An Accuracy Rating System for Discrete Probability Predictions Using Sportsbook Odds for Super Bowl LIX

Nicholas A. Beaver

NicholasABeaver.com

23 June 2025

ABSTRACT

This paper presents a practical accuracy rating system for evaluating discrete probability predictions. The approach assigns credit based on the probability given to the correct outcome, with an optional time-weighting feature that rewards earlier, riskier predictions more than late-breaking ones. The model uses Bayesian methods to estimate each predictor's skill level and determine the probability that they are doing better than random guessing. A baseline comparison against a discrete uniform probability distribution (DUPD) enables objective performance evaluation. This framework is flexible enough to manage real-time updates and live predictions, merging them with pre-event forecasts into a single, interpretable metric. The model is demonstrated by evaluating Super Bowl LIX predictions from two leading U.S. sportsbooks and the prediction market Polymarket. The result is a robust, scalable system for benchmarking predictive performance.

Keywords: Prediction accuracy, probabilistic forecasting, forecast evaluation, Bayesian inference, beta distribution, time-weighted scoring, predictive modeling, discrete probability, prediction credibility

TABLE OF CONTENTS

I. INTRODUCTION

Probability assignments are frequently seen in the realm of gambling, most prominently in events wagering (i.e., sports betting). The odds quoted by sportsbooks give implied probabilities of outcomes, and good ("sharp") bettors compare these implied probabilities with their own calculations before making a bet. Probability assignments are also used frequently in the fields of medicine, to give the probability of certain diagnoses, in political elections to predict the winning candidate, in finance to forecast the probability of macroeconomic events, and to gauge the accuracy of predictions made by logistic regression techniques and Bayesian networks. We, as the believing public, are supposed to take these predictions as expert forecasts about the probability of certain outcomes and are supposed to use these forecasts to guide our behavior accordingly.

Less frequently discussed are measures of prediction accuracy. If a predicting agent, in this paper a "predictor", assigns a probability, say "0.5491", to a certain outcome, call it "Outcome X", who among us recalls whether the predictor was correct after the event has occurred? And if the predictor was correct *this* time, what about the predictor's other predictions? What about the predictions the predictor has made in the past or has yet to make? How do these bear on the accuracy of the predictor's forecasts?

This paper outlines a method to rate the accuracy of such forecasts. It is applicable to any prediction that gives a discrete probability, such as "the probability of X is 0.5491" or "the probability that Y happens during Z time is 0.3244", and not to continuous probability assignments, such as "the probability of A is between 0.1548 and 0.2270" or "$P(B) \sim \mathcal{N}(0.2417, 0.032)$", the latter of which means that "the probability that B is true is a probability distribution taking the form of a Gaussian distribution with a mean of 0.2417 with a standard deviation of 0.0332."

With the methods outlined in this paper, the reader may be able to answer such questions as:

1. How accurate is a given prediction?
2. How accurate has a given predictor been, over time?
3. Given what we know, how likely is a given predictor to be right in the future?
4. What sort of probability distribution best describes how accurate a predictor is?
5. What happens if a predictor, or set of competing predictors, makes a *series* of predictions over time? How do these influence the predictor's overall accuracy?
6. How much should we believe a prediction from a predictor, given what we know about that predictor's accuracy?

We will return to answering these questions, directly, at the conclusion of this paper, using what has been established throughout this paper to help guide us on how to best answer them.

This paper will also give a concrete, real world example of the implementation of this accuracy rating system. The predictions, via the implied probability assignments from odds quotations, made by major sportsbooks operating in the United States and the peer-to-peer prediction network Polymarket, are evaluated. The accuracy rating system is employed to show the accuracy of the various predictors (that is, the sportsbooks and users of Polymarket) to predict the winner of Super Bowl LIX, played in February 2025. Conclusions are then drawn about the probable accuracy of these predictors predicting the outcome of this single event.

It is the hope of this paper that a use for this accuracy rating system could be found in any area where discrete predictions are made. Economic forecasts, health diagnoses, political elections, and of course, sports betting odds could all be evaluated with this system. Many more applications, it is certain, could also be found.

II. GLOSSARY OF TERMS

In this paper, we use several terms of art which are defined here.

Term	**Definition**
Predictor	A predicting agent; i.e., a person or entity that makes predictions.
Prediction	The discrete probability assignment given to an outcome by a predictor; e.g. "The Eagles winning the Super Bowl has a probability of 0.4752" is a prediction.
DUPD	The *discrete uniform probability distribution* that assigns an equal probability to each possible outcome of an event. For instance, if an event has two possible outcomes, either A or B, the DUPD is $Unif\{A, B\}$ and the probability of A is 0.5000 and the probability of B is 0.5000. (Read: *D.U.P.D.*)
Â	Prediction accuracy. This is a measure of how accurate a predictor's predictions have been as measured by the methods described in this paper. (Read: *A-hat*).
Â'	Posterior prediction accuracy beta distribution. This is a beta distribution describing the posterior probability of a predictor's prediction accuracy, formed by a prior based on a discrete uniform probability distribution and a likelihood based on the predictor's observed prediction accuracy. (Read: *A-hat prime*).

P(Â') > DUPD	Probability that a predictor's prediction accuracy beats the DUPD; e.g., if there are two possible outcomes from a contest, the DUPD is 0.5000 (that is, a 50% probability of either outcome); *P(Â') > DUPD* thus finds the probability that a predictor can beat an accuracy of 0.5000. (Read: *probability of A-hat prime greater than the D.U.P.D*).

III. RATING SYSTEM: STATIC METHOD

A. Overview of the Static Method

The rating system relies on a system of "credit." For each prediction made by a predictor, the predictor gains an amount of credit equal to the probability assigned to the real outcome. The total accrued credit is then compared to the total value of the events for which predictions were made.

The basic method of rating prediction accuracy is as follows:

$$\text{Accuracy } (\hat{A}) = \frac{\text{Credit (c)}}{\text{Events (N)}} \tag{1}$$

Let us consider an example of a predictor, called Predictor A, making a prediction about an event, called E_1. In E_1, two competitors, X_1 and Y_1, compete in a zero-sum game (meaning one will be the winner, and one will be the loser; or, possibly, both may tie, of which more, later). Predictor A assigns a probability to each X_1 and Y_1 to win:

Table 1. Predictor A's Assigned Probabilities for X_1 and Y_1 in E_1.

Competitor	Assigned Probability
X_1	0.7745
Y_1	0.2255

The outcome of the zero sum contest has a value of 1 (either X_1 or Y_1 will be given a value of 1 for a win, and the other a 0 for a loss; in the case of a tie, both will be given a value of 0.5). Suppose that X_1 wins E_1. We assign Predictor A's "credit" as follows:

Table 2. Predictor A's "Credit" After E_1 Results in a Win for X_1.

Competitor	Assigned Probability	Outcome	Credit
X_1	0.7745	1	0.7745
Y_1	0.2255	0	0.0000

Using Equation 1, we find that Predictor A's accuracy, $\hat{A}$, is 0.7745 for this run of just one event.

Let us further illustrate this basic method by evaluating Predictor A over four additional events. Each new event, denoted as E_i (where i = 2, …, 5), involves two competitors, X_i and Y_i.

Table 3. Predictor A's Assigned Probabilities for Four Additional Events.

	Assigned Probabilities	
Event	**X_i**	**Y_i**
E_2	0.5814	0.4186
E_3	0.5219	0.4781
E_4	0.6397	0.3603
E_5	0.4876	0.5124

The outcomes of these contests are as follows: E_2 is won by X_2, E_3 is won by Y_3, E_4 is won by X_4, and E_5 is won by Y_5. To figure Predictor A's accuracy ($\hat{A}$) over E_1 to E_5, we apply the following equation:

$$\hat{\mathrm{A}} = \frac{\sum c_i}{\sum N_i} \tag{2}$$

Where c_i is the credit for each contest, *i,* and N_i is the value of each contest, *i*. To compute Predictor A's accuracy, we take the data from Tables 2 and 3 detailing events E_1 through E_5 and find:

$$\hat{\mathrm{A}}_{A|E_{1\ldots}E_5} = \frac{\sum c_i}{\sum N_i} = \frac{0.7745 + 0.5814 + 0.4781 + 0.6397 + 0.5124}{5} = 0.5972 \tag{3}$$

Thus, we can say that Predictor A's accuracy ($\hat{A}$) is 0.5972 over E_1 to E_5.

B. What Happens if There Are More Than Two Outcomes?

There may be more than two anticipated outcomes. For instance, in an election, there might be more than two candidates, etc. In these cases, the static system still works the same way.

Suppose that there is a political election, E_6, with four candidates: W_6, X_6, Y_6, and Z_6. Predictor A assigns the following probabilities to each candidate to win the election:

Table 4. Predictor A's Assigned Probabilities for Four Political Candidates in an Election.

Candidate	**Assigned Probability**
W_6	0.2211
X_6	0.1458
Y_6	0.2962
Z_6	0.3369

Suppose that Y_6 wins E_6. Predictor A's accuracy for E_6 is 0.2962. This sounds bad, but recall, a discrete uniform probability distribution (DUPD), namely *Unif{W_6, X_6, Y_6, Z_6},* in this

case would assume a probability of 0.2500 for each of the four possible outcomes. Considering this one event, Predictor A has not done all that badly.

We will investigate how a discrete uniform probability distribution helps us evaluate accuracy in Section *V: Accuracy as a Beta Distribution.*

C. Why This Works

If a predictor had perfect predictive capabilities, that predictor would assign a probability of 1 to the eventual outcome of an event. This would make the predictor's accuracy equal 1 (a perfect score). Predictors, of course, are free to do this.

Predictors do not do this in practice, because the probabilities assigned are (at least in part) a function of the predictor's belief about the possible outcome of an event; the predictor hedges its bets, so to speak, by spreading its total belief across several outcomes. The more belief the predictor puts into an outcome, the higher the probability it assigns to that outcome.

In short, predictors receive credit for their level of belief in the ultimate outcome of an event. A good (or better) predictor will habitually assign more belief (a higher probability) to the outcome that proves right over time. As we will see in later sections, we also care about the probability that a predictor is really exhibiting some kind of predictive skill and not just "guessing."

IV. RATING SYSTEM: TIME-WEIGHTED METHOD

With the basic, static method established, we now add to this the issue of time-weighting.

Predictors do not always offer just one prediction for an event (one set of probabilities on the possible outcomes) but rather, several predictions that change over time as new information becomes available.

Understanding this, we seek to *weight* predictions such that the further out they are made from the event (i.e., the earlier *before* the event occurs), the more valuable they are, and contrarily, the closer they are made to the event (i.e., the nearer to the event occurring), the less valuable they are. This is because as new information becomes available, it is understood that predictions should be more accurate based on that information, and thus, the work of prediction is made easier over time. It is more a testament to the predictor's skill if the prediction made is more correct *earlier* rather than later, when information is easier to come by.

A. Time Intervals

To account for time weighting, we first think of the span of time from the first prediction made, call it P_1, to the last prediction made, call it P_n, as a series of discrete time intervals. The intervals can be days, hours, minutes, etc. as best suits the number and frequency of predictions made.

For example, if a prediction about an event is made one year before the event is to occur, and the predictor makes just a small handful of predictions over the span of that year, say eight predictions, we might make the time interval days (in which case, there will be 365 time units, one for each day of the one-year span). On the other hand, if a prediction about an event is made one week before the event is to occur, and the predictor makes a long series of predictions over that week, say twenty, we might make the time interval hours (in which case, there will be 168 time units, one for each hour of the one-week span).

After the number of time intervals, what we term *K*, has been determined, we apply the following equation to each prediction made, from P_1 to P_n. We determine a unique time-weighting factor, called *T,* for each prediction.

$$T_i = 1 - \frac{v \times k_i}{K} \tag{4}$$

Where *T* is the time-weight factor, k_i is the prediction's time interval index, and *K* is the total number of time intervals from P_1 to P_n. In Equation 4, *v* is a weighting value that the last prediction, P_n, is worth in comparison to the value of P_1. Throughout this paper, we will set the value of *v* at 0.5. This means that the last prediction in any series, P_n, will be worth half the value of the first prediction, P_1, in that same series. A discussion on the choice of *v* at 0.5 is given in the Appendix.

The sub *i* in Equation 4 is the index to the time interval. The time intervals should run from 0 (when the first prediction is made) through to *K.* For example, if the first prediction is made 1000 time units before the start of an event (that is, *K* = 1000), the intervals for *k* will run from 0, …, 1,000 in ascending order. This is demonstrated in Table 5.

Table 5. Example Time-Weighting Factor (T_i) Based on a 1,000 Time Unit Interval.

Prediction	**k_i**	**K**	**T_i**
n/a	0	1,000	1.0000
n/a	1	1,000	0.9995
P_1	2	1,000	0.9990
n/a	3	1,000	0.9985
P_2	4	1,000	0.9980
⋮	⋮	⋮	⋮
n/a	998	1,000	0.5010
P_{20}	999	1,000	0.5005
n/a	1,000	1,000	0.5000

The time-weight factor, T_i, is gradually decreasing as predictions are made and time is nearing the event being predicted. Equation 4 weights a series of predictions such that the very first prediction made in the time series is made at full value (i.e., a value of 1), while successive

predictions are made at a linearly decreasing value until the start of the event, at which point a prediction is worth v the value of the first prediction (i.e., $v = 0.5$). The series of 1,000 predictions summarized in Table 5 is illustrated in Figure 1.

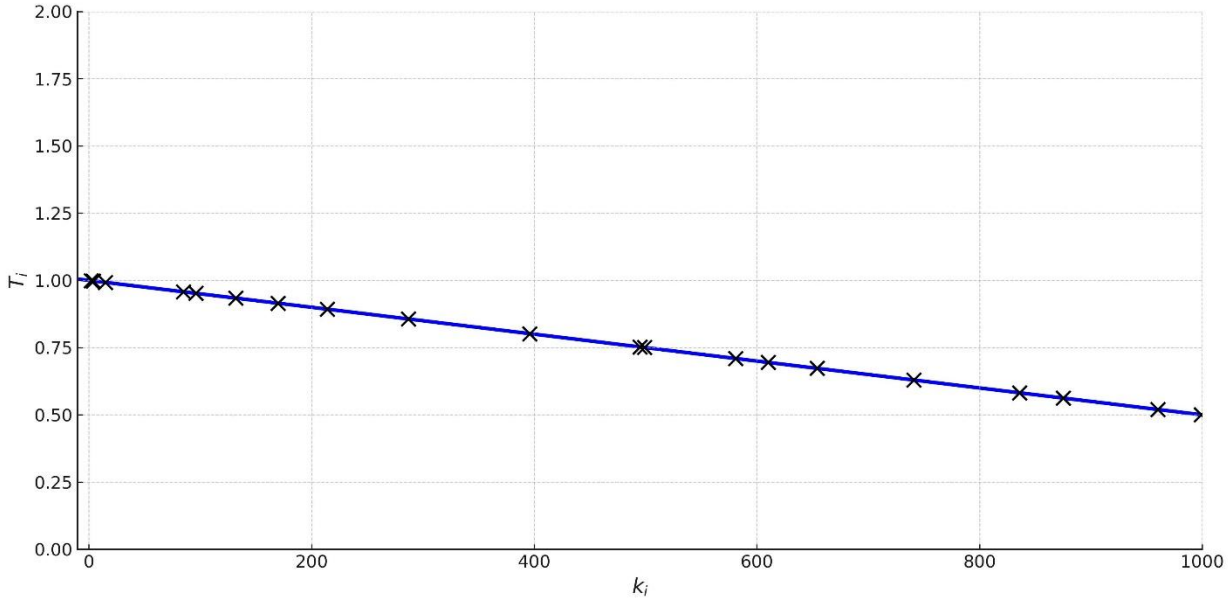

Figure 1. An illustration of Table 5. The line represents the equation $T = 1 - \frac{0.5k}{1000}$ which determines the time-weighted value of each prediction. Twenty predictions are shown as *X*'s on the plot. The first prediction is made a full value (1) while the last prediction, which is made right as the event predicted begins, is made at half value (0.5).

B. Example Time-Weighted Accuracy Rating

Consider an example: Predictor A makes five predictions about an event, E_7, in which X_7 and Y_7 are outcomes. E_7 will occur on Day_6. On each day leading up to the contest, Predictor A makes a new Prediction, as shown in Table 6.

Table 6. Predictor A's Five Predictions on E_7.

Day	**Prediction**	**X_7**	**Y_7**
1	P_1	0.7140	0.2860
2	P_2	0.6352	0.3648
3	P_3	0.6622	0.3378
4	P_4	0.5965	0.4035
5	P_5	0.6084	0.3916

On Day_6, E_7 occurs and X_7 is the outcome. We now compute Predictor A's accumulated credit for the predictions it made. To do this, we multiply the assigned probability give to the outcome, X_7, by each prediction's time-weight factor, T_i, to arrive the time-weighted credit, c_i, each prediction is worth. This is demonstrated on Table 7.

Table 7. Time-Weighting Predictor A's Five Predictions on E_7X_7.

Prediction	Probability (p_i)	T_i	p_i x T_i	Credit (c_i)
P_1X_7	0.7140	1.0	0.7140 • 1.0 =	0.7140
P_2X_7	0.6352	0.9	0.6352 • 0.9 =	0.5717
P_3X_7	0.6622	0.8	0.6622 • 0.8 =	0.5298
P_4X_7	0.5965	0.7	0.5965 • 0.7 =	0.4176
P_5X_7	0.6084	0.6	0.6084 • 0.6 =	0.3650

Figure 2 shows each of the five predictions made over the five days and their corresponding time-weight values on the trend line for Equation 4.

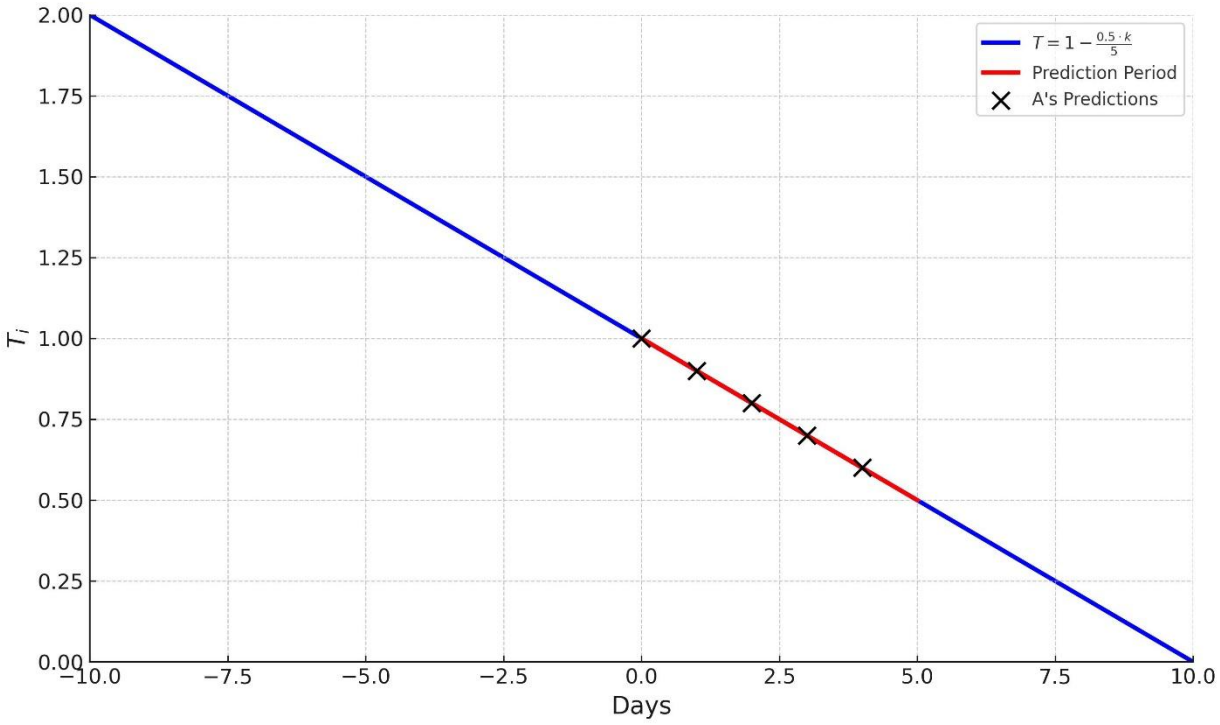

Figure 2. Time-weights for Predictions 1 through 5 (from Table 6).

Predictor A's prediction accuracy is then calculated as a ratio of total adjusted credit to total adjusted value of the contests using the following equation:

$$\hat{A} = \frac{\sum c_i}{\sum T_i} \tag{5}$$

Applying Equation 5 to Predictor A's case (i.e., the data from Table 7) in this example, we find:

$$\hat{A} = \frac{\sum c_i}{\sum T_i} = \frac{2.2764}{4.0} = 0.6495 \tag{6}$$

Predictor A's time-weighted accuracy ($\hat{A}$) was 0.6495 for E_7.

C. Normalization Among Multiple Predictors

What happens when we compare multiple predictors that make predictions at different times (i.e., if one predictor makes its first prediction much earlier or later than another)? How should their separate series of predictions be weighted? This asymmetry is solved through a normalization process in which the multiple compared predictors are put on the same time scale.

Assume that two predictors—Predictor A and Predictor B—make a series of predictions about event E_8 in which X_8 and Y_8 are outcomes. These predictions are shown on Table 8.

Table 8. Predictor A and Predictor B's Predictions about E_8.

	Predictor A		**Predictor B**	
Day	**X_8**	**Y_8**	**X_8**	**Y_8**
1	0.6489	0.3511	n/a	n/a
2	0.5887	0.4113	n/a	n/a
3	0.7495	0.2505	0.7044	0.2956
4	0.7069	0.2931	0.5775	0.4225
5	0.6154	0.3846	0.7170	0.2830
6	0.6063	0.3937	0.6618	0.3382
7	0.7182	0.2818	0.5513	0.4487

E_7 will take place on Day 8. As shown in Table 8, Predictor A began making predictions on Day 1, but Predictor B did not begin until Day 3. To find a common time-weight factor (T_i) for each prediction, we use a time-weight scale from the *earliest* prediction among all predictions (in this case, from Day 1).

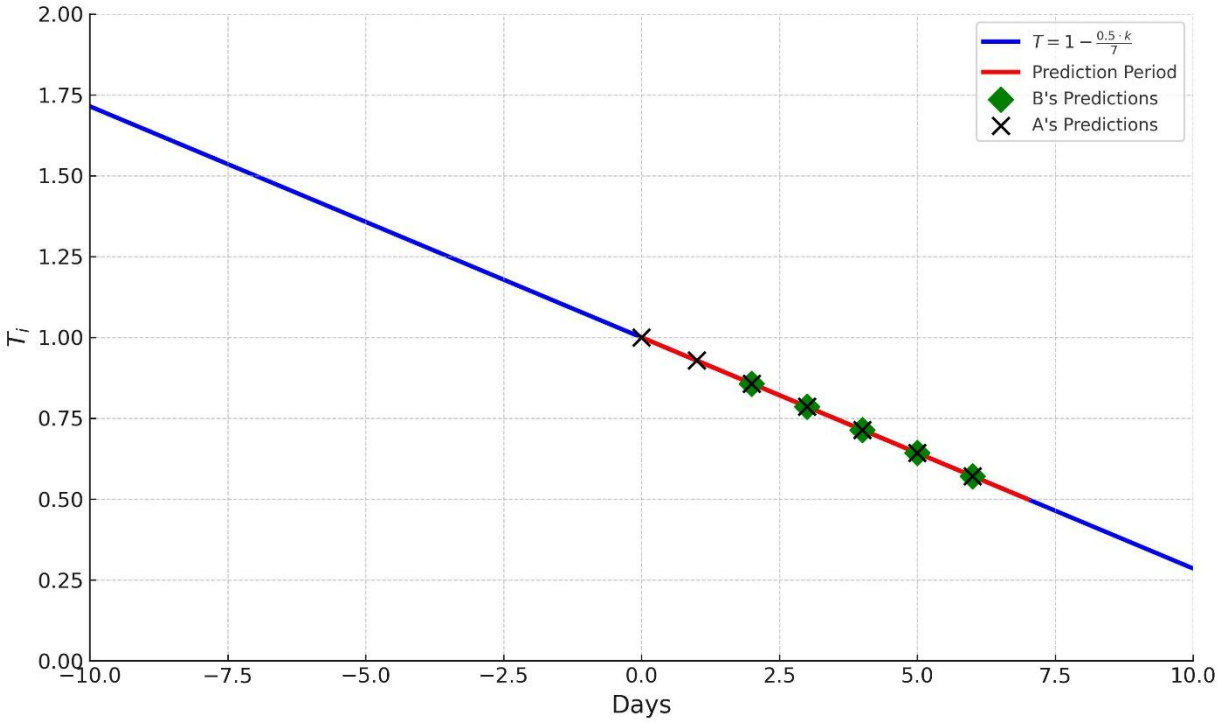

Figure 3. Predictor A and Predictor B's predictions from Table 8. Note that Predictor A makes its first predictions before Predictor B does.

Suppose that the outcome for C_8 is X_8. We now wish to determine each predictor's accuracy. Applying our previous methods, and normalizing both predictors' predictions to Day 1, we arrive at time weights as shown in Table 9.

Table 9. Credit for Predictions by Predictors A and B.

	Predictor A			Predictor B		
T_i	**Prediction**	**p_i**	**c_i**	**Prediction**	**p_i**	**c_i**
1.0000	P_1X_8A	0.6489	0.6489	n/a	n/a	n/a
0.9286	P_2X_8A	0.5887	0.5467	n/a	n/a	n/a
0.8571	P_3X_8A	0.7495	0.6424	P_1X_8B	0.7044	0.6038
0.7857	P_4X_8A	0.7069	0.5554	P_2X_8B	0.5775	0.4538
0.7143	P_5X_8A	0.6154	0.4396	P_3X_8B	0.7170	0.5121
0.6429	P_6X_8A	0.6063	0.3898	P_4X_8B	0.6618	0.4254
0.5714	P_7X_8A	0.7182	0.4104	P_5X_8B	0.5513	0.3150

Using Equation 5, we can compute each predictor's accuracy as follows:

$$\hat{A}_{A,E_8} = \frac{\sum c_i}{\sum T_i} = \frac{3.6331}{5.5} = 0.6606 \tag{7}$$

$$\hat{A}_{B,E_8} = \frac{\sum c_i}{\sum T_i} = \frac{2.3101}{3.2143} = 0.7187 \tag{8}$$

We thus find that Predictor A's accuracy ($\hat{A}_{A,E8}$) for E_8 was 0.6606 and Predictor B's accuracy ($\hat{A}_{B,E8}$) for E_8 was 0.7187. Predictor B was more accurate.

V. ACCURACY AS A BETA DISTRIBUTION

We can visualize predictors' accuracy as a beta distribution to later probe that distribution to find the probability that a given predictor is better than expected by a discrete uniform probability distribution (DUPD). We will first see how to set a prior distribution, a likelihood based on the data, and finally, a posterior distribution that we call *Â'*.

A. Example Data for Finding Â'

We assume that we wish to find Predictor A and Predictor B's *Â'* for a new event, E_9, in which there were two possible outcomes: X_9 and Y_9. We have the following known data:

Table 10. Value of Predictions Made by Predictor A and Predictor B for E_9.

	Predictor A	**Predictor B**
Predictions	15	22
Σc_i	8.8110	12.0516
ΣT_i	13.4835	20.0706

According to Table 10, Predictor A made 15 predictions about E_9 for which it received "credit" of 8.8110 and the total time-weighted value of the predictions was 13.4835. For the same event, Predictor B made 22 predictions for which it received "credit" of 12.0516 and for which the total time-weighted value of those predictions was 20.0706.

We will use the data from Table 10 to form our prior, likelihood, and posterior beta distributions.

B. Prior Distribution

The prior distribution should assume a DUPD for any set of predictions. Since E_9 had two possible outcomes, that is $Unif\{X_9, Y_9\}$, the discrete uniform probability distribution would assign a probability of 0.5000 to X_9 and a probability of 0.5000 to Y_9, meaning that the prior assumes that a predictor will be right with 0.5000 probability.

If there were three possible outcomes, instead, the prior distribution would assign a probability of 0.3333 to each outcome, meaning that the prior would assume that a predictor would be right with 0.3333 probability and wrong with 0.6667 probability. This carries for any number of outcomes in each set of predictions.

The prior should conform to the following equation:

$$\pi(\hat{A}') \sim Beta\left(\frac{\sum T_i}{\omega}, \sum T_i \times (\omega - 1)\right) \tag{9}$$

Where ω is the number of possible outcomes for the set of predictions.

The prior for E_9 for each predictor is different, based on the total value of the predictions (T_i) for each predictor. Again, we use the information from Table 10.

The prior distribution for Predictor A is:

$$\pi(\hat{A}')_{A,E_9} \sim Beta(6.7418, 6.7418) \tag{10}$$

And the prior distribution for Predictor B is:

$$\pi(\hat{A}')_{B,E_9} \sim Beta(10.0353, 10.0353) \tag{11}$$

C. Likelihood Distribution

Each predictor's likelihood distribution is based on the observed accuracy. The likelihood for each predictor takes the form of a binomial distribution:

$$L(\hat{A}') \sim Binomial(c_i, T_i - c_i) \tag{12}$$

Using the information from Table 10, we find that for E_9, the likelihood distribution for Predictor A is:

$$L(\hat{A}')_{A,E_9} \sim Binomial(8.110, 5.3735) \tag{13}$$

And the likelihood distribution for Predictor B is:

$$L(\hat{A}')_{B,E_9} \sim Binomial(12.0516, 8.0190) \tag{14}$$

These distributions count the "credit" earned by each predictor as the "successes" (i.e., the first parameter of the likelihood distribution) and the total value of the predictions minus the "credit" earned as the number of "failures" (i.e., the second parameter of the likelihood distribution) for each binomial distribution.

D. Posterior Distribution ($\hat{A}$')

With a beta prior and binomial likelihood, we combine these using Bayesian inference to arrive at a beta posterior distribution.

$$p(\hat{A}') \sim Beta(\alpha + s, \beta + f) \tag{15}$$

That is, the resulting α of the posterior distribution is a combination of the prior's α plus the number of "successes," s, from the likelihood and the resulting β of the posterior distribution is a combination of the prior's β plus the number of "failures," f, from the likelihood distribution.

Thus, the posterior distribution for Predictor A is:

$$p(\hat{A}')_{A,E_9} \sim Beta(14.8518,12.1153\ ) \quad (16)$$

And the posterior distribution for Predictor B is:

$$p(\hat{A}')_{B,E_9} \sim Beta(22.0869,18.0543\ ) \quad (17)$$

A summary of the prior, likelihood, and posterior distributions for both predictors is shown in Table 11.

Table 11. Prior, Likelihood, and Posterior Beta Distributions for Predictors A and B for E_9.

Distribution	**Predictor A**	**Predictor B**
Prior (π)	Beta(6.7418, 6.7418)	Beta(10.0353, 10.0353)
Likelihood ($\mathcal{L}$)	Binomial(8.110, 5.3735)	Binomial(12.0516, 8.0190)
Posterior (p)	Beta(14.8518, 12.1153)	Beta(22.0869, 18.0543)

With our posterior distributions, $\hat{A}'_A$ and $\hat{A}'_B$, we move now to probing these distributions to find the probability that the predictors' accuracy beats the DUPD.

VI. PROBABILITY OF BEATING A DISCRETE UNIFORM PROBABILITY DISTRIBUTION (DUPD)

With each posterior distribution determined, $\hat{A}'_A$ and $\hat{A}'_B$, we can find the probability that each predictor's accuracy beats a discrete uniform probability distribution, a term we call $P(\hat{A}') > DUPD$.

A. Computing $P(\hat{A}') > DUPD$

Again, in our example of E_9, there were two outcomes: X_9 and Y_9. This made the DUPD for E_9 0.5000 (1 divided by the number of possible outcomes). If there were three possible outcomes, say W_9, X_9, and Y_9, the DUPD would assume a probability of 0.3333 for each of these three outcomes (1 divided by the number of possible outcomes). Likewise, if there were four possible outcomes, say W_9, X_9, Y_9, and Z_9, the DUPD would assume a probability of 0.2500 for each of these four outcomes (again, 1 divided by the number of possible outcomes).

For E_9, we wish to find the probability that each predictor's prediction accuracy beats the DUPD of 0.5000. In other words, we wish to know $P(\hat{A}') > 0.5000$ for each predictor. We find that if we probe each predictor's posterior beta distribution, $\hat{A}'_A$ and $\hat{A}'_B$, respectively, we can determine this probability.

Table 12. $P(\hat{A}') > DUPD$ for Predictors A and B.

	Predictor A	**Predictor B**
$P(\hat{A}') > DUPD$	0.7039	0.7406

The two predictors' $P(\hat{A}') > DUPD$ are visualized in Figure 4.

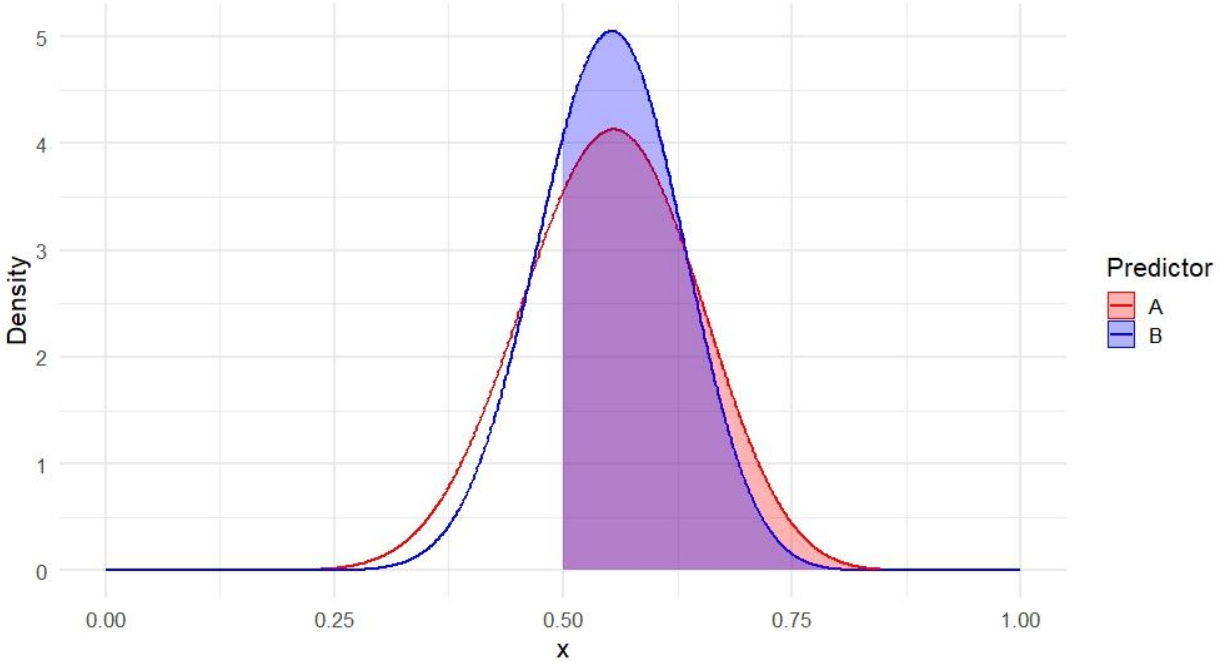

Figure 4. Visualization of the two predictors' posterior beta distributions. The shaded region in each shows $P(\hat{A}) > 0.5000$.

According to the results from Table 12, the probability that Predictor A beats the DUPD is 0.7039 and the probability that Predictor B beats the DUPD is 0.7406. Both predictors have a high probability of demonstrating real predictive skill. But, we have more confidence in Predictor B's ability to make accurate predictions because of not only its demonstrated predictive accuracy $(\hat{A})$, but the higher number of predictions Predictor B has made compared to Predictor A.

EXAMPLE IMPLEMENTATION: SUPER BOWL LIX

To demonstrate the use of this model, predictions made by the two largest U.S.-based sportsbooks and the peer-to-peer prediction market Polymarket are evaluated. The predictions come in the form of odds quotations, that is, gambling odds quotations, from these predictors centered around Super Bowl LIX.

i. Predictors & Event

We investigate the predictions made by two major U.S. sportsbooks—DraftKings and FanDuel—and the prediction market Polymarket about the winner of Super Bowl LIX played on February 9, 2025, at 23:30 hours UTC between the National Football League (NFL) teams the Kansas City Chiefs and the Philadelphia Eagles.

The Philadelphia Eagles won the contest. All accuracy rating is thus predicated on the assigned probabilities given by each predictor for the Philadelphia Eagles to win.

ii. Data

Data for the sportsbooks was obtained from historical "snapshots" of odds quoted The Odds-API [1][2]. This dataset represents 31 unique odds quotations for both participating teams. Data from Polymarket was obtained directly from Polymarket [3]. The Polymarket data includes 36 unique odds quotations for both participating teams. These 67 unique odds quotations form the full dataset used in this example analysis.

iii. Time Interval

The first prediction made by the three predictors was made by Polymarket on January 27, 2025, at 18:00 hours UTC. The event, Super Bowl LIX, began on February 9, 2025, at 23:30 hours UTC. Thus, the time interval, T_i, for this event ran from Polymarket's first prediction on January 27 at 18:00 hours UTC to the start of the event on February 9 at 23:30 hours UTC.

We set the time interval to seconds, which left us with 1,143,000 seconds between the first prediction and the start of the event. All time-weighting was done on this scale.

Figure 5 shows the times at which each predictor made its predictions on this time interval.

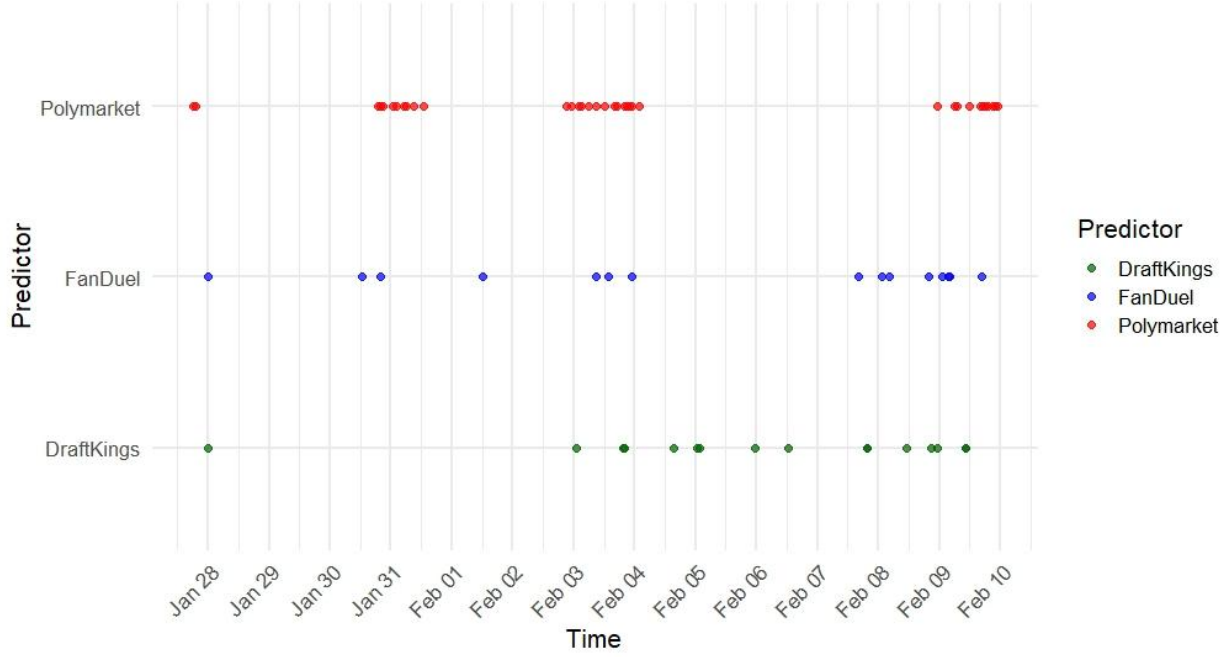

Figure 5. Unique predictions made by the evaluated predictors. Note the larger number of predictions made by Polymarket.

iv. Probability Assignments

The probability assignments for the sportsbooks were based on each predictor's quoted odds for the contest at each time interval.

To compute the implied probabilities (IP) for each game participant, the following formula was used:

$$IP \begin{cases} \frac{100}{ML+100}, \text{if ML} > 0 \\ \frac{ML \times -1}{(ML-100) \times -1}, \text{if ML} < 0 \end{cases} \quad (18)$$

Where ML are the quoted moneyline odds. These implied probabilities (IP) will sum to more than 1, due to the sportsbooks' "overrounding" of the odds to bake in a profit for the house. We remove this overround by applying the following equation to each IP:

$$Overround = IP - 1 \quad (19)$$

We then normalize each ML (ML_a for one team and ML_b for the other team) by removing the overround to arrive at the real assigned probabilities (AP):

$$AP = IP \times \frac{1}{1 + Overround} \quad (20)$$

This ensures that the assigned probabilities (AP) for both teams sum to 1, representing the predictor's predictions for the outcome.

For example, the first prediction made by FanDuel on 28 January 2025 at 23:55 UTC was: Kansas City Chiefs -124 and Philadelphia Eagles 106. Using Equations 12 through 14, we find that the assigned probabilities for this prediction are:

$$IP_{KC} = \frac{-124 \times -1}{(-124 - 100) \times -1} = 0.5536$$

$$IP_{PE} = \frac{100}{100 + 106} = 0.4854$$

$$IP_{KC} + IP_{PE} = 1.0390$$

$$\text{Overround} = 1.0390 - 1 = 0.0390$$

$$AP_{KC} = 0.5536 \times \frac{1}{1 + 0.0390} = 0.5328$$

$$AP_{PE} = 0.4854 \times \frac{1}{1 + 0.0390} = 0.4672$$

Thus, for FanDuel's first assigned probabilities (AP), we find the Kansas City Chiefs (AP_{kc}) with 0.5328 and the Philadelphia Eagles (AP_{PE}) with 0.4672 (and we find that, as expected, 0.5328 + 0.4672 = 1).

For Polymarket, no such normalization was required. The quoted probabilities at each prediction are already normalized and sum to 1 in every case (i.e., Polymarket, as a peer-to-peer network, does not "overround" its odds).

Figure 6 shows the assigned probabilities from each predictor on the Philadelphia Eagles, the winners of Super Bowl LIX, over the evaluated time interval.

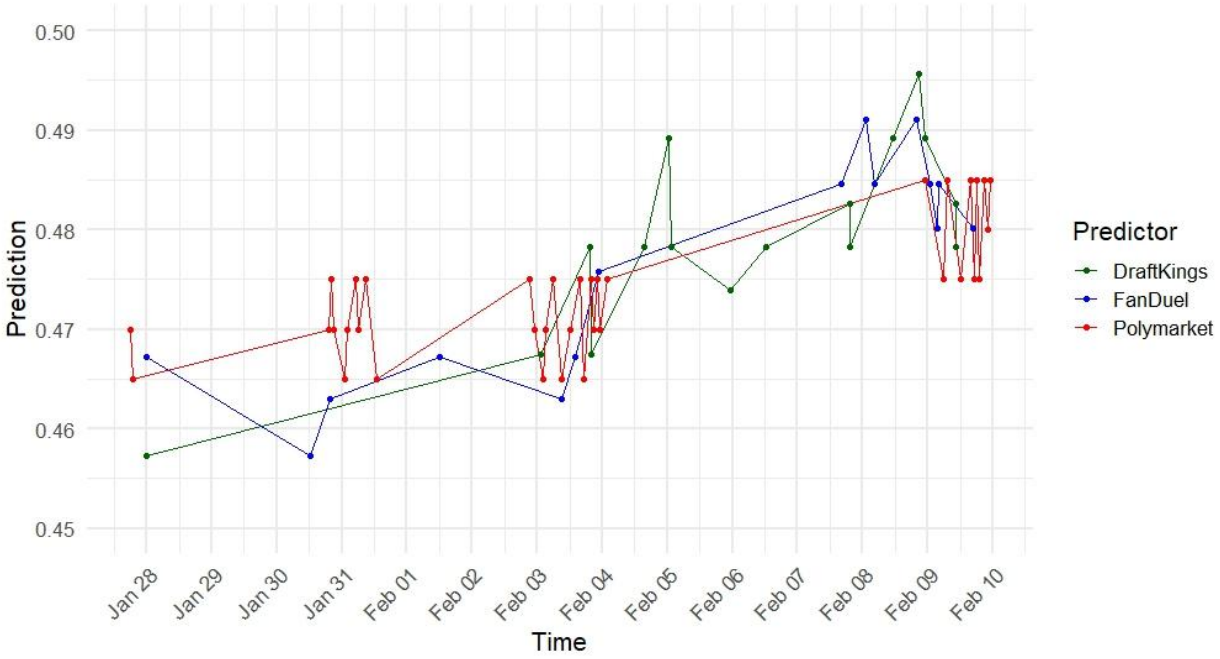

Figure 6. Assigned win probabilities for the Philadelphia Eagles to win by each predictor. Note that all predictors held the Philadelphia Eagles as the "underdog" (i.e., probability of winning <0.5) throughout the entire prediction period.

vi. Preliminary Accuracy Results

Using the methods outlined in this paper and applying them to the predictions made by the various predictors, we find *accuracy ratings* ($\hat{A}$) for each of the evaluated predictors as shown in Table 13.

Table 13. Accuracy Ratings ($\hat{A}$) for the Three Predictors.

Predictor	**Predictions**	**Σc**	**ΣT**	**Â**
DraftKings	16	4.9772	10.4194	0.4777
FanDuel	15	4.8297	10.1889	0.4740
Polymarket	18	12.2595	25.9307	0.4728

With these figures, we can now move to draw inferences about the probability that each predictor's prediction accuracy is better than the discrete uniform probability distribution (DUPD) (i.e., better than "guessing" the outcome of the contest).

VIII. INFERENCE

i. Accuracy Ratings ($\hat{A}$)

As shown in Table 13, DraftKings led the accuracy ratings at 0.4777 while Polymarket trailed the accuracy ratings at 0.4728. This is not a large disparity.

We will use the $\hat{A}$ figures from Table 13 to form prior, likelihood, and posterior beta distributions to gauge the probability that each predictor's accuracy can beat a discrete uniform probability distribution (DUPD).

ii. Prior Distributions

Each predictor is given a unique prior distribution. Each predictor's prior distribution is based on the total time-weighted value (T_i) of all predictions that predictor made. We find the prior distribution for each predictor on Table 14.

Table 14. The Three Predictors' Prior Beta Distributions.

Predictor	ΣT	Prior
DraftKings	10.4194	Beta(5.2097, 5.2097)
FanDuel	10.1889	Beta(5.0944, 5.0944)
Polymarket	25.9307	Beta(12.965, 12.965)

Each predictor's prior beta distribution assumes that half of the value of its predictions (i.e., half of its ΣT) is correct (i.e., in α) and half is incorrect (i.e., in β). For example, FanDuel's ΣT was 10.1889, its prior beta distribution is Beta(5.0944, 5.0944), since 5.0944 is half of its ΣT, and thus, the prior distribution assumes that FanDuel would predict half correctly and half incorrectly.

These beta distributions are equivalent to each predictor's discrete uniform probability distribution (DUPD).

iii. Likelihood Distributions

Each predictor's likelihood distribution is determined by the actual prediction accuracy it exhibited for the evaluated event. The likelihood distribution is simply Σc as s and $\Sigma T - \Sigma c$ as f. The likelihood distributions for each predictor are shown in Table 15.

Table 15. The Three Predictors' Likelihood Distributions.

Predictor	s	f	Likelihood
DraftKings	4.9772	5.4422	Binomial(4.9772, 5.4422)
FanDuel	4.8297	5.3592	Binomial(4.8297, 5.3592)
Polymarket	12.2595	13.6712	Binomial(12.2595, 13.6712)

iv. $\hat{A}$': Posterior Distributions

With each predictor's prior and likelihood beta distributions determined, we move next to determine each predictor's posterior beta distribution, or $\hat{A}'$.

To arrive at each predictor's posterior beta distribution, we simply add the α from the prior and the s from the likelihood distributions and the β from the prior and f from the likelihood

distributions (that is: $\alpha + s$ and $\beta + f$) to form a new posterior α and new posterior β. The posterior beta distributions are shown in Table 16.

Table 16. The Three Predictor's Posterior Beta Distributions (Â´).

Predictor	**Â´**
DraftKings	Beta(10.1869, 10.6518)
FanDuel	Beta(9.9241, 10.4536)
Polymarket	Beta(25.2249, 26.6365)

v. Probability that Â´ > DUPD

With each posterior distribution ($\hat{A}'$) calculated, we now arrive at determining the probability that each predictor was able to beat the discrete uniform probability distribution (DUPD) for the evaluated event.

Since there are two mutually exclusive outcomes for the event—i.e., either a) the Chiefs win and the Eagles lose, or b) the Eagles win and the Chiefs lose—the DUPD assigns a probability of 0.5000 to each of the two outcomes. Thus, we want to know the probability that each predictor was able to beat a probability of 0.5000. In other words, of each predictor's posterior distribution ($\hat{A}'$), we ask $P(\hat{A}') > 0.5000$. The probabilities are shown in Table 17.

Table 17. The Probability that Each Predictor Beats the DUPD.

Predictor	**Probability**
DraftKings	0.5414
FanDuel	0.5477
Polymarket	0.5783

As shown in Table 17, Polymarket leads $P(\hat{A}') > 0.5000$ with a probability of 0.5783 and DraftKings trails with a probability of 0.5414. The probabilities of $P(\hat{A}') > 0.5000$ for each predictor are visualized in Figures 6 and 7.

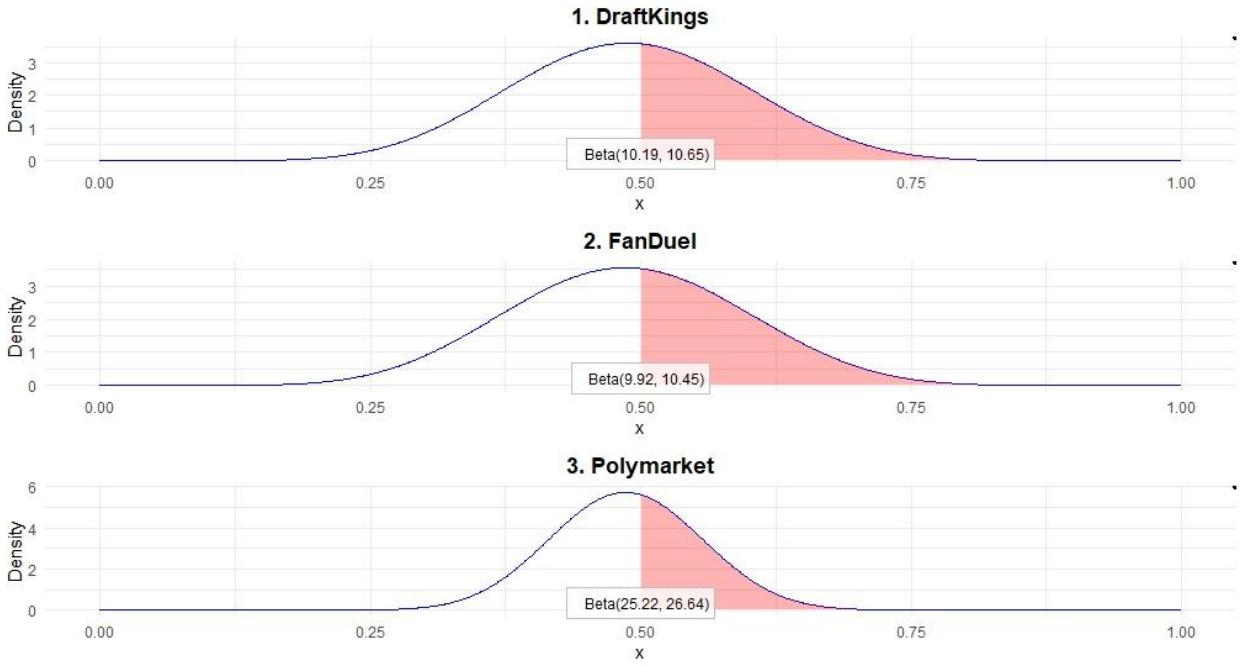

Figure 6. *P(Â΄) > 0.5000* for each predictor. The posterior beta distributions (*Â΄*) are listed alongside each predictor's individual chart. The red shaded region in each chart is the interval on which *P(Â΄) > 0.5000.*

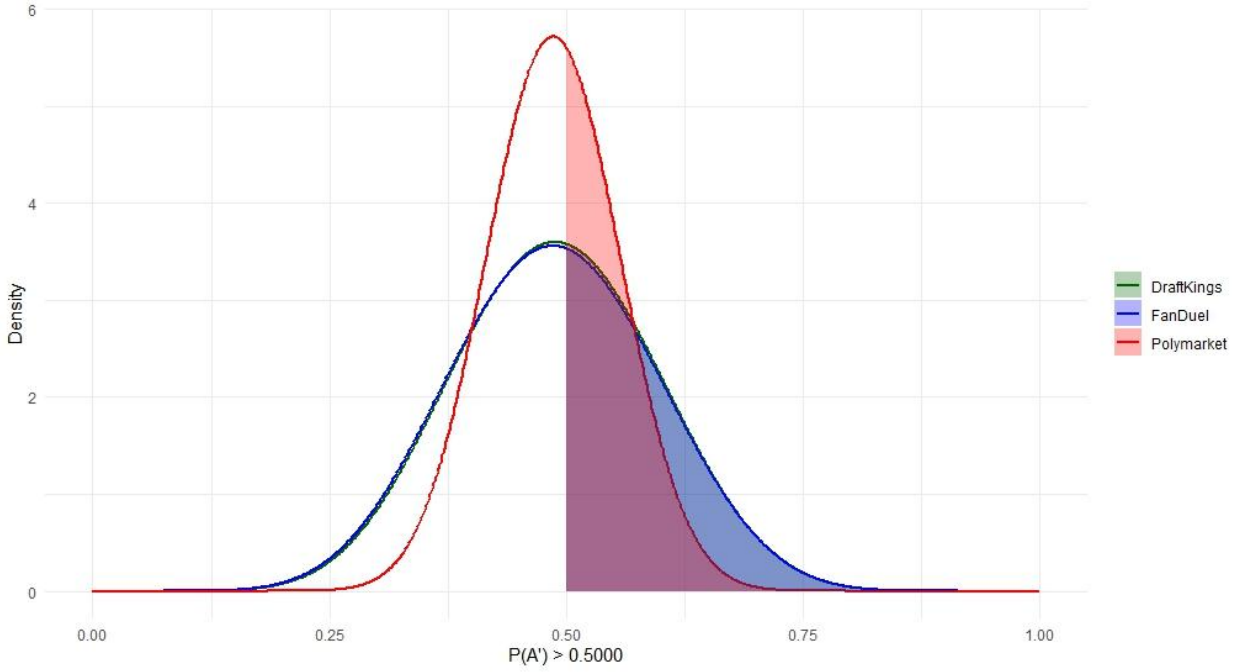

Figure 7. The *P(A') > 0.5000* for each predictor plotted together. Note that DraftKings and FanDuel have nearly identical probabilities, while that of Polymarket is far steeper (and thus, more certain).

The probability figures shown in Table 17 are not surprising. Due to the greater number of predictions made, Polymarket had much tighter posterior beta distribution; thus, the certainty about the skill exhibited by its traders is less doubtful. On the other hand, the fewer predictions made by

the two sportsbooks leaves more doubt about the skill of their traders at predicting the outcome of the event.

Table 18 compares the initial accuracy rating ($\hat{A}$) with the probability that each predictor beat the DUPD considering each predictor's posterior beta distribution:

Table 18. Summary of Predictions and Accuracy for the Predictors of Super Bowl LIX.

Predictor	Predictions	Σc	ΣT	Â	P(Â') > DUPD
DraftKings	16	4.9772	10.4194	0.4777	0.5414
FanDuel	15	4.8297	10.1889	0.4740	0.5477
Polymarket	18	12.2595	25.9307	0.4728	0.5783

Table 18 demonstrates the need for something more than a simple accuracy score ($\hat{A}$). The predictors' probable skill is both a function of how close a set of predictions was to the actual outcome of the event (i.e., at predicting the winner of the Super Bowl) as well as how many predictions were made. Any predictor can be lucky in correctly guessing an outcome: skill is demonstrated by the probability that the predictor exhibited skill over a longer and longer series of correct (or at least, more accurate) predictions.

CONCLUSIONS

This paper demonstrated a system of rating the accuracy of discrete probabilistic predictions, made as either singular predictions or as a series of updated predictions over a time interval. We return to the questions asked in the *Introduction* to summarize.

1. How accurate is a given probabilistic prediction?

To measure the accuracy of a given probabilistic prediction, we can simply evaluate the "credit" earned by the prediction and compare it to the "value" of the event predicted. This gives us a simple measurement of accuracy.

2. How accurate has a given predictor been over time?

To measure the accuracy of a given predictor, we can sum up all the "credits" of that predictor's predictions and divide them by the total "value" of the events *across multiple events.* This means we add up "credit" and event values for *n* events. Ideally, these summations reflect similar sorts of predictions (for example: predictions about the outcome of a dozen different American football games or perhaps the entire record of a given team over a single season). This gives us a measurement of that predictor's overall accuracy.

3. Given what we know, how likely is a given predictor to be right in the future?

To determine how likely a given predictor is to be correct about its predictions in the future, we can form a posterior beta distribution with a DUPD-based prior and the observed likelihood

distribution. This posterior beta distribution can answer our questions about how likely the predictor is to be right in a future prediction.

4. What sort of probability distribution best describes how accurate the predictor is?

A beta distribution, which compares how often a predictor is "right" vs. how often a predictor is "wrong" best describes a predictor's accuracy.

5. What happens if a predictor, or set of competing predictors, makes a series of predictions over time? How do these influence the predictor's overall accuracy?

To measure the accuracy of multiple predictions made by a predictor for a single event, we can time-weight these predictions in such a way that earlier predictions are "worth" more than later ones; this is so because as information becomes available closer to the event's occurrence, we expect making accurate predictions to be easier than when information is less available earlier in the time interval leading up to the event.

To compare multiple predictors' accuracies around an event, we can "normalize" this time-weighing across all predictors so that they are on the same time scale. The earliest prediction made by any predictor serves as the baseline ($Time_0$) for all prediction for that same event.

6. How much should I believe a probabilistic prediction from a given predictor, given what we know about that predictor's accuracy?

Our belief in a predictor's prediction should be based on how well that predictor has demonstrated predictive skill in the past. A predictor should be able to demonstrate skill, judged by the probability that the predator's accuracy is better than the DUPD, or mere "guessing". The higher the predictor's accuracy is above the DUPD, the more credence we should lend to that predictor's predictions.

Final Remarks

This accuracy rating system was applied to answer these questions for three predictors making discrete probabilistic predictions on Super Bowl LIX. This demonstration showed the need for considering not only a time-weighted accuracy metric, what this paper terms $\hat{A}$, but also, in forming a posterior beta distribution that helps answer the question of skill (called $\hat{A}'$). The question of skill is both a function of accurate predictions and numerous predictions: skill cannot be demonstrated from isolated correct guesses, but rather, demonstrated over a longer series of accurate predictions.

This system is thus applicable to gambling, medicine, finance, political analysis, the predictive results of logistic regression techniques or Bayesian networks, and any other application where discrete probabilities are given for future outcomes. Predictors and predictions can thus be evaluated for their accuracy and for the probability that they exhibit real skill in making predictions (that is: $P(\hat{A}') > DUPD$).

APPENDIX

A1. Predictions Made While the Event is Ongoing

It is possible in certain circumstances to continue to make predictions *after* an event has begun. Consider, for example, odds given by sportsbooks for sporting events that are live. In this section, we refer to these probability assignments as "live" for this reason. Predictions, like those we have reviewed previously, made *before* the start of the event are called *pre-event* predictions.

i. Outline of "Live" Prediction Accuracy Measurement

To determine the accuracy of "live" predictions, the assigned probabilities are weighted by a new factor called *L.* This new factor, *L,* is like *T,* the time-weight factor, but *L* is weighted by a different equation:

$$L_i = v - \frac{v \times q_i}{Q} \qquad (A1)$$

Where q_i is the prediction's time interval index for the prediction, and Q is the total number of time intervals for the event from start to finish. The term v is the identical v term from Equation 4 (see section *IV. Rating System: Time-Weighted Method,* subsection *A. Time Intervals*). Using v in Equation A1 allows for the "live" time-weighting to begin at v*'s* value and gradually reduce the time-weight of "live" predictions until they reach a value of 0 at the conclusion of the event. As discussed previously, we place a value of 0.5 for all instances of v in this paper. A discussion on the choice of v being set to 0.5 is given in in a later section of this Appendix.

L_i, q_i, and Q work like T_i, k_i, and K. Let us see an example.

ii. Example of "Live" Prediction Accuracy Measurement

Suppose Predictor A makes a series of six "live" predictions during event E_{10}: P_1 to P_6. E_{10} lasts 92 minutes and we use minutes as the relevant time interval. E_{10} has possible outcomes X_{10} and Y_{10}.

Predictor A makes the following "live" predictions about the outcome of E_{10}.

Table A1. Six "Live" Predictions Made by Predictor A During E_{10}.

			Unweighted		Weighted	
Prediction	**Time Index**	**L_i**	**X_{10}**	**Y_{10}**	**X_{10}**	**Y_{10}**
P_1	12	0.4402	0.5616	0.4384	0.2472	0.1930
P_2	29	0.3478	0.7410	0.2590	0.2577	0.0901
P_3	37	0.3043	0.9439	0.0561	0.2873	0.0171
P_4	41	0.2826	0.6311	0.3689	0.1784	0.1043
P_5	61	0.1739	0.5316	0.4684	0.0925	0.0815
P_6	74	0.1033	0.9631	0.0369	0.0995	0.0038

Suppose that the outcome of E_{10} is X_{10}. To find Predictor A's accuracy rating ($\hat{A}$), we use the following equation:

$$\hat{\mathrm{A}} = \frac{\sum c_i}{\sum L_i} \qquad (A2)$$

Equation A2 is like Equation 5, used for the time-weighted credit and value for *pre-event* predictions, except that Equation 5 and Equation A2 are on different scales.

In our example, to find Predictor A's accuracy for E_{10}, we perform the following calculation:

$$\hat{\mathrm{A}}_{A,E_{10}} = \frac{0.2472 + 0.2577 + 0.2873 + 0.1784 + 0.0925 + 0.0995}{0.4402 + 0.3478 + 0.3043 + 0.2826 + 0.1739 + 0.1033} = 0.7036 \qquad (A3)$$

The "live" prediction accuracy for Predictor A in E_{10} is 0.7036.

This "live" prediction accuracy is subject to being converted to a beta distribution (as per Section *V. Accuracy as a Beta Distribution*) and for probing this beta distribution for the probability that the predictor beat the DUPD (as per Section *VI. Probability of Beating the DUPD).*

iii. Combining "Pre-Event" and "Live" Prediction Accuracy Measurement

What if a predictor makes predictions both *before* **and** *after* an event begins? How do we combine these into a single accuracy measurement?

To compute this, we use the following equation:

$$\hat{\mathrm{A}} = \frac{\sum c_i|T_i + \sum c_i|L_i}{\sum T_i + \sum L_i} \qquad (A4)$$

That is, the sum of credits weighted by both T_i and L_i, divided by the total summed value of T_i and L_i. Equation A4 combines the time-weighted predictions made *before* the start of the event (weighted by T_i) and the time-weighted predictions made *after* the start of the event (weighted by L_i) single a single composite accuracy rating.

Suppose Predictor A makes the following predictions about E_{11}. E_{11} has possible outcomes X_{11} and Y_{11}. The first prediction given by Predictor A, P_1, is given 48 hours before the start of E_{11} and E_{11} lasts 98 minutes. We find the predictions on Table A2.

Table A2. Pre-event and Live Predictions Made for E_{11}.

				Unweighted		Weighted	
Prediction	**Time Index**	**T or L**	**Weight Factor**	**X_{11}**	**Y_{11}**	**X_{11}**	**Y_{11}**
P_1	0	T	1.0000	0.6734	0.3266	0.6734	0.3266
P_2	6,204	T	0.9551	0.6257	0.3743	0.5976	0.3575
P_3	13,566	T	0.9019	0.6133	0.3867	0.5531	0.3488
P_4	21,965	T	0.8411	0.5506	0.4494	0.4631	0.3780
P_5	30,297	T	0.7808	0.6648	0.3352	0.5191	0.2617
P_6	35,717	T	0.7416	0.6424	0.3576	0.4764	0.2652
P_7	14	L	0.4643	0.6091	0.3909	0.2828	0.1815
P_8	34	L	0.4133	0.6386	0.3614	0.2639	0.1494
P_9	46	L	0.3827	0.5835	0.4165	0.2233	0.1594
P_{10}	56	L	0.3571	0.5665	0.4335	0.2023	0.1548

E_{11} begins between P_6 and P_7, as shown in Table A2 divided by the dotted line. The portion of Predictor A's predictions made *before* E_{11} (P_1 through P_6) are on the T_i scale while the portion of Predictor A's predictions made *after* E_{11} began (P_7 through P_{10}) are on the L_i scale. The T_i scale, in this case, runs from 0 to 69,120 (the number of minutes in 48 hours) while the L_i scale runs from 0 to 98 (the full length, in minutes, of the event). Both T_i and L_i have the same time interval (i.e., minutes). Thus, E_{11} begins at the maximum interval for pre-event predictions (69,120) and at the minimum interval for live predictions (0).

Supposing that the outcome of E_{11} is X_{11}, we can now calculate Predictor A's accuracy ($\hat{A}$) for E_{11}:

$$\hat{A}_{A,E_{11}} = \frac{0.6734 + 0.5976 + 0.5531 + 0.4631 + 0.5191 + 0.4764 + 0.2828 + 0.2639 + 0.2233 + 0.2023}{1 + 0.9551 + 0.9019 + 0.8411 + 0.7808 + 0.7416 + 0.4643 + 0.4133 + 0.3827 + 0.3571} = 0.6223 \quad (A5)$$

The total prediction accuracy for Predictor A for E_{11}, considering both "pre-event" and "live" predictions, is 0.6223.

A2. Handling "Ties" in Competitive Events

In some competitive events, like sports, "ties" or "draws" between competitors are possible. In a "tie" or "draw," neither competitor wins nor loses.

To account for the possibility of ties, two methods are suggested:

1. The possibility of a tied outcome is assigned a probability by the predictor.

2. The result of a tie is "pushed."

i. Tied Outcomes Are Assigned a Probability

In competitive events like soccer (i.e., European "football"), ties are very common. In cases like these, the predictor can and should assign a probability to the possible outcome of a tie. Sportsbooks already do this. In a match between Team A and Team B, the predictor assigns a win probability to each Team A and Team B and to the possibility of a draw between the two teams; all three of these probabilities, making up the whole probability space, must sum to 1.

The accuracy rating of such predictions is thus predicated on a DUPD of *Unif{A, B, draw},* and the predictor's $\hat{A}'$ is tested against $P(\hat{A}') > 0.3333$, since the DUPD assumes three equiprobable outcomes.

ii. Results of Ties Are "Pushed"

In competitive events like American football, chess, or political elections, ties are uncommon. The prior case of assigning probabilities to ties for these events, while possible, is impractical. In these kinds of cases, it is suggested that any outcome of a tie is treated as a "push." This means that the "credit" the predictor receives for its $\hat{A}$ is equal to the DUPD.

For example, suppose that in an American football match between Team C and Team D, the predictor assigns a probability of winning of 0.6589 to Team C and 0.3411 to Team D. The match results in a tie, for which no probability of a tie was assigned. The predictor receives, then, "credit" of 0.5000, which is what the DUPD would assume the predictor would receive. The value of this "push" s then weighted by T_i for "pre-event" predictions (or L_i for "live" ones), normally.

A3. The v Variable

In Equations 4 and A1, we assigned a value of 0.5 in all cases to the v term. This term's value can be adjusted to suit the needs of the accuracy measuring model. However, it is unclear whether a different value would offer a better indication of predictive skill. What makes a prediction at the start of an event worth half the value of one made some arbitrary length away from the start of the event? Nothing. What makes it worth 0.22, 0.4999, 0.8888, or some other such value? Nothing. Weighting the value of v at 0.5 places it squarely in the middle, which, as far as this model is concerned, is more than reasonable . Adjusting the v term to a value higher than 0.5 makes the time-weighting system less valuable and makes the value of earlier predictions closer in value to later ones. Adjusting the v term to a value lower than 0.5 makes early predictions more valuable than later ones and makes the slope of the value line between the first prediction and last prediction steeper. Rather than encode a value of 0.5 into these equations, we have inserted the v variable to allow for adjustments to suit models as needed.

A4. Why Use the Discrete Uniform Probability Distribution?

The discrete uniform probability distribution is used throughout this framework to create a baseline assumption for predictive accuracy. The DUPD serves to set our prior distributions (as discussed in Section *V. Accuracy as a Beta Distribution,* subsection *B. Prior Distribution*) as well as probe the posterior distribution for the probability that a predictor's accuracy beats this baseline threshold of accuracy (as discussed in Section *VI. Probability of Beating Discrete Uniform Probability Distribution (DUPD),* subsection *A. Computing P(Â') > DUPD.*

But why should this be?

To illustrate the point, consider the following: if an event has ω possible outcomes, a predictor, call it Predictor A, must assign probabilities summing to 1 to all ω outcomes. In other words, the predictor divides the total probability space of 1 among all ω alternatives in some way. Predictor A *could* do one of three things:

1. Use some predictive model to assign probabilities to each possible outcome.
2. Assign a probability of 1 (absolute certainty of occurring) to <u>one</u> outcome and 0 (absolute certainty of *not* occurring) to all other alternatives.
3. Assign an equal probability to each possible outcome (i.e., $P(x) = \frac{1}{\omega}$).

Let us take a closer look at each of these in turn.

i. Using a Predictive Model

Predictors are assumed to use some predictive model, at the very least, their own beliefs and intuitions, to make predictions. The discussion about which predictive models should be used in which cases is beyond the scope of this paper. Suffice it to say that, absent one of the other two methods of assigning probabilities to alternative outcomes in a given event, the predictor is using *some* form of predictive model (including something like "this happened the last X times, so it will happen in event X+1").

ii. Assign 1's and 0's

A predictor could simply assign a probability of 1 to a single alternative outcome and assign a 0 to all remaining alternatives. This could be a "guess" or else based on the maximum likelihood the predictor forecasts for the event.

For instance, consider Predictor A in event E_{12}. E_{12} has W_{12}, X_{12}, and Y_{12} as possible outcomes. Predictor A could assign 1's and 0's in one of three ways:

Table A3. Three Possible States of Assigning 1's and 0's to Three Possible Outcomes.

		Assigned Probabilities	
State	**W_{12}**	**X_{12}**	**Y_{12}**
1	1	0	0
2	0	1	0
3	0	0	1

There are *n* possible states in which the predictor can assign 1's and 0's, where *n* is the number of possible outcomes.

If Predictor A is just "guessing", that is, arbitrarily assigning a 1 to one possible outcome and a 0 to the others, its long-run accuracy should converge on 0.3333. It may be true that, perhaps, Y_{12} is the most probable outcome, but if Predictor A simply assigns a 1 to each alternative arbitrarily, credit will only be received on the occasions for which Predictor A correctly places this probability of 1 in each event.

Suppose that E_{12} is replicated 100 times precisely. Suppose further that the "real" probabilities for the outcomes of E_{12} are $P(W_{12}) = 0.1547$, $P(X_{12}) = 0.2266$, and $P(Y_{12}) = 0.6187$. In the first repetition, Predictor A assigns probabilities as shown in Table A3 in State 1; in the second repetition, Predictor A assigned probabilities like in State 2, and in the third repetition, Predictor A assigns probabilities like in State 3. Predictor A repeats this pattern (States 1, 2, and then 3) across all 100 repetitions of E_{10}. What will Predictor A's accuracy be across all 100 repetitions?

To compute this, we consult Table A4.

Table A4. Computation of Predictor A's Accuracy Over 100 Repetitions of E_{12}.

Outcome	**Repetitions**	**# Times P=1 Assigned**	**Real Probability**	**Credit**
W_{12}	100	33.3333	0.1547	5.1562
X_{12}	100	33.3333	0.2266	7.5526
Y_{12}	100	33.3333	0.6187	20.6213

The "credit" that Predictor A will receive is the summed total of the number of repetitions times the number of instances in which a probability of 1 is assigned to each possible outcome times the "real" probability of each outcome. The summed credit, as shown in Table A4, is 33.33 (out of 100 repetitions). This leaves an accuracy rating of 0.3333.

The DUPD, or $\frac{1}{\omega}$ is the default "hurdle" to demonstrate that a predictor is doing something other than arbitrarily assigning 1's and 0's to possible outcomes.

iii. Assign Equal Probabilities

Lastly, a predictor could simply use the DUPD to assign probabilities in every case.

Suppose in event E_{13} there are four possible outcomes: W_{13}, X_{13}, Y_{13}, and Z_{13}. Predictor A assigns an equal probability, which is $P(x) = 0.2500$, to all possible outcomes, like shown in Table A5.

Table A5. Equiprobable Assignments for E_{13}.

	Assigned Probabilities			
	W_{13}	**X_{13}**	**Y_{13}**	**Z_{13}**
$State_n$	0.2500	0.2500	0.2500	0.2500

Predictor A does this for *n* states; that is, no matter what, Predictor A will make these equiprobable assignments every time.

Now suppose that E_{13} is repeated 1,000 times. The "real" probabilities in E_{13} are $P(W_{13}) = 0.1780$, $P(X_{13}) = 0.1936$, $P(Y_{13}) = 0.3541$, and $P(Z_{13}) = 0.2743$. For all 1,000 repetitions of E_{13}, Predictor A will assign probabilities like those shown in Table A5. After the 1,000 repetitions, what will Predictor A's accuracy be?

To compute this, we consult Table A6.

Table A6. Computation of Predictor A's Accuracy Over 1,000 Repetitions of E_{13}.

Outcome	**Repetitions**	**# Times P=1 Assigned**	**Real Probability**	**Credit**
W_{13}	1,000	0.2500	0.1780	44.5000
X_{13}	1,000	0.2500	0.1936	48.4000
Y_{13}	1,000	0.2500	0.3541	88.5250
Z_{13}	1,000	0.2500	0.2746	68.6500

The "credit" that Predictor A will receive is the summed total of the number of repetitions times the number of instances in which a probability of 1 is assigned to each possible outcome times the "real" probability of each outcome. The summed credit, as shown in Table A6 is 250 (out of 1,000 repetitions). This leaves an accuracy rating of 0.2500.

Again, the discrete uniform probability distribution (DUPD), or $\frac{1}{\omega}$ is the default "hurdle" to demonstrate that a predictor is doing something other than arbitrarily making equiprobable assignments among alternative outcomes.

iv. Putting It All Together

As we have seen in this section, the DUPD is the baseline hurdle rate. A predictor can easily achieve an accuracy rating equal to the DUPD by making arbitrary probability assignments. To

distinguish real predictive capability from simply "guessing", we maintain the DUPD as the rate of accuracy to beat.

This is seen also in the choice of prior beta distributions as reviewed in Section *V. Accuracy as a Beta Distribution,* subsection *B. Prior Distribution.*

It would be possible to use a weakly informative prior, like Jeffrey's prior, to allow the observed likelihood to better inform the posterior. We do not do this, however, for the reasons outlined in this section: we wish to distinguish between real predictive skill and simply "guessing".

Consider the information in Table A7.

Table A7. Comparison of Weakly Informative and *DUPD*-based Prior.

Scenario	**Prior**	**Likelihood**	**Posterior**	**P(x) > 0.5000**
1	Beta(33, 33)	Binomial(41, 25)	Beta(74, 58)	0.9191
2	Beta(23.5, 42.5)	Binomial (19, 28)	Beta(42.5, 51.5)	0.1754
3	Beta(14.5, 14.5)	Binomial (15, 14)	Beta(29.5, 28.5)	0.5526
4	Beta(0.5, 0.5)	Binomial (41, 25)	Beta(41.5, 25.5)	0.9760
5	Beta(0.5, 0.5)	Binomial (19, 28)	Beta(19.5, 28.5)	0.0944
6	Beta(0.5, 0.5)	Binomial (15, 14)	Beta(15.5, 14.5)	0.5734

In Scenarios 1, 2 and 3, a prior distribution is set based on the DUPD (of 0.5000, meaning there were two possible outcomes for the underlying predictions). In Scenarios 4, 5, and 6, a prior distribution based on Jeffrey's prior, *Beta(0.5, 0.5),* was chosen.

As can be seen from the *P(x) > 0.5000* column in Table A7, the DUPD-based prior used in Scenarios 1, 2, and 3 "drags" *P(x) > 0.5000* toward 0.5000. In other words, using Jeffrey's prior, as in Scenarios 4, 5, and 6, makes the extreme ends of the probability distribution more likely. Scenario 4 has a higher probability than Scenario 1 while Scenario 2 has a higher probability than Scenario 5. This is what we desire because a predictor could simply "guess" the outcomes and be "lucky" (or "unlucky"). We want to be conversative in assuming that any posterior probability is really describing underlying predictive skill and not just "guessing" (which, as we have seen, anyone can do very easily).

Using the DUPD to form a prior distribution makes measured skill tend toward the DUPD itself, but not so much so, as shown in Table A7, to dissuade us from finding real underlying predictive skill. In essence, we put a hurdle in front of believing in a predictor's predictive skill; that hurdle is the DUPD. We want to see a considerable margin above the DUPD to conclude that the predictor has predictive skill (like in Scenario 1 in Tabel A7). A *P(x) > 0.5000* somewhere closer to 0.5000 would not convince us of this (like in Scenario 3).

REFERENCES

[1] The Odds-API (n.d.). *[NFL game odds for Super Bowl LIX, 28 January 2025 to 9 February 2025].* Retrieved on 17 March 2025 from *https://the-odds-api.com/*

[2] Beaver, N.A. (2025). *DraftKings and FanDuel Predictions Made for Super Bowl LIX [.csv file].* NicholasABeaver.com. Published 23 June 2025. https://www.nicholasabeaver.com/portfolio/an-accuracy-rating-system-for-discrete-probability-predictions-using-sportsbook-odds-for-super-bowl-lix/

[3] Polymarket (2025). *Super Bowl LIX Winner.* Polymarket.com. Closed 16 February 2025. Retrieved on 17 March 2025 from https://polymarket.com/event/nfl-kc-phi-2025-02-09