

연구논문

성별, 연령 그리고 세대 구성의 동적 변화가 선거에 미치는 영향: 22대 국회의원 선거를 중심으로

홍기혜* · 민인식**

2024년 4월에 실시 예정인 22대 국회의원 선거를 앞둔 시점에서 본 연구는 전통적 투표이론인 성별, 연령, 세대(이념) 인구구성이 보수정당 후보 지지에 미치는 영향을 분석한다. 19대(2012), 20대(2016), 21대(2020) 국회의원 선거 데이터를 이용하여 읍·면·동별 인구구성과 보수후보 지지율 데이터를 활용하였다. 인구구조에 해당하는 24개 예측인자를 구성하고 머신러닝 기법인 XGBoost 알고리즘을 활용하여 각 예측인자가 선거예측에 미치는 중요도와 방향성을 도출하였다. 분석 결과에 따르면 첫째, 22대(2024) 국회의원 선거 유권자는 20대(2016) 및 21대(2020) 선거와 비교하여 30대, 40대, 50대 비율이 남녀 모두 감소하는 반면에 60대 이상과 20대(18~29세) 유권자 비율이 증가함을 확인할 수 있다. 둘째, 보수후보 지지 예측에 영향을 미치는 중요한 예측인자는 '여자-후기 MZ세대'이다. 해당 유권자 비율이 증가할수록 보수후보 지지율에 음(-)의 영향을 미친다. '남자-전통세대' 그룹 역시 보수후보 지지에 중요한 예측인자로 작동하여 해당 유권자 비율이 증가할수록 보수후보 지지율에 양(+)의 영향을 미친다. '남자-30대'는 세 번째로 중요한 예측인자이지만 이들의 비율이 증가할수록 보수후보 지지율은 음(-)으로 영향을 받게 된다. 셋째, 22대(2024) 선거의 경우, 인구구조에 의해서만 설명되는 보수후보 지지율을 21대(2020) 선거와 비교하였을 때 수도권 1,094개 읍·면·동 가운데 보수후보에 유리하게 인구구성이 변화하는 읍·면·동은 834개로 전체의 76.2%이다. 이러한 분석 결과는 상대적으로 지역주의 투표성향이 높은 수도권에서는 인구구성의 변화가 보수정당에 유리하게 작동할 가능성이 높다는 것을 시사한다.

주제어: 성별-연령-세대 효과, 22대 국회의원 선거, XGBoost, SHAP(SHapley Additive Explanations)

* 연세대학교 사회복지연구소, 각당복지재단(prd@kakdang.or.kr), 제1저자.

** 경희대학교 정경대학 경제학과 교수(imin@khu.ac.kr), 교신저자.

I. 서론

2020년 COVID-19 팬데믹 초기에 시행된 21대 국회의원 선거에서 당시 여당인 민주당이 압승한 이후 20대 대통령 선거(2022)와 8회 지방선거(2022)를 거치면서 2023년 12월 현재 정치권력이 재편되어 있는 상황이다. COVID-19 감염병과 같은 재난상황 아래에서 유권자 선택은 기존과는 다른 기준으로 후보 지지를 결정할 가능성이 있다. 신정섭(2020)은 21대 국회의원 선거 이후 설문조사 결과를 이용하여 유권자는 재난으로 인한 피해 유무로 투표 결정을 하기보다는 재난에 대한 정부의 대응을 평가하여 합리적으로 투표 결정을 한다는 결론을 얻었다. 3개월 여 남은 22대 국회의원 선거(2024년 4월 예정)는 COVID-19 재난 종료와 여·야 간 권력교체 이후 치르는 첫 번째 총선으로 그 중요성이 부각되고 있다.

국회의원 및 대통령 선거에서 후보 지지에 미치는 영향은 다양한 요인으로 분석되고 있다. 한국의 선거에서 영향력을 행사하는 전통적 요인으로는 지역주의 투표성향이 있다. 영·호남 지역분할 구도가 점차 약화되고는 있으나 지역주의는 여전히 후보 지지에서 의미 있는 영향을 미친다는 분석 결과를 얻고 있다(김태완 2012; 문우진 2017). 다만 수도권에서는 정치적 지식수준이 상대적으로 높고 지역주의의 영향을 덜 받으며 후보자의 능력, 정책, 공약 등을 중시하는 경향이 나타난다(윤종빈 2007).

국회의원 선거에서는 인물론 또는 현직자 효과 역시 중요한 요인으로 분석되고 있다. 박상운·이호준(2015)은 18, 19대 국회의원 선거에서 현직자 우위 원인을 공직자 효과와 자질효과로 구분하여 분석하였는데, 공직자 효과가 당선 가능성을 높이고 지역구의 선거경쟁도를 낮추는 것으로 파악하였다. 20대 국회의원 선거에서 나타난 후보의 지역대표성과 당선 가능성의 관계를 분석한 정수현(2017)은 지역대표성이 높을수록 후보자의 득표율이 높아지고 이러한 효과는 비수도권에서 두드러지게 나타난다는 결과를 얻었다.

전통적 투표이론인 계급투표와 달리 Lewis-Beck & Nadeu(2011)는 경제적 상황 인식이 투표 결정에 영향을 미친다고 주장한다. 한국에서도 17대 대통령 선거(2007)에서 경제투표 성향이 나타나고 이를 검증하는 연구가 활발하게 진행되고 있다(강우진 2012; 민인식 2022; 최종호 2016; 최필선·민인식 2016). 특히 최필선·민인식

(2016)과 민인식(2022) 연구에서는 주택가격 및 부동산 소유 여부가 보수후보 지지에 미치는 영향을 분석하였다.

이와 같이 다양한 요인이 한국의 선거결과에 영향을 미치고 있고 그러한 영향력은 시간에 따라 변화하고 있다. 그중에서도 전통적 투표이론에서 논의하는 성별, 연령, 세대(이념)의 효과는 국내 선거에서도 역시 중요한 요인으로 받아들이고 있다. 특히 지난 20대 대통령 선거(2022)에서는 기존의 세대-연령 지지성향과는 달리 20대 남성의 보수후보 지지가 새롭게 부각되었다. 민영(2022)은 21대 국회의원 선거에서 노년층의 투표 선택을 견인하는 요인에 연령과 세대 효과가 어떻게 다르게 작용하는지를 탐색하였으며, 선거결과에 미치는 요인을 분석하는 많은 선행연구가 세대(이념)와 연령 효과를 결합하여 분석하고 있다(문우진 2017; 정진민 2012).

본 연구에서는 후보 지지의 전통적 요인에 해당하는 성별, 연령 그리고 세대의 동적 변화를 국회의원 선거예측에 활용하고자 한다. 저출산·고령화라는 우리 사회의 급격한 인구구조 변화는 선거결과에도 영향을 미칠 것으로 예상된다. 특히 민주화 세대라고 불리는 X세대(1965~1979년생)가 사회의 중추적 위치로 자리매김하였고 이들이 장년층에 접어들면서 세대와 연령이 교차하고 있는 인구구조의 변화로 인해 연령-세대가 선거에 미치는 효과는 예전과는 다를 것으로 예상된다.

본 연구는 선거에서 성별, 연령, 세대의 효과만을 고려하기 위해 지역주의 투표성향이 상대적으로 약한 수도권 선거구만을 분석 대상으로 선택한다. 수도권 의석은 지난 21대 국회의원 지역구 가운데 48%(253석 중 121석)로 전체 선거 판세를 가늠하기에 충분하다고 할 수 있다. 실증분석 방법론은 과거의 선거결과를 통해 미래의 선거를 예측하는 것이므로 전통적 통계모형 대신 머신러닝 알고리즘에 해당하는 XGBoost를 활용하고자 한다. 이를 위해 공선성 문제를 고려하고 최대한 많은 관측치를 확보하고자 최근 선거에 해당하는 19~21대 선거에서 얻은 읍·면·동별 후보 지지율 자료를 수집하였다. 전체 유권자를 성별-연령(14개 그룹)과 성별-세대(10개 그룹)로 구분하여 전체 유권자 대비 각 그룹의 유권자 비율을 예측인자(predictor)로, 각 읍·면·동에서 보수정당 후보의 지지율을 종속변수로 설정하고 실증분석 결과를 제시한다. 과거 국회의원 선거데이터에 기초하여 22대 국회의원 선거(2024년 4월 예정)의 보수정당 후보의 지지율을 예측하는 것 외에도, 24개의 예측인자가 선거결과 예측에 기여하는 중요도와 방향성을 도출하여 논의하고자 한다.

본 연구는 다음과 같이 구성되어 있다. 이어지는 II장에서는 선거에서 연령과 세대 그리고 성별효과를 논하는 선행연구와, 선거예측에서 머신러닝 방법론의 활용과

관련된 최근 논의를 정리한다. III장에서는 실증분석에 필요한 최근 국회의원 선거시점에서 인구구성과 보수후보 지지율 데이터를 구축하는 과정을 설명한다. IV장에서는 성별-연령-세대 그룹 중에서 선거예측에 있어서 중요한 변수를 식별하고 22대(2024) 선거에서 인구구조에 근거한 읍·면·동별 보수정당 후보의 지지율을 예측하고, V장에서는 정책적 시사점 및 연구의 한계를 논의한다.

II. 선행연구 검토

1. 선거에서 연령 및 세대 그리고 성별 효과

세대는 같은 시기에 태어나서 인생의 주요한 전환기에 비슷한 사회적 사건을 겪은 사람들의 집단을 의미한다(Mannheim 1952). 세대 코호트 이론은 인간발달의 각 단계에서 겪는 사건과 그로 인한 결과에 따라 한 세대는 공통된 정서를 형성하며 이는 시간이 지나도 지속된다는 개념이다(Inglehart 1977). 이에 근거하면 특정한 세대에게 형성된 정치적 성향은 시간이 지나도 쉽게 변하지 않는다고 할 수 있으며 상이한 상황에서 성장한 각 세대는 구분되는 정치적 가치관을 갖는다고 할 수 있다. 세대 이론가들은 사회 구성원들에게 미치는 역사적 사건의 규모가 크고 사회적 변화가 신속할수록 세대는 더욱 강력한 해석 도구가 될 수 있다고 주장한다(Keniston 1969; Lambert 1972; Ryder 1965). 이러한 주장은 한국의 선거에서 세대가 높은 설명력을 가질 수 있다는 의미로 해석될 수 있다.

한국의 선거에서 세대에 관한 연구는 1987년 민주화 이후 본격화되기 시작하였다. 이전 연구에서는 한국전쟁을 기준으로 전쟁 이전과 이후 세대로 막연하게 통칭하던 코호트 집단을 전전세대, 민주세대, 신세대 등으로 구분하여 세대가 선거에 미치는 영향을 분석하는 연구들이 진행되었다(정진민 1992). 이러한 초창기 세대연구에서는 선거에서 전통적으로 영향력이 인정되는 도시화 요인이나 지역주의 요인에 비견되는 세대의 영향력을 밝히고 있다(정진민 1994). 이후 각 세대가 보이는 행태는 선거에서 관심을 두는 현상으로 다수의 연구에서 세대효과를 검증하기 시작하였다(강원택·성예진 2018; 김학량 2015; 서현진 2008; 이재철 2013; 정진민 2012; 최준영·조진만 2005).

하지만 선거에서의 차이를 설명할 수 있는 것이 연령효과인지 세대효과인지에 대한 질문이 지속되면서 연령과 세대를 구분하여 각각의 효과를 파악하고자 하는 연구들이 진행되었다. 이러한 연구들은 수년간의 자료를 결합하고(노환희 외 2013; 박명호 2009; 이갑윤 2008; 이내영·정한울 2013; 이우진 2014; 이준한 2013; 최슬기 외 2019; 허석재 2014) 관찰시점의 사회적 분위기 같은 시점효과를 통제하거나(노환희 외 2013; 최슬기 외 2019; 허석재 2014) 연령과 세대의 교호작용을 분석하며(이내영·정한울 2013) 보다 엄중하게 연령효과와 세대효과를 파악하고자 하였다. 이에 다수의 연구에서 보고하는 공통적인 결과는 나이가 들면서 정치적 성향이 보수화되는 연령효과(노환희 외 2013; 이준한 2013; 최슬기 외 2019; 허석재 2014)가 나타난다는 것이며, 일부 연구에서는 연령효과가 보다 설득력 있는 것(박명호 2009; 이준한 2013)으로 분석하고 있다.

선거연구에서 많은 학자가 관심을 보이는 또 다른 현상으로는 성별의 차이가 있다. 전통적으로 여성은 이념적으로 보수성이 강하며 보수정당을 지지하는 경향이 높다는(송선희 1986; 이남영 1995; Almond & Verba 1963; Campbell et al. 1960; Inglehart 1977; Verba et al. 1978) ‘전통적 성별의 차이(traditional gender gap)’가 강조되어 왔다. 하지만 이러한 여성의 특성은 최근 들어 교육수준 향상, 성역할 변화, 사회진출 증가 등으로 동일 조건의 남성에 비해 오히려 진보적 성향이나 진보정당을 지지하는 경향성이 높다고 밝혀지고 있는데 Inglehart & Norris(2000)는 이를 ‘현대적 성별의 차이(modern gender gap)’로 언급하기도 하였다.

한국의 경우 여성은 여전히 ‘전통적 성별의 차이’라는 특성이 나타나고 있으나 젊은 연령대에서는 ‘현대적 성별의 차이’로 전환되는 과정을 경험하고 있다고 보고한다(구본상 2017). 중년세대 연구에서는 혼인 관계를 형성하고 있는 유권자들이 해체된 유권자들보다 투표참여에 적극적인데 이는 여성에게서 더욱 두드러진다고 밝히고 있으며(김기동 외 2021), 건강상태 악화에 따른 투표참여 감소가 여성의 경우 더욱 높게 나타나는 성별 특성을 보고하고 있다(김기동·이재묵 2023). 한편, 성별 관점에서 파악하는 청년세대 연구에서는 첨예하게 형성되어 오던 젠더전선이 20대 대선에서 젠더균열로 완성되며 정치적 행동으로 연계되었다고 분석한다(김한나 2022; 라미경 2022).

이와 같은 관점은 선거에서 성별, 연령, 세대의 동적 변화가 선거예측에 구별되는 영향력이 있을 것을 시사한다. 이에 본 연구는 이러한 세 가지 예측인자의 동적 변화가 선거결과를 예측하는 모형을 통해 선거에 미치는 영향력을 파악하고자 한다.

2. 선거예측에서의 머신러닝 방법론

결과예측이 주요한 목적인 머신러닝은 의학이나 공학에서 왕성하게 활용되는 기법이었으나 사람의 행동을 예측하는 분야로 점차 확대되고 있다. 그 가운데에서도 선거결과에 대한 예측은 학계의 관심이 집중되는 영역이라고 할 수 있다.

선거와 관련해서 가장 활발하게 머신러닝 기법을 활용하고 있는 분야는 트위터, 소셜 미디어 댓글 등의 비정형 텍스트에 대해 감성 분석(sentiment analysis)을 시행하여 선거결과를 예측하는 연구이다. 많은 나라에서 이러한 방법을 통해 대통령 선거, 국회의원 선거, 지역선거 등의 결과를 예측하였는데, 구체적으로는 미국(Joyce & Deng 2017; Singh et al. 2021; Tsai et al. 2019), 한국(Park et al. 2022), 스페인(Khandelwal et al. 2017), 인도네시아(Buntoro et al. 2021), 파키스탄(Ali et al. 2022), 인도(Myilvahanan et al. 2023), 말레이시아(Balakrishnan et al. 2021), 나이지리아(Oyebode & Orji 2019) 등에서 랜덤포레스트, Word2Vec 등의 기법을 활용하여 선거결과를 예측하는 연구를 진행하였다.

한편 정형 데이터를 통해 선거의 결과를 예측한 연구들도 진행되었다. Sinha et al. (2022)은 라소(Lasso Regression)를 활용하여 대통령 선거결과를 예측하고 투표행태에 영향을 미칠 수 있는 경제적, 비경제적 요인을 파악하였다. 거시경제지표(macroeconomic indicators) 및 사회지표(social indicators)를 통해 11개 국가의 현직 정당이 권력을 유지할 것인지에 대해 예측한 연구는 다층 퍼셉트론(multi-layer perceptron)이 가장 정확한 분류를 하는 것으로 보고하였으며(Guan & Mani 2022), 선거 데이터를 사용하여 의석의 승패를 예측한 연구에서는 의사결정나무가 가장 우수한 정확도를 보이는 것으로 파악하였다(Yadav & Johari 2021). 인구통계학적 데이터를 통해 선거결과를 예측한 연구에서는 의사결정나무 등을 사용하여 투표참여 예측모델을 비교하거나(Fitrani et al. 2022) 미국의 하원 선거예측에 트리-기반(tree-based) 모델의 결과가 우수하다(Richardson & Hougen 2020)는 사실을 밝히고 있다.

국내 연구에서는 의사결정나무를 이용하여 15대 대선 결과를 예측한 연구(최종후 외 2000)와 선거에서 의사결정 유보층을 예측하는 데 4가지 트리-기반 알고리즘의 예측성능을 비교하고 평가한 연구(최종후 외 2006)를 필두로 하고 있다. 그러나 이후의 머신러닝 기법을 활용한 선거연구는 최근 들어서야 다시 시작되었다. 김정경·박민규(2023)는 20대 대통령 선거 당선자를 예측하는 데 있어 로지스틱 회귀분석과

트라-기반 알고리즘의 예측력을 비교하여 랜덤포레스트가 가장 우수한 성능을 갖는다는 것을 밝혔다. 랜덤포레스트를 활용하여 고등학생의 선거참여를 예측한 연구에서는 선거참여 요인들을 탐색하고 주요한 요인들을 인간발달의 생태체계적 관점에 따라 논의하였다(김하정·원효현 2023). 윤호영(2023)의 연구에서는 트라-기반 알고리즘과 K-평균 군집분석 등을 활용하여 유튜브 선거활동 변수로 국회의원 당선 여부를 예측하여 미디어 변수로 선거결과를 예측할 수 있다는 것을 밝혔다.

선거결과를 예측하는 이상의 국내·외 연구에서는 트라-기반 모델이 매우 활발하게 활용되고 있으며 예측력 또한 높다는 것을 파악할 수 있다. 이에 본 연구는 정형 데이터(tabular data)를 활용한 예측에 우수한 결과를 보이며(Fayaz et al. 2022; Grinsztajn et al. 2022) 다수의 선거연구에서 이용되고 있는 트라-기반 알고리즘을 통해 선거에서의 보수후보 지지율을 예측하고자 한다.

Ⅲ. 데이터 및 연구방법론

1. 읍·면·동별 인구구성 데이터

본 연구의 인구구성은 성별, 연령, 세대로 구분한다. 선거구의 가장 기초단위인 투표소 단위에서는 인구구성 데이터를 확보할 수 없으므로 본 연구에서는 다음 단계의 기초단위에 해당하는 읍·면·동 수준에서 인구구성과 그 변화를 살펴보고자 한다. 성별, 연령, 세대 구분은 <표 1>과 같이 성별-연령 그룹에서 14개 그리고 성별-세대 그룹에서 10개로 총 24개 그룹에서 전체 유권자 대비 해당 그룹의 인구비율을 상정한다.

본 연구에서 사용한 19대(2012), 20대(2016) 그리고 21대(2020) 국회의원 선거 유권자의 하한 연령은 각각 19세와 18세로 차이가 있으므로 19대, 20대 선거에서는 19세, 21대 선거에서는 18세로 설정하였다. 세대의 구성은 합의된 기준이 없으므로 본 연구에서는 선행연구에 근거한 세대 구성과 기준을 제안한다. 이에 학계에서 통용되는 구성과 기준에 따라 성인 인구집단을 전통세대(∼1953년생), 베이비붐세대(1954∼1964년생), X세대(1965∼1979년생), 그리고 MZ세대(1980∼2002년생)로 설정한다(김우성·허은정 2007; 이현아·김주희 2021; 홍기혜 2024). 다만 MZ세대는 연령 범위가 23년에 이르므로 공유하는 문화나 매체의 차이를 고려해야 한다는 주

장(김난도 외 2021)에 따라 본 연구에서는 전기 MZ(1980년~1989년생)와 후기 MZ(1990년 이후 출생)로 구분한다.

<표 1> 연령 및 세대 구성

성별-연령		성별-세대	
남자	여자	남자	여자
18~29세 age20_m	18~29세 age20_f	전통세대 ~1953년생 cohort1_m	전통세대 ~1953년생 cohort1_f
30~39세 age30_m	30~39세 age30_f	베이비붐세대 1954~1964년생 cohort2_m	베이비붐세대 1954~1964년생 cohort2_f
40~49세 age40_m	40~49세 age40_f	X세대 1965~1979년생 cohort3_m	X세대 1965~1979년생 cohort3_f
50~59세 age50_m	50~59세 age50_f	전기 MZ세대 1980~1989년생 cohort4_m	전기 MZ세대 1980~1989년생 cohort4_f
60~69세 age60_m	60~69세 age60_f	후기 MZ세대 1990년생 이후 cohort5_m	후기 MZ세대 1990년생 이후 cohort5_f
70~79세 age70_m	70~79세 age70_f	-	
80세 이상 age80_m	80세 이상 age80_f		

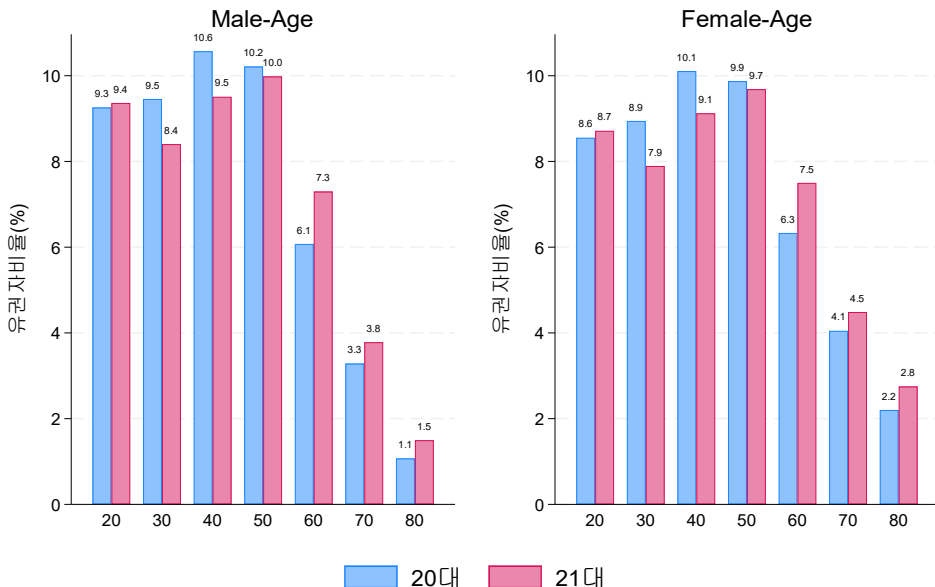
주: 19대, 20대 총선에서 age20_m과 age20_f에서 연령 하한은 19세로 설정한다.

가장 최근 선거가 치러진 21대(2020) 국회의원 선거만 분석에 사용하는 경우 연령대와 세대가 중복되는 부분이 존재할 가능성이 크다. 이러한 공선성 문제를 고려하여 본 연구에서는 가장 최근에 치러진 3번의 국회의원 선거(19대, 20대, 21대)에서 파악되는 유권자 구성을 분석에 활용하고자 한다. 특히 2020년(21대) 선거는 코로나 팬데믹 위기라는 특수한 상황에서 진행되었기 때문에 시점효과를 줄이기 위해 19대, 20대 그리고 21대 선거결과를 모두 사용하는 방식을 선택한다. 각 선거는 그 해의 4월에 치러졌으므로 같은 해 3월 시점 읍·면·동별 주민등록 인구 데이터를 사용한다. 월별 성별-연령 그룹 인구는 국가통계포털 KOSIS에서 확인할 수 있다. 인

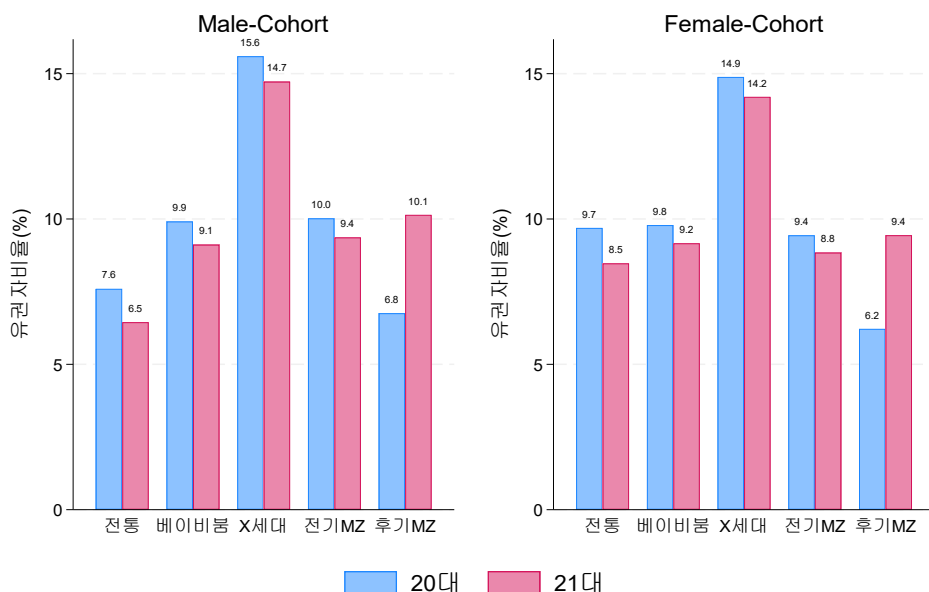
구비율의 분모에 해당하는 전체 인구는 해당 시점의 유권자(19대는 19세 이상, 21대는 18세 이상) 수로 대신한다.

서론에서 언급한 바와 같이 본 연구에서는 인구구조와 선거결과 관계를 수도권으로 한정하여 분석한다. 보수정당인 국민의힘 후보가 출마하여 득표한 기록이 있는 수도권 읍·면·동으로 한정하여 인구구조를 계산한 결과를 제시한다. 19대(2012) 선거에서는 1,075개 수도권 읍·면·동, 20대(2016) 선거에서는 1,116개 수도권 읍·면·동 그리고 21대(2020) 선거에서는 1,111개 수도권 읍·면·동이 분석에 포함되었다.

<그림 1>에서는 성별-연령 그룹별 평균 비율을 20대(2016)와 21대(2020) 선거에서 비교한다. 남·녀 공통적으로 30, 40, 50대 인구비율이 감소하고 60대 이상 그리고 20대에서 인구비율이 증가함을 확인할 수 있다. 20대 선거 대비 21대 선거에서 20대 인구비율이 사소하게 증가하는 이유는 21대 국회의원 선거에서는 18세 유권자가 포함되었기 때문으로 예상된다. 특히 60대 남성의 비율은 6.07%(20대 선거) $\Rightarrow$ 7.29%(21대 선거)로 크게 증가하였고 60대 여성의 비율 역시 6.33%(20대 선거) $\Rightarrow$ 7.50%(21대 선거)로 크게 상승하였음을 확인할 수 있다. 이러한 인구구조 변화는 60대 이상 고령층 지지를 받는 정당이 유리한 구조로 변하고 있음을 예상할 수 있다.



<그림 1> 성별-연령 인구구조 변화



<그림 2> 성별-세대 인구구조 변화

<그림 2>에서는 세대별 인구를 남자-여자 그룹별로 표현하고 있다. 남자와 여자 그룹에서 모두 공통적인 패턴이 나타난다. 전통세대, 베이비붐세대, X세대 그리고 전기 MZ세대 비율은 감소한다. 그러나 신규 유권자가 유입됨으로써 후기 MZ세대의 비율은 지속적으로 증가한다. <그림 1>과 <그림 2>의 결과를 비교하였을 때 연령구조에서는 고령 유권자가 차지하는 비중은 높지만, 성인이 된 후기 MZ세대의 비중이 점차 증가하고 있다고 판단할 수 있다. 인구와 세대에서는 어느 정도 상반된 인구구조 변화가 지난 20대와 21대 선거에서 관찰되고 있다. 이러한 인구구조 패턴은 22대(2024년 4월 예정) 선거에서는 더 뚜렷하게 나타날 가능성이 있다.

2. 국회의원 선거결과 데이터

인구구조가 보수정당 지지에 어떤 영향을 미치는지 예측하기 위해, 본 연구에서 보수정당 지지는 다음과 같이 전체 유효 득표수 대비 보수정당 후보의 득표수로 정의한다.¹⁾ 19대 선거에서는 새누리당, 20대 선거에서는 미래통합당 그리고 21대 선거에서

1) 해당 선거구의 후보자 당선 예측 대신 선거구 내 동별 득표율을 예측대상으로 선택한 이유는 다

는 자유한국당을 보수정당 후보로 정의한다. 식(1)에서 해당 읍·면·동의 투표구에서 *total vote*는 전체 투표수에서 무효투표 수를 제외한 유효 투표수만으로 정의한다.

$$conserv_{ij} = \frac{bosu_{ij}}{total\ vote_{ij}} \quad (1)$$

where *i*: 수도권 내 읍·면·동, *j*: 선거(19~21대)

19~21대 선거결과는 중앙선거관리위원회의 선거통계시스템(http://info.nec.go.kr/main/main_load.xhtml)에서 받아 데이터로 구축하였다. 관외 사전투표(국의 부재자 투표 포함)는 선거구 단위로만 공개되므로 읍·면·동별 보수정당 지지율은 관내 사전투표와 선거일투표 결과만 이용하였다.

<표 2>에서는 보수정당 후보가 출마한 수도권 내 선거구의 읍·면·동에서 *conserv* 변수의 기초통계량을 정리하여 제시한다. 19대 선거에서 1,075개 읍·면·동의 평균 득표율은 48%이다. 20대 선거에서 1,116개 읍·면·동의 평균 득표율은 41%이며 최저득표율은 8.2%(인천광역시 미추홀구 학익1동), 최다득표율은 79.5%(경기도 이천시 설성면)이다.²⁾ 21대 선거에서 1,111개 읍·면·동의 평균 득표율은 44.2%이고 최저득표율은 12.8%(인천광역시 미추홀구 학익1동), 최다득표율은 78.3%(인천광역시 강화군 서도면)이다.

<표 2> 보수정당 후보 지지율: 19대, 20대, 21대 국회의원 선거

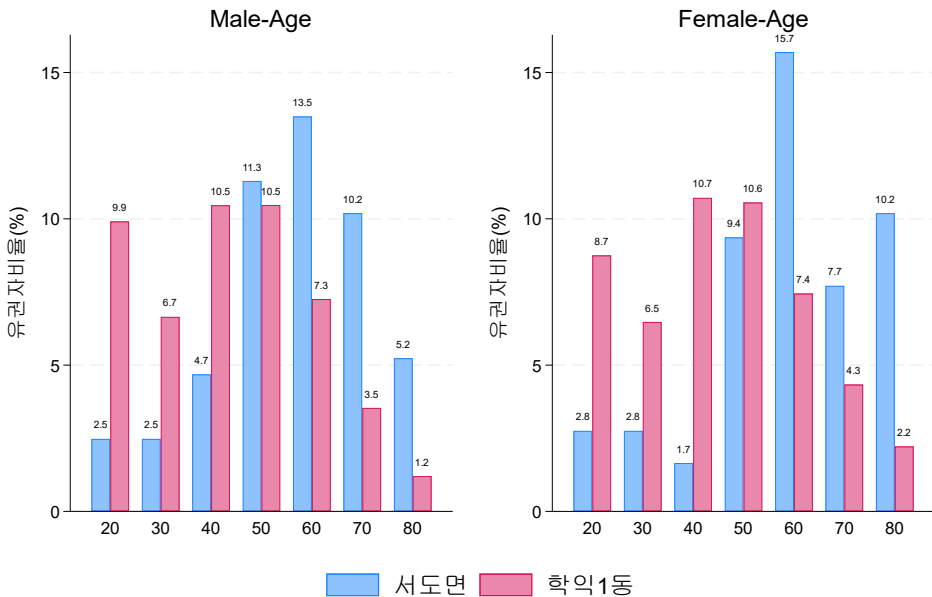
	obs	평균	표준편차	최소값	최대값
19대	1,075	0.480	0.091	0.214	0.796
20대	1,116	0.410	0.107	0.082	0.795
21대	1,111	0.442	0.090	0.128	0.783

주: 보수정당 후보가 출마한 선거구의 읍·면·동으로 한정.

읍과 같다. 첫째, 데이터 수집 단위를 선거구로 하는 경우 표본 숫자가 현저히 줄어들어 예측결과 분산을 크게 만들 수 있으므로 가능한 한 많은 수의 표본을 확보할 필요가 있다. 둘째, 지역투표 성향 및 후보자 인물론, 선거구도 역시 인구구조 못지않게 당선에 영향을 미치는 중요한 요인이다. 인구구조 변수만 가지고는 당선을 예측하는 것이 제약적이다. 따라서 당선 자체보다는 해당 읍·면·동에서 후보자의 득표율을 예측하는 것을 연구 목적으로 한다.

2) ‘인천광역시 미추홀구’는 19~20대 선거시점에서는 ‘인천광역시 남구’ 행정구역에 해당한다.

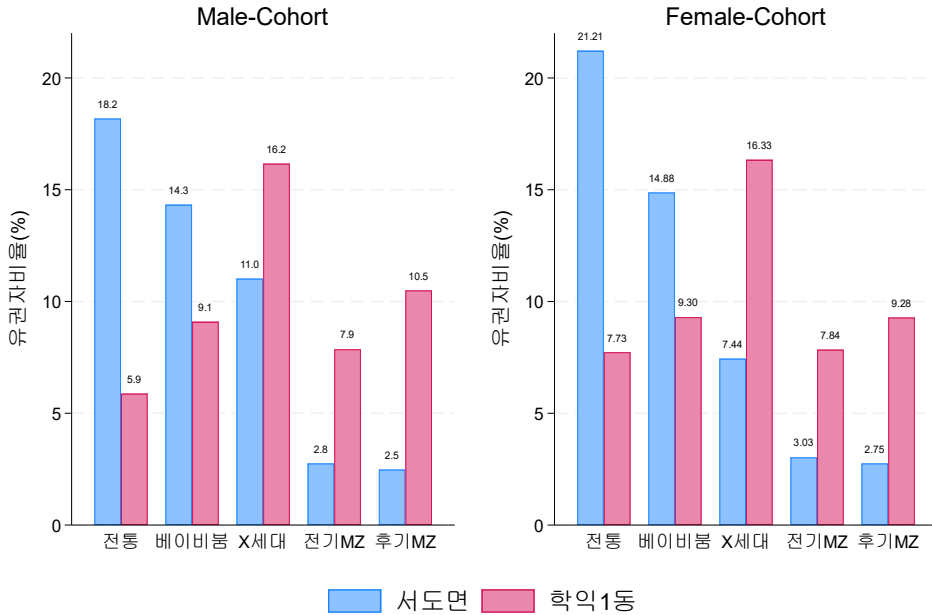
21대 국회의원 선거에 기초하여 보수후보가 최다/최저 득표율을 기록한 읍·면·동을 한정하여 인구구조에서 어떤 차이가 있는지 기술적으로 파악하고자 한다. 같은 광역시인 인천광역시에 포함되어 있음에도 불구하고 보수후보의 득표율에서 큰 차이가 있다는 것은 인구구조의 차이가 크다고 예상할 수 있다. <그림 3>에서는 21대 선거에서 최다득표율(인천 강화군 서도면)과 최저득표율(인천 미추홀구 학익1동) 읍·면·동의 성별·연령 인구구조를 비교한다. 성별과 무관하게 공통으로 20~40대 유권자 비율은 학익1동이 서도면에 비해 압도적으로 높다는 것을 확인할 수 있다. 반면 성별과 무관하게 60대 이상 연령 그룹에서는 서도면이 학익1동보다 그 비율이 압도적으로 높다. 다만 50대 남자 연령대에서는 서도면(최다)이 학익1동(최저)보다 유권자 비율이 더 높지만 50대 여자 연령대에서는 학익1동(최저)이 서도면(최다)보다 유권자 비율이 더 높다. 이러한 결과로 예상할 때 (인천광역시의 경우에는) 50대 남자가 50대 여자보다 보수후보 지지성향이 더 높다는 것을 암시한다.



<그림 3> 최다/최저 득표율 읍·면·동의 성별·연령 인구구조: 21대 국회의원 선거

<그림 4>에서는 성별·세대별 유권자 비율을 강화군 서도면과 미추홀구 학익1동을 비교한 결과를 막대그래프로 보여준다. <그림 3>과 유사하게 전통/베이비붐 세대에

서는 보수후보 지지율이 가장 높은 서도면이 학익1동에 비해 높다. 그러나 X세대, 전기와 후기 MZ세대에서는 학익1동에서 훨씬 높은 수준이다. X세대 비율의 경우 남녀 모두 학익1동이 서도면보다 높지만, 여성-X세대 그룹에서 그 차이가 더 분명하다.



<그림 4> 최다/최저 득표율 읍·면·동의 성별-세대 인구구조: 21대 국회의원 선거

3. XGBoost 머신러닝 알고리즘

본 연구의 실증분석에서는 인구구조에 의해 결정되는 보수후보 지지율 예측을 목적으로 한다. 이를 위해 전통적 통계방법론 대신 예측에 목적이 있는 기법인 머신러닝 알고리즘을 활용하고자 한다. 특히, 본 연구의 데이터는 행과 열로 이루어진 표 형식의 정형 데이터(tabular data)이므로 이를 잘 처리하는 알고리즘으로 평가받는 트라-기반 머신러닝 모델(Fayaz et al. 2022; Grinsztajn et al. 2022)에 기초하여 각 예측인자(predictor)의 가중값을 스스로 조정하는 과정을 활용하고자 한다.

트라-기반 예측기법의 기초단위는 의사결정 규칙에 따라서 가지를 뺀어 나가 트리 형태의 구조로 나타나는 의사결정나무(decision tree)이다. 의사결정나무는 이해하기

쉽고 결과해석이 용이하며 이상값 처리에 우수하고 편향(bias)³⁾도 낮다는 장점이 있다(Dangeti 2017). 그러나 주어진 데이터에 최적하게 적합화(fit) 되므로 과적합(overfitting)⁴⁾의 경향성이 크기 때문에 일반화에 어려움이 있다는 단점이 지적된다(Dangeti 2017). 이러한 단점을 보완하기 위해 단일 트리를 생성하는 대신 다수의 의사결정나무 결과를 결합하여 예측결과를 얻는 앙상블(ensemble) 모형을 사용할 수 있다. 대표적인 앙상블 기법인 부스팅(boosting)은 이전 모델의 예측결과에 따라 데이터에 가중치가 부여되어 다음 모델에게 영향을 주면서 오차를 보완하는 방식이다(Dangeti 2017). 부스팅 기법은 과적합에 강한 장점이 있어 높은 일반화 성능을 기대할 수 있으나 순차적(sequential)으로 결합해 나가는 방식이므로 훈련속도가 느리다는 단점이 있다(박혜선 2020).

이에 본 연구에서는 시스템 최적화와 머신러닝 원칙을 고려하여 개발된 XGBoost (eXtreme Gradient Boosting) 알고리즘(Chen et al. 2015)을 활용하고자 한다. 일반적으로 머신러닝 알고리즘에서는 고려되는 예측인자의 수가 증가할수록 연산의 효율성이 떨어진다. 그러나 XGBoost 알고리즘은 계산량 극대화(extreme)와 병렬학습을 지원하여 모형의 연산 효율성 및 추정력을 높였다(Chen et al. 2015). 무작위 표본 추출, 불필요한 가지(branch) 제거 등의 자체적인 규제(regularization) 장치를 통해 분산을 줄이고 과적합 문제를 개선하므로 일반화에도 유용하다(Dangeti 2017; Wade & Glynn 2020). 또한 전통적인 계량분석 기법보다 우수한 적합성을 보인다고 알려져 있다(김은미 외 2020; 홍기혜 2024). 다만 XGBoost 알고리즘도 앙상블 모형의 근본적인 문제인 해석의 어려움과 데이터가 작은 경우 여전히 과적합 가능성이 있다는 단점이 지적된다.

식(2)는 XGBoost의 목적함수를 이해하기 위한 $\hat{y}_i$ 을 정의한다. $\hat{y}_i$ 는 예측인자 x_i 가 주어졌을 때 예측값이고 K 는 트리의 개수, f 는 개별 트리 모델을 의미한다.

$$\hat{y}_i = \sum_{k=1}^K f_k(x_i), f_k \in F \quad (2)$$

XGBoost 목적함수는 식(3)과 같이 손실함수와 정규화(regularization) 함수의 합

3) 예측값이 실제값과 차이를 보이는 정도.

4) 모델이 훈련 데이터에 과하게 맞추어져 있어 훈련 데이터에만 정확한 예측을 제공하는 것.

으로 구성된다.

$$obj(\theta) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (3)$$

위 식에서 $l(y_i, \hat{y}_i)$ 는 실제값과 예측값 차이를 의미하는 손실함수를 의미한다. 정규화 함수에 해당하는 $\Omega(f_k)$ 는 다음과 같이 정의한다. T 는 복잡도를 조정하는 트리의 크기(the number of terminal nodes or leaves)이고 γ 는 트리 크기에 대한 정규화 항의 가중치 모수이다. ω_j 는 j 번째 리프(leaf)의 가중치이다. λ 는 리프 노드의 가중치에 대한 L2 정규화 항의 가중치 모수이다. 이러한 정규화 함수는 트리의 크기가 너무 커지는 것에 따라 과적합이 발생하는 경우를 방지해 준다. γ 가 클수록 트리의 크기가 작아지며 λ 가 클수록 각 리프 노드에서 가중치가 더 많이 축소된다.

$$\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T \omega_j^2 \quad (4)$$

XGBoost는 최근 들어 사회과학 분야에서 활용되기 시작하였다. 홍기혜(2024)는 개인의 기부행동을 예측하는 모형을 개발하였는데 3개의 트리-기반 분류모델 가운데 XGBoost가 가장 우수한 성능지표(accuracy 81.38%, precision 68.28%, AUC-score 85.96%)를 갖는 것으로 보고하였다. 학업성취를 판별하는 모델을 구축한 양자영 외(2022)의 연구에서는 5개의 예측모델 가운데 XGBoost가 국어(F1 78%), 영어(F1 83%), 수학(F1 82%) 점수 예측에 가장 우수한 성능을 보이는 것으로 파악되었다. 문지영·조명근(2023)의 연구에서도 XGBoost는 학업 중도탈락의 예측에 가장 우수한 결과(recall 96.9%)를 보이는 것으로 나타났다.

선거와 관련해서는 미국 대통령 선거를 예측한 연구에서 7개의 분류모델 가운데 XGBoost가 가장 우수한 성능(accuracy 96%, AUC 97%)을 보이며 다른 모델과는 달리 예측변수의 의미를 충분하게 활용한다고 밝혔다(Feng et al. 2023). Nagesh (2019)는 인도의 선거 후보자 군집을 예측하는 모형에서 XGBoost가 100%의 정확도(accuracy)를 보였다는 결과를 얻었으며, Bach et al.(2021)는 온라인 활동요인과 선거참여의 관계를 예측하는 연구에서 XGBoost 알고리즘을 활용하였다.

IV. 실증분석 결과

1. 예측성과 비교

본 연구는 19~21대 국회의원 선거결과(보수후보 지지율)를 종속변수로 설정한다. 다만 지지율은 0~1 범위를 가지고 0과 1 근처의 경계값에 대한 예측문제를 고려하여 로짓변환(logit transformation)한 값을 종속변수로 사용한다.⁵⁾ 성별·연령(14개 변수) 그리고 성별·세대(10개 변수)를 예측인자로 설정하여 XGBoost 알고리즘을 적용하였다.⁶⁾ 종속변수가 연속형이므로 파이썬의 XGBRegressor 모듈을 사용하였다.⁷⁾ 세 번의 선거에서 보수정당 후보가 출마하여 득표한 기록이 있는 3,302개 읍·면·동 표본을 이용한다. 먼저 과적합을 규제하기 위해 하이퍼-모수(Hyper-parameter)를 결정해야 한다. 5,000번의 랜덤-서치(random search)와 5겹-교차검증(cross-validation)을 통해 최적의 조합을 탐색하였다. <표 3>에서는 결정된 하이퍼 모수를 정리한다.

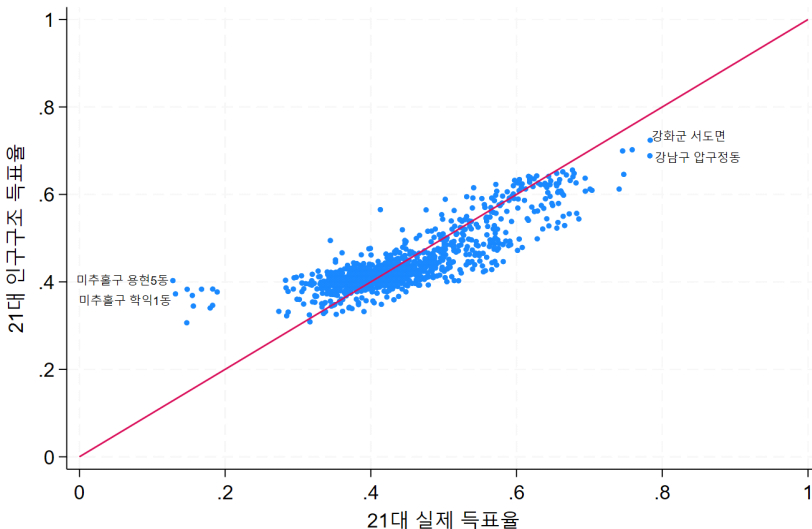
<표 3> XGBoost 알고리즘에서 하이퍼-모수 탐색

Hyper-parameter 유형	Hyper-parameter 값
max_depth: 트리의 최대 깊이(tree size)	4
n_estimators: 생성할 트리 모형의 수	44
learning_rate: 학습 단계별 이전 결과의 반영 비율	0.2
min_child_weight: 관측치에 대한 가중치 합의 최소	3
gamma: 리프(leaf) 노드의 추가 분할(split)을 결정한 최소손실 감소값	0.1

주: 하이퍼-모수 유형은 Scikit-Learn Reference와 홍기혜(2024)를 참고하여 연구자가 작성.

5) 지지율이 p 인 경우 로짓변환 종속변수 값은 $\log\left(\frac{p}{1-p}\right)$ 가 된다.
6) 예측인자로 연령·세대의 투표율은 고려하지 않고 인구구성만 사용한 이유는 다음과 같다. 종속변수는 (투표율이 반영된) 보수후보 지지율이다. 연령·세대의 투표율이 동일하지 않고 가령 고령층이 상대적으로 투표율이 높다면 이러한 투표율이 반영되고 XGBoost의 가중치가 결정될 것으로 예상된다.
7) 종속변수가 범주형인 경우에는 XGBClassifier 모듈을 사용할 수 있다.

<표 3>의 하이퍼-모수 값을 적용하여 훈련 데이터(19~21대 선거)의 예측값을 얻을 수 있다. 앞서 논의하였듯이 예측된 보수후보 지지율은 인구구조에 의해서만 결정된 값으로 이해할 수 있다. <그림 5>에서는 21대(2020) 선거의 보수후보 실제 득표율과 XGBoost 알고리즘에 의해 예측된 득표율의 차이를 보여준다. Y축은 인구구조에 의해 예측된 득표율이고 X축은 해당 읍·면·동의 실제 보수후보 득표율이다. 모든 읍·면·동이 45도 선상에 있다면 완벽하게 예측하고 있다고 해석할 수 있다. 1,111개 읍·면·동의 예측값과 실제값의 상관계수는 0.846으로 양의 상관관계가 매우 크게 나타난다. 수도권 읍·면·동에서 보수후보 지지율 예측은 대부분 0.3~0.7 범위 내에 존재한다. 인천 강화군 서도면의 경우 실제 득표율은 78.3%이지만 인구구조에 의해 예측된 득표율은 72.3%이다. 강남구 압구정동 역시 실제 득표율은 78.2%이고 예측득표율은 68.8%이다. 45도선 아래쪽에 있는 읍·면·동은 실제 득표율이 인구구조 예측득표율보다 높은 지역이다. 해당 지역에서는 인구구조 외 다른 요인(현직자 효과, 경제적 투표성향 등)이 보수후보 득표율에 긍정적인 영향을 미치고 있다고 예상할 수 있다. 반면 인천 미추홀구 용현5동과 학익1동은 실제 득표율은 13% 정도이지만 인구구조 예측득표율은 37~40%로 더 높다는 것을 알 수 있다. 45도선 위쪽에 있는 읍·면·동은 인구구조 외 다른 요인에 의해 보수후보의 득표율이 낮아지는 지역으로 이해할 수 있다.



<그림 5> 21대 선거 보수후보 실제 득표율과 예측득표율

표본 내 예측성과를 비교하기 위해 다음과 같이 인구구조 변수와 시·도/선거시점 고정효과를 포함한 로지스틱 회귀모형을 설정한다. 모형 1에서 성별-연령비율과 성별-세대비율 변수만 포함한다. 모형 2에서는 시·도(서울, 경기, 인천)와 선거시점(19대, 20대, 21대) 고정효과를 포함한다.⁸⁾ 종속변수 XGBoost 모형과 마찬가지로 보수 후보 지지율에 대한 로짓변환된 값을 사용한다.

모형 1.
$$\log\left(\frac{p}{1-p}\right) = \alpha + \sum_k \beta_k mAGE_k + \sum_h \gamma_h mCohort_h + e$$

모형 2.
$$\log\left(\frac{p}{1-p}\right) = \alpha + \sum_k \beta_k mAGE_k + \sum_h \gamma_h mCohort_h + \delta Sido + \theta Time + e$$

<표 4>에서는 본 연구에서 제시한 예측방법론인 XGBoost 및 모형 1, 2에서 도출된 21대 국회의원 선거 보수후보 지지율과 예측된 지지율 간 상관계수를 비교한다. 선형회귀모형에서는 0.649~0.650 상관계수이지만 XGBoost 알고리즘하에서는 0.846으로 크게 높다는 것을 확인할 수 있다. 인구구조 변수만 이용하더라도 본 연구의 알고리즘을 활용하였을 때 충분히 실제 보수후보 지지율과 가깝게 예측하는 것이 가능하다는 것을 보여준다.

<표 4> XGBoost와 선형회귀모형 비교

	XGBoost	모형 1	모형 2
adjusted R^2	-	0.347	0.428
$corr(p_{21}, \hat{p}_{21})$	0.846	0.650	0.649

2. 중요도 분석 및 인구구조 지수

트라-기반 알고리즘에서는 종속변수를 예측하는 데 있어 각 예측인자가 얼마나 유용한지 중요도(feature importance)를 도출하여 확인할 수 있다. 본 연구에서는 순열 중요도(permutation feature importance)보다 정확한 측정이 가능하며 예측인자의 효과(effect)

8) 회귀모형에서는 연령비율과 세대비율 합이 각각 1이 되는 점을 고려하여 female_20age(여성-20대)와 female_cohort5(여성-후기 MZ) 변수는 제외한다.

도 파악할 수 있는 SHAP(SHapley Additive Explanations) 요약도표(summary plot)를 활용한다(Lundberg & Lee 2017). SHAP 값은 특정 사례(case)에 대한 예측에 있어 각 예측인자의 공정한 기여도를 계산하기 때문에 머신러닝 모델의 예측에 이론적으로 타당한 방법을 제공하는 것으로 알려져 있다(Lundberg & Lee 2017).

특정 예측인자 x_i 의 샤플리 값 Φ_i 은 다음과 같이 쓸 수 있다. 식(5)에서 제시한 샤플리 값은 사용 가능한 모든 예측인자 x 조합에 대해 특정 예측인자 x_i 가 얼마나 예측값을 변화시키는지 평균한 것으로 예측인자의 기여도를 공평하게 평가할 수 있는 장점이 있다.

$$\Phi_i = \sum_{S \subseteq F} \frac{|S|!(|F|-|S|-1)!}{|F|!} [(f_{S \cup i}(x_{S \cup i}) - f_S(x_S))] \quad (5)$$

where S : 전체 표본에서 x_i 를 제외한 나머지 x 변수의 모든 부분집합

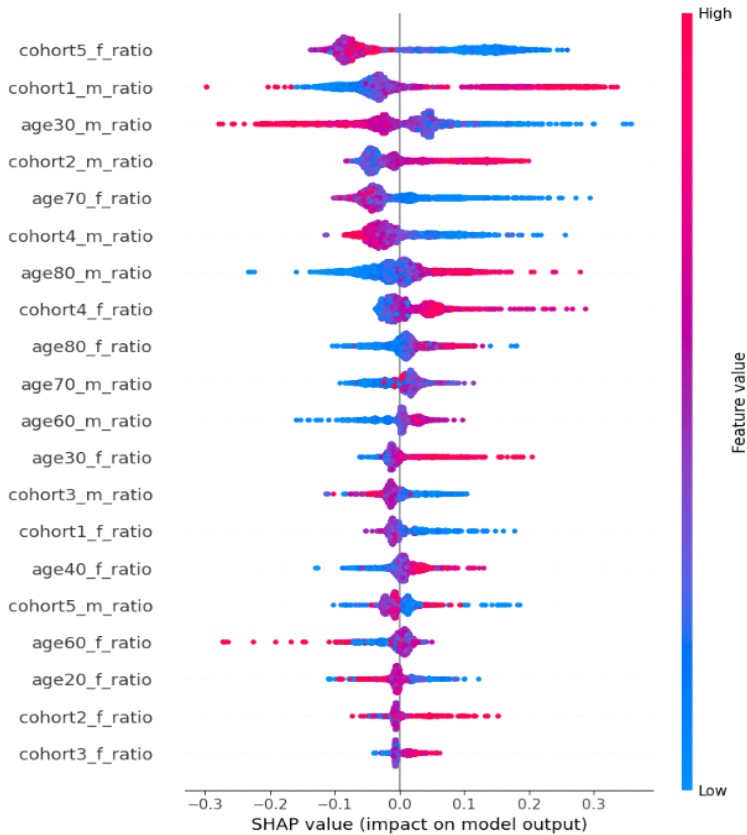
$|S|$: 집합 S 에 포함된 x 변수의 수

$|F|$: 전체 x 변수의 수

$f_{S \cup i}(x_{S \cup i})$: 집합 S 에 x_i 를 추가했을 때 모델 f 의 예측값

$f_S(x_S)$: 집합 S 만 사용했을 때 모델 f 의 예측값

SHAP 요약도표(SHAP summary plot)는 예측인자가 모델 출력에 어떤 영향을 미치는지 요약한 정보를 시각적으로 표현한다. 먼저 성별-연령, 성별-세대의 교호작용(interaction) 변수의 영향력을 분석하기에 앞서 성별, 연령, 세대의 주효과(principal effect)에 대해 XGBoost 알고리즘을 적용하였다. 이에 대한 SHAP 요약도표 결과는 부록에서 제시한다. 해당 도표에 따르면 후기 MZ세대(cohort5_ratio)가 예측에 미치는 영향력이 가장 크고 해당 세대는 보수후보 지지에 부정적 영향을 미친다. 전통세대(cohort1_ratio)와 베이비붐세대(cohort2_ratio) 역시 보수후보 지지율 예측에 있어서 중요한 변수임을 알 수 있다. 다만 같은 연령-세대일지라도 성별로 지지성향의 차이를 고려하기 위해 본 연구에서는 교호작용 변수를 이용하여 예측모형을 설정한다. 본 연구에서 사용한 예측인자 24개에서 얻은 요약도표(상위 20개 결과만 제시)는 <그림 6>에서 제시한다.



<그림 6> SHAP 요약도표: 전체 변수

인구구조 변수 가운데 보수후보 지지율에 영향을 미치는 가장 중요한 예측인자는 여자-후기 MZ세대(cohort5_f_ratio)이다. X축 0.0을 기준으로 오른쪽에 파란색으로 분포되어 있다는 것은 해당 예측인자의 비율이 커질수록(해당 유권자 비율이 증가할수록) 보수후보 지지율에 음(-)의 영향을 미친다고 해석한다. 두 번째로 중요한 인구구조는 남자-전통세대(cohort1_m_ratio)이다. 해당 인구구성이 증가할수록 보수후보 지지율에 양(+)의 영향을 미친다고 이해한다. 세 번째로 중요한 인자는 남자-30대(age30_m_ratio)이다. 여자-후기 MZ세대(cohort5_f_ratio)와 마찬가지로 해당 인구비율이 증가할수록 보수후보 지지율에 음(-)의 영향을 미친다. 이와 같이 가장 중요하게 영향을 미치는 상위 3개 변수 중 2개 요인이 남자 유권자와 관련이 있다는 것을 확인할 수 있다. 반면 여자-베이비붐세대(cohort2_f_ratio)와 여자-X세대(cohort3_f_

ratio)는 보수후보 지지율 예측에서 중요도가 매우 낮은 변수라는 것을 알 수 있다.

같은 세대임에도 불구하고 남녀 간에 차이가 있다는 것 또한 확인할 수 있다. 남자-전기 MZ세대(cohort4_m_ratio)는 보수후보에 부정적인 영향이 있지만, 여자-전기 MZ세대(cohort4_f_ratio)는 보수후보에 긍정적인 영향을 미친다는 것을 확인할 수 있다. 남자-50대(age50_m_ratio)와 남자-베이비붐세대(cohort2_m_ratio) 변수를 비교하면 남자-50대(age50_m_ratio)는 보수후보 지지율 예측에 큰 영향이 없지만, 남자-베이비붐세대(cohort2_m_ratio)는 지지율 예측에 매우 큰 영향을 미치는 것을 알 수 있다. 19대, 20대 선거에서 50대 중 후반은 21대 선거에서는 60대 그룹으로 넘어가서 50대 남자의 인구구성이 변경되기 때문에 이러한 결과가 도출된 것으로 설명할 수 있다.

3. 인구구조 변화와 22대 국회의원 선거

22대(2024) 국회의원 선거는 21대 선거에서 다시 4년이 지난 시점이기 때문에 각 읍·면·동은 인구구조 변화를 겪게 된다. 이러한 변화의 원인은 첫째, 기존 유권자가 시간이 지남에 따라 새로운 연령 그룹에 속하게 되기 때문이다. 둘째, 21대 선거에서는 연령 미달로 투표하지 못했던 새로운 젊은 유권자들이 참여하게 된다. 셋째, 지난 4년 동안 신도시 개발 등 인구이동으로 인해 읍·면·동의 인구구조 변화가 발생하게 된다. IV장 1절에서 XGBoost 알고리즘으로 도출한 24개 인구구조 변수의 가중치를 이용하여 22대 선거시점 인구구조에서 보수후보 지지율을 예측하고자 한다. 다만 2024년 4월 선거시점이 아직 도래하지 않았기 때문에 본 연구에서는 2023년 10월 시점 인구구조를 대신 사용한다.⁹⁾

21대(2020)와 22대(2024) 인구구조 변화로만 판단했을 때 보수후보에게 유리하게 변화했는지 아니면 불리하게 변화했는지 판단할 수 있다. 식(6)과 같이 $conserv_i^{adv}$ 변수를 정의하자. 이 변수는 인구구조에 의해서만 예측된 22대 선거의 보수후보 지지율과 21대 선거 보수후보 지지율의 차이로 정의한다.¹⁰⁾

$$conserv_i^{adv} = \widehat{conserv}_{22,i} - \widehat{conserv}_{21,i} \quad (6)$$

9) 2023년 10월 시점 읍·면·동별 5세 간격 성별-연령 인구 데이터는 통계청 KOSIS 사이트에서 받을 수 있다.

10) $\widehat{conserv}_{21,i}$ 는 21대 선거의 실제 보수후보 득표율이 아니라 XGBoost 알고리즘으로 예측한 21대 보수후보 지지율이다.

<표 5>에서는 21대와 22대 국회의원 선거시점을 보수후보 지지율 차이인 $conserv_i^{adv}$ 가 양(+)인 읍·면·동과 음(-)인 읍·면·동 수를 비교한다. 수도권 1,094개 읍·면·동 가운데 보수후보에 유리하게 인구구조로 변한 읍·면·동은 834개로 전체의 76.2%이다¹¹⁾. 인구구조로만 판단하였을 때 22대(2024년 4월 예정) 국회의원 선거에서는 보수정당 후보의 득표율이 21대에 비해 상승할 것으로 예상된다.

<표 5> 보수정당 후보에게 유리/불리하게 인구구조가 변한 읍·면·동 수

$conserv_i^{adv} > 0$ 보수후보에게 유리한 인구구조 변화	$conserv_i^{adv} < 0$ 보수후보에게 불리한 인구구조 변화	합계
834 (76.2%)	260 (23.7%)	1,094 (100%)

<표 6> $conserv_i^{adv}$ 변수의 상/하위 5개 읍·면·동

상위 5개 읍·면·동				하위 5개 읍·면·동			
읍·면·동	21대	22대	$conserv_i^{adv}$	읍·면·동	21대	22대	$conserv_i^{adv}$
서초구 방배본동	0.498	0.698	0.200	강화군 송해면	0.633	0.435	-0.198
성동구 응봉동	0.449	0.648	0.198	연천군 백학면	0.642	0.451	-0.191
종로구 교남동	0.426	0.623	0.196	웅진군 대청면	0.590	0.408	-0.181
강북구 삼각산동	0.395	0.589	0.193	동두천시 상패동	0.576	0.400	-0.175
수지구 상현1동	0.442	0.630	0.188	포천시 영북면	0.538	0.375	-0.162

<표 6>에서는 식(6)에서 정의한 보수후보에 유리/불리하게 인구구조가 변화한 5개 읍·면·동을 정리하여 보여준다. ‘상위 5개 읍·면·동’은 지난 4년간 보수후보에게 유리하게 인구구조가 변한 읍·면·동 5개를 제시한다. 강북구 삼각산동은 21대 시

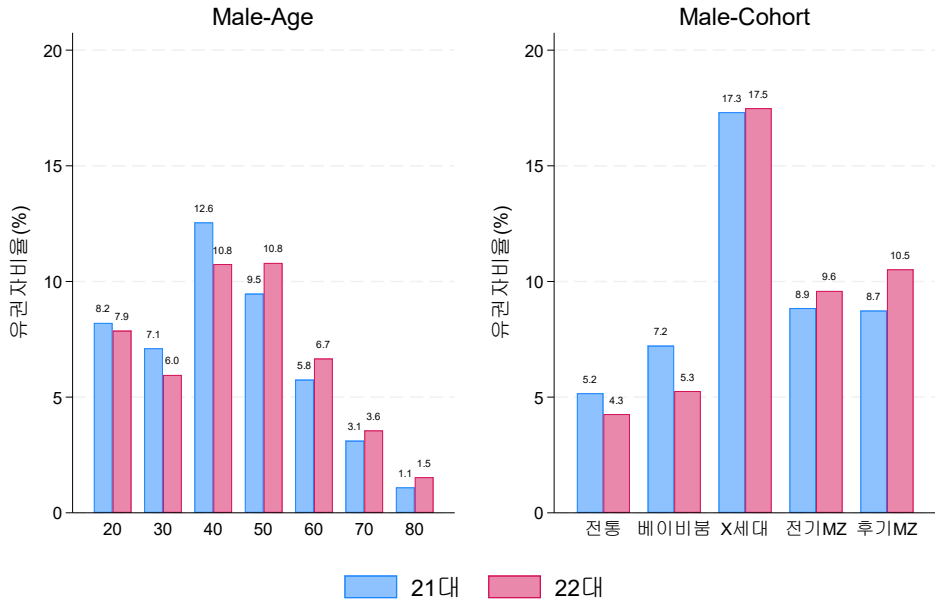
11) 1,094개 읍·면·동은 21대와 22대에 공통으로 데이터를 수집할 수 있는 읍·면·동 수이다.

점 인구구조에서는 보수후보 예측 지지율이 0.395이지만 2023년 10월 시점 인구구조에서는 보수후보 예측 지지율이 0.589로 19.3%p 상승한다. ‘하위 5개 읍·면·동’은 지난 4년간 보수후보에게 불리하게 인구구조가 변한 읍·면·동으로 이해할 수 있다. 강화군 송해면의 경우 21대 시점 인구구조에서는 0.633이지만 2023년 10월 시점에서 보수후보 지지율은 0.435로 예측된다. 따라서 인구구조 변화로만 판단하면 보수후보 지지율이 19.8%p 하락한다. 웅진군 대청면 역시 $0.590 \Rightarrow 0.408$ 로 18.1%p 하락한다.

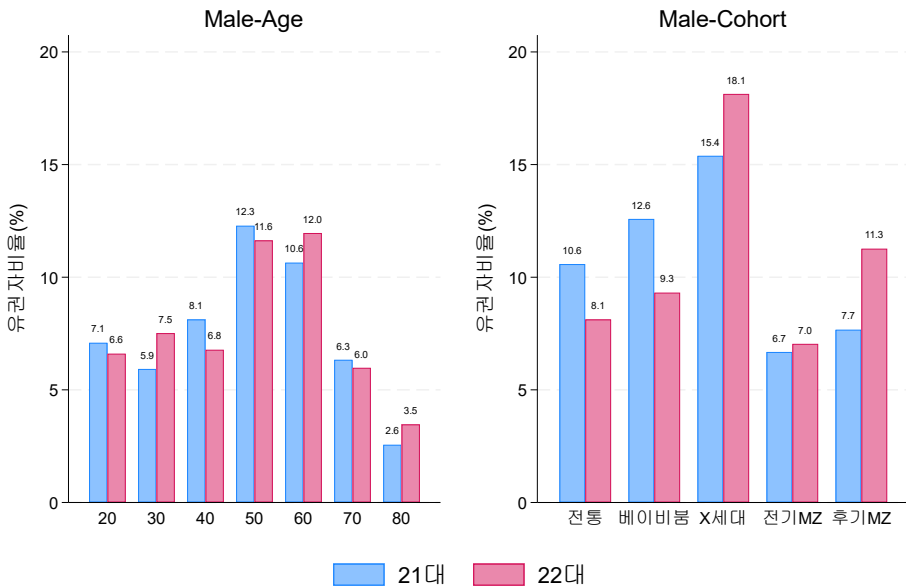
<표 6>에서 제시한 보수후보 유리/불리 5개 읍·면·동 중 1개씩을 선택하여 21대와 22대 선거시점에서 인구구조 변화의 구체적인 양상을 살펴보고자 한다. <그림 7>에서는 보수후보에 유리하게 인구구조가 변한 ‘서울특별시 강북구 삼각산동’ 인구구조를 막대그래프로 제시한다. ‘강북구 삼각산동’은 2017년 9월 우이신설선이 개통되어 교통이 개선되었고 이에 따른 인구구조 변화가 발생할 수 있다. 지면 제약 상 남자-연령과 남자-세대에 대한 인구구조 변화만 제시하고자 한다. 남자-연령 막대그래프에서 20~40대 남자 비율이 줄어들었고 50대 이상 모든 연령층에서 21대 선거 시점과 비교하여 22대 시점에서 유권자 비율이 증가하였다는 것을 알 수 있다. 남자-세대 막대그래프에서는 22대 시점에서 전통/베이비붐세대 비율이 감소하였고 전기/후기 MZ세대 비율은 오히려 증가하였음을 확인할 수 있다.

<그림 8>에서는 보수후보에 불리하게 인구구조가 변한 읍·면·동 가운데 ‘동두천시 상패동’을 선택하여 21대와 22대 인구구조를 비교한 막대그래프를 제시한다. 남자-연령 막대그래프의 특징은 30대와 60대 남자 비중이 비교적 크게 증가하였다는 것을 발견할 수 있다. 남자-세대 막대그래프에서는 전통/베이비붐세대 비율이 크게 감소한 반면 X세대, 전기/후기 MZ세대 비율이 크게 증가한 것을 확인할 수 있다.

<그림 7> 보수후보에게 유리한 인구구조 변화와 <그림 8> 보수후보에게 불리한 인구구조 변화의 남자-세대(male-cohort) 막대그래프에서 남자-전기 MZ세대는 공통적으로 증가한다. <그림 6> SHAP 요약도표로 판단할 때 남자-전기 MZ세대(cohort4_m_ratio)의 증가는 보수후보 지지율에 음(-)의 영향을 미친다. 따라서 ‘동두천시 상패동’에서는 보수후보에 불리한 인구구조가 된다. 그러나 남자-전기 MZ세대의 증가에도 불구하고 ‘강북구 삼각산동’에서는 50~80대 고령자 증가로 이와 같은 영향력이 상쇄되어 결과적으로 보수후보에 유리한 인구구조를 초래한 것으로 예상할 수 있다.



<그림 7> 보수후보에게 유리한 인구구조 변화: **강북구 삼각산동**



<그림 8> 보수후보에게 불리한 인구구조 변화: **동두천시 상패동**

V. 결론 및 시사점

20대 대통령 선거(2022) 이후 우리 사회에서 성별, 연령, 세대에 따른 지지성향은 기존과는 미세한 차이를 드러내고 있다. 저출산·고령화는 인구구조의 급격한 변화와 세대/젠더 갈등을 반영한 후보자 선택을 분석하기 위해 본 연구에서는 전통적 투표이론인 성별, 연령, 세대(이념)에 기반한 인구구성과 선거의 관계를 다시 조명하였다. 특히 최근 치러진 3번의 국회의원 선거 데이터에 기초하여 인구구조의 동적 변화에 의해 예측되는 22대(2024년 4월 예정) 선거결과를 도출하는 것이 실증분석의 주요한 목적이다.

분석 결과를 요약하면 다음과 같다. 첫째, 20대(2016)과 21대(2020) 국회의원 선거 시점을 비교하면 남·녀 공통적으로 30~50대 유권자 비율이 감소한다. 그리고 60대 이상과 20대(18~29세) 유권자 비율이 증가함을 확인할 수 있다. 특히 60대 남성의 비율은 6.07%(20대 선거) $\Rightarrow$ 7.29%(21대 선거)로 크게 증가하였고 60대 여성의 비율 역시 6.33%(20대 선거) $\Rightarrow$ 7.50%(21대 선거)로 크게 상승하였음을 확인할 수 있다. 이러한 인구구조 변화는 60대 이상 고령층 지지를 받는 정당이 유리한 구조로 변하고 있고 이러한 인구구조의 추세는 22대 선거에서도 지속될 것으로 예상된다.

둘째, 과거 3번의 국회의원 선거 데이터에 기초하면 인구구조 변수 가운데 보수후보 지지 예측에 영향을 미치는 가장 중요한 예측인자는 ‘여자-후기 MZ세대’와 ‘남자-전통세대’이다. “여자-후기 MZ세대” 비율이 증가할수록 보수후보 지지율에 음(-)의 영향을 미치며 ‘남자-전통세대’ 비율이 증가할수록 보수후보 지지율에 양(+)의 영향을 미친다. ‘남자-베이비붐세대’ 그룹 역시 보수후보 지지에 중요한 예측인자이다. ‘남자-30대’는 세 번째로 중요한 예측인자이지만 이들의 비율이 증가할수록 보수후보 지지율은 음(-)으로 영향을 받게 된다. 24개 인구구조 변수 중에서 예측중요도가 높은 4개 중 3개 변수 모두 남자 유권자와 관련이 있다는 것을 확인할 수 있다.

셋째, 머신러닝 알고리즘인 XGBoost를 활용하여 22대 국회의원 선거 보수후보 지지율을 예측하였다. 특히 인구구조에 의해서만 설명되는 보수후보 지지율을 21대 선거시점과 비교하였을 때 수도권 1,094개 읍·면·동 가운데 보수후보에 유리하게 인구구조가 바뀐 읍·면·동은 834개로 전체의 76.2%이다. 인구구조로만 판단하였을

때 수도권 선거구에서는 보수정당 후보의 득표율이 21대에 비해 22대 선거에서 더 높아질 것으로 예상한다.

최근 우리나라 선거에서는 지역주의 성향은 점차 벌어지고 있으나 성별, 연령, 세대에 의해 결정되는 인구구성은 여전히 선거결과를 예측하는 데 중요한 인자로 작동하고 있다는 것을 본 연구에서는 보여주고 있다. 특히 지역주의 투표성향이 상대적으로 적은 수도권에서 이러한 분석 결과에 기초하여 선거전략이나 공약 개발에 활용될 수 있다는 점을 시사한다. 저출산·고령화라는 인구구조 변화뿐 아니라 서울집중과 지방소멸이라는 사회적 문제에 대처하기 위한 인위적인 인구구성 변화 정책도 선거결과에 영향을 미칠 수 있다는 것을 보여준다. 연령이 높고 낮음에도 불구하고 정책선호와 투표행태가 결정되는 것은 여전히 세대의 공통된 경험이 중요하다는 것을 본 연구는 시사하고 있다. 다만 지역주의 퇴색이 점차 분명해지는 추세 속에서 세대 및 성별 갈등을 부추기는 정치 이데올로기는 정치 혐오와 정치 양극화를 가속화시킬 수 있다는 점을 경계해야 할 필요가 있다.

본 연구는 최근 3번의 국회의원 선거(2012, 2016, 2020) 데이터에 기초하여 각 읍·면·동의 선거시점 데이터를 독립적으로 가정하고 있다. 그러나 읍·면·동의 선거결과는 군집화(clustering)된 상관성이 영향을 미칠 수 있다는 점을 고려할 필요가 있다. 또한 국회의원 선거에서 지지율은 인구구조보다 지역투표 성향에 의해 더 크게 영향을 받을 것으로 예상된다. 이러한 지역투표 성향의 한계를 고려하여 수도권으로 제약하여 분석 결과를 제시하고 있지만 여전히 수도권 내에서도 이질적인 투표 성향이 나타날 수 있다. 이러한 연구 한계를 보완 및 극복하기 위한 관련 연구가 지속될 필요가 있다.

참고문헌

- 강우진. 2012. “경제정책에 대한 인식과 주택소유 형태가 투표불참에 미치는 영향 연구: 18대 총선의 경우.” 《한국정당학회보》 11(2): 67-94.
- 강원택·성예진. 2018. “2017년 대통령 선거에서 이념과 세대: 보수 성향 유권자를 중심으로.” 《한국정치연구》 27(1): 205-240.
- 구본상. 2017. “ARS 조사방식과 젊은 연령대 여성 표집의 실패: 정치적 의견이 강한 유권자들 비율에서의 성별 차이를 중심으로.” 《조사연구》 18(1): 31-60.

- 김기동·이재묵. 2023. “건강상태와 투표참여의 성별 격차: 소득의 조건효과.” 《정치정보 연구》 26(3): 1-34.
- 김기동·정다빈·이재묵. 2021. “혼인상태와 투표참여: 성별의 조건효과.” 《동서연구》 33(3): 33-56.
- 김난도·전미영·최지혜·이향은·이준영·이수진·서유현·권정운·한다혜·이혜원. 2021. 《트렌드 코리아 2022》 서울: 미래의 창.
- 김우성·허은정. 2007. “베이비붐세대, X세대, Y세대 소비자들의 소비관련 가치관과 라이프스타일의 비교.” 《소비문화연구》 10(4): 31-53.
- 김은미·김상봉·조은서. 2020. “기계학습을 활용한 주택매도 결정요인 분석 및 예측모델 구축.” 《지적과 국토정보》 50(1): 181-200.
- 김정경·박민규. 2023. “중도탈락 후보 지지층과 응답유보층 보정을 통한 선거결과 예측.” 《조사연구》 24(2): 1-15.
- 김태완. 2012. “한국 선거에서의 영호남 지역주의 투표성향에 대한 분석: 1990년대 이후 국가레벨 선거를 중심으로.” 《로컬리티 인문학》 (7): 57-104.
- 김하정·원효현. 2023. “고등학생의 선거참여 영향 요인 탐색: 랜덤 포레스트 기법을 활용 하여.” 《선거연구》 1(19): 57- 86.
- 김학량. 2015. “세대균열과 출신지별 지역주의가 투표행태에 미치는 영향 비교연구: 2014 년 서울시장 선거를 중심으로.” 《현대사회와 행정》 25(3): 77-95.
- 김한나. 2022. “20대 청년 유권자의 젠더 균열과 투표 선택: 제20대 대통령 선거를 중심으로.” 《동향과 전망》 (115): 296-338.
- 노환희·송정민·강원택. 2013. “한국 선거에서의 세대 효과: 1997년부터 2012년까지의 대선을 중심으로.” 《한국정당학회보》 12(1): 113-140.
- 라미경. 2022. “제20대 대통령선거에서 유권자의 여야성향과 투표행태 분석.” 《한국과 세계》 4(3): 87-112.
- 문우진. 2017. “지역주의 투표의 특성과 변화: 이론적 쟁점과 경험분석.” 《의정연구》 23(1): 81-111.
- 문지영·조명근. 2023. “대학생 학업 중단 예방을 위한 기계학습 기반 모형 탐색.” 《교육 평가연구》 36(2): 289-313.
- 민영. 2022. “연령과 세대의 교차: 21대 총선에서 노년층과 청년층의 투표 선택 요인.” 《정치커뮤니케이션 연구》 67: 37-73.
- 민인식. 2022. “주택소유 유형이 정치적 성향과 투표참여 의사에 미치는 영향: 분할 잠재 성장모형 활용.” 《한국정책학회보》 31(1): 231-253.
- 박명호. 2009. “2008 총선에서 나타난 세대 효과와 연령 효과에 관한 분석.” 《한국정당 학회보》 8(1): 65-86.
- 박상운·이호준. 2015. “현직자 우위의 구성요소: 한국의 제 18, 19대 국회의원선거에 대

- 한 경험적 분석.” 《사회과학연구》 31(1): 393-410.
- 박혜선. 2020. 《혼자 공부하는 머신러닝 딥러닝》 서울: 한빛미디어.
- 서현진. 2008. “17대 대통령 선거의 투표 참여와 세대에 관한 연구.” 《의정연구》 14(2): 117-142.
- 송선희. 1986. “여성의 정치참여 형태에 대한 한국과 외국의 비교고찰.” 《여성연구》 통권 12호.
- 신정섭. 2020. “코로나 19가 제21대 국회의원 선거 투표선택에 미친 영향: 정부대응 평가와 개인피해를 중심으로.” 《한국정치연구》 29(3): 155-182.
- 양자영·문경희·박성호. 2022. “인공지능을 이용한 학습부진 특성 추출 및 예측 모델 연구.” *Journal of the Korea Institute of Information & Communication Engineering* 26(4).
- 윤종빈. 2007. “2007 대선과 수도권 투표성향: 지역, 이념, 그리고 인물.” 《한국정당학회보》 6(2): 65-95.
- 윤희영. 2023. “유튜브 선거 활동은 국회의원 당선 여부를 예측할 수 있는가? : 제21대 국회의원 선거 서울시 선거구 살펴보기.” 《방송통신연구》 29-62.
- 이갑윤. 2008. “한국선거에서의 연령과 투표참여.” 《의정연구》 14(2): 93-115.
- 이남영. 1995. “성별이 투표행위에 미치는 영향-제14대 대통령선거결과를 중심으로.” 《한국과국제정치(KWP)》 11(1): 33-52.
- 이내영·정한울. 2013. “세대균열의 구성요소: 코호트 효과와 연령 효과.” 《의정연구》 19(3): 37-83.
- 이우진. 2014. “성과 세대의 정치경제.” 《재정학연구》 7(4): 1-40.
- 이재철. 2013. “유권자의 연령별 투표참여 분석: 제18대 대통령 선거를 중심으로.” 《분쟁해결연구》 11(2): 121-153.
- 이준한. 2013. “한국의 이념성향과 생애주기효과.” 《정치정보연구》 16(2): 1-26.
- 이현아·김주희. 2021. “세대별 기혼남성의 가사노동시간 연구: 베이비붐세대, X세대, Y세대를 중심으로.” 《가족자원경영과 정책 (구 한국가족자원경영학회지)》 25(4): 71-86.
- 정수현. 2017. “후보자의 지역대표성이 득표율과 당선가능성에 미치는 영향력: 제20대 국회의원 선거 결과에 대한 분석.” 《한국정치연구》 26(2): 129-157.
- 정진민. 1992. “한국선거에서의 세대요인.” 《한국정치학회보》 26(1): 145-167.
- 정진민. 1994. “정치세대와 14대 국회의원 선거.” 《한국정치학회보》 28(1): 257-274.
- 정진민. 2012. “한국 유권자들의 투표행태와 세대: 2010년 지방선거를 중심으로.” 《한국정치연구》 21(2): 1-21.
- 최슬기·이윤석·김석호. 2019. “세대별로 투표하는 정당이나 후보는 달라지는가?” 《한국사회》 20(2): 103-130.

- 최종호. 2016. “한국의 자영업자는 어떤 정책을 선호 하는가?: 2007, 2012 대선의 경험적 분석.” 《한국정치학회보》 50(1): 295-323.
- 최종호·강현철·한상태. 2006. “선거예측조사 의사결정유보층 분류 및 예측을 위한 의사결정나무모형의 비교와 평가.” *Journal of The Korean Data Analysis Society* 8(1): 167-178.
- 최종호·서두성·김유진. 2000. “의사결정나무를 이용한 15대 대선 선거예측조사 사례분석.” 《조사연구 춘계학술대회 발표논문집》 93-103.
- 최준영·조진만. 2005. “지역균열의 변화 가능성에 대한 경험적 고찰: 제17대 국회의원선거에서 나타난 이념과 세대 균열의 효과를 중심으로” 《한국정치학회보》 39(3): 375-394.
- 최필선·민인식. 2016. “읍·면·동 멀티레벨 데이터를 이용한 정당우세를 분석모형.” 《조사연구》 17(1): 85-112.
- 허석재. 2014. “세대와 생애주기, 그리고 투표선택: 1992~2012 대통령선거 분석.” 《한국과 국제정치》 30(2): 167-199.
- 홍기혜. 2024. “머신러닝기반 개인 기부행동 예측모형: 세대별 일상적 안녕에 대한 태도를 중심으로.” 연세대학교 일반대학원 사회복지정책협동과정 박사학위논문.
- Ali, H., H. Farman, H. Yar, Z. Khan, S. Habib, and A. Ammar. 2022. “Deep Learning-based Election Results Prediction Using Twitter Activity.” *Soft Computing* 26 (16): 7535-7543.
- Almond, G.A. and S. Verba. 1963. *The Civic Culture: Political Attitudes and Democracy in Five Nations*. New Jersey: Princeton University Press.
- Bach, R.L., C. Kern, A. Amaya, F. Keusch, F. Kreuter, J. Hecht, and J. Heinemann. 2021. “Predicting Voting Behavior Using Digital Trace Data.” *Social Science Computer Review* 39(5): 862-883.
- Balakrishnan, V., M. Kaity, H.A. Rahim, and N. Ismail. 2021. “Social Media Analytics Using Sentiment and Content Analyses on the 2018 Malaysia’s General Election.” *Malaysian Journal of Computer Science* 34(2): 171-183.
- Buntoro, G.A., R. Arifin, G.N. Syaifuddiin, A. Selamat, O. Krejcar, and F. Hamido. 2021. “The Implementation of the Machine Learning Algorithm for the Sentiment Analysis of Indonesia’s 2019 Presidential Election.” *IIUM Engineering Journal* 22(1): 78-92.
- Campbell, A., C. Philip, M. Warren, and S. Donald. 1960. *The American Voter*. New York: Wiley.
- Chen, T., T. He, M. Benesty, V. Khotilovich, Y. Tang, H. Cho, ..., and T. Zhou. 2015.

- “Xgboost: Extreme Gradient Boosting.” *R Package Version 0.4-2* 1(4): 1-4.
- Dangeti, P. 2017. *Statistics for Machine Learning*. Birmingham: Packt Publishing Ltd.
- Fayaz, S.A., M. Zaman, S. Kaul, and M.A. Butt. 2022. “Is Deep Learning on Tabular Data Enough? An Assessment.” *International Journal of Advanced Computer Science and Applications* 13(4): 466-473.
- Feng, G., H. Cai, K. Chen, and Z. Li. 2023. “A Hybrid Method of Sentiment Analysis and Machine Learning Algorithm for the US Presidential Election Forecasting.” *arXiv e-prints* arXiv-2312.
- Fitriani, A.S., N.E. Pratama, A.B. Raharjo, Y. Purwananto, and D. Purwitasari. 2022. “A Comparative Study on Machine Learning Based Prediction Models for Public Participation Rate in an Election Voting.” in *2022 3rd International Conference on Electrical Engineering and Informatics (ICon EEI)* (pp. 86-91). IEEE.
- Grinsztajn, L., E. Oyallon, and G. Varoquaux. 2022. “Why Do Tree-based Models still Outperform Deep Learning on Typical Tabular Data?.” *Advances in Neural Information Processing Systems* 35: 507-520.
- Guan, L. and G. Mani. 2022. “Election Forecasting Using Macroeconomic and Social Indicators via Machine Learning.” *Journal of Student Research* 11(3).
- Inglehart, R. 1977. *The Silent Revolution: Changing Values and Political Styles among Western Publics*. New Jersey: Princeton University Press.
- Inglehart, R. and P. Norris. 2000. “The Developmental Theory of the Gender Gap: Women’s and Men’s Voting Behavior in Global Perspective.” *International Political Science Review* 21(4): 441-463.
- Joyce, B., and J. Deng. 2017. “Sentiment Analysis of Tweets for the 2016 US Presidential Election.” in *2017 Ieee Mit Undergraduate Research Technology Conference (urtc)* (pp. 1-4). IEEE.
- Keniston, K. 1969. “Moral Development, Youthful Activism and Modern Society.” *Youth & Society* 1(1): 110-127.
- Khandelwal, A., S. Swami, S.S. Akhtar, and M. Shrivastava. 2017. “Classification of Spanish Election Tweets (COSET) 2017: Classifying Tweets Using Character and Word Level Features.” in *CEUR Workshop Proceedings* (Vol. 1881, pp. 49-54). CEUR-WS.
- Lambert, T.A. 1972. “Generations and Change: Toward a Theory of Generations as a Force in Historical Process.” *Youth & Society* 4(1): 21-45.
- Lewis-Beck, M.S. and R. Nadeau, 2011. “Economic Voting Theory: Testing New Dimensions.” *Electoral Studies* 30(2): 288-294.

- Lundberg, S.M. and S.I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems* 30.
- Mannheim, K. 1952. *Essays on the Sociology of Knowledge*. London: Routledge & Kegan Paul.
- Myilvahanan, K., P. Yashas, S. Pasha, M. Ismail, and V. Tharun. 2023. "A Study on Election Prediction Using Machine Learning Techniques." in *2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)* (pp. 1518-1520). IEEE.
- Nagesh, J.B.M. 2019. "Generating Political Scores Using rfm Model and Cluster Prediction by Xgboost." in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 1333-1336). IEEE.
- Oyebode, O. and R. Orji. 2019. "Social Media and Sentiment Analysis: The Nigeria Presidential Election 2019." in *2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)* (pp. 0140-0146). IEEE.
- Park, J., M. Cheon, S. Hou, and O. Lee. 2022. "Forecasting Election Result via Artificial Intelligence Approach: NLP and Machine Learning." in *Proceedings of International Conference on Communication and Computational Technologies: ICCCT 2022* (pp. 761-768). Singapore: Springer Nature Singapore.
- Richardson, B. and D.F. Hougen. 2020. "Districts by Demographics: Predicting US House of Representative Elections Using Machine Learning and Demographic Data." in *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 833-838). IEEE.
- Ryder, N.B. 1965. "The Cohort as a Concept in the Study of Social Change." *American Sociological Review* 30: 843-861.
- Singh, A., N. Dua, V.K. Mishra, D. Singh, and A. Agrawal. 2021. "Predicting Elections Results Using Social Media Activity a Case Study: USA Presidential Election 2020." in *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)* (Vol. 1, pp. 314-319). IEEE.
- Sinha, P., A. Verma, P. Shah, J. Singh, and U. Panwar. 2022. "Prediction for the 2020 United States Presidential Election Using Machine Learning Algorithm: Lasso Regression." *The Journal of Prediction Markets* 16(1).
- Tsai, M.H., Y. Wang, M. Kwak, and N. Rigole. 2019. "A Machine Learning Based Strategy for Election Result Prediction." in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 1408-1410).

IEEE.

Verba, S., N.H. Nie, and J.O. Kim. 1978. *Participation and Political Equality: A Seven-nation Comparison*. New York: Cambridge University Press.

Wade, C. and K. Glynn. 2020. *Hands-On Gradient Boosting with XGBoost and Scikit-learn: Perform Accessible Machine Learning and Extreme Gradient Boosting with Python*. Birmingham: Packt Publishing Ltd.

Yadav, A.K. and R. Johari. 2021. “Trend Analysis and Predictive Modeling Using Machine Learning Models on Indian Election Historical Dataset.” *Innovations in Information and Communication Technologies (IICT-2020)* 319.

<접수 2024.01.19; 수정 2024.03.08; 게재확정 2024.04.03>

Impact of Dynamic Changes in Gender, Age, and Generational Composition on Elections: Focusing on the 22nd National Assembly Election

Kihye Hong

(Yonsei University, Kakdang Social Welfare Foundation)

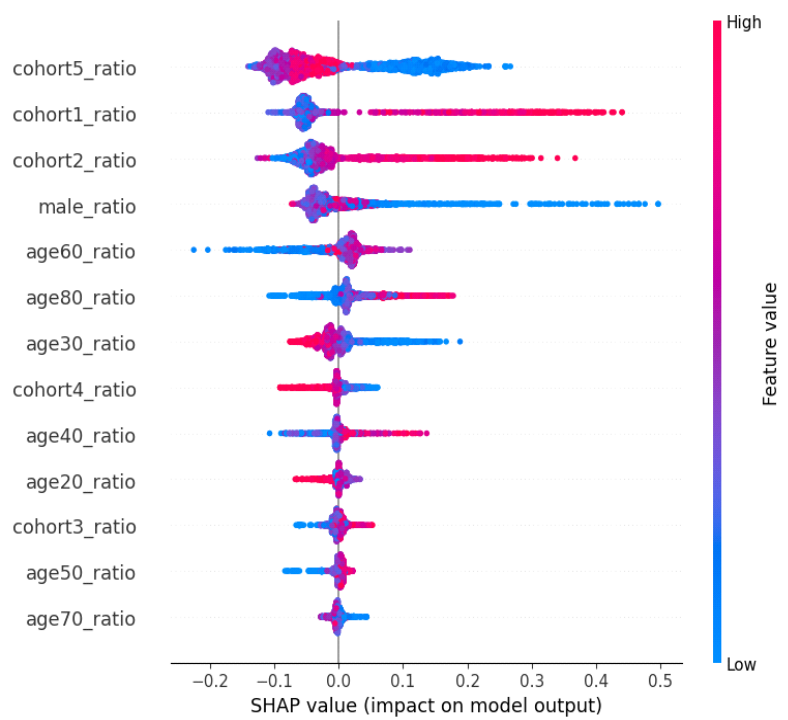
Insik Min

(Kyunghee University)

As we approach the 22nd National Assembly election in 2024, our research investigates how gender, age, and generational demographics influence support for conservative party candidates. Using data from the 2012(19th), 2016(20th), and 2020(21st) National Assembly elections, we examine population composition and support for conservative candidates at the town level. Adopting the machine learning technique XGBoost, we identify 24 predictive factors related to demographic structure and determine their importance and direction in election prediction. Our findings reveal, firstly, a notable decrease in voter ratio among both men and women aged 30-50, compared to an increase in the 60+ and 20-29 age brackets, across the 20th and 21st elections. The ‘female-late MZ generation’ demographic emerges as a key predictor, with its growth negatively correlating with support for conservative candidates. The ‘male-traditional generation’ group is identified as a critical predictor of conservative support. However, the ‘male-30s’ group, while being the third most significant predictor, shows that an increase in this demographic results in a decrease in conservative support. Further, our analysis reveals that out of 1,094 towns in the metropolitan area, 834(76.2%) towns experience demographic shifts favoring conservative candidates when comparing data between the 21st and the forthcoming 22nd elections. These results suggest that demographic changes are likely to advantage the conservative party, particularly in the Seoul-metropolitan area.

Key words: gender-age-generation effect, 22nd National Assembly Election, XGBoost, SHAP(SHapley Additive Explanations)

〈부록〉



<주효과 변수만을 고려한 SHAP 요약도표>