Predicting Donald Trump's Tweets

Adam Blakney

Abstract

In this paper we apply several statistical learning, modeling, and forecasting techniques to the problem of predicting the number of tweets Donald Trump will post in a certain amount of time. We tailor our approach to the constraings of PredictIt's Tweet markets.

1 Introduction

Prediction markets are markets that allow users to bet on the occurence of events in the future [3]. In particular, prediction markets may offer the opportunity to speculate about the outcomes of sports games, election outcomes, etc. PredictIt is one such prediction market that focuses on political events [4]. In particular, PredictIt offers markets that allow users to speculate on the number of tweets that certain political figures will post in a pre-determined amount of time. Many different factors may be considered in the task of predicting the Twitter behavior of a political figure; past Twitter behavior, current political climate, and other current events may be relevant. In this paper, we focus on the use of past Twitter behavior to predict future Twitter behavior. In particular, using the past Twitter behavior of the user @realDonaldTrump (see [1]) we aim to predict how many tweets @realDonaldTrump will post in time spans ranging from twelve hours to one week.

2 PredictIt Tweet Markets

In this section we give an brief overview of PredictIt's twitter markets, as well as a basic review of prediction markets. [3] and [2] should be consulted for further information on prediction markets and trading.

2.1 Markets, Brackets and Shares

A market is composed of brackets, which specify possible market outcomes. Users interact with the market by buying "Yes" or "No" shares in each bracket, each of which is priced between \$.01 and \$.99. When the market resolves, exactly one bracket will resolve as the winning bracket while all others resolve as losing brackets. At market closure, a winning bracket's "Yes" shares will be worth \$1; similarly a losing bracket's "No" shares will be worth \$1. Thus, generally speaking, a user is incentivized to buy "Yes" shares in brackets they believe will win and "No" shares in brackets they believe will lose.

2.2 Maximizing Profit: Probabilistic Interpratation

While the specifics of trading shares are rather involved (see [2]), here we discuss some basic facts about prediction markets, and explore how one might profit off of miscalibrated markets. Specifically, in this section, we formalize the notion of expected profit and provide a framework for interpreting prediction markets probabilistically. To formalize these concepts we first introduce the following definitions.

Definition 1. *A market M is a collection of n brackets, denoted $b_1, ..., b_n$, one of which will resolve as the winning bracket, and the rest of which will resolve as losing brackets. The index of the winning bracket is unknown until the market closure.*

Definition 2. *For each bracket b_i, denote the "Yes" price by s_i, and denote the probability of it resolving as the winning bracket p_i. (Consequently, the probability of it resolving as a losing bracket is $1 - p_i$).*

Clearly, the probabilities $p_1, ..., p_n$ are unkown: otherwise, assuming a rational market, the share prices $s_1, ..., s_n$ would converge to them and the market could not be exploited. We now consider the impact of buying "Yes" or "No" shares in a market.

Definition 3. *Define w_i to be the expected profit after buying one "Yes" share in bracket i and $\bar{w}_i$ to be the expccted profit after buying one "No" share in bracket i. Then,*

$$w_i = .9p_i - s_i,$$

and,

$$\bar{w}_i = .9(1 - p_i) - (1 - s_i),$$

where the factor of .9 arises because PredictIt takes 10% of profits. Thus we note the following:

Remark. *It is profitable to to buy a "Yes" or "No" share for bracket b_i when,*

$$.9p_i > s_i,$$

or,

$$.9(1 - p_i) > 1 - s_i,$$

respectively.

Thus, our goal is to estimate the probabilities $p_1, ..., p_n$ reliably.

2.3 @realDonaldTrump Market Example

For each selected Twitter user, a new market is opened each week, and the market brackets—which users may place bets on—are associated with the number of tweets to be posted by the account over that week. Fig.1 shows a snapshot of a @realDonaldTrump Twitter market. The left-hand side gives the bracket outcomes, which specify the number of tweets to be posted by the user. Also visible in the left-hand side is the users' stake in each bracket. Numbers boxed in green signify stake in a "yes" contract while those in red signify stake in a "no" contract. For example, we see that in Fig. 1, under the bracket title "220 - 229" is the number 100 boxed in green, indicating that the user has stake in 100 "yes" contracts for the aforementioned bracket. The remaining columns give market prices. Note that we may interpret these market prices probabilistically. For example, consider

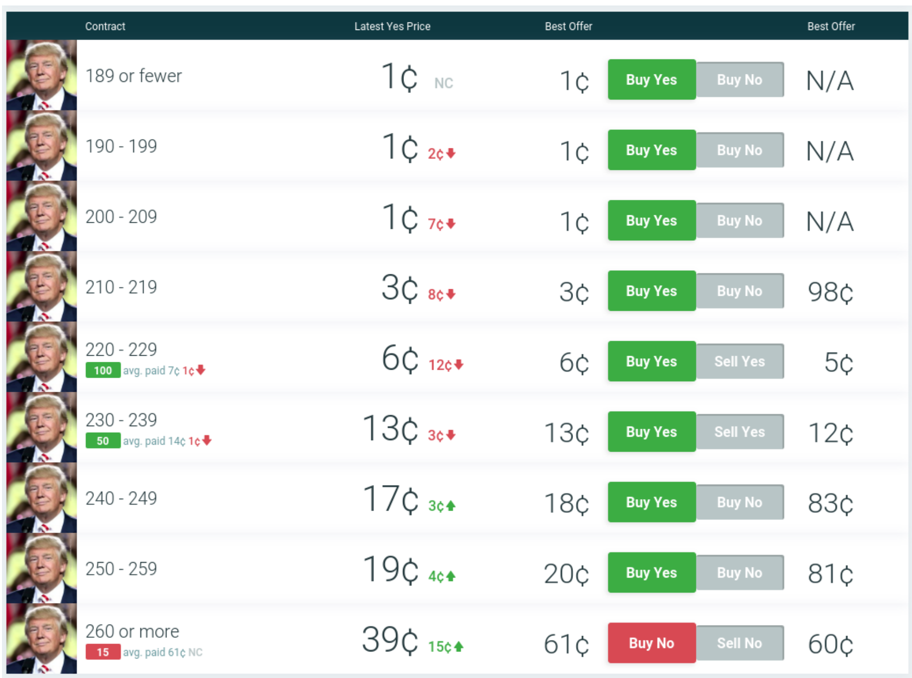

Figure 1: A snapshot of the @realDonaldTrump Twitter market with a users' stakes in various brackets.

3 Problem Approach

Recall that our goal is to estimate the probabilities $p_1, ..., p_n$ of each bracket resolving as the winning bracket. More specifically, we wish to predict the probability distribution of the number of tweets posted through the end of one week. To do so, we model the history of tweet counts as outcomes of random variables drawn from some distribution f, and use the observed values to estimate this distribution.

Consider a continuous time interval of length L partitioned into T subintervals $\tau_1, ..., \tau_T$ each of length l. Then we model the number of tweets during each of these intervals as a random variable X, with,

$$\begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} \text{number of tweets during } \tau_1 \\ \vdots \\ \text{number of tweets during } \tau_n \end{pmatrix}.$$

Thus, we assume that $X \sim f$ for some distribution f. Further we assume that Thus we assume that the number of tweets a user posts throughout a week follows some constant distribution, and we attempt to estimate this distribution through various techniques.

Given a collection of random variables $X_1, ..., X_T$ representing tweet counts during a selected time range, our goal is to predict the distribution of future tweet counts. That is, we aim to estimate the distribution of $\sum_{i=1}^{m} X_{T+i}$ for some m, with $\sum_{i=1}^{m} \hat{X}_{T+i}$, where $\hat{X}_i \sim \hat{f}$ for some

3.1 Monte Carlo Simulation

In this section we attempt to estimate the probabilities $p_1, ..., p_n$ via Monte Carlo simulation. We assume that the account's behavior is similar during similar times of the week. Thus, given W weeks of data, we partition each week into N equal length intervals, and track the number of tweets during each of these. Consider the $W \times N$ matrix of previous tweet counts X, where X_{ij} gives the number of tweets during the j-th segment of the i-th. Given X, our goal, then, is to estimate the distribution of $\hat{X}_{W+1} := (\hat{X}_{W+1,1}, ..., \hat{X}_{W+1,N})$ which gives the number of tweets during each of the N segments of the subsequent week. Then the estimate of the desired distribution is given by, $\sum_{j=1}^{N} \hat{X}_{W+1,j}$.

References

[1] Donald j. trump. abc https://twitter.com/realDonaldTrump.

[2] Donald j. trump. https://www.predictit.org/support/how-to-trade-on-predictit.

[3] Prediction market. a https://www.investopedia.com/terms/p/prediction-market.asp.

[4] Predictit. av predictit.org.