.07
40	data_rl1	69.33	70.00	77.00	73.33	77.00	64.00
41	data_rl2	70.76	72.74	72.04	69.83	72.04	50.00

42	data_rl3	69.97	68.00	74.06	64.75	76.70	53.37
43	data_rl4	75.22	65.51	75.31	68.81	75.31	62.92
44	data_rl5	64.90	59.96	56.01	56.48	56.01	50.00
45	data_rt1	67.30	65.75	61.96	64.28	59.05	57.85
46	data_rt2	64.78	62.19	57.93	62.42	57.43	49.85
47	data_sal	37.86	33.39	90.71	39.46	49.46	50.00
48	data_sfr	60.31	67.48	60.31	72.81	66.43	50.00
49	data_sps	50.39	50.49	50.00	50.74	49.26	50.00
50	data_ter	73.79	74.38	71.08	73.92	69.87	50.00
51	data_tf1	55.67	65.36	55.42	61.88	54.08	50.00
52	data_tf2	55.81	64.52	56.02	63.32	58.04	50.00
53	data_tor	57.06	59.71	82.65	55.59	74.12	50.00
54	data_trn	81.18	82.91	85.53	83.06	82.47	75.78
55	data_tsh	50.47	50.29	50.00	50.00	50.00	52.83
56	data_ucl	58.24	54.55	50.00	68.18	40.34	57.10
57	data_ufo	53.44	60.28	50.00	61.54	50.20	50.00
58	data_uni	76.50	83.51	78.76	78.76	78.76	50.00

Table B2. Balanced Accuracy with Optimized Thresholds

#	Data set	Hiatus	LR (All)	LR (R)	RF (All)	RF (R)	P/NBD
1	data_air	66.02	72.25	71.26	72.89	71.26	71.63
2	data_app	66.50	75.41	66.91	73.77	61.32	69.05
3	data_bn1	79.32	79.50	79.99	73.45	77.73	75.04
4	data_bn2	70.76	69.48	70.76	65.37	68.77	70.83
5	data_bn3	66.16	70.86	66.16	56.05	48.60	64.45
6	data_cdn	66.10	69.76	63.65	70.48	64.13	67.75
7	data_cl1	79.81	81.34	79.81	79.73	75.34	79.79
8	data_cl2	79.96	80.41	81.80	77.03	76.66	78.11
9	data_cl3	77.45	78.02	77.34	75.54	72.91	76.69
10	data_cl4	80.23	80.86	80.36	75.31	77.99	80.07
11	data_cl5	79.20	79.09	79.02	76.67	76.07	77.50
12	data_com	58.11	68.85	60.10	55.80	43.00	50.00
13	data_cpn	79.76	81.09	79.64	77.60	77.88	77.97
14	data_ctr	69.05	70.30	69.05	72.27	69.09	68.55
15	data_cz1	84.53	95.16	84.53	87.84	82.36	50.00
16	data_cz2	99.43	99.58	99.43	92.72	92.57	50.00
17	data_dmc	61.28	66.88	60.36	64.20	57.89	50.00
18	data_don	80.35	75.16	81.79	75.57	81.79	50.00
19	data_drt	58.19	54.01	57.02	49.58	55.77	60.87
20	data_eur	43.33	44.44	48.89	44.44	48.89	50.00
21	data_fa1	71.12	73.36	71.12	72.29	71.36	73.18
22	data_fa2	74.38	75.49	74.73	74.46	73.49	75.40
23	data_fat	37.50	50.00	37.50	45.83	16.67	41.67
24	data_fd1	56.93	61.04	55.16	59.01	57.27	50.00
25	data_fd2	60.96	67.15	60.85	59.50	61.86	50.00
26	data_hs1	84.16	83.67	84.96	53.77	73.21	78.90
27	data_hs2	77.77	78.03	80.88	56.37	83.70	79.15
28	data_in1	88.58	91.26	89.11	75.22	91.48	88.53
29	data_in2	89.35	92.01	89.88	79.01	91.77	88.76
30	data_jkl	72.34	64.75	72.34	56.81	71.91	73.40
31	data_kw1	54.14	60.90	55.14	63.33	55.40	57.55
32	data_kw2	62.44	67.28	62.44	65.47	56.81	75.33
33	data_nba	72.22	77.78	72.22	55.56	72.22	72.22
34	data_nhl	62.50	62.50	62.50	75.00	50.00	62.50
35	data_pap	74.81	75.39	75.11	59.75	74.81	71.10
36	data_pod	58.73	69.79	58.73	50.44	45.74	59.83
37	data_prc	65.00	45.00	25.00	53.75	32.50	57.50
38	data_red	64.97	66.56	65.08	67.82	64.56	65.41
39	data_ret	62.95	68.57	62.95	63.65	61.80	66.74
40	data_rl1	77.00	70.00	77.00	71.67	77.00	75.67
41	data_rl2	63.57	72.40	63.57	63.93	63.57	71.61

42	data_rl3	70.70	73.77	70.70	74.16	70.70	52.60
43	data_rl4	75.31	65.51	75.31	75.58	75.31	66.67
44	data_rl5	56.01	62.29	56.01	65.29	56.01	55.54
45	data_rt1	64.25	67.70	64.16	63.47	62.80	65.52
46	data_rt2	63.32	65.91	63.33	63.77	58.29	63.99
47	data_sal	9.29	38.39	70.71	37.14	49.46	50.00
48	data_sfr	67.22	73.95	67.22	69.32	67.22	50.00
49	data_sps	51.06	50.98	50.24	50.46	48.05	50.00
50	data_ter	66.08	74.69	69.87	74.14	69.87	73.56
51	data_tf1	55.76	63.39	55.76	58.15	53.90	50.00
52	data_tf2	55.39	64.21	55.39	59.46	52.51	50.00
53	data_tor	78.24	62.65	81.18	55.59	74.12	50.00
54	data_trn	83.01	84.83	83.01	84.38	79.60	84.87
55	data_tsh	50.42	50.23	50.00	50.36	50.00	50.01
56	data_ucl	47.73	65.62	47.73	68.75	44.89	47.73
57	data_ufo	55.40	58.23	55.83	57.53	50.34	50.00
58	data_uni	90.51	83.51	78.76	79.00	78.76	50.00

Appendix C: Clumpiness

Clumpiness refers to the degree to which the duration between events is identical or clumps together for certain intervals. Recent studies show that accounting for clumpiness can improve the predictive power of models (Platzer and Reutterer 2016; Zhang, Bradlow, and Small 2015). Consider Figure A1, which depicts the purchase behavior of three artificial entities. Entity I is considered to be less clumpy because there are regular time intervals between the purchases, whereas entity III is the most clumpy. Intuitively, higher event regularity should also result in higher consistency in the event occurrence across different time intervals, which could help the recency-based hiatus heuristic to perform better: a less clumpy entity is more likely to demonstrate a consistent behavior during $[T_{1/4},\ T_{1/2})$ and $[T_{1/s},\ T_1)$ and would therefore be correctly predicted by the heuristic (see entity I on Figure A1). However, the opposite case is also possible: compared with entity III, event sequence II is less clumpy but it would be incorrectly predicted as non-repeated. Therefore, the effect of event regularity on the predictive performance is uncertain. Also, the results show that clumpiness does not affect the performance of any of the models used.

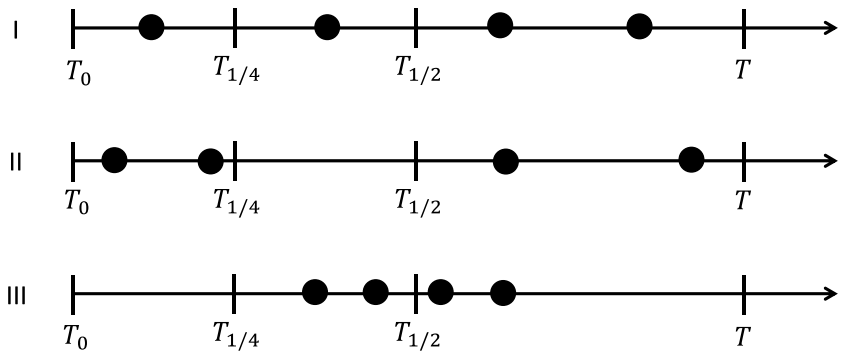

Figure A1. Clumpiness

In this study, we use a clumpiness measure suggested by Zhang et al. (2013). The clumpiness C_j for entity j is computed as:

$$C_j = \frac{\sum_{i=1}^{T+1} ln(x_i)x_i}{\ln\,(T+1)}$$

where x_i is the normalized inter-event times, *T* is the total number of time units, and *j* is the
customer index. The inter-event intervals are defined as:

$$x_i = \begin{cases} \frac{t_i}{T} \; if \; i = 1 \\ \frac{t_i - t_{i-1}}{T} \; if \; 1 < i < k \\ \frac{T - t_i}{T} \; if \; i = k \end{cases}$$

where t_i is the time of *i*-th event and k is the total number of events for this entity. The proposed clumpiness measure can take values between 0 and 1. Higher values of C_j refer to clumpier entities, that is, those with more irregular inter-event time intervals. Direct comparison of clumpiness values across customers with different purchase frequencies may be misleading given that the number of events is one of the parameters of C_j (Zhang, Bradlow, and Small 2015). The authors suggest using *p*-values that can be derived via Monte Carlo simulations for comparison. We follow their framework and compute *p*-values for the statistical test with a null hypothesis that a given event is clumpy. Finally, we compute the share of clumpy entities in a given dataset based on a 5% significance level.

Appendix D. Aggregated Results: Median Performance

This Appendix compares the median predictive performance of the hiatus heuristic and other prediction methods. Figure A2 (a) illustrates the results for 27 data sets coming from the retail domain, whereas Figure A2 (b) depicts the performance on 31 non-retail data sets.

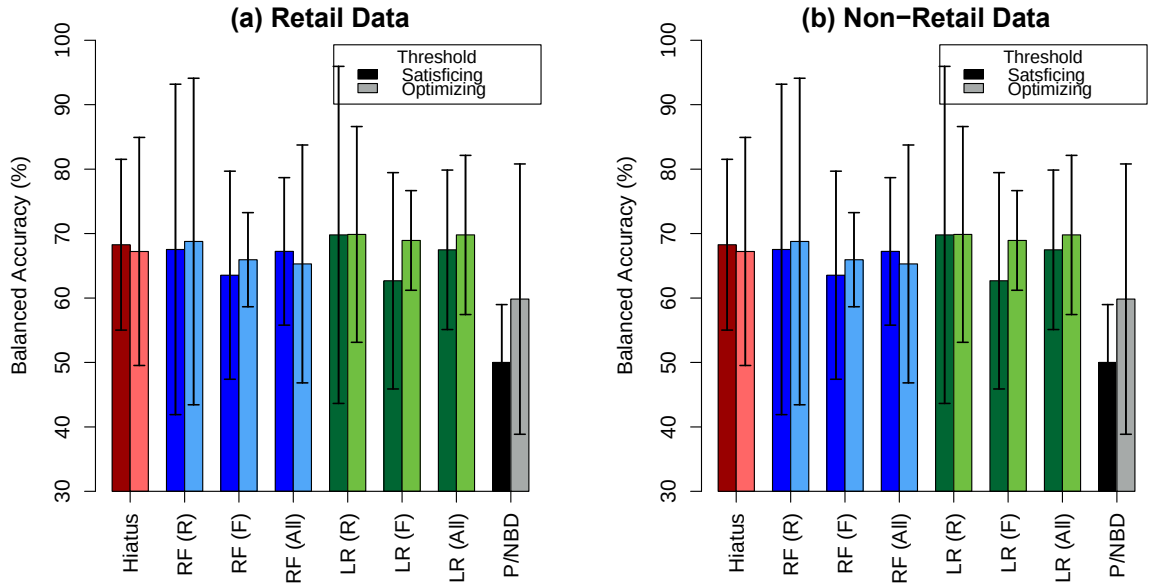

Figure A2. Comparing Median Performance

Note. The figure depicts empirical results for 27 retail data sets. Panel (a) depicts median BACC ± one interquartile range for retail data, whereas panel (b) indicates performance on non-retail data. Both satisficing and optimized thresholds are displayed. Abbreviations: RF = random forest, LR = logistic regression, P/NBD = Pareto/NBD; (R) = model that uses only recency, (F) = model that uses only frequency, (All) = model that uses all variables.

Appendix E. Aggregated Results: Optimizing

Figures A3 and A4 compare the aggregated performance of prediction methods along with the individual performance for each data set separately and display the best models with satisficing thresholds from each model family. With respect to the retail data sets depicted in Figure A3, counting how often a given model performs best reveals the following: LR (All) is best in 63% of the data sets, whereas the hiatus heuristic is best in 15% of the data sets. On non-retail data (see Figure A4), LR (All) is best in 45% of the data sets, whereas the hiatus heuristic demonstrates the best BACC in 16% of the cases.

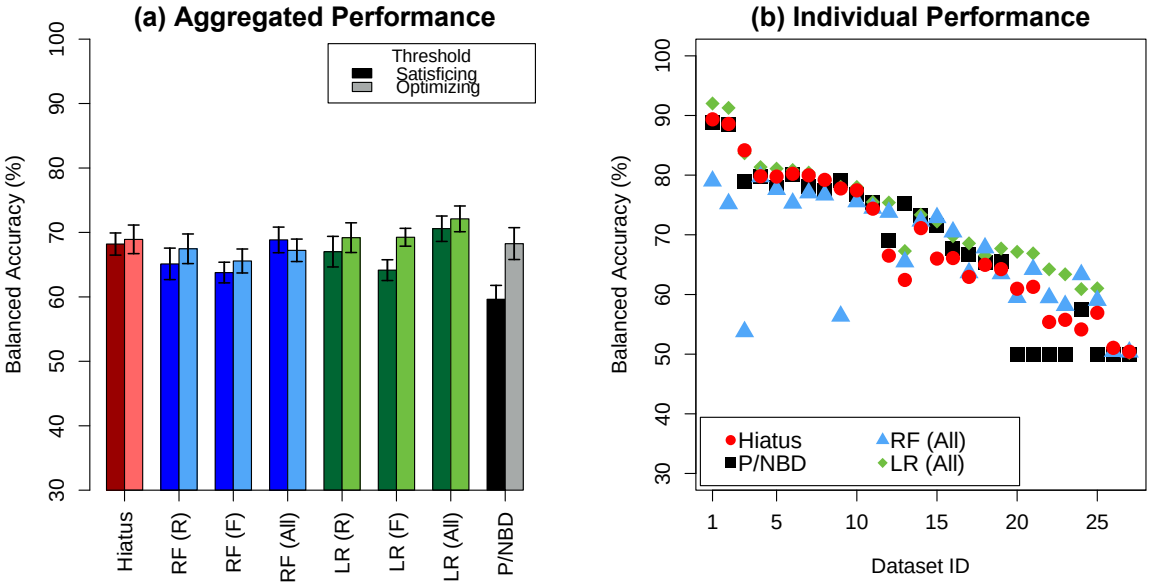

Figure A3. Comparing Performance Across 27 Retail Data Sets

Note. The figure depicts empirical results for 27 retail data sets. Panel (a) depicts mean BACC ± one standard error for each model. Both satisficing and optimized thresholds are displayed. Panel (b) plots the BACC of selected models with the optimizing thresholds for each of the 27 data sets. The data sets are sorted by the maximum observed accuracy. Abbreviations: RF = random forest, LR = logistic regression, P/NBD = Pareto/NBD; (R) = model that uses only recency, (F) = model that uses only frequency, (All) = model that uses all variables.

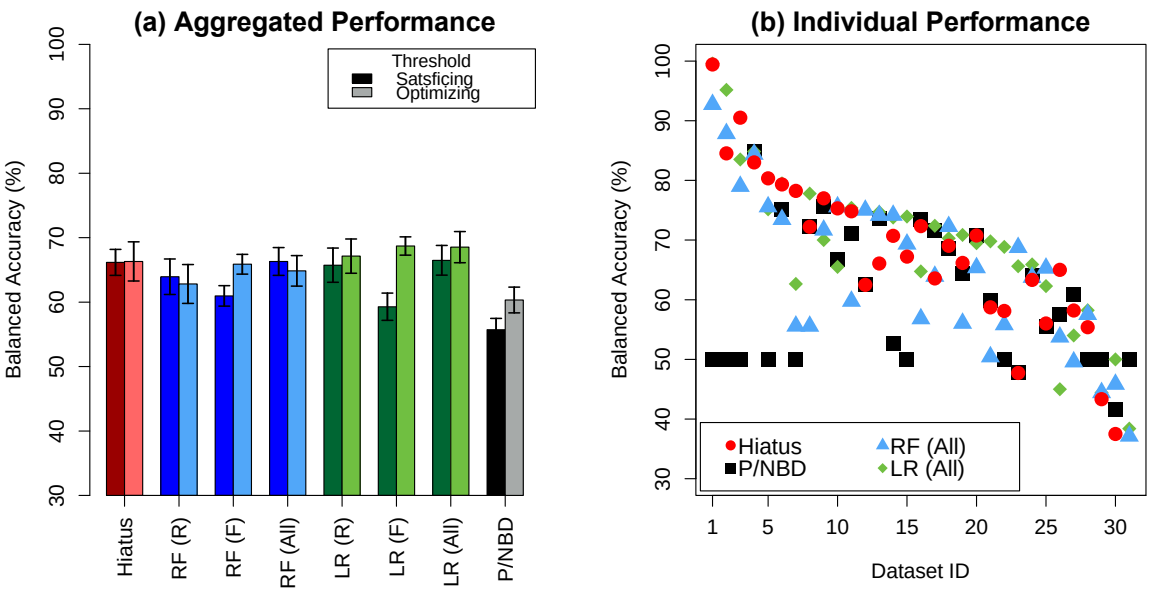

Figure A4. Comparing Performance Across 31 Non-Retail Data Sets

Note. The figure depicts empirical results for 31 non-retail data sets. Panel (a) depicts mean BACC ± one standard error for each model. Both satisficing and optimized thresholds are displayed. Panel (b) plots the BACC of selected models with the optimizing thresholds for each of the 31 data sets. The data sets are sorted by the maximum observed accuracy. Abbreviations: RF = random forest, LR = logistic regression, P/NBD = Pareto/NBD; (R) = model that uses only recency, (F) = model that uses frequency only, (All) = model that uses all variables.