Imperial College Business School

Improving Digital Advertising by Predicting Customer Check-In using Location Data

Siow Meng Low
(CID: 01255248)

Total Words Count: 5,427

Abstract

The data explosion in recent times has enabled digital marketers to perform targeted marketing. By analysing user's behavioural data, companies are able to identify who are receptive to a certain type of products or services. Inspired by the development in this area, this report studies how could the data from Location Based Social Networking (LBSN) sites, such as Foursquare, be used to achieve highly targeted advertising.

This report examines a Foursquare dataset and uses the historical check-in data to predict a user's next location. By knowing the user's next location, targeted ads and sales coupons could be sent to the user. This provides a better user experience and helps retailers achieve higher revenue.

The test results of the predictive models show that a user's most recent check-in location is highly predictive of his or her next check-in venue. Traditional machine-learning based classifiers (such as Random Forest) could achieve decent prediction accuracy. Based on these findings, LBSN service providers are encouraged to employ analytics to provide highly targeted advertisements.

Table of Contents

Abstract 1

1 Introduction 6

2 Literature Review 8

2.1 Temporal Features 8

2.2 Spatial Features 9

2.3 Spatial-Temporal Combination 9

2.4 User Preferences 10

2.5 Social Networks 11

2.6 Others 11

3 Understanding LBSN User Behaviours 12

3.1 Dataset Used 12

3.2 Features Engineered 13

3.3 Descriptive Analysis 14

3.3.1 Check-In Activities 14

3.3.2 Periodic Check-In Patterns 15

3.3.3 Distance and Time Difference between Consecutive Check-Ins 16

3.3.4 Number of Check-Ins Received by a Venue 18

3.3.5 Distinct User Preference 19

4 Methodology 21

4.1 Data Preparation 21

4.2 Data Splitting 21

4.3 Next Check-In Prediction Problem 22

4.4 Predictive Models 23

4.5 Evaluation Approach 24

4.5.1 Evaluation Metric 24

4.5.2 Model Evaluation for Cold-Start Users 24

5 Results & Discussion 25

5.1 Temporal-Based Model 25

5.1.1 Variables Used 25

5.1.2 Test Results and Discussion 25

5.2 Spatio-Temporal Model 27
5.2.1 Variables Used 27
5.2.2 Test Results and Discussion 28
5.3 Fusion Model 29
5.3.1 User Preference Score 29
5.3.2 Final Probability Estimate 30
5.3.3 Test Results and Discussion 31
6 Conclusion & Recommendation 33
References 35
Appendix 1 Conversion of Geographic Coordinates 40
Appendix 2 Imputed Values 41
Appendix 3 Multiclass Classifiers Evaluated 42
Appendix 4 Time-Series Cross Validation 43

Table of Figures

Figure 1 Sample Rows of Dataset 12
Figure 2 Distribution of Users' Active Periods 14
Figure 3 Distribution of Number of Check-Ins Performed by User 15
Figure 4 Check-Ins by Day (6 Most Popular Categories - Train Station Excluded) 15
Figure 5 Check-Ins by Hour (6 Most Popular Categories - Train Station Excluded) 16
Figure 6 Distribution of Distance between Consecutive Check-In Locations 17
Figure 7 Distribution of Time Difference between Consecutive Check-In 17
Figure 8 Distribution of Total Number of Check-Ins per Venue 18
Figure 9 Distribution of Average Number of Visits per User 19
Figure 10 Preferred Venue Categories of Top 3 Most Active Users 19
Figure 11 Test Set Accuracy@N of Temporal-Based Models 26
Figure 12 Locations of Popular Japanese Restaurants 26
Figure 13 Test Set Accuracy@N of Spatio-Temporal Models 28
Figure 14 Test Set Accuracy@N of Fusion Models 31
Figure 15 Time Series Cross Validation (4-Split) 43

Table of Tables

Table 1 Predictors Proposed by Researchers – LBSN Next Check-In Problem 8
Table 2 Data Contained in Foursquare Dataset 12
Table 3 Size of Tokyo Dataset 12
Table 4 Features Constructed from Historical Check-In Data 13
Table 5 Five Categories with Most Unpopular Venues 21
Table 6 Summary of Training & Test Sets 22
Table 7 Next Check-In Prediction Problem 23
Table 8 Target and Predictor Variables in Temporal-Based Model 25
Table 9 Target and Predictor Variables in Spatio-Temporal Model 27
Table 10 User-Venue Preference Matrix 30
Table 11 Multiclass Classifiers Evaluated 42

1 Introduction

With the proliferation of mobile devices and the widespread civil adoption of Global Positioning System (GPS) technology, vast amount of consumer location data has become available to social media companies. This has enabled them to “identify customers’ real-time intent” and deliver targeted advertisements to the users on their mobile platforms, greatly enhancing the effectiveness of the advertisements (Forbes Corporate Communications, 2016).

To tap into this new location-based advertising and marketing (LBA) market, a number of mobile platforms have been launched by Location-Based Social Networking (LBSN) sites such as Foursquare and Yelp since 2009. The global LBA market is forecast to continue growing at a compound annual growth rate of 54% and reach US$14.8 billion in 2018 (Eddy, 2014). With this huge amount of business opportunities in the LBA market, traditional key players in the digital advertising space, such as Facebook and Google, have also started offering similar services and showing location-based ads to their users in recent years.

In addition to providing location-based recommendations and advertisements, these mobile platforms also allow users in checking-in specific businesses (e.g. restaurants, attractions) and review their visits, so that other users will be able to read these reviews and discover new place of interest. As a result, these LBSN sites have gathered huge amount of users’ historical check-in data, which includes the temporal (time of visit), spatial (location of visit) information and the types of business (e.g. restaurant, bar, retailers) that a user has checked-in.

It is crucial for the LBSN providers to display the ads or promotional messages to the users who are most likely to action on them since many of the providers, for instance Foursquare and Facebook, charge the advertisers on per action basis (Foursquare, 2017; Facebook, 2017). One of the ways to improve the advertising effectiveness is to make use of the collected users’ historical check-in data to predict their possible future check-in locations. By knowing a user’s places of

interest, a more personalised and targeted advertisement can be displayed to him or her.

With the above in mind, this report aims to investigate the features that are most predictive of a user's future check-in and recommend the best analytics strategy in using historical check-in data to improve the effectiveness of location-based advertising. **Chapter 2** of this report summarises the existing studies on location prediction using user check-in data. A Foursquare dataset, containing user check-in records in the city of Tokyo (from 04 April 2012 to 16 February 2013), has been analysed and **Chapter 3** provides a brief descriptive analysis of the dataset. **Chapter 4** outlines the test methodology and results are presented in **Chapter 5**. Lastly, **Chapter 6** concludes by reiterating the recommended strategy and summarises the possible future works that could be done in this area.

2 Literature Review

Location recommendation using LBSN check-in data has attracted the attentions of the academic research community and a large number of studies have been performed in the past six years. Various set of features have been analysed and empirically tested to be quality predictors of future check-ins. These features are listed in **Table 1**.

Types	Examples

Temporal	• Time of Visit:
	  ○ Day of Week
	  ○ Time of Day

Spatial	• Location of Visit • Distance between consecutive visits
User Preferences	• Category of venues frequently visited by users (e.g. Chinese restaurants, bars)
Social Network	• Network of friends in the LBSN • Venues visited by a user's circle of friends

Table 1 Predictors Proposed by Researchers – LBSN Next Check-In Problem

2.1 Temporal Features

Based on the previous studies where users exhibit distinctly different check-in preferences at different time of the day, Gao et al (2013) proposed a matrix factorisation model in learning the temporal preferences of users. It was shown that this approach is able to achieve better recommendation performance than user-based Collaborative Filtering (CF) approach, which does not consider temporal effects.

In a separate study, Preoţiuc-Pietro & Cohn (2013) noticed that LBSN users tend to have a periodic pattern in check-in behaviours. For instance, users check-in their workplaces in weekday mornings while on Saturdays, many of the check-ins were in "Shop & Services" category.

All these findings synchronise with our intuitive understanding of human behaviours. Most people have predictable timetables (e.g. work on weekdays, go to bars at Friday nights). This pattern allows the predictive model to learn the users' individual timetables and narrow down to a smaller list of possible places of interest that could be recommended to users on different day and at different time of day.

2.2 Spatial Features

Besides the temporal patterns, spatial features also serve as important predictors of a user's next check-in. Noulas et al (2012) examined a number of mobility and temporal features. It was found that the place a user has visited in the past and its geographic distance to other check-in venues is also crucial to the prediction of next check-in location.

Manotumruksa (2015) also found strong correlations between a user's successive check-ins. Within short period of time (particularly when next check-in is within one day from the previous check-in), the next check-in location is likely to be in close proximity to the previous check-in.

In their proposed Naïve Bayesian model to estimate the probability of next check-in location, Ye et al (2011) considered the distances from all previously visited venues. Compared to other recommendation algorithms which do not consider spatial information, this model is able to more accurately predict the next venue a user will check-in next.

2.3 Spatial-Temporal Combination

Given the importance of spatial and temporal features in predicting the next check-in location, a number of researchers have proposed different predictive models in combining these features:

- Gao, Tang & Liu (2012) adopted a Bayesian approach in estimating the probability of next check-in location. They considered the "spatial prior" (the

probability of next check-in location given the location of previous check-in) and temporal effects (i.e. day of week and hour of day) in their model

- Liu et al (2016) trained a Recurrent Neural Network using two additional features: time intervals between check-ins and distance from previous check-in locations

All the models above were shown to yield better predictive performance than pure-temporal or pure-spatial approach.

2.4 User Preferences

Another group of researchers focused on recommending new locations (previously unvisited by the user) and they used recommender system approach for this purpose. For example, Berjani & Strufe (2011) proposed a recommender which uses Regularised Matrix Factorisation technique to estimate user preferences over a set of previously unvisited venues.

Bao, Zheng & Mokbel (2012) made use of user-based collaborative filtering technique to predict user receptiveness to a new location. They introduced a tweak to the algorithm where only the “local experts” who have similar preferences (e.g. someone who frequently visited Chinese restaurants in New York City) are used for the computation of location rating scores.

The recommender system approach does well in recommending novel locations to users but it does not predict the locations a user might make a return visit for. To handle both objectives, Lian et al (2015) utilised a supervised learning model to estimate whether a user’s next check-in is going to be novel (i.e. new locations) or regular (i.e. previously visited locations). Based on this estimated probability, a hybrid model, which uses classification and recommender algorithms, produces a list of both new and regular location recommendations to users.

Similar to the approach put forth by Lian et al (2015), Wang et al (2015) utilised parametric models to compute the probability of visits to new and regular locations.

A weighted objective function was then used to fine-tune the optimal trade-off between recommending new or old locations.

2.5 Social Networks

Using a user's social networking data to predict his or her next check-in location is still an active research area (Gao & Liu, 2014). Gao & Liu (2014) demonstrated that a model, which considers both historical check-ins and social networking information, is able to outperform a historical model (which bases its prediction entirely on the user's historical check-in data) in terms of prediction accuracy.

Inspired by Google's PageRank algorithm, Wang, Terrovitis & Mamoulis (2013) proposed a personalised PageRank (PPR) algorithm to model the social influence on users' check-ins. It uses recursive algorithm to identify important friends who in turn have many important friends. The venues frequently visited by important friends are recommended to the user.

Having observed that a large percentage of user check-ins are of periodic nature (e.g. commute to work every weekday, grocery shopping every Saturday noon), Cho, Myers, & Leskovec (2011) proposed a parametric model: Periodic & Social Mobility Model (PSMM). PSMM computes the probability of periodic and socially influenced check-ins and uses them to predict the most probable next check-in venues.

2.6 Others

The features mentioned earlier are the important predictors of future user check-ins, as investigated intensively by the research community. In addition to these, Lian et al (2014) discovered that user demographics information, such as gender and age, is also correlated with location predictability. LBSN service providers could also potentially make use of this information to predict next user check-in location.

3 Understanding LBSN User Behaviours

3.1 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. **Table 2** 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

Table 2 Data Contained in Foursquare Dataset

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 3**.

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

Table 3 Size of Tokyo Dataset

3.2 Features Engineered

As discussed in **Chapter 2**, 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 4.**

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**	Based on the findings of Gao & Liu (2014), 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 (**Appendix 1**) so that the models have more accurate estimate of the geographic distance between venues
User Preference	• **Visits per User Ratio for a venue v:** $\frac{Total\ Number\ of\ Check-Ins\ to\ v}{Number\ of\ Unique\ Users\ Checked-In\ to\ v}$	This quantity aims to capture the user preference factor. If the ratio were high, venue v would have received many return visits

Table 4 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.

3.3 Descriptive Analysis

3.3.1 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).

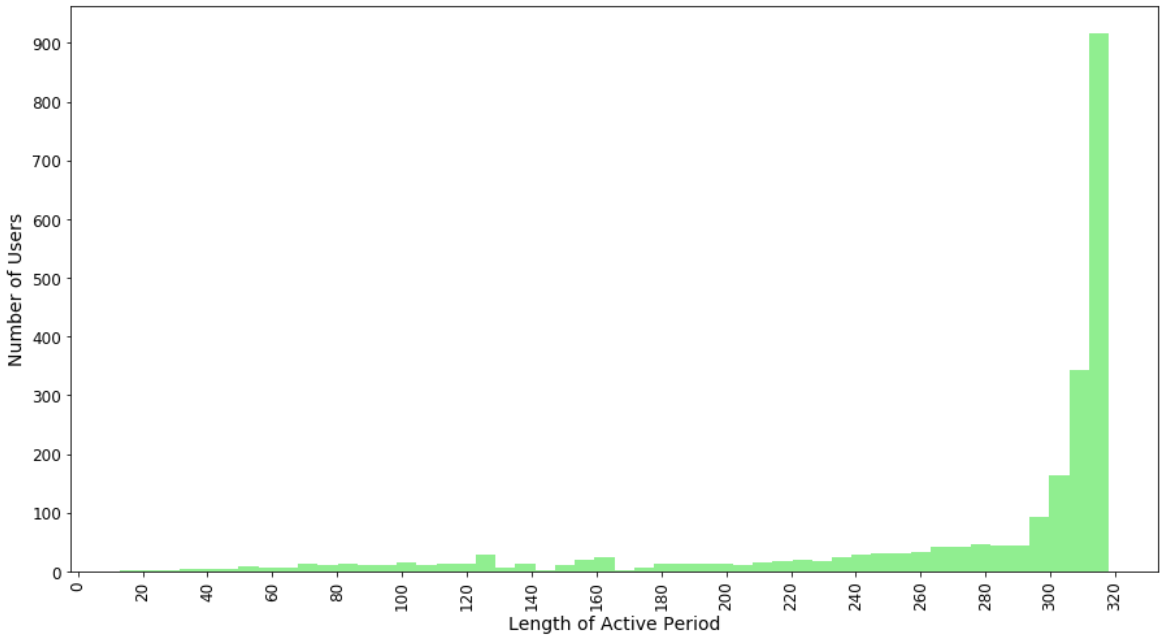

Figure 2 Distribution of Users' Active Periods

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).

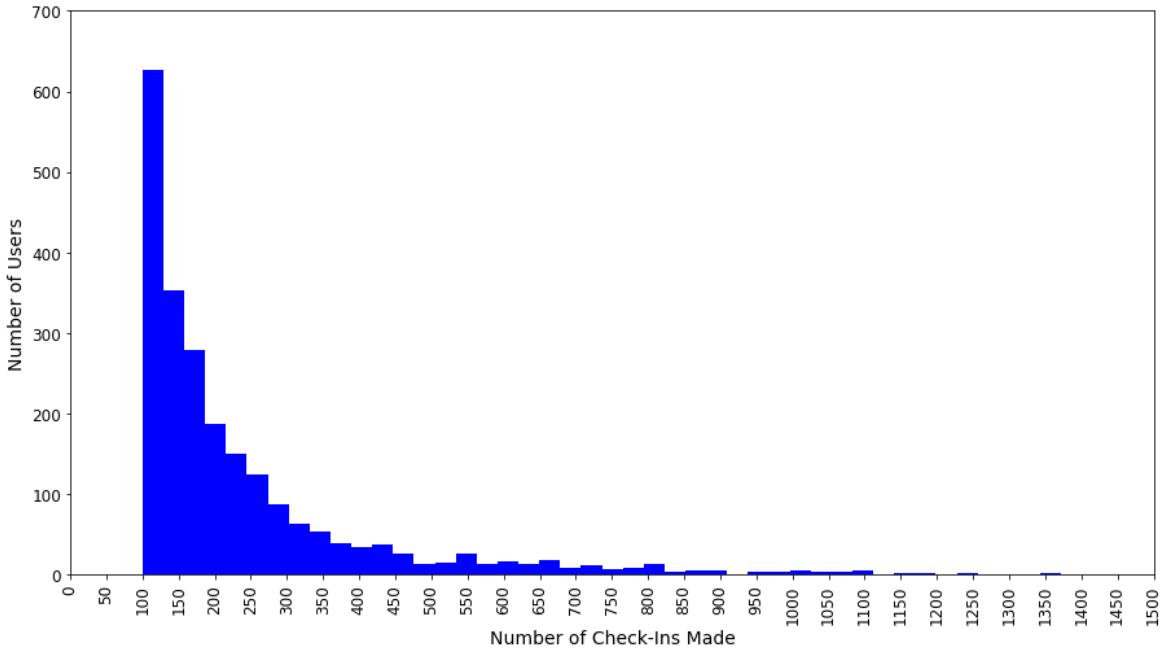

Figure 3 Distribution of Number of Check-Ins Performed by User

3.3.2 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.

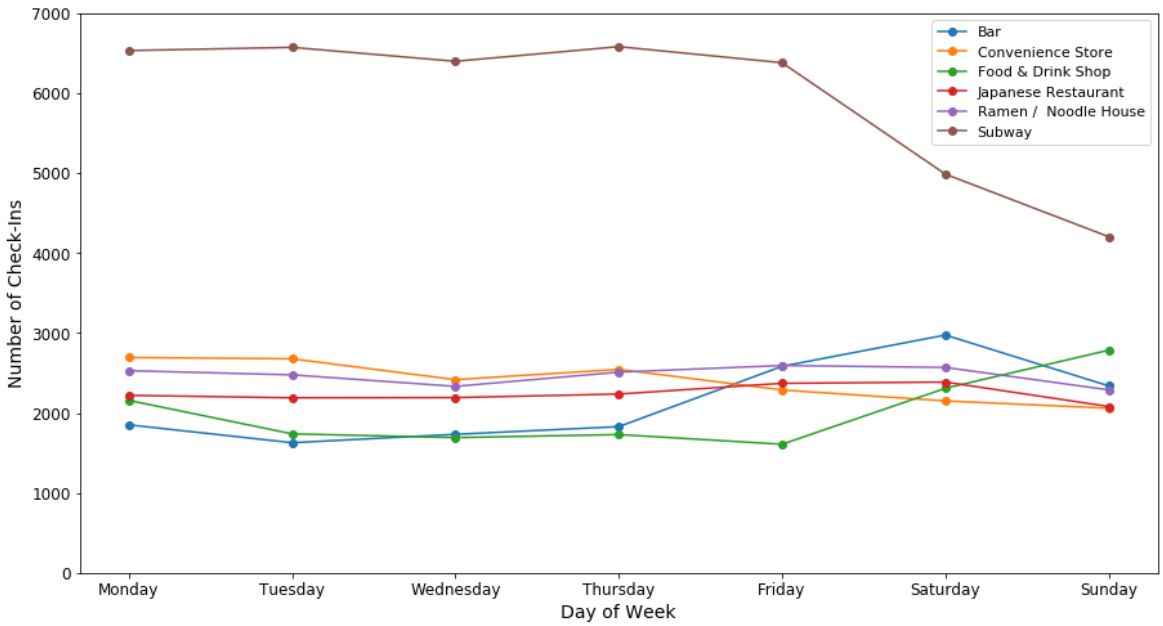

Figure 4 Check-Ins by Day (6 Most Popular Categories - Train Station Excluded)

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.

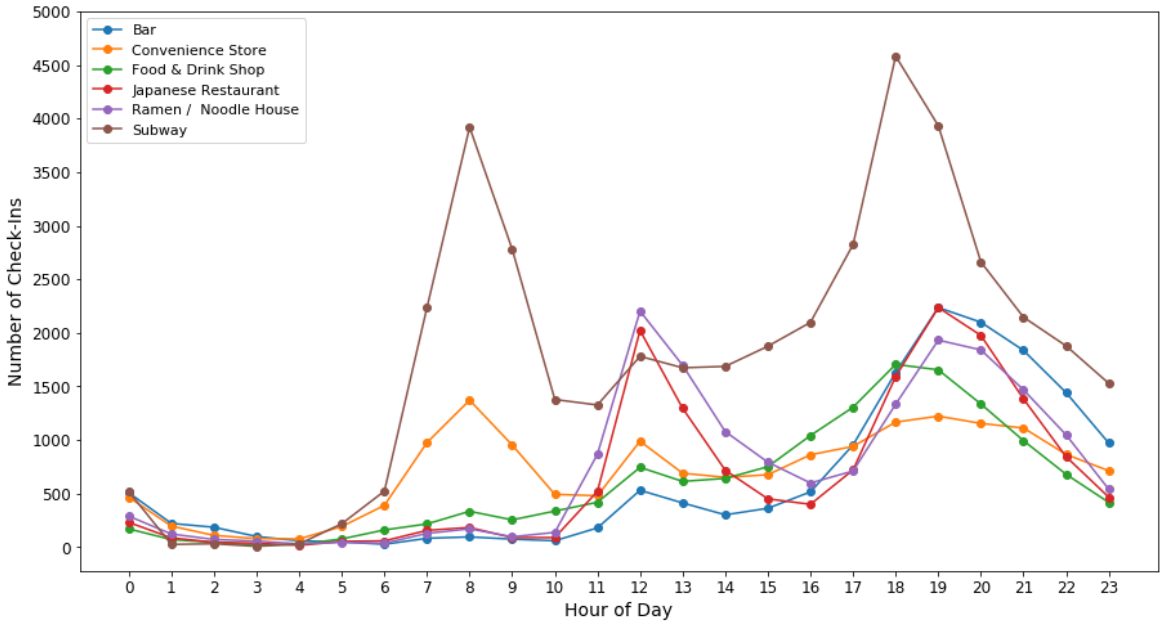

Figure 5 Check-Ins by Hour (6 Most Popular Categories - Train Station Excluded)

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.

3.3.3 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.

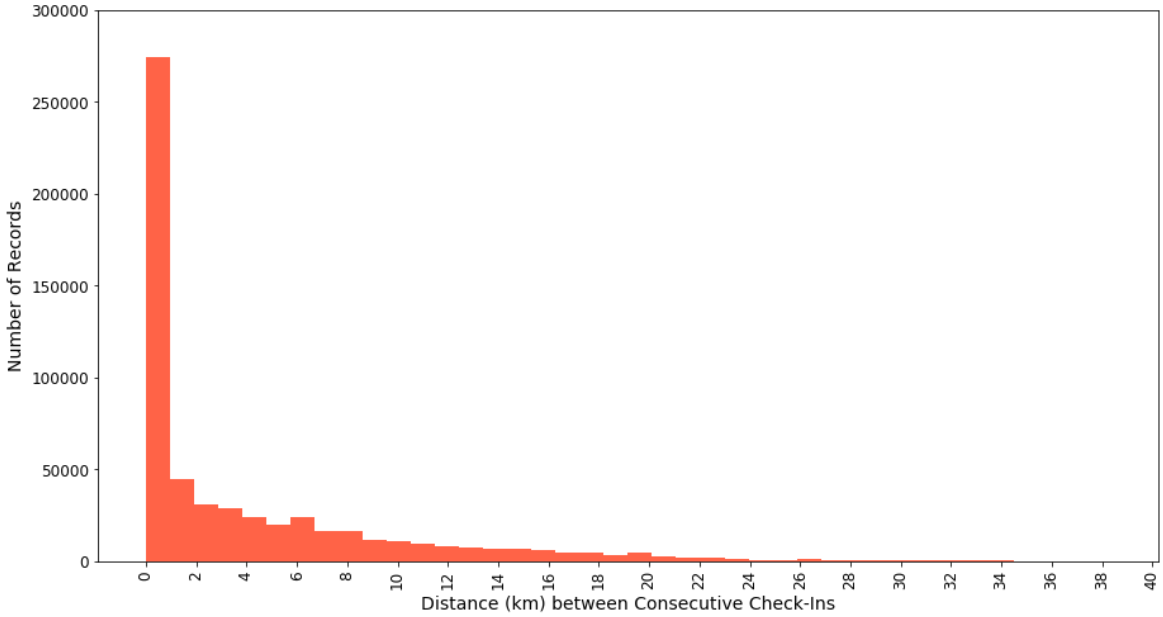

Figure 6 Distribution of Distance between Consecutive Check-In Locations

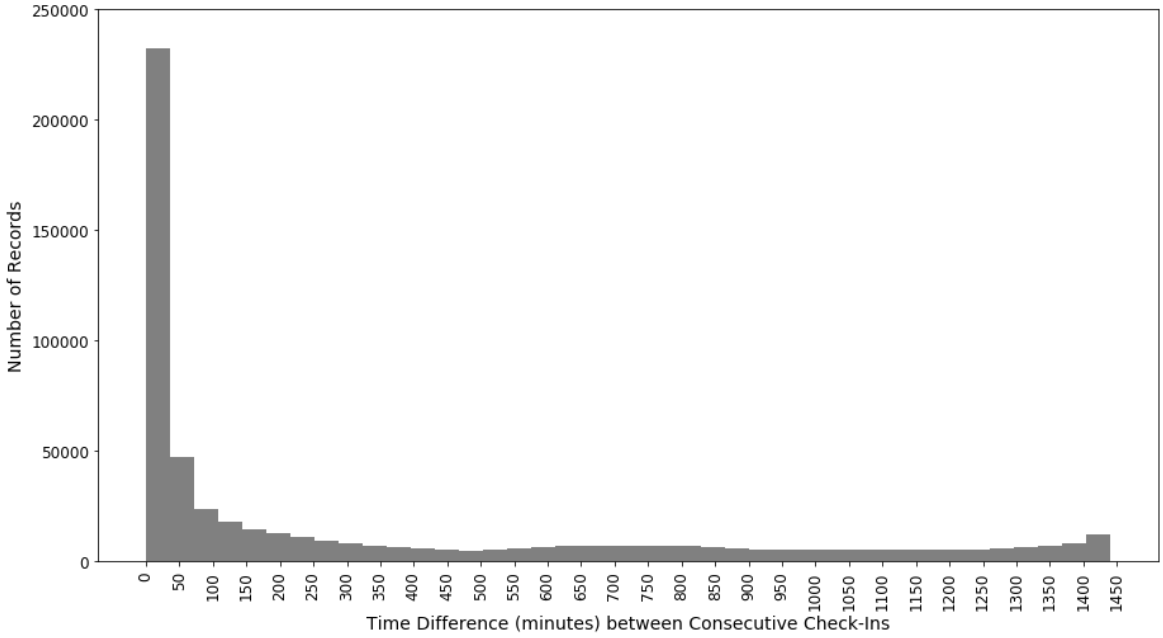

Figure 7 Distribution of Time Difference between Consecutive Check-In

3.3.4 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.

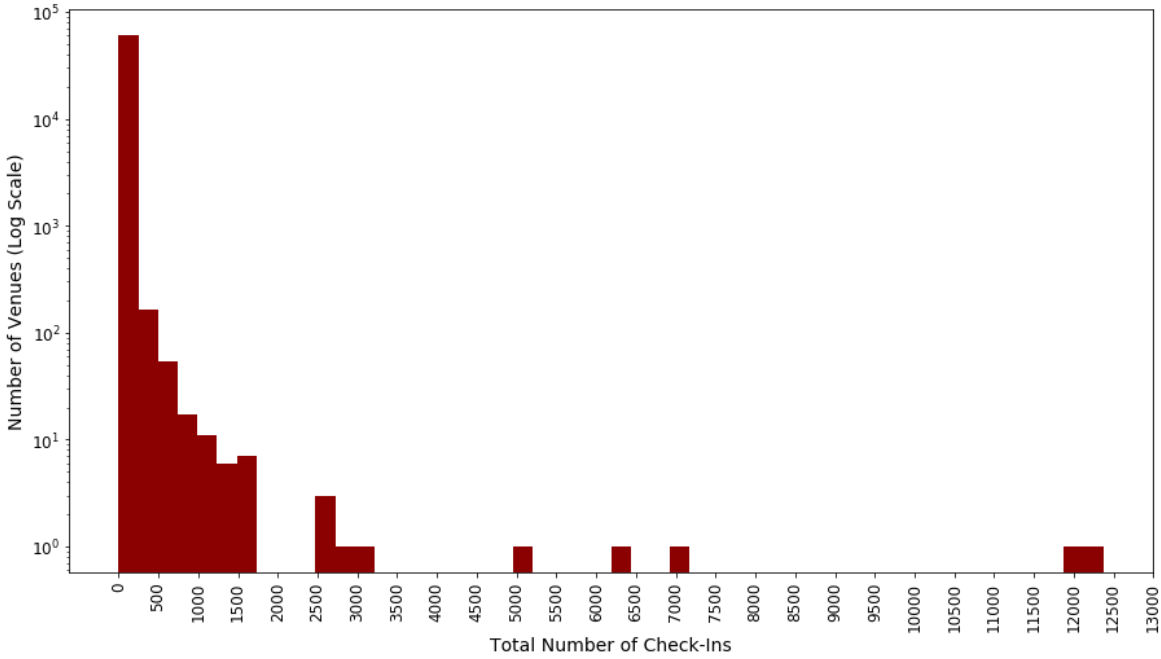

Figure 8 Distribution of Total Number of Check-Ins per Venue

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.

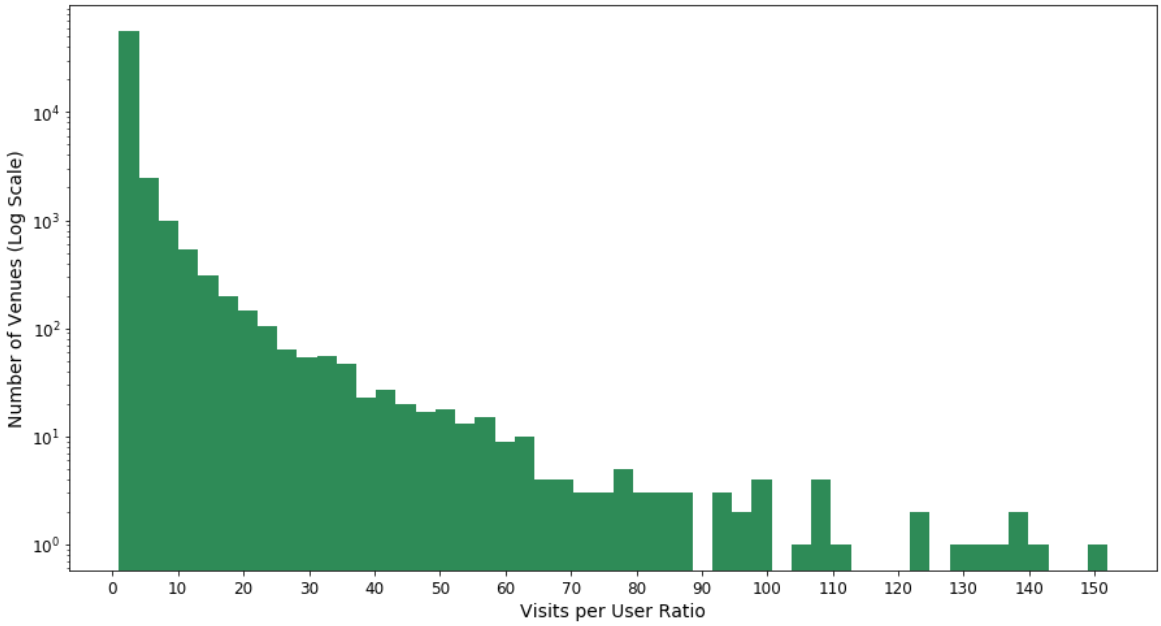

Figure 9 Distribution of Average Number of Visits per User

3.3.5 Distinct User Preference

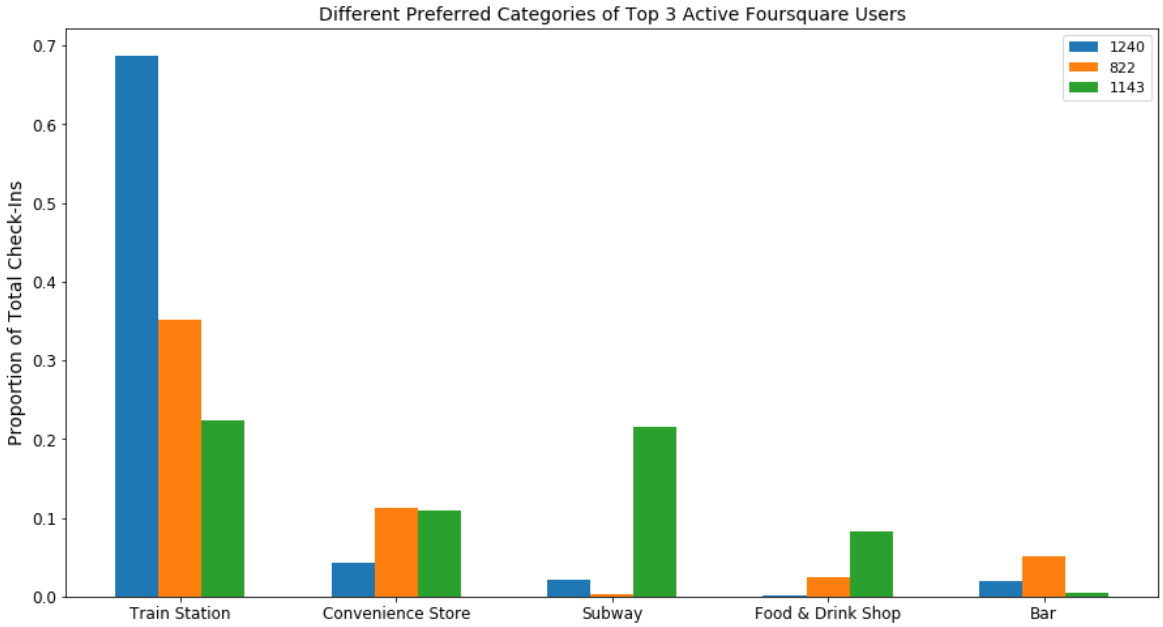

Figure 10 Preferred Venue Categories of Top 3 Most Active Users

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.

4 Methodology

4.1 Data Preparation

Table 5 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 5 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.

4.2 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, 70%-30% train-test split ratio was used and **Table 6** provides basic summaries of the data in training and test set.

Type	Training Set	Test Set
Date Range	04 Apr 2012 04:35 – 16 Nov 2012 17:03	16 Nov 2012 17:04 – 16 Feb 2013 11:34
Number of Records	185270	79401
Number of Unique Users	2246	2032
Number of Unique Venues	758	756
Number of Venue Categories	63	61

Table 6 Summary of Training & Test Sets

4.3 Next Check-In Prediction Problem

For LBSN providers to send targeted advertisements to a user, they will need to predict the user's next check-in venues. This prediction problem can be formulated as a multiclass classification problem, summarised in **Table 7**.

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 previous check-in venue
	• $t_{(u,k-1)}$**:** Timestamp of user's previous check-in
	Note
	1. Geographic coordinates, $v\_loc_{(u,k)}$ of the user's next 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 $t_{(u,k-1)}$ (**Appendix 2**)

Table 7 Next Check-In Prediction Problem

4.4 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

A detailed description of the four classifiers used can be found in **Table 11** of **Appendix 3**. Each classifier estimates the probability of a venue v being the next check-in venue. The venues with the highest estimated probabilities are used as predicted outcomes.

The hyper-parameters of Decision Tree and Random Forest were tuned using cross validation. Due to the time-based nature of the dataset, time-series cross validation technique must be used instead of k-fold cross validation. This is to ensure past data is used for training and unseen future data for validation. The detailed mechanism of time series cross validation is described in **Appendix 4**.

Global popularity method, which uses the most popular venues in the training set as the predicted outcomes, was adopted as baseline algorithm. The proposed models will be compared against the baseline algorithm in **Chapter 5**. Additionally, the predictive powers of the predictor variables listed in **Table 7** will also be evaluated.

4.5 Evaluation Approach

4.5.1 Evaluation Metric

This report uses a widely-adopted evaluation metric in location prediction tasks: *Accuracy@N* (Gao & Liu, 2014; Noulas, 2012). 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. In **Chapter 5**, the *Accuracy@N* scores will be reported for *N = {3, 5, 15, 30}*.

The metric *Accuracy@N* is especially relevant in LBSN context, where more than one venue recommendation could be displayed. The user can then pick the recommendation he or she is interested in.

4.5.2 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 (Ye et al, 2011). 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 separately in **Chapter 5**. In this report, cold-start users are defined as those who have no more than 10 check-in records in the training set.

5 Results & Discussion

Besides the test prediction accuracies of the classifiers, the predictive powers of different features (listed in **Table 7**) are also evaluated in this chapter. The classifiers were trained using different set of features and the results are presented in separate sections.

5.1 Temporal-Based Model

5.1.1 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 in **Chapter 3.3.2** could accurately predict the next check-in venue. **Table 8** lists the temporal features used as predictors.

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 8 Target and Predictor Variables in Temporal-Based Model

5.1.2 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.

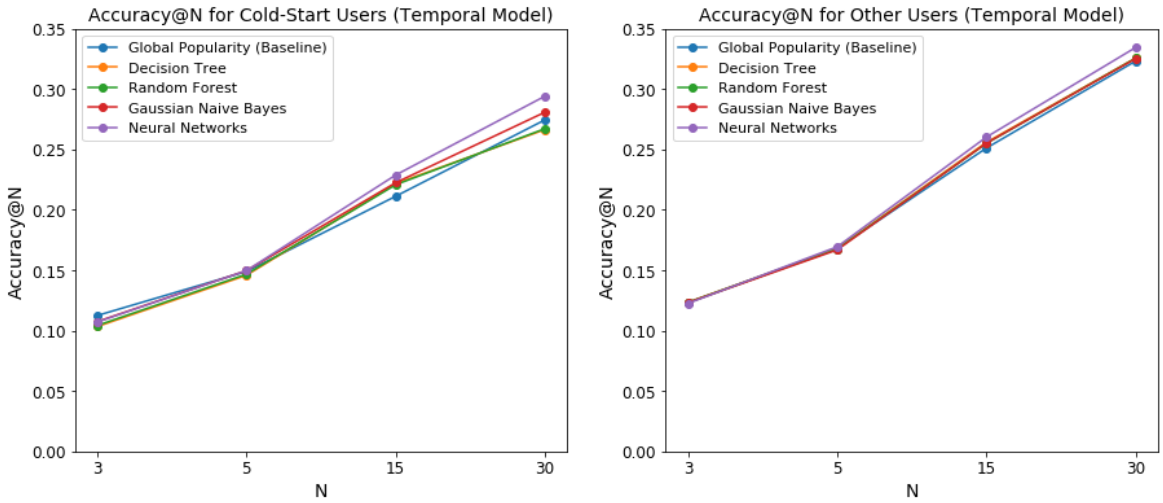

Figure 11 Test Set Accuracy@N of Temporal-Based Models

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

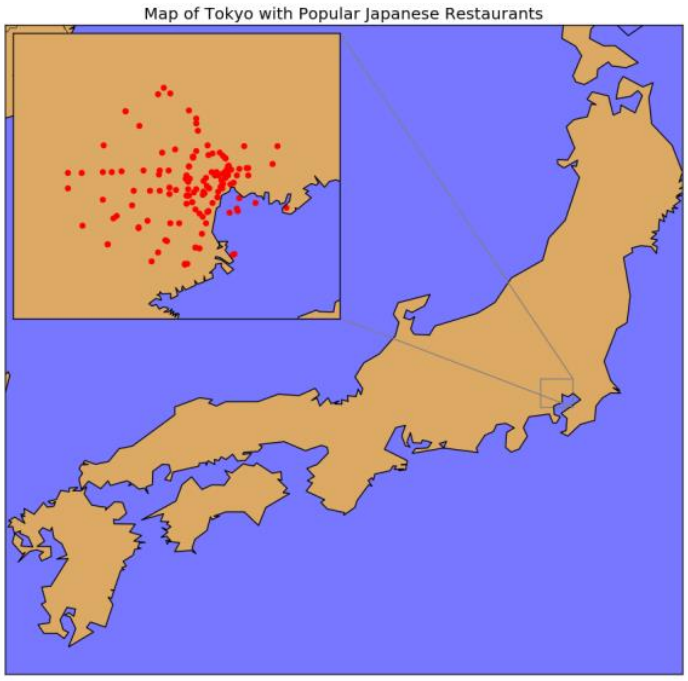

Figure 12 Locations of Popular Japanese Restaurants

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.

5.2 Spatio-Temporal Model

5.2.1 Variables Used

Gao & Liu (2014) has found that 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 9**.

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 9 Target and Predictor Variables in Spatio-Temporal Model

5.2.2 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.

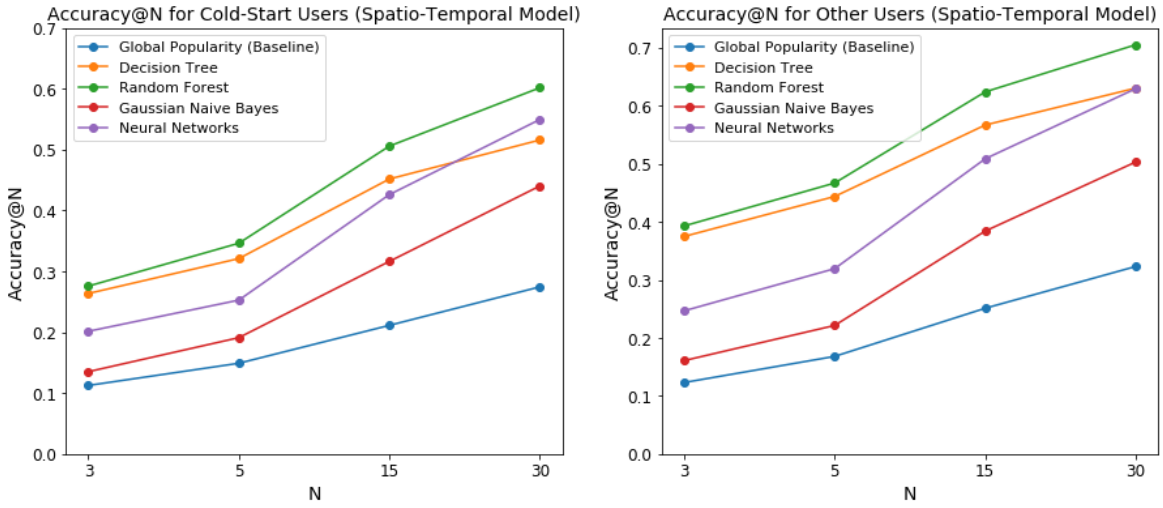

Figure 13 Test Set Accuracy@N of Spatio-Temporal Models

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*.

5.3 Fusion 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.

5.3.1 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 10**) 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 10 User-Venue Preference Matrix

The estimated probability at each cell is calculated using normalised frequency of visit:

$$\boldsymbol{P(v_j|u_i) = \frac{N_{u_i}(v_j) + 1}{\sum_{k=1}^{n}(N_{u_i}(v_k) + 1)}}$$

$\boldsymbol{N_{u_i}(v_j)}$ is the number of times a user $\boldsymbol{u_i}$ has checked-in to venue $\boldsymbol{v_j}$. Since a user might occasionally be interested in exploring new venues, Laplace smoothing is used so that a non-zero probability is assigned to previously unvisited venue.

The probability (or preference score) of a check-in venue given a user $\boldsymbol{u_i}$, $\boldsymbol{P(v_j|u_i)}$) can now be obtained by looking up the user-venue preference matrix. If $\boldsymbol{u_i}$ is not found in the matrix (i.e. user is not present in training set), each venue is assigned equal probability: $\boldsymbol{P(v_1|u_i) = P(v_2|u_i) = \cdots = P(v_n|u_i) = \frac{1}{n}}$

5.3.2 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.

Denoting the final estimated probability as $P(v_j|u_i, t_{(u_i,k)}, t_{(u_i,k-1)}, v\_loc_{(u_i,k-1)})$, the probability is proportional to the product of user preference score and Spatio-Temporal estimate:

$$P(v_j|u_i, t_{(u_i,k)}, t_{(u_i,k-1)}, v\_loc_{(u_i,k-1)}) \propto Preference\ Score \times SpatioTemporal\ Estimate$$

One of the venues must be the next check-in venue. Thus, given a user u_i, we could normalise the product to 1 to get the final probability:

$$\sum_{j=1}^{n}[P(v_j|u_i, t_{(u_i,k)}, t_{(u_i,k-1)}, v\_loc_{(u_i,k-1)})] = 1$$

5.3.3 Test Results and Discussion

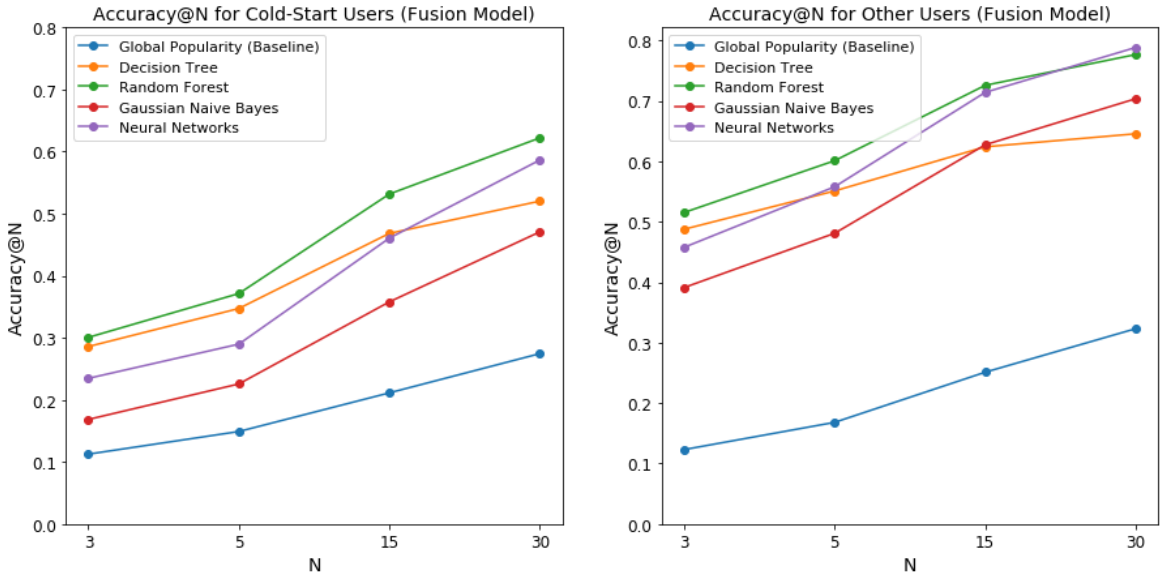

Figure 14 Test Set Accuracy@N of Fusion Models

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.

Based on these findings, LBSN service providers should encourage their new users to keep on using the app (e.g. by offering rebates or vouchers). Once the user has accumulated more check-in records, the predictive models will be able to predict his or her next check-in more accurately and more targeted advertisements can be displayed.

6 Conclusion & Recommendation

The digital marketing market has become fiercely competitive in recent years. Companies are now looking for ways to improve operational efficiency and optimise their marketing budgets by advertising to the right customers at the right time. Consequently, it has become vital for social media companies to accurately predict their user receptiveness to certain ads. In the case of LBSN providers, the historical check-in data could be used to predict future user check-ins. By knowing the users' possible next check-in venue, they will be able to display targeted advertising messages in a more effective manner.

A number of predictive techniques have been proposed by the research community in predicting future check-ins using historical check-in data. This report investigates the user check-in behaviours by analysing the Foursquare dataset and three types of features (i.e. temporal, spatial and user preference) have been used in the predictive models.

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%.

In light of these findings, LBSN providers should provide perks to new users to encourage them in using the apps more often. As they accumulate more check-ins, the predictive models will be able to predict their next check-in venues more accurately and provide more targeted advertisements. This also enhances the capability of LBSN providers to retain their users.

While this report used aforementioned features for prediction, LBSN providers could also analyse how a user's social network influences his or her check-in behaviours. A number of researchers (Cho, Myers, & Leskovec, 2011; Gao & Liu, 2014) have shown that a proportion of user check-ins were influenced by circle of friends. Analysing social network information could potentially provide high quality recommendations to users.

Next check-in prediction is still an active research topic and the data collected by LBSN providers have huge potential. The historical check-in and social network check-in data could be useful in predicting user's next check-in. By employing analytics, LBSN providers can display targeted ads to users and achieve better user experience. The main purpose of this report is to introduce the immense potentials of this area and encourage LBSN providers to consider using predictive models to improve their targeted advertising capabilities.

References

Forbes Corporate Communications (23 March 2016) Location-Based Marketing Is Fast Becoming Essential to Remain Competitive, Says New Study. Available from: https://www.forbes.com/sites/forbespr/2016/03/23/location-based-marketing-is-fast-becoming-essential-to-remain-competitive-says-new-study [Accessed 27th August 2017]

Eddy, N. (28 April 2014) Location-Based Advertising Market to Hit Nearly $15 Billion by 2018. *eWeek*. p12. Available from: http://search.ebscohost.com/login.aspx?direct=true&db=bsu&AN=95822198&site=ehost-live [Accessed 27th August 2017]

Foursquare. (2017) *Get more customers walking through your door*. Available from: http://business.foursquare.com/ads/ [Accessed 27th August 2017]

Facebook. (2017) *Buying Facebook adverts*. Available from: https://en-gb.facebook.com/business/learn/how-much-facebook-ads-cost [Accessed 27th August 2017]

Gao, H., Tang, J., Hu, X. & Liu, H. (2013) Exploring temporal effects for location recommendation on location-based social networks. In: ACM. *Proceedings of the 7th ACM conference on Recommender systems*. pp. 93-100. Available from: https://pdfs.semanticscholar.org/771c/74047c94749f718245e7ec0bc9270de143fb.pdf [Accessed 17th Aug 2017]

Preoţiuc-Pietro, D. & Cohn, T. (2013). Mining user behaviours: a study of check-in patterns in location based social networks. In: ACM. *Proceedings of the 5th Annual ACM Web Science Conference*. pp. 306-315. Available from:

http://sites.sas.upenn.edu/danielpr/files/foursq13websci.pdf [Accessed 17th Aug 2017]

Noulas, A., Scellato, S., Lathia, N. & Mascolo, C. (2012). Mining user mobility features for next place prediction in location-based services. In: IEEE. *Data mining (ICDM), 2012 IEEE 12th international conference on.* pp. 1038-1043. Available from: http://www.ccs.neu.edu/home/cbw/static/class/5750/papers/icdm12_noulas_cr.pdf [Accessed 17th Aug 2017]

Manotumruksa, J., (2015). Users location prediction in location-based social networks. In: British Computer Society. *Proceedings of the 6th Symposium on Future Directions in Information Access.* pp. 44-47. Available from: http://ewic.bcs.org/upload/pdf/ewic_fdia15_paper11.pdf [Accessed 17th Aug 2017]

Ye, M., Yin, P., Lee, W.C. & Lee, D.L. (2011). Exploiting geographical influence for collaborative point-of-interest recommendation. In: ACM. *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information* Retrieval. pp. 325-334. Available from: http://www.cse.cuhk.edu.hk/irwin.king.new/_media/presentations/p325.pdf [Accessed 17th Aug 2017]

Gao, H., Tang, J. & Liu, H. (2012). Mobile location prediction in spatio-temporal context. In: *Nokia mobile data challenge workshop*. Vol. 41, No. 2, pp. 1-4. Available from: https://www.researchgate.net/profile/Huan_Liu6/publication/265437484_Mobile_Location_Prediction_in_Spatio-Temporal_Context/links/551c9ba80cf20d5fbde557eb.pdf [Accessed 17th Aug 2017]

Liu, Q., Wu, S., Wang, L. & Tan, T. (2016). Predicting the Next Location: A Recurrent Model with Spatial and Temporal Contexts. In: *AAAI*. pp. 194-200. Available from:

http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/download/11900/11583 [Accessed 17th Aug 2017]

Berjani, B. & Strufe, T. (2011). A recommendation system for spots in location-based online social networks. In: ACM. *Proceedings of the 4th Workshop on Social Network Systems*. p. 4. Available from: http://www.academia.edu/download/32086962/A_Recommendation_System_for_Spots_in_Location-Based_Online_Social_Networks.pdf [Accessed 17th Aug 2017]

Bao, J., Zheng, Y. & Mokbel, M.F. (2012). Location-based and preference-aware recommendation using sparse geo-social networking data. In: ACM. *Proceedings of the 20th international conference on advances in geographic information systems.* pp. 199-208. Available from: https://pdfs.semanticscholar.org/4814/852815557deda9e56c9e05c233545a20d62d.pdf [Accessed 17th Aug 2017]

Lian, D., Xie, X., Zheng, V.W., Yuan, N.J., Zhang, F. & Chen, E. (2015). CEPR: A collaborative exploration and periodically returning model for location prediction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(1). p. 8. Available from: http://staff.ustc.edu.cn/~cheneh/paper_pdf/2015/Defu-Lian-TIST.pdf [Accessed 17th Aug 2017]

Wang, Y., Yuan, N.J., Lian, D., Xu, L., Xie, X., Chen, E. & Rui, Y. (2015). Regularity and conformity: Location prediction using heterogeneous mobility data. In: *ACM. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* pp. 1275-1284. Available from: http://staff.ustc.edu.cn/~cheneh/paper_pdf/2015/Yingzi-Wang-KDD.pdf [Accessed 17th Aug 2017]

Gao, H. & Liu, H. (2014). Data analysis on location-based social networks. In: *Mobile social networking*. Springer New York. pp. 165-194. Available from: http://www.nini2yoyo.com/papers/LBSN_chapter.pdf [Accessed 17th Aug 2017]

Wang, H., Terrovitis, M. & Mamoulis, N. (2013).Location recommendation in location-based social networks using user check-in data. In: ACM. *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. pp. 374-383. Available from: https://www.researchgate.net/profile/Manolis_Terrovitis/publication/260294299_Location_Recommendation_in_Location-based_Social_Networks_using_User_Check-in_Data/links/0f317530ac8b50c7bc000000.pdf [Accessed 17th Aug 2017]

Cho, E., Myers, S.A. & Leskovec, J. (2011). Friendship and mobility: user movement in location-based social networks. In: ACM. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 1082-1090. Available from: http://roke.eecs.ucf.edu/Reading/Papers/Friendship%20and%20Mobility%20User%20Movement%20In%20Location-Based%20Social%20Networks.pdf [Accessed 17th Aug 2017]

Lian, D., Zhu, Y., Xie, X. & Chen, E. (2014). Analyzing location predictability on location-based social networks. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, Cham. pp. 102-113. Available from: http://staff.ustc.edu.cn/~cheneh/paper_pdf/2014/HDefu-Lian-PAKDD.pdf [Accessed 17th Aug 2017]

Kaggle. (2017). *FourSquare - NYC and Tokyo Check-ins*. Available from: https://www.kaggle.com/chetanism/foursquare-nyc-and-tokyo-checkin-dataset [Accessed 17th Aug 2017]

Yang, D., Zhang, D., Zheng, V.W. & Yu, Z. (2015). Modeling user activity preference by leveraging user spatial temporal characteristics in LBSNs. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1), pp.129-142. Available from: http://www-public.tem-tsp.eu/~zhang_da/pub/TSMC_YANG_2014.pdf [Accessed 17th Aug 2017]

Zhang, J.D. & Chow, C.Y. (2013). iGSLR: personalized geo-social location recommendation: a kernel density estimation approach. In: ACM. *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems.* pp. 334-343. Available from: http://www.cs.cityu.edu.hk/~chiychow/papers/ACMGIS_2013a.pdf [Accessed 17th Aug 2017]

Wolfram Research, Inc. (2017). *Spherical Coordinates*. Available from: http://mathworld.wolfram.com/SphericalCoordinates.html [Accessed 25th Aug 2017]

Appendix 1 Conversion of Geographic Coordinates

The geographic coordinates in the Foursquare dataset are given as longitudes and latitudes. For the predictive model to gauge the distance between two points more accurately, they will need to be converted into 3-dimensional Cartesian coordinates.

In this report, we assume Earth to be fully spherical and the conversion formulae are as follows (Wolfram Research, 2017):

$$x = R \times \cos(latitude) \times \cos(longitude)$$

$$y = R \times \cos(latitude) \times \sin(longitude)$$

$$z = R \times \sin(latitude)$$

R is the Earth's radius: 6371 kilometres

Appendix 2 Imputed Values

If a user's previous check-in is not found (e.g. first check-in by a user ***u***), the following values are imputed for $\boldsymbol{v\_loc_{(u,k-1)}}$ *and* $\boldsymbol{t_{(u,k-1)}}$:

- $\boldsymbol{v\_loc_{(u,\,k-1)}}$: Mean geographic coordinates of all the check-in records in the training set
- $\boldsymbol{t_{(u,k-1)}}$: About 326.5 minutes (average time difference between a user's consecutive check-in records in training set) before the next check-in, $\boldsymbol{t_{(u,k)}}$

Only less than 1% of the training records do not have previous check-in records. Therefore this imputation is expected to have minimal effects on the training of predictive models.

Appendix 3 Multiclass Classifiers Evaluated

Table 11 tabulates the types of multiclass classifiers used and their parameters.

Item	Description
Decision Tree	• Based on CART (Classification and Regression Tree) algorithm, which constructs binary tree • Gini impurity is used to measure quality of node split • The hyper-parameter (minimum number of samples at leaf nodes) is tuned using time-series cross validation
Random Forest	• Average predictions over ten decision trees • Gini impurity is used to measure quality of node split • The hyper-parameter (minimum number of samples at leaf nodes of each tree) is tuned using time-series cross validation
Artificial Neural Network	• Two hidden layers (128 nodes each) to perform further feature extraction • Utilised dropout regularisation technique to reduce overfitting • Used categorical cross-entropy as the loss function to train the Neural Network for multiclass classification
Gaussian Naïve Bayes	• Parametric model • Assumes conditional independence between features • Assumes Gaussian distribution for likelihood of features • Low computational complexity and fast to train

Table 11 Multiclass Classifiers Evaluated

Appendix 4 Time-Series Cross Validation

Traditional *k*-fold cross validation technique is not suitable for time series data. At each run of *k*-fold cross validation, one single fold is used for validation and all other folds are used for training. This results in future data being used to predict past data (e.g. fold 2 is used for validation while folds 1, 3, 4, 5 are for training). Consequently the validation score calculated could be error-prone.

Since the Foursquare dataset is time-based, a variant of *k*-fold cross validation technique is used for cross validation purpose. A 4-split time-series cross validation technique is illustrated in **Figure 15**. The data is split into five folds (arranged in chronological order), with red fold indicating the fold used for training, blue fold for validation, and grey fold left unused in that run.

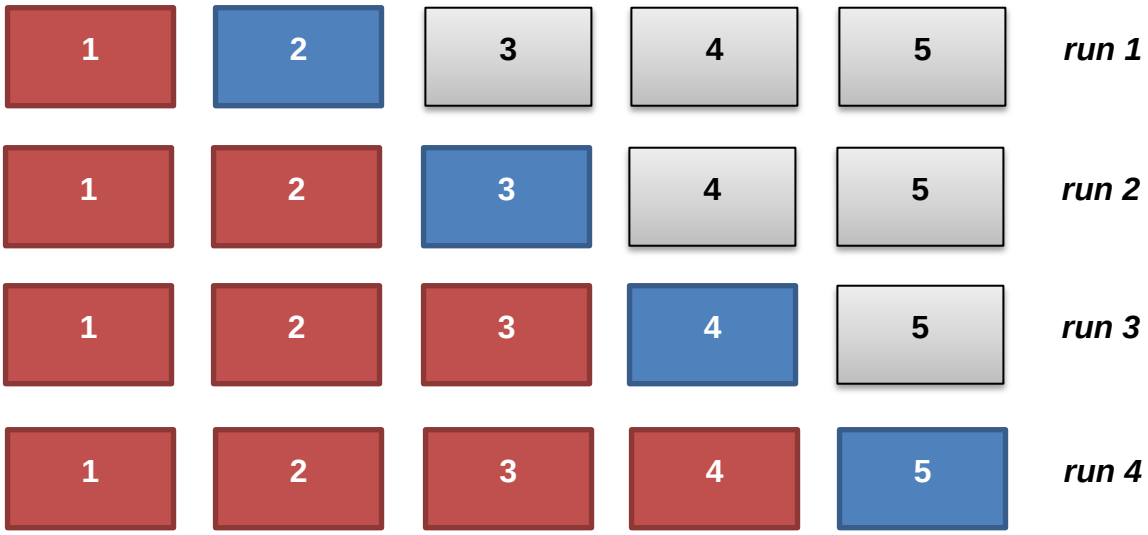

Figure 15 Time Series Cross Validation (4-Split)

In each run, only the fold prior to the validation fold is used for training. Thus, no future observations will be used for training the model. This ensures that the predictive model only picks up patterns from historical data.