2018 IEEE International Conference on Data Mining

Utilizing In-Store Sensors for Revisit Prediction

Sundong Kim
KAIST, Daejeon, Republic of Korea
sundong.kim@kaist.ac.kr

Jae-Gil Lee*
KAIST, Daejeon, Republic of Korea
jaegil@kaist.ac.kr

***Abstract*—Predicting revisit intention is very important for the retail industry. Converting first-time visitors to repeating customers is of prime importance for high profitability. However, revisit analyses for offline retail businesses have been conducted on a small scale in previous studies, mainly because their methodologies have mostly relied on manually collected data. With the help of noninvasive monitoring, analyzing a customer's behavior inside stores has become possible, and revisit statistics are available from the large portion of customers who turn on their Wi-Fi or Bluetooth devices. Using Wi-Fi fingerprinting data from ZOYI, we propose a systematic framework to predict the revisit intention of customers using only signals received from their mobile devices. 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%. Our framework showed a feasibility to predict revisits using customer mobility from Wi-Fi signals, that have not been considered in previous marketing studies. Toward this goal, we examine the effect of data collection period on the prediction performance and present the robustness of our model on missing customers. Finally, we discuss the difficulties of securing prediction accuracy with the features that look promising but turn out to be unsatisfactory.**

I. Introduction

How can we detect a customer who is willing to visit a store again, without performing extensive surveys? Is it really possible to predict a customer's intention to revisit the store without knowing their purchase history, store satisfaction, age, or even their residence location? In this study, we introduce a revisit prediction framework using only Wi-Fi signals collected by in-store sensors.

By targeting the potential loyal customers who are likely to revisit, merchants can considerably save promotion cost and enhance return on investment [1]. Many studies in recent years have focused on *online* stores and online text reviews with the help of a data provider [2], [3]. In contrast, the analysis of revisit intention in the *offline* environment has not advanced significantly over the last few decades. The main reason for this lack of progress lies in the difficulties of collecting large-scale data that is closely related to key attributes of revisiting, such as customer satisfaction with products, service quality, atmosphere, purchase history, and personal profiles [3], [4]. The first three attributes are subjective information that is difficult to capture in the offline environment, and the last two attributes are considered as confidential corporate information that is not easily accessible. Owing to these limitations, research on customer revisits in offline 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.

Recently, RFID [5], [6], Bluetooth [7], or Wi-Fi fingerprinting [8] enable us to collect high-frequency signal data without installing any applications on customer devices [9], [10]. 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 [7] and supermarkets [11], providing empirical findings of customer behavior. Nowadays, collecting data in a certain physical boundary is called as geofencing, and its market size is accounted for USD 8 billion in 2014 and is expected to reach 40 billion by 2019 [12]. 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 propritary solutions provide visitor monitoring results, such as funnel or hot-spot analysis results displayed through a dashboard.

This study moves one step forward, from visitor monitoring to customer revisit prediction. It is known that motion patterns unconsciously reflect consumer's interest in and satisfaction with the store [13]. Therefore, our key task is to find patterns that affect a customer's revisit. We systemically design features to summarize each visit as follows. 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 use the customer mobility data[1] obtained from seven flagship stores in Seoul. 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 %[2] 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

[1]For future researchers, we are negotiating to release sampled datasets with our code at *https://github.com/kaist-dmlab/revisit*.

[2]The proportion of users in their twenties who keep their Wi-Fi on is 29.2 %, according to a Korea Telecom survey [14].

∗ Jae-Gil Lee is the corresponding author.

978-1-5386-9159-5/18/$31.00 ©2018 IEEE
DOI 10.1109/ICDM.2018.00037

IEEE computer society

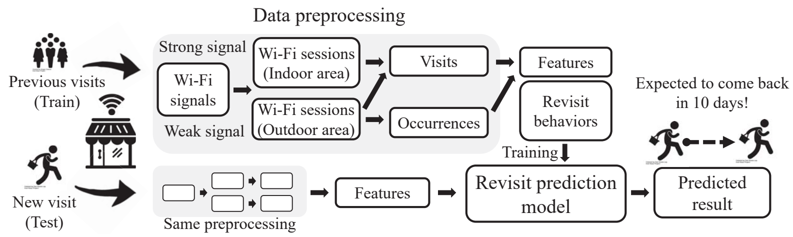


Figure 1: Revisit prediction framework architecture.

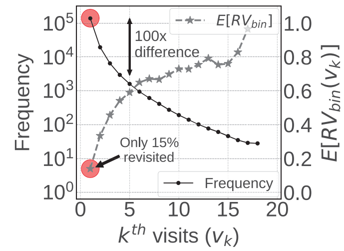


Figure 2: Revisit statistics of store E_GN.

detects his/her Wi-Fi signals, and through the data preprocessing described in Section II-B, 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 Section IV. 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 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 k^{th} visits (v_k), and the gray line denotes the average revisit rate $E[RV_{bin}(v_k)]$ 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.

As our additional contributions, we demonstrate the effectiveness of using customer mobility in comparison with baseline models considering visit distribution and temporal information. We also report the predictive power of each feature group and semantic level to show whether or not the trajectory abstraction boosts the predictability. We examine how the collection period and the volume of data affect performance. The final goal of this paper is to share the unexpected challenges faced when two groups of data show inherent differences in a statistical sense.

II. Data Description

A. Data Collection Stores

We collected data from seven flagship stores located in the center of Seoul. 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. Table I presents the statistics of the datasets.

B. Preprocessing to Generate Trajectories

1) Signal to Session Conversion: 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 RSSI level. Signals are collected continuously from each device at fairly short intervals, which are less than $1\,\text{s}$. 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.

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 $5\,\text{m}$), or far away from the store (up to $30\,\text{m}$). 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.

III. 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 (RI_{days}) and the revisit intention (RI_{bin}), which are the labels in our prediction model. Finally, we introduce the revisit prediction problem.

Table I: Statistics of the datasets.

Store ID	A_GN	A_MD	E_GN	E_SC	L_GA	L_MD	O_MD
Location	Gangnam	Myeong-dong	Gangnam	Sinchon	Garosu-gil	Myeong-dong	Myeong-dong
Collection length	222 days	220 days	300 days	373 days	990 days	747 days	698 days
# of sensors	16	27	40	22	14	11	27
# of total signals	165,443,933	890,267,554	939,815,485	632,106,890	1,935,362,316	2,818,001,166	6,499,265,088
Signal data size	15 GB	77 GB	148 GB	99 GB	164 GB	242 GB	567 GB
# of total sessions	19,937,461	33,862,373	81,449,603	34,131,034	40,282,894	74,324,676	90,713,930
# of indoor sessions ≥ 5 s	636,843	3,250,072	1,353,709	1,921,635	5,461,060	11,065,561	15,581,820
# of visits ≥ 60 s	112,672	327,940	183,246	270,366	1,062,226	1,718,359	2,008,384
# of unique visitors ≥ 60 s	100,741	232,051	147,096	186,617	846,487	1,171,583	1,065,803
Avg. revisit rate	11.73 %	31.99 %	21.18 %	36.55 %	21.22 %	32.98 %	48.73 %

A. 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 [15], 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 k^{th} 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 k^{th} 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 later.

B. 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 $RV_{days}(v) = d$, and a positive *revisit intention*, denoted by $RV_{bin}(v) = 1$, as in Definition 5.

Definition 5: If two consecutive visits of customer c_i

$$v_k = v_k(c_i, [t_{k,s}, t_{k,e}], \mathbb{T}_k)$$
$$v_{k+1} = v_{k+1}(c_i, [t_{k+1,s}, t_{k+1,e}], \mathbb{T}_{k+1})$$

meet the condition $t_{k,e} < t_{k+1,s}$, the *revisit interval* $RV_{days}(v_k)$ and the *revisit intention* $RV_{bin}(v_k)$ of the former visit v_k are as follows:

$$RV_{days}(v_k) = \text{\# days of } t_{k+1,s} - t_{k,e}$$
$$RV_{bin}(v_k) = 1$$

If a visit v_k does not have any following revisit, then

$$RV_{days}(v_k) = \infty$$
$$RV_{bin}(v_k) = 0$$ □

C. Predictive Analytics

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

IV. 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 II gives a summary of the features in our framework, which are self-explanatory. The first column describe data sources used to extract features, leading to ten different feature groups. The first six feature groups and **[TV]** are generated from the *visit* itself. **[UE]**, **[SA]**, and **[GM]** are generated using certain references: **[UE]** uses sales and holiday information

Table II: Description of the representative features according to the data sources and feature groups. 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.

Data sources	Feature groups	Twenty representative features (Among 866 features of store E_GN)	Semantic level of features					
			Sensor	Category	Floor	Gender	In/Out	None
Moving pattern of the visit	Overall statistics **[OS]** (IV-A1)	f1 = Total dwell time					✓	
		f2 = Trajectory length	✓	✓	✓	✓		
		f3 = Skewness of dwell time of each area	✓	✓		✓		
	Travel Distance/ Speed/Acceleration **[TS]** (IV-A2)	f4 = Total distance traveled inside the store		✓				
		f5 = Speed based on transition time	✓	✓	✓	✓		
		f6 = First-k HWT coefficients of acceleration	✓	✓	✓	✓		
	Area preference **[AP]** (IV-A3)	f7 = Coherency of dwell time for each level		✓	✓	✓		
		f8 = Top-k-area dwell time	✓	✓	✓	✓		
	Entrance and Exit pattern **[EE]**	f9 = Exit gate	✓					
		f10 = Number of previous re-entry on that day					✓	
	Heuristics **[HR]**	f11 = Wears clothes but does not buy		✓				
	Statistics of each area **[ST]**	f12 = Number of time sensed in the area	✓	✓	✓	✓		
		f13 = Stdev of dwell time for the area	✓	✓	✓	✓		
Temporal information of the visit	Time of visit **[TV]**	f14 = Day of the week						✓
	Upcoming events **[UE]** (IV-A8)	f15 = Remaining day until the next sale						✓
		f16 = Number of holidays for next 30 days						✓
Occurrences before the visit	Store accessibility **[SA]** (IV-A9)	f17 = Number of days since the last access					✓	
		f18 = Average interarrival time					✓	
Simultaneous visits	Group movement **[GM]** (IV-A10)	f19 = Presence of companions					✓	
		f20 = Number of companions					✓	

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: E_GN and E_SC, where we continuously tracked their floor plans during data collection periods. In Table II, 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.

Figure 3 and Figure 4 display meaningful relationships between the feature values of $f1$, $f7$, $f9$, $f15$, and $f17$ with the average revisit intention rate $E[RV_{bin}(v)]$. 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.

A. Feature Descriptions

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 $(f1)$, trajectory length $(f2)$, and average frequency of each area. We also apply skewness $(f3)$ or kurtosis to measure the asymmetric or fat-tail behavior of the dwell-time distribution of each area.

2) Travel Distance, Speed, and Acceleration ***[TS]:*** TS features are in-depth information that needs to be explored [16]. To approximate the physical distance $(f4)$ 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 $(f5)$, and we modeled the acceleration from the speed variation between consecutive areas. A time series analysis using the Haar Wavelet Transform (HWT) [17] was performed, as well as statistical analysis, to determine how the customer's interests changed with time. We included the first-16 HWT coefficients $(f6)$ in our feature set.

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 $(f8)$, 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 $(f7)$ 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) Entrance and Exit pattern ***[EE]:*** Interestingly, customers leaving through the back door $(f9)$ revisited $13.6\,\%$ more than

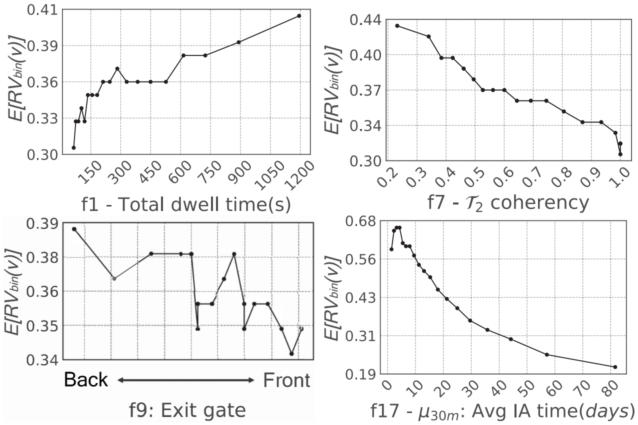


Figure 3: The relationship between the selected features and RV_{bin} in store E_SC ($E[RV_{bin}(v)\ (v \in V_{all})] = 0.3616$). Each marker point represents the average revisit intention $E[RV_{bin}(v)]\ (v \in V_b)$ of the set V_b of visits obtained by equal-frequency-binning the entire data according to feature values. Indoor moving pattern features $f1$, $f7$, and $f9$ shows at most $40\,\%$ deviation of $E[RV_{bin}(v)]$ according to the feature value. The store accessibility feature $f17$ shows $325\,\%$ deviation, which is the highest among the selected features. For $f9$, the group of customers who are most likely to use the back door are located on the left side of the x-axis.

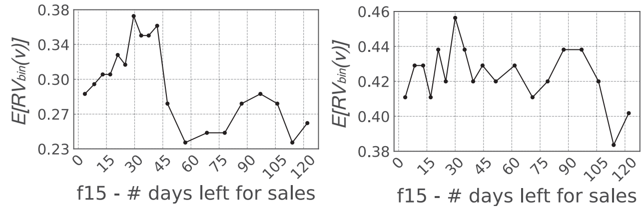
customers leaving through the front door, according to our data. Therefore, we estimated the customers' entrance and exit patterns from the sensors nearby the front and back doors. We expected that customers familiar with the store might have used a more convenient door.

5) Heuristics ***[HR]:*** We collected ideas about what kinds of patterns are likely to appear from people who are willing to revisit. In general, people commented that the dwell-time in the fitting room (FR) and the checkout counter (CS) reflected the customers' interest and their purchasing pattern. As we do not know whether the customer actually bought the item or tried it in FR, the time spent in those areas was used to make inferences. Here are two representative heuristics anticipating the revisit of customers for future purchase.

- If a customer wears clothes in the FR without purchase ($\leq 1\,\mathrm{min}$ in the CS): $f11 = 1$, for all other cases: $f11 = 0$.
- If a customer stays in the store much longer (= $10\,\mathrm{min}$) than average visitors, without purchase: $f = 1$, if not: $f = 0$.

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 ($f12$), 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 ($f13$).

7) Time of Visit ***[TV]:*** The temporal features include the time of visit such as hour of the day and day of the week ($f14$) 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.



(a) On first-time visitors (v_1). (b) On all visitors.

Figure 4: Key factors of v_1's revisit: discount and seasonality. <u>Discount-sensitive</u>: A set V_b of customers who visited between 30–45 days before a clearance sale showed a high $E[RV_{bin}(v)]$ ($v \in V_b$) compared to other customers; this difference was more apparent in first-time visitors than all visitors. <u>Seasonal-sensitive</u>: Another peak of $E[RV_{bin}(v)]$ appeared on the set of customers who made a visit between 90–105 days before the sale. It described the seasonal revisit, and it was also more noticeable to first-time visitors than all visitors.

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 store in the holiday seasons since they are outside the city. By combining simple extrinsic information, the temporal features, particularly **[UE]**, becomes the second strongest predictive feature groups. It contains six features, including a number of days left for the next clearance sale ($f15$) and a number of holidays for next 30 days ($f16$), as numeric features.

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 (IATs). We utilized two additional outdoor areas nearby the store—$5\,\mathrm{m}$ and $30\,\mathrm{m}$ zone—to detect the customer occurrences. Considering a customer's arrival process to $5\,\mathrm{m}$ zone, let us denote the time of the first occurrence by T_1. For $k > 1$, let T_k denote the elapsed time between $k-1^{\text{th}}$ and the k^{th} event. We call the sequence $\{T_k, k = 1, 2, ..., \}$ as the *sequence of IATs*. Considering the target visit as n^{th} 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 ($f17$);
- $\mathbb{1}_{n>1}$: Existence of having any occurrence before the visit;
- $\mu = \sum_{k=2}^{n} T_k/(n-1)$: Average IAT ($f18$);
- $\sigma = \sqrt{\sum_{k=2}^{n} (T_k - \mu)^2/(n-1)}$: SD of IATs;

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

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 $(f19)$ and the number of companions $(f20)$. 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 $30\,\mathrm{s}$.

B. Unused Features

1) Sequential Patterns: Sequential patterns [18], [19] 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 i^{th} patterns. We considered diverse levels of sequential patterns, as in Table III, but the result was not satisfactory. Despite that it was expensive to generate the features, their information gains were typically low.

Table III: 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.
$A \overset{*}{\dashrightarrow} B \overset{*}{\dashrightarrow} C$	A partial sequential pattern [19], an arrow $A \overset{*}{\dashrightarrow} B$ denotes that there might exist additional elements between A and B.
$A_{short} \overset{*}{\dashrightarrow} B_{long} \overset{*}{\dashrightarrow} C_{short}$	A partial sequential pattern which has a time constraint for the dwell time of each element.

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. 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.

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.

V. Evaluation Results

A. Settings

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.

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.

3) Dividing Data: Train and test data were divided through three settings:

- S1: 5-fold CV by dividing customers, where each customer data can only be included in a single fold.
- S2: 5-fold CV by dividing visits[3], 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 the S1 and S2 was insignificant to the fourth decimal place. By 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 (§ V-C1). Because of the page limit, we report the main results using the configuration S1.

4) Classifier: All results described in this section were obtained using Python API of the XGBoost [20] library that implements the gradient boosted tree [21] framework, which gave the *best* performance among various classifiers, including AdaBoost, random forest, and logistic regression implemented in the Python Scikit-learn [22] library. We also tried variants of RNN models using PyTorch [23], but the accuracy was lower than that of XGBoost and the running time took several magnitudes longer. 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

[3]As a result of § IV-B3, our model is considered to be safe to perform CV.

no longer than $1\,\mathrm{min}$ in a single machine (Intel i7-6700 with $16\,\mathrm{GB}$ RAM, without GPU).

B. Results

1) Overall Results: Table IV 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.

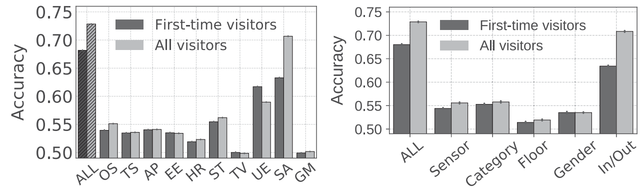
Table IV: Performance of classification and regression tasks.

Store ID	Period (days)	#features	Customer type	# data (# revisitors)	Accuracy	RMSE
A_GN	222	256	First	99,497 (9,514)	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

2) Predictive Power of Feature Groups: Figure 5(a) 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 II. For the convenience of comparison, the leftmost bar on the figure represents the results when all features in Table IV 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.

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 Figure 5(b), 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.

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



(a) On feature groups.

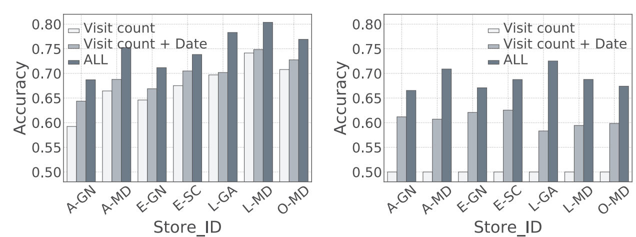
(b) On semantic levels.

Figure 5: Performance comparison on feature groups and semantic levels (store E_SC).

obtained from revisit statistics, shown in Figure 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 A.

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 6 reports the accuracy of our model[4] 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 $8.0–24.3\,\%$.



(a) On all visitors.

(b) On first-time visitors.

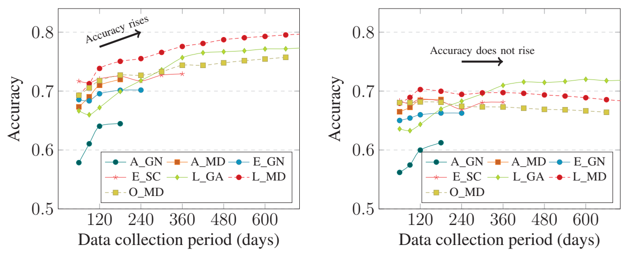
Figure 6: Effectiveness of analyzing customer trajectories.

C. Discussions

1) Importance of Data Collection Period: We are wondering how much the model's performance varies depending on the amount of data. Figure 7(a) shows that the overall prediction accuracy increases as the length of the data 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 the information on revisiting customers. Therefore,

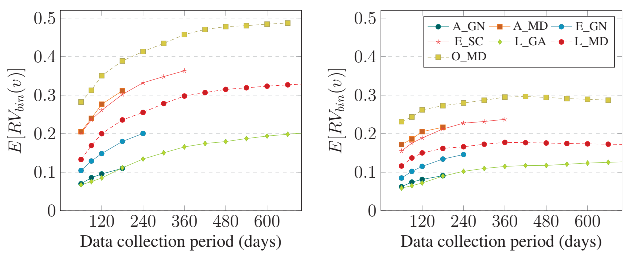
[4]For this experiment, we included visit count and date to our feature set, so the overall accuracy is slightly higher than the values reported from Table IV.

the labels in the training data could be inaccurate if we collected the information for an *insufficient period*. To confirm our conjecture, we also examined the proportion of customers' revisit intention as the data collection progressed, as in Figure 7(c). The proportion, $E[RV_{bin}(v)]$, indeed increased as the data collection period increased. However, prediction accuracy on first-time visitors did not always increase. We notice that 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 longer data collection period, performance improvement occurs by having more positive cases for regular customers.



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

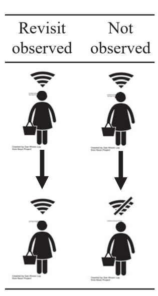
(b) On first-time visitors, accuracy does not increase after few months.



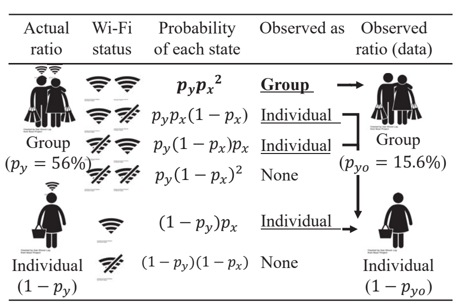
(c) On all visitors, average revisit rates keep increasing for all cases.

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

Figure 7: Impact of the data collection period.



(a) Revisit.



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

Figure 8: 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 was 56% instead of 15.6%. Researchers must not interpret the data as it is, when explaining the real behavior.

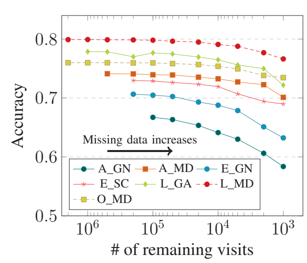
2) Real Behavior and Collected Data—Are They Same?: Noninvasively collected data is also limited, considering that not all users turn on Wi-Fi of their mobile device. Since the 4G LTE connection is very fast and ubiquitous in Korea, the proportion of 'always-on' users is just 30 % [24]. This limitation implies that the datasets were missing some customer behaviors in the real world. Figure 8(a) shows untraceable revisits due to the conditional Wi-Fi usage of the customer, and Figure 8(b) shows a gap between 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).

$$p_{yo} = \frac{Observed(Group)}{Observed(Group) + Observed(Individual)} = \frac{p_y(p_x)^2}{p_y(p_x)^2 + 2p_yp_x(1-p_x) + (1-p_y)p_x} = \frac{p_xp_y}{1+p_y-(p_x)^2} \tag{1}$$

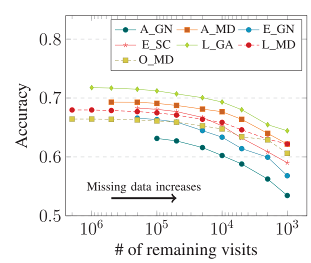
In the future, if customers' behaviors are more traceable with additional sensing technologies, we believe that *noninvasively* collected data will better reflect actual customer behaviors.

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 9 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.



(a) On all visitors.



(b) On first-time visitors.

Figure 9: Model robustness on missing customers.

4) Meaningful Insights but Low Predictability: We would like to point out that securing prediction accuracy can be difficult although the differences between re-visitors and

Table V: Statistics of feature values with revisit status, and their final predictability: statistics from the store O_MD. ($FV_1 = E[FV(v)|RV_{bin}(v) = 1]$: Average feature values of revisitors, r_{pb}: Point-biserial correlation)

Feature Name	Feature value difference by revisit status				Revisit rate difference by feature values				
	FV_1	FV_0	diff$_1$	p-value	$\max(RV_{bin})$	$\min(RV_{bin})$	diff$_2$	r_{pb}	Accuracy
f_a: Avg interarrival time (5 m)	21.8 days	44.4 days	104.2 %	0****	0.841	0.358	134.7 %	-0.207	0.7346
f_b: Total dwell time	3211 s	1612 s	99.2 %	0****	0.721	0.335	115.2 %	0.216	0.6005
f_c: Percentages of time spent in the 3rd longest area	0.112	0.087	28.5 %	0****	0.622	0.335	85.8 %	0.152	0.6035
f_d: Avg dwell time for each area	358 s	348 s	2.7 %	0****	0.588	0.410	43.5 %	0.007	0.5584

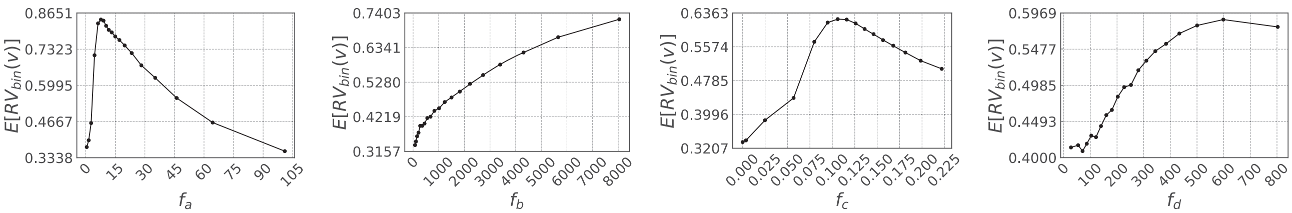


Figure 10: Detailed relationship between four features and $E[RV_{bin}(v)]$ mentioned in Table V.

non-revisitors are obvious. Some feature values significantly differ by the revisit status, each of which should be helpful to explain the difference between 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 V, 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 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 %.

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 non-linear tendency, its predictive power can be strong. The statistics of f_b and f_c in Table V confirms our argument. We assert that our high-quality prediction came from a combination of various kinds of features which behave differently.

VI. Related Work

Predictive analytics using trajectories. Using trajectories, next location prediction is the most studied topic in the computer science community. To predict the next location, frequent trajectory patterns [25], nonlinear time series analysis of the arrival and residence time [26], and HMM [27] were applied. The results support the prediction of the next location using partial trajectories is feasible, along with the regularity studies of human mobility [28]–[30]. The main difference between our study and previous studies is a prediction objective. We studied the customers' revisit intentions in the offline stores using indoor trajectories. Thus, our model focused on predicting revisits instead of locations. As far as we know, there are no studies of predicting revisit intention using trajectories captured by in-store sensors.

Customer behavior in the store. Park et al. [16] 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. [5] to find that consumers who spent more time in the grocery store become more purposeful. Although these studies did not focus on customers' revisit, they were valuable resources for us to develop features that describe customers' motion patterns.

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 [31] to rearrange the exhibits to enhance the satisfaction of visitors [32]. With the help of noninvasive monitoring, visitor studies in the museum have come to a new phase. Yoshimura et al. [7] installed eight beacons in the Louvre Museum and analyzed the most popular paths to mitigate a micro-congestion inside the museum. By tracking visitors' movements, the Guggenheim Museum [33] increased customers' engagement by making smarter curatorial decisions.

VII. Conclusions

Various retail analytics companies have set up sensors to monitor customer mobility in offline 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.

Acknowledgment

This project is funded by Microsoft Research Asia (Microsoft PO #: 97838317). We appreciate Minseok Kim for helping surveys on offline stores. We also thank ZOYI for providing additional datasets for experiments.

References

[1] D. Peppers and M. Rogers, *Managing customer experience and relationships*. Wiley, 2016.
[2] G. Liu *et al.*, "Repeat buyer prediction for E-Commerce," in *KDD*, 2016.
[3] X. Yan, J. Wang, and M. Chau, "Customer revisit intention to restaurants: Evidence from online reviews," *Information Systems Frontiers*, vol. 17, pp. 645–657, 2015.
[4] S. Um, K. Chon, and Y. Ro, "Antecedents of revisit intention," *Annals of Tourism Research*, vol. 33, no. 4, pp. 1141–1158, 2006.
[5] S. K. Hui, E. T. Bradlow, and P. S. Fader, "Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior," *Journal of Consumer Research*, vol. 36, no. 3, 2009.
[6] A. Syaekhoni, C. Lee, and Y. Kwon, "Analyzing customer behavior from shopping path data using operation edit distance," *Applied Intelligence*, 2016.
[7] Y. Yoshimura, A. Krebs, and C. Ratti, "Noninvasive bluetooth monitoring of visitors' length of stay at the louvre," *IEEE Pervasive Computing*, vol. 16, no. 2, pp. 26–34, 2017.
[8] M. Tomko, Y. Ren, K. Ong, F. Salim, and M. Sanderson, "Large-scale indoor movement analysis: the data, context and analytical challenges," in *GIScience Workshop*, 2014.
[9] Y. Ren, M. Tomko, F. D. Salim, K. Ong, and M. Sanderson, "Analyzing web behavior in indoor retail spaces," *Journal of the Association for Information Science and Technology*, vol. 68, no. 1, pp. 62–76, 2017.
[10] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann, "Tracking human mobility using WiFi signals," *PLoS ONE*, 2015.
[11] K. Yada, "String analysis technique for shopping path in a supermarket," *Journal of Intelligent Information Systems*, vol. 36, no. 3, 2011.
[12] PYMNTS. (2017) Geotracking gives brick-and-mortar a leg up on ecommerce. [Online]. Available: http://bit.ly/pymnts_geofencing
[13] T. Kim, M. Chu, O. Brdiczka, and J. Begole, "Predicting shoppers' interest from social interactions using sociometric sensors," in *CHI EA*, 2009.
[14] ZOYI, "Wi-Fi usage survey," http://bit.ly/wifi_survey_file, July 2015.
[15] Z. Yan, D. Chakraborty, C. Parent, S. Spaccapietra, and K. Aberer, "Semantic trajectories: Mobility data computation and annotation," *TIST*, vol. 4, no. 49, 2013.
[16] S. Park, S. Jung, and C. Lim, "A study on the pedestrian path choice in clothing outlets," *Korean Institute of Interior Design Journal*, vol. 28, pp. 140–148, 2001.
[17] R. S. Stanković and B. J. Falkowskib, "The Haar wavelet transform: its status and achievements," *Computers & Electrical Engineering*, vol. 29, no. 1, pp. 25–44, 2003.
[18] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, "Trajectory pattern mining," in *KDD*, 2007.
[19] J.-G. Lee, J. Han, and X. Li, "Mining discriminative patterns for classifying trajectories on road networks," *TKDE*, vol. 23, no. 5, 2011.
[20] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *KDD*, 2016.
[21] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.
[22] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
[23] PyTorch. (2018) Pytorch recurrent layers. [Online]. Available: http://pytorch.org/docs/0.3.0/nn.html#recurrent-layers
[24] OpenSignal, Inc, "Global state of mobile networks (August 2016)," Tech. Rep., 2016.
[25] A. Monreale, F. Pinelli, R. Trasarti, and F. Giannotti, "WhereNext: A location predictor on trajectory pattern mining," in *KDD*, 2009.
[26] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, "Nextplace: A spatio-temporal prediction framework for pervasive systems," in *Pervasive*, 2011.
[27] W. Mathew, R. Raposo, and B. Martins, "Predicting future locations with hidden markov models," in *UbiComp*, 2012.
[28] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific Reports*, 2013.
[29] W. Geng and G. Yang, "Partial correlation between spatial and temporal regularities of human mobility," *Scientific Reports*, 2017.
[30] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.
[31] S. S. Yalowitz and K. Bronnenkant, "Timing and tracking: Unlocking visitor behavior," *Visitor Studies*, vol. 12, no. 1, pp. 47–64, 2009.
[32] S. Jung, C. Lim, and S. Yoon, "Study on selecting process of visitor's movements in exhibition space," *Journal of the Architectural Institute of Korea Planning & Design*, vol. 27, no. 12, pp. 53–62, 2011.
[33] S. from Harvard Business School. (2015) The Guggenheim museum: Tracking visitor movements to deliver a better museum experience. [Online]. Available: http://bit.ly/guggenheim_tracking

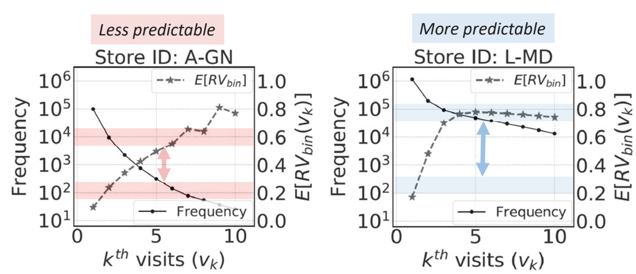
Appendix A
Lower Bounds of Prediction Accuracy

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.

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

Considering each proportion $p_k = |v_k| / \sum_k |v_k|$ and simplifying $E[RV_{bin}(v_k)]$ 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 Figure 11, 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.



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

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

Figure 11: Lower bound accuracies of two stores.