# COS 597E ASSIGNMENT 1: FORECASTING ELECTIONS

**Shreyas Gandlur, Jordan Klein, Dora Zhao**

**ABSTRACT**

In this project, we compare the accuracy of three types of election forecasters in the 2020 Presidential Election at the state level—quantitative modelers, experts, and prediction markets—on Election Day and in the months leading up it. We also analyze how the accuracy of these predictions changed over time. Using the framework described by Tetlock [9], we calculate a probability score for each prediction (uniquely identified by a forecaster-date pair) and the associated calibration and discrimination indices. We then compare these scores between and within each group of forecasters over time.

## 1 Introduction

Professional forecasters invest considerable resources into trying to predict the outcome of US Presidential Elections, often making predictions months, or even years, in advance. There are two types of mainstream forecasters: experts and quantitative modelers. Experts use their domain knowledge to classify states into discrete categories along two dimensions: whether they expect a Republican or Democrat to win, and their degree of certainty in this prediction, described qualitatively using labels such as Solid/Safe, Likely, Lean, Tilt, and Toss-up [2][3][7]. Quantitative modelers, on the other hand, build statistical models that incorporate some domain-specific knowledge on elections, but primarily rely on polling data. These models output a continuous value representing the probability that a political party wins a given state [4][5][8]. In addition to professional forecasters, the general public gets involved in making election predictions through gambling on prediction markets. In these markets, the price of a prediction corresponds to the market's estimate of the probability of that outcome occurring [6]. Using predictions for the state-level outcomes of the 2020 Presidential Election from these different types of forecasters, we attempt to answer the following two questions:

1. Which type of forecaster makes the most accurate predictions: experts, quantitative modelers, or prediction markets?
2. How do the accuracy of these forecasts change over time?

In our preregistration, we hypothesized that:

1. Quantitative modelers would make the most accurate predictions on average. We suspected this would be due to better calibration rather than higher discrimination, of their predictions compared to other forecasters.
2. We expected election predictions to become more accurate closer to the election. We suspected this would be attributable to predictions becoming better calibrated, rather than more discriminating, over time.

Our subsequent analysis of these predictions suggests the following:

1. Quantitative modelers make the most accurate predictions on average and are the best calibrated, while prediction markets are the most discriminating on Election Day. Experts perform poorly on both indices and make the least accurate predictions. Although quantitative modelers are more calibrated but less discriminating than prediction markets on Election Day itself, our findings suggest that prediction markets became more discriminating than quantitative modelers only in the final week of the election cycle. This suggests that market participants may be making more aggressive predictions in the final days before Election Day.
2. The election predictions of quantitative modelers and prediction markets did improve over time; however, expert predictions became *worse* over time. The improvements we see in quantitative modelers and prediction

markets appear to be attributable to small in magnitude but steady improvements in both calibration and discrimination over time. Experts' predictions became more discriminating over time; however, they became *less* well calibrated as the election approached.

# 2 Methods

## 2.1 Data Collection

We collect data from three different types of forecasters—quantitative modelers, experts, and prediction markets—on their predicted state-level outcomes for the 2020 Presidential Election. For quantitative modelers, we include predictions made by FiveThirtyEight [8], The Economist [5], and JHK Forecasts [4]. They give probabilistic predictions for each state. The Economist and JHK give daily predictions starting from March 1, 2020, and FiveThirtyEight gives daily predictions starting from June 1, 2020. We chose The Economist and FiveThirtyEight based on their popularity and prevalence in the media. We also include JHK Forecasts, which, while not as well-known as the other two sources, performed well in forecasting the 2018 Senate midterm races and has been featured in news sources including Newsweek and CNN. Moreover, all three forecasters' data is publicly available and their models are either open-source (The Economist) or described in detail on their websites (FiveThirtyEight, JHK Forecasts).

For the expert forecasters, we look at three outlets— Cook Political Report [2], Sabato's Crystal Ball [7], and Inside Elections [3]. Similarly to quantitative modelers, we selected these sources based on their popularity and accessibility of their data. The expert forecasts give a qualitative rating for each state, starting from as early as March 2019. To make these predictions comparable to the quantitative modelers and prediction markets, we converted the qualitative ratings into probabilistic predictions based on how quantitative modelers have assigned these qualitative labels to their numerical predictions (See Table 1).

Finally, we use data from PredictIt to represent prediction markets [6]. While other prediction markets exist, PredictIt, has the highest trade volume by far. PredictIt provides price data for a Joe Biden on Donald Trump victory in each state, corresponding to the probabilities the market assigns to each of these outcomes. PredictIt provides pricing data for the past 90 days publicly, which we use for our analysis. We note that trading volume is low until roughly September 2020, which suggests that prices prior to that may not represent a market consensus. All of the expert forecasters, quantitative modelers, and the prediction market we selected are used to generate the 270toWin's consensus electoral map, a widely-cited election forecast aggregator [1].

## 2.2 Variables

### 2.2.1 Independent Variables

For each forecaster/date pair, we have the variables:

- `dem_chance`: Quantitative modellers' and experts' predicted subjective probability of a Joe Biden victory in a given state/district (0-1). See Table 1 for methodology for converting experts' qualitative ratings to subjective probabilities.
- `dem_price`: Betting price for a Joe Biden victory in a given state/district (0-1).

From the actual election results, we have the variable:

- `outcome`: 0 = a Donald Trump victory in a given state/district, 1 = a Joe Biden victory in a given state/district.

### 2.2.2 Dependent Variables

For each forecaster/date pair, using the `dem_chance` (quantitative modelers and experts) or `dem_price` (betting markets) variables and the outcome variable, we calculate the following [9]:

- **PS (Probability Score)**: mean squared difference between `dem_chance` (or `dem_price` for PredictIt) and outcome. It measures the average deviation between subjective probability forecasters assign to events and whether (1) or not (0) they happen. A perfect probability score is 0.

| Rating (Qualitative Rating) | dem_chance (Corresponding Probability) |
|---|---|
| Solid / Safe Republican | 0.05 |
| Likely Republican | 0.25 |
| Lean Republican | 0.40 |
| Tilt (if included) Republican | 0.45 |
| Toss-up | 0.50 |
| Tilt (if included) Democrat | 0.55 |
| Lean Democrat | 0.60 |
| Likely Democrat | 0.75 |
| Solid / Safe Democrat | 0.95 |

Table 1: The corresponding probabilities with the expert forecasts for each state/district. Except for Tilt, these probabilities are given by Silver et al [8]. Tilt R/D is used by Inside Elections, and we chose to convert it to midway between Lean R/D and Toss-Up.

- **VI (Variability Index)**: the variance of the outcome variable. This measures the variability, predictability or unpredictability of an environment. The easiest to predict environments have a base rate of 0 or 1, while the most difficult to predict environments have a base rate of 0.5.
- **CI (Calibration Index)**: weighted mean of the squared differences between the proportion of predictions correct in each probability category and the probability value of that category. This is a measure of caution, how close forecasters assign probabilities to their base rates, or the correspondence between the subjective probability a forecaster assigns events and their objective frequency. Under perfect calibration, no events assigned a subjective probability of 0 occur, 10% of events assigned a subjective probability of 0.1 occur, etc.
- **DI (Discrimination Index)**: weighted mean of the squared differences the proportion of predictions correct in each probability category and the probability of the outcome occurring. This is a measurement of the ability to sort predictions into probability categories such that the proportions of correct answers across categories are maximally different from each other, or to do better than "predict the base rate strategy". Under perfect discrimination, forecasters assign a subjective probability of 0 to events that occur and 1 to those that do not.

### 2.3 Analysis

To test our hypotheses, we compare the probability predictions made by each of our forecasters with the actual outcome for each state. In particular, we are interested in seeing how the forecasters performed in comparison to each other and how they performed over time (e.g. several months before the election vs. the day of the election).

As detailed in Section 2.2.2, we calculate the calibration index, variability index, and discrimination index, as well as a probability score, for each forecaster-date pair. To study differences in accuracy within groups of forecasters, we compare the probability scores and calibration/discrimination indices for the final predictions made by each forecaster on Election Day, and to study differences between groups of forecasters, we calculate the mean of these performance metrics for each group. To study how predictive performance changes over time, we plot these performance metrics for each individual forecaster and the mean performance metrics for each group of forecasters between June 1, 2020 and November 3, 2020.

## 3 Results and Discussion

As we originally hypothesized, we find that quantitative modelers perform the best out of the three forecaster types. As seen in Figure 1, quantitative modelers have a mean probability score of $0.0358$, which is smaller than the mean probability scores of both prediction markets ($0.0381$) and experts ($0.0616$). Of the three quantitative modelers, JHK performs slightly better than FiveThirtyEight and the Economist (See Appendix Figure 1).

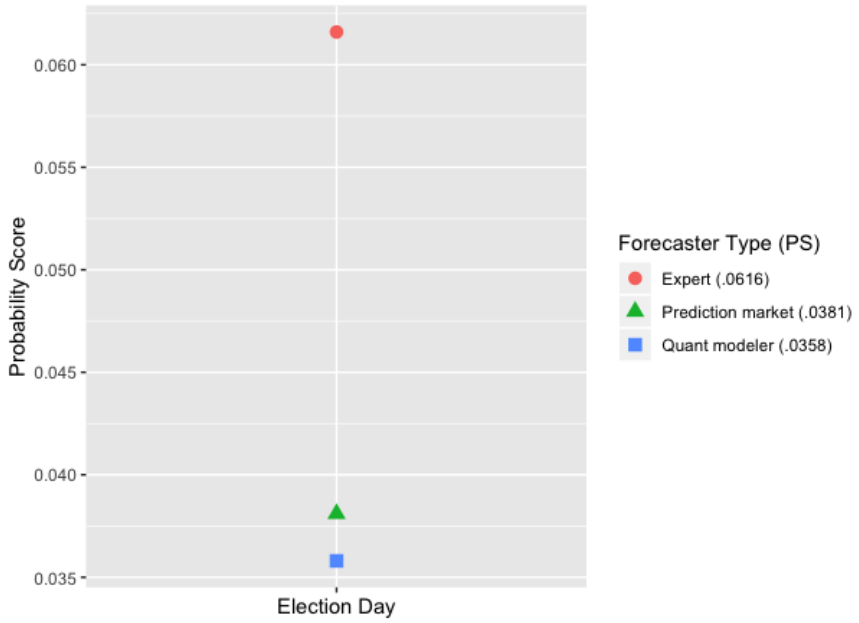


Figure 1: Mean probability scores of experts (Cook Political Report, Inside Elections, Sabato's Crystal Ball), prediction markets (PredictIt), and quantitative modelers (FiveThirtyEight, The Economist, JHK) final predictions on Election Day (November 3). Lower probability scores indicate superior performance.

Furthermore, when comparing the calibration and discrimination indices amongst the three forecaster types, we find that quantitative modelers are the best calibrated and the prediction markets are the best at discrimination, whereas experts are poor at both calibration and discrimination (See Figure 2 and Appendix Figure 2) .

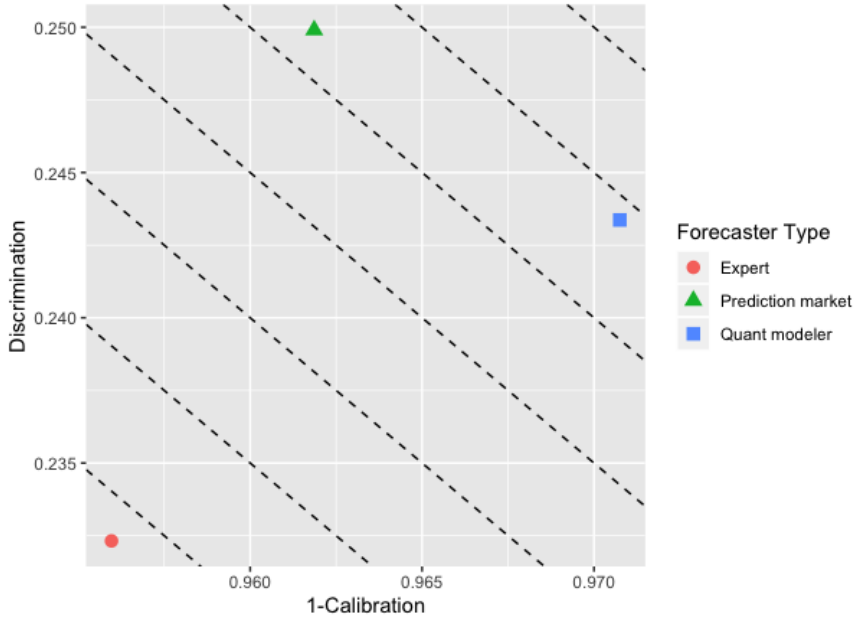


Figure 2: Mean calibration and discrimination indices of experts (Cook Political Report, Inside Elections, Sabato's Crystal Ball), prediction markets (PredictIt), and quantitative modelers (FiveThirtyEight, The Economist, JHK) final predictions on Election Day (November 3). Dashed-lines represent sets of calibration-discrimination trade-offs at which probability score is held constant. Lines closer to the top-right indicate superior performance.

When analyzing how the performances of different forecasters change over time (see Figure 3), we find that, as we expected, the performance of quantitative modelers and prediction markets improves over time. Figures 4 and 5 suggest that these can be attributed to small, cyclical, but steady, improvements to both discrimination and calibration indices over time. Additionally, we note that prediction markets look like they follow quantitative modelers in how they improve, but with a lag. We attribute this to PredictIt participants using data from models like FiveThirtyEight to inform their judgments. In fact, PredictIt contains weekly prediction markets specifically for predicting the output of FiveThirtyEight's models [6], suggesting market participants are not only well aware of them, but use them to inform their gambling strategies.

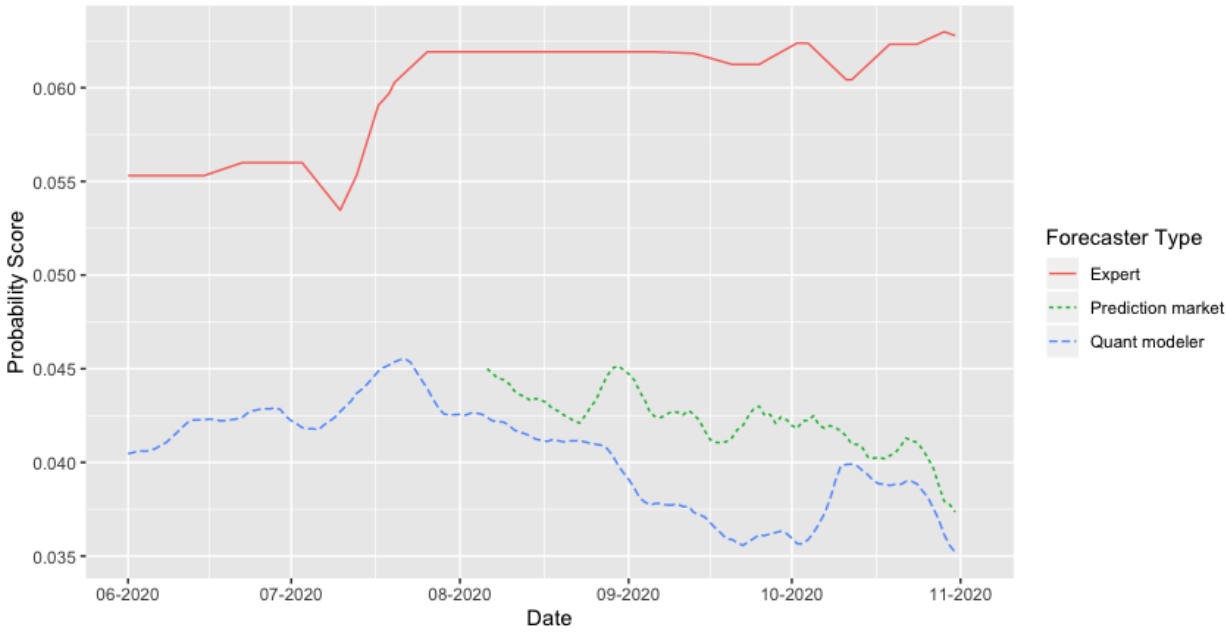


Figure 3: 7-day rolling average probability scores of experts (Cook Political Report, Inside Elections, Sabato's Crystal Ball), prediction markets (PredictIt), and quantitative modelers (FiveThirtyEight, The Economist, JHK) from June 1-November 3 (prediction market data is only available after August 2). Negative slopes indicate improving performance.

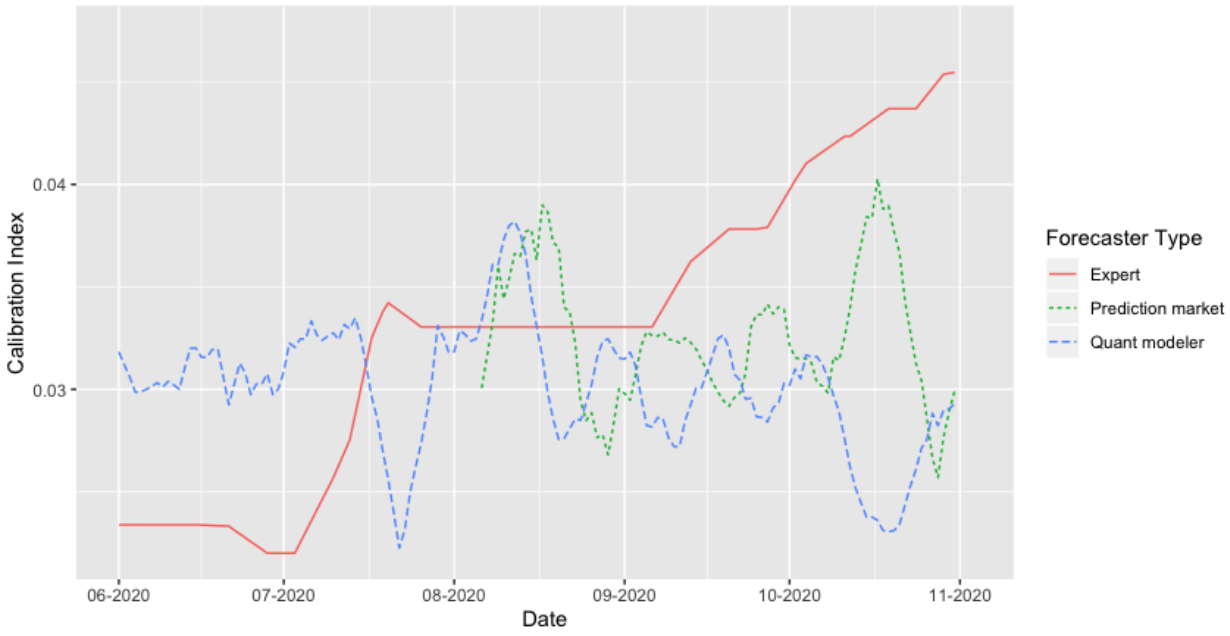


Figure 4: 7-day rolling average calibration indices of experts (Cook Political Report, Inside Elections, Sabato's Crystal Ball), prediction markets (PredictIt), and quantitative modelers (FiveThirtyEight, The Economist, JHK) from June 1-November 3. Negative slopes indicate improving calibration.

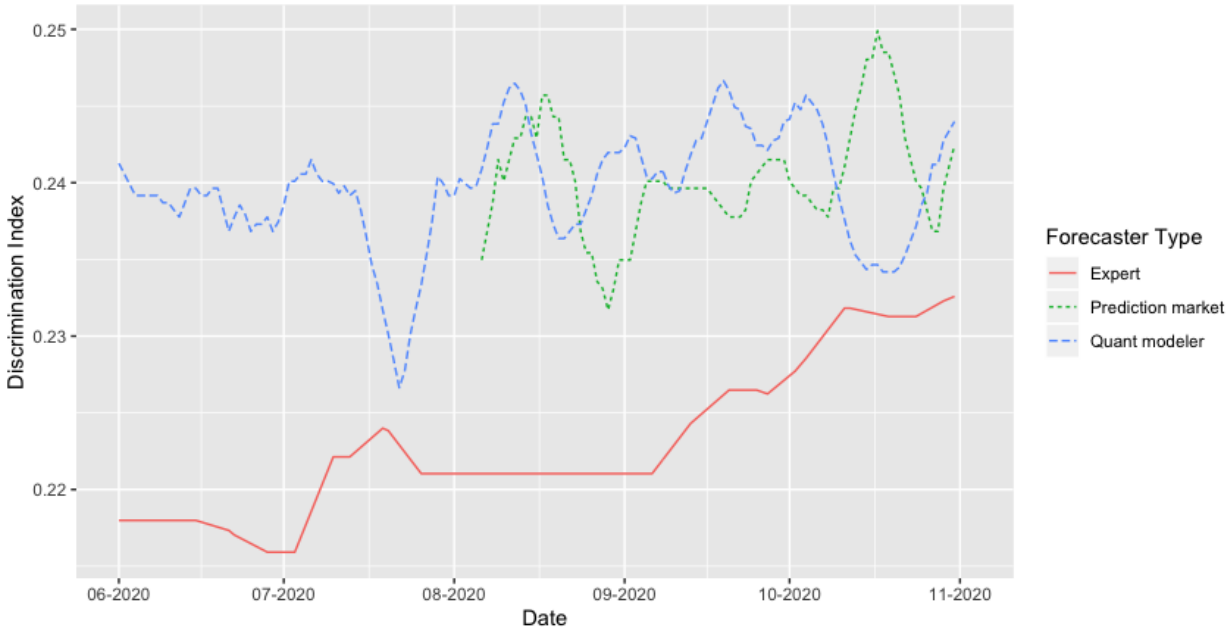


Figure 5: 7-day rolling average discrimination indices of experts (Cook Political Report, Inside Elections, Sabato's Crystal Ball), prediction markets (PredictIt), and quantitative modelers (FiveThirtyEight, The Economist, JHK) from June 1-November 3. Positive slopes indicate improving discrimination.

Unexpectedly, however, we find that the performance of experts actually diminishes over time. Figure 3 shows that this decline starts in mid-July, coinciding with a similar decline in quantitative modeler predictions. This drop coincides with mid-summer polling data that showed Biden ahead by wide margins. Indeed, Cook Political Report started

to suggest around this time that the election would be a "Democratic tsunami" instead of just a "blue wave" [10]. Quantitative modelers soon recovered and their probability scores started improving again at the end of July. However, expert predictions do not get any better, and stay relatively constant after the end of July. We attribute this to experts' overconfidence in not adjusting their predictions to new polling data showing a tightened Biden lead. Quantitative models, however, quickly adjusted to new polls showing this, resulting in their predictions recovering.

The cyclical nature of Figures 4 and 5 suggests that the discrimination and calibration of both quantitative modelers and prediction markets frequently alternate. This means it is possible that the improvement in discrimination of prediction markets is due to randomness, rather than being indicative of gamblers sharply increasing the certainty of their predictions, in the week leading up to Election Day. However, given that Figures 4 and 5 suggest that quantitative modelers were more calibrated and better at discrimination for the majority of the time frame we considered, this last-minute surge seems to support the latter.

Finally, we note that a comparison between quantitative modelers in Appendix Figures 3 and 4 indicate they are very similar in their calibration and discrimination indices. Although each model uses different methodology, each is ultimately based on the same set of polling data. When the best performing group of forecasters This suggests that, at least through polling data, there is a limit to predicting election results.

# 4 Appendix

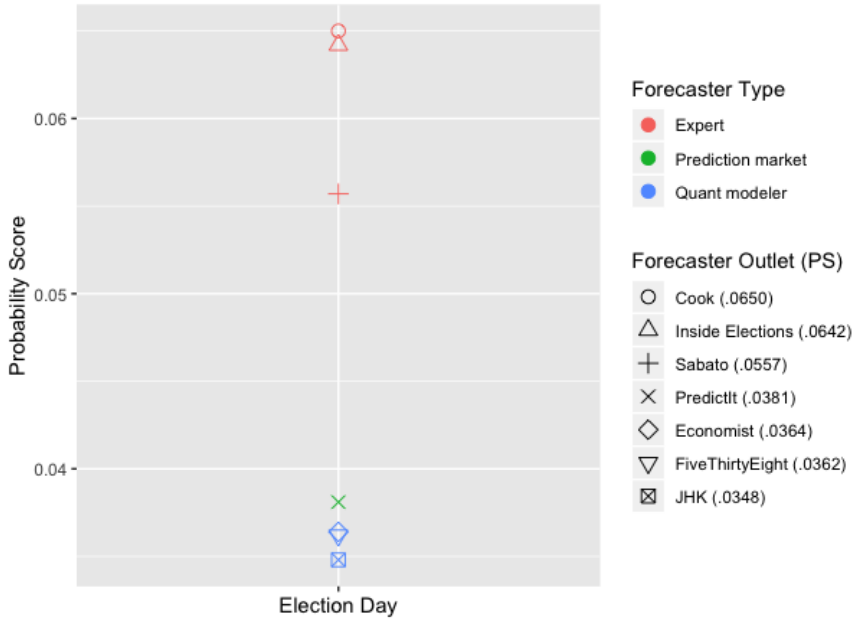


Figure 1: Mean probability scores of forecasters' final predictions on election day (November 3).

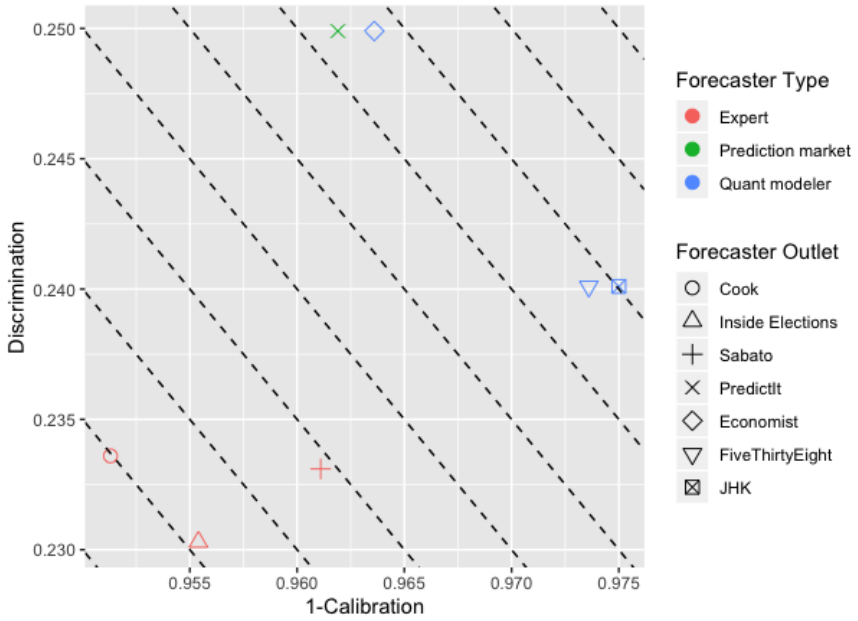


Figure 2: Mean calibration and discrimination indices of forecasters' final predictions on Election Day (November 3).

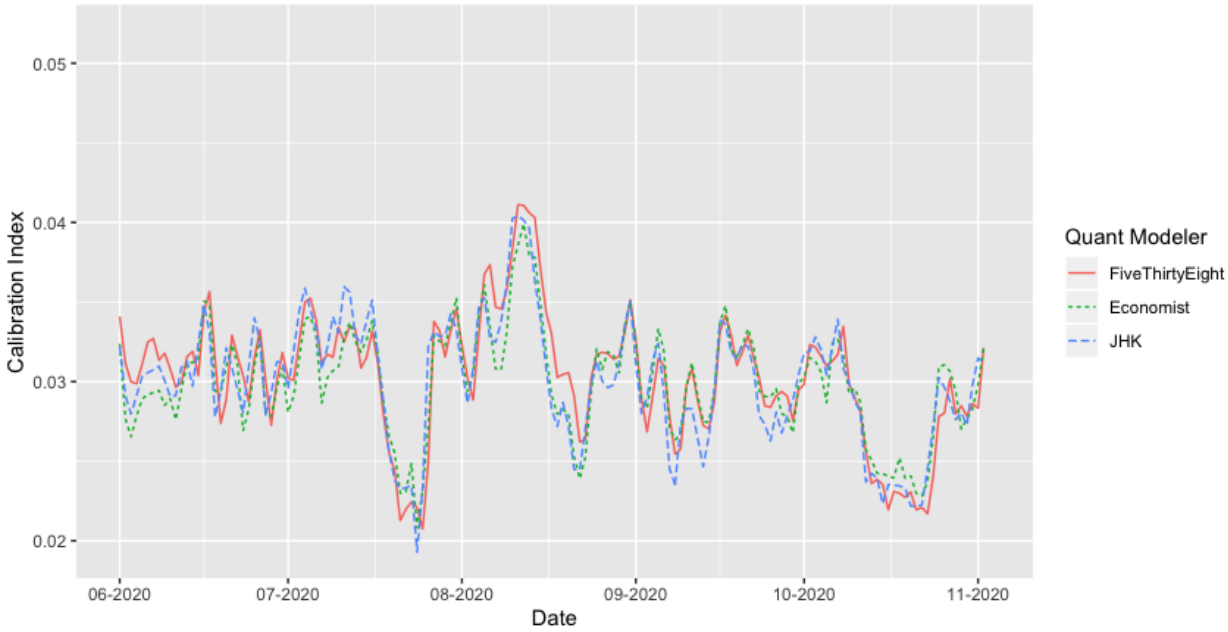


Figure 3: 7-day rolling average calibration indices of quantitative modelers from June 1-November 3.

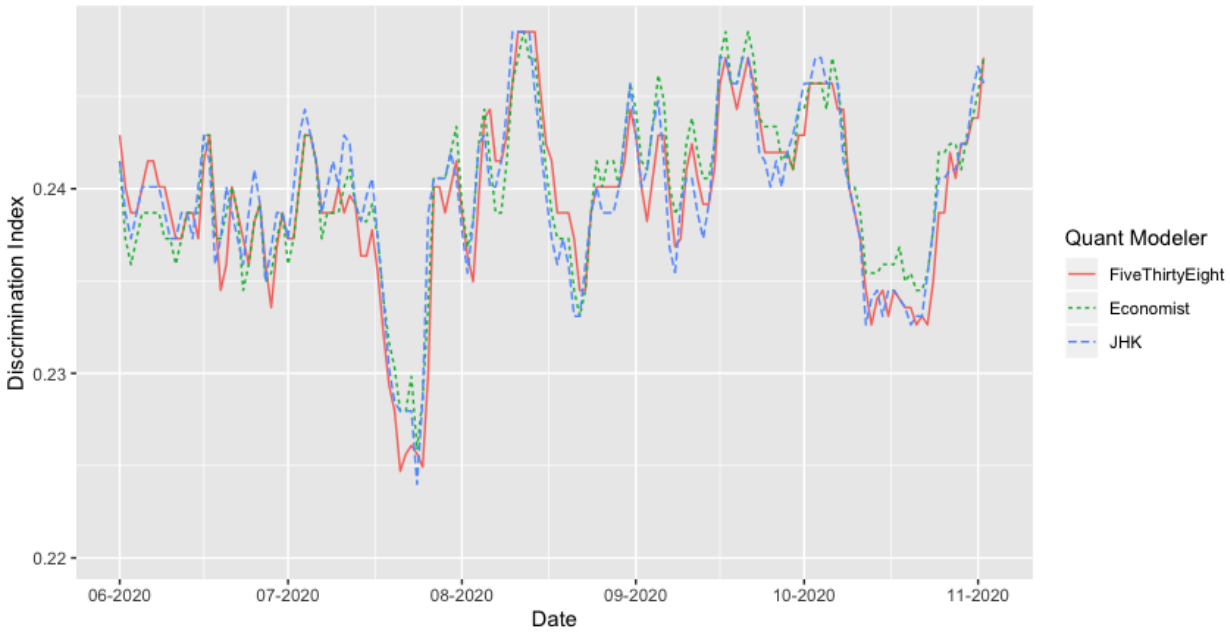


Figure 4: 7-day rolling average discrimination indices of quantitative modelers from June 1-November 3.

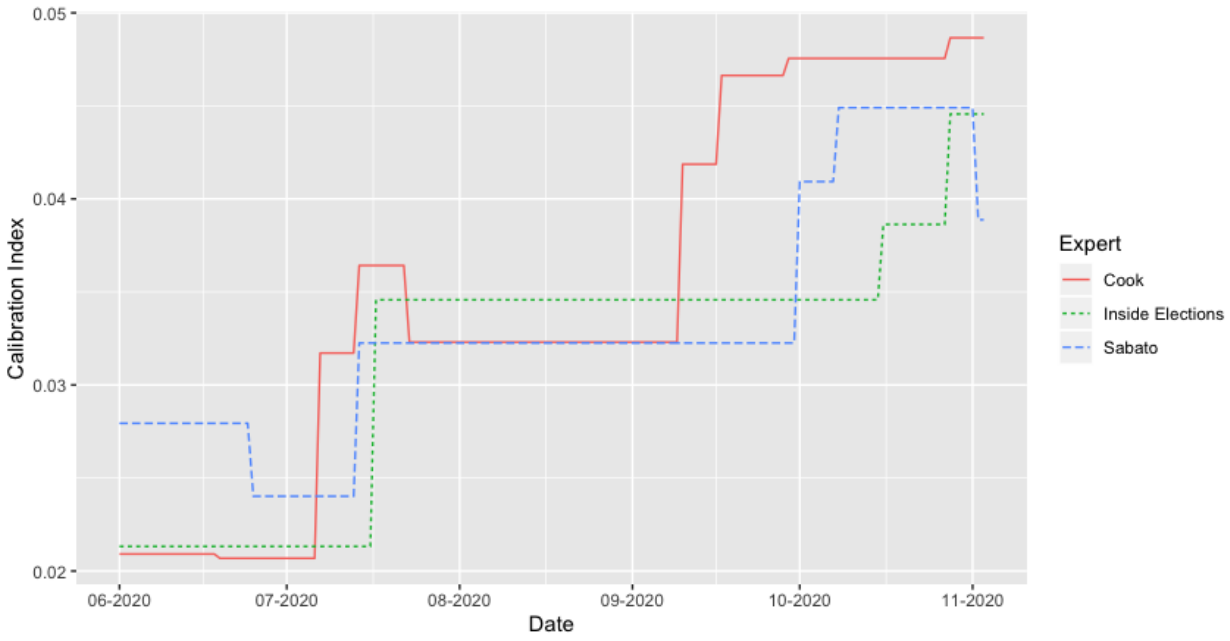


Figure 5: Mean calibration indices of experts from June 1-November 3.

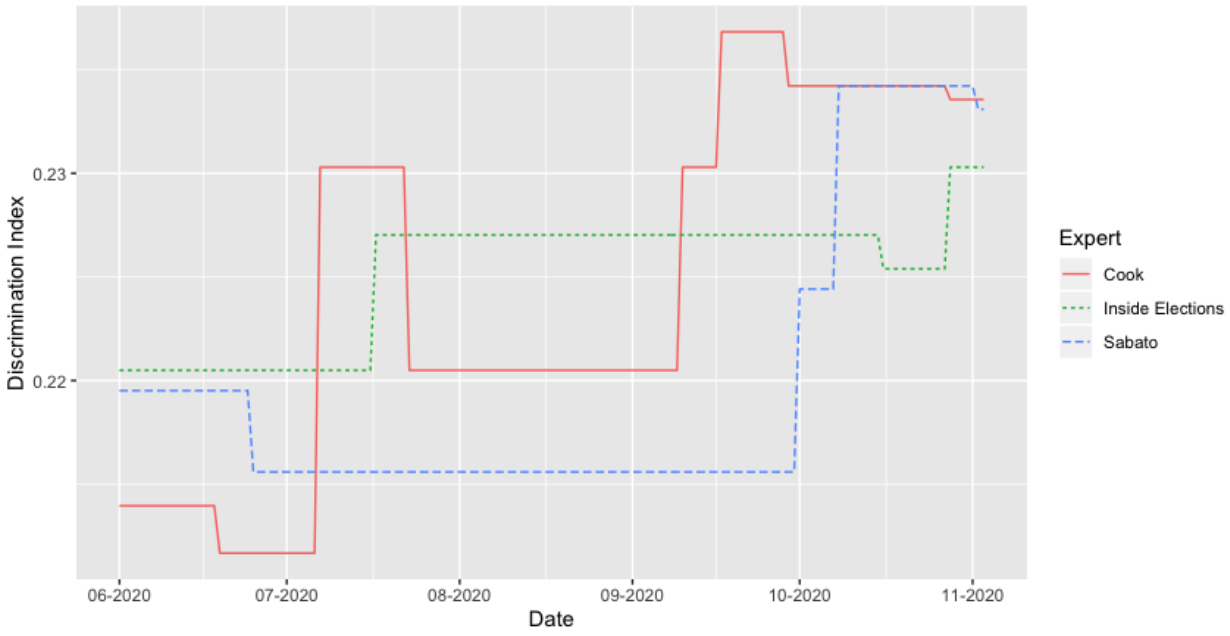


Figure 6: Mean discrimination indices of experts from June 1-November 3.

## 5 Sources

Our code is publicly available at https://github.com/jordan-klein/2020_election_predict_tourn.

Our preregistration is available on Open Science Framework at https://osf.io/vq8j2.

## References

[1] 270toWin. 2020. Consensus Forecast Electoral Map. (Nov. 2020).

[2] Charles E. Cook and Amy Walter. 2020. 2020 Electoral College Ratings. (Nov. 2020).

[3] Nathan L. Gonzales. 2020. Presidential Ratings. (Nov. 2020).

[4] Jack Kersting. 2020. 2020 Presidential Forecast. (Nov. 2020).

[5] G. Elliot Morris, Martín González, Andrew Gelman, and Merlin Heidemanns. 2020. State and national presidential election forecasting model. (Oct. 2020).

[6] PredictIt. 2020. Which party will win the Electoral College? (Nov. 2020).

[7] Larry J. Sabato, Kyle Kondik, and J. Miles Coleman. 2020. Presidential Ratings. (Nov. 2020).

[8] Nate Silver *et al.* 2020. 2020 Election Forecast. (Nov. 2020).

[9] Philip E Tetlock. 2017. *Expert political judgment: How good is it? How can we know?-New edition*. Princeton University Press, Princeton, NJ.

[10] Amy Walter. 2020. New July 2020 Electoral College Ratings. (July 2020). https://cookpolitical.com/analysis/national/national-politics/new-july-2020-electoral-college-ratings