Traveler's Next Activity Predication with Location-Based Social Network Data

Diem The To

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science in Computer Science and Software Engineering

University of Washington

2022

Committee:

Dong Si

Ying Chen

Min Chen

Marc Dupuis

Program Authorized to Offer Degree:

Computing and Software Systems

© Copyright 2022

Diem The To

University of Washington

Abstract

Traveler's Next Activity Predication with Location-Based Social Network Data

Diem The To

Chair of the Supervisory Committee:

Professor Dong Si

Computing and Software Systems

The rise of technology and the internet provides powerful means for people from all around the world to communicate and connect with one another. Online social network platforms become go-to places for users to express and share their individuality, which includes choice of activities, locations, and associated timestamps. In turn, their opinions affect the point of view of others, who are in their online friendship circle. Users' increasing usage of social networks help accumulate massive amount of data that can be further explored. Particularly, this type of data attracts and allows researchers, who are interested in studying and understanding how social factors and previous experience influence user behavior in terms of activity-related travel choice. In this paper, the goal is to utilize such rich data sources to build a model that predicts user next activity. Such model contributes a powerful tool for integrating the location prediction with transportation planning and operations processes. Besides, it is valuable in commercial applications to create

better recommendation system with higher accuracy and ultimately attract more customers to partnering businesses. By studying the dataset, which contains millions of historical check-ins from thousands of users, it is possible to derive information that is useful in predicting user next activity. The proposed approach applies machine learning techniques on the collected features to deliver highly accurate prediction results with fast training and prediction time.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank Professor Dong Si and Professor Ying Chen for their guidance and support from start to finish of this thesis. I also wish to express my appreciation to University of Washington Bothell for providing me the access to the computing resources that allowed me to work on this research project.

Chapter 1. INTRODUCTION

1.1 Rationale

Mobile devices and social networks have increased in popularity due to their convenience and the power to connect users from around the globe. Smartphones have revolutionized the way user data is collected and unlocked many new possibilities and opportunities of the type of data that can be gathered, one of which is location-related data. With user approval, mobile applications can access data associated with user whereabout and use it to enable and perform their functionalities. Google Maps [1] is a well-known example of mobile applications that use Global Positioning System (GPS) information available on smartphones to help a user navigate from their current location to the desired destination. On the other hand, social networking sites have also made their way onto smartphones. User can now send messages to friends and family members across the globe as well as sharing anything via texts, images, or videos at any time and from anywhere. Point of interest (POI) is also on the list of the many things shared via social network interactions. Their online friend circle will then be aware of the user's choice of location and activity and can write comments and share it with their online friend network. Taking advantage of this behavior, location-based social networks (LBSNs) aim to provide a platform where a location is the main focus instead of status, photo, or video. Users can add their review of places that they have visited previously, submit ratings, as well as sharing any thoughts, experience, or recommendations.

LBSNs produce extensive user-generated digital footprints, which creates an exceptional opportunity to understand the spatial and temporal aspects of user activity. In LBSNs, user activity is primarily characterized by check-in, which denotes the timestamp when the user visits a certain POI. POIs can be categorized into several categories such as entertainment, food, recreational,

shop, travel, etc. User activities can be easily characterized based on these POI categories given any LBSN check-in. For example, if Angelina, a LBSN user, check-ins at a movie theater on a Saturday night, it is most likely that she is watching a movie at the location specified by a coordinate at the time when she posts her check-in. By mining this type of activity records, user spatial and temporal activity preferences can be easily studied, which can then enable various location-based applications. The most straightforward application is POI recommendation. In the perspective of traffic manager, the social media data supplements the transportation planning and operations by being a rich data source to provide real-time information.

Another aspect that is unique to LBSNs is the availability of social network graph of participating users. It is inevitable to have a growing friend list when engaging in LBSNs, because friend-to- friend interaction as well as making new connections are two of its main usage besides sharing POIs. For example, Angelina recently moved to a new city. She will then search on an LBSN for some nearby coffee shops and read some of the reviews from past visitors. Angelina came across Betty, whose coffee preference is similar to Angelina, and they became friends. Since Betty is from the area, she can introduce Angelina to some of her favorite restaurants. In LBSNs, user check-in pattern is correlated to their friend's check-in pattern. Social friendships can explain up to 30% of activity and location choice in social networks [2]. From observation, it has been seen that users are most likely to check-in right after a friend has checked-in to the same place [2]. This means that users are more likely to visit POIs, at which one of their social connections have visited before. Chen et al. further investigate friendship influenced people's destination choices, especially in a denser social network [3].

1.2 Specific Aims

To take advantage of social media data, especially LBSNs, for transportation applications, this work is not only to understand the characteristics of the data itself, but also aims to answer the following question: which activity is a LBSN user interested in given their historical check-ins data, including timestamp, location, and activity category? In addition, it leverages user social network graph and their check-ins to model user's past behavior and predict their next choice of activity accurately. In present literature [4] [5] [6] [7], POI categories are often seen as the representation of user activities. For example, on Foursquare [8], which is a well-known LBSN, an American steak house has 'food' as its category. Therefore, it is reasonable to assume that users, who have checked-in at this venue, were here for dining. However, making activity predictions solely based on LBSN historical check-ins is not a trivial task. In order to achieve the goal of predicting user's next choice of activity, a few steps are taken:

1.2.1 Clustering

The check-in data used is from two LBSNs, Brightkite and Gowalla [2]. The first attempt is to group users in different clusters based on their entire historical check-ins pattern. Being able to visualize different clusters, which represent different mobility patterns, allows for understanding the diversity in user temporal and spatial preference. This enables the creation of a more targeted prediction model.

1.2.2 Experimentation with Naïve Neural Network

A number of additional features are extracted from the original dataset. The derived features provide more insight into user mobility pattern and activity preference. For example, from converting UTC time zone into user's local time zone, the exact day of the week that each check-in falls into could be extracted. This is important because different activities are observed

depending on the day of the week. For instance, being a student, Angelina spends most of her time on school campus during the weekdays. Her activity preference changes during the weekends because classes are not offer on Saturday and Sunday. Instead, Angelina can enjoy her time at a park, a coffee shop, or a shopping mall. Next, a simple neural network is modeled and is used to make predictions on user next activity choice. After a few initial experiments, the prediction results came out to be quite decent in terms of accuracy. However, there is still room for improving the model to produce better results.

1.2.3 Utilizing Extreme Gradient Boost (XGBoost)

In the pursue of refining the prediction accuracy, this research further utilizes XGBoost algorithm, which is a distributed gradient boosting that is optimized to be highly efficient, flexible, and portable [9] . Significant improvement in prediction accuracy is observed from the results yield when applying XGBoost.

The remainder of this paper is organized into 7 sections. Chapter 2 summarizes relevant work that has been done previously. Chapter 3 introduces and analyzes the datasets used for experimentation. Chapter 4, 5 and 6 explain in detail the experiments and their results from employing each of the following methods respectively: cluster analysis, naïve neural network, and XGBoost. Chapter 7 expands on some additional research done after the initial publication and how the new method compares to the forementioned ones. Finally, Chapter 8 concludes the paper.

Chapter 2. RELATED WORKS

Due to the increasing popularity of LBSNs, users generate tremendous number of digital footprints [7] in their daily life. Research on understanding user activity by mining these digital footprints has attracted extensive attention in recent years. Since the objective of this paper is to predict user next choice of activity in LBSNs, the following research works on user activities are first briefly reviewed from two perspectives: 1) user mobility perspective which focuses on modeling user mobility patterns by leveraging spatial and temporal regularities and 2) user preference perspective which usually focuses on inferring user preference on POIs. After that, research works considering POI categories as user activities, as well as the related works for utilizing social network aspect of LBSNs will be presented.

From user mobility perspective, various studies have been conducted for location prediction. In LBSNs, location prediction in terms of POIs aims at predicting the specific POI that users will visit next. For example, various features are incorporated in latent Dirichlet allocation model for next POI prediction [10]. Based on hierarchical Pitman-Yor process [11], a social-historical model is proposed for predicting the next check-in of a user [12]. Hierarchical Pitman-Yor process is a hierarchical Bayesian nonparametric model that is based on the Pitman-Yor process. This process is a two-parameter generalization of the Dirichlet process, and it is more suitable for application relating to natural phenomena compared to the Dirichlet process [13]. Noulas et al. extract user specific features and global mobility features and built a prediction model for next POI [14]. To improve the performance of next location prediction in LBSNs, Likhyani et al. explore coarse-grained venue categories by exploring the association between sparse location data with map information, and then leverage this auxiliary information for more accurate location prediction of the next visit [15]. A similar work is conducted by considering social friendship [16].

Based on the observation that users tend to check-in several times in a single place, a predicting model based on Markov chain is proposed [17]. Different from these works that try to predict users' future locations, the objective of this paper is to infer the users' next activity preference based on their past and current context, i.e., time and location. Similarly, another research also explores the same idea to model user activity preference [18]. However, their work does not exploit the social aspect of LBSNs. Similarly, based on the assumption that some inherent patterns from past check-ins can indicate future check-in behavior, two predictors are developed to predict both the next visited venue's category and the expected time of that check-in [19]. However, this work also does not take users' social information into consideration. A framework is created for LBSN that include a social level to describe the interaction between users and their friends [20]. They use this framework in conjunction with a supervised scoring model and a classification model to predict user's future check-in location.

Chapter 3. DATASETS

This section starts with providing information regarding the datasets that will be used. Then, different mobility patterns are explored by performing cluster analysis on mentioned datasets. The next step is to further prepare the data to extract relevant insights that could help predict user next choice of activity. Along with the original check-in attributes, the newly derived features are used to train a neural network model. To further improve the predicting power, a new model is created by taking advantage of the efficiency and flexibility of XGBoost algorithm.

3.1 Datasets

The datasets being used are publicly available real-world data from Brightkite and Gowalla, which are two of many location-based social network sites. Each dataset contains millions of check-ins from thousands of their users. Both Brightkite and Gowalla datasets have the same set of 6 features, 5 of which are user_id, timestamp, latitude, longitude, and venue_id. Timestamps are in Coordinated Universal Time (UTC). Venue_ids can only be used to identify venues in its own dataset. The 6^{th} feature, activity category, is not present in the original datasets. It is collected later using Foursquare API. Table 3.1 shows the time period during which check-ins are collected, number of check-ins per dataset, number of unique users, and the number of edges representing friendship between users. Table 3.2 shows an example of the check-in data, its format, and data type.

Table 3.1. Dataset Information

Dataset Name	Brightkite	Gowalla
Date Range	April 2008 – October 2010	February 2009 – October 2010
Number of Check-ins	4.5 million	6.1 million
Number of Users	58,228	196,591
Number of Edges	214,078	950,327

Table 3.2. Check-in Data Sample

Feature	Sample Data
User_id	0
Time_stamp	2010-10-17T01:48:53Z
Latitude	39.747652
Longitude	-104.99251
Venue_id	88c46bf20db295831bd2d1718
Activity_category	Building/default_

Chapter 4. CLUSTER ANALYSIS

The objective of this section is to group users in different clusters based on their historical check-ins. Visualizing different clusters, which represent different mobility patterns, make it possible for understanding the diversity in user temporal and spatial preference. In turn, this enables the creation of a more targeted prediction model. Firstly, cluster analysis will be done on each user's check-ins by placing the focus on user spatial displacement characteristics. Secondly, check-ins that belong to each user are stringed into a single sequence of activities. Cluster analysis is then performed on the whole dataset.

4.1 Spatial Displacement

The first step is to clean up each user's check-ins and pre-process the data so that the format is compliant with the libraries used. All data fields must be a numerical value, so activity categories are converted into number as depicted in Table 4.1. The algorithm employed in this process is mean shift because it does not require a predetermined number of clusters as an input. Mean shift is a centroid-based algorithm [21]. It works by updating candidate data points for centroids to be the mean of all points within a given region. The candidate data points then go through a filtering process to eliminate duplication. Even though the mean shift is not scalable, it is appropriate for the research purpose because the maximum number of check-ins per user is only a few thousand. For each user, their check-ins are clustered into several different groups. Results for sample users #1, #3, and #10 are plotted in Figure 4.1a, Figure 4.1b, and Figure 4.1c accordingly. Each cluster is represented by a color. For example, in Figure 4.1a, cluster 0 is denoted by red color and cluster 1 is denoted using blue color. Similarly, three colors are used to denote three different clusters in Figure 4.1b. Check-ins that belong to cluster 0, denoted in red, seem to follow a path to travel from

the Pacific Northwest to San Diego. These check-ins might be correlated based on timestamps or coordinates. They can be from a single trip, so timestamp might be close to one another. Since most of these check-ins seem to form a vertical line on the west coast, their longitudes are similar. In addition, the number of clusters varies depending on user mobility range and pattern. For example, user #10's check-ins are clustered into 5 groups (Figure 4.1c) whereas user #1's check-ins only cluster into 2 groups (Figure 4.1a). The difference in the distance traveled might have contributed to this observation because user #10 travels internationally spanning several continents whereas user 1 mainly stays in small areas of a few cities in the US. Besides, the user's city or area of residency can be inferred from check-in location and clusters on the map. For instance, a dense cluster in blue color can be observed in Figure 4.1a and there is a high possibility that this is the area where user #1 lives, goes to school or works at. Based on all plots, it is observed that there are 3 distinct spatial displacement ranges:

- High – this applies to users who travel to other continents (Figure 4.1a).
- Medium – this applies to users who travel within the country (Figure 4.1b).
- Low – this applies to users who travel within their state (Figure 4.1c).

Users, whose special displacement range is low, tend to make short-term trips both in terms of distance and time because, for shorter trips, origin and destination need to be somewhat close to one another, mostly within a state. Likewise, travelers, who have medium special displacement range tend to travel outside of their state but still inside of the US. Similarly, with high spatial displacement range, users tend to travel internationally.

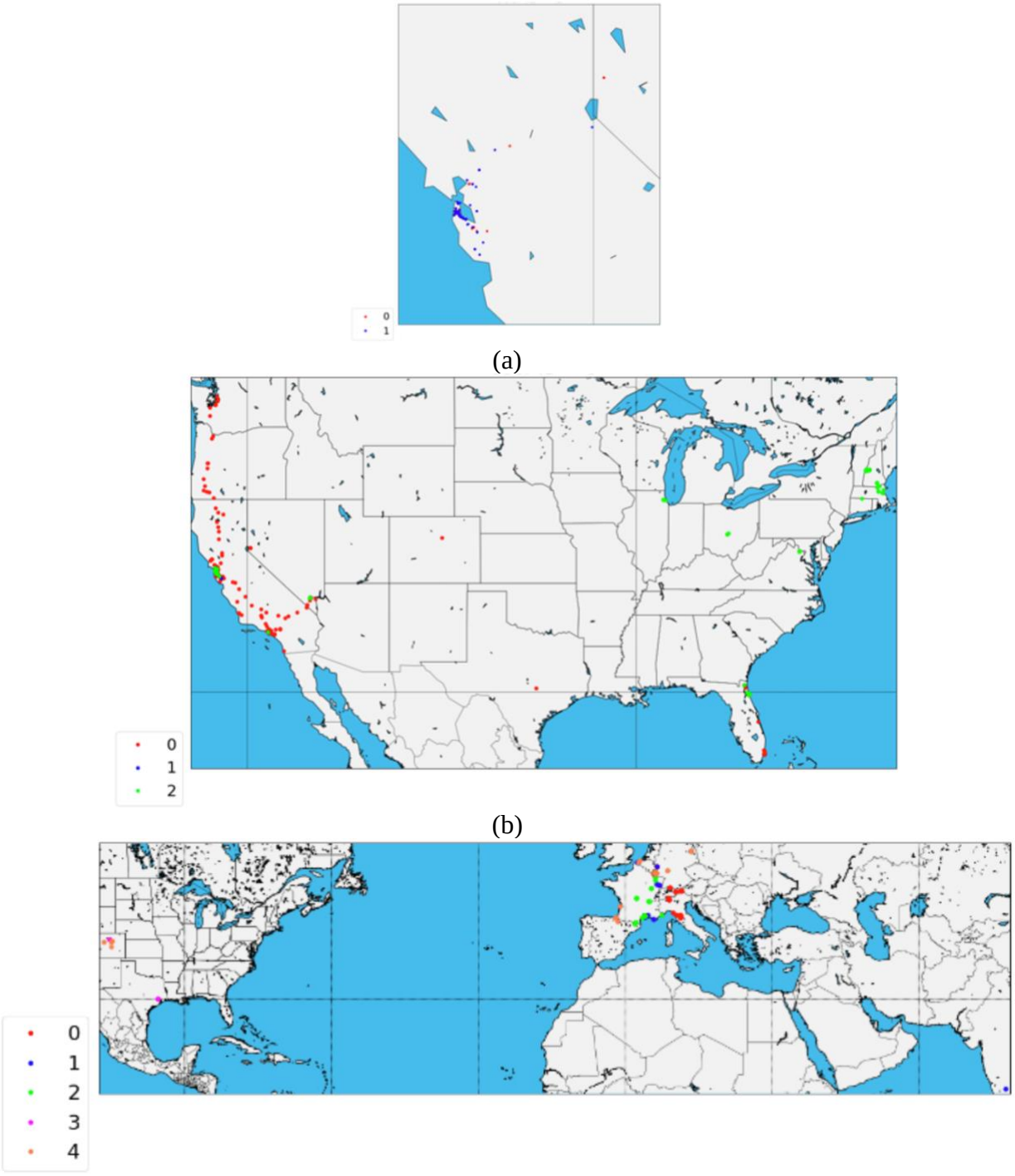


Figure 4.1 Plots of Location Clusters for User 1 (a), User 3 (b), and User 10 (c)

Table 4.1 Activity Categories and Their Corresponding Identification Number

Numerical Value	Activity Category
1	Arts entertainment
2	Office building
3	Education
4	Event
5	Food
6	Nightlife
7	Parks outdoors
8	Shops
9	Travel

4.2 Activity Sequence

The purpose of performing activity sequence cluster analysis is to understand different types of activity patterns among a population. From the original datasets, important and related attributes such as user_id, and activity_category are extracted, cleaned up and pre-processed so that they are in compliant with the library used to perform analysis.

Upon the completion of data pre-processing, each user's activities are tallied up into a single data point or activity sequence. The length of each sequence depends on the number of check-ins available for that particular user and the values vary greatly within the population. For example, user 8 only has 32 check-ins whereas user 0 has more than 2,000 check-ins. As a result, the activity sequence of user 8 is 32-character long while that of user 0 is more than 2,000-character long. It is also observed that most users do not have the same number of check-ins.

Several experiments are carried out using Brightkite dataset varying the number of clusters from 2 to 6 in order to determine the most optimal number of clusters. The result showed that keeping the number of clusters to 2 yields clusters with the most minimal distance from any points to the centroid of each cluster. However, the size of the 2 resulting clusters is highly skewed. Cluster A contains more than 93% of the data, thus leaving only less than 7% to cluster B.

In order to explore the composition of the clusters, hierarchical clustering is performed on the same dataset. Figure 4.2 shows the resulted dendrogram that is truncated to 12 leaves. From observation, the red group contains the majority of users. Specifically, cluster #7 and #8 has more than 50% the number of users, 26,200 and 10,606 accordingly.

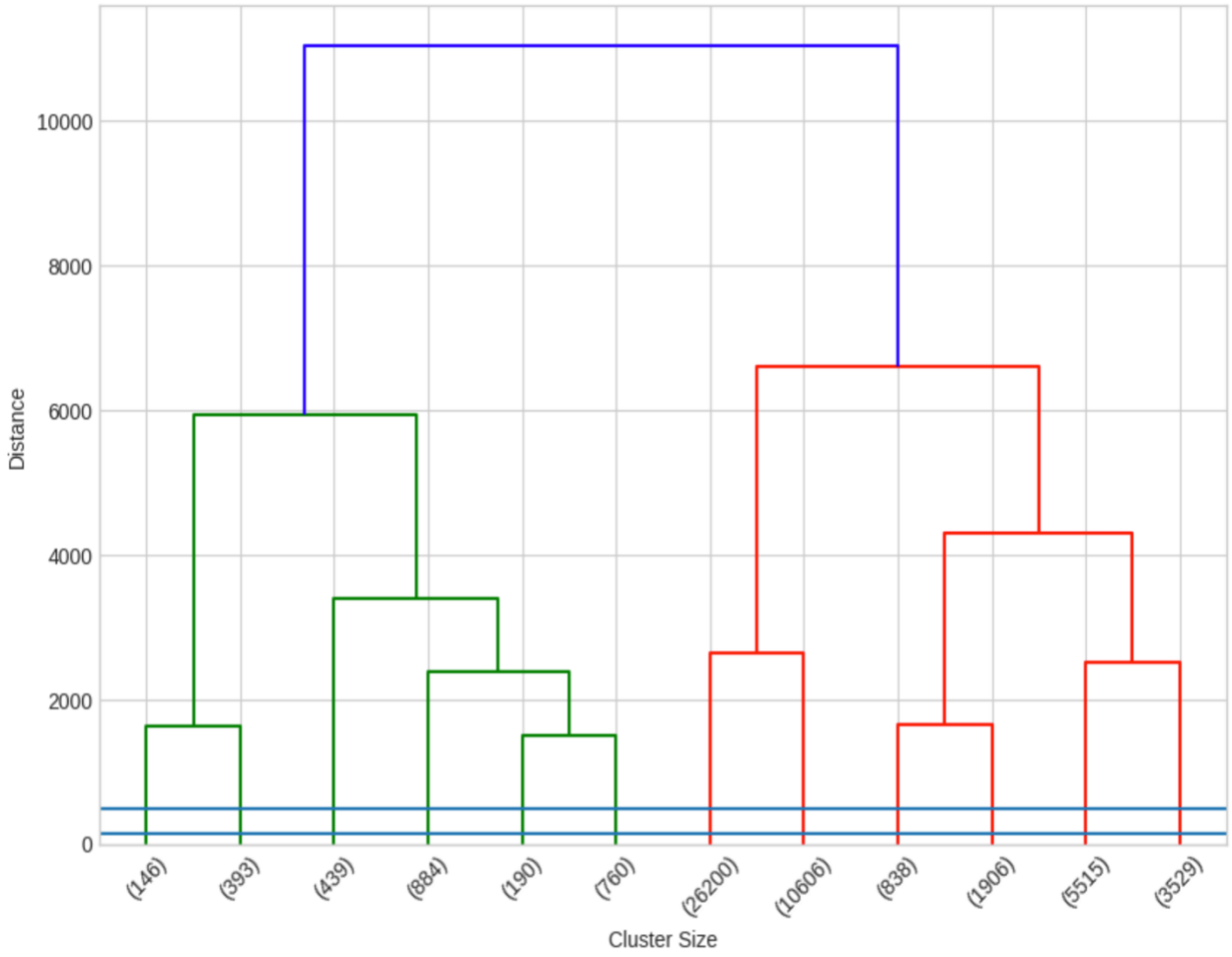


Figure 4.2 Truncated Hierarchical Clustering Dendrogram Showing the Breakdown of Clusters for Brightkite Population

Chapter 5. NAÏVE NEURAL NETWORK

The goal of training a neural network model is to predict user's next activity accurately. Utilizing this information, recommendation to nearby venues that belong to the same activity category will then be made.

5.1 Time and Activities

Converting central time to user's local time is important to compare travel pattern between users whose check-ins are in different time zones because the user's mobility patterns differ depending on the time of the day. Differences in age, gender, lifestyle, and occupation contribute to how and where a user chooses to spend their day. A person who is in their early 20s is most likely a student and tend to spend their daytime activities at an educational institution. During exam weeks, there is a high possibility that this student will spend a long hour studying at libraries at night. However, the same student can choose to spend their evening time enjoying a movie, attending a concert, or partying at a club after they have finished with their midterms or finals. On the other hand, another person who is in their 30s or 40s is most likely to spend their daytime at their work office and later enjoy their evening at home. Occupation can tell a lot about a person's mobility pattern. Employees of an airline, such as flight attendant or pilot, spend most of their time traveling long distance throughout the day. This leads to a large range of spatial displacement because they constantly travel between cities, states, or even countries. For example, users are more like to visit places such as school or work in the morning compared to bars or movie theaters, which are preferable during the nighttime for leisure.

5.2 Feature Extraction

The original datasets only provide 7 features. To obtain the information regarding each check-in, some additional features are extracted. Features are categorized into 2 major groups: user attributes

and venue attributes. Attributes that belong to users represent their travel pattern and attributes that belong to venues describe their location and activity category.

A list of user attributes and their value are shown in Table 5.1. Out of 7 attributes, the user ID is the only one that already exists in the original datasets. The rest of the attributes, such as hour of the day, day of the week, week of the month, distance traveled to next check-in location, the time elapsed from the current to next check-in and visit frequency of each category, are derived based on the given information. Table 5.2 shows a list of venue attributes and their value. Most venue attributes have already existed in the original dataset except for main category and sub-category, which are collected later using Foursquare API.

Table 5.1 User Attributes and Their Corresponding Value Added

Feature	Value Added
User ID	Uniquely identify each user
Hour of day	Activity choice differs based on time of day (morning vs nighttime)
Day of week	Activity choice differs from day to day in a week (weekday vs weekend)
Week of month	Choice of activity differs from season to season (might not be very useful since weathers are different from country to country)
Distance traveled to next check-in	How far is the user willing to travel to next activity
Time elapsed from current to next check-in	How much time is does the user wait before the next check-in/activity
Visit frequency of each category	User's preference in terms of activity type

Table 5.2 Venue Attributes and Their Corresponding Value Added

Feature	Value Added
Venue ID	Uniquely identify each venue
Venue latitude	Venue geographic location
Venue longitude	Venue geographic location
Main category	General activity type associated with this venue
Subcategory	More specific activity type associated with this venue

5.3 Neural Network

The objective is to solve a 10-class classification problem using a simple neural network. More than 25 experiments attempting to develop and train a neural network model are carried out varying the number of epochs, number of hidden layers, number of nodes per layer, activation function and early stopping. Activation functions are an essential feature of artificial neural network because they decide whether a node should be activated or not. This means that the function decides whether the information passed to the node is relevant to the given information or it should be ignored. The feature set used as training input includes all user attributes and venue attributes listed in Table 5.1 and Table 5.2. Some activation functions are employed to find the most suitable one specifically for Brightkite and Gowalla datasets. Unfortunately, after many trials using a wide range of different activation functions, the best accuracy achieved is only 47.56%.

Chapter 6. EXTREME GRADIENT BOOSTING

Due to the limited number of features, neural network is not effective in producing a model with high accuracy. Therefore, a different method known as eXtreme Gradient Boosting (XGBoost) [9] is employed. XGBoost is well-known for its scalability, which allows for faster learning by utilizing parallel and distributed computing as well as optimizing for efficient memory usage.

6.1 Data Processing

The same datasets and features used to create the neural network are used to train a model based on XGBoost method. All strings in the datasets are encoded into an integer to comply with the library used. Next, data points with missing data in any column are removed. Outliers are also removed by calculating the standard deviation of all data points. Data points that have a standard deviation larger than 3 are removed to ensure a normal distribution throughout the dataset. 70% of the data is used for training, the remaining 30% is used to evaluate the trained model.

6.2 Experiments

Besides XGBoost, a number of models are trained using different classifiers, including K-Nearest Neighbor [22], Decision Tree [23], Gaussian Naive Bayes [24], Logistic Regression [25], Linear Discriminant Analysis [26], and Random Forest [27], in order to compare and validate the effectiveness of XGBoost. Training time and predicting time are recorded to compare the performance between these models. Section 6.4 will take a closer look at the results.

6.3 Social Aspect

To further improve on the accuracy, social network graphs of all Brightkite and Gowalla users are utilized. In LBSNs, user check-in pattern is correlated to their friend's check-in pattern. Social friendships can explain up to 30% of activity and location choice in social networks [2]. From observation, it has been seen that users are most likely to check-in right after a friend has checked-in to the same place [2]. This means that users are more likely to visit POIs, at which one of their social connections has visited before. To unlock the power of social aspect, a network graph is built from each dataset. After that, the objective is to detect communities within this graph based on the density of the sub-networks.

Networks are built using the NetworkX library [28]. In a network, each node represents a user and each edge connecting 2 nodes represents a friendship. Brightkite network has 58,228 nodes and 214,078 edges. Gowalla network contains substantially more nodes and edges, which is 196,591 nodes and 950,327 edges. Since friendship is a mutual connection between two people, all edges are bidirectional and both networks are undirected. It is observed that the ratio between the number of edges and the number of nodes is lower in Brightkite compared to Gowalla. This means that the node degree, which is the number of edges connected to a node [29], of most nodes in Brightkite is also lower than that of most nodes in Gowalla. The average node degree of an undirected network can be calculated using the following formula:

$$\boldsymbol{k = \frac{2m}{n}} \tag{6.1}$$

Where *m* represents the number of edges and *n* represents the number of nodes in the network. As a result, the average node degree of Brightkite and Gowalla is calculated to be 7.35 and 9.67 respectively. This means that on average each user has about 7 friends in Brightkite dataset and about 9 to 10 friends in Gowalla dataset. To further understand the two networks, network density

is calculated. Network density is the ratio between actual connections and potential connections in a network [30]. In a network where all nodes are connected to every other node, the network density is 1. Higher network density signals a tighter-knit community and vice versa. The formula to calculate density in an undirected network is as follows:

$$\boldsymbol{d = \frac{2m}{n(n-1)}} \quad (6.2)$$

Here, *2m* represents the number of edges in the network and *n(n-1)* represents the number of possible edges. Using this formula, the calculated network density of Brightkite dataset is 0.000126 and that of Gowalla dataset is 0.000049.

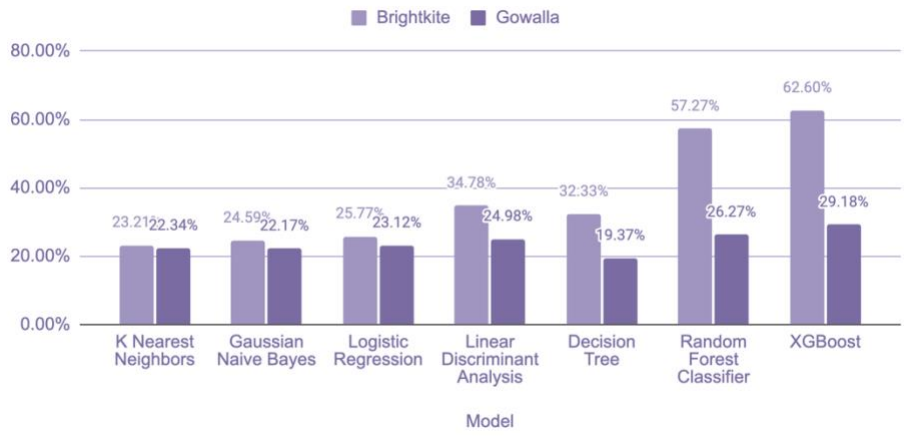
The next step is to perform community detection in order to separate nodes into their own communities. A NetworkX partition algorithm, which is called the best partition, is utilized. It partitions the graph nodes by trying to maximize the modularity of each community using Louvain Heuristic [31]. A visualization of the network graph that belongs to Brightkite community #60 is shown in Figure 6.1. This community is chosen for visualization due to its small size to distinguish individual node and edge. Brightkite community #60 only contains 319 node and 581 edges as opposed to 8779 nodes and 29,573 edges in Brightkite community #0. After the partition, communities are ranked by size and density to ensure that there are enough data points in a community for model training. Brightkite community #0 and #1, and Gowalla community #17 are selected for experimenting. The same steps described in section 6.1 and 6.2 are used to carry put the experimentation.

Figure 6.1 Brightkite Community #60 Network Graph

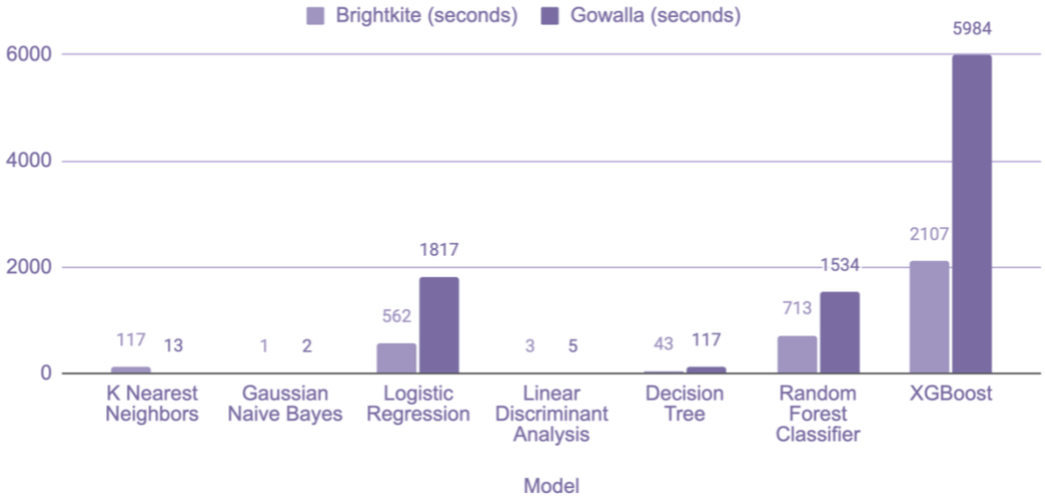
6.4 Results

6.4.1 Without Social Aspect

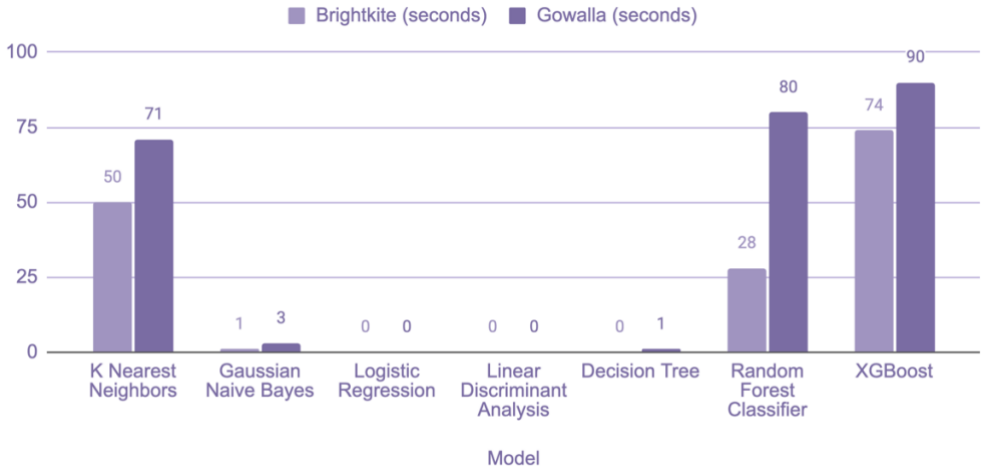
Figure 6.2a shows and compares the accuracy produced by the models on both Brightkite and Gowalla datasets. From observation, XGBoost has the highest accuracy on both Brightkite and Gowalla datasets compared to forementioned methods, 62.6% and 29.18% respectively. However, there is a discrepancy between the two datasets. Brightkite dataset accuracy is 33.42% higher than that of Gowalla dataset.



(a)



(b)



(c)

Figure 6.2 Accuracy (a), Training Time (b) and Predicting Time (c) Plots of Different Models on Brightkite and Gowalla Datasets

Figure 6.2b shows and compares the training time of the chosen classifiers on both datasets. The training time trend of the models is similar between Brightkite and Gowalla. Since Gowalla dataset has about 2 million data points more compared to Brightkite, it takes longer to complete training the models. This is reflected in the graph. A similar trend can also be seen in Figure 6.2c, which shows and compares the predicting time of the classifiers on both datasets.

Figure 6.3 shows and compares the importance of training features. Among all features, latitude, longitude, and category are the most relevant for the prediction of next activity in both datasets. In addition, user ID and distance to next check-in are also important in Brightkite dataset.

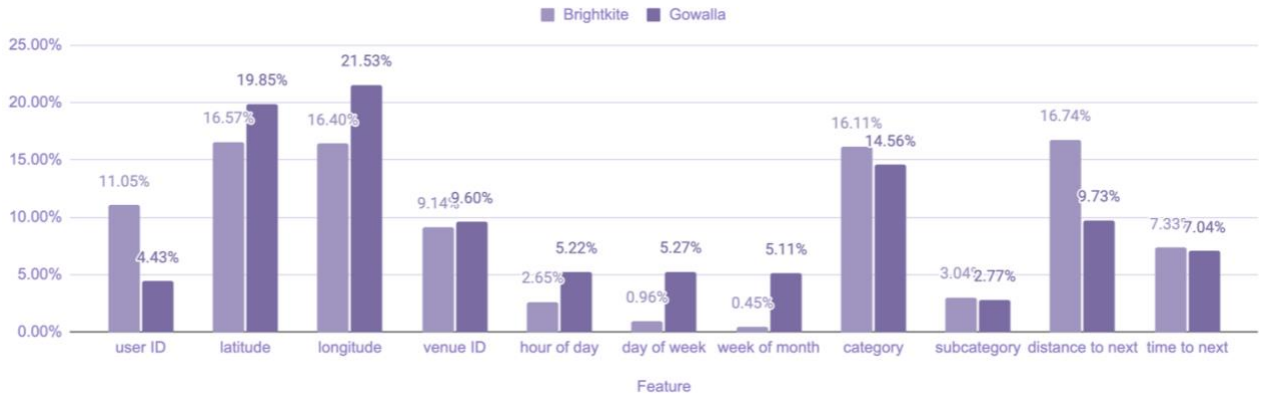
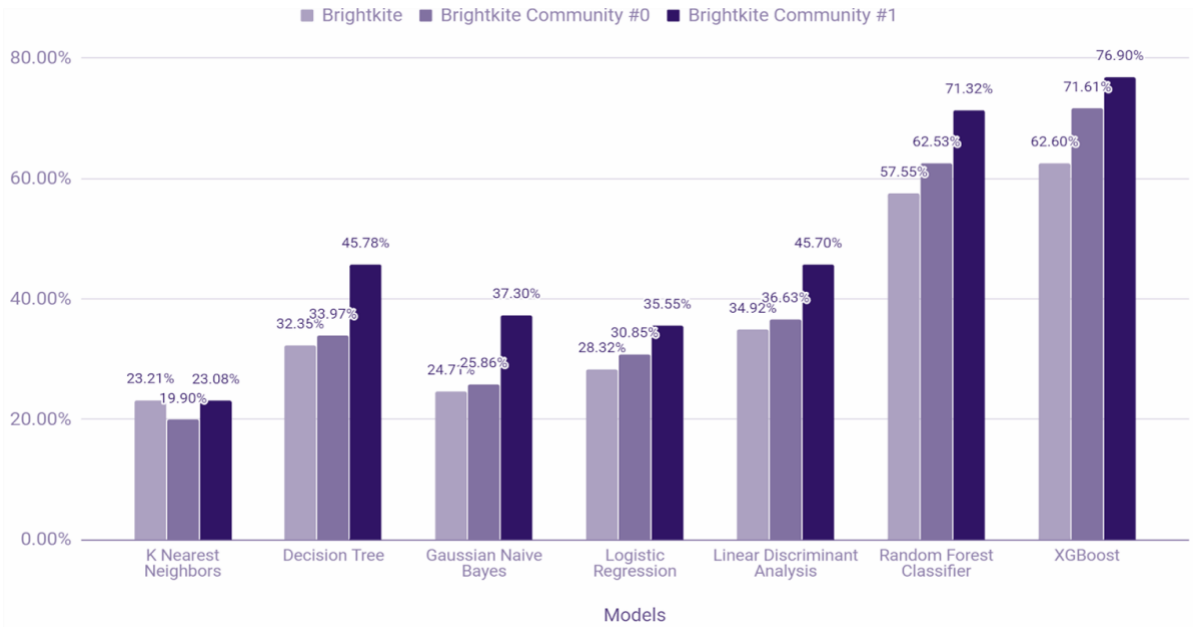


Figure 6.3 Feature Importance Plot

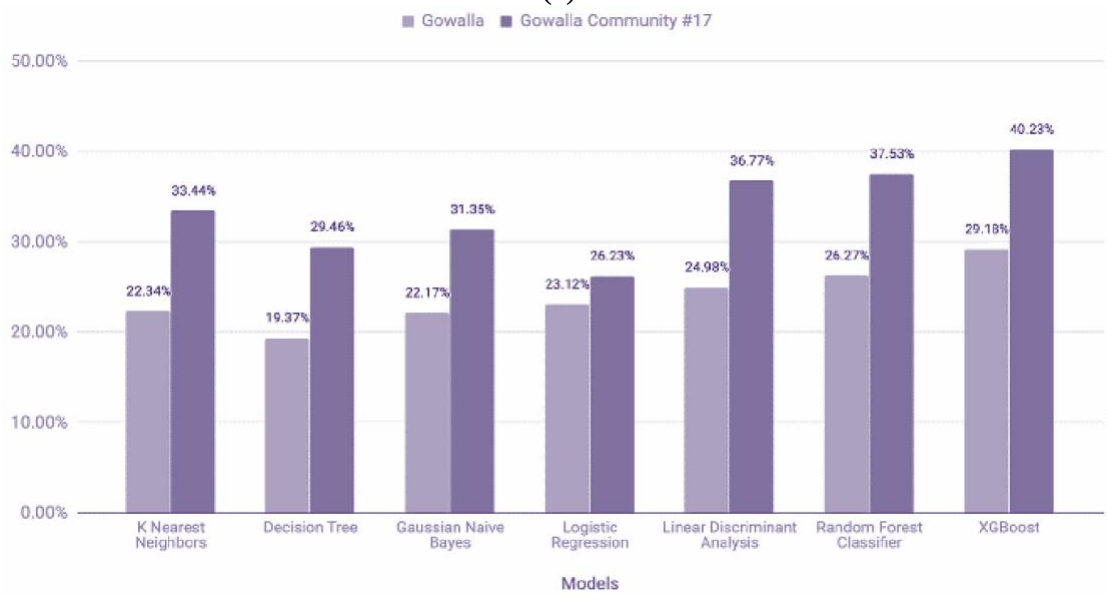
6.4.2 With Social Aspect

For Brightkite dataset, the accuracy of all models is improved after factoring in social aspect of LBSNs. Specifically, XGBoost model's accuracy is improved to 76.90% from 62.60% comparing between using the entire Brightkite dataset and only using data from community #1 (Figure 6.4a). The same results if observed on Gowalla dataset in Figure 6.4b. Social and friendship related information are confirmed to be more relevant and effective when working with a tight-knit network as opposed to a sparse one. The accuracy of all trained models is improved after only using data within a community. It is learned that choosing which communities to use as training and validating data solely based on the size of the communities is not effective. The reason being

that a large number of users within a community does not mean that there is a large number of check-ins due to the difference in check-in frequency of each user. Therefore, incorporating density alongside with community size yields better results.



(a)



(b)

Figure 6.4 Comparing Accuracy of Different Models Between Using the Whole Dataset versus Using Only Data Within a Specific Community: (a) Brightkite; (b) Gowalla

Chapter 7. DISCUSSION

In this chapter, we will discuss challenges faced with the datasets and additional work that was done to expand on the project. This includes a new method and the result from running several experiments. The reason for this new method to not be a part of the main paper is because the work is unfinished. However, it does provide new insights into the datasets and a novel way to represent the relationships between a user and a venue as well as between a venue and its activity category.

7.1 Datasets

As shown in Table 3.1 of Chapter 3, Brightkite check-ins were collected from April 2008 to October 2010 and Gowalla check-ins were collected from February 2009 to October 2010. Based on the time, we can see that both datasets are more than a decade old. With the rise in popularity of LBSNs, there should be newer check-in datasets available for experimentation. However, obtaining new datasets is challenging due to data privacy concerns. The datasets used were the best option available at the time of project inception. In an effort to extend on the datasets, we tried a few things. The first was using Google Places API [32] and Foursquare API to mine additional venue attribute so that new features can be extracted. Some of the venue attributes that we were interested in mining are city, state, country, zip code, name, business hours, ratings, reviews, and venue type. One of the issues was matching the venues in the existing datasets to Google Places API and Foursquare API venues so that we can collect venue details. Since the existing venue IDs are different from both Google Places API and Foursquare API, we tried using both coordinate and venue category/type to help with matching. However, this is not a viable solution because we were not able to find any exact match using these two pieces of information alone. Another issue faced was that there is a daily limit on the number of requests made to the APIs, so it would take

months just to collection additional venue data before we can start experimenting. In addition, we attempted to mine user data, such as gender, age group, home city, home state, interests, preferences, and similar issues were faced. We found that user data has an even higher level of privacy concerns compared to venue data because we are not allowed to access any user information without user permission. As a result, we decided to proceed with the already available information from the original datasets.

7.2 Graph Convolutional Network and Heterogeneous Graph

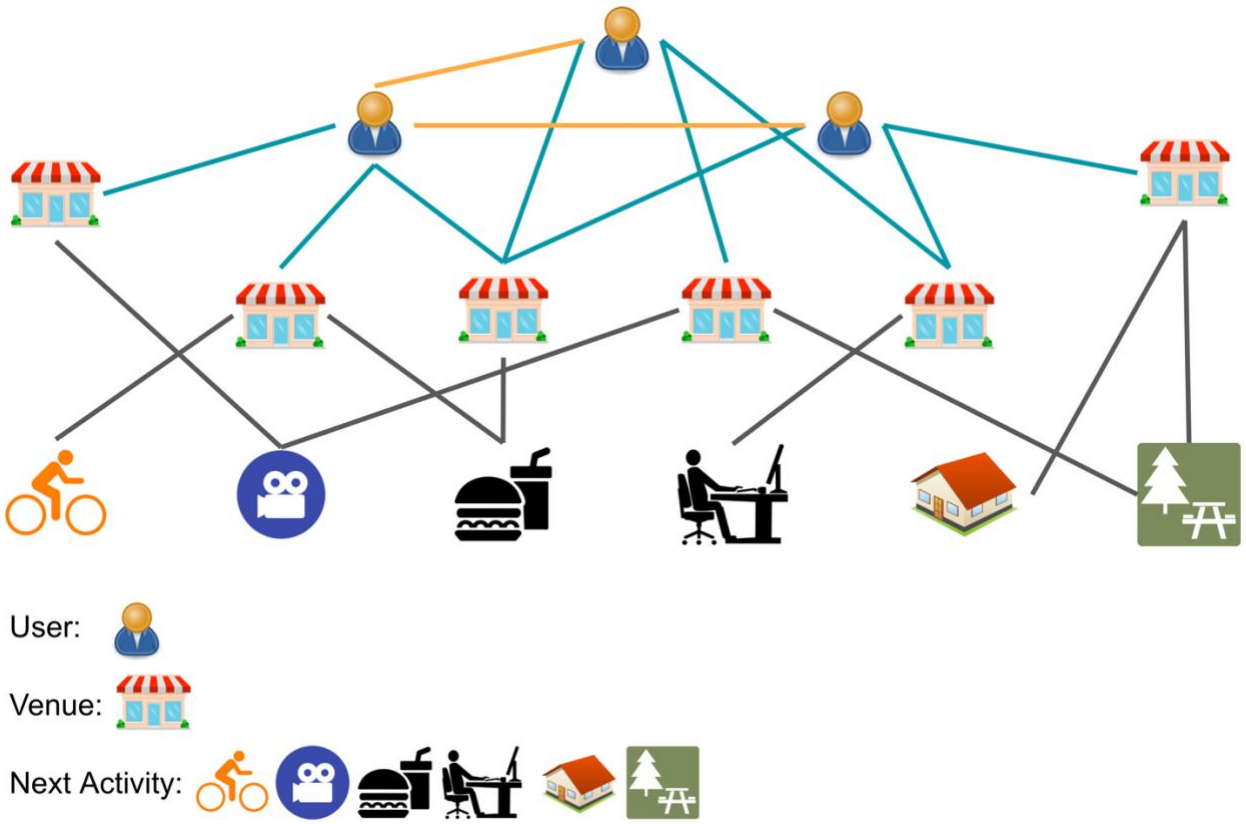


Figure 7.1 Heterogeneous Graph Design

In an effort to further improve on the prediction accuracy and challenge ourselves to look at the problem from a different angle, we set out to review some additional literature. Inspired by a paper by Wu et al., which discusses a model that utilizes Graph Convolutional Networks (GCNs) on a

user mobility heterogeneous graph to infer social relationship [33], we investigated applying similar idea to infer user next activity choice.

7.2.1 Heterogeneous Graph and Graph Convolution Network

A heterogeneous graph is an information network that contains different types of nodes and/or multiple types of edges [34]. Each of these different nodes and edges tends to have different types of attributes that represents its own characteristics. As a result, the model that represents each type of nodes or edges can have a different number of dimensions. Figure 7.1 contains 3 types of nodes that represent users, venues, and next activity categories. Edges that connect these entities represent the following types of relationships:

- Users with other users corresponding to friend-of relationships
- Users with venues corresponding to visited-by relationships
- Venues with next activity categories corresponding to next-of relationships

To implement a heterogeneous graph, we use a library called Deep Graph Library (DGL) [35]. The first step in creating the graph is creating three dictionaries that represent the 3 types of relationships mentioned. They are then passed into a DGL API to create a heterogeneous graph. Each node already contains its own ID since each edge is represented by grouping an ID of one node to an ID of another node in a tuple. To add to the graph the information about the characteristics of each type of nodes and edges, we add the attributes to their corresponding entities. Table 7.1 contains a list of node types and their attributes. Table 7.2 contains a list of edge types and their attributes.

Table 7.1 Heterogeneous Graph Node Attributes

Node Type	Attributes
User	• User ID
Venue	• Place ID • Latitude • Longitude • Activity Category
Next Activity Category	• Activity ID • Activity Category

Table 7.2 Heterogeneous Graph Edge Attributes

Edge Type	Attributes
Friend-of	• Number of venues visited together
Visit-by	• Hour of day • Day of week
Next-of	• User ID • Venue ID • Distance to next location • Time elapse from current venue to next location

GCN is a type of Graph Neural Network that enables convolutions on arbitrary structured graphs [33]. The ability of propagating information from one layer to another enables GCNs to learn localized patterns at multiple levels. By combining both the graph structure and node attributes, they can learn predictive embeddings. In the next sections, we will go over the model training setup, different experiments ran and the results.

7.2.2 Experiments and Results

Several experiments were carried out to test the effectiveness of the combination of GCN and heterogeneous graph. We utilize an example from the DGL documentation website to construct a Relational GCN [34] by passing in the heterogeneous graph built in section 7.2.1. Data from 1 Brightkite community and 3 Gowalla communities were used for training, validation, and testing.

As showed in Table 7.3, the overall accuracy for Brightkite #0, Gowalla #0, Gowalla # 5 and Gowalla #9 are 55.07%, 45.29%, 57.97% and 51.67% respectively. Comparing to XGBoost performance on Brightkite #0 at 71.61%, heterogeneous graph method's accuracy of 55.07% is substantially lower.

The same experiment set up is applied to each user within the mentioned community to see how the model would perform. By doing this, we recognize that the social aspect is remove from the model because the graph now only contains information of a single user, whom has no connection to other users. The only information that this graph has is the venues that this user visited and the activity the user chose to do after each visit. Our reasoning for doing this is that this will help us filtering out individuals, whom we do not have enough information about to be able to make accurate predictions on. One of the reasons could be that these users' check-ins do not have any correlation. Once this experiment has completed, a list of accuracy per user is produced. We decided to keep only users, whose prediction accuracy is above 70%, to see if it will increase the overall accuracy. We will refer to these users as "high accuracy users" from now on. We see that the overall accuracy increases. The community with the most significant increase in accuracy is Gowalla #0 at 93.35%, followed by Gowalla #9 at 72.84%, Brightkite #0 at 62.64%, and finally Gowalla #5 at 61.41%. Among, the four communities, Gowalla #0 has the highest median of number of check-ins, which is 69, compared to Brightkite #0 at 45, Gowalla #5 at 50 and Gowalla #9 at 54. Having a high median number of check-ins could be the factor that contribute to having the highest accuracy.

Table 7.3 Heterogeneous Graph Experiments and Results

	Brightkite #0	Gowalla #0	Gowalla #5	Gowalla #9
Overall accuracy	55.07%	45.29%	57.97%	51.67%
Total number of users	3,900	2,808	1,609	1,446
Number of users with high accuracy	1,838	490	264	257
High accuracy user percentage	47.18%	17.77%	16.41%	17.77%
Accuracy of keeping only high accuracy users	69.20%	93.35%	61.41%	72.84%
Accuracy of keeping only high accuracy users and removing venues with only 1 visit	62.64%	92.8%	43.11%	51.37%
Average number of Check-ins	201	188	131	134
Min number of Check-ins	10	10	10	10
Max number of Check-ins	2,100	2,100	1,877	2,133
Median number of Check-ins	45	69	50	54
Mode number of Check-ins	11	25	25	25

The next thing that we tried was removing check-ins where the associated venue has only been visited once by a user in the entire dataset. We thought that this would help us with improving the accuracy. However, we observe that the accuracy drops, and it drops quite a bit in a couple of communities. While the accuracy only drops 0.55% for Gowalla #0 and 6.56% for Brightkite #0, it drops 18.3% for Gowalla #5 and 21.47% for Gowalla #9. Removing these data points seem to be a significant loss when training using data from Gowalla #5 and #9 when considering accuracy. The reason for this decrease in accuracy was probably because that we remove some useful data points that was helping the model. By removing venues that a user only visits once, we left out the sentiment suggesting that this user might not like this venue and its associated activity category. With this information, the model could have been discouraged to predict the next activity that is offered at the venue that the user does not want to visit again.

To conclude, the combination of using GCN and heterogeneous graph design along with keeping only high accuracy users seems promising based on the fact that we can achieve a record high accuracy of 93.35% for Gowalla #0 among all things that we have tried so far. On the other hand, heterogeneous graph approach does not seem to work as well as XGBoost because compared to the accuracy of 69.20% when using heterogeneous graph for Brightkite #0, the accuracy of using XGBoost is 71.61% for the same community. At this point, the only thing that can be said is that the result really depends on the community, and it would be helpful to figure out the reasons for such high accuracy in Gowalla #0.

Chapter 8. CONCLUSIONS

LBSNs provides a vast amount of user spatial-temporal and mobility data. This paper presents an approach that utilizes machine learning techniques along with taking advantage of the social network graphs of two LBSN datasets to accurately predict the user's next choice of activity. Firstly, cluster analysis is performed to group users in different clusters based on their entire historical check-ins pattern. Being able to visualize different clusters, which represent different mobility patterns, allows for understanding the diversity in user temporal and spatial preference. This enables the creation of a more targeted prediction model. The next step is experimenting with Naïve Neural Network. Taking it a step further, some additional features are extracted from the original datasets. The derived features provide more insight into user mobility pattern and activity preference. Then, a simple neural network is modeled and is used to make predictions on a user's next activity choice. After a few initial experiments, the prediction results came out to be quite decent in terms of accuracy. However, there is still room for improving the model to produce better results. In the pursuit of refining the prediction accuracy, this research further utilizes XGBoost algorithm. Significant improvement in prediction accuracy compared to previous models is observed from the yielded results. Going into the discussion section of this paper, we attempt to improve the accuracy by introducing the use of heterogeneous graph design in solving this prediction problem. We saw a record high accuracy in one community among those that were examined, even when comparing to XGBoost. Experiments on other communities return mediocre results. Therefore, it is inconclusive whether this new approach is better than XGBoost.

In the future, there are a number of works that can be done. The first improvement is to run experiments on the same set of communities using both XGBoost and GCN with heterogeneous graph so that we can compare them fairly. Secondly, figuring out the reason behind the discrepancy

in accuracy between Brightkite and Gowalla is beneficial because one of the goals is to be able to apply the method to any LBSN datasets. The third research problem that we want to explore in the future is looking at the precision and recall of all forementioned approach because we want to look at the models holistically and accuracy alone does not represent all aspect of any one method. Lastly, it would be informative to understand the reason why heterogeneous graph was able to achieve such high accuracy on Gowalla community #0. One way to achieve this is to compare Gowalla #0 to others to determine unique characteristics that could attribute to the vast improvement on accuracy.

BIBLIOGRAPHY

[1] Google, "Google Maps," Google, [Online]. Available: https://www.maps.google.com. [Accessed 6 February 2022].

[2] E. Cho, S. A. Myers and J. Leskovec, "Friendship and Mobility: User Movement in Location-Based Social Networks," in *ACM*, San Diego, 2011.

[3] Y. Chen, A. Frei and H. S. Mahmassani, "Exploring activity and destination choice behavior in social networking data," in *Transportation Research Board 94th Annual Meeting*, Washington DC, 2015.

[4] D. Lian and X. Xie, "Collaborative activity recognition via check-in history," in *LBSN '11: Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, Chicago, 2011.

[5] H. Cheng, J. Ye and Z. Zhu, "What's your next move: User activity prediction in location-based social networks," in *Proceedings of the 2013 SIAM International Conference on Data Mining (SDM)*, Austin, 2013.

[6] F. Pianese, X. An, F. Kawsar and H. Ishizuka, "Discovering and predicting user routines by differential analysis of social network traces," in *2013 IEEE 14th International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Madrid, Spain, 2013.

[7] A. Noulas, C. Mascolo and E. Frias-Martinez, "Exploiting foursquare and cellular data to infer user activity in urban environments," in *2013 IEEE 14th International Conference on Mobile Data Management*, Milan, Italy, 2013.

[8] "Foursquare," Foursquare, [Online]. Available: www.foursquare.com. [Accessed 20 July 2019].

[9] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, California, 2016.

[10] J. Chang and E. Sun, "Location 3: How users share and response to location-based data on social networking sites," *ICWSM*, pp. 74-80, 2011.

[11] J. Pitman and M. Yor, "The two-parameter poisson-dirichlet distribution derived from a stable subordinator," *The Annals of Probability*, vol. 25, no. 2, pp. 855-900, 1997.

[12] H. Gao, J. Tang and H. Liu, "Exploring social-historical ties on location-based soial networks," in *Vol. 6 No. 1 (2012): Sixth International AAAI Conference on Weblogs and Social Media*, Dublin, Ireland, 2012.

[13] W. Fan, H. Sallay and N. Bouguila, "Online learning of hierarchical Pitman-Yor process mixture of generalization Dirichlet distributions with feature selection," *IEEE Transactions of Neural Networks and Learning Systems,* vol. 28, no. 9, pp. 2048-2061, 2016.

[14] A. Noulas, S. Scellato, N. Lathia and C. Mascolo, "Mining user mobility features for next place predicton in location-based services," in *2012 IEEE 12th International Conference on Data Mining*, Brussels, Belgium, 2012.

[15] A. Likhyani, D. Padmanabhan, S. Bedathur and S. Mehta, "Inferring and exploiting categories for next location prediction," in *WWW '15 Companion: Proceedings of the 24th International Conference on World Wide Web*, Florence, Italy, 2015.

[16] K. Wang, L. Gu, S. Guo, H. Chen, V. C. M. Leung and Y. Sun, "Crowdsourcing-based content-centric network: a social perspective," *IEEE Network,* vol. 31, no. 5, pp. 28-34, 2017.

[17] W. Li, C. Eickhoff and A. P. d. Vries, "Want a coffee?: predicting users' trails," in *SIGIR '12: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, Portland, Oregon, 2012.

[18] D. Yang, D. Zhang, V. Zheng and Z. Yu, "Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 45, no. 1, pp. 129-142, 2015.

[19] V. Kounev, "Where will I go next?: Predicting future categorical check-ins in location based social networks," in *8th International Conference on Collaborative Computing: Networking, Applications and Worksharing*, Pittsburgh, Pennsylvania, 2012.

[20] J. Cao, S. Xu, X. Zhu, R. Lv and B. Liu, "Effective fine-grained location prediction based on user check-in pattern in LBSNs," in *2017 Fifth International Conference on Advanced Cloud and Big Data (CBD)*, Shanghai, China, 2017.

[21] M. Nedrich, "Mean Shift Clustering," Atomic Object, 26 May 2015. [Online]. Available: https://spin.atomicobject.com/2015/05/26/mean-shift-clustering/. [Accessed 6 February 2022].

[22] L. E. Peterson, "K-nearest neighbor," Scholarpedia, 21 February 2009. [Online]. Available: http://scholarpedia.org/article/K-nearest_neighbor. [Accessed 7 February 2022].

[23] T. Hastie, R. Tibshirani and J. Friedman, Elements of Statistical Learning, Springer, Springer, 2009.

[24] H. Zhang, "The Optimality of Naive Bayes," in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, Miami Beach, Florida, 2004.

[25] S. Swaminathan, "Logistic Regression - Detailed Overview," Medium, 15 March 2018. [Online]. Available: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc. [Accessed 7 February 2022].

[26] S. Raschka, "Linear Discriminant Analysis," 3 August 2014. [Online]. Available: https://sebastianraschka.com/Articles/2014_python_lda.html. [Accessed 7 February 2022].

[27] T. Yiu, "Understanding Random Forest," Medium, 12 June 2019. [Online]. Available: https://towardsdatascience.com/understanding-random-forest-58381e0602d2. [Accessed 7 February 2022].

[28] "Network X Network Analysis in Python," [Online]. Available: https://networkx.org/. [Accessed 2019].

[29] N. DQ, "Node degree definition," Math Insight, [Online]. Available: http://mathinsight.org/definition/node_degree. [Accessed 2019].

[30] "What is Network Density - and How Do You Calculate It?," The Vital Edge, 13 June 2013. [Online]. Available: https://www.the-vital-edge.com/what-is-network-density/.. [Accessed 2019].

[31] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment,* vol. 2008, 2008.

[32] "Google Places API," Google, [Online]. Available: https://developers.google.com/maps/documentation/places/web-service. [Accessed 2 May 2018].

[33] Y. Wu, D. Lian, S. Jin and E. Chen, "Graph Convolutional Networks on User Mobility Heterogeneous Graphs for Social Relationship Inference," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19),* Macao, China, 2019.

[34] Q. Gan, M. Wang, M. Li, G. Karypis and Z. Zhang, "Working with Heterogeneous Graphs," 2018. [Online]. Available: https://docs.dgl.ai/en/0.6.x/tutorials/basics/5_hetero.html. [Accessed 2019].

[35] M. Wang, D. Zheng, G. Gan, M. Li, Z. Ye, C. Ma, J. Zhou, X. Song, T. Xiao, T. He, W.-m. Ye, G. Karypis and Z. Zhang, "Deep Graph Library," 07 12 2018. [Online]. Available: dgl.ai. [Accessed 23 01 2022].