공 학 석 사 학 위 논 문

머신러닝을 활용한 부산시 부동산 지수 분석 및 예측

2025년 2월

국 립 부 경 대 학 교 대 학 원

ICT교통융합전공

이 학 만

공 학 석 사 학 위 논 문

머신러닝을 활용한 부산시 부동산 지수 분석 및 예측

지도교수 김 동 재

이 논문을 공학석사 학위논문으로 제출함.

2025년 2월

국 립 부 경 대 학 교 대 학 원

ICT교통융합전공

이 학 만

이학만의 공학석사 학위논문을

인준함.

2025년 2월 21일

위 원 장 공학박사 정 연 호 (인)

위 원 공학박사 류 지 열 (인)

위 원 공학박사 김 동 재 (인)

목 차

목 차

표 목 차

그 림 목 차

머신러닝을 활용한 부산시 부동산 지수 분석 및 예측

이 학 만

부 경 대 학 교　대 학 원　ICT교통융합전공

요　약

국내에서 가구 평균 자산 중 부동산이 73%로 매우 큰 비중을 차지하고 있으며, 국민들은 투자자산으로 인식하여 가격상승을 통한 자산 증식을 위해 부동산을 활용하고 있는 실정이다. 이러한 사실이 경제에 미치는 영향을 감안하여 정치·경제적 분야에서 주요정책으로 다뤄지고 있지만 전국적으로 아파트 가격은 이러한 다양한 정책에도 불구하고 불안정한 추세를 보이고, 향후 대책을 수립하기 위한 가격 예측은 매우 중요하게 지속적으로 연구하고 분석할 주제이다.

부동산 시장의 예측 선행연구는 계량경제학 모델을 활용한 부동산 지수 예측부터 이들 모델과 머신러닝 모델에 대해 구분한 연구로 요약할 수 있다. 이러한 연구들은 연구자에 따라 다양한 머신러닝 기법을 활용하고, 이들의 예측력을 분석하는 다양한 연구의 필요성이 제기되고 있다.

이에 본 연구에서는 부산시를 대상으로 부동산 관련 지수 4개를 선정하여 머신러닝 모델인 LSTM 모델을 이용한 다변량 연구를 진행하였다. 부동산 관련 4개 지수에 대한 자료는 한국부동산원에서 제공하는 것으로, 이중 입력변수 3개는 부산 지가지수, 부산 전세가격지수, 부산 부동산 소비심리지수, 출력변수 즉, 예측력 측정 대상 변수는 부산 아파트매매지수를 선정하여 연구를 진행하였다.

연구결과, 다른 연구와 마찬가지로 LSTM 모델에서 부산 아파트매매지수의 예측정확도가 0.98, 오차율은 0.06으로 예측력이 매우 우수하게 나타났다.

향후 연구에서는 지역을 부산시의 구 단위로 세분화한 비교연구 및 미시적·거시적 경제지표 등을 더 많은 변수로 활용한 추가 연구 및 좀 더 다양한 머신러닝 모델을 활용한 연구를 진행한다면 더욱 풍부한 연구결과가 도출될 것으로 판단된다.

Analysis and Forecast of Real Estate Index in Busan
Using Machine Learning

Hag Man Lee

Major of ICT and Transportation Convergence, The Graduate School,
Pukyong National University

Abstract

In Korea, real estate accounts for 73% of the average household assets, and people perceive it as an investment asset, using it to increase their wealth through price appreciation. Considering this impact, it is treated as a major policy in political and economic fields, and predicting real estate prices is a very important topic that requires continuous research and analysis.

This study selected four real estate-related indices in Busan and conducted a multivariate study using the LSTM (Long Short-Term Memory) machine learning model. Among the four real estate-related indices, three were used as input variables: the Busan Land Price Index, the Busan Rental Price Index, and the Busan Real Estate Consumer Sentiment Index. The output variable, or the target variable, was the Busan Apartment Sales Price Index.

The results showed that, similar to other studies, the prediction accuracy of the Busan Apartment Sales Price Index using the LSTM model was very high at 0.98.

Future research is expected to yield richer results by utilizing more economic variables and various machine learning models.

제 1 장 서 론

1.1. 연구의 배경 및 목적

가계금융복지조사에 따르면 2023년 기준 국내에서 가구 평균 자산의 73%에 해당하는 부동산은 매우 큰 비중을 차지하고 있다. 국내 가구의 자산 대부분이 부동산에 편중되었으며, 대부분의 국민들은 부동산을 투자자산으로 인식하여 가격상승을 통한 자산 증식을 위해 활용하고 있다. 이러한 사실이 경제에 미치는 영향을 감안하여 정치•경제적 분야에서 주요 정책으로 다뤄지고 있다.

국내 부동산 시장은 2008년 글로벌 금융위기로 침체되었다가 2013년 이후 주택경기 활성화 정책으로 상승으로 전환하고 지속적 상승을 보였다. 이에 정부에서는 주택가격 안정화를 위해 2017년부터 정책을 여러 차례 발표했다. 하지만 주택도시보증공사(HUG)가 발표한 2024년 6월 말 기준 부산 민간아파트의 최근 1년간 1㎡당 평균적인 분양가격은 646만원으로 집계되었다. 이를 3.3㎡당 기준으로 환산해 보면 2,131만원으로 국민평형이라고 불리는 전용 84.3㎡ 기준으로 7억 원이 넘는 금액이다. 이렇게 높은 가격과 불안정한 부동산 시장은 국가 거시경제를 불안하게 만들 수 있는 중요한 요인이기 때문에 미래 부동산 시장에 대한 예측은 높은 관심을 받았고, 예측 모델의 연구가 활발하게 진행되어 왔다.

이 중에서 부동산 지수 예측은 부동산 가격을 기초로 지수화되어 부동산 가격과 매우 밀접한 관련이 있으며 부동산 가격 추이나 변동을 예측에 도움을 줄 수 있다. 이러한 사실로 인해 부동산 가격지수나 부동산 소비심리

지수 등에 대한 예측연구가 지속적으로 진행되었지만, 다양한 부동산 지수들에 대한 고찰과 예측 비교연구는 아직도 부족한 실정이다.

전국적으로 아파트 가격은 이러한 다양한 정책에도 불구하고 불안정한 추세를 보이고, 향후 대책을 수립하기 위한 가격변동 예측은 중요하고, 지속적으로 연구하고 분석할 주제이다.

부동산 시장 예측 선행연구는 계량 경제학 모델을 활용한 부동산 지수 예측부터 머신러닝 모델에 대한 연구로 구분하여 요약할 수 있다. 따라서, 연구자에 따라 다양한 머신러닝 기법을 활용하고, 이들의 예측력을 분석하는 연구의 필요성이 제기된다.

기존의 부동산 가격을 평가 및 예측하는 방법 중 전통적인 통계 기법으로 과거 가격 변화 추이를 분석하여 미래의 가격을 예측하는 시계열 모델이 있다. 시계열 모델에는 단변량 시계열 모형 즉, 가격의 변동 추세 변수만을 이용하는 자기 회귀누적이동평균(ARIMA: Auto Regressive Integrated Moving Average)모델과 다변량 모형 즉, 가격의 추세와 가격에 영향을 주는 거시경제지표 또는 공급 및 수요 변수를 포함하는 벡터자기회귀(VAR: Vector Auto Regressive)모델로 구분할 수 있다[1].

전통적인 시계열분석 모델은 통계적 논리로 부동산 시장의 현상을 설명하기 위해 활용되어 왔지만 부동산 시장의 변화에 따라서 고려해야 하는 정성적 및 정량적 변수들이 증가하여 모델에 직접적으로 반영하기 어려워졌다. 이에 기계학습(ML: Machine Learning)의 발전으로 인공지능(AI: Artificial Intelligence)을 활용한 시계열 예측 방법이 기존 계량경제 학적 시계열 모델의 한계를 보완하며 활발히 도입되었다[2].

머신러닝을 이용한 시계열 예측 방법은 전통적인 시계열 분석법과는 다르게 변수들 간의 선형성 가정, 변수들 간의 상호 독립성 가정 등의 제약을 받지 않고 분석 및 예측이 가능해진다. 그렇기 때문에 모델의 적용할 수

있는 범위가 넓고 전통적인 시계열 분석법보다 상대적으로 높은 예측력을 가지는 장점이 있다. 이와 같은 인공신경망은 자료들을 자체 학습해 미래를 예측하면서 복잡한 비선형 문제를 연구하는 최적의 도구가 될 수 있다[3]. 따라서 인공신경망 알고리즘인 장단기기억(LSTM: Long Short Term Memory) 알고리즘을 통한 부동산 지수 예측을 진행한다.

본 연구에서는 부동산과 관련된 지수를 대상으로 이론적 고찰 및 머신러닝을 활용한 예측을 진행하며, 그 결과를 통해 1) 머신러닝 모델의 예측력(정확도)을 분석하고 2) 부동산 관련 지수를 활용한 예측과 관련하여 머신러닝 활용 방향을 제시한다.

연구 목적은 다음과 같다.

첫째, 머신러닝을 활용한 부동산 지수 예측을 진행하여 부동산 지수 예측에서 머신러닝의 활용성을 검증하고, 예측 관련 변수를 4가지의 다양한 부동산 관련 지수들로 구성하여 기존 연구들과 차별화된 예측을 진행한다.

둘째, 머신러닝 모델의 예측 정확도가 높은가를 측정하고 그 결과를 도출한다.

마지막으로, 부동산 지수 예측 결과에 따른 연구결과 및 향후 연구과제를 도출한다.

1.2. 연구의 범위 및 방법

본 연구의 범위는 부산광역시의 부동산 지수 중 아파트매매지수를 예측하는 것을 목적으로 한다. 부산광역시는 서울특별시, 경기도 다음으로 약 328만명이 거주하고 있는 지역이다[4]. 또한, 부산광역시는 저출생과 초고령화 여파로 광역시 가운데 처음으로 소멸 위험단계에 진입했다[5]. 이는 부산광역시의 부동산 가격을 기초로 한 부동산 지수는 서울, 경기 등 수도권의

인구 밀집 지역을 제외한 나머지 국내 지역 또는 지방의 부동산 가격을 예측하는데 매우 중요한 배경이 될 수 있다.

예측에는 한국부동산원에서 제공하는 2014부터 2023년까지 10년간의 부동산 관련 지수를 대상으로 하며 다양한 부동산 지수중에서도 부동산 가격과 밀접한 연관을 가지고 기존 연구에서도 부동산 가격 예측에 활용되었던 아파트 매매지수, 지가지수, 전세가격지수, 부동산 소비심리지수로 범위를 한정하여 예측을 진행한다.

이를 위해 본 연구는 첫째, 부동산 관련 지수 결정 및 예측 선행연구의 이론적 고찰과 부동산 지수 예측모델 관련하여 이론 및 선행연구 검토, 둘째, 머신러닝 모델을 활용한 부동산 지수 중 아파트매매지수 예측, 마지막으로 본 연구결과 및 시사점을 제시하는 순으로 진행된다.

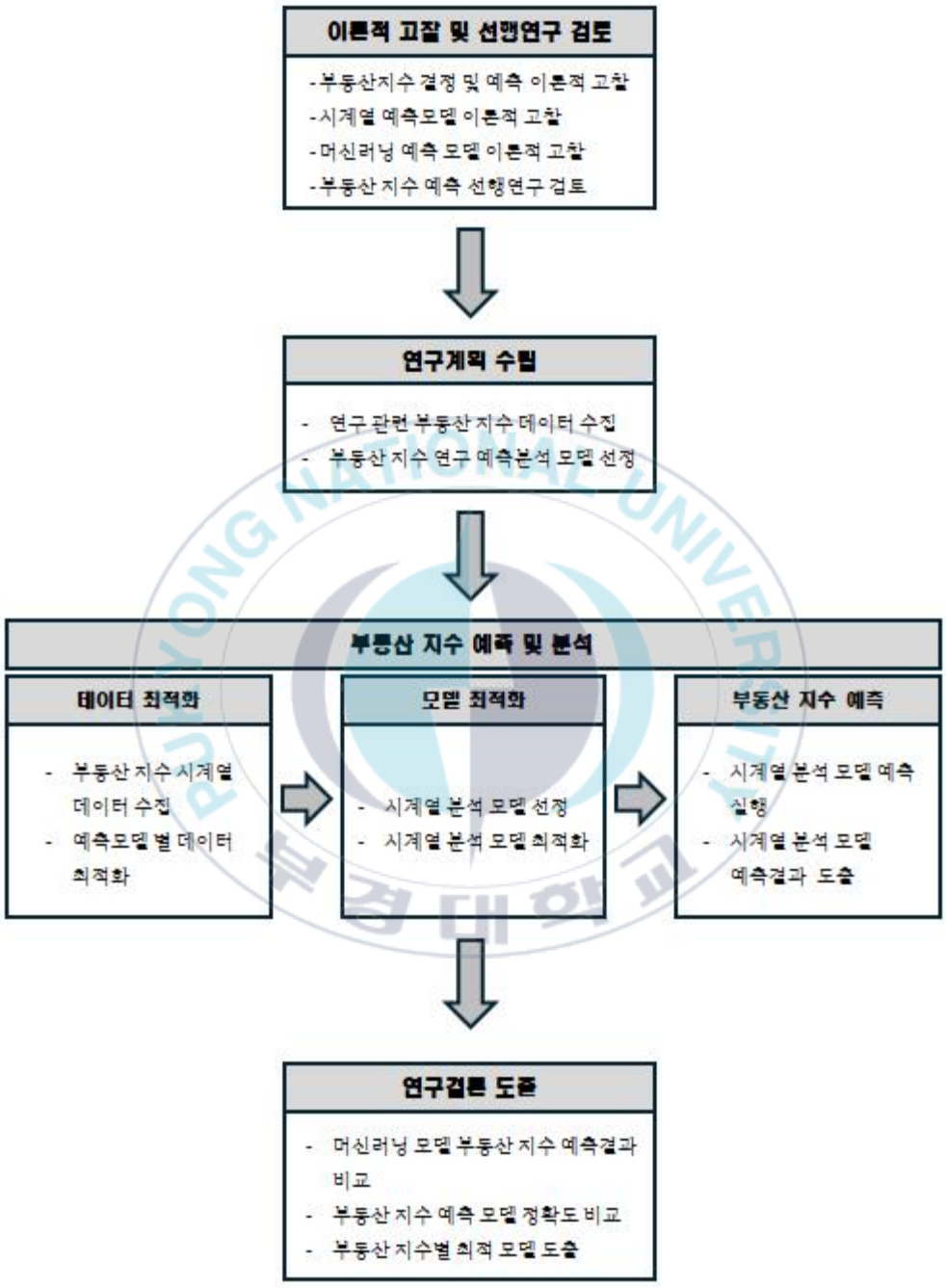
이론적 분석 및 선행연구 검토는 부동산 지수 예측 이론 및 모델, 기존의 계량 경제학적 방법론인 시계열 예측 모델, 그리고 최근 부동산 지수 예측에 활용되고 있는 머신러닝 모델에 대해 알아본다. 또한, 선행연구 검토를 위해 계량경제학 모델로 다변량 자료를 이용하여 부동산 가격을 예측 분석했던 연구 및 논문들과 머신러닝 모델을 이용하여 부동산 가격을 예측 분석하고 이를 계량경제학 모델과 비교한 논문들을 살펴본다. 아울러 머신러닝 모델을 활용하여 시계열 자료를 분석한 논문 중 다변량 자료의 예측력을 비교한 연구 및 논문들을 참고한다.

실증분석에서 데이터 선정은 10년간의 부동산 지수 4가지를 대상으로 하며 부동산 지수 예측과 연관되는 경제학의 설명변수를 추가해 부동산 지수 예측 결과를 설명하고자 한다. 머신러닝 예측 방법으로는 기존 연구에서 부동산 지수 예측에 활용되었으며 높은 정확도를 보인 LSTM 머신러닝 모델을 활용한다. 머신러닝 모델은 1차적으로 부동산 관련 지수를 데이터별로 최적화하고 2차적으로 최적화 모델을 활용해 부동산 지수의 예측을 진

행한다.

본 연구는 선행연구들과의 차별성은 첫째, 본 연구에서는 인공지능 시계열 예측 분야에서 예측력을 인정받고 있는 LSTM을 활용하여 다변량 부동산 관련 지수 예측 모델을 구축하고 머신러닝 모델의 활용성을 확인하고자 한다. 둘째, 머신러닝 모델의 설계, 모델 최적화 과정 등 우수한 예측 모델 구축을 위한 과정 등을 더욱 구체적으로 제시하고자 한다. 머신러닝의 경우 같은 알고리즘을 사용하더라도 최적 모델을 찾는 과정이 다를 수 있기 때문에 본 연구의 최적화 과정이 유사 연구에서 예측 모델 구축에 참고가 될 수 있을 것이다. 마지막으로, 서울 등의 수도권이 아닌 부산시의 부동산 관련 지수의 시계열데이터를 활용하여 최적화 예측 모델의 성능을 확인하고자 한다.

본 연구의 범위 및 방법을 토대로 작성한 연구의 흐름은 아래 <그림 1>과 같다.


이론적 고찰 및 선행연구 검토
-부동산지수 결정 및 예측 이론적 고찰
-시계열 예측모델 이론적 고찰
-머신러닝 예측 모델 이론적 고찰
-부동산 지수 예측 선행연구 검토
연구계획 수립
- 연구 관련 부동산 지수 데이터 수집
- 부동산 지수 연구 예측분석 모델 선정
부동산 지수 예측 및 분석
데이터 최적화
- 부동산 지수 시계열 데이터 수집
- 예측모델별 데이터 최적화
모델 최적화
- 시계열 분석 모델 선정
- 시계열 분석 모델 최적화
부동산 지수 예측
- 시계열 분석 모델 예측 실행
- 시계열 분석 모델 예측결과 도출
연구결론 도출
- 머신러닝 모델 부동산 지수 예측결과 비교
- 부동산 지수 예측 모델 정확도 비교
- 부동산 지수별 최적 모델 도출

<그림 1> 연구흐름도

제 2 장 이론적 분석 및 선행연구 검토

2.1. 부동산 관련 지수 결정 및 방법론

부동산 관련 다양한 지수가 존재하고 있다. 한국부동산에서는 부동산 지수로 지가지수, 매매지수, 전세가격지수, 월세가격지수, 실거래가격지수, 임대가격지수를 조사 및 제공하고 있다. 또한 관련 지수로는 경기종합지수, 소비자물가지수, 생산자물가지수, 부동산 소비심리지수를 제공하고 있다. 이와 같은 부동산 지수들은 부동산 가격 즉 지가, 매매가격, 전세가격, 월세가격, 실거래가격을 기초로 지수화한 자료이며 부동산 관련 지수는 부동산과 연관되는 경제 요소를 기초로 지수화한 자료이다.

아파트 매매지수와 전세가격 지수가 포함되는 공동주택 실거래 가격지수는 일정 시점을 기준시점(=100)으로 하고 반복 매매 모형 추정산식을 기반으로 전국의 아파트 및 연립·다세대를 대상으로 실제 거래를 통해 신고된 아파트 및 연립·다세대 거래 가격수준 및 변동률을 파악하여 보다 정확한 시장동향을 국민에게 제공하고 정부 정책 수립에 참고 자료로 활용한다. 지가지수가 포함되는 전국지가변동률 조사는 기준시점(=100)으로 하고 Laspeyres 수정산식을 이용하여 전국의 지가변동 상황을 조사하고 토지정책수행 및 감정평가 시 시점수정 등을 위한 자료로 활용된다[6].

부동산 소비심리지수가 해당되는 부동산시장소비자 심리조사는 부동산 시장 소비자의 행태변화 및 인지수준을 설문조사를 통해 파악하고, 부동산 시장 분석을 위한 주요 기초자료로 활용하며 부동산 소비심리지수 = (주택시장 소비심리지수 * 0.9) + (토지시장 소비심리지수 * 0.1)와 같은 지수산식을 가지는데 여기서 주택시장 소비심리지수는 (주택매매가격지수 + 주택

전세가격지수 + 주택거래지수) / 3 산식에 의해 산출되고 토지시장 소비심리지수는 (토지가격지수 + 토지거래지수) / 2 식에 의해 결정된다[7].

2.2. 부동산 지수 예측 모델

2.2.1. 정성적 부동산 지수 예측 모델

부동산 가격을 정확하게 예측하기 위해 다양한 방법들이 연구 및 시도되어왔다. 전통적인 예측법은 정성적 분석 방법, 정량적 분석 방법으로 2가지로 나눌 수 있다.

정성적 분석 방법은 신뢰성이 있는 부동산 가격의 시계열 자료가 없거나 인과 분석법의 적용이 어려워서 수치적 분석이 불가능한 경우에 활용된다. 정성적 분석 방법으로는 브레인스토밍 기법, 델파이 기법, 설문조사 기법, 유추기법, 시나리오 기법 등이 있다. 이는 개략적으로 시장의 동향이나 특성 등을 분석할 수 있는 분석 방법이다[2].

브레인스토밍 기법은 전문가들이 자유토론을 통해 의견을 나누고 종합하여 부동산 가격에 대한 합의를 도출해내는 방법이다. 장점으로 전문가들의 다양한 아이디어를 도출할 수 있다는 것이 있다. 델파이 기법은 전문가들에게 면적과 서면조사를 반복적으로 진행하여 미래의 부동산 가격에 대해 합의를 이끌어내는 방법이다. 전문가들의 참여 정도에 따라 통계적 자료를 만들 수 있다[8]. 하지만 시간이 많이 소요되며 참여 의지가 낮을 경우 유의미한 결과를 얻기 어렵다는 단점이 있다. 설문조사 기법은 부동산 시장의 공급자 및 수요자를 대상으로 설문조사를 진행하여 부동산 가격을 예측하는 방법이다. 이는 시장 참여자에 대한 직접적인 조사이기 때문에 신뢰성이 높아 많이 활용된다[9]. 유추 기법은 유사한 사례를 기준으로 부동산

가격을 유추해 내는 방법이다[10]. 시나리오 기법은 과거와 현재의 상황을 바탕으로 미래를 가상 시나리오를 통해 예측하는 방법이다[11]. 이외에도 형태학적 분석법(Morphological analysis), 트리 분석법(Tree analysis) 등의 정성적 분석 방법이 활용되고 있다[2].

2.2.2. 정량적 부동산 지수 예측 모델

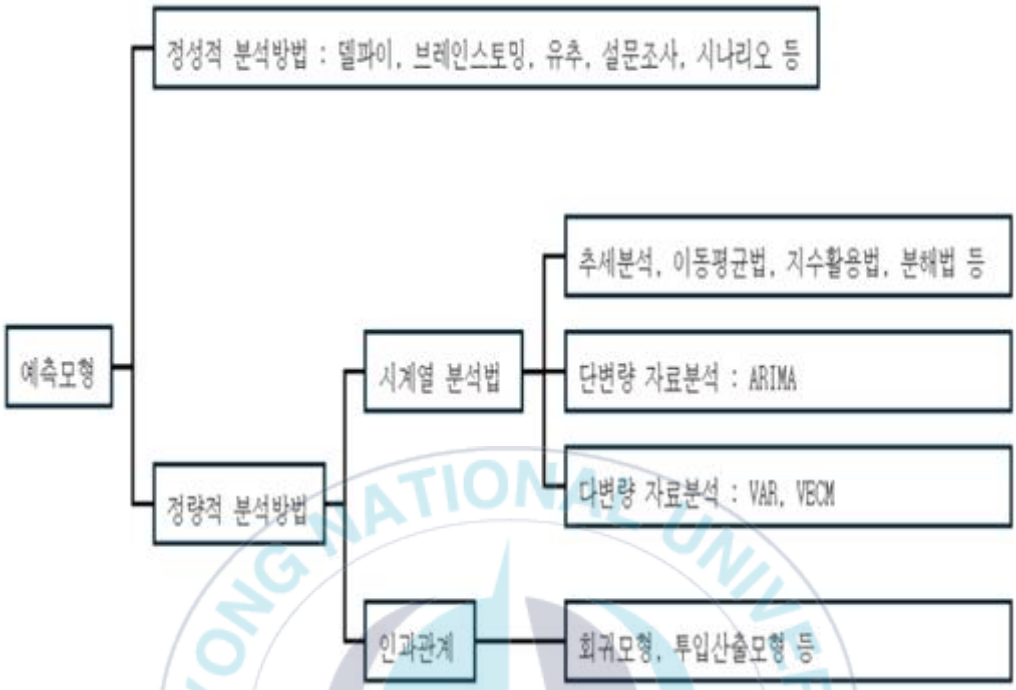
정량적 분석 방법은 수치화된 자료를 통계적으로 처리하여 분석하거나 미래를 예측하는 방법이다. 여기에는 크게 시계열 분석법과 인과 분석 방법으로 나눌 수 있다[12].

시계열 분석법은 과거 주택시장의 가격 자료를 분석하여 미래를 예측하는 분석 방법이며 기본적으로 자료의 추세가 선형적이고 안정적일 때 효과적인 방법이다. 이러한 방법은 미래의 변동은 과거의 추세를 반복한다는 원리를 기본바탕으로 분석하기 때문에 과거 자료에 미래의 변화에 대한 정보가 들어 있다고 전제한다. 또한 실증적 자료를 활용하기 때문에 설득력이 높고 설명이 용이하다는 장점과 분석을 위해서는 시간 자료가 충분히 축적되어야 하고 변동성이 높은 경우에는 예측이 어렵다는 한계점이 있다[2].

시계열 자료에는 계절변동(Seasonal pattern), 순환변동(Cycle pattern), 추세변동(Trend pattern), 불규칙변동(Random variation) 등에 대한 값이 포함되어 있다. 시계열 모델은 주로 시계열 자료에서 계절성, 순환 주기, 추세와 같은 요소들을 분리하여 그것들이 상호적으로 어떻게 관계하는 가를 규명하여 수학적으로 표현한다[13]. 이를 분리하는 방법에는 계절 요인 분석법, 이동평균법, 추세분석법, 지수평활법, 분해법 등이 있다. 단변량 자료 분석에는 이동평균법의 ARIMA모델과 다변량 자료의 분석에는 VAR 모델이 주로 활용되고 있다[2].

인과 분석 방법이란 부동산 시장의 설명 변수들과 종속변수 간의 인과관계를 이용하는 계량적 분석 기법이며, 이는 다른 예측 방법과 비교하여 정교하다. 인과 분석 방법의 인과관계는 수학적으로 표현되기에 종속변수에 영향을 주는 요소들의 현재와 과거의 영향 요인을 포함시킬 수 있다. 인과 분석 방법에는 회귀 모델(Regression model), 연립 방정식 모델(Simultaneous equations model, 계량경제모델), 투입산출모델(Input-output matrix model), 선행지수 모델(Leading composite index model) 등의 다양한 방법이 있다. 부동산 가격의 분석에서는 주로 다양한 변수를 활용하는 다중회귀모델이 사용된다. 다중회귀모델은 종속변수에 영향을 주는 환경요인 간의 회귀방정식을 구하고 환경요인이 변화함에 따라 독립변수가 어떻게 변화하는지 예측한다[2].

아래 <그림 2>는 본 논문에서 설명한 정성적 및 정량적 예측 모델에 대한 분류도이다.



<그림 2> 예측 모델 분류도[14]

2.3. 시계열 예측 모델

2.3.1. 시계열 예측 모델 이론

시계열(Time series)이란, 시간 흐름에 따라 연속하여 관측된 값들을 정리한 통계 계열을 의미하고 시계열 자료란, 시계열에 의해 수집된 통계자료를 말한다[15]. 시계열 자료를 활용해 예측하는 방법으로는 추세분석(Trend analysis), 평활법(smoothing method), 분해법(decomposition method),자기회귀누적이동평균(ARIMA: Auto Regressive Integrated Moving Average) 모델에 의한 방법 등이 있다[14].

시계열 분석기법은 과거 변수 형태가 미래에도 반복된다는 원리에 기반하기 때문에 종속변수와 관련된 환경요인들이 안정적일 경우에는 높은 예측결과를 얻을 수 있지만 변화가 클 경우 오차가 커지는 한계를 가진다[10].

전통적인 시계열 자료 분석에는 분해법이 있다. 이는 여러 성분을 분해하고 각 성분들을 추정하여 본래의 시계열을 해석하는 방법이다. 성분에는 체계적 성분인 추세 성분, 계절성뿐, 순환성분과 불규칙성분으로 나눠진다[10]. 여기서 추세 성분(T: Trend component)이란 시계열 자료의 기본적인 변동 추세이며 장기간에 걸친 경향을 의미한다. 계절 성분(S: Seasonal component)은 1년 이내에서 일정 주기로 반복되는 변동이며, 순환성분(C: Cyclical component)은 1년 이상의 주기적인 변동 경향이고 관측값의 상하 변동 주기가 규칙성과 관계없이 반복되는 경향이다. 불규칙성분(I: Irregular component)은 우발적으로 불규칙하게 나타나는 변동으로 체계적 성분들을 측정하기 위해서 제거해야 한다[16].

2.3.2. ARIMA(Auto Regressive Integrated Moving Average)

ARIMA모델이란 단일 시계열 변수의 현재 관측치 Zn과 시간이 흐름에 따라 일정한 규칙성을 갖는다고 가정하고 미래를 예측하는 방법이다. 현재의 관측값 Zn을 과거의 관측값{Zt, t 〈 n}들과 백색 잡음(white noise)인 오차{εt, $Z_n = f(Z_1, Z_2, \cdots, E_1, E_2 \cdots, E_{n-1}) + E_n, t \leq n$}의 선형 결합 형태로 표현한다. 이와 같은 단변량 시계열 분석 방법은 모델 설정이 용이하지만 변수들 사이의 상호작용은 반영하지 못한다. 단변량 시계열 분석 방법에서는 시계열 자료의 과거 값들이 현재의 관측값에 영향을 미치고 이러한 상관관계를 해석하는 것을 자기 회귀(AR: Auto Regressive) 과정이라고 한다. 현재 관측값에 대한 이전의 관측값을 lag 1의 관측값이라 하는데 lag의 값이

커질수록 현재의 관측값이 먼 과거의 값에 영향을 받는다고 여긴다. 하지만 과거 추세를 정확하게 예측하려 해도 미랫값은 백색 잡음이라고 하는 오차에 의해 영향을 받는다. 여기서 오차들은 독립적이고 동일한 분포를 따르는 확률변수들이며, 백색 잡음에 의한 영향을 분석하는 과정을 이동평균(MA: Moving Average) 과정이라 한다[14].

또한 ARIMA 모델은 각 관측값에서 평균과 분산을 일정하게하기 위해 미분값을 활용한다. 평균과 분산이 일정해진 자료는 안정된(stationary) 자료라고 부르며 안정된 자료는 일정한 평균과 분산을 가져 분석과 예측이 용이해진다. 미분을 통해 얻어진 안정된 시계열 자료는 AR모델, MA모델, 자기회귀이동평균(ARMA: Auto Regressive Moving Average)모델로 표현할 수 있게 된다[14].

2.3.3. VAR(Vector Auto Regressive)

벡터회귀(VAR: Vector Autoregressive)모델은 인과관계가 있는 k개의 현재 변수를 종속변수로 설정하고 설명변수는 이들의 과거 값들로 하여 선형회귀방정식을 통해 시계열의 확률과정(stochastic process)을 추정하는 방법이다[15]. 이는 회귀분석과 시계열 분석 방법의 결합 형태이며 단기 예측을 목적으로 개발되어 여러 개의 시계열 자료에 대한 분석을 위해 변수 상호 간에 영향을 주는 동적 연립 방정식 모델이다[17].

VAR 모델은 구조모델과 달리 선험적 경제이론을 배제하고 변수 간의 상관관계 및 시차 상관관계를 활용하여 구성된 다변량 시계열 모델이다. 이는 인과관계가 있는 변수들의 현재 관측치를 종속변수로 설정하고 본 변수들과 타 변수들의 과거 관측치들을 설명변수로 설정하여 n개의 선형 회귀방정식을 통해 시계열 프로세스를 추정하는 방법이다[18]. VAR 모델은 다

항 회귀 모델 즉, 다변량 시계열 자료를 분석하는 모델로 특별한 제약이 없어 시계열 분석에 자주 이용된다. 변수들의 동적 관계가 단기적 지연(lag)에는 영향을 받지만, 장기적으로는 받지 않는다는 것을 기본 가정으로 한다[19].

2.4. 머신러닝 예측 모델

2.4.1. 머신러닝 예측 모델 이론

부동산 가격에 대한 예측에서는 입력변수를 부동산의 미시적 요소로 하여 회귀분석을 하거나 거시적 요소의 시계열 자료로 하여 시계열 분석을 한다. 전통적 계량 통계학의 회귀분석에서는 공간계량경제모델(Spatial Econometrics Model), 최소자승법(OLS: Ordinary Least Square), 공간지리가중회귀모델(GWR: Geographically Weighted Regression)등이 활용되었다. 시계열 분석에서는 단변량일 경우 AR, ARIMA모델을 활용하고 다변량일 경우 VAR, VECM모델을 활용하였다[14].

이와 더불어 최근 금융에서는 머신러닝의 도입이 늘어나고 있는 추세이다. 머신러닝이란 인공 지능(artificial intelligence)을 구현하는 하나의 방법이며 인간이 학습하는 것과 같이 컴퓨터가 알고리즘, 프로그램을 이용한 학습을 통해 의사결정을 하거나 새로운 정보를 도출하는 것을 말한다(배성완, 2019). 머신러닝의 학습방법은 지도학습(supervised learning), 비지도학습(unsupervised learning), 강화학습(reinforcement learning)으로 나뉜다.[14].

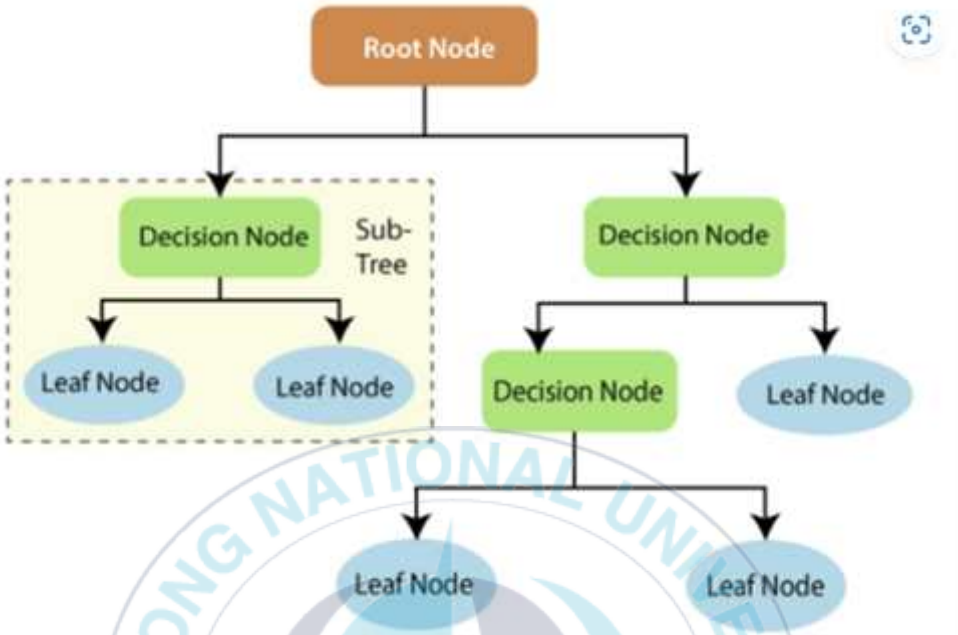
지도학습이란 입력값과 결괏값을 가지고 있는 데이터를 통한 학습 방법이다. 이는 학습된 모델로 미지의 결괏값을 예측하는 것을 목적으로 한다. 주

로 구체적인 수치를 예측하는 회귀(regression)에 활용된다. 지도학습의 알고리즘은 비용 함수(cost function)의 값 그리고 예측된 결괏값의 오류를 줄여나가는 방법으로 구동한다. 지도학습을 활용하는 방법론으로는 서포트 벡터 머신, 인공신경망, 심층신경망, 랜덤 포레스트, 부스팅 등이 있다. 비지도 학습이란 결과값이 없는 데이터를 통한 학습 방법이다. 컴퓨터가 스스로 학습하며 데이터의 구조나 패턴 등을 찾아낸다. 차원 축소(dimensionally reduction)나 클러스터링(clustering)에 주로 활용된다. 비지도 학습이 지도학습과 가장 다른 점은 주어진 데이터의 성질을 직접적으로 추측해 낸다는 것이다[14]. 주성분분석(principal component analysis), k-means 군집화(k-means clustering) 등은 대표적인 비지도 학습 알고리즘이다. 마지막으로, 강화학습이란 행동의 주체인 에이전트(agent)가 현재 상태(state)를 인식해 선택 가능한 행동(action) 중에 최대 보상(maximum reward)을 가지는 행동을 선택하는 방법을 말한다. 에이전트가 관측하고 행동하며 보상받는 상호작용을 경험하고, 시행착오를 통해 학습해 나가는 방식이 인간의 지식 습득 과정과 닮아있다. 대표적으로 Game AI, Robot Navigation 등에서 활용된다[14]. 이와 같은 머신 러닝은 사람이 풀 수 있는 수준의 문제이지만 학습 규모가 방대한 경우이거나, 수학적으로 정의할 수 있는 문제이지만 너무 복잡하여 사람이 수학적으로 명확하게 정의하는 것이 어려울 때 유용하게 쓰인다[20].

2.4.2. 랜덤 포레스트(RF, Random Forest)

랜덤 포레스트(RF, Random Forest)는 부트스트랩(bootstrap)을 이용하여 여러 개의 표본을 생성하고 의사결정나무(Decision Tree)모델을 적용해 결과를 종합하는 앙상블 알고리즘 중 하나이다[21]. 회귀트리 모델은 설명변

수 $X_1, X_1, \cdots, X_n$를 J개의 지역(region) $R_1, R_2, \cdots, R$에 겹치지 않게 분할하고 R에 속하는 관찰 치에 대해 R의 관찰치 평균값을 예측치로 제시한다. R는 잔차제곱합(residual sum of squares)이 최소화 되도록 분할하며 과적합의 문제를 해결하기 위해서는 트리의 규모를 가장 크게 해두고 가지를 쳐내면서 적정규모의 트리를 결정해 나가는데 이는 $\sum_{m=1}^{|t|}\sum_{x_i \in R_m}(y_i - \hat{y}_{R_m})^2 + a|T|$을 최소화하는 과정이라 할 수 있다[22]. 식에서 $|T|$는 트리 T의 가지(terminal node) 수를 의미하고 R_m은 m번째 가지의 분할지역을 의미한다. a는 동조 파라미터(tuning parameter)로 a=0일 때 패널티가 없어 최대 트리가 되고 반대로 a가 커지면 트리규모가 작아지게 된다[23]. 이와 같은 랜덤 포레스트를 활용하는 이유는 주택가격에 미치는 영향 요인은 매우 다양하기 때문에 해당되는 요인을 모두 고려할 시 비선형 충격함수를 추정하는데 제약이 있기 때문이다. 변수 선택 시 무작위 포레스트를 사용하면 종속변수와 설명변수 간 중요도 점수를 확인하여 더 중요한 관계에 있는 변수들을 선별해 내는 것이 가능하다[24]. 또한 랜덤 포레스트는 수천 개의 트리구조를 만들기 때문에 최종 모델을 제시할 수는 없지만 분류모델 성능이 뛰어나다는 장점이 있으며[25], 이상치에 크게 영향을 받지 않는다[26].



<그림 3> 랜덤 포레스트

2.4.3. XGBoost(eXtreme Gradient Boosting)

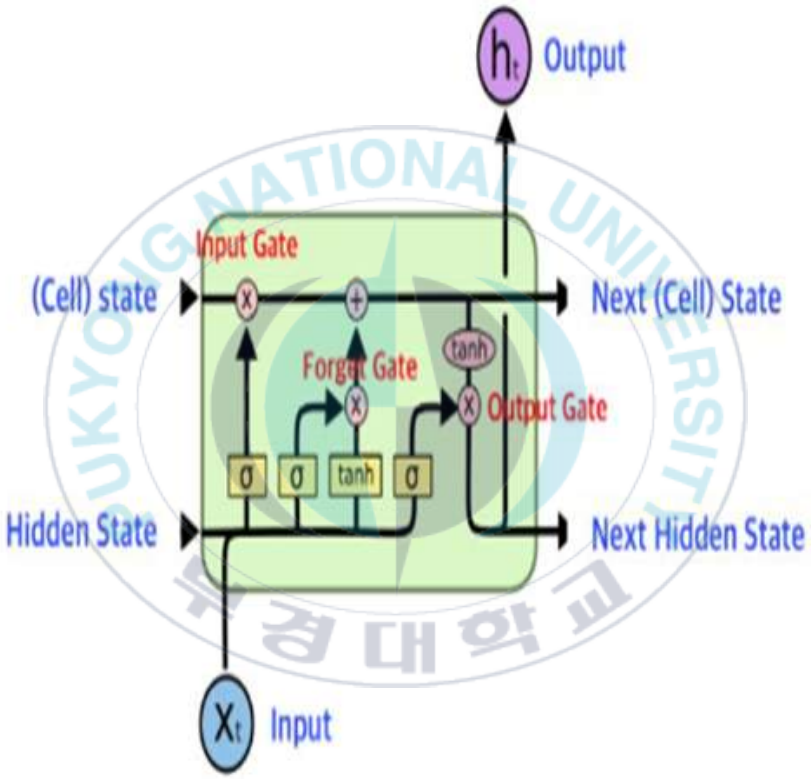
랜덤 포레스트(RF, Random Forest)와 마찬가지로 XGBoost(eXtreme Gradient Boosting) 모델은 다수의 결정트리를 결합시킨 앙상블 방법이다. 의사결정트리를 차례로 학습하며 선행 트리의 오류를 개선하는 앙상블 기법의 알고리즘으로[27] 이전 단계의 트리 모양에 많은 영향을 받게 된다. XGB 모델은 의사결정나무의 데이터 전처리과정이 단순하다는 장점과 다른 의사결정나무 방법에 비해 빠른 속도와 확장성을 가지고 있다는 장점을 가지고 있다. XGBoost의 부스팅(Boosintg)을 이용하여 여러 개의 회귀나무를 이용해 오차를 줄여나가면서 최적 트리를 찾는다. 가지치기 과정에서

획득한 정보를 순서대로 연산하고 점수가 음의 값이 될 때까지 가지를 제거한다. Gain값이 최대가 되도록 과정을 반복한 후 최종적으로 높은 점수의 트리들을 조합하여 모델을 만든다[28]. 상수항만으로 구성된 $F_0(x) = \arg\gamma \sum_{i=1}^{n} L(y_i, \gamma)$는 초기 모델이며 x는 설명변수, y는 종속변수를 의미하고 L(y,F(x))는 미분 가능한 손실함수(loss function)이며 $\gamma_{im} = -\left[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)}\right] F(x) = F_{m-1}(x)$와 같이 유사잔차(pseudo-residuals)를 M번 반복한다. 이후 계산된 유사잔차에 대하여 기본 학습자(base learner)인 $h_m(x)$를 적합한 후 $\gamma_m = \arg\gamma \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)$의 γ_m을 계산하고 $F_m = F_{m-1}(x) + \gamma_m h_m(x)$로 잔차를 업데이트하고 M번 반복한다[23].

2.4.4. 장단기 메모리(LSTM : Long Short Team Memory)

LSTM(Long Short Team Memory) 모델은 장기패턴을 학습할 때 발생하는 기울기 소멸 문제를 극복할 수 있어 어려운 시퀀스 문제나 큰 규모의 순환 네트워크 문제들을 다룰 수 있다. 이는 RNN의 특별한 형태라고 할 수 있다. 표준 RNN은 피드백 루프가 단순한 신경망이지만 LSTM 모델은 신경망 층 대신 메모리 셀(또는 블록)로 구성되어 있고 각 셀은 3개의 게이트를 가져 셀의 상태에 따라 정보의 흐름이 조절된다. 반복 모듈은 표준 RNN 모델에서는 단일 층에서 이루어지지만, LSTM 모델에서는 상호 반응하는 네 개의 층을 포함해 이루어진다[2]. LSTM 네트워크는 과 같은데 Unit의 맨 위의 선은 내부 메모리를 나타내는 셀 상태 C_t이고 아래의 선은 은닉 층의 상태 h_t이다. 하나의 메모리 블록의 망각(f), 입력(i), 출력(o)의 세 개 게이트는 상태와 출력을 조절한다. 각 게이트는 시그모이드(σ) 신경

망 층과 곱셈(×) 연산으로 구성되어 있으며 망각 게이트는 블록에서 버릴 정보를 결정하고 입력 게이트는 입력값을 메모리에 업데이트를 결정한다 (박성훈, 2020). 출력 게이트는 입력과 메모리를 통해 출력을 결정하며 메모리에 기억시킬 값의 범위를 조절한다[29].



<그림 4> LSTM 네트워크

2.5. 선행연구 이론적 고찰

2.5.1. 시계열 분석 부동산 지수 예측

기존 연구에서는 부동산 지수와 시계열 분석을 통해 부동산 가격을 예측한 다양한 연구가 진행되어 왔다. 시계열 분석에는 주로 ARIMA 모델이 사용되었으나, 예측력을 높이기 위해 개입분석 모델이나 국면전환 모델, 비관측 요인 모델, 요인분석 모델, 자기회귀오차 모델, GARCH 모델, IGARCH 모델 등이 함께 활용되기도 하였다[14].

먼저 부동산 가격 예측에 가장 일반적으로 활용되는 시계열 모델인 ARIMA 모델을 활용한 연구들을 검토하고자 한다. 두바이에서는 ARIMA 모델을 활용해 87개월간의 주택 매매가격지수를 분석하여 12개월간의 월간 주택 매매가격지수를 예측하였고 그 결과 실제 값과 2.4% 차이를 보였다[30]. 또한 ARIMA모델로 우리나라 서울과 전국의 31년간 분기별 주택매매가격지수를 분석하여 실제 값과 비교한 RMSE(Root Mean Square Error)값이 서울은 0.4937, 전국은 0.3539가 도출되었다[31]. 이들은 데이터의 지역이나 특성에 따라 예측력의 차이가 있음을 보여주었다. 또한 개입효과를 가진 시계열 자료 분석에서는 ARIMA보다 개입분석모델을 활용한 분석이 높은 예측력을 보였다[32]. 이를 통해 시계열 자료의 특성과 모델에 따라 예측력이 달라지는 사실을 알 수 있다.

ARIMA 모델 이외에 국면전환 모델, 비관측요인 모델, IGARCH모델을 활용하여 우리나라의 전국 아파트가격지수를 예측하는 연구도 있었다. 전국 아파트가격지수를 활용하여 36개월까지의 예측력을 RMSE, 평균절대오차(MAE)로 비교하였다. 표본 내 예측에서, 단기 예측력은 국면전환 모델이 예측력이 높게 나타났다. 장기 예측력은 비관측 요인 모델이 가장 좋은

성능을 보였다. 하지만 표본 밖 예측에서, 비 관측 요인 모델의 예측력이 가장 높았고 국면전환 모델의 예측력이 가장 낮았다[33].

2.5.2. 머신러닝 부동산 지수 예측

기존 시계열 분석 모델의 예측 연구 이후, 최근에는 머신러닝 모델이 등장하며 이를 활용한 부동산 시장에 관한 예측 연구가 활발하게 진행되고 있다. 선행연구를 보면 시계열 부동산 가격 데이터를 활용하는 단변량 분석에서는 머신러닝을 활용한 예측이 기존 시계열 분석 모델인 ARIMA에 비해 우수하였다. 머신러닝분석 결과 비교에는 주로 평균제곱오차(MSE), RMSE가 활용되었다. 민성욱의 딥러닝을 이용한 주택가격 예측연구에서는 다양한 머신러닝 모델을 통해 부동산 지수를 예측하였다[34]. 실거래가 지수에서 SVM(Support Vector Machine)보다 MLP(Multi-layer Perceptron) 모델과 RF 모델의 예측력이 높았고 불규칙한 형상의 주택매매량에서는 RF 모델이 가장 좋은 결과를 보였다. 배성완과 유정석의 연구에서는 10년간의 서울 아파트 매매실거래가격지수, 아파트 매매가격지수, 아파트 전세가격지수, 지가지수를 활용하여 DNN, LSTM 모델로 부동산가격지수를 예측하였다. 예측력의 비교대상은 ARIMA로 하였으며 RMSE를 기준으로 비교를 진행하였다. 결과적으로 DNN과 LSTM의 예측력이 ARIMA 모델에 비해 우수하였다. 그리고, 아파트 매매 실거래 가격지수를 시계열 분석 모델의 단변량에 ARIMA, 다변량에 VAR, BVAR를 활용하여 예측하였고 머신러닝 방법으로 SVM, RF, GBRT, DNN, LSTM을 활용하였다. 결과적으로는 머신러닝 모델이 시계열 분석 모델보다 예측력이 우수하였으며 머신러닝 모델 간에는 단변량 변수의 예측력이 높다는 분석을 보여주었다[23]. 또한 머신러닝을 활용한 예측은 부동산 지수의 비선형성이 높은 곳에서 더

욱 좋은 성능을 발휘한다고 하였다[2]. 하지만 Cao의 연구에서는 다변량 모델의 예측력이 높은 결과를 보이며 활용하는 모델과 데이터에 따라 예측력이 달라지는 것을 알 수 있었다[35]. 이와 관련해 배성완과 유정석은 머신러닝 공동주택 가격 예측에서 앙상블 방법 계열의 GBRT와 RF의 예측력이 상대적으로 높았으며 데이터나 초모수 조합에 따라 다른 결과값을 가질 수 있음을 시사하였다[23].

2.5.3. 본 연구의 차별성과 의의

머신러닝 모델을 활용한 부동산 지수 예측 선행연구를 요약하면 다음과 같다. 첫째, 예측 모델의 발전에 따라 활용되는 모델과 알고리즘이 다양하게 변화하였다. 머신러닝 모델 활용 초반의 2000년대에는 MLP 신경망이 주로 활용되었지만 이후 DNN, RNN, LSTM과 같은 알고리즘을 활용한 연구가 증가하였다. 둘째, 대부분의 국내외 연구에서 머신러닝 모델과 다중회귀모델, ARIMA와 VAR 등 전통적인 통계 모델과의 예측력 비교 분석을 통해 머신러닝 모델의 우수성을 실증적으로 검증하였다. 셋째, 적절한 모델과 변수를 선정하는 작업이 예측력 높은 결과를 도출하기 위해 중요하다는 것을 확인할 수 있다.

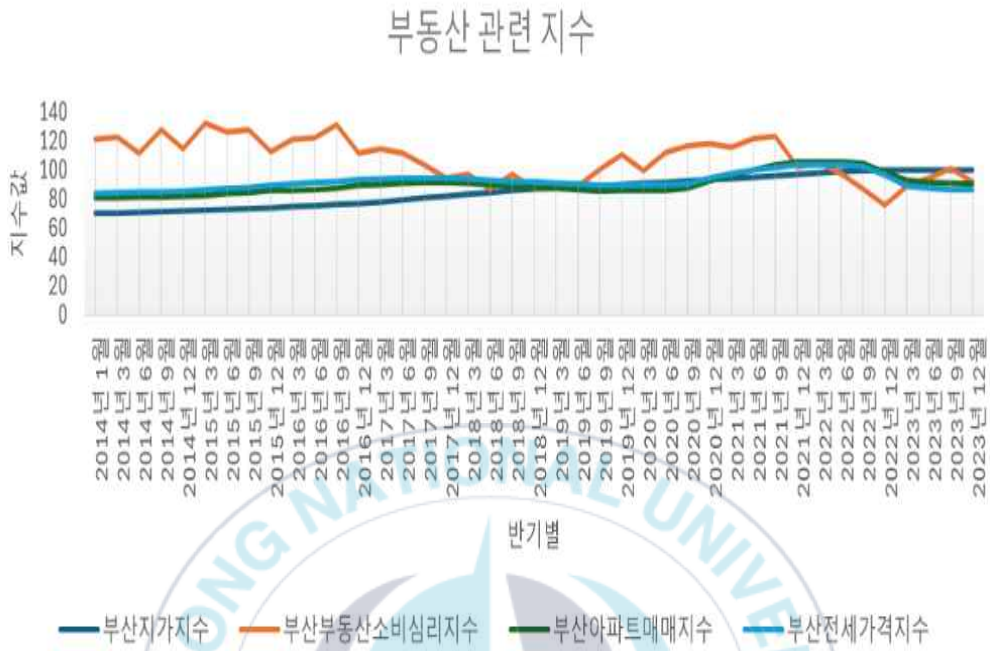
본 연구는 선행연구와 다음과 같은 차별성과 의의를 가진다. 첫째, 본 연구에서는 인공지능 시계열 예측분야에서 예측력을 인정받고 있는 LSTM을 활용하여 다변량 부동산 관련 지수 예측 모델을 구축하고 머신러닝 모델의 활용성을 확인하고자 한다. 국내에서는 배성완과 유정석[23]의 연구와 이태형[2]이 본 연구와 유사한 자료를 활용하여 단변량과 다변량 머신러닝 모델을 구축하였다. 본 연구에서는 머신러닝 모델의 설계, 모델 최적화 과정 등 우수한 예측 모델 구축을 위한 과정 등을 더욱 구체적으로 제시하고자

한다. 머신러닝의 경우 같은 알고리즘을 사용하더라도 최적 모델을 찾는 과정이 다를 수 있기 때문에 본 연구의 최적화 과정이 유사 연구에서 예측 모델 구축에 참고가 될 수 있을 것이다. 또한 모델의 코딩에 대한 설명을 추가하여 향후 연구 수행에 도움을 주고자 한다. 마지막으로, 부동산 관련 지수의 시계열 데이터를 활용하여 예측력을 높이고, 최적화 예측 모델의 성능을 확인하고자 한다. 본 연구에서는 분석 대상 범위를 부산광역시로 하고 입력지수는 지가지수, 전세가격지수, 부동산 소비심리지수로, 출력지수는 아파트 매매지수 분석하여 예측력을 분석하고자 한다. 이는 자료의 특성에 따른 예측정확도 차이를 알 수 있을 것이다.

제 3 장　부동산 지수별 데이터 분석

3.1. 분석 데이터 설명

머신러닝을 이용하여 다양한 부동산 지수를 예측하고 머신러닝의 활용성을 확인해 보고자 다음과 같은 데이터를 분석 대상으로 하였다. 우선, 머신러닝을 통한 예측은 활용하는 변수에 따라 예측력이 다르게 나타나는 특성을 가진다. 배성완과 유정석[3]은 머신러닝 모델 예측에서 다변량 변수의 예측력이 단변량 변수보다 우수하다는 분석을 제시하였다. 이에 본 연구에서는 머신러닝의 예측력을 높이기 위해 단변량 변수를 활용해 머신러닝을 통한 부동산 지수 예측을 진행하고자 하였다. 또한 데이터는 부산을 범위로 하였으며 부동산 지수의 종류로는 한국부동산원에서 제공하는 1) 아파트매매지수, 2) 지가지수, 3) 전세가격지수, 4) 부동산 소비심리지수를 활용하였다. 데이터 기간은 2014년부터 2023년까지 10년간의 데이터를 대상으로 하며 공간적 범위는 부산광역시로 한정하였다.



<그림 5> 부동산 관련 지수 그래프

위의 <그림 5> 부동산 관련 지수 그래프에서 각 데이터의 특징으로 지가지수는 선형적인 데이터 형상을 가지고 있고, 아파트매매지수, 전세가격지수는 비교적 선형적이지만 등락이 있는 데이터 형상을 가지고 있다. 마지막으로 부동산 소비심리지수는 가장 비선형적인 불규칙적인 데이터 형상을 가지고 있으며, 다른 지수의 움직임과는 다르게 3~6개월 선행적으로 움직이는 경향을 보이고 있다.

3.2. 기초통계 분석

3.2.1 데이터 분석 및 처리

데이터 처리 시, 장애가 발생할 것으로 예상되어, 한글 변수명을 영문 변수명으로 아래 <표 1>과 같이 변경하였다.

<표 1> 변수명 변경

변수명	변경 후 변수명
부산지가지수	Land Price
부산부동산소비심리지수	Consumption Sentiment
부산전세가격지수	Rental Price
부산아파트매매지수	Sales

각 변수의 데이터와 관련하여, count, mean, std, min, max 값은 아래 <표 2> 데이터 분석과 같으며, 데이터의 결측값은 없는 것으로 확인하였다.

<표 2> 데이터 분석

구분	Land Price	Consumption Sentiment	Rental Price	Sales
count	120.0000	120.0000	120.0000	120.0000
mean	85.7791	108.6175	92.1688	90.2845
std	10.2813	15.5037	5.4204	7.0514
min	70.1890	75.7000	83.6370	85.7590
max	100.0000	135.9000	103.6220	105.8120

3.2.2 왜도 및 첨도 확인

머신러닝에서는 입력변수(feature) 별 값이 정규분포를 따르는지 체크하는 것이 매우 중요하다. 이를 간접적으로 확인할 수 있는 통계량으로 왜도와 첨도가 있다. 이를 확인하기 위해 Python 언어에서 skew와 kurtosis 명령어로 확인하였다.

<표 3> 왜 도

변수명	Value
Land Price	-0.0603
Consumption Sentiment	-0.1325
Rental Price	0.6974
Sales	0.9479
dtype : float64	

<표 4> 첨 도

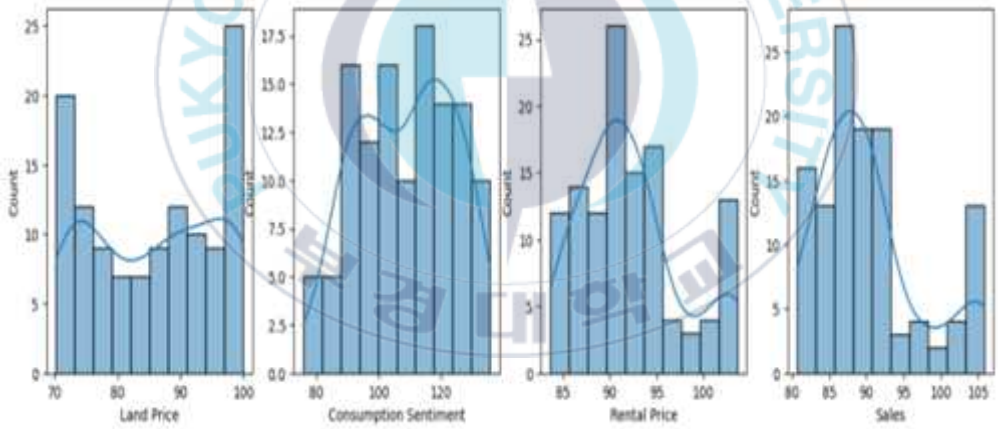
변수명	Value
Land Price	-1.4451
Consumption Sentiment	-1.0512
Rental Price	-0.2415
Sales	0.0639
dtype : float64	

위의 <표 3> 왜도 및 <표 4> 첨도의 값을 확인하였다. 허용 가능한 범위인 왜도 ±3 이하, 첨도 ±10 이하를 초과하는 변수들이 없는 것으로 확인되었다.

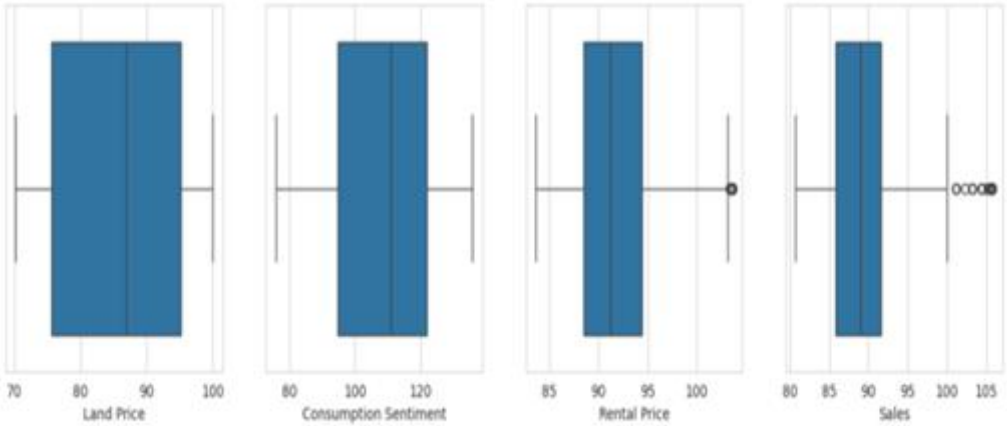
3.2.3 이상값 확인

구간 변수별 히스토그램을 이용하여 이상값을 확인하였으며, 확인방법은 Python 언어에서 맷플롯립(matplotlib) 라이브러리와 시본(seaborn) 라이브러리를 활용하였다.

아래 <그림 6> 분포도 및 <그림 7> 상자 그림에서 확인한 바와 같이 각 변수별 데이터값이 전체적으로 상한값과 하한값에서 특별하게 문제가 있는 이상 값이 존재하지 않는 것으로 확인되었다.



<그림 6> 부동산 관련 지수 분포도



<그림 7> 부동산 관련 지수 분포 상자그림

3.2.4 상관계수 검토

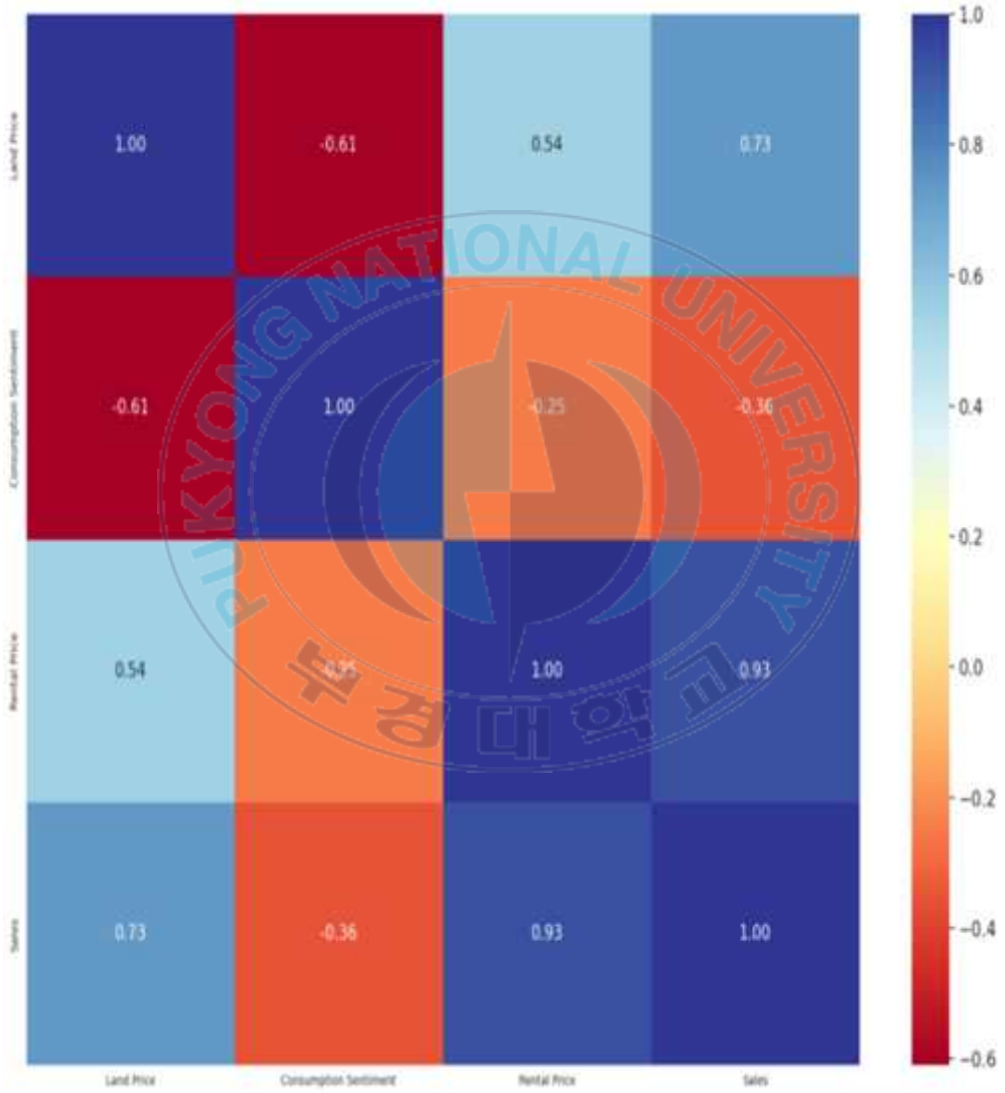
본 연구의 각 변수 간 상관계수를 확인하였다. 상관계수는 두 변수의 신형 종속성(linear dependence)을 나타내는 계수로서, -1과 1사이의 값을 갖는다. 한 변수의 값이 커질 때 나머지 변수의 값도 커지는 것을 양의 상관관계라고 하고, 그 반대의 경우를 음의 상관관계라고 한다. 양의 상관관계가 커질수록 상관계수는 1에 가까워지고, 음의 상관관계가 커질수록 상관계수는 -1에 가까워진다.

<표 5> 변수 간 상관계수

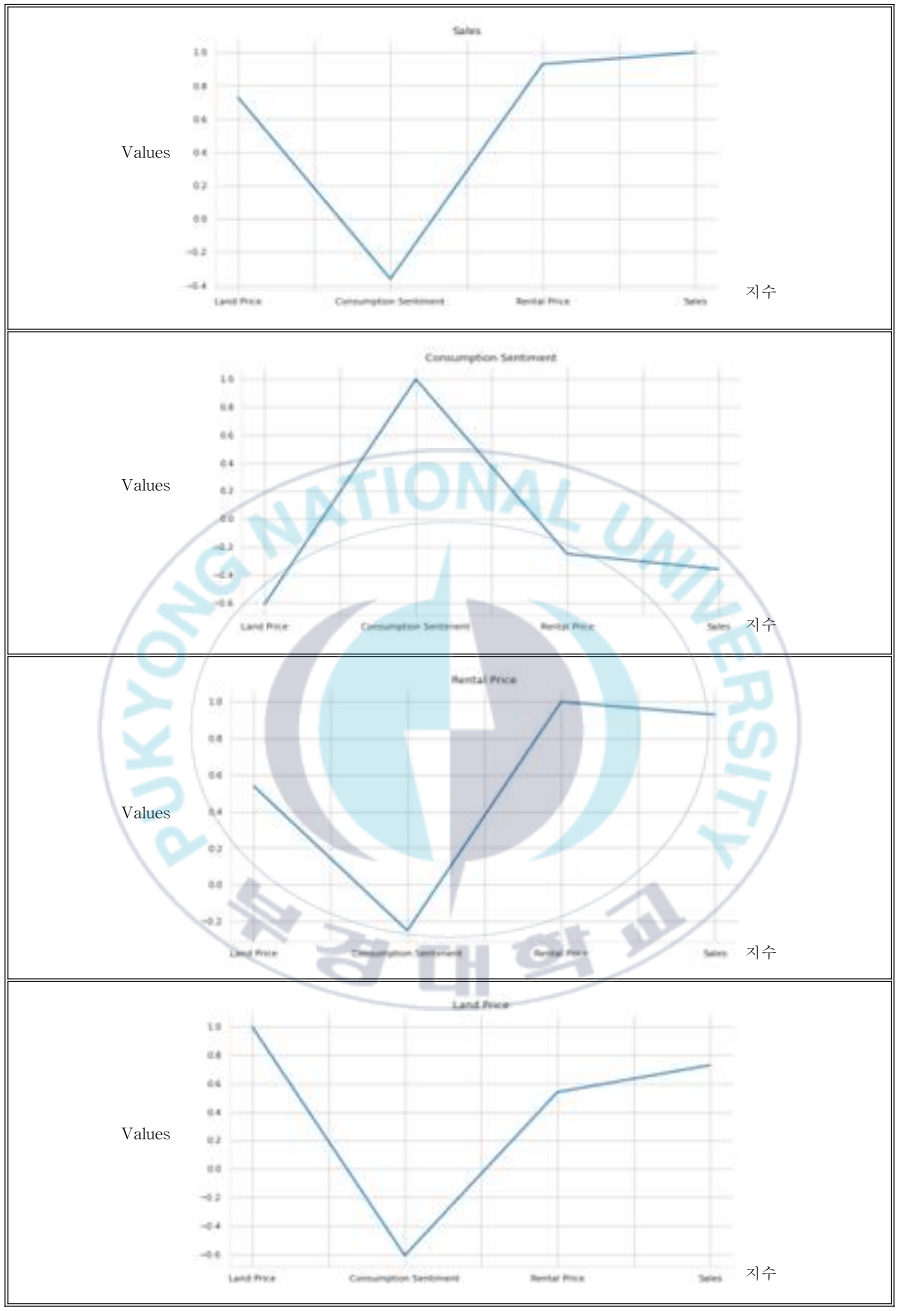
구분	Land Price	Consumption Sentiment	Rental Price	Sales
Land Price	1.0000	-0.6100	0.5400	0.7300
Consumption Sentiment	-0.6100	1.0000	-0.2500	-0.3600
Rental Price	0.5400	-0.2500	1.0000	0.9300
Sales	0.7300	-0.3600	0.9300	1.0000

위 <표 5> 변수 간 상관계수의 결과를 살펴보면, 부산 지가지수(Land Price)와 부산 전세가격지수(Rental Price), 부산 아파트매매지수(Sales)는 양의 상관관계를 나타내고 있다. 특히, 부산 전세가격지수(Rental Price)와 부산 아파트매매지수(Sales)는 매우 강한 양의 상관관계를 있으며, 부산 소비심리지수(Consumption Sentiment)는 각 변수들과 음의 상관관계를 가지는 특징이 있다.

이를 한눈에 이해하도록 아래 <그림 8> 변수 간 상관계수 도식 및 <그림 9> 각 변수별 Values를 더 자세히 확인할 수 있다.



<그림 8> 변수 간 상관계수 도식



<그림 9> 각 변수별 Values

제 4 장 부산 아파트 매매지수 예측

각 변수의 데이터를 인공 신경망 모델인 LSTM(Long Short Term Memory)에 대해 적용하여 모델을 활용하였다. 부산지가지수(Land Price), 부산부동산소비심리지수(Consumption Sentiment), 부산전세가격지수(Rental Price), 부산아파트매매지수(Sales) 등 총 4개의 변수 중 3개인 부산지가지수(Land Price), 부산부동산소비심리지수(Consumption Sentiment), 부산전세가격지수(Rental Price)를 독립변수로, 부산아파트매매지수(Sales)를 종속변수로 구성하여 연구를 진행하였다. 세부적인 연구방법은 Python언어에서 PyTorch를 활용하여 구현하였다. 이후, 데이터를 받고 로드하였으며, 변수 값 지정, LSTM Cell 네트워크 구축, 옵티마이저(Optimizer) 및 손실함수를 정의하였다. 모델의 옵티마이저(Optimizer) 결과는 다음 <표 6>과 같다.

<표 6> Optimizer 결과

Iteration	Loss	Accuracy
500	2.159	29.709
1000	0.701	79.230
1500	0.298	89.349
2000	0.111	94.279
2500	0.059	95.199
3000	0.072	96.279
3500	0.169	95.919
4000	0.143	97.120
4500	0.062	97.190
5000	0.043	97.230
5500	0.005	97.559
6000	0.087	97.400
6500	0.013	97.970
7000	0.031	97.680
7500	0.267	97.150
8000	0.113	97.769
8500	0.105	97.889
9000	0.015	98.349

모델 학습 및 성능 체크, 테스트 데이터셋 검증을 하였으며, 예측 결과는 아래 <표 7>과 같이 Test Loss는 0.06, Test Accuracy는 0.98로 측정되어 매우 높은 예측 성능을 보여주었다.

<표 7> 예측 결과

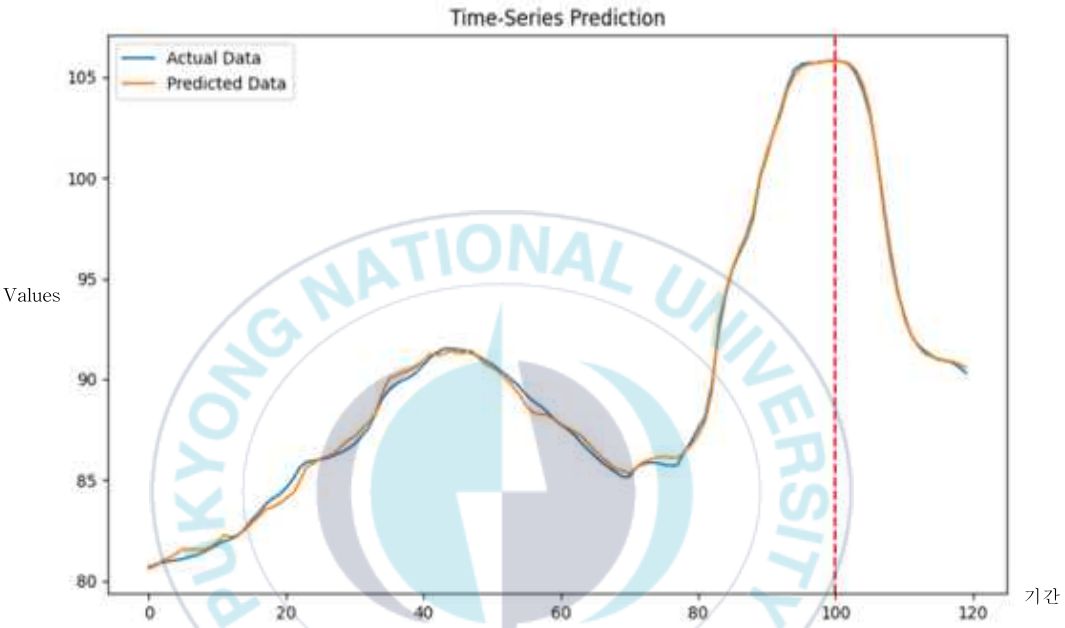
Test Loss	Test Accuracy
0.06	0.98

모델의 네트워크 정의 및 변수값을 설정하고 학습하였다. 아래는 <표 8> Epoch에 따른 Loss 결과이다. Epoch(인공신경망에서 전체 데이터 셋에 대해 학습한 forward pass, backward pass 과정을 거친 것으로, 전체 데이터 셋에 대해 한번 학습을 완료한 상태)가 900 및 1000에서 Loss는 0.00011의 값이 나왔다.

<표 8> Epoch 결과

Epoch	Loss
0	0.18388
100	0.00079
200	0.00024
300	0.00017
400	0.00015
500	0.00014
600	0.00013
700	0.00012
800	0.00012
900	0.00011
1000	0.00011

최종적으로 모델을 학습 및 예측결과, 아래 <그림 10>과 같이 예측력이 매우 우수한 것으로 결과가 나타났다.



<그림 10> 예측 결과

제 5 장 결 론

5.1. 연구결과 요약

부산 지가지수(Land Price), 부산 부동산소비심리지수(Consumption Sentiment), 부산 전세가격지수(Rental Price)를 입력변수, 부산 아파트매매지수(Sales)를 출력변수로 구성하여, 인공신경망 모델인 LSTM(Long Short Term Memory) 모델을 이용하여 연구한 결과, 시계열의 부동산지수 예측 성능의 정확도가 매우 높게 나타났다.

5.2. 연구의 한계 및 향후 연구과제

본 논문의 변수는 부동산 관련 변수인 부산 지가지수(Land Price), 부산 전세가격지수(Rental Price), 부산 아파트매매지수(Sales), 부산 부동산소비심리지수(Consumption Sentiment)로만 구성하였다. 이에 따라, 변수들의 숫자가 적고, 이들 변수들의 상관관계가 매우 높아 연구 결과에서 예측의 정확도가 매우 높았을 것으로 생각된다.

향후에는 부산시 구별 구분 연구와 원유가격, 주가, 달러 환율, 물가상승률 등의 거시적 및 미시적 경제지표를 더 많은 변수로 활용하여 연구하면 좀 더 풍부한 연구 결과를 도출할 수 있을 것이라고 생각된다.

또한 본 연구에서 다루지 않은 랜덤 포레스트(RF, Random Forest) 및 XGBoost(eXtreme Gradient Boosting) 등의 머신러닝 모델을 활용하여 진행 및 비교 연구할 필요성을 것으로 생각된다.

참고문헌

[1] 문권순. (1997). “벡터자기(VAR)모형의 이해”, 통계분석연구 제2권 제1호,pp. 23-56.

[2] 이태형. (2019). “인공신경망을 활용한 주택가격지수 예측에 관한 연구: 서울 주택가격지수를 중심으로”, 중앙대학교대학원, 박사학위논문.

[3] Jain, K. & Payal. (2011). “A review study on urban planning & artificialintelligence”, International Journal of Soft Computing and Engineering(IJSCE) Vol.1 No.5, pp.101-104.

[4] 국가통계포털. (2024).『인구총조사』

[5] 한국고용정보원. (2024).『지방 소멸 2024 : 광역 대도시로 확산하는 소멸 위험』보고서.

[6] 한국부동산원. (2023). 통계정보 보고서.

[7] 국토연구원. (2020). 부동산시장 소비자 심리조사 지침서.

[8] 이종열, & 이원곤. (2010). 부동산 가치형성에 미치는 요인에 관한 연구.한국정책연구, 10(3), 285-299.

[9] 김경환, 이한식. (2000). “부동산 가격 거품과 가격 전망”,대한부동산학회지 제18권, pp. 59-81.

[10] 강병기, 김선주, 신광식, 이국철. (2011). “부동산투자 및 시계량·전략적 접근방법”, 박문사.

[11] 김용창, & 이성호. (2010). 미래예측기법을 이용한 주택의 미래분석.지리학논총.

[12] 김종욱. (2004). “엑셀을 활용한 경영의사결정”, 박영사.

[13] Uysal, M., & Crompton, J.L. (1985). “An overview of approaches used toforecast tourism demand”, Journal of Travel Research Vol. Spring, pp.7-15.

[14] 박성훈. (2020). 머신러닝을 이용한 서울특별시 부동산 지수 예측 모델 비교, 석사학위논문, 한양대학교, 서울.

[15] 박헌수, 안지아. (2009). “VAR 모형을 이용한 부동산가격 변동 요인에 관한 연구”, 부동산연구 제19권, pp. 27-49.

[16] 조신섭, 손영숙, 성병찬. (2015). “SAS/ETS를 이용한 시계열 분석 4판”, 율곡출판사.

[17] Sims, C. A. (1980). Macroeconomics and reality. Econometrica: journal ofthe Econometric Society, 1-48.

[18] 김동환. (2015). VECM 모형을 이용한 주택시장과 거시경제변수 관계 분석. 대한부동산학회지, 33(2), 179-203.

[19] 이태형, & 전명진. (2018). 딥러닝 모형을 활용한 서울 주택가격지수예측에 관한 연구: 다변량 시계열 자료를 중심으로. 주택도시연구, 8(2),39-56.

[20] 문성은, 장수범, 이정혁, 이종석. (2016). “기계학습 및 딥러닝 기술동향”.정보와 통신, 33(11), pp. 49-56.

[21] 민성욱, & 서충원. (2017). 부동산학 연구 흐름과 특성 분석. 부동산학보,69, 102-115.

[22] 이창로. (2015). 비모수 공간모형과 앙상블 학습에 기초한 단독주택가격추정 (Doctoral dissertation, 서울대학교 대학원).

[23] 배성완, 유정석. (2018). “머신 러닝 방법과 시계열 분석 모형을 이용한부동산 가격지수 예측”, 주택연구 제26권 제1호, pp. 107-133.

[24] 장영재. (2018). “주택매매가격 영향요인의 비선형적 효과 분석”,한국자료분석학회지 제20권 제6호, pp. 2953-2966.

[25] 나성호, & 김종우. (2019). 공공데이터를 활용한 아파트 매매 가격 결정모형의 예측능력 비교: 서울 강남구 지역을 중심으로. 한국지적정보학회지,21(1), 3-12.

[26] 배성완. (2019). "머신 러닝을 이용한 주택 가격 예측력 비교", 단국대학교대학원, 박사학위논문.

[27] 김인호, & 이경섭. (2020). 트리 기반 앙상블 방법을 활용한 자동 평가 모형 개발 및 평가: 서울특별시 주거용 아파트를 사례로.한국데이터정보과학회지, 31(2), 375-389.

[28] 박성호. (2019). 기계 학습을 이용한 아파트 가격결정요인 분석,석사학위논문, 동아대학교, 부산.

[29] Brownlee, J. (2017). “Long short-term memory networks with python”,Machine learning mastery (Ebook Edition: v1.2).

[30] Hepsen, A., & Vatansever, M. (2011). “Forecasting future trends inDubai housing market by using Box-Jenkins autoregressive integratedmoving average”, International Journal of Housing Markets and AnalysisVol.4 No.3, pp.210-223

[31] 김동환. (2014). “ARIMA 모형을 이용한 주택시장의 가격 예측 분석”, 대한부동산학회지 제32권, pp. 277-294.

[32] 임성식. (2016). “주택가격지수 모형의 비교연구”, 한국데이터정보과학회지제27권 제6호, pp. 1573-1583

[33] 이영수. (2014). “단일 변수 시계열 모형들의 주택가격지수 예측력 비교”,부동산학연구 제20권 제4호, pp. 75-94

[34] 민성욱. (2016). “딥 러닝을 이용한 주택가격 예측 모형 연구”, 강남대학교대학원, 박사학위 논문.

[35] Cao, Q., Ewing, B.T., & Thompson, M.A. (2012). “Forecasting windspeed with recurrent neural networks”, European Journal of OperationalResearch Vol.221 No.1, pp. 148-154.