Progress Report

GOVT-653

Kiernan Nicholls

March 7, 2019

Update

I have not made significant changes to the research question I initially proposed. I am trying to determine the predictive capabilities of markets as they compare to the more popular mathematical forecasting models. Are prediction markets as accurate as forecasting models? Under what conditions does each method perform? What role might prediction markets play in the Congressional campaigns?

In statistical terms: I propose the null hypothesis of no difference in proportion races correctly called by markets and models. Ultimately, I hope to perform a test of equal proportion on a string of prediction outcomes (correct or incorrect predictions).

Market Data

PredictIt.org was launched in late 2014 to host prediction markets, primarily on American politics. PredictIt is owned and operated by Victoria University of Wellington with support from Aristotle, Inc.. PredictIt partners with academic researchers, providing trading data for research purposes. After signing a data use agreement with the site, the provided me with trading data from 118 markets pertaining to 2018 Midterm elections.

The raw data spans 675 days from January 1, 2017 to December 12, 2018. There are 44,711 observations of the following 11 variables:

1. Market ID
2. Market question
3. Market symbol
4. Contract name
5. Contract symbol
6. Prediction date
7. Opening contract price
8. Low contract price
9. High contract price
10. Closing contract price
11. Volume of shares traded

Table 1: 10 of 44,711 observations with 7 of 11 variables

ID	Market	Contract	Date	Open	Close	Volume
4030	CA39.2018	GOP.CA39.2018	2018-02-26	0.31	0.31	0
4255	MN03.2018	GOP.MN03.2018	2018-09-26	0.19	0.21	2
3812	AZSEN18	DEM.AZSEN18	2017-11-16	0.48	0.57	41
3532	LEWI.MN02.2018	n/a	2018-05-09	0.36	0.36	0
4271	PA17.2018	DEM.PA17.2018	2018-03-30	0.79	0.87	16
2941	MANCHIN.WVSENATE.2018	n/a	2018-04-09	0.73	0.70	116
4156	FL17.2018	DEM.FL17.2018	2018-10-07	0.02	0.02	0
3767	NH01.2018	DEM.NH01.2018	2018-06-12	0.83	0.83	0

ID	Market	Contract	Date	Open	Close	Volume
3503	KING.MESENATE.2018	n/a	2017-11-06	0.87	0.87	0
4257	IL12.2018	DEM.IL12.2018	2018-04-26	0.53	0.53	0

Each market poses a question (Which party will win the 2018 House of Reps race in Texas's 21st district?). The possible answers to that question (Democratic or Republican) are the contracts that comprise the market. When a trader is interested in buying shares of a contract (100 shares of a Democratic party winning the Texas 21st for $0.69 each), they make an open offer on the market. A corresponding trader agrees to buy the converse contract (100 shares of a Republican party winning the Texas 21st for $0.31 each). Those traders can buy or sell these shares throughout the election at whatever price another trader agrees on. After the election, each correct contract executes at $1.00, with a 10% fee going towards the operational costs of the exchange.

The price of a contract is directly proportional to the trader's probabilistic interpretation of the election outcomes. If a trader believes a party has a high chance of winning an election, he will not place a bet without a low amount of risk. The binary outcome of the futures contracts allow for a direct probabilistic interpretation of the election results.

In market theory, the volume of shares traded plays a crucial role in proper price discovery; too few shares traded and the market may not properly react to changes in the election circumstances. In my analysis, a market price over $0.50 indicates a prediction of that candidate winning the election. The closing price of a contract represents that day's final market prediction. We can compare each day's prediction with the eventual winner to assess the accuracy. The proportion of all races correctly predicted represents the accuracy of the markets method.

Model Data

FiveThirtyEight.com was launched in 2008 by Nate Silver to aggregate polls of the Democratic Presidential Primary to better forecast the winner. In the decade since, the model used by FiveThirtyEight has grown in complexity. For the 2018 Midterm elections, FiveThirtyEight published models for House, Senate, and Governors races. The models incorporate quantitative inputs (primarily polling) to simulate the election and produce a probabilistic view of the election.

The team at FiveThirtyEight makes public the top-line output of their models as four separate `.csv` files on their website:

1. `senate_national_forecast.csv`
2. `senate_seat_forecast.csv`
3. `house_national_forecast.csv`
4. `house_district_forecast.csv`

The Senate seat and House district level forecasts will be used in this project. Each observation represents one day's probability of victory for one candidate. There are 28,353 observations at the Senate seat level and 302,859 at the House district level. Together, There are about 3,380 unique daily predictions from (97 days).

The raw data spans 97 days from August 1st to November 5th. Together, the Senate and House data sets contain 328,113 observations of 12 variables:

1. Prediction date
2. Election state
3. Election Congressional district
4. Whether the election is a "special election"
5. Candidate's full name
6. Candidate's political party
7. Whether the candidate is an incumbent

8. Model version (classic, lite, or deluxe)
9. Candidate's probability of victory
10. Candidate's expected share of the vote
11. Candidate's approx. minimum share of the vote
12. Candidate's approx. maximum share of the vote

Table 2: 10 of 299,760 observations with 9 of 12 variables

Date	State	District	Special	Party	Incumbent	Model	Win Probability	Expected Share
2018-10-01	MD	4	NA	R	FALSE	deluxe	0.000	18.67
2018-10-08	NC	7	NA	R	TRUE	lite	0.778	52.06
2018-08-20	MI	6	NA	R	TRUE	deluxe	0.892	53.47
2018-08-13	CA	11	NA	D	TRUE	deluxe	1.000	77.44
2018-10-26	GA	10	NA	R	TRUE	deluxe	1.000	67.81
2018-10-19	AL	7	NA	D	TRUE	lite	1.000	100.00
2018-09-05	TX	8	NA	R	TRUE	deluxe	1.000	75.15
2018-08-04	TX	11	NA	LIB	FALSE	deluxe	0.000	3.31
2018-09-28	TX	7	NA	D	FALSE	classic	0.480	49.80
2018-09-18	SC	7	NA	D	FALSE	lite	0.036	39.82

Tidy Data

The data from the markets and model can be combined and cleaned to produce a single data frame with 26,778 observations of 10 variables:

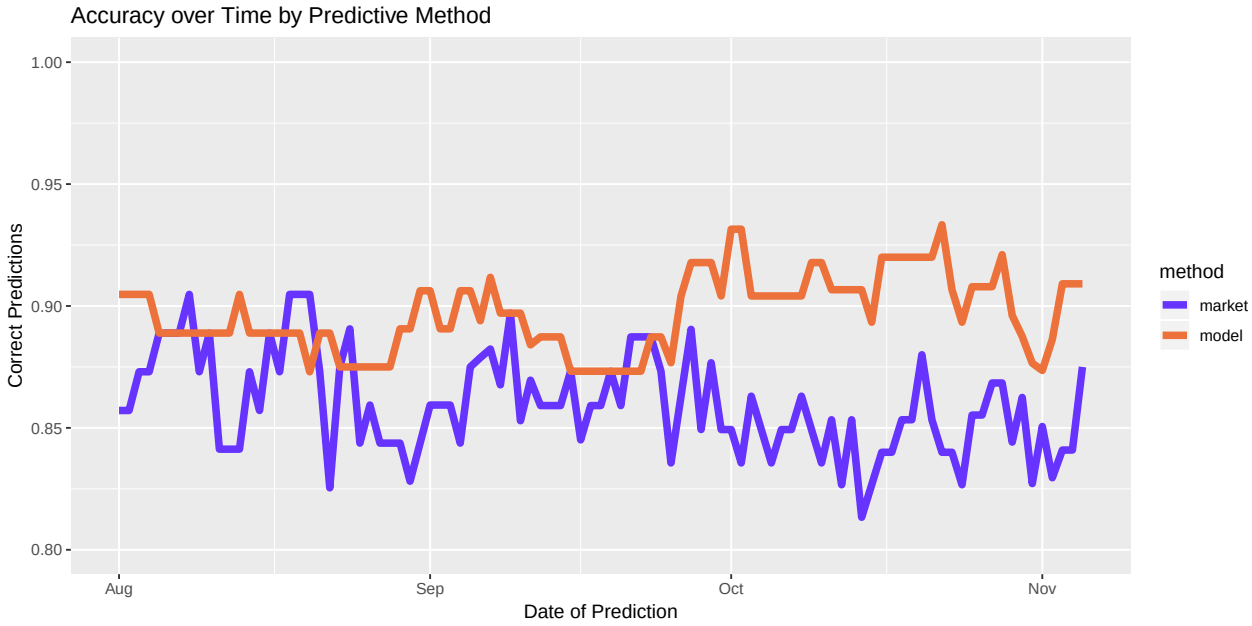
1. Prediction date
2. Election code
3. Candidate's name
4. Election chamber
5. Candidate's party
6. Whether the election is a "special election"
7. Whether the candidate is an incumbent
8. Whether the prediction comes from the markets or model
9. Candidate's probability of victory

Table 3: 10 of 26,778 observations with 9 variables

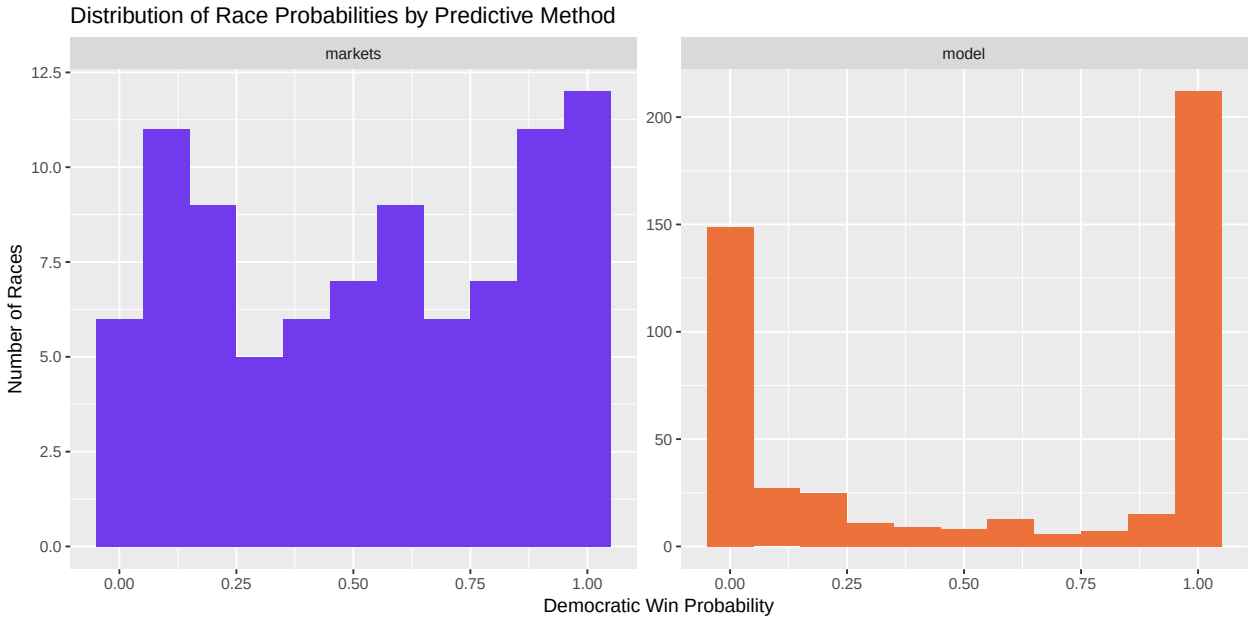
Date	Race	Name	Chamber	Party	Special	Incumbent	Method	Probability
2018-09-26	CT-05	Hayes	house	D	FALSE	FALSE	market	0.950
2018-08-09	NC-13	Manning	house	D	FALSE	FALSE	market	0.550
2018-08-09	NY-01	Zeldin	house	R	FALSE	TRUE	market	0.710
2018-10-17	CA-10	Denham	house	R	FALSE	TRUE	market	0.290
2018-08-27	MA-99	Warren	senate	D	FALSE	TRUE	model	0.999
2018-09-02	NV-03	Lee	house	D	FALSE	FALSE	model	0.687
2018-09-25	MD-06	Trone	house	D	FALSE	FALSE	market	0.990
2018-10-27	IA-01	Blum	house	R	FALSE	TRUE	model	0.034
2018-08-23	PA-15	Boser	house	D	FALSE	FALSE	market	0.040
2018-09-23	VA-06	Cline	house	R	FALSE	FALSE	market	0.920

Exploratory Plots

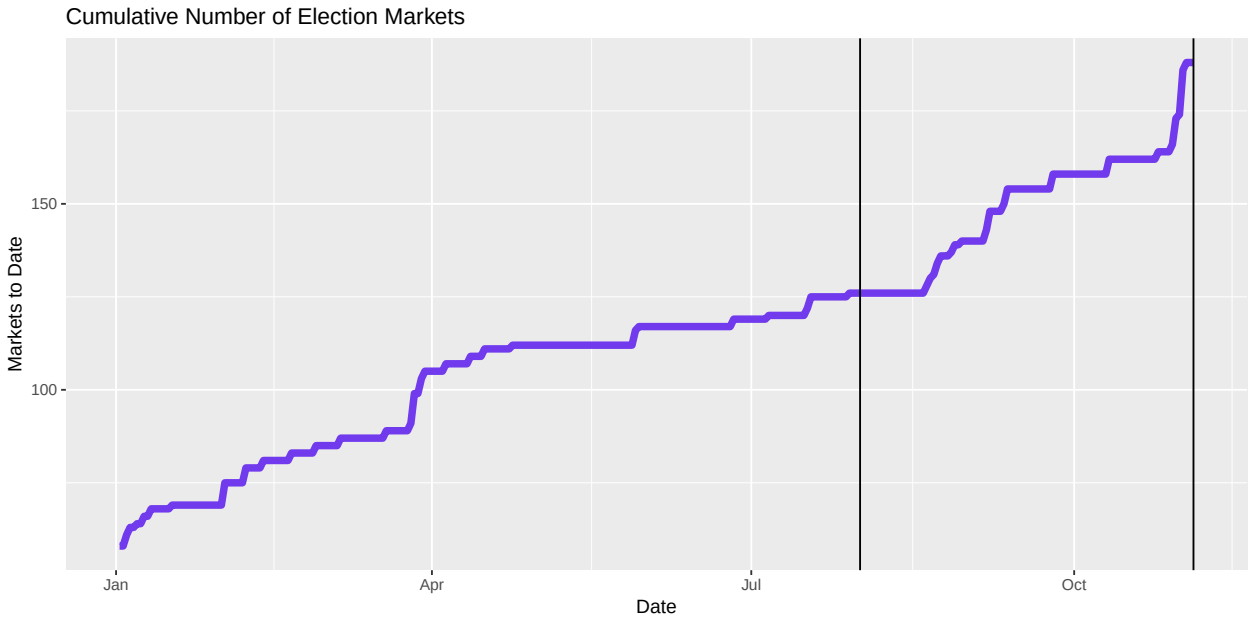
Below is a preliminary comparison of of each method's percentage of correct predictions over time. More work needs to be done to properly account for all the differences in the 111 races.



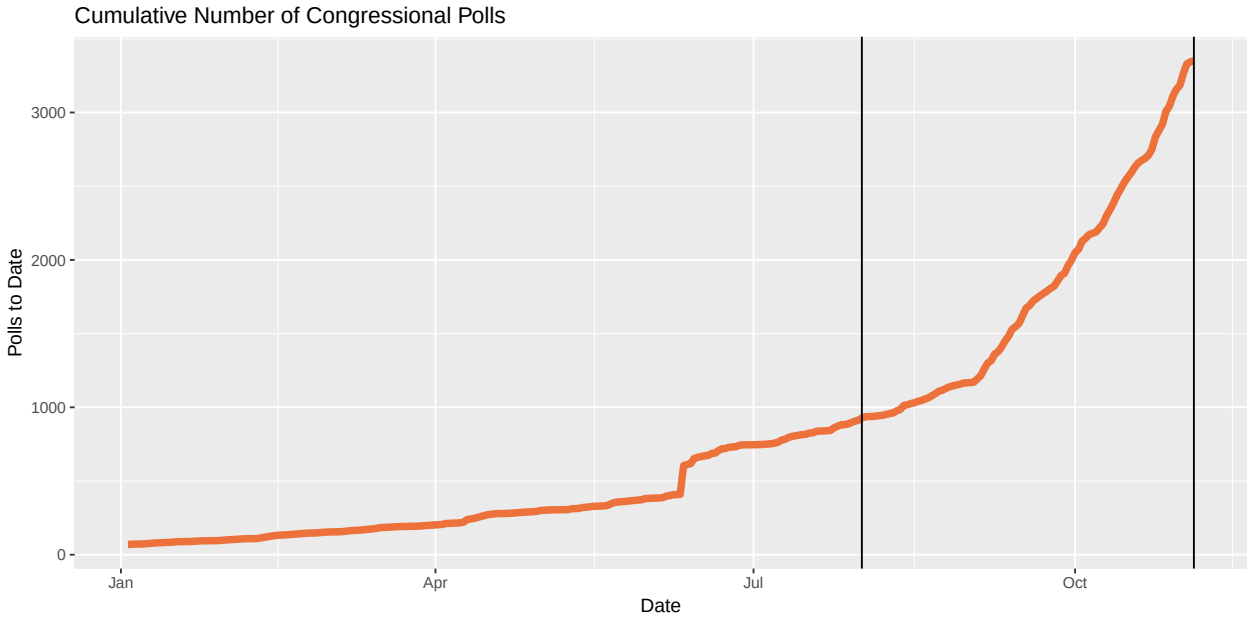
Below are histograms comparing the distribution of probabilities in the raw data sets from PredictIt and FiveThirtyEight the day before the election. Note how the model predicts every race every day while the markets only focus on races of interest to the traders. The races of interest are those where the traders believe they can make a successful bet against the market, meaning a greater proportion of the races predicted by markets are toss-ups. Both methods will have to be assessed on these shared races of interest.

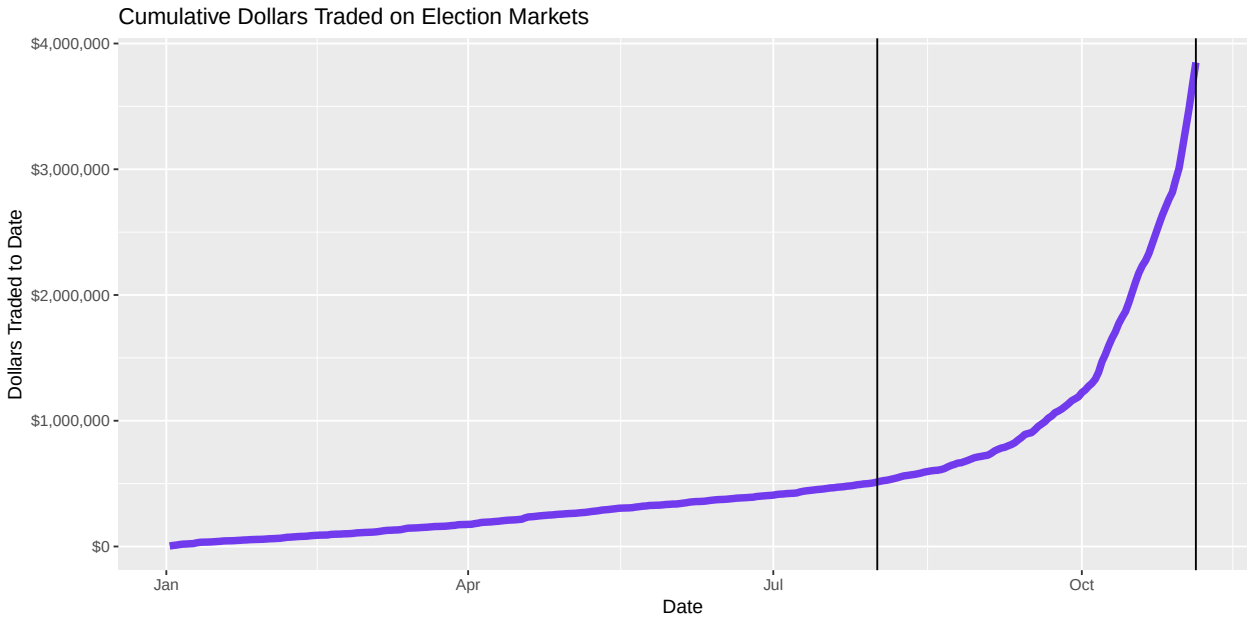


Another point of difference is the number of predictions made by each method over time. The model model predicts every race every day, whereas more markets are added to the exchange every day. Below is a plot showing the number of open midterm election markets hosted on the PredictIt exchange in 2018.



Below are plots showing the the increase in the two underlying inputs in each method. While the model takes into account a number of quantitative factors, polling still plays the largest and more useful roll in predicting an election. Similarly, the accuracy of prediction markets relies on the volume of the markets; the greater the volume of money being exchanged on the markets, the greater the economic forces of price discovery and can capture the underlying probabilities associated with the market equilibrium.




Cumulative Dollars Traded on Election Markets
$4,000,000
$3,000,000
$2,000,000
$1,000,000
$0
Dollars Traded to Date
Jan
Apr
Jul
Oct
Date