

M.SC. ENGG. THESIS

Developing Techniques for Weather-Aware Prediction of User Activities and Future Visits

by
Samia Nawshin

Submitted to

Department of Computer Science and Engineering
in partial fulfilment of the requirements for the degree of
Master of Science in Computer Science and Engineering



Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology (BUET)

Dhaka - 1000

September 2018

Dedicated to my loving parents


AUTHOR'S CONTACT

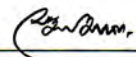
Samia Nawshin


Email: samia.nawshin@gmail.com


The thesis titled “Developing Techniques for Weather-Aware Prediction of User Activities and Future Visits”, submitted by Samia Nawshin, Roll No. **0412052033 P**, Session April 2012, to the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, has been accepted as satisfactory in partial fulfillment of the requirements for the degree of Master of Science in Computer Science and Engineering and approved as to its style and contents. Examination held on September 29, 2018.


Board of Examiners

1. 

Dr. Mohammed Eunus Ali
Professor (Supervisor)
Department of CSE, BUET, Dhaka - 1000.
2. 

Dr. Md. Mostofa Akbar
Head and Professor Member
(Ex-Officio)
Department of CSE, BUET, Dhaka - 1000.
3. 

Dr. M. Sohel Rahman
Professor Member
Department of CSE, BUET, Dhaka - 1000.
4. 

Dr. Rifat Shahriyar
Assistant Professor Member
Department of CSE, BUET, Dhaka - 1000.
5. 

Dr. Nova Ahmed
Associate Professor Member
Department of EEE (External)
North South University, Dhaka - 1219.

Candidate's Declaration

This is hereby declared that the work titled “Developing Techniques for Weather-Aware Prediction of User Activities and Future Visits” is the outcome of research carried out by me under the supervision of Dr. Mohammed Eunos Ali, in the Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, Dhaka - 1000. It is also declared that this thesis or any part of it has not been submitted elsewhere for the award of any degree or diploma.



Samia Nawshin

Candidate

Acknowledgment

First and foremost I offer my sincerest gratitude to my supervisor, Dr. Mohammed Eunos Ali, who has supported me throughout my thesis with his patience, motivation, enthusiasm, and immense knowledge. He helped me a lot in every aspect of this work and guided me with proper directions whenever I sought one. I could not have imagined having a better supervisor and mentor for my M.Sc. study and research. His patient hearing of my ideas, critical analysis of my observations and detecting flaws (and amending thereby) in my thinking and writing have made this thesis a success.

I would also want to thank the members of my thesis committee for their valuable suggestions. I thank Dr. Md. Mostofa Akbar, Dr. M. Sohel Rahman, Dr. Rifat Shahriyar and specially the external member Dr. Nova Ahmed.

In this regard, I remain ever grateful to my beloved parents, who always exist as sources of inspiration behind every success of mine I have ever made. I would also like to express my sincere thank to my loving spouse for his support and inspiration.

Abstract

In recent years, the accelerated escalation of smart phones has led to the increasing popularity of Location-Based Social Networking(LBSN) sites such as Foursquare, Facebook Places, Twitter etc. LBSNs allow and encourage users to publish information about their current location or visiting places through check-ins and offer them to associate their posts and photos with their check-ins and share with their friends and family as well as tagging them. These produce fast growing, fine-grained and vast in volume data and provide a means of user profiling and modeling. Huge volume of user generated data of social media presents an opportunity to find interesting insights about users' preferences of places at different times. Prediction of users' daily routine, finding users' location preference, identification of users' mobility patterns from the check-in datasets covers the current state of the art. Plethora of works have been done to find such comprehensions about user activity from check-in dataset of social media by considering various aspects such as, frequency of check-ins, time of check-ins, venue of check-ins etc. The knowledge of such intuitions about users' preferences of places or activities has wide range of applications covering social media commerce, targeted advertisement, influencer marketing etc. Among all the works done so far, no one considers the influence of weather on human life while predicting or finding users' mobility or activity pattern, though its effect is enormous. Earlier psychological studies show that weather has a strong influence on human life and its consideration for users' whereabouts and whatabouts prediction is more constructive and pragmatic. Motivating from all these observations we propose the first approach to find user activity and mobility pattern from social media data based on weather condition. In this thesis, we develop several machine learning based models to predict future activity, visiting places and travel mode of users' from previous users' check-ins and travel patterns on a given weather condition, for example, a user may prefer to visit sea beaches on a sunny weather, whereas an indoor entertainment on a rainy weather. Again he or she may prefer cycling on a clear weather whereas taxi or private car on a rainy weather.

Table of Contents

<i>Board of Examiners</i>	ii
<i>Candidate's Declaration</i>	iii
<i>Acknowledgment</i>	iv
<i>Abstract</i>	v
1 Introduction	1
1.1 Motivation and Application	3
1.2 Research Objectives	5
1.3 Research Challenges and Solution Overview	6
1.4 Contributions	7
1.5 Outline	7
2 State of the Art	9
2.1 Collection of dataset	9
2.2 Discovering and predicting users daily routine and life-style patterns	10
2.3 Predicting future activities or inferring activity preferences	12
2.4 Clustering socially relevant venues	12
3 Problem Formulation	14
3.1 Overview of our system architecture	14
3.2 Our research problem	15

4	Methodology	16
4.1	Overview of Our Approach	17
4.2	Extracting location based dataset	18
4.3	Extracting weather information	19
4.4	Data Preprocessing	19
4.5	Computing statistical significance	20
4.6	Building classification models	21
5	Building Model for Transport Mode Prediction	22
5.1	Feature selection	23
5.1.1	Feature selection for categorical weather information and categorical transport mode	23
5.1.2	Feature selection for numerical weather information and categorical transport mode	25
5.1.3	Feature selection using automated subset selection method	27
5.2	Building the Classification Model	28
5.3	Limitation of the Build Model	29
5.4	Handling Class Imbalance Problem Using SMOTE Algorithm	31
5.5	Re-Building the Classification Model	34
5.6	Discussion	34
6	Building Model for Day-Time Activity Prediction	37
6.1	Feature selection	38
6.1.1	Feature selection for categorical weather information and categorical day-time activity	38
6.1.2	Feature selection for continuous weather information and categorical day-time activity	39
6.1.3	Feature selection using automated subset selection method	41
6.2	Building the Classification Model	41
6.3	Discussion	43

7	Building Model for Night-Time Activity Prediction	45
7.1	Feature selection	45
7.1.1	Feature selection for categorical weather information and categorical night-time activity	46
7.1.2	Feature selection for continuous weather information and categorical night-time activity	47
7.1.3	Feature selection using automated subset selection method	48
7.2	Building the Classification Model	48
7.3	Discussion	50
8	Building Model for Future Visit Prediction	52
8.1	Feature selection	53
8.1.1	Feature selection for categorical weather information and categorical visiting place	53
8.1.2	Feature selection for continuous weather information and categorical visiting place	54
8.1.3	Feature selection using automated subset selection method	56
8.2	Building the Classification Model	56
8.3	Handling Class Imbalance Problem Using SMOTE Algorithm	58
8.4	Re-Building the Classification Model	59
8.5	Discussion	59
9	Conclusions	62
	References	64

List of Figures

1.1	User activities on different weather condition	2
1.2	Users' food preferences on different weather condition	4
3.1	System architecture	14
3.2	Research problem	15
4.1	Overview of our approach	17
4.2	Extracting weather information	19
4.3	Merging of similar weather categories	20
5.1	Generation of synthetic instances with the help of SMOTE	32

List of Tables

5.1	Dataset of Tokyo for building the classification model for preferable transport mode	22
5.2	Chi-Square test between transport mode and weather summary	24
5.3	Chi-Square test between transport mode and weather icon	24
5.4	Pooled Within-Groups Matrices	26
5.5	Fisher’s linear discriminant function coefficients between weather attributes and transport mode	27
5.6	Feature selection in R using exhaustive selection algorithm	28
5.7	Selected features for the classification model	29
5.8	Performance of different classifiers for building the model preferable transport mode	29
5.9	Classification results of the model preferable transport mode build using Random Forest algorithm	30
5.10	Up-Sampled dataset after applying SMOTE	33
5.11	Classification results of the model preferable transport mode after up-sampling the dataset using SMOTE algorithm	33
5.12	Dataset of New York city used for building the classification model for preferable transport mode	34
5.13	Performance of different classifiers for building the model preferable transport mode on New York dataset	35
5.14	Classification results of the model preferable transport mode build using Random Forest algorithm on New York dataset	35
6.1	Dataset of Tokyo for building the classification model for preferable day-time activity	37
6.2	Chi-Square test between preferable day-time activity and weather summary	38

6.3	Chi-Square test between preferable day-time activity and weather icon	39
6.4	Pooled Within-Groups Matrices	40
6.5	Fisher’s linear discriminant function coefficients between weather attributes and day-time activity	40
6.6	Selected features for the classification model	41
6.7	Performance of different classifiers for building the model preferable day-time activity	42
6.8	Classification results of the model preferable day-time activity build using Random Forest algorithm	42
6.9	Dataset of New York city for building the classification model for preferable day-time activity	43
6.10	Performance of different classifiers for building the model preferable day-time activity on New York dataset	43
6.11	Classification results of the model preferable day-time activity build using Random Forest algorithm on New York dataset	44
7.1	New York city dataset for building the classification model for preferable night-time activity	45
7.2	Chi-Square test between preferable night-time activity and weather summary	46
7.3	Chi-Square test between preferable night-time activity and weather icon	46
7.4	Fisher’s linear discriminant function coefficients between weather attributes and night-time activity	48
7.5	Selected features for the classification model	49
7.6	Performance of different classifiers for building the model preferable night-time activity	49
7.7	Classification results of the model preferable night-time activity build using Random Forest algorithm	49
7.8	Tokyo city dataset for building the classification model for preferable night-time activity	50
7.9	Performance of different classifiers for building the model preferable night-time activity on Tokyo city dataset	50
7.10	Classification results of the model preferable night-time activity build using Random Forest algorithm on Tokyo city dataset	51
8.1	Tokyo city dataset for building the classification model for preferable visiting places .	52

8.2	Chi-Square test between preferable visiting place and weather summary	53
8.3	Chi-Square test between preferable visiting place and weather icon	54
8.4	Pooled Within-Groups Matrices	55
8.5	Fisher’s linear discriminant function coefficients between weather attributes and visiting place	55
8.6	Selected features for the classification model	56
8.7	Performance of different classifiers for building the model preferable visiting place . . .	57
8.8	Classification results of the model preferable visiting place build using Random Forest algorithm	57
8.9	Up-Sampled dataset after applying SMOTE	58
8.10	Performance of different classifiers for building the model preferable visiting place after up-sampling	58
8.11	Classification results of the model preferable visiting place after up-sampling the dataset using SMOTE algorithm	59
8.12	Dataset of New York city used for building the classification model for preferable visiting place	60
8.13	Performance of different classifiers for building the model preferable visiting place on New York dataset	60
8.14	Classification results of the model preferable visiting place build using Random Forest algorithm on New York dataset	61

Chapter 1

Introduction

In recent years social networking sites such as Foursquare, Flickr, Facebook Places, Twitter are becoming increasingly popular with the proliferation of smart phones. These social medias facilitate the sharing of ideas and information and the building of virtual networks and communities. Though social media originated as a tool that people used to interact with friends and family but later adopted by businesses that wanted to take advantage of a popular new communication method to reach out to customers. Social media has the ability to connect and share information among the people coming from anywhere on earth. There is so much information on the Internet about everything and anything but the most valuable information are getting generated by social media data. Recently the development and growth of social media data rises to a point where it is now regarded as ubiquitous. Facebook has more than 2.2 billion monthly active users as of January 2018 while Twitter has more than 500 million tweets sent each day. All the status updates, pictures, and videos posted and shared by people on their social media contains prodigious information. Sometimes the information are hidden and sometimes evident. They contain rich information about users demographics, likes, dislikes, etc. Social media data has a range of attributes associated with them which are not found in “offline” data. The variety and potential size of social media data has a great source or mine of wide range of intuitions. As a result of this procreation of easily and quickly accessible social media data, analysts and policymakers in both government and private sectors have begun to consider how such data can be harnessed to support robust evidence-based policymaking. They are continuously interrogating the datasets and their findings are unprecedentedly compelling and useful. One of the biggest applications of analyzing user data is in business for social media marketing. Targeted advertisements are already

pretty commonplace which is the outcome of social media data analysis. It provides an indispensable tool for finding and engaging with customers, sales, advertising and promotion, gauging trends and offering customer service by analyzing and extracting the hidden contents. Predictive analytics is also going to become increasingly popular. By analyzing the big data from social media, it is possible to predict the future of a particular event or attention, which gives a good hand for influencer marketing.

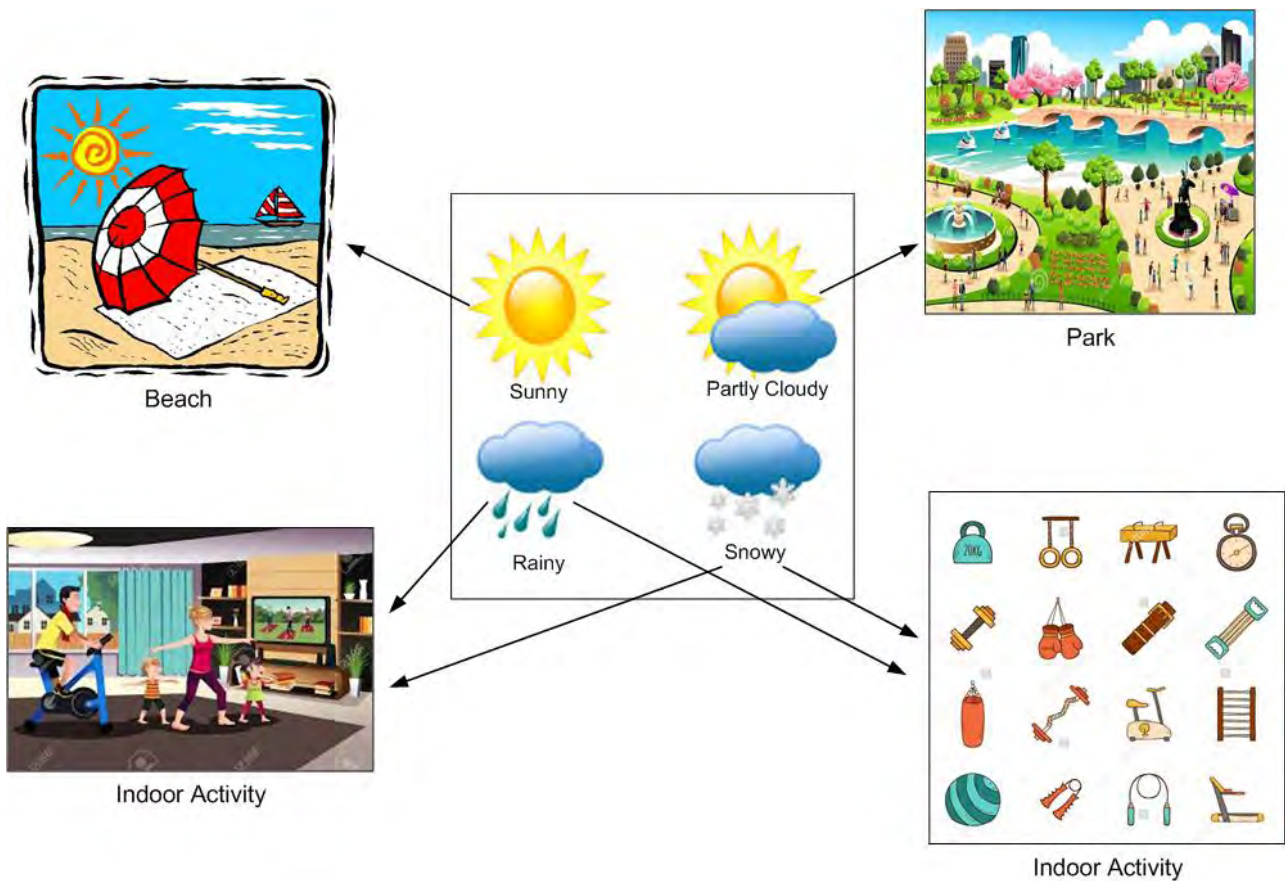


Figure 1.1: User activities on different weather condition

There are numerous types of social media platforms. Among them location-based social networking(LBSN) sites are social networks that use GPS features to locate people and let them broadcast their location and other content from their mobile devices. LBSN sites allow users to share their location information of visiting different venues or places through “check-ins” provides geo-location data. These geo-location data offers information in new ways to understand people’s attitude and engrossment through their activity choices. The “check-in” data contains information about when, why,

where, how people visits, provides a source for mining users' preferences of places or transport mode at different times. Many works have been done to find intriguing information from user check-in datasets of various social networks for example Foursquare, Gowalla, Twitter, Facebook etc. Users share their current location or the venue they have visited earlier with their friends through "check-ins". Prediction of users' daily routine [11], identifying individuals' life-style pattern [12], inference of urban activity pattern [13], finding individuals' mobility pattern [14] from check-in datasets are common research activities now a days. Other stimulating works done on geo-location data are the prediction of future activities [15] [16] [17], finding socially relevant venues of a city [18] [19] [20] etc, cover the present state of the art. Though several works are done on users' activity analysis and future activity prediction from the analysis of check-in datasets no work done so far consider the influence of weather condition on users' behavior or activity. Weather condition affects human life significantly. Years of research in many fields show that weather has a strong influence on human behavior[2][3][4][5]. People may choose to visit places because of a particular weather condition in order to gain personal satisfaction and for a sense of well-being. Motivating from these engaging and insightful works of users' activity or mobility pattern analysis we decide to find peoples' or users' activity preferences, preferable visiting places and modes of transportation from check-in datasets based on a given weather condition.

The rest of the chapter is organized as follows. Section 1.1 briefly discusses about the motivation of our proposed work with its application. Section 1.2 outlines the objectives of our research. Section 1.3 projects our research challenges and solution overview. Then Section 1.4 highlights the contribution of our thesis. Finally, an organization of the remaining chapters are given in Section 1.5.

1.1 Motivation and Application

In our day to day life weather affects practically every single activity whether it is the place we want to visit or the food we want to consume or even the product we want to purchase [1]. In reality, weather affects practically every consumer purchase decision. The food we eat, the clothes we wear, what transportation we use to travel and even what type of house we prefer to live in, can all be influenced by weather condition. Understanding this relationship can pay huge dividends to marketers. By executing weather based marketing campaigns, brands can gain a real competitive advantage. For that, marketers need to know users preferences on different weather conditions. Location Based Social

Networking (LBSN) sites allow users to share their location information of visiting different venues or places, mode of traveling, activity choices through check-ins which produces an enormous size of user generated data. These check-in data provides a great means for understanding people's behavior and concernment. By analyzing user generated check-in datasets, plethora of works have been done to find users' daily routine [11], their life style pattern [12], their places of interests [13] etc.,. Years of research in many fields show that weather has a strong relation to, and influence on, a number of phenomena related to human behavior [2][3][4][5].



Figure 1.2: Users' food preferences on different weather condition

While LBSN sites generate huge volume of check-in data containing many interesting insights about users' activity choices and patterns, weather condition has a very influential effects on user's choices of activities. People generally choose visiting places according to particular weather conditions in order to gain special experiences. For example people prefers to go to coffee shop on cold weather rather than very hot weather, whereas nobody will prefer to go to an ice cream parlor on a snowy day. Particular places may also be attractive to people because of a particular weather condition, for example, a user may prefer to visit sea beaches in a sunny weather, whereas the same user may prefer for an indoor entertainment in a rainy weather. The knowledge of users' preferences of places and activities based on weather conditions, sellers or marketers can decide their policies to attract customers and promote their products. The previous knowledge of users

visited places in a given weather condition, travel agents or planners can predict the future attention of people's visited area, and they can offer or plan different skins to attract user's contemplation.

Changes in weather and climate also have potential affect on the travel behavior of all kind of transport users. Several research efforts studied the effects of weather condition on users' travel pattern or behavior, or the mode of transportation they usually use to travel [6][7][8][9][10]. Travel demand is an important issue in transportation. Weather can influence travel demand in various ways, including diversions of trips to other paths or modes and cancellation of trips. A deeper understanding of how weather conditions affect traffic is essential for policy making. The knowledge of users' preferable transport mode on a given weather condition can give a hand to policy maker in both governmental and nongovernmental sectors.

Motivating from these observations in our thesis we build several machine learning based models to predict future activity, visiting place and preferable transport mode of users' from previous users' check-ins and given weather condition.

1.2 Research Objectives

From previous discussions we have identified the following objectives of our research.

- To build models to find the correlation between weather condition and users' preferred activity, visiting place and transport mode.
- To predict the future activity of user's for a given weather condition from previous users' check-ins.
- To predict user's probable visiting place from previous check-ins and given weather condition.
- To predict user's preferable mode of transport from previous transport choice history and given weather condition.

The possible outcomes from this research are listed below.

- Classification models to identify human activity and mobility pattern or visiting places on different weather condition.

- Classification model to predict users' preferred transport mode for a future visit.
- A tool which will give various business organizations or company to predict user's preferences at a given weather condition and promote their business to targeted customers.

1.3 Research Challenges and Solution Overview

To build the classification models we need to overcome some challenges. Our first challenge is to build the link between check-in dataset with corresponding weather information because no check-in datasets contain weather information. Our collected dataset have users' check-in information only. Each check-in data is associated with its time stamp, its GPS coordinates and its semantic meaning (represented by fine-grained venue-categories). So, we need to collect weather information of every single check-in by using an weather forecast service via an API. We are the first to create a check-in dataset collected from social media having weather information of every single check-in.

Another major challenge of our work is data sparsity problem. There are 251 different category of venues and 38 different weather conditions. As a result instances per venue category has very few number of instances per weather category. For this reason it is very hard to find the correlation between weather and venues. So, we use only those venue categories having a moderate number of instances which are suitable for machine learning based classification model building approach. We also merge same type of weather categories by considering them as similar. For example, weather condition "Clear" and "Partly Cloudy" are almost similar, so we merge this two categories into one class level. Our dataset also suffers from class imbalance problem. We solved this problem using special data re-sampling technique.

Our third challenge is that, our independent variable weather condition has several attributes mixed of both categorical and numerical variable where our dependent variable venue category is categorical variable. So, we face a problem for feature selection as it is different for categorical and numerical independent variables. To overcome this challenges, we use multiple feature selection processes and merge their results to obtain the total set of features for building our classification models.

1.4 Contributions

To the best of our knowledge, we propose the first approach for weather aware prediction of users' activity, visiting place and preferred transport mode. In summary, the contributions of this thesis are as follows:

- We create a check-in dataset associated with the weather information of every single check-in when there is no such datasets available. We collected the weather information of every single check-in of two different datasets of Tokyo and New York city containing 5,73,703 and 2,27,428 instances respectively. Then cross linking these two datasets of check-ins and weather information we create a single dataset for using in our research.
- We handle the class imbalance problem of the dataset for building high accuracy classification model.
- We build models to find the correlation between weather condition and users' activity, visiting place and travel mode by handling the mixed mode(categorical and numerical) attributes of the independent variables of our dataset.
- We build four different machine learning based classification model for the,
 - i. Prediction of preferable transport modes of user's.
 - ii. Prediction of user activities at day-time.
 - iii. Prediction of user activities at night-time.
 - iv. Prediction of user's visiting places.
- We build all four models using two different datasets of two different cities, New York and Tokyo.
- We propose a new approach for finding users' interest based on weather condition which will promote the "target marketing" and "influencer marketing" in business.

1.5 Outline

The remaining part of the thesis is organized as follows:

In Chapter 2, we outline the research work related to our thesis.

In Chapter 3, we formulate our research problem.

In Chapter 4, we briefly discuss the solution overview and introduce the 4 prediction models we build in our thesis.

In Chapter 5, we describe the complete process of building the first model for transport mode prediction. We also analyze the performance of our built model in this chapter.

In Chapter 6, we describe the process of building our second classification model for the prediction of day-time activity. This chapter also contains the performance analysis of the model.

In Chapter 7, we present our third model for the prediction of night-time activity of users' with performance analysis of the model.

In Chapter 8, we present our last model for the prediction of future visiting place of user. The chapter concludes with the performance analysis of the model.

In Chapter 9, we conclude the thesis with possible directions for future work.

Chapter 2

State of the Art

In this chapter, we discuss the work related to our research problem. First, we describe about various check-in datasets. We divide the research works related to our thesis in three groups: discovering and predicting users' daily routine and life-style patterns, predicting future activities or inferring activity preferences and clustering relevant venues from users' check-in datasets of LBSN sites. In Section 2.1, we discuss about the check-in datasets collected and used in various research. In Section 2.2, we discuss the works done to discover and predict users' daily routine and life-style patterns from users' check-in data. In Section 2.3, works on the prediction of future activities and the inference of activity preferences of users' from check-in dataset are discussed. In Section 2.4 we discuss about the clustering of socially relevant venues from users' check-in data.

2.1 Collection of dataset

Location Based Social Networking (LBSN) sites allow users to share their location information of visiting different venues or places through check-ins. These huge volume of user data provides an opportunity to find interesting insights about users preferences of places or transport mode at different times. Many works has been done to find interesting information from user check-in datasets of various social networks for example Foursquare, Gowalla, Twitter, Facebook etc. Users share their current location or the venue they have visited earlier with their friends. Generally LBSNs give unique IDs to all venues or locations. A user “checks in” to a specific venue by using a smartphone or tablet to choose from a list of venues near their current location. These locations are determined by Wi-Fi or

GPS. This information is sent to the LBSN server and shared with their friends. A user can check-in to a venue during each visit. Check-ins at different venues or places often encouraged through incentives. Check-in data consists of check-in history of the users, where each check-in is described by a user id, venue id, venue category id, and the time of the check-in. In addition, most LBSN services also provide secondary data that describe the underlying social network of the users. The use of dedicated mobile application for the collection of check-in dataset is a common practice. Very often the check-in dataset also contains information about the review of the venues users visit. Most commonly the check-in dataset contains the GPS coordinates of every single check-in and the time stamp of that check-in but it varies from set to set depending on the data collection platform or according to the necessity of the research purpose. Sometime datasets contain information about the semantic meaning of the venues, users experience or reviews about the visiting venues or location, user id etc.

2.2 Discovering and predicting users daily routine and life-style patterns

The study of human activity patterns from user check-in datasets is gaining attention rapidly now a days. Several works are done on it. These studies traditionally relies on the continuous tracking of user location.

In [11] the authors predict users daily routine from their check-ins. Here daily routine means when and where the user used to take break-fast, lunch, where do a user goes every day etc. Authors consider the activity pattern discovery from a new perspective. Instead of actively sampling increasing volumes of sensor data, they explore the participatory sensing potential of multiple mobile social networks. In multiple mobile social networks users often disclose information about their location and the venues they visit. The work of [11] presents automated techniques for filtering, aggregating, and processing combined social networking traces with the goal of extracting descriptions of regularly-occurring user activities, which refers as user routines. They use two localized data sets about a single pool of users. The first one contains public geotagged Twitter messages and the second one is Foursquare check-ins that provide meaningful venue information about the locations people visit.

Geo-location data from social media offers new ways, to understand people's attitudes and interests through their activity choices. [12] Explores the idea of inferring individual life-style patterns from activity-location choices revealed in social media. It presents a model to understand life-style choices of users from the contextual information or location categories of check-ins performed by users. Through building probabilistic topic models they infer individual geo life-style patterns. Here geo life style patter is pictured from two different perspective. The former is, i) to characterize the patterns of user interests to different types of places and the later one is, ii) to characterize the patterns of user visits to different neighborhoods. They use dataset of Foursquare check-ins of the users from New York City. They infer users' life-style pattern from their activity-location choices i.e., what kind of locations user used to visit for what kind of activity for example what shopping places they used to shop, what restaurants they used to visit, what transportation mode they used to travel etc.

Location-based social network generated data contains rich information on the whereabouts of urban dwellers. Such data reveals who spends time where, when, and on what type of activity for example, where do they go for shopping, what type of restaurant they used to visit etc. These type of information can be used to describe city region in terms of activity that takes place therein. In [13] authors make a probabilistic model with minimal assumptions about the data using Foursquare check-in dataset. They extract many interesting information about urban activity pattern from users check-ins of visiting locations. These interesting information are about the place or regions of the city, which places are similar to each other in the city, what are the features which distinguish one region from another. Foursquare dataset is used on this research.

It is possible to study individuals' mobility patterns at a fine-grained level and to see how they are impacted by social factors through location based social networks. In [14] authors analyze the check-in patterns of users' in LBSN. By analyzing users' mobility patterns they found that users' mobility pattern is correlated with social interactions. Authors observe significant temporal clustering within check-in activities. Human mobility exhibits structural regularities though they change over time. Authors include three mechanism to describe these check-in dynamics. They found that, (1) users' behavior is strongly influenced by his/her own recent activity, (2) Social influence for example, a visit by a user triggers future visits by his friends and (3) Exogenous effects, which include external events (such as releasing new software for the service or a promotion campaign) that modulate the attendance

rates. In this work authors are especially interested in assessing social influence on visitation patterns of users'. They do their research on Gowalla dataset.

2.3 Predicting future activities or inferring activity preferences

In LBSNs users interact with physical points of interest (POIs) by showing their presence in real-time and leaving their comments. These large-scale user generated digital footprints bring an opportunity to understand the spatial and temporal features of user activity where user activities are represented by check-ins. In [15] authors proposed an approach to predict users activity preferences by mining the spatial and temporal features of user activity. At first they model the spatial and temporal activity preference separately, and then uses a principle way to combine them for activity preference inference.

In [16] authors propose another way to predict venues that a user will likely visit, given historical information of his/her other or previously visited venues using Foursquare dataset. They cast this as a ranking problem. Given a list of candidate venues for each user, their methodology rank venues such that high ranking venues are more likely to be visited by the user. They explore Latent Dirichlet Allocation (LDA) topic models for venue prediction.

Another work is done for the prediction of user's next location using Twitter dataset. In [17] authors introduce a new methodology to predict individual's next location based on sparse footprints accumulated over a long time period using social media networks. Many other works are also done on next venue prediction or time-aware prediction on check-in datasets. Here we address the most relevant ones.

2.4 Clustering socially relevant venues

Understanding individual and collective mobility patterns is important for many applications. It also provides a great scope for the researchers. Many work is also done on it. In [18] authors examine the similarity of users based on the venues they have visited in the past. They cluster venues in such a

way which reflects the similarity of users who have not necessarily visited the exact same venues in the past. They use network structure information to cluster venues so that a venue's group reflects its functionality and based on the functionality of the venues' group they can find the similarity of users. Two different users may not visit the same venue or place but if they visit different venues within same cluster it is very normal that they have common interest. For example, university, library or book stores are different places but they are related to the activity "studying". So, users' associated with anyone of the three place must have similarities or common interests. They use Gowalla dataset. Another very interesting work of trade area analysis from user generated mobile location data is done in [19].

The identification of places with similar usage in urban region is an interesting topic for authorities, urban analysts and residents, as it provides valuable insights. In [20] authors present an approach to obtain this highly valuable knowledge. They segment city areas into clusters based on activity profiles from LBSN's data. A segment is represented by different locations sharing the same temporal distribution of check-ins. They find out how to describe the topic of the determined segments by modeling the difference to the overall temporal distribution of check-ins of the region. They use Foursquare dataset.

All the works done so far finds many interesting and insightful contents from social media data. Plethora of works have been done to predict users' activity from check-in data collected from social media. Among all the researches on the prediction of user activity no one addresses the weather condition for user activity prediction. In [2] authors explores the effects of weather on people's everyday activity. They characterize user activities using GPS traces of mobile phone users and considering temperature, rainfall and wind speed as weather parameter. Though the last work addresses the weather issue but they do not use social media data. They just uses mobile call record data to locate a user's presence in a location and they also assume the activity for the user as the most probable activity of that location using activity aware map. We are the first to find user activity from social media data based on weather conditions.

In this chapter we have discussed the works done so far which are related to our thesis. In the next chapter we will formulate our research problem, we have solved in this thesis.

Chapter 3

Problem Formulation

In this chapter we formulate our research problem. In Section 3.1 we give an overview of our system architecture with the output of our thesis. Then in Section 3.2 we present the basic steps we follow to solve our research problem.

3.1 Overview of our system architecture

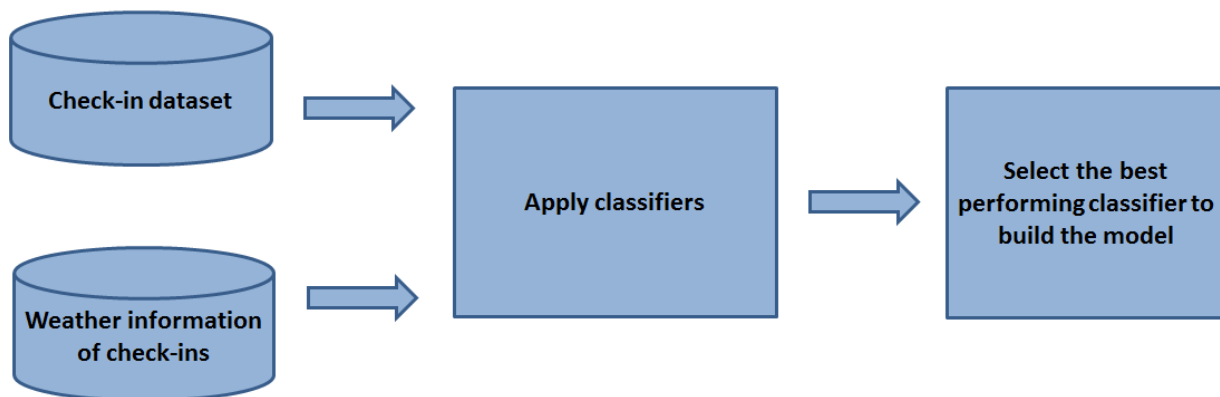


Figure 3.1: System architecture

The goal of our thesis is to predict users' activity, visiting location and preferable transport mode for a visit. We divide user activities in two different time slot, i) day-time activity and ii) night-time activity. The output of our research is 4 prediction model and they are the following:

- i. Model for the prediction of preferable transport mode of user's.
- ii. Model for the prediction of users' activities at day time.
- iii. Model for the prediction of users' activities at night time.
- iv. Model for the prediction of users' visiting places.

Figure 3.1 shows an overview of our system architecture. At first check-in dataset are collected. Then weather information of every single check-in is collected. Then for building the 4 classification model several classification algorithms are applied and from them the best performing classification algorithm is selected for building the final model.

3.2 Our research problem

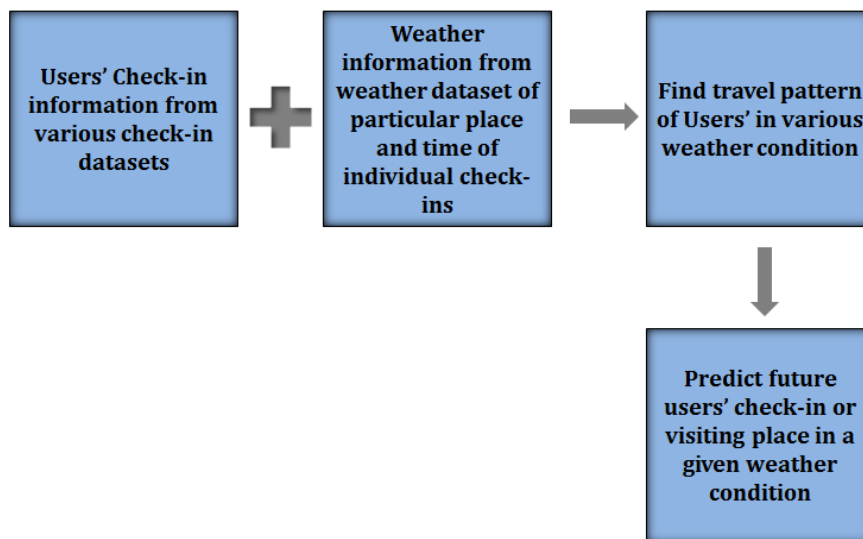


Figure 3.2: Research problem

Figure 3.2 shows the basic steps we follow to solve our research problem. After the collection of check-in dataset and weather information, the two datasets are cross-linked. Then machine learning based classification algorithms are applied on the dataset to learn previous knowledge and generate activity pattern or visiting pattern of users'. Finally the learned knowledge are used to build the prediction models.

Chapter 4

Methodology

In this chapter, we present our approach to develop machine learning based models to predict future activity, visiting place and transport mode of users' from previous users' check-ins and given weather condition. In Chapter 2 we discuss about the existing works done so far to find various interesting information about user, from their check-in datasets of various social networks. Among all the works of human activity prediction from social media data, no works address the issue of weather condition. In this thesis we propose the first approach to find user activity from social media data based on weather condition.

At first we extract users' location based dataset in terms of check-ins and weather information of individual check-ins. Then we select features for our model by computing various statistical significance tests between weather conditions and check-in places. Then we build the classification models based on the selected features and investigate the prediction potential of our classification models. Finally we predict future visits or activities of users' using our built model.

Rest of this chapter is organized as follows. In Section 4.1 we give the overview of our approach. Section 4.2 describes how we extract the location based dataset for our research. Then in Section 4.3 we discuss the process of weather information extraction of the location based dataset. Section 4.4 describes the data preprocessing step and Section 4.5 describes our feature selection process. Finally in Section 4.6 we introduce the models we build throughout our thesis.

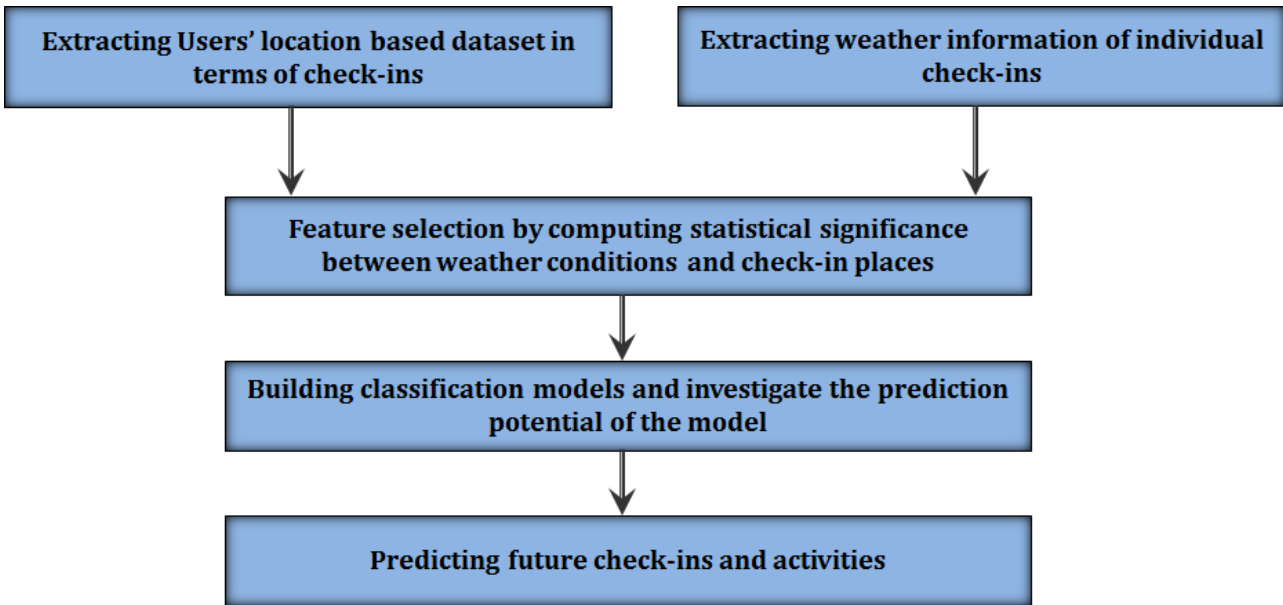


Figure 4.1: Overview of our approach

4.1 Overview of Our Approach

Outline of our proposed approach or methodology of building classification models for user’s future activity, visiting place and transport mode prediction is shown in Figure 4.1. The whole process can be summarized in the following steps.

- i. Extracting location based dataset: We collect Yang’s Foursquare dataset [15] of New York city and Tokyo city. Both dataset contains check-ins of individual users of different venues of the city. Each check-in data is associated with its time stamp, its GPS coordinates and its semantic meaning (represented by fine-grained venue-categories).
- ii. Extracting weather information: We extract weather information of every single check-in from a weather forecast service, *forecastio*¹. In the location based dataset every check-in data contains the latitude and longitude of every checked-in location along with the time of the check-in. *forecastio* provides the weather information of that particular location and time.
- iii. Data Preprocessing: In our dataset there are two types of variables i) weather information and ii) check-in places. We consider weather information as independent variable and check-in places

¹<https://darksy.net/dev/>

as dependent variable. The independent variable weather condition has two nominal attributes i) summary and ii) weather icon. Again the attribute summary has 38 different values. For building the classification models we narrow down the 38 weather category into 6 class levels.

- iv. Computing statistical significance: To build the classification models we need to select appropriate features. For feature selection we perform different statistical significance tests between weather attributes i.e., temperature, humidity, wind speed, condition summary, etc. and location categories or check-ins for example coffee shop, food place, art and entertainment place, etc. As our independent variable is a mixed type variable having both categorical and numerical attributes we need to follow multiple statistical significance test for feature selection.
- v. Building classification models: To build suitable classification models to predict users' activity, visiting place and transport mode preferences based on given weather attributes we use several classifiers and they are NaiveBayes, RandomTree, RandomForest and REPTree. Then from a comparative analysis of their performance we select the most suitable classifier for building our model. In these models, we consider weather attributes as independent variable and venue category as dependent variable.
- vi. Predicting future check-ins and activities: Finally we predict preferable activity or visiting places or preferable transportation mode of a user based on the given weather condition using the built models.

The following sections present the details of above mentioned steps of our proposed approach.

4.2 Extracting location based dataset

We have collected Yang's Foursquare dataset [15] of New York city and Tokyo city. The New York dataset contains 227,428 check-ins of 1083 individual users and the Tokyo dataset contains 573,703 check-ins of 2293 individual users of 251 different venues. Each check-in data is associated with anonymized user id, Foursquare venue id, Foursquare venue category id, Foursquare venue category name, latitude and longitude of the venue or check-in place, timezone offset in minutes between when the check-in occurred and the same time in coordinated universal time (UTC) and UTC time.

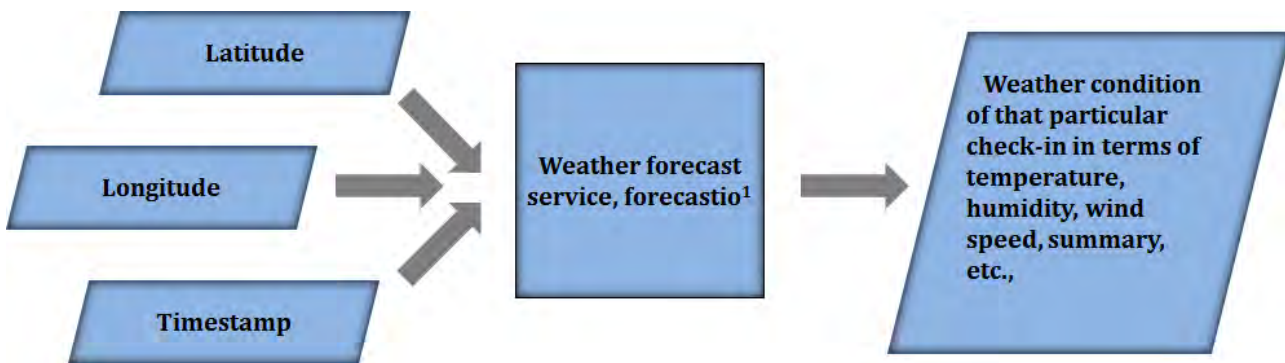


Figure 4.2: Extracting weather information

4.3 Extracting weather information

Then we collect weather information of every single check-in by using an weather forecast service via an API named *forecastio*². This API receives latitude, longitude and timestamp of every single check-in as parameter and provide us with the weather information of that particular check-in in terms of summary, weather icon, precipitation intensity, precipitation probability, temperature, apparent temperature, dew point, humidity, wind speed, wind bearing, visibiity, cloud coverage and air pressure which is shown in Figure 4.2. The *forecastio* API answers at most 1000 requests of weather information under one API key per day. So it took a long time to collect the weather information of 227,428 and 573,703 individual check-ins of both New York and Tokyo city datasets. Weather information was unavailable for some tuples and finally we get 227,427 and 570,901 instances for New York and Tokyo city respectively.

4.4 Data Preprocessing

Our independent variable weather condition has 13 different attributes. Among those 13 attributes first two attributes *weather summary* and *weather icon* are categorical. The attribute *weather summary* has 38 different values and the attribute *weather icon* has 9 different values. For building the classification model we narrow down the 38 different weather categories into 6 class levels. The reason behind this is the real feel or user experience of many weather condition is almost same, so there effects on user behavior or activity will be similar. For example there are two categories “Light Rain” and “Drizzle”.

²<https://darksky.net/dev/>

We put both of them in the same class level “Light Rain” because the user experience doesn’t vary from one another. Figure 4.3 shows how we merge various weather categories into one particular class. We also calculate their correlation and we find that weather categories lies in the same class has strong correlation with each other in terms of visiting places on that weather.

Breezy Windy Windy and Mostly Cloudy Windy and Overcast Windy and Partly Cloudy Breezy and Humid Breezy and Overcast Breezy and Mostly Cloudy Breezy and Partly Cloudy	Breezy	Heavy Snow Flurries Heavu Snow and Breezy Light Snow Snow Snow and Breezy	Snow
Clear Dry Dry and Partly Cloudy Partly Cloudy	Clear	Foggy Humid and Mostly Cloudy Humid and Overcast Dry and Mostly Cloudy Mostly Cloudy Overcast Humid and Partly Cloudy	Mostly Cloudy
Heavy Rain Heavy Rain and Breezy Heavy Rain and Windy Rain Rain and Breezy Rain and Windy	Heavy Rain	Drizzle Drizzle and Breezy Light Rain Light Rain and Breezy Light Rain and Windy Drizzle and Windy	Light Rain

Figure 4.3: Merging of similar weather categories

4.5 Computing statistical significance

For building the classification models we need to select appropriate features from a set of features. Our independent variable weather condition has 13 different attributes, which are *weather summary*, *weather icon*, *precipitation intensity*, *precipitation probability*, *temperature*, *apparent temperature*, *dew point*, *humidity*, *wind speed*, *wind bearing*, *visibility*, *cloud coverage* and *air pressure*. Here, *precipitation intensity*, *precipitation probability*, *temperature*, *apparent temperature*, *dew point*, *humidity*, *wind speed*, *wind bearing*, *visibility*, *cloud coverage* and *air pressure* are numerical or continuous variables

and *weather summary* and *weather icon* are categorical variables. Our dependent variable visited location is also a categorical variable. As our independent variable is a mixed type of variable having both nominal and numerical attributes, so we compute several statistical significance test for feature selection. When our independent variable is numerical we have used *Fishers Linear Discriminant Analysis (LDA)* Test and when our independent variable is categorical we have used *Chi-Square* Test. Then we choose only those attributes having statistical significance or correlation with the dependent variables as our feature for computing visiting location or future activity preferences.

4.6 Building classification models

Our main objective of this thesis is to build various suitable classification models for the prediction of user's activity and future visit or prediction of user's preferable transport mode on a given weather condition. In this thesis we build 4 different models. They are,

- i. Model for the prediction of preferable transport mode of user's: The model has five different class levels. They are *Bus*, *Light rail*, *Private transport*, *Subway* and *Train*. The model will predict among these five modes of transport, which transport mode will be preferable for a user on a given weather condition.
- ii. Model for the prediction of users' activities at day-time: This model has four different class levels. They are *Traveling*, *Shopping at mall*, *Watching movie in theater* and *Staying at home*. The classification model will predict which day-time activity an user may prefer on a given weather condition among these four types of day-time activities.
- iii. Model for the prediction of users' activities at night-time: This model has two different class levels and they are *Visiting nightlife spot* and *Staying at home*. Our model will predict the preferable night-time activity of an user on a given weather condition among this two night-time activities.
- iv. Model for the prediction of users' visiting places: Our last model has four different class levels of visiting places and they are *Park*, *Harbor / Marina*, *Indoor Museum* and *Sea Beach*. The model will predict users' preferable visiting place among these four classes of visiting places on a given weather condition.

In the following four chapters, we describe the process we follow to build the above mentioned models.

Chapter 5

Building Model for Transport Mode Prediction

Our first model is for the prediction of preferable transport mode of user's on various weather condition. Our model has five different class levels or transport mode. They are *Bus*, *Light rail*, *Private transport*, *Subway* and *Train*. Table 5.1 shows the dataset of Tokyo city we used for building the model. The dataset contains 2,54,335 instances. The class *Train* is the majority class having 78.4% of instances, where *Private Transport* is the minority class having only 1% of instances among the full dataset.

Table 5.1: Dataset of Tokyo for building the classification model for preferable transport mode

Class	Number of instances	Percentage
Train	199456	78.4%
Subway	41460	16.3%
Light Rail	2972	1.17%
Bus	7930	3.12%
Private Transport	2516	1%
Total	254335	100%

According to Section 4.5 for building the model we need to select the appropriate features. The process is described in the following sections.

5.1 Feature selection

There are 14 different attributes in our dataset. Among them first 13 attributes contain the weather information of a particular check-in and the 14th attribute is the transport mode. Here weather information is independent variable having 13 different attributes and transport mode is the dependent variable. The weather attributes are *weather summary*, *weather icon*, *precipitation intensity*, *precipitation probability*, *temperature*, *apparent temperature*, *dew point*, *humidity*, *wind speed*, *wind bearing*, *visibility*, *cloud coverage* and *air pressure*. Here *weather summary* and *weather icon* are categorical variables while the rest 11 attributes are numerical or continuous variable. Dependent variable transport mode is categorical. So, for feature selection we need to follow two different approaches as there is mixed type of attributes in the independent variables and they have both categorical and numerical or continuous values.

5.1.1 Feature selection for categorical weather information and categorical transport mode

We conducted Chi-Square(χ^2) test to check the correlation between *preferable transport mode* and weather attribute *weather summary* and *weather icon*. The Chi-square test(χ^2) statistic is calculated as,

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} \quad (5.1)$$

$$\text{Expected} = \frac{\text{marginal column frequency} - \text{marginal row frequency}}{\text{total sample size}} \quad (5.2)$$

We found that users' preferable transport mode and weather conditions are correlated. To find the correlation we need two values. One is degrees of freedom (*df*) and the other is Chi-square (χ^2) value. Then from Chi-square table¹ we find the critical value for given *df*. The *df* is basically a number that

¹<http://www.z-table.com/chi-square-table.html>

determines the exact shape of the distribution of a dataset. For the attribute *weather summary* we found the value $\chi^2 = 81.808$ and the $df = 20$ and for the attribute *weather icon* we found the value $\chi^2 = 166.065$ and the $df = 32$. df is calculated as,

$$df = (i - 1).(j - 1) \quad (5.3)$$

here, i is the number of attributes of the independent variables and j is the number of class levels in the dataset used for building the model. As our dataset contains 5 class levels, so j is equal to 5. Attribute *weather summary* has 6 distinct values, so i is equal to 6. Thus degrees of freedom is $(6-1).(5-1) = 20$. In the same way for attribute *weather icon* there are 9 distinct values, so i is equal to 9 and degrees of freedom is $(9-1).(5-1) = 32$. Table 5.2 shows the result of chi-square test between transport mode and *weather summary* and Table 5.3 shows the result of chi-square test between transport mode and *weather icon*. Both the tables show the critical value for given df . As we find both the calculated χ^2 values are greater than the critical values, so the attributes *weather summary* and *weather icon* are correlated with preferable transport mode and statistically significant. So primarily we select this two attributes as feature variable for building our classification model.

Table 5.2: Chi-Square test between transport mode and weather summary

	Value	df	Critical Value	Asymp. Sig. (2-sided)
Pearson Chi-Square	81.808	20	31.410	0.000
Likelihood Ratio	84.279	20		0.000
N of Valid Cases	254334			

Table 5.3: Chi-Square test between transport mode and weather icon

	Value	df	Critical Value	Asymp. Sig. (2-sided)
Pearson Chi-Square	166.065	32	46.194	0.000
Likelihood Ratio	163.868	32		0.000
N of Valid Cases	254334			

5.1.2 Feature selection for numerical weather information and categorical transport mode

Our independent or predictor variable weather condition has 11 numerical or continuous variables. They are *precipitation intensity*, *precipitation probability*, *temperature*, *apparent temperature*, *dew point*, *humidity*, *wind speed*, *wind bearing*, *visibility*, *cloud coverage* and *air pressure*. For ranking the predictors importance or determine factors that discriminate between classes we use Fisher's linear discriminant analysis(LDA). Discriminant analysis can handle only one outcome variable having two or more than two class levels. Our only dependent variable preferable transport mode has 5 different class levels.

A discriminant analysis calculates the probability of group membership based on a series of independent or predictor variables. The predictor variables are measured in scaled level of dimensions and the dependent variable is categorical. The assumptions for discriminant analysis are strict. If data does not meet the assumptions we can not apply discriminant analysis on the dataset.

For discriminant analysis dependent variables' category must be mutually exclusive. Our dataset meets this constraint. All the five class levels *Train*, *Subway*, *Light Rail*, *Bus* and *Private Transport* of our dependent variable preferable transport mode are mutually exclusive. The other assumptions for discriminant analysis are all the predictor variables must be independent to each other, normally distributed and there should be absence of outliers. There is no outliers in our dataset. All the 11 predictor variables are independent to each other and normally distributed. In terms of sample size there should be 5 times as many observations as predictor variables. Our sample size also meets this assumption. Finally, predictor variables could not be highly correlated with one another. If two variables have correlation coefficient larger than 0.8 then the dataset has multicollinearity problem and LDA cannot be applied on the dataset. We test our dataset and remove this multicollinearity [21] problem.

We compute a multi-level discriminant analysis using SPSS where 11 weather attributes are predictors or independent variable and preferable transport mode is the categorical dependent variable. We calculate the Pooled Within-group Correlation matrix which provide bivariate correlations

Table 5.4: Pooled Within-Groups Matrices

Correlation	Temperature	Apparent Temperature	Dew Point
Temperature	1.000	0.995	0.882
Apparent Temperature	0.995	1.000	0.888
Dew Point	0.882	0.888	1.000

between our 11 predictors. We find that the 3 attributes *temperature*, *apparent temperature* and *dew point* are highly correlated with each other having correlation coefficients larger than 0.8. Table 5.4 shows the coefficient values. All other coefficients are less than 0.8. In fact they are less than 0.65. At this point we need to choose only 1 predictor from these 3 predictors. To do so we compute LDA 3 times individually using each of the 3 predictors along with the other 8 predictors each time. We find that the predictor *apparent temperature* affects most in the classification. So we choose *apparent temperature* among the 3 predictors and discard the attribute *temperature* and *dew point* as our feature to solve this multicollinearity problem. We can also say that among these 3 predictors *apparent temperature* or real feel temperature is more perceiving to an user then just temperature of a day or dew point. So, it has the highest effects on users' experience.

Table 5.5 shows Fisher's linear discriminant function coefficients between weather attributes and transport mode. These coefficients are helpful in deciding which variable affects more in classification. If we compare the values between groups, the higher coefficient means the variable attributes more for that group. The equation of linear discriminant function or the classification function is as follows,

$$C_k = C_{k0} + C_{k1}X_1 + C_{k2}X_2 + \dots + C_{km}X_m \quad (5.4)$$

here, C_k is the classification score for group K

C_{ki} are the coefficients in the Table 5.5 where, $i = 0, 1, \dots, m$

m is the number of attributes

For every single observation classification score is calculated for each group by the coefficients according of the Table 5.5 and observation is assigned in the highest score group. From Table 5.5 we can see that the attribute *precipitation probability* has a very low coefficient value, so we discard it

Table 5.5: Fisher’s linear discriminant function coefficients between weather attributes and transport mode

Predictors	Bus	light Rail	Private Transport	Subway	Train
Precipitation intensity	182.531	181.187	182.004	181.934	182.969
Precipitation probability	-19.503	-19.833	-19.605	-19.511	-19.507
Apparent temperature	24.165	24.308	24.268	24.165	24.306
Humidity	209.324	210.008	209.912	208.647	208.953
Wind speed	8.261	8.369	8.261	8.236	8.249
Wind bearing	0.282	0.283	0.282	0.283	0.282
Visibility	-0.177	-0.183	-0.130	-0.182	-0.185
Cloud coverage	-0.698	-0.295	-0.703	-0.237	-0.508
Air pressure	23.946	23.950	23.942	23.940	23.943

from our feature variables. Finally, we select 8 attributes from 11 numeric attributes as features for building our classification model and they are *precipitation intensity*, *apparent temperature*, *humidity*, *wind speed*, *wind bearing*, *visibility*, *cloud coverage* and *air pressure*.

5.1.3 Feature selection using automated subset selection method

To validate our feature selection process we also use an automated subset selection method using *leaps* [22] R package implementation. We select the best subsets of features of size 7, 8 and 9 using the exhaustive selection algorithm [23]. *Leaps* package performs an exhaustive search to find out best subset of weather attributes using an efficient branch and bound algorithm. Table 5.6 shows the attribute subsets of various sizes selected by the exhaustive selection algorithm of R. From the table we can see that the selected feature sets are almost similar to the feature set we selected using Fisher’s linear discriminant analysis. Finally we choose 8 numeric attributes for building our classification model and they are *precipitation intensity*, *apparent temperature*, *humidity*, *wind speed*, *wind bearing*, *visibility*, *cloud coverage* and *air pressure*. We also use 2 nominal or categorical variable *weather summary* and *weather icon* as our feature and our dependent variable is only categorical variable

preferable transport mode.

Table 5.6: Feature selection in R using exhaustive selection algorithm

Size	Selected Attribute Subset
7	precipitation intensity, precipitation probability, humidity, wind speed, wind bearing visibility and air pressure
8	precipitation intensity, precipitation probability, humidity, wind speed, wind bearing visibility, cloud coverage and air pressure
9	precipitation intensity, precipitation probability, apparent temperature, humidity wind speed, wind bearing, visibility, cloud coverage and air pressure

5.2 Building the Classification Model

To build the classification model finally we choose 11 independent or predictor variables and one dependent variable transport mode. Table 5.7 shows the selected features for building our classification model and Table 5.1 shows our dataset. Then we apply Naive Bayes, Random Forest, Random Tree and RepTree classifier in our dataset by using WEKA [24] machine learning toolkit. Table 5.8 shows the performance of all the classifiers we used to build the classification model to predict users' preferable transport mode. The table presents the classification results of our built model in terms of true positive rate(TPR), false positive rate(FPR), area under the ROC curve(AUC), mean absolute error(MAE) and root mean squared error(RMSE) for predicting preferable transport mode of user from given weather condition. TPR defines how many instances are correctly classified as positive among all positive instances and FPR defines how many instances are incorrectly classified as positive among all negative instances available during the test. We calculate the performance of the classifier by using AUC values under 10-fold cross validation. From the table we can see that the performance of Random Forest Tree Ensemble is best. So, finally we choose Random Forest Tree Ensemble as our classifier. Table 5.9 presents the classification results of our built model. We find that on an average the AUC value of our classifier is 66.6%. We also find that mean absolute error(MAE) of our model is 0.1279 and root mean squared error(RMSE) is 0.2628.

Table 5.7: Selected features for the classification model

Predictors	Dependent Variable
<i>Weather Summary</i>	Preferable Transport Mode
<i>Weather Icon</i>	
<i>Precipitation Intensity</i>	
<i>Apparent Temperature</i>	
<i>Humidity</i>	
<i>Wind Speed</i>	
<i>Wind Bearing</i>	
<i>Visibility</i>	
<i>Cloud coverage</i>	
<i>Air Pressure</i>	

Table 5.8: Performance of different classifiers for building the model preferable transport mode

Classifiers	AUC	TPR	FPR	MAE	RMSE
NaiveBayes	0.519	0.784	0.784	0.146	0.2677
RandomForest	0.666	0.777	0.655	0.1279	0.2628
RandomTree	0.572	0.692	0.548	0.1231	0.3508
RepTree	0.578	0.777	0.753	0.1381	0.2704

5.3 Limitation of the Build Model

Though the AUC value is showing quite moderate performance but the classifier suffers from major limitations. Standard machine learning algorithms have a bias towards classes which have large number of instances and they tend to predict only the majority class, considering the features of minority classes as noise. So, there is a high probability of misclassification of minority classes as compared to the majority classes. In our built model, if we look at the TPR rate of each class levels we can see that except the class *Train*, other 4 class levels have a very poor TPR rate. The reason behind is the class imbalance problem of our dataset. From Table 5.1 we can see that the percentage of the class *Train*

Table 5.9: Classification results of the model preferable transport mode build using Random Forest algorithm

Class	TPR	FPR	AUC
Train	0.948	0.825	0.659
Subway	0.166	0.044	0.699
Light Rail	0.055	0.002	0.678
Bus Station	0.032	0.005	0.622
Private Transport	0.547	0.001	0.807
Weighted Avg.	0.777	0.655	0.666

is 78.4%. So, it is the major class having most of the instances of the dataset. The second largest class is *Subway* and it is 16.3% of the total dataset. The rest of the 3 classes are very low in number of instances. The FPR rate of the class *Train* is 0.825 which is very high because as the number of instances for this class is very large in compare to other classes so, all other classes are misclassified as *Train* most of the time. Another observation is except the class *Train*, among other 4 classes, the class *Private Transport* has a TPR rate of 0.547 which is comparably higher than other 3 classes. The reason behind is, except the class *Private transport* other 4 classes are the public transport of Tokyo city. So their characteristics or relationship with weather condition have some common features but *Private Transport* is completely different kind of vehicles and should have a different types of relationship with weather conditions. So, the class *Private Transport* can be correctly classified due to it's unique features than other classes. All other 4 classes except *Train* are not misclassified also due to their lack of instances. So, our build model performs very poorly. It can only classify the class *Train* and most of the time it misclassifying all other classes as *Train*. Only it can classify the class *Private Transport* with a rate of 54.7%. So, we need to handle this class imbalance problem of our dataset to improve the accuracy of our model.

5.4 Handling Class Imbalance Problem Using SMOTE Algorithm

We notice that our above built model is biased and inaccurate due to the imbalance of class distribution of our dataset. Machine learning algorithms do not take into account the class proportion or balance of classes. So we need to follow an approach for solving such class imbalance problem using sampling technique. The main objective of balancing classes is to obtain a dataset having approximately the same number of instances for all the classes by re-sampling techniques. It is done by either increasing the frequency of minority classes or decreasing the frequency of majority classes. There are several re-sampling techniques practiced commonly in machine learning and data science. Among them the technique *Random Under-sampling* aims to balance class distribution by randomly eliminating majority class examples. This technique improves the run time and storage problems by reducing the number of training data samples when the training dataset is huge in size but it can discard potentially useful information which could be important for building classification model. Another re-sampling technique *Random Over-Sampling* increases the number of instances in the minority class by randomly replicating them in order to present a higher representation of the minority class in the sample. Though this technique outperforms the previous technique and does not leads to any information loss, it increases the likelihood of over-fitting since it replicates the minority class events. In *Cluster-Based Over-Sampling* techniques, K-means clustering algorithm is independently applied to minority and majority class instances. This identifies clusters in the dataset. Then each cluster is oversampled in a way that all clusters of the same class have an equal number of instances and all classes have the same size. This technique overcome the challenge of class imbalance but has a drawback of over-fitting the training set. Another re-sampling technique *Synthetic Minority Over-sampling Technique* (SMOTE) uses a subset of data from the minority class and creates new synthetic similar instances. Then add the newly created data into the original data to train the models. SMOTE avoids to make exact replicas of minority class instances to overcome the over-fitting problems. It also does not loss any useful information.

After weighing the pros and cons of various re-sampling techniques we choose *Synthetic Minority Over-sampling Technique* (SMOTE) to deal with our class imbalanced dataset. Figure 5.1 shows the

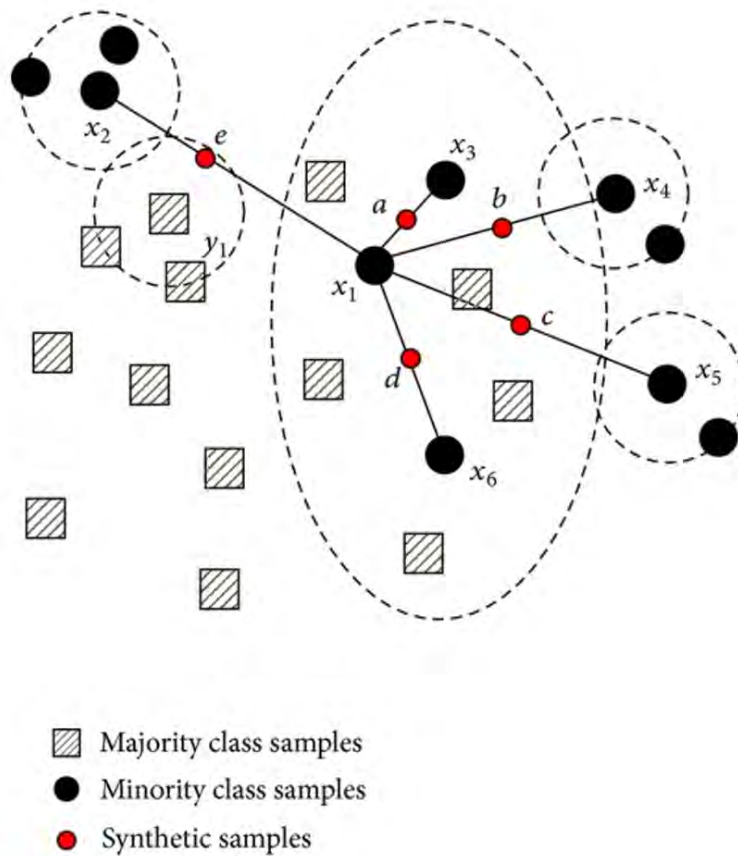


Figure 5.1: Generation of synthetic instances with the help of SMOTE

generation of synthetic instances using SMOTE re-sampling algorithm. SMOTE algorithm is also applied in various real life applications to solve the class imbalance problems of real datasets. In detection of fraudulent transactions of bank credit cards, identifying customer churn of various companies, predicting the possibilities of natural disaster like Earthquakes, identifying various rare diseases in medical diagnostics, in all the mentioned scenarios, datasets suffer from class imbalance problems. In real life though these are the target events but the datasets starve for the instances because these are the rare events. So, all these real applications must need to deal with the class imbalance problems and SMOTE algorithms is very commonly used in such kind of applications [25] [26] [27] [28] [29].

SMOTE creates synthetic observations of the minority class by finding the k -nearest-neighbors for minority class observations. Then it randomly choose one of the k -nearest-neighbors and creates

Table 5.10: Up-Sampled dataset after applying SMOTE

Class	Number of instance	Percentage
Train	199456	63%
Subway	41460	13%
Light Rail	23776	7.5%
Bus	31720	10%
Private Transport	20128	6.4%
Total	316540	100%

a similar but randomly tweaked new observation from it. We apply SMOTE on our dataset and up-sample our minority classes. Table 5.10 shows the class distribution we find after applying SMOTE on our class imbalanced dataset. We increase the number of instances up to 6 times than previous, for 3 minority classes of our dataset. For example, previously the class Private Transport had 1% of instances of the full dataset. We up-sample this class 6 times and finally we get this class having 6.4% of instances of the total dataset.

Table 5.11: Classification results of the model preferable transport mode after up-sampling the dataset using SMOTE algorithm

Class	TPR	FPR	AUC
Train	0.933	0.448	0.834
Subway	0.173	0.037	0.740
Light Rail	0.804	0.006	0.974
Bus Station	0.569	0.012	0.929
Private Transport	0.862	0.004	0.985
Weighted Avg.	0.783	0.289	0.851

5.5 Re-Building the Classification Model

Again we apply Random Forest classifier in our dataset of Table 5.10 by using WEKA [24] machine learning toolkit. Table 5.11 presents the classification results of our newly built model in terms of true positive rate (TPR), false positive rate (FPR) and area under the ROC curve (AUC) for predicting preferable transport mode. We calculate the performance of the classifier by using AUC values under 10-fold cross validation. We find that on an average the AUC value of our classifier is 85.1%. We also find that mean absolute error (MAE) of our model is 0.1371 and root mean squared error (RMSE) is 0.2581. This time the TPR rate of all the classes except *Subway* is quite high. The class *Subway* is misclassified as *Train* because in Tokyo *Subway* and *Train* are two most commonly used public transport for all the time, so both the classes should share common features for weather. The class size *Train* is almost 5 times as large as the class *Subway*, so the instances of *Subway* are misclassified as the class *Train* at a high rate. All other 3 classes have moderate TPR, FPR and AUC values.

5.6 Discussion

Table 5.12: Dataset of New York city used for building the classification model for preferable transport mode

Class	Number of instances	Percentage
Train	6408	27.1%
Subway	9348	39.6%
Light Rail	733	3.1%
Bus	4474	18.9%
Private Transport	2645	11.2%
Total	23608	100%

We have another dataset of 23,608 instances of New York city. Table 5.12 shows the number of instances per each class level. We apply Naive Bayes, Random Forest, Random Tree and RepTree classifiers on these dataset. Table 5.13 shows the performance of all the classifiers we use to build the model to predict users' preferable transport mode. From the table we see that Random Forest

Table 5.13: Performance of different classifiers for building the model preferable transport mode on New York dataset

Classifiers	AUC	TPR	FPR	MAE	RMSE
NaiveBayes	0.530	0.376	0.367	0.291	0.3871
RandomForest	0.646	0.495	0.253	0.2541	0.3748
RandomTree	0.579	0.395	0.238	0.2418	0.4916
RepTree	0.564	0.378	0.317	0.2766	0.395

Table 5.14: Classification results of the model preferable transport mode build using Random Forest algorithm on New York dataset

Class	TPR	FPR	AUC
Train	0.470	0.224	0.613
Subway	0.587	0.415	0.630
Light Rail	0.307	0.011	0.718
Bus Station	0.368	0.123	0.642
Private Transport	0.494	0.041	0.766
Weighted Avg.	0.495	0.253	0.646

Tree Ensemble shows the best performance. So, finally we use it for building our classification model. Table 5.14 shows the performance of our build model. From Table 5.12 we can see that here *Subway* is the majority class unlike to the dataset of Table 5.1 having 39.6% instances of total dataset. Unlike to the Tokyo city *Subway* is the main public transport in New York city. Then the second largest class is *Train*. Here, *Light Rail* is the minority class having 3.1% of instances. Other two class *Bus* and *Private Transport* have 18.9% and 11.2% of instances respectively. Unlike to the dataset of Table 5.1 in New York city dataset *Private Transport* is not the minority class. 48% of household of the city own private transport [30]. The dataset of Table 5.12 is not that imbalance same as the dataset of Table 5.1. So, we don't apply SMOTE algorithm on it. We directly build the model using the basic dataset and our build model shows moderate performance shown in Table 5.14. The model has an AUC value of 64.6%. All the classes have moderate TPR and low FPR except the class *Subway* having

FPR of 41.5%. As it is the majority class so very often other classes got misclassified by this class. *Subway* is the main and most common public transport of the city. So, it is used in almost all kind of weather. So, it overlaps with all kind of weather conditions. Private transports are preferable in rainy or snowy weather for convenience [30]. Heavy rain and snow fall causes difficulty to use public transport as users' need to go outside from their living place to another place to use public transport.

Chapter 6

Building Model for Day-Time Activity Prediction

Our second model is for the prediction of user's preferable day-time activity on various weather condition. This model has four different class levels or activity mode. They are *Traveling*, *Shopping at mall*, *Watching movie in theater* and *Staying at home*. Table 6.1 shows the dataset we used for building the model. The dataset contains 21,755 instances. All classes have almost same number of instances except the class *Staying at home* having only 13.2% of instances among the whole dataset. All other 3 classes have around 30% of instances of the total dataset.

Table 6.1: Dataset of Tokyo for building the classification model for preferable day-time activity

Class	Number of instances	Percentage
Traveling	6683	30.7%
Shopping at mall	7069	32.5%
Watching movie in theater	5130	33.6%
Staying at home	2873	13.2%
Total	21755	100%

Now we need to select the appropriate features for building the classification model as stated in Section 4.5. The process is described in the following sections.

6.1 Feature selection

There are 14 different attributes in this dataset. First 13 attributes contain the weather information of a particular check-in and the 14th attribute is the check-in place or venue category represented in the dataset as day-time activity mode. Here weather information is independent variable having 13 different attributes and day-time activity mode is the dependent variable. Weather information has both categorical and numerical values where day-time activity contains categorical value. So, we need to follow two different techniques for feature selection as we have mixed type of independent variables.

6.1.1 Feature selection for categorical weather information and categorical day-time activity

We conduct Chi-Square test to check the correlation between *preferable day-time activity* and weather attribute *weather summary* and *weather icon*. We found that users' preferable day-time activity and weather conditions are correlated. For the attribute *weather summary* we found the value $\chi^2 = 38.683$ and the $df = 15$. For the attribute *weather icon* we found the value $\chi^2 = 227.048$ and the $df = 24$. It's calculated using Equation 5.3. Table 6.2 shows the result of chi-square test between day-time activity and *weather summary* and Table 6.3 shows the result of chi-square test between day-time activity and *weather icon*. We check the Chi-Square table for critical values and find both the attributes *weather summary* and *weather icon* are statistically significant as we did for building the previous model. So primarily we select this two attributes as feature variable for building our classification model.

Table 6.2: Chi-Square test between preferable day-time activity and weather summary

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	38.683	15	0.001
Likelihood Ratio	35.930	15	0.002
N of Valid Cases	21755		

Table 6.3: Chi-Square test between preferable day-time activity and weather icon

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	227.048	32	0.000
Likelihood Ratio	225.064	32	0.000
N of Valid Cases	21755		

6.1.2 Feature selection for continuous weather information and categorical day-time activity

In our dataset we have 11 numerical or continuous variables same as the earlier dataset used to build the previous model. For ranking the predictors' importance or determine factors that discriminate between classes, again we use Fisher's linear discriminant analysis (LDA). Our dependent variable preferable day-time activity has 4 different class levels. Our second dataset also meets all the assumptions of discriminant analysis, otherwise we can not apply LDA on it. We discussed the assumptions in Section 5.1.2. All the four class levels, *Traveling*, *Shopping at mall*, *Watching movie in theater* and *Staying at home* of our dependent variable preferable day-time activity are mutually exclusive. All the 11 predictor variables are independent to each other, normally distributed and there is no outliers in our dataset. Sample size also meets the assumption. We test our dataset and remove the multicollinearity [21] problem.

We compute a multi-level discriminant analysis using SPSS where 11 weather attributes are predictors and preferable day-time activity is the dependent variable. We calculate the Pooled Within-group Correlation matrix which provide bivariate correlations between our 11 predictors. We find that same as the previous model the 3 attributes *temperature*, *apparent temperature* and *dew point* are highly correlated with each other having correlation coefficients larger than 0.8. Table 6.4 shows the coefficient values. All other coefficients are less than 0.8. So, at this point we need to choose only 1 predictor from these 3 predictors. To do so we compute LDA 3 times individually using each of the 3 predictors along with the other 8 predictors each time. We find that the predictor apparent temperature affects most in the classification as before. So we choose apparent temperature among the 3 predictors and discard the attribute *temperature* and *dew point* as our feature to solve this

multicollinearity problem. It is also intuitive that among these 3 predictors apparent temperature is the most perceiving attribute to an user then just temperature of a day or dew point.

Table 6.4: Pooled Within-Groups Matrices

Correlation	Temperature	Apparent Temperature	Dew Point
Temperature	1.000	0.994	0.882
Apparent Temperature	0.994	1.000	0.887
Dew Point	0.882	0.887	1.000

Table 6.5: Fisher's linear discriminant function coefficients between weather attributes and day-time activity

Predictors	Traveling	Shopping at mall	Watching movie in theater	Staying at home
precipitation intensity	693.344	692.687	692.790	695.254
precipitation probability	4.403	4.574	4.470	4.362
Apparent temperature	92.518	92.477	92.446	92.525
Humidity	3347.929	3345.944	3345.350	3345.179
Wind speed	2.542	2.535	2.528	2.569
Wind bearing	0.263	0.263	0.264	0.262
Visibility	16.272	16.246	16.209	16.122
Cloud coverage	-79.307	-79.872	-79.434	-79.233
Air pressure	26.953	26.956	26.954	26.948

Table 6.5 shows Fisher's linear discriminant function coefficients between weather attributes and day-time activity. These coefficients are helpful in deciding which variable affects more in classification. If we compare the values between groups, the higher coefficient means the variable attributes more for that group. From table 6.5 we can see that the attribute cloud coverage has a very low coefficient value, so we discard it from our feature variables. Finally, we select 8 attributes from 11 numeric attributes as features for building our classification model.

6.1.3 Feature selection using automated subset selection method

To validate our feature selection process again we use the automated subset selection method using *leaps* [22] R package implementation. We select the best subsets of features of size 7, 8 and 9 using the exhaustive selection algorithm [23]. Finally we choose 8 numeric attributes for building our classification model and they are *precipitation intensity*, *precipitation probability*, *apparent temperature*, *humidity*, *wind speed*, *wind bearing*, *visibility* and *air pressure*. We also use 2 nominal or categorical variables *weather summary* and *weather icon* as our feature and our dependent variable is only categorical variable *preferable day-time activity*.

6.2 Building the Classification Model

Table 6.6: Selected features for the classification model

Predictors	Dependent Variable
<i>Weather Summary</i>	Preferable Day-Time Activity
<i>Weather Icon</i>	
<i>Precipitation Intensity</i>	
<i>Precipitation Probability</i>	
<i>Apparent Temperature</i>	
<i>Humidity</i>	
<i>Wind Speed</i>	
<i>Wind Bearing</i>	
<i>Visibility</i>	
<i>Air Pressure</i>	

To build the classification model finally we choose 11 independent or predictor variables and one dependent variable day-time activity. Table 6.6 shows the selected features for building our classification model and table 6.1 shows our dataset. Then we apply Naive Bayes, Random Forest, Random Tree and RepTree classifiers in our dataset by using WEKA [24] machine learning toolkit. Table 6.7 shows the performance of all the classifiers we use to build the classification model to predict users'

Table 6.7: Performance of different classifiers for building the model preferable day-time activity

Classifiers	AUC	TPR	FPR	MAE	RMSE
NaiveBayes	0.548	0.317	0.292	0.3578	0.4321
RandomForest	0.651	0.522	0.226	0.3239	0.4215
RandomTree	0.567	0.371	0.237	0.3146	0.5608
RepTree	0.580	0.364	0.263	0.3436	0.4462

preferable day-time activity. The table presents the classification results of our built model in terms of true positive rate(TPR), false positive rate(FPR), area under the ROC curve(AUC), mean absolute error(MAE) and root mean squared error(RMSE)for predicting preferable day-time activity of user from given weather condition. We calculate the performance of the classifier by using AUC values under 10-fold cross validation. From the table we can see that the performance of Random Forest Tree Ensemble is best. So, finally we choose Random Forest Tree Ensemble as our classifier. Table 6.8 shows the performance of our build model.

Table 6.8: Classification results of the model preferable day-time activity build using Random Forest algorithm

Class	TPR	FPR	AUC
Traveling	0.589	0.273	0.654
Shopping at mall	0.557	0.291	0.626
Watching movie in theater	0.469	0.169	0.662
Staying at home	0.376	0.065	0.682
Weighted Avg.	0.522	0.226	0.651

We find that on an average the AUC value of our classifier is 65.1%. We also find that mean absolute error(MAE) of our model is 0.3239 and root mean squared error(RMSE) is 0.4215. The AUC value is showing quite moderate performance and the TPR rate of all the classes except *Staying at home* are moderate. *Staying at home* is the minority class having fewer instances. Also people generally stays at home in any weather, so it overlaps with other classes. People tends to travel more in clear weather. The class *Traveling* is more correlated with clear weather summary. During rain or snow *Watching*

movie in theater is more common.

6.3 Discussion

We have another dataset of 28,649 instances of New York city. Table 6.9 shows the number of instances per each class level. We apply Naive Bayes, Random Forest, Random Tree and RepTree classifiers on these dataset. Table 6.10 shows the performance of all the classifiers we use to build the model to predict users' preferable day-time activity mode. From the table we see that Random Forest Tree Ensemble shows the best performance. So, again we use it for building our classification model. Table 6.11 shows the performance of our build model.

Table 6.9: Dataset of New York city for building the classification model for preferable day-time activity

Class	Number of instances	Percentage
Traveling	7439	25.9%
Shopping at mall	7285	25.4%
Watching movie in theater	5634	19.7%
Staying at home	8291	28.9%
Total	28649	100%

Table 6.10: Performance of different classifiers for building the model preferable day-time activity on New York dataset

Classifiers	AUC	TPR	FPR	MAE	RMSE
NaiveBayes	0.585	0.340	0.233	0.3568	0.4371
RandomForest	0.696	0.439	0.193	0.314	0.4179
RandomTree	0.595	0.394	0.205	0.3029	0.5503
RepTree	0.633	0.385	0.214	0.3348	0.4391

On an average the AUC value of the classifier is 69.6%. Its performance is pretty good. All the classes

Table 6.11: Classification results of the model preferable day-time activity build using Random Forest algorithm on New York dataset

Class	TPR	FPR	AUC
Traveling	0.331	0.195	0.606
Shopping at mall	0.409	0.203	0.671
Watching movie in theater	0.411	0.134	0.723
Staying at home	0.581	0.221	0.781
Weighted Avg.	0.439	0.193	0.696

have moderate AUC values. Similar to the previous model we find that the class *Traveling* has good correlation with clear *weather summary*. The class *Staying at home* is more common in rain and snow *weather summary*.

Chapter 7

Building Model for Night-Time Activity Prediction

Our third model is for the prediction of user's preferable night-time activity on given weather condition. Our model has two different class levels or activity mode. They are *Visiting nightlife spot* and *Staying at home*. Table 7.1 shows the dataset we used for building the model. The dataset contains 20,849 instances.

Table 7.1: New York city dataset for building the classification model for preferable night-time activity

Class	Number of instances	Percentage
Visiting nightlife spot	11647	55.9%
Staying at home	9202	44.1%
Total	20849	100%

Same as the previous two model building process now we need to select the appropriate features for building the classification model. The process is described in the following sections.

7.1 Feature selection

This dataset also contains 14 different attributes. First 13 attributes contain the weather information and the 14th attribute is the check-in place represented as night-time activity. Weather information

is independent variable and night-time activity is the dependent variable. As independent variable weather information is mixed type of variable containing both categorical and numerical values, we follow 2 different techniques for feature selection.

7.1.1 Feature selection for categorical weather information and categorical night-time activity

Among 13 independent variables of weather information *weather summary* and *weather icon* are categorical. So, we use Chi-Square test to find the statistical significance. We conduct Chi-Square test to check the correlation between *preferable night-time activity* and weather attribute *weather summary* and *weather icon*. We found that users' preferable night-time activity and weather conditions are related. For the attribute *weather summary* we found the value $\chi^2 = 27.752$ and the $df = 5$. For the attribute *weather icon* we found the value $\chi^2 = 35.933$ and the $df = 8$. It's calculated using Equation 5.3. Table 7.2 shows the result of chi-square test between night-time activity and *weather summary* and Table 7.3 shows the result of chi-square test between night-time activity and *weather icon*. As we find both the attributes *weather summary* and *weather icon* are statistically significant so primarily we select this two attributes as feature variables for building our classification model.

Table 7.2: Chi-Square test between preferable night-time activity and weather summary

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	27.752	5	0.000
Likelihood Ratio	27.613	5	0.000
N of Valid Cases	20849		

Table 7.3: Chi-Square test between preferable night-time activity and weather icon

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	35.933	32	0.000
Likelihood Ratio	35.831	32	0.000
N of Valid Cases	20849		

7.1.2 Feature selection for continuous weather information and categorical night-time activity

In this dataset we also have 11 numerical or continuous variables same as the earlier dataset used to build the previous two models. For ranking the predictors importance or determine factors that discriminate between classes, again we use Fisher's linear discriminant analysis (LDA). Our dependent variable preferable night-time activity has 4 different class levels. Our third dataset also meets all the assumptions of discriminant analysis, otherwise we can not apply it. We discussed the assumptions in Section 5.1.2. All the two class levels, *Visiting night life spot* and *Staying at home* of our dependent variable preferable night-time activity are mutually exclusive. All the 11 predictor variables are independent to each other, normally distributed and there is no outliers in our dataset. Sample size also meets the assumption. We test our dataset and remove the multicollinearity [21] problem as we did in Section 5.1.2.

We compute a multi-level discriminant analysis using SPSS where 11 weather attributes are predictors and preferable night-time activity is the dependent variable. We calculate the Pooled Within-group Correlation matrix which provide bivariate correlations between our 11 predictors. We find that same as the previous 2 models the 3 attributes *temperature*, *apparent temperature* and *dew point* are highly correlated with each other having correlation coefficients larger than 0.8. All other coefficients are less than 0.8. So, at this point we choose only apparent temptation same as previous times.

Table 7.4 shows Fisher's linear discriminant function coefficients between weather attributes and night-time activity. If we compare the values between groups, the higher coefficient means the variable attributes more for that group. From Table 6.5 we can see that the attribute wind bearing has a very low coefficient value, so we discard it from our feature variables. Finally, we select 8 attributes from 11 numeric attributes as features for building our classification model.

Table 7.4: Fisher’s linear discriminant function coefficients between weather attributes and night-time activity

Predictors	Visiting nightlife spot	Staying at home
precipitation intensity	857.925	859.457
precipitation probability	75.368	75.609
Apparent temperature	81.011	80.986
Humidity	3373.513	3374.414
Wind speed	27.994	27.977
Wind bearing	0.495	0.495
Visibility	14.212	14.261
Cloud coverage	9.283	9.312
Air pressure	29.698	29.677

7.1.3 Feature selection using automated subset selection method

To validate our feature selection process again we use the automated subset selection method using *leaps* [22] R package implementation. We select the best subsets of features of size 7, 8 and 9 using the exhaustive selection algorithm [23]. Finally we choose 8 numeric attributes for building our classification model and they are *precipitation intensity*, *precipitation probability*, *apparent temperature*, *humidity*, *wind speed*, *visibility*, *cloud coverage* and *air pressure*. We also use 2 nominal or categorical variables *weather summary* and *weather icon* as our feature and our dependent variable is only categorical variable *preferable night-time activity*.

7.2 Building the Classification Model

To build the classification model finally we choose 11 independent or predictor variables and one dependent variable night-time activity. Table 7.5 shows the selected features for building our classification model and Table 7.1 shows our dataset.

Then we apply Naive Bayes, Random Forest, Random Tree and RepTree classifiers in our dataset

Table 7.5: Selected features for the classification model

Predictors	Dependent Variable
<i>Weather Summary</i>	Preferable Night-Time Activity
<i>Weather Icon</i>	
<i>Precipitation Intensity</i>	
<i>Precipitation Probability</i>	
<i>Apparent Temperature</i>	
<i>Humidity</i>	
<i>Wind Speed</i>	
<i>Visibility</i>	
<i>Cloud Coverage</i>	
<i>Air Pressure</i>	

Table 7.6: Performance of different classifiers for building the model preferable night-time activity

Classifiers	AUC	TPR	FPR	MAE	RMSE
NaiveBayes	0.526	0.559	0.536	0.4801	0.5155
RandomForest	0.729	0.668	0.354	0.3931	0.4602
RandomTree	0.618	0.625	0.390	0.3751	0.6123
RepTree	0.622	0.609	0.418	0.4351	0.5147

Table 7.7: Classification results of the model preferable night-time activity build using Random Forest algorithm

Class	TPR	FPR	AUC
Visiting night life spot	0.751	0.437	0.729
Staying at home	0.563	0.249	0.729
Weighted Avg.	0.668	0.354	0.729

by using WEKA [24] machine learning toolkit. Table 7.6 shows the performance of all the classifiers we use to build the classification model to predict users' preferable night-time activity. The table

presents the classification results of our built model in terms of true positive rate(TPR), false positive rate(FPR), area under the ROC curve(AUC), mean absolute error(MAE) and root mean squared error(RMSE) for predicting preferable night-time activity of user from given weather condition. We calculate the performance of the classifier by using AUC values under 10-fold cross validation. From the table we can see that the performance of Random Forest Tree Ensemble is best. So, finally we choose Random Forest Tree Ensemble as our classifier. Table 7.7 shows the performance of our build model. We find that on an average the AUC value of our classifier is 72.9%. We also find that mean absolute error (MAE) of our model is 0.3931. The AUC value is showing moderate performance.

7.3 Discussion

Table 7.8: Tokyo city dataset for building the classification model for preferable night-time activity

Class	Number of instances	Percentage
Visiting nightlife spot	1882	72.7%
Staying at home	705	27.3%
Total	2587	100%

Table 7.9: Performance of different classifiers for building the model preferable night-time activity on Tokyo city dataset

Classifiers	AUC	TPR	FPR	MAE	RMSE
NaiveBayes	0.537	0.695	0.705	0.43	0.4664
RandomForest	0.652	0.712	0.602	0.3563	0.4423
RandomTree	0.576	0.667	0.516	0.3328	0.5767
RepTree	0.667	0.714	0.670	0.3774	0.4551

We have another dataset of 2587 instances of Tokyo city. Table 7.8 shows the number of instances per each class level. We apply Naive Bayes, Random Forest, Random Tree and RepTree classifiers on these dataset. Table 7.9 shows the performance of all the classifiers we used to build the model to predict users' preferable day-time activity mode. From the table we see that Random Forest Tree Ensemble

shows the best performance. So, finally we use it for building our classification model. Table 6.11 shows the performance of our build model.

Table 7.10: Classification results of the model preferable night-time activity build using Random Forest algorithm on Tokyo city dataset

Class	TPR	FPR	AUC
Visiting night life spot	0.900	0.790	0.652
Staying at home	0.210	0.100	0.652
Weighted Avg.	0.712	0.602	0.652

This dataset is imbalanced and we can see that while the class *Visiting nightlife spot* has 72.7% of instances of the total dataset, the class *Staying at home* has only 27.3% of dataset. So classifier performs poorly and having TPR rate of 9% for one class and 2.1% for another class. Again the FPR rate of the first class is 7.9% while it is only 1% for the another class. That means the model is classifying almost all the instances as the majority class and it is a very common phenomena for the classifiers for such kind of imbalanced datasets. We observed that our model performs pretty well for the previous dataset of Table 7.1 which is a balanced dataset.

Chapter 8

Building Model for Future Visit Prediction

Our last model is for the prediction of user's preferable visiting place on different weather conditions. Our model has four different class levels or visiting places. They are *Park*, *Harbor / Marina*, *Indoor Museum* and *Sea Beach*. Table 8.1 shows the dataset we used for building the model. All the three classes except the class *Sea Beach* has almost similar number of instances. Only the class *Sea Beach* has very few number of instances of only 0.2%. So, the dataset is imbalanced for this class.

Table 8.1: Tokyo city dataset for building the classification model for preferable visiting places

Class	Number of instances	Percentage
Park	7206	30.8%
Harbor / Marina	9051	38.7%
Indoor Museum	7077	30.3%
Sea Beach	44	0.2%
Total	23378	100%

Now we need to select the appropriate features from 13 attributes of our independent variable weather condition for building the model as discussed in the previous three chapters. The process is described in the following sections.

8.1 Feature selection

This dataset also contains 14 different attributes. First 13 attributes contain the weather information and the 14th attribute is the check-in place represented as visiting place. Weather information is independent variable and visiting place is the dependent variable. As independent variable weather information is mixed type of variable containing both categorical and numerical values, we need to follow two different techniques for feature selection.

8.1.1 Feature selection for categorical weather information and categorical visiting place

We conduct Chi-Square test to check the correlation between *preferable visiting place* and weather attribute *weather summary* and *weather icon*. We found that users' preferable visiting place and weather conditions are related. For the attribute *weather summary* we found the value $\chi^2 = 26.793$ and the $df = 15$. For the attribute *weather icon* we found the value $\chi^2 = 723.629$ and the $df = 24$. It's calculated using Equation 5.3. Table 8.2 shows the result of chi-square test between visiting place and *weather summary* and Table 8.3 shows the result of chi-square test between visiting place and *weather icon*. As we find both the attributes *weather summary* and *weather icon* are statistically significant so primarily we select this two attributes as feature variable for building our classification model.

Table 8.2: Chi-Square test between preferable visiting place and weather summary

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	26.793	15	0.030
Likelihood Ratio	27.031	15	0.028
N of Valid Cases	23378		

Table 8.3: Chi-Square test between preferable visiting place and weather icon

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	723.629	24	0.000
Likelihood Ratio	742.546	24	0.000
N of Valid Cases	23378		

8.1.2 Feature selection for continuous weather information and categorical visiting place

In our dataset we have 11 numerical or continuous variables same as the earlier datasets used to build the previous models. For ranking the predictors importance or determine factors that discriminate between classes, again we use Fisher's linear discriminant analysis (LDA). Our dependent variable preferable day-time activity has 4 different class levels. Our last dataset also meets all the assumptions of discriminant analysis, otherwise we can not apply it. We discussed the assumptions in Section 5.1.2. All the four class levels, *Park*, *Harbor / Marina*, *Indoor Museum* and *Sea Beach* of our dependent variable preferable visiting place are mutually exclusive. All the 11 predictor variables are independent to each other, normally distributed and there is no outliers in our dataset. Sample size also meets the assumption. We test our dataset and remove the multicollinearity [21] problem.

Then we compute multi-level discriminant analysis using SPSS where 11 weather attributes are predictors and preferable visiting place is the dependent variable. We calculate the Pooled Within-group Correlation matrix which provide bivariate correlations between our 11 predictors. We find that same as the previous 3 models the 3 attributes *temperature*, *apparent temperature* and *dew point* are highly correlated with each other having correlation coefficients larger than 0.8. Table 8.4 shows the coefficient values. All other coefficients are less than 0.8. So, at this point we need to choose only 1 predictor from these 3 predictors. To do so we compute LDA 3 times individually using each of the 3 predictors along with the other 8 predictors each time. We find that the predictor *apparent temperature* affects most in the classification as before. So we choose *apparent temperature* among the 3 predictors and discard the attribute *temperature* and *dew point* as our feature to solve this multicollinearity problem.

Table 8.4: Pooled Within-Groups Matrices

Correlation	Temperature	Apparent Temperature	Dew Point
Temperature	1.000	0.995	0.890
Apparent Temperature	0.995	1.000	0.896
Dew Point	0.890	0.896	1.000

Table 8.5: Fisher's linear discriminant function coefficients between weather attributes and visiting place

Predictors	Traveling	Shopping at mall	Watching movie in theater	Staying at home
precipitation intensity	264.612	262.562	263.131	264.969
precipitation probability	2.194	2.449	2.461	1.904
Apparent temperature	4.193	4.198	4.197	4.221
Humidity	211.886	210.541	209.407	210.525
Wind speed	8.633	8.641	8.611	8.705
Wind bearing	0.346	0.346	0.346	0.344
Visibility	20.170	20.270	20.309	20.333
Cloud coverage	-72.028	-71.734	-71.251	-71.755
Air pressure	23.723	23.721	23.725	23.717

Table 8.5 shows Fisher's linear discriminant function coefficients between weather attributes and visiting place. If we compare the values between groups, the higher coefficient means the variable attributes more for that group. From Table 8.5 we can see that the attribute *cloud coverage* and *wind bearing* have a very low coefficient value, so we discard them from our feature variables. Finally, we select 7 attributes from 11 numeric attributes as features for building our classification model.

8.1.3 Feature selection using automated subset selection method

Again, to validate our feature selection process we use the automated subset selection method using *leaps* [22] R package implementation. We select the best subsets of features of size 7, 8 and 9 using the exhaustive selection algorithm [23]. Finally we choose 7 numeric attributes for building our classification model and they are *precipitation intensity*, *precipitation probability*, *apparent temperature*, *humidity*, *wind speed*, *visibility* and *air pressure*. We also use 2 nominal or categorical variable *weather summary* and *weather icon* as our feature and our dependent variable is only categorical variable preferable visiting place.

8.2 Building the Classification Model

To build the classification model finally we choose 9 independent or predictor variables and one dependent variable. Table 8.6 shows the selected features for building our classification model and Table 8.1 shows our dataset.

Table 8.6: Selected features for the classification model

Predictors	Dependent Variable
<i>weather summary</i>	Preferable Visiting Place
<i>weather icon</i>	
<i>precipitation intensity</i>	
<i>precipitation probability</i>	
<i>apparent temperature</i>	
<i>humidity</i>	
<i>wind speed</i>	
<i>visibility</i>	
<i>air pressure</i>	

Then we apply Naive Bayes, Random Forest, Random Tree and RepTree classifiers in our dataset by using WEKA [24] machine learning toolkit. Table 8.7 shows the performance of all the classifiers we

use to build the classification model to predict users' preferable visiting place. The table presents the classification results of our built model in terms of true positive rate(TPR), false positive rate(FPR), area under the ROC curve(AUC), mean absolute error(MAE) and root mean squared error(RMSE) for predicting preferable visiting place of user from given weather condition. We calculate the performance of the classifier by using AUC values under 10-fold cross validation. From the table we can see that the performance of Random Forest Tree Ensemble is best. So, finally we choose Random Forest Tree Ensemble as our classifier. Table 8.8 shows the performance of our build model. We find that on an average the AUC value of our classifier is 67.1%. We also find that mean absolute error(MAE) of our model is 0.2871.

Table 8.7: Performance of different classifiers for building the model preferable visiting place

Classifier	AUC	TPR	FPR	MAE	RMSE
NaiveBayes	0.564	0.391	0.327	0.3251	0.408
RandomForest	0.671	0.492	0.262	0.2871	0.3956
RandomTree	0.582	0.445	0.281	0.2775	0.5268
RepTree	0.604	0.434	0.295	0.3041	0.4212

Table 8.8: Classification results of the model preferable visiting place build using Random Forest algorithm

Class	TPR	FPR	AUC
Park	0.390	0.235	0.620
Harbor / Marina	0.549	0.332	0.658
Indoor Museum	0.527	0.202	0.739
Sea Beach	0.023	0.000	0.641
Weighted Avg.	0.492	0.262	0.671

The AUC value is showing moderate performance. The TPR rate of all the classes except *Sea Beach* is moderate. This is the minority class having very few number of instances. From Table 8.1 we see that the class *Sea Beach* has only 0.2% of instances where all other classes have more than 30% of instances. This dataset is class imbalanced. So, we need to handle this class imbalance problem of our

dataset to improve the accuracy of our model.

8.3 Handling Class Imbalance Problem Using SMOTE Algorithm

Now we need to follow an approach for solving the class imbalance problem using re-sampling technique. As we have discussed in Section 5.4, we use SMOTE re-sampling technique to solve this problem.

Table 8.9: Up-Sampled dataset after applying SMOTE

Class	Number of instances	Percentage
Park	7206	30.4%
Harbor / Marina	9051	38.2%
Indoor Museum	7077	29.9%
Sea Beach	352	1.5%
Total	23686	100%

Table 8.9 shows the class distribution we find after applying SMOTE on our class imbalanced dataset. We increase the number of instances of the minority class *Sea Beach* up to 3 times than previous, of our dataset. Previously the class had 0.2% of instances of the full dataset. We up-sample this class 3 times and finally we get this class having 1.5% of instances of the total dataset.

Table 8.10: Performance of different classifiers for building the model preferable visiting place after up-sampling

Classifiers	AUC	TPR	FPR	MAE	RMSE
NaiveBayes	0.563	0.374	0.304	0.3361	0.415
RandomForest	0.680	0.498	0.254	0.286	0.3941
RandomTree	0.589	0.451	0.272	0.2746	0.524
RepTree	0.608	0.440	0.285	0.3047	0.4222

Table 8.11: Classification results of the model preferable visiting place after up-sampling the dataset using SMOTE algorithm

Class	TPR	FPR	AUC
Park	0.387	0.228	0.626
Harbor / Marina	0.551	0.325	0.665
Indoor Museum	0.527	0.203	0.739
Sea Beach	0.781	0.001	0.980
Weighted Avg.	0.498	0.254	0.680

8.4 Re-Building the Classification Model

Again we apply Naive Bayes, Random Forest, Random Tree and RepTree classifiers in our dataset by using WEKA [24] machine learning toolkit. Table 8.10 shows the performance of all the classifiers we used to build the classification model to predict users' preferable visiting place. The table presents the classification results of our built models in terms of true positive rate(TPR), false positive rate(FPR), area under the ROC curve(AUC), mean absolute error(MAE) and root mean squared error(RMSE) for predicting preferable visiting places of user from given weather condition. We calculate the performance of the classifier by using AUC values under 10-fold cross validation. From the table we can see that the performance of Random Forest Tree Ensemble is best. So, finally we choose Random Forest Tree Ensemble as our classifier. Table 8.11 shows the performance of our build model. We find that on an average the AUC value of our classifier is 68.0%. We also find that mean absolute error(MAE) of our model is 0.286. All classes have moderate TPR rate and low FPR rate. The class *Sea Beach* has good correlation with Clear weather. *Indoor Museum* has a correlation with Cloudy weather.

8.5 Discussion

We have another dataset of 9,537 instances of New York city. Table 8.12 shows the number of instances per each class level. We apply Naive Bayes, Random Forest, Random Tree and RepTree classifiers on these dataset. Table 8.13 shows the performance of all the classifiers we used to build the model to

predict users' preferable transport mode. From the table we see that Random Forest Tree Ensemble shows the best performance. So, finally we use it for building our classification model. Table 8.14 shows the performance of our build model.

Table 8.12: Dataset of New York city used for building the classification model for preferable visiting place

Class	Number of instances	Percentage
Park	4804	50.4%
Harbor / Marina	2906	30.5%
Indoor Museum	1252	13.1%
Sea Beach	575	6.0%
Total	9537	100%

Table 8.13: Performance of different classifiers for building the model preferable visiting place on New York dataset

Classifiers	AUC	TPR	FPR	MAE	RMSE
NaiveBayes	0.547	0.457	0.415	0.3194	0.4205
RandomForest	0.627	0.514	0.355	0.2855	0.3962
RandomTree	0.565	0.448	0.320	0.2757	0.525
RepTree	0.557	0.491	0.437	0.3046	0.4113

From Table 8.14 we can see that the performance of this model is not good though the AUC value is showing moderate result of 62.7% on an average. Only the class *Park* has good TPR rate of 71.1% though its FPR rate is also high and it is 56.5%. All other classes have low TPR rate. The dataset is imbalanced and except the class *Park* all other classes have also low FPR rate dew to the lack of instances.

From the last four chapters we are discussing the details of our model building approach and how we build all the four classification models to solve our formulated four prediction models. We also discuss the performance results of our classification models. In the next chapter, we will conclude our

Table 8.14: Classification results of the model preferable visiting place build using Random Forest algorithm on New York dataset

Class	TPR	FPR	AUC
Park	0.711	0.565	0.611
Harbor / Marina	0.359	0.204	0.625
Indoor Museum	0.218	0.055	0.637
Sea Beach	0.292	0.017	0.757
Weighted Avg.	0.514	0.355	0.627

dissertation by recapitulating our work, contributions and findings.

Chapter 9

Conclusions

In this thesis we have developed techniques for weather-aware prediction of users' activities, future visiting places and preferable modes of transport from user generated geo-tagged data created by location based social networking sites. We formulate a new problem of finding users' activity based on weather condition from user generated social media data, as it is an indispensable tool for extracting latent contents about users. We find it very interesting that if users' activity or mobility pattern can be analyzed and if we can relate how these are influenced by weather condition, it will introduce a new dimension to the target marketing for businesses. Our thesis has the following contributions:

First, though check-in datasets of various LBSN sites like Facebook, Foursquare, Twitter are available but no check-in dataset contains weather information. We have created 2 new datasets containing weather information of each individual check-ins while there was no such dataset available for research. We have exploited the data fusion of two different data source, i) Foursquare and ii) forecastio weather service's weather information to build the models that can accurately predict various kind of preferences of user regarding activity, visiting place or transport mode.

Second, we have built several models to find the correlation between weather condition and users' activity, visiting place and transport mode. Our dataset contains mixed type of variables containing both categorical and numerical values. So we have used two different approaches to handle such mixed type of independent variables for feature selection.

Third, the dataset we have used, has data sparsity problem. It was a big challenge for us to find the appropriate class levels for building the models. We have chosen only the classes having moderate number of instances. So, our models are not representing all possible types of activities, visiting places or transport modes. But our proposed approach is applicable for building models having other types of visiting locations, activities or transport modes. Our datasets also have class imbalance problem and we have solved this issue.

Fourth, we have formulated four different problems to solve. Our first problem is to predict users' preferable transport mode on various weather condition. Then second problem is to predict users' day-time activity, and the third problem is to predict user's night-time activity based on a given weather condition. Finally the last problem is to predict users' preferable visiting place on a given weather condition. To address the problems we have proposed machine learning based classification model building approach and built four different prediction models showing moderate performances. We have created eight different datasets for building the four models from the original two datasets of New York and Tokyo city. We have observed that among all the datasets, balanced ones build high performance models. We have exploited four different techniques of building classification models, Naive Bayes, Random Forest, Random Tree and RepTree and found that ensemble model performs better than independent models. Our prediction models for future visit shows AUC of 68% and 62.7% for Tokyo and New York city dataset respectively. The prediction models for day-time activity shows AUC of 65.1% and 69.6% and prediction model for night-time activity shows AUC of 72.9% and 65.2% for New York and Tokyo city dataset respectively. Prediction model for preferable transport mode shows AUC of 85.1% after handling class imbalance problem on Tokyo city dataset and 64.6% on New York city dataset.

Our models have wide range of real life applications including target marketing, traveling place recommendation and policy making for tourism management. In future we plan to integrate our model with a real recommendation application system used by any kind of business, travel agent or tourism management system.

References

- [1] Yong Liu, Vassilis Kostakos, and Hongxiu Li. Climatic effects on planning behavior. *PloS one*, 10(5):e0126205, 2015.
- [2] Teerayut Horanont, Santi Phithakkitnukoon, Tuck W Leong, Yoshihide Sekimoto, and Ryosuke Shibasaki. Weather effects on the patterns of people’s everyday activities: a study using gps traces of mobile phone users. *PloS one*, 8(12):e81153, 2013.
- [3] Z Spasova. The effect of weather and its changes on emotional state–individual characteristics that make us vulnerable. *Advances in Science and Research*, 6(1):281–290, 2012.
- [4] Kevin E Trenberth, Kathleen Miller, Linda Mearns, and Steven Rhodes. *Effects of changing climate on weather and human activities*. University Science Books Sausalito, CA, 2000.
- [5] Edgar Howarth and Michael S Hoffman. A multidimensional approach to the relationship between mood and weather. *British Journal of Psychology*, 75(1):15–23, 1984.
- [6] Farhana Ahmed, G Rose, and C Jacob. Impact of weather on commuter cyclist behaviour and implications for climate change adaptation. In *Australasian Transport Research Forum (ATRF), 33rd, 2010, Canberra, ACT, Australia*, volume 33, 2010.
- [7] Sui Tao, Jonathan Corcoran, Mark Hickman, and Robert Stimson. The influence of weather on local geographical patterns of bus usage. *Journal of transport geography*, 54:66–80, 2016.
- [8] Mario Cools, Elke Moons, Lieve Creemers, and Geert Wets. Changes in travel behavior in response to weather conditions: do type of weather and trip purpose matter? *Transportation Research Record: Journal of the Transportation Research Board*, (2157):22–28, 2010.

-
- [9] Chengxi Liu, Yusak O Susilo, and Anders Karlström. Weather variability and travel behaviour—what we know and what we do not know. *Transport reviews*, 37(6):715–741, 2017.
- [10] Muhammad Sabir, Jv Ommeren, Mark J Koetse, and Piet Rietveld. Impact of weather on daily travel demand. *Work Pap Dep Spat Econ VU Univ Amsterdam*, 2010.
- [11] Fabio Pianese, Xueli An, Fahim Kawsar, and Hiroki Ishizuka. Discovering and predicting user routines by differential analysis of social network traces. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2013 IEEE 14th International Symposium and Workshops on a*, pages 1–9. IEEE, 2013.
- [12] Samiul Hasan and Satish V Ukkusuri. Location contexts of user check-ins to model urban geo life-style patterns. *PloS one*, 10(5):e0124819, 2015.
- [13] Emre Çelikten, Géraud Le Falher, and Michael Mathioudakis. Extracting patterns of urban activity from geotagged social data. *arXiv preprint arXiv:1604.04649*, 2016.
- [14] Yoon-Sik Cho, Greg Ver Steeg, and Aram Galstyan. Where and why users’ check in”. In *AAAI*, pages 269–275, 2014.
- [15] Dingqi Yang, Daqing Zhang, Vincent W Zheng, and Zhiyong Yu. Modeling user activity preference by leveraging user spatial temporal characteristics in lbsns. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 45(1):129–142, 2015.
- [16] Wen-Haw Chong, Bing-Tian Dai, and Ee-Peng Lim. Prediction of venues in foursquare using flipped topic models. In *European Conference on Information Retrieval*, pages 623–634. Springer, 2015.
- [17] Qunying Huang. Mining online footprints to predict user’s next location. *International Journal of Geographical Information Science*, pages 1–19, 2016.
- [18] Yoon-Sik Cho, Greg Ver Steeg, and Aram Galstyan. Socially relevant venue clustering from check-in data. Chicago, IL, 08/2013 2013. URL http://snap.stanford.edu/mlg2013/submissions/mlg2013_submission_22.pdf.
- [19] Yan Qu and Jun Zhang. Trade area analysis using user generated mobile location data. In *Proceedings of the 22Nd International Conference on World Wide Web, WWW ’13*, pages 1053–

- 1064, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-2035-1. doi: 10.1145/2488388.2488480. URL <http://doi.acm.org/10.1145/2488388.2488480>.
- [20] Roberto Rösler and Thomas Liebig. Using data from location based social networks for urban activity clustering. In *Geographic Information Science at the Heart of Europe*, pages 55–72. Springer International Publishing, 2013.
- [21] O. corporation., interpreting results of discriminant analysis. <https://www.originlab.com/doc/Origin-Help/DiscAnalysis-Result>. Accessed: 2018-05-26.
- [22] Thomas Lumley and A Miller. Leaps: regression subset selection. r package version 2.9. 2009.
- [23] Variable selection using automatic methods. <https://www.r-bloggers.com/variable-selection-using-automatic-methods/>. Accessed: 2018-05-26.
- [24] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18, November 2009. ISSN 1931-0145. doi: 10.1145/1656274.1656278. URL <http://doi.acm.org/10.1145/1656274.1656278>.
- [25] Runtao Yang, Chengjin Zhang, Lina Zhang, and Rui Gao. A two-step feature selection method to predict cancerlectins by multiview features and synthetic minority oversampling technique. *BioMed research international*, 2018, 2018.
- [26] Renhao Liu, Lawrence O Hall, Kevin W Bowyer, Dmitry B Goldgof, Robert Gatenby, and Kaoutar Ben Ahmed. Synthetic minority image over-sampling technique: How to improve auc for glioblastoma patient survival prediction. In *Systems, Man, and Cybernetics (SMC), 2017 IEEE International Conference on*, pages 1357–1362. IEEE, 2017.
- [27] Yu-Dong Zhang, Yin Zhang, Preetha Phillips, Zhengchao Dong, and Shuihua Wang. Synthetic minority oversampling technique and fractal dimension for identifying multiple sclerosis. *Fractals*, 25(04):1740010, 2017.
- [28] Yu-Dong Zhang, Guihu Zhao, Junding Sun, Xiaosheng Wu, Zhi-Heng Wang, Hong-Min Liu, Vishnu Varthanan Govindaraj, Tianmin Zhan, and Jianwu Li. Smart pathological brain detection by synthetic minority oversampling technique, extreme learning machine, and jaya algorithm. *Multimedia Tools and Applications*, pages 1–20, 2017.

-
- [29] Cangzhi Jia and Yun Zuo. S-sulfpred: A sensitive predictor to capture s-sulfenylation sites based on a resampling one-sided selection undersampling-synthetic minority oversampling technique. *Journal of theoretical biology*, 422:84–89, 2017.
- [30] Transportation in new york city. https://en.wikipedia.org/wiki/Transportation_in_New_York_City. Accessed: 2018-09-23.