

Polls, Punditry, or Prediction Markets: An assessment of election forecasting

Harry Crane*

November 9, 2018

Abstract

I compare forecasts of the 2018 U.S. midterm elections based on (i) probabilistic predictions posted on the FiveThirtyEight blog and (ii) prediction market prices on PredictIt.com. Based on empirical forecast and price data collected prior to the election, the analysis assesses the calibration and accuracy according to Brier and logarithmic scoring rules. I also analyze the performance of a strategy that invests in PredictIt based on the FiveThirtyEight forecasts.

1 Introduction

Citizens are inundated with analysis, data, and prognostication in the lead-up to an election. Polls measure the sentiment of voters at some place and time prior to the election. Pundits interpret the polls in the broader context of the economy and society. Prediction markets trade on the basis of the information provided by polls, pundits, and other news, thus putting a price on the various outcomes.

The validity of these indicators for projecting the outcome has long been a source of confusion and debate. Polls are riddled with uncertainty due to sampling error and possible bias caused by non-response and possibly dishonest responses. Pundits are compromised, intentionally and unintentionally, by their own biases and preferences. It is less clear how prediction markets, comprised of a large number of individuals making their own subjective assessments based on news and polls, are impacted by these same errors and biases.

Though long a topic of debate, doubts about the reliability of polling and expert opinion have increased in the aftermath of the 2016 election. Overall, the consensus of the polls, pundits, and markets gave Hilary Clinton a clear advantage over Donald Trump. When Trump eventually won, many were quick to criticize this as a failure of experts, analytics, and markets. On the other side, the experts, principally proponents of statistical methods and prediction markets, argued that these criticisms stem from a misperception of what forecasts represent. Though FiveThirtyEight assessed Clinton's chances at a promising 72%, they also recognized a non-negligible 28% chance for Trump. On the day of the election, shares of a Hilary Clinton victory sold for \$0.81 on PredictIt, which translates to a forecasted 81% chance for Clinton but also a 19% chance for Trump.

*Rutgers University, Department of Statistics and Biostatistics

After the election, *SimplyStatistics*, a blog run by academic statisticians, stated plainly that Nate Silver “once again got it right”,¹ in part because the true vote margins in 15 out of 16 states fell within Silver’s credible intervals, which is “exactly what we expect since 15/16 is about 95%”.²

As many others have pointed out, Trump’s election is not inconsistent with either forecast: assuming the FiveThirtyEight and PredictIt forecasts were more-or-less accurate, they rated a Trump win at a live possibility, in the range of 19%-28%. Quoting Erik Brynjolfsson, the *New York Times* wrote on November 10, 2016, “people often do not understand that if the chance that something will happen is 70 percent, that means there is a 30 percent chance it will not occur. The election performance, [Brynjolfsson] said, is ‘not really a shock to data science and statistics. It’s how it works.’ ”

While there is some validity to claims that probabilistic forecasts can’t be definitively validated or falsified by any single outcome, that doesn’t mean that probabilistic forecasts cannot be tested. To the contrary, probabilistic forecasts are meaningful only if they can be tested, against each other and against reality. Below I discuss how to administer these tests, and to demonstrate what they say about the relative insights provided by analytics and prediction markets. My goal is not to determine, once and for all, whether poll aggregators or prediction markets are useless or useful, or in the latter case to argue in favor of which is more useful. A number of confounding factors make such an assessment far beyond the scope of this analysis, as I discuss in Section 4.

The paper is organized as follows. In Section 2, I review a number of statistical measures for evaluating probabilistic forecasts. These measures can be used to assess a range of statistical properties of forecasts, including calibration, accuracy, and profitability when used as the basis of a betting strategy. In Section 3, I analyze the performance of FiveThirtyEight and PredictIt forecasts against these metrics. Importantly, these tests are valid only if identified prior to conducting the test (i.e., prior to election day), for otherwise the analysis can easily be distorted by *post hoc* cherry-picking of favorable or unfavorable metrics, by supporters and critics alike. In Section 4, I discuss key takeaways from this analysis, which hopefully will prove helpful in future election cycles.

Prior to the 2018 U.S. midterm elections, I compiled forecasts for a number of Senate and House of Representatives races from the FiveThirtyEight website. For comparison, I also collected the market prices for the corresponding prediction market on PredictIt.com. I “pre-registered” my analysis on Twitter³ and solicited suggestions for appropriate metrics on which to base the assessment. The results of this analysis are below.

2 How to evaluate forecasts

Just as there are different methods to making forecasts, there are several ways to evaluate their performance. Statistical performance, such as the coverage probabilities of credible intervals, is a way to assess model soundness. For the sake of assessing the reliability of the actual forecasts, however, we are interested in how well the forecasts lineup with what actually happened. For this, statisticians have a number of metrics to assess the

¹<https://simplystatistics.org/2016/11/09/not-all-forecasters-got-it-wrong/>

²There are issues with this analysis. Aside from possibly cherry-picking the 16 states under consideration, there’s the more obvious problem that FiveThirtyEight’s intervals were designed for 80% coverage, not 95% as the *SimplyStatistics* authors assumed. From this, the 15/16 coverage was higher than would be expected of 80% credible intervals.

³Thread available at <https://twitter.com/HarryDCrane/status/1056643258158211072>.

different aspects of a forecast. Here I discuss three such measures: calibration, accuracy, and profit/loss in a betting game. The first two of these, calibration and accuracy, are statistical properties that assess how well the forecasts perform on average. The last of these is less forgiving, in that it allows an adversary to exploit perceived weaknesses in a forecast. I discuss each in turn.

2.1 Calibration (Unbiasedness)

A forecaster is called *calibrated* if, roughly, his 10% forecasts occur 10% of the time, 20% forecasts occur 20% of the time, and so on. Calibration is thus measure of unbiasedness, in the sense that when a calibrated forecaster states a probability of p , the associated outcome tends to occur with frequency p . Note that calibration is not a property of a single forecast, but rather of a collection of forecasts, and therefore of a forecaster.

Formally, calibration is assessed by comparing the empirical frequencies with which forecasted outcomes occurred, denoted as $\hat{F}(p)$ for each forecast $0 \leq p \leq 1$, against the frequencies of an ideal calibrated forecaster, for which $F(p) = p$ for $0 \leq p \leq 1$. To do this, we smooth the forecasts using a kernel function $K(p, p')$, which quantifies how much a forecast of p' should be counted toward the performance of forecasts of p . Given a collection of forecasts $\mathbf{p} = \{p_i\}_{i=1, \dots, N}$, outcomes $\mathbf{y} = \{y_i\}_{i=1, \dots, N}$, and a kernel $K(p, p')$, we define

$$\hat{F}(p) = \frac{1}{N} \sum_{i=1}^N K(p, p_i)(y_i - p_i), \quad 0 \leq p \leq 1. \quad (1)$$

For a simple example,

$$K(p, p') = \begin{cases} 1, & p = p', \\ 0, & \text{otherwise,} \end{cases} \quad (2)$$

offers one such possibility, which is too fine-grained for practical purposes as, for all or almost all values of p , any given forecaster will have forecasted at most 1 outcome to have probability p . To address this, it's natural to smooth out the weights assigned by K to share information among nearby forecasts (e.g., 50% and 51%). The natural next step beyond (2) is to aggregate all forecasts within some region of p by selecting a small value for ϵ (e.g., $\epsilon = 0.02$) and putting

$$K(p, p') = \begin{cases} 1, & |p - p'| < \epsilon, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

Though addressing the issue posed by (2), the kernel given in (3) has issues of its own, especially due to its sensitivity to the sharp cutoff at the arbitrarily chosen ϵ .

We can address this further by choosing a continuous kernel, which for a given p assigns some weight (however small) to all forecasts p' . Ideally, $K(p, p')$ should be decreasing in $|p - p'|$, so that as the forecast p' gets further from p , the contribution of p' to the forecaster's calibration at p becomes negligible. Thus, under such a kernel, both 1% and 49% forecasts affect the calibration at 50%, but the weight assigned to the latter overwhelms the former. The Gaussian kernel is a specific example that we use here, which for $\sigma > 0$ is defined as

$$K(x, x'; \sigma) = \exp\left(-\frac{(x - x')^2}{2\sigma^2}\right), \quad -\infty \leq x, x' \leq \infty \quad (4)$$

2.2 Caveats of calibration

At first glance, calibration seems like a minimal desirable property of probabilistic forecasts. At minimum, it aids the interpretability of a forecaster’s output: when a forecaster puts a 40% probability on something, it can be expected to happen 40% of the time. It’s important, however, to keep in mind that calibration is a property of the forecaster, not the events being forecast. While good forecasts tend to be calibrated, not all calibrated forecasts are good, and as I discuss a bit later not all uncalibrated forecasts are necessarily bad.

In fact, it is easy for a forecaster to be calibrated without doing any actual “forecasting”. For a simple example, imagine forecasting the outcome of a horse race with 10 horses. Without any information about the horses in the race, we know that 1 horse will win and the other 9 will lose. In other words, exactly 10% of the horses in the race will win. Assigning a uniform 1/10 probability to each horse guarantees calibration, but is not a viable strategy for making money at the track. So while these forecasts are calibrated, the forecasting strategy requires no apparent forecasting expertise and provides no insight into the eventual outcome. It’s also a forecasting approach that is known intuitively to the man on the street who instinctively answers “fifty-fifty” to any yes-no question.

Defensive forecasting is a more sophisticated technique that guarantees calibration without the need to do any real “forecasting”.⁴ In defensive forecasting, the forecaster iteratively corrects for past deviations from calibration, based only on past forecasts. In effect, the forecaster who knows how he is being evaluated can overcorrect for past deviations when making future forecasts by thinking of his forecasts as a repeated game against a bettor who makes money whenever the forecaster deviates from calibration. The forecaster’s goal in this game is to prevent the bettor from making money, and he can set his forecasts to ensure that doesn’t happen.

For illustration, let $K(\cdot, \cdot; \sigma)$ be the Gaussian kernel in (4). Given a collection of $N \geq 1$ previous forecasts $\mathbf{p} = \{p_i\}$ and outcomes $\mathbf{y} = \{y_i\}$, define the score at p by

$$S_N(p) = \frac{1}{N} \sum_{i=1}^N K(p, p_i; \sigma)(y_i - p_i), \quad 0 \leq p \leq 1. \quad (5)$$

In the above game-theoretic framework, $S_N(\cdot)$ is interpreted as a betting strategy for a gambler who is attempting to exploit a forecaster’s deviations from calibration. Given a forecaster’s past performances, $S_N(p)$ is the amount of capital that the gambler will bet for (if $S_N(p) > 0$) or against (if $S_N(p) < 0$) an outcome to which the forecaster assigns probability p . Knowing the gambler’s strategy, the forecaster can easily defend against it, and prevent the forecaster from making money, by choosing his next forecast p_{N+1} to be any p^* for which $S_N(p^*) \approx 0$. In doing so, the forecaster avoids losing money (i.e., becoming less calibrated), thus guaranteeing the long-run calibration of his forecasts according to the metric in (5). Figure 1 shows an implementation of this defensive forecasting protocol for samples of various sizes, showing that as the sample size increases, the forecaster becomes better calibrated by using this strategy, even though his forecasts don’t account for any information about the events being forecasted.

In his book, Silver calls calibration “the single most important” test of a forecast, claiming “calibration is difficult to achieve in many fields”. He reiterated this point on a November 2 episode of the FiveThirtyEight politics podcast,⁵ when he identified cali-

⁴V. Vovk, A. Takemure and G. Shafer. Defensive forecasting.

⁵<https://fivethirtyeight.com/features/politics-podcast-how-to-judge-our-forecasts/>

bration as the best metric to judge the FiveThirtyEight forecasts. Defensive forecasting highlights the error in these claims. While uncalibrated forecasts may signal a deficiency, calibrated forecasts do not indicate accuracy or that the forecasts provide any insight into the events under discussion.

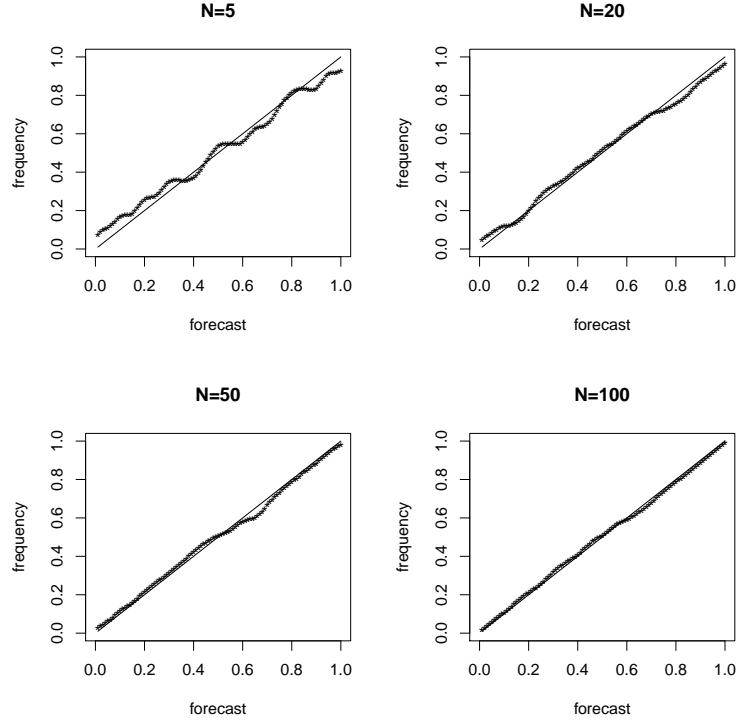


Figure 1: Plot of forecast against frequency for the defensive forecasting strategy with $\sigma^2 = 0.001$ and different values of sample size. The plots show how defensive forecasting progresses toward better calibration as N increases.

2.3 Accuracy

Good forecasts should be accurate. The most accurate forecasts are perfect predictions—they assign probability 1 to events that happen and probability 0 to events that don’t happen. The least accurate forecasts assign probability 1 to events that don’t happen and probability 0 to events that do happen. Most forecasts lie somewhere in between, assigning an intermediate probability $0 < p < 1$ to an event which will eventually be determined.

To formally assess the accuracy of probabilistic forecasts, we define a *scoring rule* $S(P, y)$, which assigns a value to all possible predictive distributions P and outcomes y . Intuitively, $S(P, y)$ can be thought of as the reward (or loss) that a forecaster with predictive distribution P receives (or incurs) when y occurs. A series of forecasts, represented by a set of predictive distributions $\mathbf{P} = \{P_i\}_{i=1, \dots, N}$, and outcomes $\mathbf{y} = \{y_i\}_{i=1, \dots, N}$, is assigned a score

$$S(\mathbf{P}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^N S(P_i, y_i). \quad (6)$$

In this discussion of election forecasts, we specialize to the case of binary forecasts, for which each outcome y_i either occurs ($y_i = 1$) or not ($y_i = 0$). Writing p_i to denote the forecast of the outcome $y_i = 1$, and thus $1 - p_i$ for the forecast of $y_i = 0$, some common scoring rules are

$$\text{Brier : } S_B(\mathbf{p}, \mathbf{y}) = -\frac{1}{N} \sum_i (p_i - y_i)^2$$

$$\text{Logarithmic : } S_{\log}(\mathbf{p}, \mathbf{y}) = \frac{1}{N} \sum_i y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \text{ Divergence from } Q : S_q(\mathbf{p}, \mathbf{y}) = \frac{1}{N} \sum_i$$

where $\mathbf{p} = \{p_i\}_{i=1,\dots,N}$ and $\mathbf{y} = \{y_i\}_{i=1,\dots,N}$. Each of these is an example of a proper scoring rule, meaning that expected accuracy under true outcome distribution \mathbf{q} is maximized at $\mathbf{p} = \mathbf{q}$. A forecaster being evaluated by a proper scoring is thus incentivized to assign forecasts \mathbf{p} that align with his true beliefs about the outcome distribution \mathbf{q} .

2.4 Fundamental Principle of Probability

While calibration and accuracy are accepted statistical measures of forecasts, participants in prediction markets are evaluated directly by their profit and loss. In writing about alternative approaches to resolving the replication crisis,⁶ I defined the *Fundamental Principle of Probability* (FPP):

If you assign a probability to an outcome happening, then you must accept a bet on the other side (that the outcome will not happen) at the correct implied odds.

For example, a forecaster who assigns a probability of 75% that Ted Cruz will win his 2018 Senate race should offer 3-to-1 odds to anyone who wants to bet that Beto O'Rourke will win.

When we extract probabilities from prediction market prices, we are applying the FPP. When someone trades a share of Ted Cruz's successful re-election at 0.75, they are asserting that his probability of re-election is at least 75%. This principle thus provides a straightforward way to compare forecasts by regarding probabilities as prices and computing the profit and loss of each forecast against the prices set by the other. For example, suppose that Forecaster 1 assigns probability p_1 and Forecaster 2 assigns probability p_2 to the same outcome y . By the FPP, Forecaster 1 is offering odds of $p_1/(1 - p_1)$ for a bet on $y = 0$ and $(1 - p_1)/p_1$ for a bet on $y = 1$. Similarly, Forecaster 2 is offering odds of $p_2/(1 - p_2)$ for a bet on $y = 0$ and $(1 - p_2)/p_2$ for a bet on $y = 1$. Assuming that $p_1 > p_2$, Forecaster 1 would accept odds offered by Forecaster 2 on $y = 1$, and Forecaster 2 would accept the odds offered by Forecaster 1 on $y = 0$.

Given this setup, the question turns to how much each forecaster should risk on their respective bets. In the case of just a single outcome, this choice doesn't affect the outcome: as Forecaster 1 and Forecaster 2 are on opposite sides, one will win and the other will lose regardless of outcome. The question is pivotal, however, when we consider multiple forecasts at a time. Suppose, for example, that the same two forecasters are considering two independent events $y(1)$ and $y(2)$, for which they have respective probability $p_1(1)$ and $p_1(2)$ (for Forecaster 1) and $p_2(1)$ and $p_2(2)$ for (Forecaster 2). Assume once again that

⁶<https://www.researchers.one/article/2018-08-16>

$p_1(1) > p_2(1)$ and $p_1(2) > p_2(2)$. Except in cases where the relative odds are equivalent, the individual expected values will differ for the two bets between the two forecasters, in such a way that the maximum expected value will be attained by Forecaster 1 betting all of his capital on one of the outcomes and Forecaster 2 betting all of his capital on the other outcome.

Maximizing expected profit is just one consideration that gamblers take into account. Another is volatility of their wealth, and in particular their risk of being ruined. Individuals who participate in betting markets must allocate their available bankroll among the markets offered on the site. Maximizing expected profit is among their objectives. Avoiding ruin is another. Assuming each forecaster has a total capital of 1 unit, then by their own assessment of the probabilities, their expected profits would be maximized by risking the entire unit on the outcome that offers the largest expected payout. This approach, however, also maximizes the probability of ruin.

An alternative approach that divides the bankroll into equally pieces $1/N$ to bet on each of the N outcomes diversifies the portfolio while reducing the overall expected profit. While the division of capital intuitively risks the chance of ruin, the choice to divide the bankroll evenly into segments of size $1/N$ is arbitrary, and doesn't take into account the relative advantage of each available bet. Consider, for example, an offering of 2 bets, both paying 1-to-1 odds, the first event is guaranteed to happen with probability 100% and the second has a 51% chance. Both are favorable bets, but the first dominates the second. Spreading bets evenly between these two options is a bad strategy for the bettor. It also fails to properly exploit the weaknesses in the offered forecasts.

Compared to the accuracy measures in the previous section, the FPP is a stricter way to evaluate forecasts because of the optionality it affords would-be bettors. A casino offers a large number of bets at pre-determined odds, but the gambler has the choice of which bets to take. A casino that offers 99 perfectly priced bets and 1 bet that gives the player a 5% advantage would perform reasonably well by most standard measures of accuracy, by which the 1 bad offering contributes just 1/100th to the total loss, but would likely lose a lot of money, and may be ruined, in an actual betting scenario because smart gamblers will forgo the 99 accurately priced bets in favor of the one in which they have an advantage.

On the same November 2 podcast cited above, Silver noted the discrepancy between his forecasts and the prediction markets, such as PredictIt. When asked if he was invested in these markets, Silver stated that he can't for "ethical reasons". From the viewpoint of the FPP, however, Silver's lack of exposure to the same downside risks as his readers should his forecasts turn out to be wrong should be considered unethical. As Silver himself describes Chapter 4 of his book when discussing weather forecasting, there are a number of reasons why weather forecasts may be intentionally biased towards forecasting more rain than the model predicts (say, 20% when the model calls for 5%), or otherwise skewing their forecasts in favor of presentation over accuracy. By offering bets in accordance with his forecasts, Silver would put to rest concerns that he engages in the same practices. Below we show how the FiveThirtyEight forecasts compare to the PredictIt forecasts based on a betting portfolio determined by the Kelly criterion.

2.5 Kelly criterion for multiple simultaneous bets

First, observe that payoff odds of $b \geq 0$ on the outcome of an event translates to a forecasted probability $q = 1/(1 + b)$ that the event will occur. Supposing we forecast the

outcome to have probability p , we should not bet if $p < 1/(1+b)$ and we should bet if $p > 1/(1+b)$. The question, in the latter case, is how much? Betting the entire bankroll on a positive expected value outcome maximizes expected profit for that bet but also maximizes the risk of ruin over the long run. To balance these, we look for the optimal fraction f of the bankroll to bet in order to maximize growth of capital while avoiding the risk of ruin.

Let $K_N(f)$ denote our wealth after betting a fraction f of our bankroll on each of N bets. For each bet $i = 1, \dots, N$, let y_i indicate whether the i th outcome resulted in a win ($y_i = 1$) or loss ($y_i = 0$). Total wealth after outcomes $\mathbf{y} = \{y_i\}_{i=1, \dots, N}$ is given by

$$\begin{aligned} K_N(f) &= K_{N-1}(f)(1+bf)^{y_N}(1-f)^{1-y_N} \\ &= K_0 \prod_{i=1}^N (1+bf)^{y_i}(1-f)^{1-y_i} \\ &= K_0(1+bf)^{\sum_{i=1}^N y_i}(1-f)^{N-\sum_{i=1}^N y_i}. \end{aligned}$$

Normalizing so that $K_0 \equiv 1$ and defining $S_N = \sum_{i=1}^N y_i$ for the number of successful (winning) bets and $F_N = N - S_N$ for the number of failed (lost) bets gives

$$K_N(f) = (1+bf)^{S_N}(1-f)^{F_N}.$$

By monotonicity, the maximum of $K_N(f)$ and $\log K_N(f)$ occurs at the same value of f , allowing us to find f by maximizing $\log K_N(f)$. We define the growth rate of f by

$$G_N(f) = \frac{1}{N} \log K_N(f) = \frac{S_N}{N} \log(1+bf) + \frac{F_N}{N} \log(1-f),$$

and we seek to maximize its expectation

$$\mathbb{G}(f) = \mathbb{E}G_N(f) = p \log(1+bf) + (1-p) \log(1-f).$$

Taking the derivative with respect to f and setting equal to 0 gives the Kelly criterion

$$f^* = \frac{bp + p - 1}{b}. \quad (7)$$

These calculations generalize to an offering of $B \geq 1$ simultaneous bets at odds of $b_1, \dots, b_B \geq 0$. In general, let $p(y_1, \dots, y_B)$ be the (joint) probability of outcome (y_1, \dots, y_B) . A betting portfolio $\mathbf{f} = (f_1, \dots, f_B)$ which allocates fraction $f_i \geq 0$ to a bet on $y_i = 1$ at odds b_i , subject to the constraint that the total bets cannot exceed the entire bankroll, i.e., $\sum_i f_i \leq 1$, gives

$$\mathbb{G}(\mathbf{f}) = \sum_{y_1, \dots, y_B \in \{0,1\}} p(y_1, \dots, y_B) \log \left(1 + \sum_{i=1}^B (y_i b_i f_i - (1-y_i) f_i) \right).$$

If these B bets are assumed to be independent, then the joint probabilities satisfy the product rule,

$$p(y_1, \dots, y_B) = \prod_{i=1}^B p_i^{y_i} (1-p_i)^{1-y_i},$$

and maximizing expected growth is obtained by solving the system of B equations given by $\partial \mathbb{G}(\mathbf{f})/\partial f_i \equiv 0$ for $i = 1, \dots, B$. In particular, for $i = 1, \dots, B$,

$$\frac{\partial \mathbb{G}(\mathbf{f})}{\partial f_i} = \sum_{y_1, \dots, y_B \in \{0,1\}} \frac{(y_i b_i f_i - (1 - y_i) f_i) \prod_{i=1}^B p_i^{y_i} (1 - p_i)^{1-y_i}}{1 + \sum_{i=1}^B (y_i b_i f_i - (1 - y_i) f_i)} = 0$$

for each $i = 1, \dots, B$.

At least two practical considerations should be accounted for when applying this betting strategy to election forecasting. First, in the case of election forecasting, the assumption of independence is known to be violated. If a Democratic candidate wins a close race in one state, then the conditional probability of the Democratic candidate winning in other close races increases; and similarly for Republican candidates. Second, and distinct from the issue of dependence, is that the probabilities forecast by both FiveThirtyEight and PredictIt are estimates based on polling data and other information available prior to the election. The forecasted probability may overstate the advantage of betting at the offered odds. In particular, since it can be presumed that the party on the other side is using their own forecasts, bettors might infer that their own forecasts may be erring too far in one direction or another.

For example, suppose odds of b are offered on an outcome which is forecast to have probability p . For $q = 1/(1 + b)$, suppose $p > q$ has a spread of $p - q > 0$. Because both p and q are based on incomplete information, these diverging estimates may suggest that p overestimates the probability, q underestimates the probability, or both. However, the advantage of the player is that the odds b based on the forecast q are fixed, while the bettor can adjust his bet sizing to account for possible overestimation. As long as q is an underestimate, it is advantageous to bet, the only question is what fraction to risk.

In this case, a margin of safety $0 < m < 1$ adjusts the estimate p to $p^* = p - m(p - q) = (1 - m)p + mq$, the linear interpolation between p and q . For this adjusted forecast, the growth rate becomes

$$\mathbb{G}(f) = ((1 - m)p + mq) \log(1 + bf) + ((1 - m)(1 - p) + m(1 - q)) \log(1 - f).$$

Taking the derivative and setting to 0 gives

$$\frac{\partial \mathbb{G}(f)}{\partial f} = \frac{p^* b}{1 + bf} - \frac{1 - p^*}{1 - f} = 0,$$

so that

$$f^* = \frac{p^*(b + 1) - 1}{b},$$

as usual. Noting that $p^* = (1 - m)p + mq$ and $b = (1 - q)/q$, the above calculation corresponds to a $(1 - m)$ -fractional Kelly strategy,

$$f^* = (1 - m)f,$$

where $f = (p(b + 1) - 1)/b$ is the usual Kelly criterion for a probability p outcome offered at b odds. Because fractional Kelly merely scales the computed fractions by a pre-determined constant, applying the fractional Kelly correction doesn't affect our comparison of the FiveThirtyEight and PredictIt forecasts.

3 Analysis

I compiled the FiveThirtyEight forecasts and market prices on PredictIt for all House, Senate, and Gubernatorial races for which there was an open market on PredictIt from August 8, 2018 until November 6, 2018.

Below, I report the relative performance of 538 and PredictIt based on the above metrics and the availability of results as of November 8, 2018. The analysis will be updated as the results become final.

3.1 Senate

Table 1 shows the forecasts and results for the Senate. A calibration plot is shown in Figure 2.

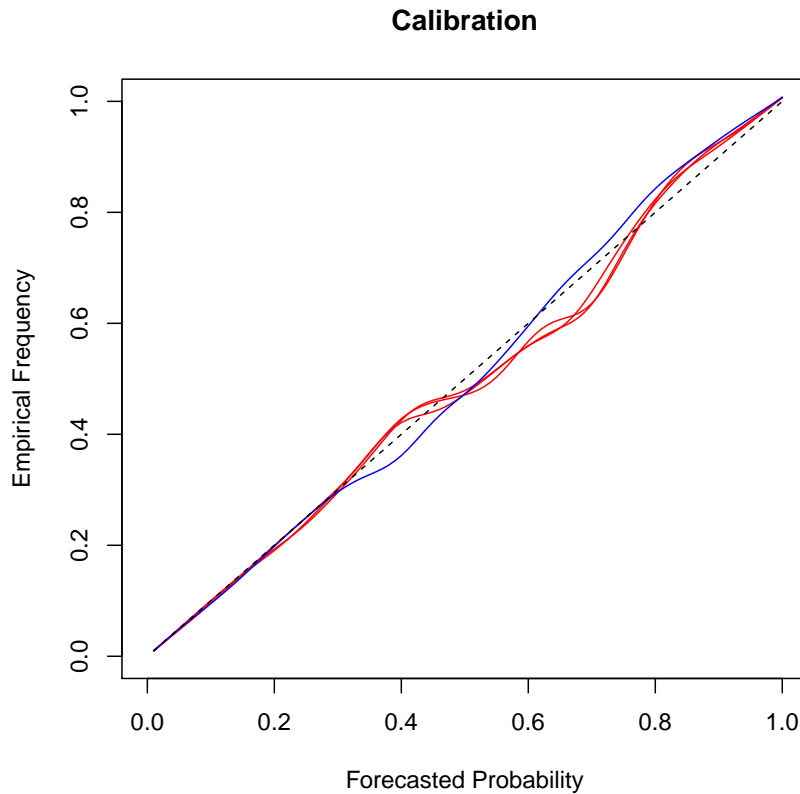


Figure 2: Plot of forecast theoretical versus empirical frequencies. The green line represents the forecasts of PredictIt and the three overlapping red lines correspond to FiveThirtyEight forecasts

The performance of these based on Brier, Log, and Divergence scores are shown in Table 5. As this shows, PredictIt was more accurate than all three FiveThirtyEight models according to both metrics.

We next pit the forecasts against each other according to a betting game. As noted above, there is difficulty to account for dependence in calculating the Kelly criterion. As an initial test, we weight bets according to the standard Kelly criterion for a single bet, as

Outcoome	538 (Classic)	538 (Lite)	538 (Deluxe)	PredictIt	Outcome
Cruz (TX)	0.788	0.814	0.810	0.790	YES
AZ Senate (Repub)	0.414	0.399	0.394	0.566	YES
TN Senate (Repub)	0.839	0.834	0.816	0.863	YES
Tester (MT)	0.871	0.842	0.841	0.670	YES
McCaskill (MO)	0.601	0.564	0.553	0.390	NO
Manchin (WV)	0.866	0.853	0.865	0.770	YES
Donnelly (IN)	0.689	0.686	0.644	0.510	NO
Heitkamp (ND)	0.247	0.095	0.222	0.130	NO
Stabenow (MI)	0.978	0.983	0.990	0.830	YES
Nelson (FL)	0.695	0.707	0.681	0.570	NO
Menendez (NJ)	0.907	0.904	0.920	0.790	YES
Heller (NV)	0.514	0.524	0.466	0.400	NO
Feinstein (CA)	0.963	0.913	0.964	0.930	YES
Warren (MA)	0.998	0.996	0.998	0.960	YES
Brown (OH)	0.960	0.972	0.978	0.880	YES
Baldwin (WI)	0.975	0.978	0.985	0.910	YES
Sanders (VT)	0.999	0.999	0.999	0.980	YES
Smith (MN)	0.892	0.864	0.916	0.850	YES
Heinrich (NM)	0.990	0.989	0.993	0.930	YES
Casey (PA)	0.969	0.974	0.984	0.930	YES
King (ME)	0.988	0.980	0.993	0.960	YES

Table 1: FiveThirtyEight and PredictIt forecasts of Senate as of November 2.

	Brier	Log	Q -Divergence
538 (Lite)	-0.099	-0.294	-0.020
538 (Classic)	-0.101	-0.298	-0.022
538 (Deluxe)	-0.093	-0.277	-0.017
PredictIt	-0.070	-0.259	0.006

Table 2: Brier, Log, and Q -Divergence scores for Senate races.

in (7). Table ?? shows the profit/loss of FiveThirtyEight vs. PredictIt for different values of a threshold T , which filters out any races for which both PredictIt and FiveThirtyEight forecasts are larger than T or less than $1 - T$. This threshold is chosen to account for known market inefficiencies such as longshot bias and transaction fees. Indeed, our analysis finds that PredictIt outperformed FiveThirtyEight in a betting match on Senate races, but that this difference is understated because of market inefficiencies for extreme probabilities.

3.2 House of Representatives

Table 6 shows the forecasts and results for the House of Representatives.

T	FiveThirtyEight	PredictIt
1	-1.8%	+32.3%
0.95	-3.7%	+37.2%
0.90	-3.7%	+37.2%
0.85	-14.8%	+57.5%

Table 3: Return on investment (ROI) for Senate races as a function of threshold T .

T	FiveThirtyEight	PredictIt
1	+8.0%	-32.3%
0.95	+12.1%	-26.8%
0.90	+15.2%	-22.6%
0.85	+17.5%	-20.0%

Table 4: Return on investment (ROI) for House races as a function of threshold T .

4 Discussion

The above analysis gives a preliminary look at the relative performance of FiveThirtyEight and PredictIt for forecasting the 2018 midterm congressional elections. This analysis is based on the projected outcomes as of November 8, 2018 and will be updated with further details, including gubernatorial and historical forecasting information, as the results become finalized.

All of the accuracy metrics used here agree with the general conclusion that PredictIt outperformed FiveThirtyEight in the Senate races, while FiveThirtyEight outperformed PredictIt in the House. What conclusions ought to be drawn from this is less clear. I make just a few preliminary observations here.

1. The day after the election, Nate Silver referred to the above analysis as “very cherry-picked” and “stupid”.⁷ To the contrary, the methodology carried out above was posted only prior to the election results for the purpose of avoiding charges of cherry-picking. For his own part, Nate Silver has been justifiably charged with cherry-picking. In the linked Twitter threads in the footnotes, Silver suggests that the House races were more important to look at than the Senate, and that the analysis involving betting is inaccurate because he wouldn’t have made all of those bets. Prior to the election, Silver identified calibration as the metric against which his forecasts should be evaluated, but afterwards evaluated on predictive accuracy, boasting of a 90+% figure.

A primary objective of the analysis presented here is to highlight the value and importance in pre-registering the metrics by which forecasts are to be assessed. Without a well-established metric ahead of time, suspicion of cherry-picking is inevitable.

2. In addition to the well-established accuracy measures for forecasts using proper scoring rules, I have proposed here a metric in terms of a betting portfolio where

⁷See Twitter correspondence <https://twitter.com/NateSilver538/status/1060180845678522369> and <https://twitter.com/NateSilver538/status/1060196243786350593>.

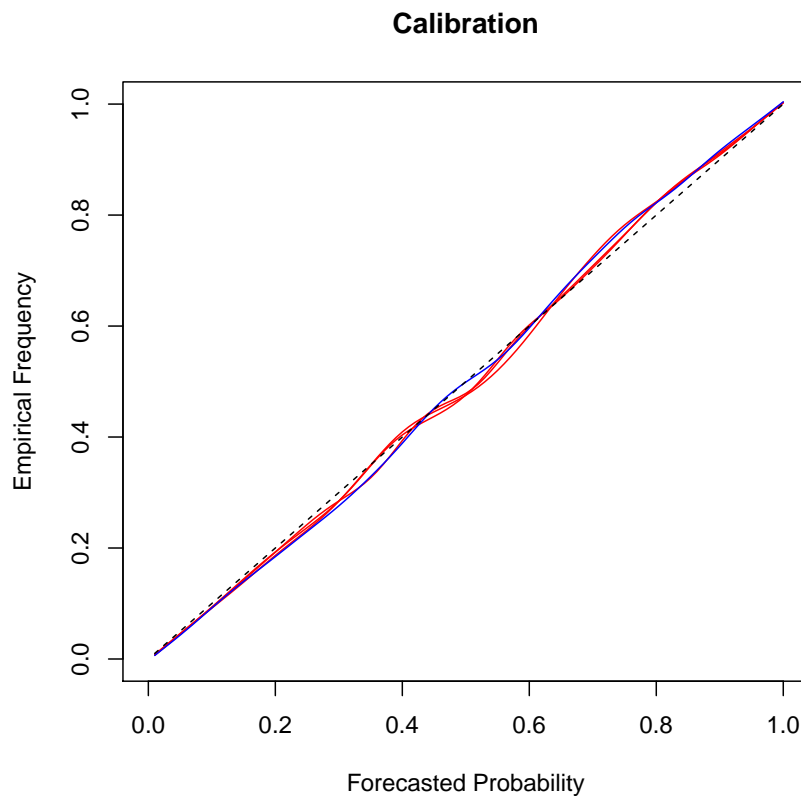


Figure 3: Plot of forecast theoretical versus empirical frequencies for the House of Representatives. The green line represents the forecasts of PredictIt and the three overlapping red lines correspond to FiveThirtyEight forecasts.

the two forecasters compete against each other in a betting market in which the odds are set by their respective forecasts. This measure of accuracy offers a concrete interpretation, especially in light of the real-world mechanism by which prediction market prices are set. In this way, the FiveThirtyEight forecasts could test the prediction market by implementing a betting strategy along the lines suggested. If FiveThirtyEight's forecasts are superior to the prediction market, then they should be able to turn a profit in these markets.

3. The analysis shows that FiveThirtyEight would have turned a small overall profit over PredictIt in a hypothetical game without transaction fees. If transaction fees were added, however, the advantage diminishes. There are known inefficiencies in prediction markets due to long-shot biases and transaction fees. As a result, the price of an unresolved market can never go above 99% or below 1%, and more often these extreme probability events stay in the range 95-97% due to long-shot bias, transaction costs, and opportunity cost of investing in a low-payoff event. The prediction markets performance improves when these extreme probability events are removed, suggesting that the markets are relatively more reliable on events with greater uncertainty.
4. Silver has objected to the comparison of prediction markets and forecasts, call-

	Brier	Log	Q -Divergence
538 (Lite)	-0.10.1	-0.328	-0.006
538 (Classic)	-0.096	-0.305	-0.004
538 (Deluxe)	-0.087	-0.282	0.001
PredictIt	-0.094	-0.319	-0.006

Table 5: Brier, Log, and Q -Divergence scores for House races.

ing it “dumb”, at least in part because participants in the prediction markets use FiveThirtyEight as part of their decision making. Though undoubtedly true that some participants in the market use FiveThirtyEight, it is not immediately clear that those who use FiveThirtyEight are the veritable “smart money” that makes the prediction market prices reliable. In fact, as the output in Tables 1 and 1 show, many markets show more than 15-20% discrepancy between PredictIt and FiveThirtyEight, while others show virtual agreement between the two. If FiveThirtyEight were as influential in the prediction markets as suggested, then (i) there wouldn’t be such a large discrepancy in price and (ii) whatever discrepancy in price existed should be more consistent between markets. Finally, since FiveThirtyEight primarily aggregates polling data, it isn’t necessarily clear that bettors who use FiveThirtyEight are gleaning any more than they would from the raw polling data. Evidently, the precise forecasts at FiveThirtyEight do not have a systematic effect on the prediction market prices. And given that markets would be affected from polls in any case, it could be that FiveThirtyEight provides a convenient service in poll aggregation but adds no value beyond that.

Outcome	538 (Classic)	538 (Lite)	538 (Deluxe)	PredictIt	Outcome
Devin Nunes (CA-22)	0.955	0.901	0.975	0.890	YES
Barbara Comstock (VA-10)	0.114	0.142	0.120	0.150	NO
Chris Collins (NY-27)	0.756	0.743	0.795	0.750	YES
Steve Knight (CA-25)	0.370	0.690	0.403	0.500	YES
David Brat (VA-7)	0.572	0.583	0.541	0.630	NO
Nancy Pelosi (CA-12)	0.999	0.999	0.999	0.980	YES
Rodney Davis (IL-13)	0.725	0.783	0.775	0.750	YES
David Valadao (CA-21)	0.798	0.743	0.840	0.890	YES
Pete Sessions (TX-32)	0.729	0.554	0.655	0.550	NO
Doug Lamborn (CO-5)	0.980	0.989	0.988	0.970	YES
Rod Blum (IA-1)	0.048	0.113	0.062	0.190	NO
Will Hurd (TX-23)	0.781	0.961	0.792	0.830	TBA
Jason Lewis (MN-2)	0.148	0.203	0.136	0.190	NO
Steve Chabot (OH-1)	0.834	0.852	0.836	0.870	YES
Mimi Walters (CA-45)	0.382	0.481	0.364	0.450	TBA
Brian Fitzpatrick (PA-1)	0.581	0.483	0.547	0.480	YES
John Culberson (TX-7)	0.490	0.598	0.512	0.560	NO
Peter Roskam (IL-6)	0.516	0.456	0.383	0.310	NO
Jeff Denham (CA-10)	0.233	0.395	0.301	0.430	TBA
Don Bacon (NE-2)	0.590	0.819	0.681	0.830	YES
Duncan Hunter (CA-50)	0.782	0.742	0.817	0.750	YES
Carlos Curbelo (FL-26)	0.457	0.520	0.508	0.470	NO
Tom MacArthur (NJ-3)	0.455	0.546	0.430	0.470	TBA
MI-11 (Rep)	0.224	0.252	0.188	0.279	NO
NY-1 (D)	0.068	0.066	0.041	0.101	NO
NJ-7 (D)	0.773	0.751	0.703	0.693	YES
IL-12 (D)	0.279	0.169	0.242	0.248	NO
PA-06 (D)	0.989	0.949	0.988	0.897	YES
CT-05 (D)	0.972	0.902	0.985	0.950	YES
VA-05 (D)	0.282	0.506	0.339	0.337	NO
NY-09 (D)	0.999	0.999	0.999	0.960	YES
GA-07 (D)	0.134	0.061	0.133	0.233	TBA
SC-04 (D)	0.001	0.020	0.001	0.050	NO
MD-06 (D)	0.993	0.979	0.996	0.947	YES
NY-14 (D)	0.999	0.999	0.999	0.968	YES
FL-17 (D)	0.002	0.035	0.001	0.059	NO
NM-2 (D)	0.413	0.420	0.456	0.438	YES
TX-6 (D)	0.056	0.141	0.020	0.061	NO
NJ-02 (D)	0.979	0.967	0.984	0.961	YES
NV-03 (D)	0.736	0.595	0.778	0.644	YES
ME-02 (D)	0.629	0.551	0.594	0.515	TBA
NC-13 (D)	0.321	0.270	0.351	0.305	NO
AZ-2 (D)	0.962	0.858	0.957	0.902	YES
MN-1 (D)	0.539	0.470	0.535	0.509	NO
NY-22 (D)	0.555	0.481	0.576	0.471	YES
NY-19 (D)	0.587	0.561	0.603	0.571	YES
NH-1 (D)	0.838	0.774	0.889	0.890	YES
VA-6 (D)	0.005	0.046	0.003	0.062	NO
FL-15 (D)	0.424	0.440	0.366	0.396	NO
IA-3 (D)	0.688	0.591	0.641	0.636	YES
MS-03 (D)	0.003	0.042	0.003	0.058	NO
PA-09 (D)	0.002	0.035	0.001	0.052	NO
KS-2 (D)	0.622	0.510	0.594	0.481	NO
MN-07 (D)	0.928	0.714	0.958	0.788	YES
OH-12 (D)	0.341	0.309	0.348	0.394	NO
NV-04 (D)	0.827	0.655	0.853	0.784	YES
NC-09 (D)	0.456	0.470	0.535	0.525	TBA
MI-08 (D)	0.492	0.411	0.540	0.500	YES
CA-48 (D)	0.596	0.443	0.606	0.410	TBA
TX-29 (D)	0.999	0.998	0.999	0.970	YES
KS-03 (D)	0.843	0.888	0.858	0.762	YES
KY-06 (D)	0.506	0.583	0.516	0.475	NO
WV-03 (D)	0.072	0.277	0.096	0.225	NO
FL-27 (D)	0.841	0.848	0.844	0.667	YES
PA-17 (D)	0.954	0.891	0.960	0.880	YES

Table 6: FiveThirtyEight and PredictIt forecasts of House of Representatives as of November 2