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 (or more) accurate that forecasting models, and if so, under what conditions? What role might prediction markets play in the American Congressional campaign?

In statistical terms: I propose the null hypothesis of no difference in proportion races correctly called by markets and models. Over one hundred races of interest will are predicted daily, from August 1st to November 5th, by both the markets on PredictIt and the model by FiveThirtyEight.

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
4341	LAMB.CO05.2016	n/a	2018-04-24	0.31	0.26	121
4638	CA48.2018	DEM.CA48.2018	2018-10-26	0.52	0.56	154
2941	MANCHIN.WVSENATE.2018	n/a	2018-08-13	0.76	0.83	2074
4255	MN03.2018	GOP.MN03.2018	2018-10-07	0.26	0.14	3
4177	PASEN18	DEM.PASEN18	2018-10-04	0.88	0.93	1
2940	SANDERS.VTSENATE.2018	n/a	2017-06-24	0.89	0.88	1
2928	CRUZ.TXSENATE.2018	n/a	2018-02-21	0.76	0.75	118
4831	NY22.2018	DEM.NY22.2018	2018-09-30	0.79	0.79	0

ID	Market	Contract	Date	Open	Close	Volume
2940	SANDERS.VTSENATE.2018	n/a	2018-06-26	0.92	0.93	38
4015	MD06.2018	DEM.MD06.2018	2018-01-21	0.91	0.91	1

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 aggrigate 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 incorperate quantitative inputs (primarily polling) to simulate the election and produce a probabalistic 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 candiate 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
2018-09-28	MD	4	NA	R	FALSE	deluxe	0.000	18.61
2018-09-21	IL	3	NA	D	TRUE	classic	1.000	68.07
2018-10-13	MO	4	NA	LIB	FALSE	deluxe	0.000	3.28
2018-09-11	NC	2	NA	D	FALSE	lite	0.309	46.36
2018-09-22	TN	1	NA	D	FALSE	deluxe	0.000	23.91
2018-09-11	IL	17	NA	R	FALSE	deluxe	0.000	34.42
2018-08-09	TX	3	NA	LIB	FALSE	classic	0.000	3.43
2018-09-23	AZ	9	NA	R	FALSE	classic	0.007	39.61
2018-09-15	AL	1	NA	D	FALSE	classic	0.001	35.06
2018-08-03	CA	37	NA	R	FALSE	deluxe	0.000	10.17

Tidy Data

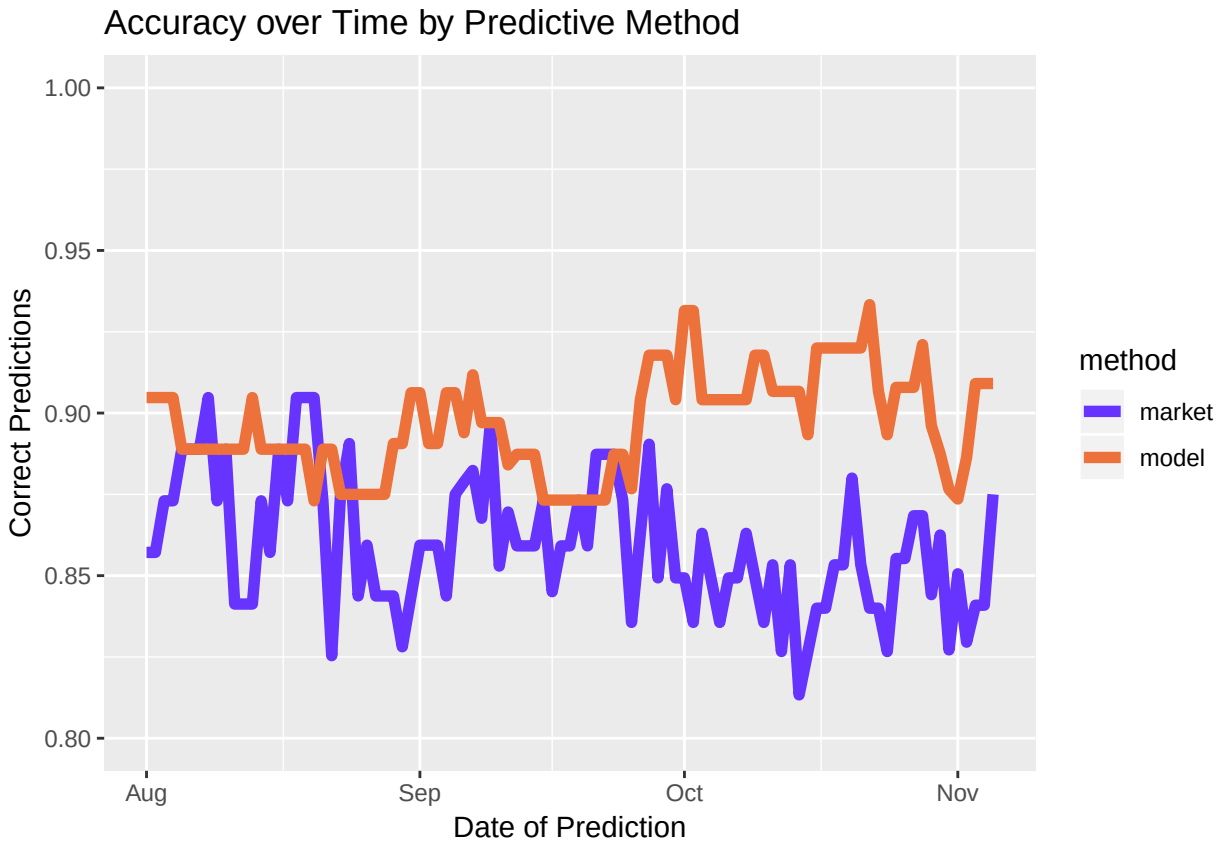
The data from the markets and model can be combined and cleaned to produce a single data frame with 26,778 observations of 10 variables:

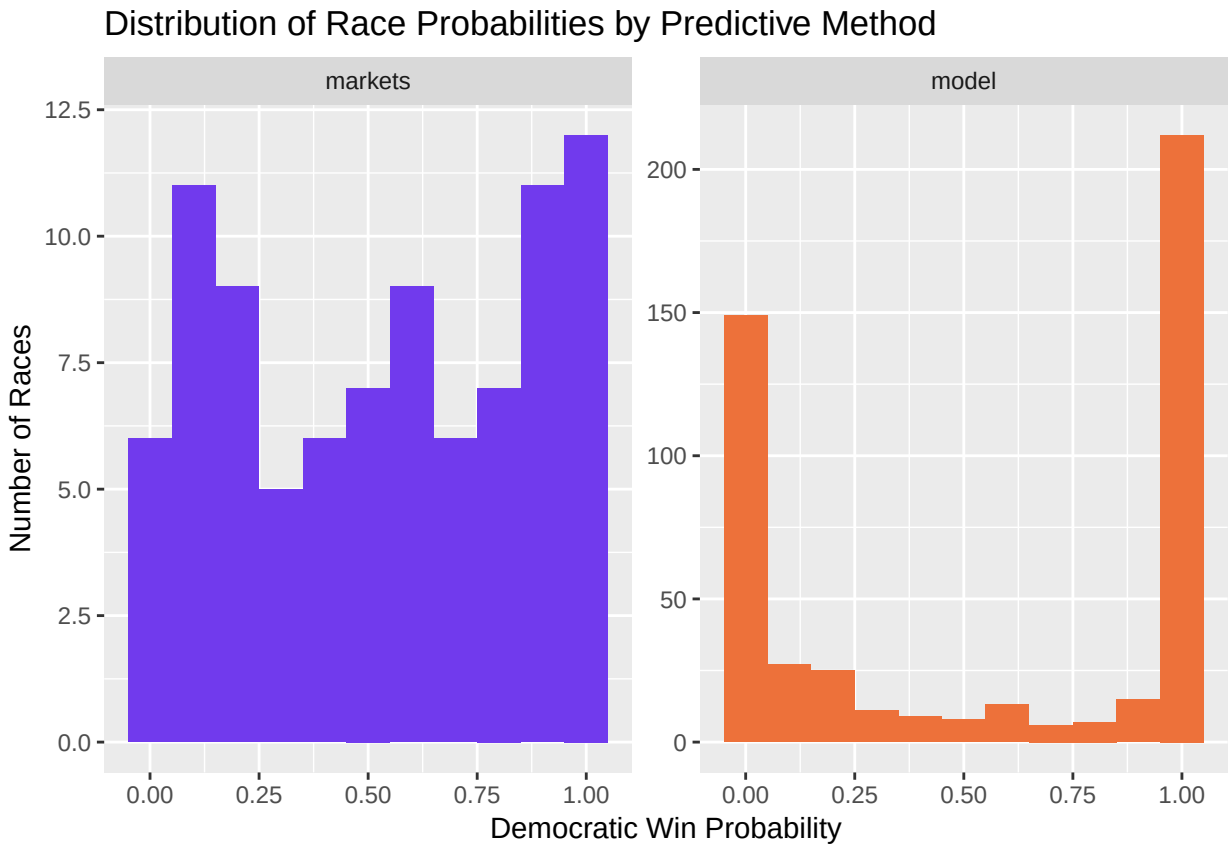
1. Prediction date
2. Election code
3. Candidate's name
4. Election chamber
5. Candidate's party
6. Whether the election is a "special election"
7. Whether the candiate is an incumbent
8. Whether the prediction comes from the markets or model
9. Candidate's probability of victory
10. Whether the prediction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

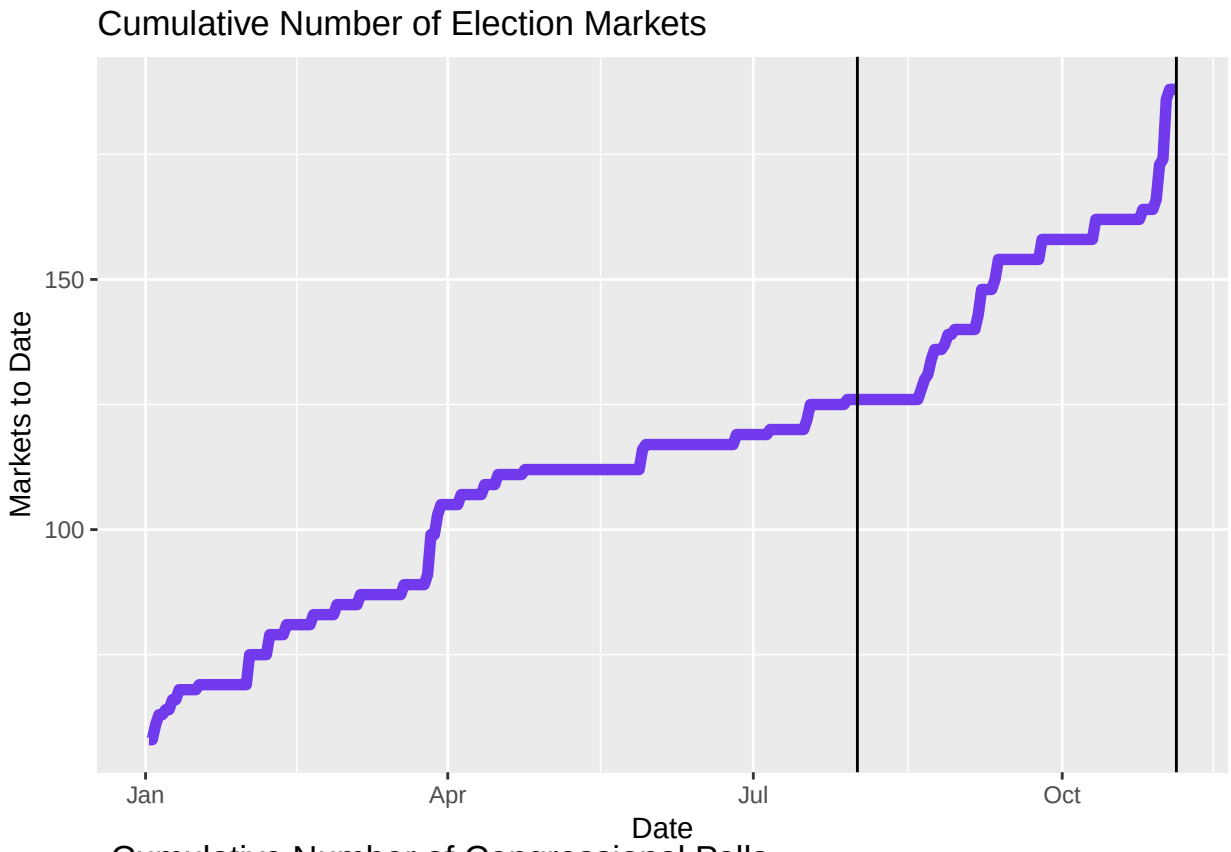
Table 3: 10 of 26,778 observations with 9 of 10 variables

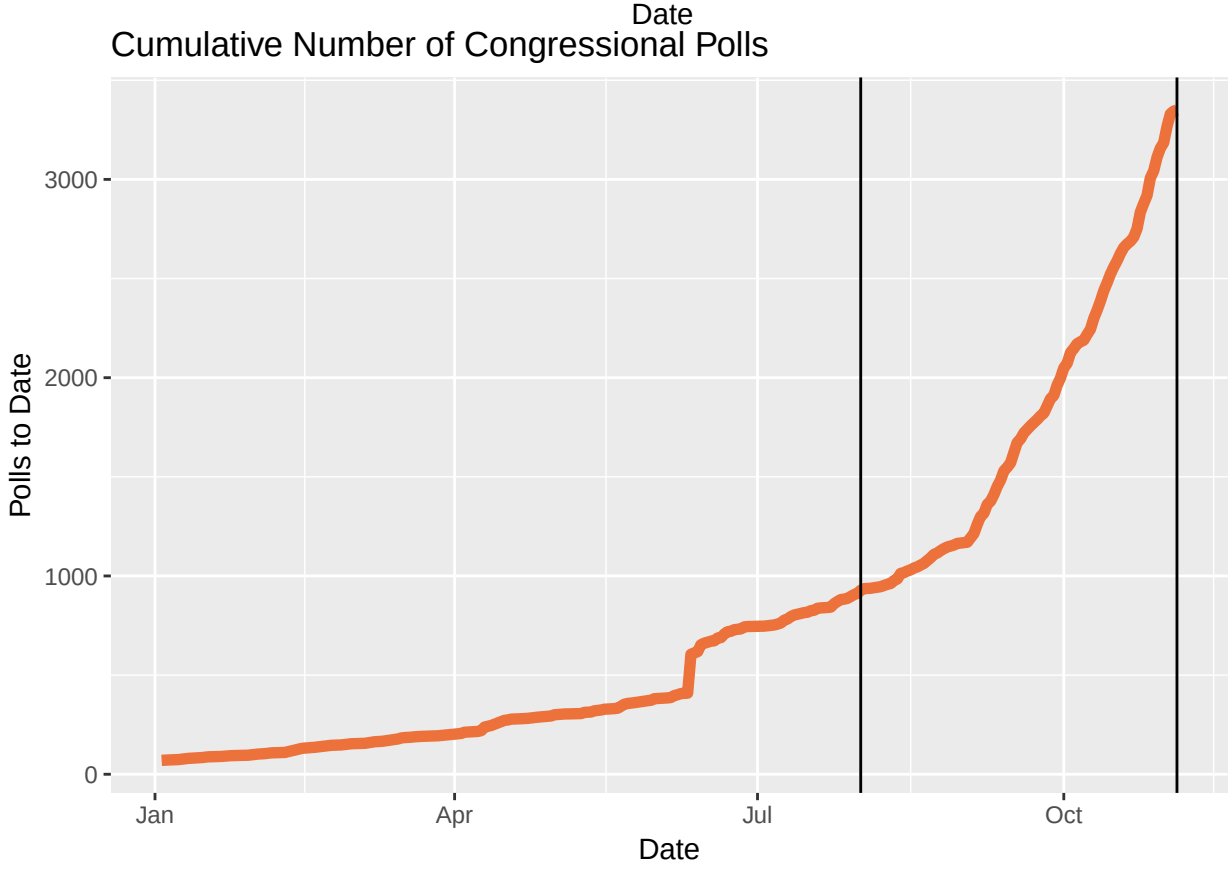
Date	Race	Name	Chamber	Party	Incumbent	Method	Probability	Outcome
2018-09-19	MI-99	Stabenow	senate	D	TRUE	model	0.992	TRUE
2018-08-01	PA-15	Boser	house	D	FALSE	market	0.070	TRUE
2018-10-21	OH-01	Pureval	house	D	FALSE	model	0.224	TRUE
2018-10-02	MN-08	Stauber	house	R	FALSE	market	0.550	FALSE
2018-10-13	WV-99	Manchin	senate	D	TRUE	model	0.867	TRUE
2018-11-05	MI-99	Stabenow	senate	D	TRUE	model	0.960	TRUE
2018-08-31	WA-08	Schrier	house	D	FALSE	market	0.550	TRUE
2018-10-25	NY-09	Gayot	house	R	FALSE	model	0.000	FALSE
2018-10-29	PA-99	Casey	senate	D	TRUE	market	0.920	TRUE
2018-08-16	WV-03	Ojeda	house	D	FALSE	model	0.060	TRUE

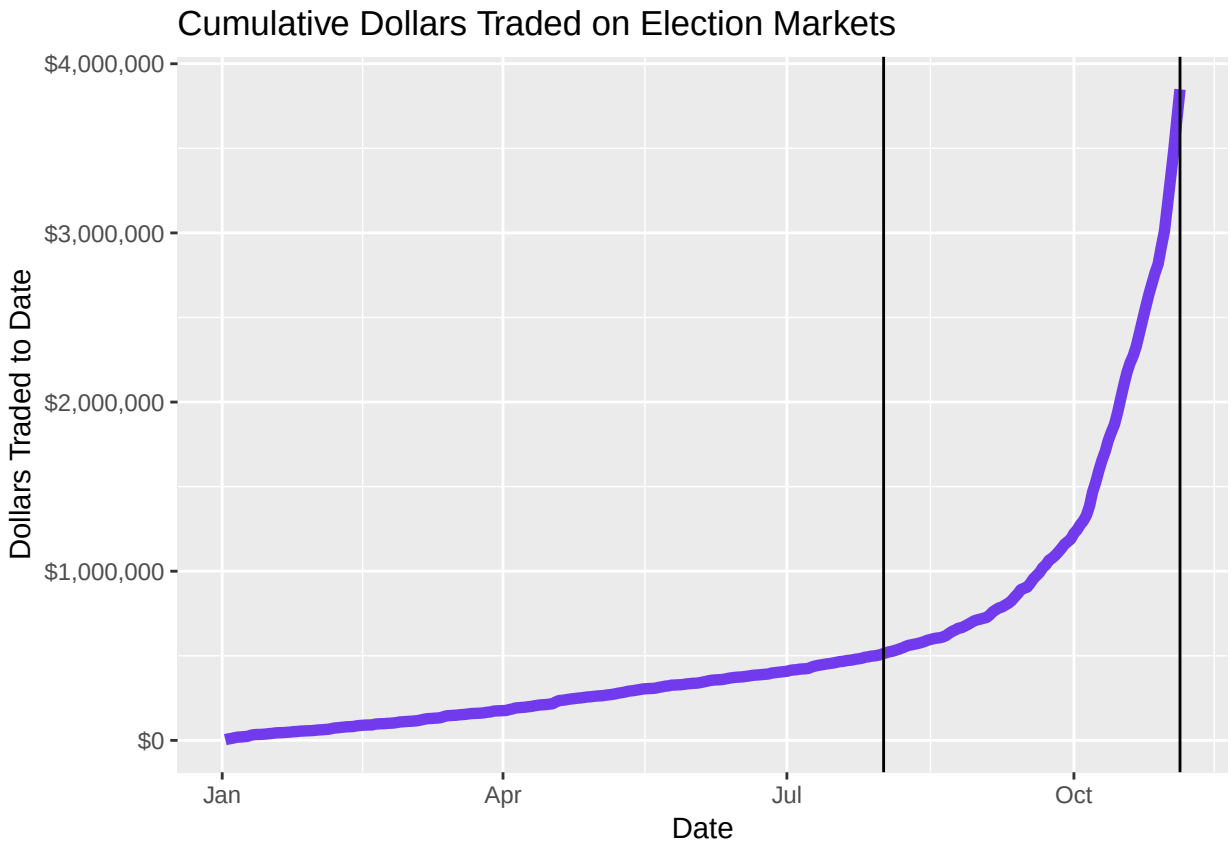
Exploratory Plots






Cumulative Number of Election Markets
Markets to Date
150
100
Jan
Apr
Jul
Oct
Date


Cumulative Number of Congressional Polls
Polls to Date
3000
2000
1000
0
Jan
Apr
Jul
Oct
Date


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