Dataset Used

A Foursquare dataset from Kaggle (2017) was used to evaluate the effectiveness of the predictive models. This dataset was collected by Yang et al (2015) and contains the Foursquare user check-in information between 04th April 2012 and 16th February 2013 in Tokyo. it summarises the information provided in the dataset.

Data	Remarks
User ID	A number which uniquely identifies a user
Venue ID	A character string which uniquely identifies a venue
Venue Category	Type of venue visited, e.g. Chinese Restaurant, Bars
Location of Visit	Latitude and Longitude coordinates of the visited venue
Time of Visit	Timestamp of check-in

Figure 1 shows the first five rows in the dataset. Each row represents a check-in event.

	userId	venueId	venueCategoryId	venueCategory	latitude	longitude	timezoneOffset	utcTimestamp
0	1541	4f0fd5a8e4b03856eeb6c8cb	4bf58dd8d48988d10c951735	Cosmetics Shop	35.705101	139.619590	540	Tue Apr 03 18:17:18 +0000 2012
1	868	4b7b884ff964a5207d662fe3	4bf58dd8d48988d1d1941735	Ramen / Noodle House	35.715581	139.800317	540	Tue Apr 03 18:22:04 +0000 2012
2	114	4c16fdda96040f477cc473a5	4d954b0ea243a5684a65b473	Convenience Store	35.714542	139.480065	540	Tue Apr 03 19:12:07 +0000 2012
3	868	4c178638c2dfc928651ea869	4bf58dd8d48988d118951735	Food & Drink Shop	35.725592	139.776633	540	Tue Apr 03 19:12:13 +0000 2012
4	1458	4f568309e4b071452e447afe	4f2a210c4b9023bd5841ed28	Housing Development	35.656083	139.734046	540	Tue Apr 03 19:18:23 +0000 2012

Figure 1 Sample Rows of Dataset

The size of the Tokyo dataset is tabulated in **Table 1**.

Type	Number
Number of Check-In Records	573703
Number of Unique Users	2293
Number of Unique Venues	61858
Number of Venue Categories	247

Table 1 Size of Tokyo Dataset

Features Engineered (Attention Model):

the research community has found that the strongest predictors of future check-in include the spatial and temporal dimensions of current and previous visits, as well as the individual user preferences. As a preparatory step to model building, relevant features has been derived from the dataset and listed in Table 2.

Attention Type	**Feature**	**Rationale**
Temporal	• **Day of Week (of current check-in)** • **Hour of Day (of current check-in)**	The regular temporal patterns exhibited by users may be useful in predicting next check-in location
Spatio-Temporal	• **User's Last Check-In Coordinates** • **Geographic Distance between Consecutive User Check-In Locations** • **User's Last Check-In Timestamp** • **Time Difference between Consecutive User Check-Ins**	a user's previous check-in has strong influence on his or her next check-in location. Note: Longitude and latitude data have been converted to 3-dimensional spherical coordinates so that the models have more accurate estimate of the geographic distance between venues
User Preference	• **Visits per User Ratio for a venue :**	This quantity aims to capture the user preference factor. If the ratio were high, venuewould have received many return visits

Table 2 Features Constructed from Historical Check-In Data

The descriptive analysis in subsequent sections will inspect which of these features have high predictive power and should be included in the predictive models.

Descriptive Analysis

Check-In Activities

Figure 2 depicts the distribution of users' active period (number of days between first and last check-in). Around 80% of the users remain active throughout the entire 10-month period. It is deduced that most users reside in Tokyo during this period and their check-in patterns are expected to exhibit certain regularities (e.g. check-in to workplace during weekdays).

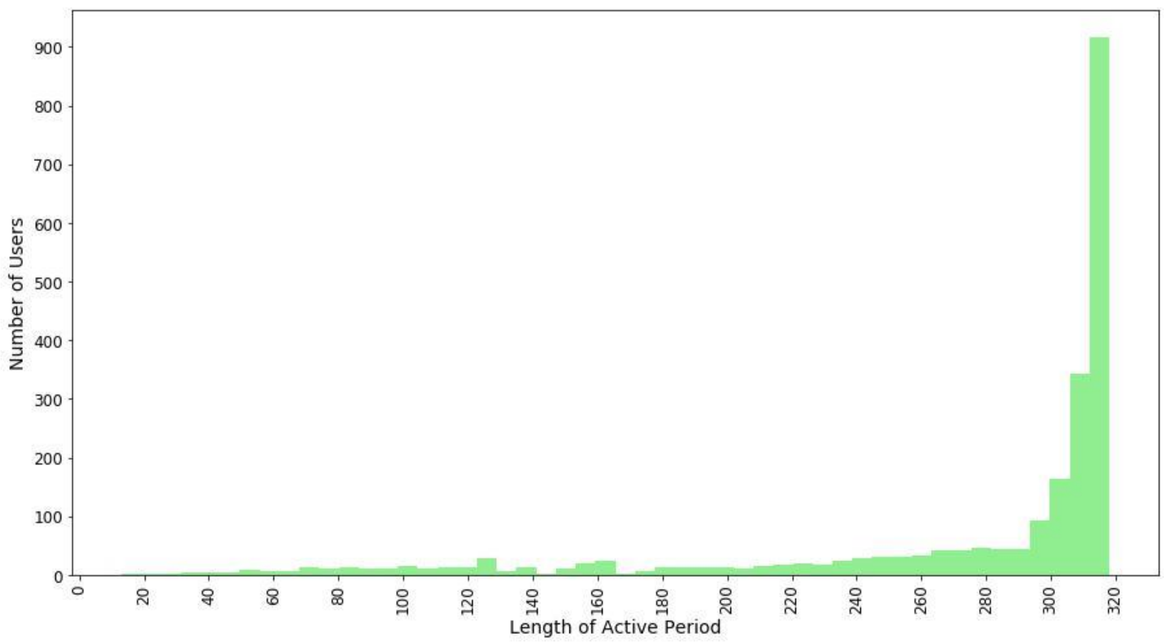

Figure 2 Distribution of Users' Active Periods

Figure 3 illustrates the distribution of the number of check-ins per user. All users have performed more than 100 check-in actions throughout the whole period and roughly 80% of them have less than 300 check-in records. The long tail shows that a small percentage of users generated a lot of check-ins. For this group of hyper active users, we can expect to observe high degree of recurring check-in patterns (e.g. check-in to homes every night).

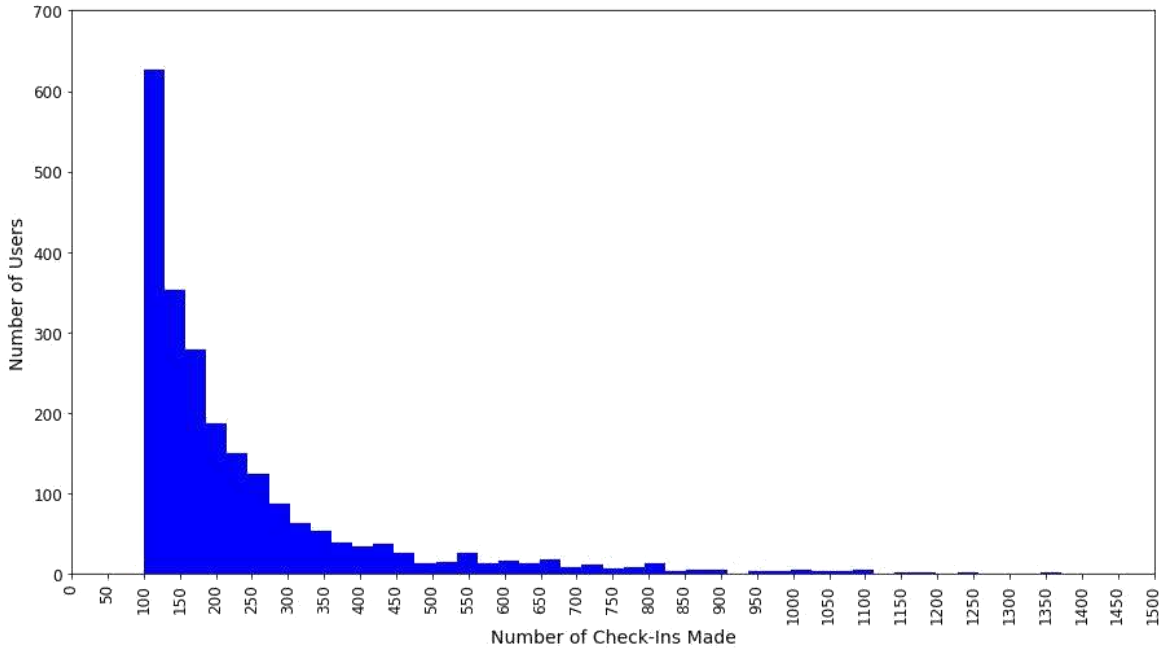

Figure 3 Distribution of Number of Check-Ins Performed by User

Periodic Check-In Patterns

Figure 4 and **Figure 5** outline the temporal trends of check-ins by day and hour respectively. Train station category is excluded from the diagrams because of its disproportionately large number of check-ins.

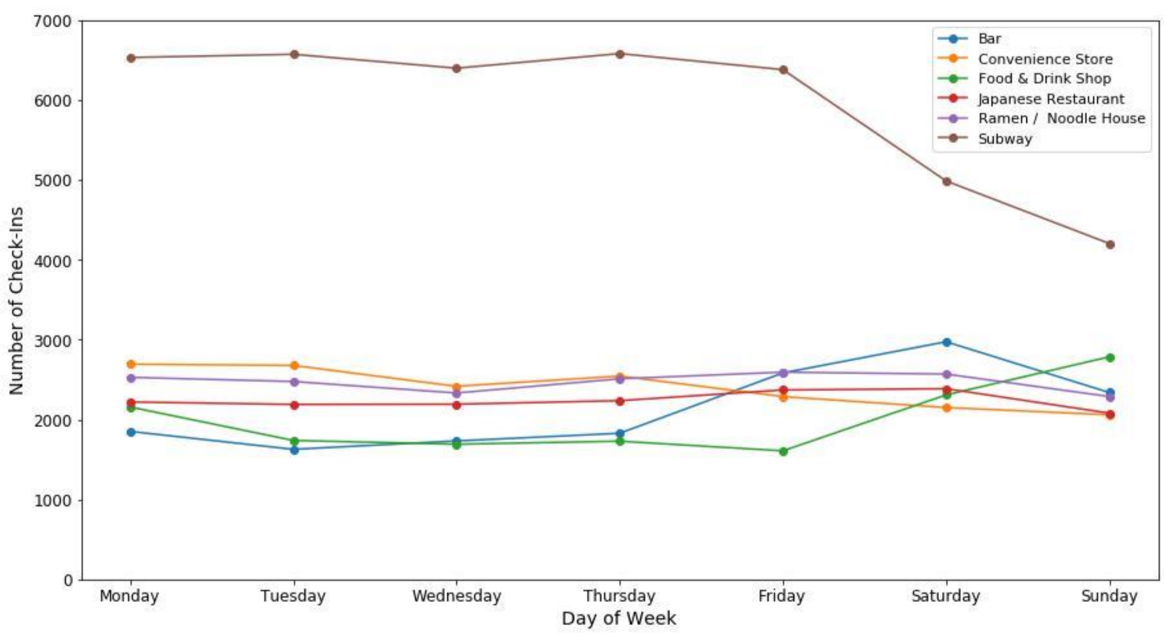

Figure 4 Check-Ins by Day (6 Most Popular Categories - Train Station Excluded)

From **Figure 4**, it can be observed that the Subway usage remained high on weekdays and dropped by 20% - 30% during weekends. Meanwhile, the number of check-ins to bars increased significantly on Friday and weekends.

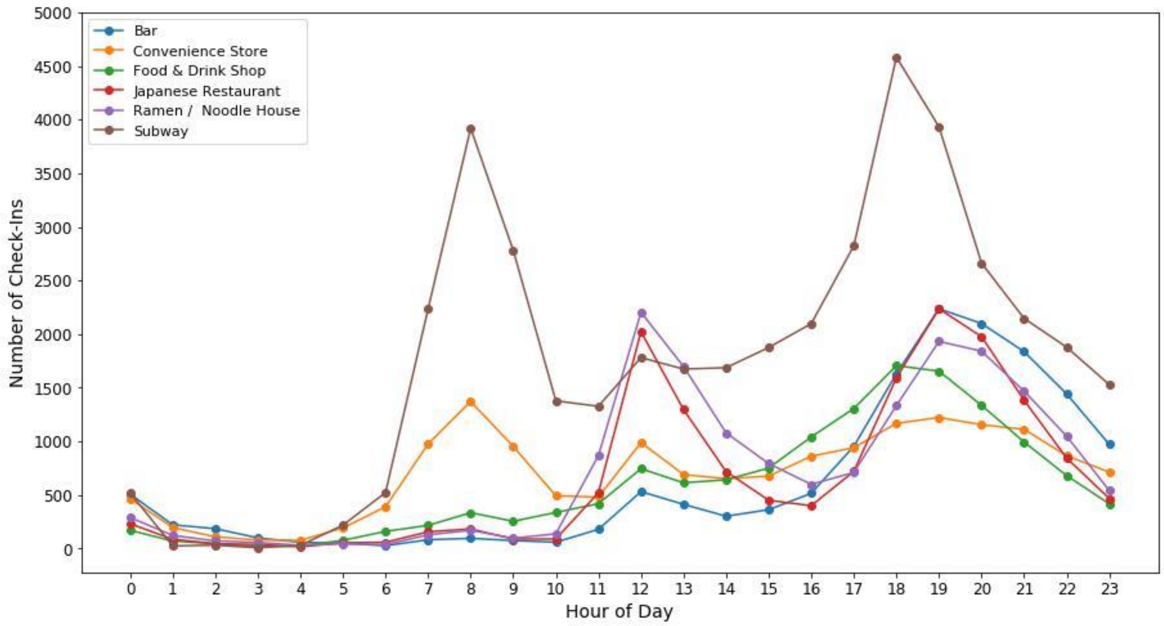

Figure 5 Check-Ins by Hour (6 Most Popular Categories - Train Station Excluded)

The daily pattern in **Figure 5** indicates that the highest Subway usage was in the morning (7AM-9AM) and evening (5PM-8PM) when users commute between workplaces and homes. At noon, the number of check-ins to “Japanese Restaurants” and “Ramen / Noodles House” surged. These periodic patterns imply that temporal features are correlated with the user check-ins and can potentially be good predictors.

Distance and Time Difference between Consecutive Check-Ins

The distribution of geographic distance between consecutive check-ins **(Figure 6**) shows that almost 50% of the check-ins were within 1 kilometre of previous check-in location. The number of check-ins drops exponentially as distance increases.

The distribution of time difference between consecutive check-ins (**Figure 7**) also exhibits similar behaviour where numerous check-ins were performed within 30 minutes of previous check-in. These observations suggest that the previous check-in location could be a high quality predictor since a user could not possibly travel long distance within a short timeframe.

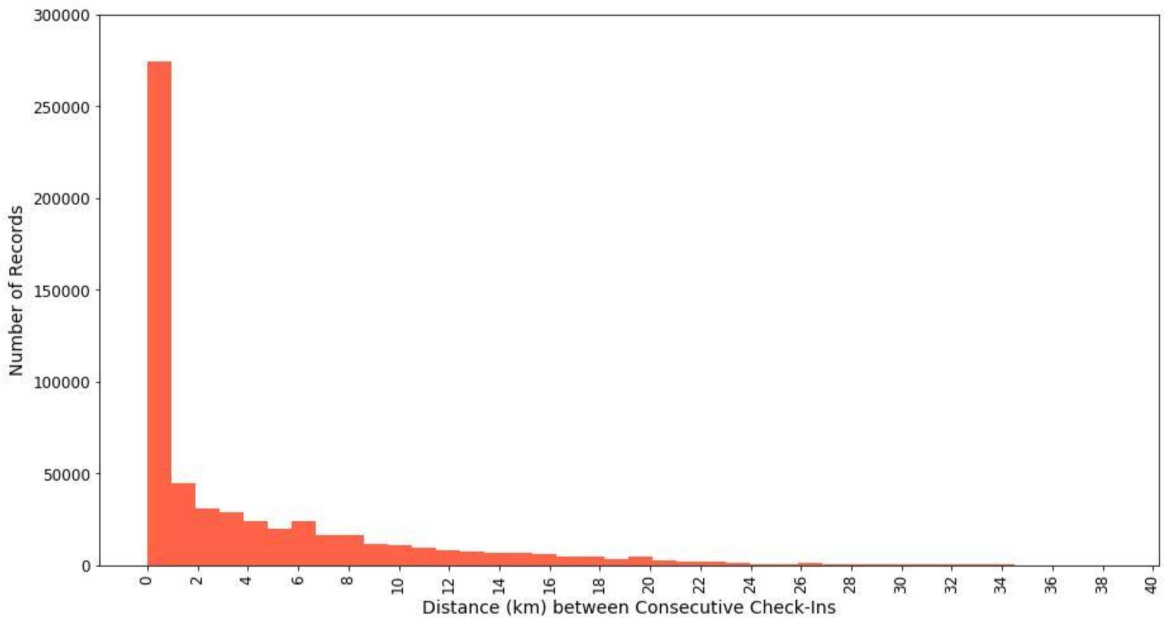

Figure 6 Distribution of Distance between Consecutive Check-In Locations

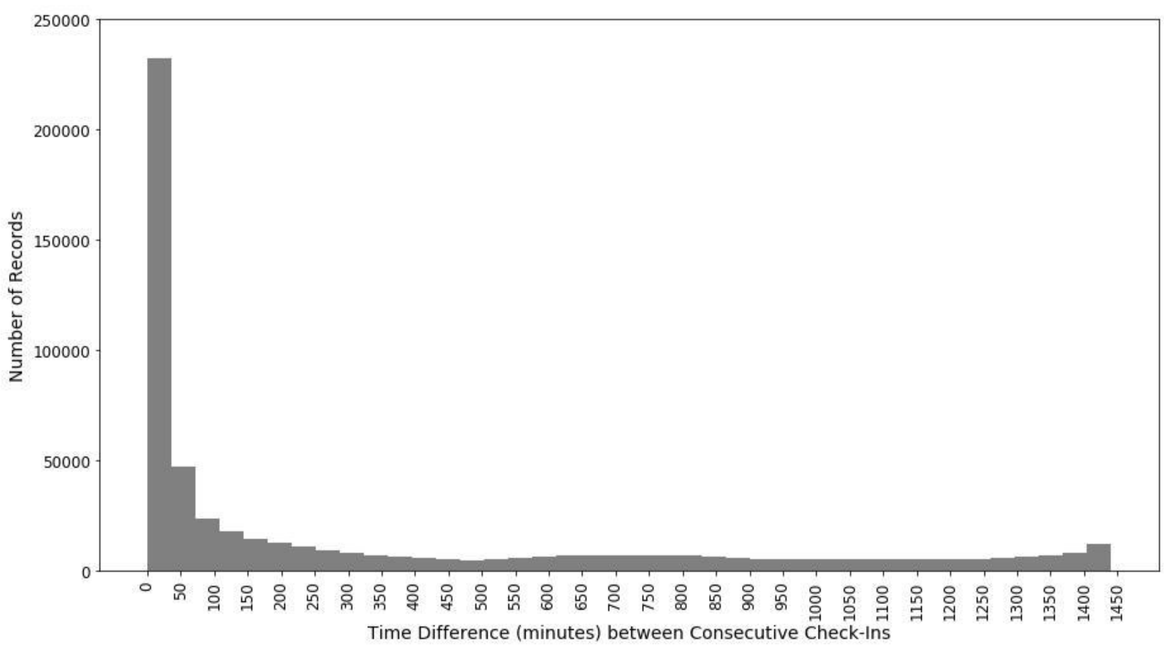

Figure 7 Distribution of Time Difference between Consecutive Check-In

Number of Check-Ins Received by a Venue

The histogram in **Figure 8** (y-axis in log scale) specifies that extremely high number of venues were visited very infrequently. More than 60,000 venues were visited less than 250 times while only a few popular venues accumulated more than 2,000 user check-ins.

Based on this observation, it can be deduced that users were prone to return to the highly popular venues. Thus, a naïve algorithm, which always recommends the most popular venues, can be used as baseline algorithm and compared against other predictive models for performance evaluation purpose.

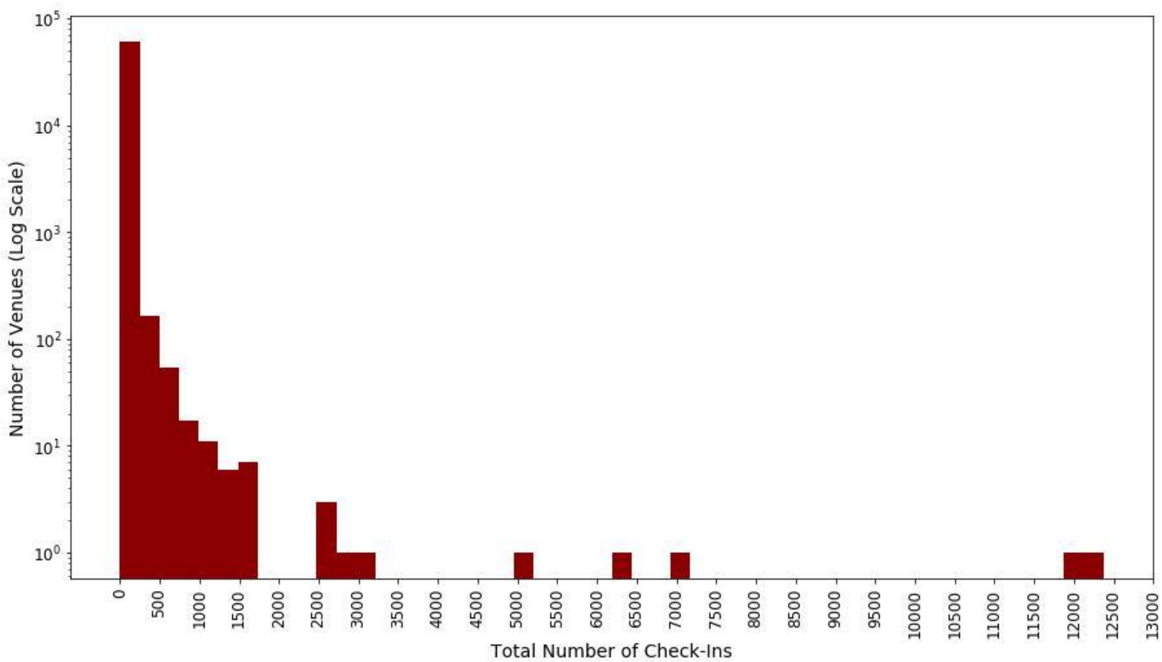

Figure 8 Distribution of Total Number of Check-Ins per Venue

Visits per user ratio (defined in **Table 4**) quantify the returning frequency and its distribution in **Figure 9** (y-axis in log scale) demonstrates that users frequently revisited their favourite venues. Although a dominant percentage of the venues had less than 3 average visits per user, the fat tail signals that a sizeable number of venues were very frequently revisited by the same users.

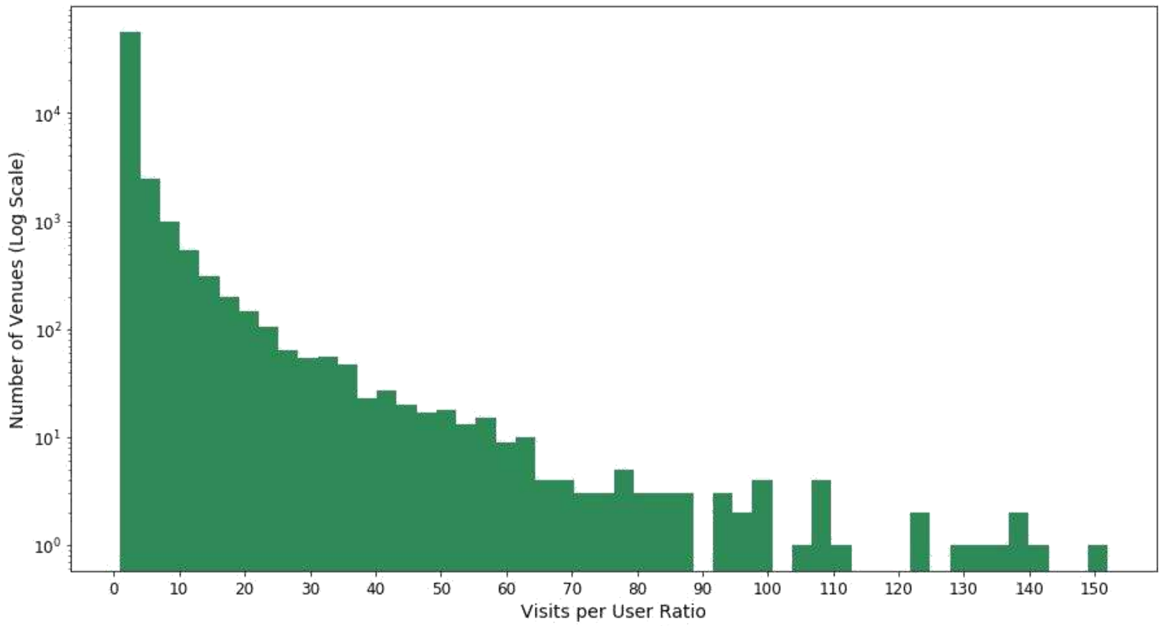

Figure 9 Distribution of Average Number of Visits per User

Distinct User Preference

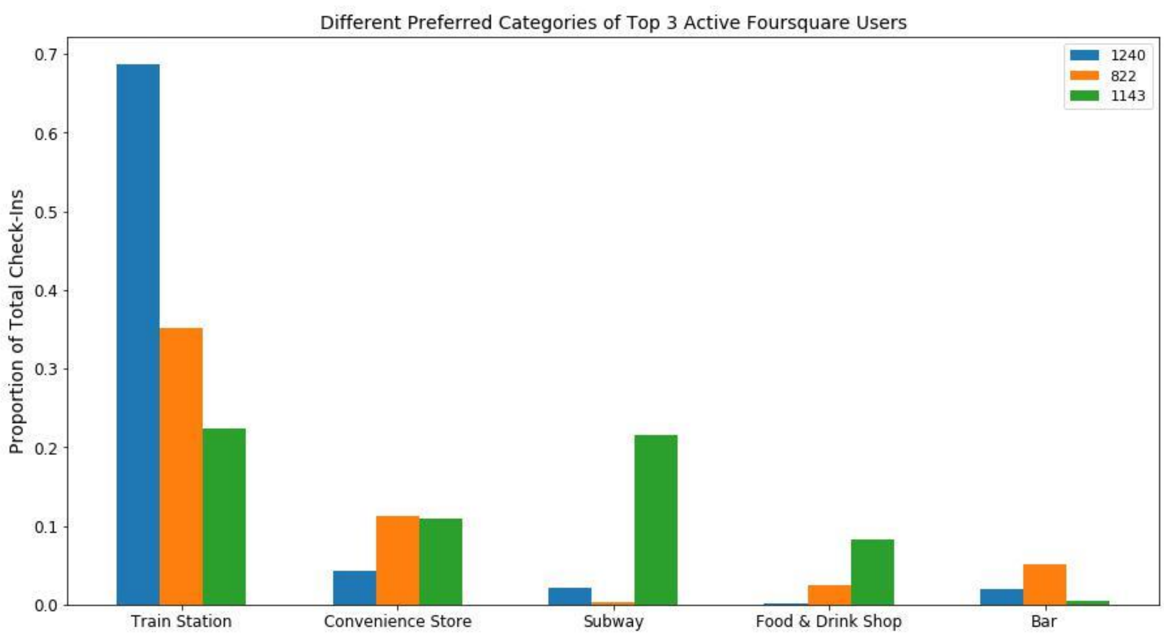

Figure 10 Preferred Venue Categories of Top 3 Most Active Users

Figure 10 depicts the three most active users' check-in frequencies to five different venue categories. User 1240 was a heavy train user (nearly 70% of his or her check-ins were at Train Stations) and hardly used subway. The other user 1143 took both subway and train regularly and visited "Food & Drink Shop" more frequently than the other two. Clearly, each user displayed distinct check-in preferences and individual user preference could be highly predictive of the next check-in venue.

Methodology

Data Preparation

Table 3 lists the categories with most number of unpopular venues (venue with less than 100 check-ins). Most of them were from food industry.

Category	Number of Unpopular Venues
Japanese Restaurant	5511
Bar	4002
Ramen / Noodle House	3599
Convenience Store	3138
Café	2182

Table 3 Five Categories with Most Unpopular Venues

The two most likely reasons behind this low number of visits are:

- Visitors did not like the venue after their initial visits and hence hardly returned
- The venue was not as popular or had received mediocre ratings by visitors, thus it could not attract many new visitors

Since these venues were likely to be low quality recommendations, their corresponding check-in records should be removed from the training set so that the predictive models would not recommend them. Similarly, their check-in records were removed from test set too.

Data Splitting

Before building the predictive models, a separate training set and test set are required. Since the dataset is time-based, the data must be split by time to ensure that the model is trained using historical data and tested using unseen future data, in order to provide an unbiased estimate of the model performance.

In this report, 80%-20% train-test split ratio was used

Next Check-In Prediction Problem

This prediction problem can be formulated as a multiclass classification problem, summarised in **Table 4**.

Item	Description
Prediction Problem	Given a user u accessing the mobile app at time t_k, predict the venue $v_{(u,k)}$ which u is going to check-in
Target Variable	$v_{(u,k)}$ **:** next check-in venue of u

Predictor Variables	• $t_{(u,k)}$**:** Time of user accessing the app, further split into:
	  ○ Day of week
	  ○ Hour of day
	• $v\_loc_{(u,k-1)}$**:** Geographic coordinates of the user's <u>previous</u> check-in venue
	• $t_{(u,k-1)}$**:** Timestamp of user's <u>previous</u> check-in
	<u>Note</u>
	1. Geographic coordinates, $v\_loc_{(u,k)}$ of the user's <u>next</u> check-in venue must not be used as predictors because they are characteristics of target variable $v_{(u,k)}$

	2. If previous check-in is not available, values will be imputed for ***v_loc**$_{(u,k-1)}$* and ***t**$_{(u,k-1)}$*

Table 4 Next Check-In Prediction Problem

Predictive Models

Four machine-learning based multiclass classifiers were selected as the predictive models for the next check-in prediction problem:

- Decision Tree
- Random Forest
- Gaussian Naïve Bayes
- Artificial Neural Network (MLP)

Evaluation Approach

Evaluation Metric

For each prediction task, the classifier returns a list of *N* venues with the highest estimated probabilities. The prediction is deemed as success if the actual check-in venue is within the list. The final score *Accuracy@N* is the ratio of number of successful predictions to total number of prediction tasks. **the *Accuracy@N* scores will be reported for *N = {3, 5, 15, 30}*.**

Model Evaluation for Cold-Start Users

One key challenge of next check-in prediction problem is to accurately predict the next check-in locations for cold-start users .Cold-start users are the users with few check-in records. Due to little historical information, the prediction accuracy for this group of users is likely to be lower since the predictive model would heavily rely on other users' behavioural data in estimating cold-start users' next check-in venues.

Taking this difficulty into consideration, this report will assess the model performance for cold-start and non-cold users.

Results & Discussion

Besides the test prediction accuracies of the classifiers, the predictive powers of different features are also evaluated in this chapter. The classifiers were trained using different set of features and the results are presented in separate sections.

Temporal-Based Attention Model

Variables Used

The first model to be tested is temporal based model. The test accuracy of this model will determine whether the temporal patterns observed.

Item	Description
Target Variable	$v_{(u,k)}$: next check-in venue of u
Predictor Variables	$t_{(u,k)}$: Time of user accessing the app, further split into: • Day of week • Hour of day

Table 5 Target and Predictor Variables in Temporal-Based Model

Test Results and Discussion

The prediction accuracies over test set are depicted in **Figure 11**. Accuracies for cold-start users and other users are reported in two separate graphs. It is immediately obvious that none of the classifiers was able to comfortably outperform baseline algorithm (i.e. Global Popularity Method). All classifiers achieved relatively low accuracies. At *N = 30* (top-30 most probable locations were returned), only around 30%-35% of the prediction tasks were classified as success for “Other Users” group.

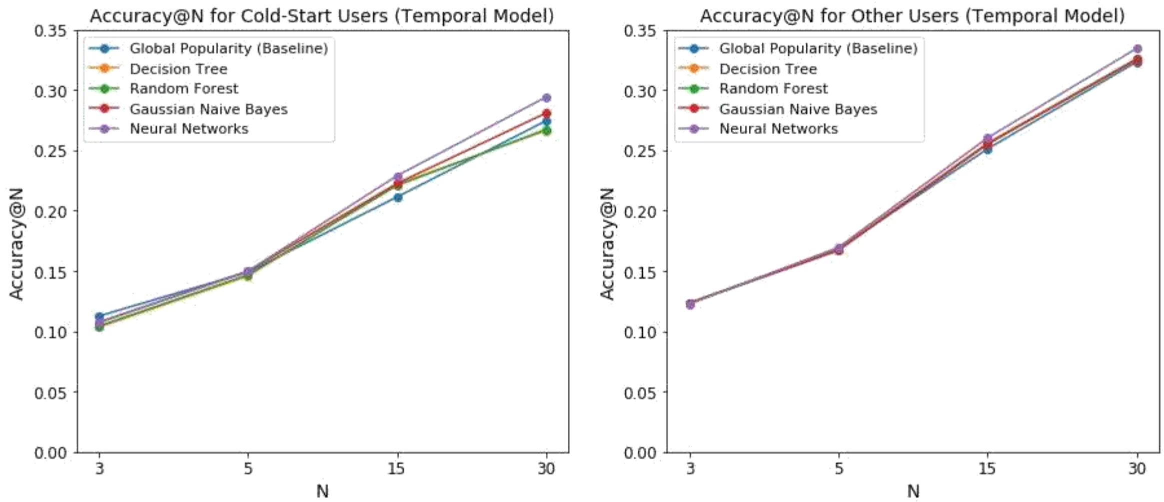

Figure 11 Test Set Accuracy@N of Temporal-Based Models

It can be deduced that temporal information does not yield much predictive power by itself, in predicting check-in venue. Temporal information might give us some higher-level insights, such as users tend to go to food places at noon time (**Figure 5**). However, it has difficulties in predicting lower-level details such as the exact check-in location.

To illustrate this, **Figure 12** shows the geographical locations of the Japanese Restaurants with more than 250 check-ins in the training set. They are scattered around the whole city of Tokyo. Even though the classifiers were able to learn from training data that user tends to go to Japanese Restaurants at noon time, there were simply too many candidates for the classifiers

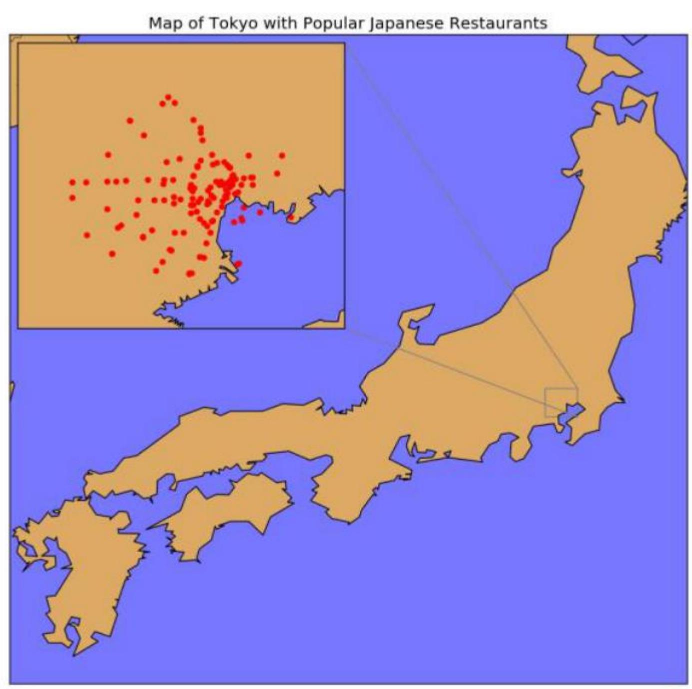

Figure 12 Locations of Popular Japanese Restaurants

to pinpoint the exact venue.

However, the predictive power of temporal features can be enhanced when coupled with other information (such as location and demographics). Next, we shall examine how temporal and spatial features can be combined to make predictions on next check-in venue.

Saptio-Temporal Attention Model

Variables Used

a user's historical check-in influences his or her next check-in location. This effect was found to be short-term and hence the most recent check-in location should have the strongest influence over the next check-in venue.

Based on the above findings, a Spatio-Temporal model is constructed where the user's previous check-in location is used as predictor variable (alongside temporal features). The list of predictor variables used is tabulated in **Table 6**.

Item	Description
Target Variable	$v_{(u,k)}$: next check-in venue of ***u***
Predictor Variables	• $t_{(u,k)}$: Time of user accessing the app, further split into: ○ Day of week ○ Hour of day • $v\_loc_{(u,k-1)}$: Geographic coordinates of the user's previous check-in venue • $t_{(u,k)} - t_{(u,k-1)}$: Time difference (in seconds) between the user's previous check-in and time of accessing the app

Table 6 Target and Predictor Variables in Spatio-Temporal Model

Test Results and Discussion

In **Figure 13**, the most striking difference compared with the temporal models is that the Spatio-Temporal models achieved much higher prediction accuracy, across all types of classifiers. The user's previous check in details (i.e. geographical location and time of previous check-in) help the classification algorithms in narrowing down to a list of more probable venues, given the observation that the next check-in venue tends to be in close proximity to the previous one (**Figure 6**).

We can also witness a significant improvement over the baseline model, even for the cold-start users with no more than 10 historical check-in records. This indicates that even though a user may not have rich history of check-ins, his or her previous check-in location is still highly predictive of the next check-in venue. In other words, the next check-in venue of a "Cold-Start User" can also be predicted with decent accuracy given his or her previously checked-in venue.

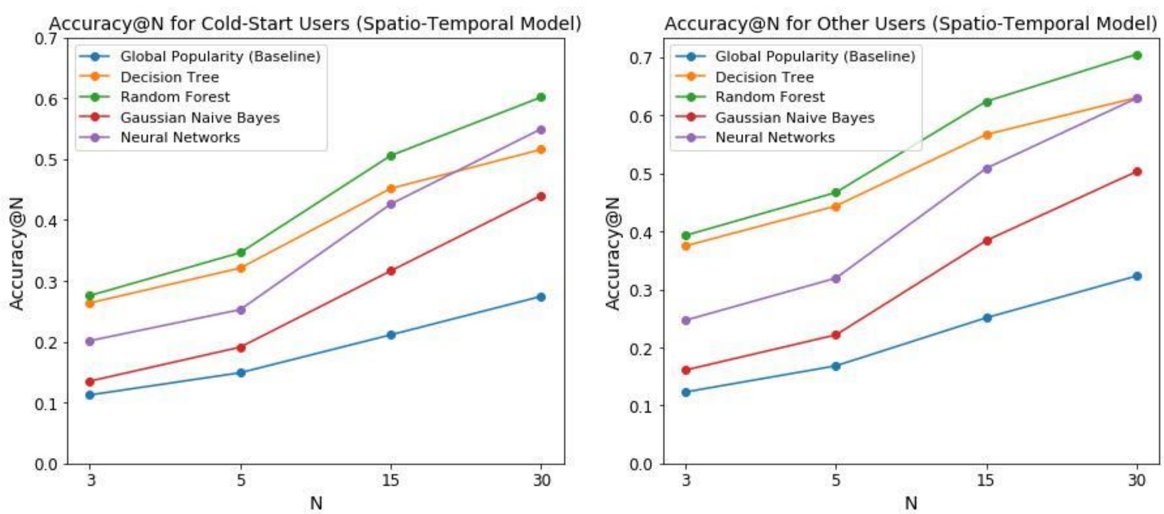

Figure 13 Test Set Accuracy@N of Spatio-Temporal Models

Lastly, all the classifiers outperformed the baseline model by comfortable margin. Relative to other classifiers, Gaussian Naïve Bayes classifier was the worst-performing classifier. This could be attributed to the conditional independence assumptions of Naïve Bayes classifier. This assumption does not hold in our selected features. Conditioning on current check-in venue, the time of visit is not entirely independent of the user’s previous location. For example, in the scenario where a user checks-in to a Sushi restaurant at Ginza, the knowledge of the time of visit affects the probability distribution of the user’s previous check-in location. If the time of visit is weekday evening, the user is more likely to be coming from his workplace. If the time of visit is Saturday noon, the user is more likely to be travelling from home. The violation of this assumption resulted in a biased estimate; hence the prediction accuracy of Naïve Bayes is relatively lower.

The non-parametric classifiers, such as Random Forest and Decision Tree, performed the best out of the classifiers tested. These models were able to learn the decision boundary using training data and they generalised well to the unseen data in test set. With the use of bootstrap aggregation, Random Forest was able to outperform a single Decision Tree across all *N values*. Compared with these two models, the accuracy of Neural Network was lower, particularly at lower *N*.

User preference attention Model

Thus far the predictive models were based on temporal and spatial features and did not consider the individual user preference. We know from **Figure 10** that each user had distinct check-in preference and **Figure 9** tells us that some popular venues enjoyed high number of repeat visits. All these signify that a user might have a unique preference and tend to regularly check-in to his or her favourite venue.

In order to leverage on user preference information, a “Fusion Model” is proposed. Fusion model considers the individual user’s venue preference and adjust the probability estimate accordingly.

User Preference Score

Fusion Model considers user check-in preference by computing the users’ historical visit frequency to a particular venue. With the training data, the user-venue preference matrix (**Table 07**) is built.

	v_1	v_2	v_3	v_4	v_5	v_6	…	v_n
u_1	$P(v_1\|u_1)$	$P(v_2\|u_1)$	$P(v_3\|u_1)$	$P(v_4\|u_1)$	$P(v_5\|u_1)$	$P(v_6\|u_1)$	…	$P(v_n\|u_1)$
u_2	$P(v_1\|u_2)$	$P(v_2\|u_2)$	$P(v_3\|u_2)$	$P(v_4\|u_2)$	$P(v_5\|u_2)$	$P(v_6\|u_2)$	…	$P(v_n\|u_2)$
u_3	$P(v_1\|u_3)$	$P(v_2\|u_3)$	$P(v_3\|u_3)$	$P(v_4\|u_3)$	$P(v_5\|u_3)$	$P(v_6\|u_3)$	…	$P(v_n\|u_3)$
…	…	…	…	…	…	…	…	…
u_m	$P(v_1\|u_m)$	$P(v_2\|u_m)$	$P(v_3\|u_m)$	$P(v_4\|u_m)$	$P(v_5\|u_m)$	$P(v_6\|u_m)$	…	$P(v_n\|u_m)$

Table 7 User-Venue Preference Matrix

Final Probability Estimate

The preference scores obtained from the user-venue preference matrix are used to adjust the probabilities estimated by Spatio-Temporal model. Similar to the approach adopted by Zhang & Chow (2013) in fusing location rating with probability of location, the probability estimated by Spatio-Temporal model can be fused with the user preference estimate using product rule.

Test Results and Discussion

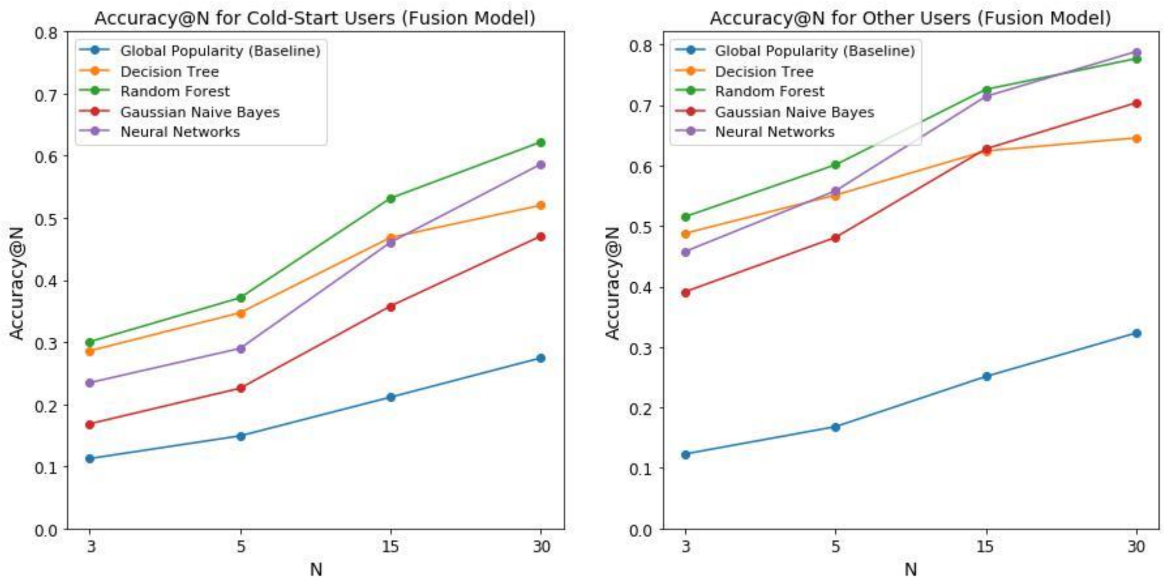

Figure 14 Test Set Accuracy@N of Fusion Models

For "Cold-Start Users", the test accuracies of the fusion models (left diagram of **Figure 14**) were almost identical to the Spatio-Temporal models. This is expected because these users had not checked-in many venues yet, hence the user-venue preference matrix might not accurately reflect their true preferences. Consequently, the predictions produced by fusion models might not consist of the places that they would revisit.

On the other hand, there is a noticeable increase in prediction accuracies across all *N* values for “Other Users” group. At *N = 3*, the *Accuracy@3* for Random Forest has increased from 0.37 to almost 0.50 (~35% increase). For this group of users, the fusion model was able to gauge the user’s preferred venues more accurately and make adjustment accordingly, resulting in more accurate predictions.

Conclusion

Temporal models, which utilised time of app usage information, did not significantly outperform the baseline model which always suggests the most popular venues of all times. This implies that the time of app usage alone is not helpful in predicting the check-in venue.

Next, the Spatio-Temporal models considered user's previous check-in details and were able to predict the next check-in venue with decent accuracy. Even for "Cold-Start" users who have few check-in records, the Random Forest based model outperformed the baseline model by more than 100% in terms of prediction accuracy.

Finally, in addition to the spatio-temporal features, the fusion models also factored in individual user preference. This further improved the prediction accuracy (for "Other Users" group) by around 10-20%.