Big Data
Volume 5 Number 4, 2017

DOI: 10.1089/big.2017.0047

ORIGINAL ARTICLE

Improving Predictive Accuracy in Elections

David Sathiaraj,[1,2,*] William M. Cassidy Jr.,[3] and Eric Rohli[2,4]

Abstract

The problem of accurately predicting vote counts in elections is considered in this article. Typically, small-sample polls are used to estimate or predict election outcomes. In this study, a machine-learning hybrid approach is proposed. This approach utilizes multiple sets of static data sources, such as voter registration data, and dynamic data sources, such as polls and donor data, to develop individualized voter scores for each member of the population. These voter scores are used to estimate expected vote counts under different turnout scenarios. The proposed technique has been tested with data collected during U.S. Senate and Louisiana gubernatorial elections. The predicted results (expected vote counts, predicted several days before the actual election) were accurate within 1%.

Keywords: political big data; predictive analytics; voter scores; predict election outcomes; behavioral analytics; computational social sciences; machine learning; data science

Introduction

This article introduces a machine learning approach to predicting vote counts and vote shares in political elections. The article describes the development of the Campaign-Specific Voter Score Algorithm (CVS Algorithm) and the generation of individualized voter scores for every individual present in a large voter registration database. The CVS algorithm provides a quantitative representation of an individual voter's preference for a candidate over other candidates in a political campaign.

In most political data sets, such as a voter file, an individual's registered party is recorded. One might think that this is sufficient information to make a prediction on the preferred candidate. In reality, a prediction on a preferred candidate is anything but trivial. For example, in a primary election where people from the same political party compete to earn their party's nomination for the general election, it is impossible to predict an individual's preference for one of the candidates using only the individual's party of registration. In addition, an individual's registered party is often not a true indicator of their perceptions of a candidate. For example, the state of Louisiana has a large proportion of residents who are registered as Democrats, but tend to vote for Republicans.

One may also argue that a repeated number of survey polls are sufficient to depict the accurate mood of the electorate. However, even with statistically representative samples, the sample size is too small to accurately pinpoint individual voter behavior—the overall election outcome predictions can be error prone. In the FiveThirtyEight blog soon after the results of the 2014 midterm elections[1] and the 2016 elections,[2,3] bloggers such as Nate Silver, Carl Bialik, Harry Enten and Dhurmil Mehta describe and document skews in small sample polls yielding wrong expectations of election outcomes.

The inadequacy of polls in forecasting elections is widely recognized. Previous academic research has documented that survey polls only provide broad-brush estimates but that, in most elections, individual voter scores can provide a more microscopic view of the electorate.[4] In another review of the literature, authors Nickerson and Rogers[5] describe a data science arms race within political campaigns. In their telling, there is a widespread usage of traditional statistical techniques such as ordinary least squares and logistic regression, but also an understanding that these methods are both too dependent on the skill of the analyst

[1]*Department of Geography and Anthropology, Louisiana State University, Baton Rouge, Louisiana.*
[2]*NOAA Southern Regional Climate Center, Louisiana State University, Baton Rouge, Louisiana.*
[3]*Booth School of Business, University of Chicago, Chicago, Illinois.*
[4]*College of Engineering, Louisiana State University, Baton Rouge, Louisiana.*

*Address correspondence to: *David Sathiaraj, Department of Geography and Anthropology, E335 Howe Russell, Tower Drive, Louisiana State University, Baton Rouge, LA 70803-2804,* E-mail: davids@srcc.lsu.edu

and may not apply to different regions, issues, or campaigns. This has led to an increased interest in supervised learning but, as of now, limited to no adoption within political campaigns.

There is another problem with just using surveys for predicting election outcomes. Authors Ansolabehere and Hersh[6]describe in detail problems associated with surveys such as misreporting by respondents. This can lead to poor turnout model estimations and erroneous election outcome scenarios. Fulgoni et al.[7] provide creative strategies that combine data analytics and social media to maximize marketing and messaging of candidates and their stances during live election campaigns.

More recent elections, such as the U.S. Presidential Election in 2016, highlight the failure of survey-based polls in predicting the correct outcome.[8] Most polls wrongly predicted that Hillary Clinton would win the U.S. Presidential election—Trump won[9]. Recent work has called polling an "imperfect science" and lists scenarios where random sampling (and small samples) can introduce biases that lead to inaccurate predictions.[8] Possibilities cited for wrong predictions in the 2016 election[8,9] include an underrepresentation of true Trump supporters in polling samples and respondents being reluctant to make a decision for whom they plan to vote.

Since relying solely on polls can lead to inaccurate predictions, the authors propose an algorithm that relies not only on polls but also on other disparate sources of dynamic campaign data that capture a real-time pulse of the electorate. By doing so, the proposed method can collect signals from a wide swath of the voting electorate and thereby generate accurate predictions. This work introduces a technique that combines static pieces of information with dynamic campaign information to generate real-time machine learning models. Static pieces include data such as voter registration data, optional consumer data sets (that describe at a household level behavioral consumer habits such as buying or spending habits, hobbies, and recreational interests), and demographic attributes associated with an individual or household. Dynamic pieces of information include data such as polls, donation data, grassroots surveys, and field-based information. The real-time predictive models generated are then used to predict individualized voter scores for the entire electorate. These scores are used to predict vote counts, which can be adapted to model various voter turnout scenarios. Due to the dynamic input from recent polls and grassroots data, the algorithm is sensitive to minute changes in perceptions of the electorate. This enables the CVS algorithm to be highly accurate in predicting vote counts and percentages.

Background and related work

While political data analytics is not a new field, it has grown substantially over the past 10 years. This growth is primarily a result of two factors. First, there has been a substantial decrease in the cost of computing power and maintaining databases. Most political campaigns have binding time and budget constraints, so the cost of procuring data and developing sophisticated analyses is no longer as prohibitive as in the past. National political parties now maintain detailed databases with behavioral, commercial, and political characteristics of voters. Instead of constructing and extending their own databases, political candidates are able to use existing data resources for targeting and outreach efforts.

The second factor is a recent increase in political consultants with backgrounds in fields such as computer science, data science, and statistics. In the past, most political consultants had backgrounds in law or the humanities. A lack of human capital at the intersection of politics and more quantitatively inclined fields limited the availability of sophisticated data-driven computational and statistical methods. This lack of human capital was compounded by a disinterest among many campaign managers and candidates who did not see the utility in a computational approach. This viewpoint has changed, in large part due to the high profile presidential campaigns by President Barack Obama in 2012 and Senator Ted Cruz in 2016, both of which relied heavily on data analytics to guide their campaigns.[5,10]

Recent work at the intersection of machine learning and social data includes sentiment analysis of candidates' perceptions[11] and mining Facebook pages to predict political affiliation of a Facebook user.[12] Additional work in this area includes creation of a voting recommendation tool that matches voter preferences and positions of political parties and candidates.[13]

One of the primary goals of political data analytics is to develop individualized scores for voters. These scores usually range from 0 to 1 and predict the likelihood of a voter supporting a particular candidate. These scores are typically derived using common statistical methods, such as multivariate regression. Such methods have some drawbacks. They assume that the variables or attributes conform to a simplistic distribution (in most cases, normal distribution) or assume linear dependencies. These problems guided the exploration of the proposed

machine learning approach. The proposed methodology does not make such assumptions and applies machine learning techniques that do not assume an underlying distribution. Machine learning approaches have been used in a live election environment to derive decision or action rules.[14] This research is similar to the cited work, which was tried on three live election environments. One distinction between this work and earlier machine learning research[14] is that this work is a more comprehensive framework that predicts which candidate an individual is likely to support or not support. Although prior work has implemented data mining techniques and scores,[4,5,10] the data-driven aspects were limited to isolated portions of the campaign and not an integral part of a campaign operation. A literature review also revealed work such as Jungherr[15] that compares data-driven campaign efforts in the national German election in 2013 and that describes data-driven efforts in the 2015 election in the United Kingdom.[16] More recently, there has also been an analysis of television viewing habits of consumers to predict election outcomes.[17]

The description of these data-driven campaigns is based on interviews conducted with campaign staff and not a general purpose algorithmic methodology. This work provides a generic, theoretical algorithmic framework that provides a basis for integration into all aspects of a campaign and applicable to a wide spectrum of political campaigns.

Key feature

The key feature in this technique is to combine static and historical voter specific information with dynamic, real-time campaign specific information to generate rich insightful data stores that can be used for predictive modeling. The static historical data points used in the model include voter registration files, voting histories, and consumer data sets. These static data sets are typically available within a political campaign or accessible to a campaign via their state or national party. Work by Ansolabehere and Hersh[6] describes in great detail how private data vendors such as Catalyst LLC have created national voter registration data. These vendors extend these baseline files with elaborate census data information and data on the consumer habits of the voter. This regularly updated voter database is then sold to a national party (in this case, the Democratic Party). On the Republican Party side, there are similar firms that sell similar voter data sets. The national or state party officials then make the data sets available to their party candidate or campaign. Authors Nickerson and Rogers[5] describe the source of such data and why campaigns need such extensive data profiles on voters: for accurate, current voter contact information. Some campaigns supplement the voter registration information and consumer data with their own voter contacts based on past supporters and donors from previous campaigns.

It is important to address privacy concerns with the use of such data sets. Work by Rubinstein[18] describes voter privacy issues in the age of Big Data. The cited work also describes additional dangers as different data sets are aggregated and that individuals may object to the use of their data in forecasting their behavior. The proposed CVS algorithm meets the standards laid out in Rubenstein[18] because it only seeks to identify likely supporters. This work does not change the role of a typical political campaign to persuade or the individual's right to vote or not to vote. In a second relevant article, ethical issues in the Big Data industry are discussed, particularly pertaining to consumer data sets.[19] It describes downstream uses of consumer data (such as this work) and outlines use cases that are harmful, that is, when value is destroyed (not created) for a stakeholder, diminished rights (instead of realization of rights) for a stakeholder, and disrespectful treatment of a consumer. In this work, none of the harmful use cases apply. The sources for these data are discussed in detail below.

This work used a voter data set and a consumer data set that was made available to the campaigns by the national Republican Party (a downstream use). However, in the development of the predictive model, the use of the consumer data set was optional for two reasons. First, consumer information is typically available at a generic household level and not at an individual level. An example of this is a field "hunting_fishing" that described using a boolean value to indicate an interest of someone in the household for hunting or fishing. Another reason for the use of consumer data to be optional is that the data can be sparse with several column attributes having no information. In the proposed methodology, the voter registration data were more dense and were sufficient to meet the need as a static data source and to obtain a highly accurate predictive model along with the dynamic data sources.

However, using only static data sets such as the ones described above, one cannot get real-time predictive insights. This is where dynamic information sources and data sets come into play (Recent work such as Nickerson and Rogers[5] describes a similar need and the dynamic data sources that campaigns use to get real-time

insights). This dynamic, campaign-specific information includes aggregates of survey polls, data collected from phone calls, emails, and past donations, and support tags collected from door-to-door canvassing efforts. Data are internally matched to an individual record in the voter file, and the relevant attributes of individuals are aggregated. Attribute counts can exceed 300 columns of data with demographic markers, household information, and past voting histories included. Similarly, individual profiles from the dynamic pieces of information (from multiple surveys, door-to-door collected tags, email and phone contacts, and donors) are aggregated row-wise to develop a sample size that is 10–15 times larger than a typical survey sample size of 800. These aggregated profiles are used as training sets in the development of an automated, machine learning predictive model.

The predictive models are dynamic and feedback driven. A political campaign has the ability to collect new pieces of information such as a new survey poll or a daily log of grassroots information collected by canvassing volunteers. These new information inputs are processed by revising the training set and rerunning the predictive model. The predictive models are dynamically updated with the new information and a new set of revised scores can be generated.

The predicted scores are valuable pieces of information. The individualized scores facilitate the creation of targeted campaigns. The scores are useful in predicting how independent registered voters are likely to vote in an electoral campaign. The scores also help pinpoint accurate ground conditions in an electoral field. This information enables campaigns to identify strengths and weaknesses and enables campaigns to recalibrate their campaign messaging and targeting. In the next section, a theoretical algorithmic framework that predicts individualized scores is developed to provide quantitative insights on whether an individual is likely to support a candidate or not.

The uniqueness of this method is the dynamic generation of model scores based on any new campaign specific data inputs such as polling data or tags collected from canvassing efforts. The model can be run with each new set of inputs, and scores can be generated within a few minutes.

Problem Description

Formal notation

Consider a large voter file database V_f that comprises m rows (voter file records) and n columns (attributes) (size of m varies from 3 to 190 million rows and number of columns ranges from 250 to 300). This setup typically falls under the Big Data domain. Optionally, also consider an associated consumer database C (containing behavioral attributes), each row of which is linked to the voter database V_f with a unique household id attribute *hid*. Databases V_f and C will also be referred as "static" pieces of information in this work, as they do not change often over time.

Also consider dynamic pieces of information (sources of data that are more prone to change) such as polls (referred to as P), data collected by political grassroots work (referred to as G), and donation data (referred to as D).

The following notation will be used in the subsequent descriptions of the algorithm:

- p: number of polls conducted.
- P: set of polling data sets, $P = \{P_1, P_2, \ldots, P_p\}$, so P_i corresponds to poll i and $i \in [1, p]$
- g: the number of grassroots collected data sets.
- G: set of grassroots data sets collected, $G = \{G_1, G_2, \ldots, G_g\}$, so G_i corresponds to poll i and $i \in [1, g]$
- d: the number of donation data sets collected.
- D: set of donor data sets collected, $D = \{D_1, D_2, \ldots, D_d\}$, so D_i corresponds to poll i and $i \in [1, d]$
- DC: a set representing dynamic pieces of information with $DC = P \cup G \cup D$
- T_r: Derived training set for CVS algorithm
- T_e: Derived test set
- c_1: Label representing class 1 in T_r if support is for candidate c_1
- c_2: Label representing class 2 in T_r if support is for candidate c_2
- ScoreSet: Set of individual voter scores. It is a vector of size $1 \times n$
- classLabels: A vector, corresponding to labels assigned to T_r
- t: boundary threshold that separates the derived scores into two classes
- V_f: voter file, in the form of a matrix $m \times n$, m rows and n attributes.
- $V_f[id]$: vector of the shape $1 \times n$, referencing the id column in V_f
- $V_f[phonenum]$: vector of the shape $1 \times n$, referencing the phone number column, *phonenum* in V_f
- $V_f[race]$: vector of the shape $1 \times n$, referencing the gender column, race in V_f

- $V_f[age]$: vector of the shape $1 \times n$, referencing the age column, age in V_f
- $V_f[gender]$: vector of the shape $1 \times n$, referencing the gender column, gender in V_f
- M_0: predictive model that classifies individuals into c_1 and c_2
- *tp%*: projected turnout percentage, used for estimating election vote counts

Matching records (record linkage)

The first step in this process is using record linkages to generate a data set T_r that can be used to train a machine learning model. Record linkage is essentially the linking of information from two or more data sources (databases) to define a single entity (in this case, a single individual).[20] This linking operation in big data sets can be computationally complex due to the absence of a unique identifier information (such as a national identification number or a social security number in the United States). Record linkage finds applications in diverse fields such as healthcare,[21] business,[20] and government agencies.[20] Record linkage also helps in the removal of duplicate entries. It can help save money in political campaigns such as removing of duplicate mailing addresses during a targeted mail operation. There exists an extensive survey on the use of machine learning, rule-based and similarity metric-based techniques to identify duplicate records.[22] In political science, some work have applied record linkage techniques to link campaign contribution databases and identify contributions made by the same donor.[23] Work such as Bilenko et al.[24] describes linking of information using names. A recent technical report describes an algorithm for linking records based on age, date of birth, gender, and name.[25] In this work, linking of information across multiple data sources to a single individual is important to derive unique individual profiles for training our machine learning model. A schema for generating the training set is depicted in Figure 1. The key to building this training set is to develop individual records using dynamic information (from say a poll) and link the information with the voter registration database, V_f, and the consumer set, C.

Consider a dynamic data set such as a polling survey, P_i. Raw polling data typically consist of a number of individual responses to survey questions and meta information such as the tuple containing age, gender, phone number, race, and, in rare cases, a unique voter id. After several small pilot studies, it was found that the attribute columns of phonenum, race, age, and gender were sufficient to provide a high match rate with the large voter file V_f. Using a rule-based record linkage method, a scan is run for the same combination of phone number, race, age, and gender in the more comprehensive and complete voter file, V_f. Once a match and a unique voter id are found, columns from the voter file V_f and the consumer file C are combined, and the resulting attributes are appended to the training set T_r.

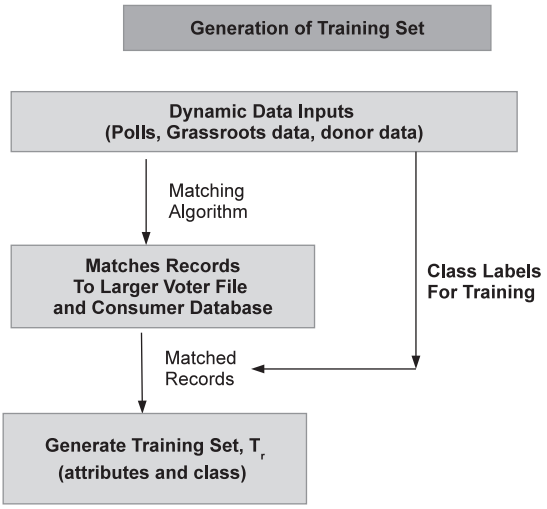

FIG. 1. Pipeline for generating the training data.

The next piece of information needed is the class attribute of the matched individual. Surveys typically include questions about a campaign and an individual's preference for a candidate. The key survey question is often referred to as a "ballot" question; survey respondents are asked which candidate they are most likely to support. Responses to this question form the class attribute in the training set T_r. For the sake of simplicity and for developing the classification framework as a two-class problem, the "ballot" survey question with multiple candidates (>2) can be reduced to considering either the top two candidates or the two most dominant parties (such as Republican and Democratic in the U.S. political system). Depending on the response to the "ballot" question, an appropriate class label is assigned to the matched record (so in a two-candidate scenario, if the survey response is in favor of candidate c_1, then a class label of c_1 is assigned to the matched record, and if the survey response is in favor of candidate c_2, then a class label of c_2 is assigned). This process of class labeling is repeated for all the matched records until all matched records are exhausted. The procedure can also be applied

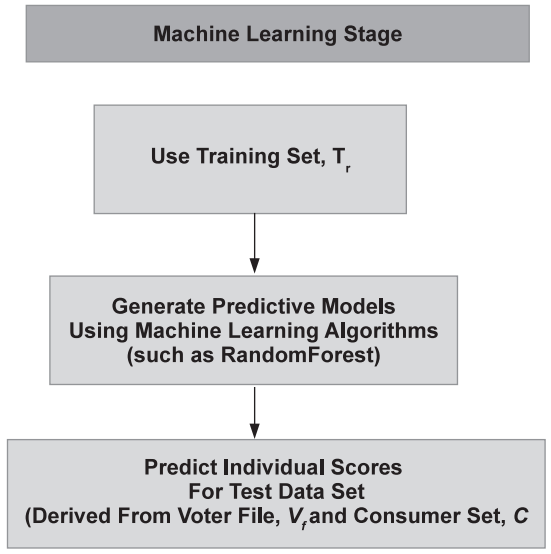

FIG. 2. Pipeline for generating individualized voter scores.

to additional polling data, grassroots collected data, and other miscellaneous sources such as donation data.

CVS algorithm and voter scores

The next step in the process is to use the derived training set to develop a classification model. Classification techniques such as Random Forest[26] and Support Vector Machines[27] were used to develop a predictive model M_0. A 10-fold cross validation is applied to ensure that there are no biases in the fitted model. The test data set T_e is derived from the entire voter file V_f. The test data set is initially unlabeled. The attributes for the test data set T_e will be derived similar to the training set T_r by combining the attribute columns associated with V_f and consumer file C. Using the predictive model M_0, class labels and associated probabilities are predicted for the test data set T_e. The schema of generating individualized voter scores is outlined as Figure 2. The algorithm is outlined in Figure 3.

The voter scores are outputted as ScoreSet and lie in the range of [0, 1]. Supporters of candidate c_1 will have voter scores that lie in the range of $[0, t]$ and supporters of candidate c_2 will lie in the range of $(t, 1]$.

In an ideal case, the threshold t that forms the boundary between the two classes will be equal to 0.5—the halfway point between 0 and 1. However, since political campaigns vary across regions and demographics, an empirical method is needed to identify the boundary threshold t. In the set of polling data sets P, identify records that marked the ballot question in the poll as "undecided." Those records are then matched with larger voter file V_f. Using the CVS ScoreSet, find the scores of the "undecided" individuals. The median of those scores forms the boundary threshold t.

Feedback driven and real-time models

The CVS algorithm is a feedback-driven methodology where scores are constantly refreshed as new data are

Require: V_f and C (static components) and dynamic component set $DC = \{P,G,D\}$.
1: Initialize array *MatchedIds* = ϕ
2: Initialize array classLabels = ϕ
3: **for** each dataset *dc* in *DC* **do**
4: **for** each row iterator *r* in *dc* **do**
5: Find V_f [id] such that V_f [*phonenum*] = *r*[*phonenum*] and V_f [gender] = *r*[gender] and V_f [race] = *r*[race]
6: **if** V_f [id] is found **then**
7: *MatchedIds* = *MatchedIds* + V_f [id]
8: classLabels = classLabels + *Label* (c_1 or c_2) {based on response question in dataset}
9: **end if**
10: **end for**
11: **end for**
12: Initialize matrix $T_r = \phi$
13: **for** each id in *MatchedIds* **do**
14: $T_r = T_r + (V_f$ [id] + C[id]) {matrices V_f and C are collated column-wise, matrix T_r is appended row-wise}
15: **end for**
16: Append classLabels vector to T_r
17: Train a Classification Algorithm ML using T_r to get fitted model M_0
18: Generate unlabeled, test data T_e, by column-wise collation of matrices V_f and C.
19: Use Model M_0 on test data set T_e to predict probabilistic scores ScoreSet.
20: **return** ScoreSet.
21: Repeat algorithm (steps 1–20) as new time-relevant datasets (such as new polls, grassroots data, updated donor files etc) are added in the set *DC*

FIG. 3. The *CVS* algorithm. CVS, Campaign-Specific Voter Score.

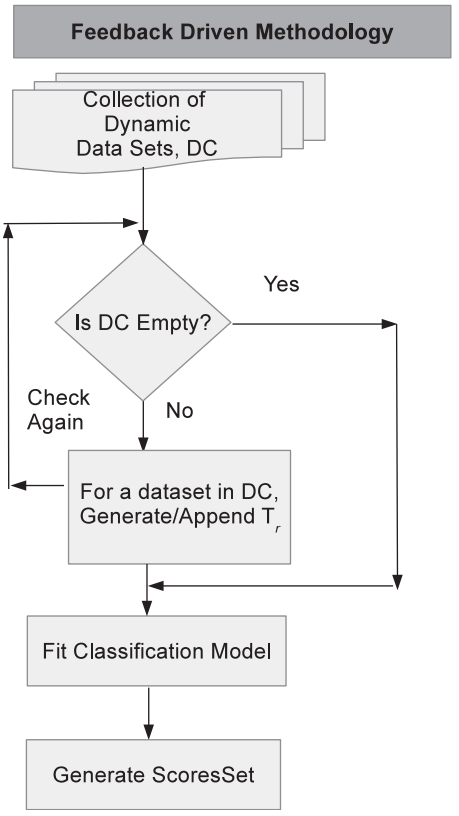

FIG. 4. Feedback-driven model generation.

collected, which is typical during the course of an election campaign. Using the notation outlined, the dynamic data store *DC* will be refreshed or appended frequently. Each data set in the data store *DC* is used for matching or record linking. Once all the data sets in *DC* are exhausted and training set T_r is generated, machine learning models are fitted, and scores are generated. Since the data set store *DC* is frequently updated with newly collected campaign data the generated machine learning models and the subsequent voter scores provide a real-time granular pulse of the electorate. The process of aggregating the training set T_r provides the machine learning model with a large set of training examples. This is a significant advantage over solely using polling data since polls typically have small sample sizes and extrapolating this small-sized information to a larger voter file can lead to larger predictive errors. This feedback-driven model generation is depicted as Figure 4.

Explanation on the real-time models. During the 2014 election campaign for the U.S. senate in Louisiana and the 2015 election campaign for the Louisiana governor's race, this predictive technology was tested in a live election environment. The focus was on Louisiana due to the campaign based in Louisiana and the consequent availability of data catered to that political jurisdiction. Both campaigns had access to two static data sets from the Republic National Party—one was voter registration data ($\sim$2,930,000 rows and 168 columns) and the other a consumer data set ($\sim$1,930,000 rows and 90 columns). To build the dynamic data sets, the 2014 senate campaign would collect grassroots data from a host of volunteers spread across the state of Louisiana. The volunteers would go door-to-door in neighborhoods, knock on each household, and conduct a campaign-specific survey if permitted by the household owner. The data would come back on a daily basis in comma-separated text files. The authors then applied a series of record linkage procedures to match individual voter responses back to voter file to gauge voter perceptions on the candidate. The authors achieved a success match rate of nearly 99% on the data that were collected by the door-to-door volunteers. The greater the response rate of the surveyed households, the richer the training set became for the machine learning model.

At the same time, on a weekly basis, the campaign would receive polling results from a contracted pollster. Although the polling data contained merely phone numbers and the individual responses to questions, using record linkage procedures, the authors could match the survey respondents to the voter and consumer data bases. The success rate for matching a poll response to an individual voter was nearly 94% (this is similar to that reported in this technical report[25]). On a daily basis, the campaign would also receive phone-based voter contact information where voters would respond on whom they plan to vote for and their positions on various issues that matter to them. This information was also collected and matched to the static data sources and appended to the training data set. As new inputs came into the campaign in the form of phone contacts, door-to-door grassroots efforts or polling information on a periodic basis (daily, weekly, or sub-daily when closer to election day), the model is retrained using constantly refreshed training data. While the data sets were limited to Louisiana, the algorithmic framework outlined in this work is a generic procedure that is applicable to any election campaign that has access to data sets described in this study.

Some notable predictive aspects of the models derived are described in this study. A key variable that emerged was a variable called household mix (hmix). The household mix defines a political partisan mix

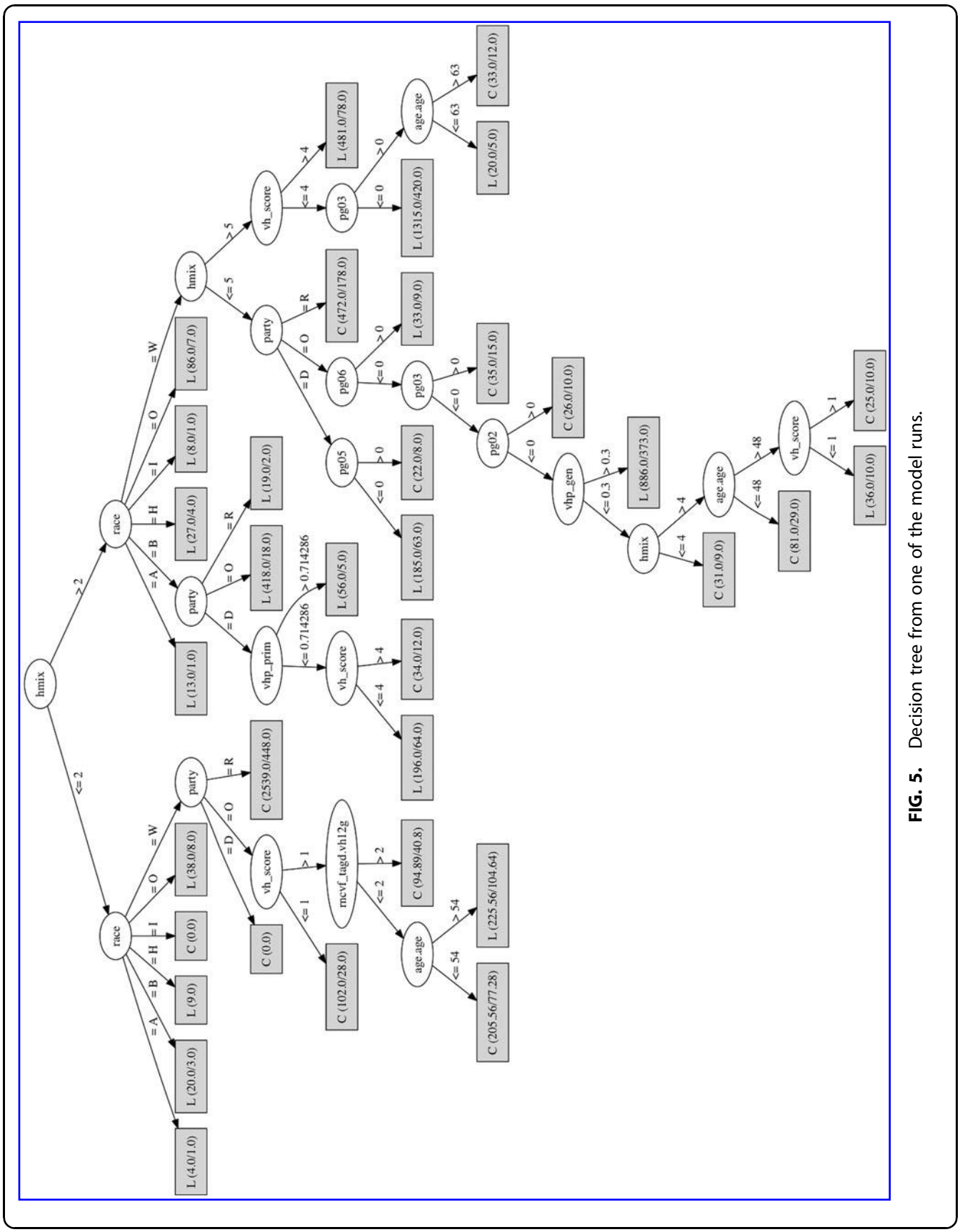

FIG. 5. Decision tree from one of the model runs.

Table 1. Important decision variables

Key variable	*Description*	*Mean Gini index score*
Hmix	Household Mix	838.3
Party	Party of Registration	708.9
Cd	Congressional District	319.7
Race	Race	233.7
vh_score	Voting History Score (or how many times has someone voted)	188.4
vh12g	Voted in 2012 General Election (0 or 1)?	143.4
vh11g	Voted in 2011 General Election (0 or 1)?	37.3
sex	Gender/Sex	130.8
vh08g	Voted in 2008 General Election (0 or 1)?	124.2

within each household. As an example, if a household had four members with three members registered as Republican and one member registered as an independent, the model predicted the independent member as leaning Republican. The idea is somewhat intuitive: all members of a household tend toward the majority view to avoid unpleasant within-household political disputes. This is similar to that evidenced in prior work[28]—political agreement among spouses in two national U.S. elections. The other defining variables were age, race, gender, and voter scores based on participation in previous elections. A sample decision tree model of one such tree constructed during Random Forest construction[26] is depicted in Figure 5. One can see from the tree that the first decision split happens with the hmix variable, followed by race and the party of registration. Each of the nodes in the decision tree (rectangular boxes) indicates the first initial of the candidate names. Using such a decision model, the proposed framework derives predictive scores on each individual voter.

Table 1 provides a listing of important predictor variables in the generation of the model. Important variables are household mix (hmix), the party of registration, the congressional district that a voter belongs to, the race of the individual, and series of indices that indicate voting in recent elections (for example, vh12g is a boolean variable with a value of 1 if they voted in the 2012 U.S. Presidential election and 0 if they did not). Notice that these predictor variables are not generic to one area (or Louisiana—the location used for this study), but can scale to any national or international election campaign. The CVS model predicted the correct voting result on all three live elections and was highly accurate in predicting actual vote shares.

Applications

The CVS algorithm and its scores have two significant applications. One application is the prediction of vote counts under various turnout scenarios, and the other is for campaign resource allocation through voter file segmentation and focused microtargeting.

Predict vote counts

One of the key components in a campaign is the ability to estimate potential vote shares before the actual election. This estimation, when combined with good planning, enables proper resource allocation needed to maximize vote count. Campaigns can identify demographics that have the potential to switch their vote and increase resource flow to improve their share among that particular demographic. Accurate estimation of vote counts before the election also helps campaigns differentiate certain win regions from certain loss regions. An estimation of vote count could be the difference between a narrow loss and a narrow win on Election Day.

Predicting vote counts before an election is nontrivial as voter turnout percentages can be unpredictable. This unpredictability can be due to a variety of factors such as misreported data,[6] weather,[29] or differing voting behavior depending on the type of election.[30] Hence, campaigns have to estimate using multiple turnout scenarios. Using scores derived from the CVS algorithm, one can estimate expected vote counts under different turnout percentages. For a given turnout percentage *tp%*, a random sample of size $tp\% \times |V_f|$ is drawn from the voter file V_f. The associated CVS derived scores for this sample are tabulated. For each record, if the associated CVS score is less than or equal to the boundary threshold t, then that record is tallied as a vote for candidate c_1. If the score is greater than the boundary threshold t, then it is counted as a vote for candidate c_2. This rule is repeated for all the records in the random sample, and the tally counts for c_1 and c_2 are tabulated to get vote counts for each candidate.

Voter file segmentation

One other application of CVS scores is the ability to segment and filter a large voter file into smaller targeted segments for effective microtargeting during a political campaign. The usefulness of predictive CVS scores in political campaigns is also highlighted by prior work such as Nickerson and Rogers.[5] A subset of voters that match a given demographic profile can be identified and CVS scores can be used to identify the most likely supporters within that subset. This makes resource allocation more efficient.

As an example, Louisiana has a large proportion of white voters who are registered as Democrats (about

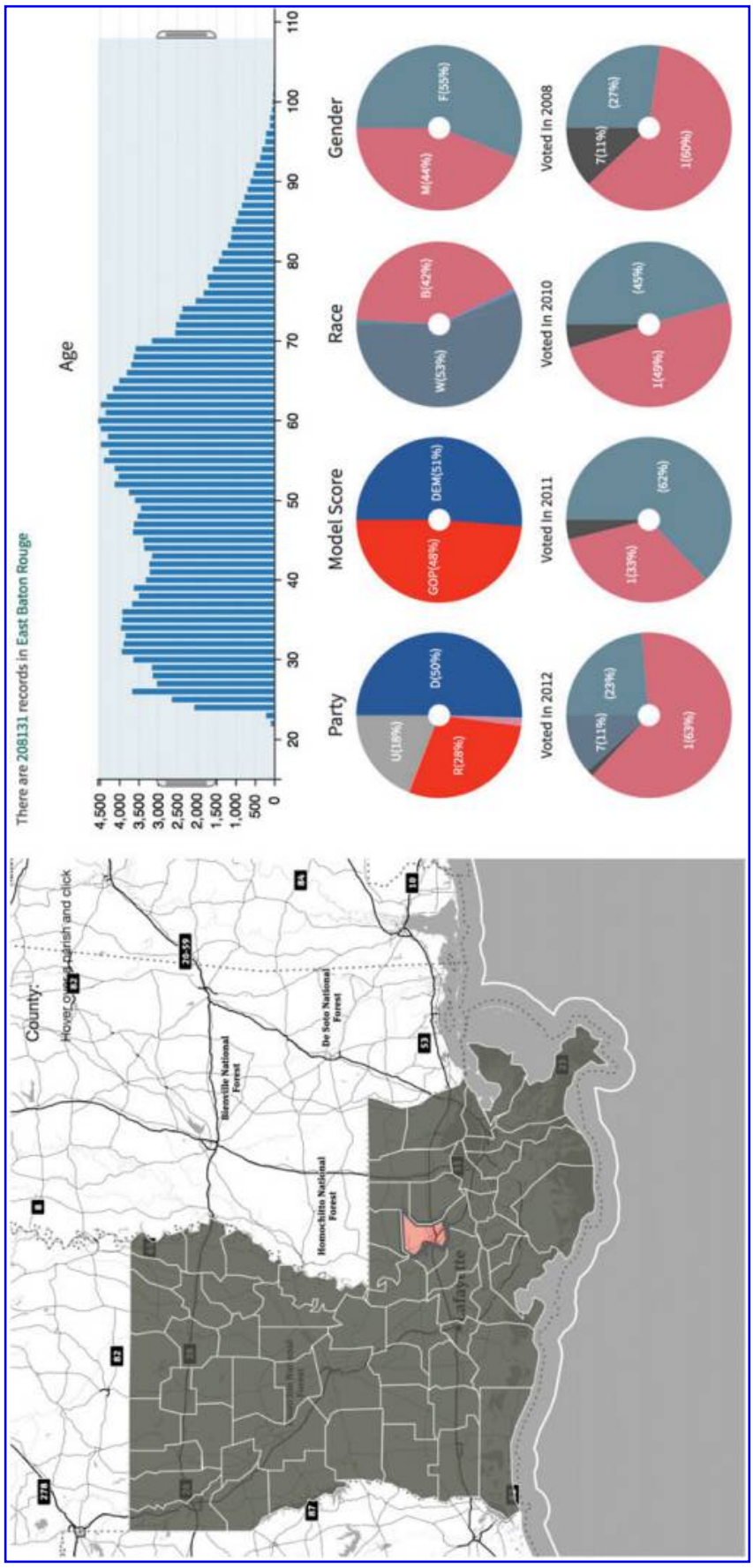

FIG. 6. Screenshot of the Big Data voter file segmentation using predictive scores.

Table 2. Early vote predictions using CVS algorithm during the 2014 U.S. Senator election in Louisiana

Election	*Model predicted, %*	*Actual result, %*
Primary—Republican Party vote share	52.3	52.8
Runoff—Republican Party vote share	57.3	58.9

CVS, Campaign-Specific Voter Score.

30.3% based on the Louisiana Secretary of State data[31]), but historically tend to vote for a Republican candidate. One example of this is older voters whose voting habits have changed, but who have not changed their registered party affiliation. This has been noted in both previous academic work[5] and in the press.[32]

This is a valuable test case for the CVS algorithm as the party of registration does not help in predicting actual voting behavior. The CVS model scores from this methodology can predict this segment with a high degree of accuracy. The CVS model scores correctly predict this group of Republican-leaning registered Democrats representing 14.7% of the total white voter registrations ($\sim$400,000 voters in a 3,000,000 voter file). This shows the value of how these scores provide pinpointed targeted information that can help guide campaigns in reaching their likely supporters.

The authors also developed a visualization framework that utilizes the voter file information and the predictive scores to effectively segment voters based on various covariates such as age, race, gender, and voting behavior. The visualization framework is available at the following URL, note that voter contact information anonymized: http://datadecisions.lsu.edu/vfdemo This visualization serves as an application of the predictive scores to dissect big data voter files and create custom audiences for campaign outreach events such as direct mail, tele-town hall meetings, and phone banking. Creating custom audiences using these predictive scores can result in large savings for small, medium, and large-sized political campaigns. The savings come by minimizing wastage and streamlining outreach audiences.[5] During the 2014 Louisiana Senate campaign nearly $100,000 could have been saved over just 10 direct mail outreach efforts using predictive scores to develop in-house voter contact lists instead of renting lists from external vendors. The predictive scores are grouped under the pie chart labeled "Model Score" in the visualization. A screenshot of the tool is provided as Figure 6.

Table 3. Model predictions versus actual results (2014 U.S. Senator runoff election)

Election	*Republican, %*	*Democratic, %*	*Difference from actual result, %*
FiveThirtyEight analysis (December 5, 2014)	57.8	42.2	1.9
CVS model predicted (December 5, 2014)	56.1	43.9	0.2
Actual result	55.9	44.1	—

Table 4. Model predictions versus actual results (2015 Louisiana Governor's runoff election)

Election	*Republican, %*	*Democratic, %*	*Difference from actual result, %*
CVS model predicted (November 6, 2015)	44.8	55.2	0.9
Actual result	43.9	56.1	—

Results

The CVS algorithm was tested during three live election campaigns. The first two elections were the 2014 primary and runoff elections for the U.S. Senate seat from Louisiana. The other election was the 2015 Louisiana Governor's election. For each of the elections, the CVS algorithm generated scores for $\sim$3 million individuals in the Louisiana voter file. The scores were applied to predict the early vote outcomes soon after the early voting period ended. Typically, the early voting period lasts about 6 days, ending about a week before the actual election day. In the 2014 U.S. Senate Primary, the early voting period lasted from October 21 to October 28 (October 26 was not included as this date was a Sunday). Election Day for the 2014 Primary was November 4.

Table 2 compares the CVS model prediction of early vote results with the actual results for both the primary and runoff elections for Louisiana's U.S. Senate elections of 2014. In the primary election, the model predicted a Republican vote share of 52.3% (and a Democratic vote share of 47.7%). The actual early voting result for the Republican vote share, declared a week after early voting had been completed and after the CVS model had made its prediction, was 52.8%. The predicted result was 0.5% off the actual result. In the runoff election, the model predicted result was off by a mere 1.6%.

Table 3 compares the performance of the CVS Model with the actual 2014 Louisiana's U.S. Senate Runoff Election result and to an analysis conducted by the nationally renowned prediction blog http://fivethirtyeight.com By this metric, the CVS Model prediction of the result (predicted a day before the actual election) was a

mere 0.2% from the actual result. In contrast, the FiveThirtyEight blog analysis was off by 1.9%.

The CVS model was also applied to the runoff election of the 2015 Louisiana Governor's race. Table 4 provides model derived vote share estimates (model run 3 weeks before the actual election). The table shows that the model was off from the actual result by a mere 0.9%.

In summary, Tables 2–4 provide strong evidence of the accuracy of the CVS model derived predictions under live election scenarios.

Conclusion

A hybrid machine learning approach was developed to predict election outcomes more accurately. The methodology is based on individualized scores that give an indication of an individual's preference for a particular candidate or election issue. The proposed method accurately predicted vote counts for three Louisiana elections. The accuracy of this predictive methodology was further validated by the closeness of predicted vote counts to the actual election results.

Author Disclosure Statement

No competing financial interests exist.

References

1. Silver N. The polls were skewed toward democrats. November 5, 2014. Available online at http://goo.gl/kgNN5Z (last accessed November 2, 2017).
2. Bialik C, Enten H. The polls missed Trump. We asked pollsters why. November 9, 2016. Available online at https://fivethirtyeight.com/features/the-polls-missed-trump-we-asked-pollsters-why (last accessed November 2, 2017).
3. Mehta D. How much the polls missed by in every state. December 2, 2016. Available online at https://fivethirtyeight.com/features/how-much-the-polls-missed-by-in-every-state (last accessed November 2, 2017).
4. Issenberg S. How President Obama's campaign used big data to rally individual voters. Technol Rev. 2013;116:38–49.
5. Nickerson DW, Rogers T. Political campaigns and big data. J Econ Perspect. 2014;28:51–73.
6. Ansolabehere S, Hersh E. Validation: What big data reveal about survey misreporting and the real electorate. Polit Anal. 2012;20:437–459.
7. Fulgoni GM, Lipsman A, Davidsen C. The power of political advertising: Lessons for practitioners how data analytics, social media, and creative strategies shape US presidential election campaigns. J Adv Res. 2016; 56:239–244.
8. Valentino NA, King JL, Hill WW. Polling and prediction in the 2016 presidential election. Computer. 2017;50:110–115.
9. Why 2016 election polls missed their mark. Pew Research Center. Available online at http://pewrsr.ch/2fmNGH6 (last accessed on October 19, 2017).
10. Hamburger T. Cruz campaign credits psychological data and analytics for its rising success. December 13, 2015. Available online at http://wpo.st/Qzl41 (last accessed November 2, 2017).
11. Bhattacharya S, Yang C, Srinivasan P, Boynton P. Perceptions of presidential candidates' personalities in Twitter. J Assoc Inf Sci Technol. 2016;67:249–267.
12. David E, Zhitomirsky-Geffet M, Koppel M, Uzan H. Utilizing Facebook pages of the political parties to automatically predict the political orientation of Facebook users. Online Inf Rev. 2016;40(5, SI):610–623.
13. Tsapatsoulis N, Agathokleous M, Djouvas C, Mendez F. On the design of social voting recommendation applications. Int J Artif Intell Tools. 2015; 24:1550009.
14. Holubiec J, Szkatula G, Wagner D. A knowledge-based model of parliamentary election. Inf Sci. 2012;202:24–40.
15. Jungherr A. Four functions of digital tools in election campaigns: The German Case. Int J Press-Politics. 2016;21(3, SI):358–377.
16. Anstead N. Data-driven campaigning in the 2015 United Kingdom general election. Int J Press-Politics. 2017;22:294–313.
17. Barfar A, Padmanabhan B. Predicting presidential election outcomes from what people watch. Big Data. 2017;5:32–41.
18. Rubinstein IS. Voter privacy in the age of big data. Wisconsin Law Rev. 2014(5):861–936.
19. Martin KE. Ethical issues in the big data industry. Mis Q Exec. 2015;14:67–85.
20. Herzog TH, Scheuren F, William E. Winkler. Record linkage. Wiley Interdiscip Rev Comput Stat. 2010;2:535–543.
21. Siegert Y, Jiang X, Krieg V, Bartholomaeus S. Classification-based record linkage with pseudonymized data for epidemiological cancer registries. IEEE Trans Multimed. 2016;18:1929–1941.
22. Elmagarmid AK, Ipeirotis PG, Verykios VS. Duplicate record detection: A survey. IEEE Trans Knowl Data Eng. 2007;19:1–16.
23. Giraud-Carrier C, Goodliffe J, Jones BM, Cueva S. Effective record linkage for mining campaign contribution data. Knowl Inf Syst. 2015;45:389–416.
24. Bilenko M, Mooney R, Cohen W, et al. Adaptive name matching in information integration. IEEE Intell Syst. 2003;18:16–23.
25. Ansolabehere S, Hersh ED. Adgn: An algorithm for record linkage using address, date of birth, gender and name. Statistics and Public Policy 2017 (in press); DOI: 10.1080/2330443X.2017.
26. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
27. Hearst MA, Dumais ST, Osman E, et al. Support vector machines. IEEE Intell Syst Appl. 1998;13:18–28.
28. Smith AE. The diverse impacts of politically diverse networks: Party systems, political disagreement, and the timing of vote decisions. Int J Public Opin Res. 2015;27(4, SI):481–496.
29. Lee J, Hwang W. Weather, voter turnout and partisan effects in Korea, 1995–1999. Asian J Soc Sci. 2017;45:507–528.
30. Gerber AS, Huber GA, Biggers DR, Hendry DJ. Why don't people vote in US primary elections? Assessing theoretical explanations for reduced participation. Elect Stud. 2017;45:119–129.
31. La secretary of state—registration statistics—statewide. Available online at https://goo.gl/ZGLkRF (last accessed October 31, 2017).
32. Louisiana votes red even as democrats, republicans lose sway. nola.com. Available online at https://goo.gl/fRwRrH (last accessed October 31, 2017).

Cite this article as: Sathiaraj D, Cassidy WM Jr., Rohli E (2017) Improving predictive accuracy in elections. *Big Data* 5:4, 325–336, DOI: 10.1089/big.2017.0047.

Abbreviation Used

CVS = Campaign-Specific Voter Score