ISSN 2733-6239(Print) ISSN 2733-6247(Online)

https://doi.org/10.5659/JAIK.2024.40.1.35

Transformer와 LSTM을 순차적으로 적용하여 분석한 주택 가격 지수 예측 연구
- 문제점 분석 및 해결 전략을 중심으로 -

Investigating Problem Analysis and Solution Strategies in Predicting Housing Price Index Through Sequential Application of Transformer and LSTM Models

윤 도 경* 신 동 윤**
Yoon, Do-Kyung Shin, Dong-Yoon

* 단국대 건축학과 석사과정, M.A Candidate, Dept. of Architecture, Dankook University, Korea
** 단국대 건축학과 교수, Associate Professor, Dept. of Architecture, Dankook University, Korea
(Corresponding author : sindy@dankook.ac.kr)

Abstract

This study aims to review previous research on factors affecting the housing price index and construct a prediction model for the index using Long Short-Term Memory (LSTM) and Transformer models. Specifically, it combines LSTM, specialized in processing time-series data, and DistilBERT, specialized in handling text data, to utilize both historical housing price index data and relevant news articles. The experimental results of the proposed model confirmed significant accuracy when comparing predicted values in each region (J, S, G) with the actual values. However, some clusters displayed relatively high errors, indicating a need for additional analysis and improvement. Additionally, it was observed that subjective elements could significantly impact the interpretation of clustering results, highlighting the necessity for further analysis. Result visualization and statistical analysis were conducted, confirming their accurate reflection of housing price fluctuation trends in each region. This study introduces a novel approach to predicting the housing price index using deep learning models like LSTM and DistilBERT, providing valuable insights into real estate market trend predictions. The approaches and findings from this research are anticipated to provide valuable starting points for further exploration of creative solutions and the development of effective problem-solving strategies.

키워드 : LSTM, Transformer, 주택 가격 예측, 문제점 분석, 해결 전략, 머신러닝, 예측 모델링, 오차 분석
Keywords : LSTM, Transformer, Housing Price Prediction, Problem Analysis, Solution Strategies, Machine Learning, Predictive Modeling, Error Analysis

1. 서 론

주택은 주거서비스를 제공하는 내구재로서 역할을 함과 동시에 가치저장의 수단으로서의 기능을 갖추고 있다. 이로 인해 2000년대부터 선진국 포함 많은 국가들의 부동산, 주식 등 자산가격의 급등락 현상 과 거시경제의 안정화에 대한 문제가 대두되어 왔다. 이에 따라 학계에서는 자산가격의 변동 원인에 대한 연구가 다양하게 이루어지고 있으며(Lee, 2023), 또한 정부에서는 부동산정책을 경제정책을 위한 중요한 수단으로 활용하고 있으며, 부동산 가격에 영향을 미치는 요인을 분석하려는 연구도 활발하게 이루어지고 있다(Jo & Kim, 2013). 주택가격은 근본적으로 주택의 수요와 공급에 의해 결정되지만 다양한 요인들이 복합적으로 작용하여 접근성, 주거 공간, 환경, 교통, 조망, 근린시설 등의 영향을 받으며(Cheon, 2007), 소비자의 심리지수요인 또한 많은 영향을 받고 있다(Kim, Fang, & Yoo, 2013). 다양한 시장 참여자들이 투자 및 거래 결정을 내릴 때 중요한 변수인 주택과 관련된 주요 지표들을 예측하는 연구는 활발하게 진행되고 있다. 이러한 모델링은 부동산 시장의 동향 및 가격 변동을 파악하고, 이에 대응하여 현명한 투자 및 거래 결정을 내릴 수 있는 유용한 자료로 활용될 수 있다.

국내 선행연구를 살펴보면 이처럼 주택과 관련된 다양한 요인들을 분석하고, 다양한 주택가격의 요인분석 및 예측을 시도하기 위해 전통적으로 시계열 자료를 바탕으로 VAR(Vector AutoRegressions), ARIMA(AutoRegressive Integrated Moving Average)등의 시계열분석모형(Time Series Analysis)을 사용했지만, 최근에 들어서 머신러닝

(Machine Learning)기법을 이용한 분석 및 예측모델이 많아지는 추세이다. 그러나 단일모형으로 분석 후 예측 및 비교하는 연구는 많이 진행되었지만, 여러 개의 다른 예측 모델을 결합하여 하나의 최종 모델을 구축하는 방법을 시도한 국내사례는 많지 않다.

이에 본 연구의 목적은 LSTM과 Transformer 모델을 활용하여 각각 시계열 자료와 비선형 자료의 분석을 수행하고, 이를 결합하여 주택 가격을 예측하는 과정에서 발생한 문제점 및 해결 전략을 체계적으로 분석하고자 한다. 이는 주택 가격 예측 모델 개발에 참여하는 초보 연구자들이 유사한 문제에 직면했을 때, 본 연구를 통한 인사이트를 바탕으로 효과적인 해결 방안을 모색하고 적용할 수 있게 돕는 것을 목표로 한다. 본 연구가 실험의 성공과는 별개로, 실험 과정에서 발생하는 다양한 문제들을 분석하고 이를 해결하기 위한 전략을 제시함으로써, 이와 같은 연구를 처음 시작하는 연구자들에게 실질적인 도움을 제공하고자 한다. 이를 통해, 실험의 성공 여부를 넘어서 실험 과정 자체의 이해와 그에 따른 문제 해결 능력을 향상시키는 데에 초점을 맞춘 연구를 진행하고자 한다. 정형데이터는 국민은행(한국부동산원)에서 발표하는 전국주택동향조사의 매매가격지수를 수집하였으며, 비정형데이터는 뉴스 RSS피드의 데이터를 웹크롤링(Web Crawling)을 사용하여 수집하였다.

본 연구의 구성은 다음과 같다. 2장은 관련된 선행연구를 고찰하였으며, 3장에서는 데이터의 수집과 전처리 과정에 대해 기술하였다. 4장에서는 실험 설계에 대해 기술하며, 5장에서는 실험 결과를 분석하였고, 6장에서는 연구의 한계와 일반화를 기술하였다. 마지막으로 7장에서는 본 연구를 진행하는 데 참조한 문헌들을 나열하였다.

2. 선행연구

본 장에서는 주택가격지수의 변동요인에 관한 연구 현황과 함께 머신러닝 기법을 활용한 주택 가격 지수 예측 등에 대한 선행연구를 기술한다.

2.1 주택가격지수의 변동요인 연구

조건부 변동성 모형인 GARCH 모형을 사용하여, 아파트 매매가격 지수와 아파트 전세 가격 지수를 대상으로 국내 주택 시장의 가격 변동성과, 변동성 결정 요인을 파악하고자 하였다(Kim & Yoo, 2014). 분석 결과, 변동성의 크기에 따라 높은 국면과 낮은 국면에서 영향을 주는 요인들은 서로 달랐음을 확인했으며, 주택 가격과 전세 사격 모두 변동성이 높은 국면에서 영향을 미치고 있음을 확인하였다.

기존의 연구에서는 주택시장의 상황과 각 지역의 고유한 특성 요인을 고려하여 주택 가격에 영향을 미치는 변수를 파악하고, 이를 바탕으로 지역별 영향 요인을 분석하는 방식이 제안되었다(Kim, Song, & Lee, 2010). 분석 결과, 수도권 지역에서는 ‘대졸이상인구비율’, ‘인구천명당 주택수’, ‘가구수’가 유일한 독립변수로 선택되었으며, 비수도권 지역에서는 ‘재정자립도’등 이 유의한 의미를 가졌음을 밝혔다.

주택시장에서 거시경제와 주택가격 간의 관계를 분석하기 위해, Ham & Son(2022)은 기존의 VAR 모형에 베이지안 추론을 적용하였다. 또한, 다수의 사전정보와 초모수(Hyper parameters) 조합을 고려한 Bayesian VAR 모형이 주택가격 예측력을 향상시키는 것을 평가하였다. 분석 결과, 충격반응함수를 살펴보았을 때 이자율 상승은 매매가격과 전세가격을 하락시키고 물가와 생산의 상승은 매매가격과 전세가격을 상승시키는 것으로 밝혔다.

주택 매매 가격에 영향을 미치는 요인을 찾아내어, 장기 및 단기 예측에 활용 가능한 계량경제모형을 구축하였으며, 정부의 금리정책, 금융 기관의 의사결정, 그리고 경제 여건 변화가 지역 주택 매매 가격에 어떠한 영향을 미치는지 분석한 결과는 Yoon, Son, & Lee(2016)의 연구에서 확인하였다. 분석 결과, 한국은행의 기준금리, 예금은행의 주택대출액, KOSPI 종합지수 등의 전국 수준의 변수들과 지역경제 수준의 변수들이 지역주택매매가격과 공적분 관계에 있는 것을 발견하였으며, 장기적으로 모든 설명변수들이 통계적으로 유의한 영향을 미치는 것을 확인하였다.

Table 1. Summary of research on factors affecting housing price index

Author	Analysis Method	Analysis Variables
Kim & Yoo(2014)	GARCH, ARDL, MSAR, Markov Switching AR	Inflation rate, Employment growth rate, Population growth rate, Interest rate growth
Kim, Song, & Lee(2010)	Multiple Regression Analysis	Including real estate, Number of households, Population density, Number of marriages per 1,000 people, and 14 others
Ham & Son(2022)	Bayesian VAR	Consumer prices, Interest rate, Production, Jeonse price, Sales Price
Yoon, Sohn, & Lee(2016)	PanelUnitRoot Test, Panel Cointegration Test and Relationship Estimation	Interest rate, Stock price, Housing loan, Economic conditions, Consumer prices, Housing guarantee, Jeonse price, Employment index, Number of households, Elderly population proportion

2.2 머신러닝 기법을 이용한 예측연구

Chun & Yang (2019)는 딥 러닝 알고리즘의 다양성을 인지하면서도, 주택 가격 예측에 초점을 맞추었다. 시계열 예측에 적합하다고 알려진 RNN, LSTM, GRU 등의 모델들을 사용하여 예측력을 비교 분석하였으며, 이를 통해 가장 우수한 예측 성능을 보여주는 모델을 식별하고자 했다.

서울 아파트 가격 지수 예측 모델을 구축하기 위해 RNN 및 LSTM 알고리즘을 활용하였고, 이는 ARIMA 모델과 비교 분석한 결과는 Lee(2019b)의 연구에서 확인하였다. 분석 결과, 기존 시계열 모형보다 더 높은 예측력을 가지고 있음에도 불구하고, 기계학습을 위한 충분한 자료 확보 및 객관적 기준 부재로 인한 최적 모형을 찾는데 한계가 존재한다는 분석결과를 제시하였다.

2017년 1월부터 12월까지 서울시 강남구, 경기도 고양시 일산서구의 실거래자료를 활용하여 부동산 가격에 대한 예측에 있어서 머신 러닝의 활용 가능성을 검토한 결과는 Bae(2019)의 연구에서 확인되었다. 분석 결과, 머신러닝 방법이 전통적인 주택 가격 평가 모형보다 더 우수한 예측력을 보임을 확인하였고, 부동산 가격 평가에 있어서 높은 활용성을 가지고 있는 것으로 판단하였다. 부동산 가격은 개별성(個別性)이 강하고 여러 가지 데이터화하기 어려운 요인들에 의해 영향을 받고 있기 때문에 모든 부동산 가격 추정에서 머신러닝이 의미 있는 성과를 보여줄 수 있을 것으로 단정하기는 어렵다고 밝혔다.

2006년 1월부터 2018년 12월까지 서울 수도권의 아파트 시장 데이터를 분석하여 비슷한 가격과 가격변화 패턴을 보이는 수 개의 하위시장으로 나누고, 각 하위시장의 미래 가격변화를 예측하기 위해 시계열 군집분석과 인공신경망 모형을 사용한 결과는 Lee(2019a)의 연구에서 확인되었다. 분석 결과, 2018년도 말 이후 주택가격과 거시경제 지표와의 관계에 있어 실제 값과 예측 값과 사이의 차이가 상대적으로 크게 관측되었다고 하였다.

아파트 매매실거래가격지수 데이터로 시계열분석 모형과 머신 러닝 방법을 활용하여 부동산 가격지수 예측력을 비교 분석한 결과, 시계열분석모형(ARIMA, VAR, BVAR)보다 머신러닝 모형(SVM, RF, GBRT, DNN, LSTM)의 예측력이 더 우수하며, 시계열분석 모형의 경우 선형 모형을 가정하기 때문에 급변하는 비선형 형태의 시장을 예측하는 경우에는 한계점이 존재한다고 밝혔다(Bae & Yu, 2018).

전통적인 시계열 모델과 비교하여 머신러닝과 딥러닝이 주택가격 예측에 어떠한 효과를 미치는지에 대해 알아보기 위해, Milunovich(2020)는 호주의 주택 가격 예측을 중점으로 두고 있으며, 머신러닝(ML)과 딥러닝(DL) 알고리즘을 활용한 예측 능력을 탐구하였다. 분석 결과, 단변량 선형 모델이 단기적인 주택 가격 예측에 최적이며, 복잡한 비선형 모델이 장기적인 예측에 효과적임을 밝혔다.

Table 2. Summary of prediction research using machine learning techniques

Author	Analysis Method	Analysis Variables
Chun & Yang(2019)	RNN, LSTM, GRU	Apartment price, CD interest rate, Household loans, Building permit area, Consumer Price Index(CPI)
Lee(2019b)	ARIMA, VAR, VECM, RNN, LSTM	Apartment price index, Expected inflation rate, Apartment lease price index, Loan interest rate, Stock market index, Consumer Price Index(CPI), Unemployment rate
Bae(2019)	SVM, RF, GBRT, DNN, LSTM	Actual transaction apartment price index, Corporate bond yield, Money supply, Mining and manufacturing index, Lease price index
Lee(2019a)	Division Method, GAK, PAM, LSTM	Apartment price index, Seoul apartment price index, KOSPI index, 3-year government bond yield, Money supply growth, Inflation rate
Bae & Yoo(2018)	ARIMA, VVAR, BVAR, SVM, GBRT	Actual transaction apartment price index, Corporate bond yield, Consumer Price Index(CPI), Money supply
George (2020)	ARMA, VAR, SVR	Consumer Price Index(CPI), Real Gross Disposale Icome per Capita(Real GDI), Real Equities, Real Rent, Unemployment Rate, Real Mortgage Rate, Australians $)

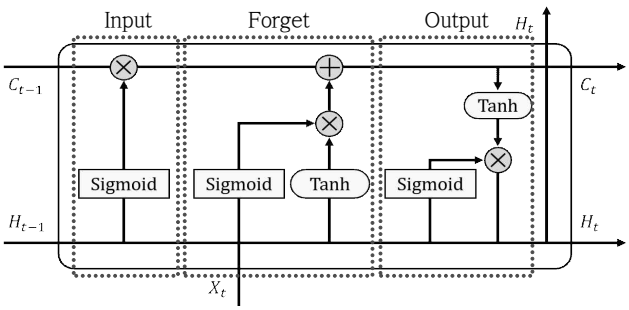


Figure 1. Structure of the LSTM model

LSTM(Long Short-Term Memory networks)는 RNN(Recurrent Neural Network)의 한 분류로, 과거 데이터가 미래의 결과에 영향을 주는 구조적 특징으로 인해 주식 및 주택가격과 같이 시계열 자료를 기반으로 한 예측에 탁월한 성능을 보이는 모델이다(Hong, 2020). 기존의 RNN은 계층의 수가 증가하면서 기울기가 불안정해지는 기울기 소실과 기울기 폭주의 문제가 발생한다. 해당 문제를 해결하기 위해 Hochreiter and Schmidhuber는 1977년에 LSTM 모델을 제기했으며, 여러 분야에서 정확 기록을 세운 바 있다. <Figure 1.>와 같이 LSTM은 3개의 주요 게이트를 가지며, Forget Gate 덕분에 학습 과정에서 기울기의 수렴을 조절하여 기울기 소실 또는 기울기 폭주 문제를 완화시키며, 동시에 장기적 기억도 유지할 수 있다(Yue, Kim, & Cho, 2023).

<Table 2.>를 보면 공통적으로 LSTM 모델을 연구에 적용한 것을 볼 수 있는데, 이는 다양한 예측연구에서 우수한 성능을 입증한 모델임을 확인할 수 있다.

2.3 Transformer Model

Vaswani et al.(2017)에서 Transformer는 2017년 구글에서 발표한 논문 “Attention is all you need”에서 제안된 자연어 처리 모델(NLP Model)로 <Figure 2.>와 같이 기존의 Seq2Seq 모델과 같이 인코더(Encoder)-디코더(Decoder) 구조를 따르면서도 순환 신경망(Recurrent Neural Network, RNN)을 사용하지 않고 오직 어텐션(Attention)만으로 구현된 모델이다(Han, 2021). 또한 내부 연산을 병렬 처리함으로써 순환신경망 보다 학습 속도가 빠르다는 이점을 가지고 있다(Ma, 2020).

2.4 선행연구 고찰

본 연구는 주택 가격 지수의 변동요인과 이를 예측하는 머신러닝 기법에 대한 선행연구들을 고찰하였다. Kim & Yoo(2014)와 Kim, Song, & Lee(2010) 등의 연구에서는 다양한 변수들이 주택 가격 지수의 변동성에 영향을 미친다는 결과를 도출하였다.

또한, 시계열 데이터를 처리하는 데 있어 LSTM(Long Short-Term Memory)와 같은 순환 신경망 계열의 알고리즘이 널리 사용되어 왔으며, 이러한 모델들은 시간적 순서 정보를 학습하는 능력으로 인해 주택 가격 예측 등에서 우수한 성능을 보여주었다.

그러나 LSTM 등의 구현 과정에서는 다양한 어려움이 발생할 수 있다.

이와 별개로 Transformer 모델은 RNN 없이 오직 Attention 메커니즘 만으로 입력 데이터의 시간적 순서 정보를 처리할 수 있으며, 병렬 처리 가능성으로 인해 학습 속도 면에서 장점이 있다. 그러나 Transformer 모델의 경우, 시계열 데이터 예측 문제에 대한 적용 사례와 검증 결과가 아직 부족하다.

이에 따라, 본 연구는 이전 연구들이 탐색한 여러 접근 방식과 그 한계점들을 바탕으로, 주택 가격 지수 예측에 실패하는 경우에 대해 실패 원인을 파악하기 위해 데이터 전처리 과정, 모델 학습 방법, 하이퍼파라미터 설정 등 다양한 요소들을 검토하며, 집중적으로 분석한다. 특히 실패 원인 파악 및 개선 전략 수립에 초점을 맞추어, 이를 통해 얻은 인사이트를 공유함으로써 다른 연구자들이 유사한 문제를 해결하는데 도움을 제공하려고 한다.

3. 데이터 수집 및 전처리

3.1 데이터 소스 및 설명

본 연구의 종속 변수인 주택 가격 지수는 국민은행(한국부동산원)에서 발표하는 전국 주택 동향 조사의 매매가격 지수를 사용하였다. 공간적 범위로는 전국, 수도권, 지방으로 나누어 수집하였다. 수도권은 서울, 인천, 경기도를 포함하며, 그 외의 지역은 지방으로 분류하였다. 유형은 아파트, 연립 및 다세대 주택, 단독주택을 통합한 수치로 집계하였다. 시간적 범위로는 2014년 06월부터 2023년 5월까지의 데이터를 사용하였다. 주택 가격 지수는 2021년 6월의 주택 가격을 기준으로 한 상대적 수치이며, 통계청에서 수집한 통계 자료를 연구 목적에 맞게 재구성하였다.

본 연구에서 사용한 독립 변수로서의 비선형 데이터는 ’Google Trend’에서 주택 가격과 가장 연관이 깊은 10개의 검색어를 추출하여 재구성하였다. 추출된 검색어에 대한 내용은 <Table 3.>에 기록되어 있다.

지상파 방송 3사(MBC, SBS, KBS) 중 MBC의 뉴스 데이터에서 해당 검색어가 포함된 뉴스의 RSS 피드를 수집하였다. 이는 MBC 웹페이지의 구조가 비교적 단순하여 웹 크롤링이 용이했기 때문이다. 수집된 RSS 피드를 바탕으로 웹 크롤링을 통해 자연어 데이터(Date, Title, Description, SearchTitle)를 추출하였으며, 이에 대한 내용은 <Table 4.>에 기재되어 있다.

Table 3. Description of variable data

Category	Variable Name	Variable Description	From-To
Dependent Variable	HPI	Housing Price Index	2014. 06. ~ 2023. 05.
Independent Variables	JH	Joint Housing	
	PP	Public Price	
	JHP	Joint Housing Price	
	REP	Real Estate Price	
	RE	Real Estate	
	CH	Country House	
	CHP	Country House Price	
	PHP	Prefabricated House Price	
	PH	Prefabricated House	
	AP	Apartment Price	

Table 4. The number of variable data

Variable \ Name	No. of articles (No.)
	MBC
JH	3,525
PP	2,080
JHP	680
REP	8,660
RE	34,705
CH	915
CHP	95
PHP	0
PH	205
AP	6,860

3.2 텍스트 데이터 처리 및 클러스터링

(1) 텍스트 전처리

텍스트 데이터의 전처리는 자연어 처리 작업의 핵심적인 초기 단계이다. 뉴스기사 사이트로부터 수집된 텍스트 데이터는 주택 가격에 영향을 미칠 수 있는 다양한 요인

들을 반영하고 있으며, 이를 적절하게 처리하는 것이 중요하다.

첫째로, 본 연구에서는 불용어 제거 과정을 실시하였다. HTML 태그, 숫자, 구두점 등과 같은 불용어를 제거하여 모델이 핵심 내용에만 집중할 수 있도록 하였다.

둘째로, 토큰화 과정을 거쳤다. 본 연구에서 사용된 DistilBERT 모델에 적합하게 Hugging Face의 Transformers 라이브러리에 내장된 BERT tokenizer를 활용하여 문장을 개별 토큰으로 분리하였다. DistilBERT 모델에 대한 설명은 4장에서 계속한다. 마지막으로 정규화 과정을 진행하였다. 설정된 최대 길이를 초과하는 단어들은 잘라내었으며, 위와 같은 전처리 과정을 거친 결과 얻어진 정제된 데이터 셋은 후속 작업인 임베딩 추출 및 클러스터링에 활용되었다.

(2) k-means 클러스터링

k-means 클러스터링을 사용하여 문장 임베딩들을 그룹화 하였다. k-means 알고리즘은 분할 클러스터링 알고리즘 중에서 일반적으로 가장 많이 사용되는 알고리즘 중 하나이다(Lee & Lee, 2011).

scikit-learn 라이브러리의 KMeans 클래스를 사용하여 임베딩 벡터들을 클러스터링 하였으며 'K'값은 사용자가 정의하는 파라미터로, 본 연구에서는 10개의 클러스터를 생성하도록 설정하였다.

3.3 LSTM 모델 구성 및 초기화

LSTM 모델은 다양한 구조와 설정으로 구성될 수 있다. 본 연구에서는 하나의 LSTM 층과 함께 Dense(완전 연결) 층을 추가하여 간단한 아키텍처를 구성하였다. 편의상 전국권에 대한 분석은 J, 수도권에 대한 분석은 S, 지방에 대한 분석은 G로 통일하였다.

Dense 층은 출력 차원을 3으로 설정하였으며, 이는 J, S, G 각각에 대한 예측 값을 생성하는 역할을 수행한다. LSTM 층은 시간적인 의존 관계를 학습하기 위해 사용되었으며, Dense 층은 LSTM 층의 출력을 기반으로 최종 예측 값을 생성하는 역할을 수행한다. 모델의 입력 데이터로는 각 뉴스 기사의 클러스터 ID와 주택 가격 지수 데이터가 함께 사용되었다. 클러스터 ID는 범주형 변수로서 원-핫 인코딩 방식으로 변환되어 입력으로 사용되었다. 이렇게 함으로써 각 기사가 어느 클러스터에 속하는지 정보가 모델에 반영되었다.

모델의 초기화 과정에서는 Xavier 가중치 초기화(Glorot & Bengio, 2010) 방법을 적용하여 모델의 성능을 향상시켰다. 이 방법은 신경망의 가중치를 적절한 값으로 설정함으로써 처음에 학습 속도를 높이고 최적의 솔루션에 수렴할 가능성을 높이는데 도움을 준다.

4. 실험 설계

본 장에서는 실험 설계의 내용을 기술하며 <Figure 2.>는 전체적인 실험의 흐름을 나타내고 있다.

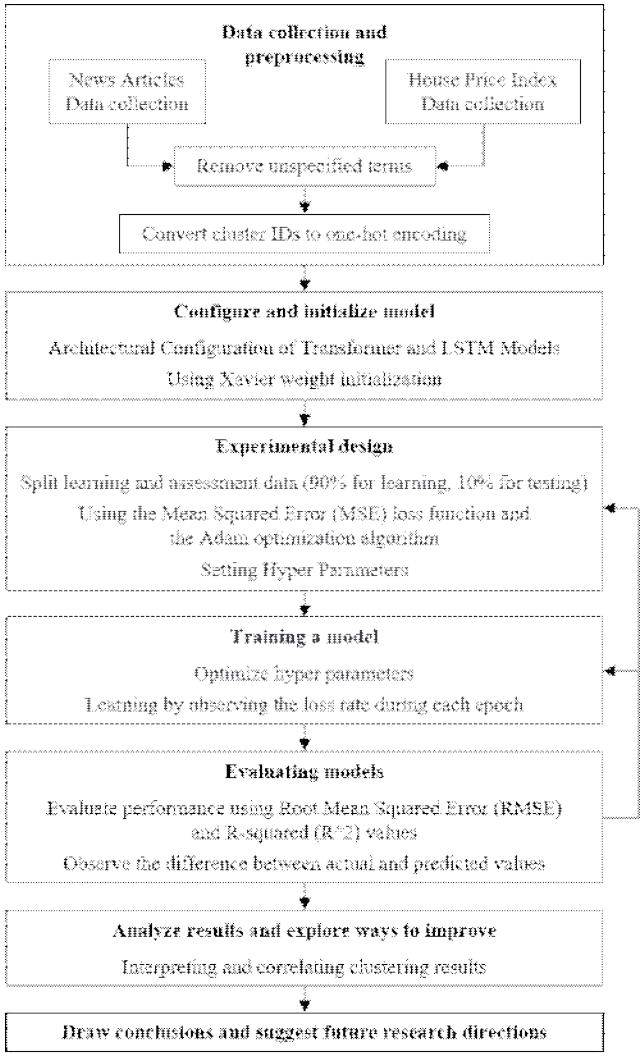


Figure 2. Flowchart of the experiment

4.1 모델 및 알고리즘

본 연구에서 사용된 LSTM과 DistilBert 모델은 각각 시계열 데이터 처리와 텍스트 데이터 처리에 특화된 딥러닝 모델이다. DistilBert는 Transformer 기반의 언어 처리 모델로, Google의 BERT(Bidirectional Encoder Representations from Transformers)를 경량화 하여 학습 속도를 개선하고 메모리 사용량을 줄인 버전이다. Transformer는 입력 데이터의 시퀀스 간 상호 작용과 종속성을 학습하는 데 특히 유용한 구조로, 자연어 처리(NLP) 작업에서 널리 사용된다. DistilBERT와 LSTM을 선택한 이유는 다음과 같다.

첫째, 원래 BERT와 비교하여 학습 속도가 빠르며 성능 저하가 상대적으로 적기 때문이다. 둘째, Transformer 기반으로 문장 내 단어들 사이의 문맥적 관계를 잘 파악하며, 복잡한 자연어 정보를 벡터 형태로 잘 추출할 수 있기 때문이다(Sanh et al. 2019).

둘째, 주택 가격 지수처럼 과거 패턴과 미래 추세가 중요한 시계열 데이터에 대해 LSTM은 매우 적합하므로 LSTM을 선택하였다.

DistilBert와 LSTM을 결합하여 수집된 텍스트 데이터와 과거 주택 가격 지수 데이터를 활용하여 주택 가격 지수 예측 모델을 구축하였다.

4.2 데이터 분할 및 학습/평가 방법

주택 가격 지수 예측 모델을 훈련하고 평가하기 위해 사용된 데이터를 분할하는 방법과, 이를 바탕으로 진행된 학습 및 평가 과정에 대해 설명한다.

데이터는 총 10년의 범위에서 전체 데이터 셋의 90%를 훈련용으로, 나머지 10%를 테스트용으로 분할하였다. 이러한 비율은 일반적인 머신러닝 프로젝트에서 널리 사용되는 비율로, 충분한 양의 훈련 데이터와 함께 모델의 실제 성능을 검증하기 위한 적절한 양의 테스트 데이터를 확보할 수 있게 해준다. 모델 학습 과정에서는 Mean Squared Error(MSE) 손실 함수와 Adam 최적화 알고리즘을 사용하였다. MSE 손실 함수는 예측 값과 실제 값 사이의 차이(오차)를 제곱하여 평균을 도출한 값으로, 회귀 문제에서 자주 사용된다(Jo & Park, 2023). Adam 최적화 알고리즘은 경사 하강법을 기반으로 하며 효과적인 최적화 접근 방법 중 하나이며, 학습률 파라미터가 어떻게 조정되어야 하는지 자동으로 결정한다(Jais, Ismail, & Nisa, 2019).

4.3 Hyper Parameter 설정과 최적화

<Table. 5>를 보면 총 5개의 Hyper Parameter 값을 설정하였음을 알 수 있다.LSTM 모델의 입력 크기를 결정하는 'input_size'는 스케일링된 학습 데이터 셋의 형태에 기반한다. 또한, LSTM 셀 내부의 hidden state와 cell state의 차원 수를 정의하는 'hidden_size'는 모델이 학습하는 능력과 밀접한 관련이 있다. 과도하게 큰 hidden_size는 overfitting을 초래할 위험이 있으며, 반대로 과소한 값은 underfitting 문제를 야기할 수 있다. 세 가지 주택 가격 지수('J', 'S', 'G') 값을 동시에 예측하기 위해, LSTM Model에서 출력 차원인 'output_dim' 파라미터를 3으로 설정하였다.

모델 학습 과정에서 한 번에 처리하는 데이터 샘플의 수인 'batch_size' 역시 중요한 역할을 한다. 이 값은 메모리 사용량과 학습 속도, 그리고 최적화 품질 간의 균형을 맞추는 데 필요하다. Adam 최적화 알고리즘에서 사용된 학습률은 각 epoch 반복마다 가중치 업데이트 양을 결정한다. 이 값이 과도하게 높으면 최소 손실 값을 지나칠 위험이 있으며, 반대로 너무 낮으면 학습 속도가 느려질 수 있다.

학습 과정은 총 200회(epoch) 동안 진행되었다. <Figure 3.>를 보면 Epoch이 진행됨에 따른 Loss rate를 보여주고 있다. 검증 셋을 사용하여 모델 성능을 평가하였으며, 가장 낮은 Loss Rate를 보인 Hyper Parameter 값을 선택하였다. 모델 성능 평가에는 Root Mean Squared Error (RMSE), 그리고 R-squared (결정 계수) 값을 사용하였으며, RMSE는 실제 값과 예측 값 사이의 오차 크기를 적절히 반영하는 지표이다.

Table 5. Value of hyper parameter

Hyper parameter	Value
Hidden Size	50
Learning Rate	0.001
Batch size	256
Epoch	200
output_dim	3

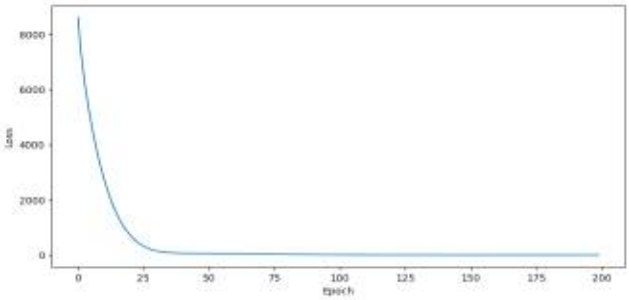


Figure 3. Loss rate of LSTM

5. 실험 결과

5.1 주택 가격 지수 예측 모델의 성능 평가

모델의 성능 평가를 위해 Root Mean Squared Error (RMSE)와 R-squared(R^2)를 채택하였다. RMSE는 실제 값과 예측 값 사이의 차이를 제곱하여 평균화한 후, 루트를 씌운 값으로, 이 값을 최소화하는 방향으로 모델을 학습시켰다. R^2는 회귀 분석에서 종속 변수의 총 변동성 중에서 독립 변수가 설명하는 변동성의 비율을 나타내며, 1에 가까울수록 데이터를 잘 설명하는 모델임을 의미한다.

<Table 6>를 보면 학습된 모델은 테스트 데이터 셋에 대해 RMSE 값 J: 27.46, S: 26.60, G: 28.29와 R^2 값 J: -63.30%, S: -37.78%, G: -113.36%를 기록하였다.

R^2 값이 음수로 나타난 것은 모델이 주어진 데이터를 적절히 처리하지 못했다는 것을 보여주며, 이는 복잡한 패턴을 포착하기 위해 필요한 특성들이 누락되었거나, 모델 구조가 너무 단순하여 발생할 가능성이 높다. 이러한 한계점은 모델의 예측 정확도를 저하시키는 주요 요인으로 작용하였다. 이를 개선하기 위해, 더 많은 특성을 포함하거나 복잡한 패턴을 더 잘 포착할 수 있는 모델 구조를 탐색할 필요가 있으며, 하이퍼파라미터 값을 조정할 필요가 있다. 또한, 모델 훈련 과정에서 발생할 수 있는 과적합 문제를 방지하기 위한 방안도 고려해야 한다.

이에 본장에선 이 문제에 대응하기 위해 다수의 하이퍼파라미터 조정 실험을 진행하였다. hidden size, learning rate, batch size, epoch 및 output_dim 등의 파라미터를 하나씩 변경하면서 loss rate 줄일 수 있도록 값을 조정시켜 각각의 파라미터가 모델 성능에 어떤 영향을 미치는지 관찰하였다. 이 과정에서 loss rate는 안정적으로 점차 줄어들었다. 그러나 이런 개선에도 불구하고 RMSE 와 R^2의 값은 여전히 개선되지 않았다. 이러한 결

과는 모델이 아직 데이터 내부의 복잡성을 충분히 학습하지 못함을 시사한다. 따라서 추가적인 연구 방법론, 데이터 전처리 방법 개선, 다른 라이브러리 사용 등 다양한 접근 방식을 고려하는 것도 중요할 것으로 보인다.

Table 6. Values of metric

Category	Metric	value
J	RMSE	27.462244033813477
	R-squared	-63.29602581541168
S	RMSE	26.59834861755371
	R-squared	-37.78104621601104
G	RMSE	28.288055419921875
	R-squared	-113.36483416284044

5.2 결과 시각화 및 통계적 분석

연구에서 k-means 알고리즘을 활용하여 클러스터링을 수행한 결과, 각 클러스터 내 주요 단어들이 유사한 패턴을 보이는 것으로 관찰되었다. 특히 '%'와 '부동산'과 같은 단어가 대부분의 클러스터에서 빈번하게 등장하는 현상이 발견되었다. 이는 모든 뉴스기사가 부동산 관련 이슈에 중점을 두고 있음을 나타내며, 해당 단어들은 문장의 의미를 해석하는데 중요한 역할을 수행한다. 이러한 점은 클러스터링의 유용성을 보여주지만, 동시에 클러스터링 결과의 해석에 있어서의 한계점을 드러낸다. 더욱이, <Table 7.>과 같이 Cluster 4와 5에서는 '오늘'이라는 단어가 상대적으로 많이 나타나며, 이로부터 일일 부동산 동향에 대한 뉴스일 가능성을 추측해볼 수 있다. 하지만 초기 데이터 전처리 과정에서 기본 한국어 불용어 리스트를 활용하여 제거하였음에도 위 언급된 패턴이 나타난 것으로 보아 해당 단어들은 기본 불용어 리스트에 포함되지 않았다.

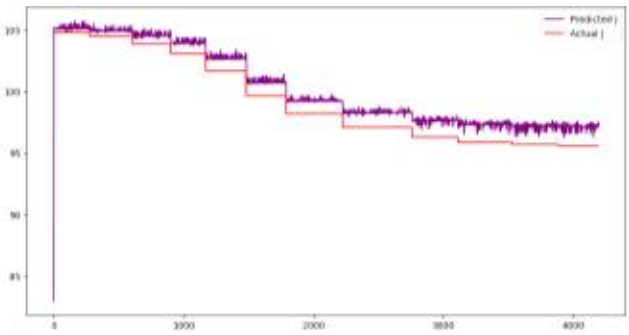
따라서 앞서 언급된 단어들은 문서의 주제 파악에 일부 도움이 되지만, 반대로 클러스터링 결과 해석과정에서 혼란 요소로 작용할 가능성도 있다. 본 연구의 목적과 데이터 특성 등 종합적인 요소를 고려하여 추가적인 불용어 정의 및 전처리 과정 개선의 필요성을 제언한다. 그럼에도 불구하고 클러스터링 결과를 해석하는 것은 주관적일 수 있으며, 특정 주제나 패턴을 찾기 위해서는 추가적인 분석이 필요하다. 이에 따라 본 연구의 클러스터링 결과는 초기 단계의 패턴 인식에 그치며, 보다 구체적인 주제 파악을 위해서는 더욱 세밀한 분석이 요구된다.

Table 7. Result clustering

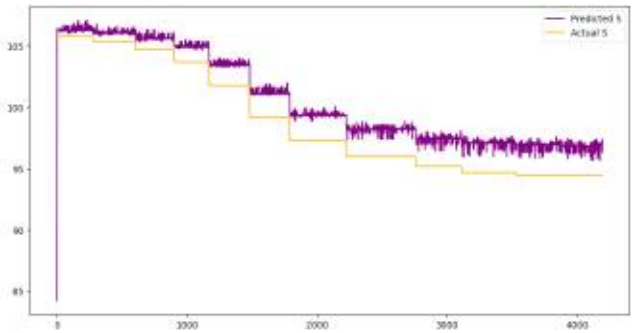
Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
%	%	%	%	%
2552	1773	1783	712	2701
있습니다	‘	‘	것으로	부동산
1776	745	537	532	1564
앵커	앵커	부동산	‘	것으로
1680	730	458	517	1137
부동산	수	것으로	부동산	‘
1649	503	435	463	1081
수	것으로	[	아파트	앵커
1333	496	294	388	1037
[	부동산	]	서울	밝혔다
1306	491	294	284	836
]	있습니다	아파트	따르면	있습니다
1306	454	273	255	829
‘	있는	서울	올해	오늘
1191	404	245	191	801
#	한	밝혔다	이후	수
1150	381	213	166	678
한	)	가장	?	아파트
1086	341	209	164	586

Cluster5	Cluster6	Cluster7	Cluster8	Cluster9
부동산	앵커	앵커	%	부동산
2552	1773	1783	712	2701
%	?	?	부동산	%
1776	745	537	532	1564
밝혔다	수	%	것으로	것으로
1680	730	458	517	1137
것으로	%	수	아파트	밝혔다
1649	503	435	463	1081
오늘	[	‘	서울	수
1333	496	294	388	1037
서울	]	있는	따르면	‘
1306)	491	294	284	836
‘	있습니다	있습니다	밝혔다	있습니다
1306	454	273	255	829
있습니다	있는	한	지난	오늘
1191	404	245	191	801
아파트	한	지금	‘	말했다
1150	381	213	166	678
수	‘	합니다	올해	아파트
1086	341	209	164	586

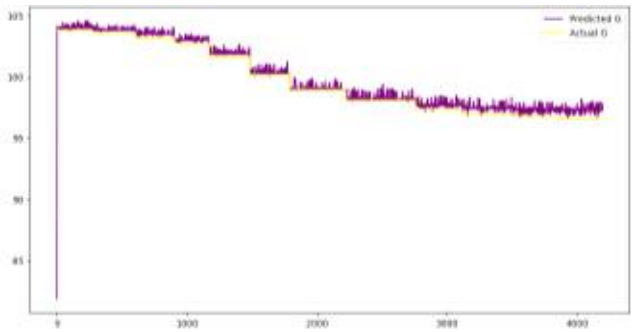
나머지 10%(1년)의 데이터 셋 동안의 주택 가격 지수 예측 값과 실제 값을 그래프로 비교하였다. 이 과정에서 J, S, G 각 지역에 대한 예측 그래프를 별도로 생성하였다. <Figure 4.>를 보았을 때 각 지역의 주택 가격 지수에 대한 예측 그래프와 실제 그래프를 비교해 보았을 때, 두 그래프는 전반적으로 유사한 추세를 보이거나, 큰 차이가 없다. 이는 모델이 각 지역의 주택 가격 변동 추세를 잘 반영하고 있음을 의미한다. 따라서 위 세 개의 지역권(J,S,G)에서 각각 생성된 예측 값과 실제 값 비교 그래프들은 제안된 모델이 주택 가격 지수 변동률을 상당 정확도로 예상할 수 있음을 확인시켜 준다.



a) Comparison of actual and predicted housing price index(J)



b) Comparison of actual and predicted housing price index(S)

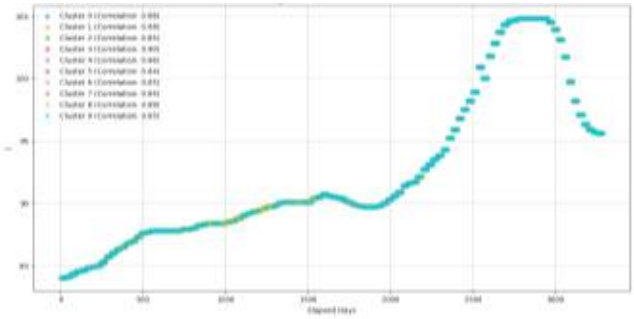


c) Comparison of actual and predicted housing price index(G)

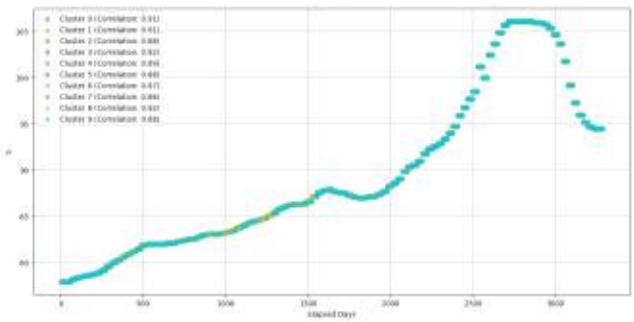
Figure 4. Comparison of actual and predicted housing price index (J and S and G)

또한 주택가격지수와 각 클러스터 간의 상관관계를 계산하고 시각화를 진행하였다. J, S, G 세 분류지역의 주택가격 지수 데이터에 대해 10개의 클러스터마다 상관 계수를 산출하였으며, 이를 그래프로 표현하여 상호간의 관계를 파악하고자 했다. 각 그래프에서 x축은 경과된 일자('Elapsed_days')을, y축은 해당 지역의 주택 가격 지수 값을 나타내며, 각 클러스터 내의 개별 데이터 포인트를 각 점으로 표현했고 각 클러스터를 다른 색상으로 표시하여 구분이 용이하게 하고자 했다. <Figure 5.>를 보면 그래프로만 보았을 땐 점들이 이어져서 실제 주택가격지수와 매우 유사한 그래프 형상을 띠고 있어 해당 클러스터가 주태가격지수의 변동률을 잘 반영하고 있음을 나타낸다. 다시 말해, 해당 클러스터에 속하는 뉴스 기사들의 내용이 주택가격지수에 큰 영향을 미치며, 해당 기사들은 부동산 시장 동향 예측에 중요한 정보를 제공함을 알 수 있다.

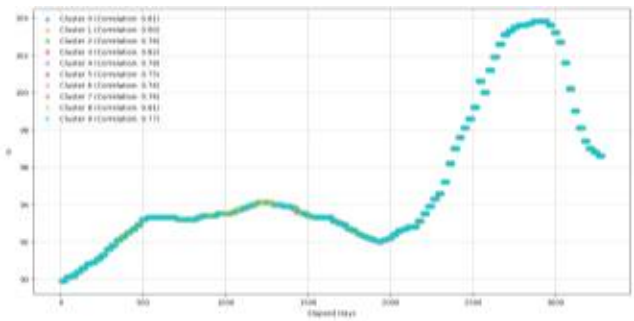
하지만 이는 클러스터링의 문제점을 보여준다. 점들이 겹쳐 보이는 현상은 클러스터 내의 데이터 포인트들이 비슷한 특성을 가지고 있다는 것을 의미할 수도 있지만, 이 연구에서는 이를 불완전한 클러스터링 결과로 해석한다. 즉, 데이터 포인트들이 너무 밀접하게 위치해 있는 것은 해당 클러스터가 너무 광범위하게 정의되었다는 것을 나타내며, 반대로 점들이 이어져 실제 주택 가격 지수와 유사한 형태를 띠는 것은 특정 클러스터가 과적합 되었다고 볼 수 있다. 이를 개선하기 위해, 클러스터링 알고리즘의 파라미터를 조정하거나 다른 클러스터링 방법을 탐색할 필요가 있다. 또한, 클러스터링 결과의 해석에 있어서의 한계점도 고려해야한다. 예를 들어, 클러스터링 결과가 실제로 어떤 의미를 가지는지 해석하는 것은 주관적인 요소가 포함될 수 있으므로, 이에 대한 명확한 기준 설정이 필요하다. 따라서, 이런 관찰 결과를 바탕으로 데이터 분석 방법론에 대한 개선 필요성을 제시한다.



a) Correlation of Housing Price Indices (J) with Elapsed Days across Clusters



b) Correlation of Housing Price Indices (S) with Elapsed Days across Clusters



c) Correlation of Housing Price Indices (G) with Elapsed Days across Clusters

Figure 5. Correlation of Housing Price Indices (J and S and G) with Elapsed Days across Clusters

6. 연구의 한계와 일반화

이 연구의 한계점을 더욱 명확하게 인식하는 것이 중요하다. 첫째, 사용된 모델의 성능이 완벽하지 않았다는 점을 지적할 수 있다. 특히, R^2 값이 음수로 나타난 것은 모델이 주어진 데이터를 적절히 처리하지 못했다는 것을 보여주며, 이는 복잡한 패턴을 포착하기 위해 필요

한 특성들이 누락되었거나, 모델 구조가 너무 단순하여 발생할 가능성이 높다. 이는 모델의 예측 정확도를 저하시키는 주요 요인으로 작용하였다.

둘째, 이 연구에서는 클러스터링의 문제점을 보여주고 있다. 점들이 겹쳐 보이는 현상은 클러스터 내의 데이터 포인트들이 비슷한 특성을 가지고 있다는 것을 의미할 수도 있지만, 이 연구에서는 이를 불완전한 클러스터링 결과로 해석하였다. 이를 개선하기 위해, 클러스터링 알고리즘의 파라미터를 조정하거나 다른 클러스터링 방법을 탐색할 필요가 있다.

셋째, 이 연구는 특정 지역(전국권, 수도권, 지방권)의 주택 가격 지수에 초점을 맞추었다. 이는 이 연구의 결과가 해당 지역들에 한정적일 수 있음을 의미하며, 각각의 지역이나 다른 시장 상황에서는 다른 결과가 나올 수 있다. 따라서, 이 연구의 결과를 다른 지역이나 다른 시장 상황에 그대로 일반화하거나 반영하는 것은 주의가 필요하다.

마지막으로, 이 연구는 주택 가격 지수에 영향을 미치는 다양한 요인들 중 뉴스 기사의 내용만을 분석하였다. 이는 뉴스 기사가 주택 가격 지수에 미치는 영향을 분석하는 데에 중요하지만, 다른 중요한 요인들을 무시한 채로 모델을 구축하였다는 한계점을 가진다.

이러한 한계점들을 고려하여, 이 연구의 결과를 일반화하거나 다른 상황에 적용할 때는 주의해야 한다. 이 연구의 한계점을 극복하고, 더욱 정확하고 신뢰성 있는 모델을 개발하기 위해서는 앞서 언급한 한계점들을 개선하는 추가적인 연구가 필요하다. 이는 더 복잡한 모델 구조의 탐색, 더 많은 특성의 포함, 다른 지역이나 시장 상황에 대한 고려, 그리고 주택 가격에 영향을 미치는 다양한 요인들을 고려한 모델 구축 등을 포함할 수 있다.

REFERENCES

1. Bae, S. W. (2019). *Forecasting Property Prices Using the Machine Learning Methods : Model Comparisons*, Ph. D. Dissertation, Dankook University.
2. Bae, S. W., & Yu, J. S. (2018). Predicting the Real Estate Price Index Using Machine Learning Methods and Time Series Analysis Model. *Housing Studies*, 26(1), 107-133.
3. Cheon, I. H. (2007). A Study on the Effect of Yang Tak Factor in the Housing Price- A Case of Haeundae New Town -. *Korean Association for Housing policy Studies*, 15(1), 99-126.
4. Chun, H. J., & Yang, H. S. (2019). A Study on Prediction of Housing Price Using Deep Learning. *Journal of The Residential Environment Institute of Korea*, 17(2), 37-49.
5. Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep forward neural networks. *In Proceedings of the International Conference on Artificial Intelligence and Statistics*, Society for Artificial Intelligence and Statistics.
6. Ham, J. Y., & Son, J. Y. (2022). Analysis of Housing Price Fluctuations Using Bayesian Approach. *Journal of The Residential Environment Institute of Korea*, 20(1), 1-21.
7. Han, C. J. (2021). *Study on Interpretable Transformer Model for Multi-step Stock Price Movement Forecasting*, Thesis, Seoul National University.
8. Hong, S. H. (2020). A Study on Stock Price Prediction System Based on Text Mining Method Using LSTM and Stock Market News. *Journal of Digital Convergence*, 18(7), 223-228.
9. Jo, M. S., & Kim, T. H. (2013). A Study on Characteristics of Hedonic Price Models Based on Meta-regression Analysis. *Journal of the Korean Data Analysis Society*, 15(5), 2765-2780.
10. Jais, I. K. M., Ismail, A. R., & Nisa, S. Q. (2019). Adam Optimization Algorithm for Wide and Deep Neural Network. *Knowledge Engineering and Data Science*, 2(1), 41-46.
11. Jo, L. J., & Park, W. S. (2023). Effects of Loss Functions on the Performance of NeRF. *Korean Institute of Next Generation Computing*, 202-205.
12. Kim, D. W. & Yoo, J. S. (2014). The Determinants of Housing Price Indices Volatility Using a Switching Regression Analysis. *Housing Studies*, 22(3), 69-99.
13. Kim, G. G., Song, H. C., & Lee, J. H. (2010). A Study on the Determinants of the Change Rate of Housing Price by Areas. *Review of Real Estate and Urban Studies*, 3(1), 101-115.
14. Kim, S. Y., Fang, X. Z., & Yoo, S. J. (2013). Causality Between Sales Price of Housing Index and Consumer Sentiment Index : Impact on Korea(South) and China. *The Journal of Modern China Studies*, 15(1), 175-210.
15. Lee, M. S. (2023). A Study on the Relation between Inflation Expectations and House Prices: Comparison between South Korea and the U.S.. *Journal of KREAA*, 29(1), 7-36.
16. Lee, S. W., & Lee, W. H. (2011). Refining Initial Seeds using Max Average Distance for K-Means Clustering. *Jounal of Internet Computing and Services*, 12(2), 103-111.
17. Lee, S. J. (2019a). Segmentation of Housing Submarkets and Housing Price Prediction Through Data Mining. *Journal of Environmental Studies*, 64, 176-177.
18. Lee, T. H. (2019b). *Prediction of Seoul House Price Index Using Artificial Neural Network,* Ph. D. Dissertation, Chung-Ang University.
19. Ma, B. H. (2020). *Natural Language Generation Using GAN and Transformer Models*, ph. D. Dessertation, Ajou University.

20. Milunovich, G. (2020). Forecasting Australia's real house price index: A comparison of time series and machine learning methods. *Journal of forecasting*, 1098-1118.
21. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108.
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
23. Yoon, S. M., Sohn, S. H., & Lee, J. I. (2016). Empirical Analysis on the Long-Run and Short-Run Determinants of Regional House Price Dynamics. *Korea Real Estate Academy Review*, (67), 198-211.
24. Yue, Y., Kim, W. H., & Cho, Y. S. (2023). Stock Market Prediction Based on LSTM Neural Networks. *The Journal of International Trade & Commerce*, 19(2), 391-407.

(Received Oct. 27, 2023/ Revised Nov. 29, 2023/ Accepted Dec. 23, 2023)