applied sciences

MDPI

Article

Influence of Social Overhead Capital Facilities on Housing Prices Using Machine Learning

Juryon Paik [1], Seung-June Baek [1], Jun-Wan Kim [1] and Kwangho Ko [2,*]

[1] Department of Data Information and Statistics, Pyeongtaek University, Pyeongtaek-si 17869, Gyeonggi-do, Republic of Korea; jrpaik@ptu.ac.kr (J.P.); nabsj@ptu.ac.kr (S.-J.B.); rlawnsdhks7@ptu.ac.kr (J.-W.K.)

[2] Department of Smart Mobility, Pyeongtaek University, Pyeongtaek-si 17869, Gyeonggi-do, Republic of Korea

* Correspondence: kwangho@ptu.ac.kr

Featured Application: Real Estate Price Prediction, Housing Price Prediction Applications, Establishment of a Stable Real Estate Policy.

Abstract: The South Korean residential real estate market is influenced by both the traditional dynamics of demand and supply and external factors such as housing policies and macroeconomic conditions. Considering the proportion of housing assets in individual wealth, market fluctuations can have significant implications. While previous studies have utilized variables such as GDP growth rate, patent issuance, and birth rate, and employed models such as LSTM and ARIMA for housing price predictions, many have overlooked the influence of local factors. In particular, there has been insufficient investigation into the impact of subway stations and living social overhead capital facilities on housing prices, especially in metropolitan areas. This study seeks to bridge this gap by analyzing the usage trends of subway stations, evaluating the impact of living social overhead capital facilities on housing values, and deriving the optimal machine learning model for price predictions near subway stations. We compared and analyzed a total of eight machine learning regression models, including Linear Regression, Decision Tree, Random Forest, LightGBM, Ridge, Lasso, Elastic Net, and XGBoost, all of which are popular regression models, especially in the context of machine learning and data science. Through comparative analysis of these machine learning techniques, we aim to provide insights for more rational housing price determinations, thereby promoting stability in the real estate market.

Keywords: feature extraction; housing price prediction; living social overhead capital facilities; subway proximity; machine learning

Citation: Paik, J.; Baek, S.-J.; Kim, J.-W.; Ko, K. Influence of Social Overhead Capital Facilities on Housing Prices Using Machine Learning. *Appl. Sci.* **2023**, *13*, 10732. https://doi.org/10.3390/app131910732

Academic Editor: José Salvador Sánchez Garreta

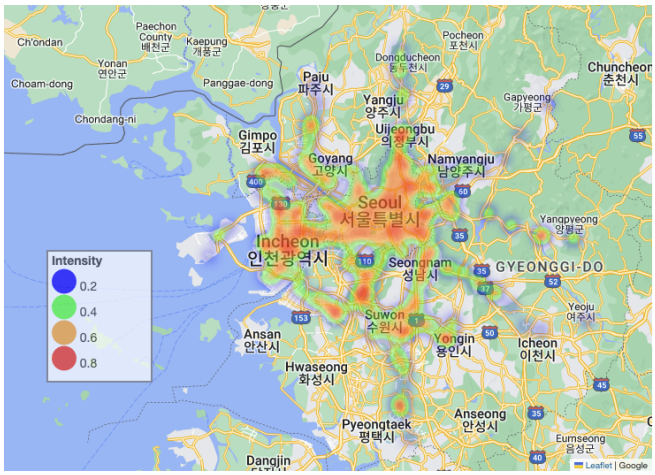
Received: 27 August 2023
Revised: 20 September 2023
Accepted: 25 September 2023
Published: 27 September 2023

1. Introduction

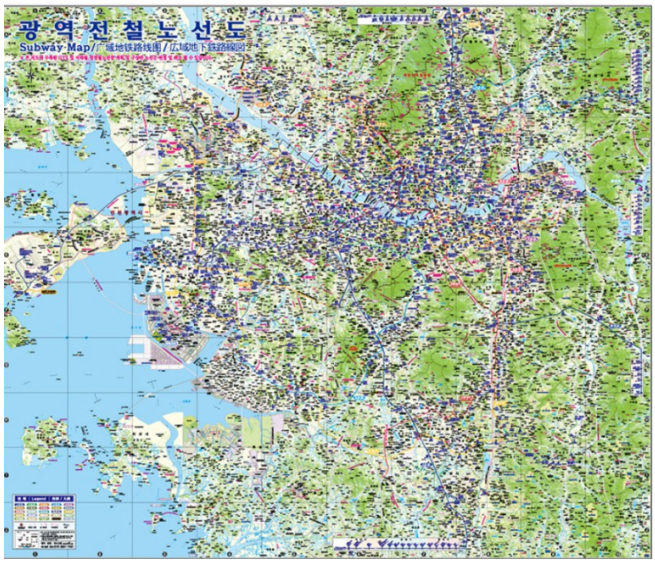
The housing market, an intricate nexus of activities including purchasing and selling of homes, real estate transactions, and price determination, significantly contributes to the dynamics of a nation's economy [1–3]. It is notably susceptible to an array of determinants such as overarching economic health, fluctuating interest rates, consumer demand, and shifts in government policies. As such, the condition of the housing market can be interpreted as a barometer of a country's economic vitality, with a robust housing market often signaling a prosperous economy and a frail one suggesting impending economic downturn [2,4–6].

In South Korea, the housing market is integral to the overall economy, with its stability resonating across a wide spectrum of stakeholders, not just homeowners. This ripple effect touches landlords, construction firms, lenders, and policymakers, and has broader implications for the community at large. Within this context, areas close to subway stations in South Korea, known as 'station proximity areas', have garnered significant attention. To further illustrate the close relationship between these areas and housing transactions, Figure 1

provides a visual representation. Figure 1a visualizes the number of housing transactions in South Korea's metropolitan area over a year. The broad orange region represents the metropolitan area centered around Seoul, with the narrower orange extensions depicting areas connected by subway lines emanating from the center. Figure 1b displays a South Korean metropolitan subway route map. The densely packed blue dots represent subway stations in the metropolitan area; similarly, the lines of blue dots extending from the center show the subway routes. Through this visual representation, it is evident that housing transactions in South Korea predominantly occur centered around the metropolitan area and major subway stations.



(**a**)



(**b**)

Figure 1. Visualization of the relationship between areas where housing transactions occurred in South Korea from November 2021 to November 2022 and the distribution of metropolitan and regional subway routes. (**a**) Visualization of housing transaction counts centered in the metropolitan area; (**b**) Visualization of the metropolitan and regional subway routes (Source: jido45.com).

Building on this observation, it is clear that these transit-oriented developments play a pivotal role in shaping the real estate landscape of South Korea. The convenience and accessibility offered by these areas results in heightened preference, significantly impacting housing prices in South Korea [7–14]. In light of global urbanization trends and the push towards sustainable urban living, such station proximity areas in South Korea not only offer logistical advantages, they contribute to urban efficiency. In light of the importance of the

housing market in the national economy of South Korea, price stability in station proximity areas becomes even more paramount. Understanding and predicting the dynamics of housing prices in these areas is crucial for all relevant stakeholders and for the broader national economy of South Korea.

Modeling housing price forecasting in densely populated areas of South Korea offers significant research benefits. The implications of price forecasting extend beyond just the housing and real estate markets, encompassing factors such as transportation hubs (stations), transportation infrastructure (railways, highways), potential traffic bottlenecks, utilities (electricity, water, sewage systems), and educational institutions. Accurate forecasting can furnish individuals and organizations with critical insights, aiding informed decisions when purchasing, selling, or investing in real estate. Moreover, this research can bolster the case for new investments in rail and other infrastructure. Effective forecasting models can provide transactional advantages while tempering the volatility inherent to housing prices [15–21]. Such models are pivotal for regulators and policymakers, allowing them to pinpoint, monitor, and address anomalies in the housing market, thereby fostering a more stable and predictable market landscape.

In this study, we aim to construct a comprehensive prediction model for housing prices in South Korean station proximity areas. We employ machine learning methods predicated on regression models. Regression models, statistically sophisticated tools that predict a dependent variable based on the values of one or more independent variables, have proven effective in identifying and mapping complex interdependencies among variables [22,23]. By harnessing the capabilities of these models, we endeavor to build a robust and accurate housing price prediction model that uses actual data from areas surrounding subway stations.

A distinctive aspect of our research is the integration of living Social Overhead Capital (SOC) facilities as stable factors into the predictive modeling process. SOC facilities represent core infrastructures and services, including transportation systems, communication networks, energy and water supply, schools, and hospitals. They are provided by both the governmental and private sectors to buttress economic growth and enhance societal welfare [24–26]. As these facilities are less susceptible to rapid change and upheaval, incorporating them into our model as stable predictors can enhance the model's predictive accuracy and reliability.

In pursuit of a more stable housing market, this study sets out to develop an advanced and precise housing price prediction model using stable factors, such as facilities with low volatility, along with distinct regional characteristics. By accomplishing this, we hope to provide valuable insights and a blueprint for future research aimed at enhancing the predictability and stability of housing prices in station areas in South Korea. In turn, this could contribute to a more predictable and stable housing market, better informed stakeholders, and a more resilient economy overall.

2. Related Research

The housing market, along with its myriad influencing factors, has been the subject of extensive academic exploration, especially around station proximity areas [27–35]. To provide a comprehensive context for our research within this vast domain, it is imperative to delve into the methodologies, findings, and implications of prior studies. This section aims to review the relevant literature, highlighting the evolution of predictive modeling in the housing market and the integration of machine learning techniques.

Historically, predictions in the housing market have been deeply rooted in traditional econometric models [36,37]. These models primarily emphasize macroeconomic indicators. One notable study [36] delved into the complexities of the Japanese housing market, shedding light on the challenges faced by the average Japanese individual when purchasing a house. This research introduced the innovative "asset market approach", which conceptualizes houses as unique asset classes. By combining the consumption capital asset pricing model with housing and residential land supply functions, a robust theoretical

framework was crafted. However, the study's focus on data from the 1970s and 1980s and its primary emphasis on the Japanese context limit its applicability to contemporary global housing markets.

Mikhed and Zemčík's research [38] provided a fresh perspective on the U.S. housing market. They meticulously investigated whether house prices genuinely reflected the underlying economic fundamentals. Their findings, derived from both aggregate and panel data, revealed significant discrepancies between house prices and their determinants, which were especially evident before the 2006 market correction. While their insights were invaluable, the study's reliance on traditional statistical methodologies suggests potential enhancements through the integration of modern machine learning techniques.

Genesove and Han's work [39] is particularly noteworthy for its in-depth exploration of housing market liquidity. By leveraging a unique dataset spanning diverse geographical areas and timeframes, they offered a panoramic view of market dynamics. Their innovative approach to understanding demand shocks and their subsequent impact on liquidity was groundbreaking. However, the study's reliance on data from the National Association of Realtors and its traditional methodologies indicate areas for potential improvement.

The hedonic pricing model, a cornerstone in real estate economics, has been instrumental in understanding how various factors, such as location and amenities, influence property values. Building on this foundation, several studies have investigated the relationship between housing prices and transportation infrastructure [12,32,40–44]. One study in Beijing [32] emphasized the positive correlation between proximity to rail transit systems and elevated property values, while another study in Tianjin [40] highlighted the transformative impact of subway systems on urban landscapes and housing prices.

Machine learning's integration into housing price predictions has been a game changer [45–52]. One recent study [45] has showcased the potential of various algorithms, with the RIPPER algorithm standing out for its superior predictive capabilities. This research underscored the transformative power of machine learning, suggesting a paradigm shift from traditional methodologies.

By 2018, the field of housing market research had matured considerably, with researchers employing more intricate datasets and refining their methodologies. A testament to this progression is the seminal work by [46]. This study provided a fresh perspective on the subject, specifically focusing on Ames, Iowa. The authors meticulously dissected the Ames Housing dataset, employing regression-based supervised learning methodologies to predict housing prices. Through a rigorous comparative analysis of multiple models, they identified an optimal model, which they then used as a foundation for amalgamating predictions. Their innovative approach to feature engineering and categorization stands out, offering a blueprint for future research in the domain. They delved deep into the intricacies of the dataset, exploring factors such as neighborhood characteristics, property age, and amenities. Despite securing the 35th position out of 2221 entries on Kaggle.com's public leaderboard, a prominent platform for data science competitions [47,48], this study acknowledged its limitations while emphasizing the need for broader validation and exploration. This work underscores the burgeoning potential of machine learning in real estate economics, reinforcing the notion that the horizon of possibilities in housing market research continues to expand as datasets grow and computational techniques advance.

Fast-forwarding to 2023, the research landscape has witnessed further advancements. A study by H. Peng et al. [51] introduced LUCE, a novel predictive model tailored for the Toronto housing market. LUCE was designed to address two pivotal challenges in real estate evaluation: the scarcity of recent sales prices, and the sparsity of housing data. This model's ingenuity lies in its ability to structure housing data in a Heterogeneous Information Network (HIN), where graph nodes represent crucial housing entities and attributes pivotal for price evaluation. By leveraging Graph Convolutional Networks (GCNs), LUCE extracts spatial information, such as the geographical locations of housesm from the HIN. Subsequently, the model employs Long Short-Term Memory (LSTM) networks to capture the temporal dependencies in housing transaction data over time. This dual

approach allows LUCE to provide a comprehensive and up-to-date housing evaluation dataset, significantly simplifying downstream appraisal tasks.

In the same year, another groundbreaking study [52] presented at the European Conference on Social Media delved into the nexus between social media sentiment and housing prices. This research probed the influence of Twitter sentiment, specifically pertaining to the COVID-19 pandemic, on the resale prices of Housing Development Board (HDB) apartments in Singapore. The study utilized the VADER lexicon-based tool for sentiment analysis and employed the Granger Causality method to discern the relationship between sentiment scores and reported COVID-19 cases. The research harnessed the power of neural networks for prediction, emphasizing the advantages of using Twitter sentiment over traditional predictors. The findings revealed that the incorporation of Twitter sentiment can augment prediction accuracy, surpassing models that rely solely on traditional predictors. This study underscored the pivotal role of sentiment analysis derived from Twitter data in urban economics, shedding light on the profound capability of social media platforms to encapsulate the behavioral economic nuances of a populace.

The progress in the housing market research domain over the years highlights the significance of exploring diverse determinants impacting housing prices. Beyond the conventional indicators, recent investigations have particularly gravitated towards more contemporary factors. These include aspects such as the influence of transportation infrastructure, notably rail transit systems, and the burgeoning relevance of social media sentiment.

Building upon these existing findings, our research takes a nuanced approach. Rather than studying transportation infrastructure and SOC facilities in isolation, our study uniquely combines these elements. By integrating data related to SOC facilities with subway station information, we present a more nuanced perspective on the influencers of housing prices. The distinctiveness of our research is further amplified through our methodological approach, which leverages advanced machine learning techniques to further refine the understanding of this complex domain.

Next, Section 3 provides detailed insights into our data-centric approach. We assembled a varied dataset featuring both SOC facilities and subway stations. Ensuring the quality and relevance of this data was paramount; thus, rigorous preprocessing was undertaken. This involved rectifying missing values, addressing outliers, and eliminating potential biases, ensuring a dataset that aptly mirrors urban housing market influences.

The next phase involved a comprehensive analysis deploying eight advanced machine learning models. With the intention of determining the model that best fits our combined dataset, each model was meticulously trained and tested. Through detailed evaluations and cross-validation, we zeroed in on a model that demonstrated unparalleled predictive precision, emphasizing the potential of machine learning in offering refined insights into the housing market's intricate dynamics.

3. The Proposed Scheme

3.1. Raw Data

In this study, we aim to construct a model to predict housing prices in areas in close proximity to subway systems, a significant component of living SOC facilities. We utilized the Metro-Adjacent Residential Transaction data, provided by the Korea Real Estate Board and available on the National Transportation Data Open Market, as our primary input variables. This dataset, shown in Table 1, comprises actual housing transaction data extracted from the Ministry of Land, Infrastructure, and Transport's real estate transaction disclosure system, specifically, transactions occurring within 500 m of subway stations. Based on this dataset, we incorporated additional data on SOC facilities and life convenience facilities as independent variables in our analysis, considering them as constant factors that could influence housing prices.

Table 1. Dataset specification of metro-adjacent residential transactions provided by the Korea Real Estate Board.

Feature	Feature Description
SIGUNGU_CD	Municipality Code [1]
EMDL_CD	Submunicipality Code [2]
CLL	Land Lot Classification (1: Regular, 2: Mountain)
MNO	Land Lot Number (Main) [3]
SNO	Land Lot Number (Sub) [3]
ADRES	Address Name (Legal District)
HUS_TP	Type of Multi-unit Housing (Apartment, Multi-family, Studio)
COMP_NM	Complex (Building) Name
BLDG_YEAR	Year of Construction
FLR	Floor Information
XUAR	Exclusive Area (m^2)
CTRT_YRMTH	Contract Year and Month
CTRT_DAY	Contract Day
TRANSCT_TYPE	Transaction Type (Sale, Jeonse [4], Monthly Rent)
DLNG_AMOUNT	Sale Price (in 10,000 KRW)
GRNTE_AMOUNT	Security Deposit (in 10,000 KRW)
MTHRNT_AMOUNT	Monthly Rent (in 10,000 KRW)
NEAR_SUBW_NM	Nearest Subway Station Name
NEAR_SUBW_DIST	Straight-line Distance to the Nearest Subway Station

[1] The 'Municipality Code' in this context is a unique identifier used specifically for designating administrative divisions in South Korea. [2] The 'Submunicipality Code' refers to the hierarchical levels of administrative divisions in South Korea, from smaller (township/town/neighborhood) to larger (city/county/district). This is used to identify specific areas within a city or county. [3] These terms refer to the identifiers used in the address system in Korea, similar to street names and house numbers in Western address systems. [4] In South Korea, this denotes a distinctive rental system in which a large lump sum deposit is made in lieu of monthly rent.

For data partitioning, we utilized approximately one year's worth of data, from November 2021 to November 2022, as our training dataset, while we used about one month of data for December 2022 as our testing dataset. Considering the scope and objectives of our study, we chose to focus solely on sales data. Hence, out of the total 360,084 instances in our training dataset, we utilized 58,342 instances of sales data, and out of the 25,622 instances in our testing dataset we utilized 2804 instances of sales data. We noted that there were missing values observed in our dataset, specifically, in the 'BLDG_YEAR' (Building Years) variable. In the training dataset, we identified 810 missing instances, while in the testing dataset we found 35 missing instances. The specific approach for handling these missing values is detailed in the subsequent methodology subsection of this study.

3.2. Preprocessing Data

This subsection discusses the preprocessing steps performed on the data prior to analysis. Data preprocessing is a crucial step in any data-driven project, helping to clean, normalize, and transform the raw data into a format suitable for further analysis or model training. These steps are essential to ensure the quality and reliability of the results derived from the data. We outline the specific preprocessing techniques used in our study, detailing why each step was necessary and how it contributed to the overall analysis.

3.2.1. Variable Modification and Reduction (in This Paper, the Terms 'Variable' and 'Feature' Are Used Interchangeably)

During preprocessing, we conducted both variable elimination and modification to optimize our dataset for further analysis. We identified a set of variables, as presented in Table 1, that either contained duplicate information or were not applicable to our study. The variables 'SIGUNGU_CD' (Si-Gun-Gu Code), 'EMDL_CD' (Eub-Myeon-Li Code), 'CLL' (Classification of Land Lot), 'MNO' (Main Number), and 'SNO' (Sub Number) encapsulate address-related information, which is already sufficiently represented by the

'ADRES' (Address) variable. To avoid redundancy, these overlapping variables were eliminated. Additionally, we removed variables such as 'GRNTE_AMOUNT' (Guarantee Amount) and 'MTHRNT_AMOUNT' (Monthly Rent Amount) which were irrelevant to sale transactions from our analysis. This streamlined the dataset and ensured that our model would only be trained on features pertinent to our research objectives.

After variable elimination, we modified the 'HUS_TP' (House Type) variable, which represents the property type of each transaction. 'HUS_TP' consists of three unique categories: 'Apartment', 'Studio', and 'Multi-family residential'. To allow the machine learning algorithms to process these categorical data more effectively, we applied the one-hot encoding technique. This process converts each category into a separate column assigned a binary value of 1 (presence of the feature) or 0 (absence of the feature). This transformation allowed our model to utilize the property type data effectively without any inherent order or priority.

3.2.2. Log-Scaling of Monetary Variables

Utilizing the 'DLNG_AMOUNT', which represents the property sale price, as a direct target variable poses certain challenges. Individuals with substantial financial resources are more likely to reside in larger properties, potentially inflating the 'DLNG_AMOUNT'. Conversely, those with limited resources might correspond to a comparatively lower 'DLNG_AMOUNT'. To circumvent this potential bias and adopt a more objective metric, we introduce the Square Meter Unit Price (SMUT), computed by dividing the 'DLNG_AMOUNT' by the property size, 'XUAR'.

Building on this, in order to examine regional variations in SMUP we conducted clustering based on the 'ADRES' feature. Figure 2 graphically depicts the SMUP of station proximity areas, with the monetary values benchmarked to 10,000 Korean Won units. We observe that the highest prices are more than double the lowest, and the majority of these regions are concentrated within Seoul.

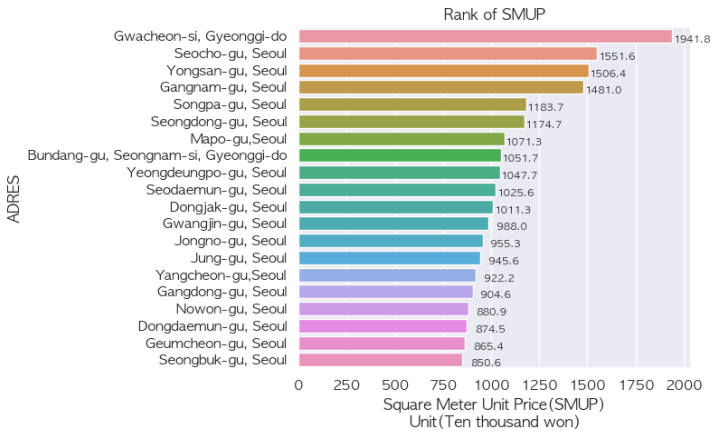


Figure 2. Bar chart illustrating the ranking of real estate prices per square meter in different regions based on actual transaction data. Each bar represents a different region, while the length of the bar denotes the price per square meter. The regions are ranked from highest to lowest based on their respective prices per square meter. This visualization highlights the spatial variations in real estate prices.

As a foundational step in our methodology, it is crucial to log-scale the 'DLNG_AMOUNT'. The need for this transformation emerges from the observed disparities and potential skewness in the distribution of property prices. By implementing log-scaling, we aim to reduce the influence of large or severely skewed values, striving for a more normalized distribution that is favorable for machine learning model performance. The intricacies of this log-scaling process can be visualized in Figure 3. After determining 'SMUT' using the relation $\text{SMUT} = \frac{\text{DLNG\_AMOUNT}}{\text{XUAR}}$, we subsequently applied a log-scaling transformation to SMUT itself. This transformation is an integral part of our data preprocessing, ensuring that our

target variable is aptly processed for the subsequent analytical phases in alignment with our foundational methodologies.

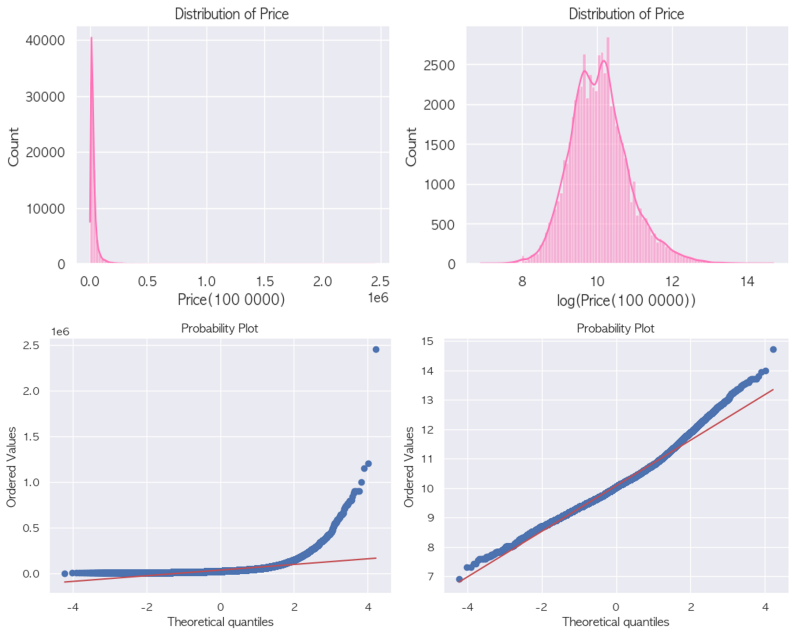


Figure 3. Logarithmic scaling of the target variable DLNG_AMOUNT (sales price). The **left** figures represent the data before scaling, while the **right** figures illustrate the data after the application of logarithmic scaling.

It is worth noting that when the log-transformed 'DLNG_AMOUNT' is divided by 'XUAR', the resultant 'SMUT' inherently undergoes log-scaling. This ensures that the distribution of 'SMUT' is further normalized, setting the stage for potential improvements in our model's performance.

3.2.3. Geocoding and Distance Calculation

Following log-scaling, we utilized geocoding to pinpoint the geographic locations related to the housing transactions. Geocoding converts addresses into their corresponding positions on the Earth's surface (i.e., latitude and longitude), enabling the mapping and analysis of geographical data [53–55]. For the geocoding process, we leveraged the AI·NAVER API from the Naver Cloud Platform to convert the 'ADRES' variable from our base dataset into geographical coordinates.

With the geocoded locations of the housing transactions, we used the Haversine formula to determine the distance between the houses and the surrounding SOC facilities [56,57]. The Haversine formula calculates the great-circle distance between two points on the surface of a sphere, for instance the Earth, taking its curvature into account. This makes it more suitable than regular flat-plane distance calculations. The formula is as follows:

$$a = \sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_1)\cos(\phi_2)\sin^2\left(\frac{\Delta\lambda}{2}\right)$$
$$c = 2 \cdot atan2(\sqrt{a}, \sqrt{1-a})$$
$$d = R \cdot c$$

where ϕ is the latitude, λ is the longitude, and R is Earth's radius (mean radius = 6371 km). For instance, this formula can be used to find the distance between Seoul (coordinates 37.5665° N, 126.9780° E) and Busan (coordinates 35.1796° N, 129.0756° E) in South Korea. For our study, we aggregated facilities within a 500 m radius, typically considered a 5–10 min walk, which is within walking distance. As we used geocoded coordinates for our calculations, the Haversine formula was essential for accurate distance calculations.

3.2.4. Socio-Environmental Determinants of Housing Prices

The socio-environmental context of a region plays a pivotal role in determining real estate values. Specifically, properties in areas that boast high quality public services, such as esteemed educational institutions and healthcare facilities, and have robust transportation networks often fetch higher prices than those in areas lacking these amenities. Conversely, areas deficient in public transportation or quality public services typically experience diminished real estate demand, leading to reduced property prices [58–60]. Recognizing the importance of these determinants, we have incorporated the condition of local infrastructure and amenities surrounding a property into our dataset under the term 'local living facilities'.

Educational institutions are paramount in homebuying decisions, especially for families with school-age children. The proximity of schools to a residence is often a primary consideration for such families. In light of this, our dataset includes the count of elementary and middle schools within a 500 m radius of the property, represented as 'ELET_SCH' (Elementary School) and 'MDL_SCH' (Middle School). Furthermore, urban amenities such as parks and libraries, which elevate quality of life, have a bearing on property prices. We have quantified the number of urban parks and libraries within a 500 m radius of the property and integrated these data as variables named 'CNT_PARK' (Count of Park) and 'CNT_LIB' (Count of Library) in our dataset.

After incorporating these socio-environmental determinants, another crucial variable that comes to mind is the average income by region. It is generally understood that owning property in affluent areas requires a proportionate income. Thus, we expected a strong correlation between the average regional income and property prices. To substantiate this, we analyzed the total reported salary by region based on the end-of-year tax settlement of earned income. This analysis yielded the average annual salary per person by region, which we represented in our dataset as the variable 'INC_BY_REG' (Income By Region). This addition further enriched our understanding of the interplay between property prices and regional income.

3.2.5. The Influence of Housing Brands on Property Value

For potential homebuyers, the brand reputation of a residential property often plays a pivotal role in the decision-making process. Specifically, properties developed by renowned construction firms are typically associated with higher quality and standards, which can subsequently influence market prices. This sentiment is corroborated by data from the Korea Corporate Brand Reputation's apartment brand reputation analysis in January 2023 [61], which ranked Hillstate as the top brand, followed by Prugio and Xi.

Considering the evident significance of housing brands in the real estate market, our study sought to incorporate this factor into our analysis. We referenced the 'Top 20 Apartment Brand Preferences' as provided by the Korea Corporate Reputation Research Institute. Using the 'COMP_NM' (Company Name) variable, we identified whether a particular property was developed by one of the major construction companies. This binary representation serves to capture the brand value of the property, indicating whether it was constructed by a leading developer in the industry.

3.2.6. Assessing the Importance of Subway Stations and Data Processing

Subway station passenger traffic varies based on diverse user intentions, reflecting the significance of stations that serve various destinations, including offices, residences, hospitals, entertainment venues, and more. In this study, our objective is to gauge the importance of each subway station by analyzing passenger entry and exit volumes.

Our primary data source was the 'Standard Station Information' from the Rail Portal, which covered the entire year of 2022 as of January 2023. During the data collection process, we encountered several challenges, notably the disparity in the operating institutions across different subway lines within the metropolitan area. Acquiring data from subway lines operated by the private sector was especially challenging, as these entities are not mandated

to disclose such information. Furthermore, several public datasets contained inaccuracies attributed to issues such as turnstile errors. To address these challenges, we compiled a comprehensive list of subway line-specific operating institutions in the metropolitan area, which isdetailed in Table 2.

Table 2. Subway operating institutions by line.

Line	Operating Institution
Incheon Subway Line 1 & 2, Urban Railway Line 7	Incheon Transit Corporation
Everline	Yongin Light Rail Corporation
Incheon International Airport Line	Airport Railroad Corporation
Ui LRT	Ui LRT Corporation
Shinbundang Line	DX Line
Greater Capital Area Light Rail Sillim Line	ROTEM SRS Co., Ltd.
Maglev	Incheon International Airport Terminal
Gimpo Gold Line	Gimpo Gold Line Co., Ltd.
Jinjeop Line	Namyangju City Corporation
Seoul Metro Lines 1–8 (part of Line 9)	Seoul Metro
Greater Capital Area Metro Line 9	Seoul Metro Line 9 Corporation
Uijeongbu LRT	Uijeongbu Light Rail Co., Ltd.
Gyeonggang Line, Gyeongbu Line, Janghang Line, Gyeongwon Line, Gyeongui Central Line, Gyeongin Line, Gyeongchun Line, Bundang Line, Suin Line, Gwacheon Line, Ansan Line, Ilsan Line, West Sea Line, Itx-Gyeongchun Line	Korea Railroad Corporation

Building on this, in order to understand the usage volumes of each subway station, we collected data on the total number of boardings and alightings for 2022, as provided by a public agency portal. We initially supplemented the missing values with data from the 'Integrated Transport Card Big Data System' operated by the Ministry of Land, Infrastructure, and Transport. This system provides data based on transportation card usage during boarding and alighting. For stations with persisting missing values, we utilized 2021 data with adjustments based on the usage trend of the nearest station in 2022.

During data preprocessing, we made transformations using the 'NEAR_SUBW_NM' (Near Subway station Name) variable to enhance its linkage with the passenger volume data. This involved reconciling station names that had changed over time and resolving duplicate station names across different lines. We standardized these names to a consistent 'Line_StationName' format. Lastly, to evaluate the significance of subway stations, the primary variable (passenger volum) was normalized using the MinMaxScaler. This technique, widely used in machine learning, scales and normalizes data to fall between 0 and 1.

3.2.7. Interplay between Housing Prices and Financial Market Dynamics

The dynamics of housing prices are influenced by a myriad of factors, one of which is the prevailing financial climate [1–6]. Among various indicators representing the financial health of an economy, the policy rate emerges as a pivotal gauge reflecting a nation's monetary policy and the broader economic sentiment. In South Korea, the Bank of Korea is responsible for setting and adjusting this rate. Recognizing the significance of the policy rate in our study, we incorporated it as a primary variable to understand the intricate interplay between the monetary policy landscape and housing prices.

In light of the time-series nature of our foundational dataset, it is essential to consider the timing of housing transactions. By combining the 'CTRT_YRMTH' (Contract

Year and Month) and 'CTRT_DAY' (Contract Day) variables, we introduce a new variable, 'PLC_RATE' (Policy Rate), which corresponds to the policy rate prevailing on the specific date of the transaction. This integration offers a nuanced reflection of the financial context at the time of each housing transaction. However, an inherent limitation must be acknowledged in that the funding mechanisms employed for housing purchases may not always align with the contemporaneous policy rate; for instance, prospective homeowners might secure financing well in advance, potentially under different interest rate conditions. Thus, solely examining the interplay between the policy rate and housing prices might provide an oversimplified perspective, highlighting the need for more comprehensive studies that delve deeper into this complexity.

3.2.8. Addressing Missing Data and Incorporating Building Age

The concurrent development of housing and associated infrastructure, such as roads and sewage systems, provides distinct efficiency benefits compared to sequential development. This integrated approach in housing projects streamlines approval processes, eliminating the need for multiple approval stages and accelerating the overall development timeline [62,63]. Against this backdrop, it is logical to infer missing values by observing analogous developmental patterns within the dataset.

In relation to the 'BLDG_YEAR' variable from Table 1, we identified a total of 845 missing values across both the training and testing datasets. To address this, we employed the KNN imputer method [64], leveraging inherent data similarities. The KNN imputer works by pinpointing the K-nearest neighboring data points and then imputing the missing value based on the average of these neighbors. For our study, we utilized the KNN Imputer from scikit-learn, setting it to consider the five nearest data points for imputation purposes. The age of a building is undeniably a pivotal factor in real estate purchasing decisions. To encapsulate this aspect, we introduced the 'AGE_BLDG' variable, which is derived by subtracting the building's construction year from the contract year.

3.3. Modeling Process

In this section, we delve into the intricacies of our modeling process, starting with an assessment of the input variables. Our model incorporates a total of fifteen input variables. Among these, three variables, the 'NEAR_SUB_DIST', 'FLR' (Floor), and 'XUAR' (Exclusive Area), are directly extracted from Table 1 of our foundational dataset. They respectively represent the distance from a subway station (within 500 m), the floor number of the building, and the exclusive area of the house. The remaining variables have been refined through various preprocessing steps. For instance, the 'WTD_SUBW_RANK' variable was derived by summing the boarding and alighting counts at subway stations and then normalizing these values. A significant aspect of this preprocessing involved geocoding based on the 'ADRES' variable, which allowed us to utilize the resulting latitude and longitude coordinates. In addition, we introduced the 'INC_BY_REG' variable, which is based on the 2021 wage income settlement details specific to the housing's region, be it a city or district. The 'PLC_RATE' variable reflects the benchmark interest rate set by the Bank of Korea on the transaction day. Furthermore, the variables 'ELET_SCH' and 'MDL_SCH' denote the count of elementary and middle schools, respectively, within a 500 m radius. In a similar vein, the 'CNT_PARK' and 'CNT_LIB' variables capture the number of parks and libraries, respectively, within the same proximity. Lastly, the 'COMP_NM' variable serves as an indicator signifying whether the housing was constructed by a renowned construction company. A comprehensive list of the input variables utilized in our final model can be found in Table 3.

3.3.1. Multicollinearity and Its Mitigation

As we progress through the modeling process, it is imperative to ensure the validity and reliability of the model. One potential pitfall in multiple regression models that can compromise model reliability is the presence of multicollinearity. Multicollinearity is a

phenomenon observed in multiple regression analysis when several independent variables are highly correlated with each other. In models burdened with high multicollinearity, deciphering the distinct influence of each independent variable becomes intricate, potentially leading to unreliable coefficient estimates [65,66]. Such scenarios can inflate the standard errors of the regression coefficients, rendering it very challenging to attain statistically significant results.

Table 3. Summary of input features utilized in the modeling process after complete preprocessing.

Feature	Feature Description
WTD_SUBW_RANK	Ranking of subway stations based on weighted passenger volume
NEAR_SUBW_DIST	Distance between the property and the nearest subway station
INC_BY_REG	Ranked region based on weighted average income
PLC_RATE	Policy interest rate corresponding to the contract date
ELET_SCH	Number of elementary schools
MDL_SCH	Number of middle schools
CNT_PART	Number of parks
CNT_LIB	Number of libraries
FLR	Floor Information
XUAR	Exclusive Area (m^2)
COMP_NM	Branded construction company status
HUS_TP_APT	Apartment status
HUS_TP_STD	Studio status
HUS_TP_MUTF	Multi-Family status
LOG_PRICE	Log-transformed sale price
AGE_BLDG	Building age

To address the issue of multicollinearity in our analysis, we employed the Variance Inflation Factor (VIF). The VIF calculates the magnitude of multicollinearity among independent variables. Specifically, it sets each independent variable as the dependent variable and conducts a regression analysis against other variables, using the R^2 value for its computation. Typically, a VIF value exceeding 10 indicates significant multicollinearity between the variable in question and others.

Utilizing the statsmodels.api library in Python, VIF values for all independent variables were determined. The computed VIF values for the modeling input variables are depicted in Figure 4. However, after performing one-hot encoding (OHE) on the original 'HUS_TP' variable, we observed potential multicollinearity. Considering the inherent nature of OHE, which transforms a single variable into multiple binary columns, it is not uncommon to encounter high multicollinearity among the generated features. Recognizing this, we excluded the highly correlated variables and re-evaluated the VIF. Upon inspecting the revised VIF values for each feature, all variables demonstrate a VIF less than 10.

	VIF_Factor	Feature
0	2.644417	FLR
1	4.592661	XUAR
2	4.160502	NEAR_SUBW_DIST
3	3.041231	ELET_SCH
4	1.901695	MDL_SCH
5	2.738005	CNT_PARK
6	1.755285	CNT_LIB
7	4.842190	INC_BY_REG
8	1.229727	COMP_NM
9	2.343413	WTD_SUBW_RANK
10	3.055039	AGE_BLDG
11	7.322778	PLC_RATE

Figure 4. Variance Inflation Factor (VIF) values for all features, assessing multicollinearity among input variables in the modeling process.

3.3.2. Optimal Model Determination

Upon addressing multicollinearity, our next step involved evaluating the efficacy of eight distinct machine learning models using cross-validation. A concise overview of these models and their unique characteristics is provided in Table 4.

Table 4. Overview of the eight machine learning models employed in the study, highlighting their unique characteristics and features.

Model	Description
Linear Regression	A general linear regression learning model that reflects the correlation between explanatory and dependent variables.
Ridge Regression	A linear regression learning model with L1 regularization.
Lasso Regression	A linear regression learning model with L2 regularization.
ElasticNet Regression	A linear regression learning model that combines both L1 and L2 regularization.
Decision Tree	A tree-based learning model that branches in the direction of lower impurity and learns to minimize this impurity.
Random Forest	Utilizing Bagging, this tree-based learning model selects variables randomly, preventing overfitting typically seen in Decision Trees.
XGBoost	An ensemble tree-based learning model that utilizes boosting techniques. It inputs the loss of the previous model into the learning data and uses the gradient method to correct errors.
LightGBM	A tree-based learning model that minimizes error loss by employing methods like GOSS, EFB, and Leaf Wise.

Model validation is an integral component of the modeling process. Cross-validation, a foundational technique in machine learning, gauges a model's performance by partitioning the dataset into multiple training and test subsets. This approach is instrumental in preventing overfitting. However, our dataset presents a unique challenge due to its inherent time series nature. Resorting to standard cross-validation methods would be problematic, as it poses the risk of unintentionally leveraging future data to forecast past or present events. To navigate this intricacy, we adopt the Time Series Nested Cross-Validation (TS-Nested CV) strategy [67]. This specific validation approach is tailored for data with temporal dependencies. It ensures that the training dataset always precedes the validation set in chronological order. As such, instead of the conventional k-fold techniques, the TS-Nested CV is preferred. We implemented this process using the Time Series Split tool available in the scikit-learn library. A schematic depiction of the approach can be found in Figure 5.

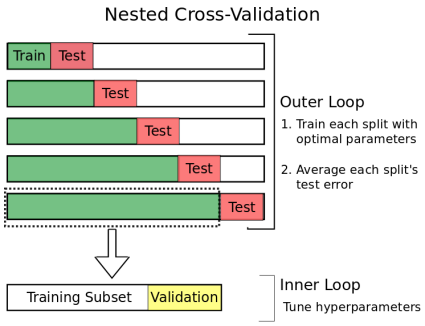


Figure 5. Nested cross-validation process: illustration of the step-by-step splitting of training and test datasets, emphasizing the hierarchical structure of hyperparameter tuning within the inner loop and performance evaluation in the outer loop.

Choosing the right evaluation metric is crucial for assessing a model's performance accurately. These metrics offer quantitative insights into a model's predictive accuracy. In this study, the target variable 'SMUP' is derived from the housing sale price variable 'DLNG_AMOUNT'. As depicted in Figure 2, there exists a notable discrepancy exceeding a two-fold difference between the highest and lowest prices, indicating potential outliers that could significantly influence the sale price. When the Root Mean Squared Error (RMSE) is employed as a loss function, its inherent sensitivity to outliers is augmented, as it squares the residuals between the actual and predicted values. This can obscure whether predictions systematically underestimate or overestimate the actual values. To address this limitation, we applied the Root Mean Squared Logarithmic Error (RMSLE) as our evaluation metric. By computing the logarithm of both the actual and predicted values and then determining their difference, the RMSLE effectively ensures that smaller and larger errors are treated more uniformly, rendering it robust against outliers.

Additionally, we employed the Relative Root Mean Squared Error (RRMSE) as another performance metric. The RRMSE is a nuanced variant of the RMSE designed to appraise the precision of predictive models concerning the span of the target variable. Unlike the RMSE, which is bound by the original measurement scale, the RRMSE affords the flexibility to compare various measurement techniques. Consequently, any inaccuracies in predictions manifest as elevated RRMSE values.

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1) - \log(a_i+1))^2} \tag{1}$$

$$\text{RRMSE} = \sqrt{\frac{1}{n}\frac{\sum_{i=1}^{n}(a_i - p_i)^2}{\sum_{i=1}^{n} p_i^2}} \tag{2}$$

In the above equations, p_i represents the predicted values, a_i denotes the actual values, and n is the number of data points.

Having established the evaluation metrics crucial for our analysis, we proceeded to evaluate our models under a rigorous validation scheme. Using the TS-Nested CV method, we compared the eight models based on the RMSLE and RRMSE metrics. For this assessment, we primarily used the default hyperparameters for each model. Figure 6 presents the RMSLE and RRMSE values for the eight models after training them to predict the target value, 'SMUP'. As evidenced by the results shown in Figure 6a, the LightGBM model boasts the smallest relative error, or RMSLE, between the actual and predicted values. For instance, when 'SMUP' was set at KRW 10 million, the Linear Regression model approximated it to be around KRW 5.3 million and LightGBM estimated it as KRW 7.6 million. In terms of RRMSE, which evaluates the accuracy of different predictive models relative to the range of 'SMUP' values, LightGBM again showed better performance than the other models. Based on these findings, we identified LightGBM as the most apt predictive model for our study.

3.4. Hyperparameter Tuning of LightGBM and Feature Importance Analysis

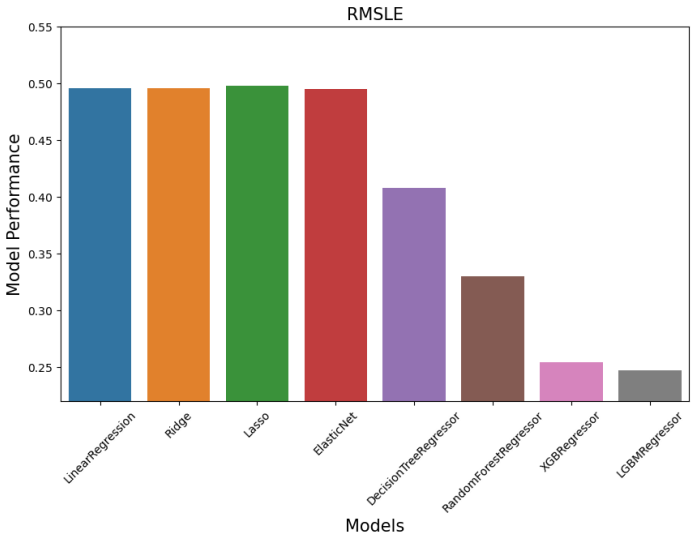
To further refine the performance of the chosen LightGBM model, hyperparameter tuning is essential. For this task, we utilized the OPTUNA library, an open-source Python tool designed specifically for hyperparameter optimization. While it shares similarities with GridSearch, RandomSearch, and BayesianOptimization, OPTUNA offers a more streamlined and efficient approach to optimize both machine learning and deep learning models. It significantly reduces the time and effort typically required for model selection and hyperparameter adjustment.

Considering that a lower RMSLE signifies better accuracy, our optimization objective was set to 'minimize'. In order to balance optimization quality with computational efficiency, we limited the number of iterations to $n\_trial = 100$. Post-optimization, the model's accuracy saw a marked improvement, with the RMSLE value decreasing from the initial 0.24230 to 0.21999. The reduction of the RMSLE value by 0.02 after logarithmic

transformation of the actual values through hyperparameter tuning signifies that if the pre-adjustment prediction was made with an error rate of 24% at KRW 100 million, the post-adjustment prediction would be at an error rate of 22%, resulting in a value of KRW 104.5 million. This equates to a difference of approximately KRW 4.5 million per SMUP, potentially leading to substantial financial losses for homeowners due to this error. The accuracy of our predictions improved with hyperparameter tuning.

	Model	Average RMSLE Score_(Error Difference)	Average Relative RMSE_(Error Ratio)
0	LinearRegression	0.475909	0.074206
1	Ridge	0.475838	0.074194
2	Lasso	0.477934	0.074980
3	ElasticNet	0.477575	0.074515
4	DecisionTreeRegressor	0.366252	0.057326
5	RandomForestRegressor	0.303261	0.047464
6	XGBRegressor	0.248270	0.038771
7	LGBMRegressor	0.242301	0.037831

(**a**)



(**b**)

Figure 6. Cross-validation results for the target value SMUP: average RMSLE and RRMSE values for the eight regression models, complemented with a bar chart highlighting the differences in average RMSLE values among the models. (**a**) Average RMSLE and RRMSE values for the eight regression models; (**b**) Bar chart comparison of the average RMSLE values across eight models.

In the LightGBM-based analysis, our primary objective was to pinpoint the salient features influencing housing prices. To gauge the specific impact of subway stations on the 'SMUP' prediction, we conducted a separate performance assessment across the eight models, excluding the two subway-related features, 'WTD_SUBW_RANK' and 'NEAR_SUBW_DIST'. Figure 7 depicts a discernible decrease in predictive accuracy for all models, indicating the significant role these subway attributes play. It can be inferred that the proximity to subway stations, coupled with their high usage, exerts a positive influence on housing price predictions.

In our prediction analysis, we scrutinized the influence of these two features and subsequently assessed their importance. We divided the feature importance analysis into two scenarios, one including these two features and the other excluding them; the results are presented graphically in Figure 8. In the importance analysis graph which includes the two features, the feature with the most significant influence is WTD_SUBW_RANK, representing the subway stations frequently used by many people, followed by INC_BY_REG,

indicating purchasing power, and then AGE_BLDG, representing the age of the building. The other graph showcases the influence of features when the same two specific features are excluded. Interestingly, while the housing area (XUAR) has a lesser influence than AGE_BLDG in the top graph, it ranks second in the bottom graph. This can be inferred as those with greater purchasing power tending to prefer houses with a larger area. In both graphs, the influence of the construction company's brand was the least impactful. Thus, it can be discerned that features related to proximity to subway stations and socio-infrastructure facilities exert the most substantial influence on housing price predictions.

	Model	Average RMSLE Score
0	LinearRegression	0.498259
1	Ridge	0.498159
2	Lasso	0.491364
3	ElasticNet	0.491929
4	DecisionTreeRegressor	0.377632
5	RandomForestRegressor	0.316555
6	XGBRegressor	0.265644
7	LGBMRegressor	0.263464

(**a**)

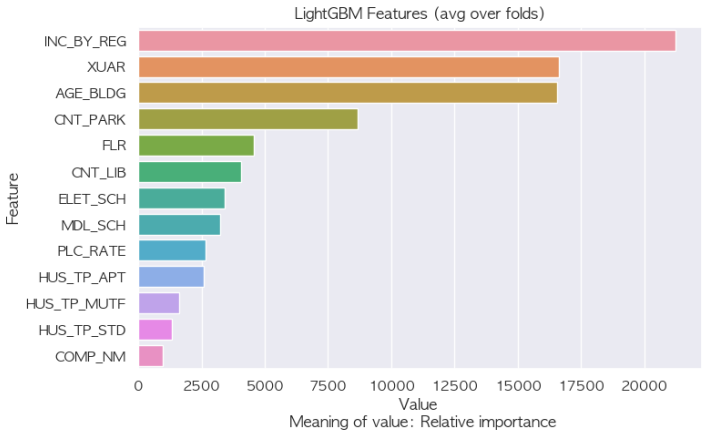
	Model	Average RMSLE Score
0	LinearRegression	0.475909
1	Ridge	0.475838
2	Lasso	0.477934
3	ElasticNet	0.477575
4	DecisionTreeRegressor	0.366252
5	RandomForestRegressor	0.303261
6	XGBRegressor	0.248270
7	LGBMRegressor	0.242301

(**b**)

Figure 7. To understand the impact of subway stations on housing price predictions, the performance metrics of the eight regression models were evaluated while excluding the WTD_SUBW_RANK and NEAR_SUBW_DIST features prior to hyperparameter tuning: (**a**) performance metrics when excluding the two features and (**b**) performance metrics when including the two features.

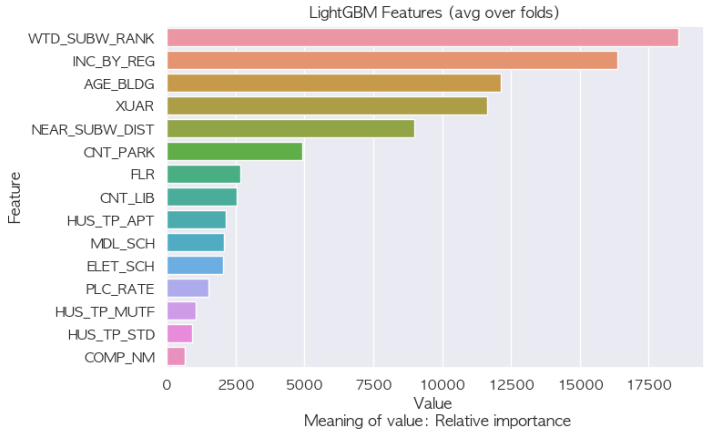
These findings suggest that location-related external factors, especially transit accessibility, have a more pronounced impact on housing prices than the inherent attributes of the properties. Additionally, the age and size of a property play crucial roles in its valuation. This underscores the importance of location and surrounding amenities, particularly transportation links, in urban real estate valuation. Further exploration of the diverse factors influencing housing prices presents a valuable direction for future research.

Armed with these insights, stakeholders in the real estate market, both buyers and sellers, can make more informed decisions, especially concerning location and transportation accessibility. As we transition to our concluding remarks, it is vital to contemplate the broader implications of our findings, especially in the realms of urban planning, policy formulation, and emerging real estate trends. This sets the foundation for the discussion in the forthcoming section.



(**a**)

Figure 8. *Cont.*



(**b**)

Figure 8. Comparison of feature importance in the LightGBM model based on the inclusion or exclusion of the WTD_SUBW_RANK and NEAR_SUBW_DIST features: (**a**) feature importance when excluding WTD_SUBW_RANK and NEAR_SUBW_DIST and (**b**) feature importance when including WTD_SUBW_RANK and NEAR_SUBW_DIST.

4. Discussion

In the rapidly changing field of housing market research, our study offers a fresh perspective by examining the interplay between traditional econometric models and the latest machine learning techniques. By combining data from SOC facilities and subway stations, we present a comprehensive method to understand the multifaceted factors influencing housing prices.

Our integration of diverse datasets provides a holistic view of urban amenities and their impact on housing prices. The rigorous preprocessing and data integration methods we employ can serve as a reference point for future studies, emphasizing robustness and reliability. Moreover, our comparative evaluation of eight advanced machine learning models highlights the transformative potential of machine learning in reshaping housing price predictions. Our model not only delivers noteworthy predictive accuracy, it illuminates the intricate relationships between various determinants.

Although many studies have focused on either transportation infrastructure or the significance of SOC facilities, our research stands out through its integrated approach. We move beyond traditional approaches, offering a deeper understanding of housing prices by considering both subway accessibility and the importance of SOC facilities. This comprehensive perspective ensures a more accurate prediction model, setting our research apart from others.

We recognize that the vast landscape of urban economics holds many areas ripe for exploration. In particular, national policies related to real estate, especially housing, play a pivotal role in influencing prices. The implementation of transportation infrastructure, expansion of social amenities, land development, and other state-driven initiatives form the backbone of these influences. We believe that integrating information from news reports, articles, and social networks related to these policy measures with the data we have gathered could yield more accurate and trustworthy predictions for real estate prices, including housing. As a future direction, we aim to collate policy data in order to examine its impact on price forecasting. Moreover, we intend to delve deeper into the influence of proximity to urban centers, schools, and public facilities in subsequent studies, harnessing these data to refine our predictive models. Building on these efforts, we lastly aim to perform comparisons with other machine learning-based prediction studies while leveraging our improved predictive model. We intend to compare and analyze the results using predictive models from other studies, employing features extracted by integrating housing transaction data near subway stations with SOC-related data.

Our research provides valuable insights and charts a path for future inquiries. By emphasizing the importance of a comprehensive approach and the potential of machine learning, we aspire to stimulate continued innovation in housing market research.

Author Contributions: Conceptualization, J.P. and S.-J.B.; methodology, J.P., S.-J.B. and J.-W.K.; software, S.-J.B. and J.-W.K.; validation, S.-J.B., J.-W.K. and K.K.; formal analysis, J.P.; investigation, S.-J.B. and J.-W.K.; resources, S.-J.B. and J.-W.K.; data curation, S.-J.B., J.-W.K. and K.K.; writing—original draft preparation, S.-J.B. and J.-W.K.; writing—review and editing, J.P. and K.K.; visualization, S.-J.B. and J.-W.K.; supervision, J.P.; project administration, J.P.; funding acquisition, J.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1F1A1064073).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Metro-Adjacent Residential Transaction data, available on the National Transportation Data Open Market, https://www.bigdata-transportation.kr/frn/prdt/detail?prdtId=PRDTNUM_000000020052, Apartment brand January 2023 big data analysis results, available on Korea Corporate Brand Reputation, https://brikorea.com/bbs/board.php?bo_table=rep_1&wr_id=2126, Korea Railroad Corporation, https://info.korail.com/info/selectBbsNttView.do;jsessionid=Zr2b7ZEC5zImD31IuWd5VAK72TuaxwQfUAppAfGsTzlsebEobzZT8I13TT5RPdpD?key=867&bbsNo=425&nttNo=15252&searchCtgry=&searchCnd=all&searchKrwd=&integrDeptCode=&pageIndex=1, Seoul Metro Corporation, http://www.seoulmetro.co.kr/kr/board.do?menuIdx=548&bbsIdx=2215619, Incheon Metro Corporation, https://www.ictr.or.kr/main/bbs/bbsMsgDetail.do?msg_seq=253&bcd=data, Namyangju City Urban Corporation, https://www.ncuc.or.kr/jj_line/1734?action=read&action-value=313f7ca5e6d4b76452aef867a71a1f2e, Rail Portal, https://data.kric.go.kr/rips/M_01_01/detail.do?id=32, Traffic Card Big Data System, https://www.stcis.go.kr/wps/main.do, Railway industry Information Center, http://www.kric.go.kr/jsp/industry/rss/citystapassList.jsp?q_org_cd=A010010042&q_fdate=2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, W.; Chen, S.; Guo, D.; Li, B. The impact of internet real estate intermediary platform on the real estate market. In Proceedings of the 4th International Conference on Crowd Science and Engineering, Jinan, China, 18–21 October 2019; pp. 132–139.
2. Chen, C.-J. ; Zhai, H.; Wang, A.; Ma, S.; Sun, J.; Wu, C.; Zhang, Y. Experimental research on the impact of interest rate on real estate market transactions. *Discret. Dyn. Nat. Soc.* **2022**, *2022*, 9946703. [CrossRef]
3. Li, Y. The impact of COVID-19 on China's real estate industry and the outlook for industry trends. *Bcp Bus. Manag.* **2022**, *34*, 337–343. [CrossRef]
4. Kurihara, Y. Demand for money under low interest rates in Japan. *J. Econ. Financ. Stud.* **2016**, *4*, 12–19. [CrossRef]
5. Ofor, T. N. ; Alagba, O. S.; Ifurueze, M.S. Housing finance market and economic growth of West Africa region: A study of Nigeria and Ghana. *Int. J. Bus. Econ. Dev.* **2018**, *6*, 49–60.
6. Ayan, E.; Eken, S. Detection of price bubbles in Istanbul housing market using LSTM autoencoders: A district-based approach. *Soft Comput.* **2021**, *25*, 7957–7973. [CrossRef]
7. Bajic, V. The effects of a new subway line on housing prices in metropolitan Toronto. *Urban Stud.* **1983**, *20*, 147–1583. [CrossRef]
8. Agostini, C.A.; Palmucci, G.A. The anticipated capitalisation effect of a new metro line on housing prices. *Fisc. Stud.* **2008**, *29*, 233–256. [CrossRef]
9. Wang, L. Impact of urban rapid transit on residential property value. *Chin. Econ.* **2010**, *43*, 33–52. [CrossRef]
10. Sun, W.; Zheng, S.; Wang, R. The capitalization of subway access in home value: A repeat-rentals model with supply constraints in Beijing. *Transp. Res. Part A Policy Pract.* **2015**, *80*, 104–115. [CrossRef]
11. Trojanek, R.; Gluszak, M. Spatial and time effect of subway on property prices. *J. Hous. Built Environ.* **2018**, *33*, 359–384. [CrossRef]
12. Wen, H.; Gui, Z.; Tian, C.; Xiao, Y.; Fang, L. Subway opening, traffic accessibility, and housing prices: A quantile hedonic analysis in Hangzhou, China. *Sustainability* **2018**, *10*, 2254. [CrossRef]
13. Li, S.; Chen, L.; Zhao, P. The impact of metro services on housing prices: A case study from Beijing. *Transportation* **2019**, *46*, 1291–1317. [CrossRef]
14. Zhou, Z.; Chen, H.; Han, L.; Zhang, A. The effect of a subway on house prices: Evidence from Shanghai. *Real Estate Econ.* **2021**, *49*, 199–234. [CrossRef]

15. Choi, M.; Byeon, S. Comparison on forecasting performance of housing price prediction models in Seoul. *Seoul Stud.* **2016**, *17*, 75–89.
16. Lee, T. H.; Jun, M. Prediction of Seoul house price index using deep learning algorithms with multivariate time series data. *SH Urban Res. Insight* **2018**, *8*, 39–56. [CrossRef]
17. Bae, S.Y.; Chung, E.-C.; Lee, S.Y. Effects of urban railway transportation services on housing prices: Case of apartments in Gyeonggi Province. *J. Korea Real Estate Anal. Assoc.* **2018**, *24*, 85–98.
18. Bae, S.W. Forecasting Property Prices Using the Machine Learning Methods: Model Comparisons. Ph.D. Thesis, Department of Urban Planning and Real Estate, Dankook University, Yongin-si, Republic of Korea, 2019.
19. Kim, H.; Kwon, Y.; Choi, Y. Assessing the impact of public rental housing on the housing prices in proximity: Based on the regional and local level of price prediction models using long short-term memory (LSTM). *Sustainability* **2020**, *12*, 7520. [CrossRef]
20. Song, Y.S.; Kim, H.; Cho, O.-S. Investigation of prediction of house price change in Seoul based on demographics with back propagation algorithm. *J. Inst. Elec. Inf. Eng.* **2020**, *57*, 27–33.
21. Kim, H.-S. Machine Learning Forecasting of Residential Market: The Case of Innovation Clusters. Master's Thesis, Department Business Adminidtration, Hanyang University, Seoul, Republic of Korea, 2021.
22. Snee, R.D. Validation of regression models: Methods and examples. *Technometrics* **1977**, *19*, 415–428. [CrossRef]
23. Fahrmeir, L.; Kneib, T.; Lang, S.; Marx, B. *Regression*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 21–72.
24. Koukouli, S.; Vlachonikolis, IG.; Philalithis, A. Socio-demographic factors and self-reported funtional status: The significance of social support. *BMC Health Serv. Res.* **2002**, *2*, 20. [CrossRef]
25. Chiesura, A. The role of urban parks for the sustainable city. *Landsc. Urban Plan.* **2004**, *68*, 129–138. [CrossRef]
26. Vera-Toscano, E.; Teca-Amestoy, V. The relevance of social interactions on housing satisfaction. *Soc. Indic. Res.* **2008**, *86*, 257–274. [CrossRef]
27. Cervero, R.; Duncan, M. Benefits of proximity to rail on housing markets: Experiences in Santa Clara County. *J. Public Transp.* **2002**, *5*, 1–18. [CrossRef]
28. McMillen, P.D.; McDonald, J. Reaction of house prices to a new rapid transit line: Chicago's Midway Line, 1983–1999. *Real Estate Econ.* **2004**, *32*, 463–486. [CrossRef]
29. Andersson, D.E.; Shyr, O.F.; Fu, J. Does high-speed rail accessibility influence residential property prices? Hedonic estimates from southern Taiwan. *J. Transp. Geogr.* **2010**, *18*, 166–174. [CrossRef]
30. Debrezion, G.; Pels, E.; Rietveld, P. The impact of rail transport on real estate prices: An empirical analysis of the Dutch housing market. *Urban Stud.* **2010**, *48*, 997–1015. [CrossRef]
31. Efthymiou, D.; Antoniou, C. How do transport infrastructure and policies affect house prices and rents? Evidence from Athens, Greece. *Transp. Res. Part A Policy Pract.* **2013**, *52*, 1–22. [CrossRef]
32. Dai, X.; Bai, X.; Xu, M. The influence of Beijing rail transfer stations on surrounding housing prices. *Habitat Int.* **2016**, *55*, 79–88. [CrossRef]
33. Tan, R.; He, Q.; Zhou, K.; Xie, P. The effect of new metro stations on local land use and housing prices: The case of Wuhan, China. *J. Transp. Geogr.* **2019**, *79*, 102488. [CrossRef]
34. Berawi, M.A.; Miraj, P.; Saroji, G.; Sari, M. Impact of rail transit station proximity to commercial property prices: Utilizing big data in urban real estate. *J. Big Data* **2020**, *7*, 71. [CrossRef]
35. Yang, L.; Liang, Y.; He, B.; Yang, H.; Lin, D. COVID-19 moderates the association between to-metro and by-metro accessibility and house prices. *Transp. Res. Part D Transp. Environ.* **2023**, *114*, 103571.
36. Okumura, T.; Ueda, K.; Iwamoto, Y.; Kanemoto, Y.; Shibata, A.; Yoshida, A.; Maquito, F. Housing investment and residential land supply in Japan:an asset market approach. *J. Jpn. Int. Econ.* **1997**, *11*, 27–54.
37. Gonzalez, A. *Resilience of Microfinance Institutions to National Macroeconomic Events: An Econometric Analysis of MFI Asset Quality*; MIX Discussion Paper No. 1; SSRN: Rochester, NY, USA , 2007. [CrossRef]
38. Mikhed, V.; Zemčík, P. Do house prices reflect fundamentals? aggregate and panel data evidence. *J. Hous. Econ.* **2009**, *18*, 140–149.
39. Genesove, D.; Han, L. Search and matching in the housing market. *J. Urban Econ.* **2012**, *72*, 31–45.
40. Sun, H.; Wang, Y.; Li, Q. The impact of subway lines on residential property values in Tianjin: An empirical study based on hedonic pricing model. *Discret. Dyn. Nat. Soc.* **2016**, *2016*, 1478413.
41. Hawkins, J.; Habib, K.N. Spatio-temporal hedonic price model to investigate the dynamics of housing prices in contexts of urban form and transportation services in Toronto. *Transp. Res. Rec.* **2018**, *2672*, 21–30.
42. Lisi, G. Hedonic pricing models and residual house price volatility. *Lett. Spat. Resour. Sci.* **2019**, *12*, 133–142. [CrossRef]
43. Lieske, N.S.; Nouwelant, R.; Han, J.H.; Pettit, C. A novel hedonic price modelling approach for estimating the impact of transportation infrastructure on property prices. *Urban Stud.* **2021**, *58*, 182–202.
44. Luo, H.; Zhao, S.; Yao, R. Determinants of housing prices in Dalian city, China: Empirical study based on hedonic price model. *J. Urban Plan. Dev.* **2021**, *147*, 05021017. [CrossRef]
45. Park, B.; Bae, J.K. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Syst. Appl.* **2015**, *42*, 2928–2934. [CrossRef]
46. Fan, C.; Cui, Z.; Zhong, X. House prices prediction with machine learning algorithms. In Proceesdings of the 37th International Conference on Machine Learning Conference, Vienna, Austria, 25–31 July 2020; pp. 6–10.

47. Quaranta, L.; Calefato, F.; Lanubile, F. KGTorrent: A dataset of python jupyter notebooks from kaggle. In Proceedings of the IEEE/ACM 18th International Conference on Mining Software Repositories, Madrid, Spain, 17–19 May 2021; pp. 550–554.
48. Miller, C.; Picchetti, B.; Fu, C.; Pantelic, J. Limitations of machine learning for building energy prediction: ASHRAE great energy predictor III kaggle competition error analysis. *Sci. Technol. Built Environ.* **2021**, *28*, 610–627. [CrossRef]
49. Varma, A.; Sarma, A.; Doshi, S.; Nair, R. House price prediction using machine learning and neural networks. In Proceedings of the Second International Conference on Inventive Communication and Computational Technologies, Coimbatore, India, 20–21 April 2018; pp. 1936–1939.
50. Zhang, Q. Housing price prediction based on multiple linear regression. *Sci. Prog.* **2021**, *2021*, 7678931. [CrossRef]
51. Peng, B.; Li, J.; Wang, Z.; Yang, R.; Liu, M.; Zhang, M.; Yu, P.S.; He, L. Lifelong property price prediction: A case study for the Toronto real sstate market. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 2765–2780. [CrossRef]
52. Durai, S.A.; Wang, Z. Resale HDB price prediction considering covid-19 through sentiment analysis. In Proceedings of the 10th European Conference on Social Media, Krakow, Poland, 18–19 May 2023; pp. 276–285.
53. Fan, C.; Wu, F.; Mostafavi, A. A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters. *IEEE Access* **2020**, *8*, 10478–10490. [CrossRef]
54. Monir, N.; Rasam, A.; Ghazali, R.; Suhandri, S.F. Address geocoding services in geospatial-based epidemiological analysis: A comparative reliability for domestic disease mapping. *Int. J. Geoinform.* **2021**, *17*, 156–166.
55. Panecki, T. Mapping imprecision: How to eeocode data from inaccurate historic maps. *ISPRS Int. J. Geo-Inform.* **2023**, *12*, 149. [CrossRef]
56. Chopde, N.R.; Nichat, M.K. Landmark based shortest path detection by using a* and haversine formula. *Int. J. Innov. Res. Comput. Commun. Eng.* **2013**, *1*, 298–302.
57. Alam, C.N.; Manaf, K.; Atmadja, A.R.; Aurum, D.K. Implementation of haversine formula for counting event visitor in the radius based on Android application. In Proceedings of the International Conference on Cyber and IT Service Management, Bandung, Indonesia, 26–27 April 2016; pp. 1–6.
58. Aliyu, A.; Kemiki, O.; Bello, M. Transportation accessibility benefit and the dynamic pattern of real estate prices: Emerging literature. *Path Sci.* **2018**, *4*, 1001–1016. [CrossRef]
59. Liu, F.; Chen, K.; Zhang, T.; Zhang, Y.; Song, Y. Will good service quality promote real estate value? Evience from Beijing, China. *Land* **2022**, *11*, 166. [CrossRef]
60. Gupta, A.; Nieuwerburgh, S.V.; Kontokosta, C. Take the Q train: Value capture of public infrastructure projects. *J. Urban Econ.* **2022**, *129*, 103422. [CrossRef]
61. Goo, C.-H. Apartment Brand January 2023 Big Data Analysis Results. Available online: https://brikorea.com/bbs/board.php?bo_table=rep_1&wr_id=2126 (accessed on 26 August 2023).
62. Kamp, H. Transport infrastructures and sustainability of urban development. *J. Irish Urban Stud.* **2002**, *1*, 37–46.
63. Lieske, S.N.; McLeod, D.M.; Coupal, R.H. Infrastructure development, residential growth and impacts on public service expenditure. *Appl. Spat. Anal. Policy* **2015**, *8*, 113–130. [CrossRef]
64. Juna, A.; Umer, M.; Sadiq, S.; Karamti, H.; Eshmawi, A.A.; Mohamed, A.; Ashraf, I. Water quality prediction using KNN imputer and multilayer perceptron. *Water* **2022**, *14*, 2592. [CrossRef]
65. Farrar, D.E.; Glauber, R.R. Multicollinearity in regression analysis: The problem revisited. *Rev. Econ. Stat.* **1967**, *49*, 92–107. [CrossRef]
66. Dupuis, D. J.; Victoria-Feser, M.-P. Robust VIF regression with application to variable selection in large data sets. *Ann. Appl. Stat.* **2013**, *7*, 319–341. [CrossRef]
67. Donate, J.P.; Cortez, P.; Sánchez, G.G.; Miguel, A.S. Time series forecasting using a weighted cross-validation evolutionary artificial neural network ensemble. *Neurocomputing* **2013**, *109*, 27–32. [CrossRef]