Knowledge and Information Systems (2020) 62:1005–1035
https://doi.org/10.1007/s10115-019-01373-y

REGULAR PAPER

A systematic framework of predicting customer revisit with in-store sensors

Sundong Kim[1] · **Jae-Gil Lee**[1,2]

Received: 4 January 2019 / Revised: 27 May 2019 / Accepted: 10 June 2019 / Published online: 29 June 2019
© Springer-Verlag London Ltd., part of Springer Nature 2019

Abstract
Recently, there is a growing number of off-line stores that are willing to conduct customer behavior analysis. In particular, predicting revisit intention is of prime importance, because converting first-time visitors to loyal customers is very profitable. Thanks to noninvasive monitoring, shopping behaviors and revisit statistics become available from a large proportion of customers who turn on their mobile devices. In this paper, we propose a systematic framework to predict the revisit intention of customers using Wi-Fi signals captured by in-store sensors. Using data collected from seven flagship stores in downtown Seoul, we achieved 67–80% prediction accuracy for all customers and 64–72% prediction accuracy for first-time visitors. The performance improvement by considering customer mobility was 4.7–24.3%. Furthermore, we provide an in-depth analysis regarding the effect of data collection period as well as visit frequency on the prediction performance and present the robustness of our model on missing customers. We released some tutorials and benchmark datasets for revisit prediction at https://github.com/kaist-dmlab/revisit.

Keywords Revisit prediction · Retail analytics · Predictive analytics · Feature engineering · Marketing · Mobility data

1 Introduction

1.1 Motivation

By identifying potentially loyal customers who are more likely to revisit, merchants can considerably save on promotion cost and enhance return on investment [27]. Many studies in recent years have focused on *online* stores and online text reviews with the help of a data provider [18,42]. In contrast, the analysis of revisit intention in the *off-line* environment has not been carried out. The main reason lies in the difficulties of collecting large-scale data that is closely related to key attributes of revisiting, such as customer satisfaction with

✉ Jae-Gil Lee
jaegil@kaist.ac.kr

1 Graduate School of Knowledge Service Engineering, KAIST, Daejeon, Republic of Korea

2 Department of Industrial and Systems Engineering, KAIST, Daejeon, Republic of Korea

products, service quality, atmosphere, purchase history, and personal profiles [37,42]. Those attributes are either subjective or confidential, which are not easily accessible. Owing to these limitations, research on customer revisits in off-line stores has been conducted through surveys. These studies help us gain an understanding of underlying hypotheses that affect customer satisfaction. However, their findings cannot be easily generalized because of a small sample size.

With the advance of sensing technologies such as radio-frequency identification (RFID) [8,35], Bluetooth [45], and Wi-Fi fingerprinting [36], we are capable of collecting high-frequency signal data without installing any applications on customer devices [29,30]. These signals can be converted to fine-grained mobility data. Using such data, noninvasive monitoring of visitors has been carried out in different settings, such as in museums [45] and supermarkets [40], providing empirical findings of customer behaviors. Nowadays, collecting data in a certain physical boundary is called as geofencing [32] and its market size is increasing rapidly. Companies such as ZOYI, VCA, RetailNext, Euclid, ShopperTrak, and Purple installed their own sensors to geotrack real-time mobility patterns of customers in their clients' stores. Their proprietary solutions provide visitor monitoring results, such as funnel or hot-spot analysis results displayed through a dashboard. In addition, it is expected that huge amounts of shopping behaviors will be generated in cashier-less stores introduced by the enterprises such as Alibaba and Amazon.

1.2 Contribution

In this paper, we propose a systematic framework for *predicting the revisit intention* of customers using Wi-Fi signals captured by in-store sensors. Our framework includes the entire procedure for revisit prediction—from data preparation to model learning. The key challenge is how to generate the most effective set of features from the Wi-Fi signals. We systemically design the features to summarize each visit in two aspects. First, we interpret the device location at various semantic levels to understand user behaviors. Second, we utilize weak signals usually captured outside a store to expand our trajectory to the widest possible range. Using this information, we are able to track a customer's behavior outside the store even if they did not enter the store.

We also benefit from large-scale customer mobility data captured by in-store sensors. Seven flagship stores in downtown Seoul were carefully selected for data collection to cover various shop categories. The number of unique customers collected in the seven stores reaches 3.75 million. The data is very attractive because we can capture approximately 20–30%[1] of customer mobility without any intervention. Furthermore, the data collection period is 1–2 years, which is long enough to study revisit behaviors.

Figure 1 illustrates the overall procedure of our prediction framework. If a customer comes into a store, the framework detects his/her Wi-Fi signals, and through the data preprocessing described in Sect. 2.2, transforms the signals to a visit and an occurrence. From the customer's visit and previous occurrences, extensive features are derived to describe his/her motion patterns, as discussed in Sect. 4. In this regard, our framework relies upon the belief that motion patterns unconsciously reflect consumer's satisfaction with the store [13]. Finally, we can predict his/her revisit behavior, using a trained model.

Our experiments demonstrate that our revisit prediction framework achieves up to 80% accuracy of the binary revisit classification of all trajectories. Additionally, it successfully

[1] The proportion of users in their twenties who keep their Wi-Fi on is 29.2%, according to a survey by Korea Telecom (July 2015).

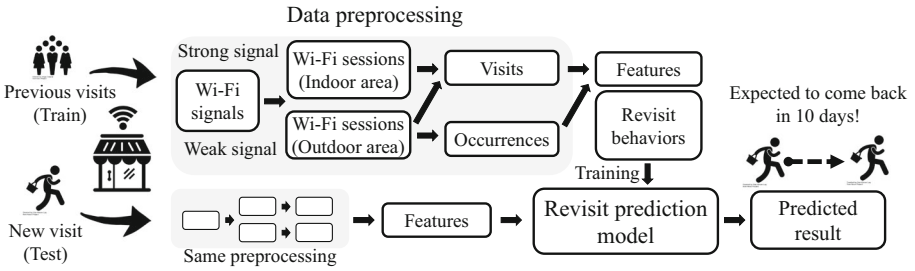


Fig. 1 Revisit prediction framework architecture

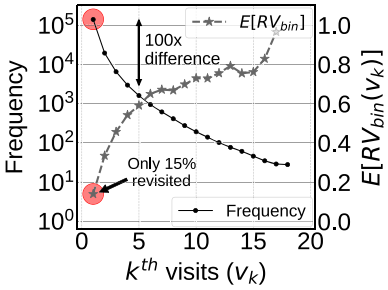


Fig. 2 Revisit statistics of store E_GN. $E[RV_{\text{bin}}(v_k)]$ denotes the average revisit rate of the group of visitors who visit k times

predicts the revisit of first-time visitors by up to 72% accuracy. In the case of actual apparel stores, it is very useful to predict the revisits of first-time customers, because they account for more than 70% of all visitors. Most importantly, our 80% accuracy is achieved by features, all derived from Wi-Fi signals with minimal external information (dates of public holidays, clearance sales). Thus, we expect that the prediction power will rise significantly by adding private data such as personal profiles and purchasing patterns.

Figure 2 illustrates the observed revisit statistics during the data collection period in store E_GN. The black line denotes the number of observations $|v_k|$ of kth visits(v_k), and the gray line denotes the average revisit rate $E[RV_{\text{bin}}(v_k)]$ of all v_k's. The fact that the $|v_5|$ is 100 times less than $|v_1|$ implies that it is very difficult to retain first-time visitors as regular customers. It also describes how valuable it is to raise the revisit rate of first-time visitors that account for 70% of all customers,[2] thereby emphasizing again the importance of our work. Along with the model accuracy, we report the predictive power of each feature group and semantic level to show whether or not the trajectory abstraction boosts the predictability. We also demonstrate the effectiveness of using customer mobility features in comparison with baseline models considering visit distribution and temporal information. We discuss how the collection period and the volume of data affect performance. Another important goal of this paper is to share the unexpected challenges faced when two groups of data show inherent differences in a statistical sense.

This paper extends our earlier work [12] presented at IEEE ICDM 2018 and also selected as one of the best papers. In particular, the evaluation of our framework has been significantly improved by addressing the comments to our earlier work. In this extended version, we empirically show that mobility features are effective even with a few records, by tracking the predictive power of our model conditioned on the number of previous visits. We also

[2] In Fig. 2, the ratio of the first-time visitors in store E_GN is over 70%. We made a few assumptions to interpret the data as it is and will discuss them in "Appendix D".

Table 1 Statistics of the datasets

Store ID	A_GN	A_MD	E_GN	E_SC	L_GA	L_MD	O_MD
Category	(Footwears)		(Fast-fashion)		(Character shop)		Cosmetics
Length (days)	222	220	300	373	990	747	698
Sensors	16	27	40	22	14	11	27
Total signals	165M	890M	940M	632M	1.94B	2.82B	6.50B
Total sessions	19.9M	33.9M	81.4M	34.1M	40.3M	74.3M	90.7M
Indoor sessions ≥ 5 s	637K	3.25M	1.35M	1.92M	5.46M	11.1M	15.6M
Visits ≥ 60 s	113K	328K	183K	270K	1.06M	1.72M	2.01M
Unique visitors ≥ 60 s	101K	232K	147K	187K	846K	1.17M	1.07M
Avg. revisit rate	11.7%	32.0%	21.2%	36.6%	21.2%	33.0%	48.7%

test various machine learning techniques and their stacked ensemble model, in addition to XGBoost [4] used in our earlier work.

The remainder of this paper is organized as follows. In Sect. 2, we describe the datasets used in this paper. After introducing the main concepts and formalizing the problem in Sect. 3, we describe the characteristics of the features in Sect. 4. In Sect. 5, we explain the experiment settings and present overall prediction results. Also, we discuss the lessons and challenges obtained through the experiments. After reviewing related work in Sect. 6, we conclude this study in Sect. 7.

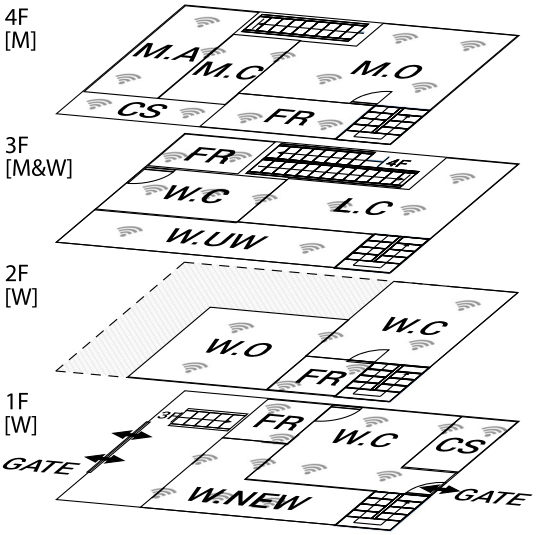
2 Data description

In this section, we introduce our customer mobility data captured from off-line stores. The number of customers in our data is very high, and the collection period is long enough to obtain reliable results. Throughout this section, we share some statistics of our datasets and introduce necessary preprocessing to find meaningful semantics from the raw Wi-Fi signals.

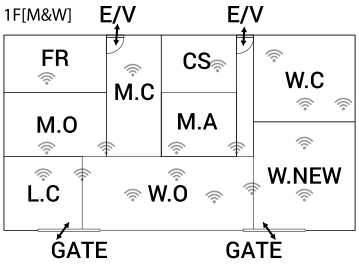
2.1 Data collection stores

We collected data from seven flagship stores located in the center of Seoul. Each of these stores is one of the largest stores of each brand, consisting of several floors. These stores are known to be the busiest stores in Korea. Because of their location and size, these stores have up to 10,000 daily visitors. For example, our target store E_GN is a four-story building located on the side of a Gangnam boulevard where two million people walk by each month. Store E_SC is located on the ground floor of a major department store in the downtown Sinchon area, which is also connected to one of the busiest subway stations used by college students. Table 1 presents the statistics of the seven datasets, and Fig. 3 illustrates the location of sensors and categories of two stores E_GN and E_SC.[3]

[3] Owing to a nondisclosure agreement, additional store information cannot be disclosed. We encourage readers to think that dozens of sensors cover the other stores in a similar manner.



(a) Floor plan of store E_GN. Store E_GN has five semantic levels: sensor level, category level, floor level, gender level, and in/out level.



(b) Floor plan of store E_SC.

Area	Description
M.O	Men's office wear
M.C	Men's casual wear
M.A	Men's accessories
W.O	Women's office wear
W.C	Women's casual wear
W.UW	Women's underwear
W.NEW	Women's new arrival
L.C	Limited collection
FR	Fitting room
CS	Checkout counter

(c) Area explanation.

Fig. 3 Location of sensors and categories of two stores E_GN and E_SC. Wi-Fi icons indicate the location of the sensors, and the category names for each section are described in (**c**)

2.2 Preprocessing to generate trajectories

2.2.1 Signal-to-session conversion

To collect Wi-Fi signals, we utilized ZOYI Square sensors developed by WalkInsights.[4] The installed sensors enable us to collect Wi-Fi signals from any device that turns on its Wi-Fi. A single Wi-Fi signal includes an anonymized device ID, sensor ID, timestamp, and its received signal strength indicator (RSSI) level, which is a measurement of the power present in a received radio signal. Signals are collected continuously from each device at fairly short intervals, which are less than 1 s. To understand customer mobility, we carry out a conversion process to remove redundant signals and combine them into Wi-Fi session logs. Each session includes a device ID, area ID, and dwell time, and it becomes an element of a semantic trajectory. Predefined RSSI thresholds are utilized for signal-to-session conversion. These values guarantee that the device is in the vicinity of a sensor. The logic of this conversion is simple. For instance, a new session is created when a sufficiently strong RSSI is received for the first time. The session continues if the sensor receives consecutive strong signals, and it ends if the sensor no longer receives strong signals. The session also ends if another sensor receives a strong RSSI from that device.

[4] https://walkinsights.com/sensors.

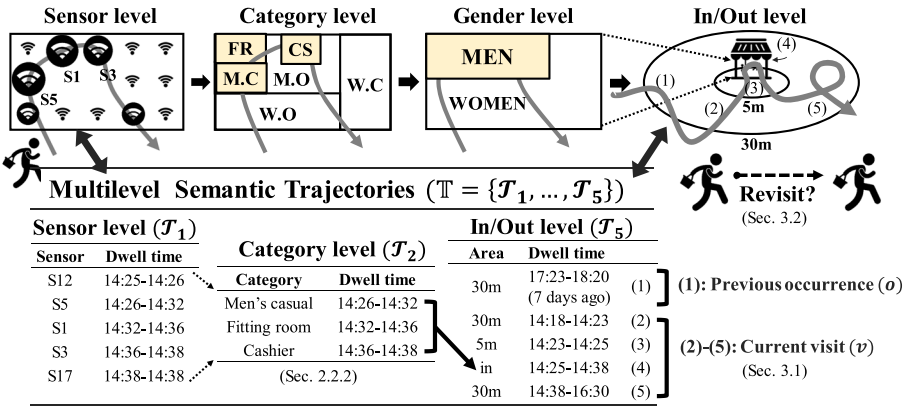


Fig. 4 Generation of multilevel trajectories to predict a customer's revisit intention: Using noninvasive monitoring, customer Wi-Fi signals are collected. These are then transformed into a sensor-based trajectory, and further summarized into categories, floors, genders, and surrounding areas. We extracted features from these multilevel trajectories to determine the characteristics related to customer behavior

2.2.2 Location semantics

It is also possible to detect the semantic location of a customer by taking advantage of the semantic coherency of contiguous sensors. For example, we can identify if the customer is looking at daily cosmetics or she is in a fitting room. Additionally, we can describe a customer's location to floor-level or gender-level semantic areas. Moreover, we generate in-/out-level areas by examining whether the customer is inside the store, nearby the store (up to 5 m), or far away from the store (up to 30 m). This becomes possible by controlling multiple RSSI thresholds to activate detection with weaker signals. Therefore, an entity of Wi-Fi session data encompasses a customer's dwell time not only in the area corresponding to sensors but also in the wider semantic areas. By integrating the Wi-Fi sessions with different semantics, we construct a multilevel semantic trajectory to describe each visit as illustrated in Fig. 4.

3 Problem definition

In this section, we formally define the main concepts introduced in our paper. First, we define a multilevel semantic trajectory ($\mathbb{T}$) that expresses a customer's motion pattern, and define visit (v) and occurrence (o) using $\mathbb{T}$. Next, we define the revisit interval (RV_{days}) and the revisit intention (RV_{bin}), which are the labels in our prediction model. Finally, we introduce the revisit prediction problem.

3.1 Key terms and concepts

Definition 1 A *semantic trajectory* $\mathcal{T}$ is a structured trajectory of size n ($n \geq 1$) in which the spatial data (the coordinates) are replaced by semantic areas [43], that is, $\mathcal{T} = \{s_1, \ldots, s_n\}$, where each element (= a session) is defined by $s_i = (sp_i, t_{in}^{(sp_i)}, t_{out}^{(sp_i)})$. Here, sp_i represents the semantic area, $t_{in}^{(sp_i)}$ is a timestamp for entering sp_i, and $t_{out}^{(sp_i)}$ is a timestamp for leaving sp_i. □

If a session length $t_{out}^{(sp_i)} - t_{in}^{(sp_i)}$ is shorter than 5 s considering walking speed and the distance between adjacent sensors, a customer is likely to pass that area without consideration, and thus, we delete the element from the trajectory.

Definition 2 A *multilevel semantic trajectory* $\mathbb{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_l\}$ is a set of semantic trajectories with l ($l \geq 1$) different semantic levels. Each semantic trajectory $\mathcal{T}_i$ represents a customer's trajectory using semantic areas of level i. □

For our indoor environment, we utilized semantic levels inside the store, except for the highest level l indicating the in/out level. The total dwell time of $\mathcal{T}_l$ is always longer than $\mathcal{T}_1, \ldots, \mathcal{T}_{l-1}$, because the in/out mobility utilizes weak signals that can be captured for a longer period than the strong signals used for indoor behavior.

Definition 3 A *visit* v is a unit action of entering the store. $v_k(c, [t_s, t_e], \mathbb{T})$ is a kth visit by customer c, who is sensed from t_s to t_e, of which the motion pattern is described with a multilevel semantic trajectory $\mathbb{T}$. □

We consider only the visits that are long enough to represent meaningful behavior. For the sensor-level trajectory $\mathcal{T}_1$, the total dwell time $t_e - t_s$ should be greater than 1 min, because it takes less than 1 min to go through the store. The data preprocessing thresholds are empirically configured depending on the size of a store and the number of sensors.

Definition 4 An *occurrence* o is a special case of a visit that represents a unit action of lingering around the store without entrance. $o_k(c, [t_s, t_e], \mathbb{T})$ is a kth occurrence by customer c, who is sensed from t_s to t_e, of which the mobility is only captured in the outdoor area with $\mathbb{T} = \{\emptyset, \ldots, \emptyset, \mathcal{T}_l\}$. □

Although we did not have any personal information such as the customer's residence, we could measure his/her accessibility to the store through the occurrence. For each visit, we use a set of previous occurrences as a reference to generate store accessibility features [SA], which will be explained in Sect. 4.1.9.

3.2 Prediction objectives

If a customer revisits the store after d days, the previous visit v of the customer has a d-day *revisit interval*, denoted by $RV_{\text{days}}(v) = d$, and a positive *revisit intention*, denoted by $RV_{\text{bin}}(v) = 1$, as in Definition 5.

Definition 5 If two consecutive visits of customer c_i, $v_k = v_k(c_i, [t_{k,s}, t_{k,e}], \mathbb{T}_k)$ and $v_{k+1} = v_{k+1}(c_i, [t_{k+1,s}, t_{k+1,e}], \mathbb{T}_{k+1})$, meet the condition $t_{k,e} < t_{k+1,s}$, the *revisit interval* $RV_{\text{days}}(v_k)$ and *revisit intention* $RV_{\text{bin}}(v_k)$ of the former visit v_k are $RV_{\text{days}}(v_k) =$ #days of$t_{k+1,s} - t_{k,e}$ and $RV_{\text{bin}}(v_k) = 1$. If a visit v_k does not have any following revisit, then $RV_{\text{days}}(v_k) = \infty$ and $RV_{\text{bin}}(v_k) = 0$. □

3.3 Predictive analytics

Our problem is now formally defined as follows:

Customer Revisit Prediction: Given a set of visits $V_{\text{train}} = \{v_1, \ldots, v_n\}$ with *known* revisit intentions $RV_{\text{bin}}(v_i)$ and revisit intervals $RV_{\text{days}}(v_i)$ ($v_i \in V_{\text{train}}$), build a classifier C that predicts *unknown* revisit intention $RV_{\text{bin}}(v_{\text{new}})$ and revisit interval $RV_{\text{days}}(v_{\text{new}})$ for a new visit v_{new}.

4 Feature engineering

To have a multiperspective view of customer movements, we construct each visit as a five-level semantic trajectory, $\mathbb{T} = \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4, \mathcal{T}_5\}$, where the levels correspond to *sensor*, *category*, *floor*, *gender*, and *in/out*, respectively. We expect the pattern captured using multiple levels can be helpful in predicting customer revisits. Thus, some features were created for each semantic level.

Table 2 gives a summary of the *representative* features in our framework. The first column describes the ten different feature groups categorized by their characteristics. The first seven feature groups are generated solely from the *customer mobility* itself. The last three feature groups: Upcoming Events [UE], Store Accessibility [SA], and Group Movement [GM] are generated using certain references: [UE] uses sales and holiday information for the near future, [SA] uses the *occurrences* of the customer before making this visit, and [GM] considers other visits at the same time.

For seven stores, the total number of generated features varies from 220 to 866 depending on the number of areas and the number of semantic levels used. $\mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4$-level features are generated only for two stores: Store ID of E_GN and E_SC, where we continuously

Table 2 Description of the representative features according to the data sources and feature groups

Feature groups	Twenty representative features (Among 866 features of store E_GN)	Semantic level of features					
		$\mathcal{T}_1$	$\mathcal{T}_2$	$\mathcal{T}_3$	$\mathcal{T}_4$	$\mathcal{T}_5$	$\emptyset$
Overall statistics	f_1 = Total dwell time					✓	
	f_2 = Trajectory length	✓	✓	✓	✓		
	f_3 = Skewness of dwell time of each area	✓	✓		✓		
Travel dist, speed, acceleration	f_4 = Total distance traveled inside the store		✓				
	f_5 = Speed based on transition time	✓	✓	✓	✓		
	f_6 = First-k HWT coeff. of acceleration	✓	✓	✓	✓		
Area preference	f_7 = Coherency of dwell time for each level		✓	✓	✓		
	f_8 = Top-k-area dwell time	✓	✓	✓	✓		
Entrance and exit pattern	f_9 = Exit gate	✓					
	f_{10} = Number of prev. reentry on that day					✓	
Heuristics	f_{11} = Wears clothes but does not buy		✓				
Statistics of each area	f_{12} = Number of time sensed in the area	✓	✓	✓	✓		
	f_{13} = Stdev of dwell time for the area	✓	✓	✓	✓		
Time of visit	f_{14} = Day of the week						✓
Upcoming events	f_{15} = Remaining day until the next sale						✓
	f_{16} = Number of holidays for next 30 days						✓
Store accessibility	f_{17} = Number of days since the last access					✓	
	f_{18} = Average interarrival time					✓	
Group movement	f_{19} = Presence of companions					✓	
	f_{20} = Number of companions					✓	

The ✓ indicates the best semantic level to describe each feature. For features with multiple ✓, the values of the features at each level are different, thus having different meanings. Features f_{14}, f_{15} and f_{16} have no corresponding semantic level so we denote their semantic level as $\emptyset$

tracked their floor plans during data collection periods. Among all features, we introduce 20 *representative* features to best describe the characteristic of each feature group. On the right side of the table, the corresponding semantic level for each feature is marked.

Figures 5 and 6 display meaningful relationships between the feature values of f_1, f_7, f_9, f_{15}, and f_{17} with the average revisit intention rate $E[RV_{\text{bin}}(v)]$. By dividing total visits into 20 equal bins according to feature values, we can identify the association between feature values and revisit rates without being affected by outliers.

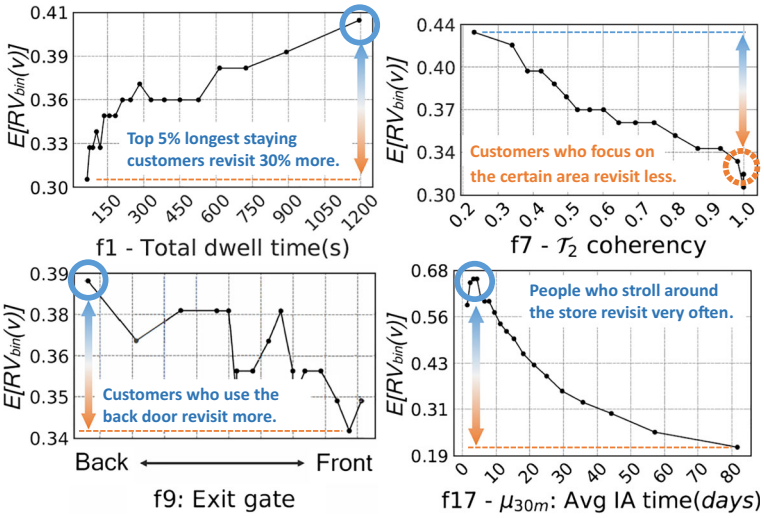
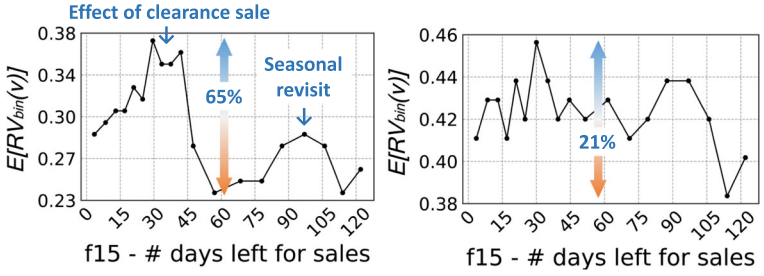


Fig. 5 The relationship between the selected features and RV_{bin} in store E_SC ($E[RV_{\text{bin}}(v)\ (v \in V_{\text{all}})] = 0.3616$). Each marker point represents the average revisit intention $E[RV_{\text{bin}}(v)]\ (v \in V_b)$ of the set V_b of visits obtained by equal-frequency-binning the entire data according to feature values. Indoor moving pattern features f_1, f_7, and f_9 shows at most 40% deviation of $E[RV_{\text{bin}}(v)]$ according to the feature value. The store accessibility feature f_{17} shows 325% deviation, which is the highest among the selected features. For f_9, the group of customers who are most likely to use the back door are located on the left side of the x-axis



(a) First-time visitors (v_1): Prone to events. **(b)** All visitors: Indifferent to events.

Fig. 6 Key factors of v_1's revisit: discount and seasonality. *Discount-sensitive*: A set V_b of customers who visited between 30 and 45 days before a clearance sale showed a high $E[RV_{\text{bin}}(v)]\ (v \in V_b)$ compared to other customers; this difference was more apparent in first-time visitors than all visitors. *Seasonal-sensitive*: Another peak of $E[RV_{\text{bin}}(v)]$ appeared on the set of customers who made a visit between 90 and 105 days before the sale. It described the seasonal revisit, and it was also more noticeable to first-time visitors than all visitors

4.1 Feature descriptions

In this section, we introduce the detail of each feature group used in our model. With the background information for designing each feature, we show some correlations between features and customer revisits.

4.1.1 Overall statistics [OS]

[OS] features represent the high-level view of a customer's indoor movement patterns, and therefore, the predictive power is relatively strong. By considering the trajectory as a whole, we can extract features such as total dwell time (f_1), trajectory length (f_2), and average frequency of each area. We also apply skewness (f_3) or kurtosis to measure the asymmetric or fat-tail behavior of the dwell-time distribution of each area.

4.1.2 Travel distance, speed, and acceleration [TS]

[TS] features are in-depth information that needs to be explored [25]. To approximate the physical distance (f_4) traveled by the customer, we created a network based on the physical connectivity between areas. We used the transition time to obtain the shopping speed (f_5), and we modeled the acceleration from the speed variation between consecutive areas. A time series analysis using the Haar Wavelet Transform (HWT) [34] was performed, as well as statistical analysis, to determine how the customer's interests changed with time. We included the first-16 HWT coefficients (f_6) in our feature set.

4.1.3 Area preference [AP]

With [AP] features, it is possible to identify the difference between a customer viewing a specific area with high concentration and a person shopping lightly throughout the store. The area name and dwell time (f_8), and its proportion over the total dwell time of the top-3 areas at each level are included in the basic features. The coherency of each level (f_7) determines the consistency of the customer's behavior. The definition of the coherency metric is the proportion of time spent in the longest staying area. This metric is effective to capture regular customers who know the store's layout and go directly to the desired area.

4.1.4 Entrance and exit pattern [EE]

Interestingly, customers leaving through the back door (f_9) revisited 13.6% more than customers leaving through the front door, according to our data. Therefore, we estimated the customers' entrance and exit patterns from the sensors nearby the front and back doors. We expected that customers familiar with the store might have used a more convenient door.

4.1.5 Heuristics [HR]

To fully exploit the relation between customer trajectories and revisits, we interviewed the managers and part-timers of the stores to get intuitions on what kinds of patterns are likely to appear from the customers who are willing to revisit. In general, the interviewees agreed that staying in certain areas, trying an item, and purchasing or postponing the item can reflect

customers' interest and purchase pattern that lead to revisits. These steps of actions, in fact, correspond to online shopping activities—i.e., browse, add to cart, checkout, and then revisit or churn [18]. As we do not know whether a customer actually tried an item in the fitting room or purchased it, we inferred those actions by tracking the dwell time in the fitting room and the checkout counter. Here are two representative heuristics anticipating the revisit of customers for future purchase.

- If a customer wears clothes in the fitting room without purchase (≤ 1 min in the checkout counter): $f_{11} = 1$, for all other cases: $f_{11} = 0$.
- If a customer stays in the store much longer (= 10 min) than average visitors, without purchase: $f = 1$, if not: $f = 0$.

The reasons for these associations are as follows. If the customer tries an item or stays in the store for a long time, he/she is prone to purchase the item. However, the fact that the customer does not purchase the item right away implies that there is a possibility of purchasing that item at the next visit.

4.1.6 Statistics of each area [ST]

If a certain semantic area is highly relevant to revisit, the statistics from that area have higher predictability. For all semantic areas, we created six features including the number of times it was sensed (f_{12}), the percentage of the total time spent in the area (that is used for developing the coherency feature), and the standard deviation of the times sensed in the area (f_{13}). As explained before, the difference in the total number of features is mainly due to the difference in the number of areas that each store has.

4.1.7 Time of visit [TV]

The temporal features include the time of visit such as the hour of the day and day of the week (f_{14}) as basic features. The values of the features described above are ordinal and thus were transformed into multiple binaries by one-hot encoding. The value of a temporal feature is determined by the entrance time.

4.1.8 Upcoming events [UE]

Customers are more likely to visit a store in the period of a clearance sale. However, they are less likely to visit the fashion district in the holiday seasons(e.g., Spring Festivals, Thanksgiving week) since they are out of the city center. For example, customers who visited one month before the clearance sale have higher chance to revisit since they would like to get a discount during the upcoming sales. By combining simple extrinsic information, the temporal features, particularly [UE], become the second strongest predictive feature groups. It contains six features, including a number of days left for the next clearance sale (f_{15}) and a number of holidays for next 30 days (f_{16}), as numeric features.

4.1.9 Store accessibility [SA]

When installing sensors inside the store, could you imagine that the weak noise collected outside the store would provide the most important clue to predict revisit? Surprisingly, the revisit predictability increased dramatically when we included [SA] features using weak

signals, which could have been overlooked as mere noises. The following settings are expected to be applicable to many studies when conducting research using in-store signals that do not contain customer address information.

The features are designed to capture various aspects from interarrival times. We utilized two additional outdoor areas nearby the store—5 m and 30 m zone—to detect the customer occurrences. Considering a customer's arrival process to 5 m zone, let us denote the time of the first occurrence by T_1. For $k > 1$, let T_k denote the elapsed time between $k-1$th and the kth event. We call the sequence $\{T_k, k = 1, 2, ...,\}$ as the *sequence of interarrival times*. Considering the target visit as nth event of the arrival process, we use the following features:

- $n-1$: Number of occurrences before the visit;
- T_n: Number of days from the last occurrence (f_{17});
- $\mathbb{1}_{n>1}$: Existence of having any occurrence before the visit;
- $\mu = \sum_{k=2}^{n} T_k/(n-1)$: Average interarrival time (f_{18});
- $\sigma = \sqrt{\sum_{k=2}^{n}(T_k - \mu)^2/(n-1)}$: Standard deviation of interarrival times;

In addition to these five features from T_k, we added the average sensed time for previous occurrences.

4.1.10 Group movement [GM]

Unlike previous features, [GM] features were extracted by considering multiple trajectories. This is a representative feature that can only be captured by analyzing surrounding trajectories that happened simultaneously with the main trajectory. In our feature extraction framework, we considered the presence of companions (f_{19}) and the number of companions (f_{20}). One of the biggest characteristics of judging whether or not to be a companion is to enter the store at the same time. Based on the information obtained through the field study, we considered that two visitors are in a group when their entrance time and exit time are both within 30 s. Additional information related to this feature can be found in Sect. 5.3.2 and "Appendix D".

4.2 Unused features

Some potentially useful features were not included in our final model because their effect on the accuracy was marginal. However, we would like to mention them since they could be useful in other types of predictive analytics [14,18].

4.2.1 Sequential patterns

Sequential patterns [7,14] were not effective for the revisit prediction task on our datasets, so we omitted them from the final framework. To briefly describe our approach, we retrieved top-k discriminative sequential patterns by the information gain and generated k features. Each feature $f_i(v)$ denotes the number of times a trajectory of visit v contains ith patterns. We considered diverse levels of sequential patterns, as in Table 3, but the result was not satisfactory. Despite that it was expensive to generate the features, their information gains were typically low.

Table 3 Types of sequential patterns

Pattern type	Description
$A \rightarrow B \rightarrow C$	A sequential pattern having an order, where the following element appears immediately after the previous element
$A \overset{*}{\rightarrow} B \overset{*}{\rightarrow} C$	A partial sequential pattern [14], an arrow $A \overset{*}{\rightarrow} B$ denotes that there might exist additional elements between A and B
$A_{\text{short}} \overset{*}{\rightarrow} B_{\text{long}} \overset{*}{\rightarrow} C_{\text{short}}$	A partial sequential pattern which has a time constraint for the dwell time of each element

4.2.2 Past indoor information

We excluded the features that average up the customer's previous indoor mobility statistics, as well as those that represent the amount of changes from past statistics [18]. By nature, the number of features becomes doubled per revisit by considering that information. However, they were not a strong indicator of revisits unlike [SA] and thus were removed.

4.2.3 Features that may interfere with fair evaluation

Since most customers have a small number of visits, we developed a general model that considers the mobility of the entire set of customers. According to this principle, we considered each visit separately, by removing customer identifiers. In this way, we can also ensure that our model is robust to general cross-validation settings. We excluded the visit date to avoid a biased evaluation that favors the customers who visited in an earlier period. We also ignored the explicit visit count information.

5 Evaluation results

In our experiments, we verify that our feature set designed from customer mobility patterns is effective in predicting customer revisit, especially for newcomers. In addition, we verify the performance of individual feature groups and semantic levels. Throughout the discussion section, we provide more detailed analyses regarding the revisit prediction. The key contents include the demonstration of the performance change over the length of data collection period and model robustness on missing customers. We conclude this section by sharing the difficulties of securing accuracy in line with the gap between the predictive power and the statistical significance of each feature.

5.1 Settings

5.1.1 Prediction tasks

We designed prediction tasks to explore customers' revisit behaviors. The first task is a binary classification task to predict customers' revisit intention RV_{bin}. The second task is a regression task to predict the revisit interval RV_{days} between two consecutive visits. For each task, we conducted experiments on two different data subsets. First, we see the performance of our model on the entire customer dataset. Second, we used a dataset consisting of only

the first-time visitors to show that our prediction framework is effective in determining the willingness of first-time visitors to revisit.

5.1.2 Scoring metrics

We used two scoring metrics: *accuracy* and *root-mean-squared error (RMSE)* for the classification and regression tasks, respectively.

- The *accuracy* is the ratio of the number of correct predictions to that of all predictions. We used it for the classification task because it is considered to be the most intuitive metric for store managers and practitioners. To fairly compare the model performance in seven imbalanced datasets with different revisit rates, we downsampled non-revisited customers for each dataset. In this way, we designed the task as a binary classification on balanced classes having 50% as a random baseline. To mitigate the risk of the sampling bias, we prepared *ten* different downsampled train/test sets with random seeds. The averages of ten executions were reported in the paper.
- The *RMSE* is measured between the actual interval and the predicted interval. To make the RMSE values of seven stores with different data collection periods comparable, a RMSE value was normalized by the length T of the data collection period. Because we cannot calculate the revisit interval for the last visit, we excluded the customers' last visits for the regression task.

5.1.3 Data preparation

The training and testing data were prepared with three settings:

– S1: Fivefold cross-validation by dividing *customers*, where each customer data can be included only in a single fold.
– S2: Fivefold cross-validation by dividing *visits*,[5] where each visit is handled independently.
– S3: First 50% visits as the training data, and other 50% as the testing data.

The accuracy difference between S1 and S2 was insignificant to the fourth decimal place. In S3, there was an accuracy loss of about 2.5% on average compared to S1 and S2, due to floor plan changes of the stores and inaccurate labels caused by truncation in time (Sect. 5.3.1). Because of the page limit, we report the main results using the configuration S1.

5.1.4 Classifier

All results described in this section were obtained using Python API of the XGBoost [4] library that optimized the gradient boosting tree [5] framework. XGBoost gave the *best* performance among logistic regression, decision trees, random forests, AdaBoost, and gradient boosting trees implemented in the Python Scikit-learn [26] library. For this manuscript, we also compared the performance with up-to-date boosting classifiers such as LightGBM [11] and CatBoost [28], and LightGBM was 5.7 times faster than CatBoost with similar performance. To further improve performance, we also tried a two-level stacking by incorporating the top-3 individual models, but the performance improvement was marginal. We add the results of the non-best models in "Appendices A and B" to avoid breaking the original flow.

[5] As a result of Sect. 4.2.3, our model is considered to be safe to perform cross-validation.

Table 4 Performance of classification and regression tasks

ID	Length	# features	Cust-type	# data (# revisitors)	Accuracy	RMSE
A_GN	222	256	First	99,497 (9514)	0.6336	0.2132
			All	112,672 (13,222)	0.6689	0.2000
A_MD	220	328	First	223,103 (47,917)	0.6930	0.1865
			All	327,940 (104,913)	0.7412	0.1622
E_GN	300	866	First	144,610 (21,701)	0.6663	0.1862
			All	183,246 (38,817)	0.7050	0.1627
E_SC	373	663	First	172,551 (41,036)	0.6818	0.1824
			All	270,366 (98,818)	0.7288	0.1475
L_GA	990	244	First	838,241 (107,925)	0.7173	0.1403
			All	1,062,226 (225,409)	0.7789	0.1244
L_MD	747	220	First	1,154,486 (197,476)	0.6799	0.1416
			All	1,718,359 (566,701)	0.7991	0.1146
O_MD	698	316	First	1,033,253 (294,949)	0.6645	0.1311
			All	2,008,384 (978,699)	0.7599	0.1028

We used *all* features for training and testing the model, since using all features gives the best performance and the boosting tree classifier is robust to potential correlations between features. The elapsed time for each fold with 200,000 visits and 660 features took no longer than 1 min in a single machine (Intel i7-6700 with 16 GB RAM, without GPU).

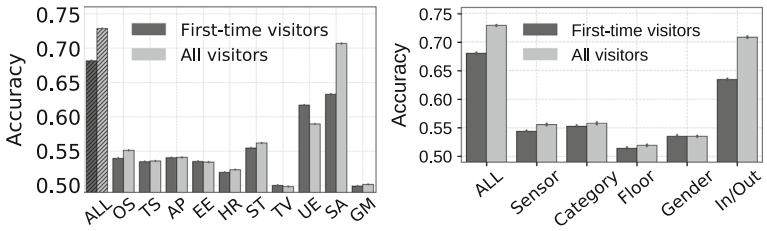
5.2 Results

5.2.1 Overall results

Table 4 shows the overall accuracy and RMSE. First, the prediction accuracy for first-time visitors is 67% averaged over seven stores. By only using mobility data captured by in-store sensors, *two* out of *three* customer's revisit is predictable without having *any* historical data in the store. Second, the average prediction accuracy increases to 74% by considering all customers. Third, the stores with a long data collection period and abundant user logs generally show high performance, while this trend might not happen depending on the characteristics of the stores.

5.2.2 Predictive power of feature groups

Figure 7a investigates the predictive power of each group of features in store E_SC. Each bar corresponds to the prediction results using the features of only a specific group. The labels of the x-axis are the abbreviations of the feature groups categorized in Table 2. For the convenience of comparison, the leftmost bar on the figure represents the results when all features in Table 4 are used. It was observed that the *store accessibility* [SA] features have the strongest predictive power, especially for the prediction with all visitors, followed by the *upcoming event* [UE] features for the first-time visitors.



(a) On feature groups. **(b)** On semantic levels.

Fig. 7 Performance comparison on feature groups and semantic levels (store E_SC)

5.2.3 Predictive power of semantic levels

As opposed to our intuition, a performance of semantic levels inside the store did not boost the performance that much. As in Fig. 7b, the performance of the features generated from the category level ($\mathcal{T}_2$) only beats the features from the sensor level ($\mathcal{T}_1$). Besides, the semantic trajectories generated from the floor-level ($\mathcal{T}_3$) and the gender level ($\mathcal{T}_4$) were not effective to predict customer revisit in the store E_SC. We can conclude that finding effective trajectory abstraction is difficult even if the hierarchical information is provided.

5.2.4 Performance improvement by analyzing trajectories

To measure the performance improvement using our features, we developed two different baselines for comparison. The first baseline is a theoretical lower bound of the prediction accuracy obtained from revisit statistics, shown in Fig. 2. Since we fully ignored any other information here, the prediction accuracy with this limited information must be lower than that of using full trajectories. The procedure of deriving lower bounds is given in "Appendix C".

The second baseline is a model to which the visit date is added. Since our task utilizes finite time series datasets with time-dependent objectives, the earlier collected logs tend to have a relatively high revisit rate. Therefore, by including a visit date as an additional feature, the baseline accuracy improves naturally. If there existed infinite data, the performance increase by this factor would disappear. The benefit of using customer mobility can be considered as the gap between our final model and the second baseline.

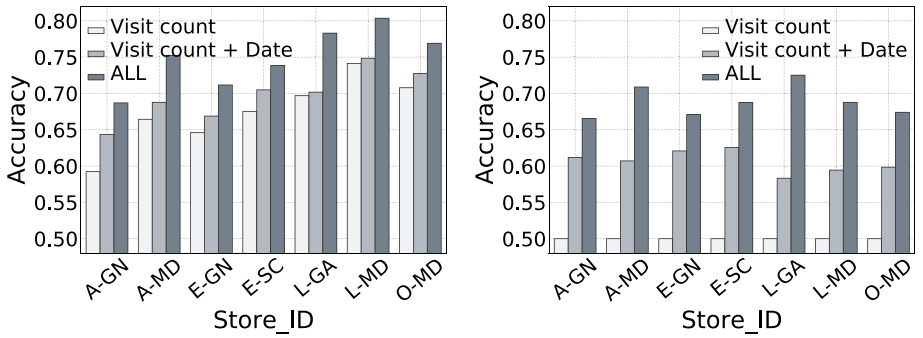
Figure 8 reports the accuracy of our model[6] against two baselines. We note that our final model is more effective than the second baseline by 4.7–11.6% in terms of accuracy. Among the first-time visitors, the effectiveness of trajectory analysis increases, showing a performance improvement of 8.0–24.3%.

5.2.5 Prediction accuracy according to the number of visits

For further analysis, we measured the prediction accuracy for each customer group determined by their number of visits. For this experiment, we used the model trained on all customers.

Customers who visit more than a certain number of times usually have a high chance to revisit. Thus, we expect that our model can predict their revisits with high accuracy. The

[6] For this experiment, we included visit count and date to our feature set, so the overall accuracy is slightly higher than the values reported from Table 4.



(a) On all visitors. **(b)** On first-time visitors.

Fig. 8 Effectiveness of analyzing customer trajectories

Table 5 Prediction accuracy (%) conditionally measured on groups of customers with the same number of visits

Store ID	A_GN	A_MD	E_GN	E_SC	L_GA	L_MD	O_MD
# Visits							
v_1	0.661	0.741	0.681	0.716	0.763	0.778	0.758
v_2	0.732	0.735	0.716	0.691	0.795	0.773	0.706
v_3	0.824	0.786	0.791	0.751	0.840	0.848	0.757
v_4	0.856	0.808	0.845	0.800	0.848	0.879	0.801
v_5	–	0.803	0.865	0.831	0.847	0.885	0.820
v_6	–	0.810	0.884	0.852	0.846	0.883	0.829
v_7	–	0.808	0.907	0.861	0.856	0.879	0.834
v_8	–	0.814	0.911	0.866	0.836	0.878	0.838
v_9	–	0.802	0.903	0.875	0.863	0.874	0.837
v_{10}	–	0.789	–	0.900	0.867	0.870	0.839

We only reported the result where $|v_n| \geq 50$ on the test set

results in Table 5 confirm this expectation. As customers visited more often, the prediction accuracy tended to increase in all stores. Interestingly, we found that the prediction accuracy sometimes was the lowest in the case of v_2 since those groups of customers seemed to have the most uncertain behavior on their revisits.

Table 6 shows the improvement of our model compared with the two baselines in Sect. 5.2.4 for each customer group. It indicates that our model is more effective than the baselines by over 10%, especially on v_1 and v_2. Thus, our feature set is shown to be effective in predicting customers' revisits even when they are newcomers.

5.3 Discussions

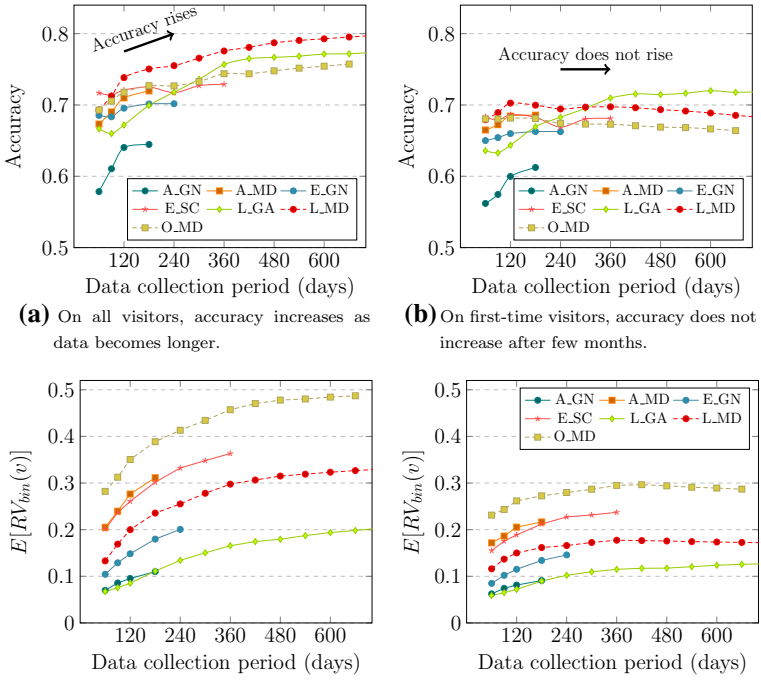
5.3.1 Importance of data collection period

We are wondering how much the model's performance varies depending on the amount of data. Figure 9a shows that the overall prediction accuracy increases as the length of the data

Table 6 Improvement of our model against the two baselines

Store ID	A_GN	A_MD	E_GN	E_SC	L_GA	L_MD	O_MD
# Visits							
v_1	18.6/7.7	17.1/14.7	12.9/9.1	10.4/7.1	18.2/17.6	10.5/10.4	7.6/7.4
v_2	4.9/1.2	13.5/5.0	7.5/2.0	15.1/3.1	4.6/3.0	18.4/12.5	29.7/13.0
v_3	1.7/0.4	4.2/1.3	3.0/0.4	7.5/1.3	0.9/0.3	2.5/1.2	8.0/3.5
v_4	1.3/0.3	3.5/0.5	2.8/1.1	5.5/0.7	1.0/0.1	0.9/0.2	3.7/1.0
v_5	–	3.2/0.3	1.3/−0.4	3.8/0.8	1.1/0.1	0.7/0.0	2.7/0.5
v_6	–	2.3/0.2	1.6/0.8	3.3/0.4	1.3/0.2	0.8/0.0	2.4/0.2
v_7	–	3.8/0.8	1.8/−0.1	2.7/1.0	1.3/0.3	0.8/0.0	2.2/0.2
v_8	–	4.0/-0.2	1.7/0.5	2.4/0.0	1.4/0.2	1.2/0.0	2.2/0.2
v_9	–	3.6/0.0	1.5/0.9	3.2/0.6	1.8/0.6	1.4/0.2	2.0/0.0
v_{10}	–	3.1/0.0	–	2.1/0.2	0.9/0.2	1.6/−0.1	2.5/0.2

The first number represents the improvement of prediction accuracy over the first baseline, and the second number represents the improvement over the second baseline



(a) On all visitors, accuracy increases as data becomes longer.

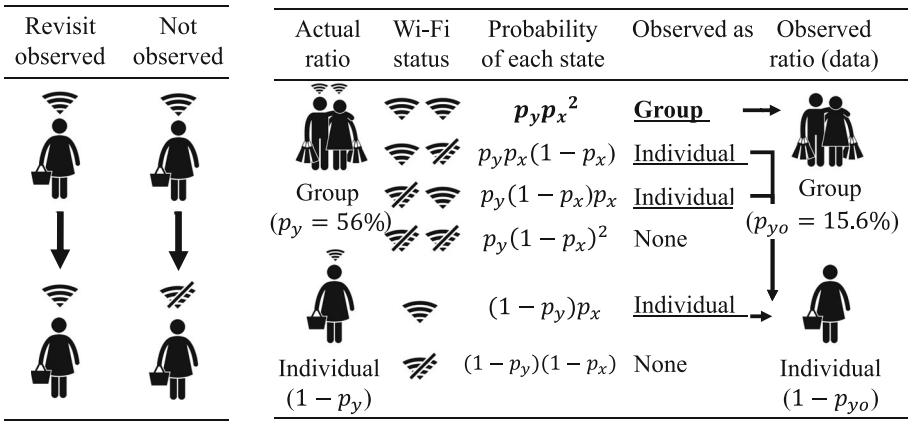
(b) On first-time visitors, accuracy does not increase after few months.

(c) On all visitors, average revisit rates keep increasing for all cases.

(d) On first-time visitors, average revisit rates decrease in some cases.

Fig. 9 Impact of the data collection period

collection period increases. The performance rapidly increases over the first few months, and then the increment is getting smaller. The main reason for the poor performance in the first few months is the lack of information on revisiting customers. Therefore, the labels in the training data could be inaccurate if we collected the information for an *insufficient period*. To



(a) Revisit. **(b)** Group movement ($p_x = 0.392, p_y = 0.56$).

Fig. 10 Missing behaviors in noninvasively collected data. **a** Customers' revisits were untraceable if they did not have Wi-Fi turned on. **b** The actual group movement ratio observed from the store was 56% instead of 15.6% observed in the data. Researchers must not interpret the data as it is, when explaining the real behavior

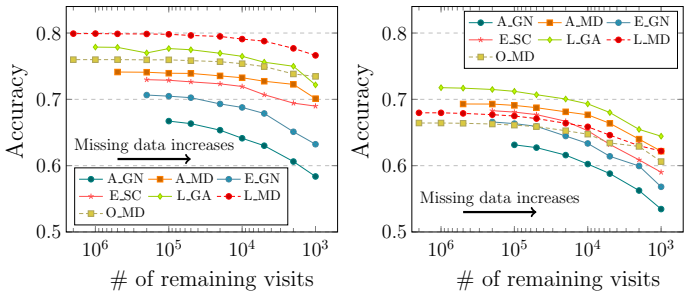
confirm our conjecture, we also examined the proportion of customers' revisit intention as the data collection progressed, as in Fig. 9c. The proportion, $E[RV_{\text{bin}}(v)]$, indeed increased as the data collection period increased. However, prediction accuracy on first-time visitors did not always increase. We notice that the average revisit rate also decreases for those cases, i.e., O_MD and L_MD, which implies that recently visited customers do not tend to revisit the store. Overall, with a longer data collection period, performance improvement occurs by having more positive cases for regular customers.

5.3.2 Real behavior and collected data—Are they same?

Noninvasively collected data is also limited, considering that not all users turn on Wi-Fi of their mobile device. Since the 4G LTE connection is very fast and ubiquitous in Korea, the proportion of 'always-on' users is just 30% [24]. This limitation implies that the datasets were missing some customer behaviors in the real world. Figure 10a shows untraceable revisits due to the conditional Wi-Fi usage of the customer, and Fig. 10b shows a gap between the actual/observed proportion of group movements caused by low Wi-Fi usage. The reason for the difference is that both companions must use Wi-Fi to verify the accompanying records on the data. p_x denotes the probability of customers who turn on Wi-Fi on-site (including 'conditionally-on' users), and p_y denotes the actual proportion of customers in a group of size two. Here we ignore groups more than two customers, which are not that common. Then the proportion p_{yo} of group customers observed in the data can be represented as Eq. (1).

$$\begin{aligned} p_{yo} &= \frac{\text{Observed(Group)}}{\text{Observed(Group)} + \text{Observed(Individual)}} \\ &= \frac{p_y(p_x)^2}{p_y(p_x)^2 + 2p_y p_x(1-p_x) + (1-p_y)p_x} = \frac{p_x p_y}{1 + p_y - (p_x)^2} \end{aligned} \tag{1}$$

Therefore, readers should recognize that the observed movement ratio can be very different from the actual movement ratio. We leave additional details in "Appendix E" and briefly



(a) On all visitors. **(b)** On first-time visitors.

Fig. 11 Model robustness on missing customers

introduce how to utilize this gap to decide the 30 s threshold to determine group movements. In the future, if customers' behaviors are more traceable with additional sensing technologies, we believe that *noninvasively* collected data will better reflect actual customer behaviors.

5.3.3 Performance on incomplete data

Assuming that some of the customers' data are completely gone, is the performance of our model reliable? We confirmed that over 95% of the performance of our model is maintained with a very small fraction of the dataset (e.g., 0.5% for L_MD). For each store, we randomly removed the records of a set of customers and measured the model performance using the remaining data. Figure 11 shows the averages for 20 different executions. The accuracy loss of the model was within 3% if 10,000 visits were secured. This observation can be also interpreted as follows:

- For large-scale mobility data, a comparable prediction model can be built by using small data subsets.
- On the other hand, we can estimate the prediction performance when all customer data becomes traceable.
- High prediction accuracy of some stores may not be due to their large number of visitors.

5.3.4 Meaningful insights but low predictability

We would like to point out that securing prediction accuracy can be difficult although the differences between revisitors and non-revisitors are obvious. Some feature values significantly differ by the revisit status, each of which should be helpful to explain the difference between the two groups. But from the perspective of a prediction task, the correlation coefficient was relatively small, and the prediction accuracy using the feature was not very high.

In Table 7, the relative difference diff_1 of feature values depending on the future revisit status is noticeable (2.7–104.2%). Besides, the p value ($p < 10^{-100}$) from Mann-Whitney U test shows that the feature values of the two groups are from different distributions. From another perspective, the relative difference diff_2 in the average revisit rate between the top 5% and the bottom 5% of customers in terms of feature values also shows clear distinction by 43.5–134.7%.

Table 7 Statistics of feature values with revisit status, and their final predictability: statistics from the store O_MD

Name	Feature description	Feature value difference by revisit status			
		FV_1	FV_0	diff_1 (%)	p value
f_a	Avg interarrival time (5 m)	21.8 days	44.4 days	104.2	0****
f_b	Total dwell time	3211 s	1612 s	99.2	0****
f_c	Dwell time proportion in 3rd area	0.112	0.087	28.5	0****
f_d	Avg dwell time for each area	358 s	348 s	2.7	0****

	Revisit rate difference by feature values			r_{pb}	Accuracy
	$\max(RV_{\text{bin}})$	$\min(RV_{\text{bin}})$	diff_2 (%)		
f_a	0.841	0.358	134.7	−0.207	0.7346
f_b	0.721	0.335	115.2	0.216	0.6005
f_c	0.622	0.335	85.8	0.152	0.6035
f_d	0.588	0.410	43.5	0.007	0.5584

$FV_1 = E[FV(v)|RV_{\text{bin}}(v) = 1]$: Average feature values of revisitors, r_{pb}: Point-biserial correlation. The details of the four feature values can be found in Fig. 12

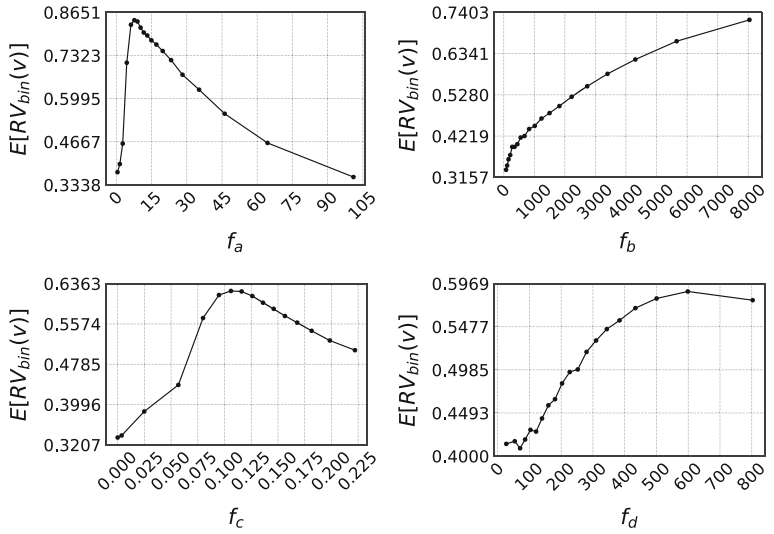


Fig. 12 Detailed relationship between four features and $E[RV_{\text{bin}}(v)]$ mentioned in Table 7

However, the correlation coefficient and the final prediction accuracy using the feature are not as impressive as diff_1 and diff_2. Practitioners should note that the behavioral difference between the two groups is obvious and the p value is high, but not in terms of the metric of correlation and prediction accuracy. Also, the feature should not be discarded because of the low correlation coefficient. If the feature has a nonlinear tendency, its predictive power can be strong. The statistics of f_b and f_c in Table 7 confirms our argument. We assert that our high-quality prediction came from a combination of various kinds of features which behave differently.

6 Related work

Predictive analytics using trajectories. Next location prediction using trajectories is one of the most actively studied topics in the computer science community. To predict the next location, frequent trajectory patterns [7,23], nonlinear time series analysis of the arrival and residence time [31], the hidden Markov model (HMM) [22], and cluster-based prediction with semantic features [44] were applied. Performances of many approaches were compared by Baumann et al. [1], and the data sparsity problem was handled by Xue et al. [39]. The results support that the prediction of the next location using partial trajectories is feasible, along with the regularity studies of human mobility [6,19,33]. Within the subject of predicting the next location, the prediction of the final destination of a taxi [2,3,20] has been also actively studied since the 2015 ECML/PKDD competition.[7] The main difference between our study and previous studies is the prediction objective. We studied the customers' revisit intentions in the off-line stores using indoor trajectories. Thus, our model focused on predicting revisits instead of locations. As far as we know, there is no study of predicting revisit intention using large-scale trajectories captured by in-store sensors.

Customer behavior in the store. Park et al. [25] examined the factors of route choice in three clothing outlets by tracking 484 customers. They considered spatial characteristics of the outlet, types of customers, and their shopping behaviors. In the grocery store, an RFID-based tracker system with shopping carts enabled Hui et al. [8] to find some interesting causality such as consumers who spent more time in the grocery store become more purposeful, or after purchasing virtue categories, the presence of other shoppers attract consumers yet reduce their tendency for purchase. Yada [40] applied a character string analysis technique, EBONSAI, originally developed in the field of molecular biology. They converted each shopping area into a character to applied their algorithm in order to discover purchasing behaviors. Hwang and Jang [9] introduced process mining techniques to understand customer pathways. The Petri-net model learned by inductive learning algorithms provides a formal representation of the shopping path of customers. With the collaboration between sensor providers and their clients, they showed that customers' behavioral patterns and sales revenue changed in accordance with process models and store layouts. This study also utilized the Kolon store dataset collected by ZOYI, a data provider of our seven stores. Although these studies did not focus on customers' revisit, they were valuable resources for us to develop the features that describe customers' motion patterns. Currently, Alibaba's Hema Xiansheng[8] and Amazon Go[9] are the most widely known future retailers, breaking the traditional retail experience. Because of the abundant in-store data from these retailers, we expect that there will be tremendous opportunities to study customer behavior patterns during their shopping time.

Indoor analysis in other places. Traditionally, the analysis of customers' indoor movement and connections to space has been conducted in the area of architecture or interior design. Especially for museums, various movement patterns were tracked manually [16,41] to rearrange the exhibits to enhance the satisfaction of visitors [10]. For example, the extent of visibility of the display was studied [17] to arrange the main display by using the behavior of passive visitors [10]. They concluded that visitors are influenced by the continuity in display within their view. With the help of noninvasive monitoring, visitor studies in the museum have come to a new phase. Yoshimura et al. [45] installed eight beacons in the Louvre Museum

[7] http://bit.ly/kaggle_taxi_interview_1st_nn.

[8] https://www.freshhema.com/.

[9] https://en.wikipedia.org/wiki/Amazon_Go.

and analyzed the most popular paths to mitigate a micro-congestion inside the museum. By tracking visitors' movements, the Guggenheim Museum[10] increased customers' engagement by making smarter curatorial decisions. Both museums and stores are the places where customers' indoor mobility data can be meaningful for the study of customer satisfaction. Thus, we expect that our framework is also applicable to the museum visitor studies.

7 Conclusions

Various retail analytics companies have set up sensors to monitor customer mobility in off-line stores. Although it was difficult to connect with other kinds of data because of legal issues, we confirmed that customer mobility indeed involves diverse meanings. Without having access to customer purchase data or customer profile, we have found that revisit intention of customers are predictable by up to 80%, using only Wi-Fi signals collected by in-store sensors. Toward this goal, we suggested guidelines for setting the collection period of indoor data for revisit prediction. We also showed our model is robust even if a majority of customer data is missing. Moreover, we demonstrated that significant observations may be in disagreement with the final predictive power. We expect that our findings will help data scientists and marketers from both retail analytics companies and their clients make important decisions. In the future, we plan to discover additional aspects of revisits from inter-store mobility with an advanced model to learn the customer revisit mechanism.

Acknowledgements This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) (No. 2017R1E1A1A01075927). We appreciate Minseok Kim for helping surveys on off-line stores and drawing floor plans. We also thank ZOYI for providing active discussion in regard to the datasets.

A. Comparison on various classifiers

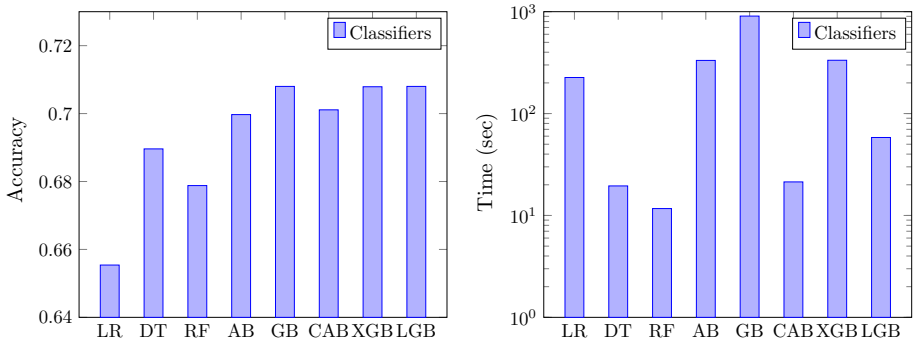
We compared the performances of *eight* classifiers. We used default parameter settings for classifiers and some tuned parameters are listed below.

- Classifiers provided by Scikit-learn [26].[11] The parameters used are summarized as follows.
  - LR (Logistic Regression): default settings.
  - DT (Decision Tree): max_depth = 4.
  - RF (Random Forests): n_estimator = 10.
  - AB (AdaBoost): default settings.
  - GB (Gradient Boosting): max_depth = 4.
- Up-to-date boosting classifiers:
  - CAB (CatBoost): depth = 4, learning_rate = 0.1, iterations = 30.
  - XGB (XGBoost): max_depth = 4, learning_rate = 0.1.
  - LGB (LightGBM): max_depth = 4, learning_rate = 0.1.

Figure 13 summarizes the comparison results for the eight classifiers in terms of prediction accuracy and running time. To obtain stable results, we repeated fivefold cross-validation 25

[10] http://bit.ly/Guggenheim_App.

[11] Scikit-learn 0.20, which is the latest version at the time of this submission, was used for the experiments.



(a) Performance comparison. **(b)** Running time comparison.

Fig. 13 Comparison between classifiers. LGB turns out to be the most effective among all classifiers. **a** Average accuracy on all experiments, **b** average running time on all experiments

times and then reported the averages by aggregating the results of the seven stores. As a result, LGB turned out to be the fastest classifier among the three best-performing classifiers—GB, XGB, and LGB. CAB was very fast as well as gave comparable results. Interestingly, DT took more time than RF and showed a better result in the default setting. Table 8 shows the details of Fig. 13 by showing the accuracy for each of the seven stores. The mean and standard deviation were calculated from the average accuracies of 25 different fivefold cross-validations.

B. Comparison on stacking models

To achieve additional performance improvement, we applied stacking (meta ensembling) with eight strategies. *Stacking* is a model ensembling technique used to combine multiple predictive models to generate a better model [38]. Usually, the stacked model is known to outperform each of the individual models owing to its smoothing nature and its ability to highlight each base model. The main point of the stacking is to utilize the prediction results of the base models as features for the stacking model in the second layer.

To do this, we selected CAB, XGB, and LGB as the base models. We further separated a training set into three subsets and used two subsets to make the prediction labels for the remaining subset. The prediction labels for the testing set were also calculated together three ($=_3 C_2$) times, and the three sets of the labels for the testing set were averaged for the final use. In this way, we generated the label features for both training and testing sets. These additional features are fed to the final LGB stacking model. We followed a general procedure from the reference[12] and added three options. Figure 14 illustrates the process of creating eight stacking models (M_1–M_8) through the choice of the three options. The description of the three options is as follows.

– Sampling strategy: A parameter that determines whether to use either random oversampling [15] or downsampling. This option is not directly related to the stacking, but we added it to improve the accuracy by treating the class imbalance problem.

[12] http://bit.ly/Kaggle_Guide_Stacking.

Table 8 Prediction accuracy (%) of various classifiers for the revisit prediction task

Store ID	A_GN	A_MD	E_GN	E_SC	L_GA	L_MD	O_MD
(a) *Experimental results on the models trained by first-time visitors*							
LR	59.56 ± 0.27	66.39 ± 0.13	61.80 ± 0.22	60.94 ± 0.21	69.08 ± 0.08	65.11 ± 0.09	63.48 ± 0.07
DT	62.42 ± 0.26	66.42 ± 0.11	64.97 ± 0.25	66.31 ± 0.09	69.98 ± 0.06	65.33 ± 0.05	63.94 ± 0.04
RF	61.63 ± 0.25	66.74 ± 0.13	61.84 ± 0.32	62.50 ± 0.20	69.34 ± 0.16	65.57 ± 0.11	63.58 ± 0.13
AB	62.51 ± 0.31	68.52 ± 0.12	65.39 ± 0.16	66.83 ± 0.14	71.05 ± 0.06	67.26 ± 0.05	65.68 ± 0.03
GB	63.13 ± 0.20	69.30 ± 0.10	66.69 ± 0.19	68.29 ± 0.10	71.83 ± 0.06	67.77 ± 0.05	66.21 ± 0.04
CAB	63.12 ± 0.27	68.43 ± 0.11	65.78 ± 0.18	67.44 ± 0.10	70.84 ± 0.07	66.94 ± 0.05	65.32 ± 0.05
XGB	63.14 ± 0.23	69.29 ± 0.10	66.67 ± 0.15	68.28 ± 0.10	71.79 ± 0.06	67.76 ± 0.04	66.19 ± 0.03
LGB	63.18 ± 0.25	69.31 ± 0.11	66.68 ± 0.18	68.28 ± 0.11	71.80 ± 0.06	67.77 ± 0.05	66.19 ± 0.03
(b) *Experimental results on the models trained by all visitors*							
Store ID	A_GN	A_MD	E_GN	E_SC	L_GA	L_MD	O_MD
LR	61.58 ± 0.34	69.30 ± 0.16	62.43 ± 0.48	60.15 ± 1.43	72.64 ± 0.05	75.41 ± 0.12	69.69 ± 0.20
DT	66.10 ± 0.27	72.18 ± 0.05	68.30 ± 0.11	70.85 ± 0.05	76.38 ± 0.04	78.29 ± 0.04	73.98 ± 0.01
RF	65.13 ± 0.26	71.38 ± 0.10	66.91 ± 0.24	68.84 ± 0.20	75.68 ± 0.11	77.74 ± 0.20	73.47 ± 0.18
AB	66.25 ± 0.25	73.19 ± 0.07	69.78 ± 0.10	72.02 ± 0.04	76.85 ± 0.05	79.12 ± 0.02	75.07 ± 0.01
GB	66.67 ± 0.21	74.11 ± 0.05	70.69 ± 0.09	73.06 ± 0.05	77.87 ± 0.03	79.75 ± 0.07	75.82 ± 0.01
CAB	66.62 ± 0.23	73.53 ± 0.05	69.96 ± 0.10	72.15 ± 0.06	77.14 ± 0.04	79.11 ± 0.09	75.16 ± 0.01
XGB	66.69 ± 0.21	74.09 ± 0.06	70.67 ± 0.07	73.05 ± 0.05	77.85 ± 0.03	79.74 ± 0.08	75.81 ± 0.01
LGB	66.70 ± 0.20	74.10 ± 0.05	70.69 ± 0.09	73.05 ± 0.06	77.86 ± 0.04	79.74 ± 0.08	75.81 ± 0.01

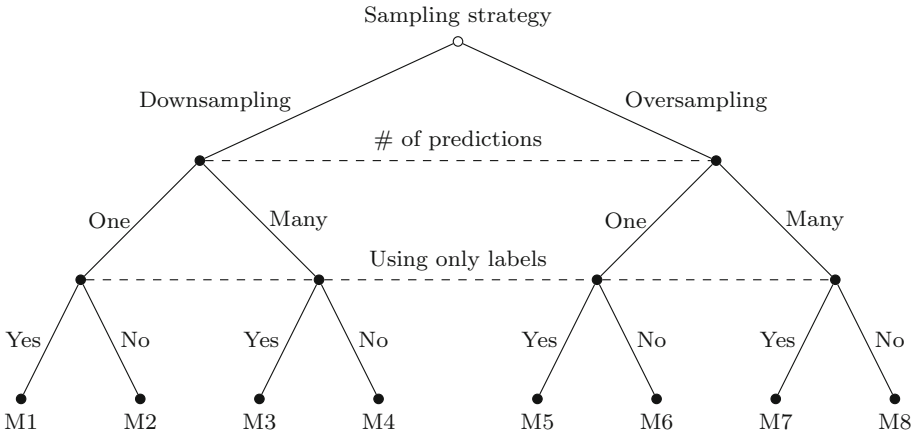


Fig. 14 Stacking options

– # of predictions: A parameter that determines whether to use one model or multiple models for each fold. The former case generates a single additional feature, and the latter case generates three additional features.
– Using only labels: A parameter that determines whether to use only the prediction labels (one or three features) or to use all existing features with the prediction labels (n+1 or n+3 features where n is the total number of hand-engineered features used).

Table 9 shows the average accuracy results obtained for each of the seven stores in details.[13] We observed that the performance improvement was not so high despite the long running time of the stacking model. Thus, we conjecture that each of the best-performing classifiers achieved almost the highest accuracy by itself.

C. Lower bounds of prediction accuracy

The visit logs v_k with the same visit count k are considered to have the same information. To maximize the accuracy, we must predict the label l of v_k by the following criteria:

$$\forall v : l(v \in v_k) = \begin{cases} 1, & \text{if } E[RV_{\text{bin}}(v_k)] \geq 1/2 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

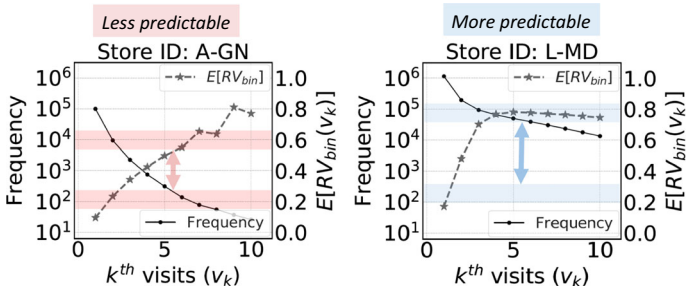
Considering each proportion $p_k = |v_k| / \sum_k |v_k|$ and simplifying $E[RV_{\text{bin}}(v_k)]$ as r_k, the lower bound accuracy of a model can be represented as $LB = \sum_k p_k \cdot \max(r_k, 1 - r_k)$. In the experiment of only first-time visitors, $LB = 1/2$ since $p_1 = 1$ and $r_1 = 1/2$.

The interpretation with the lower bound is as follows. For higher predictability, the revisit tendency of each v_k should be homogeneous. In Fig. 15, we can notice that store L_MD is more predictable than A_GN, because $|r_k - 0.5|$ of L_MD is larger than that of A_GN for the majority of k.

[13] We ran another five sets of fivefold cross-validation for this experiment. Thus, the values of the baselines in Table 9 are slightly different from those in Table 8 within the margin of error.

Table 9 Prediction accuracy (%) of stacking models for the revisit prediction task with the data of all visitors

Store ID	A_GN	A_MD	E_GN	E_SC	L_GA	L_MD	O_MD
Single model							
LR	61.57 ± 0.17	69.33 ± 0.07	62.61 ± 0.43	60.92 ± 0.96	72.65 ± 0.04	75.34 ± 0.14	69.64 ± 0.18
DT	66.17 ± 0.26	72.21 ± 0.03	68.33 ± 0.17	70.86 ± 0.06	76.37 ± 0.03	78.27 ± 0.01	73.98 ± 0.01
RF	65.17 ± 0.16	71.36 ± 0.13	66.84 ± 0.26	68.63 ± 0.24	75.68 ± 0.12	77.68 ± 0.17	73.37 ± 0.39
AB	66.40 ± 0.29	73.18 ± 0.05	69.82 ± 0.07	72.00 ± 0.05	76.82 ± 0.03	79.12 ± 0.01	75.06 ± 0.01
CAB	66.84 ± 0.11	73.80 ± 0.02	70.44 ± 0.15	72.61 ± 0.06	77.39 ± 0.03	79.38 ± 0.01	75.51 ± 0.01
XGB	66.78 ± 0.15	74.10 ± 0.04	70.67 ± 0.12	73.03 ± 0.07	77.83 ± 0.04	79.71 ± 0.01	75.81 ± 0.01
LGB	66.88 ± 0.23	74.11 ± 0.03	70.64 ± 0.13	73.04 ± 0.07	77.83 ± 0.02	79.71 ± 0.01	75.82 ± 0.02
Stacking model							
M1	66.56 ± 0.12	73.88 ± 0.03	70.49 ± 0.07	72.91 ± 0.09	77.56 ± 0.02	79.52 ± 0.01	75.66 ± 0.01
M2	66.70 ± 0.13	73.95 ± 0.03	70.52 ± 0.08	72.95 ± 0.08	77.62 ± 0.02	79.59 ± 0.01	75.69 ± 0.01
M3	66.57 ± 0.15	74.01 ± 0.02	70.55 ± 0.11	72.97 ± 0.08	77.79 ± 0.02	79.66 ± 0.01	75.77 ± 0.01
M4	66.78 ± 0.22	74.07 ± 0.02	70.65 ± 0.11	73.07 ± 0.06	77.82 ± 0.02	79.69 ± 0.01	75.80 ± 0.01
M5	67.04 ± 0.19	73.91 ± 0.05	70.62 ± 0.13	72.95 ± 0.10	77.58 ± 0.02	79.52 ± 0.01	75.65 ± 0.01
M6	67.00 ± 0.28	73.96 ± 0.04	70.64 ± 0.14	72.99 ± 0.09	77.64 ± 0.02	79.58 ± 0.01	75.69 ± 0.01
M7	66.88 ± 0.19	74.06 ± 0.04	70.67 ± 0.11	73.01 ± 0.08	77.80 ± 0.02	79.66 ± 0.01	75.77 ± 0.01
M8	66.97 ± 0.15	74.10 ± 0.04	70.71 ± 0.11	73.10 ± 0.07	77.83 ± 0.02	79.70 ± 0.01	75.80 ± 0.01



(a) The case of a less predictable store with LB 0.595.

(b) The case of a more predictable store with LB 0.741.

Fig. 15 Lower bound accuracies of two stores

D. Assumptions to interpret the data

Here, we would like to clarify how we count the first-time visitors and explain several underlying assumptions to consider.

- Assumption 1: Because we do not know whether customers visited a store before data was collected, we simply assume that the customers did not visit before the collection period. We believe that this assumption is reasonable because the stores in which we collected the data were relatively new at that time we began data collection.
- Assumption 2: Because customers are captured only when they turn on the Wi-Fi of their mobile device, we assume that the customers' Wi-Fi turn on behavior is consistent when they visit the store. Also, we assume that there is no correlation between Wi-Fi usage and customer groups (first-time visitors and VIP customers).
- Assumption 3: We assume that customers visit the store with a device having the same MAC address. For this purpose, we retained only Android devices but removed Apple devices in the preprocessing step, because the later versions of iOS 8.0 follow a MAC-address randomization policy [21] which makes infeasible to identify the same customer.

Rigorously speaking, the proportion of true first-time visitors would be less than 70% by considering all the effects explained above. Nevertheless, these customers are also likely to be early stage visitors.

E. Deciding the group movement threshold

We decided 30 s group movement threshold by the following logic. According to our observation at store E_GN in the afternoon of June 24 and June 26, 2017, 56% of 105 customers entered the store with their companions, which was more than half. Considering $p_x = 39.2\%$ as the on-site Wi-Fi turn on rate (Always-on: 29.2%, Conditionally-on: 10%) [24] and $p_y = 56\%$ as the actual proportion of customers in a group, we expected that $p_{yo} = 15.5\%$ of the total visitors were represented as having companions in our collected data of store E_GN (by Eq. 1 in Sect. 5.3.2). By setting 30 s as a threshold of accompaniment, we also obtained 15% of the total visitors were considered as having companions in the same data. By considering a gap between actual group ratio and observed group ratio, we claim that 30 s is an appropriate threshold to distinguish group movement.

References

1. Baumann P, Kleiminger W, Santini S (2013) The influence of temporal and spatial features on the performance of next-place prediction algorithms. In: Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing. ACM, pp 449–458
2. Besse PC, Guillouet B, Loubes J-M, Royer F (2017) Destination prediction by trajectory distribution based model. IEEE Trans Intell Transp Syst 99:1–12
3. Brébisson A, Simon É, Auvolat A, Vincent P, Bengio Y (2015) Artificial neural networks applied to taxi destination prediction. In: Proceedings of the 2015 ECML/PKDD discovery challenge. Springer, pp 40–51
4. Chen T, Guestrin C (2016) XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 785–794
5. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
6. Geng W, Yang G (2017) Partial correlation between spatial and temporal regularities of human mobility. Sci Rep 7:6249
7. Giannotti F, Nanni M, Pinelli F, Pedreschi D (2007) Trajectory pattern mining. In: Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 330–339
8. Hui SK, Bradlow ET, Fader PS (2009) Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior. J Consum Res 36(3):478–493
9. Hwang I, Jang Y (2017) Process mining to discover shoppers' pathways at a fashion retail store using a wifi-base indoor positioning system. IEEE Trans Autom Sci Eng 14:1786–1792
10. Jung S, Lim C, Yoon S (2011) Study on selecting process of visitor's movements in exhibition space. J Archit Inst Korea Plan Des 27(12):53–62
11. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: a highly efficient gradient boosting decision tree. In: Advances in neural information processing systems, vol 30. Curran Associates, Inc, pp 3146–3154
12. Kim S, Lee J-G (2018) Utilizing in-store sensors for revisit prediction. In: IEEE international conference on data mining. IEEE, pp 217–226
13. Kim T, Chu M, Brdiczka O, Begole J (2009) Predicting shoppers' interest from social interactions using sociometric sensors. In: CHI'09 extended abstracts on human factors in computing systems. ACM, pp 4513–4518
14. Lee J-G, Han J, Li X (2011) Mining discriminative patterns for classifying trajectories on road networks. IEEE Trans Knowl Data Eng 23(5):713–726
15. Lemaître G, Nogueira F, Aridas CK (2017) Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning. J Mach Learn Res 18(17):1–5
16. Lim C, Park H, Yoon S (2013) A study of an exhibitions space analysis according to visitor's cognition. J Archit Inst Korea Plan Des 29(8):69–78
17. Lim C, Yoon S (2010) Development of visual perception effects model for exhibition space. J Archit Inst Korea Plan Des 26(5):131–138
18. Liu G, Nguyen TT, Zhao G, Zha W, Yang J, Cao J, Wu M, Zhao P, Chen W (2016) Repeat buyer prediction for E-commerce. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 155–164
19. Lu X, Wetter E, Bharti N, Tatem AJ, Bengtsson L (2013) Approaching the limit of predictability in human mobility. Sci Rep 3:2923
20. Lv J, Li Q, Sun Q, Wang X (2018) T-CONV: a convolutional neural network for multi-scale taxi trajectory prediction. In: Proceedings of the 2018 IEEE international conference on big data and smart computing. IEEE, pp 82–89
21. Martin J, Mayberry T, Donahue C, Foppe L, Brown L, Riggins C, Rye EC, Brown D (2017) A study of MAC address randomization in mobile devices and when it fails. Proc Priv Enhanc Technol 2017(4):365–383
22. Mathew W, Raposo R, Martins B (2012) Predicting future locations with hidden Markov models. In: Proceedings of the 2012 ACM conference on ubiquitous computing. ACM, pp 911–918
23. Monreale A, Pinelli F, Trasarti R, Giannotti F (2012) WhereNext: a location predictor on trajectory pattern mining. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 637–646
24. OpenSignal, Inc (2016) Global state of mobile networks (August 2016). Technical report
25. Park S, Jung S, Lim C (2001) A study on the pedestrian path choice in clothing outlets. Korean Inst Inter Des J 28:140–148

26. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
27. Peppers D, Rogers M (2016) Managing customer experience and relationships. Wiley, New York
28. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A (2018) CatBoost: unbiased boosting with categorical features support. In: Advances in neural information processing systems, vol 31. Curran Associates, Inc, pp 6639–6649
29. Ren Y, Tomko M, Salim FD, Ong K, Sanderson M (2017) Analyzing web behavior in indoor retail spaces. J Assoc Inf Sci Technol 68(1):62–76
30. Sapiezynski P, Stopczynski A, Gatej R, Lehmann S (2015) Tracking human mobility using WiFi signals. PLoS ONE 10(7):e0130824
31. Scellato S, Musolesi M, Mascolo C, Latora V, Campbell AT (2011) Nextplace: a spatio-temporal prediction framework for pervasive systems. In: Proceedings of the 9th international conference on pervasive computing. Springer, pp 152–169
32. Sheth A, Seshan S, Wetherall D (2009) Geo-fencing: confining Wi-Fi coverage to physical boundaries. In: Proceedings of the 7th international conference on pervasive computing, pp 274–290
33. Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. Science 327(5968):1018–1021
34. Stanković RS, Falkowskib BJ (2003) The Haar wavelet transform: its status and achievements. Comput Electr Eng 29(1):25–44
35. Syaekhoni A, Lee C, Kwon Y (2018) Analyzing customer behavior from shopping path data using operation edit distance. Appl Intell 48:1912–1932
36. Tomko M, Ren Y, Ong K, Salim F, Sanderson M (2014) Large-scale indoor movement analysis: the data, context and analytical challenges. In: Proceedings of analysis of movement data, GIScience 2014 workshop
37. Um S, Chon K, Ro Y (2006) Antecedents of revisit intention. Ann Tour Res 33(4):1141–1158
38. Wolpert DH (1992) Stacked generalization. Neural Netw 5:241–259
39. Xue AY, Zhang R, Zheng Y, Xie X, Huang J, Xu Z (2013) Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. In: Proceedings of the 29th IEEE international conference on data engineering. IEEE, pp 254–265
40. Yada K (2011) String analysis technique for shopping path in a supermarket. J Intell Inf Syst 36(3):385–402
41. Yalowitz SS, Bronnenkant K (2009) Timing and tracking: unlocking visitor behavior. Visit Stud 12(1):47–64
42. Yan X, Wang J, Chau M (2015) Customer revisit intention to restaurants: evidence from online reviews. Inf Syst Front 17:645–657
43. Yan Z, Chakraborty D, Parent C, Spaccapietra S, Aberer K (2013) Semantic trajectories: mobility data computation and annotation. ACM Trans Intell Syst Technol 4(3):1–38
44. Ying JJC, Lee WC, Weng TC, Tseng VS (2011) Semantic trajectory mining for location prediction. In: Proceedings of the 19th ACM SIGSPATIAL international conference on advances in geographic information systems. ACM, pp 34–43
45. Yoshimura Y, Krebs A, Ratti C (2017) Noninvasive bluetooth monitoring of visitors' length of stay at the louvre. IEEE Perv Comput 16(2):26–34

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sundong Kim is currently a Ph.D. Candidate in the Graduate School of Knowledge Service Engineering/Department of Industrial & Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST). He received a B.S. and M.S. from KAIST, in 2013 and 2015, respectively. His research interests include predictive analytics, user modeling, and deep learning.

Jae-Gil Lee is an Associate Professor at Graduate School of Knowledge Service Engineering/Department of Industrial & Systems Engineering, Korea Advanced Institute of Science and Technology (KAIST) and is leading Data Mining Lab. Before that, he was a Postdoctoral Researcher at IBM Almaden Research Center and a Postdoc Research Associate at Department of Computer Science, University of Illinois at Urbana-Champaign. At IBM, he was one of the key contributors to IBM Smart Analytics Optimizer. At University of Illinois, he worked on spatiotemporal data mining with Prof. Jiawei Han. He earned the Ph.D. degree under the supervision of Prof. Kyu-Young Whang from KAIST in 2005. His research interests encompass spatiotemporal data mining, stream data mining, and big data analysis.