pISSN : 2465-9754 eISSN : 2508-1292 https://doi.org/10.30902/jrea.2023.9.1.23

Journal of Real Estate Analysis

http://www.ejrea.org *April 2023, Vol.9, No.1, pp. 23~42*

프롭테크의 소비자 평가와 머신러닝을 이용한 아파트 매매 가격 분석

Analysis of Apartment Sale Prices with Consumer Reviews in Proptech and Machine Learning

이해인* · 황현준**
Haein Lee · Hyeonjun Hwang

Abstract

In order to analyze apartment sales prices this study extracted user-centric and atypical data from the proptech platform and used the topic modeling covering consumer-centric variables of the hedonic price model. Analysis models included the linear regression and the random forest. According to the results from the linear regression for the hedonic price model about the sales prices per supply size of apartments, estimated coefficients of topics in terms of living areas and education marked 11.6% and 11.5% respectively, which was statistically significant within the margin of 1%. In the case that the random forest would be used instead of the linear regression 0.71%, a coefficient of determination, could go up to approximately 0.93. This study utilized the information of the proptech platform and the machine learning technology in order to statistically identify data in terms of demand for apartments sales prices. Thus, this study was very helpful in suggesting a tangible methodology. This study also indicated that the government should consider the information including the proptech platform produced by consumers in order to come up with measures for real estates.

Keywords: Proptech, Hedonic price model, Latent dirichlet allocation, Linear regression, Random forest

* 경북대학교 데이터사이언스대학원 석사과정(주저자) | Master Student, Graduate School of Data Science, Kyungpook National University | First Author | kokogodls@knu.ac.kr |
** 경북대학교 데이터사이언스대학원 조교수(교신저자) | Assistant Professor, Graduate School of Data Science, Kyungpook National University | Corresponding Author | hhwang@knu.ac.kr |

Ⅰ. 서론

부동산 시장은 높은 거래 비용, 낮은 유동성 및 높은 정보 불평등을 특징으로 한다(Ibbotson and Siegel, 1984). 부동산 시장의 정보 불평등 및 비대칭으로 인하여, 공급자가 소비자보다 많은 정보를 알고 있으며 소비자가 접근할 수 있는 정보는 제한되어 있다는 특징을 가지고 있다. 또한, 부동산 시장에서 통용되는 정보 대부분은 공급자적 측면에서 생산되며, 공급자에게서 소비자에게로 일방향적 흐름을 보인다.

한편, 프롭테크(proptech)[1] 시장의 성장으로 인해 플랫폼을 중심으로 한 온라인 시장이 성장하고 있다. 부동산 플랫폼 기업은 플랫폼이라는 속성상 공급자 측면에서 사용자에게 정보를 제공하는 기능도 있지만, 상호관계로서 사용자로부터 누적적으로 획득하는 정보의 양이 상당하다(이정윤 외, 2021). 이에 프롭테크 플랫폼은 기존의 구조화된 공급자적 정보가 주류를 이루던 오프라인 시장과는 달리 소비자적 측면이 강하며 정형화되지 않은 사용자 중심의 정보가 대두되고 있는 것이 특징이다.

부동산에 대한 관심과 더불어 과거부터 시계열 기법이나 회귀 모형 등을 활용하여 부동산 상품의 가격을 추정해왔다. 그 중, 헤도닉 가격 모형(hedonic price model)이 부동산 분야에서 가장 광범위하게 사용되고 있다. 한편, 최근에는 가격 추정에 있어 분석모형으로 머신러닝 기법을 도입하여 예측 성능을 높이려는 시도가 있었다.

본 연구는 머신러닝을 이용하여 1) 헤도닉 가격 모형의 특성변수에 프롭테크의 소비자 중심 비정형 정보를 반영하고, 더불어 2) 추출된 비정형 소비자 정보를 포함한 랜덤 포레스트 기반의 비선형 가격 모형을 통해 대구시 부동산 가격을 분석하는 것을 목적으로 한다. 이를 위하여 대구광역시의 8개 구군의 2018년부터 2022년까지 5년간의 아파트 거래내역과 프롭테크 플랫폼의 아파트 단지 정보와 리뷰 기반의 비정형 정보를 이용할 것이다.

본 연구의 구성은 다음과 같다. 2장 선행연구는 헤도닉 가격 모형과 프롭테크 관련한 선행연구에 대해 살펴보고, 선행연구와 차별되는 본 연구의 기여점을 제시한다. 3장 자료 및 분석모형은 두 가지 분석모형인 헤도닉 가격 모형과 랜덤 포레스트에 이용되는 소비자 중심 특성 변수 추출 및 가공 방식, 최종 데이터셋 선정에 대해 다룬다. 4장 분석 결과는 두 분석모형의 분석결과를 각각 제시하고 두 결과를 비교한다. 마지막 5장에서는 연구 결과를 요약하고 연구의 한계점을 제시한다.

Ⅱ. 선행연구

1. 선형 인과관계 분석 모형

사회과학 분야에서 널리 쓰이는 헤도닉 가격 모형(hedonic price model)은 상품의 가격 가치가 그 상품의 특성으로부터 기인한다는 가정을 바탕으로 하여 개별 특성에 가격에 미치는 영향을

1) 부동산(property)과 기술(technology)의 합성어.

 https://doi.org/10.30902/jrea.2023.9.1.23

추정하는 모형이다(Brown and Rosen, 1982). 헤도닉 가격 모형을 이용한 국내 주택 시장 및 주택 특성 추정에서 허세림·곽승준(1994)은 다음과 같이 분석하였다. (ⅰ) 주택은 그 주택이 보유하는 도심, 직장까지의 거리, 교통 혼잡도, 대기오염, 학군, 근린시설 등의 지역 특성과 방의 개수, 화장실 겸 욕실의 개수, 주택규모 등의 주거특성으로 구성된 주택 특성으로부터 주택 서비스를 제공한다. (ⅱ) 주택 특성들은 각각의 잠재가격(implicit price)을 가지고 있으며 주택의 가격은 이들 주택 특성들의 잠재가격의 총합으로 정의된다. (ⅲ) 소비자들은 주택 구조물 그 자체보다는 주택이 생산하는 주택 서비스를 소비하고 그 서비스에 대해 가격을 지불한다.

주택의 가치는 주택구성요소들의 합성물로 이루어진 효용의 크기로 평가되어 개개의 요소들이 명시적으로 거래되지는 않지만 주택구입자(임대자)는 개별요소에 대해 주관적으로 잠재적 가치를 평가한다(임재현, 1998). 주택 특성들은 각각의 잠재가치를 가지고 있으며 주택 가치는 이들 주택 특성들의 잠재가치 총합으로 정의된다(허세림 · 곽승준, 1994).

헤도닉 가격 모형은 잠재가치의 총합으로 평가되는 가격을 종속변수로, 특성 변수를 설명변수로 하는 선형 다중 회귀분석을 통해 추정된다. 잠재가치가 특성 변수들에 의해 결정되기 때문에(이용만, 2008), 어떤 특성을 반영하여 모형을 추정하느냐가 가장 중요한 문제이다(조민서 외, 2011). 헤도닉 가격 모형의 회귀추정에서 도출되는 특성 변수들의 추정계수(coefficient)가 특성 가격(hedonic price)이다(이용만, 2008). 따라서, 각 특성 변수의 추정계수를 통해 가격에 미치는 영향과 정도를 파악할 수 있다.

한편, 조민서 외(2011)는 2000년부터 10년간 국내 부동산 분야에서 적용되고 있는 헤도닉 가격 모형의 특성 변수를 분석하여 세대 특성, 단지 특성, 입지 특성, 환경 특성으로 나누고, 추정에 이용되는 구체적인 변수를 〈표 1〉과 같이 제시하였다.

선형 다중회귀를 통해 추정되는 헤도닉 가격 모형은 회귀모형이 갖는 통계적 한계를 포함하고 있다. Chau and Chin(2003)은 다양한 독립변수의 사용으로 인해 독립변수 간의 공선성에 따른 추정 불안정성, 포함되지 못한 특성에 의한 생략변수 편의 가능성을 지적하였다. 이러한 문제 때문에, Chau and Chin(2003)은 독립변수 선정에 있어서 기술력뿐만 아니라 경험 및 판단력을 강조하고 있다. 비슷하게 헤도닉 가격 모형은 특성 변수의 영향 범위 및 경계를 파악하기 힘들어 회귀모형의 해석력이 떨어지는 한계를 지닌다(구

〈표 1〉 국내 헤도닉 가격 모형 세부 변수

구분	세부 변수
세대 특성	규모(전용면적), 출입 방식, 주차, 층, 향, 경관 및 조망, 소음, 난방방식
단지 특성	건축적 특성(단지 규모 및 용적률), 용도 지역 특성, 단지 규모, 건설사, 노후도
입지 특성	도심 접근성, 교통, 교육, 학군, 상업시설, 의료기관, 공공기관, 뉴타운
환경 특성	자연환경, 골프장, 유흥시설, Shift[2], 공해, 위험 · 혐오시설

자료 : 조민서 외(2011)의 재구성.

2) 민선4기 서울특별시의 주택정책(장기전세주택).

본상 · 신병진, 2015).

또한, 헤도닉 가격 모형은 선형 모형을 이용하기 때문에 특성 변수들의 관계를 지나치게 단순화하는 문제가 있다. 이러한 문제는 소비자 효용함수의 구체적인 형태를 직접 관찰할 수 없고, 시장의 복잡성을 야기하는 모든 특성을 관측할 수 없기 때문에 발생한다(Hong et al., 2020). 다시 말하자면, 전통적으로 사용되는 선형 회귀모형은 변수 간의 관계를 단순 선형화하여 직관적인 해석이 가능하지만 현실 복잡성을 반영하기 쉽지 않고 모형의 한계로 인해 가치 추정의 정확도가 낮아지는 문제가 발생한다.

전통적인 헤도닉 가격 모형 기반의 부동산 분석의 단점을 보완하기 위해 머신러닝을 이용한 부동산 가격 분석이 활용되고 있다. 〈표 2〉는 부동산 분석에서 선형 회귀모형과 보완적으로 이용된 머신러닝에 대한 연구를 종합하고 있다. 〈표 2〉의 부동산 가격 예측력의 비교에 따르면, 선형 회귀모형보다 비선형 머신러닝 모형의 성능이 전반적으로 우수함을 확인할 수 있다.

특히 특성변수가 많이 포함되어 다중공선성 문제가 우려될 경우에는 트리 기반의 머신러닝 모형이 불안정성을 줄일 수 있는 것으로 보인다(김인호 · 이경섭, 2020). 그 중에서도 랜덤 포레스트(random forest) 모형이 주택 시장의 비선형성과 입지효과의 포착에 효과적이다(홍정의, 2021).

한편, 〈표 2〉에서 살펴본 머신러닝 기법을 이용한 관련 연구 동향은 크게 기존 분석 방법과 머신러닝 모형의 비교, 머신러닝 모형 간의 비교로 나눌 수 있다. 기존 분석과 머신러닝 기법을 비교한 연구는 선형회귀모형(김이환 외, 2022; 김인호 · 이경섭, 2020; 양건필 · 전해정, 2022; Hong et al., 2020), 다중회귀모형(배성완 · 유정석, 2018), 시계열 분석 모형(김승현 외, 2022)과 다양한 머신러닝 모형을 비교하였다. 머신러닝 모형을 분석하여 비교한 연구는 랜덤 포레스트, XGBoost, LightGBM과 같은 트리 기반의 머신러닝(김승현 외, 2022; 김이환 외, 2022; 김인호 · 이경섭, 2020; 배성완 · 유정석, 2018; 양건필 · 전해정, 2022; 이주미 외, 2021; Hong et al., 2020), 인공신경망 기반의 머신러닝 및 딥러닝(김이환 외, 2022; 배성완 · 유정석, 2018), SVM과 같은 분류 목적의 머신러닝(배성완 · 유정석, 2018)을 부동산 분야에 적용하여 모형별 특성과 각 성능을 비교하였다.

본 연구에서는 머신러닝 기법을 이용해 생략변수 편의 해소를 위한 소비자 특성 변수 도출 그리고 선형성 문제를 완화하기 위한 트리 기반 랜덤 포레스트 모형 적용을 통해서 부동산 시장 분석에서 헤도닉 가격 모형의 한계를 완화할 수 있는 방편으로 머신러닝 활용 방식을 제안한다.

2. 프롭테크

프롭테크(proptech)는 부동산(property)과 기술(technology)의 합성어로, 부동산 분야에 빅데이터, 인공지능, 블록체인, 확장 현실, 드론 등 다양한 기술을 활용하는 부동산산업, 서비스, 기업 등을 통칭하는 용어이다(경정익 · 권대중, 2022). 부동산 관련 사업 영역에서는 부동산 중개 및 임대, 개발부터 관리, 투자 및 자금조달(한국프롭테크포럼, 2023)에 이르기까지 다양한 영

 https://doi.org/10.30902/jrea.2023.9.1.23

〈표 2〉 국내 부동산 분야 비선형 머신러닝 관련 연구

논문명	저자	예측 성능 비교			
기계 학습을 이용한 공동주택 가격 추정: 서울 강남구를 사례로	배성완 · 유정석 (2018)	구분	MAE	RMSE	
		MRA	121.778	177.385	
		SVM	112.342	160.312	
		RF	108.033	153.806	
		GBRT	102.925	148.996	
		DNN	113.022	160.723	
A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea	Hong et al. (2020)	구분	R-squared		
		OLS	0.726056		
		RF	0.976198		
트리 기반 앙상블 방법을 활용한 자동 평가 모형 개발 및 평가: 서울특별시 주거용 아파트를 사례로	김인호 · 이경섭 (2020)	구분	R-squared		
		OLS	0.427		
		RF	0.940		
		XGBoost	0.959		
		LightGBM	0.958		
		Stacked(XG)	0.959		
		Stacked(LGBM)	0.931		
머신러닝을 이용한 부동산 지수 예측 모델 비교	이주미 외 (2021)	구분	RMSE		
		RF	0.1299		
		XGBoost	0.1236		
		LSTM	0.0978		
기계학습 방법론을 활용한 아파트 매매가격지수 연구	김이환 외 (2022)	구분	R-squared	RMSE	MAPE
		헤도닉	0.848	0.167	12.852
		인공신경망	0.937	0.107	7.763
		RF	0.996	0.066	4.432
기계학습과 XAI를 활용한 아파트 가격과 지역 특성과의 관계 분석	양건필 · 전해정 (2022)	구분	MAE	MAPE	RMSE
		OLS	1.997	29.936	3.126
		XGBoost	1.994	28.646	3.194
		RF	1.854	28.440	2.970
머신러닝과 패널고정효과를 활용한 아파트 실거래가 예측	김승현 외 (2022)	구분	MAE	MAPE	RMSE
		FE	2,487.82	4.7934	13,262.35
		RF	2,788.53	3.7794	9,329.08
		MARS	853.37	1.3443	5,064.80

주 : MRA, multiple regression analysis; SVM, support vector machine; RF, random forest; GBRT, gradient boosted regression tree; DNN, deep neural network; OLS, ordinary least squares; FE, fixed effects model; MAE, mean absolute error; RMSE, root mean squared error; MAPE, mean absolute percentage error.

역을 아우르고 있다. 국내에서는 2015년 부동산 관련 공공데이터 개방을 기점으로 프롭테크 관련 기업이 크게 성장하고 있다. 한국프롭테크포럼[3]에 따르면, 국내 프롭테크 시장의 누적 투자 금액이 2022년 기준 5조 원을 초과(한국프롭테크포럼, 2022)하였으며, 한국프롭테크포럼의 2018년 11월 기준 26개였던 회원사 수는 2023년 1월 기준 390개 회원사로 15배 성장하였다.

국내 주요 프롭테크 서비스 중 직방, KB부동산, 호갱노노, 아실 등은 온라인 플랫폼을 기반으로 부동산 관련 정보를 제공하는 서비스를 시행 중이다. 플랫폼 내 부동산 공급과 관련 정보는 실거래가, 위치 및 주변 시설, 분양 정보, 학군 정보부터 평면도, 외관 및 내부시설 사진, 상세 설명 텍스트까지 다양한 정보를 포함한다(〈표 3〉 참조).

프롭테크 플랫폼 서비스가 가지고 있는 주요한 특징은 앞선 부동산 공급 관련 정보에 더해 프롭테크 플랫폼 자체에서 생성되는 정보를 폭넓게 포함하고 있다는 점이다. 대표적으로 플랫폼에 참여하는 사용자의 리뷰, 연령대 및 방문 횟수 분석, 인기 순위 등이 사용자 생산 정보에 해당한다. 이러한 사용자 생산 정보는 부동산 소비자의 경향이나 평가, 심리, 입소문 등을 반영하고 있다.

프롭테크가 제공하는 자료의 성질과 성장세를 통해 짐작할 수 있는 점은, 부동산 공급 관련 정보를 플랫폼에 모았다는 이점과 더불어 사용자 생산 정보에 대한 소비자들의 관심이 높고 이러한 정보가 부동산 시장에서 중요한 역할을 하고 있다는 것이다.

프롭테크와 소비자의 주관을 반영하는 부동산 연구는 양적으로 많지 않고 방법론적인 측면에서

〈표 3〉 국내 주요 프롭테크 플랫폼 서비스 제공 정보

서비스명	서비스 내 제공 정보	
	업체 및 외부 생산 정보	사용자 생산 정보
직방[4]	단지 정보, 직방 시세, 실거래가, 학군 정보, 교통정보, 지역 분양단지 등	거주민 리뷰(사용자 별점 및 리뷰), 단지 인기 순위
KB부동산[5]	요약정보, KB시세, 실거래가, 평면도, 예산, 학군/교통, 중개사무소 등	전국/우리동네 커뮤니티(사용자 게시판), 지역 조회수 순위
호갱노노[6]	실거래가, 세금/대출 계산기, 평면도, 주차 공간, 빠른 배송 생활권, 주변 상권/교통/학군/입주 예정 아파트 정보 등	이야기(사용자 게시판), 사용자 관심 기반 추천, 실시간 방문자 분석
아실[7]	거래현황, 중개사, 단지 기본정보, 학군 정보, 주변 정보 등	아파트 거주자 커뮤니티(사용자 게시판), 사용자 조회수, 관심 사용자 수, 단지 인기 순위

3) http://proptech.or.kr

4) https://www.zigbang.com

5) https://kbland.kr

6) https://hogangnono.com

7) https://asil.kr

 https://doi.org/10.30902/jrea.2023.9.1.23

도 단편화되어 있는 것으로 보인다. 국내의 프롭테크 관련 연구는 프롭테크 업계 종사자들의 실태 조사 기반의 질적 연구를 통해 프롭테크 산업의 정책 제언을 모색하는 정책 연구(이정윤 외, 2021)가 있다. 그리고 부동산 시장에 대한 소비자의 경향, 심리, 평가 등을 반영하고자 시도한 대부분의 국내 연구는 뉴스 기사나 SNS 분석을 통해 이루어졌다(김다니 · 김란, 2021; 문태헌 · 김혜림, 2020; 박재수 · 이재수, 2021; 윤성진 외, 2021; 이종민 외, 2017; 장몽현 · 김한수, 2019).

본 연구에서는 머신러닝 기법 중 토픽모델링을 이용하여 프롭테크 플랫폼 내의 사용자 생산 정보를 헤도닉 가격 모형의 특성 변수로 구성하고자 한다. 이를 통해 소비적 측면의 정보를 가격 추정에 투입함으로써 부동산 시장의 소비자 중심적 변화와 현실성을 직접적으로 반영할 수 있을 것으로 기대한다.

3. 선행연구와의 차별점

본 연구의 기여점은 크게 두 가지로, 프롭테크 플랫폼 서비스의 사용자 정보를 가격 추정의 특성 변수로 사용하여 부동산 시장에서의 소비자 중심 정보의 유의성을 확인했다는 것과 부동산 시장 가격 추정 분야에서 연구 전반에 머신러닝 기법을 사용했다는 것이다.

첫째로, 본 연구는 사용자 생산 정보를 헤도닉 가격 모형의 특성 변수로 활용하여, 가격 추정에 있어 최근 부동산 시장의 현실성을 반영할 수 있고, 생략변수 편의 등의 구조적 문제 또한 개선 가능하다. 프롭테크 플랫폼 서비스의 사용자 생산 정보는 부동산 시장에서 통용되는 소비자 중심의 정보에 해당한다. 반면, 기존의 헤도닉 가격 모형의 특성 변수는 대부분 공급자 위주의 정보로 구성되어 있다. 그러나 최근 프롭테크 시장과 사용자 생산 정보의 수요 성장으로 인해 가격 추정에 있어 소비자 중심의 정보를 고려할 필요가 있다. 이러한 소비자 중심 정보가 생략되면 선형 회귀모형 기반의 헤도닉 가격 모형에서 생략변수 편이가 발생할 우려가 크다.

추가로 본 연구는 헤도닉 가격 모형의 특성 변수로 가공하여 가격 추정에 적극적으로 활용한다는 점에서 의의가 있다. 본 연구에서는 프롭테크 플랫폼 서비스의 사용자 생산 정보 중 리뷰 텍스트를 이용하여 비정형 데이터로 특성 변수를 구성한다. 부동산 시장 분야에서 비정형 데이터를 활용한 사례는 뉴스 기사와 SNS를 분석하여 시장 동향을 살피거나 가격과의 상관을 살핀 경우가 주를 이룬다. 이와 같은 선행 연구는 비정형 데이터를 가격 추정에 있어 보완적 지표로 이용했다.

두 번째로, 본 연구는 특성 변수 추출과 분석모형에 머신러닝을 동시에 적용하여 연구 전반에 머신러닝 기법을 활용한다는 점에서 선행연구와는 차별된다. 최근 머신러닝 기법의 발달로, 부동산 시장의 가격 추정 분야 전반에서 머신러닝을 활용한 사례가 늘고 있다. 다만 머신러닝이 분석모형에 적용되는 경우가 대부분이며, 분석모형 외에 적용되더라도 분석모형과 동시에 머신러닝을 적용한 연구는 드물다.

이에 본 연구는 부동산 시장 분야에 있어 헤도닉 가격 모형의 특성 변수로 프롭테크 플랫폼 서비스의 사용자 생산 및 비정형 데이터를 활용하

고, 특성 추출과 분석 모형에 동시에 부동산 가격 예측에 머신러닝 기법을 적용한 첫 연구라고 할 수 있다.

Ⅲ. 자료 및 분석모형

1. 특성 추출

특성 추출을 설명하기에 앞서, 특성 추출에 사용한 프롭테크 플랫폼의 사용자 리뷰 텍스트에 대해 간략히 설명하고자 한다. 본 연구를 위하여 '직방'의 '거주민 리뷰' 텍스트를 사용하였다. 해당 텍스트는 "리뷰해주신 아파트를 주변 사람에게 추천하는 이유를 말씀해주세요."라는 안내에 따라 작성되었으며, 실제 내용 또한 해당 아파트의 추천할만한 긍정적인 요소에 대하여 주로 서술하고 있다.

이와 같은 프롭테크 플랫폼의 사용자 리뷰 텍스트로부터 소비자 중심 특성 변수를 추출하기 위하여, 토픽 모델링을 실시하였다. 리뷰 텍스트는 소비자가 직접 부동산의 특성에 대해 기술하고 있으므로, 토픽 모델링을 통해 소비자가 생각하는 부동산의 특성을 추출하고, 유형화하여 변수로 생성할 수 있다.

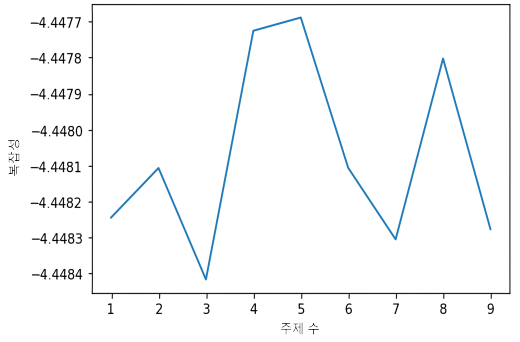
이를 위하여 토픽 모델링의 대표적인 유형인 LDA(잠재 디리클레 할당, latent dirichlet allocation)을 사용하였다. LDA는 텍스트와 같은 이산 데이터 수집에 대한 생성 확률 모델로(Blei et al., 2003), 전체 문서들에서 잠재적인 주제를 찾아내는 확률 기반의 알고리즘이다. 다시 말해, LDA는 문서에 혼합된 주제들로부터 관찰된 각 단어들이 확률분포에 의해 생성된다고 가정하고, 단어와 주제 간의 조건부 확률을 통해 각 주제에 대한 가장 일반적인 단어를 결정할 수 있다. 문서마다 토픽을 할당한다는 측면에서 머신러닝의 다양한 클러스터링 기법과 비슷한 측면이 있지만, 하나의 문서 내에 다양한 토픽이 존재할 수 있다고 가정한다는 점에서 차이를 보인다(윤성진 외, 2021).

LDA 모델의 사전을 구성하기 위하여, 전체 리뷰 텍스트마다 명사를 추출하여 각 문서를 구성하고 있는 단어를 도출한다. 이때 하나의 문서는 하나의 아파트 단지에 대한 리뷰의 모음으로 구성된다. 단순히 문서 내에 언급이 많은 단어가 아닌, 문서 내에서 의미를 가지는 단어에 대해 사전을 구성하기 위해 TF-IDF(term frequency-inverse document frequency) 기준, 유의미한 상위 100개 단어로 사전을 구성한다. TF-IDF란, 텍스트 마이닝을 위해서 문서 내부의 단어가 얼마나 중요한지를 평가하기 위한 표현 방식이다(박종영 · 서충원, 2015).

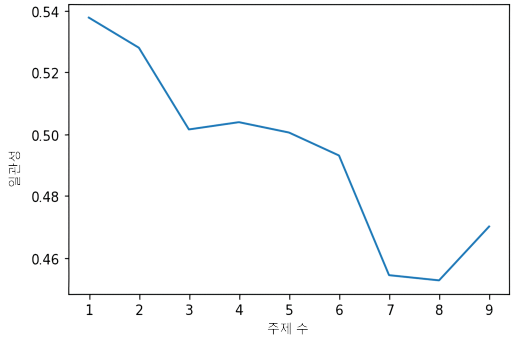
구성한 사전을 이용하여 전체 문서에 대해 학습시키고, 전체 문서에 대하여 적절한 주제의 수를 탐색한다. 토픽 모델링에서 최적의 주제 수를 탐색하기 위한 기준은 복잡성(perplexity)과 일관성(coherence)이 있다(윤성진 외, 2021). 첫째로, 복잡성이란 주제를 구성하는 단어들이 확률적으로 분포하고 있다는 데에 착안하여 확률 모델이 결과를 얼마나 정확하게 예측하는지를 말해준다(Blei et al., 2003). 이에 복잡성이 낮을수록 단어가 가진 확률이 주제를 정확하게 예측한다고

해석한다. 둘째로, 일관성이란 토픽과 그 하위 집단에 포함된 단어들의 적합성을 나타낸다(Röder et al., 2015). 이에 일관성이 높을수록 하나의 토픽이 의미론적으로 일관되게 구성되었다고 해석한다. 따라서, 토픽모델링 시에는 낮은 복잡성과 높은 일관성을 보이는 주제 수를 선택해야 한다.

주제의 복잡성과 일관성을 고려하여 적절한 주제의 수를 탐색하기 위하여 〈그림 1〉, 〈그림 2〉와 같이 주제 수 변화에 따른 복잡성과 일관성 변화 그래프를 생성하였다. 〈그림 2〉에서 주제 수가 2개보다 많아질 경우 일관성이 급격히 낮아지지만, 〈그림 1〉에서 복잡성은 주제 수에 따라 변동이 크다. 따라서 상대적으로 높은 일관성을 유지하면서도 가장 낮은 복잡성을 가진 주제 수인 3개로 결정하였다. 결정된 주제는 〈표 4〉에 구분되어 있다. 또한, 주제 간 단어 분포가 이질적이고 주제 내 단어 분포가 동질적이며, 주제별 특성이 명확히 드러나는 모델을 최종 모델로 선정하였다. 최종 선정된 LDA 모델을 적용하여 각 문서에 대해 각 주제에 속할 확률을 계산하고, 각 문서마다 가장 확률이 높은 주제를 할당하였다.



〈그림 1〉 주제 수에 따른 복잡성 변화



〈그림 2〉 주제 수에 따른 일관성 변화

2. 변수 구성

본 연구의 분석 대상은 대구광역시의 8개 구군(중구, 동구, 서구, 남구, 북구, 수성구, 달서구, 달성군)에서 2018년부터 2022년까지 5년간 거래된 아파트 개별 세대를 대상으로 한다. 분석 대상을 추정하기 위한 특성 변수는 선행연구를 참고하되 수집할 수 있는 범위 내에서 선정하였다. 또한, 대구광역시의 지역적 특성을 살리기 위하여 대구광역시를 분석 대상으로 한 이소영 외(2020), 장몽현 · 김한수(2020)의 연구를 참고하여 교육과

〈표 4〉 주제별 주요 단어 및 문서 수

주제	주요 단어	문서 수(개) (%)
생활권	마트, 지하철, 교통, 환경, 백화점, 출퇴근, 병원	20,239 (75.6)
교육	학군, 학교, 초등학교, 학원, 중학교, 고등학교, 교육	3,697 (13.8)
실속	가격, 대비, 투자, 재개발, 평수, 신혼부부	2,830 (10.6)
전체		26,766

관련된 특성 변수인 초등학교, 중학교, 고등학교, 학원을 추가로 입지 특성에 구성하였다. 본 연구에서는 프롭테크 플랫폼의 소비자 생산 정보를 이용하여 헤도닉 가격 모형의 변수로 구성하는 것을 목적으로 한다. 따라서 선행연구에서 살펴본 기존 특성 변수에 더하여 특성 추출에서 생성한 소비자 중심 특성을 추가하였다.

〈표 5〉와 같이 구성한 변수를 생성하기 위하여 공공데이터 및 프롭테크 플랫폼에서 수집한 데이터를 이용하였다. 본 연구에서 이용한 데이터의 출처 및 활용 데이터는 〈표 6〉과 같다.

종속변수인 가격과 세대 및 단지 특성의 층은 국토교통부의 아파트 매매 실거래 상세자료의 거래금액과 전용면적, 층을 이용하였다. 대부분의 아파트 매매 가격 추정은 세대별 시세를 이용하지만, 층수나 평형이 다른 경우의 용이한 비교를 위해 '공급 면적당 가격'(거래가격 / 공급면적)을 별도로 계산하여 종속변수로 사용하였다.

세대 및 단지 특성에 해당하는 노후도, 층, 난방유형, 규모, 시공사, 용적률 6개 변수는 직방 홈페이지의 아파트 단지별 정보에서 소비자가 직접 접근하여 확인할 수 있는 정보로 구성하였다. 입지 특성에 해당하는 지하철, 공원, 초등학교, 중학교, 고등학교, 학원의 6개 변수를 생성하기 위해 각 분야에 해당하는 공공데이터를 활용하였으며, 도로명주소 및 위도와 경도를 기준으로 각 아파트 단지로부터 각 시설까지의 직선거리를 측정하여 변수로 생성하였다.

소비자 중심 특성 변수를 생성하기 위하여 프롭테크 플랫폼 서비스 직방에서 제공하는 사용자 생산 정보인 '거주민 리뷰'를 수집하였다. 직방에서 제공하는 '거주민 리뷰'는 타 플랫폼에서 제공하는 사용자 생산 정보에 비해 '리뷰'라는 주제에

〈표 5〉 특성 변수 및 세부 변수

특성	세부 변수
세대 및 단지 특성	노후도, 층, 난방유형, 규모, 시공사, 용적률
입지 특성	지하철, 공원, 초등학교, 중학교, 고등학교, 학원
소비자 중심 특성	생활권_주제, 교육_주제

〈표 6〉 분석 데이터 출처 및 활용 데이터

구분	출처	활용 데이터
아파트매매 실거래 상세자료	국토교통부	거래금액, 도로명, 법정동, 전용면적, 단지 이름, 건축 연도, 층 등
아파트 단지 데이터	직방	단지 이름, 건축 연도, 세대수, 최고층수, 난방유형, 리뷰 텍스트 등
건설업체 시공 능력 평가 공시	국토교통부	아파트 공사실적 순위(2017~2021년)
전국 도시철도역사 정보 표준데이터	국가철도공단	위도, 경도, 역사 도로명주소
전국 도시 공원 정보 표준데이터	국토교통부	위도, 경도, 소재지 도로명주소
전국 초중등학교 위치 표준데이터	교육부	위도, 경도, 소재지 도로명주소
학원 및 교습소 현황	대구광역시교육청	도로명주소

 https://doi.org/10.30902/jrea.2023.9.1.23

맞게 작성되고 있으며, 아파트추천(총평), 교통여건, 주변환경, 단지관리, 거주환경으로 정리된 각 50자 이상의 텍스트를 제공하고 있는 것이 장점이다.

특성 추출 절에서 설명한 바와 같이, LDA 모델을 활용하여 '거주민 리뷰' 텍스트로부터 각 문서마다 할당된 주제를 기준으로 헤도닉 모형에 포함시키기 위하여 2개 더미변수를 생성하였다. 2개 더미변수는 생활권_주제, 교육_주제의 해당 여부에 대한 더미변수이다. 3개로 할당된 주제들 중 특히 실속 관련 주제의 경우 가격의 측면을 반영하여 교육이나 생활권과 같은 특징적인 분류를 나타내는 항목들을 제외한 나머지 성격을 갖는다. 본 분석이 매매 가격 분석이기 때문에 가격적 측면에 대한 상대적인 교육 및 생활권의 효과를 통계적으로 모형화 하기 위하여 생활권_주제와 교육_주제를 더미변수로 취급하여 모형에 포함하였다.

아파트 매매 실거래 상세 자료 중 대구광역시 8개 구군에서 2018년 1월부터 2022년 11월까지 5년간 발생한 아파트 매매 거래 내역 28,289개를 수집하였다. 이 중, 직방 데이터 기준 대구광역시 아파트 단지 기본 정보 및 사용자 리뷰가 존재하는 1,136개 내역과 매칭하여, 최종 26,766개 거래 내역을 기반으로 데이터셋을 구성하였다. 최종 26,766개 데이터에 대하여 1개 종속변수, 14개 특성 변수를 구성하였으며, 최종 데이터셋과 기초통계량은 〈표 7〉, 〈표 8〉과 같다.

〈표 7〉 최종 데이터셋 및 변수 구성

구분		변수명	단위	세부 내용
종속변수		가격	만 원	아파트 공급면적당 매매가격
특성변수	세대 및 단지 특성	노후도	연	경과 연수
		층	층	거래 층수
		난방유형	-	개별난방=1, other=0
		규모	개	단지 내 세대 수
		시공사	-	아파트 공사실적 순위 10순위 이내=1, other=0
		용적률	퍼센트	대지면적에 대한 연면적의 비율
	입지 특성	지하철	미터	최근접 지하철역 직선거리
		공원	미터	최근접 공원 직선거리
		초등학교	미터	최근접 초등학교 직선거리
		중학교	미터	최근접 중학교 직선거리
		고등학교	미터	최근접 고등학교 직선거리
		학원	개	1km 반경 이내 사설학원 수
	소비자 중심 특성	생활권_주제	-	생활권 주제 문서=1, other=0
		교육_주제	-	교육 주제 문서=1, other=0

〈표 8〉 최종 데이터셋 연속형 변수 기초 통계량

구분	평균	표준편차	최솟값	최댓값
가격	468	204	119	2,099
노후도	20	12	0	46
층	9.57	7.28	1	53
규모	601	458	32	1,999
용적률	270	130	58	826
지하철	608	322	65	17,066
공원	351	207	40	1,060
초등학교	377	178	90	1,200
중학교	1,926	2,603	92	10,300
고등학교	747	414	92	5,800
학원	59	58	0	508
N	26,766			

3. 분석모형

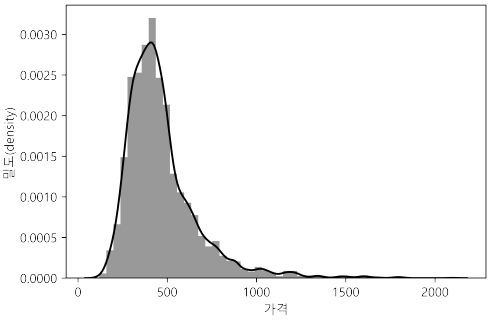
1) 선형회귀모형

$$\ln P_i = D_i'\beta + W_i'\gamma + X_i'\delta + \epsilon_i \qquad \text{(식 1)}$$

헤도닉 가격 모형은 (식 1)과 같다. 종속변수는 로그로 변환된 공급 면적당 가격($\ln P_i$)이고 D, W, X 모두 특성 변수이다. 이 중 본 연구의 주관심은 소비자 중심 특성을 나타내는 D의 추정 계수인 소비자 중심 특성 가격(β)이다. W는 세대 및 단지 그리고 입지 특성들을 포함하는 특성 변수이다. X는 나머지 통제변수에 해당하는 특성 변수로 상수항을 포함하여 연도별 효과, 시군구별 효과 등을 의미한다. 끝으로 ϵ은 오차항이다.

앞서 설명한 소비자 중심 특성 변수와 연관되는 기타 세대, 단지, 입지 특성들(W)은 기존 연구를 참고하여 통제함으로써 기존의 전통적인 헤도닉 모형에서 생략변수 편의를 발생시킬 수 있는 통제변수를 모형에 포함할 수 있도록 구성하였다. 특히 선행연구를 참고하여 노후도(경과 연수), 층(거래 층수)에 대하여 각각 제곱하였다. 김태훈(2004)은 경과 연수는 아파트 가격과 비선형 관계(nonlinear relationship)를 가지며, 경과 연수가 오래된 아파트일수록 가격이 감소하여야 하지만 재건축 사업 효과 등의 기대심리로 인해 가격은 오히려 상승한다는 것을 제시하였다. 더 나아가, 김태훈(2004)은 아파트 가격에 대한 비선형 관계의 추정을 위해 경과 연수를 제곱 및 세제곱하여 다항회귀모델에서의 유의성을 확인하였다. 또한, 정수연 외(2009)는 해외 사례와 달리 국내 아파트의 경우 고층부보다는 중간층에 대한 선호가 상대적으로 크고, 이에 따라 아파트 가격과 층수가 비선형 관계를 가지는 것을 확인하였다.

종속변수는 로그 변환을 하였는데 〈그림 3〉과 같이, 종속변수인 가격에 대하여 데이터의 분포가 오른쪽으로 꼬리가 긴 분포를 가진 것을 확인하여, 로그 변환을 통해 정규성을 확보하기 위함이다.



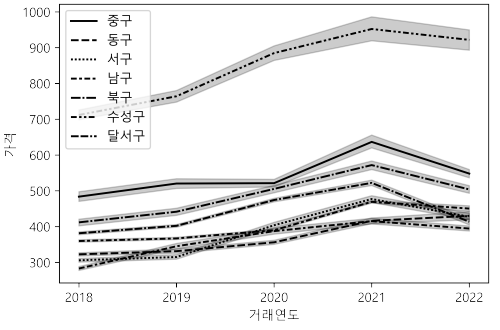
〈그림 3〉 아파트 매매 가격 히스토그램 및 커널 밀도

추가적인 통제 변수 특히, 시간과 시군구의 효과에 따른 영향력을 통제하기 위해 다음과 같이 선정하였다. 아파트 매매 가격은 매물이 위치한 지역구나 거래하는 시기에 따라 많은 영향을 받는다. 거래 시점의 영향을 통제하는 방법 중 시간 더미변수를 이용하는 방법이 가장 높은 설명력을 가진다는 박종기 외(2011)의 연구를 참고하여, 거래 구군과 거래 연도를 통제하기 위하여 구군 더미변수, 시간 더미변수를 추가하였다. 통제변수는 5년간의 거래 연도와 대구광역시 8개 구군의 거래 지역에 대한 더미변수로 이루어져 있다.

헤도닉 가격 모형은 지역별로 주요 특성이 다를 수 있고, 또 특성 가격이 지역별로 차이가 날 수 있기 때문에 이분산(heteroskedasticity) 현상이 종종 나타난다(이용만, 2008). 따라서 〈그림 4〉와 같이 공간적 분석 대상인 대구광역시 8개 구군의 연간 가격 변화가 상이한 것을 확인하여, 이분산을 고려하는 분산 추정량을 이용하였다.

2) 랜덤 포레스트

랜덤 포레스트는 각 의사결정나무(decision tree)가 독립적으로 샘플링된 랜덤 벡터의 값에 의존하고 모든 의사결정나무에 대해 동일한 분포를 갖는 예측 변수의 조합이다(Breiman, 2001). 이때 의사결정나무란 데이터의 규칙과 패턴을 파악하고 데이터셋을 분류 혹은 예측하는 알고리즘 기법이다. 랜덤 포레스트의 절차(Breiman, 2001)는 다음과 같다. 우선 전체 데이터셋으로부터 학습 데이터와 평가 데이터를 분할한다. 학습 데이터로부터 각 N개의 데이터를 B번 추출하여 B개의 부스트랩(bootstrap) 데이터셋을 생성한다. 부스트랩이란 중복을 허용하여 샘플을 추출하는 방법을 말한다. 다음으로 각 부스트랩의 데이터를 이용하여 각 의사결정나무 B개를 학습시킨다. 미리 설정한 초매개변수(hyper-parameter)에 따라 의사결정나무의 형태가 달라진다. 초매개변수는 의사결정나무의 분할 횟수(max_depth), 리프 노드의 최소 데이터 수(min_samples_leaf), 분할 시 최소 데이터 수(min_samples_split), 의사결정나무의 개수(n_estimators)로 이루어진다. 학습된 각 의사결정나무는 평가 데이터를 입력하여 서로 다른 예측 결과를 산출할 수 있다. 마지막으로 전체 의사결정나무의 결과들에 대하여 투표를 통해 최종 결과가 결정된다. 회귀 모형의 경우, 각 결과값의 평균으로 최종 결과가 산출된다.



〈그림 4〉 거래 지역별 연간 아파트 매매 가격 변화

랜덤 포레스트는 선형 회귀모형과 달리 다중공선성과 변수 간 비선형성을 고려하지 않아도 되는 장점을 가진다. 비모수적 방법으로서 선형 모델보다 더 복잡한 패턴을 발견하기 쉽다. 다만, 수많은 의사결정나무를 계산하기 위하여 학습 시간이 높고 연산량이 많은 것이 단점이다.

랜덤 포레스트는 연구자가 직접 입력해야 하는 초매개변수에 따라 모형의 예측 성능이 달라진다. 최적의 모형을 구성하기 위하여 그리드 서치(grid search) 기법을 통해 가장 좋은 예측 성능을 가지는 초매개변수를 탐색하였다. 그리드 서치란, 초매개변수에 들어갈 수 있는 값에 대해 모든 경우마다 예측 성능을 계산하여 최적의 값을 탐색하는 방식이다. 의사결정나무의 분할 횟수(max_depth)는 높아질수록 학습 데이터에 과적합되어 예측 성능을 떨어뜨리기 쉬우므로 이를 고려하여 조정하였다. 그리드 서치를 통해 도출한 초매개변수는 〈표 9〉와 같다.

〈표 9〉 랜덤 포레스트 모형의 초매개변수

구분	설정값
max_depth	14
min_samples_leaf	8
min_samples_split	8
n_estimators	100

Ⅳ. 분석 결과

1. 선형회귀모형

선형회귀모형 추정과 성능 평가를 위하여 전체 데이터를 학습용 70%, 평가용 30%의 비율로 나누어 학습과 예측을 별도의 데이터로 실시하였다. 학습 데이터만으로 모형 학습을 실시한 뒤, 평가 데이터에 학습된 모형을 적용하여 예측값 및 예측성능을 도출하는 방식이다. 〈표 10〉의 분석 결과는 학습 데이터로 도출되었으며, 〈표 11〉의 예측성능은 평가 데이터로 도출되었다.

〈표 10〉 헤도닉 가격 모형 추정 결과

변수	추정계수 (표준오차)	변수	추정계수 (표준오차)
생활권_주제	0.116** (0.008)	층 /10,000	−7.0 (0.001)
교육_주제	0.115** (0.01)	난방유형	−0.014* (0.005)
시공사	0.085** (0.006)	노후도	−0.046** (0.001)
노후도_제곱 /10,000	8.0** (0.212)	2021년	0.357** (0.005)
공원 /10,000	3.0** (0.138)	2022년	0.234** (0.005)
학원 /10,000	3.0** (0.469)	2020년	0.199** (0.004)
층_제곱 /10,000	2.0** (0.177)	2019년	0.073** (0.004)
규모 /10,000	0.970* (0.046)	수성구	0.464** (0.012)
중학교 /10,000	−0.334** (0.030)	동구	0.197** (0.026)
지하철 /10,000	−0.386* (0.127)	북구	−0.068** (0.007)
고등학교 /10,000	−1.0** (0.072)	서구	−0.084** (0.0011)
초등학교 /10,000	−2.0** (0.0115)	달서구	−0.097** (0.01)
용적률 /10,000	4.0** (0.210)	남구	−0.107** (0.008)
상수항	6.53** (0.017)	달성군	−0.291** (0.006)
N	26,766		
Adj. R^2	0.721		

주 : * p〈0.05, ** p〈0.01.

소비자 중심 특성 변수를 포함한 헤도닉 가격 모형의 회귀분석 추정결과는 〈표 10〉에 제시되어

〈표 11〉 선형 회귀모형 예측 성능 비교

구분	전체 특성 변수	소비자 중심 특성 변수 제외
Adj. R^2	0.714	0.637
RMSE	0.200	0.225

주 : RMSE, root mean squared error.

있다. 주관심 변수인 '생활권_주제'와 '교육 주제' 모두 1%의 통계적 유의수준 하에서 기준변수로 설정된 '실속_주제'에 비해 각각 약 11.6%, 11.5%의 한계 증가 효과가 있는 것으로 분석되었다.

소비자 중심 특성 변수와 통제 변수 외 14개 변수 중 5%, 1% 유의수준을 만족하는 경우의 추정결과는 다음과 같다. 세대 및 단지 특성 중, '시공사'(시공사 10순위 이내)와 '난방유형'(개별난방)은 각 기준 상태에 비하여 약 8.5%, 약 -1.4%의 한계 증가 효과가 있다. 다음으로 '노후도_제곱', '층_제곱', '규모', '용적률'이 각 1단위 변화할 때에 종속변수인 '가격'이 평균적으로 각각 0.08%, 0.02%, 0.00097%, -0.04% 변화하였다. 입지 특성의 경우, '공원', '학원', '중학교', '지하철', '고등학교', '초등학교'가 각 1단위 변화할 때에 종속변수인 '가격'이 평균 0.03%, 0.03%, -0.0033%, -0.0039%, -0.01%, -0.02% 변화하였다.

소비자 중심 특성 변수를 포함하는 것에 대한 영향력을 확인하기 위하여 이들을 제외한 경우를 포함하여 회귀분석 모형의 성능 평가 지표인 조정된 결정계수(adjusted R-squared, Adj. R^2), RMSE(root mean squared error)를 비교하였다. 포함되는 변수의 수가 변하기 때문에 비교를 위해 조정된 결정계수를 이용하였다. 〈표 11〉의 비교 결과에 따르면 소비자 중심 특성 변수를 제외한 경우에 조정된 결정계수가 0.077 감소하고 RMSE가 0.025 증가하였다.

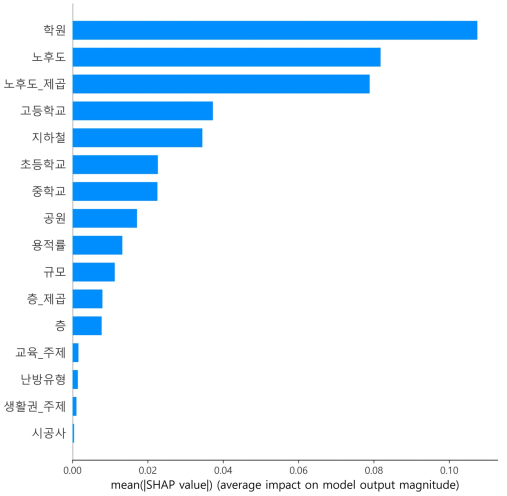
선형회귀모형의 예측 성능은 약 71%로 선행연구에서 살펴본 머신러닝을 적용한 경우에 비하여 상대적으로 낮은 예측 성능을 보인다. 그러나 추정계수를 통해 개별 독립변수가 종속변수에 미치는 영향력을 정량적으로 파악할 수 있으며, p-value를 통해 개별 독립변수의 통계적 유의성을 확인할 수 있다는 데에 의의가 있다.

2. 랜덤 포레스트

선형회귀모형과 동일하게, 과적합을 방지하기 위하여 데이터셋을 분리하였으며, 소비자 특성 변수의 포함 여부에 따라 별도로 예측 성능을 비교하였다.

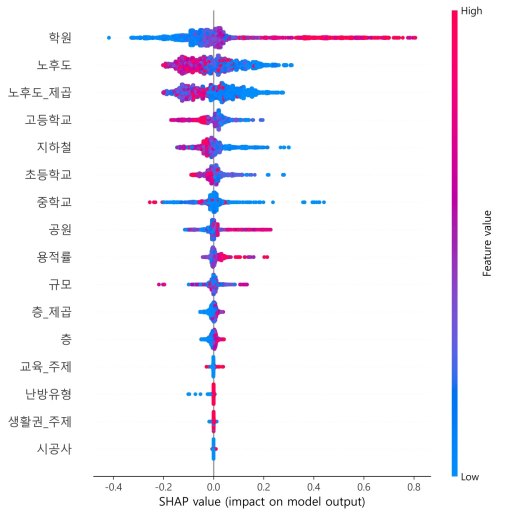
랜덤 포레스트 분석 결과를 해석하기 위하여, 예측 해석을 위한 통합 프레임워크인 SHAP (SHapley Additive exPlanations; Lundberg and Lee, 2017)를 이용하였다. SHAP는 모델을 구성하는 각 특성에 예측에 대한 중요도를 할당하여 앙상블 또는 딥러닝 모델과 같이 복잡한 모델을 보다 쉽게 해석할 수 있다.

본 연구의 랜덤 포레스트 모델에 대하여 SHAP의 변수 중요도와 예측 영향도를 출력한 결과는 〈그림 5〉, 〈그림 6〉과 같다. 〈그림 5〉는 각 특성이 예측에 미치는 절대적인 영향을 나타낸다. 이때 학원, 노후도, 노후도_제곱, 고등학교, 지하철, 초등학교, 중학교, 공원, 용적률, 규모, 층_제곱, 층, 교육_주제, 난방유형, 생활권_주제, 시공사 순으로 변수 중요도가 높았다.



주 : SHAP, shapley additive explanations.

〈그림 5〉 랜덤 포레스트 SHAP 변수 중요도



주 : SHAP, shapley additive explanations.

〈그림 6〉 랜덤 포레스트 SHAP 예측 영향도

〈그림 6〉은 〈그림 5〉와 같은 내용을 다시 표현한 것이다. y축은 각 특성 변수, x축은 SHAP 값, 색은 특성 변수의 값을 나타낸다. 빨간색일수록 높은 값이고, 파란색일수록 낮은 값이다. y축은 변수 중요도에 따라 내림차순으로 정렬되어 있다. 가장 변수 중요도가 높은 학원의 경우, 아파트 주변의 학원 수가 많을수록 예측에 긍정적인 영향을 미쳤다고 해석할 수 있다. 소비자 중심 특성 변수의 경우, 교육_주제에 해당하는 빨간색 점이 0 이상의 SHAP 값을 보이며 예측에 긍정적인 영향을 미쳤다. 그러나 생활권_주제는 그렇지 않아 예측에 큰 영향을 미치지 못한 것으로 나타났다.

〈표 12〉에 제시된 소비자 중심 특성 변수 포함 여부에 따른 예측 성능 비교를 보면, 소비자 중심 특성 변수를 제외한 경우에 조정된 결정계수가 소폭 감소하고 RMSE가 소폭 증가하였다. 소비자 중심 특성 변수의 포함 여부에 따라 큰 차이를 보이지는 않지만 소비자 중심 특성 변수가 포함된 경우, 약 93%의 예측력을 보이고 있다.

랜덤 포레스트 모형과 선형회귀 모형의 비교를 위해 예측성능을 이용한다면 랜덤 포레스트의 예측 성능은 약 93%로 선형 회귀모형보다 성능이 우수하다고 평가할 수 있다. 만약 아파트 매매 가격에 대한 예측이 목적이라면 선형 회귀모형에 비하여 랜덤 포레스트를 통한 아파트 매매 가격 예측이 더 나은 성능을 보일 가능성이 높다. 그러나 개별 독립변수가 종속변수에 미치는 영향을 변수

〈표 12〉 랜덤 포레스트 예측 성능 비교

구분	전체 특성 변수	소비자 중심 특성 변수 제외
Adj. R^2	0.9279	0.9278
RMSE	0.10063	0.10069

주 : RMSE, root mean squared error.

 https://doi.org/10.30902/jrea.2023.9.1.23

중요도를 통하여 대략적으로만 파악할 수 있으며, 개별 독립변수의 통계적 유의성을 확인할 수 없다는 한계가 있다. 즉, 아파트 매매 가격에 미치는 요인에 대한 통계적 분석은 선형 회귀분석을 이용하는 것이 나을 것이다.

Ⅴ. 결론

본 연구는 프롭테크 플랫폼 서비스 직방의 '거주민 리뷰' 텍스트를 이용하여 헤도닉 가격 모형의 특성 변수로 구성하였으며, 이를 통해 소비자 중심의 정보가 아파트 가격 추정에 얼마나 영향을 미치는지 분석하였다. 또한, 특성 추출과 분석 모형에서 머신러닝 기법을 사용하여 연구 전반에 머신러닝을 적용하였다. 이를 통해 부동산 가격 분석을 위한 전통적인 헤도닉 모형에서 기존 통제변수들과 소비자 정보의 상관관계에서 발생할 수 있는 생략변수 편의 가능성을 소비자 특성 정보를 모형에 포함함으로써 내생성의 문제를 완화할 수 있었고, 더불어 선형 모형의 추정의 한계를 머신러닝 적용을 통해 일부 극복할 수 있었다.

비정형 구조인 '거주민 리뷰' 텍스트를 변수화하기 위하여 머신러닝 토픽 모델링 기법 중 LDA (잠재 디리클레 할당) 모델을 이용하였다. 이에 대구광역시 8개 구군의 '거주민 리뷰'를 이루고 있는 주제를 '생활권', '교육', '실속'으로 도출하였으며, LDA 모델 학습 결과로 아파트 단지별로 주제를 할당하여 더미변수의 형태로 소비자 중심 특성 변수 '생활권_주제, '교육_주제'를 구성하였다.

국내 및 대구광역시 관련 선행연구를 참고하여 세대 및 단지 특성, 입지 특성으로 헤도닉 가격 모형의 특성 변수를 구성하였으며, 기존 선행연구에서 참고한 특성 변수에 더하여 프롭테크 플랫폼 서비스의 사용자 생산 정보에서 추출한 소비자 중심 특성 변수를 추가하였다.

헤도닉 가격 모형에서 전통적으로 사용된 선형 회귀모형 분석 결과, 소비자 중심 특성 변수가 높은 유의함을 보였으며 생활권_주제, 교육_주제 순으로 종속변수에 대하여 각 11.6%, 11.5%의 영향력을 보였다. 또한, 소비자 중심 특성 변수가 포함되었을 경우, 그렇지 않은 경우보다 조정된 결정계수가 0.077만큼 향상되어 0.714에 해당하는 예측 성능을 보인다.

랜덤 포레스트를 시행한 결과, 0.928의 높은 예측 성능을 보였다. 다른 특성 변수에 비하여 소비자 중심 특성 변수가 높은 변수 중요도를 보이지 않았으나, 교육_주제에 해당할수록 예측에 긍정적인 영향을 미치는 것을 확인하였다. 다만, 선형회귀 모형의 결과와는 달리 소비자 중심 특성 변수가 포함되었을 경우, 그렇지 않은 경우보다 조정된 결정계수 측면의 예측 성능이 0.0001 향상되어 모형 간 예측 성능에서 큰 차이를 보이지 않았다. 이는 새롭게 도출된 변수로서 소비자 중심 특성 변수의 정의나 측정 등에서 한계가 있었던 것으로 보인다. 향후 연구에서는 소비자 중심 특성 변수의 고도화 및 다양화를 통해 변수의 안정된 성능을 얻는 과정이 필요하다.

한편, 본 연구에서 사용한 아파트 실거래 데이터는 내재된 공간적 자기 상관으로 인해 공간적 종속성 및 이분산성 문제가 있다. 본 연구에서는 헤도닉 가격 모형을 사용함과 동시에 공간적 자기

상관을 해소하기 위하여 종속변수를 로그 변환한 준로그모형 사용, 일부 변수의 제곱항 사용, 이분산을 고려한 분산 추정량 이용 등의 노력을 하였다. 그럼에도 불구하고 선형 회귀모형은 비선형 데이터를 반영하기 어려운 모형 자체의 한계로 인해 상대적으로 낮은 예측 성능을 보였다. 한편, 랜덤 포레스트는 비모수 모형으로서 오차의 분포 가정으로부터 자유로운 특징으로 인해 상대적으로 우수한 예측 성능을 보였다. 향후 연구에서는 아파트 실거래 데이터에 대하여 공간적 자기 상관 문제에 특화된 모형을 사용할 필요가 있으며, 다양한 형태의 비선형 모형을 사용하여 성능을 비교할 필요가 있다.

본 연구는 다음의 추가적인 한계점들이 존재한다. 먼저, 본 연구가 활용한 데이터가 전국이 아닌 대구광역시의 8개 구군만을 대상으로 하여 분석하였다는 한계가 있다. 다음으로, 아파트 가격을 형성하는 많은 특성 중 본 연구에서 구성한 특성 변수 이외에 누락된 특성이 존재할 가능성이 있다. 또한, 생략변수 편의를 줄이기 위해 소비자 중심 특성 변수를 도입하였으나, 소비자의 주관적 의견에서 추출된 소비자 중심 특성과 객관적으로 측정된 입지 특성 간의 상관 및 영향이 검증되지 않았다. 추후 연구에서는 소비자 중심 특성 변수의 고도화를 위하여 기존 특성 변수와 소비자 중심 특성 변수 간의 상관을 측정하고, 변수 간의 상관을 고려할 필요가 있다. 마지막으로 프롭테크 플랫폼 서비스에서 제공하고 있는 사용자 생산 정보는 신뢰성이 보장되지 않았다는 한계가 있다. 리뷰 작성에 있어 사용자에 대한 별도의 인증이 없어 텍스트 데이터에 대해서도 완전히 신뢰할 수 없었으나, 광고 글이나 관련 없는 게시물의 빈도가 낮고 작성 예시를 세부적으로 제공하여 리뷰 텍스트의 질이 높은 플랫폼 서비스인 직방을 선택하여 이러한 한계를 극복하고자 하였다. 향후 연구에서는 더욱 다각적인 특성 변수에 대하여 더욱 검증된 데이터를 바탕으로 분석을 진행할 필요가 있다.

ORCID

이해인 https://orcid.org/0000-0002-6866-2065
황현준 https://orcid.org/0000-0002-2192-0290

참고문헌

1. 경정익 · 권대중, 2022,「프롭테크 동향과 진화방향에 대한 소고」,『부동산융복합연구』, 2(2):5-23.
2. 구본상 · 신병진, 2015,「능형회귀분석을 활용한 부동산 헤도닉 가격모형의 정확성 및 해석력 향상에 관한 연구: 서울시 구로구 아파트를 대상으로」,『한국건설관리학회 논문집』, 16(5):77-85.
3. 김다니 · 김란, 2021,「부동산 정책에 대한 시민 인식 구조와 그 시기별 변동에 관한 트위터 분석: 토픽 모델링 기법을 활용하여」,『융합사회와 공공정책』, 15(1):33-63.
4. 김승현 · 김원혁 · 이윤수, 2022,「머신러닝과 패널 고정효과를 활용한 아파트 실거래가 예측」,『주택연구』, 30(1):43-69.
5. 김이환 · 김형준 · 류두진 · 조훈, 2022,「기계학습 방법론을 활용한 아파트 매매가격지수 연구」,『부동산분석』, 8(3):1-29.

6. 김인호 · 이경섭, 2020,「트리 기반 앙상블 방법을 활용한 자동 평가 모형 개발 및 평가: 서울특별시 주거용 아파트를 사례로」,『한국데이터정보과학회지』, 31:375-389.
7. 김태훈, 2004,「재건축 특성에 따른 아파트 가격 변화에 관한 연구」,『부동산연구』, 14(2):179-200.
8. 문태헌 · 김혜림, 2020,「신문기사 감성분석과 아파트 매매실거래가의 관계 분석: 부산지역 사례」,『주거환경』, 18(3):135-149.
9. 박재수 · 이재수, 2021,「기계학습 기술을 이용한 부동산 감성지수 개발 모형 연구」,『부동산학연구』, 27(2): 47-62.
10. 박종기 · 이상경 · 강승일, 2011,「오피스 가격 결정 요인에 관한 연구: 거래특성과 공간자기상관을 중심으로」,『부동산연구』, 21(3):91-108.
11. 박종영 · 서충원, 2015,「TF-IDF 가중치 모델을 이용한 주택시장의 변화특성 분석」,『부동산학보』, 63:46-58.
12. 배성완 · 유정석, 2018,「기계 학습을 이용한 공동주택 가격 추정: 서울 강남구를 사례로」,『부동산학연구』, 24(1):69-85.
13. 양건필 · 전해정, 2022,「기계학습과 XAI를 활용한 아파트 가격과 지역특성과의 관계 분석」,『부동산연구』, 32(3):7-24.
14. 윤성진 · 이관용 · 홍정기, 2021.「토픽 모델링을 활용한 부동산 대책 및 사회적 이슈 분석」,『부동산정책연구』, 22(1):15-37.
15. 이소영 · 김명연 · 김은정, 2020,「아파트 가격에 영향을 미치는 근린환경 요인 분석: 서울시 강남구와 대구시 수성구를 대상으로」,『부동산분석』, 6(1): 19-37.
16. 이용만, 2008,「헤도닉 가격 모형에 대한 소고」,『부동산학연구』, 14(1):81-87.
17. 이주미 · 박성훈 · 조상호 · 김주형, 2021,「머신러닝을 이용한 부동산 지수 예측 모델 비교」,『대한건축학회논문집』, 37(1):191-199.
18. 이정윤 · 오경주 · 안재준, 2021,「국내 프롭테크 기업의 발전방향에 대한 연구: 부동산 플랫폼 정보 제공 기능을 중심으로」,『지식경영연구』, 22(2): 55-76.
19. 이종민 · 이종아 · 정준호, 2017,「뉴스 빅데이터를 이용한 전세 가격 예측: 토픽모형 분석을 중심으로」,『부동산학보』, 69:43-57.
20. 임재현, 1998,「주택특성가격이론의 발전 모색」,『한국행정학보』, 32(1):247-261.
21. 장몽현 · 김한수, 2019,「텍스트 마이닝을 활용한 주택가격 변동에 관한 연구」,『한국주거학회논문집』, 30(2):35-42.
22. _____, 2020,「공간계량모형을 활용한 아파트가격 영향요인 분석 연구: 대구시 수성구를 중심으로」,『한국주거학회논문집』, 31(1):79-86.
23. 정수연 · 이성원 · 박홍희, 2009,「위계선형모형을 이용한 서울시 아파트 층별 가격 분석」,『감정평가학논집』, 8(2):43-52.
24. 조민서 · 정삼화 · 김태훈, 2011,「특성가격모형의 분석결과를 종합한 주택가격 결정요인에 관한 연구」,『주택연구』, 19(4):49-78.
25. 한국프롭테크포럼, 2022,「한국프롭테크포럼 회원사 편람」, 서울: 한국프롭테크포럼.
26. _____, 2023, "Korea Proptech Forum Member's Map," Accessed February 5, 2023, http://proptech.or.kr/map.
27. 허세림 · 곽승준, 1994,「헤도닉가격기법을 이용한 주택특성의 잠재가격 추정」,『주택연구』, 2(2): 27-42.
28. 홍정의, 2021,「랜덤 포레스트 알고리즘을 통한 주택 대량평가모형 연구」,『부동산분석』, 7(1): 1-28.
29. Blei, D. M., A. Y. Ng, and M. I. Jordan, 2003, "Latent dirichlet allocation," *Journal of*

machine Learning Research, 3:993-1022.

30. Breiman, L., 2001, "Random forests," *Machine Learning*, 45(1):5-32.
31. Brown, J. N. and H. S. Rosen, 1982, "On the estimation of structural hedonic price models," *Econometrica*, 50(3):765-768.
32. Chau, K. W. and T. L. Chin, 2003, "A critical review of literature on the hedonic price model," *International Journal for Housing Science and Its Applications*, 27(2):145-165.
33. Hong, J., H. Choi, and W. S. Kim, 2020, "A house price valuation based on the random forest approach: The mass appraisal of residential property in South Korea," *International Journal of Strategic Property Management*, 24(3):140-152.
34. Ibbotson, R. G. and L. B. Siegel, 1984, "Real estate returns: A comparison with other investments," *Real Estate Economics*, 12(3): 219-242.
35. Lundberg, S. M. and S. I. Lee, 2017, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, 30:4765-4774.
36. Röder, M., A. Both, and A. Hinneburg, 2015, "Exploring the space of topic coherence measures," In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Shanghai, China, 399-408.

논문접수일: 2023년 2월 21일
심사(수정)일: 2023년 4월 7일
게재확정일: 2023년 4월 13일

국문초록

본 연구는 프롭테크 플랫폼으로부터 사용자 중심의 비정형 데이터를 추출한 뒤, 토픽 모델링 기법을 이용하여 헤도닉 가격 모형의 소비자 중심 특성변수로 구성하고, 아파트 매매 가격에 미치는 영향을 분석하고자 한다. 분석모형으로는 선형 회귀방법과 랜덤 포레스트를 함께 사용한다. 아파트 공급 면적당 거래가격에 대한 헤도닉 가격 모형의 회귀분석 결과에 따르면 생활권, 교육과 관련된 토픽의 추정 계수는 각각 11.6%, 11.5%로 1% 유의수준 하에서 통계적으로 유의한 것으로 추정된다. 회귀분석 대신 랜덤 포레스트를 이용할 경우 약 0.71의 결정계수는 약 0.93으로 개선된 예측을 보인 것으로 확인된다. 본 연구는 프롭테크 플랫폼 서비스의 정보와 머신러닝 기술을 적용하여 부동산 시장의 아파트 가격 분석에서 수요 측면의 정보가 포함되어야 함을 통계적으로 밝히고, 그 방법을 구체적으로 제안했다는 기여가 있다. 우리의 연구 결과는 정부가 부동산 관련 정책을 추진함에 있어서 프롭테크 서비스 등을 포함하여 소비자가 생성하는 정보를 고려해야 한다는 정책적 시사점을 제시한다.

주제어 : 프롭테크, 헤도닉가격모형, 잠재디리클레할당, 선형회귀모형, 랜덤포레스트