조사연구(2023. 5) 24권 2호: 1−15
DOI http://dx.doi.org/10.20997/SR.24.2.1

연구논문

중도탈락 후보 지지층과 응답유보층 보정을 통한 선거결과 예측*

김정경** · 박민규***

본 연구에서는 NBS 선거여론조사 결과를 활용하여 2022년 3월에 실시한 20대 대통령선거의 당선자를 예측하는 통계적 방안을 살펴보았다. 20대 대통령선거의 경우 투표일 1주일 전까지 지지 gnqh자를 결정하지 못한 응답유보자가 약 20%에 달하였으며, 선거일 직전에 안철수 후보의 사퇴가 있었던 점을 고려하여 응답유보자와 안철수 후보 지지자를 지도학습 방법을 통해 분류한 후 각 후보의 득표율을 예측하였다. 지도학습 도구로는 로지스틱 회귀분석, 의사결정나무 그리고 랜덤포레스트 방안을 고려하였다. 득표율의 예측에 있어서는 랜덤포레스트 방안이 가장 좋은 성능을 보였으며, 모든 방안이 1, 2위 후보자 득표율의 비율을 유사하게 예측하였다. 선거일에 가까울수록 예측의 정확도는 대체로 높아졌으나, 응답률이 상대적으로 낮았던 시점의 예측오차는 매우 크게 나타났다. 여론조사 공표보도 금지기간이 해제되어 투표일 직전 자료를 활용한 응답유보자 분류가 가능하다면 선거여론조사를 활용한 득표율 예측은 더욱 정교해질 것으로 기대된다.

주제어: 대통령선거, 지도학습, 의사결정나무, 랜덤포레스트

* 본 연구는 김정경의 석사학위논문의 내용을 바탕으로 이루어졌으며, 고려대학교 연구비에 의하여 수행되었음(K2209861).
** 고려대학교 통계학과 석사 과정 졸업(jk8370@korea.ac.kr).
*** 고려대학교 통계학과 교수(mpark2@korea.ac.kr), 교신저자.

Ⅰ. 서론

대통령선거를 포함한 전국단위 선거가 있는 해에는 투표일 이전 일정 기간 많은 수의 선거여론조사가 수행된다. 선거여론조사를 위한 조사도구로 흔히 사용되는 전화조사는 오랜 기간을 통해 고도화되어 왔다. KT에서 제공하는 전화번호부에 의존한 최초의 전화조사가 포함오차에 취약한 점을 고려하여 유선전화번호 기반 임의전화걸기(Random Digit Dialing: RDD) 기법이 많은 연구자에 의하여 제안되었다. 이러한 전화번호부 기반 조사의 문제점은 김선웅(2004), 강현철 외(2008) 허명회·김영원(2008)의 연구에서 다루어졌으며, 이를 보완하기 위한 RDD 기법의 원리와 응용 분야에 관한 연구로는 Waksberg(1978), Lepkowski(1988) 그리고 Casady & Lepkowski(1988)가 있다. 다양한 통신기술, 특별히 무선전화 사용자가 급격하게 증가하고 유선전화 보유율 및 재택율은 상대적으로 낮아짐에 따라 포함오차를 최소화하기 위하여 고안된 유선전화 기반 RDD기법은 그 한계를 드러내기 시작했다. 이를 보완하려는 방안으로 유선전화와 휴대전화를 함께 사용하는 혼합 RDD 기법이 2011년 서울시장 보궐선거부터 국내에서는 꾸준히 사용되고 있다.

과거 국내에서의 선거여론조사를 위한 RDD 기법은 할당추출법과 함께 적용되었다. 즉, 사전에 할당을 위한 셀을 정의하고, 각 셀에 배정된 표본 수의 조건을 만족하는 표본이 추출되도록 제한적으로 RDD 기법을 적용하였으며, 이는 엄격한 의미에서 확률추출법(probability sampling)으로 분류할 수 없는 표본설계 방안이다. 따라서 할당추출법이 사용된 RDD 기반 조사결과의 분석 시 사용되는 추정량의 비편향성과 분산과 같은 통계적 성질은 전통적인 디자인 기반 추론(design-based inference)의 측면에서는 계산할 수 없다. 대신 관심 변수에 대한 적절한 모형을 가정하고 주어진 가정하에서 추정량 및 이의 성질을 유도하는 모형 기반 추론(model-based inference)이 사용되었다. 실제 주로 사용되었던 모형은 각 할당 셀(층) 내에서 개체들의 값들은 같은 평균과 분산을 갖는다는 셀 평균 모형(cell mean model)이다.

모집단의 특성을 나타내는 모수의 추정이 가장 중요한 목적인 조사 통계학 분야에서 디자인 기반 추론이 선호되는 이유는 모형 기반 추론의 경우 가정한 모형이 적절하지 않을 때 추정량의 통계적 성질이 왜곡되기 때문이다. 특별히 기댓값에 대한 모형의 가정에 오류가 있는 경우에 추정량의 편향이 발생하여 그 결과를 신뢰할

수 없다. 할당추출법이 갖는 이러한 문제로 인하여 선거여론조사를 위해서 휴대전화 가상번호를 활용한 층화추출법이 2017년 이후 사용되기 시작하였다.

유권자를 대상으로 하는 선거여론조사는 선거결과 예측을 위해 투표자를 대상으로 이루어지는 출구조사와 그 목적이 다름에도 불구하고 실제 선거결과를 주기 때문에 조사 기반 선거결과 예측값과 실제 결과와의 직접적인 비교를 통해 조사의 품질이 평가되고 있다. 선거결과와의 직접적인 비교를 통한 조사의 품질을 나타내기 위하여 사용되는 도구로는 다음의 Martin et al.(2005)이 제시한 척도 A가 있다.

$$A = \log\left(\frac{d/t}{D/T}\right) \tag{1}$$

여기서 d와 t는 조사에 근거한 두 후보의 예측 득표율을 나타내며 D와 T는 투표결과 실제 두 후보가 획득한 득표율을 나타낸다. Arzhemier & Evans(2014)는 척도 A의 일반화된 형식의 척도를 제안하였다. 척도 A는 두 후보 간의 득표율 차이에 대한 예측값의 정확성을 평가하기 위한 지표로서 두 후보의 득표율이 같은 방향과 같은 규모로 잘못 예측되었을 때도 0에 가까운 값을 갖는다. 즉, 조사의 목적이 1위 후보를 예측하는 것에 있다면 척도 A는 예측 정확성을 나타내는 좋은 척도가 될 수 있다.

언급한 바와 같이 선거여론조사는 출구조사와는 달리 선거일 이전에 유권자를 대상으로 이루어지기 때문에 선거결과 예측을 위해 사용되는 것은 바람직하지 않다. 그 이유로는 먼저 조사 모집단인 실제 투표자의 집단과 목표 모집단인 유권자 집단의 차이로 인해 추출틀 오차가 발생하기 때문이다. 두 번째로는 투표 이전에 이루어진 조사로 지지 후보를 결정하지 못한 응답 유보층 및 여론조사 공표금지 기간인 투표일 1주일 이내에 사퇴한 후보자를 지지한 유권자의 응답으로 인하여 발생하는 오차 때문이다. 추출틀 오차를 줄이는 방안으로는 투표하겠다고 응답한 유권자만을 대상으로 성별과 연령 그룹과 같은 주요 인구학적 변수를 이용하여 수정하는 방안이 흔히 사용되며, 부분적으로나마 이러한 보정을 통해 추출틀 오차로 발생하는 편향이 감소하는 것으로 알려져 있다. 응답 유보층의 투표 행위 예측을 위해서는 로지스틱 회귀모형을 포함한 다양한 분류모형이 적용될 수 있으며, 이를 이용한 결과가 단순 집계 기반 예측보다 그 오차를 줄일 수 있음을 김정훈 외(2017)과 곽은선·김영원(2022)은 보였다.

투표일 전 1주일은 여론조사 공표금지 기간으로 실제 여론조사가 이루어지지 않

으며 20대 대통령선거에서와 같이 이 기간에 사퇴한 후보가 있는 경우에는 이의 영향이 반영되지 못한 예측 결과가 제공된다. 이러한 문제를 해결하기 위하여 본 연구에서는 선거여론조사에서 응답유보 유권자와 사퇴 후보 지지자를 다양한 분류기법을 통해 분류하여 각 후보의 최종 득표율을 산출하는 통계적 방안을 고려하였다. 분류기법으로는 로지스틱 회귀모형과 의사결정나무 그리고 랜덤포레스트 기법을 적용하였으며 서로 다른 방법론의 비교를 위해서는 예측 득표율과 실제 득표율의 단순 차이 및 척도 A을 사용하였다.

Ⅱ. 분류모형

지도학습(supervised learning)을 통해 모형을 적합하고 이를 이용하여 새로운 개체를 특정한 그룹으로 분류하는 분류(classification)모형 기법은 전통적으로 통계학에서는 판별분석 기법으로 소개되었으며, 최근에는 기계학습(machine learning)이라는 이름으로 다양한 방안들이 소개되고 있다. 선거여론조사에서 지지하는 후보가 없다고 응답하거나 투표일 전 1주일 이내에 사퇴하는 후보를 지지한 응답자를 특정 후보자의 지지자로 분류하는 방안으로 판별분석 혹은 기계학습 방안을 적용할 수 있다. 본 연구에서는 분류모형으로 다항 로지스틱 회귀모형, 의사결정나무 그리고 랜덤포레스트 방안을 고려하였다.

다항 로지스틱 회귀모형에서 반응변수 Y는 K개의 가능한 범주를 갖는 확률변수로 i번째 개체가 범주 k에 속할 확률은 다음과 같이 정의할 수 있다.

$$\log\left[\frac{P(Y_i=k)}{P(Y_i=K)}\right]=\boldsymbol{x}_i'\boldsymbol{\beta}_k,\quad k=1,2,\cdots,K-1 \tag{2}$$

여기서 K는 마지막 범주를 나타내며 $\boldsymbol{x}_i=(1,x_{i1},\cdots,x_{ip})'$는 설명변수의 열벡터를, $\boldsymbol{\beta}_k=(\beta_{k0},\beta_{k1},\ \cdots\ \beta_{kp})'$는 범주별로 정의되는 회귀 계수를 나타낸다. 식 (2)에서 정의한 회귀 계수 벡터는 $(\boldsymbol{x}\,,\,Y)$가 측정된 개체들의 집합을 통해 추정되며, 설명변수 벡터 $\boldsymbol{x}_0$을 갖는 새로운 개체가 범주 k에 속할 확률은 식 (3)에 의하여 산출되며, 그 확률이 가장 큰 범주에 새로운 개체는 분류된다.

$$P(Y_0 = k) = \frac{\exp(\boldsymbol{x}'\widehat{\boldsymbol{\beta}}_k)}{1 + \sum_{i=1}^{K-1} \exp(\boldsymbol{x}'\widehat{\boldsymbol{\beta}}_k)} \tag{3}$$

의사결정나무는 분류 규칙을 나무 구조로 나타내어 전체 자료를 몇 개의 소집단으로 분류(classification)하는 분석 방법이다. 나무 구조에 의해서 모형이 표현되기 때문에 직관적으로 이해하기 쉽다는 장점이 있다. 의사결정나무는 반응변수의 종류와 분류 방법에 따라 다양한 알고리즘이 존재하는데, 그중에서 본 연구에 사용한 CART 알고리즘은 다음과 같다.

① 각 설명변수에 대해 모든 가능한 분할 지점을 조사한다.
② ①에서 구한 분할 지점 중 불순도의 향상도가 가장 큰 지점을 분류 집합으로 선택한다.
③ 각 관측치에 대해 ②에서 선택한 분류 집합에 속하면 왼쪽 자식노드로, 아니면 오른쪽 자식노드로 보낸다.
④ 위 과정을 모든 하위 노드에서 반복 시행하여 나무 구조를 구성한다.

불순도의 향상도(goodness of split)는 부모노드에 비해 자식노드들의 불순도(impurity)가 얼마나 감소하였는지를 나타내는 척도이다. 노드 내에 관측치들이 같은 반응변수 값을 가질수록 불순도는 낮아지며, 불순도가 많이 개선되는 방향으로 분할 규칙을 선택한다. 상위 노드 R에서 두 개의 영역 R_l과 R_r로 나누어지며, 이때 불순도 $I(R)$, $I(R_l)$, 그리고 $I(R_r)$을 각각 계산할 수 있다. 이를 이용한 불순도의 향상도 공식은 식 (4)와 같다.

$$IG(R,s) = I(R) - p_L \times I(R_l) - p_R \times I(R_r) \tag{4}$$

$I(R)$는 노드R의 불순도 함수이며, s는 분할변수의 분할 지점을 의미한다. p_L과 p_R은 각각 부모노드 대비 왼쪽 자식노드와 오른쪽 자식노드의 비율을 의미한다. 불순도 함수는 지니 지수(Gini index)를 사용하며, 이는 식 (5)와 같다.

$$I(t) = Gini(t) = 1 - \sum_{i=1}^{K} \pi_{ti}^2 \quad (5)$$

π_{ti}는 노드 t에서 범주 i에 속하는 관측치의 비율을 의미하고, K는 반응변수 범주의 수이다. 지니 지수는 관측치들이 동일한 반응변수 값을 가질수록 값이 작아진다.

랜덤포레스트는 Breiman(2001)에 의해 제안된 분석 방법으로, 학습 과정에서 구성한 여러 개의 의사결정나무를 이용하여 분류를 수행하는 앙상블(Ensemble) 기법이다. 앙상블 기법이란 주어진 데이터로부터 여러 개의 예측 모형을 생성한 후 이 모형들을 조합하여 최적의 예측 모형을 생성하는 기법을 말한다.

랜덤포레스트는 의사결정나무가 갖는 과대 적합 문제를 최소화하여 모델의 정확도를 향상한다는 장점을 가지고 있으며, 본 연구에서 사용한 알고리즘은 다음과 같다.

① $b = 1, \cdots, B$ 에 대하여

가) 학습 데이터에서 크기가 N 인 부트스트랩 표본$Z^* = (\boldsymbol{x}^*, Y^*)$을 추출한다. 여기서 N은 학습 데이터의 크기를 나타낸다.

나) 부트스트랩 표본 Z^*에서 의사결정나무의 각 종단 노드(terminal node)가 최소 노드 크기인 n_{min}에 이를 때까지 다음 과정을 반복하며 의사결정나무 C_b을 생성한다.

- p개의 설명변수에서 m개를 무작위로 선택한다($m < p$).
- 선택한 m개의 설명변수 중 최적의 설명변수와 분기점(split point)을 찾는다.
- 노드를 두 개의 자식노드(child node)로 나눈다.

② B 개의 C_b을 앙상블 한 결과인 $\{C_b\}_1^B$ 을 도출한다.

③ 분류에 있어 $C_b(x)$을 b번째 의사결정나무의 분류 예측이라고 한다면, $\boldsymbol{x}_0$값을 갖는 새로운 개체에 대한 예측은 다음과 같이 가장 높은 빈도를 갖는 범주로 결정된다.

$$C_{rf}^B(x) = majority\ vote\ \{C_b(x)\}_1^B$$

랜덤포레스트는 무작위성(randomness)을 최대로 주기 위해서 각 부트스트랩 표

본에서 설명변수를 무작위로 추출한다. 따라서 관측치뿐만 아니라 변수에도 임의성을 적용함으로써 의사결정나무를 구축하는 데 다양성을 확보하게 된다. 따라서 의사결정나무의 특징인 분산이 크다는 단점을 보완할 수 있다.

위 알고리즘에서 랜덤포레스트의 의사결정나무의 수 B, 각 의사결정나무의 최소 노드 크기 n_{min}과 각 노드에서 임의로 선택할 설명변수의 수 m은 적절한 값을 선택해야 하는 초모수이다. 의사결정나무의 수는 랜덤포레스트의 성능을 결정하는 주요 모수로 클수록 모형의 성능이 좋아진다. 최소 노드 크기는 나무의 깊이를 설정하는 모수로 클수록 얕은 나무를 생성한다. 각 노드에서 임의로 선택할 설명변수의 수는 랜덤포레스트의 의사결정나무 간의 연관성을 설정하는 모수로 클수록 비슷한 나무들을 생성한다. 랜덤포레스트를 이용한 분류의 경우, 일반적으로 n_{min}은 1을 사용하고 m은 $\sqrt{p}$를 사용하며 본 연구에서도 이 값을 이용하였다. 다양한 분류모형에 대한 자세한 내용은 박유성(2022)를 참고하면 된다.

Ⅲ. 분류모형을 이용한 20대 대통령선거결과 예측

2022년 3월 9일 실시된 제20대 대통령선거에서는 윤석열 후보가 48.56%의 득표율로 47.84%의 득표를 한 이재명 후보를 제치고 대통령으로 당선되었다. 투표일 이전에 많은 수의 선거여론조사가 실시되었으나 본 연구에서는 매주 통신 3사로부터 제공받은 가상번호에서 전화번호를 추출하여 조사가 이루어진 전국지표조사(NBS: National Barometer Survey)자료를 활용하였다.

전국지표조사는 엠브레인퍼블릭, 케이스탯리서치, 코리아리서치, 한국리서치가 외부 기관의 의뢰를 받지 않고 자체적으로 시행·공표하는 정기 전화 여론조사이다. NBS 전화조사는 국내 전화조사에서 일반적으로 사용하는 할당추출(quota sampling) 방법이 아닌 모수추정이 가능한 확률적 표본추출방법인 층화확률추출(stratified random sampling) 방법을 사용하였다. 각 시·도 구분과 성/연령 그룹을 이용한 세부 층화를 통해 전체 192개 층을 구성하고 유권자 수 기준 비례배분을 통해 세부 층별 목표 표본크기를 산출하였다. 조사방법으로는 컴퓨터를 이용한 전화면접조사(CATI: Computer Assisted Telephone Interview)를 실시하였고, 전화를 받지 않거

나 통화 중이면 접촉률을 높이기 위해 요일과 시간대를 달리하여 5차례 재통화를 실시하였다. 재통화 실시에도 불구하고 조사할 수 없는 경우, 같은 층 내에서 다른 휴대전화 가상번호 1개를 무작위 추출하여 조사를 시도하였다.

본 연구에서는 주요 정당의 대통령 후보가 결정된 후인 2021.11~2022.3에 수행된 53~67차 조사자료를 사용하였다. 67차를 제외한 조사에서의 표본크기는 대략 1,000명 수준이며, 67차에는 2,013명을 조사하였다. 전체적인 응답률은 20.3%~32.5%였다. NBS 전화조사의 설문 문항은 응답자의 기본정보에 해당하는 거주지역, 성별, 연령, 직업, 학력을 묻는 문항과 응답자의 정치성향을 파악할 수 있는 대통령의 국정운영에 대한 의견, 지지하는 정당, 대통령선거에서 기대하는 결과 등의 문항을 포함하고 있다.

<표 1>은 각 조사 시점별 공표된 NBS 결과를 나타내고 있다. 지지 후보를 결정하지 않은 응답유보자의 비율이 10%~25%로, 두 후보자의 실제 득표율의 차이가 1% 미만임을 고려하면 응답유보자에 대한 예측은 최종 투표결과 예측에 매우 중요한 요소임을 알 수 있다. 또한 주요 후보였던 안철수 후보가 마지막 67차 NBS 조사가 이루어진 시점 이후인 3월 3일에 사퇴함에 따라 NBS 조사 시에 안철수 후보를 지지했던 응답자의 실제 투표 행위 역시 최종 선거결과를 예측하는 데에 매우 중요한 요소로 작용하였다. 이러한 상황을 반영하여 본 연구에서는 응답유보자 그리고 안철수 후보 지지자를 2장에서 언급한 분류모형을 이용하여 재분류한 후 투표결과를 예측하였다.

분류모형의 적용을 위하여 거주지역, 성별, 연령대, 직업, 학력, 경제적 계층 인식, 이념성향, 대통령 국정운영 평가, 지지 정당, 투표 참여 의향, 대선 당선 전망, 대통령선거 인식이 설명변수로 사용되었다. 분석을 위해서는 R이 사용되었으며, 의사결정나무 분석을 위해서는 rpart, 랜덤포레스트 분석을 위해서는 randomForest 패키지가 사용되었다. 분류를 위해서는 지지하는 후보 '없음' 혹은 '모름'으로 응답한 대상자를 테스트 데이터(test data)로 특정 후보를 지지한 응답자를 학습 데이터(training data)로 정의하고, 학습 데이터를 이용하여 분류모형을 구축하고 이를 테스트 데이터에 적용하여 분류를 시행하였다. 이 과정에서 지지하는 후보 '없음' 혹은 '모름'으로 응답한 대상자는 이재명 지지자, 윤석열 지지자, 심상정 지지자 그리고 기타 후보 지지자 중 하나의 경우로 분류되었다. 67차 자료 기준 각 분류 방법의 오분류율은 다항 로지스틱 회귀모형 4.92%, 의사결정나무 방법 6.35% 그리고 랜덤

포레스트 0.6%로 랜덤포레스트의 오분류율이 가장 낮게 나타났다.

<표 1> NBS 조사결과 (단위: %)

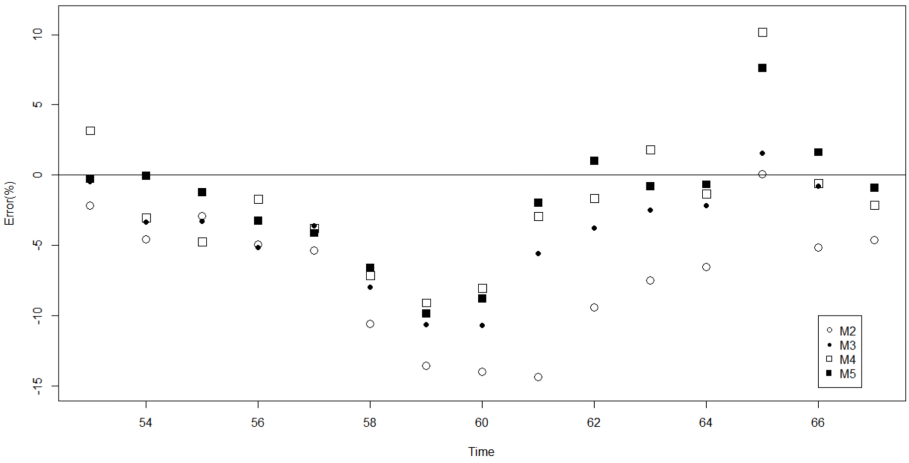
회차	윤석열	이재명	기타	응답유보
53차	38.67	32.47	12.25	16.62
54차	36.27	34.67	11.48	17.59
55차	35.26	32.47	9.54	22.73
56차	33.64	33.07	10.40	22.89
57차	35.69	37.68	9.25	17.38
58차	28.80	35.05	11.49	24.22
59차	28.17	39.07	13.25	19.50
60차	27.69	36.48	15.92	19.90
61차	28.13	36.57	17.58	17.73
62차	32.72	34.08	16.74	16.48
63차	33.75	35.05	13.40	17.81
64차	35.29	34.51	14.17	16.02
65차	40.05	30.90	11.40	17.65
66차	38.65	37.18	13.18	10.99
67차	38.22	37.97	10.72	13.09

<표 2>는 다양한 통계적 방법을 통해 선거결과를 예측한 결과이다. 방법 1(M1)은 NBS 조사결과이며, 방법 2(M2)는 NBS 전화조사 결과에서 응답 유보층을 제외하고 나머지 후보자들의 지지율 합이 100%가 되도록 조정한 것이다. 방법 3(M3), 방법 4(M4) 그리고 방법 5(M5)는 응답유보자와 안철수 지지자를 각각 다항 로지스틱 방법, 의사결정나무 그리고 랜덤포레스트 방법을 통해 분류한 후 선거결과를 예측한 결과이다.

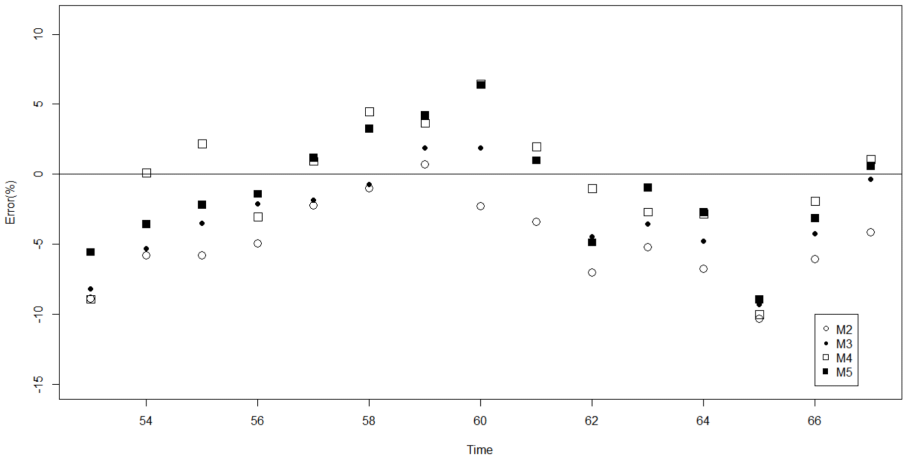
<표 2> 예측 결과[후보1: 윤석열(48.56%), 후보2: 이재명(47.83%)] 단위: %)

차수	방법 1(M1)		방법 2(M2)		방법 3(M3)		방법 4(M4)		방법 5(M5)	
	후보1	후보2	후보1	후보2	후보1	후보2	후보1	후보2	후보1	후보2
53	38.67	32.47	46.37	38.94	48.10	39.64	51.72	38.90	48.26	42.26
54	36.27	34.67	44.01	42.07	45.22	42.53	45.52	47.91	48.49	44.24
55	35.26	32.47	45.63	42.02	45.30	44.31	43.82	50.02	47.34	45.65
56	33.64	33.07	43.63	42.89	43.42	45.74	46.86	44.80	45.29	46.41
57	35.69	37.68	43.20	45.60	44.95	46.01	44.78	48.75	44.43	48.98
58	28.80	35.50	38.00	46.84	40.57	47.11	41.41	52.28	41.98	51.05
59	28.17	39.07	35.00	48.54	37.92	49.69	39.47	51.47	38.68	52.01
60	27.69	36.48	34.58	45.54	37.85	49.71	40.53	54.28	39.75	54.18
61	28.13	36.57	34.19	44.45	42.99	48.90	45.63	49.78	46.59	48.81
62	32.72	34.08	39.17	40.80	44.78	43.37	46.90	46.82	49.58	42.95
63	33.75	35.05	41.06	42.64	46.06	44.26	50.35	45.13	47.76	46.88
64	35.29	34.51	42.03	41.10	46.40	43.03	47.20	45.00	47.89	45.12
65	40.05	30.90	48.64	37.52	50.11	38.52	58.74	37.80	56.15	38.89
66	38.65	37.18	43.42	41.77	47.79	43.59	47.98	45.89	50.18	44.68
67	38.22	37.97	43.97	43.69	47.54	47.50	46.41	48.87	47.65	48.42

또한 <그림 1>과 <그림 2>는 각 방법론을 통해 예측된 두 후보의 지지율과 실제 투표에서 얻는 지지율의 차이로 계산된 단순 예측오차이다. 응답유보자와 안철수 지지자를 고려하지 않은 방법의 지지율 예측오차가 가장 크게 나타나고 있는 것을 확인할 수 있으며 조사 기간의 중반에는 윤석열 후보의 약세가 두드러지게 나타나고 있음을 확인할 수 있다. 지지율 예측에 있어서는 전체적으로 랜덤포레스트 방법이 다른 방법에 비하여 작은 단순 예측오차를 나타내고 있다.



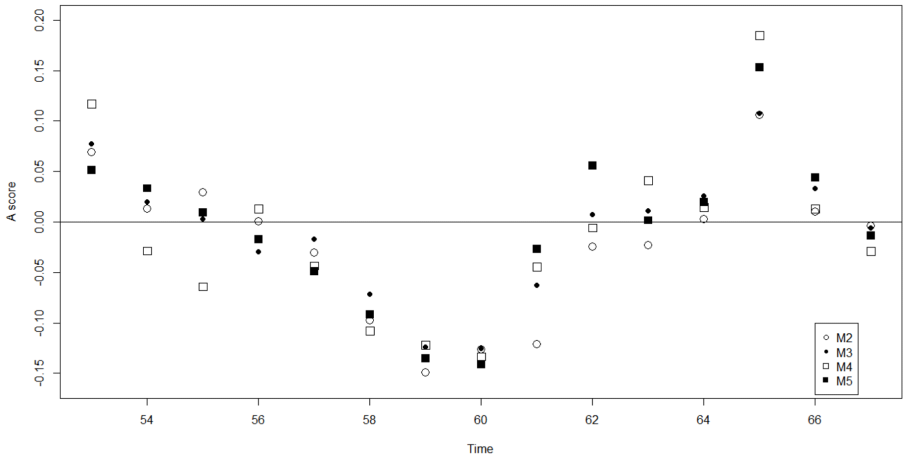
<그림 1> 윤석열 후보자 지지율의 단순 예측오차



<그림 2> 이재명 후보자 지지율의 단순 예측오차

<그림 3>은 식 (1)에서 정의한 척도 A의 결과를 보여주고 있다. 응답유보자와 안철수 후보 지지자를 제외하고 산출한 1, 2위 지지율의 비율을 통해 산출한 척도 A의 값과 다양한 분류 방법을 적용하여 계산한 척도 A 값이 대부분은 매우 유사

하게 나타나고 있다. 특별히 마지막 조사인 67차의 경우에는 분류 나무 방법론을 제외하고 나머지 모든 방안이 1, 2위 지지율의 비율을 거의 정확하게 예측한다. 이는 1위와 2위의 지지율 차이를 기준으로 각 방법을 평가할 때 응답유보자나 안철수 후보 지지자의 투표가 최종 결과에 미치는 영향이 적었음을 의미하며, 더 나아가서는 응답유보자나 안철수 후보 지지자의 평균적인 투표성향이 조사 과정에서 지지 후보를 나타낸 응답자군과 유사함을 의미한다. 타 차수에 비하여 65차의 결과가 선거일이 가까움에도 불구하고 분류 방법을 적용한 결과의 예측오차가 크게 나타나고 있다. 주목할 점은 해당 차수의 응답률이 20.3%로 현저히 낮았고 65차 조사 기간 중 공식 선거운동이 시작되었다는 점이며, 이러한 비표본 오차와 선거환경 변화가 낮은 수준의 예측 정확성에 대한 가능한 하나의 설명이 될 수 있을 것이다.



<그림 3> 척도 A

Ⅳ. 결론

본 연구에서는 지난 20대 대통령선거 과정에서 수행된 선거여론조사 중 NBS 조사자료를 활용하여 선거결과를 예측하는 방안을 고려하였다. 이를 위하여 평균 20% 수준인 응답유보층과 더불어 투표일 직전에 사퇴한 안철수 지지자를 다양한

지도학습 방법을 통해 분류하였다. 선거 전 5개월 동안의 지지율 예측값 분석을 통해서 시간에 따른 선거의 변화 양상을 파악할 수 있었다. 윤석열 후보는 2021년 12월부터 22년 1월 초까지 약세를 보였으나 이후 반등하였다. 공식 선거운동 기간이 시작된 이후에는 두 후보의 지지율이 매우 근접해 있으나 윤석열 후보가 약간 앞서는 패턴을 보인다. 5개월의 기간 동안 두 후보의 지지율 변동이 크게 나타났으며, 이는 부동층의 의견이 시점에 따라 매우 유동적이었음을 나타내고 있다.

전체적으로 각 후보의 지지율 예측에 있어서는 랜덤포레스트 기법을 통해 응답유보자와 안철수 후보 지지자를 분류한 결과가 상대적으로 다른 방법들에 비하여 작은 예측오차를 제공하고 있다. 마지막 선거여론조사인 67차 결과에 지도학습 방안을 적용하여 산출한 두 후보 지지율의 오차는 최대 2% 내로 매우 정확한 예측을 한 것으로 판단된다. 이러한 결과는 사전 여론조사 금지기간이 실제 적용되지 않으리라고 기대되는 차기 선거에서는 적절한 지도학습 방법을 적용한 분류기법을 통해 투표일에 근접한 시점의 선거여론조사를 통해서 선거결과를 예측할 수 있음을 보여주고 있다. 다만 본 연구의 결과는 지난 20대 대선 자료만을 분석한 사후연구 결과이기 때문에 이의 직접적인 활용을 위해서는 후보자들의 구성이나 투표율 등의 외적인 요소를 고려하여 추후 적용되어야 할 것이다. 특별히 본 연구에서 고려한 방안들을 선거일 이전에 적용하여 선거결과를 예측한 후 실제 결과와 비교하는 방법 역시 고려되어야 할 것이다.

참고문헌

강현철·한상태·김지연·정용찬·허명회. 2008. “RDD 전화조사와 주요결과.” ≪조사연구≫ 9(1): 1-22.

곽은선·김영원. 2022. “현행 선거여론조사 방법의 정확성과 선거결과 예측 가능성.” ≪조사연구≫ 23(1): 131-153.

김선웅. 2004. “이동전화 확산에 따른 유선전화 가구보유율의 변화: 한국을 포함한 주요 국가들을 중심으로.” ≪조사연구≫ 5(1): 27-49.

김정훈·김지연·권은혜·김혁·강현철. 2017. “여론조사를 통한 선거예측에서 무응답층 보정의 효과.” *Journal of the Korean Data Analysis Society* 19(1): 107-117.

박유성. 2022. ≪파이썬을 이용한 통계적 머신러닝≫. 자유아카데미.

허명회·김영원. 2006. “RDD 표본 대 전화번호부 표본: 2007년 대통령 선거 예측 사례.” ≪조사연구≫ 9(3): 55-69.

Arzheimer, K. and J. Evans. 2014. “A New Multinomial Accuracy Measure for Polling Bias.” *Political Analysis* 22: 31-44.

Breiman, L. 2001. “Random Forests.” *Machine Learning* 45(1), 5-32.

Casady, R.J. and J. Lepkowski. 1999. “Telephone Sampling.” in P.S. Levy and S. Lemeshow. *Sampling of Populations*. New York: Wiley. Chapter 13 pp. 455-479.

Lepkowski, J. 1988. “Telephone Sampling Methods in the United States.” in R.M. Groves et al(Eds). *Telephone Survey Methodology*. New York: Wiley. Chapter 5 pp. 73-98.

Martin, E.A., M.W. Traugott, and C. Kennedy. 2005. “A Review and Proposal for a New Measure of Poll Accuracy.” *Public Opinion Quarterly* 69(3): 342-369.

Waksberg, J. 1978. “Sampling Methods for Random Digit Dialing.” *Journal of the American Statistical Association* 73: 40-46.

<접수 2023.03.08; 수정 2023.04.17; 게재확정 2023.05.01>

Predicting Election Results by Correcting for Retained Supporters and Respondents of Candidates Who are Eliminated in the Middle

Jeong-Kyung Kim
(Korea University)
Mingue Park
(Korea University)

This study examines a statistical plan to predict the winners of the 20th presidential election held in March 2022 using the results of the NBS election poll. About 20% of respondents in the 20th presidential election could not decide their supporters until a week before the voting date. Considering that Ahn Cheol Soo resigned just before the election day, the percentage of votes for each candidate was predicted after classifying them through supervised learning. The supervised learning tools considered were logistic regression analysis, decision tree, and random forest measures. The Random Forest method showed the best performance in predicting the vote rate, and all measures similarly predicted the ratio of the vote rate of the first and second candidates. The accuracy of the prediction increased as the election date approached, but the prediction error at the time when the response rate was relatively low was very large. If the prohibition period for reporting public opinion polls is lifted and thus it is possible to classify respondents using data just before the voting day, the prediction of the vote rate using election polls is expected to be more sophisticated.

Key words: presidential election, supervised learning, decision tree, random forest