한국정치학회보, 59집 2호 2025 여름, 75-108

투표율과 정당 편향:
머신러닝을 활용한 예측 시뮬레이션

조진희*

논문요약

본 논문은 21대 총선과 20대 대선 유권자 조사를 분석하여 투표율과 정당 편향의 관계를 알아본다. 널리 사용되는 머신 러닝 방법을 활용한 예측 시뮬레이션을 통해 총 투표율, 사전투표율, 선거당일 투표율의 상승이나 하락이 정당 득표율과 선거 결과에 어떤 영향을 미치는지를 분석한다. 분석 결과, 투표율 상승이나 사전투표율 상승이 진보 정당에 더 유리하다는 통념은 부분적으로만 옳은 것으로 나타난다. 투표율만을 가지고 특정 정당에 유리하거나 불리하다는 일관된 결론을 내리는 것은 매우 어려우며, 선거 자체의 특징과 유권자의 선호 변화에 따라 어떤 투표방법의 참여율 변화가 어떤 정당에게 더 유리할지는 달라질 수 있다.

주제어: 투표율, 머신러닝, 시뮬레이션, 정당 편향, 선거

* 경희대학교 정치외교학과 교수
https://doi.org/10.18854/kpsr.2025.59.2.004

원고접수일 2025년 04월 05일 | **심사시작일** 2025년 04월 25일 | **게재확정일** 2025년 06월 13일

Ⅰ. 서론

투표율 상승 또는 하락은 특정 정당에 더 유리할까? 사전 투표율, 또는 선거 당일 투표율의 상승이나 하락은 특정 정당에 더 유리할까? 특정 투표 방법은 특정 정당에 더 유리할까? 투표 참여가 가장 기본적인 정치참여에 해당된다는 점을 생각할 때, 투표율 및 투표방법과 정당편향의 관계를 밝히는 것은 민주적 반응성 제고의 측면에서 매우 중요한 문제이다. 더욱이, 제21대 국회의원 선거 이후 나타난 선거제도에 대한 다양한 논쟁들은 이러한 질문의 중요성을 더욱 부각시킨다.

전통적인 통념은 낮은 투표율이 보수 정당에 유리한 반면, 높은 투표율이 진보 정당에 더 유리하다는 것이다. 그 기저의 논리는 젊은 층의 투표 참여와 관련되어 있다. 구체적으로, 연령대가 높은 경우 통상 높은 확률로 선거에 참여하기 때문에, 선거에서 투표율이 높아진다면 젊은 층이 평소보다 더 높은 확률로 투표했다는 의미라는 것이다. 연령대가 높을수록 보수정당 지지자일 가능성이 더 높기 때문에, 젊은 층의 투표율 상승은 상대적으로 진보정당에 더 유리하게 작용할 것이라는 논리이다.

그러나 기존 연구들에 따르면 이러한 통념은 경험적 근거가 부족한 것으로 나타난다. 가령, 지병근(2012)은 19대 총선을 분석하여 투표율이 올라가면 오히려 보수정당의 당선 가능성이 높아진다고 주장하였다. 그와 비슷하게 박경미(2012)도 13대 총선부터 19대 총선까지를 대상으로 하여 서울시 선거구를 분석하고, 투표율 상승이 전통적 통념과는 다른 효과를 가진다는 것을 보였다. 그에 따르면 투표율 상승은 보수정당의 상대적 득표율에는 긍정적 영향을 미치지만, 그렇다고 해서 진보정당의 당선 가능성이나 득표율 등에 부정적 영향을 미치지는 않는다.

최근의 선거에서 더욱 두드러지는 논란은 사전투표와 관련된 부정선거 음모론이다. 해당 논란은 온라인 매체를 중심으로 제기되면서 특히 보수 성향의 일반 국민들에게도 큰 호응을 얻었으며, 심지어 학계에서도 부정선거 여부를 통계적으로 탐지하려는 노력이 이루어진 바 있다(민인식, 유경준 2021; Mebane 2020). 부정선거론에 찬성하지 않는 유권자의 경우에도, 해당 논란으로 인해 사전투표에 대한 신뢰도가 낮아졌음은 물론, 사전투표가 진보계열 정당에 더 유리하다는 통념이 새롭게 등장하여 강화되는 추세이다(김준석 외 2022).

투표율과 정당 편향, 더 구체적으로 사전 투표율 혹은 당일 투표율과 정당 편향의 문제는 향후 선거결과의 민주적 반응성과 대표성 강화를 위한 제도 개편의 측면에서도 반드시 다루어져야 할 주제이다. 선거의 공정성에 대한 논란으로 인해 사전 투표제도를 더욱 확대 개편할 것인지, 아니면 축소할 것인지, 그리고 본투표일을 확대할 것인지 여부에 대한 다양한 논의 뿐 아니라 법안 발의까지 이루어지고 있는 현 시점에서 과학적인 근거를 바탕으로 한 학술적 연구가 절실하다.[1]

이 논문에서는 머신러닝 예측 모델을 활용하여 투표율, 사전투표율, 선거 당일 투표율의 정당 편향에 대해 다각도로 검토한다. 다양한 목적으로 널리 활용되는 소프트맥스 회귀(Softmax Regression with

[1] 2025년 3월 4일 국민의힘 장동혁 의원(충남 보령 · 서천)이 사전투표 폐지를 골자로 하는 개정 법안을 대표발의 하였다.

Regularization), 랜덤포레스트(Random Forest), XG부스트(eXtreme Gradient Boosting), 그리고 딥뉴럴네트워크(Deep Neural Network)를 제21대 국회의원 선거와 제20대 대통령 선거 유권자 설문조사 자료에 적용하여 투표참여와 투표선택에 대한 다양한 반사실적(counterfactual) 시나리오들을 분석한다. 구체적으로, 첫째, 투표에서 기권한 유권자의 투표 선택을 예측하여 기권자가 선거에 참여했다면 선거결과가 어떻게 바뀌었을지 분석한다. 둘째, 지난 21대 총선과 20대 대통령 선거에서 총 투표율, 사전투표율, 선거당일 투표율이 더 상승했거나 더 하락했다면 정당 득표율의 격차가 어떻게 변화했을지 예측한다. 셋째, 어느 정당이나 후보에 투표했는지, 혹은 어느 정당이나 후보에 투표했을 것으로 예측되는지에 따라 유권자의 투표 참여 형태가 달라지는지 분석한다.

21대 총선과 20대 대선은 투표율 변화와 정당 편향을 알아보기에 유용한 조건을 제공한다. 21대 총선을 분석 대상으로 삼은 이유는 해당 선거에서 사전투표 논란이 시작되었기 때문이다. 21대 총선 자료를 분석함으로써 사전 투표와 당일 투표의 투표율 변화가 과연 선거결과에 논란이 될 만큼의 영향을 미쳤는지 알 수 있다. 20대 대선은 21대 총선으로부터 약 2년 뒤에 치러진 선거인데, 21대 총선에서와는 달리 보수 정당의 윤석열 후보가 승리를 거두었다. 승리한 정당이 서로 다른 두 인접 선거를 비교함으로써 투표율과 정당편향이 비슷한 방향으로 일관된 관계를 갖는지 파악할 수 있다.

이 논문의 공헌은 크게 두 가지로 정리할 수 있다. 먼저, 기존 연구가 선거구를 분석단위로 이루어진 데에 반해, 이 논문에서는 개인 유권자를 분석단위로 한다. 선거구 단위의 분석이 많은 유용한 시사점을 주기는 하지만, 투표를 결정하는 주체는 선거구가 아니라 유권자 개인인 만큼 생태학적 오류(ecological fallacy)에 빠질 위험이 있다. 선거구 단위의 분석에서는 보통 투표율이 높은 선거구와 낮은 선거구에서의 정당 및 후보 득표율을 비교하게 되는데, 만일 투표율이 높은 선거구와 낮은 선거구 간 유권자 개인들의 특징이 서로 다르다면 이런 식의 비교가 설득력을 잃게 된다. 어떤 사람이 선거에서 기권했는지 집단분석으로는 알아내기 어렵기 때문에, 투표율이 오른다고 하더라도 그 결과를 예측하기가 쉽지 않다. 또한 기권자 집단과 투표 참여자 집단이 매우 다른 정치적 특징을 지니고 있는 경우에도 투표율 상승의 효과를 제대로 알아내기가 어려워진다. 선거구 단위의 분석에 대해 기존 연구에서 제기되는 또 다른 문제는 내생성(endogeneity) 문제이다. 정당의 투표독려 캠페인이 득표율과 투표율에 동시적으로 영향을 주어 내생성 문제가 발생한다는 것이다(Arnold and Freier 2016). 이 논문에서는 개인 유권자를 분석단위로 한 예측모델을 활용하여 선거구 단위 분석의 문제점을 보완할 수 있다.

다음으로, 이 논문은 기존의 전통적 통계학적 접근에서 벗어나, 최근 발전하고 있는 머신러닝 예측모델을 활용하여 투표율과 정당편향의 문제를 다룬다. 투표율과 관련한 예측 문제는 전통적 회귀분석보다 머신러닝 모델로 더 잘 해결할 수 있다. 전통적 회귀분석 방식에서는 독립변수가 종속변수에 어떤 식으로 영향을 미치는지를 알고자 하지만, 이 논문에서 다루는 문제는 종속변수를 잘 예측하는 것과 관련이 있다. 즉 예측을 주된 목적으로 하는 머신러닝 기법이 가장 적합한 연구문제라고 할 수 있다. 머신러닝을 기반으로 한 시뮬레이션을 통해 투표율이 어떤 식으로 선거결과에 영향을 미치는 지에 대해 다양한 분석 결과를 제시한다.

Ⅱ. 선행 연구

투표율과 정당편향에 대한 기존의 연구는 상당 부분 미국정치의 상황에 집중되어 있다. 전통적으로 공화당 지지자들은 정당 충성도와 투표 참여율이 높으며, 상대적으로 이에 반해 민주당 지지자들의 충성도 및 투표 참여율은 낮은 것으로 여겨져 왔다. 따라서 높은 투표율은 민주당에 더욱 유리하고 낮은 투표율은 공화당에 유리하다는 통념이 존재하는데, 많은 학술적 연구가 이러한 통념을 뒷받침하고 있다.

가령, 고메즈 외(Gomez et al. 2007)는 미국 대선에서 날씨가 투표율에 미친 효과를 분석하였다. 궂은 날씨 하에서의 투표율은 통상의 투표율보다 의미있는 수준으로 낮아지는데, 그 결과 공화당의 득표율이 높아지는 경향이 있음을 밝혔다. 이와 비슷하게 마티네즈와 힐(Martinez and Hill 2007)은 2000년과 2004년의 미국 대선을 분석하여 높은 투표율이 민주당에게 더 유리하다고 주장하였다. 그에 의하면, 공화당 후보였던 부시(Bush)는 높은 투표율에도 불구하고 선거에서 승리했다는 것이다. 다만, 대통령 선거가 아닌 주지사 선거에서는 그와 같은 일관된 정당 편향이 나타나지 않았다.

이러한 연구들에 대한 주요한 비판 중 하나는 바로 내생성 문제이다. 선거구의 투표율은 외생적으로 주어진 것이 아니며, 정당의 동원 노력에 의한 효과로 인해 득표율에 대한 독립적 효과를 추정하는 것이 어렵다는 비판이다 (Hansford and Gomez 2010; Arnold and Freier 2015). 이러한 문제점을 해결하기 위해, 한스포드와 고메즈(Hansford and Gomez 2010)는 날씨를 도구변수로 사용하여 도구변수추정(Instrumental variable estimation)을 활용한 분석 결과를 제시한다. 도구변수 추정 결과에 따르면, 투표율이 높아질 때 여전히 민주당 후보가 더 유리해지지만, 그러한 효과는 늘 일정한 것이 아니라 선거구민의 선호 정당 구성과 현재 대통령의 정당이 어디냐에 따라 조건 지어진다. 아놀드와 프라이어(Arnold and Freier 2015)도 날씨를 도구 변수로 활용하되, 독일의 노르트라인-베스트팔렌(North-Rhine Westphalia) 주의 선거를 분석하였다. 미국에서의 많은 연구결과와 마찬가지로 투표율이 낮아지면 보수 정당에 더 유리하다는 결과를 보여준다.

이들 연구와는 다르게 개인을 분석단위로 하는 연구로는 시트린 외(Citrin et al., 2003)의 연구를 들 수 있다. 이 연구에서는 주 단위의 출구조사 결과와 센서스(Census) 데이터를 결합하여 모든 사람이 투표를 한다면 선거 결과가 어떻게 바뀌는지에 대한 시뮬레이션을 실시하였다. 연구 결과, 선거에 참여하지 않은 유권자들은 민주당 지지자일 가능성이 더 높았지만, 많은 경우 선거결과를 뒤집을 만큼의 영향력은 없었을 것으로 예측되었다. 즉, 투표율 상승이 민주당에게 유리한 것이 사실이라고 하더라도 그 정도가 선거 결과를 뒤집을 정도는 아니라는 것이다. 그러나 기권한 유권자들과 투표에 참여한 유권자들의 정당지지 성향이 얼마나 다른지는 선거에 따라서, 지역에 따라서 크게 달라지기 때문에 일률적인 영향을 단정 지을 수 없다는 결론을 제시한다. 이와 비슷하게 루벤슨 외(Rubenson et al., 2007)의 연구도 2000년 캐나다 연방 선거 설문데이터를 분석하여 모든 사람이 투표를 했다고 하더라도 선거 결과가 바뀌지 않는다는 결론을 제시한다. 한편, 베른하겐과 마쉬(Bernhagen and Marsh 2007)의 경우에는 기권자의 투표선택을 데이터 결측치 문제로 치환하여 다중 대체 기법을 활용한다. 이들의 분석

에 따르면 모든 사람이 투표한다고 하더라도 반드시 중도좌파 정당에게 이득이 되기보다는 소수 정당 및 비현직자 후보들에게 유리해지는 결과가 나타난다.

투표율과 정당 편향에 대한 한국의 연구는 주로 사회적으로 활발한 투표독려운동이 이루어졌던 2012년 총선에 초점이 맞추어져 있다. 주로 야당이었던 민주통합당 지지자들을 중심으로 투표독려운동이 벌어졌기 때문에, 투표율 상승이 정말 진보정당에 유리한지 알아보고자 하는 노력이 이루어졌다. 흥미롭게도, 연구의 결과는 주로 전통적 통념에 반하는 것이다. 지병근(2012)는 19대 총선의 결과를 선거구 수준에서 분석하고, 투표율이 높은 선거구에서 새누리당 후보의 득표율과 당선가능성이 더 높다는 경험적 결과를 제시한다. 박경미(2012)는 서울시 선거구의 13대-19대 총선 결과를 분석하여 투표율이 새누리당의 당선가능성과 득표율에 미치는 긍정적 효과가 제한적 조건하에서만 성립한다고 주장한다. 또한 투표율이 상승한다고 해서 진보정당이 불리해지는 것도 아니라는 분석결과를 제시한다.

보다 최근의 연구로는 머신러닝 기법을 활용한 가상준(2024)을 들 수 있다. 이 연구에서는 20대 대선과 22대 총선에서 투표 기권자가 만일 투표에 참여했더라면 어떤 정당에 투표하였을지를 분석하였는데, 분석 결과 기권자들은 선거에서 패배한 정당을 선택했을 것으로 예측되었다. 즉, 투표율 상승이 특정 정당에 유리하다기 보다는 선거에 따라서 투표율에 따른 정당 유불리가 달라졌다는 결론이다.

투표율과 정당 편향에 대한 한국의 연구는 주제의 중요성에도 불구하고 그 숫자가 매우 제한적이다. 지병근(2012)과 박경미(2012)의 연구가 많은 유용한 시사점을 제공하지만, 선거구 단위의 집합적 데이터를 주로 분석했다는 점, 그리고 해외 연구에서 지적된 내생성 문제를 다루지 못하고 있다는 점에서 보완이 이루어질 여지가 있다. 또한, 가상준(2024)에서는 기권자에 대한 분석만이 이루어지고 있는데, 투표자의 투표참여 행태도 함께 분석하면 투표율 상승 뿐 아니라 하락에 따른 결과도 살펴볼 수 있다.

투표율과 정당편향에 직접적인 관계는 없지만, 사전투표에 대한 한국의 연구도 살펴볼 필요가 있다. 사전투표에 대한 선행 연구는 주로 사전투표의 동원효과와 편의효과를 분석하는 데에 집중한다. 즉, 사전투표가 없었다면 투표하지 못했을 유권자가 사전투표로 인해 투표에 참여할 가능성이 높아지는 지가 연구의 주된 관심사라고 할 수 있다. 이를 분석하기 위해 기존 연구에서 취하는 주된 접근 방법은 사전투표유권자가 기권자와 당일 투표자 중 어느 집단과 더 유사한지 살펴보는 것이다. 20대 총선까지의 많은 연구들이 사전투표자와 당일투표자는 동질적인 집단이며, 따라서 사전투표제도는 동원효과는 약하고 유권자에게 편의를 제공할 뿐이라는 결론을 내린다(가상준 2016, 2018; 강신구 2016; 김도경 2014).

흥미로운 점은, 20대 총선까지의 연구들을 보면 사전투표자와 당일투표자를 구분하기 위한 주요 변수 중에 이념적 방향성은 포함되지 않는다는 것이다. 정당일체감이 있는 유권자인지, 극단적 이념성향을 지닌 유권자인지 등이 포함되는 경우도 있지만, 진보와 보수를 가르는 설명변수들은 분석에 포함되지 않았다.

사전 투표가 논란이 된 21대 총선 이후에는 이념적 방향성이 주요한 설명변수들로 연구에 등장하고 있지만 다양한 결과가 혼재되어 있다. 예를 들어, 이재묵(2020)의 연구에서는 자기이념과 문재인 정부

에 대한 평가가 주요 변수로 포함되었으며, 당일투표자가 사전투표자에 비해 더 보수적이고 문재인 정부에 더 비판적이라는 분석 결과를 제시한다. 반면, 가상준(2021)에서는 이념성향이 주요 변수로 포함되었으나 사전투표자와 당일투표자 간에 유의미한 차이를 나타내지 못하는 것으로 나타난다. 박상훈과 허재영(2023)의 연구에서도 20대 대통령 선거를 분석하며 주요 정당에 대한 일체감 및 이념성향을 설명변수로 포함시켰으나 역시 사전투표자와 당일투표자 간에 유의미한 효과를 지니지 못하는 것으로 분석되었다. 사전투표에 대한 많은 논란에도 불구하고 최근 선거에서 정당편향성과 관련된 간접적인 증거조차 일관되지 않다는 점은 주목할 만하다.

Ⅲ. 데이터와 방법론

이 논문에서는 21대 총선과 20대 대선의 유권자 사후 설문조사 데이터를 분석에 활용한다. '제 21대 국회의원 총선거 관련 유권자 정치의식조사'는 2020년 4월에 한국사회과학데이터센터와 한국선거학회가 공동으로 실시하였으며, 총 1200명의 응답 자료를 포함하고 있다. '제 20대 대통령 선거 관련 유권자 의식조사'는 2022년 4월에 한국사회과학데이터센터가 실시하였으며, 총 1250명의 응답자료를 포함하고 있다.

머신러닝 기반의 예측모델에서는 통계적 접근방법과는 달리 매우 많은 변수들을 모형에 포함시키는 것이 일반적이다. 통계적 접근에서는 종속변수의 예측보다는 참된 모형(true model)에 기반한 독립변수의 유의성이 주된 관심사가 되기 때문에, 올바른 모형(correct model)을 선택하고 필수적인 설명변수를 선별하여 분석에 포함시키는 것이 매우 중요하다. 그러나 머신러닝 기반의 예측모델에서는 타겟(target) 변수의 예측 정확도를 높이는 것이 주된 관심사이며, 따라서 인과적 관계에 기반하지 않은 변수라고 하더라도 예측 정확도를 높일 수 있다면 모형에 포함시켜 분석을 진행하게 된다.

이 논문에서도 가능한 모든 설문문항을 분석에 활용하였다. 구체적으로, 타겟 변수 중 하나인 투표에 참여했는지를 묻는 문항만을 예외로 하고, 모든 응답자가 대상이 되는 모든 문항을 모형에 포함시켰다. 즉, 1. 어느 정당이나 후보에 투표했는지를 묻는 투표 선택을 비롯해서 투표에 참여한 사람만 응답하게 되어 있는 문항, 2. 투표를 하지 않은 이유 등을 비롯해서 기권한 사람만 응답하게 되어 있는 문항, 그리고 3. 투표에 참여했는지를 묻는 문항의 세 가지 경우를 제외한 모든 문항이 분석에 포함된다.[2]

[2] 변수 선택(feature selection)에 관한 최근의 추세는 연구자의 자의적 변수 배제보다는 규제(regularization) 등을 활용한 가중치 조정을 통해 모델 훈련 과정에서 변수 선택이 이루어지는 것이 더 바람직하다고 본다. 피처가 10,000개 이상이라던지, 타겟 변수와 상관없는 변수들이 포함되어 있는 경우에는 여전히 불필요한 변수를 배제하는 것이 계산 비용을 줄이고 모델 성능을 향상시킬 수 있겠지만, 이 논문의 데이터는 선거 사후 설문조사로 이런 경우에 해당하지 않는다. 특히, 이 논문에서 주요하게 활용하는 규제를 포함한 소프트맥스 회귀와 트리 계열(tree-based) 모델들의 경우 사전적 변수 선택이 불필요하다는 것이 잘 알려져 있다(Islam et al. 2024; Ruczynski and Kozak 2024).
중복적 변수가 있다면 제거하는 것이 계산비용을 절약할 수 있기 때문에 피처 간 상관행렬을 계산하여 상관계수가 일정 기준(0.7,

그 결과 21대 총선의 경우에는 총 183개의 피쳐가 사용되었고, 20대 대선에서는 총 210개의 피쳐가 사용되었다. 이 중에 범주형 변수(categorical variables)는 원-핫-인코딩(one-hot-encoding)하여 분석에 활용하였다.[3] 분석에 포함된 피쳐들은 기존 연구에서 중요하게 다뤄지는 핵심적인 변수(선거에 대한 관심, 나이, 지역, 과거 투표 이력, 이념성향, 후보자나 정당에 대한 평가, 호감도 등)들은 물론 다양한 정치적 의견과 성향에 대한 내용을 포함하고 있다.

이 논문에서는 널리 쓰이는 머신러닝 모델 중 규제를 포함한 소프트맥스 회귀(Softmax Regression with Regularization), 랜덤포레스트(Random Forest), XG부스트(eXtreme Gradient Boosting), 그리고 딥뉴럴네트워크(Deep Neural Network)를 활용하여 예측모델을 훈련하고, 그 중 가장 뛰어난 성능을 보인 모델을 최종 분석에 활용하였다.[4]

소프트맥스 회귀는 다중분류에 널리 쓰이는 머신러닝 모델로, 너무 복잡하지 않은 분류 문제에서 뛰어난 성능을 보인다. 피쳐의 개수가 너무 많을 때에 나타날 수 있는 과대적합 문제를 해결하기 위해 엘라스틱넷(Elastic Net) 규제를 적용하였다.[5] 엘라스틱넷 규제는 L1규제(Lasso)와 L2규제(Ridge)를 혼합하여 변수의 가중치에 패널티를 가하게 된다. 엘라스틱넷 규제를 통해 불필요한 피쳐들은 0 또는 0에 가까운 가중치를 받게 되어 과대적합 문제를 해결할 수 있다.

다음으로, 랜덤포레스트와 XG부스트는 같은 결정트리 계열의 모델이다. 트리계열의 모형들은 직관적인 해석이 가능하고 복잡한 데이터에서도 잘 작동한다는 장점을 지니고 있어서 정치학 연구에서도 널리 활용되고 있다(e.g., Gohdes 2020; Wäckerle and Silva 2023). 특히 XG부스트는 첸과 구에스트린(Chen and Guestrin 2016)이 개발한 부스팅 모델로 많은 머신러닝 대회에서 우승을 거둘 정도로 분류 문제에 성능이 뛰어나다. 두 모델 모두 결정트리를 기반으로 하지만, 사용하는 앙상블 학습 방법에 차이가 있다. 먼저, 랜덤포레스트의 경우에는 같은 데이터를 여러 번 샘플링해서 서로 다른 데이터셋을 만들고 각각의 데이터셋으로 결정트리를 학습한다. 이를 배깅(bagging) 방식이라 부른다. 랜덤포레스트는 이렇게 학습된 결정 트리의 예측을 모아서 다수결 방식으로 분류 문제를 예측하기 때문에 하나의 트리에 비해 훨씬 안정적이고 과대적합 문제도 완화된다. 한편, XG부스트의 경우에는 간단한 하나의 트리로 예측을 시작하여, 그 트리의 예측 오차를 줄이는 방향으로 새로운 트리를 추가해나가는 학습방식을 사용한다. 예측 오차를 줄이기 위해 손실함수(loss function)를 기준으로 한 오차의 변화율

0.8, 0.9를 활용)을 넘는 경우 분석에서 제외해 보았다. 그러나 이렇게 자의적인 기준으로 피쳐를 제거하면 오히려 모델의 전반적인 예측 성능이 저하되는 것을 확인할 수 있었다. 따라서 이 논문에서는 가능한 많은 피쳐를 포함시키되 다양한 규제와 하이퍼파라미터 튜닝을 통해 과다적합 등의 문제를 최소화하는 접근방법을 선택하였다.

3 원-핫-인코딩은 범주형 변수의 각 범주로부터 이진지표(binary indicator)를 생성하는 것이다. 21대 총선 설문 데이터의 경우, 결측치(NA)가 있는 문항들에 대해서는 결측치도 하나의 범주로 처리하여 원-핫-인코딩 하였다. 20대 대선 설문 데이터에는 결측치가 없었다.

4 논문에서 활용한 파이썬 패키지는 다음과 같다:
사이킷런(scikit-learn) 버전 1.3.2 (https://github.com/scikit-learn/scikit-learn),
XG부스트(xgboost) 버전 1.7.2 (https://github.com/dmlc/xgboost),
텐서플로(TensorFlow) 버전 2.18.0 (https://github.com/tensorflow/tensorflow).

5 과대적합이란 모형이 훈련데이터에서만 뛰어난 예측성능을 보이고, 테스트데이터에서는 성능이 떨어지는 일반화문제를 말한다.

(gradient) 정보를 활용하기 때문에 이를 그래디언트 부스팅(gradient boosting) 방법이라고 부른다. XG부스트는 랜덤포레스트에 비해 더 정밀하고 복잡한 관계도 잘 처리하는 장점이 있다. 과대적합을 방지하기 위해 랜덤포레스트에서는 트리의 깊이와 계수의 개수를 조정하였으며, XG부스트에서는 트리 분할 파라미터인 감마(gamma)와 L1 및 L2 규제, 행 단위 및 열 단위 샘플링(subsample, colsample_bytree) 등도 추가적으로 적용하였다.

마지막으로, 딥뉴럴네트워크는 인공신경망 모델 중 하나로 복잡한 데이터, 특히 이미지와 같은 비정형 데이터를 다루는데 뛰어난 성능을 보인다.[6] 이 논문에서는 2개에서 5개의 사이의 은닉층을 활용하여 예측모델을 훈련시켰으며, 과다적합을 해결하기 위해 드롭아웃(dropout), L1 및 L2 규제, 조기종료(early stopping) 등을 활용하였다.

이 논문에서 설정한 타겟변수는 크게 1. 투표 참여(기권/참여), 2. 투표 참여(기권/사전/당일), 3. 투표 선택의 세 가지이다. 투표 참여 모델을 훈련시키기 위해서 전체 데이터에서 무작위로 80%의 데이터를 추출하였는데, 이 때 레이블(label)의 분포가 훈련데이터와 테스트데이터 간에 동일하게 되도록 층화추출을 적용하였다. 투표 선택 모델을 훈련시키기 위해서는 투표에 참여한 사람들의 데이터만을 활용할 수밖에 없다. 기권자 데이터에는 레이블이 없기 때문이다. 따라서, 투표선택 모델에서는 투표자 데이터의 80%를 무작위로 추출하여 훈련에 사용하였으며, 훈련데이터와 테스트데이터 간에 레이블의 분포가 동일하도록 층화추출을 적용하였다. 21대 총선 투표선택 모델에서는 지역구와 비례 모두 총 6개 범주(지역구는 더불어민주당/미래통합당/민생당/정의당/기타정당/무소속, 비례는 민주당계열/미래한국당/국민의당/정의당/민생당/기타정당)를, 20대 대선에서는 총 4개의 범주(이재명/윤석열/심상정/기타후보)를 예측하도록 하였다. 모든 모델의 훈련 과정에서 최적의 하이퍼파라미터를 탐색하기 위해 훈련데이터를 5개의 부분집합으로 나누어 교차검증(cross-validation)을 활용했다.[7]

6 딥뉴럴네트워크 모델은 2012년 이후로 가장 각광받는 머신러닝 모델 중 하나이지만, 정형 데이터(tabular data)에서는 트리기반의 모형보다 성능이 저하되는 경우가 많다 (Grinsztajn 외 2022).

7 교차검증이란 모델의 최적 하이퍼파라미터 선택을 위해서 훈련데이터를 다시 하위 집단으로 나누어 각각의 하위 집단을 한 번씩 검증용(validation)으로 나머지를 학습용으로 사용하는 과정을 반복하는 기법이다. 교차검증을 통해 주어진 하이퍼파라미터 조합의 일반화 성능을 평가하여, 여러 조합의 하이퍼파라미터 중 가장 좋은 성능을 내는 조합을 찾을 수 있다.
이 논문에서는 교차검증을 위해 훈련데이터를 5개로 나누었다. 훈련데이터를 너무 작은 사이즈로 나누게 되면 타겟 변수의 특정 클래스 데이터포인트가 너무 적어지는 문제가 있다. 이 경우 제대로 학습이 이루어지지 않을 수 있기 때문에 모든 클래스에 일정 수 이상의 데이터가 분포될 수 있도록 5개 집단으로 나누고 각 집단마다 클래스 비율이 동일하게 유지되도록 하였다.

Ⅳ. 분석결과

1. 모델 선택

머신러닝 예측 모델을 훈련시킨 후에 테스트데이터를 활용하여 그 성능을 평가하고, 그 중 가장 뛰어난 모델을 선정하여 최종 분석에 활용하였다. 훈련 데이터에서는 높은 성능을 보이더라도 테스트데이터에서의 성능이 낮은 경우에는 과다적합으로 인해 일반성이 떨어지는 모델이 되기 때문이다.

아래 <표 1>은 각 타겟별로 예측 모델의 평가지표를 보여준다. 정확도(accuracy)는 전체 예측 중에서 올바르게 예측된 비율을 나타낸다. 재현율(recall)은 각 클래스 별로 해당 클래스에 실제로 속하는 데이터포인트 중 몇 개나 제대로 예측해 내었는지를 의미한다. 정밀도(precision)는 각 클래스 별로 해당 클래스로 예측된 것 중 실제로 그 클래스에 해당하는 비율을 의미한다. F1 점수(F1-score)는 각 클래스 별로 정밀도와 재현도의 조화평균을 구한 것이다.[8] 재현율, 정밀도, F1 점수는 타겟의 각 클래스 별 값을 제시하는 것이 바람직하지만, 타겟의 클래스 수가 너무 많기 때문에 편의상 각 클래스별 자료 수의 비중으로 가중평균한 값을 제시하였다.[9] 전반적으로 볼 때, 두 선거 모두에서 투표참여(기권/참여)의 경우 재현율은 참여 클래스에서 더 높고, 정밀도와 F1 점수는 기권 클래스에서 더 높은 양상을 보인다. 투표참여(기권/사전/당일)의 경우에도 두 선거 모두에서 정밀도는 기권 클래스에서 가장 높지만, 사전투표와 당일투표 클래스와의 차이는 그리 크지 않다. 반면 재현율이나 F1 점수의 경우 21대 총선에서는 기권 클래스가 나머지 두 클래스보다 더 높지만, 20대 대선에서는 사전/당일 클래스가 기권 클래스보다 약간 더 높으며, 사전/당일 클래스 간에는 거의 차이가 없다. 투표선택 모델들의 경우, 모든 선거 모든 모형에서 거대 양당 클래스의 지표는 비슷한 수준의 매우 큰 값을 갖고 나머지 클래스의 지표는 대체적으로 작은 값을 갖는다. 특기할만한 점은, 21대 총선 비례대표 선거의 경우, 정의당 클래스의 정밀도는 거대 양당과 큰 차이가 없으며, 재현율도 거대 양당만큼은 아니지만 60%를 상회하는 좋은 분류 성능을 보여준다는 점이다.

<표 1>을 볼 때, 두 선거 모두에서 투표선택 모형은 매우 높은 성능을 보여주지만, 투표 참여에 대한 모형은 그보다는 낮은 수준의 성능을 보인다.[10] 특히 투표 참여를 기권/참여의 두 분류로 예측하였을

8 각 지표를 수식으로 나타내면 다음과 같다. 각 클래스에 대해 해당 클래스로 예측된 경우를 '양성(Positive)', 해당 클래스가 아니라고 예측된 경우를 '음성(Negative)'라고 부르면, 총 네 가지의 경우가 있을 수 있다. 실제로 양성인데 양성으로 예측된 경우(TP: True Positive), 음성인데 양성으로 잘못 예측된 경우(FP: False Positive), 양성인데 음성으로 잘못 예측된 경우(False Negative), 음성인데 음성으로 예측된 경우(TN: True Negative)이다. 이 때, 각 지표는 아래와 같이 계산할 수 있다:
정확도$=\frac{TP+TN}{TP+TN+FP+FN}$, 재현율$=\frac{TP}{TP+FN}$, 정밀도$=\frac{TP}{TP+FP}$, F1점수$=2\times\frac{\text{정밀도}\times\text{재현율}}{\text{정밀도}+\text{재현율}}$.

9 재현율을 가중평균하면 정확도와 같은 값을 갖는다.

10 투표 선택의 경우 예측범주를 거대 양당 및 그 후보로만 한정하면 모든 모델의 성능이 크게 올라간다(가령, 정확도의 경우 21대 총선은 89.2%-95.4%, 20대선은 95.8%-97.4%). 그러나 범주의 수를 양당으로 제한하게 되면, 기권자의 투표선택 예측에 있어서 거대 양당의 예상 득표율이 과장될 수 있다. 만일 기권자가 실제로 군소 정당에 투표한 사람들과 비슷한 특징을 가지고 있는 경우, 이들이 투표에 참여한다면 군소 정당에 투표할 것으로 예측되어야 타당할 것이다. 애초에 레이블의 범주를 양당으로

때는 85% 이상의 정확도를 보이며 다른 지표의 값도 85%를 상회하지만, 기권/사전/당일의 세 분류로 예측하였을 때는 모든 지표의 값이 60% 내외이다.

| 표 1 | 머신러닝 예측 모델의 평가지표

	21대 총선				
	평가지표	소프트맥스 회귀	랜덤포레스트	XG부스트	딥뉴럴네트워크
투표참여 (기권/참여)	정확도/재현율	0.829	0.833	**0.854**	0.812
	정밀도	0.829	0.833	**0.853**	0.81
	F1 점수	0.829	0.828	**0.853**	0.807
투표참여 (기권/사전/당일)	정확도/재현율	0.583	0.604	**0.621**	0.563
	정밀도	0.578	0.602	**0.609**	0.574
	F1 점수	0.579	0.597	**0.609**	0.562
투표선택 (지역구)	정확도/재현율	0.834	0.821	**0.841**	0.821
	정밀도	**0.789**	0.736	0.756	0.736
	F1 점수	**0.81**	0.775	0.795	0.776
투표선택 (비례)	정확도/재현율	**0.729**	0.652	0.708	0.632
	정밀도	**0.748**	0.6	0.675	0.542
	F1 점수	**0.712**	0.588	0.679	0.561
	20대 대선				
	평가지표	소프트맥스 회귀	랜덤포레스트	XG부스트	딥뉴럴네트워크
투표참여 (기권/참여)	정확도/재현율	**0.892**	0.864	0.864	0.876
	정밀도	**0.885**	0.747	0.825	0.858
	F1 점수	**0.867**	0.801	0.826	0.839
투표참여 (기권/사전/당일)	정확도/재현율	**0.6**	0.528	0.576	0.484
	정밀도	**0.608**	0.601	0.59	0.558
	F1 점수	**0.597**	0.491	0.571	0.444
투표선택	정확도/재현율	**0.925**	0.91	0.91	0.905
	정밀도	**0.923**	0.86	0.86	0.855
	F1 점수	**0.908**	0.884	0.884	0.879

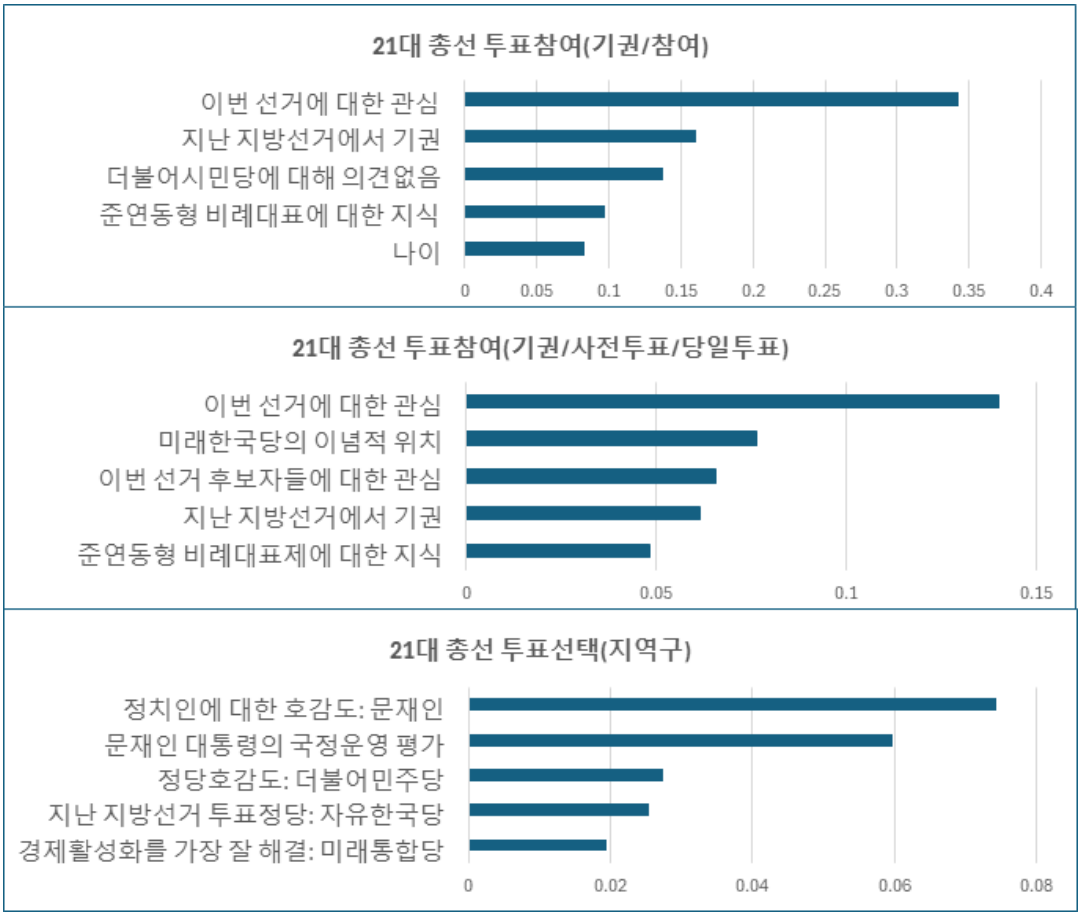
기권/사전/당일투표의 낮은 예측 성능은 기존 연구에서 발견했듯이 사전투표자와 당일 투표자가 인구통계학적으로나 사회경제적 지위, 정치적 성향에서도 크게 구분되지 않는 특징을 지니고 있다는 점에 기인한다(가상준 2016, 2018; 강신구 2016). 보다 최근의 연구들은 사전투표자가 정치이념이나 정부에 대한 평가 측면에서 당일 투표자와 구분된다고 주장하지만, 이 연구들에서도 한 두 개의 독립변수를 제외한 나머지 독립변수들에서는 유의미한 차이가 나타나지 않는다(가상준 2021; 이재묵 2020). 기존 연구에서는 계수 추정에 사용된 데이터 (즉, 훈련데이터) 내에서도 분류 정확도가 76.5%에 머무르는 것을 볼 때 (강신구 2016), 테스트데이터에서 60% 정도의 정확도는 나쁘지 않은 수준이다.[11] 다만,

제한하는 경우에는 이런 기권자도 양당 중 한 당을 선택할 것으로 예측할 수밖에 없기 때문에 결과에 왜곡이 생긴다. 다행히 범주를 양당으로 제한하지 않은 경우에도 투표선택 모델의 정확도는 상당히 높은 편이기 때문에 다중 범주 모델을 선택하였다.

이후의 분석에서 사전투표자와 당일투표자의 분류가 기권자와 투표자의 분류보다는 정확도가 떨어진다는 점에 유의할 필요가 있다.

<표 1>에 나타난 예측성능을 기준으로 21대 총선의 투표 참여와 지역구 투표선택 분석에는 XG부스트를, 그리고 21대 총선의 비례대표 투표 선택과 20대 대선 분석에는 소프트맥스 회귀를 선택하였다. 이후의 분석에서는 훈련이 끝난 모델을 전체데이터에 적용하여 투표 참여 및 선택 확률을 예측하였다. 투표선택에 대한 분석 시에, 일부 응답자는 투표에 참여했다고 응답했으나 투표 선택에 대해 응답하지 않은 경우가 있다. 이러한 응답자들은 투표에 참여한 것으로 간주하되, 투표 선택에 대해서는 모델의 예측값으로 대체하여 분석한다. 21대 총선의 경우 투표자 783명 중 63명이 어느 정당이나 후보자에게 투표했는지 밝히지 않았다. 20대 대선의 경우 투표에 참여했다고 응답한 1,078명 중 80명이 누구에게 투표했는지 밝히지 않았다.

| 그림 1 | XG부스트 피쳐 중요도(Feature Importance)



11 만일 과다적합문제를 고려하지 않고 훈련데이터에서의 예측성능만을 높이고자 한다면 이 논문에서 사용한 대부분의 모델에서 분류 정확도가 99%를 상회한다.

본격적인 분석에 앞서, 머신러닝모델 훈련과정에서 어떤 피쳐들이 예측 향상에 가장 많이 기여했는지 살펴보자. 먼저, 21대 총선의 투표 참여와 지역구 투표 선택을 분석하기 위해 선정된 XG부스트 모형에서 예측에 가장 많이 기여한 5개의 피쳐를 <그림 1>로 정리하였다.

투표 참여에 대한 예측에서 가장 중요했던 피쳐는 '이번 선거에 대한 관심'으로 나타났다. 이는 기존 연구에서도 잘 알려진 투표 참여를 설명하는 주요변수 중 하나이다. 그 밖에도 나이, 이전 선거 참여 등 통상적으로 투표참여를 잘 설명하는 것으로 알려진 피쳐들이 포함되었다. 특기할만한 점은, 21대 총선에서 새롭게 등장한 준연동형 비례대표제에 대한 지식, 그리고 그에 따른 위성정당들에 대한 의견이 주요 피쳐로 등장했다는 점이다. 이는 정치지식이 투표참여에 중요한 역할을 한다는 기존 연구의 주장과 부합하는 결과이다 (Palfrey and Poole, 1987; Moon, 2011).

지역구의 투표 선택에 대한 예측에서는 문재인 대통령과 집권당인 더불어민주당에 대한 호감도 및 평가가 중요한 피쳐로 나타났다. 흥미로운 점은, 주요 야당인 미래통합당에 대한 호감도는 포함되지 않은 대신, 특정 정책분야에서의 미래통합당 선호가 중요한 것으로 나타난다는 점이다. 미래통합당에 대한 호감도보다는 경제활성화 분야에서 미래통합당이 가장 낫다는 선택을 했는지의 여부가 유권자의 선택에 더 중요한 영향을 미친 것으로 풀이된다.

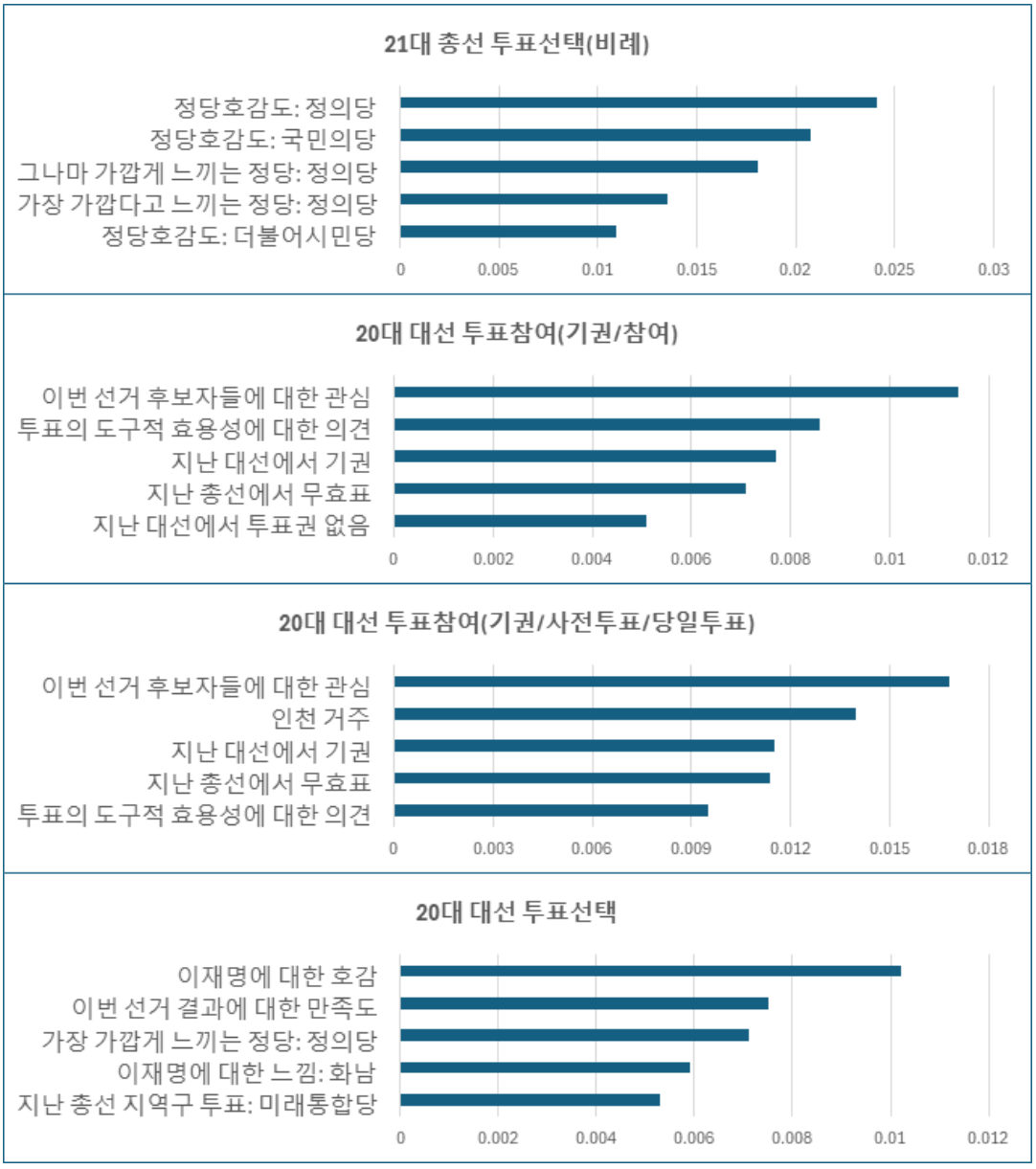
21대 총선 비례대표 선택 및 20대 대선을 분석하는 모델로 선정된 소프트맥스 회귀에서는 순열 중요도(permutation importance)를 살펴보았다. 결정트리 기반 모형에서는 예측에 기여한 정도를 통해 피쳐 중요도를 쉽게 계산할 수 있지만, 소프트맥스 회귀 모형에서는 같은 방식의 계산을 할 수 없다. 그를 대신하여 널리 쓰이는 피쳐 평가 방법 중 하나가 순열 중요도이다. 순열 중요도에서는 피쳐의 값을 무작위로 섞어서 모델 성능에 미치는 영향을 계산한다. 만약 해당 피쳐의 값을 섞었을 때 모델의 성능에 큰 영향이 있다면, 그 피쳐는 중요한 것으로 평가할 수 있다.

<그림 2>를 살펴보면 21대 총선 비례대표 선택 예측에서 중요한 역할을 한 변수들은 주로 정의당과 관련된 변수들, 그리고 국민의 당과 더불어 시민당에 대한 호감도로 나타났다. 앞서 언급하였듯이, 투표 선택 예측 결과를 보면 거대 양당의 선택은 예측 성능이 좋지만 다른 정당들을 선택한 경우에는 상대적으로 예측이 어려웠는데 정의당만이 그 예외에 해당된다. 지역구 투표선택에서는 거대 양당을 지지했던 유권자들 중에 비례대표에서는 군소 정당을 지지하는 경우가 있는데, 이들을 구분해 내는 데에 정의당에 대한 문항들이 중요한 역할을 한 것으로 풀이된다.

20대 대선 투표 참여에서 중요한 피쳐로 선정된 것은 해당 선거 후보자들에 대한 관심과 과거 투표 이력이다. 대통령 선거의 특징 상, 선거 자체보다는 후보자들에 대한 관심이 더 높은 중요도를 갖는 것으로 보인다. 그 밖에 흥미로운 피쳐는 바로 '투표의 도구적 효용성에 대한 의견'이다. 이 변수는 "만약 내가 지지하는 후보(또는 정당)가 선거에서 이길 확률이 없다면, 내가 투표하는 것은 별로 의미가 없다."에 대해 얼마나 공감하는지를 나타낸 것이다. 투표참여 자체에서 의미를 찾기보다는 투표를 통해 달성할 수 있는 효용으로 투표행위를 평가한다는 점에서 도구적 효용성에 대한 의견이라고 볼 수 있다. 20대 대선에서 선거 전까지 이루어진 여론 조사 중 과반이 넘는 수에서 윤석열 후보의 우세를

예측했다는 점을 생각하면(중앙선거관리위원회 2022), 이재명 후보 지지자들의 투표동기가 저해됐을 가능성을 점쳐볼 수 있다.

| 그림 2 | 소프트맥스 회귀 순열 중요도 (Permutation importance)



2. 기권한 유권자의 투표선택 예측 결과

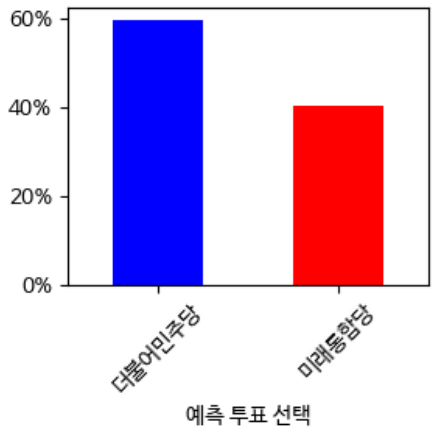
기권한 유권자가 만일 투표에 참여했다면 어느 정당이나 후보에 투표했을까? 기권자의 정당 또는 후보자 선호는 투표자의 선호와 큰 차이가 있을까? 기권한 유권자가 투표에 참여한 유권자와 얼마나 비슷한 선호를 가지고 있는지를 분석함으로써, 투표율과 선거결과의 관계, 그리고 유권자의 투표참여 동인에 대한 이해를 높일 수 있다.

만일 기권자의 정당 및 후보자 선호가 투표자와 비슷하다면, 기권자가 투표에 참여했다고 하더라도 선거 결과에는 큰 영향이 없었을 수 있다. 이 경우, 투표율이 민주적 반응성 및 대표성과 직접적인 관계가 없으며, 투표율이 낮다고 하더라도 최적의 선거결과를 달성하는 데에는 문제가 없을 수 있다 (Feddersen and Pesendorfer 1996; McMurray 2013).

반대로, 만일 기권자의 선호가 투표자의 선호와 상당히 다르다면, 기권자를 투표장으로 불러내는 것이 선거 결과에 큰 영향을 미칠 수도 있다. 특히, 선거가 접전이라면 투표율에 따라 선거의 승자가 뒤집힐 수도 있다는 의미가 된다. 한편, 선거가 접전이 아닌 경우라면, 다수 유권자들과 선호가 상당히 다르다는 사실 자체가 기권자들의 선거 참여 동기를 저해했을 수 있다. 즉, 다른 유권자들이 선호하지 않는 정당이나 후보를 선호하는 경우, 자신이 선호하는 정당이나 후보가 승산이 없다고 판단했기 때문에 투표에 참여하지 않았을 수 있다는 설명이 가능하다.

아래 <그림 3>은 21대 총선 기권자의 투표선택을 XG부스트 모델로 예측한 결과이다. 21대 총선 설문조사에서 기권자는 총 417명으로 전체 설문대상자 1200명 중 34.75%에 해당한다. 21대 총선에서 지역구 후보를 낸 주요 정당은 더불어민주당, 미래통합당, 민생당, 정의당이다. <그림 3>에 따르면, 기권자는 그 중 더불어민주당 후보자에게 59.5%, 그리고 미래통합당 후보자에게 40.5%가 투표했을 것이라고 예측되었다.

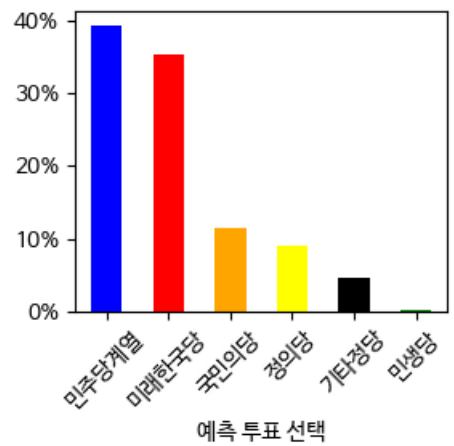
| 그림 3 | 21대 총선 기권자의 예측 투표 선택(지역구)



이러한 예측 결과는 기권자의 선호가 투표참여자의 선호와 크게 다르지는 않다는 것을 보여준다. 투표 참여자의 경우, 53.6%가 더불어민주당, 36.5%가 미래통합당, 그리고 나머지가 정의당과 국민의 당에 약 4%씩 투표한 것으로 설문문항에 응답하였다. 기권자에서 투표자보다 주요 양당에 대한 선호가 더 크게 나타나지만, 방향성이나 상대적인 격차가 매우 다르게 나타나는 것은 아니다.

다음으로, 21대 총선 기권자의 비례대표 선거 투표예측 결과를 살펴보자. <그림 4>에 따르면, 기권자의 39.3%가 더불어시민당과 열린민주당에, 35.3%가 미래한국당에, 11.5%가 국민의 당에, 그리고 나머지 정당에 10% 미만이 투표하는 것으로 예측되었다. 투표자의 경우, 36.7%가 더불어시민당과 열린민주당에, 30.4%가 미래한국당에, 13.2%가 정의당에, 11.7%가 국민의 당에 투표한 것으로 응답하였다. 이로 볼 때, 비례대표 선거의 경우에는 국민의 당과 정의당 간에 더 많은 표를 받은 정당 순위가 달라지기는 하지만, 그 차이가 2% 정도에 불과하여 역시 기권자와 투표자 간에 큰 차이를 보였다고 하기는 어렵다.

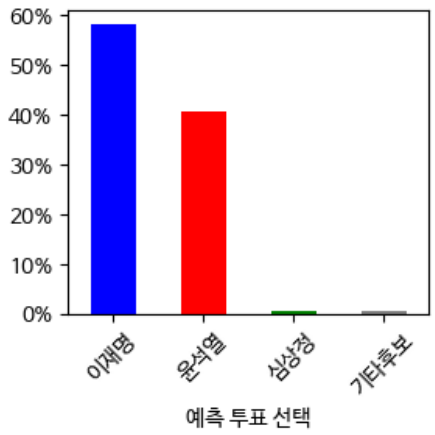
| 그림 4 | 21대 총선 기권자의 예측 투표 선택(비례)



(참고: 민주당 계열에는 더불어시민당과 열린민주당이, 기타정당에는 우리공화당, 민중당,한국경제당, 친박신당 등이 포함됨)

이에 반해, 20대 대선의 경우에는 기권자와 투표자의 선호에 의미 있는 차이가 나타난다. 20대 대선 유권자 조사 응답자 중 기권자는 172명으로 설문대상자 1250명 중 13.76%에 해당한다. <그림 5>에서 볼 수 있듯이, 기권자의 58.1%는 이재명 후보, 40.7%는 윤석열 후보, 0.6%는 심상정 후보를 선택하는 것으로 예측되었다. 그러나 투표자의 경우는 이재명 후보 47.5%, 윤석열 후보 47.5%로 동률을 이루고, 그 뒤를 이어 심상정 후보 4.3%의 선택을 한 것으로 나타났다. 유권자와 기권자의 투표선택이 크게 차이가 나고, 동시에 초박빙의 접전으로 치러진 20대 대선과 같은 경우에는 투표율 변화가 선거결과에 민감한 영향을 줄 수밖에 없다.

| 그림 5 | 20대 대선 기권자의 예측 투표 선택



종합적으로 살펴볼 때, 기권자와 유권자의 투표선택은 선거의 특성에 따라서 크게 달라진다는 점을 알 수 있다. 21대 총선에서는 두 집단 간 선택에 큰 차이가 없었던 반면, 불과 2년 뒤에 치러진 20대 대선에서는 유의미한 차이가 나타났다. 앞서 보았듯이 지난 선거에서의 투표여부가 다음 선거의 참여를 예측하는 데에 큰 역할을 한다는 점에 주목하면, 늘 선거에 참여하는 집단과 그렇지 않은 집단이 이질적으로 존재한다고 생각하기 쉽다. 그러나 기권자의 예측 투표선택 분석을 통해, 선거에 따라 투표자와 기권자 집단의 특성이 달라진다는 점을 확인할 수 있다.

3. 투표율과 정당 득표율 시뮬레이션

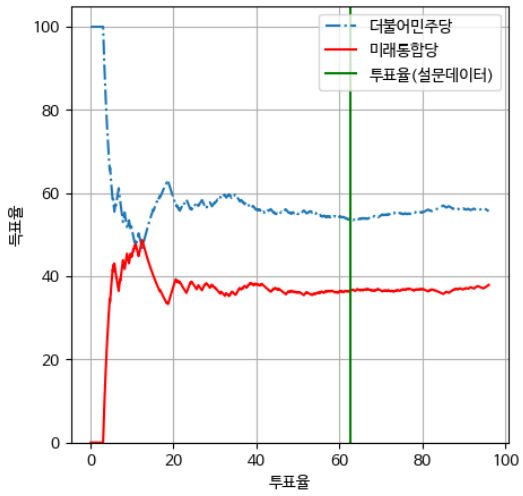
투표율 상승이나 하락이 어떤 정당에 더 유리한지 살펴보기 위해서는, 투표율이 상승하거나 하락할 때 어떤 정당의 지지자가 투표에 추가로 참가하거나 기권할지 알아야 한다. 예를 들어, 두 정당 A와 B가 있다고 하자. 만일 투표율이 하락할 때 A 정당 지지자가 B 정당 지지자에 비해 더 이탈하기 쉬운 특징을 지니고 있다면, 투표율 하락은 A 정당에게 더 불리할 가능성이 높은 것이다. 이와 같은 분석을 하기 위해 머신러닝 예측모델로부터 얻은 투표참여확률(기권/참여)과 투표선택확률을 활용하여 다음과 같은 시뮬레이션을 실시하였다.

(1) 투표에 참여했다고 응답한 사람들을 가장 투표참여확률이 높은 사람부터 차례로 줄 세운다.
(2) 그 뒤를 이어, 투표에 참여하지 않았다고 응답한 사람들을 가장 투표참여확률이 높은 사람부터 이어서 줄 세운다.
(3) 줄 세운 순서대로 투표에 참여했다고 가정하고 투표율 변화에 따른 정당득표율을 계산한다. 이때 투표에 참여했다고 응답한 사람들은 설문지 응답의 투표선택을 기준으로, 기권했다고 응답한

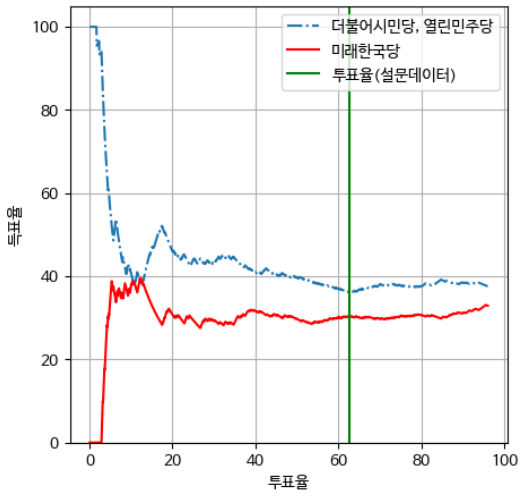
사람들은 예측모델의 투표선택 결과를 기준으로 투표에 임했다고 가정한다.

투표에 기권했다고 응답한 사람들 중에도 예측모델에서는 매우 높은 투표확률을 가진 것으로 나타나는 경우가 있다. 21대 총선 기권자 중에서 가장 높은 값을 가진 사람의 투표확률은 0.92, 20대 대선 기권자 중에서는 0.96으로 예측되었다. 투표에 참여했다고 응답한 사람 중에 이보다 낮은 투표참여확률을 가진 경우도 많지만, 설문응답자의 응답을 진실한 것으로 간주하고 시뮬레이션에 반영했다.

| 그림 6 | 21대 총선 총투표율과 정당득표율(지역구)



| 그림 7 | 21대 총선 총투표율과 정당득표율(비례)



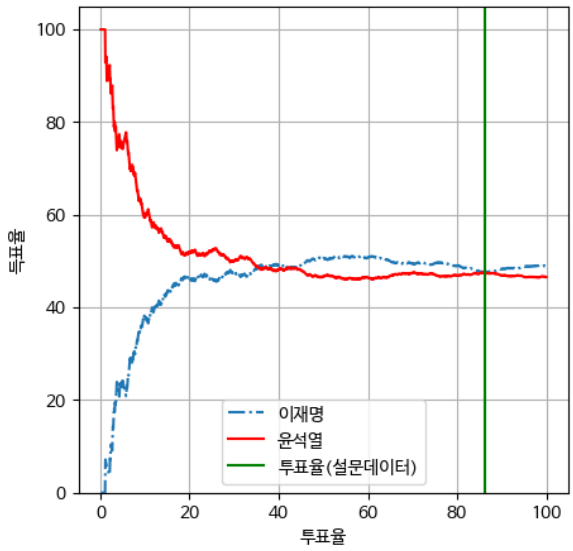
먼저, 21대 총선의 총투표율과 정당득표율을 살펴보자. <그림 6>에는 지역구 선거, <그림 7>에는 비례대표 선거의 시뮬레이션 결과가 나타나 있다. 그림에서 초록색 실선으로 표시된 부분은 설문에서 투표에 참여했다고 응답한 사람들의 줄이 끝나는 부분이다. 즉 설문데이터 상의 투표율이라고 할 수 있다. 설문데이터 상의 투표율은 실제 투표율 66.2%에 매우 근접한 65.3%이다.

그림으로부터 알 수 있듯이, 21대 총선에서는 투표율 상승이 특정 정당에 특별히 유리하지는 않은 것으로 나타났다. 지역구와 비례 모두 투표율이 더 상승한다고 해도 정당득표율의 격차가 크게 좁혀지거나 벌어지지 않는다. 이는 앞서 분석했던 기권자의 예측투표선택이 투표자의 투표선택과 크게 다르지 않았다는 데에서도 유추해볼 수 있는 결과이다.

투표율 하락의 경우에는 흥미로운 부분이 있다. 첫째로, 가장 높은 투표참여 의사를 지닌 사람들은 보수 정당이 아니라 진보 정당 지지자들이었다는 것이다. 투표율이 0에서 10퍼센트 초반에 이르기까지 더불어민주당의 정당득표율이 미래통합당의 정당 득표율보다 훨씬 높은 수준을 유지하고 있다. 둘째로, 비례대표 선거의 경우, 투표율 하락이 더불어민주당에게 약간 더 유리한 것으로 나타난다. <그림 7>에서 볼 수 있듯이, 투표율이 초록색 실선 왼쪽으로 움직이는 경우, 두 정당간의 격차가 다소 늘어난다는

사실을 확인할 수 있다. 이는 낮은 투표율이 보수 정당에게 더 유리하다는 통념과는 거리가 있는 결과이다. 다만, 지역구 선거의 경우에는 그와 같은 경향성이 나타나지 않는다.

| 그림 8 | 20대 대선 총투표율과 주요 정당 후보 득표율



다음으로, 20대 대선 시뮬레이션 결과를 살펴보자. 20대 대선 설문데이터 상의 투표율은 86.24%로 실제 투표율인 77.1%보다 훨씬 높은 수준으로 나타났다. 다만, 투표에 참여했다고 응답했으나 누구에게 투표했는지 밝히지 않은 80명을 기권자로 간주하면 실제에 근접한 79.84%의 투표율을 얻을 수 있다. 그러나 설문 응답자의 응답을 진실한 것으로 간주하고 86.24%의 투표율을 기준으로 시뮬레이션을 실시하였다.

<그림 8>은 21대 총선과는 매우 다른 결과를 보여준다. 첫째, 21대 총선에서와는 달리, 가장 높은 투표참여의사를 가진 사람들은 진보정당이 아닌 보수정당 후보 지지자들이었다. 투표율이 30퍼센트 중반을 넘어갈 때 까지도 윤석열 후보의 득표율이 더 높다는 것을 알 수 있다. 둘째, 투표율이 현재보다 더 낮아졌다면, 이재명 후보의 득표율이 윤석열 후보의 득표율보다 더 높아졌을 것이다. 그러한 우위는 투표율이 40%에 근접하도록 내려갈 때까지 유지되지만, 그 이하의 투표율에서는 다시 윤석열 후보가 우위에 있는 것으로 나타난다. 셋째, 앞서 살펴본 기권자의 예측투표선택에서 유추해 볼 수 있듯이, 투표율이 현재보다 더 높아졌을 경우에도 이재명 후보의 득표율이 더 높아졌을 것으로 예측되었다.

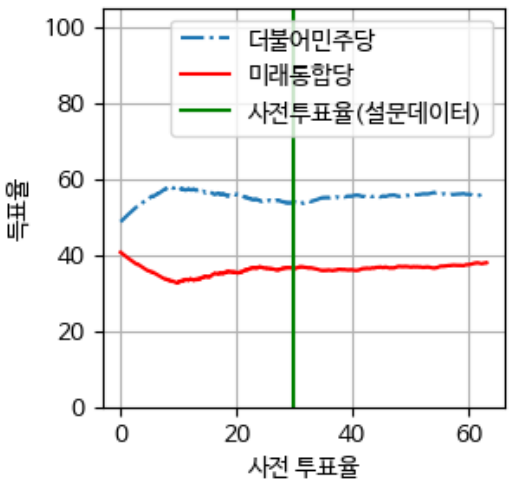
투표참여(기권/사전투표/당일투표) 예측 모델을 활용하여 사전투표와 당일투표에 대해서도 비슷한 시뮬레이션을 실시할 수 있다. 만일 선거 당일에 투표한 사람들의 투표선택은 그대로 두고, 사전투표율만 변화한다면 선거 결과가 어떻게 변화할까? 이를 알아보기 위해 사전투표 시뮬레이션을 다음과 같이 실시하였다.

(1) 투표에 참여했다고 응답한 사람들 중 당일에 투표했다고 응답한 사람들의 투표 선택은 주어진 것으로 본다. 즉 양 당의 득표율은 당일 투표에서 주어진 차이를 기준으로 출발한다.

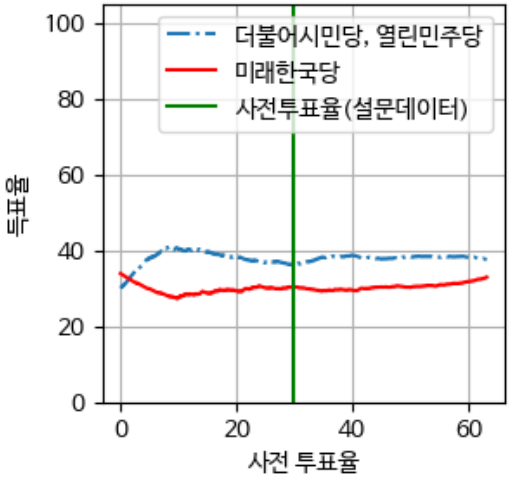
(2) 투표에 참여했다고 응답한 사람들 중 사전투표에 참여했다고 응답한 사람들을 사전투표확률이 가장 높은 사람부터 차례로 줄을 세운다. 그 뒤를 이어, 투표에 참여하지 않았다고 응답한 사람들을 가장 사전투표참여확률이 높은 사람부터 이어서 줄 세운다.

(3) 줄 세운 순서대로 투표에 참여했다고 가정하고 사전투표율 변화에 따른 정당득표율을 계산한다. 이 때 사전투표에 참여했다고 응답한 사람들은 설문지 응답의 투표선택을 기준으로, 기권했다고 응답한 사람들은 예측모델의 투표선택 결과를 기준으로 투표에 임했다고 가정한다.

| 그림 9 | 21대 총선 사전투표율과 정당 득표율(지역구)



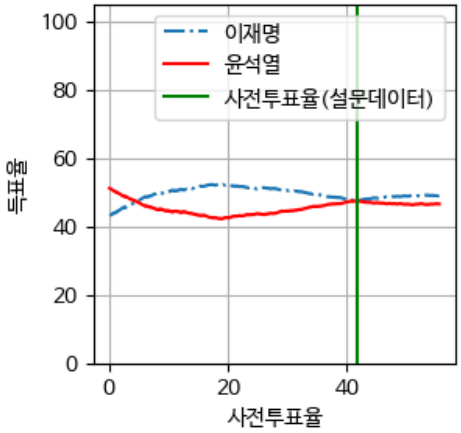
| 그림 10 | 21대 총선 사전투표율과 정당득표율(비례)



21대 총선의 사전투표율 시뮬레이션의 결과는 <그림 9>와 <그림 10>에 나타나 있다. <그림 9>의 지역구 선거를 살펴보면, 당일투표에서부터 더불어민주당이 더 우세하게 출발하여 사전투표율 상승 초반에 정당 간 득표율 격차가 크게 벌어진다는 것을 알 수 있다. 즉, 사전투표율이 0이었다고 하더라도 지역구 선거에서는 더불어 민주당이 더 높은 정당 득표율을 보였을 것이라는 의미이다. 득표율 격차는 사전투표율이 10프로를 넘어가면서부터는 약간 줄어드는 경향을 보이며, 사전투표율이 설문 데이터상에서보다 더 높아지는 경우에는 별다른 변화를 보이지 않는다. <그림 10>의 비례대표 선거에서는 미래한국당이 민주당 계열 위성 정당보다 약간 더 높은 득표율로 출발한다. 그러나 지역구 선거에서와 마찬가지로 사전투표율 상승 초반에 우세가 뒤집히고 득표율 격차가 크게 벌어진다. 이후 사전투표율이 10프로를 넘어서 상승함에 따라 정당 간 격차가 약간 줄어드는 경향이 보이며, 기권자들이 투표에 참가하면서부터는 다시 격차가 약간 증가하는 양상을 보인다.

20대 대선에 대한 사전투표 시뮬레이션 결과는 <그림 11>에 제시되어 있다. 당일 투표에서는 윤석열 후보가 약 7% 정도의 우위를 보인다. 그러나 21대 총선에서와 비슷하게, 사전투표 상승 초반부에 그러한 우세가 뒤집히고 이후 정당 격차가 점차 증가하게 된다. 그러나 사전투표율이 계속 증가하여 약18%를 넘어가게 되면 다시 정당 간 득표율 격차가 점차 줄어들기 시작하여 설문데이터상의 사전투표율에 도달하면 두 후보의 득표율이 동률을 이룬다. 만일 사전투표가 현재보다 더 증가한다면 이재명 후보의 득표율이 상대적으로 더 많이 증가하는 것으로 예측된다.

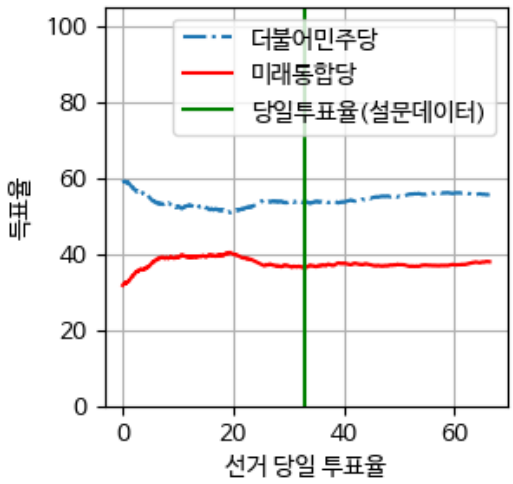
| 그림 11 | 20대 대선 사전투표율과 주요 정당 후보 득표율



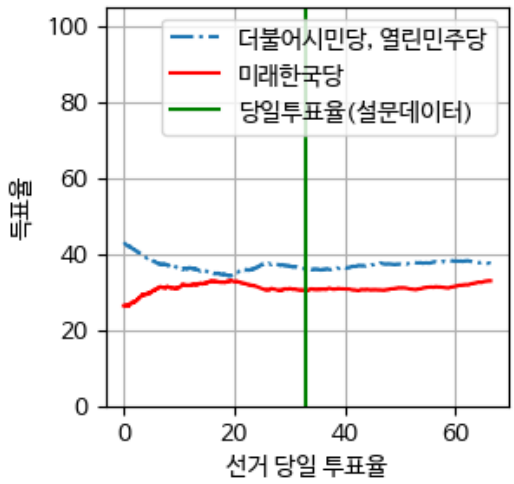
마지막으로, 당일 투표에 대해서도 아래와 같이 비슷한 방식의 시뮬레이션 분석을 실시하였다.

(1) 투표에 참여했다고 응답한 사람들 중 사전에 투표했다고 응답한 사람들의 투표 선택은 주어진 것으로 본다. 즉 양당의 득표율은 사전투표에서 주어진 차이를 기준으로 출발한다.
(2) 투표에 참여했다고 응답한 사람들 중 선거 당일투표에 참여했다고 응답한 사람들을 당일투표확률이 가장 높은 사람부터 차례로 줄을 세운다. 그 뒤를 이어, 투표에 참여하지 않았다고 응답한 사람들을 가장 당일투표참여확률이 높은 사람부터 이어서 줄 세운다.
(3) 줄 세운 순서대로 투표에 참여했다고 가정하고 당일투표율 변화에 따른 정당득표율을 계산한다. 이 때 당일투표에 참여했다고 응답한 사람들은 설문지 응답의 투표선택을 기준으로, 기권했다고 응답한 사람들은 예측모델의 투표선택 결과를 기준으로 투표에 임했다고 가정한다.

| 그림 12 | 21대 총선 선거 당일 투표율과 정당 득표율(지역구)

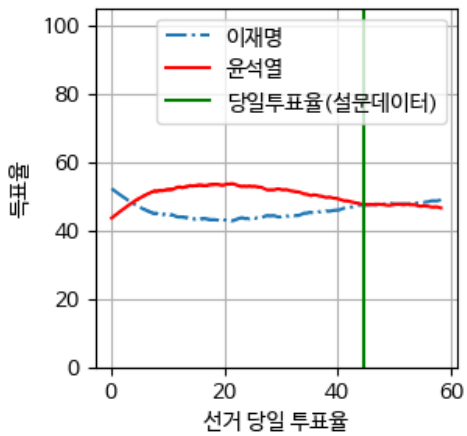


| 그림 13 | 21대 총선 선거 당일 투표율과 정당 득표율(비례)



21대 총선의 당일투표 시뮬레이션 결과는 <그림 12>와 <그림 13>에 제시되어 있다. 지역구와 비례 선거 모두에서 사전투표의 결과로 더불어민주당이 우위로 출발한다. 선거 당일 투표율 증가 초반에 이 격차가 줄어들지만 우세가 뒤집힐 정도로 줄어들지는 않는다. 즉, 당일투표 참여의사가 높은 사람들은 미래통합당 지지자인 경우가 더 많지만, 그 정도가 사전투표에서의 열위를 뒤집을 정도는 아니다. 당일 투표율이 20퍼센트를 넘어가면서 정당 간 격차가 다시 증가하지만, 이후 일정한 수준을 유지하게 된다. 당일 투표율의 경우는 현재보다 더 높아지는 경우에도 정당 간 득표율 격차에 큰 변화는 없는 것으로 나타난다.

| 그림 14 | 20대 대선 선거 당일 투표율과 주요 정당 후보 득표율



마지막으로 20대 대선의 당일투표율 시뮬레이션 결과를 살펴보자. 사전투표에서 이재명 후보의 득표율이 약 9.6% 더 높게 출발하지만, 이 격차는 당일 투표율 상승 초반부에 뒤집힌다. 이후 득표율 격차가 꾸준히 증가하지만 당일 투표율이 20%를 넘어가게 되면 다시 줄어드는 양상을 보인다. 흥미로운 것은, 당일 투표율이 얼마간 더 증가한다고 하더라도 특정후보의 득표율이 상대적으로 더 많이 증가하지는 않는다는 것이다.

투표율과 정당득표율 시뮬레이션 결과를 종합적으로 살펴볼 때, 투표율, 사전투표율, 당일투표율의 변화가 특정 정당에 어떤 식으로 유리한지는 일률적으로 특정할 수 없다는 결론을 내릴 수 있다. 특히, 총 투표에서는 투표방법과 상관없이 가장 높은 투표참여 의사를 가지는 유권자들의 성향이 선거에 따라 달라진다는 것을 알 수 있다. 반면, 사전투표에 가장 높은 참여 확률을 보이는 사람들은 두 선거 모두에서 더불어민주당 지지자들이며, 반대로 당일투표에 가장 높은 참여확률을 보이는 사람들은 두 선거 모두에서 보수정당 지지자들이라는 것을 확인할 수 있다.

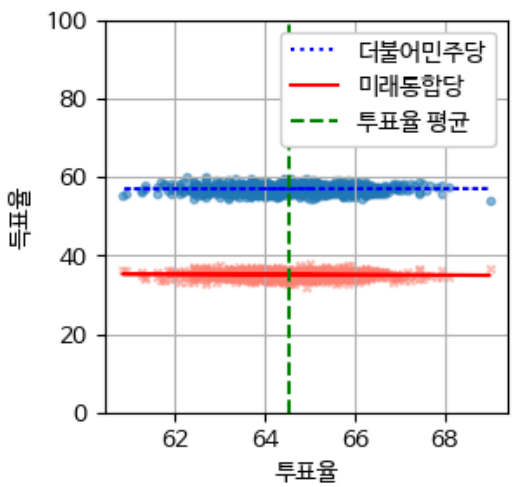
사전투표나 당일 투표에서 나타난 이러한 경향 때문에 마치 특정 투표 방법이 특정 정당에 더 유리하다는 인상을 주기 쉬우나, 시뮬레이션 결과는 그러한 판단이 옳지 않다는 점을 시사한다. 위에서 보았듯이, 사전투표나 당일투표 모두 투표율이 일반적이지 않을 만큼 낮은 수준에서는 그러한 판단이 적용될 수 있지만, 어느 수준 이상의 투표율을 넘어가게 되면 그렇지 않다는 것이다. 위에서 보았듯이, 사전투표에 가장 높은 참여 확률을 보이는 사람들은 두 선거 모두에서 더불어민주당 지지자들이지만, 20대 대선의 경우 사전투표율이 실제보다 훨씬 더 낮았다면 오히려 이재명 후보에게 유리했을 것이다. 이러한 결과는 투표율 정보만으로 특정 정당의 우세를 판단하기는 매우 어렵다는 것을 의미한다.

4. 투표율과 정당 득표율에 대한 확률적 투표 참여 시뮬레이션

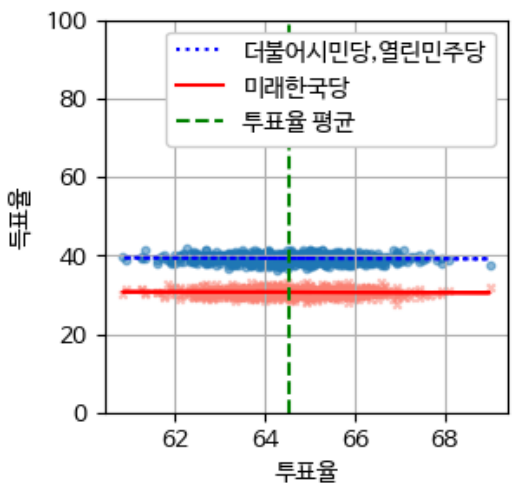
앞서 실시한 시뮬레이션에서는 특정 투표율 하에서 어떤 유권자가 투표에 참여하게 될 지를 투표 참여 확률에 기반하여 판단하고, 이를 바탕으로 정당 득표율을 예측하였다. 그러나 현실에서는 투표 참여 확률이 높은 성향의 유권자가 반드시 항상 투표에 참여하는 것이 아니며, 마찬가지로 투표 참여 확률이 낮은 유권자라도 항상 투표에 불참하는 것은 아니다. 정당의 동원 전략이나, 우발적인 사건, 날씨나 건강 등의 환경, 주변 사람의 권유 등에 따라 유권자의 참여 여부가 달라질 수 있다.

이러한 문제를 고려하기 위한 방법으로 확률적 투표 참여 시뮬레이션을 생각해 볼 수 있다. 즉, 예측된 투표 참여 확률을 기반으로 각 유권자의 투표 참여 여부를 무작위로 추출한 뒤, 해당 유권자들의 투표 예측 결과를 바탕으로 투표율 및 정당 혹은 정당 후보 득표율이 어떤 영향을 받는지 살펴보는 것이다. 이는 기저에 바탕하고 있는 유권자들의 성향에는 변화가 없지만, 현실의 다양한 요소에 의해 약간의 충격(perturbation)이 주어질 때 선거 결과가 달라졌을지를 알아보기 위한 분석이다.

| 그림 15 | 21대 총선 투표율과 정당 득표율 (지역구)



| 그림 16 | 21대 총선 투표율과 정당 득표율 (비례)

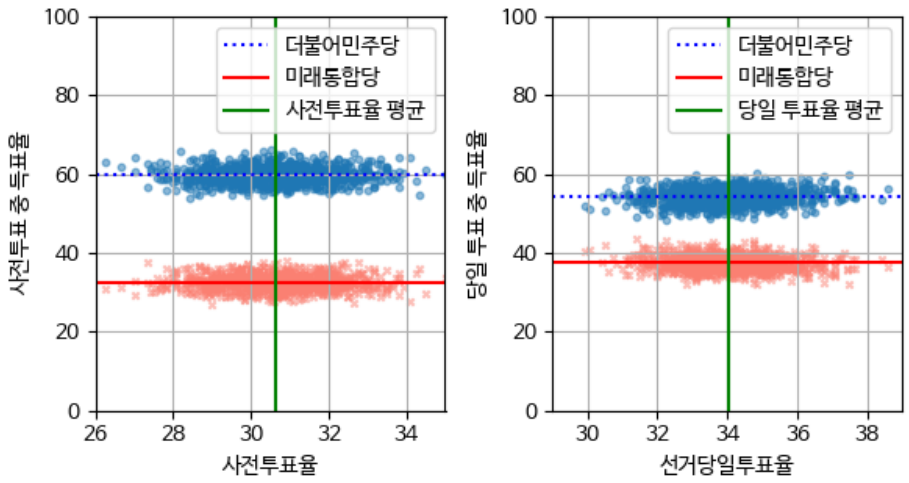


<그림 15>과 <그림 16>은 투표참여(기권/참여) 확률을 기반으로 하여 유권자를 추출하고, 추출된 유권자들의 투표선택을 기반으로 득표율을 계산하는 방식의 시뮬레이션을 1000번 시행한 결과이다. 가령, 어떤 유권자의 기권 확률이 0.2, 참여 확률이 0.8로 예측되었다면, 한번 시뮬레이션을 할 때마다 1이 나올 확률이 0.8인 베르누이 분포로부터 독립적 추출을 시행하여 해당 유권자의 투표 참여 여부를 결정하는 것이다. 이 때, 실제로 투표에 참여했던 사람이 투표에 참여하는 것으로 뽑힌 경우에는 실제 투표 선택을 득표율 계산에 사용하였으며, 투표에 참여하지 않은 사람이 뽑힌 경우에는 투표 선택 예측값으로 대신하였다.

앞서 살펴보았던 바와 마찬가지로, 21대 총선에서는 투표 참여자에 변화가 생기고 이로 인해 투표율이 약간 변화하는 경우에도 양당의 득표율에 눈에 띌만한 변화가 없는 것으로 나타난다. 참여하는 유권자 집단이 얼마간 변화하더라도 그에 따라서 선거 결과에는 영향이 없다는 것을 확인할 수 있다.

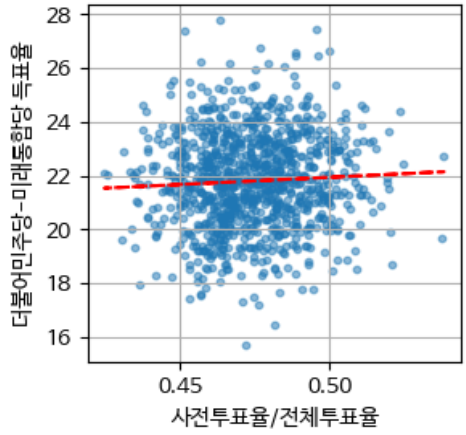
이러한 결론은 투표 참여를 사전 투표와 선거 당일 투표로 나누어 살펴보아도 동일하다. 이 경우에는 만일 기권/사전투표/당일투표의 확률이 0.2, 0.3, 0.5로 예측된 유권자가 있다면, 매 시뮬레이션 시행마다 세 선택 중 하나를 해당 확률로 추출하도록 하여 해당 유권자의 투표 참여 방법을 결정하는 것이다. 아래 <그림 17>에서는 이렇게 결정된 21대 총선의 사전 투표율과 선거 당일 투표율에 따른 정당 득표율의 시뮬레이션 결과를 보여준다. 앞선 시뮬레이션에서와 마찬가지로, 투표율 변화 자체에 따른 득표율 차이는 거의 없지만, 사전투표에서의 정당 득표율의 격차가 선거 당일 투표율에서의 격차보다 평균적으로 훨씬 더 크다는 것을 확인할 수 있다.

| 그림 17 | 21대 총선 사전 투표율 및 선거 당일 투표율과 정당 득표율(지역구)



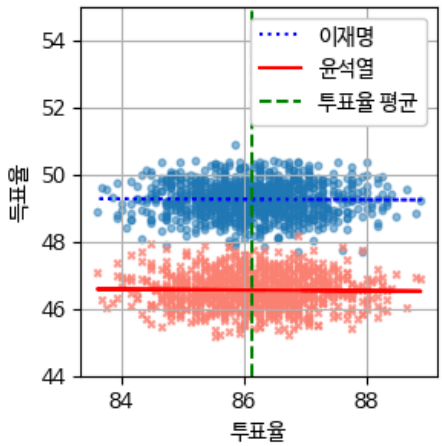
그렇다면, 유권자 집단의 구성 변화로 총 투표 참여 중 사전투표의 비율이 높아질 때에는 평균적으로 더불어민주당의 상대적인 득표율이 더 높아질까? 이를 알아보기 위해 <그림 18>에서는 위에서 행한 1000번의 시뮬레이션에서 각각 총 투표율에서 사전투표율이 차지하는 비중을 계산하고 이에 따라 양당의 득표율 차이가 어떻게 분포되어 있는지를 제시하였다. 그림에서 알 수 있듯이, 사전투표 비중이 올라갈 때 양당의 득표율 격차는 올라가는 추세를 보여주기는 하지만, 그 정도는 매우 미미하다. 더불어민주당 지지자가 사전투표에 더 많이 참여하는 경우 당일 투표에서는 상대적으로 미래통합당 지지자가 더 많았다는 것을 알 수 있다. 비례대표 투표 선택도 지역구 투표선택과 비슷한 결과를 보여주기 때문에 해당 시뮬레이션 결과는 생략하였다.

| 그림 18 | 21대 총선 사전투표 비율과 정당 득표율 차이(지역구)

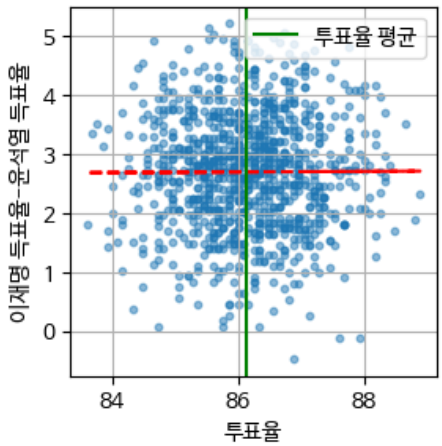


21대 총선과는 달리 박빙이었던 20대 대선의 경우에는 앞선 분석에서도 선거 결과가 투표율에 매우 민감하게 좌우된다는 결론을 얻었는데, 확률적 투표 참여 시뮬레이션에서도 그와 일관된 결과가 도출된다. 아래 그림의 시뮬레이션 결과에 따르면 투표에 참여한 유권자 집단이 약간만 변하는 경우에도 대부분 선거 결과가 뒤집힌다는 사실을 알 수 있다. 우선 <그림 19>를 살펴보면 대부분의 경우 이재명 후보의 득표율이 윤석열 후보의 득표율을 크게 상회한다. <그림 20>에서 이를 한 눈에 알아볼 수 있는데, 1000번의 시뮬레이션 중에 이재명 후보의 득표율이 윤석열 후보의 득표율보다 더 작은 경우는 단 세 번에 불과하다. 이는 투표율이 약간만 올라가거나 내려갔더라도 이재명 후보가 선거에서 이겼을 것이라는 이전 시뮬레이션의 분석과도 맥을 같이하는 결과이다. 특히, 선거에서 실제로는 투표에 참여했지만 예측된 투표 참여 확률은 낮았던 집단이 주로 윤석열 후보 지지자 중에 많았고, 선거에 기권했지만 예측된 투표 참여율이 높았던 집단은 주로 이재명 후보 지지자였기 때문에 확률적 투표 참여 시뮬레이션에서는 이와 같이 실제 선거와는 다른 결과가 높은 비율로 나타났다고 해석할 수 있다.

| 그림 19 | 20대 대선 총투표율과 후보 득표율

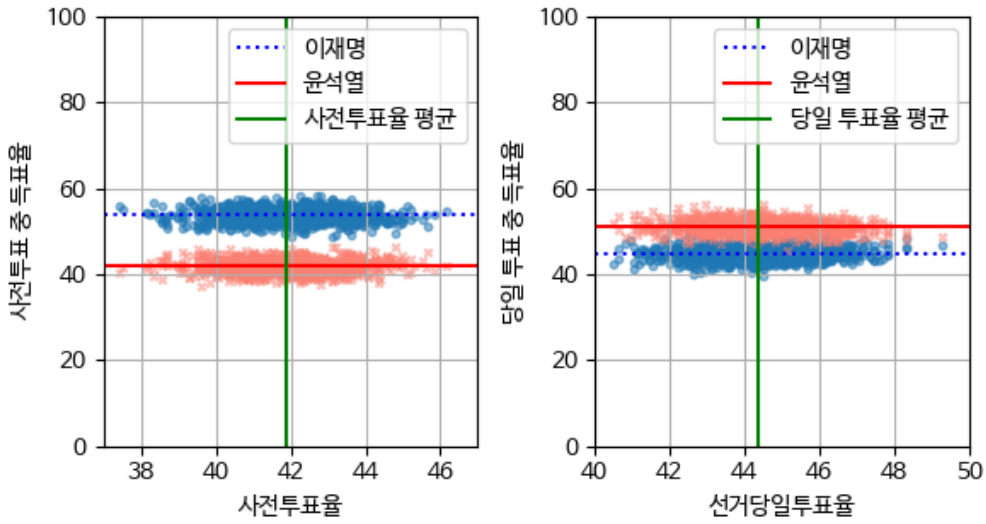


| 그림 20 | 20대 대선 총투표율과 후보 득표율 차이

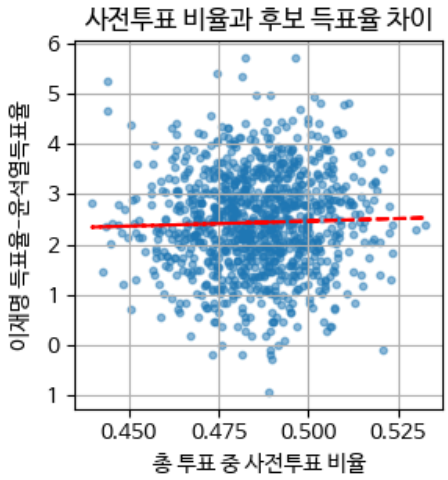


20대 대선을 사전 투표와 선거 당일 투표로 나누어 살펴보아도 같은 결론을 내릴 수 있다. <그림 21>을 보면 이전 시뮬레이션 결과와 마찬가지로 사전 투표에서는 이재명 후보의 득표율이 더 높고, 선거 당일 투표에서는 대체로 윤석열 후보의 득표율이 더 높다는 것을 확인할 수 있다. <그림 22>를 살펴보면, 총 투표에서 사전투표의 비중이 늘어날수록 이재명 후보와 윤석열 후보의 득표율 격차가 더 벌어지지만, 그 정도는 미미하다. 또한 총투표율 분석결과와 비슷하게 1000번의 시뮬레이션 중에 7번을 제외하고는 이재명 후보가 선거에서 승리할 것으로 예측되었다.

| 그림 21 | 20대 대선 사전 투표율 및 선거 당일 투표율과 후보 득표율



| 그림 22 | 20대 대선 사전투표 비율과 후보 득표율 차이



5. 투표 선택과 투표 참여

선거에서 특정 투표방법이 특정 정당에 일률적으로 유리하다거나 불리하다고 볼 수는 없지만, 그럼에도 불구하고 유권자들의 지지 성향에 따라 투표방법에 대한 선호가 달라질 수 있다. 특히, 21대 총선에서 사전투표가 부정선거 논란을 일으킬 만큼 이슈화되었기 때문에, 2년 후 치러진 20대 대선에서는 그러한 경향성이 더욱 강화되었을 가능성도 있다. 실제로 김준성 외(2022)의 연구는 20대 대선에서

보수적 성향의 유권자들이 사전투표에 대해 낮은 수준의 신임을 보인다고 밝힌다.

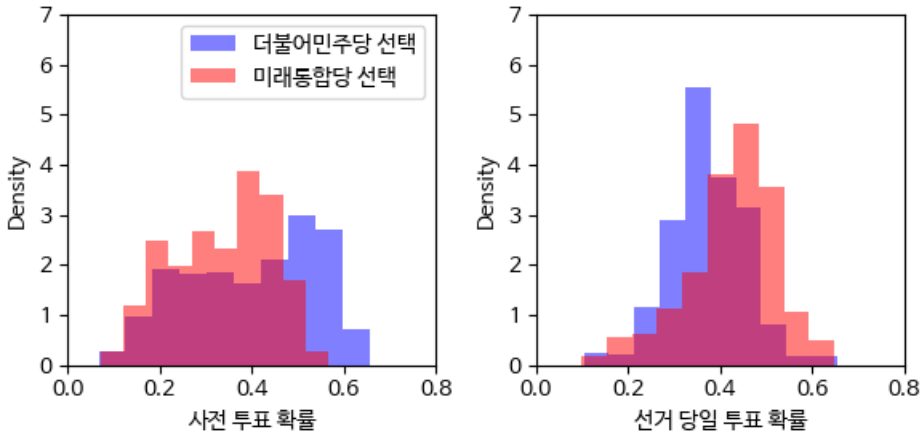
유권자들의 투표선택과 투표참여방법에 일관된 방향성이 있는지 알아보기 위해 투표 선택에 따른 투표참여확률의 분포를 살펴볼 수 있다. 일반적인 회귀분석 방법을 활용한 분석에서는 실제 투표 선택을 설명변수로, 투표 방법을 종속변수로 사용하는 것이 불가능하다. 개념적으로 투표 선택이 투표 방법의 원인변수가 될 수 없을 뿐만 아니라, 기권자의 경우 투표 선택 변수가 결측치이기 때문이다. 따라서 간접적으로 유권자의 지지 정당이나 후보, 정당일체감, 이념성향 등을 설명변수로 하여 비슷한 분석을 하게 된다. 반면, 이 논문의 분석에서는 투표선택변수를 제외한 다른 문항들을 활용하여 투표 참여 확률을 예측하고, 이 확률이 실제 투표 선택 변수에 따라 어떻게 달라지는지를 살펴본다. 기권자의 경우에는 실제 투표 선택 대신, 투표 선택에 대한 예측 결과를 활용한다.

<그림23>부터 <그림28>까지는 21대 총선 지역구와 비례대표 선거, 20대 대선을 각각 투표자와 기권자로 나누어 살펴본 것이다. 투표자의 경우 그림을 통해서 일관되게 확인할 수 있는 경향은, 진보 정당이나 그 정당의 후보를 선택한 사람들은 사전투표에 참여할 확률이 보수 정당이나 보수 정당 후보 지지자들에 비해 확연히 높다는 점이다. 반대로, 선거당일투표 확률의 경우에는 보수 정당이나 보수 정당 후보 지지자들의 확률이 일관되게 더 높은 것으로 나타났다.

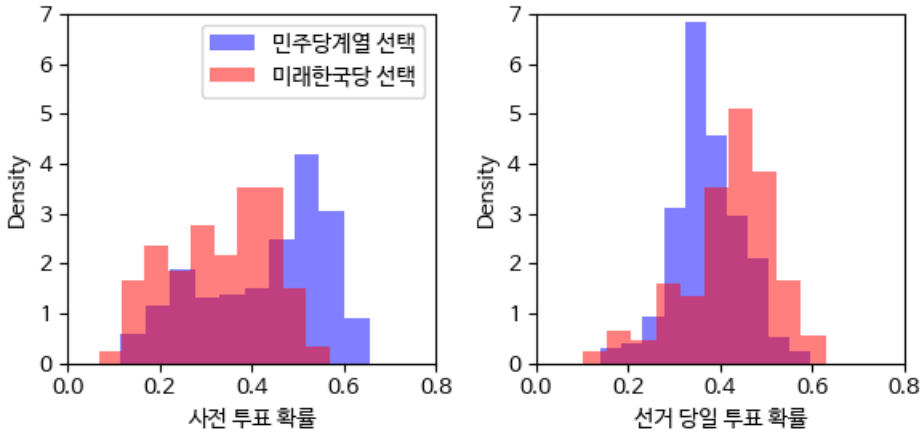
그러나 이러한 경향이 21대 총선에 비해 20대 대선에서 의미있는 수준으로 강화되었다고 보기는 어렵다. 분포의 꼬리가 더 오른쪽 극단값으로 넓어지기는 했지만, 20대 대선이 21대 총선에 비해 박빙의 선거였다는 점, 그리고 훨씬 높은 투표율을 보였다는 점을 고려하면, 꼬리 부분의 모양보다는 분포가 서로 겹치는 부분인 보라색 부분의 넓이에 주목해야 한다. 양 선거 간 보라색 부분의 넓이를 비교해 보면 큰 차이를 보이지 않는다는 것을 알 수 있다.

마지막으로, 투표자에 비해 기권자의 경우 이러한 경향성이 덜 두드러지기는 하지만, 비슷한 경향이 관찰된다는 것을 알 수 있다. 당연하게도, 투표자에 비해 기권자의 투표 참여 확률은 낮은 수준에 많이 분포되어 있다. 특히 21대 총선 기권자의 경우, 상당 수 보수 정당 예측 지지자들의 투표확률이 매우 낮은 수준에 머무르는 것을 확인할 수 있다. 그러나, 대체적으로 여전히 사전투표에서는 더불어민주당/이재명 후보 예측 지지자가, 당일투표에서는 미래통합당/윤석열 후보 예측 지지자가 오른쪽 꼬리에서 더 높은 분포를 갖는다.

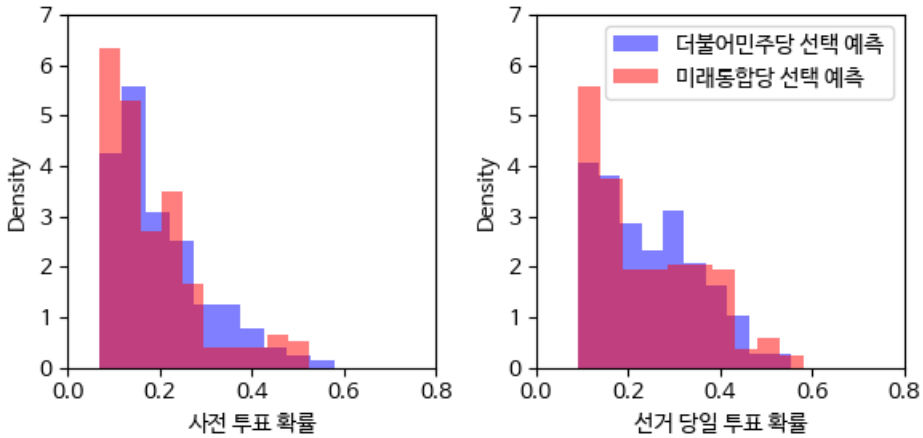
| 그림 23 | 21대 총선 지역구 투표 선택에 따른 투표 확률 (투표자)



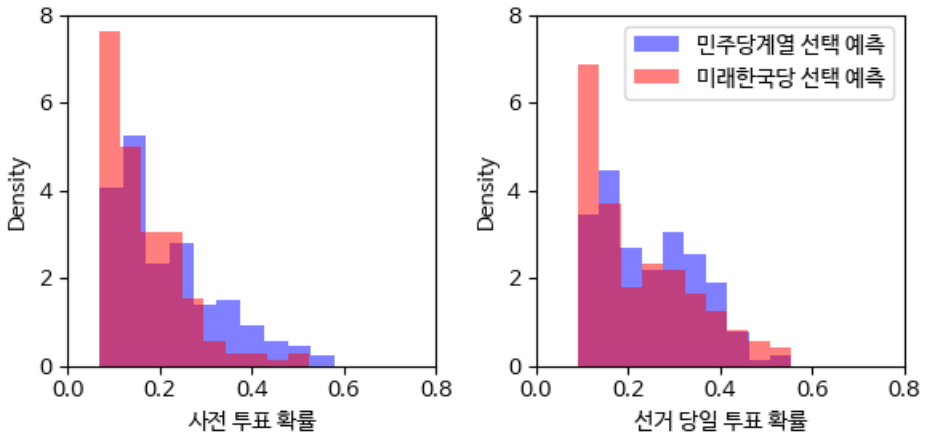
| 그림 24 | 21대 총선 비례대표 투표 선택에 따른 투표 확률 (투표자)



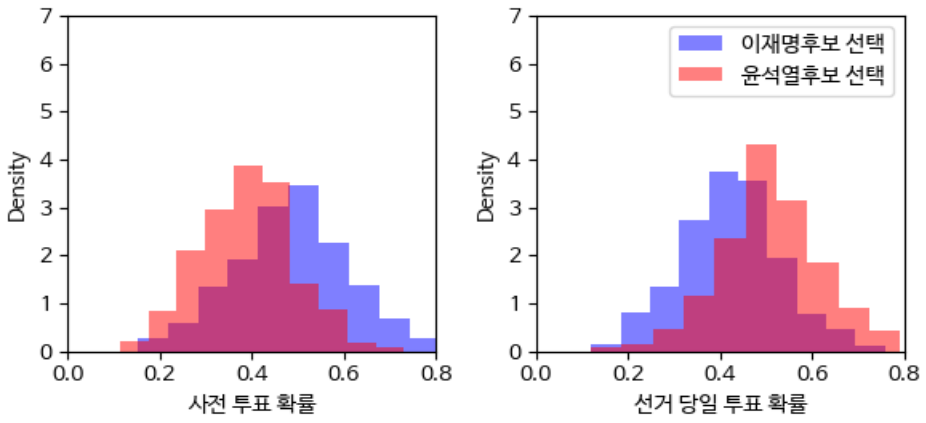
| 그림 25 | 21대 총선 지역구 투표 선택 예측에 따른 투표 확률 (기권자)



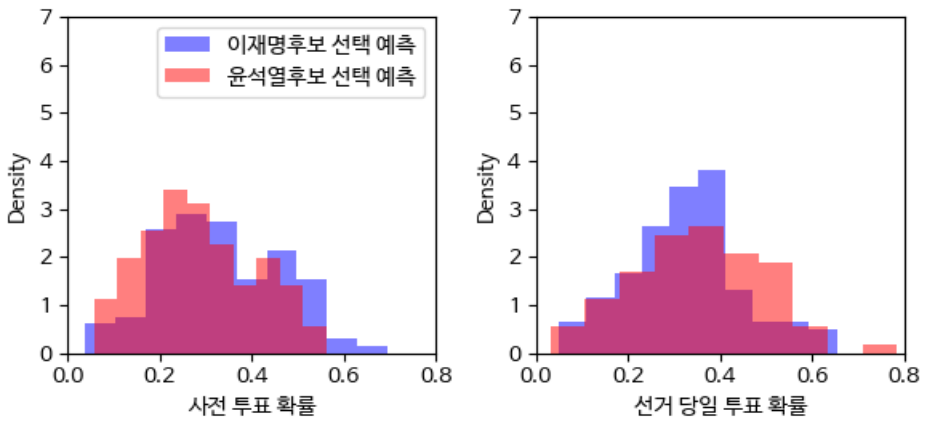
| 그림 26 | 21대 총선 비례대표 투표 선택 예측에 따른 투표확률(기권자)



| 그림 27 | 20대 대선 투표선택에 따른 투표확률(투표자)



| 그림 28 | 20대 대선 투표선택 예측에 따른 투표확률(기권자)



V. 결론

이 논문에서는 머신러닝 기반 예측 시뮬레이션을 통해 총 투표율, 사전투표율, 선거 당일 투표율과 정당편향에 대해 분석하였다. 분석의 결과, 전통적 통념과 부합하는 부분이 어느 정도 존재하지만, 투표율만을 가지고 특정 정당에 유불리를 따지는 것은 매우 어렵다는 결론을 내릴 수 있다. 선거 자체의 특징과 유권자의 선호 변화에 따라 어떤 투표방법의 참여율이 어느 정당에게 더 유리할 지는 달라질 수 있다. 이는 시트린 외(2003)의 연구와도 일맥상통하는 결과이다.

기권자의 특성을 분석한 결과, 선거에 따라서 기권자 집단이 투표자 집단과 비슷한 정치적 선호를 가질 수도 있지만 그렇지 않은 경우도 있다는 것을 알 수 있다. 따라서 투표율 상승이 한쪽 정당에 늘 유리한 결과를 가져다주기는 어렵다는 결론을 도출할 수 있다.

투표율과 정당 득표율에 대한 시뮬레이션 결과는 일정 부분 통념과 일치하는 측면이 있지만, 정당편향과 관련해서는 배치되는 편에 가깝다. 먼저, 21대 총선에서는 가장 높은 투표참여의사를 지닌 유권자가 진보성향의 유권자임을 확인할 수 있는데 이는 기존의 통념과는 다른 결과라고 할 수 있다. 반드시 사전투표율이 높다고 해서 더불어민주당에게 유리하지 않고, 당일투표율이 더 높아진다고 해서 미래통합당에게 더 유리하지도 않았다는 결과도 마찬가지이다. 그러나 가장 높은 사전투표 참여 의사를 가진 사람들이 진보정당 지지자라는 점과, 가장 높은 당일투표 참여의사를 가진 사람들이 보수정당 지지자라는 점은 어느 정도 기존의 통념과 부합한다.

마지막으로, 진보정당에 투표하는 유권자가 사전투표 참여확률이 더 높고, 보수정당에 투표하는 유권자가 당일 투표 참여확률이 더 높다는 것도 기존 통념에 부합하는 결과이다. 다만, 그러한 경향이 21대 총선보다 20대 대선에서 더 두드러진다고 보기는 어렵다.

논문의 분석 결과가 시사하는 바는 사전투표에서 선거부정이 없는 경우에도 당일 투표의 결과에 비해 사전투표의 결과가 진보정당에 훨씬 더 유리하게 나타날 개연성은 충분히 있다는 것이다. 진보정당 선택자가 보수정당 선택자보다 사전투표 참여확률이 높은 반면, 보수정당 선택자가 진보정당 선택자보다 당일투표 참여확률이 더 높다는 사실은 이와 같은 결론을 뒷받침한다. 따라서, 당일투표와 사전투표 결과가 크게 차이가 난다고 하더라도 그것이 곧바로 부정선거의 가능성을 시사한다고 보기는 어렵다는 것이다.

다음으로, 사전투표율과 선거당일 투표율이 일정 수준을 넘기는 경우에는, 상대적으로 투표참여확률이 낮은 유권자도 선거에 참여했다는 의미이기 때문에 단지 투표율의 수준만을 가지고 정당 유불리를 따질 수는 없다는 것이다. 20대 대선에서 사전투표율이 더 낮았다면 오히려 이재명 후보의 득표율이 더 높았을 것이라는 사실이 이를 뒷받침한다.

여러 가지 유용한 시사점에도 불구하고 이 논문의 한계 역시 존재한다. 먼저, 데이터의 한계를 들 수 있다. 머신러닝 방법은 대규모의 데이터에서 더 좋은 성능을 발휘할 수 있다. 그러나 이 논문에서는 1200명 남짓의 설문조사 데이터를 활용하는데에 머물러서 그 정확도 면에서 한계가 있었다. 특히, 투표

참여를 기권/사전투표/당일투표로 나누어 예측했을 때의 정확도는 아쉬운 측면이 있다. 추후 더 큰 규모의 데이터가 마련되거나, 장기간에 걸친 일관된 설문조사가 마련된다면 이러한 단점을 보완할 수 있을 것으로 기대한다. 또한, 사후 선거조사에 의존했다는 점도 한계로 지적될 수 있다. 선거 결과를 알고 난 이후에 여론조사가 이루어지기 때문에, 응답의 진실성 측면에서 왜곡이 있을 수 있다. 20대 대선 설문 조사상의 투표율이 실제 투표율과 큰 괴리가 있다는 점도 설문 응답을 얼마나 신뢰할 수 있을지에 대해 의구심을 품게 한다. 그러나 이 경우에도 설문응답 상의 후보 득표율이 실제 후보 득표율과 큰 차이를 보이지 않기 때문에, 특정 성향 유권자가 투표 참여 여부를 과대보고하는 등의 체계적 편향은 심각하지 않은 것으로 판단된다. 향후 더 정교한 설문기법을 활용한 조사를 통해 후속 연구에서 더 신뢰할만한 분석 결과를 내놓을 수 있기를 기대한다.

보다 근본적으로는 대표성 있는 표본의 확보가 어렵고 투표 참여 여부에 대해 진실하지 않은 응답의 가능성이 있다는 점에서 설문조사가 투표율 예측에 적합하지 않다는 비판이 있을 수 있다. 그러나 앞서 지적한 바와 같이 집단 분석의 문제를 보완하기 위해 유권자 개인을 분석단위로 하려면 설문조사 외의 다른 대안을 찾기는 어렵다. 현재로서는 선거구 단위의 집합적 분석결과와 개인 유권자 설문조사를 분석한 결과를 상호보완적으로 활용하여 연구 문제에 대한 답을 찾아가는 것이 최선의 방안이다.

또한, 이 논문은 투표율에 따른 정당 편향 여부를 다각도로 분석하는 데 주력하여 한국 정치의 맥락에서 중요하게 논의되어 온 여러 요인들에 대한 고려가 미흡하였다. 특히 세대별 투표 성향, 지역주의, 정당 지지 성향, 갈등적 정책 이슈에 대한 유권자들의 입장 변화 등이 투표율 및 득표율에 어떤 영향을 미칠 수 있는지를 분석하는 것은 향후 연구에서 다루어져야 할 중요한 과제로 보인다.

참고문헌

가상준. 2016. “사전투표제는 투표율을 제고하는가?”『한국정당학회보』15권 1호, 5-28.

_____. 2018. “사전투표 유권자의 특징 변화.”『한국정당학회보』17권 4호, 99-120.

_____. 2021. “2020 년 국회의원선거에서 사전투표 유권자의 특징과 투표선택.”『한국정치학회보』55권 2호, 89-108.

_____. 2024. “투표 기권자는 누구에게 투표하였을까?”『한국정당학회보』23권 3호, 5-31.

강신구. 2016. “사전투표제도와 투표율: 제 20 대 국회의원선거 유권자 조사 자료분석.”『한국정치연구』25권 3호, 225-251.

김도경. 2014. “누가 사전투표제에 참여했는가?: 부산지역 대학생을 중심으로: 부산지역 대학생을 중심으로.”『21 세기정치학회보』24(3): 485-510.

김준석, 구본상, 최준영. 2022. “사전투표, 당일투표, 그리고 유권자의 투표신임: 제 20 대 대통령선거를 중심으로.”『선거연구』1권 17호, 5-32.

문우진. 2011. “정치의식과 불평등 투표.”『국가전략』17권 3호, 73-93.

민인식, 유경준. “사전투표 분석방법에 관한 탐색적 연구-계량경제학적 관점.”『한국정책학회보』30권 1호, 61-86.

박경미. 2012. “투표에 관한 세 가지 통념에 대한 경험적 분석: 서울시 선거구의 총선 결과를 중심으로.”『의정연구』20권 3호, 154-183.

이재묵. 2020. “21 대 국회의원 선거에서의 사전투표 유권자 특징 분석.”『글로벌정치연구』13권 2호, 1-23.

지병근. 2012. “투표율 상승이 민주통합당에게 이로울까?: 제 19 대 총선에서 나타난 투표율의 정당편향.”『한국정치연구』21권 3호: 127-153.

중앙선거관리위원회. 2022. “제20대 대통령선거 선거여론조사결과 추이.” 중앙선거여론조사 심의위원회 홈페이지, 2022년 7월. https://shipvote.nec.go.kr/site/nec/ex/bbs/View.do?cbIdx=1129&bcIdx=188078(검색일: 2025년 4월 4일)

Arnold, Felix, and Ronny Freier. 2016. “Only conservatives are voting in the rain: Evidence from German local and state elections.” *Electoral Studies* 41: 216-221.

Bernhagen, Patrick, and Michael Marsh. 2007. “The partisan effects of low turnout: Analyzing vote abstention as a missing data problem.” *Electoral studies* 26(3): 548-560.

Chen, Tianqi, and Carlos Guestrin. 2016. “Xgboost: A scalable tree boosting system.” *In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* 785-794.

Citrin, Jack, Eric Schickler, and John Sides. 2003. “What if everyone voted? Simulating the impact of increased turnout in senate elections.” *American Journal of Political Science* 47(1): 75-90.

Feddersen, Timothy J. and Wolfgang Pesendorfer. 1996. “The Swing Voter’s Curse.” *American Economic Review* 86(3):408-424.

Gohdes, Anita R. 2020. “Repression technology: Internet accessibility and state violence.” *American*

Journal of Political Science 64(3): 488-503.

Gomez, Brad T., Thomas G. Hansford, and George A. Krause. 2007. “The Republicans should pray for rain: Weather, turnout, and voting in US presidential elections.” *Journal of Politics* 69(3): 649-663.

Grinsztajn, Léo, Edouard Oyallon, and Gaël Varoquaux. 2022. “Why do tree-based models still outperform deep learning on typical tabular data?.” *Advances in neural information processing systems* 35: 507-520.

Hansford, Thomas G., and Brad T. Gomez. 2010. “Estimating the electoral effects of voter turnout.” *American Political Science Review* 104(2): 268-288.

Islam, Raisa, Subhasish Mazumdar, and Rakibul Islam. 2024. “An Experiment on Feature Selection using Logistic Regression.” *2024 5th Information Communication Technologies Conference (ICTC):* 319-324. IEEE.

Martinez, Michael, and David Hill. 2007. “Was the joke on the democrats again? Turnout and partisan choice in the 2004 US election.” *American Review of Politics* 28: 81-95.

Mebane Jr, Walter R. 2020. “Anomalies and Frauds in the Korea 2020 Parliamentary Election.” publié le.

McMurray, Joseph C. 2013. “Aggregating information by voting: The wisdom of the experts versus the wisdom of the masses.” *Review of Economic Studies* 80(1): 277-312.

Palfrey, Thomas R., and Keith T. Poole. 1987. “The relationship between information, ideology, and voting behavior.” *American journal of political science* 31(3): 511-530.

Rubenson, D., Blais, A., Fournier, P., Gidengil, E., & Nevitte, N. 2007. “Does low turnout matter? Evidence from the 2000 Canadian federal election.” *Electoral Studies* 26(3): 589-597.

Ruczyński, Hubert, and Anna Kozak. 2024. “Do Tree-based Models Need Data Preprocessing?” *AutoML Conference* (Workshop Track).

Wäckerle, Jens, and Bruno Castanho Silva. 2023. “Distinctive voices: Political speech, rhetoric, and the substantive representation of women in European Parliaments.” *Legislative Studies Quarterly* 48(4): 797-831.

Voter Turnout and Party Advantage: A Machine Learning-based Approach

Jinhee Jo (Kyung Hee University)

Abstract

This paper analyzes the relationship between voter turnout and party advantage by examining voter surveys from the 21st National Assembly election and the 20th Presidential election. Employing several popular machine learning algorithms, the paper investigates how changes in total voter turnout, early voting turnout, and election day turnout impact party vote shares and election outcomes. The analysis reveals that the common belief that higher voter turnout or early voting turnout benefits liberal parties is only partially true. It is difficult to draw consistent conclusions about whether a higher turnout benefits a specific party based solely on voter turnout, as the impact of participation in different voting methods may vary depending on the characteristics of the election itself and changes in voter preferences.

Key words: voter turnout, machine learning, simulation, party advantage, election