주택금융연구 (2020) 4권, 33-64
ISSN: 2586-4629

Housing Finance Research
https://www.hf.go.kr

기계학습 알고리즘을 이용한 주택가격감정 시스템의 구축 및 평가: XGBoost, LightGBM, CatBoost 알고리즘에 기반하여

홍정의*

요 약

주택 가격 감정 모형은 아주 작은 비용으로 대량의 부동산 감정을 동시에 수행할 수 있기 때문에, 모기지 담보가치 추정, 주택가격지수 산출, 재산세 추정 등과 같이 대규모 자산 가치평가를 빈번하게 수행해야 하는 모든 활동 영역에서 다양하게 활용될 수 있다. 최근에는 급격히 성장하는 데이터 수집·분석 기법들을 주택 감정 모형의 정확성을 상승시키는 데에 활용하는 연구들이 늘어가고 있다. 본문의 목적은 효율성과 예측력이 높은 것으로 알려진 세 가지 알고리즘(XGBoost, LightGBM, CatBoost)을 통해 주택 감정 모형을 구축하고, 그 성능과 특징 및 활용방법을 분석하는 것이다. 본문은 2009년부터 2019년까지 서울에서 거래된 아파트 매매 데이터 620,617건을 통해 헤도닉 모형과 기계 학습 모형 기반의 주택 가치 감정 모형의 예측력을 비교하였다. 분석 결과는 다음과 같다; 첫째, 기계 학습 모형의 예측력은 상대적인 측면 (헤도닉 가격 모형에 비해) 뿐 아니라, 절대적인 측면 (모형의 실용적 활용 가능성)에서도 상당히 높게 나타났다. 헤도닉 모형의 경우, 시장 가격에 대한 예측의 평균 백분율 오차가 약 11.5% 내외인 반면, XGBoost · LightGBM · CatBoost는 각각 3.7%, 3.8%, 3.6%에 불과했다. 두 번째로, CatBoost 알고리즘이 평균 예측력에서나, 이상치 발생 빈도에서나 다른 두 알고리즘에 비해 더 우수한 것으로 나타났다. 세 번째로, 소프트 보팅을 통한 세 알고리즘의 앙상블 모형을 구축하는 경우, 개별 알고리즘보다 더 예측력을 상승시킬 수 있음을 확인하였다.

핵심어 : 기계학습, XGBoost, LightGBM, CatBoost, 대량 주택 감정

* 홍정의, 주저자, 한동대학교 경영경제학부 교수, 경제학 박사, hwgh024@handong.edu

Ⅰ. 서론

이른바 4차 산업혁명 시기에 접어들면서, 기계학습, 인공 신경망 기법 등을 통한 각 분야의 혁신이 가시화되고 있다. 이러한 데이터 과학 기술을 주택 시장 연구 및 관리에 접목한 프롭테크(proptech)는 당분간 흐름을 이어갈 것으로 예상된다. 프롭테크는 다양하게 적용될 수 있지만, 특히 빅 데이터와 그에 대한 다양한 기계 학습 기법을 바탕으로 전통적 모형에 비해 예측력이 뛰어난 주택 가치 감정 모형을 만드는 데에도 활용될 수 있다는 점에서 흥미롭다. 주택 가치 감정 모형 또는 대량 감정 모형이란 일정한 수리·통계 모형 또는 컴퓨터 공학적 기법을 적용하여 부동산의 해당 시점 시장 가격을 예측하는 시스템 (Standard of Mass Appraisal of Real Property)을 의미한다.[1] 전통적으로 주택의 감정평가는 감정평가사를 통한 추정을 의미하는데, 이는 직접적이고 정교한 반면, 상대적으로 비용이 많이 드는 방식이다. 한편, 대량 감정 모형의 경우 매우 작은 비용으로 대량의 부동산 감정을 동시에 실시할 수 있기 때문에, 모기지 담보가치 추정, 건설 투자 입지 분석, 주택가격지수 산출, 재산세 산정 등과 같이 대규모의 가치평가를 빈번하게 수행해야 하는 모든 활동 영역에 다양하게 활용될 수 있다(Zhou, Ji, et al., 2018, Wang and Li, 2020).

최근의 데이터 분석 기술이 특히 주택 가격 감정 모형 분야에서 두각을 드러낼 수 있는 이유는 크게 두 가지가 있다. 첫째는, 근래의 데이터 수집 및 가공 기술의 발전으로 인해 종래에는 관찰하지 못했던 변수들을 관찰할 수 있게 되었으며, 수집할 수 있는 표본의 수 또한 대폭 늘어났다는 것이다. 특히 GIS를 통한 입지 정보를 이용할 수 있게 되면서부터 이로 인한 주택 가격의 차이를 효과적으로 포착할 수 있는 다양한 모형들을 활용할 수 있게 되었다. 표본 수의 확보 문제는 모형의 복잡성이 높을수록 중요해지는데, 이는 복잡한 모형일수록 과적합 문제에 취약해지기 때문이다. 과적합이란 특정 표본 집단의 특성을 지나치게 적응하는 상황으로, 결과적으로 분석 대상 전반에 대한 모형의 예측력을 훼손시킨다. 이른바 빅데이터 기술로 불리는 데이터 수집 가공 능력의 상승은 이처럼 보다 복잡하고 비선형성이 높은 모형을 활용할 수 있는 길을 열어주고 있다. 두 번째로, 데이터 분석 모형 기법 자체의 비약적인 발전이다. 이는 앞서 언급한 데이터 수집 가공 기술의 발전과 맞물려 최근 폭발적인 추세로 성장하고 있다. 이러한 모형들은 전통적 계량모형에 비해 훨씬 높은 복잡성과 비선형성을 다룰 수 있으며, 그에 따라 특히 변수 예측 문제에서 높은 성능을 가지게 되었다. 주택가격 감정 모형에서 가장 중요한 것 역시 예측의 정확성과 효율성, 즉, 가능한 한 적은 종류의 변수를 통해 가능한 한 정확하고 안정적인 예측을 수행하는 것이다. 그러므로 최근의 데이터 분석 기법들은 주택 감정 시스템에서도 일대 혁신을 불러올 수 있을 것으로 기대된다.

---

1) IAAO. Standard on Mass Appraisal of Real Prop.; IAAO: Kansas City, MO, USA, 2017.

근래 다양한 데이터 분석 기법이 소개되면서, 이를 주택 가치 감정 모형에 적용한 연구들도 늘어가는 추세이다. 대표적으로, 의사결정나무 (Fan et al., 2006), 랜덤포레스트 (Antipov and Pokryshevskaya, 2012; Hong et al., 2020), 서포트 벡터 머신 (Gu, Zhu, and Jiang, 2011), 인공신경망 (McCluskey and Anand, 1999; Selim, 2009), 그래디언트 부스트 (Park and Bae, 2015) 등이 주택 감정 모형에 사용될 수 있음이 알려져 있다.

본문의 목적은 최근 그래디언트 부스트를 기반으로 효율성과 예측력을 상승시키는 것으로 알려진 세가지 알고리즘(XGBoost, LightGBM, CatBoost)을 통해 대량 주택 감정 모형을 구축하고, 그 성능과 특징 및 활용방법을 분석하는 것이다. 이러한 알고리즘들은 수행시간이 느리고, 과적합 규제가 없는 그래디언트 부스트의 한계를 극복한 것으로 평가되며, 특히 분류 문제에 있어서 다른 기계 학습 기법에 비해 예측 성능이 높은 것으로 알려져 있다. 이 세 알고리즘은 비교적 최근에 소개되었기 때문에, 랜덤 포레스트나 인공 신경망 등과 같이 더 일찍 알려진 알고리즘들에 비해 주택 가격 감정 모형 분야에서 얼마나 효율적으로 사용될 수 있는지 상대적으로 덜 알려져 있다.

또한, 본문은 이 세 알고리즘의 예측력이 전통적인 방식(특히 헤도닉 모형)에 비해 얼마나 상승하는지를 비교 분석하였다. 전통적인 선형 헤도닉 모형은 몇 가지 가정 하에서 자산의 가격을 설명변수의 회귀식으로 표현하는데(Rosen, 1974), 이러한 방식의 장점은 간단한 선형 다변량 회귀분석을 통해서 감정모형 시스템을 구축할 수 있다는 데에 있다. 즉, 변수 간 관계를 직관적으로 해석하거나 대중에게 설명하기 쉬운 특징을 가지고 있다는 것이다. 그러나, 이와 같은 단순성과 직관성이 동시에 약점이 되기도 한다. 완전 경쟁, 분리 가능한 효용구조, 완전히 통합된 시장구조 등의 엄격한 경제학적 가정이 동시에 만족되지 않는 경우, 회귀식으로 표현 가능한 단순한 함수형태는 실제의 부동산 시장이 가지고 있는 복잡성을 포착하기에 충분하지 않기 때문이다(Sheppard, 1999; Malpezzi, 2002). 결과적으로 헤도닉 모형에 기반한 부동산 감정은 함수 형태의 단순성이나 관찰 가능한 변수의 제한으로 인한 설명력 저하 문제를 갖게 된다. 한편, 기계학습 기반의 모형은 기본적으로 변수 간 관계를 하나의 함수 형태로 묘사하는 것이 아니기 때문에, 실제 시장의 비선형성을 포착하는 데에 있어 헤도닉 모형보다 유리한 특징을 가지고 있다. 무엇보다도, 헤도닉 모형에서는 입지 효과를 포착하기 위해서 주요 시설과의 거리 또는 이동 시간을 정교하게 추정해야 하는 데에 반해, 위 알고리즘들은 주택의 입지 정보를 직접 이용하여 주택 가치 간의 비선형적 차이 자체를 학습하는 데에 사용할 수 있다. 입지 효과는 주택 가치 간 차이를 설명하는 데에 있어 가장 중요한 요소일 뿐 아니라 그 복잡성도 높기 때문에, 이를 효과적으로 포착하는 만큼 전통적 방식에 비해 더 높은 예측력을 갖는 것으로 나타나게 될 것이다.

본문은 2009년부터 2019년까지 서울에서 거래된 아파트 매매 데이터 620,617건을 통해 헤도닉

모형과 세 알고리즘 (XGBoost · LightGBM · CatBoost) 기반의 주택 가치 감정 모형을 비교 분석하였다. 결과를 간략히 요약하면 다음과 같다; 첫째, 헤도닉 모형에 비해 기계 학습 알고리즘의 예측력이 월등히 높게 나타났다. 헤도닉 모형의 경우, 시장 가격에 대한 예측의 평균 백분율 오차가 약 11.5% 내외인 반면, XGBoost · LightGBM · CatBoost는 각각 3.7%, 3.8%, 3.6%에 불과했다. 이는 기계학습 기반의 예측력이 다른 모형과의 상대적인 비교뿐 아니라, 절대적 수치 측면에서도 유의미하게 낮아질 수 있다는 것을 의미한다. 즉, 절대적인 예측력이 다양한 실용적 목적에 사용되기에 적합할 정도로 정교화될 수 있음을 보여준다. 두 번째로, 약간이나마 CatBoost 알고리즘의 예측력이 다른 두 알고리즘에 비해 더 높게 나타났다. 평균적 예측력 (평균 백분율 오차나 R-squared)에서 뿐 아니라, 이상치의 발생빈도 등 모형의 안정성 측면에서도 CatBoost가 더 우수한 것을 확인할 수 있다. 세 번째로, 소프트 보팅을 통한 세 알고리즘의 앙상블 모형을 구축하는 경우, 개별 알고리즘보다 더 예측력을 상승시킬 수 있음을 확인하였다. 보팅은 서로 다른 알고리즘을 통해 얻어진 예측을 결합하는 방식으로, 본 연구에서는 평균을 취하는 방식을 택했다. 이는 개별 알고리즘을 통한 예측 과정에서 발생한 고유의 소음이 다른 알고리즘의 예측과 결합되는 과정에서 약화될 수 있음을 의미하는 것이다.

본 연구의 구성은 다음과 같다. II장에서는 주택감정모형 분야에서 최근 기계학습 기법이 적용된 사례를 소개하였다. III장은 본 연구에서 사용한 세 알고리즘의 원리와 특징을 개략적으로 서술하였다. IV장은 분석자료와 기초통계를 설명하고, 그것을 통해 구축한 주택감정모형의 예측력을 비교 · 분석하였다. V장에서는 앞의 논의를 정리하며, 추후의 연구과제 등에 관해 서술하였다.

II. 이론 및 선행연구 검토

주택가격 추정 시 사용하는 가장 전통적이고 잘 알려진 모형은 헤도닉 가격 모형이다. Lancaster (1966)와 Rosen (1974)의 모형에서 시작된 헤도닉 모형은 주택의 시장 가치를 주택의 속성에 대한 회귀분석으로 묘사할 수 있는 이론적 근거를 제공한다. Lancaster의 소비자 이론에서 소비자는 상품 자체가 아니라, 속성에서 효용을 얻는 것으로 묘사된다. 그러므로 어떤 재화의 소비는 재화에 담긴 속성의 복합을 소비하는 것으로 표현될 수 있다. Rosen (1974)은 이러한 소비자 이론을 헤도닉 가격 모델로 확장했는데, 그는 재화의 가치를 Lancaster (1966)에서 묘사된 바처럼 재화에 포함된 각 속성의 가치로 환원할 수 있다고 제시했다. 각 속성이 완전 경쟁 시장에서 각각의 암묵적 가격을 갖는다는 가정하에 재화의 시장 가치는 그 재화에 포함된 속성 가치의 합으로 해석될 수 있으며, 결과적으로 이는 재화의 시장 가치가 그 재화의 속성에 대해 회귀될 수 있음을 의미한다.

그러나, 실제 회귀분석 과정에서 가정되어야 하는 함수의 형태이나 변수 목록들은 헤도닉 가격 이론 자체에 의해서는 특정될 수 없기 때문에(Malpezzi, 2002), 회귀분석의 적용 과정에서 단순화되거나 포착되지 않은 변수 또는 다중공선성 등에 의해 모형의 설명력이 왜곡될 가능성이 항상 존재한다. 특히, 헤도닉 가격 모형은 일반적으로 변수 간의 선형 관계와 독립성을 가정하는데, 다양한 이유로 실제 시장에서는 이러한 가정이 위배될 수 있음을 고려해야 한다.

이러한 헤도닉 모형의 한계를 극복하려는 방향은 다양하게 이루어져 왔는데, 크게 분류하면 다음의 두 가지를 생각할 수 있다. 첫 번째는 전통적 선형 헤도닉 모형의 가정을 완화하거나 입지효과를 통제하는 방식으로 모형 해석성은 유지하면서도 설명력을 높이는 것이다. 예를 들어, 변수관계를 Box-Cox 함수로 표현하거나(Goodman, 1978; Halvorsen and Pollakowski, 1981; Rasmussen and Zuehlke, 1990 등), k-근접 이웃 모형(Meese and Wallace, 1991; McCluskey and Anand, 1999 등)을 사용하는 방식은 단순히 선형 함수로 표현된 헤도닉 모형에 비해 변수 관계에 내재된 비선형성을 더 효과적으로 포착할 수 있는 것으로 알려져 있다. 또한, 측정 가능한 입지 변수들만으로 포착될 수 없는 입지 효과를 모형에 반영하기 위한 공간 상관 모형(대표적으로, Dubin, 1992; Anselin and Bera, 1998; Basu and Thibodeau, 1998; Anselin, 2001; Conway et al., 2010 등) 등도 이에 해당할 수 있다.

두 번째는 실제 시장의 복잡성과 비선형성을 보다 유연하게 포착할 수 있는 새로운 모형을 사용하는 것이다. 특히 최근 데이터 분석 기법, 특히 기계학습 기법의 폭발적인 성장과 함께 다양한 기계학습 모형이 주택 가격 감정 모형 분야에 적용되기 시작했는데, 이런 기법은 헤도닉 모형에 비해 직관성과 해석성이 떨어지지만, 예측력 자체는 유의미하게 상승하는 것으로 알려져 있다. 최근 데이터 과학 분야의 폭발적 성장과 더불어 다양한 기계학습 기법이 소개되었는데, 그 중 주택가격감정에 사용되는 모형은 분류 문제뿐 아니라, 회귀 문제에 있어서도 안정적으로 높은 예측력을 갖는 모형들이다. 대표적인 알고리즘들을 나열하면, 의사결정트리 (Fan, et al., 2006), 랜덤 포레스트 (Antipov and Pokryshevskaya, 2012; Hong, et al., 2020), 서포트 벡터 머신 (Gu, et al., 2011; Zurada, et al., 2011; Mu, et al., 2014) 그래디언트 부스트 (Park, B., and Bae, J. K., 2015), 인공신경망 (McCluskey and Anand, 1999; Limsombunchai, 2004; Selim, 2009) 등이 있다.

이러한 연구들은 주로 기계학습 등을 이용한 모형이 주택 감정 시스템으로 활용되는 경우, 예측력이 얼마나 상승하는지를 보여준다. 대부분의 연구에서 기계학습 기반의 모형은 전통적 모형에 비해 월등한 예측력을 보이는 것으로 조사된다. 예를 들어, Hong, et al.(2020)은 의사결정나무의 앙상블 기법인 랜덤 포레스트 알고리즘을 통해 우리나라 서울 강남구 아파트 가격에 대한 예측 모형을 추정하였는데, 이들은 기계학습 기반 모형의 예측력이 헤도닉 모형에 비해 상대적으로 뛰어날 뿐만

아니라, 평균 백분율 오차가 5~6% 정도에 불과할 정도로 정확성이 상승할 수 있음을 보여주었다.

최근 그래디언트 부스트를 기반으로 그 효율성과 예측력을 증가시키는 알고리즘들이 소개되었는데, 본 연구에서 주목하고 있는 XGBoost, LightGBM, CatBoost가 그것이다. 이러한 알고리즘들은 수행시간이 느리고 과적합 규제가 없는 그래디언트 부스트의 한계를 극복한 것으로 평가되며, 특히 분류 문제에 있어서 다른 기계 학습 기법에 비해 예측 성능이 높은 것으로 알려져 있다. 이러한 기계 학습 기법은 비교적 최근에 소개되었으며, 주택 가격 감정 문제에 이 알고리즘들이 얼마나 효율적으로 사용될 수 있는지 기존의 알고리즘 (랜덤 포레스트나 인공 신경망 등)에 비해 상대적으로 아직 덜 알려진 것으로 보인다.

Ⅲ. 분석 기법의 소개

1. 그래디언트 부스트 알고리즘

본문에서 사용된 XGBoost, LIghtGBM, CatBoost는 모두 그래디언트 부스트 방식에 기반하고 있으므로, 이 장에서는 먼저 이에 대해 간략히 서술한다. 그래디언트 부스트는 대표적인 기계학습 알고리즘 중 하나로, 주로 분류 문제 또는 값 예측 문제 (회귀)에 사용된다. 기계학습에서 부스팅(boosting)이란 단순하고 약한 학습자(weak learner)를 오차를 최소화하는 방향으로 결합시켜 보다 정확한 학습자(strong learner)를 만드는 앙상블 형식의 알고리즘을 의미한다. 개별 예측기로는 설명력이 낮지만 가벼운 예측기를 먼저 만들고, 이를 통해 나타난 오류는 그 다음 예측기가 보완한다. 이러한 방식으로 각 예측기의 예측력을 합치면 기존보다는 정확한 모델이 만들어지고, 다시 남아 있는 예측 오류를 다음 예측기를 통해 보완하여 기존 모형에 더하는 과정을 반복하는 방식이다.

그래디언트 부스트 방식은 손실함수(loss function)을 정의하여, 경사하강법을 통해 이를 최소화하는 상태를 찾는 방식을 택한다. 이때 손실함수는 일반적으로 실제 값과 예측 값 간 차이의 제곱의 합으로 정의되나, 그래디언트 부스트로 강점은 미분 가능한 함수라면 다른 형식의 손실함수를 신축적으로 사용할 수 있다는 것이다. 경사하강법은 이러한 손실함수의 크기를 줄이는 방향으로 모형을 지속적으로 변화시켜 나가는 방식을 의미한다. 손실함수의 값을 고도가 있는 언덕에 비유한다면, 기계학습의 목표는 가능한 한 저지대로 이동하는 것일 것이다. 이때 어느 방향으로 얼마만큼 이동해야 아래로 내려가는 지는 계수 변화 또는 모형 추가에 따른 손실함수 값의 변화 방향 및 크기를 추적하면 될 것이다. 만약 손실 함수가 예측 오류의 제곱합인 경우, 손실 함수의 경사값 (미분값)은 실제에 대한 예측의 잔차항이 된다. 이런 경우, 알고리즘은 이 잔차를 목표값으로 하는

예측자를 다시 학습하여 기존 모형에 더하는 방식으로 예측력을 높이게 된다. 그래디언트 부스트는 이처럼 예측오차를 감소시키는 방향을 따라 계수를 수정하거나 모형을 누적시키는 과정을 더이상 손실함수의 값을 하락시킬 수 없을 때까지 (즉, 가장 높은 예측력에 달할 때까지) 반복하는 것이다.

이를 기술적으로 표현하면 〈표 1〉과 같다.

2. XGBoost 알고리즘

XGBoost는 Extreme Gradient Boost의 약자로, 기본적으로 그래디언트 부스트의 방식을

〈표 1〉 그래디언트 부스트 알고리즘

투입요소 : 학습 표본 $(x_i, y_i)_{i=1}^{n}$, 미분 가능한 손실함수 $L(y, F(x))$, 최대반복횟수 M

알고리즘 :

1. 다음을 만족시키는 상수 γ를 찾아 $F(x)$의 초기값으로 부여한다.

$$F_0(x) = \arg\min \sum_{i=1}^{n} L(y_i, \gamma).$$

2. m에 대하여 1부터 M까지 아래를 반복한다.

2-1. 다음과 같은 유사잔차(pseudo-residuals)를 계산한다.

$$r_{i,m} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, ..., n.$$

2-2. 약한 학습자 $h_m(x)$를 위의 유사잔차에 대하여 학습한다.

2-3. 아래의 극대화 문제의 해를 통해 승수 γ_m로 계산한다.

$$\gamma_m = \arg\min \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)).$$

2-4. 모형 $F(x)$를 아래와 같이 업데이트한다.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. $F_M(x)$를 최종 반환한다.

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 기계 학습자에 비해 예측 성능이 뛰어나며, 병렬 CPU를 통한 학습이 가능해 그래디언트 부스트에 비해 빠른 수행시간을 갖는 것으로도 평가받는다.

3. LightGBM 알고리즘

그래디언트 부스트 계열의 알고리즘은 반복 누적하는 약한 학습자로 의사결정트리를 사용한다. 이때 LightGBM은 일반적인 트리 분할 방법과 다르게 리프 중심 트리 분할 방식을 사용한다. 보통의 의사결정트리 기반 알고리즘은 개별 트리가 과적합되지 않도록 균형 트리 분할 방식을 사용하여 트리의 길이를 줄인다. 즉, 최대한 트리의 균형을 맞추면서 개별 트리의 깊이를 최소화하는 것이다. 하지만 LightGBM은 리프 중심 트리 분할 방식을 사용하는데, 이는 개별 의사결정트리가 비대칭적이 되더라도 (즉, 과적합 가능성이 있더라도) 최대한의 효율적 학습을 수행하도록 유도하는 것이다. LightGBM 알고리즘은 이러한 약한 학습자를 쌓아갈수록 결국은 최종적으로 균형 트리 분할 방식에 비해 예측 오류를 줄어들 수 있다는 것에 착안하고 있다. 이러한 방식의 대표적인 강점은 예측 성능을 유지하거나, 오히려 강화시키면서도 수행시간이 더 빠르다는 것이다. 일반적으로 XGBoost가 그래디언트 부스트에 비해 빠르다고 알려져 있지만, LightGBM은 한층 더 가벼운 연산을 수행할 수 있다. LightGBM은 XGBoost와 마찬가지로 병렬 연산 기능을 제공함으로써 큰 데이터에 대한 효율적인 연산을 가능하게 한다.

4. CatBoost 알고리즘

그래디언트 부스트를 비롯한 여타 많은 기계 학습 알고리즘(랜덤 포레스트 등)에서는 범주형

변수를 처리하기 위해 정수화하거나 더미변수화한다. 이러한 방식은 약점을 가지고 있는데, 먼저 범주형 변수를 정수화하는 경우 실제로는 카디널리티(cardinality)가 없는 변수를 카디널리티가 있는 변수로 변환하는 것으로 의도치 않은 정보의 왜곡을 낳을 수 있다. 더미변수화하는 방식은 범주형 변수의 범주 개수 만큼의 변수를 추가하는 것이 되므로, 범주의 개수가 많으면 많을수록 모형의 복잡성을 높이고, 연산 효율성을 훼손시키게 된다. 특히 주택 감정 모형에서와 같이 행정동 또는 각 시점 등과 같이 다수의 범주값을 갖는 범주형 변수를 포함하는 연산 과정에서는 알고리즘의 복잡성이 너무 높아져 더미변수화 된 변수들(즉, 범주값의 구분자들) 중 상당수가 연산에 반영되지 못할 수 있다.

CatBoost 알고리즘은 정보 획득(information gain)량이 동일한 여러 속성을 하나의 속성으로 묶어 버림으로써, 결과적으로 범주형 변수들을 효과적으로 알고리즘 연산에서 반영할 수 있게 설계되었다. 즉, 범주형 변수의 성질(즉, 다른 변수값 간에 대소비교는 불가능하지만, 하나의 변수로 묶이는 성질)을 유지하면서 모형을 학습할 수 있다는 것을 의미한다. 결과적으로 CatBoost는 변수 숫자의 확장으로 인한 정보의 탈락을 최소화할 수 있으며, 그로 인한 예측력과 연산속도의 상승을 기대할 수 있다.

Ⅳ. 분석 데이터

본 연구의 분석 대상은 2009년 7월부터 2019년 12월까지 서울시에서 일어난 아파트 거래로, 총 표본 개수는 620,617개이다. 아파트는 우리나라의 대표적인 주거 방식으로 많은 표본을 수집할 수 있을 뿐 아니라, 단독주택 등에 비해 비관측 요소가 상대적으로 적고 데이터로 집계되는 주택 속성이 균일하기 때문에, 관찰 가능한 주택 속성을 통해 가격을 예측해야 하는 본문의 분석의 특성상 더 적합한 분석 대상일 수 있다(Hong et al., 2020). 본 연구에서 사용한 표본 크기가 일반적인 헤도닉 모형 분석의 사례에 비해 상당히 크기 때문에, 복잡성이 높은 기계학습 알고리즘을 학습하기에도 충분할 뿐더러 연구 결과의 일반성을 주장하는 데에도 강점이 있을 것으로 보인다.

아래 〈표 2〉는 수집된 자료의 정의를 소개하고 있으며 〈표 3〉은 그에 대한 기초통계를 제공한다. 수집 자료에 대한 개략은 다음과 같다. 먼저, 아파트 매매가격은 국토교통부에서 공개하는 아파트 매매 실거래가를 활용하였다. 주택 속성은 크게 구조적 속성과 입지적 속성이 있는데, 본 연구에서 포함하고 있는 구조적 속성으로는 아파트의 크기(전용 면적), 방 개수, 화장실 개수, 층 수, 난방 방식(개별난방, 중앙난방, 지역난방), 복도 형태(복도식, 계단식, 혼합식), 경과 년 수, 아파트 단지 내 세대수, 동수, 가구 당 평균 주차 대수, 단지 내 최고층의 높이, 최저층의 높이, 건폐율, 용적률이 있다. 입지적 속성으로는 가장 가까운 지하철역로부터의 거리, 주요 편의시설(공원, 박물관)

〈표 2〉 분석 자료의 정의와 단위

분류	변수	설명	단위
종속변수	가격	거래시점의 매매가격	원
구조적 속성	크기	전용면적	m^2
	방 수	방의 개수	개
	화장실 수	화장실의 개수	개
	층	위치한 층의 높이	층
	난방	개별/중앙/지역난방의 여부	범주형
	경과년수	준공년도와 거래년도의 차이	년
	세대수	아파트 단지에 존재하는 세대의 수	세대수
	동수	아파트 단지에 존재하는 건물(동)의 수	동수
	용적률	(건물 연면적 / 대지 면적)×100	%
	건폐율	(건물 면적 / 대지 면적)×100	%
	주차대수	아파트 단지 주차가능 대수 / 세대 수	대수
	최고층	아파트 단지 내 건물 중 최고층의 높이	층
	최저층	아파트 단지 내 건물 중 최저층의 높이	층
입지적 속성	행정구역	행정동의 구분	범주형/더미 변수
	위치	위도, 경도	좌표값
	지하철로부터의 거리	가장 가까운 지하철역로부터의 거리	미터
	공원으로부터의 거리	가장 가까운 공원으로부터의 거리	미터
	초등학교로부터의 거리	가장 가까운 초등학교로부터의 거리	미터
	중학교로부터의 거리	가장 가까운 중학교로부터의 거리	미터
	고등학교로부터의 거리	가장 가까운 고등학교로부터의 거리	미터
	대학교로부터의 거리	가장 가까운 대학교로부터의 거리	미터
	박물관으로부터의 거리	가장 가까운 박물관으로부터의 거리	미터
	주민센터로부터의 거리	가장 가까운 주민센터로부터의 거리	미터

으로부터의 거리, 교육기관(초등, 중등, 고등, 대)으로부터의 거리, 행정시설(주민센터로부터의 거리)이 있다. 이때 거리는 GIS 자료를 통해 유클리드 거리(m)를 계산하였다. 또한, 입지적 속성값

〈표 3〉 기초통계

변수	평균	표준편차	최소값	최대값
가격	54,043.88	36,949.47	700	700,000
크기	79.49762	28.52365	11.95	325.39
방개수	2.969445	0.667722	1	8
화장실수	1.659708	0.492461	1	5
세대수	1,032.104	1,153.402	5	9,510
동수	11.88267	14.04946	1	122
주차대수	1.142896	0.479476	0.02	11.95
용적률	287.9124	126.9176	2	1,477
건폐율	24.90165	27.63889	2	2,457
최고층수	19.18334	6.811636	4	69
최저층수	12.16432	5.424588	1	54
층수	9.445466	6.190082	-4	68
경과년수	13.78384	7.813038	0	49
지하철로부터의 거리	770.5905	618.0864	2.645571	5,583.588
공원으로부터의 거리	1,036.23	526.1945	55.73893	3,268.238
초등학교로부터의 거리	337.1795	169.516	10.59662	1,810.038
중학교로부터의 거리	471.2093	252.5819	2.587437	2,130.155
고등학교로부터의 거리	579.8585	333.7149	24.61597	2,837.36
대학교로부터의 거리	1,867.998	1,190.588	50.28124	7,111.538
박물관으로부터의 거리	1,829.797	1,068.645	35.34001	6,839.077
주민센터로부터의 거리	1,938.611	992.8313	16.82066	6,521.695

자체는 아니지만 입지적 차이를 구분할 수 있는 변수로써 행정구역(동)과 위치값(위도, 경도)도 수집되었다.

Ⅴ. 정량분석결과

1. 예측력 비교 분석

1) 비교대상: 헤도닉 모형의 설정

본 연구는 앞서 소개한 XGBoost, LIghtGBM, CatBoost이 얼마나 실제 주택 시장의 복잡성과 비선형성을 효율적으로 포착할 수 있는지 알아보기 위해, 이러한 기계학습 모형들과 선형 헤도닉 가격 모형의 예측력을 비교하였다. 본문에서 사용한 헤도닉 모형은 다음과 같다.

$$\ln P_i = \beta_0 + X_{s,i}\beta_s + X_{l,i}\beta_l + T\beta_t + \epsilon_i \qquad \langle 식\ 1 \rangle$$

위에서 $\ln P_i$는 주택매매가격의 자연로그이며(하첨자 i는 i번째 주택을 의미한다), β_0는 상수항이다. $X_{s,i}$와 $X_{l,i}$는 각각 i번째 주택의 구조적 속성들과 입지적 속성들을 담은 벡터를 의미한다. β_s와 β_l는 각각 구조적 속성들과 입지적 속성들의 계수값을 담은 벡터이다. 본문의 모형에서 구조적 속성으로는 크기, 방 개수, 화장실 개수, 난방 방식, 층수, 경과년수, 세대 당 평균 주차가능 대수, 용적률, 건폐율, 단지 내 세대수, 단지 내 동 수, 단지 내 최고층 높이, 단지 내 최저층 높이가 포함되었다. 아파트가 분석대상인 경우, 개별 주택 자체의 특성 이상으로 해당 아파트 단지의 특성이 유의미한 영향을 미칠 것으로 보고 다양한 변수를 포함하였다. 입지적 속성으로는 가까운 지하철역으로부터의 거리, 주요 편의시설(공원, 박물관)로부터의 거리, 교육기관(초등, 중등, 고등, 대)으로부터의 거리, 행정시설(주민센터로부터의 거리)에 더해 행정동을 구분하는 더미변수가 포함되었다. 거리는 GIS 자료를 통해 유클리드 거리를 계산하였다. 행정동 더미가 포함된 이유는 주요 입지로부터의 거리뿐 아니라, 다양한 요소(예를 들어, 입지 주변 거주자들의 평균소득 등)들도 입지 가치에 큰 영향을 미칠 수 있기 때문이다. 이러한 입지별 차이의 상당 부분은 행정동의 구분과 연관되어 있기 때문에, 이를 포착하는 더미 변수를 모형에서 입지적 속성에 추가로 고려하였다. 즉, 행정동 더미의 추정 계수값에는 이처럼 직접 관찰할 수 없는 동 단위의 효과가 집약되어있는 것으로 이해할 수 있을 것이다. T는 거래 시점에 대한 더미변수이며, β_t는 그에 대한 계수이다. 거래 시점에 따라 이자율, 거시경제적 환경 등 다양한 요소가 복합적으로 가격에 영향을 미치므로, 다양한 개별 요소를 직접 통제하는 것보다는 이를 시점에 대한 더미변수의 형태로 한꺼번에 포착하는 것이 효과적일 것이다. 마지막으로 ϵ_i는 오차항으로 평균이 0인 정규분포를 따른다고 가정하였다.

본문에서는 헤도닉 모형의 독립변수로 상당히 다양한 변수를 포함하고 있는데, 한 가지 고려해야 할 점은 본문의 맥락에서 헤도닉 모형을 사용하는 이유가 종속변수 값의 예측 자체에 있다는 것이다. 일반적으로 변수를 많이 포함하게 되는 경우, 다중공선성 등의 문제로 인해 개별 변수의 영향력이 왜곡되는 경우가 나타날 수 있다. 그러나, 중요한 점은 종속변수의 예측에 대해서는 설사 다중공선성 문제가 있더라도 예측력이 저하되지 않는다는 것이다 (Greene, W. H., 2003). 이는 개별 변수의 계수가 변수간 공선성의 존재로 인해 왜곡되더라도, 그러한 관계를 포함한 최종결과가 종속변수에 대한 예측값으로 반환되기 때문이다. 이런 경우 모형의 예측력을 상승시키기 위해서는 가능한 다양한 변수를 포함하도록 설계할 수 있다. 다시 말해, 본문에서는 다소간 공선성을 일으킬 수 있는 변수들이 포함되어 있더라도 회귀분석 기반 모형의 예측력을 극대화하는 데에 집중하였다는 것이다. 이처럼 예측력이 극대화 된 헤도닉 모형을 기계학습 모형과 비교하는 것이 모형 간 예측력의 차이가 (변수 선정의 차이가 아닌) 각 모형의 비선형성과 복잡성의 포착력에 있다는 것을 보여줄 수 있을 것이다.

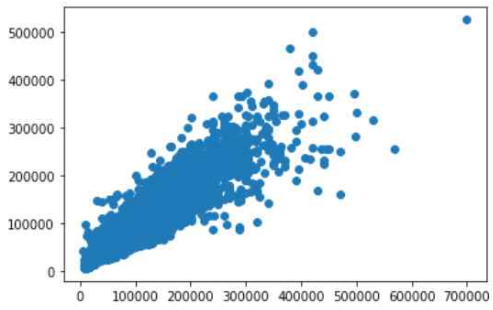
2) 모형 간 예측력 비교

이 장에서는 헤도닉 가격 모형, XGBoost, LightGBM, CatBoost 알고리즘을 통해 주택가격감정 모형을 측정하고, 그 예측력을 비교하였다. 모형 간 예측력을 객관적으로 비교하기 위해 수집된 전체 표본 620,617개는 학습 표본과 평가 표본으로 분할되었다. 이때 학습 표본은 전체의 90%인 558,555개, 평가 표본은 10%인 62,062개로 임의 할당되었다. 본문은 먼저 학습 표본을 통해 각 모형을 추정하였다. 그리고 이렇게 측정된 모형들에 대해 평가 표본들의 속성값만을 관찰하고, 매매가격을 예측하도록 하였다.[2)]

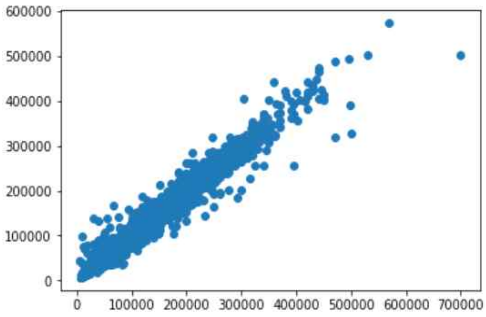
〈그림 1〉은 평가 표본의 속성값을 통해 각각의 모형이 예측한 주택 가치와 실제 시장에서 거래된 주택 가격의 산점도를 나타낸 것이다. 가로축은 실제 주택 매매가격을 의미하며, 그에 대한 예측치는 세로축에 나타나 있다. 만약 완벽하게 추정값과 실제값이 일치한다면 모든 점이 45도 선에 위치하게 될 것이며, 상관계수는 1이 될 것이다. 반면 추정치와 실제값이 아무런 관계가 없다면, 산점도에는 일정한 규칙이 나타나지 않을 것이며, 상관계수 역시 0에 가까울 것이다. 그림에 나타난 모든 모형에서 뚜렷한 정의 관계가 관찰된다. 한 가지 확인할 수 있는 것은 헤도닉 모형의 산점도가 다른 기계 학습 알고리즘 기반의 모형의 경우에 비해 분산이 크게 나타난다는 것이다. 특히 이러한 분산의 정도는 저가 주택에서보다 고가 주택에서 더 심하게 나타난다. 반면 XGBoost, LightGBM, CatBoost

2) 본 연구에서는 1. 전체 표본의 수가 충분히 크고, 2. 각 기계 학습자들이 앙상블 기법에 기반하기 때문에 위처럼 임의분할을 하더라도 그로부터 얻은 분석 결과의 질적인 해석은 거의 동일하다.

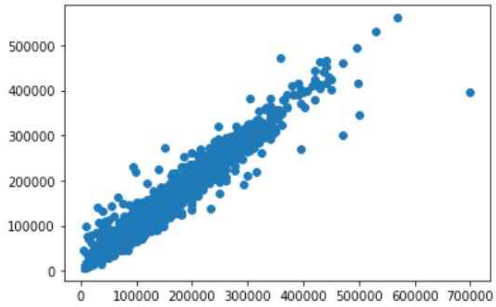
〈헤도닉 모형〉



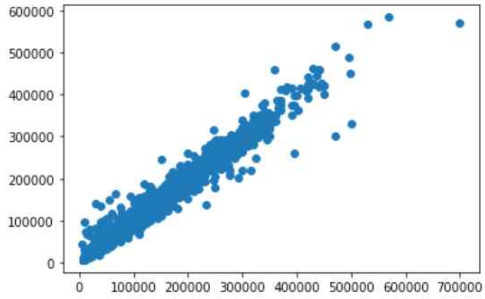
〈XGBoost〉



〈LightGBM〉



〈CatBoost〉



주: 가로축은 실제 주택매매가격, 세로축은 그에 대한 예측값.

〈그림 1〉 주택감정=모형의 산점도 비교

모형의 산점도에서는 이러한 현상이 나타나지 않는 것을 확인할 수 있다.

본문을 각 모형의 정확성을 다음의 두 가지 기준을 통해 평가하였다. 첫째는 평균 절대 백분율 오차(Mean Absolute Percentage Error, 이하 MAPE)이다. 이는 개별 추정치가 실제값에서 몇 퍼센트만큼 평균적으로 이탈하고 있는 지를 보여주는 가장 직접적인 평가 기준으로, 다음과 같이 정의된다.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\left|\frac{\hat{P}_i - P_i}{P_i}\right| \qquad \langle 식\ 2 \rangle$$

이때 P_i와 $\hat{P}_i$는 각각 i번째 실제 주택매매가격과 예측값을 의미한다.

두 번째는 R-squared로, 이는 모형을 통해 예측할 수 있는 표본 간 차이 (분산)과 실제 차이의 비율을 의미한다. R-squared는 아래와 같이 정의된다.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(P_i - \hat{P}_i)^2}{\sum_{i=1}^{n}(P_i - \overline{P})^2}, \qquad \langle 식\ 3 \rangle$$

이때 $\overline{P}$는 평가 표본의 평균을 의미한다.

아래 〈표 4〉는 각 모형들의 MAPE, 〈표 5〉는 R-squared를 비교하여 보여준다. 먼저, MAPE와 R-squared 모두에서 헤도닉 모형과 기계 학습 기반 알고리즘의 예측력 차이가 상당하다는 것을 확인할 수 있다. 헤도닉 모형의 경우 평균 백분율 오차는 11.5% 가량으로, 평가 표본의 개수가 62,206개에 달한다는 것을 고려할 때 낮은 수치라고 평가할 수는 없을 것이다. 그러나 기계학습 기반의 예측력은 모두 4% 이내로, 데이터 자체의 관측 오류나 변수로 관찰 불가능한 특이 표본(예를 들어, 판매자의 사정으로 실제 가치에 비해 크게 저평가된 거래 등)으로 인한 불가피한 설명력 손상이 있음을 감안할 때, 본문에서 사용된 알고리즘들은 시장 자체의 비선형성이나 입지 가치로 인한 주택 가치 결정구조의 복잡성 등 다양한 요소로 인한 가격 차이를 상당 부분 포착할 수 있다는 것을 의미한다. 또한, 기존의 다른 기계 학습(랜덤 포레스트나 서포트 벡터 머신) 또는 인공신경망 기반

〈표 4〉 Mean average percentage error (MAPE)의 비교

	MAPE
헤도닉 모형	11.508
XGBoost	3.728
LightGBM	3.854
CatBoost	3.609

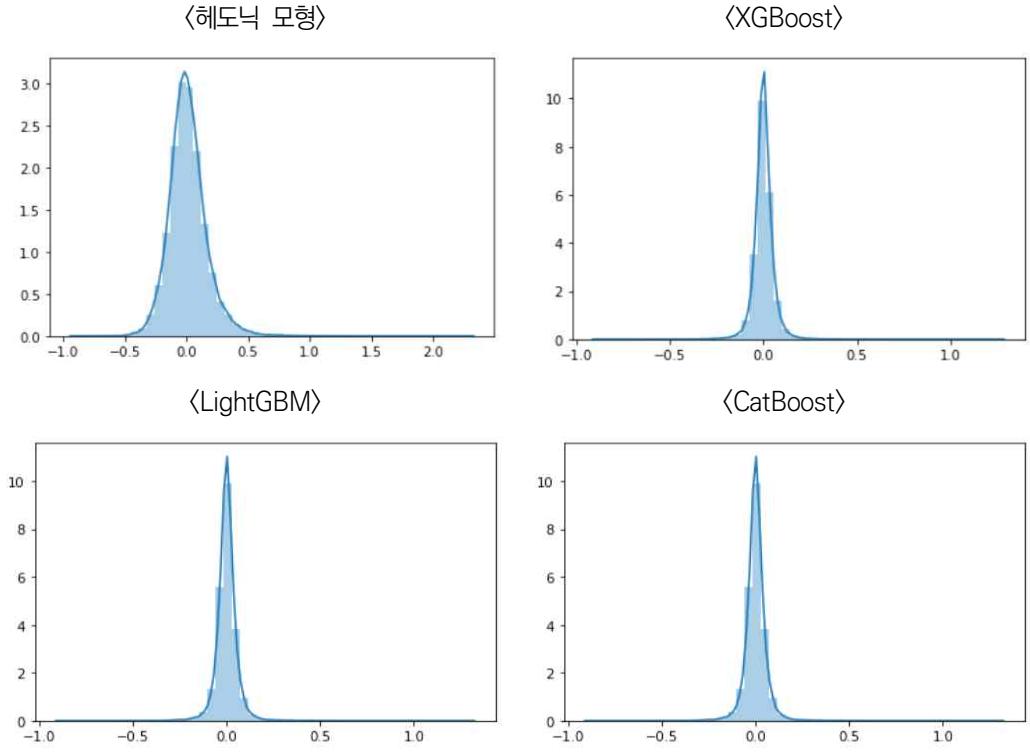
〈표 5〉 R-squared의 비교

	R-squared
헤도닉 모형	0.8962
XGBoost	0.9856
LightGBM	0.9836
CatBoost	0.9878

주택 감정모형의 평균 백분율 오차가 낮은 경우에도 5% 내외였다는 것을 생각해 보면, 그래디언트 부스트에 기반한 세 알고리즘의 활용성이 상당히 높다는 것을 알 수 있다. 본문에서 추정된 모형 중에서도 CatBoost의 예측력이 다소 더 높게 나타나는데, 이 모형은 평균 오차에서 뿐만 아니라, 다음 장에서 설명될 이상치의 발생빈도에서도 더 우수한 성능을 나타내는 것으로 확인된다.

3) 오차의 분포 및 성질

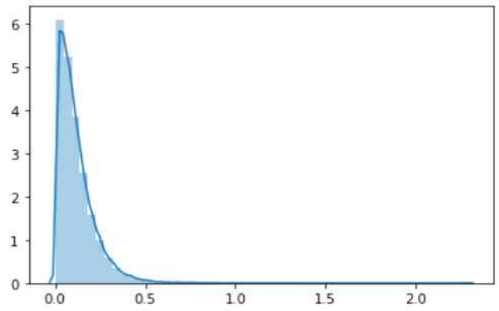
이 장에서 우리는 각 모형들을 통한 예측의 분포 및 이상치의 발생빈도를 비교분석하였다. 먼저 예측 오차(백분율 기준)의 분포를 통해 각 모형을 통한 예측의 특성을 간단히 비교할 수 있다. 다음 〈그림 2〉와 〈그림 3〉은 각 모형의 예측 오차와 그 절대값의 히스토그램을, 〈표 6〉은 표준편차와 분위값을 보여주고 있다. 그래프 형태와 표준편차 값의 차이에서 확인할 수 있듯이 헤도닉 모형에



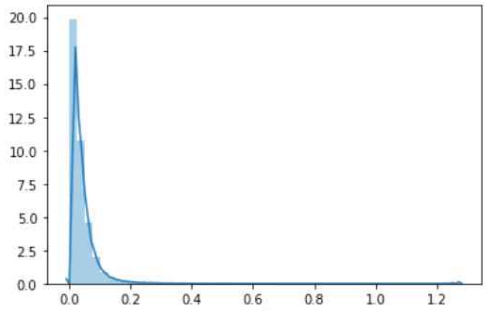
주: 가로축은 실제값과 예측값 간의 % 격차 ($= \frac{P_i - \hat{P}_i}{\hat{P}_i}$), 세로축은 그에 대한 빈도.

〈그림 2〉 예측 오차의 히스토그램

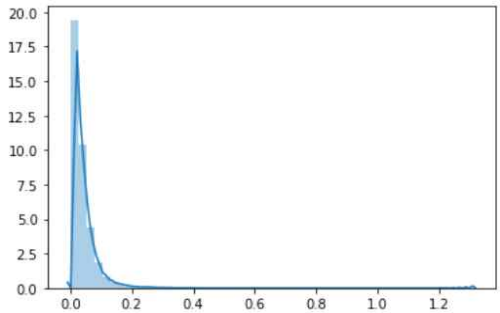
〈헤도닉 모형〉



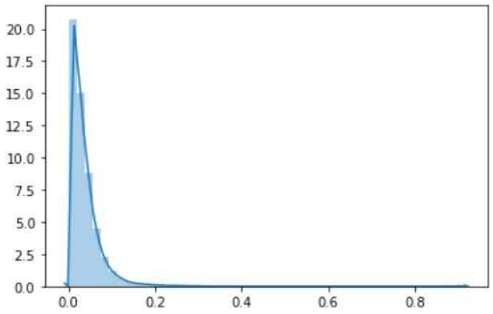
〈XGBoost〉



〈LightGBM〉



〈CatBoost〉



주: 가로축은 실제값과 예측값 간 % 격차의 절대값 ($= \left| \frac{P_i - \hat{P}_i}{\hat{P}_i} \right|$), 세로축은 그에 대한 빈도.

〈그림 3〉 예측 오차(절대값)의 히스토그램

〈표 6〉 예측 오차의 분포

	헤도닉 모형	XGBoost	LightGBM	CatBoost
표준편차	0.156275	0.056608	0.05787	0.052316
1사분위 값	-0.0837	-0.02357	-0.02508	-0.02443
2사분위 값 (중위값)	-0.00032	0.001036	0.000459	0.000626
3사분위 값	0.092038	0.026486	0.026019	0.026064
표본 수	62,062			

비해 세 기계학습 알고리즘을 통한 예측이 더 조밀하게 형성되는 것을 확인할 수 있다. 표준편차 값은 헤도닉 모형, XGBoost, LightGBM, CatBoost 순으로 각각 0.156, 0.056, 0.057, 0.052로,

앞에서 확인한 R-squared 및 MAPE의 순과 동일한 것을 발견할 수 있다. 〈표 6〉은 1,2,3 사분위 값도 제공하고 있는데, 이를 통해 전체의 상위 50%가 1사분위(25%) 값과 3사분위(75%) 값 사이에 존재한다는 것을 확인할 수 있다. 즉, 예측력이 가장 높은 CatBoost의 경우, 절반 정도의 예측이 −2.4%와 2.6%의 예측 오차 범위에 들어간다는 것이다.

다음으로, 각 모형들을 통한 예측에서 나타나는 오류 중 이상치의 발생 빈도에 대해 알아본다. 실제 값과 예측값 사이의 차이는 모형의 불완전성이나 관측 불가능한 속성들(매매 또는 구매자의 특성 등) 또는 자료 자체의 결함으로 인해서 불가피하게 발생할 수밖에 없으나, 종종 모형을 통한 예측에서는 추정값이 실제값에 비해 지나치게 크거나 작은 경우가 발견될 수 있다. 평균적 예측 오차 자체는 낮더라도, 이러한 이상치가 빈번하게 발생하는 경우, 모형의 신뢰성이 크게 떨어지기 때문에 가능한 한 이를 낮추는 것이 중요하다.

예측 이상치의 발생은 예측 모형의 복잡성 및 비선형성과 관련이 있을 수 있다. 선형 회귀분석 기반의 예측 모형의 경우, 종속변수에 대한 추정치는 속성값의 선형투사(linear projection)에 의해 이루어진다. 그러므로 계수가 과대 평가되거나 과소 평가된 속성의 속성값이 특정 표본에서 매우 큰 경우에만 큰 편차가 발생할 것이다. 반면, 그래디언트 부스트 기반의 세 알고리즘의 경우, 이상치의 발생 메커니즘을 정형화하기 어렵다. 이는 모형의 과적합 문제와도 연관되어 있을 수 있다. 세 알고리즘 모두 단순하고 약한 예측 모형을 오차를 수정하는 방향으로 반복 누적시키는 방식을 택하는데, 이때 약한 예측 모형으로는 의사결정트리가 사용된다. 이때 만약 의사결정트리들의 예측방향이 학습표본 자체에 포함된 특이성에 과도하게 적응하는 경우 평가표본의 예측에 대해서는 이상치를 발생시킬 가능성이 있다. 하지만 만약 기계학습 알고리즘 기반 모형들이 실제 시장의 비선형적 구조를 더 효과적으로 포착할 수 있다면, 모형에 내재된 비선형성이 이상치의 발생으로 연동되지는 않을 것이다.

다음의 〈표 7〉은 이러한 이상치의 발생빈도를 비교하고 있다. 본문 맥락에서 예측 이상치는 추정값이 실제값에 비해 지나치게 크거나 작은 경우로, 여기서는 다음의 세 가지를 기준으로 정의하였다; 1. 실제값이 예측값의 50%이상 멀어진 경우, 2. 실제값이 예측값의 75%이상 멀어진 경우, 3. 실제값이 예측값의 100%이상 멀어진 경우. 표의 결과를 요약하면 다음과 같다. 첫째, 기계학습 알고리즘의 이상치 발생 빈도가 헤도닉 모형에 비해 현저히 낮다. 헤도닉 모형의 경우, 620,617개의 평가 표본 중 예측값이 실제의 50% 이상 벗어난 사례가 545건으로 전체의 약 0.87% 정도가 이에 해당한다. 반면, XGBoost, LightGBN, CatBoost의 경우 각각 58, 43, 31건으로 선형회귀 기반 모형에 비해 10~15분의 1 이하로 감소한 것을 확인할 수 있다. 모형뿐 아니라 데이터 자체의 결함도 존재할 가능성이 있음을 감안하면, 기계 학습 알고리즘의 복잡성으로 인한 불안정성

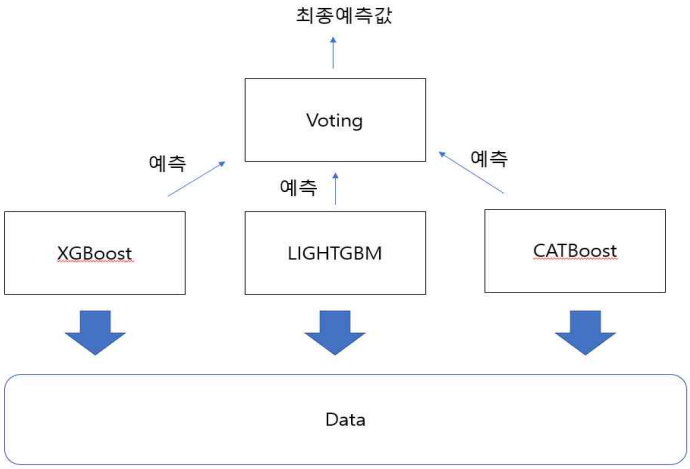
〈표 7〉 이상치 발생 비율

	오차 50% 이상	오차 75% 이상	오차 100% 이상
헤도닉 모형	0.00878	0.00209	0.00058
XGBoost	0.00085	0.00021	0.00005
LightGBM	0.00077	0.00016	0.00002
CatBoost	0.0005	0.00011	0

문제는 크지 않음을 알 수 있다. 둘째, 기계 학습 알고리즘 중에서도 CatBoost 알고리즘 기반 모형의 예측 이상치 발생율이 다른 두 알고리즘에 비해 더 낮은 것으로 확인된다. 특히 CatBoost에서는 예측치가 실제의 100% 이상 벗어나는 큰폭의 예측 오류가 620,617건 중 단 한 건도 발생하지 않았다. 또한, 평균적 예측력은 XGBoost가 LightGBM보다 다소 높지만(〈표 4, 5〉), 이상치의 발생빈도는 LightGBM가 더 낮은 것으로 나타났다.

2. 기계학습 예측기의 결합

이 장에서는 위에서 제시된 기계 학습 알고리즘을 얻은 모형의 결합을 통해 더 높은 예측력을 가진 모형을 구현할 수 있는지에 대해 간략히 알아보았다. 모형의 결합은 보팅(voting)이라고도 하는데, 말 그대로 서로 다른 알고리즘을 통해 얻은 예측값들을 투표(다수결) 또는 평균하여 개별 알고리즘에 포함될 수 있는 오류의 가능성을 중화하는 앙상블 기법이다 (〈그림 4〉). 이때 투표의 경우 하드보팅,

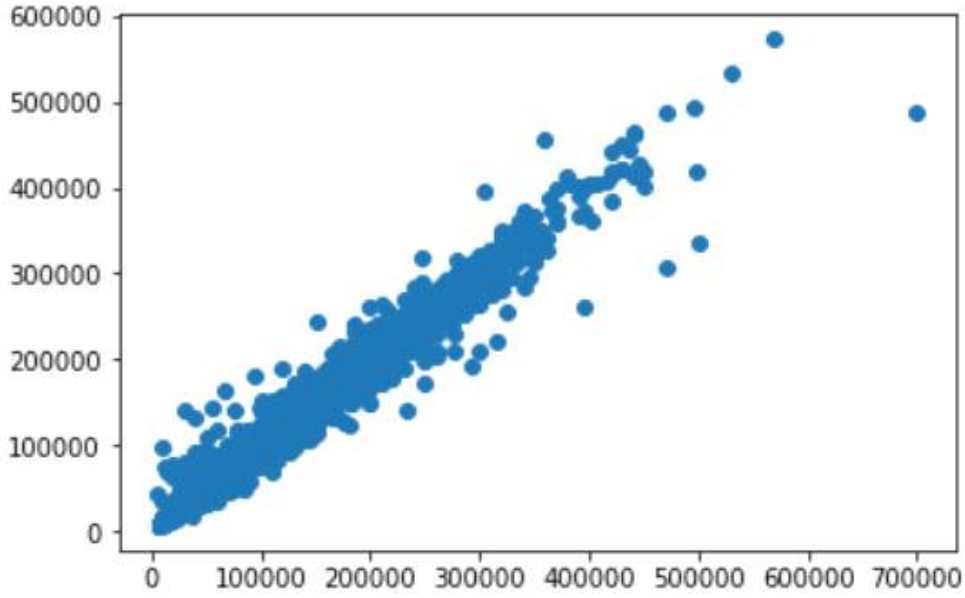


〈그림 4〉 서로 다른 알고리즘의 결합

평균의 경우 소프트 보팅이라고 부른다. 하드 보팅의 경우, 최종 예측값이 범주변수의 형태인 분류 문제에 적용될 수 있기 때문에, 본 연구에서는 위의 세 알고리즘(XGBoost, LightGBM, CatBoost)을 통한 소프트 보팅의 예측 성능을 측정해 보았다. 본문에서는 가장 직관적인 형태의 소프트 보팅인 세 알고리즘의 예측값 평균을 사용하였다.

〈그림 5〉는 이러한 앙상블 모형의 예측치와 실제 값 간 산점도이며, 〈그림 6〉은 백분율 오차의 히스토그램이다. 앙상블 모형의 산점도와 히스토그램은 본질적으로 세 알고리즘의 산점도와 히스토그램과 유사할 수밖에 없을 것이다. 그러나, 세 알고리즘의 산점도와 앙상블 모형의 산점도를 자세히 비교해 보면, 각각의 세 알고리즘의 산점도에서 상대적으로 예측오차가 벌어졌던 구간(45도 선에서 분산된 구간)들이 약간이나마 조밀해진 것을 확인할 수 있다.

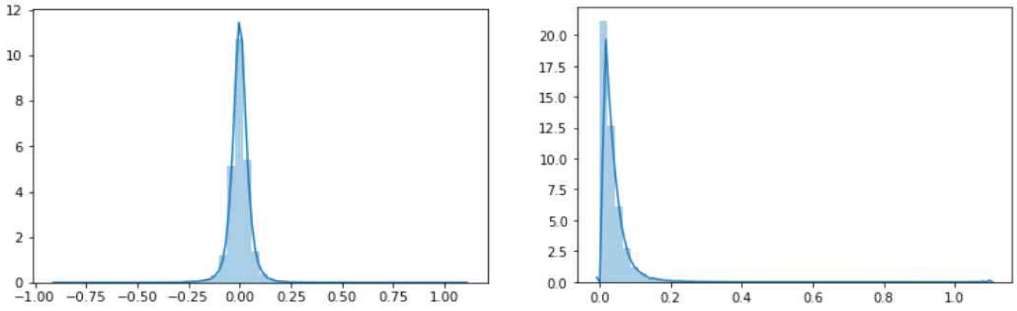
〈표 8〉은 예측 성능을 평가했던 기준(MAPE, R-squared)을 앙상블 모형에 적용한 것으로, 이러한 예측 성능의 상승 효과를 보다 직접적으로 드러내고 있다. 흥미로운 것은 R-squared와 MAPE 기준 모두에서 약간이나마 앞서 소개된 알고리즘들보다 예측 성능이 상승하는 것으로 나타난다는 것이다. 〈표 8〉과 〈표 4〉를 비교해 보면, XGBoost, LightGBM, CatBoost의 MAPE는 각각 3.72, 3.85, 3.60인 데에 반해, 앙상블 모형의 경우 동일한 표본에 대한 MAPE가 3.54로 더 낮아지는 것으로 확인된다. 이는 각각의 알고리즘 고유의 특성으로 인해 생겼던 소음들이 다른 알고리즘에는 나타나지



주: 가로축은 실제 주택매매가격, 세로축은 그에 대한 예측값.

〈그림 5〉 결합된 모형을 통한 예측의 산점도

〈백분율 오차〉 〈절대값 백분율 오차〉



주1: 첫 번째 그래프의 가로축은 실제값과 예측값 간 % 격차 ($= \frac{P_i - \hat{P}_i}{\hat{P}_i}$), 세로축은 그에 대한 빈도.

주2: 첫 번째 그래프의 가로축은 실제값과 예측값 간 % 격차의 절대값 ($= \left| \frac{P_i - \hat{P}_i}{\hat{P}_i} \right|$), 세로축은 그에 대한 빈도.

〈그림 6〉 결합 모형의 예측 오차 히스토그램

〈표 8〉 결합된 모형의 예측 성능

R-sqaured	MAPE
0.9871	3.549

않기 때문에, 이를 평균하는 과정에서 예측에 포함된 소음의 크기가 작아진 효과로 해석할 수 있다. 즉, 세 알고리즘 모두 기존의 모형에 비해 뚜렷한 성능 향상이 나타나지만, 앙상블 기법을 통해 이를 결합하는 경우 예측 성능을 더 상승시킬 수 있다는 것이다. 이러한 예측 성능의 향상은 〈표 9〉와 〈표 6〉의 비교에서도 발견할 수 있다.

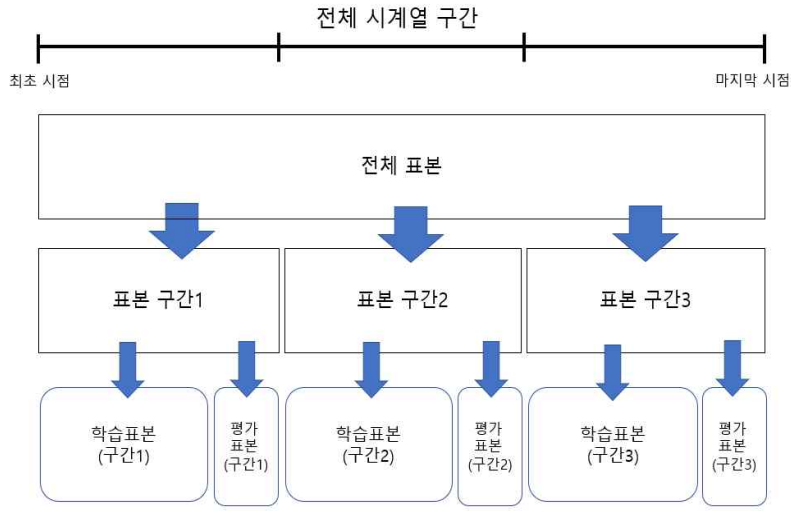
〈표 9〉 결합된 모형의 백분율 예측 오차 분포

표준편차	0.052749
1사분위 값	-0.02319
2사분위 값(중위값)	0.000745
3사분위 값	0.024807
표본 수	62,062

3. 시계열 순서에 의한 표본 분할 검증

앞의 기본 정량분석에서는 전체 표본을 9:1의 비율로 임의분할하여 학습 데이터와 평가 데이터를 구분하였다. 이는 본문의 맥락에서 감정 모형의 활용 가능성이 현시점만의 자산가치 감정에 국한되지 않기 때문에, 가능한 한 많은 표본을 학습함으로써 예측력을 극대화하려는 시도로 볼 수 있다. 그러나 이처럼 시계열 순서를 무시하는 경우, 현 시점에서 유사한 다른 자산의 가치들이 아직 관찰되지 않은 자산에 대한 예측력은 제대로 평가될 수 없다.[3] 따라서 이 장에서는 시계열 순서를 고려하여 학습 표본과 평가 표본을 분할하고, 각 모형의 예측력을 비교한 결과를 수록하였다.

분할 방식은 아래와 같다. 먼저, 분석 대상의 전체 시계열 구간(본 연구에서는 월 단위까지 관측)을 균등하게 분할한다.[4] 다음으로는 분할된 각각의 구간 표본을 시계열 순서로 나열하고, 순서가 빠른 90%는 학습 표본으로, 나머지 10%는 평가 표본으로 사용한다 (〈그림 7〉). 이처럼 시계열 구간 분할을 이용하는 경우, 모든 학습 표본은 평가 표본에 비해 과거값이므로 결과적으로 과거 관측값을 토대로



〈그림 7〉 시계열 구간에 기반한 표본 분할

3) 시계열 순서를 무시하고 학습하는 경우(아직 관찰되지 않은), 해당 시점 또는 그 이후의 변화를 학습할 수 있기 때문이다.

4) 그런데, 주택 거래 빈도는 매년 일정하지 않으므로 이 경우 시계열을 균등분할하는 경우, 각 구간에 포함된 표본의 개수는 달라질 수 있다. 반면, 표본 개수를 동일하게 분할하는 경우, 시계열 구간의 길이가 균등하지 않을 것이다. 본 연구에서는 맥락상 분할 된 시계열의 길이를 균일하게 맞추는 쪽에 초점을 맞추었다.

미래 자산가치를 예측하는 모형이 된다.

한 가지 고려해야 할 점은 표본의 분할이 곧 표본의 크기가 줄어든다는 것을 의미하므로, 자연스럽게 모형의 예측력 저하를 피할 수 없다는 것이다. 즉, 앞 장에서 제시된 예측력과의 차이가 시계열 순서를 고려함으로써 발생할 뿐 아니라, 표본 수 감소에 의해서도 나타날 수 있다는 것이다. 이로 인한 영향을 줄이기 위해서는 분할된 각 구간에 충분한 수의 표본이 포함될 수 있을 정도로 분할이 이루어져야 할 것이다. 반면, 구간 분할 없이 시계열 순서로 전체 표본을 학습 표본과 평가 표본으로 할당하는 경우, 평가 표본과 학습 표본의 시간 격차가 지나치게 벌어지게 된다. 본문에서는 이 두 가지 측면을 고려하여, 전체 시계열 구간을 3개의 하위 구간으로 분할한 경우의 결과를 수록하였다.[5)]

아래 〈표 10〉은 각 모형들의 MAPE, 〈표 11〉은 R-squared를 비교하여 보여준다. 먼저 확인할 수 있는 것은, 표본수가 약 3분의 1로 감소하였음에도 불구하고, 예측력 저하가 심하지 않다는 것이다. 예를 들어 CatBoost의 경우, MAPE 기준 정확성이 4.2~5.9로 3배의 표본 수를 사용한 앞 장의 분석 결과에 비해 약 1~2% 정도의 평균 오차 증가만이 있는 것으로 나타났다. 이는 시계열 배열 순서를 예측에 고려하더라도 모형의 정확성이 상당한 수준으로 유지될 수 있음을 의미한다. 둘째, 앞 장에서와 마찬가지로 XGBoost와 LightGBM에 비해 근소하지만 CatBoost의 예측성능이 더 우수한

〈표 10〉 각 시계열 구간 별 MAPE의 비교

	구간 1	구간 2	구간 3
XGBoost	6.18	4.62	5.66
LightGBM	6.93	4.61	5.53
CatBoost	5.93	4.2	4.82

〈표 11〉 각 시계열 구간 별 R-squared의 비교

	구간 1	구간 2	구간 3
XGBoost	0.943	0.977	0.97
LightGBM	0.938	0.973	0.973
CatBoost	0.946	0.978	0.981

5) 본 연구의 자료는 2009년 7월부터 2019년 12월까지를 포함하므로, 3개로 분할하는 경우 각 구간이 약 3년 반 정도의 기간을 포함한다고 볼 수 있다.

것으로 나타났다. MAPE 기준 XGBoost는 4.62~6.18, LightBoost는 4.61~6.93의 정확성을 보이고 있는데, CatBoost는 4.2~5.93으로 약 평균 1% 가량 예측력이 더 높은 것으로 나타났다.

Ⅵ. 결론 및 정책적 시사점

본 연구는 2009년부터 2019년까지 서울에서 거래 된 아파트 매매 데이터를 통해 헤도닉 모형과 세 알고리즘 (XGBoost · LightGBM · CatBoost) 기반의 주택 가치 감정 모형을 비교 · 분석하였다. 분석 결과, 전통적인 선형 헤도닉 모형에 비해 기계 학습 알고리즘의 예측력이 월등히 높게 나타났다. 회귀분석 기반 모형의 경우, 예측의 평균 백분율 오차는 약 11.5% 내외인 반면, XGBoost · LightGBM · CatBoost는 각각 3.7%, 3.8%, 3.6%에 불과했다. 이는 기계 학습 기반의 예측력이 다른 모형과의 상대적인 비교뿐 아니라, 절대적 수치 측면에서도 유의미하게 낮아질 수 있다는 것을 의미한다. 본 연구에서 사용한 세 알고리즘 중 미세하나마 CatBoost 알고리즘의 예측력이 다른 두 알고리즘에 비해 더 높게 나타나는 것도 확인할 수 있었다. 특히, 이상치의 발생빈도가 CatBoost에서 더 낮았으며, 이는 모형의 평균적 정확성뿐 아니라, 안정성이 다른 알고리즘에 비해 우수하다는 것을 의미한다.

흥미로운 것은, 세 알고리즘의 앙상블 모형이 개별 알고리즘보다 더 높은 예측 정확성을 가지는 것으로 나타났다는 것이다. 이는 본 연구가 기계학습을 통한 주택감정모형의 구축 과정에서, 예측력이 높은 특정 알고리즘을 선별해 활용하는 것뿐만 아니라, 다양한 알고리즘의 조합을 통해 정확도를 더 상승시킬 여지가 있음을 보여주고 있음을 의미한다.

이처럼 기계학습 알고리즘을 통해 감정 모형의 정확도가 상승한다는 것은 무엇을 의미할까? 현재도 다양한 분야에서 주택 가격 감정 시스템 (또는 모형)을 활용하고 있다. 예를 들어, 주택가격지수나 조세액의 추정 시 직접 관찰되지 않은 (시장에서 거래되지 않은) 주택의 가격은 모형을 통해 추정되고 있다. 그러나 본 연구에서도 알아본 바와 같이 선형 헤도닉 모형 기반의 가격 추정은 약 10% 내외의 상당한 예측오차를 동반할 수 있으므로, 보다 정교한 추정이 요구되는 실용적 활용 가능성은 제약되어 있다고 볼 수 있다. 기계학습을 이용해 모형의 정확도를 실용적으로 활용할 수 있는 일정 수준 이하로 끌어올리는 경우, 기존의 활용되던 분야에서의 생산성과 정확성이 상승할 뿐 아니라, 주택감정모형을 통한 다양한 활동(사업성 분석이나 자산 관리 등)이 본격적으로 가능해질 것으로 예상된다.

그럼에도 불구하고 이와 같은 기계학습 기반의 모형들은 모형의 해석 가능성이 매우 낮다는 단점을 가지고 있다. 전통적인 헤도닉 모형은 쉽게 해석 가능한 특정 형태의 수식으로 변수 간 관계를 표현하는 반면, 본 연구에서 분석한 기계학습 모형들은 변수와 가격 간의 관계를 직접적으로 표현하는

수식을 갖지 않기 때문에 종합적인 예측력은 상승하지만 그 가치 결정의 구조를 직관적으로 분석하는 데에는 한계가 있다. 후속 연구들에서는 기계학습을 통해 얻은 예측자가 변수 변화의 구조에 어떤 식으로 예측값을 변경하는지를 2차적으로 분석하는 등의 시도들도 가능할 것으로 보인다. 이후에는 기계학습 모형의 높은 예측 정확성을 통해 수행할 수 있는 다양한 활용법뿐만 아니라, 기계학습을 통해 실제 부동산 시장의 가치 결정구조에 어떤 비선형성이 존재하는지를 파악하는 모형들도 더 제시될 것으로 기대한다.

참고문헌

Anselin, L., 2001, "Spatial econometrics". A Companion to Theoretical Econometrics, 310330.

Antipov, E. A. and Pokryshevskaya, E. B., 2012, "Mass appraisal of residential apartments: An application of random forest for valuation and a CART-based approach for model diagnostics," Expert Systems with Applications, 39(2), 1772-1778.

Basu, S. and Thibodeau, T. G., 1998, "Analysis of spatial autocorrelation in house prices," The Journal of Real Estate Finance and Economics, 17(1), 61-85.

Conway, D., Li, C. Q., Wolch, J., Kahle, C. and Jerrett, M., 2010, "A spatial autocorrelation approach for examining the effects of urban greenspace on residential property values," The Journal of Real Estate Finance and Economics, 41(2), 150-169.

Dubin, R. A., 1992, "Spatial autocorrelation and neighborhood quality," Regional Science and Urban Economics, 22(3), 433-452.

Fan, G. Z., Ong, S. E., and Koh, H. C., 2006, "Determinants of house price: A decision tree approach," Urban Studies, 43(12), 2301-2315.

Goodman, A. C., 1978, "Hedonic prices, price indices and housing markets," Journal of Urban Economics, 5(4), 471-484.

Greene, W. H., 2003, Econometric analysis. Pearson Education India.

Gu, J., Zhu, M. and Jiang, L., 2011, "Housing price forecasting based on genetic algorithm and support vector machine," Expert Systems with Applications, 38(4), 3383-3386.

Halvorsen, R. and Pollakowski, H. O., 1981, "Choice of functional form for hedonic price equations," Journal of Urban Economics, 10(1), 37-49.

Hong, J., Choi, H. and Kim, W. S., 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.

IAAO. Standard on Mass Appraisal of Real Prop.; IAAO: Kansas City, MO, USA, 2017.

Lancaster, K. J., 1966, "A new approach to consumer theory," Journal of Political Economy, 74(2), 132-157.

Limsombunchai, V., 2004, "House price prediction: Hedonic price model vs. artificial neural network," In New Zealand Agricultural and Resource Economics Society Conference (pp. 25-26).

Malpezzi, S., 2002, "Hedonic pricing models: A selective and applied review," Housing Economics and Public Policy, 67-89.

McCluskey, W., and Anand, S., 1999, "The application of intelligent hybrid techniques for the mass appraisal of residential properties," Journal of Property Investment & Finance.

Meese, R. and Wallace, N., 1991, "Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices," Real Estate Economics, 19(3), 308-332.

Mu, J., Wu, F. and Zhang, A, 2014, "Housing value forecasting based on machine learning methods," In Abstract and Applied Analysis (Vol. 2014). Hindawi.

Park, B., and Bae, J. K., 2015, "Using machine learning algorithms for housing price prediction: The case of fairfax county, virginia housing data," Expert Systems with Applications, 42(6), 2928-2934.

Rasmussen, D. W. and Zuehlke, T. W., 1990, "On the choice of functional form for hedonic price functions," Applied Economics, 22(4), 431-438.

Rosen, S., 1974, "Hedonic prices and implicit markets: Product differentiation in pure competition," Journal of Political Economy, 82(1), 34-55.

Selim, H., 2009, "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network," Expert Systems with Applications, 36(2), 2843-2852.

Sheppard, S., 1999, "Hedonic analysis of housing markets," Handbook of Regional and Urban Economics, 3(1), 595-1635.

Wang, D. and Li, V. J., 2019, "Mass appraisal models of real estate in the 21st century: A systematic literature review," Sustainability, 11(24), 7006.

Zhou, G., Ji, Y., Chen, X. and Zhang, F., 2018, "Artificial neural networks and the mass appraisal of real estate," International Journal of Online and Biomedical Engineering (IJOE), 14(03), 180-187.

Zurada, J., Levitan, A., and Guan, J., 2011, "A comparison of regression and artificial

intelligence methods in a mass appraisal context", Journal of Real Estate Research, 33(3), 349-387.

(논문 접수일: 2020.10.17. 수정논문 접수일: 2020.11.30. 논문 채택일: 2020.12.11.)

부록. 주요 하이퍼파라메터 설정 내용

〈부록 표 1〉 XGBoost의 하이퍼파라메터

하이퍼파라메터명	설정내용
base_score	0.5
gamma	10
learning_rate	0.1
max_depth	20
min_child_weight	1
n_estimators	100
n_jobs	1
nthread	None
objective	reg:linear
random_state	0
reg_alpha	0
reg_lambda	2
scale_pos_weight	1
colsample_bytree	1
subsample	1
verbosity	1

〈부록 표 2〉 LightGBM의 하이퍼파라메터

하이퍼파라메터명	설정내용
boosting_type	gbdt
class_weight	None
colsample_bytree	1.0
importance_type	split
learning_rate	0.1
max_depth	200
min_child_samples	20
min_child_weight	0.001
min_split_gain	0.0
n_estimators	3000
n_jobs	−1
num_leaves	20000
random_state	None
reg_alpha	0.0
reg_lambda	0.0
subsample	1.0
n_estimators	3000
min_data_in_leaf	5

〈부록 표 3〉 CatBoost의 하이퍼파라메터

하이퍼파라메터명	설정내용
nan_mode	Min
eval_metric	RMSE
iterations	4000
sampling_frequency	PerTree
fold_permutation_block	0
leaf_estimation_method	Newton
grow_policy	SymmetricTree
penalties_coefficient	1
boosting_type	Plain
model_shrink_mode	Constant
feature_border_type	GreedyLogSum
bayesian_matrix_reg	0.10
l2_leaf_reg	1
random_strength	1
rsm	1
boost_from_average	True
max_ctr_complexity	4
model_size_reg	0.5
subsample	0.80
use_best_model	False
depth	12
sparse_features_conflict_fraction	0
leaf_estimation_backtracking	Anyimprovement
best_model_min_trees	1
model_shrink_rate	0
min_data_in_leaf	1
loss_function	RMSE
learning_rate	0.1
score_function	Cosine
leaf_estimation_terations	1
bootstrap_type	MVS
max_leaves	64

Housing Finance Research (2020) vol. 4, 33-64
ISSN: 2586-4629 https://www.hf.go.kr

An Application of XGBoost, LightGBM, CatBoost Algorithms on House Price Appraisal System

Jengei Hong*

Abstract

This paper compares the predictive power of a conventional hedonic pricing model and three machine learning algorithm (XGBoost, LightGBM, CatBoost) based models by using 620,617 apartment data in Seoul from 2009 to 2019. The results are summarised as follows; First, the predictive power of the machine learning models are significantly high not only in the comparison to the conventional model but also in the absolute accuracy related to its practical usefulness. The mean percentage error of XGBoost, LightGBM, and CatBoost were only, respectively, 3.7%, 3.8%, and 3.6% while those of the hedonic model was around 11%. Second, we found that CatBoost algorithm is slightely more performative to the other two algorithms in terms of overall predictive power and frequency of outlier occurrences. Third, this paper show that an ensemble model of the three algorithms can raise the predictive power further.

Keywords : Machine Learning, XGBoost, LightGBM, CatBoost, Mass Appraisal

* Jengei Hong, Corresponding author, School of Management & Economics, Handong Global University, Assistant Professor, hwgh024@handong.edu