「정보시스템연구」 제29권 제3호
한국정보시스템학회
2020년 9월, pp. 119~144

http://dx.doi.org/10.5859/KAIS.2020.29.3.119

기계학습 알고리즘을 활용한 지역 별 아파트 실거래가격지수 예측모델 비교: LIME 해석력 검증

조보근* · 박경배** · 하성호***

〈목 차〉

Ⅰ. 서 론

부동산은 개인과 조직, 정부가 소유하고 있는 자산 중 가장 높은 비중을 차지하며, 부동산 가격의 변동은 경제 주체들의 경제 상황에 큰 영향을 미친다. 2006년 이후 미국이 초 저금리 정책을 종료함에 따라 미국의 서브프라임 모기지 대출의 금리는 상승하고 부동산 버블이 꺼지며 주택 가격은 급락하였다. 또한, 서브프라임 모기지 대출의 부실화는 대출 서비스를 제공하는 금융기관들의 부실을 야기하여 미국의 대형 금융사, 증권회사의 파산으로 이어졌다. 2008년 이후 서브프라임 모기지 사태는 세계 경제시장에도 영향을 주어 세계금융위기를 불러 일으켰다(신용상 등, 2007).

한국의 경우, 1998년 IMF 외환위기 이후 부동산 가격이 급락하여 자산디플레이션 우려가 제기되었다. 경기가 안정화되며 부동산의 가격

* 빅웨이브에이아이 이사, iambgjo@gmail.com(주저자)
** 인텔리콘 법률 AI 미디어랩 소장, kbp@intellicon.co.kr
*** 경북대학교 경영학부, hsh@knu.ac.kr(교신저자)

은 빠른 속도로 회복되었으나, 이후 세계 경기와 정부의 정책 및 금리의 변화 등 다양한 요인들에 의해 부동산 가격은 크고 작은 변동을 겪어 왔다. 이와 같이 급변하는 시장 환경 속에서 부동산 가격을 예측하고자 하는 시도는 국가, 기업, 가계에 있어, 행정상의 편리, 투자 및 재산 관리 등의 목적으로 큰 관심을 받고 있다. 또한, 최근 각광받고 있는 기계학습 기술의 발전에 따라 더욱 정밀하고 높은 예측력을 가진 모델의 등장에 대한 기대감이 높아지고 있다(정훈・김주원, 2017).

부동산 가격 예측을 위한 전통적인 접근 방법은 설명변수와 종속변수 간의 선형 모델을 기반으로 한 선형회귀모형을 활용하는 것이었다. 기존에 부동산 가격 지수를 예측하기 위하여 주로 사용된 모델로는 자기회귀이동평균모형(Auto-Regressive Integrated Moving Average, ARIMA), 벡터자기회귀모형(Vector Auto-Regression, VAR), 벡터오차수정모형(Vector Error Correction Model, VECM) 등이 있다(배성완・유정석, 2017; 임병진・한성윤, 2009; 한경수, 2011). 최근 기계학습 기법을 활용한 예측모델이 여러 분야에서 적용되고 있으며, 부동산 가격 예측을 위한 시도 또한 증가하고 있는 추세이다. 비선형 모델링이 가능한 기계학습을 활용한 모델은 기존의 선형 기반의 예측모형 보다 정밀한 예측이 가능하다(배성완・유정석, 2018). 그러나 특정 기계학습 기법은 내부적으로 어떠한 변수가 모델에 영향을 미치는지 파악하는 것이 쉽지 않다는 한계가 있다.

본 연구는 부동산 가격 예측에 적합한 예측모형을 확인하고 지역적으로 부동산 가격 예측에 중요한 변수를 규명하여 예측모델의 결과에 대한 해석을 하고자 한다. 이를 위해 본 연구에서는 첫째, 부동산 가격 예측을 위해 전통적으로 활용되는 선형 기반의 회귀모델 중 선행연구에서 활용되는 대표적인 단변량 시계열 예측모형인 ARIMA 모형과 기계학습 기법의 예측모델인 랜덤 포레스트(Random Forest), 순환신경망(Recurrent Neural Network, RNN) 기반의 장단기 메모리(Long Short-Term Memory, LSTM) 모델들의 예측 성능을 기반으로 평가 및 비교 분석한다. 둘째로, 선정된 최종 예측모델을 기반으로 도시 별로 적용한 후 지역적 대리모형인 LIME(Local Interpretable Model-agnostic Explanation) 알고리즘을 통해 지역 별 예측모델에서 주요 변수들을 규명한다. 마지막으로, 최종 분석된 지역 별 특징을 비교 분석하여 부동산 가격이 급변하는 지점의 동인을 도출한다.

Ⅱ. 이론적 배경

2.1 경영정보학의 시계열 예측 연구

최근 경영정보학 분야의 다양한 적용 가능성에 힘입어 여러 분야에서 시계열 예측 연구가 진행되고 있다. 선일석(2015)의 연구에서 창고업의 관리 및 정책 수립에 시사점을 제공하기 위하여 ARIMA 모형을 활용하여 보관 및 창고업 생산지수의 예측을 진행하였다. 보관 및 창고업 생산지수의 경우 1년 단위로 등락이 반복되는 계절성을 가진 시계열 자료로 1차 차분 후 ARIMA(1,1,0)(2,1,0) 12 모형을 통해 향후 생

산지수의 예측을 시도하였다. 예측 결과, 보관 및 창고업 생산지수가 하락하는 추세를 가지며 이에 대비할 수 있는 전략의 필요성을 제기하였다.

신이레와 윤상후(2016)는 특정 시간대를 참조 선으로 하여 전력수요를 예측하는 연구를 진행하였다. 분석에 활용된 모형은 주간 계절성과 연간 계절성을 고려할 수 있는 이중 계절성 Holt-Winter 모형과 ARMA 모형 기반의 TBATS(Trigonometric Exponential Smoothing State Space Model with Box-Cox Transformation, ARMA Errors, Trend and Seasonal Components) 모형을 적용하여 각 모형의 예측 성능을 비교하였다. 평균제곱근오차(Root Mean Square Error, RMSE)와 절대평균백분율오차(Mean Absolute Percentage Error, MAPE)를 활용하여 예측 성능을 비교한 결과 TBATS 모형의 예측 성능이 높은 것으로 나타났다.

서종덕(2016)의 연구에서는 우리나라의 환율 예측을 위한 기계학습 기반의 모형과 계량경제 모형의 혼합 모델을 제시하였다. 랜덤 포레스트 모형과 금융시장의 변동성을 고려할 수 있는 GARCH 모형을 결합한 RF-GARCH 모형의 경우 기존의 GARCH 모형에 비해 높은 예측력을 가지는 것을 검증하였다. 또한, 예측 과정에서 다양한 변수들을 선택하며 진행한 결과 미국 달러 지수, KRX100 지수, KOSPI 지수 순으로 외환시장에서 영향을 미치는 것을 확인하였다.

최가영 등(2017)은 자연 휴양림 이용 수요를 예측하기 위해 계절 ARIMA 모형과 지수평활모형을 활용하였다. 자연휴양림 이용객의 수의 경우 계절성이 뚜렷하게 나타나 는 것으로 나타났으며, 예측 결과, 계절 ARIMA 모형에 비해 지수평활모형인 Winters 가법 모형의 예측력이 상대적으로 높은 것으로 나타났다. 또한, 점진적인 상승 추세를 확인하였으며, 여름과 가을의 이용객이 집중되는 것을 확인하여 해당 연구가 관광업의 효율적인 관리정책의 수립에 도움을 줄 것으로 기대하였다.

이낙영과 오경주(2019)의 연구에서 KOSPI200 선물지수 예측을 위하여 디노이징 필터(denoising filters)와 LSTM 모형을 적용하였다. 데이터의 노이즈를 제거한 후 LSTM 모형을 적용한 결과 예측모형의 성능이 더 우수하게 나타났으며, 디노이징 필터의 경우에 이동평균 필터(moving average filters)보다 사비츠키-골레이 필터(Savitzky-Golay filters)를 활용하였을 때 예측 성능이 향상되는 것을 확인하였다.

이와 같이, 국내의 경영정보학 관련 연구의 경우 시계열 예측 연구를 다양한 분야에 적용하고 있는 것으로 확인되었다. 예측모형의 경우 ARIMA 모형이 많이 활용되고 있으며, 최근 기계학습 관련 방법론을 적용한 연구가 활발히 진행되고 있다. 본 연구의 경우 아파트실거래가격지수 예측을 위해 적합한 모형을 찾기 위해 ARIMA 모형과 랜덤 포레스트, LSTM 모형을 RMSE와 MAPE 기반으로 각 모델의 성능을 비교한 후 최적의 예측 성능을 가진 모델을 활용하여 최종 비교 분석에 사용하였다.

2.2 부동산 가격지수 예측모델 연구

우리나라의 부동산 관련 연구에서는 거시경제지표와 부동산 가격 간의 관계를 분석하거나 부동산 가격 예측을 시도하는 연구가 주를 이

루어 왔다. 특히 2000년대 초반까지의 부동산 가격 지수 예측에 관한 연구에서는 주로 ARIMA, VAR, VECM 등 선형 기반의 회귀모형을 이용하여 부동산 가격의 예측을 시도하였다(손경환, 1991; 김근용, 1998; 김갑성・서승환, 1999; 윤주현・김혜승, 2000).

박헌주와 박철(2001)은 1987년부터 2000년까지 거시경제 변수와 지가 변동률을 활용하여 VAR과 ARIMA 모형을 통해 2001년 지가 변동률의 예측을 시도하였다. VAR 모형의 설명변수로는 그랜저 인과관계(Granger Casuality) 검정 결과, 지가 변동률에 있어 인과관계가 가장 높은 실질 GDP 성장률, 회사채 수익률 등을 활용하였다. 이를 통해 2001년의 지역 별, 분기별 토지 시장을 전망하였다. 분석 결과, 국제유가의 상승, 반도체 가격의 하락, 선진국의 경기후퇴 등 외부환경에 의한 경기 침체, 정책적 요인, 토지의 수요자나 공급자의 주관적인 기대 등 토지 가격 하락에 영향을 주는 다양한 상황이 회복될 경우 전국 토지 시장 가격은 점진적으로 상승할 것으로 예측하였다.

손정식 등(2003)은 주택매매가격 변동률과 전세 가격 변동률을 예측하기 위해 ARIMA 모형과 VAR 모형을 사용하였다. 주택매매가격 예측을 위한 VAR 모형에 활용된 설명변수는 그랜저 인과관계 검정을 통해 회사채수익률, 종합주가지수, 실질 GDP 성장률, 주택매매가격을 내생 변수로 선정하고 달러환율과 토지가격을 외생변수로 활용하였다. 동일 기간을 대상으로 분석한 결과 VAR의 예측 정확성이 비교적 높은 것으로 나타났다. 또한, 토지 가격이나 전세가격과 달리 주택매매가격에 있어서 실질 GDP 성장률과 같은 기본 요인의 계수 보다 주가 변동률과 같은 자산선택행위의 계수가 높게 나온 것을 확인하였다.

2000년대 중반부터 기존에 회귀분석에 편중된 연구 방법들에서 벗어나 기계학습 기법을 포함하여 다양한 기법을 적용한 부동산 가격 예측 연구들이 시도되었다. 남영우와 이정민(2006)은 서울시 아파트 분양가격에 영향을 미치는 거시경제변수를 활용한 모형을 설정 후 다중 회귀분석 모형과 인공신경망 모델을 활용하여 분양 가격의 예측을 시도하였다. 예측모델의 설명변수로는 소비자물가지수, 총 통화량, 국민총생산, 국제원유도입단가, 환율, 건축원자재도입단가, 지가변동률지수 등 거시경제변수들과 지역특성변수인 지역 더미변수, 평형을 적용하였다. 분석 결과 MSE 기준으로 다중 회귀분석의 경우 7.34, 인공신경망의 경우 0.77로 나타나 다중 회귀분석 모형보다 인공신경망 모형의 예측력이 훨씬 우수한 것으로 나타났다.

정원구와 이상엽(2007)은 공동주택가격지수 예측을 위해 거시경제변수와 공동주택가격지수 등 총 60개의 입력변수를 이용하여 2개의 은닉층으로 구성된 심층신경망을 이용하여 공동주택 가격지수 예측을 시도하였다. 분석 결과 금융자산 관련 분야에서 입증된 인공신경망 모델이 부동산 시장에서도 안정적인 예측력을 나타내며 부동산 투자자의 위험 부담을 줄이고 정부의 정책 수립에 있어 도움을 주는 것으로 판단되었다. 또한, 부동산 정책변수 발표 전후 오차가 커지는 것을 통해 정책 관련 변수의 추가 필요성을 제기하였다.

임성식(2014)은 가격이 관측 값의 변동이 일어나는 개입 효과를 가진 시계열자료의 분석에 적합한 예측모형을 찾기 위해 아파트 규모별

주택가격지수 데이터를 활용하여 모형별 예측력 정도를 비교하였다. VAR 모형의 예측력보다 ARIMA 모형의 예측력이 더 높은 것으로 분석되었고, 시계열 분석 자료에 개입효과에 대한 정보를 추가하는 개입분석모형의 예측 정확성이 더 높은 것을 확인하였다.

배성완과 유정석(2018)의 연구에서는 부동산 가격지수 예측에 기계학습 기반 예측모델의 적용 가능성을 확인하기 위하여 서울특별시의 아파트실거래가격지수를 6가지 전통적인 시계열분석 모형과 5가지 기계학습 방법의 예측력을 각각 단변량 변수와 다변량 변수를 활용하여 비교 분석하였다. 안정적인 시장 상황의 경우 시계열 분석 모형과 기계학습 방법 모두 유의미한 추세 예측을 하였다. 그러나 시장 상황이 급변하는 경우 기계학습을 적용한 모형은 유사하게 가격지수를 예측하는 반면, 시계열분석 모형은 시장 추세를 전혀 예측하지 못하는 것을 확인하였다. 또한, 기계학습을 적용한 모형 중에서 LSTM 모형의 예측력이 가장 우수하게 나왔지만, 기계학습 모형들의 결과 값이 산출되는 근거를 확인할 수 없는 한계를 제시하였다. <표 1>은 국내의 부동산 가격지수 예측모델에 관련된 선행연구의 연구 대상, 연구 모형, 활용된 변수 및 모델의 성능에 대해 요약하였다. 국외 연구에서도 부동산 관련 지수 예측에 관련된 연구가 활발히 진행되어 왔다. Raymond(1997)의 연구에서 ARIMA 모형을 활용하여 홍콩 지역의 산업 관련 부동산 가격지수에 대한 예측을 시도했으며, 부동산 가격 예측 분야에 있어 기술적인 방법론의 활용 가능성을 제시하였다. 결과의 해석에 있어 정성적인 분석이 필요할 수 있지만, 기술적인 방법론을 통해 전체적인 관계를 파악할 수 있으며 경제 이론에 근거를 제시할 수 있다. 그러나 이전의 목표변수 값만을 활용하는 단변량 시계열 예측모형이라는 점에서 결과 해석에 있어 다른 모델과 결합할 경우 더 나은 예측을 제공할 수 있을 것으로 판단하였다.

Zhou and Sornette(2008)는 1983년부터 2005년까지 미국 라스베이거스 주의 27개 구역의 주택 가격지수 예측 및 부동산 거품 존재 유무를 확인하기 위해 주택 가격지수 성장률을 예측하는 연구를 진행하였다. 분석 방법으로는

<표 1> 국내의 부동산 가격지수 예측모델 선행 연구

선행 연구	연구 대상	연구 모형	활용 변수	비고
남영우와 이정민 (2006)	서울시 아파트 분양 가격	회귀분석, 인공신경망	거시경제지표, 지역 더미변수, 평형	인공신경망 모형 우수
정원구와 이상엽 (2007)	공동주택가격지수	은닉층 2개의 심층 신경망	거시경제지표 및 공동주택 가격지수 등 60개의 변수	정책변수의 추가 필요성 제시
임성식(2014)	아파트 규모별 주택 가격지수	VAR, ARIMA, 개입 분석 모형		개입분석모형, ARIMA, VAR 순으로 성능 우수
배성완과 유정석 (2018)	서울특별시 아파트 실거래가격지수	6가지 시계열 분석 모형과 5가지 기계 학습모형	거시경제지표 5개 변수 및 아파트실거래가격지수	LSTM 모형의 예측력이 가장 우수

로그주기 멱함수법칙(log-periodic power-law)과 연내 구조 모델(intra-year structure)을 통해 주택가격지수 성장률을 예측하였으며, 가격이 급격히 증가하는 부동산 거품의 존재를 확인하였다. 월별 예측을 통해 연도별 예측에 비해 실시간으로 대처할 수 있는 모델을 제시하고 지역 단위를 세분화하여 지역 별 특징을 확인하였다.

Wu and Brynjolfsson(2014)은 구글 트렌드의 주택검색지수(House Searching Index, HSI)를 활용한 다중회귀분석을 통해 미국 51개 주의 주택가격지수 및 거래량의 예측을 시도하였다. 분석 결과, HSI를 활용한 가격 예측모형이 전문가들의 미래 가격 예측보다 23.6% 나은 정확성을 보이는 것을 확인하며 간단한 회귀 모델에서도 HSI가 주요한 변수임을 검증하였다. 또한, 현대 사회에서 소비자들의 구매 행태가 변화하고 있으며, 데이터가 산출되기까지 시간이 소요되는 정부 데이터에 비교하여 실시간으로 수집 및 적용 가능한 새로운 초미시경제 데이터의 활용 가능성을 제시하였다.

Li and Chu(2017)는 대만 타이페이의 주택가격지수인 Cathay 주택가격지수와 Sinyi 주택가격지수를 예측하기 위해 GDP, 통화량, 경제성장률, 소비자물가지수 등 11개의 거시경제지표를 활용하여, 역전파 신경망(Back Propagation Neural Network, BPN)과 방사기저함수 신경망(Radial Basis Function Neural Network, RBF)의 두 가지 인공신경망 예측모델을 비교 분석하였다. 분석 결과 RBF 모델과 BPN 모델이 두 가지 주택가격지수를 예측함에 있어 성능 면에서 유의미한 차이가 없음을 확인했다.

부동산 가격지수 예측에 관련된 선행연구에서 전통적인 선형 기반의 시계열 예측모형인 ARIMA, VAR, VECM 등을 사용한 분석이 주를 이루고 있으며, 최근 SVM, 랜덤 포레스트, 인공신경망, 심층신경망, 순환신경망 등 기계학습 알고리즘을 활용한 부동산 가격지수 예측에 관한 연구가 활발히 진행되고 있다. 전반적으로 선형 기반의 예측모형에 비해 기계학습을 이용한 예측모형의 예측 정확성이 우수하였으며, 부동산 가격지수를 예측하기 위한 설명변수로 금리, 주가지수, 환율, 국내총생산, 경제성장률, 소비자 물가지수, 전세 가격지수, 통화량 등의 거시경제지표가 주를 이루는 것으로 나타났다. 또한, 현대 사회의 정보 습득이나 구매 형태의 변화에 따라 부동산 가격 예측 과정에 온라인 검색지수와 같은 새로운 지표의 적용 필요성을 확인할 수 있다.

본 연구에서는 거시경제지표, 부동산 관련 지표, 이전 아파트실거래가격지수 및 구글 검색지수를 활용하여 아파트실거래가격지수를 예측하고자 한다. 이를 위하여 학습데이터 셋에서 목표변수를 포함하고 예측모델을 설계하는 지도학습방식의 기계학습 기법 중 ARIMA, 랜덤 포레스트, LSTM을 선택하여 예측 정확도를 비교하였다.

2.3 LIME(Local Interpretable Model-agnostic Explanation)

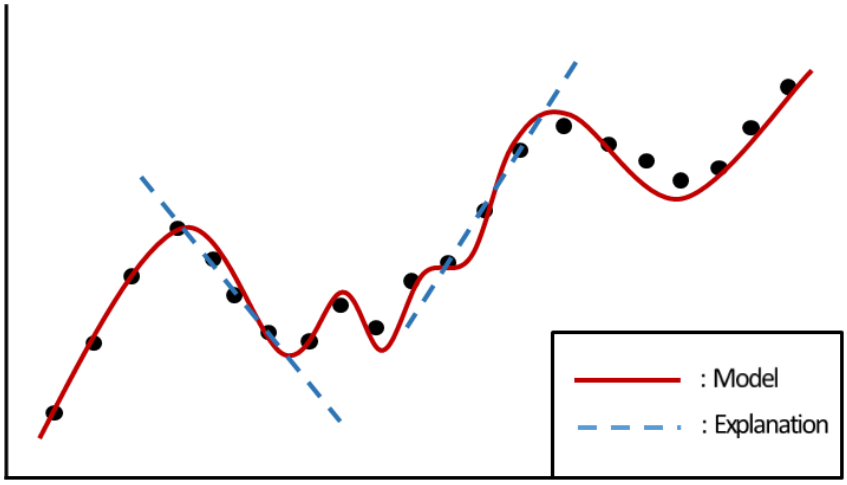
기계학습에는 의사결정나무(decision tree) 모델과 같이 결과를 도출하는 과정을 확인할 수 있는 모델이 있는 반면, 심층신경망, 순환신경망, 랜덤 포레스트 등과 같은 대부분의 모델

에서는 결과의 도출 과정을 알 수 없는 소위 '블랙 박스' 모델이라는 지적을 받는다. 기계학습 모델의 해석력을 강화시키기 위한 방법으로 기계학습 모델을 통해 도출 된 결과를 의사결정나무, 라소 모델이나 회귀분석과 같이 설명력이 높은 모델에 결합하는 대리 모형(surrogate model), 그리고 설명하고자 하는 예측 값 주변에 지역적으로 대리 모형을 국한 시킨 후 변수를 조절해가며 모델의 민감도를 분석할 수 있는 LIME 알고리즘 등이 있다(Audet et al., 2000; Ribeiro et al., 2016).

LIME 알고리즘의 목표는 기계학습 모델이 특정한 예측을 어떻게 만들어냈는지 해석하는데 있다. LIME 알고리즘은 기존에 학습된 기계학습 모델의 학습데이터를 변형하여, 변형된 데이터와 기계학습 모델의 예측 값으로 이루어진 새로운 데이터 셋을 생성한다. 이후 관측하고자 하는 관측 값과 샘플링 된 관측 값들의 근접성을 기반으로 가중치를 매기고, LASSO나 의사결정나무와 같은 해석 가능한 모델로 새로운 예측을 진행한다. 이렇게 구해진 설명 모델을 해석하여 변수들의 상대적인 중요성 및 예측결과에 미치는 영향 등 기계학습 모델에서 결과가 도출되는 과정에 대한 해석을 시도할 수 있다. 관측하고자 하는 인스턴스 x에 대한 설명 모델인 explanation(x)은 다음과 같이 나타낼 수 있다.

$$Explanation(x) = \arg\min L(f, g, \pi_x) + \Omega(g)$$

설명 모델 g는 모델 복잡도 Ω(g)를 낮게 유지하면서 설명 모델 g의 예측이 원래 모델 f의 예측과 유사한지를 측정하는 손실함수 L을 최소화하는 모델이다. G는 가능한 설명 모델의 집합이며, 근접성 측도 π_x는 인스턴스 x 주변의 범위가 얼마나 큰지 정의한다. LIME알고리즘은 손실함수 L의 최적화를 하는 역할을 하며 설명 모델이 사용할 최대 변수 수와 같은 모델의 복잡도는 분석가가 결정해야 한다. <그림 1>에서는 시계열 회귀 모형에서 LIME 알고리즘이 적용되는 모습을 시각화하였다.



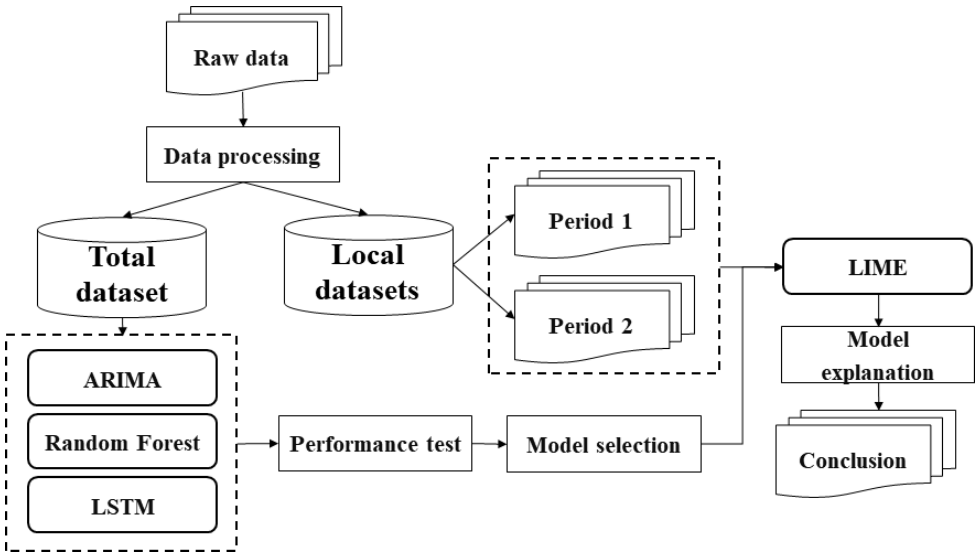
<그림 1> 회귀 모형에 대한 LIME 분석

Ⅲ. 연구 방법

3.1 연구 모형

본 연구의 지역 별 아파트실거래가격지수 예측모델 비교를 위한 전체적인 연구모형은 <그림 2>와 같다. 예측모델의 설명변수로는 이론적 배경을 기반으로 기존 연구에서 실증적으로 분석한 변수인 거시경제지표, 부동산 관련 지표와 함께 미시경제지표인 구글 트렌드의 관련 검색어 검색 지수를 활용하였다. 구글 트렌드의 관련어 검색 지수는 초미시 경제지표로 활용되며 다양한 분야의 시장 분석, 혹은 가격 예측에 활용될 수 있다(Choi and Varian, 2012; Wu and Brynjolfsson, 2015).

본 연구에서는 각 지역적 특성을 포함시키기 위하여 각 도시 명을 검색어로 하여 부동산 관련 검색 지수를 설명변수로 활용하였다. 대표적인 통계 및 기계학습 프로그래밍 언어인 R 3.6.1을 통해 원본 데이터에서 결측치 및 이상치를 제외하고 분석모델에 적합한 데이터 구조로 변형하는 데이터 전처리 과정을 수행하였다. 또한, 다양한 기계학습 라이브러리 및 패키지를 지원하는 프로그래밍 언어인 파이선(Python) 3.6.3의 사이킷런(Scikit-learn), 케라스(Keras) 패키지를 통해 전국 단위 아파트실거래가격지수에 대한 ARIMA, 랜덤 포레스트, LSTM을 적용한 실거래가격지수 예측모델을 만들고, 각 모델 별로 RMSE 및 MAPE 값을 비교하여 최적의 예측모델을 선정한다. 최종 예측모델이 선정되면 서울시, 부산시, 인천시의 아파트실거래가격지수에 대하여 변곡점과 개입 효과가 일어나는 부분을 분리하여 각 시점별로 모델을 학습시켜 예측 결과를 확인한다. 마지막으로 학습된 모델은 LIME 분석을 통해 가격 지수의 증감에 영향을 미치는 설명변수의 중요도 및 효과를 해석하고 지역 별 특성을 도출하며 최종적으로 비교 분석을 진행한다.



<그림 2> 연구 모형

3.2 데이터 수집

본 연구에서는 각 지역 별 아파트실거래가격지수를 예측하기 위해 8가지 설명변수를 적용하였으며 설명변수에 대한 상세 설명은 <표 2>와 같다. 주택 시장을 예측하기 위한 거시경제지표로 선행 연구에서 주로 사용되는 변수는 금리, 주가지수, 환율, 국내총생산, 경제성장률, 소비자 물가지수, 전세 가격지수, 통화량 등이 있다(송인호, 2015; 배성완・유정석, 2017). GNP나 GDP는 선행 연구에서 설명변수로 중요하게 여겨진 거시경제지표였으나, 월별 데이터를 통해 예측하는 현재 연구의 특성상 분기별로 제공되는 데이터는 예측모델 생성 과정에 적합하지 못하여 설명변수에서 제외하였다. 시계열 예측을 하는 과정에서 변하지 않는 값은 다음 값을 예측하는 모델의 성능을 저하시킬 수 있다. 모든 설명변수는 2006년 1월부터 2019년 5월까지의 데이터를 각 통계청, 한국감정원 부동산통계정보, 구글 트렌드에서 수집하였다.

<표 2> 설명변수 및 출처

구분	설명변수	데이터 출처
거시경제지표	지역 별 소비자물가지수	통계청(http://kosis.kr)
	정기예금금리	
	통화량(M2)	
부동산 관련 지표	지역 별 아파트실거래가격지수	한국감정원 부동산통계정보 (http://r-one.co.kr)
	지역 별 지가지수	
	지역 별 아파트매매현황	
	주택담보대출금리	
미시경제지표	지역 별 부동산 관련 검색 지수	구글 트렌드 (https://trends.google.co.kr)

<표 3> 전국 데이터 기초 통계량

	최소값	최대값	평균	중앙값	표준편차
전국 소비자물가지수	79.31	105.65	94.57	97.00	7.75
정기예금금리	1.30	6.28	3.06	2.88	1.31
통화량(M2)	1027697	2771633	1860440	1819290	485641.44
전국 아파트실거래가격지수	59.50	102.60	84.48	83.60	11.12
전국 지가지수	81.39	110.55	93.97	92.33	6.65
전국 아파트매매현황	12300	117812	52264	51406	15128.65
주택담보대출금리	2.66	7.580	4.495	4.290	1.30
전국 부동산 관련 검색 지수	16.00	93.00	32.18	26.00	14.60

<표 3>은 전국 아파트실거래가격지수를 예측하기 위해 사용된 각 설명변수들의 기초통계량을 나타낸 것이다. 설명변수들 가운데 정기예금금리 및 주택담보대출금리는 하락하는 추세를 나타내며 소비자 물가지수, 통화량, 지가지수는 2006년 이후로 비교적 꾸준한 상승 곡선을 그리는 것을 확인하였다. 아파트 매매 현황과 구글 트렌드의 지수는 뚜렷한 상향이나 하향 곡선을 그리고 있진 않는 것으로 나타났다. 종속변수인 아파트실거래가격지수의 경우 꾸준한 상승 곡선을 그리고 있기 때문에 간단한 추세 분석은 단순 회귀를 활용해도 충분하다. 그러나 본 연구에서는 지수의 등락을 포함한 정밀한 예측모델을 찾고자 기계학습 기반의 알고리즘을 활용하여 아파트실거래가격지수의 예측을 시도하였다.

3.3. 분석 모델 선정

최종 분석 모델의 선정에 앞서 각 모델의 예측 성능 평가를 위해 각 변수들을 0에서 1까지 정규화한 후 전체 시계열의 70%에 해당하는 2015년 4월을 기준으로 학습을 진행하는 학습 데이터 셋 109건과 테스트 데이터 셋 48건을 구분하였다. 기준 시점 이전의 학습데이터 셋으로 학습된 각 모델 사이의 평가는 학습에 사용되지 않은 테스트 데이터 셋에 대한 모델의 예측 값과 실제 값 사이의 평균제곱근오차(RMSE)와 절대평균백분율오차(MAPE)의 평균을 산출하여 예측 성능을 비교한다. 시계열 예측모델의 특성과 정밀한 예측을 위하여 본 연구에서는 이전 4개월의 실제 값을 반영하여 학습 및 평가를 진행하였다.

3.3.1. ARIMA(AutoRegressive Integrated Moving Average model)

시점을 고려하지 않는 일반적인 회귀분석과 달리 시계열 예측은 예측변수에 대하여 과거 관측 값들을 수집 및 분석하여 숨겨진 관계를 설명하고 미래의 값을 예측할 수 있는 모델을 개발하는 과학 분야이다. 선형 회귀 기반의 단변량 시계열 예측모형인 ARIMA는 과거의 관측 값들과 오차를 통해 미래를 예측하는 방법론인 자기회귀이동평균 모형(ARMA)을 일반화하여 자기회귀모형(AR)과 이동평균모형(MA)을 결합한 Box-Jenkins의 알고리즘에 과거 관측치의 추세를 포함시켜 일반화한 모델이다(Box and Jenkins, 970; Zhang, 2003). ARIMA 모형의 시계열 생성 모델은 p개의 과거 관측 값과 q개의 오차항들의 선형결합 형태인 다음의 식과 같다:

$$y_t = \theta_0 + \sum_{i=1}^{p} \phi_i y_{t-i} - \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

ϵ_t는 평균이 0이고 완전 독립적인 랜덤 분포를 따를 때, y_t와 ϵ_t는 각각 t 시점에서의 종속변수의 값과 랜덤 오차이다. p와 q는 ARIMA 모형의 매개변수로 q가 0이면 p의 순서를 따르는 AR모형이 되고, 반대로 p가 0이면 q의 순서를 따르는 MA모형이 된다. d는 일반 차분 회수이다. 그러므로 ARIMA 모형에서는 적정 p, d, 그리고 q의 값을 발견해야 이상적인 결과를 얻을 수 있다.

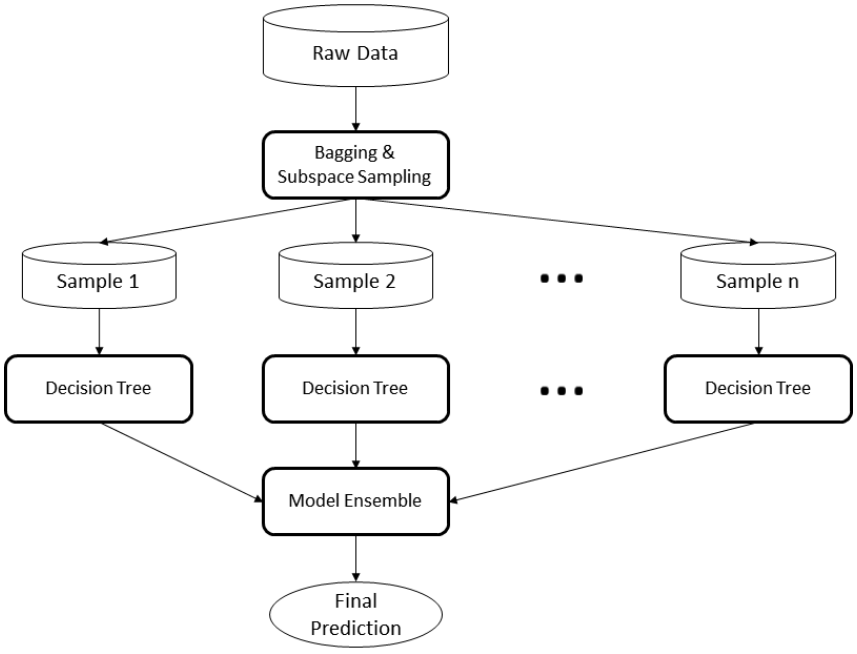
본 연구에서는 전국 아파트실거래가격지수를 예측하기 위한 최적의 매개변수 p와 q를 발견하기 위하여 아카이케의 정보 기준(Akaike's information criterion, AIC)와 베이지안 정보

기준(Bayesian information criterion, BIC)을 적용하였다. 1차 차분 후 각각의 파라미터를 0에서 3까지 순차적으로 검증하여 AIC와 BIC를 각각 -643.508과 -629.961로 최소화 한 p=2, q=1인 ARIMA(2,1,1)모델을 최종 모형으로 선정하였다.

3.3.2. 랜덤 포레스트

랜덤 포레스트(Random Forest)는 브레이먼(Breiman)이 2001년 고안한 예측 분석 방법론으로 배깅(bagging, 부트스트랩 결합), 아공간 샘플링(subspace sampling) 및 복수 개의 의사결정나무를 결합하여 분류 및 회귀 문제에 대한 예측모델을 제시하는 기계학습 기반의 앙상블 형이다(Breiman, 2001; Kelleher et al., 2015).

랜덤 포레스트 알고리즘은 부트스트랩 결합과 아공간 샘플링을 통해 관측치와 입력변수가 무작위로 추출된 n개의 샘플에 대하여 각각 의사결정나무를 학습시켜 완성된 n개의 의사결정나무의 예측 값을 통합하는 과정을 거친다. 부트스트랩 결합은 원본 데이터 셋에서 랜덤 샘플링으로 추출한 여러 부트스트랩 샘플을 생성하여 모델링한 후 결합하여 최종 예측모형을 만드는 알고리즘이다. 부트스트랩 결합에서 모델링 결과를 결합할 때 목표변수가 연속형 변수이면 각 예측 결과의 평균, 범주형 변수이면 투표 방식을 이용하여 최종 예측 결과를 출력한다. 부트스트랩 결합을 이용하면 원본 데이터 셋으로부터 여러 번의 샘플링을 통해 예측모형의 분산을 줄여 모형의 변동성을 감소시킬 수 있다. 랜덤 포레스트 모델의 구성은 <그림 3>과 같다.

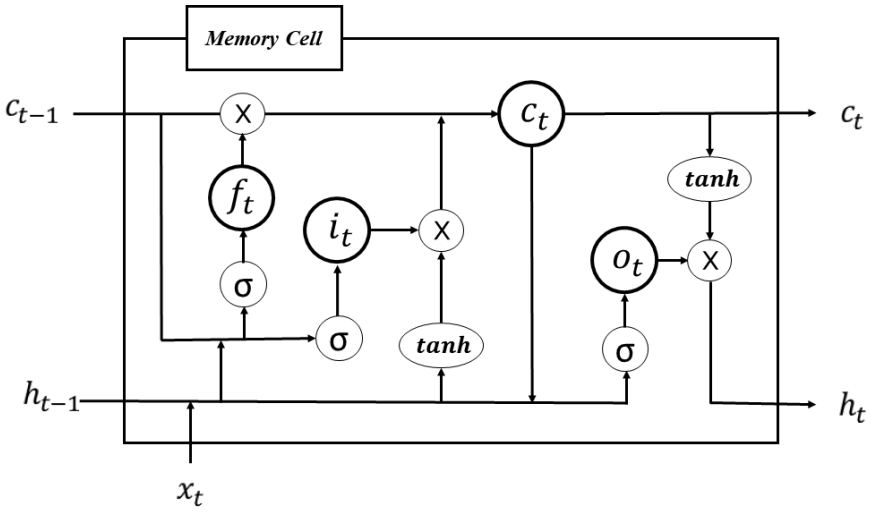


<그림 3> 랜덤 포레스트 모형

랜덤 포레스트를 이용한 회귀모형을 학습하기 위해서 본 연구에서는 각 노드의 샘플링 과정에서 모든 변수를 고려하였으며, 의사결정나무의 개수를 50에서 200까지 변화시켜가며 실험하였다. 테스트 데이터와 예측 값 사이의 RMSE와 MAPE가 최소화 되는 150개의 의사결정 나무를 가진 랜덤 포레스트 모델을 선정하여 분석에 사용하였다.

3.3.3. LSTM(Long Short-Term Memory networks)

LSTM은 시퀀스 생성이나 시계열 형태의 데이터의 예측을 위한 RNN 모델 중 하나로, Hocheriter and Schmidhuber(1997)에 의해 처음 고안되었고, 이후 Gers et al.(2000), Graves and Schmidhuber(2005) 등에 의해 추가 연구되었다(Fischer and Krauss, 2018). LSTM은 은닉상태(hidden state)를 포함시켜 장기 의존성을 학습하도록 설계된 순환신경망 모델이며, 기존 순환신경망의 한계인 학습이 진행될수록 기울기가 급격히 커지거나 사라지는 한계를 메모리 셀 내부의 입력게이트(input gate)와 망각게이트(forget gate)를 통해 극복한 모델이다. <그림 4>는 LSTM의 구조도를 나타낸 것이며 구성함수는 다음과 같다.



$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o)$$

$$h_t = o_t \tanh(c_t)$$

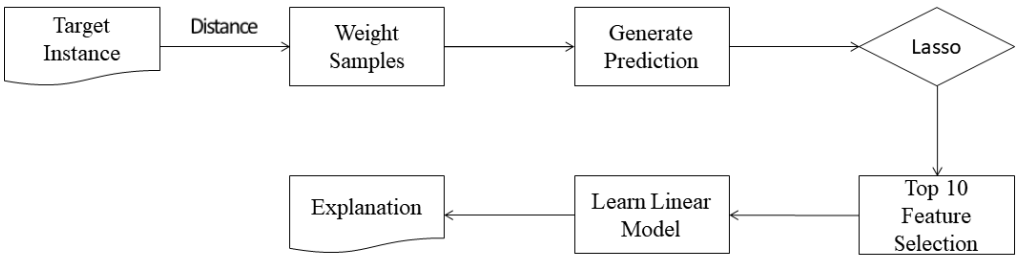
<그림 4> Long Short-term Memory Cell

위 수식에서 σ는 활성화 함수인 시그모이드이고, i, f, o, c, h는 각각 input gate, forget gate, output gate, cell, hidden state를 의미하며, 메모리 셀의 주요 입출력 벡터들의 크기는 h_t의 크기와 같다. W는 가중치 행렬을 의미하며 편향은 b로 나타내었다. LSTM 학습 과정에서 메모리 셀의 도함수의 크기를 폭발적으로 커지는 것을 방지하기 위해 하이퍼볼릭탄젠트 함수(hyperbolic tangent function)를 활성화 함수로 활용하여 hidden state를 계산한다. LSTM은 텍스트 생성과 스텝별 시계열 분석에 적합한 순환신경망 모델로, 본 연구에서는 1개의 은닉 LSTM 층에 8개의 입력변수를 4개월씩 30번 학습하며 Hidden State를 업데이트하는 모델로 설계하였다. 매개변수는 은닉층에 16,650개와 출력층에 61개를 포함해 총 16,621개의 매개변수가 활용되었다. 활성화 함수는 reLu 함수를 사용하였고, Adam(adaptive moment estimation) optimizer를 이용하여 모델의 최적화를 진행하였다. 은닉층의 노드 수를 조정하며 RMSE 값이 최소로 나오는 60개 노드를 가진 LSTM 모델을 최적의 모델로 선정하였다.

3.4. 지역 별 예측모델 비교 분석

지역 별 예측모델 비교 분석에 앞서 전국 단위 예측모델 성능 비교를 통해 최종 선정된 예측모형을 기반으로 서울, 부산, 인천 3개 도시의 아파트실거래가격지수를 학습시킨다. 이때 각 예측모델은 목표변수인 도시 별 아파트실거래가격지수 데이터를 대상으로 도시 별 설명변수를 입력변수로 활용한다. 각 도시 별 예측모델이 학습이 끝나면 LIME 알고리즘을 통해 각 지역 별 예측모델에서 나타나는 특징을 비교 분석한다.

각 도시 별, 지역 별 부동산 가격의 상승과 하락에 영향을 주는 중요 설명변수를 찾기 위해 지역 별로 아파트실거래가격지수 예측 값이 가장 크게 증가하는 지점과 감소하는 지점에 대하여 해당 지점과의 거리를 기반으로 가중치를 가지는 데이터 셋을 구성하여 지역적인 특색을 가진 예측 값을 새로 구한다. 예측모델에서 해당 지점의 주요 변수를 결정하기 위해서 라쏘 모델(least absolute shrinkage and selection operator, Lasso)을 활용한다. 4개 시점의 8개 설명변수, 총 32개의 입력변수 중에서 각 지점의 예측에 작용하는 상위 10개의 주요 변수들은 라쏘 모델 적합을 통해 선정한다. 선정된 변수들을 기준으로 선형 기반의 지역적 대리모형을 적용하여 각 변수가 모델의 결과에 미치는 영향을 설명하고, 각 결과에 대해 3개 도시 별 비교 분석을 진행하여 아파트실거래가격지수에 영향을 미치는 지역 별 특징을 확인한다.

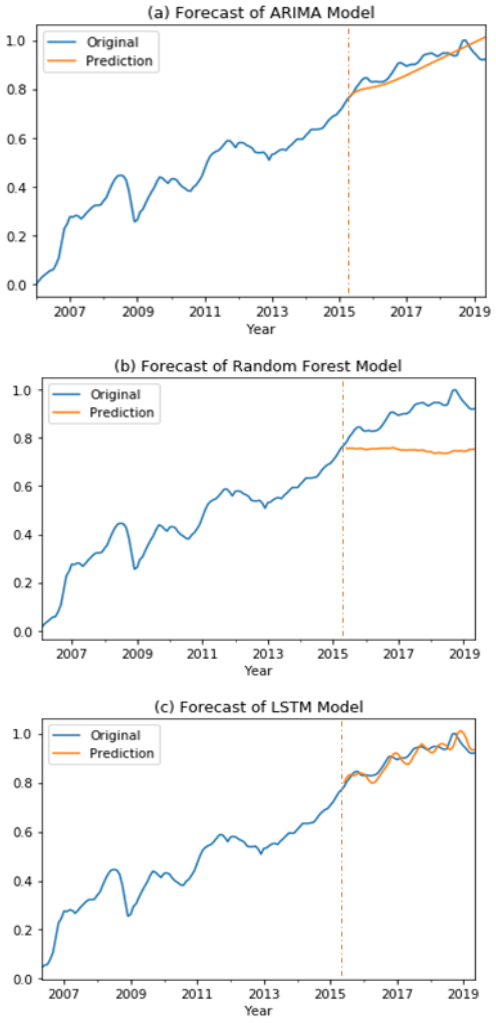


<그림 5> LIME 분석 흐름도

Ⅳ. 분석 결과

4.1 분석모델 선정 결과

전국 모델을 통해 아파트실거래가격지수의 예측에 적합한 모델을 선정하기 위해 ARIMA, 랜덤 포레스트, LSTM 모델을 각각 학습한 후



<그림 6> 전국 아파트실거래가격지수에 대한 (a) ARIMA, (b) 랜덤 포레스트, (c) LSTM의 예측 결과

<표 4> 시계열 예측모형 비교

	RMSE	MAPE	평균
ARIMA	0.03768	0.03481	0.03625
랜덤 포레스트	0.10356	0.10191	0.10274
LSTM	0.02376	0.02207	0.02292

테스트 데이터 셋을 활용해 각 모델의 RMSE와 MAPE를 계산하였다. <그림 6>은 실제 예측을 한 결과를 그래프로 나타낸 것이며, <표 4>는 각 모델들의 예측에 대한 RMSE, MAPE 및 오차율의 평균을 나타낸 것이다.

<표 4>에서 확인할 수 있듯이 예측 성능이 가장 좋은 모델은 RMSE와 MAPE가 각 0.023763, 0.022069으로 실제 값에 대한 예측 값의 전반적인 오차가 가장 낮은 LSTM으로 분석되었다. ARIMA는 전반적인 추세를 반영하여 향후 예측 값의 흐름을 읽는데 유의미한 결과를 나타내었지만 세부적인 정확한 예측 값을 기대하기 힘든 한계를 가지고 있다(배성완 · 유정석, 2018). 랜덤 포레스트의 경우 트리 기반의 지도학습 모델이기 때문에 이미 학습이 된 범위 내의 설명변수와 종속변수에 대해서는 높은 정확성을 가지지만 본 연구와 같이 지속적으로 증가하거나 감소하여 학습이 되지 못한 설명변수나 종속변수가 포함되거나 주를 이룰 경우 유의미한 예측을 할 수 없는 것으로 나타났다. 반면, 시계열을 적용한 LSTM 모델의 경우 새로운 데이터의 증감분에 대해서도 비교적 안정적인 예측을 하였다.

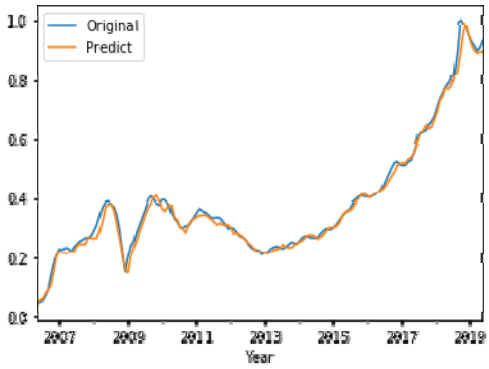
지역 별 아파트실거래가격지수를 위한 예측 모델로 최종 선정된 LSTM 모델을 활용하여 3개 도시의 데이터 별로 학습된 모형을 바탕으로 지역 별 비교 분석을 진행하였다. 분석에 활

용된 설명변수는 전국 아파트실거래가격지수의 예측과 마찬가지로 4개월 분량의 거시경제지표, 부동산 관련 지표, 구글 트렌드의 검색지수 8가지 데이터를 통합한 총 32개를 활용하였다. 지역 별로 차이가 없는 정기예금금리, 통화량(M2), 주택담보대출금리는 모든 도시에 동일하게 입력변수로 사용하였고, 소비자 물가지수, 아파트실거래가격지수, 지가지수, 아파트 매매현황, 부동산 관련 검색 지수는 지역 별 데이터를 입력변수로 사용하였다.

지역 별로 오차율의 근소한 차이는 있지만 전반적으로 오차율이 4% 미만으로 나타났다. 이는 선행 연구에서 LSTM을 이용한 아파트실거래가격지수의 예측모델에서 나타난 5% 미만의 오차율에 벗어나지 않는 것을 확인하였으며(배성완 · 유정석, 2018), 특정 지역에서는 더 높은 예측력을 보이는 것을 확인하였다. 이를 통해 각 지역 별 예측모델의 학습이 잘 이루어졌음을 확인할 수 있다.

4.2. 서울의 아파트실거래가격지수

LSTM을 적용한 서울시의 아파트실거래가격지수 예측모델은 RMSE가 0.01893, MAPE가 0.02911로 수치적으로 매우 안정적으로 학습이 된 것을 확인할 수 있다. 최종 학습된 서울시의 예측모델의 예측 결과는 <그림 7>과 같다. 2008년부터 서울시의 아파트 가격이 급락한 원인으로는 2008년 미국 서브프라임 모기지 사태로 인한 세계 금융 위기의 여파에 따른 투자 심리의 위축을 의심해볼 수 있다. 2009년부터 회복세를 보인 서울시의 아파트 가격은 2013년 이후 지속적인 상승세를 나타내고 있다.



<그림 7> 서울 아파트실거래가격지수 예측모델

위 예측모델의 예측 값을 기반으로 서울의 아파트실거래가격지수의 예측 값이 가장 큰 폭으로 증가한 지점(t)은 예측 값이 119.18인 2018년 10월로 전월 대비 4.62 증가하였으며 가장 큰 폭으로 감소한 지점(t')은 예측 값이 71.37인 2008년 11월로 전월 대비 3.43 감소한 것으로 나타났다. <표 5>는 해당 시점들에 대한 상위 10개 변수의 LIME 설명을 나타낸 것이다. 즉, LIME 설명이 0.05가 나온다면 해당 변수가 한 단위 증가할 때마다 예측확률이 5% 증가함을 의미한다.

시계열 분석의 특성상 이전의 실거래가격지수의 영향이 높게 나오지만 그 외에 서울시의 아파트실거래가격지수의 상승과 감소에 영향을 미치는 주요 변수로는 전 월의 아파트 매매현황, 통화량, 주택담보대출금리, 구글 트렌드의 서울시 부동산 관련 검색지수, 정기예금금리가 공통적으로 나타났다.

흥미로운 점은 대체로 비슷한 추세를 보이는 주택담보대출금리와 정기예금금리가 서울시의 부동산 가격의 예측모델에서는 반대로 작용하는 것을 확인할 수 있었다. 주택담보대출금리의 증가는 실거래가격지수의 상승에는 부정적인

<표 5> 서울 아파트실거래가격지수 LIME 분석 결과

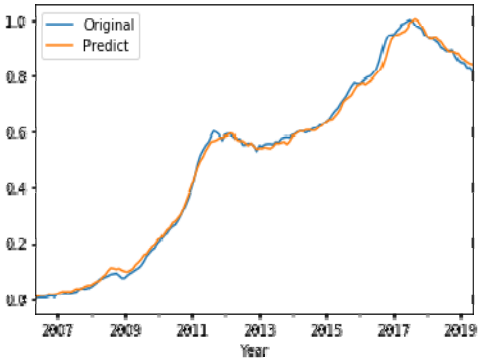
최대 증가 지점(t): 2018년 10월		최대 감소 지점(t'): 2008년 11월	
변수 명	설명	변수 명	설명
아파트실거래가격지수(t-1)	0.09	정기예금금리(t'-1)	-0.03
아파트실거래가격지수(t-2)	0.06	아파트실거래가격지수(t'-1)	-0.03
아파트실거래가격지수(t-3)	0.05	서울 지역 아파트매매현황(t'-1)	-0.02
아파트실거래가격지수(t-4)	0.04	주택담보대출금리(t'-2)	0.02
서울 지역 아파트매매현황(t-1)	0.03	통화량(t'-1)	-0.02
통화량(t-1)	0.02	주택담보대출금리(t'-1)	0.02
주택담보대출금리(t-1)	-0.02	정기예금금리(t'-2)	-0.02
서울 부동산 관련 검색 지수(t-1)	0.02	주택담보대출금리(t'-3)	0.02
통화량(t-2)	0.02	서울 부동산 관련 검색 지수(t'-1)	-0.01
정기예금금리(t-2)	0.02	통화량(t'-2)	-0.01

영향, 하락에는 긍정적인 영향을 나타내며, 정기예금금리는 이와 반대로 작용하는 것으로 분석되었다. 주택담보대출금리가 증가하면, 소유주의 부담감이 증가하여 부동산 시장에 매물이 증가하고 이에 따라 부동산 가격이 감소하는 것이 일반적인 상식이지만, 이와 비슷한 추세를 보이는 정기예금금리가 반대로 작용한다는 것은 상당히 이례적인 결과로 볼 수 있다.

시점별로 보았을 때, 아파트실거래가격지수를 제외하고 4개월 전의 데이터는 아파트 가격 예측에 중요한 설명변수로 활용되지 않았으며, 1개월 전과 2개월 전의 데이터가 예측에 유효한 것으로 나타났다. 서울시 가격이 가장 크게 감소하는 t' 지점에서 아파트실거래가격지수는 전반적으로 상승하므로 가격지수가 급락하는 지점에서 이전의 아파트실거래가격지수가 상승하는 정보는 미래의 아파트실거래가격지수 예측에 큰 영향을 주지 못한 것으로 볼 수 있다. 이에 반해, 전월의 정기예금금리가 큰 영향을 주는 것으로 분석되었으며, 주택담보대출금리는 시점에 상관없이 가격 예측에 중요한 변수로 나타났다. LIME 분석 결과를 통해, 주택담보대출금리가 인하되고 정기예금금리가 상승할 경우, 두 달 안에 아파트실거래가격지수의 하락을 의심해 볼 수 있음을 알 수 있다.

4.3. 부산의 아파트실거래가격지수

부산시의 아파트실거래가격지수의 경우 2009년 이후로 아파트 가격이 큰 폭으로 상승하여 서울시에 비교하여 상대적으로 2008년 세계 금융위기의 영향을 적게 받은 것으로 나타났다. 2012년과 2018년 이후에 하락세가 나타나며 이러한 원인으로 전년도의 급격한 가격 상승, 공급 과잉과 부산시의 인구 감소를 고려해 볼 수 있다. 최종 학습된 부산시의 예측모델에서의 예측 결과는 <그림 8>과 같다.



<그림 8> 부산 아파트실거래가격지수 예측모델

부산시의 경우, 아파트실거래가격지수의 예측 값이 가장 큰 폭으로 증가한 지점(t)과 가장 큰 폭으로 감소한 지점(t')은 각각 2016년 11월과 2012년 5월로 나타났다. t 시점에서의 예측 값은 96.24로 전월 대비 1.96 증가하였으며 t' 시점에서의 예측 값은 77.88로 전월 대비 1.18 감소한 것으로 나타났다. <표 6>은 해당 시점들에 대한 상위 10개 변수의 LIME 설명을 나타낸 것이다.

<표 6> 부산 아파트실거래가격지수 LIME 분석 결과

최대 증가 지점(t): 2016년 11월		최대 감소 지점(t'): 2012년 5월	
변수 명	설명	변수 명	설명
아파트실거래가격지수(t-1)	0.12	아파트실거래가격지수(t'-1)	0.05
아파트실거래가격지수(t-2)	0.10	아파트실거래가격지수(t'-2)	0.04
아파트실거래가격지수(t-3)	0.09	아파트실거래가격지수(t'-3)	0.04
아파트실거래가격지수(t-4)	0.07	아파트실거래가격지수(t'-4)	0.03
통화량(t-1)	0.03	부산 지역 아파트매매현황(t'-3)	-0.02
부산 지가지수(t-1)	0.02	부산 지가지수(t'-1)	-0.01
통화량(t-3)	0.02	부산 지가지수(t'-2)	-0.01
통화량(t-2)	0.02	부산 부동산 관련 검색 지수(t'-2)	0.01
통화량(t-4)	0.02	부산 지역 아파트매매현황(t'-4)	-0.01
부산 지가지수(t-2)	0.02	부산 부동산 관련 검색 지수(t'-3)	0.01

부산시의 아파트실거래가격지수의 예측에 주요한 영향을 미치는 변수는 이전의 아파트실거래가격지수, 통화량, 부산의 지가지수, 아파트 매매 현황, 부산시의 부동산 관련 검색지수로 나타났다. 부산시의 아파트실거래가격지수 상승에 대한 예측에 영향을 미치는 주요 변수는 통화량과 지가지수로 나타났으며, 가격지수가 감소하는 지점에서 아파트 매매현황의 증가와 부산 지역의 지가지수의 상승이 아파트실거래가격지수에 감소에 부정적인 영향을 주는 것으로 확인되었다.

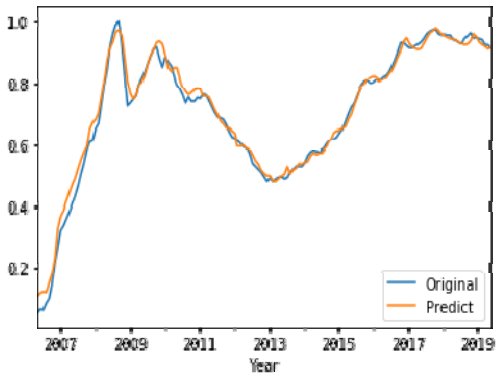
특히, 아파트 매매현황의 경우 3개월과 4개월 전의 데이터가 이전 2개월의 아파트 매매현황보다 예측에 더 영향력이 있는 변수로 나타났다. 실제로 부산의 경우, 매년 1월과 2월에 아파트매매 거래량이 크게 감소하는 특징이 있으

며, 이에 따라 5월이 되면 아파트실거래가격지수가 감소하거나 상대적으로 낮은 상승을 보이는 계절성이 있다.

이를 통해 통화량이 상승하고 있고, 3개월, 4개월 전의 아파트 매매 현황이 증가하며 지가지수가 최근 상승하고 있다면 부산시의 아파트실거래가격지수는 비교적 안전한 투자요건을 갖추고 있다고 고려해 볼 수 있다. 또한, 부산지역의 부동산 관련 검색 지수의 증가는 부동산 가격 결정에 있어 양의 영향을 미치는 것으로 분석되어 부동산 관련 검색 지수의 모니터링을 통해 부동산 가격의 예측을 시도할 수 있음을 확인하였다.

4.4. 인천의 아파트실거래가격지수

인천시의 아파트실거래가격지수의 추세를 살펴보면, 2008년 국제 금융위기 이후 2009년 이후 잠시 회복세를 보이다가 2010년부터 2012년까지 지속적으로 감소하였다. 그 이후에 2013년 이후부터 다시 상승세를 나타내는 것을 확인할 수 있다. 인천의 경우, 2000년대 경제자유구역 조성, 공항 및 항만 건설 등 대규모 개발사업으로 인해 부동산 시장의 호황을 맞이하였으나, 2008년 국제 금융위기 이후 개발 사업이 지연되거나 중단이 되면서 침체기를 겪은 것으로 볼 수 있다. 2014년 이후, 정부의 적극적인 부동산 시장 활성화 정책 및 부동산 경기회복에 따라 주택시장을 중심으로 호조세를 이어가고 있다. <그림 9>는 최종 학습된 인천시의 아파트실거래가격지수 예측모델에서의 예측 결과를 나타낸 것이다.



<그림 9> 인천 아파트실거래가격지수 예측모델

인천의 아파트실거래가격지수의 예측 값이 가장 큰 폭으로 증가한 지점(t)과 가장 큰 폭으로 감소한 지점(t')은 각각 2006년 12월과 2008년 12월로 나타났으며, t 시점에서의 예측 값은 73.31로 전월 대비 4.24 증가하였으며 t' 시점의 경우 예측 값이 93.64으로 전월 대비 3.20 감소한 것으로 나타났다. LIME 분석을 통해 각 시점에 대한 예측 결과에 주요한 영향을 미치는 상위 10개의 변수는 다음 <표 7>과 같다.

인천시의 아파트실거래가격지수의 상승과 감소에 영향을 미치는 주요 변수로는 아파트실거래가격지수를 포함하여 정기예금금리, 소비자 물가지수, 아파트매매현황, 부동산담보대출금리로 나타났다. 인천시의 실거래가격지수는 다른 지역에 서울시와 마찬가지로 정기예금금리의 영향을 많이 받는 것으로 분석되었으며, 소비자 물가지수와 부동산담보대출금리가 아파트 가격 결정의 중요한 변수로 나타났다. 특히, 부동산 가격 감소의 예측에 주요한 변수로 나타난 소비자 물가지수의 경우, 직전 월의 데이터보다 4개월 전의 데이터가 예측 결과에 훨씬 영향이 높은 것으로 나타났다. 이를 통해 시

<표 7> 인천 아파트실거래가격지수 LIME 분석 결과

최대 증가 지점(t): 2006년 12월		최대 감소 지점(t'): 2008년 12월	
변수 명	설명	변수 명	설명
아파트실거래가격지수(t-1)	-0.12	아파트실거래가격지수(t'-2)	0.07
아파트실거래가격지수(t-2)	-0.09	아파트실거래가격지수(t'-3)	0.07
아파트실거래가격지수(t-3)	-0.08	아파트실거래가격지수(t'-4)	0.05
아파트실거래가격지수(t-4)	-0.07	정기예금금리(t'-1)	-0.05
정기예금금리(t-1)	-0.05	소비자물가지수(t'-4)	0.05
소비자물가지수(t-4)	0.05	아파트실거래가격지수(t'-1)	0.04
인천 지역 아파트매매현황(t-1)	0.04	소비자물가지수(t'-3)	0.04
소비자물가지수(t-3)	0.04	소비자물가지수(t'-2)	0.03
정기예금금리(t-2)	-0.03	부동산담보대출금리(t'-1)	0.03
부동산담보대출금리(t-1)	0.03	정기예금금리(t'-2)	-0.03

계열 예측모델의 입력변수 선정에 있어 다양한 시점을 고려해 볼 필요성을 확인하였다.

인천시의 아파트실거래가격지수의 예측에서 전월 대비 가장 높은 증가가 생긴 2006년 12월에 이전의 아파트실거래가격지수가 음의 설명력을 가지는 원인은 예측 결과가 전체 예측 범위에 비교해서 낮은 예측 값을 가지는 것을 생각해 볼 수 있다. 지역적 대리모형인 LIME 알고리즘 특성상 해당 시점에서의 인과관계에 대해서만 고려하게 되는데 매우 낮은 예측 값의 경우 이전에 비해 증가한 영향보다는 낮은 값을 가지게 된 원인을 찾도록 학습이 되었을 가능성이 있다. 또한, 아파트실거래가격지수의 증가와 감소에서 정기예금금리와 부동산담보대출금리가 똑같은 영향을 주었다는 점에서 예측모델 해석에서의 한계점을 고려해 볼 수 있다. 이를 극복하기 위해 예측 값의 증가와 감소에 작용하는 변수의 차이를 명확히 구분할 수 있는 새로운 설명변수 추가를 통해 해석력을 증가시킬 수 있을 것으로 기대된다.

4.5. 지역 별 비교 분석

분석 결과를 바탕으로 지역 별 예측모델에서 나타나는 특징들을 종합해보면 다음 <표 8>과 같다. 아파트 매매현황은 모든 관측 도시에서 아파트실거래가격지수의 변동에 주요 변수로 선정되었으며, 통화량과 정기예금금리, 주택담보대출금리가 그 뒤를 이어 부동산 가격 결정에 가장 중요한 설명변수로 나타났다. 또한, 구글 트렌드 검색지수가 서울, 부산의 부동산 가격 변동의 주요 변수로 나타났으며, 이를 통해 예측하고자 하는 값의 변동 폭이 크거나 추세가 일정하지 않을 경우 구글 트렌드의 검색지수를 설명변수로 활용한다면 예측모델의 예측력을 향상 시킬 수 있을 것으로 기대된다.

<표 8> 지역 별 주요 변수

지역	거시경제지표	부동산 관련 지표	구글 트렌드
서울	통화량, 정기예금금리	아파트매매현황, 주택담보대출금리	O
부산	통화량	아파트매매현황, 지가지수	O
인천	소비자물가지수, 정기예금금리	아파트매매현황, 주택담보대출금리	X

부산의 경우 다른 도시에 비해 해당 도시의 지가지수의 영향을 많이 받는 것으로 나타났으며, 특히 다른 도시에 비해 거시경제지표보다는 부동산 관련 지표나 검색지수와 같은 지역 별 특성이 드러나는 설명변수들의 영향력이 높은 것으로 나타났다.

Ⅴ. 결 론

본 연구는 2006년 1월부터 2019년 5월까지의 표본기간 동안 전국 3개 도시 별 아파트실거래가격지수를 수집하여, 다양한 예측모델 간의 성능 비교 및 예측 결과에 미치는 주요 설명변수를 규명하여 부동산 가격 결정에 있어 지역적 특징을 비교 분석한 연구이다. 이를 위해서 전통적인 시계열 분석 모형인 ARIMA 모형, 기계학습 기반의 랜덤 포레스트, 순환 신경망 기반의 LSTM을 활용한 예측모형 간의 예측 성능을 비교하여 지역 별 비교 분석에 적용할 최종 예측모델을 선정하였다. 선정된 최종 모델을 기반으로 지역 별 예측모델을 적합 후 LIME 알고리즘을 적용하여 지역 별로 예측 값의 급격한 변동이 오는 지점에 영향을 주는 주요 설명변수를 규명하여, 결과에 대한 해석 및 지역 별 비교 분석을 진행하였다. 또한, 구글 트렌드의 검색지수를 설명변수로 활용하여 국내 부동산 연구에 있어 초 미시경제지표의 활용 가능성을 검토하였다.

분석 결과를 요약하면, 우선 아파트실거래가격지수 예측을 위한 세 가지 예측모형 중 LSTM, ARIMA, 랜덤 포레스트 순으로 예측 효과성이 높았다. ARIMA 모형의 경우 전체적인 추세를 예측하기에는 적합한 모형이지만 실제 가격이 어느 정도 오르거나 내릴지에 대한 정확성을 요구하는 예측모델로서는 적합하지 못한 결과를 나타냈다. 반면, LSTM의 경우 전체적인 추세 뿐 아니라 기존에 학습되지 못한 결과에 대해서도 우수한 예측 결과를 가지는 것으로 확인되었다. 기계학습 기반의 랜덤 포레스트의 예측 효과성이 상대적으로 높지 못한 원인으로는 학습 데이터 셋을 기반으로 회귀나무(Regression Tree)를 만드는 알고리즘의 특성상, 지속적으로 증가하거나 감소하여 학습되지 못한 범위 밖의 설명변수나 결과 값에 대한 올바른 학습이 되지 못하는 한계점을 상정해 볼 수 있다.

최종 모델인 LSTM 모형으로 적합한 3개 도시의 아파트실거래가격지수 예측모델을 통해 LIME 알고리즘을 바탕으로 각 지역 별 부동산 가격 변동에 영향을 주는 주요 설명변수 및 영향력을 확인하였다. 지역 별 아파트 매매현황과

통화량, 금리의 경우 전 지역에 걸쳐 주요 변수로 규명되어 부동산 예측모델에서 주요 변수로서의 활용 가능성을 확인하였으며 서울, 부산 지역에서는 구글 트렌드의 지역 별 부동산 관련 검색지수가 주요 설명변수로 확인되었다. 특히, 부산의 경우 다른 도시와 달리 지역 별 지가지수의 영향을 크게 받는 것으로 분석되었다.

본 연구는 부동산 관련 연구뿐만 아니라 기계학습 분석기법을 적용하고자 하는 사회과학 연구 분야에서 예측모델 간 비교와 모델을 해석하는 방법론에 있어 학술적인 시사점이 있다. 첫째로, 부동산 실거래가격지수 예측에 있어 전통적인 시계열 모형과 기계학습 기법 사이의 예측 성능을 비교 및 실제 예측 결과를 시각화하여, 사회과학 분야의 시계열 예측 연구에서 기계학습 기법 활용의 중요성을 확인하였다. 전체적인 추세를 예측에 대해서는 전통적인 시계열 예측모형인 ARIMA 모형도 유의미한 결과를 나타냈지만 정밀한 예측을 위해서는 기계학습 기반의 시계열 예측모형이 더욱 우수한 것으로 나타났다. 이러한 결과는 시계열 예측 관련한 선행연구들의 결과와 일부 일치하는 것이다(배성완 · 유정석, 2018).

둘째로, 시계열 예측모형에서 학습된 범위 밖의 설명변수나 목표변수가 포함되었을 경우 랜덤 포레스트 기반의 예측모형은 이상적인 결과를 가져오지 못한 점을 통해 데이터의 특징에 따라 다른 방법론을 적용할 필요성을 제시하였다. 지속적인 증가를 하거나 감소를 하는 데이터에 대한 예측 시도 시, LSTM을 활용한 예측모형은 이전의 시계열 예측 및 일부 기계학습 모형보다 더욱 정교하고 활용도가 높은 결과를 가져올 수 있을 것으로 기대된다.

셋째, 지역적 대리모형인 LIME 알고리즘을 적용하여 기계학습 기반의 예측모델의 결과에 대한 해석을 시도하였다. 이는 기존 기계학습 방법론을 이용한 선행연구에서 높은 예측 성능에도 불구하고, 지속적으로 한계점으로 지적된 복잡한 기계학습 모델 내부의 인과를 알 수 없는 블랙 박스 이슈에 대한 해결 방안을 제시하는 것이다(배성완 · 유정석, 2017).

마지막으로, 선행연구에서 주로 활용되는 거시경제지표 외에도 구글 트렌드 검색지수의 설명변수로서 활용 가능성을 검토하였다. LIME 분석을 통해 서울, 부산, 인천 등과 같이 예측하고자 하는 아파트실거래가격지수가 일정한 추세를 가지지 않는 경우 구글 트렌드의 검색지수가 주요 변수로서 실제 지수 예측에 도움이 된다는 것을 확인하였다.

본 연구의 실무적인 시사점은 다음과 같다. 본 연구는 지역 별 비교 분석을 통해 부동산 시장 예측모델에 있어 전반적으로 중요한 설명변수 뿐 아니라 지역 별로 중요한 설명변수를 비교하고 지역적인 특징을 포착하였다. 지역 별로 부동산 가격의 증가와 감소에 영향을 주는 주요 변수를 규명하고자 하는 시도는 개인 및 기업의 투자자들이나 정책 결정권자의 의사결정에 도움이 되는 지표를 제공할 수 있으며 더 나아가 부동산 시장에서의 지역적 정보 격차를 여러 방면에서 줄여주는 역할을 기대할 수 있다. 또한, 부동산 시장은 특성상 국제 정세와 같은 외부환경과 정책적인 요인에 영향을 많이 받기 때문에 부동산 시장에 대한 전망을 하기 위해서는 다양한 요인을 고려해야 한다. 부동산 가격 예측에 있어 데이터 기반의 정량적인 분석은 정성적인 방법에 비해 비용과 시간을 절

감할 수 있으며, 전체적인 시장 상황을 이해하여 투자와 정책 수립 등을 최종적으로 결정하는 데 있어 도움이 될 수 있다.

본 연구에서 사용되는 기계학습 방법론의 경우, 투입되는 변수나 매개변수의 값에 따라 다른 결과를 가질 수 있다. 또한, 지역적 대리모형인 LIME 알고리즘은 특성상 특정 시점에서 설명변수의 영향력을 해석하는데 용이하지만 모델 전체에 작용하는 입력변수의 영향을 확인하기에는 한계가 있다. 본 연구의 분석 결과, 회귀모델의 해석 시 이전 시점에 비해 크게 증가하거나 감소한 지점을 고려하지 않고 단순히 큰 값이나 작은 값이 결정되는 원인을 분석하는 것도 LIME 알고리즘의 한계점으로 드러났다.

또한, 본 연구에서는 총 8개의 입력변수를 활용하여 예측모델을 구성하였지만, 부동산 정책을 포함하여 지역 별 특징을 포함할 수 있는 다양한 입력변수를 활용한다면 더욱 유의미한 결과를 얻을 수 있을 것으로 기대된다. 향후 연구의 발전 방향은 다양한 출처에서 부동산 시장 예측에 활용 가능한 데이터나 지역적 특징을 포착할 수 있는 데이터를 포함하여 부동산 시장 가격의 형성에 영향을 미치는 주요 변수들을 새로 규명하고, LIME 분석을 통해 다양한 시점에서 주요 설명변수에 대한 일반화를 시도하고 지역적으로 나타나는 특징을 서로 비교해 보는 것이다.

참고문헌

김근용, “주택 가격 예측을 위한 모형 설정과 검정”, *국토*, 제197권, 1998, pp. 54-61.

남영우, 이정민, “아파트시장 예측을 위한 신경망 분석 적용 가능성에 관한 연구”, *한국건설관리학회*, 제7권, 제2호, 2006, pp. 162-170.

박헌주, 박철, “시계열 모형에 의한 토지 시장의 예측 연구”, *주택 연구*, 제9권, 제1호, 2001, pp. 27-55.

배성완, 유정석, “딥 러닝을 이용한 부동산가격지수 예측”, *부동산 연구*, 제27권, 제3호, 2017, pp. 71-86.

배성완, 유정석, “머신 러닝 방법과 시계열 분석 모형을 이용한 부동산 가격지수 예측”, *주택 연구*, 제26권, 제1호, 2018, pp. 107-133.

서종덕, “데이터 마이닝 기법을 이용한 환율예측: GARCH와 결합된 랜덤 포레스트 모형”, *산업경제연구*, 제29권, 제5호, 2016, pp. 1607-1628.

선일석, “ARIMA 모형을 이용한 보관 및 창고업의 예측 연구”, *기업경영리뷰*, 제6권, 제1호, 2015, pp. 77-91.

손경환, “통화량 변동이 주택가격에 미치는 영향”, *국토 연구*, 제15권, 1991, pp. 35-43.

손정식, 김관영, 김용순, “부동산 가격 예측모형에 관한 연구”, *주택 연구*, 제11권, 제1호, 2003, pp. 49-76.

송인호, “주택 시장과 거시경제의 관계”, *부동산 도시연구*, 제8권, 제1호, 2015, pp. 47-65.

신용상, 구본성, 하준경, 이규복, 송재은, 이윤석, “서브프라임 모기지 사태의 분석과 전망”, 2007, *KIF 금융리포트*.

신이레, 윤상후, “특정 시간대 전력 수요예측 시계열모형”, *한국데이터정보과학회지*, 제27권, 제2호, 2016, pp. 275-284.

윤주현, 김혜승, “주택 시장의 경기 동향 및 단기 전망 연구”, *안양국토연구원*, 2000.

이낙영, 오경주, “디노이징 필터와 LSTM을 활용한 KOSPI200 선물지수 예측”, *한국데이터정보과학회지*, 제30권, 제3호, 2019, pp. 645-654.

임병진, 한성윤, “주식시장 지수와 부동산 시장 지수의 시계열 특성 비교와 관계에 관한 실증적 연구”, *산업경제연구*, 제22권, 제4호, 2009, pp. 2065-2083.

임성식, “주택 가격지수 예측모형에 관한 비교연구”, *한국데이터정보과학회지*, 제25권, 제1호, 2014, pp. 65-76.

정원구, 이상엽, “인공신경망을 이용한 공동 주택 가격 지수 예측에 관한 연구-서울지역을 중심으로”, *주택 연구*, 제15권, 제3호, 2007, pp. 39-64.

정훈, 김주원, “A Machine Learning Approach for Mechanical Motor Fault Diagnosis”, *산업경영시스템학회지*, 제40권, 제1호, 2017, pp. 57-64.

최가영, 이정희, 유리화, “시계열분석을 통한 자연휴양림 계절별 이용수요 예측: 계절 ARIMA 모형과 지수평활 모형을 중심으로”, *관광경영연구*, 제21권, 제3호, 2017, pp. 271-289.

한경수, “부동산 가격영향 요인이 주택매매가격지수에 미치는 영향”, *경영교육연구*, 제26권, 제2호, 2011, pp. 547-565.

한국감정원 부동산통계정보, Retrieved from http://www.r-one.co.kr

한국 통계청, Retrieved from http://www.kosis.kr

Alpaydin, E., “*Introduction to machine learning*”, MIT press, 2009.

Audet, C., Denni, J., Moore, D., Booker, A., and Frank, P., “A surrogate-model-based method for constrained optimization”, *8th Symposium on Multidisciplinary Analysis and Optimization*, 4891, 2000.

Bourassa, S. C., Cantoni, E., and Hoesli, M., “Spatial dependence, housing submarkets, and house price prediction”, *The Journal of Real Estate Finance and Economics*, 35(2), 2007, 143-160.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M., *Time series analysis: Forecasting and control*, John Wiley and Sons, 2015.

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

Choi, H., and Varian, H., “Predicting the present with Google Trends”, *Economic Record*, 88, 2012, 2-9.

Fischer, T., and Krauss, C., “Deep learning with long short-term memory networks for financial market predictions”, *European Journal of Operational Research*, 270(2), 2018, 654-669.

Gers, F. A., Schmidhuber, J., and Cummins, F., “Learning to forget: Continual prediction

with LSTM", *Neural Computation*, 12(10), 2000, 2451-2471.

Graves, A., "Generating sequences with recurrent neural networks", *arXiv preprint arXiv:1308.0850*, 2013.

Graves, A., and Schmidhuber, J., "Framewise phoneme classification with bidirectional LSTM and other neural network architectures", *Neural Networks*, 18(5-6), 2005, 602-610.

Hochreiter, S., and Schmidhuber, J., "Long short-term memory", *Neural Computation*, 9(8), 1997, 1735-1780.

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.

Li, L., and Chu, K.-H., "Prediction of real estate price variation based on economic parameters", *2017 International Conference on Applied System Innovation(ICASI)*, 2017, 87-91.

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

Marsland, S., *Machine learning: An algorithmic perspective*, Chapman and Hall/CRC, 2014.

Negnevitsky, M., *Artificial intelligence: A guide to intelligent systems*, Pearson Education, 2005.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., and Dubourg, V., "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, 12(10), 2011, 2825-2830.

Ribeiro, M. T., Singh, S., and Guestrin, C., Model-agnostic interpretability of machine learning, *arXiv preprint arXiv:1606.05386*, 2016a.

Ribeiro, M. T., Singh, S., and Guestrin, C., "Why should I trust you?: Explaining the predictions of any classifier", *22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016b, 1135-1144.

Ribeiro, M. T., Singh, S., and Guestrin, C., "Anchors: High-precision model-agnostic explanations", *32nd AAAI Conference on Artificial Intelligence*, 2018.

Sutton, R. S., and Barto, A. G., *Reinforcement learning: An introduction*, MIT Press, 2018.

Tse, R. Y., "An application of the ARIMA model to real-estate prices in Hong Kong", *Journal of Property Finance*, 8(2), 1997, 152-163.

Wu, L., and Brynjolfsson, E., "*The future of prediction: How Google searches foreshadow housing prices and sales*", In *Economic analysis of the digital economy*, University of Chicago Press,

2015.

Zhou, W.-X., and Sornette, D., “Analysis of the real estate market in Las Vegas: Bubble, seasonal patterns, and prediction of the CSW indices”, *Physica A: Statistical Mechanics and its Applications*, 387(1), 2008, 243-260.

조 보 근 (Jo, Bo-Geun)

미국 Eastren Michigan University에서 Computer Information System 전공 학사학위를 취득하고, 경북대학교 경영학부에서 석사학위를 취득하였다. 현재 경북대학교에서 경영학부 박사과정 중에 있으며, ㈜빅웨이브에이아이에 이사로 재직 중이다. 주요 연구 분야는 기계학습, 딥러닝, 텍스트 마이닝 등이다.

박 경 배 (Park, Kyung-Bae)

미국 University of Texas at Dallas에서 경영정보학 석사학위를 취득하고, Virginia Tech.에서 박사과정 후, 경북대학교 경영학부에서 박사를 취득하였다. 현재 인텔리콘 법률AI미디어랩 소장으로 재직 중이다. 주요 연구 분야는 토픽모델링, 텍스트마이닝, 자연어처리 등이다.

하 성 호 (Ha, Sung-Ho)

한국과학기술원에서 박사학위를 취득하고 경북대학교 경영학부에 재직 중이다. 국내외 유수 학술지의 편집위원을 역임하였으며, 데이터마이닝, 기계학습, 지능정보시스템에 대한 연구를 진행 중이다.

<Abstract>

Comparative Analysis for Real-Estate Price Index Prediction Models using Machine Learning Algorithms: LIME's Interpretability Evaluation

Jo, Bo-Geun · Park, Kyung-Bae · Ha, Sung-Ho

Purpose

Real estate usually takes charge of the highest proportion of physical properties which individual, organizations, and government hold and instability of real estate market affects the economic condition seriously for each economic subject. Consequently, practices for predicting the real estate market have attention for various reasons, such as financial investment, administrative convenience, and wealth management. Additionally, development of machine learning algorithms and computing hardware enhances the expectation for more precise and useful prediction models in real estate market.

Design/methodology/approach

In response to the demand, this paper aims to provide a framework for forecasting the real estate market with machine learning algorithms. The framework consists of demonstrating the prediction efficiency of each machine learning algorithm, interpreting the interior feature effects of prediction model with a state-of-art algorithm, LIME(Local Interpretable Model-agnostic Explanation), and comparing the results in different cities.

Findings

This research could not only enhance the academic base for information system and real estate fields, but also resolve information asymmetry on real estate market among economic subjects. This research revealed that macroeconomic indicators, real estate-related indicators, and Google Trends search indexes can predict real-estate prices quite well.

Keyword: Machine learning, Real-estate market, Time series analysis, LIME, RNN, LSTM

* 이 논문은 2020년 8월 20일 접수, 2020년 9월 6일 1차 심사, 2020년 9월 16일 게재 확정되었습니다.