Urban Commercial Site Selection and Station Passenger Flow Prediction Based on Machine Learning: A Case Study of Tokyo's Yamanote Line

Chih-Hung Lin[1], Yu-Hao Ting[2], Li-Chun Chen[3], Ting-Wei Tsai[4]*

[1] School of Accounting and Finance, National Taipei University of Business, Taiwan
[2] Department of Finance, National Taipei University of Business, Taiwan
[3] Institute of Industrial Economics, National Central University, Taiwan
[4] Department of Digital Media Design, National Yunlin University of Science and Technology, Taiwan
11188001@ntub.edu.tw, howardting17@gmail.com, ab20020612@gmail.com, danieltsai816@gmail.com

Abstract

This study investigates the application of machine learning techniques to analyze passenger flow data and predict optimal commercial store types at stations along Tokyo's Yamanote Line. The Yamanote Line connects major commercial districts and bustling urban areas, with high passenger volumes and diverse consumer activity, making it a prime region for strategic retail deployment. However, determining the optimal store type and location remains a major challenge for businesses due to limited data-driven tools.

To address this, we propose a predictive framework that integrates open datasets from the Japanese Statistics Bureau, Seikatsu Guide.com and e-Stat. These datasets include station-level traffic, census demographics, consumer expenditure, and geographic information within a one-kilometer radius of each station. Feature selection was conducted using the Random Forest algorithm to identify key determinants influencing commercial placement. Subsequently, multiple classifiers—including Gradient Boosting, Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost)—were evaluated for prediction performance.

Experimental results show that XGBoost outperformed other models, achieving an F1-score of 0.642 and accuracy of 0.68. By constructing a data-driven location intelligence model, this research supports evidence-based commercial site selection and enhances decision-making in competitive urban retail environments. The proposed framework not only improves prediction accuracy but also offers practical guidance for retail site planning in smart cities globally. Future research could explore the incorporation of real-time dynamic data and deep learning techniques to further improve accuracy and adaptability.

Keywords: Machine learning, Location intelligence, Commercial site selection, Yamanote Line, XGBoost

1 Introduction

1.1 Research Background and Motivation

Urban commercial site selection is a critical decision that significantly impacts a company's success and competitiveness in the market.

Tokyo's Yamanote Line, one of the city's most important transportation arteries, connects major commercial districts and bustling urban centers. Its high passenger flow and diverse commercial activities make it a prime location for new store openings. However, effectively leveraging data analytics to determine the optimal store type and location remains a major challenge for many businesses.

Currently, many companies still rely on traditional site selection methods, such as field surveys, expert judgment, and geographic search tools like Yelp or NAVITIME. These approaches often suffer from fragmented information, a lack of quantitative support, and insufficient consideration of competition and market trends. With the rapid development of big data and machine learning technologies, improving the precision of site selection decisions through intelligent methods has become a key research focus.

This study proposes a machine learning–based predictive model for commercial site selection. By integrating multi-dimensional data—such as station passenger volume, demographic information, consumer spending, and geographic attributes—and applying advanced machine learning techniques including Gradient Boosting, Random Forest, Logistic Regression, and Extreme Gradient Boosting (XGBoost), this research aims to predict the most suitable store types for each station, thereby enhancing the accuracy and feasibility of business location strategies.

Moreover, studies on high-density commercial zones like Tokyo's Yamanote Line offer valuable implications for other global metropolitan areas seeking data-driven retail planning strategies [1].

1.2 Research Objectives

This study aims to develop an intelligent predictive model for commercial site selection by leveraging open data and machine learning techniques. The primary objectives are as follows:

1.2.1 Data Integration and Analysis

To construct a comprehensive database by integrating

*Corresponding Author: Chih-Hung Lin; Email: 11188001@ntub.edu.tw
DOI: https://doi.org/10.70003/160792642026072704013

multiple open data sources, including the Japanese Statistics Bureau, e-Stat, and Seikatsu Guide.com. The database includes information on train stations, passenger flow, census data, consumer expenditure, and geographic attributes.

1.2.2 Application of Machine Learning

To identify key variables that significantly influence commercial site selection. Feature selection is conducted using the Random Forest algorithm to extract the most representative decision-making factors. Various classification models, such as Gradient Boosting and XGBoost, are employed and evaluated to compare their prediction accuracy and performance.

1.2.3 Decision Support for Site Selection

To build a data-driven Location Intelligence Model that supports retail businesses in making evidence-based site selection decisions across different station areas, thereby reducing investment risks and enhancing the likelihood of market success.

Location Intelligence, as defined by Wolfe and Moon [2], refers to the strategic use of spatial data and analytics for informed business decision-making. This study adopts this perspective to guide model construction and variable selection.

2 Literature Review

With the rapid advancement of smart city development, the spatial layout and site selection strategies for commercial facilities have increasingly shifted toward data-driven and intelligent decision-making approaches. In particular, urban transportation hubs—such as areas surrounding metro and railway stations—have emerged as highly competitive locations for new store openings due to their high foot traffic and diverse consumer potential. Effectively utilizing big data and artificial intelligence (AI) technologies to accurately analyze the multifaceted factors influencing location decisions and to forecast the most suitable commercial types has become a key issue of concern for both academia and industry.

Accordingly, this study adopts a perspective grounded in machine learning and location intelligence to review and synthesize recent literature. The goal is to examine the current applications and development potential of various analytical methods in commercial site selection, thereby laying the theoretical foundation for the construction of our proposed predictive model.

2.1 Applications of Machine Learning in Commercial Site Selection

In recent years, the development of machine learning (ML) techniques has significantly advanced data-driven decision-making in the areas of commercial site selection and market analysis. Machine learning enables automatic classification and pattern recognition from large volumes of data, improving both the accuracy and efficiency of decision-making processes.

Traditionally, research on commercial location selection has relied heavily on statistical models—such as logistic regression—to analyze the factors influencing store placement decisions. However, these methods often assume the absence of multicollinearity among variables and struggle to capture complex nonlinear relationships. Although traditional models (such as logistic regression) offer strong interpretability, ensemble learning techniques (e.g., Random Forest and XGBoost) often demonstrate superior predictive accuracy in high-dimensional or non-linear research contexts [11-13].

In contrast, ensemble learning methods—such as Random Forest (RF) and Gradient Boosting (GB)—enhance predictive performance through multiple decision trees and can identify the key factors in location decision-making processes [1]. Among predictive modeling techniques, Extreme Gradient Boosting (XGBoost) has demonstrated outstanding performance in variable selection and accuracy across diverse domains, including finance, retail, and healthcare [3]. For instance, Noorunnahar et al. (2023) applied XGBoost to COVID-19 transmission forecasting, demonstrating its robustness in handling high-dimensional and temporal data, while other studies have validated its applicability in consumer demand prediction and retail location analysis.

Therefore, this study investigates the feasibility of applying XGBoost to commercial site prediction and compares its performance with other machine learning models.

2.2 Research on Site Selection Decisions and Influencing Factors

The success of commercial site selection depends on a combination of factors, including consumer behavior, traffic volume, rental costs, and the competitive landscape. Although prior studies have proposed various approaches to site selection decision-making, several limitations remain.

2.2.1 GIS-Based Site Selection Analysis

The application of Geographic Information Systems (GIS) in commercial site selection has been widely studied. For example, Kao et al. (2013) employed GIS to analyze optimal locations for convenience stores in Hsinchu City, while Meng-Hung Xie (2009) applied GIS techniques to department store site selection. While these methods offer valuable spatial visualization capabilities, they often fail to incorporate real-time market data and dynamic consumer behavior patterns.

2.2.2 Recommendation System-Based Commercial Location Selection

In recent years, some studies have attempted to integrate recommendation systems into the site selection process. For instance, Mao et al. (2019) adopted collaborative filtering techniques to recommend suitable store types based on consumer preferences and geographic proximity. However, such methods are typically reliant on historical data, making them less adaptable to rapidly changing market conditions and thus limited in their practical applicability in dynamic commercial environments.

2.2.3 Competition Analysis-Based Site Selection

Li et al. (2017) proposed a social path-based recommendation mechanism, analyzing the relationship

between competing stores and consumer preferences to identify optimal store locations. Similarly, Pan et al. (2023) examined spatial correlations between cafés and teahouses in Shanghai, identifying public transportation access, policy direction, and market positioning as key factors affecting store distribution. However, these studies largely focus on specific industries and lack comparative analysis across different sectors.

2.3 Limitations of Existing Research

Based on the reviewed literature, this study identifies the following research gaps:

2.3.1 Lack of Cross-Industry Commercial Site Prediction

Most existing studies focus on a single industry—such as food and beverage or retail—and thus fail to provide comprehensive site selection recommendations across multiple industry sectors.

2.3.2 Limited Capability of Traditional Methods in Handling Multidimensional Data

Conventional approaches, such as GIS-based analysis and statistical regression models, struggle to integrate and account for multiple influencing factors simultaneously, including traffic volume, consumer purchasing power, and competitive dynamics.

2.3.3 Insufficient Comparative Analysis of Machine Learning Models

Although some studies have applied machine learning to commercial site selection, few have conducted comprehensive performance comparisons among different algorithms, such as Gradient Boosting, Logistic Regression, Random Forest, and XGBoost.

2.4 Comparison of Prior Approaches

Table 1 summarizes the key differences between this study and existing literature in terms of site selection methodologies, data integration strategies, and applied models.

The comparison highlights this study's more comprehensive use of machine learning techniques and open data in contrast to traditional statistical or GIS-based approaches.

By positioning this research within the methodological landscape, Table 1 illustrates the advancement and novelty of the proposed modeling framework.

2.5 Contributions of This Study

The key contributions of this research are as follows:

2.5.1 Development of a Data-Driven Location Intelligence Model

This study develops a data-driven location intelligence model that integrates multidimensional datasets to enhance the accuracy of commercial site selection decisions.

2.5.2 Comparative Evaluation of Machine Learning Models

By comparing the predictive performance of various machine learning algorithms—namely Random Forest, Logistic Regression, Gradient Boosting, and XGBoost—this study identifies the most effective model for supporting commercial location decisions.

2.5.3 Application to Tokyo's Yamanote Line

The proposed approach is empirically validated using data from Tokyo's Yamanote Line, demonstrating the effectiveness of machine learning techniques in high-density urban commercial areas and offering insights for future research.

3 Data Source and Processing

3.1 Research Framework and Process

This study applies machine learning techniques to develop a predictive model for commercial site selection. By integrating multidimensional datasets-including passenger traffic, geographic information, census data, and consumer expenditure—the model assists businesses in making more accurate location decisions. The research process is illustrated in Figure 1.

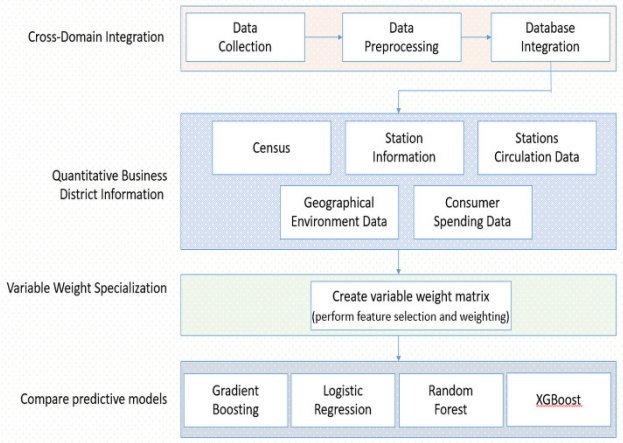

Figure 1. Research framework and workflow

3.1.1 Data Collection and Integration

Open data sources were aggregated, including governmental statistical data, commercial datasets, and GIS-based spatial information.

3.1.2 Data Preprocessing

This step involved handling missing values, standardizing data, and selecting relevant features for analysis.

3.1.3 Machine Learning Modeling

Random Forest was first used for feature selection. Subsequently, multiple classification algorithms—including Gradient Boosting, Logistic Regression, Random Forest, and XGBoost—were applied to develop and evaluate predictive models.

3.1.4 Results Analysis and Validation

The performance of different models was compared to validate prediction outcomes and provide strategic recommendations for retail location planning.

3.2. Data Sources

The primary data used in this study were obtained from publicly available government datasets, transportation traffic records, consumer behavior statistics, and geographic information systems. These datasets cover areas within a one-kilometer radius around 30 major

stations along Tokyo's Yamanote Line and serve as the foundation for model development and analysis.

The dataset used in this study consists of four major categories sourced from government and commercial open data platforms. First, passenger flow data were collected from JR East Open Data, providing daily boarding and alighting records for all 30 stations along the Yamanote Line in the year 2022. Second, census data were obtained from the Japanese Statistics Bureau, based on the latest 2020 national survey conducted every five years, offering demographic information such as population, age distribution, and household structure. Third, consumer spending data covering the years 2019 to 2022 were retrieved annually from e-Stat and Seikatsu Guide, including detailed area-based statistics on retail consumption. Finally, geographic information, including land use classifications and commercial zoning boundaries, was provided by the Tokyo Metropolitan Government's 2022 open GIS datasets.

This paragraph-based presentation integrates data type, source institution, coverage period, and update frequency, ensuring transparency and reproducibility.

3.3 Data Preprocessing

Prior to analysis, data cleaning and transformation were conducted to ensure consistency and reliability across datasets. The preprocessing steps are as follows:

3.3.1 Handling Missing Values

For stations with missing passenger flow data, historical averages were used to fill in the gaps. In cases where geographic information was incomplete, missing data were supplemented using the Google Maps API.

3.3.2 Data Normalization

Since the variables (e.g., rental costs and passenger volumes) are measured in different units, all numerical features were normalized using Min-Max scaling, transforming values to a range between 0 and 1. This prevents scale-related biases in model training.

3.3.3 Feature Engineering

3.3.3.1 Feature Selection

The Random Forest algorithm was used to compute feature importance scores and identify the most influential variables affecting commercial site selection.

3.3.3.2 Construction of NewVariables

Several composite indicators were created to enhance the model's interpretability and predictive capability:

- Competition Intensity Index = Number of similar stores around the station / Total passenger flow
- Commercial Potential Index = (Population density × Average income) / Rental cost
- Transportation Accessibility Index = Station classification (major/minor) × Average daily passenger flow

These engineered features allow the model to better capture the underlying commercial potential of each station area.

3.4 Data Integration and Application

After data preprocessing, the final dataset was compiled with a set of key variables used as input features for the machine learning models to predict optimal store types at each station. The main variables included in the dataset are as Table 3 follows:

Table 3. Key variables for model input

Variable name	Description
Passenger Flow	Daily number of passengers entering and exiting each station
Population Density	Number of residents per square kilometer within a 1 km radius of the station
Average Consumer Expenditure	Monthly average spending per capita in the corresponding area
Rental Cost	Average commercial rent in the target area
Competition Intensity Index	Ratio of similar stores to total passenger flow within the station vicinity
Commercial Potential Index	Composite index calculated as (population density × average income) / rent
Transportation Accessibility Index	Combined indicator of station type (major/minor) and average daily traffic

These features were selected to reflect both socio-economic and spatial characteristics of the station surroundings, allowing the predictive model to capture regional commercial potential more effectively.

To highlight the comprehensiveness of this study's feature selection, a comparison of key variable dimensions adopted across major related studies is presented in Table 2. This table demonstrates that the proposed model integrates multiple commercial, demographic, and flow-related indicators more extensively than prior approaches.

4 Model Development and Evaluation

4.1 Feature Selection

In developing our predictive model for optimal store type and location, we drew on multiple data streams. We selected features that capture the annual station passenger counts from 2018 to 2022 (including both habitual and occasional riders), the density of public amenities such as parks and libraries, average residential and commercial land values, and property vacancy rates (including rental vacancies). The model's outcome variable classifies each station's most suitable store category, spanning a range of options from izakayas, sushi bars, yakiniku, buffets, and curry shops to cafés and dessert outlets, ramen shops, family restaurants, fast-food chains, and various entertainment and dining venues (e.g., karaoke, banquet halls, Japanese and Western cuisines, hotpot, Asian eateries, bars/pubs).

4.2 Model Selection

To construct a reliable and accurate prediction model for commercial site selection, this study evaluates and compares multiple machine learning algorithms based on the following criteria:

4.2.1 Suitability for Classification Problems

The objective is to classify the most appropriate commercial type per station; thus, classification models are selected.

4.2.2 Balance of Interpretability and Accuracy

Algorithms are selected based on prediction accuracy, computational efficiency, and model interpretability.

4.2.3 Capability to Handle High-Dimensional Data

Models capable of managing heterogeneous and multidimensional data (e.g., passenger flow, geographic information, consumer behavior) are prioritized.

4.2.4 The Following Models are Employed in this Study

4.2.4.1 Gradient Boosting (GB)

- Advantages: Ensemble of decision trees that improves prediction accuracy and captures nonlinear relationships.
- Limitations: Computationally intensive and sensitive to hyperparameter tuning.

4.2.4.2 Logistic Regression (LR)

- Advantages: High interpretability; useful for identifying the influence of key decision variables.
- Limitations: Assumes no multicollinearity and is limited in capturing nonlinear effects.

4.2.4.3 Random Forest (RF)

- Advantages: Automatically identifies important variables; robust in handling large and heterogeneous datasets.
- Limitations: Slower when processing large-scale data; high computational cost.

4.2.4.4 Extreme Gradient Boosting (XGBoost)

- Advantages: Handles missing values automatically, performs well with high-dimensional and temporal data, and reduces overfitting via regularization and parallel computation.
- Limitations: Requires careful tuning of hyperparameters to optimize performance.

4.3 Parameter Configuration

4.3.1 Logistic Regression (Logit Model)

In the logistic model, the target variable is binary. The logarithm of the ratio of the counts for the two binary outcomes is called the log odds. This ratio, known as the odds, represents the likelihood of an event occurring versus not occurring. The logistic model is derived from the log probability function and has two possible outcomes: event occurrence (Yi = 1) or non-occurrence (Yi = 0). The probability of event occurrence is denoted as p, and non-occurrence as 1-p. The calculation equations are given in formulas (1) and (2).

$$\text{occurrence } P_i = prob\left(Y_i = 1\right) = \frac{e^{f(x)}}{1+e^{f(x)}} \tag{1}$$

$$\text{non-occurrence } 1-P_i = prob\left(Y_i = 0\right) = \frac{1}{1+e^{f(x)}} \tag{2}$$

When the probability of an event occurring is $P(x)$, using Logistic Regression, equation (3) is formulated as follows:

$$P(x) = \ln\left[\frac{p}{1-p}\right] = \ln\left[e^{f(x)}\right] = f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{3}$$

Where P = the probability of success; β_0 = the constant term; β_i = coefficient; X_i = independent variables; i = 1, 2, 3, …, k.

In this study, the Logistic Regression model adopted the "lbfgs" solver, which is suitable for large datasets and ensures faster convergence compared to "liblinear." The lbfgs solver, based on a quasi-Newton method, also supports multinomial classification, whereas liblinear is limited to binary classification. Thus, lbfgs was selected to enhance computational efficiency and maintain robustness under our data scale.

Furthermore, L2 regularization was applied to prevent overfitting, and the regularization strength parameter (C) was set to 1.0, which follows common defaults and balances bias and variance.

4.3.2 Decision Tree

A hierarchical model composed of nodes and branches used for classification and regression. Common algorithms include ID3, CART, and C5.0. Tree pruning is applied to prevent overfitting.

In this study, the Decision Tree model was configured with the default setting for max_depth, which allows the tree to grow fully in the initial stage. This approach ensures that potential non-linear feature interactions can be captured without prematurely limiting the model's representational capacity. To address the risk of overfitting associated with deep trees, cost-complexity pruning (post-pruning) was applied to remove branches with limited contribution to generalization. This strategy balances interpretability and predictive accuracy, while preventing overfitting to the training data.

4.3.3 Random Forest

Introduced by Breiman [4], this ensemble model builds multiple decision trees using the bagging method. Random feature subsets and error estimation ensure diversity and model robustness.

In this study, the number of estimators for the Random Forest model was set to 100, balancing between predictive power and computational efficiency. The max_depth parameter was left as default to allow the model to grow trees fully and capture potential non-linear interactions.

4.3.4 XGBoost

Developed by Chen Tianqi at the University of Washington, XGBoost is a Gradient Boosting Decision Tree (GBDT) that iteratively improves prediction by correcting residuals from prior models. It has achieved outstanding results in time-series forecasting and medical prediction (e.g., COVID-19 spread [5]).

In this study, the max_depth for XGBoost was set to 6, balancing model complexity and overfitting risk. The number of estimators was set at 100 based on preliminary

cross-validation results to ensure convergence and manageable training time.

4.3.5 Gradient Boosting Classifier

In this study, the number of estimators for the Gradient Boosting Classifier was set to 100 and the learning rate to 0.1, following common defaults to balance accuracy and convergence. The max_depth was set to 3 to prevent overfitting and encourage generalization.

Data preprocessing incorporated imputing missing entries by their mean or median, scaling all features using Min–Max normalization, and converting categorical variables via one-hot encoding. Key hyperparameters comprised the number of boosting iterations (n_estimators), the update step size to guard against overfitting (learning_rate), the maximum depth of each tree (max_depth), and the fraction of training samples employed in each boosting round (subsample).

4.4 Model Training and Evaluation

4.4.1 Training Strategy

To mitigate the risk of overfitting, a 5-fold cross-validation procedure was employed during the training phase. This approach partitions the dataset into five mutually exclusive subsets, iteratively using four folds for training and one for validation, thereby ensuring that the evaluation is not biased toward a particular data partition. The results demonstrated that the XGBoost model yielded stable predictive performance across folds, with the standard deviation of the F1-score remaining below 0.02, indicating robustness and consistency of the model.

The dataset was randomly split into 70 % for training and 30 % for testing, and model performance was evaluated using accuracy (overall correct classifications), precision (true positives among predicted positives), recall (true positives among actual positives), and the F1 score (the harmonic mean of precision and recall).

4.4.2 Evaluation Metrics

Model performance was assessed using overall accuracy, the F1 score to balance precision and recall, and confusion matrices to examine class-specific errors.

5 Result and Discussion

5.1 Comparison of Model Performance

To evaluate the predictive performance of different machine learning models, this study adopts Accuracy and F1 Score as primary evaluation metrics. The results are summarized in Table 4.

Table 4. Prediction performance of different machine learning models

Model type	F1 score	Accuracy
Gradient Boosting	0.6061	0.65
Logistic Regression	0.6284	0.65
Random Forest	0.6289	0.66
XGBoost	0.6422	0.68

The XGBoost model achieved the highest F1 Score (0.6422) and Accuracy (0.68), indicating superior predictive capability. This model integrated multi-source data—including station information, census data, passenger volume, consumer expenditure, and geographic characteristics—and used weekly passenger flow data collected via Google Maps crawlers from 2018 to 2022.

These results demonstrate the practical effectiveness of XGBoost in forecasting suitable store types around Tokyo's Yamanote Line stations.

5.2 Model Comparison and Discussion

To further analyze the prediction results of different models, Figures 2 through 9 present the confusion matrices and heatmap analysis corresponding to each classifier.

5.2.1 Gradient Boosting

The confusion matrix of the Gradient Boosting classifier (Figure 2) shows 130 samples in the True Positive (TP) region and 181 samples in the True Negative (TN) region, indicating a relatively acceptable performance for this model.

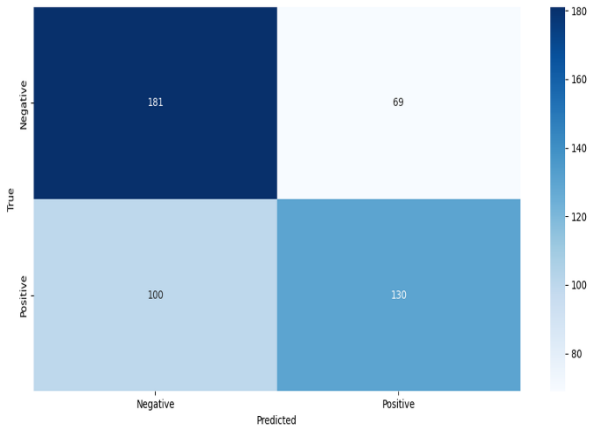

Figure 2. Confusion matrix-gradient boosting

As shown in the heatmap for this model (Figure 3), most features demonstrate negative correlations. With an F1 Score of 0.606061 and an Accuracy of 0.647917, Gradient Boosting exhibits the lowest performance among all evaluated models, suggesting relatively weak predictive accuracy.

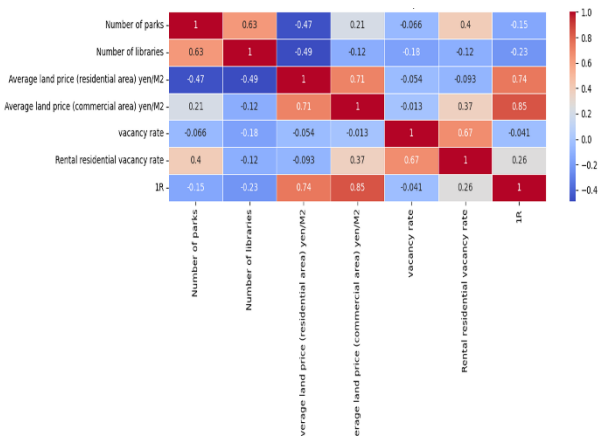

Figure 3. Heatmap-gradient boosting

5.2.2 Logistic Regression

The confusion matrix for the Logistic Regression model (Figure 4) shows 134 samples classified as True Positives (TP) and 177 samples as True Negatives (TN). Compared with other models, this indicates that Logistic Regression yields relatively lower accuracy within the context of this study.

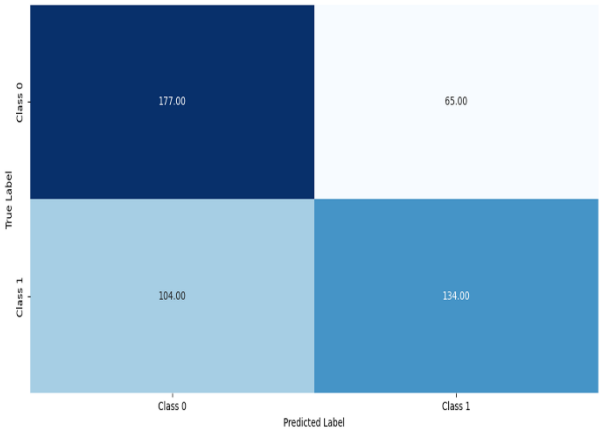

Figure 4. Confusion matrix-logistic regression

As shown in the heatmap (Figure 5), certain features appear to influence model accuracy. However, most correlations among variables are either negative or weak. With an F1 Score of 0.628384 and an Accuracy of 0.65—similar to that of Gradient Boosting—Logistic Regression is considered one of the relatively less accurate models in this comparison.

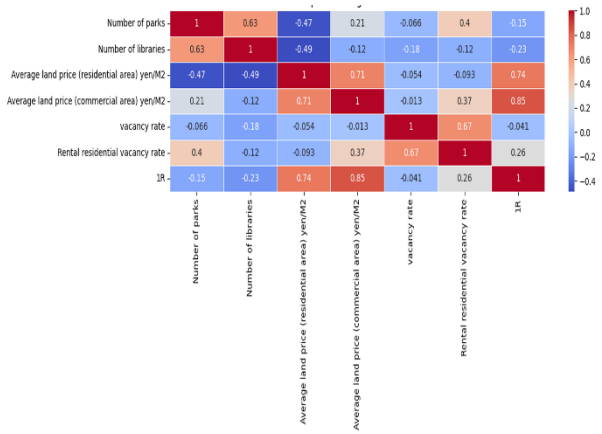

Figure 5. Heatmap-logistic regression

5.2.3 Random Forest

As shown in Figure 6, the Random Forest model correctly identified 139 samples as True Positives (TP) and 179 samples as True Negatives (TN), making it the second most accurate model in this study.

The heatmap for the Random Forest model (Figure 7) illustrates strong correlations among multiple features. This suggests that adjustments to one variable—such as the number of libraries-may influence other associated factors, which in turn could affect consumer spending behavior and product preferences. These interdependencies may lead to changes in prediction outcomes. With an F1 Score of 0.628895 and an Accuracy of 0.66, Random Forest demonstrated the second-best performance among the evaluated models.

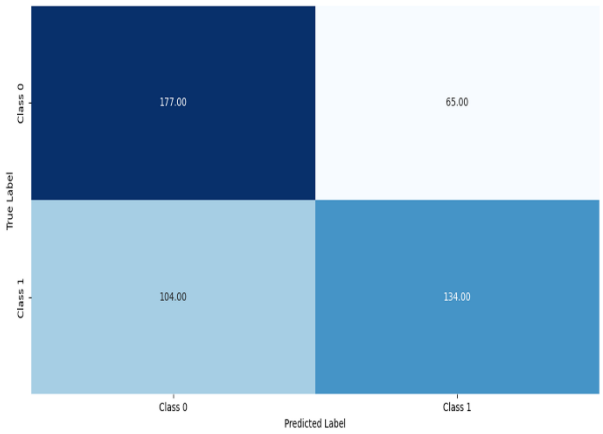

Figure 6. Confusion matrix-random forest

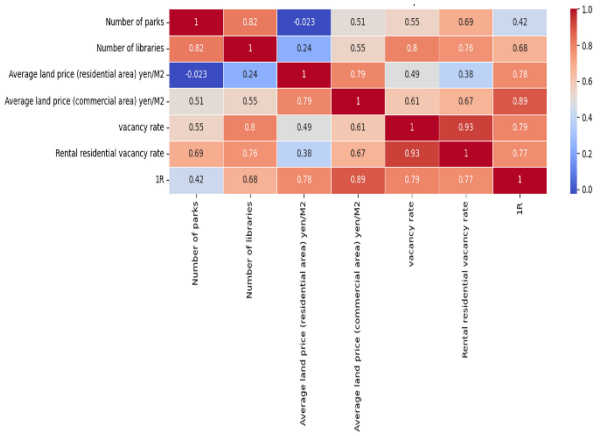

Figure 7. Heatmap-random forest

5.2.4 Extreme Gradient Boosting (XGBoost)

The confusion matrix for the XGBoost classifier (Figure 8) shows 138 samples as True Positives (TP) and 186 samples as True Negatives (TN), indicating that this model achieved the highest predictive accuracy among all the models evaluated in this study.

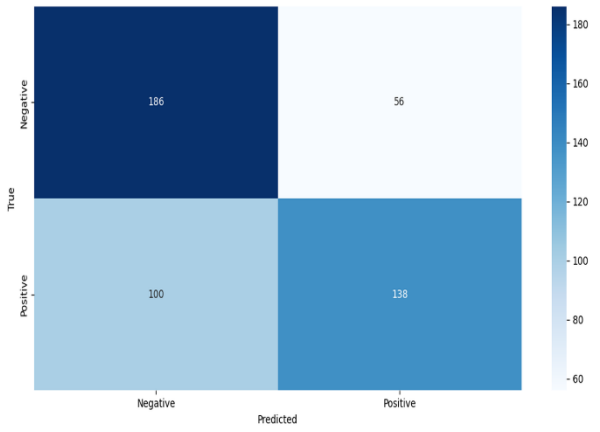

Figure 8. Confusion matrix-XGBoost

The heatmap for XGBoost (Figure 9) reveals strong correlations among various features. This suggests that changes to one variable—such as the number of libraries-can influence other associated factors, which may affect consumer spending behavior and product preferences, ultimately altering prediction outcomes. With an F1 Score of 0.642166 and an Accuracy of 0.68, XGBoost demonstrated the best performance, making it the recommended model for commercial site prediction in this study.

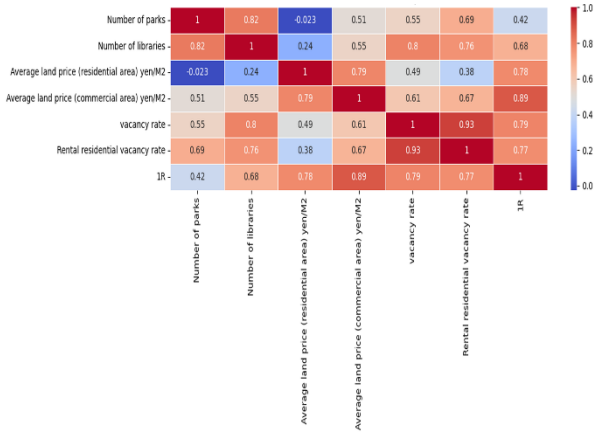

Figure 9. Heatmap-XGBoost

The superior performance of XGBoost can be attributed to its ability to effectively capture non-linear relationships and complex feature interactions, which are prevalent in urban commercial site selection contexts. Its built-in regularization further mitigates overfitting, ensuring robust generalization across diverse station environments. Moreover, interpretability analyses indicate that passenger volume consistently emerged as the most influential predictor, followed by income level, suggesting that both demand intensity and consumer purchasing power jointly drive optimal site recommendations.

5.3 Feature Importance Analysis

In the XGBoost model, feature importance refers to the contribution of each variable to the model's predictive performance. These importance scores serve as the foundation for interpretability and guide further feature selection. The key influential factors identified are as follows:

5.3.1 Passenger Flow (Daily Inflow and Outflow per Station)

Daily passenger traffic is the most important variable in the model. High foot traffic stations are strongly associated with higher commercial potential and business value.

5.3.2 Population Density (Population within a 1-Kilometer Radius / Area)

Population density has a moderate impact on prediction results and aligns with overall retail demand patterns in urban areas.

5.3.3 Rental Cost (Average Commercial Rent per Unit Area)

Rental cost is a critical factor in determining investment risk at a specific location.

5.3.4 Competitive Environment (Number of Similar Stores / Passenger Flow)

The level of competition directly influences investment risk, as areas with higher saturation may reduce profitability for new entrants.

5.4 Conclusion

5.4.1 Research Findings

This study demonstrates that machine learning approaches can more accurately predict optimal commercial site locations than traditional selection methods, with XGBoost outperforming other models due to its ability to handle high-dimensional data and maintain strong generalization.

5.4.2 Implications for Commercial Site Selection

Businesses ought to prioritize data-driven analyses—especially of passenger flow and consumer behavior—over intuition-based strategies, recognizing that high foot traffic alone does not guarantee profitability and that investment decisions must also account for competitive intensity and rental costs.

5.4.3 Academic and Practical Contributions

This study contributes not only a predictive model with superior accuracy, but also a demonstration of how integrating cross-domain open data—such as station flow, demographic indicators, and consumer behavior—can meaningfully enhance commercial location decision-making in dense urban contexts. In practice, the model offers actionable insights for retail strategists, urban developers, and policymakers, enabling more informed and data-driven spatial planning. Our findings suggest that ensemble learning methods, particularly XGBoost, provide more reliable predictions than traditional statistical models, enabling organizations to make data-driven and resilient site selection decisions. For retail enterprises, this translates into reduced investment risks and optimized store placement, while for the insurance industry, it supports strategic branch network planning and enhances customer accessibility within smart city ecosystems. Furthermore, the framework may be adapted to other metropolitan areas worldwide, promoting scalable applications in smart city initiatives. For instance, our results suggest that in high-footfall areas such as Shinjuku and Shibuya, convenience stores and fast-service restaurants are better aligned with predicted consumer flows, whereas in high-income districts such as Meguro and Bunkyo, specialty retail and premium dining options are more strongly supported by the model outputs.

5.4.4 Limitations and External Validity

While our empirical case on Tokyo's Yamanote Line provides a rigorous testbed for the proposed framework, we acknowledge that its focus on a single urban rail corridor may limit the generalizability of our findings. Commercial environments, consumer behaviors, and transportation infrastructures can vary substantially across different cities and regions. Therefore, we recommend a more detailed discussion of external validity, noting that adaptations—such as recalibrating feature sets or retraining models on local open data—would be required to apply this framework elsewhere. Future studies should explore

cross-city validations and assess how regional differences (e.g., land use patterns, fare structures, or demographic profiles) influence the model's predictive accuracy and practical recommendations.

Future research could further incorporate real-time smart card transaction data to capture dynamic fluctuations in urban rail passenger flow, combined with deep learning models (e.g., LSTM or GRU) to enhance time-series predictive capabilities. Such an approach would not only better reflect short-term seasonality and sudden changes in ridership but also provide more timely and evidence-based guidance for commercial site selection, thereby strengthening the practical implications of this study for urban retail planning.

References

[1] B. Mahesh, Machine learning algorithms – A review, *International Journal of Science and Research (IJSR)*, Vol. 9, No. 1, pp. 381–386, January, 2019. https://doi.org/10.21275/ART20203995

[2] D. Wolfe, G. Moon, Location intelligence, in: S. Shekhar, H. Xiong (Eds.), *Encyclopedia of GIS*, New York, NY: Springer, 2008, pp. 627–630.

[3] Y.-N. Tu, W.-T. Hsu, H.-C. Huang, M.-K. Hsu, Y.-H. Lin, J.-J. Hung, Combing the Government Open Datasets to Build the Location Selection Decision Support System -The Local Consumption Ability and Neighboring Industries Perspectives, *Instruments Today*, Vol. 215, pp. 4–21, June, 2018.

[4] L. Breiman, Random forests, *Machine Learning*, Vol. 45, No. 1, pp. 5–32, October, 2001. https://doi.org/10.1023/A:1010933404324

[5] M. Noorunnahar, A. H. Chowdhury, F. A. Mila, A tree based eXtreme Gradient Boosting (XGBoost) machine learning model to forecast the annual rice production in Bangladesh, *PLoS ONE*, Vol. 18, No. 3, Article No. e0283452, March, 2023. https://doi.org/10.1371/journal.pone.0283452.

[6] Z. Yu, M. Tian, Z. Wang, B. Guo, T. Mei, Shop-Type Recommendation Leveraging the Data from Social Media and Location-Based Services, *ACM Transactions on Knowledge Discovery from Data*, Vol. 11, No. 1, Article No. 1, February, 2017. https://doi.org/10.1145/2930671

[7] W. Li, L. Sui, M. Zhou, H. Dong, Short-Term Passenger Flow Forecast for Urban Rail Transit Based on Multi-Source Data, *Journal on Wireless Communications and Networking*, Vol. 2021, Article No. 9, January, 2021. https://doi.org/10.1186/s13638-020-01881-4

[8] R. Conforti, M. de Leoni, M. La Rosa, W. M. P. van der Aalst, A. H. M. ter Hofstede, A recommendation system for predicting risks across multiple business process instances, *Decision Support Systems*, Vol. 69, pp. 1–19, January, 2015. https://doi.org/10.1016/j.dss.2014.10.006

[9] Y.-M. Li, L.-F. Lin, C.-C. Ho, A Social Route Recommender Mechanism for Store Shopping Support, *Decision Support Systems*, Vol. 94, pp. 97–108, February, 2017. https://doi.org/10.1016/j.dss.2016.11.004

[10] C.-C. Kao, G.-W. Lee, Convenience store site selection analysis in Hsinchu City, *Geographic Information System*, Vol. 7, No. 4, pp. 25–28, October, 2013. https://doi.org/10.6628/GIS.2013.7(4).6

[11] K. Janowicz, S. Gao, G. McKenzie, Y. Hu, B. Bhaduri, GeoAI: Spatially Explicit Artificial Intelligence Techniques for Geographic Knowledge Discovery and Beyond, *International Journal of Geographical Information Science*, Vol. 34, No. 4, pp. 625–636, 2020. https://doi.org/10.1080/13658816.2019.1684500

[12] Y. Sadahiro, H. Matsumoto, Market Area Analysis with a Focus on the Spatial Relationship Between Sites and Their Visitors, *Transactions in GIS*, Vol. 28, No. 5, pp. 1111–1129, August, 2024. https://doi.org/10.1111/tgis.13167

[13] J. Lu, X. Zheng, E. Nervino, Y. Li, Z. Xu, Y. Xu, Retail Store Location Screening: A Machine Learning-Based Approach, *Journal of Retailing and Consumer Services*, Vol. 77, Article No. 103620, March, 2024. https://doi.org/10.1016/j.jretconser.2023.103620

Biographies

Chih-Hung Lin holds an M.S. in Information Management from NCU and is a Ph.D. student at the School of Accounting and Finance, National Taipei University of Business, Taiwan. His research spans AI, fintech, information management, cybersecurity, and marketing. He holds PMP, PMI-ACP, ISO/IEC 27001, and ISO 22301 certifications as well.

Yu-Hao Ting graduated from the Department of Finance at National Taipei University of Business. His research interests include financial econometrics, data visualization, machine learning, and natural language processing (NLP). He is currently working as a business development specialist.

Li-Chun Chen is a master's student at the Institute of Industrial Economics, National Central University. Her research focuses on outward foreign direct investment, international trade, and heterogeneous impacts on R&D and firm performance in Taiwan's biotechnology and medical sector.

Ting-Wei Tsai is currently studying in the Department of Digital Media Design at National Yunlin University of Science and Technology. His research interests include digital content monetization and design strategies, as well as design psychology and purchasing decision-making.

Appendix

Table 1. Methodological comparison between this study and prior research

Study & Year	Model & Method	Data & Features	Application domain	Key findings
This study (2025)	GB, Logit, RF, XGBoost	Consumer preferences, geolocation	Retail store type prediction (Tokyo, Yamanote Line)	Integrated open data and ML to optimize site selection and store type prediction
Yu et al. [6], 2017	Shop-Type Recommendation, LBS + Social Media	Social media, LBS, POI	Retail shop type recommendation	Integrated social media and location-based information for shop-type recommendation
Li et al. [7], 2021	SARIMA-SVM Hybrid	Passenger flow, IoT sensors	Urban rail passenger flow prediction	Improved short-term passenger flow forecasting using multi-source data
Conforti et al. [8], 2015	Risk Prediction System	Business process instances	Operational risk in business processes	Proposed predictive model for cross-instance risk management
Li et al. [9], 2017	Context-aware Route Recommendation	Shopping trajectory, social network, Markov chain	Retail shopping support	Personalized shopping route recommendation using contextual information
Kao & Lee [10], 2013	GIS-based Analysis	POI, road network	Convenience store site in Hsinchu	Evaluated store proximity and accessibility using GIS
Janowicz et al. [11], 2020	GeoAI	Spatial big data, GIS	Location intelligence	Established GeoAI framework for spatial AI applications
Sadahiro & Matsumoto [12], 2024	Market Area Analysis	Visitor distribution, spatial interaction	Retail market area	Analyzed visitor-site spatial relationships for site planning
Lu et al. [13], 2024	Machine Learning Screening	POI, competitor, demographic, TGI	Retail store location	ML-based retail location screening using public data

Table 2. Comparison of feature variables across studies

Study & Year	Multi-store types	Passenger flow included	ML-based method used	Competition analyzed
This study (2025)	✓	✓	✓	✓
Yu et al. [6], 2017	✓	✗	✓	✗
Li et al. [7], 2021	✗	✓	✓	✗
Conforti et al. [8], 2015	✗	✗	✓	✓
Li et al. [9], 2017	✓	✗	✓	✓
Kao & Lee [10], 2013	✗	✓	✗	✗
Janowicz et al. [11], 2020	✗	✓	✓	✗
Sadahiro & Matsumoto [12], 2024	✓	✓	✗	✓
Lu et al. [13], 2024	✓	✓	✓	✓