Utilizing In-Store Sensors for Revisit Prediction

Sundong Kim and Jae-Gil Lee
Korea Advanced Institute of Science and Technology

https://github.com/kaist-dmlab/revisit
While You Are Shopping
Collecting Data with Wi-Fi APs
Retail Analytics

- Provide a dashboard, as well as consultancy services
- Data-driven monitoring examples:

Visitors

Outside Traffic by Hour

Utilizing In-Store Sensors for Revisit Prediction (by Sundong Kim)
Related Works & Our Study

Indoor Tracking
- Interior design, Museum
- Visitor locations → Measure interests → Display plan

Prediction (x)

Predictive Analytics
- Next location
- Customer life-time value
- Churn (On-line)

Off-line Revisit (x)

Revisit Studies
- Marketing, tourism
- Questionnaire
- Qualitative Factors

Mobility (x)

To Discover the Relation between
Customer Revisit and their Mobility
More than 85% of retail purchases still happen offline. [Link]

Retaining customer is very important. [Article]
(5% more retention → 25-95% more profit)

More than 70% of visits are first-time visits

Rate of revisit

More than 85% of retail purchases still happen offline. [Link]

Retaining customer is very important. [Article]
(5% more retention → 25-95% more profit)
RQ1: How to predict customer revisits?
   → Using a GBT model with carefully designed features.

RQ2: How much effect of trajectory has on prediction performance?
   → Accuracy improves by 5-12% compared to LBs.
Outline

• Introduction
• **Prediction Framework**
• Features
• Performances
• Conclusion
Our Framework

Data preprocessing

Previous visits (Train)

New visit (Test)

Strong signal
Wi-Fi sessions (Indoor area)
Visits
Features
Expected to come back in 10 days!

Weak signal
Wi-Fi sessions (Outdoor area)
Occurrences
Revisit behaviors

Same preprocessing

Features
Revisit prediction model
Predicted result
Multi-level Trajectories

- Multi-level descriptions of the customer visit
Outline

• Introduction
• Prediction Framework
• **Features <<**
• Performances
• Conclusion
Feature Engineering

• Considered feature groups:
  • Overall statistics
  • Travel distance/speed/acceleration
  • Area preference
  • Entrance and exit pattern
  • Heuristics
  • Statistics of each area
  • Time of visit
  • Upcoming events
  • Store accessibility
  • Group movement
Feature value & revisit rate (1)

- $T_2$ level:
  - Customers who focus on the certain area revisit less!
  - Top 5% longest staying customers revisit 30% more!
Feature value & revisit rate (2)

Customers who use back door revisit more!

People who stroll around the area revisit very often!

- AVG IA time: Average Interarrival time
"Sale" for first-time visitors

Number of days left for sales:
- Feature with non-linear relationship

Effect of clearance sale

(a) First-time visitors: Prone to special events.  (b) All visitors: Indifferent to events.
Outline

• Introduction
• Prediction Framework
• Features
• Performances
• Conclusion
Utilizing In-Store Sensors for Revisit Prediction (by Sundong Kim)

Mobility Data from In-Store Sensors

- 7 Flagship stores
- 110K-2M visits/store
- 220-990 days collected
- Avg. traj length = 6.56

<table>
<thead>
<tr>
<th>Shop ID</th>
<th>A_GN</th>
<th>A_MD</th>
<th>E_GN</th>
<th>E_SC</th>
<th>L_GA</th>
<th>L_MD</th>
<th>O_MD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Location</td>
<td>Seoul, Korea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Length (days)</td>
<td>222</td>
<td>220</td>
<td>300</td>
<td>373</td>
<td>990</td>
<td>747</td>
<td>698</td>
</tr>
<tr>
<td># sensors</td>
<td>16</td>
<td>27</td>
<td>40</td>
<td>22</td>
<td>14</td>
<td>11</td>
<td>27</td>
</tr>
<tr>
<td>Data size</td>
<td>15GB</td>
<td>77GB</td>
<td>148GB</td>
<td>99GB</td>
<td>164GB</td>
<td>242GB</td>
<td>567GB</td>
</tr>
<tr>
<td># visits &gt; 60s</td>
<td>0.11M</td>
<td>0.33M</td>
<td>0.18M</td>
<td>0.27M</td>
<td>1.06M</td>
<td>1.72M</td>
<td>2.01M</td>
</tr>
<tr>
<td>Revisit rate</td>
<td>11.73%</td>
<td>31.99%</td>
<td>21.18%</td>
<td>36.55%</td>
<td>21.22%</td>
<td>32.98%</td>
<td>48.73%</td>
</tr>
</tbody>
</table>
Results: Prediction Accuracy

<table>
<thead>
<tr>
<th>Store</th>
<th>Accuracy (First)</th>
<th>Accuracy (All)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A_GN</td>
<td>0.6336</td>
<td>0.6689</td>
</tr>
<tr>
<td>A_MD</td>
<td>0.6930</td>
<td>0.7412</td>
</tr>
<tr>
<td>E_GN</td>
<td>0.6663</td>
<td>0.7050</td>
</tr>
<tr>
<td>E_SC</td>
<td>0.6818</td>
<td>0.7288</td>
</tr>
<tr>
<td>L_GA</td>
<td>0.7173</td>
<td>0.7789</td>
</tr>
<tr>
<td>L_MD</td>
<td>0.6799</td>
<td>0.7991</td>
</tr>
<tr>
<td>O_MD</td>
<td>0.6645</td>
<td>0.7599</td>
</tr>
</tbody>
</table>

Accuracy of 7 stores using a XGBoost Classifier.
Effectiveness of the Feature Set

- Baseline 1 (LB): By only knowing the **number of visits**
- Baseline 2: By knowing the **number of visits & date of the visit**
- By utilizing features derived from **Wi-Fi signals**, we achieved **significant** performance improvement on **revisit prediction**.

(a) On all visitors

(b) On first-time visitors
Data Collection Period

- To find the **right amount of data** to study revisit
- To maintain sensors for **securing enough profit**
  → Find the **minimum sufficient** amount of data $T$ to predict revisit without accuracy loss

### Data Length

**Short**
- Low cost & Small effort
- Less evidence
  - Changes in revisit rate

**Long**
- More evidence
- Closer to the steady state
- Capture most of revisits
- Hard to persuade clients

![Graph showing data collection period vs. frequency](image)
Impact on Prediction Accuracy

1) On all visitors
- # Regular customers ↑
- Accuracy **gradually increases.**

2) On first-time visitors
- Cover longer timeframe
- Accuracy **reaches a plateau.**
Robustness on Missing Customers

• Over 95% of the performance is maintained with a very small fraction of the dataset (e.g., 0.5% for L_MD)

(a) On all visitors

(b) On first-time visitors
## Real Behavior vs. Collected Data

**Significantly different!**

<table>
<thead>
<tr>
<th>Actual ratio</th>
<th>Wi-Fi status</th>
<th>Probability of each state</th>
<th>Observed as</th>
<th>Observed ratio (data)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group (p_y = 56%)</td>
<td>![Wi-Fi symbols]</td>
<td>p_y p_x^2</td>
<td>Group</td>
<td>![Wi-Fi symbols]</td>
</tr>
<tr>
<td>Individual (1 - p_y)</td>
<td>![Wi-Fi symbols]</td>
<td>p_y p_x (1 - p_x)</td>
<td>Individual</td>
<td>![Wi-Fi symbols]</td>
</tr>
<tr>
<td>p_y (1 - p_x) p_x</td>
<td>Individual</td>
<td>![Wi-Fi symbols]</td>
<td>![Wi-Fi symbols]</td>
<td></td>
</tr>
<tr>
<td>p_y (1 - p_x)^2</td>
<td>None</td>
<td>![Wi-Fi symbols]</td>
<td>![Wi-Fi symbols]</td>
<td></td>
</tr>
<tr>
<td>(1 - p_y) p_x</td>
<td>Individual</td>
<td>![Wi-Fi symbols]</td>
<td>![Wi-Fi symbols]</td>
<td></td>
</tr>
<tr>
<td>(1 - p_y) (1 - p_x)</td>
<td>None</td>
<td>![Wi-Fi symbols]</td>
<td>![Wi-Fi symbols]</td>
<td></td>
</tr>
</tbody>
</table>

\( p_x \) = Wi-Fi turn on rate (39.2%)
\( p_y \) = Ratio of customers with companion (Observed at the spot)
\( p_{yo} \) = Ratio of customers with companion (To be observed in the data)
Conclusions

• **Goal:** To discover the relation between *customer revisit* and their *mobility*

• **Data:**
  - Customer mobility data captured in seven stores

• **Findings:**
  - Prediction models using handcrafted features
  - Predictive powers of each feature groups
  - Performance improvement by utilizing indoor trajectories
  - Predictive powers by collecting longer period
  - Robustness on missing data
Thank you!

Scan me for details 😊
(Paper, Slides, Datasets, Tutorial)  https://github.com/kaist-dmlab/revisit