제 1 주제

머신러닝과 XAI를 활용한 주택 가격과 지역 특성과의 관계 분석

Using machine learning and XAI, An Analysis of the Relationship between Housing Price and Regional Characteristics

양 건 필* · 전 해 정**
Geonpil Yang · Haejung Chun

차 례

ABSTRACT

1. CONTENTS

(1) RESEARCH OBJECTIVES

his study aims to predict housing prices through regional characteristics through the machine learning model and apply the XAI methodology to infer the relationship between independent and dependent variables used in the model.

(2) RESEARCH METHOD

The Linear Regression, XGBoost in the boosting model, RandomForest in the bagging model, and the SHAP model in the XAI model were used.

(3) RESEARCH FINDINGS

The application results for the linear regression, XGBoost, and RandomForest models were compared through MSE, MAP, and RMSE, and the model application result showed the highest accuracy of the random forest model. In addition, the importance of variables was high in similar variables in both XGBoost and RandomForest.

* 주저자 : 상명대학교 부동산학과, 박사과정, jpfeel1986@gmail.com

** 교신저자 : 상명대학교 일반대학원 부동산학과 교수, hjchun6807@smu.ac.kr

2. RESULTS

The variables that have a positive effect on apartment sales prices are the proportion of the high-income population, rental income, number of apartments, detached houses, commercial districts, bus stops, commercial workers, proportion of male/female workers.

3. KEY WORDS

XAI, SHAP, Boosting, Bagging, XGBoost, RandomForest

국문초록

주택 가격과 지역특성과의 관계를 분석하기 위해 시설, 인구, 직종, 소득/소비의 항목별로 독립 변수 데이터셋을 구축했으며 종속 변수는 국토교통부의 아파트 실거래가 데이터를 활용하였다. 구축된 데이터셋에 선형회귀 모형과 부스팅 모형 중 XGBoost, 배깅 모형 중 랜덤포레스트를 적용해 지역 특성을 통한 주택 가격을 예측하고 XAI 방법론 중 SHAP모형을 적용해 독립 변수와 종속 변수 간 관계를 추론하였다. 분석의 공간적 범위는 서울특별시를 국가 기초 구역단위로 나누어 설정하고 내용적 범위는 주택 중 아파트로 시간적 범위는 2021년으로 실증분석하였다. 주택 가격에 영향을 미칠 수 있는 다양한 요인들이 있지만 본 연구에서는 지역 특성에 초점을 맞추고 지역 특성 중 시설, 인구, 직종, 소득/소비의 관점에서 분석을 수행하였다. 분석 결과, 아파트매매가격에 정(+)의 영향을 미치는 변수들은 고소득 비중, 임대 소득자 거주 비중, 아파트수, 다가구주택수, 단독주택수, 상권활성화도로 나타나며, 부(−)의 영향을 미치는 변수들은 저소득 비중, 버스정류장수, 일반기업체 종사자 거주 비중, 상가수, 공무원 거주 비중, 교육계 종사자 거주 비중, 언론계 종사자 거주 비중, 자영업(건설/제조) 종사자 거주 비중, 다세대주택수, 대학교수 등으로 나타났다.

핵심어 : XAI, SHAP, 부스팅, 배깅, XGBoost, 랜덤포레스트

Ⅰ. 서 론

주택 가격은 다의적인 개념으로 주택 그 자체의 가격과 그 주택을 둘러싸고 있는 환경의 가격이 합해진 것이다. 토지의 가격을 지대 수입의 관점에서 보면 토지의 소유 및 기타 권리이익으로부터 발생되는 장래수익에 대한 현재가치라고 할 수 있는데, 장래가치의 현재적 묘사의 성격으로 주택 가격은 그 본질이 추상성을 내포하고 있으며, 주택 가격은 주택의 수요와 공급에 의해 결정되고 일단 주택 가격이 결정되면 그 가격은 주택의 수요와 공급에 영향을 미쳐 수급을 조절하는 가격의 이중성을 갖고 있다.

이러한 부동산 가격이 이루어지는 형성요인은 일반적 요인과 지역적 요인 개별적 요인으로 구분할 수 있는데 일반적 요인으로는 인구 가구수, 공공시설, 토지거래 관행, 교육 및 복지수준 등 일반적 요인 중 사회적 요인과 일반적 요인 중 경제적 요인인 소득 및 소비수준, 물가수준 및 통화량, 금융 재정정책, 기술혁신 및 산업구조, 경제성장과 국제수지 등이 있고, 일반적 요인 중 행정적 요인인 법적 행정적 조치 정책결정 등 공적 규제, 토지소유 및 거래에 대한 규제, 토지이용에 대한 규제와 완화, 부동산 세제의 상태, 주택 가격의 통제, 토지의 선매제 등이 있다.

주택 가격을 형성하는 요인 중 지역적 요인은 위에서 설명한 일반적 요인과 각 지역의 자연적 조건의 상관 결합에 의해 그 지역의 규모, 구성기능 등에 영향을 미쳐 각 지역의 특성을 형성하고, 그 지역에 속하는 주택의 가격형성에 영향을 주는 요인으로는 해당 지역의 시설 특성, 인구 특성, 소득 특성, 소비 특성, 직종 특성 등이 있다.

주택 가격을 형성하는 요인을 발견하고 예측하기 위해서 전통적인 통계모형부터 최신 머신러닝 모형까지 다양한 방법론을 적용되고 활발히 연구되고 있다. 전통적인 통계모형의 분석의 초점은 독립 변수와 설명 변수의 관계를 추론하는 것에 집중되어 있고 반면 머신러닝 모형을 활용한 분석의 경우 예측 능력 제고에 초점을 맞추고 있다. 머신러닝 모형을 활용한 연구들이 많아진 배경에는 기존 통계모형에서의 구조적 문제로 인한 예측력 부진이 있으며, 특히 복잡한 요인들로 인해 가격이 형성되는 주택시장의 경우 낮은 예측력으로 인해 본래의 목적인 독립 변수와 설명 변수의 관계 추론의 당위성을 잃게 되는 경우가 발생했다.

따라서, 부동산 분야뿐만 아니라 다양한 분야에서도 변수 간의 설명력과 예측력 두 가지 측면에서 일정 수준 이상 보장할 수 있는 방법론을 개발하기 위해 활발한 연구가 이루어지고 있다. 이러한 배경에서 등장한 방법론이 XAI(eXplainable AI) 방법론이다. XAI는 머신러닝 방법론을 통해 예측력은 향상시키면서 동시에 변수 간 설명이 불가능했던 블랙박스(Black Box) 영역을 해석하여 의사결정을 지원할 수 있도록 지원하는 방법론이다. 현재 XAI는 의료, 금융, 보안 등 다양한 분야에서 연구 및 적용되고 있으며 더 많은 분야로 적용 영역을 넓혀가고 있으며 부동산 분야에서도 부동산 가치 평가 등의 영역에 XAI가 적용되고 있다.

이에 본 연구는 머신러닝 모형을 통해 지역 특성을 통해 주택 가격을 예측하고 XAI 방법론을 적용해 모형에 활용된 독립 변수와 종속 변수 간 관계를 추론하고자 한다. 주택 가격에 영향을 미칠 수 있는 다양한 요인들이 있지만 본 연구에서는 지역 특성에 초점을 맞추고 지역 특성 중 시설, 인구, 직종, 소득/소비의 관점에서 분석을 수행하였다.

본 연구의 공간적 범위는 서울특별시를 국가기초 구역[3]단위로 나누어 설정하고 내용적 범위는 주택 중 아파트로 시간적 범위는 2021년으로 설정 후 머신러닝 모형인 선형회귀, 부스팅, 배깅 모형을 이용해 실증분석하였다. 그리고 XAI 모형 DALEX와 SHAP 패키지를 활용해 변수 간 관계 추론을 수행하였다.

독립 변수는 시설, 인구, 직종, 소득/소비의 항

3) 도로명주소를 기반으로 국토를 읍면동의 면적보다 작게 일정한 경계를 정하여 나눈 구역(전국 : 34,349개)이다.

목별로 데이터셋을 구축했으며 주요 변수인 인구, 직종, 소득/소비는 'NICE평가정보[4]'의 데이터를 활용했으며, 종속 변수는 국토교통부의 아파트 실거래가 데이터를 활용하였다.

본 연구의 구성은 다음과 같다. 2장은 관련된 선행연구를 살펴본다. 3장은 머신러닝 모형과 XAI 모형에 대해 알아본다. 4장은 실증분석으로 변수 탐색 결과 및 모형 적용 결과 그리고 변수 간 관계 추론 결과에 대해 구술한다. 마지막 결론으로는 연구결과를 최종 요약하고 결과에 따른 시사점을 제시한다.

Ⅱ. 선행연구

국내에서 주택 가격 결정 요인에 대한 연구는 80년대부터 시작되었으며 연구 목적과 대상에 따라서 다양한 변수들을 이용하여 분석하였다.[5] 주택가격의 상승이나 하락은 사회적으로 경제성장률, 물가지수 등과 같은 복합적이고 거시적인 요인들이 작용한 결과이나 특정 시점에서의 하위시장에서는 주택 주변의 특성인 인구, 교통, 소득, 소비 등과 같은 요인을 고려할 수 있다. 주택 가격결정 요인에 관한 연구는 앞서 언급한 거시적 관점, 미시적 관점 등 다양한 요소들을 고려하여 연구되어 오고 있다.

인구사회 및 사회경제적 특성에 주목한 연구들을 살펴보면 다음과 같다. 김주영·김주후(2006)은 자가비율, 대졸 이상 비율, 소득수준 등 가구 단위의 사회경제적 변수들을 사용했으며, 김동중·임덕호(2009)는 총인구수를 이준용·손재영(2013)은 전입건수 등 상주인구변화와 관련된 변수들을 사용하였다. 근린 입지특성에 주목한 연구들을 살펴보면 박헌수·김정훈(2004)은 지역 내 아파트의 주거특성, 도심 접근성 변수를 김준현(2012)은 지역의 주거, 상업 지구 등 토지이용현황, 형태, 경사도의 토지 특성, 그린벨트의 위치, 혐오 시설물로부터의 거리 등의 변수를 사용하였다.

거주자의 소득 특성에 주목한 연구들을 살펴보면 다음과 같다. Adams and Füss(2010), Mikhed and Zemcik(2009), Iacoviello and Neri (2010) 등 다수의 선행연구에서 소득과 주택가격 간 양(+)의 관계가 입증되었으며, Miles and Pilonca(2008)에 따르면 부동산 가격 결정요인을 분해(decomposition)한 결과 1인당 실질소득 변동이 평균적으로 42%를 차지하는 것으로 나타났다고 하였다. 또한, Mckenzie(1933)는 미국의 15개 도시를 대상으로 분석한 결과 소도시일수록 지가와 인구의 변동이 비례하는 경향을 나타낸다고 하였고, Amato(1969)는 콜롬비아 수도 보고타를 대상으로 인구밀도와 지가와의 관계 분석 연구에서 상대적으로 소득이 높은 사람은 도심에 가까운 곳에, 상대적으로 소득이 낮은 사람은 교외에 분포한다고 보았고, Goldberg (1972)는 연구결과 도시의 지가와 인구밀도 간에 상관관계가 높다고 하였다.[6]

머신러닝을 활용한 주택 가격 예측 분야의 선행연구를 살펴보면 다음과 같다. 머신러닝을 활용한 아파트 가격 예측과 관련한 최근 연구는 다음과 같다. Park and Bae(2015)는 버지니아주 Fairfax 카운티의 5,359개 타운하우스의 거래 자료를 활용하여 Ripper, Baive Batesuan과 AdaBoost 방식을 적용, 주택 가격 예측을 수행하였고 Ripper 알고리즘이 가장 높은 정확도를 나타냄을 확인하였다. Pai and Wang(2020)은 주택 가격 예측을 위해 실거래데이터를 머신러닝 모델인 LSSVR, GRNN, BPNN, CART 모형을 적용하여 분석을 수행하였으며, 적용 모형 중 LSSVR이 가장 우수한 결과를 도출하였다고 하였다. 특히 LSSVR의 적용 결과는 과거의 부동산 가격 예측 관련 연구보다 평균절대비율오차(MAPE) 측면에서 오차가 더 낮은 것으로 나타났다고 하였다.

4) NICE평가정보(NICE Information Service)는 대한민국의 개인 및 기업 신용정보회사이다.

5) 임종현, 유진호, 이주형, "주변지역 토지이용특성이 주택 가격결정에 미치는 영향 : 일산신도시 공동주택을 중심으로", 국토연구, 국토연구원, 2008, 제57권, pp.49-63.

6) 신영재, "서울시의 인구, 산업별 종사자 및 최고지가의 분포와 상관관계에 대한 연구", 대한지리학회지, 대한지리학회, 2014, 제49권, 제4호, pp.509-524.

배성완 · 유정석(2018)은 시계열 모형과 머신러닝 모형을 아파트매매실거래가격지수를 대상으로 예측력을 비교분석하였다. 시계열 모형은 단변량 시계열 중 ARIMA 모형, 다변량 시계열 중 VAR과 베이지언 VAR 모형을 이용하였으며, 머신러닝 모형 중 서포트 벡터머신(SVM), RF, 그래디언트 부스팅 회귀트리(GBRT), DNN, LSTM 모형을 이용하여 예측력을 비교분석하였다. 시장 상황이 일정한 추세를 띄고 있는 경우 시계열 모형 또는 머신러닝 모형 모두 일정 수준 이상의 예측력을 보여주고 있지만, 시장이 비선형으로 변할 경우 시계열 모형이 아닌 머신러닝 모형만 의미 있는 예측력을 보여주고 있는 것으로 나타났다고 말했다.

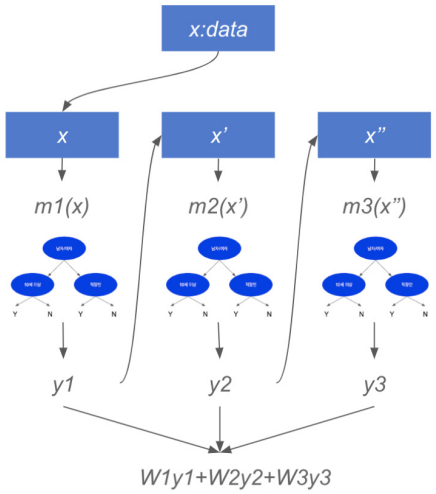
Ⅲ. 분석모형

1. XGBoost

XGBoost는 Extreme Gradient Boost의 약자로, 기본적으로 그래디언트 부스트의 방식을 따르지만 그래디언트 부스트의 단점 중 하나인 느린 수행 시간과 과적합 문제를 해결한 알고리즘이다. 그래디언트 부스트의 경우, 손실함수를 감소시키는 최적의 함수를 찾기 위해 가능한 경우의 수를 모두 탐색한다. 이때 만약 고려되는 변수의 수가 많은 경우 연산 효율성은 급격히 떨어질 수 있다. 특히, 각각의 범주형 변수를 알고리즘을 통해 연산하기 위해 각 범주형 변수에 포함된 범주값을 더미변수화하는 경우가 많으므로, 결과적으로 많은 범주를 포함한 범주형 변수가 소수만 포함되어도 극단적인 비효율성으로 인한 연산력의 저하에 노출될 수 있다. XGBoost는 변수의 분포를 고려하여 이런 비효율적 탐색 과정을 간략화하여 결과적으로 모형의 연산 효율성과 추정력을 상승시키는 알고리즘이다. 일반적인 그래디언트 부스트의 경우, 과적합에 대응하는 기능이 별도로 존재하지 않지만, XGBoost에서는 과적합에 대한 규제를 통해 보다 안정적인 예측이 가능한 것으로 알려져 있다. XGBoost는 다른 기계학습자에 비해 예측 성능이 뛰어나며, 병렬 CPU를 통한 학습이 가능해 그래디언트 부스트에 비해 빠른 수행시간을 갖는 것으로도 평가받고 있으며 XGBoost의 목적 함수는 식(1)과 같다.[7]

$$\text{obj}(\theta) = \underbrace{\sum_i^n l(y_i, \hat{y}_i)}_{\text{▲training loss}} + \underbrace{\sum_i^K \Omega(f_k)}_{\text{▼complexity of tree}} \qquad (1)$$

〈그림 1〉 Boosting 개념도



2. 랜덤포레스트

랜덤포레스트(Random Forest)는 브레이먼(Breiman)이 2001년 고안한 예측 분석 방법론으로 배깅(bagging), 아공간 샘플링(subspace sampling) 및 다수의 의사결정나무를 결합하여 분류 및 회귀 문제에 대한 예측모델을 제시하는 머신러닝 기반의 앙상블 모형이다(Breiman, 2001; Kelleher et al., 2015).

랜덤포레스트 알고리즘은 부트스트랩 결합과 아공간 샘플링으로 관측치와 입력변수가 무작위로 추출된 특정 n개의 샘플에 대해 각각 의사결정나무를 학습시켜 완성된 n개의 의사결정나무의 예측값을 통합하는 과정을 거친다. 부트스트랩 결합은 원본 데이터셋에서 랜덤 샘플링으로 추출한 여러개의

7) 홍정의, "기계학습 알고리즘을 이용한 주택가격감정 시스템의 구축 및 평가: XGBoost, LightGBM, CatBoost 알고리즘에 기반하여", 주택금융연구, 한국주택금융공사, 2020, 제4권, pp.33-64.

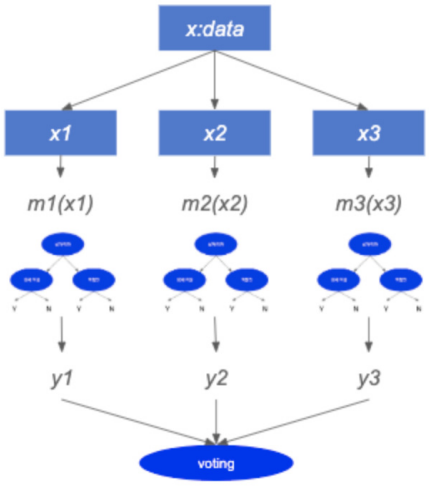
부트스트랩 샘플을 생성하여 모델링 후 결합하여 최종 예측모형을 생성하는 알고리즘이다. 부트스트랩 결합에서 모델링 결과를 결합할 때 연속형 변수의 목표변수이면 각 예측 결과의 평균을 이용하고, 범주형 변수의 목표변수이면 투표 방식을 이용하여 최종 예측 결과를 출력하는 방식을 가지고 있다. 부트스트랩 결합을 이용하면 원본 데이터셋으로부터 다수의 샘플링으로 예측모형의 분산을 낮춰 모형의 변동성을 감소시킬 수 있다는 장점이 있다.

랜덤포레스트는 Out-of-bag(OOB) 데이터를 활용하여 변수 중요도에 대한 추정치를 제공하는데 부트스트랩은 표본을 복원 추출하므로 전체 데이터 중 약 1/3 가량이 모형 적합에 이용되지 않는다. 이를 OOB 데이터라 하며 OOB 데이터를 다시 검증 데이터로 활용하여 개별 트리의 OOB 오차를 계산할 수 있다. n개의 데이터셋에 대한 랜덤포레스트의 평균 OOB 오차는 식(2)로 구할 수 있다.[8)]

$$\text{OOB error} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad (2)$$

변수 중요도는 오차 간의 차이 계산을 통해 정량화할 수 있으며 랜덤포레스트는 특정 변수값을 무작위로 치환한 후 두 번째 OOB 오차를 구한다. 이 과정에서 치환된 변수값과 실제값 간의 상관관계는 제거되므로 일반적으로 예측 오차는 치환하기 전의 OOB 오차에 비해 높다. 따라서 두 값의 차이가 클수록 변수의 중요도는 높게 나타난다.

〈그림 2〉 Bagging 개념도



3. SHAP(SHapley Additive exPlanations)

SHAP 기법은 게임 이론에 등장하는 Shapley Value를 이용하는 기법이다. Shapley Value는 전체 성과를 만들어내는 데 변수별로 어느 정도 기여했는지를 수치로 나타낸 값이다. 여기서 특정 변수의 기여도는 모든 변수를 조합한 경우 계산되는 성과에서 해당 변수를 제외한 나머지 변수들을 조합한 경우 계산되는 성과를 빼 차이를 계산함으로써 측정할 수 있다[9)]. 즉, 전체 성과가 특정 특성을 제외하면 어떻게 변화하는지 정도를 계산해 내어, 해당 특성의 기여도를 계산해 낼 수 있다.[10)]

SHAP 기법은 도출한 예측 모델의 결과를 예측에 사용한 특성들이 각각 어떠한 영향을 미쳤는지를 나타내는 기여도로 분해한다. 이때 Shapley Value는 양의 값뿐만 아니라 음의 값을 가질 수도 있으며, Shapley Value가 음수라는 것은 예측에 있어서 해당 특성이 부(-)의 영향을 미쳤다고 판단할 수 있다. SHAP는 서로 영향을 미치는 특성들이 존재할 경우 해당 특성들 사이의 의존도를 고려하여 각 특성의 영향력을 계산한다.[11)]

SHAP 기법은 예측 모델에 사용되는 특성들이

8) 장동률, "랜덤 포레스트를 활용한 작품 가격 예측 모형 연구", 신뢰성응용연구, 한국신뢰성학회, 2020, 제20권, 제1호, pp.34-42

9) A. Doniec, S. Lecoeuche, R. Mandiau, and A. Sylvain, "Purchase intention-based agent for customer behaviours." information Sciences, June 2020, Vol. 521, pp. 380-397.

10) S. Kim, W. Kim, Y. Jang, and H. Kim, "Development of Explainable AI-Based Learning Support System," The Journal of Korean Association of Computer Education, August 2021, Vol. 24, No. 1, pp. 107-115.

11) J. Ahn, "XAI, Examine the Inside of Artificial Intelligence," Wikibooks, 2020.

예측에 미치는 평균 영향도를 계산해 준다. SHAP는 특성과 특성 사이의 의존도를 고려한다. 양의 영향력뿐만 아니라 음의 영향력도 반영하여, 사용된 특성들이 예측에 어느 정도의 영향을 미치는지 그 정도에 따라 나타내준다. 따라서 부의 영향은 반영되지 않는 특성 중요도 기법보다 영향력이 더욱 정확하게 계산되어 나온다. 또한, SHAP 기법은 지역적인 설명(local interpretation)도 가능하다. 각각의 사례에 대해 어떤 특성이 예측에 긍정적 또는 부정적인 영향을 주었는지 나타내 줌으로써 사례별로 예측 결과에 대한 설명이 가능하다는 장점이 있다. 본 연구에서는 트리 기반 학습 모델에 적합하게 변형된 treeSHAP 기법을 사용한다. treeSHAP 기법은 트리 기반 모델에 적용할 때 일반적인 kernelSHAP 기법보다 계산 비용이 적게 들어 빠르고 정확한 계산이 가능한 모델이다.

Ⅳ. 실증분석

1. 변수설명

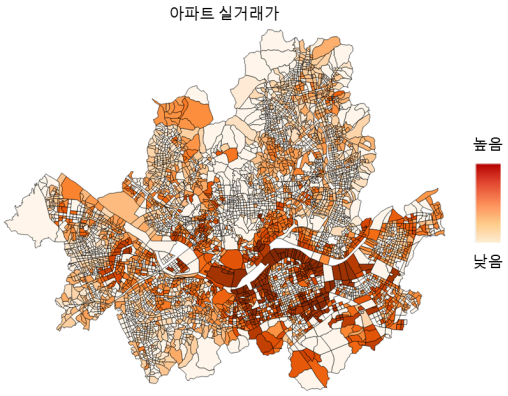
본 연구의 공간적 범위는 서울특별시를 국가 기초구역단위로 나누어 설정하고 내용적 범위는 주택 중 아파트로 시간적 범위는 2021년으로 설정 후 선형회귀, 부스팅, 배깅 모형을 이용해 실증분석하였다.

독립 변수는 시설, 인구, 직종, 소득/소비의 항목별로 데이터셋을 구축했으며 주요 변수인 인구, 직종, 소득/소비는 신용평가정보사의 데이터를 활용했으며, 종속 변수는 국토교통부의 아파트 실거래가 데이터를 활용하였다.

〈표 1〉 변수 기초통계량

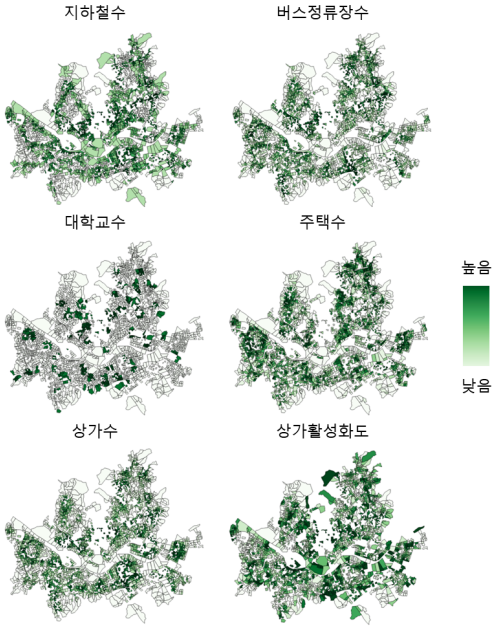
	변수명	평균	표준편차	최소값	최대값
시설	지하철수	0.00002	0.00003	0.00065	0
	버스정류장수	0.00188	0.00166	0.02473	0.00001
	대학교수	0.00002	0.00006	0.00093	0
	총 주택수	0.00642	0.00448	0.03198	0
	주택수 (아파트)	0.00387	0.00420	0.02620	0
	주택수 (빌라)	0.00215	0.00281	0.02326	0
	주택수 (단독주택)	0.00005	0.00013	0.00171	0
	주택수 (다가구주택)	0.00049	0.00063	0.00406	0
	상가수	0.01527	0.01262	0.19282	0.00003
	상권활성도	43.3099	7.61304	82.2500	6.06789
인구	유동인구수 (남성 20대)	0.00245	0.00520	0.10290	0
	유동인구수 (여성 20대)	0.00330	0.00773	0.15757	0
	유동인구수 (남성 30대)	0.00213	0.00356	0.03839	0
	유동인구수 (여성 30대)	0.00228	0.00400	0.04416	0
	유동인구수 (남성 40대)	0.00226	0.00372	0.04107	0
	유동인구수 (여성 40대)	0.00256	0.00465	0.06712	0
	유동인구수 (남성 50대)	0.00335	0.00529	0.06723	0
	유동인구수 (여성 50대)	0.00284	0.00471	0.07143	0
	거주인구수 (20대)	0.00496	0.00319	0.03264	0
	거주인구수 (30대)	0.00486	0.00325	0.04024	0
	거주인구수 (40대)	0.00494	0.00319	0.03988	0
	거주인구수 (50대)	0.00488	0.00303	0.03302	0
직종	공무원 거주 비중	6.02803	3.19236	75.25	0
	교육계 거주 비중	4.00806	1.66450	15.18	0
	금융계 거주 비중	3.34397	2.00924	25.3	0
	언론계 거주 비중	0.42388	0.43092	4.09	0
	의료계 거주 비중	2.95668	1.70885	27.18	0
	일반기업체 거주 비중	30.2815	4.81259	47.34	9.27
	임대소득자 거주 비중	26.6253	4.57414	45.6	6.18
	연금소득자 거주 비중	0.24002	0.23458	2.56	0
	자영업 (건설/제조) 거주 비중	3.28155	1.41035	14.64	0
	자영업 (숙박/음식) 거주 비중	7.23041	2.07488	25	0
	전문직 거주 비중	0.70926	0.45183	5.55	0
소득/소비	2-3천만원 소득 비중	32.9081	10.0557	70.23	5.69
	3-4천만원 소득 비중	24.2217	4.15395	48.46	7.79
	4-5천만원 소득 비중	11.8261	2.70453	25.37	2.94
	5-6천만원 소득 비중	11.0912	3.94355	27.27	2.07
	6-7천만원 소득 비중	3.52834	2.14685	13.54	0
	7천만원 이상 소득 비중	7.80949	6.70627	38.11	0
	월평균카드 소비(천원)	1,689	846	27,897	632
	아파트매매가 (천원)	894,503	584,017	6,220,326	103,080

〈그림 3〉 서울특별시 아파트매매가격 분포



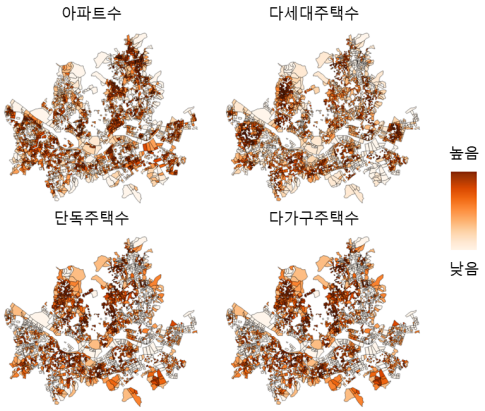
〈그림3〉은 서울특별시 국가기초구역별 아파트매매가격 분포지도로 강남과 송파, 한강 주변 지역에서 매매가가 높게 나타났다. 상대적으로 강북과 강서 지역에서의 매매가격이 낮게 나타났다.

〈그림 4〉 서울특별시 주요 시설 분포



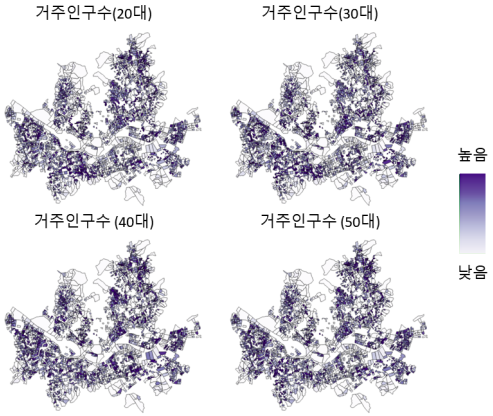
〈그림4〉은 서울특별시 주요 시설 분포지도로 교통시설과 주택 및 상가수는 전반적으로 유사한 지역에 나타났다. 절대적인 주택수는 강남지역보다 강서 및 강북지역에 더 많이 분포했다.

〈그림 5〉 서울특별시 유형별 주택수



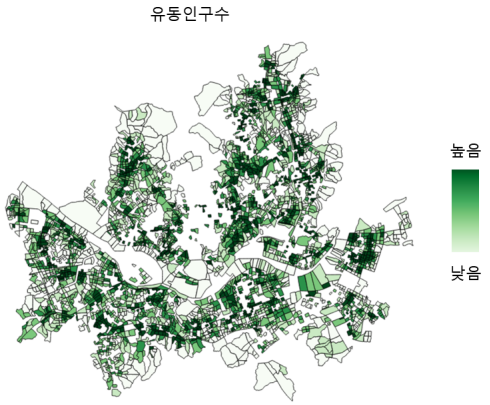
〈그림5〉은 서울특별시 유형별 주택수 분포지도로 아파트수는 상대적으로 도봉구, 노원구, 강남구, 송파구에 많이 분포했으며, 은평구의 경우 아파트는 적고 그 외 주택유형(다세대, 단독, 다가구)이 많이 분포했다. 또한, 강남구에 속한 내곡동, 세곡동의 경우 단독주택과 다가구주택수가 많이 분포했다.

〈그림 6〉 서울특별시 연령대별 거주인구수



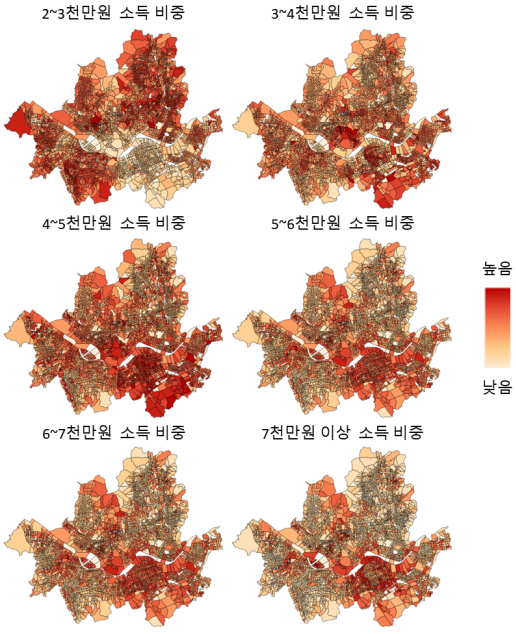
〈그림6〉은 서울특별시 연령대별 거주인구수 분포지도로 20-30대 거주인구는 상대적으로 관악구, 동작구에 40-50대 거주인구는 강남구에 더 많은 인구분포를 나타냈다.

〈그림 7〉 서울특별시 유동인구수



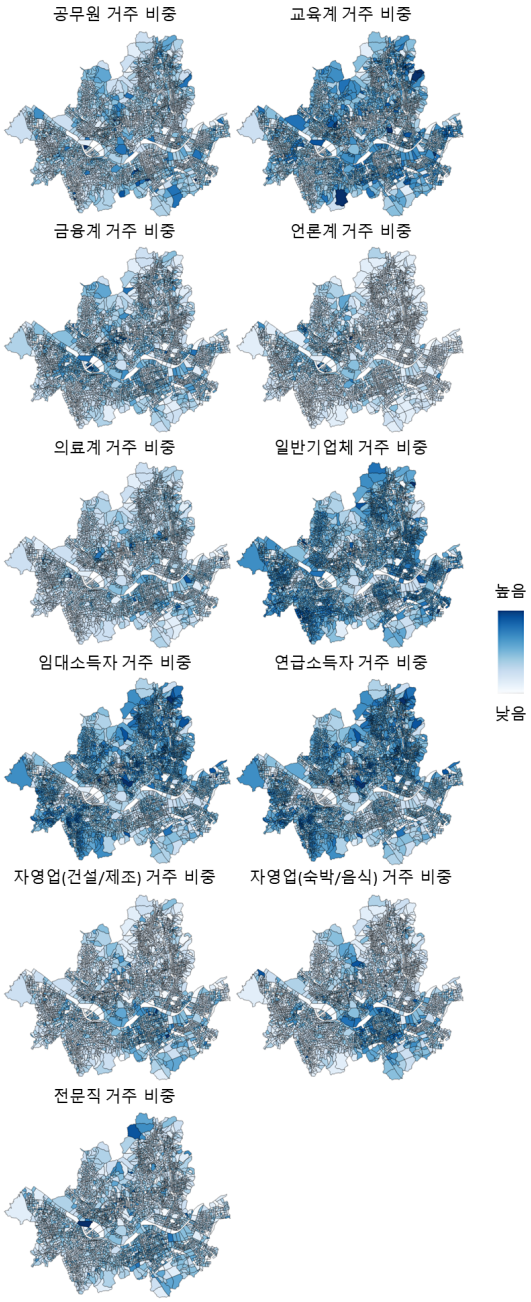
〈그림7〉은 서울특별시 유동인구수 분포지도로 서울특별시 외곽지역들을 제외하고는 비교적 전체 지역에 고루 분포했다.

〈그림 8〉 서울특별시 금액대별 소득 비중



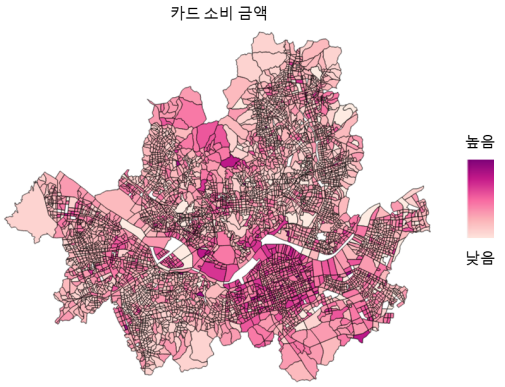
〈그림8〉은 서울특별시 금액대별 소득 비중으로 2–3천만원 소득 비중과 그 외의 소득 비중에서 지역적 차이가 크게 나타났다.

〈그림 9〉 서울특별시 직종별 거주 비중



〈그림9〉은 서울특별시 직종별 거주 비중으로 직종별로 거주 지역이 상이하게 분포했다. 자영업의 경우 상권이 발달한 지역 주변에서의 거주 비중이 높은 특징을 나타냈고 그 외 직종의 이와 달리 전 지역에 고루 분포했다. 그리고 교육계, 일반기업체, 임대소득자, 일반기업체, 전문직의 경우 주요 도심에서 떨어진 서울 외곽지역에 거주하는 특징을 나타냈다.

〈그림 10〉 서울특별시 카드 소비 금액



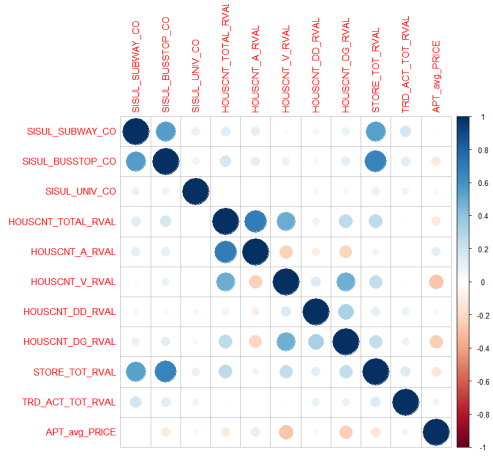
〈그림10〉은 서울특별시 카드 소비 금액 분포 지도로 강남구, 서초구, 용산구, 중구 등에서의 소비액이 많이 나타났다. 서울 외곽지역인 도봉구, 노원구, 구로구, 관악구 등에서는 소비가 상대적으로 적게 나타났다.

2. 상관관계 분석

모형 적용 전 변수 간 상관관계를 탐색하기 위해 피어스 상관계수인 식(3)을 적용해 상관분석을 수행했다.

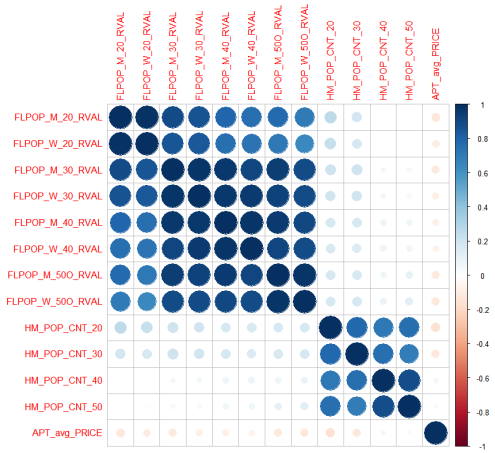
$$r_{XY} = \frac{\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_i^n \left(X_i - \overline{X}\right)^2}\sqrt{\sum_i^n \left(Y_i - \overline{Y}\right)^2}} \quad (3)$$

〈그림 11〉 서울특별시 주요시설 간 상관계수



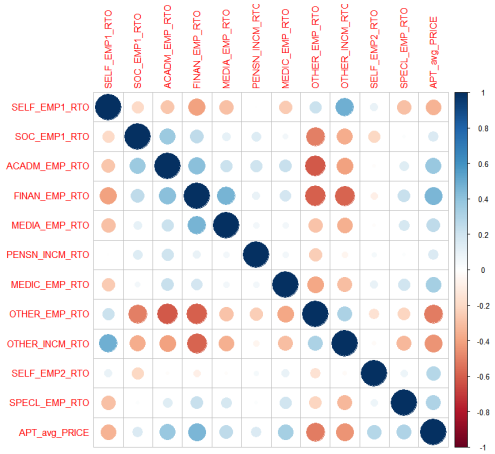
아파트매매가격과 주요 시설 간 상관계수를 살펴보면 버스정류장수, 총주택수, 다세대주택, 다가구주택 및 상가수는 아파트매매가격과 음의 상관관계를 나타냈다. 반면, 아파트수는 아파트매매가격과 양의 상관관계를 나타냈다.

〈그림 12〉 서울특별시 인구특성 간 상관계수



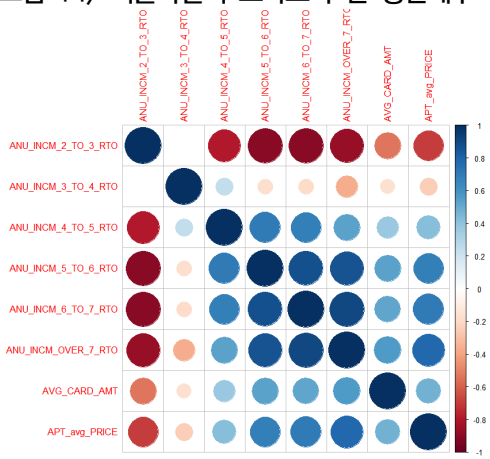
아파트매매가격과 인구 간 상관계수를 살펴보면 유동인구와 아파트매매가격은 약한 음의 상관관계를 나타내며 거주인구 중 40대의 거주인구수와 약한 양의 상관관계를 나타냈다.

〈그림 13〉 서울특별시 직종 간 상관계수



아파트매매가격과 직종 간 상관계수를 살펴보면 금융, 교육, 의료, 전문직, 자영업(숙박/음식) 직종의 거주 비중과 양의 상관관계를 나타냈으며 연금, 임대 및 자영업(제조/건설) 직종의 거주 비중과는 음의 상관관계를 나타냈다.

〈그림 14〉 서울특별시 소득소비 간 상관계수



아파트매매가격과 소득소비 간 상관계수를 살펴보면 소득 4천만원 미만 소득을 가진 거주 인구의 비중과는 음의 상관관계를 4천만원 이상 소득을 가진 거주 인구의 비중과는 양의 상관관계를 나타냈다. 또한, 카드 소비 금액과 아파트매매가격은 양의 상관관계를 나타냈다.

3. 모형 적용

주택 가격과 지역특성과의 관계 분석을 위해 선형회귀, 부스팅, 배깅 모형을 적용했으며 부스팅 모형 중 XGBoost 모형, 배깅 모형 중 랜덤포레스트 모형을 적용해 비교 분석하였다. 독립 변수 간 발생하는 다중공선성은 모형의 설명력이 왜곡될 가능성이 있는 선형회귀 모형에 국한해서 처리하여 적용하였다.

XGBoost 모형은 분류와 회귀의 앙상블 모델로, 파라미터 설정이 모델 학습에 유의미한 영향을 끼칠 수 있다. 본 연구에서는 독립 변수를 활용해 종속 변수인 아파트매매가격 예측에 초점을 두었다.

XGBoost 모형의 파라미터는 크게 트리 수(n-estimator), 최대깊이(max_deep), 학습률(Learning Rate)에 따라 모형 결과가 달라질 수 있다. 본 연구에서는 트리 수를 500, 최대깊이를 10, 학습률을 0.3으로 설정하여 수행하였다. 학습데이터(Train data)와 테스트데이터(Test data)를 5:5의 비율로 나누어 연구를 진행하였다.

랜덤포레스트 모형도 XGBoost 모형과 동일하게 분류와 회귀의 앙상블 모델로써 파라미터 설정이 모형 학습에 유의미한 영향을 끼친다. 랜덤포레스트 모형의 파라미터는 일반적으로 트리수(n-estimator) 및 최대깊이(max-deep)에 따라 결과값이 달라진다. 본 연구에는 트리 수 500, 최대깊이 10으로 XGBoost 모형과 동일한 값을 설정하여 모델학습을 진행하였다. 학습 데이터(train data)와 테스트 데이터(test data) 또한, XGBoost 모형과 동일하게 5:5로 나누어 연구를 진행하였다.

3개의 모형에 대한 적용 결과는 RMSE, MAE, MAPE를 통해 비교했으며 모형 적용 결과 〈표2〉에서와 같이 랜덤포레스트 모형의 정확도가 가장 높게 나왔다.

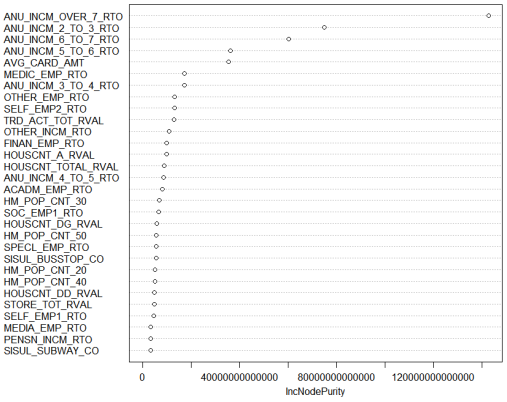
〈표 2〉 모형별 예측 정확도

모형	RMSE	MAE	MAPE
LM	3.12647	1.99644	29.936
XGBoost	3.19387	1.99371	28.646
RF	2.96973	1.85402	28.44

4. 모형 해석 결과

〈그림15〉은 랜덤포레스트의 변수 중요도를 시각화한 것으로 IncNodePurity값이 클수록 중요도가 커짐을 나타낸다. 7천만원 이상 소득 비중 변수가 가장 중요한 변수로 이용되고 있으며, 2-3천만원 소득 비중 변수가 두 번째로 중요한 변수가 됨을 나타내고 있다. 6-7천만원 소득 비중, 5-6천만원 소득 비중, 월평균 카드소비금액이 그다음으로 중요한 변수이며 그 외에는 서로 미미한 차이를 보이고 있다.

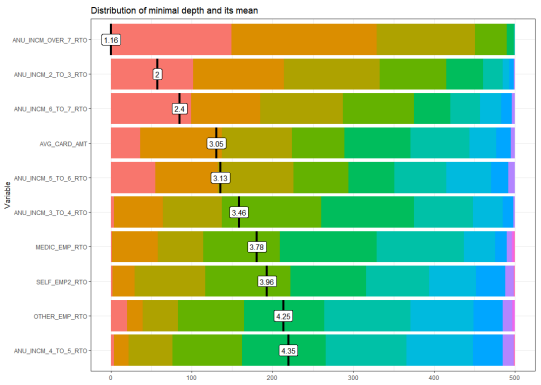
〈그림 15〉 랜덤포레스트 변수 중요도



기본 랜덤포레스트 패키지 외에도 랜덤포레스트의 구조와 변수의 역할을 이해를 돕기 위한 XAI를 적용한 randomForest Explainer 패키지가 개발되어 있다. 이 패키지를 통해 다양한 유형의 변수 중요도를 확인할 수 있으며 변수들 간의 교호작용에 대해 파악할 수 있다.

Min_depth_distribution 함수는 예측 변수들의 평균 최소 깊이의 분포를 조사할 때 사용된다. Depth는 각 트리에서 예측 변수들이 처음 나타나는 지점을 나타내는 것으로 root node 근처에 있을수록 depth는 0에 가까워지며 중요변수로 판단할 수 있다. 〈그림16〉은 500개의 tree 중 각 나무에서의 최소 깊이의 분포와 평균값을 나타내고 있다. 7천만원 이상 소득 비중 변수와 2-3천만원 소득 비중 변수의 평균 최소 깊이는 각각 1.16, 2로 두 독립 변수 모두 depth 1, 2에서 대부분 나타나고 있으며 아파트매매가격을 설명하는 데에 중요한 역할을 하고 있음을 알 수 있다. 이는 변수 중요도에서의 결과와 같다.

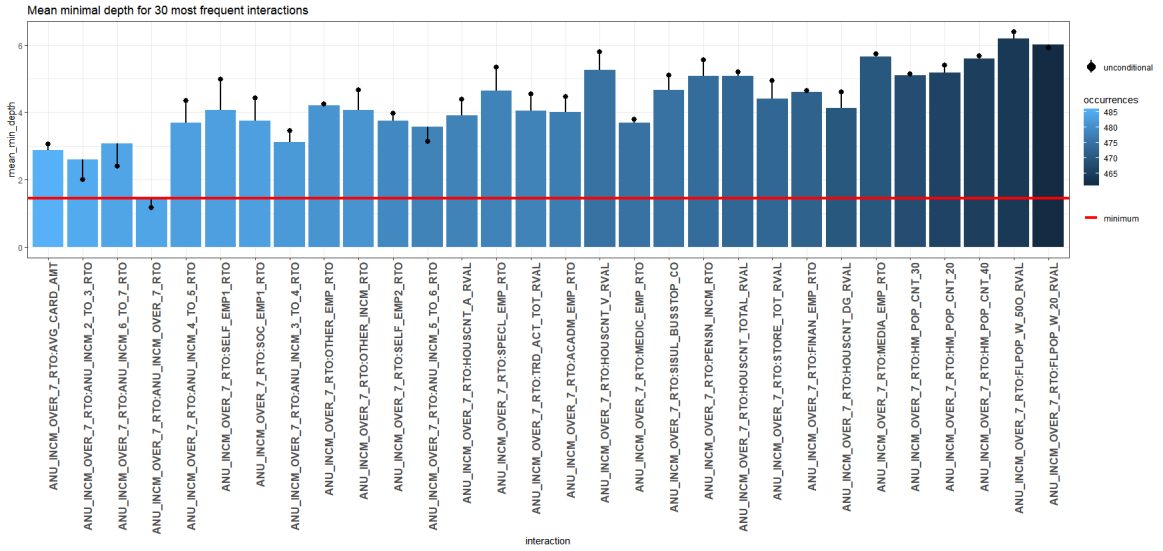
〈그림 16〉 최소 깊이의 분포와 평균



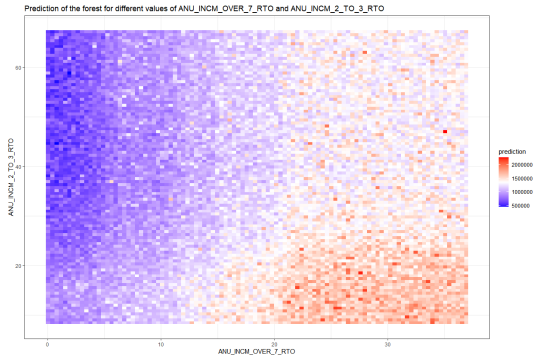
교호작용은 두 변수 사이 최소 깊이의 차이로 나타낼 수 있고 이들의 평균을 그림으로 나타낸 것을 최소 깊이 상호작용이라 한다 〈그림17〉에서의 최소 깊이의 평균(mean_min_depth)이 작을수록 두 변수가 차례로 나오는 경향이 있으므로 교호작용이 크다고 판단할 수 있다.

Occurrence는 tree에 두 변수가 함께 나오는 횟수를 의미한다. 저소득 비중 변수와 고소득 비중 변수 그리고 월평균카드 소비금액 변수는 최소 깊이 상호작용 그림에서 볼 수 있는 가장 빈번한 상호작용이라고 할 수 있다.

〈그림 17〉 Mean minimal depth for 30 most frequent interactions



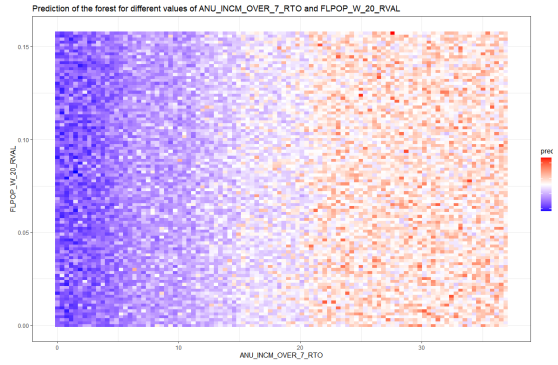
〈그림 18〉 소득구간에 따른 아파트매매가격 예측값



〈그림18〉은 변수 간 교호작용을 확인하기 위하여 7천만원 이상의 소득 비중 변수와 2-3천만원의 소득 비중 변수에 따른 아파트매매가격의 예측값을 색으로 나타낸 것으로 7천만원 이상의 소득 비중 값이 크고 2-3천만원의 소득 비중 값이 작을수록 아파트매매가격이 높게 나타나는 것을 알 수 있다.

반면 상호작용이 가장 적다고 나타난 7천만원 이상의 소득 비중 변수와 20대 여성 유동인구수 변수의 경우 〈그림19〉에서와 같이 특정 패턴을 찾아볼 수 없다.

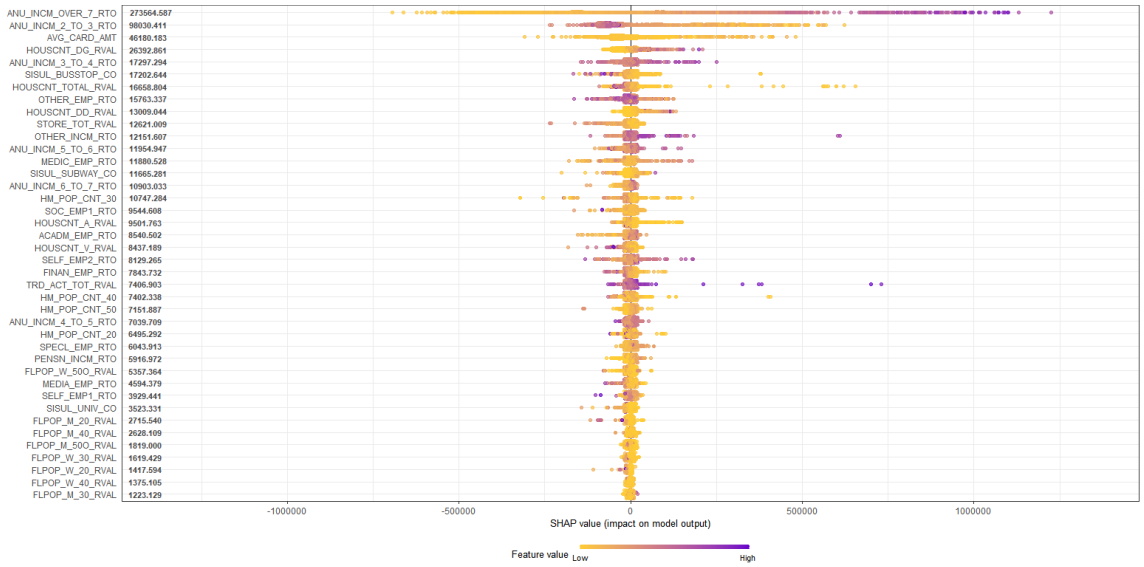
〈그림 19〉 소득구간과 유동인구 수에 따른 아파트 매매가격 예측값



XGBoost에서는 변수 중요도로 상대적 영향력을 계산해 준다. 상대적 영향력의 값이 클수록 감소량이 큰 것을 나타낸다. 아파트매매가격에 대한 변수 중요도의 상위권은 랜덤포레스트의 결과와 비슷하게 7천만원 이상 소득 비중 변수와 2-3천만원 소득 비중 변수가 차지하고 있다〈그림20〉.

그러나 랜덤포레스트와는 달리 주택수와 버스정류장, 지하철수가 보다 높은 위치를 차지하고 있다.〈그림20〉은 각 변수들이 아파트매매가격 예측에 미치는 영향을 정(+)과 부(−)의 효과로 측정한 SHAP(SHAPley Additional exPlanations) 방법을 적용한 시각화 결과이다. 아파트매매가격에 정(+)의 영향을 미치는 변수들은 6천만원 이상 소득 비중, 임대 소득자 거주 비중, 아파트수, 다가구주택수, 단독주택수, 상권활성화도로 나타나며, 부(−)의 영향을 미치는 변수들은 3천만원 미만 소득 비중, 버스정류장수, 일반기업체 종사자 거주 비중, 상가수, 공무원 거주 비중, 교육계 종사자 거주 비중, 언론계 종사자 거주 비중, 자영업(건설/제조) 종사자 거주 비중, 다세대주택수, 대학교수, 20대 남성/여성 유동인구수로 나타났다.

〈그림 20〉 Mean minimal depth for 30 most frequent interactions



Ⅴ. 결론

본 연구는 주택 가격과 지역 특성과의 관계를 머신러닝과 XAI 방법론을 적용하여 실증분석하였다. 독립변수는 시설, 인구, 직종, 소득/소비의 항목별로 데이터셋을 구축했으며 주요 변수인 인구, 직종, 소득/소비는 신용평가사의 데이터를 활용했으며, 종속 변수는 국토교통부의 아파트 실거래가 데이터를 활용하였다.

머신러닝 모형 적용 전 EDA(Exploratory Data Analysis) 관점에서 종속 독립 변수와 종속 변수의 GIS 상에서의 분포도를 살펴봤으며, 독립 변수별로 지역적 특징들이 상이하다는 것을 관찰할 수 있었다. 주택 가격과 지역 특성과의 관계를 정량화하기 위해 선형회귀, XGBoost, 랜덤포레스트 모형을 적용하였고 모형 결과의 정확도를 비교 후 모델에 따른 결과값을 XAI 관점에서 해석 적용하였다.

예측모형 적용 결과 랜덤포레스트, XGBoost, 선형회귀 순으로 모든 평가지표에서 예측력이 우수한 것으로 나타났다. 변수 중요도 관점에서는 XGBoost와 랜덤포레스트의 우선순위가 전반적으로 유사하게 나왔지만 일부 변수들에서는 우선순위의 차이가 발생하였다. 가장 중요도가 높은 변수들은 소득 관련 변수였고 중요도가 낮은 변수들은 유동인구 관련 변수였다.

아파트매매가격에 정(+)의 영향을 미치는 변수들은 6천만원 이상 소득 비중, 임대 소득자 거주 비중, 아파트수, 다가구주택수, 단독주택수, 상권활성화도로 나타나며, 부(−)의 영향을 미치는 변수들은 3천만원 미만 소득 비중, 버스정류장수, 일반기업체 종사자 거주 비중, 상가수, 공무원 거주 비중, 교육계 종사자 거주 비중, 언론계 종사자 거주 비중, 자영업(건설/제조) 종사자 거주 비중, 다세대주택수, 대학교수, 20대 남성/여성 유동인구수로 나타났다.

본 연구결과에 따르는 시사점은 최근 머신러닝을 활용해 주택 가격을 예측하는 연구들이 크게 증가하고 있는 상황에서 대부분의 머신러닝 예측 모형 내의 독립 변수와 종속 변수의 관계성을 설명하는 해석 부분이 설명이 불가능한 블랙박스 영역으로 남아 있다. 블랙박스 영역(Black Box)을 해결하기 위해 설명가능한 인공지능(XAI)이라는 용어로 유용한 방법론들이 제시되고 있으며 XAI의 필요성을 크게 2가지로 분류할 수 있다. 첫째, '일반화'로 훈련 데이터셋으로 구축된 모형이 실제 환경 데이터에서 좋은 결과를 내지 못하는 경우 이를 '모형이 일반화되지

않았다'라고 판단하다. 모형이 동작하는 과정에서 데이터의 어떤 부분에 집중하는지를 설명할 수 있다면, 모형 개발자는 검증 데이터셋에 대해서 동일한 성능을 보이는 여러 모델들 중에서 더 '일반화'의 가능성이 높은 모델을 선택할 수 있을 것이다. 또한 모델 디버깅 및 개선으로 예측 모형의 아키텍처를 변경해야 할지, 특정 카테고리의 훈련 데이터셋을 더 확보해야 할지 등의 의사결정을 하기 위해서 모형 해석을 통한 일반화가 필요하다.

둘째, '새로운 지식에 대한 가설과 발견'으로 예측 모형의 결과에 대한 '설명'이라는 시도 자체가 새로운 가설을 만들어내는 데 중요한 역할을 담당하여 평가와 검증할 수 있는 가설의 발견이라는 면에서 모형 설명력이 중요한 역할을 담당하게 된다. 이러한 관점에서 XAI가 주택시장을 예측하고 예측한 결과를 객관적 데이터로 설명하는 과정이 보다 강화된다면 주택시장을 이해하고 새로운 트렌드를 발견하는데 있어 의미가 클 것으로 보인다.

본 연구의 한계점으로는 시간적 측면에서는 특정기간(2021년)으로 제한하고 분석함에 따라 독립변수들이 종속 변수에 미치는 영향의 변화들을 측정하지 못했다는 점과 공간적인 측면에서 범위를 서울특별시로 한정하여 발생적 지역적 한계점이 있다. 그리고 변수의 측면에서 더 많은 거시변수 및 미시변수들을 활용하지 못했다는 점이 본 연구의 한계점이며 이를 보완한 연구는 추후 과제로 남긴다.

참고문헌

김동중·임덕호, 2009, “지역 기반산업이 주택가격에 미치는 영향,” 주택연구, 17(3), 83-105.

김주영·김주후, 2006, “위계선형모델을 적용한 근린특성의 지가영향 분석,” 국토계획, 41(5), 33-43.

김준현, 2012, “혐오시설이 주변 지가에 미치는 영향 분석: 서울시립승화원 사례,” 지방행정연구, 26(4), 275-296.

박헌수·김정훈, 2004, “시공간자기회귀모형을 이용한 서울 아파트가격지수 추정에 관한 연구,” 국토연구원국토연구, 42, 125-140.

배성완·유정석, “머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측”, 주택연구, 제26권 제1호, 한국주택학회, 2018, pp. 107-133.

신영재, “서울시의 인구, 산업별 종사자 및 최고지가의 분포와 상관관계에 대한 연구”, 대한지리학회지, 대한지리학회, 2014, 제49권, 제4호, pp.509-524.

이준용·손재영, 2013, “패널분석을 이용한 대도시 주택가격 추이 분석,” 부동산학연구, 19(4), 71-86.

임종현, 유진호, 이주형, “주변지역 토지이용특성이 주택 가격결정에 미치는 영향 : 일산신도시 공동주택을 중심으로”, 국토연구, 국토연구원, 2008, 제57권, pp.49-63.

장동률, “랜덤 포레스트를 활용한 작품 가격 예측 모형 연구”, 신뢰성응용연구, 한국신뢰성학회, 2020, 제20권, 제1호, pp.34-42

홍정의, “기계학습 알고리즘을 이용한 주택가격감정 시스템의 구축 및 평가: XGBoost, LightGBM, CatBoost 알고리즘에 기반하여”, 주택금융연구, 한국주택금융공사, 2020, 제4권, pp.33-64.

A. Doniec, S. Lecoeuche, R. Mandiau, and A. Sylvain, "Purchase intention-based agent for customer behaviours." information Sciences, June 2020, Vol. 521, pp. 380-397.

Adams, Zeno and Roland Füss, “Macroeconomic Determinants of International Housing Markets,” Journal of Housing Economics, 2010, Vol. 19, pp. 38-50.

Amato, P. W., Population Densities, Land Values, and Socioeconomic class in Bogota, Colombia, Land Economics, 1969, 45, 66-73

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

Goldberg, M. A., An Evaluation of the Interaction between Urban Transport and Land Use Systems, Land Economics, 1972, 48, 338-346.

Iacoviello, Matteo and Stefano Neri, “Housing Market Spillovers: Evidence from an Estimated DSGE Model,” American Economic Journal: Macroeconomics, 2010, Vol. 2, pp. 125-64.

J. Ahn, "XAI, Examine the Inside of Artificial Intelligence," Wikibooks, 2020.

Kelleher, J. D., Mac Namee, B., and D'arcy, A., Fundamentals of machine learning for

predictive data analytics: Algorithms, worked examples, and case studies, MIT Press, 2015.

Mckenzie, R. D., 1933, e Metropolitan Community, Ressell, New York, 37.

Mikhed, Vyacheslav and Petr Zemcik, “Do House Prices Reflect Fundamentals? Aggregate and Panel Data Evidence,” Journal of Housing Economics, 2009, Vol. 18, pp. 140-149.

Miles, David and Vladimir Pilonca, “Financial Innovation and European Housing and Mortgage Markets,” Oxford Review of Economic Policy, 2008, Vol. 24, pp. 145-175

Pai, Ping-Feng. and Wen-Chang, Wang, “Using Machine Learning Models and Actual Transaction Data for Predicting Real Estate Prices”, Applied Sciences, 2020, Vol. 10 No. 17, pp.1-5832.

Park, Byeonghwa and Jae Kwon, Bae, “Using Machine Learning Algorithms for Housingprice Prediction: The Case of Fairfax County, Virginia Housing Data”, Expert Systems with Applications, 2015, Vol. 42 No. 6, pp. 2928-2934.

S. Kim, W. Kim, Y. Jang, and H. Kim, "Development of Explainable AI-Based Learning Support System," The Journal of Korean Association of Computer Education, August 2021, Vol. 24, No. 1, pp. 107-115.