# FINDING SUITABLE LOCATIONS FOR LIVE CAMPAIGNS USING DIFFERENT MACHINE LEARNING TECHNIQUES

**SUBMISSION DATE: 12.10.17**

## SUBMITTED BY:

Mohiuddin Md. Abdul Qudar(15101014)

Tahrat Tazrin(15101131)

Kazi Zainab Khanam (15101119)

Department of Computer Science and Engineering

**Supervisor:**

**AmitabhaChakrabarty, Ph.D**

Assistant Professor

Department of Computer Science and Engineering

# Declaration

We, hereby declare that this thesis is based on results we have found ourselves. Materials of work from researchers conducted by others are mentioned in references.

**Signature of Supervisor**

**AmitabhaChakrabarty, Ph.D**
Assistant Professor
Department of Computer Science and Engineering
BRAC University

**Signature of Authors**

**Tahrat Tazrin(15101131)**

**Mohiuddin Md Abdul Qudar**
**(15101014)**

**Kazi Zainab Khanam (15101119)**

# ABSTRACT

For business booming, in recent years the idea of find an ideal location for reaching potential customers has been the focus of past research. Traditional approaches have faced many negative responses, now to make business profitable the new marketing strategy is live campaigns. With the growth of location-based social networks collecting data of user mobility and popularity of places has recently become attainable, but not without analyzing to find the optimal location and time for live campaigns with greater accuracy. In this paper, we study the predictive power of various machine learning and mining features on finding suitable location for live campaigns through the use of a dataset collected from Foursquare in New York. We selected 10 candidate areas where the data was preprocessed according to the feature, a score is computed on the candidate areas to do live campaigns based on the features using most suitable algorithm with the accuracy. The results with Random Forest and Decision Tree are shown at the end of the report. Lastly, our proposed model shows how performance varies when using different features and predicting the suitable locations for live campaigns. We achieve 88.25% accuracy in Decision Tree regression model and an accuracy of 88.48% and 70.04% in Support Vector Machine (SVM) and Random Forest respectively.

**Keywords**- LBSN, Live Campaigns, Location, algorithms, SVM, Random forest, Decision Tree

# Acknowledgement

First of all, we would like to express our gratitude to Almighty Allah for keeping us safe and sound to initiate the research work and to put our best efforts and successfully complete it.

I would also like to thank my supervisor, Amitabha Chakrabarty, Ph.D, for guiding me to my aim, patiently handling each and every problem I faced. Most importantly, I would like to thank sir for giving me the opportunity to do thesis under his guidance.

Lastly I would also like to thank my family, friends, and all the teachers of School of Engineering and Computer Science of BRAC University for relentless support and motivation.

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# CHAPTER 1

## Introduction

Our social options are endless, human relationships are constrained in many ways. They take time, energy, and often money to maintain. Even after considering these barriers people interact with each other regularly and mostly through online. In practice these highly dynamic real time social sites like Facebook, Twitter, YouTube are producing tremendous amount of data. Together with services for location sharing like Foursquare, Facebook places and Google Latitude we can now activities, actions, agility and aisle of millions of people [1]. In terms of scale, the Foursquare service alone claims over 6 million registered users (Foursquare 2011) and nearly 1 million check-ins per day (Grove 2010). Like similar services, Foursquare allows users to "check in" at different venues (e.g., grocery stores, restaurants), write tips, and upload pictures and videos [2]. Due to exponential increase in smartphones around the world, location based services (LBS) has become tremendously popular. Smart phones with Global Positioning System (GPS) modules have the ability to produce the location to its user. This took a turn in the applications of numerous field involving location based services. For example a photo with phone's camera now show the geographical position, which can easily be uploaded in any social site. People are more willing to share their location even when not with friends, in the social sites. These Exabyte of social, temporal and spatial data gives us huge opportunity in various fields of research like business, tourism, marketing, and planning for urban and rural city. With the growth of location-based social networks collecting data of user mobility and popularity of places has recently become attainable, but not without analyzing to find the optimal location and time for live campaigns with greater accuracy. To get the peak results in any fields we must be able to take the paramount benefit from these data. Countless exhaustive studies have been conducted to enquire about the customer’s behavior and their acceptance towards mobile and internet advertising. Most studies have concluded that majority of the public find advertising information useless and extremely annoying [4]. As advertisement is getting a negative response so a primary way of reaching the customers is through live campaigns in suitable locations. Traditional television and internet advertisements are no longer effective now. For increasing business value it is necessary to analyses this strategy by using machine learning for finding a suitable place for live campaigning. With the above in mind, the report is aimed to investigate the features that are most predictive to finding suitable location and recommending the best analytical strategy in using the check-in data of Foursquare to improve the effectiveness of location-based advertising.

## 1.1 Motivation

Machine learning is the most demanding topic in computer science. More than half of all current ongoing researches are on this machine learning field. Numerous researches have also been conducted in improving the field of advertisement as well as business is hugely important in a country's economy because it is the main economic engine for the country and advertising helps to keep the people informed about whatever new products or services are available in market.

## 1.2 Contribution Summary:

Our overall contributions are summarized as below:

- At first, a problem is formulated for finding the suitable locations for live campaigns
- Then several features are extracted and preprocessed, for testing and training purpose, the features are extracted from the available dataset
- After this, the impact of this features are shown by implementing various machine learning algorithm that consists of SVM, Decision Tree learning and Random Forest.
- Finally, the results from the exhaustive experiments of implementing the features with the algorithms shows the effectiveness of this approach.

## 1.3 Methodology

At first, a set of candidate areas from a set of areas where the commercial enterprises might be interested in arranging the campaign will be selected. The data of the candidate area such as its venue Id, check-in user Id, venue category Id, the time of the check-in and the time zone offset of the respective location is collected from available dataset of any check-in based social media like Foursquare. Then the dataset is analyzed with respect to various spatial and temporal features such as check-in count, check-in consistency, check-in identity, openness, density and temporal signal. A new dataset is then prepared after the respective data of each of the features are calculated, where the data needs to be converted into numerical values and also all the null values of the x attributes are eliminated with the minimum value. Finally, machine-learning algorithms are implemented on the extracted pre-processed data implementing various test and train size and finally a score is computed of the candidate areas based on the most suitable algorithm with the best accuracy result.

### 1.4 Thesis Overview

The organization of this report is as follows. Chapter 2 is Literature Review presented Related Works similar to our work. Chapter 3 covers our proposed Model and its methodology. Chapter 4 covers the details of Algorithms used in this thesis. Chapter 5 is designed with the Methodology of implementation. Result Analysis and Comparison of accuracy of the predicted results using all algorithms are described in Chapter 6. Lastly, Chapter 7 represents our Future plan and concludes of the report.

# CHAPTER 2

# Literature Review

This chapter contains literature review related with Background which includes Machine Learning Algorithm (ML) and Global Positioning System (GPS).This chapter also refers Related Works. Besides, this chapter will also give information about our research ac

## 2.1 Finding suitable locations for live Campaign

Designing an efficient recommender system has been a major research area in the field of Machine Learning of computer vision. Multiple Machine learning algorithms have been implemented in order to execute a productive recommender system. We are greatly inspired by the work of Rahman and Nayeem [1]. Our paper is closely related to their work. They have used (SVM) Regression model to identify the optimal location and time to carry a live campaign. In their paper, they have analyzed spatial and temporal features which they have extracted and how those features influence the location based social services.

## 2.2 Event Recommender System

Georgiev, Noulas and Mascolo took the initiative of designing an event recommender system to determine the predictability of a person attending a future event or not based on the check-in data collected from the Foursquare application [5]. In their work they have mainly focused on multiple social, temporal and spatial features which can influence a user's choice towards any event, so, they found out similarities between the user's profile data in contrast with the recorded data of the previous check-in activities of the highly potential event participants. Hence, they formulated an event prognosis task which would rank the events for each individual user based on multiple prediction features. In order to prove one of their hypothesis that – the user profile whose behavior is closest to the mass number of targeted participants in any event, then it is more likely for the user to get attracted to the event, to materialize this hypothesis, for the preference of the place type of each user, they treated users as document and place type as term into a vector of real numbers.in Fig 2.2.1 shows  They shows a directed weight graph (called socio-spatial graph) that they have designed to measure the degree of how friends attending events can influence the user, in which they synthesized both the social and spatial factors which connects the users, place types and events. On this graph, they implemented random walks with restart to evaluate the chances of participation for a user-pair event.

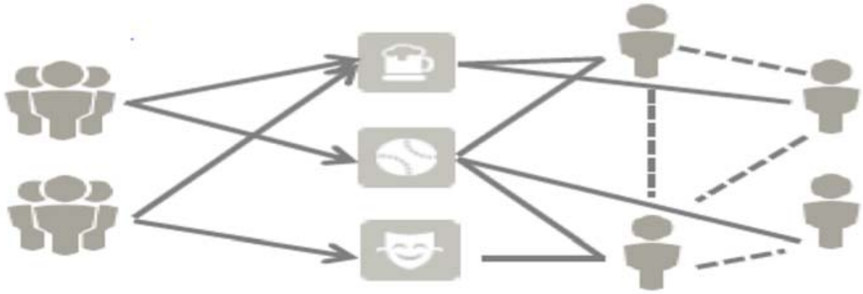


**Fig. 2.2.1 a directed weight graph [5]**

Sklar, Shaw and Hogue have designed a real-time event detection engine that is based on a probabilistic model which provides the negative binomial distribution over the number of people checked-in at any given time, for measuring how unusually busy a place becomes [6]. Quercia et al. provided cold-start event recommendations for users whose home location is known, when they studied the event prediction problem, they have implemented their studies by using the K – Nearest Neighbor machine learning algorithm [12]. However, they did not address the personalization in their work. Lee combined trust relations together with explicit user feedback to recommend cultural events as users cannot choose or evaluate the recommendation which are based on collaborative filtering recommenders, hence they have proposed a trust based cultural recommender in which users can rate the trustworthiness of the other users, a recommender is generated based on their respective ratings[13].During which Minkov et al showed in his work successfully that collaborative ranking predictions of future events are much more productive than pure content-based recommendation[14]. He combined content based with collaborative filtering approaches to draw user selections on latent topics which were hidden based on the scientific talk announcements. Firstly, they examined that whether collaborative LowRank method can enhance the performance for an event recommendation system, then they compared their LowRank against RankSVM method, for which a separate model was also generated for each user which is displayed in fig 2.2.2 Their results showed that LowRank using TFIDF features provides high level performance over the entire range of the inspected training set size, as they had more sparse features in their case LowRank method had a more direct control.

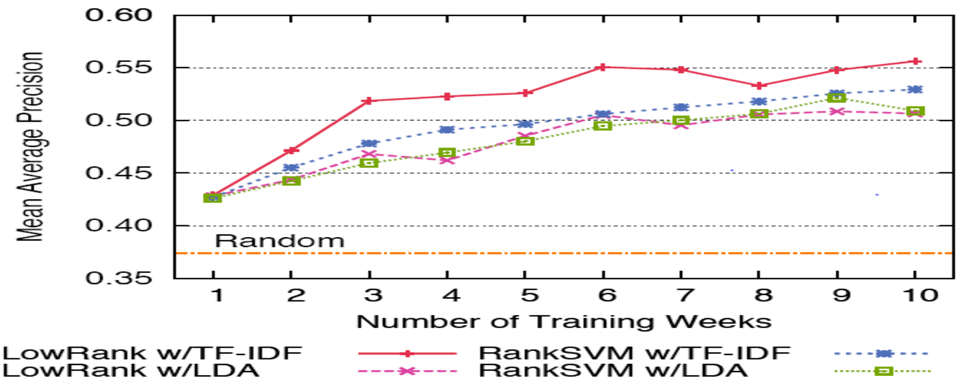


**Figure. 2.2.2 Comparison of RankSVM and LowRank algorithm. [14]**

Later on, Liao et al. designed a system based on offline spontaneous interactions and co-attendance information, Liao et al designed a latent model that would suggest the related events in offline ephemeral social networks formed around conference talks [15].

### 2.3 Finding optimal locations for Commercial Activities

Jensen looked for the suitable geographical positioning of the retail stores, in his work, he implemented a spatial network based formulation for locating retail stores, where the nodes are 55 different types of retail stores and weighted signal links are defined to model attraction and repulsion attributes in the network[16][17]. Porta et al proposed an approach, where the authors' studies were based on the exhaustive analysis of the spatial distribution of commercial activities; they analyzed the relationship between street centrality and retail store density, validating how the former achieves an important role in the development of urban structure and land usage[7][8]. Karamshuk et al. extended the results by combining the geographical and mobility features, which showed better results on the quality of an area for becoming a potential spot to be opened as new retail store [9]. They specifically studied the problem of finding an optimal location for a new retail store; they addressed a diverse set of data mining features, designing the spatial and semantic information about patterns and places of users' movements in the nearby areas.

Song et al. studied the duration of peoples' activity based on each user's most visited location [10]. They exploited the limitation of predictability in human dynamics by focusing on the mobility patterns of anonymized mobile phone users. Hence, they measured the entropy of each individual's track.

### 2.4 Check-in pattern Analysis

Cheng et al. and Ye et al. showed that semantic information can be released based on the daily and weekly check-in patterns which can enhance location based search and location recommendation[18][19]. Hasan and Ukkusuri's approach was to analyze the macro – scale geo-location data from social networking sites to conclude the individual user's activity patterns. They have exploited the data from traditional surveys, their model provided an activity generation mechanism which is a convenient tool for an activity-travel simulator [21].

### 2.5 Population based Crowd model

Liang et al. proposed a time evolving population model for measuring the number of people departing from a crowd that allows them to predict the number of people remaining in a crowd [20]. They have used check-in data and event related tweets in order to design their crowd based population. They evaluated the period of time each user is spending on the crowd and also the number of people leaving the crowd at any period and the number of posts formed by the crowd. Lastly, they verified their model by predicting a number of posts for some particular events and by predicting the traffic volume for Manhattan. They considered both the event driven crowds and location driven crowds when focusing on the statistical and probabilistic analysis.

## 2.6 Microblogging Social Networks

Lee identified how microblogging social networks can be used as a well-grounded source of data of upcoming events, his approached to extract their spatio-temporal from the messages to improve event awareness [22]. Author has also implemented a density-based online clustering method for investigating microblogging text streams, so that temporal and geospatial features of real-world events can be obtained.

# CHAPTER 3

## Proposed Model

### 3.1 Workflow

In our proposed model, at first, 10 candidate areas from a set of areas where the commercial enterprises might be interested in arranging the campaign will be selected. Then the information of the respective area such as its venue Id, check-in user Id, venue category Id, the time of the check-in and the time zone offset of the respective location is collected from available dataset of any check-in based social media like Foursquare. For each of the candidate area, every location in within the 400m radius will be considered for our calculation. Then the dataset is exhaustively analyzed with respect to various spatial and temporal features like check-in count, check-in consistency, check-in identity, openness, density and temporal signal. The check-in count depicts the number of check-in of the users at a particular day and at a particular hour of the week in each of the unique venues of the candidate areas. The check-in consistency will be calculating the correlation between the weekly and monthly check-in count to ensure consistency in the check-in data. Check-in identity shows the number of unique check-ins i.e the check-in of one user at a particular venue is only once considered. Density denotes the number of unique venue-ids at a particular area and the openness conveys the amount of open space available at a location which is basically the inverse of the density. The temporal signal shows the number of check-in at a particular location in a specific time of day (morning, noon, afternoon, evening). A new dataset is then prepared after the respective data of each of the features are calculated, where the data needs to be converted into numerical values and also all the null values of the x attributes needs to be eliminated by filling up the null space with the minimum value. Lastly machine-learning algorithms are executed on the extracted pre-processed data implementing various test and train size and finally a score is computed of the candidate areas based on the most suitable algorithm with the best accuracy result.

## 3.2 Flowchart

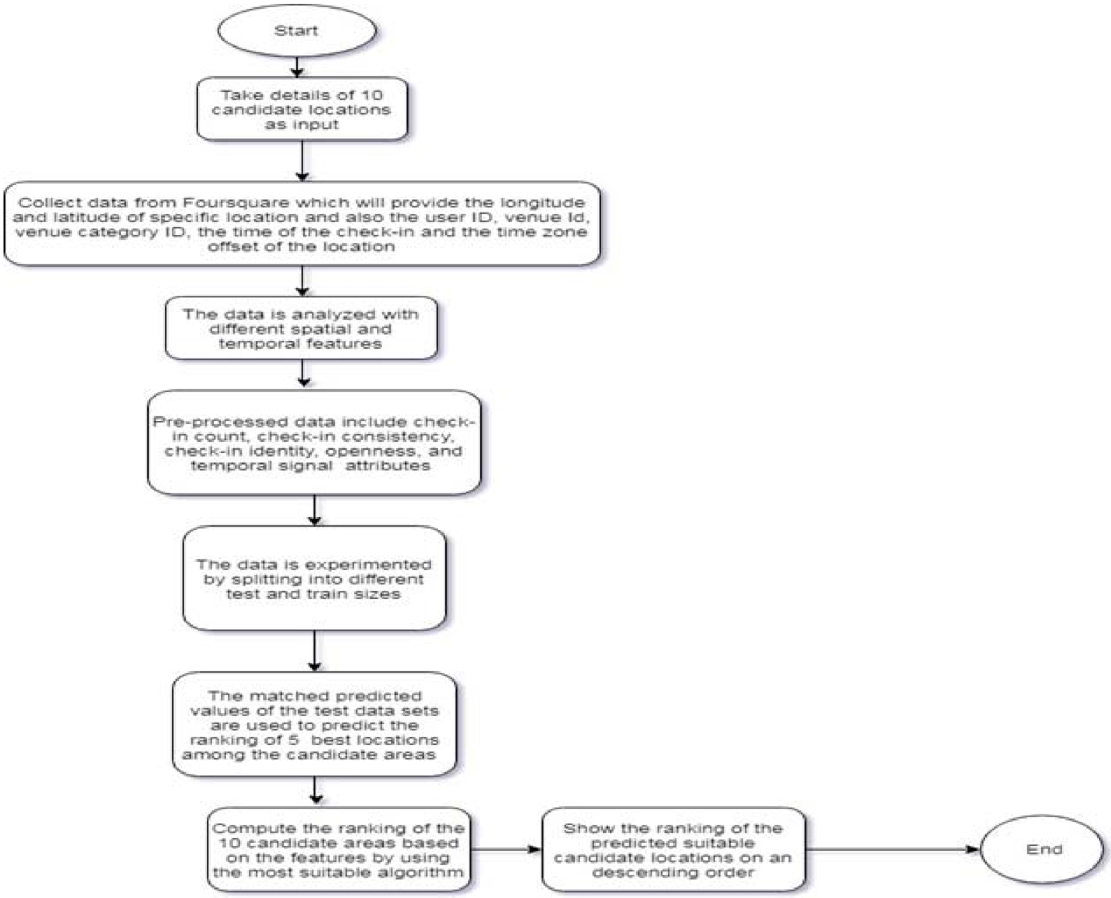

Start
Take details of 10 candidate locations as input
Collect data from Foursquare which will provide the longitude and latitude of specific location and also the user ID, venue Id, venue category ID, the time of the check-in and the time zone offset of the location
The data is analyzed with different spatial and temporal features
Pre-processed data include check-in count, check-in consistency, check-in identity, openness, and temporal signal attributes
The data is experimented by splitting into different test and train sizes
The matched predicted values of the test data sets are used to predict the ranking of 5 best locations among the candidate areas
Compute the ranking of the 10 candidate areas based on the features by using the most suitable algorithm
Show the ranking of the predicted suitable candidate locations on an descending order
End

**Figure 3.2.1 shows the Block diagram of our proposed model**

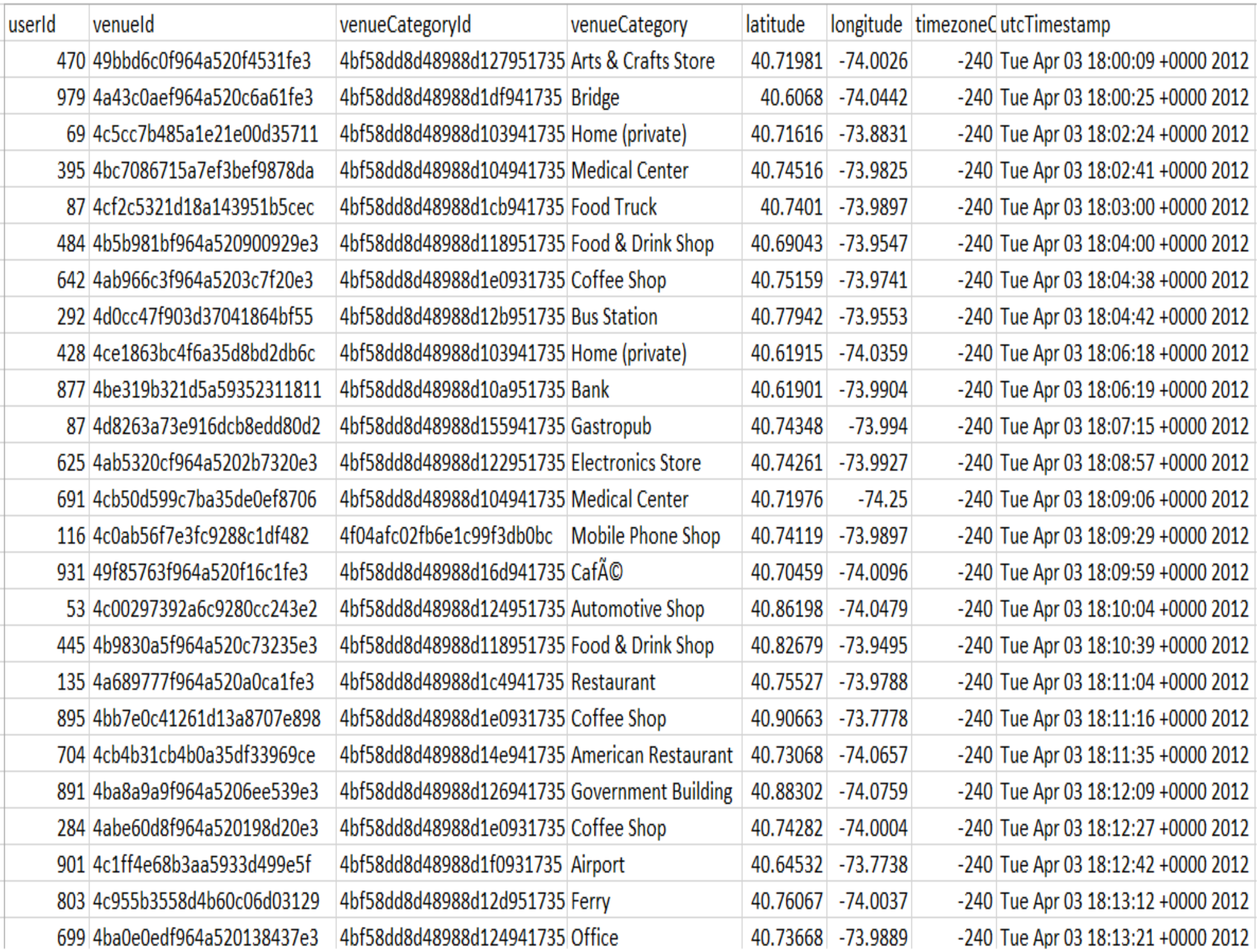
# CHAPTER 4

## Data preprocessing

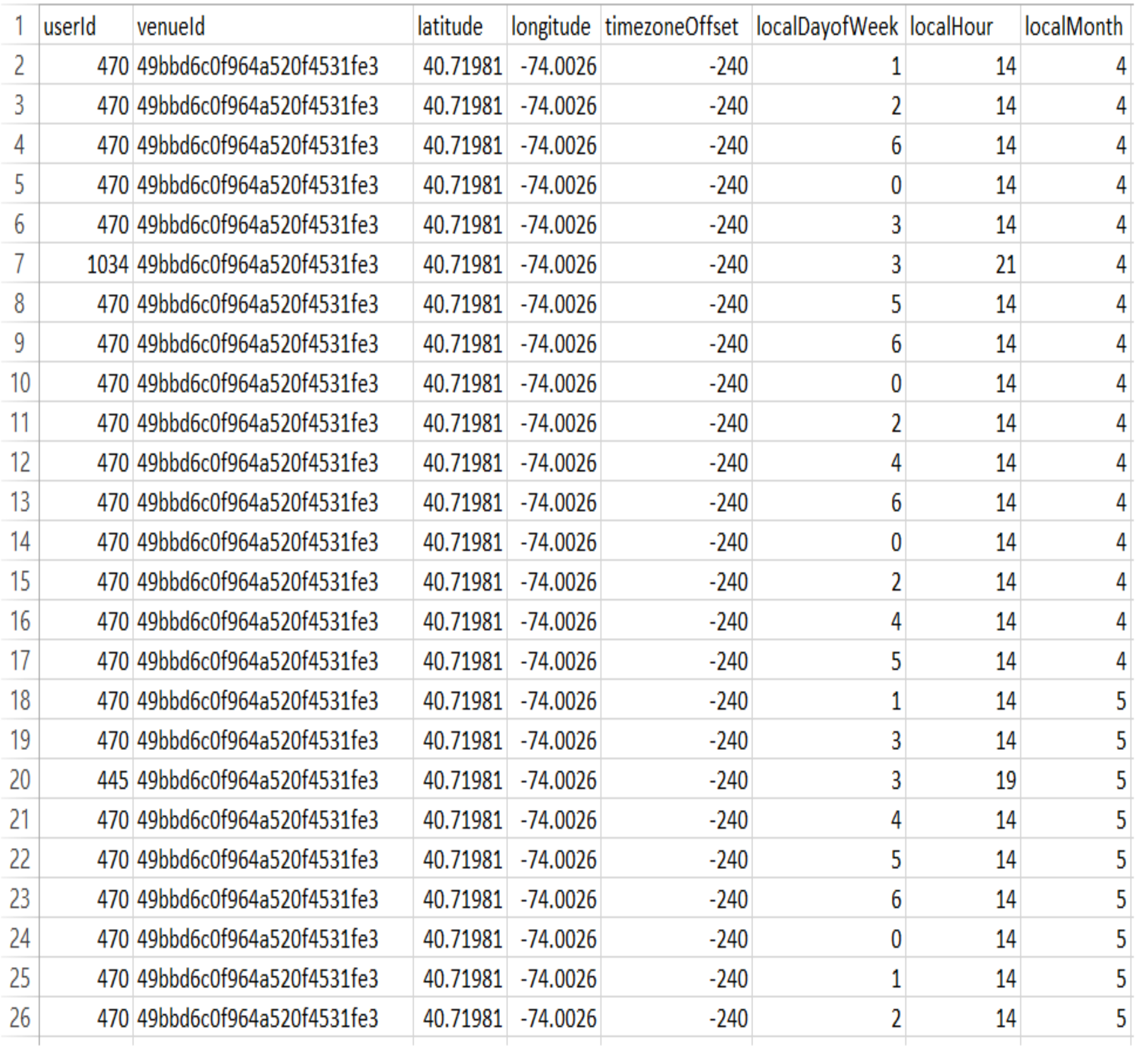
This chapter contains details about the dataset we have prepared to run the algorithms and the features we used for preparing. The original dataset we used is collected from a widely used social media called Foursquare which contained long term (10 months) check in data for New York City.  The available information were User ID, Venue ID, Latitude, Longitude, Venue Category and Time zone Offset of 227428 check-ins. Figure 4.1shows a screenshot of the first 25 entries of raw data. Each row represents information about a single check-in and the column represents the attributes of the check-in information. The user id denotes a unique id for every user of Foursquare, the venue id gives a unique id to the locations where check-in occurred, the category id and name states the type of the venue, each category has a unique name and id, although the names might be similar. The latitude and longitude gives the coordinate of the location of check-in and the UTC Timestamp and Time zone offset gives the time of visit. There were 1083 unique users, 38333 unique venues, 400 unique categories of venues and 91024 unique check-ins of the user that is only the check-in of one user in one location is considered. This way, any redundant check-ins that might cause noise can be ignored.

Table 4.1.2 shows the information of the below table, six features were implemented and the value of each feature was calculated. The result were all numerical values which was processed to form a new dataset that was used as the final dataset. Ten random candidate areas were selected, from where all the locations within a radius 400m of the area were considered for the research.  After refining the data and processing it into features, a total of 8673 data were found.

**Table 4.1.1 First 25 entries of Raw Data**



| userId | venueId | venueCategoryId | venueCategory | latitude | longitude | timezoneC | utcTimestamp |
|---|---|---|---|---|---|---|---|
| 470 | 49bbd6c0f964a520f4531fe3 | 4bf58dd8d48988d127951735 | Arts & Crafts Store | 40.71981 | -74.0026 | -240 | Tue Apr 03 18:00:09 +0000 2012 |
| 979 | 4a43c0aef964a520c6a61fe3 | 4bf58dd8d48988d1df941735 | Bridge | 40.6068 | -74.0442 | -240 | Tue Apr 03 18:00:25 +0000 2012 |
| 69 | 4c5cc7b485a1e21e00d35711 | 4bf58dd8d48988d103941735 | Home (private) | 40.71616 | -73.8831 | -240 | Tue Apr 03 18:02:24 +0000 2012 |
| 395 | 4bc7086715a7ef3bef9878da | 4bf58dd8d48988d104941735 | Medical Center | 40.74516 | -73.9825 | -240 | Tue Apr 03 18:02:41 +0000 2012 |
| 87 | 4cf2c5321d18a143951b5cec | 4bf58dd8d48988d1cb941735 | Food Truck | 40.7401 | -73.9897 | -240 | Tue Apr 03 18:03:00 +0000 2012 |
| 484 | 4b5b981bf964a520900929e3 | 4bf58dd8d48988d118951735 | Food & Drink Shop | 40.69043 | -73.9547 | -240 | Tue Apr 03 18:04:00 +0000 2012 |
| 642 | 4ab966c3f964a5203c7f20e3 | 4bf58dd8d48988d1e0931735 | Coffee Shop | 40.75159 | -73.9741 | -240 | Tue Apr 03 18:04:38 +0000 2012 |
| 292 | 4d0cc47f903d37041864bf55 | 4bf58dd8d48988d12b951735 | Bus Station | 40.77942 | -73.9553 | -240 | Tue Apr 03 18:04:42 +0000 2012 |
| 428 | 4ce1863bc4f6a35d8bd2db6c | 4bf58dd8d48988d103941735 | Home (private) | 40.61915 | -74.0359 | -240 | Tue Apr 03 18:06:18 +0000 2012 |
| 877 | 4be319b321d5a59352311811 | 4bf58dd8d48988d10a951735 | Bank | 40.61901 | -73.9904 | -240 | Tue Apr 03 18:06:19 +0000 2012 |
| 87 | 4d8263a73e916dcb8edd80d2 | 4bf58dd8d48988d155941735 | Gastropub | 40.74348 | -73.994 | -240 | Tue Apr 03 18:07:15 +0000 2012 |
| 625 | 4ab5320cf964a5202b7320e3 | 4bf58dd8d48988d122951735 | Electronics Store | 40.74261 | -73.9927 | -240 | Tue Apr 03 18:08:57 +0000 2012 |
| 691 | 4cb50d599c7ba35de0ef8706 | 4bf58dd8d48988d104941735 | Medical Center | 40.71976 | -74.25 | -240 | Tue Apr 03 18:09:06 +0000 2012 |
| 116 | 4c0ab56f7e3fc9288c1df482 | 4f04afc02fb6e1c99f3db0bc | Mobile Phone Shop | 40.74119 | -73.9897 | -240 | Tue Apr 03 18:09:29 +0000 2012 |
| 931 | 49f85763f964a520f16c1fe3 | 4bf58dd8d48988d16d941735 | CafÃ© | 40.70459 | -74.0096 | -240 | Tue Apr 03 18:09:59 +0000 2012 |
| 53 | 4c00297392a6c9280cc243e2 | 4bf58dd8d48988d124951735 | Automotive Shop | 40.86198 | -74.0479 | -240 | Tue Apr 03 18:10:04 +0000 2012 |
| 445 | 4b9830a5f964a520c73235e3 | 4bf58dd8d48988d118951735 | Food & Drink Shop | 40.82679 | -73.9495 | -240 | Tue Apr 03 18:10:39 +0000 2012 |
| 135 | 4a689777f964a520a0ca1fe3 | 4bf58dd8d48988d1c4941735 | Restaurant | 40.75527 | -73.9788 | -240 | Tue Apr 03 18:11:04 +0000 2012 |
| 895 | 4bb7e0c41261d13a8707e898 | 4bf58dd8d48988d1e0931735 | Coffee Shop | 40.90663 | -73.7778 | -240 | Tue Apr 03 18:11:16 +0000 2012 |
| 704 | 4cb4b31cb4b0a35df33969ce | 4bf58dd8d48988d14e941735 | American Restaurant | 40.73068 | -74.0657 | -240 | Tue Apr 03 18:11:35 +0000 2012 |
| 891 | 4ba8a9a9f964a5206ee539e3 | 4bf58dd8d48988d126941735 | Government Building | 40.88302 | -74.0759 | -240 | Tue Apr 03 18:12:09 +0000 2012 |
| 284 | 4abe60d8f964a520198d20e3 | 4bf58dd8d48988d1e0931735 | Coffee Shop | 40.74282 | -74.0004 | -240 | Tue Apr 03 18:12:27 +0000 2012 |
| 901 | 4c1ff4e68b3aa5933d499e5f | 4bf58dd8d48988d1f0931735 | Airport | 40.64532 | -73.7738 | -240 | Tue Apr 03 18:12:42 +0000 2012 |
| 803 | 4c955b3558d4b60c06d03129 | 4bf58dd8d48988d12d951735 | Ferry | 40.76067 | -74.0037 | -240 | Tue Apr 03 18:13:12 +0000 2012 |
| 699 | 4ba0e0edf964a520138437e3 | 4bf58dd8d48988d124941735 | Office | 40.73668 | -73.9889 | -240 | Tue Apr 03 18:13:21 +0000 2012 |



| 1 | userId | venueId | latitude | longitude | timezoneOffset | localDayofWeek | localHour | localMonth |
|---|---|---|---|---|---|---|---|---|
| 2 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 1 | 14 | 4 |
| 3 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 2 | 14 | 4 |
| 4 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 6 | 14 | 4 |
| 5 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 0 | 14 | 4 |
| 6 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 3 | 14 | 4 |
| 7 | 1034 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 3 | 21 | 4 |
| 8 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 5 | 14 | 4 |
| 9 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 6 | 14 | 4 |
| 10 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 0 | 14 | 4 |
| 11 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 2 | 14 | 4 |
| 12 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 4 | 14 | 4 |
| 13 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 6 | 14 | 4 |
| 14 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 0 | 14 | 4 |
| 15 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 2 | 14 | 4 |
| 16 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 4 | 14 | 4 |
| 17 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 5 | 14 | 4 |
| 18 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 1 | 14 | 5 |
| 19 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 3 | 14 | 5 |
| 20 | 445 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 3 | 19 | 5 |
| 21 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 4 | 14 | 5 |
| 22 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 5 | 14 | 5 |
| 23 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 6 | 14 | 5 |
| 24 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 0 | 14 | 5 |
| 25 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 1 | 14 | 5 |
| 26 | 470 | 49bbd6c0f964a520f4531fe3 | 40.71981 | -74.0026 | -240 | 2 | 14 | 5 |

**Table 4.1.2 shows the six features implemented**

## 4.1 Features

The six features that were used to construct the dataset are: check-in count, check-in consistency, density, openness, check-in identity and temporal signal. Each result of the feature will return a numeric value that will be proportional to the quality of each candidate location. In order to reduce noise and create a more valid dataset, two noise sensitive features were used, the check-in consistency and check-in identity. This was important since the dataset deals with user provided data and the information might not always be correct or valid. Most of the features calculated were spatial-temporal based since time and space are the two most important attributes of the research. The description of each of the features we considered are given below:

### 4.1.1 Check-in count

It shows the number of check-in of an area in a specific day at a specific time. The results of each area are shown as a three dimensional graph where hour, day and check-in are in x, y and z axis respectively. An example of such a 3D graph is shown below in figure 4.3. It is done for a particular area of the dataset, named Canal Subway Street, where check-in of different hour of day with respect local day of the week are shown. The data is grouped by time and user id, so that only the user check-in having a unique time at a unique location is considered, and it was then normalized considering a maximum and a minimum value and then calculating by the formula

$$normalization = \frac{x - min_x}{max_x - min_x} \qquad (1)$$

### 4.1.2 Check-in consistency

The check-in consistency is one of the noise sensitive features. It represents the measure of change of check-in pattern of an area. The less the change in check-in pattern, the more is the value of check-in consistency, hence indicating the stability of check-in of an area. This reduces the noise because it is giving low scores to the areas that are occasionally crowded at a particular day on a particular time instead of all year long. Therefore, the commercial enterprise will not face an unexpected audience response at their desired location. In order to calculate check-in consistency we measured the correlation between two check-in and also the overall check-in correlation. The consistency is then calculated by averaging these correlation values. For example at a particular day there can be a concert at an area, so this will be a noise issue, as in that place it is not crowded throughout the whole year but only for a day, so at that particular it will be inefficient to carry on a live advertising campaign. Also, there can be a feast or festival going at only a specific time of the year and the event does not occur every day on the same place so, it taking such places into account into the  preprocessing dataset  will actually cause an noise error. So, in order to eliminate such a noise issue such places cannot be taken into account into the dataset so it is very important to find a stability in the check-in pattern by com paring the weekly and monthly check-in count.

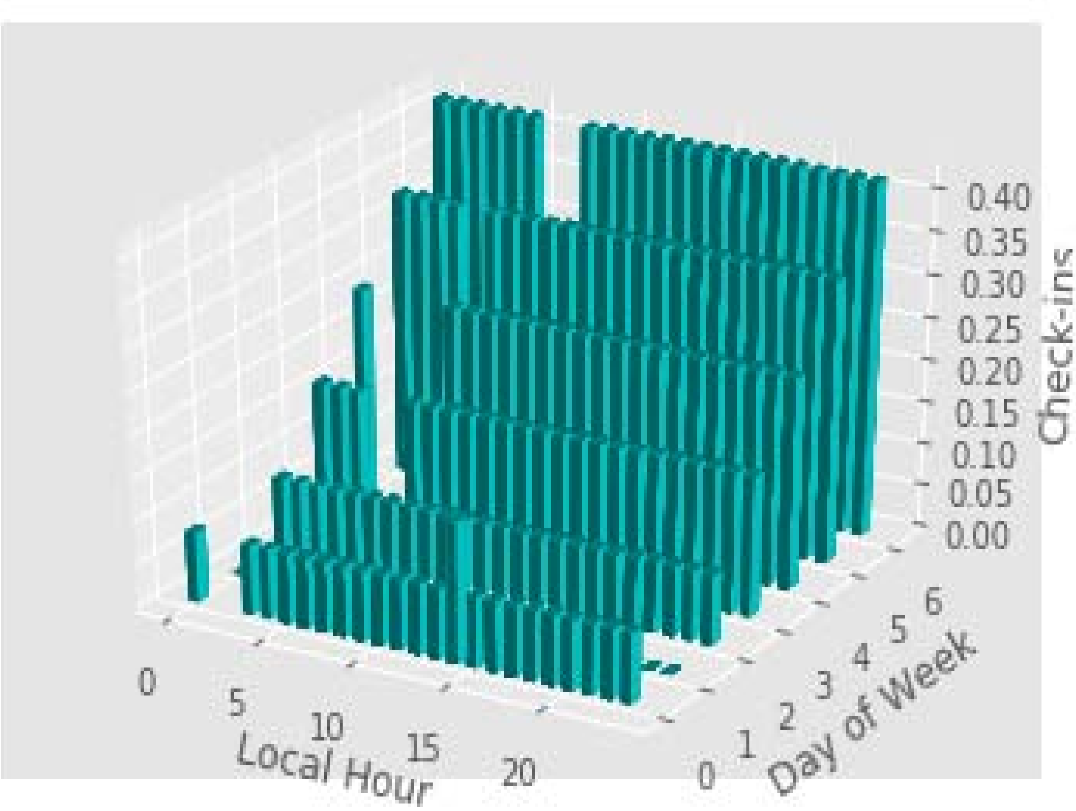


**Figure 4.1.1 shows the check-in count of one of the candidate areas – Canal Subway Street**

### 4.1.3 Density

It denotes the number of location we have in a certain area. We have considered a radius of 400m for each area. So the density will give us the number of venues for each area within the 400m radius. We considered one of the location to be the center of area and calculated distance of all the neighboring locations, and the locations that were within 400m were counted to find the total density of the area.

### 4.1.4 Openness

Openness signifies the amount of free space available in a location. The more open the space is, the easier it is to arrange a campaign and gather people due to all the space. Openness is considered to be the inverse of the value of density because the less dense area will be the more open space will it have. Therefore, [1]

$$openness = \frac{1}{density} \tag{2}$$

### 4.1.5 Check-in Identity

Check-in identity considers only the unique check-ins of an area and therefore eliminates multiple check-ins of the same person in the same area. This reduces the noise since an individual's regular check-in at the same place may increase the popularity of that area and hence give an incorrect result. Therefore, it is our second noise sensitive feature. We calculated the check-in identity by the formula [1]:

$$Identity = \frac{Numberofunique-che\check{c}k\imath ns}{Numberoftotal-che\check{c}k\imath ns} \tag{3}$$

### 4.1.6 Temporal signal

The temporal signal measures the particular time of the day when the people are willing to go to a place. This is an important feature because it makes our result even more accurate by determining the exact location to choose for different time of the day. We divided 24 hours of the day into four slots, morning, noon, afternoon and evening. We considered the morning time to be from 12 am – 11.59 am, the noon at exactly 12 pm, afternoon from 12.01pm – 5.59 pm and evening from 6pm to 11.59 pm.Then after categorizing each check-in according to the time slots, we calculated the temporal signal by the following formula [1] below where we have used this function as a feature, because a specific location is not busy round a clock but at a specific period of time. For example, coffee shops normally gets busy during the afternoon, so if will be more efficient to advertise during the evening than any other time as such places stay crowded during that period of time. Also, during weekends, a park will have a higher tendency to be crowded that on any week day so it is highly important to consider the check-in pattern analysis of a particular period of time. Hence, such a factor plays an important role in finding the optimal location for live campaign.

$$\text{Temporal Signal} = \sum_{i\in m,n,a,e} \frac{number\ of\ unique\ check\ -\check{\imath}ns}{total\ number\ o\check{f}\ check-\imath ns} \qquad (4)$$

A graph was plotted that showed the number of check-ins of top 6 categories at a specific hour of the day shown in figure 4.2.6.

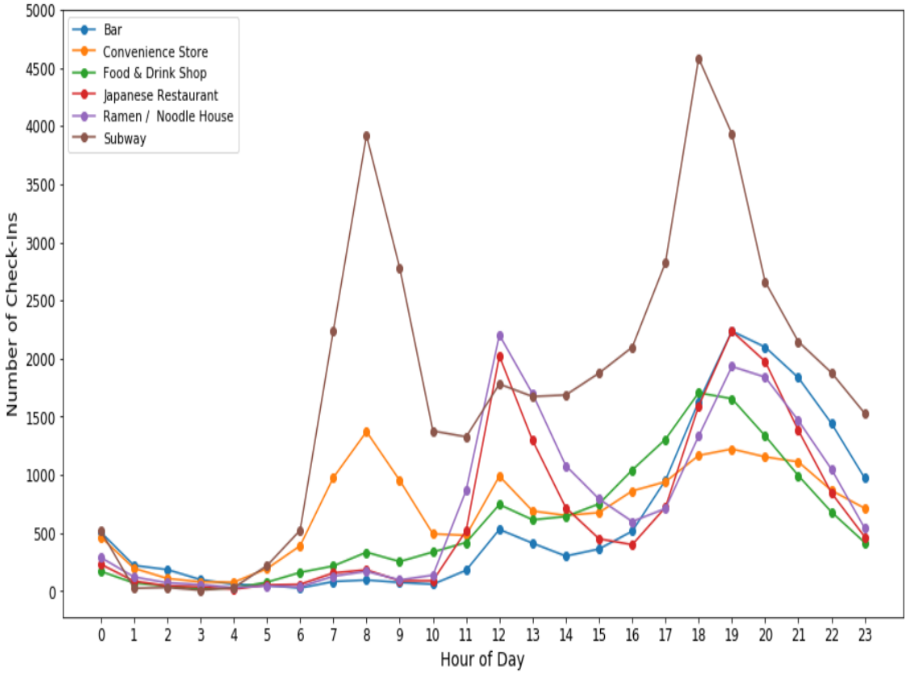
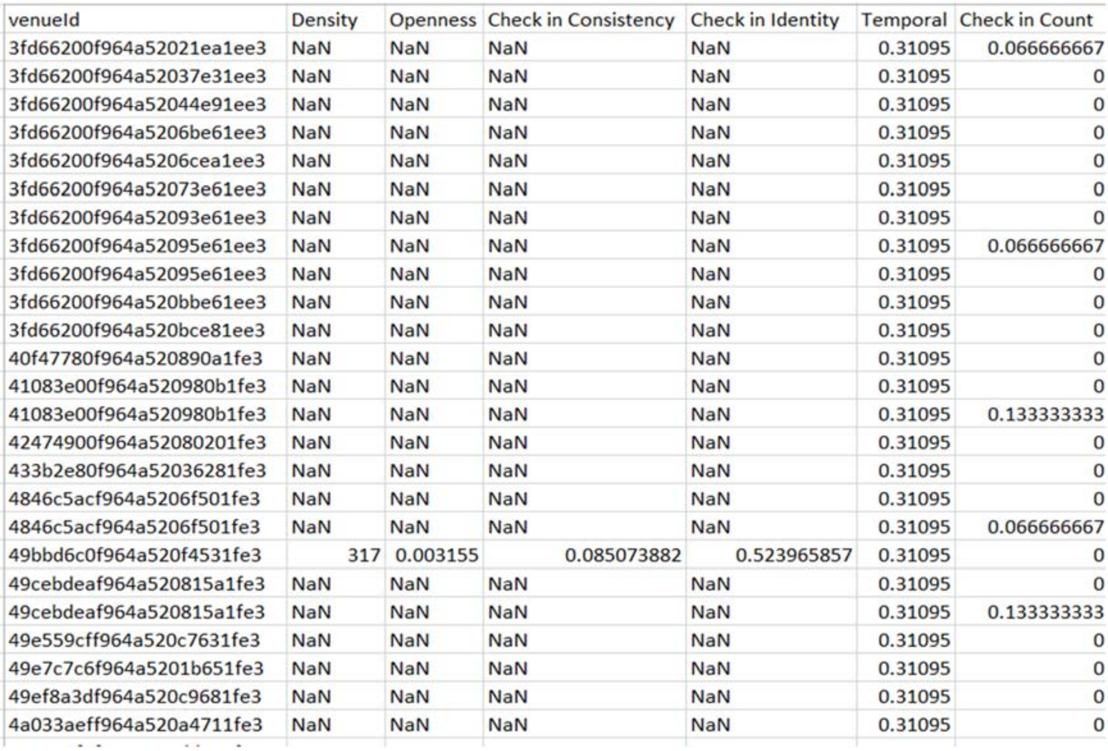


**Figure 4.1.6 Number of check-ins of top 6 categories at a specific hour of the day**

## 4.2 Data Arrangements

The output of our research is the most suitable location for a live campaign. So, location ID is the first column of our dataset. The features were calculated and also put as columns of our dataset. The rows of our dataset are the unique check-ins of each user at a specific time. Therefore, the numbers of columns were 7 and the numbers of rows were 8673 for 10 candidate areas.

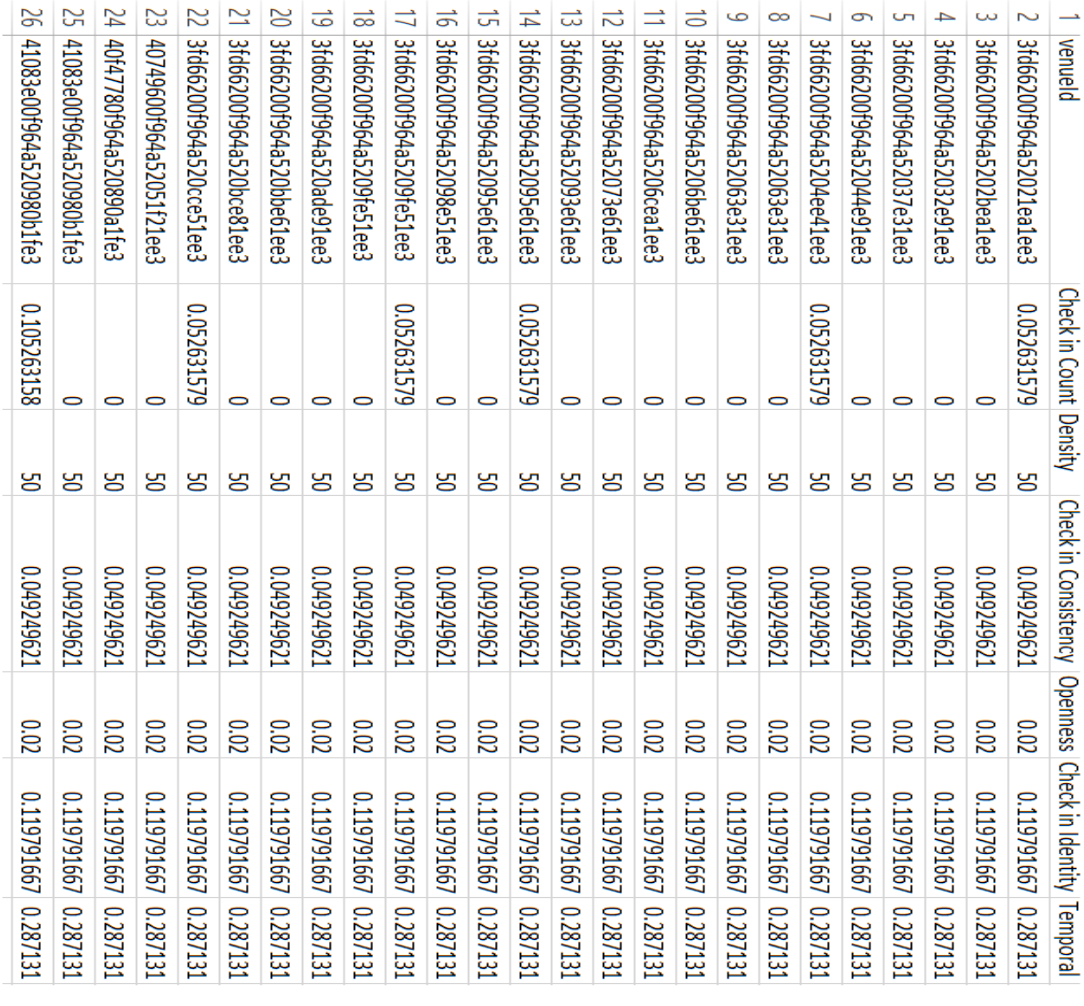
While calculating the features like density, openness, check-in identity and check-in consistency we found some of the values to be null. Table 4.2.1 represents the dataset of values that contained invalid data.



| venueId | Density | Openness | Check in Consistency | Check in Identity | Temporal | Check in Count |
|---|---|---|---|---|---|---|
| 3fd66200f964a52021ea1ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0.066666667 |
| 3fd66200f964a52037e31ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 3fd66200f964a52044e91ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 3fd66200f964a5206be61ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 3fd66200f964a5206cea1ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 3fd66200f964a52073e61ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 3fd66200f964a52093e61ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 3fd66200f964a52095e61ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0.066666667 |
| 3fd66200f964a52095e61ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 3fd66200f964a520bbe61ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 3fd66200f964a520bce81ee3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 40f47780f964a520890a1fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 41083e00f964a520980b1fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 41083e00f964a520980b1fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0.133333333 |
| 42474900f964a52080201fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 433b2e80f964a52036281fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 4846c5acf964a5206f501fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 4846c5acf964a5206f501fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0.066666667 |
| 49bbd6c0f964a520f4531fe3 | 317 | 0.003155 | 0.085073882 | 0.523965857 | 0.31095 | 0 |
| 49cebdeaf964a520815a1fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 49cebdeaf964a520815a1fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0.133333333 |
| 49e559cff964a520c7631fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 49e7c7c6f964a5201b651fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 49ef8a3df964a520c9681fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |
| 4a033aeff964a520a4711fe3 | NaN | NaN | NaN | NaN | 0.31095 | 0 |

**Table 4.2.1: Dataset of features with invalid (NaN) values**

For check-in identity and check-in consistency, we put the minimum value of the column as our base value for each of the null boxes and for density we assumed the numbers of neighboring locations in an area have the same density, and therefore the value of openness changed accordingly, since it is the inverse of the density value. This was done to validate the dataset and get a more accurate result. Table 4.2.2 shows the final dataset we used after refining our data.



| 1 | venueId | Check in Count | Density | Check in Consistency | Openness | Check in Identity | Temporal |
|---|---|---|---|---|---|---|---|
| 2 | 3fd66200f964a52021ea1ee3 | 0.052631579 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 3 | 3fd66200f964a5202bea1ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 4 | 3fd66200f964a52032e91ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 5 | 3fd66200f964a52037e31ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 6 | 3fd66200f964a52044e91ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 7 | 3fd66200f964a5204ee41ee3 | 0.052631579 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 8 | 3fd66200f964a52063e31ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 9 | 3fd66200f964a52063e31ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 10 | 3fd66200f964a5206be61ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 11 | 3fd66200f964a5206cea1ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 12 | 3fd66200f964a52073e61ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 13 | 3fd66200f964a52093e61ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 14 | 3fd66200f964a52095e61ee3 | 0.052631579 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 15 | 3fd66200f964a52095e61ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 16 | 3fd66200f964a52098e51ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 17 | 3fd66200f964a5209fe51ee3 | 0.052631579 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 18 | 3fd66200f964a5209fe51ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 19 | 3fd66200f964a520ade91ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 20 | 3fd66200f964a520bbe61ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 21 | 3fd66200f964a520bce81ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 22 | 3fd66200f964a520cce51ee3 | 0.052631579 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 23 | 40749600f964a52051f21ee3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 24 | 40f47780f964a520890a1fe3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 25 | 41083e00f964a520980b1fe3 | 0 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |
| 26 | 41083e00f964a520980b1fe3 | 0.105263158 | 50 | 0.049249621 | 0.02 | 0.119791667 | 0.287131 |

**Table 4.2.2 First Twenty Five Entries of the final dataset**

# CHAPTER 5
# Algorithms

In this chapter, we will talk about the algorithms we used for our research purpose. We used Support Vector Machine (SVM), Decision tree and Random Forest algorithms and compared the accuracy of results of each algorithm. From the Foursquare dataset, the features for X were- We used the cross validation method to divide the data into two groups, test and train, to implement the algorithms where we tried different test sets (5%, 10%, 15%, 20%, 25% & 30%) and the one resulting in the highest accuracy was considered.

Our research topic is a classification predictive model and we have used supervised learning approach. Due to this, we used some of the supervised learning algorithm like SVM, random forest and decision tree to calculate the results. The input of the algorithms are all the columns of our final dataset and the output is the location id, which needed to be encoded to numeric value for it was represented as a string in the dataset and the algorithms require a numeric value to calculate.

The detailed description of the algorithms used are given below:

## 5.1 Support Vector Machine(SVM)

Support Vector Machine are supervised learning algorithm in machine learning which generates a hyperplane or a set of hyperplanes, where a hyperplane denotes a geometrical subspace in which a dimension is one less than that of its ambient space. SVM constructs such a hyperplane or a collection of hyperplanes in a high or infinite dimensional space which is mainly used for classification and regression. When the nearest training data set has the largest distance from the hyperplane, generally a good separation can be achieved.

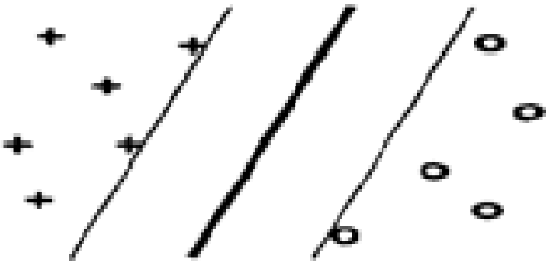


**Fig 5.1.1 shows how a hyper plane is of SVM [23-24]**

Generally, in most scenarios the problem lies in a finite dimensional space, as a result the sets are not differentiated in a linearly separable in their respective space, so, in order to make the separation more easier the original finite-dimensional space is mapped into a much higher-dimensional space, the dot products are generated in terms of the variables in the original space [27], by initializing the kernel function k(x,y). In the high dimensional space, the hyperplanes are defined as the collection of points whose dot product with the vector in the particular space is constant. The vectors defining the hyperplanes can be transformed into linear combination with parameters $\alpha(i)$of feature vectors of xi of the database where the points x of the feature space are defined with following relation-

$$\sum_i \alpha(i)k(x(i),x) = constant \qquad (5)$$

Hence, in the scenario of this research topic, it fall into a classification problem. As we had multiple features of x so our problem was defined by the kernel function to transform the non-linear SVM into a linear SVM. Linear kernels were also used because it deals with the outliers of the dataset more efficiently. SVM linear is used as with the help of Scikit Learn, we were able to use classifier and different function to run the algorithms.

### 5.2 Decision Tree

On the other hand, the other supervised learning algorithm used was Decision Tree learning, based on individual input variables, it predicts the value of the target variable. A decision tree is such that it consists of branches where each interior nodes points to children for each of the possible input variables, the leaf of the tree represents the value of the target variables where the values of the input variables is denoted by the route from the root to the leaf. Source set is splitted into subsets which is based on the test result of the attribute value. In a recursive manner, the process is repeated on each of the derived subset which is also known as recursive partitioning, that constructs a decision tree that correctly classifies members of the population by dividing it into subpopulations based on several dichotomous independent variables [28].When the splitting variable no longer is adding value to the predictions or when the subset nodes all have the same target variable value, at that time the recursive process is completed. The following metrics are used. For this algorithm we used the decision tree classifier from Scikit learn and other related functions.

Metrics generally measure the similarity of the target variables with the subsets. The resulting values are combined after the metrics are applied to each of the subsets in order to measure the quality of the split. The following metrics are-

Gini impurity- measures how often does a randomly selected element from the set is incorrectly labeled, if it was labeled randomly based on the distributions of the subset labels. In order to calculate Gini impurity for a set of items of J classes that belongs to such that i BELONGS TO {1,2,....J} and let ***pi*** be the fraction of items labeled with the class ***i*** in the sets as follows:

$$Ig(p) = \sum_{i=1}^{J} pi \sum_{k\neq 1} pk = \sum_{i=1}^{J} pi(1-pi) = \sum_{i=1}^{J}(pi - pi^2) = \sum_{i=1}^{J} pi - \sum_{i=1}^{j} pi^2 = 1 - \sum_{i=1}^{J} pi^2 \qquad (6)$$

This algorithm can handle both categorical and numerical data efficiently [28]. As the Preprocessed Foursquare Dataset have both the numerical and categorical data, where the number of check-in is numerical but the temporal feature which is the highest number of check-in a categorical venue id at a particular time is a categorical data. Hence, Decision Tree was implemented.

### 5.3 Random Forest

Random Forest also known as random decision forest are an ensemble learning method in which multiple learning algorithms are used to attain better predictive performance, it operates by generating a multitude of decision trees during the training time and evaluating the classification of the individual trees. [25-26].The random forest applies the general method of bootstrap aggregating which is basically the bagging of the tree learners. For a training set of X = x1,………., xn which corresponds to Y=y1,.........,yn, where the bagging is repeated B times which chooses a random sample and replaces with previous chosen sample for b = 1 ,.........,B,so, after the training process is completed, the prediction of the unseen sample is calculated by taking the maximum number of votes in the case of classification as shown:

$$f = \frac{1}{B}\sum_{b=1}^{B} fb(x') \qquad (7)$$

Random forest is also based on Classification and Regression Tree. The prediction was made averaging the result of 100 decision trees. However, for our dataset the result of random forest was less than that of the other two algorithms.

# CHAPTER 6

# Results and Discussion

In this chapter, we will discuss and compare between the accuracy results of all the algorithms using the same dataset. As mentioned in earlier chapters, three algorithms were used to find the ranking and accuracy of the algorithms which are Support Vector Machine, Decision Tree and Random Forest. We used different sizes of train and test datasets for e.g. 5%-95%, 20%-80%, 30%-70% etc. The input of all the algorithms are the features found as well as the venue id, which needed to be encoded to integer value. The output of our model is the unique venue ID and its respective longitude and latitude values which are the target attribute in our case.

## 6.1 Result Ranking

The ranking of the locations were calculated based on the prediction frequency of the matched results of the test and train datasets. We have extracted the matched results of test and train datasets into a new data frame to calculate the frequency of the occurrence of unique venue ID. After that, we have normalized the score to rescale the range of the score since the range of our score for ranking can vary. For the test size of 5%, the accuracies of each of the algorithm were highest hence we showed the output of our results of the ranking of respective locations for each of the algorithm with the test size of 5%. The following tables below 6.1.1, 6.1.2, 6.1.3 shows the ranking of the venue id and its respective latitude and longitude from most suitable to least, according to SVM, Decision tree and Random Forest respectively. As, our target attribute was the venue ID and its respective longitude and latitude locations, so we matched the predicted result with the tested result of the features with respect to  unique venue ID, then we extracted  the result of the matched and predicted result into another data frame and extracting the matched results, we found the mostly occurred result of the unique venue ID, and normalized the value to bring it to rescale the range and we then listed the mostly occurred result in a descending and found its locations by the respective latitude and longitude and its location name through the aid of Google Map.

| Venue ID | latitude | longitude |
|---|---|---|
| 4ae8906ef964a52071b021e3 | 40.71267093 | -74.01193142 |
| 4b40cf3df964a5206bbb25e3 | 40.71834203 | -74.00061817 |
| 4ae8906ef964a52071b021e3 | 40.71267093 | -74.01193142 |
| 4a954328f964a520cf2220e3 | 40.70972827 | -74.01076978 |
| 4b02c41cf964a520c14922e3 | 40.71178449 | -74.00208235 |

**Table: 6.1.1 ranking of venue ID for SVM**

| Venue ID | latitude | longitude |
|---|---|---|
| 4ae8906ef964a52071b021e3 | 40.71267093 | -74.01193142 |
| 4b40cf3df964a5206bbb25e3 | 40.71834203 | -74.00061817 |
| 4b02c41cf964a520c14922e3 | 40.71178449 | -74.00208235 |
| 4b40cf3df964a5206bbb25e3 | 40.71838403 | -74.00052317 |
| 4b02c41cf964a520c14922e3 | 40.71178449 | -74.00208235 |

**Table: 6.1.2 ranking of venue ID for Decision Tree**

| Venue ID | latitude | longitude |
|---|---|---|
| 4ae8906ef964a52071b021e3 | 40.71267093 | -74.01193142 |
| 4b40cf3df964a5206bbb25e3 | 40.71834203 | -74.00061817 |
| 4a954328f964a520cf2220e3 | 40.70972827 | -74.01076978 |
| 4ae8906ef964a52071b021e3 | 40.76467093 | -74.00193245 |
| 4b40cf3df964a5206bbb25e3 | 40.71834203 | -74.00061527 |

**Table: 6.1.3 ranking of venue ID for Random Forest**

The figure 6.1.1 shows the graphical representation of the output of the ranking of the locations with respect to score, where we have implemented it with SVM, where the x axis shows the ranking of the unique venue IDS and the y axis shows the score of the top 5 locations that will be recommended to the user. In this graph we can see that Greenwich St has the highest score which is close to 1 so we can deduce that all the features were favorable for selecting Greenwich St as the optimal location for live Campaigns.

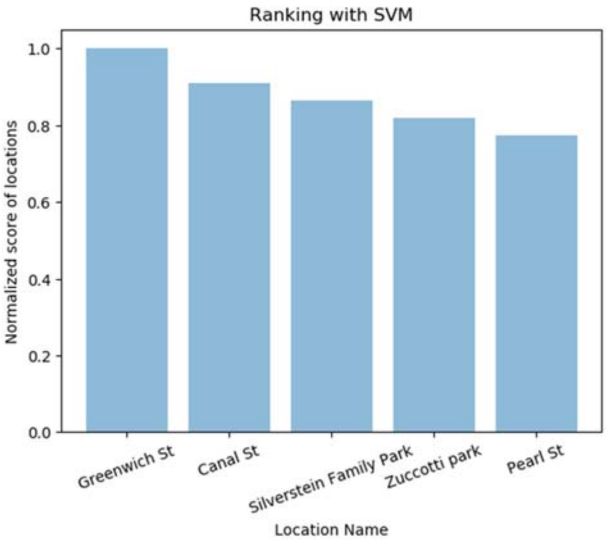


**Figure 6.1.1 shows the ranking with the implementation of SVM**

The figure 6.1.2 shows the graphical representation of the output of the ranking of the locations with respect to score, where we have implemented it with Decision Tree Algorithm, in this graph we can see that the 1st and 2nd place recommended by Decision Tree matches with that predicted  result by SVM and also the 5th place matches as well, except for the 3rd and the 4th place where the place recommended at the 1st place is also Greenwich St hence it shows that Greenwich St is very much popular and favorable for live campaign features such as the check-in count, density has favored for Greenwich St.. Noise feature has proved to be eliminated for this location as check-in is consistent and unique check-in identity feature. Hence, we can conclude that the performance level of both the SVM and Decision tress were very similar to each other.

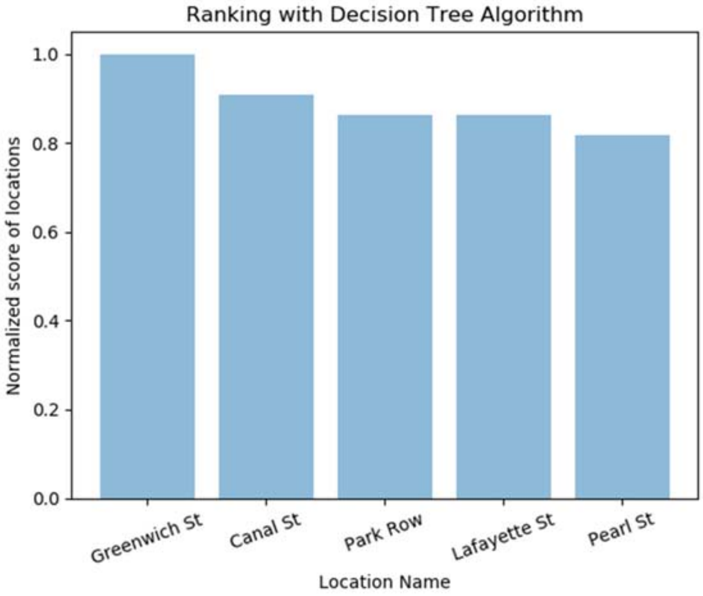


**Figure 6.1.2 shows the ranking with the implementation of Decision Tree**

Figure6.1.3 shows the graphical representation of the output of the ranking of the locations with respect to score, where we have implemented it with Random Forest Algorithm. Among the candidate locations and its neighbors, Canal St is most popular and favorable for live advertising. However, in the implementation of all three algorithms the 3rd and the 4th location names varies but again the last location name has matched by using All Three algorithms. In the Figure 6.1.3 the score for 2nd and 3rd ranking is lower than the ranking implemented with SVM and Decision Tree. So we can come to conclusion that the performance, level of random forest is lower than other two algorithm. Also, with the help of Google Map we have evaluated the location name which corresponded to the respective latitude and longitude.

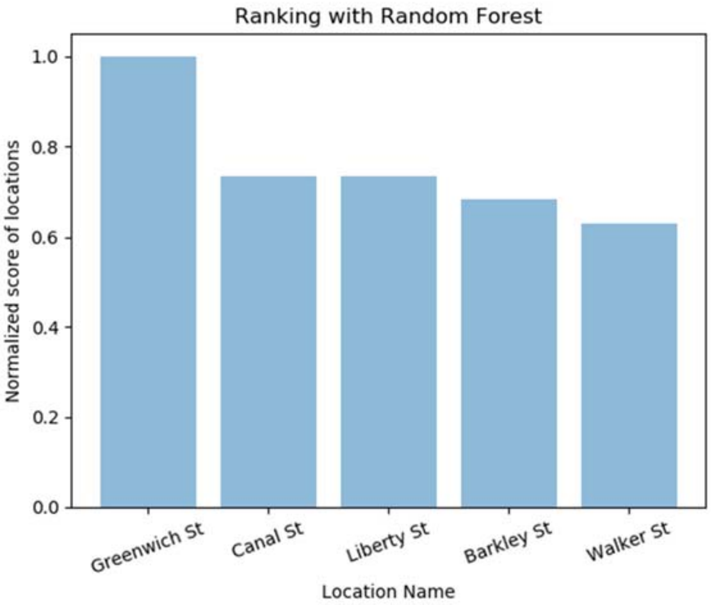


**Figure 6.1.3 shows the ranking with the implementation of Random Forest**

## 6.2 Result Accuracy

The accuracy of the three algorithms with respect to different test and train sets were calculated in order to find the best accuracy value. Below is the result of the graphs we plotted with the different values we found for accuracy. Due to the similarity of accuracy values of SVM and Decision tree, the graphs looked ambiguous merged in one, so we had to plot three different graphs for the three algorithms.

The figure 6.2.1 below shows the accuracy graph of the support vector machine algorithm with respect to various test sizes. We have taken test sets of size 5, 10, 15, 20, 25 and 30. As we can see from the graph, the accuracy of SVM is decreasing with increasing test size. For example when the test set size was 5, the accuracy was 88.5%, whereas when test size was 15, accuracy was 84.9% and when it was 30, the accuracy was 83.9%. It dropped by a large amount from test size 5 to 20 and then increased a little at 25 and then decreased slightly again at test size 30.

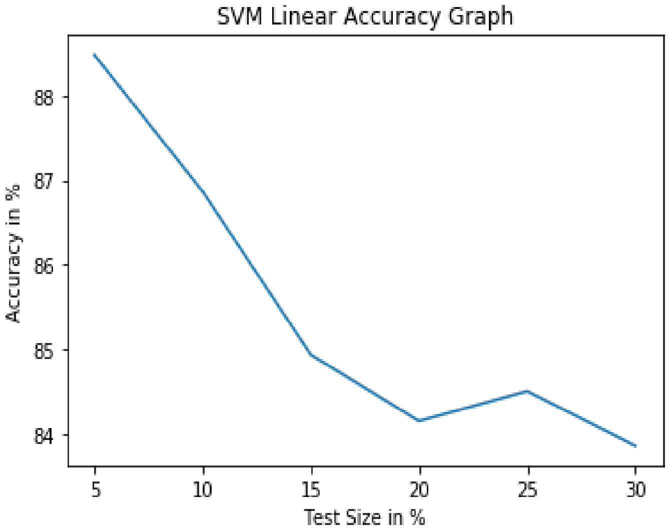


**Figure 6.2.1 SVM Linear Accuracy Graph**

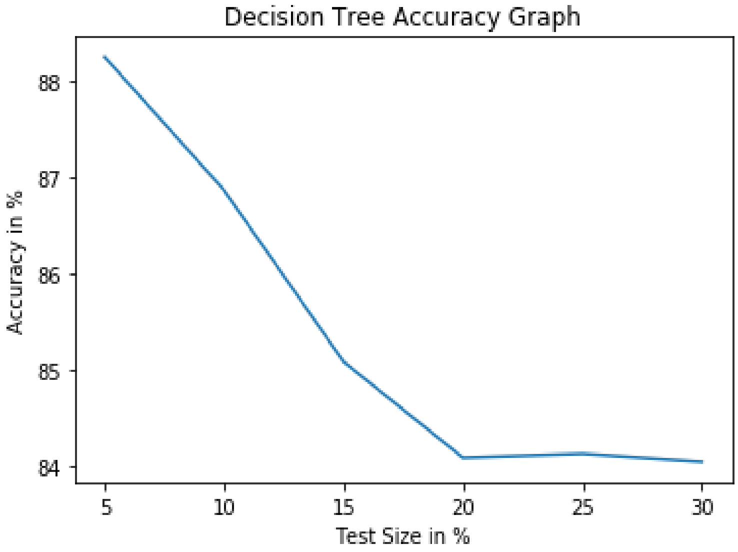


**Figure 6.2.2 Decision Trees Accuracy Graph**

The above figure 6.2.2 shows the accuracy graph of decision tree algorithm with respect to test size. Like SVM, we have taken test size of 5, 10, 15, 20, 25 and 30 for calculating the accuracy. The accuracy for decision tree algorithm also decreases with increasing test size. And their rate of change is quite the same. It drops significantly from test size 5 to 20. The accuracy at test size 5 was 88.2% from which it dropped to 84.1% at test size. Then it kept dropping till 30 unlike SVM which had a slightly larger value at 25.

The below figure 6.2.3 below shows the accuracy graph of Random Forest Algorithm with respect to various test sizes. The random forest has comparatively has smaller accuracy values than both SVM and decision tree but the pattern of accuracy is the same. It decreases drastically from test size 5, when it was 70.0% to 65.1% at test size 20. Then it changed slightly at 25 and then dropped to 63.8% again at 30. The test sizes takes here were also 5, 10, 15, 20, 25 & 30.

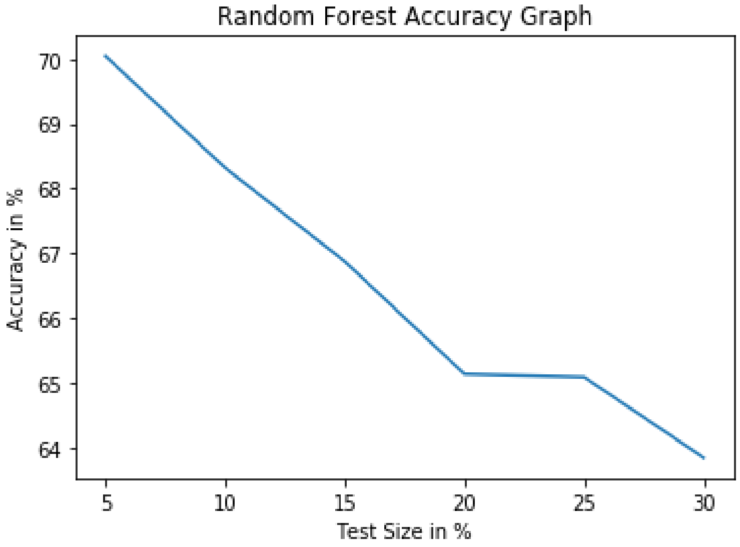


**Figure 6.2.3 Random Forest Accuracy graph**

The node sizes of the trees were also altered to find the accuracy change with respect to node size. Figure 6.2.4 below shows the accuracy graph of random forest algorithm having different node size.A test size of 5% was used for this graph. As we can see, the accuracy of graph increases with respect to the node size up to 50 from 68.7% to 70.0%. After even if the number of nodes are increased the accuracy reaches a constant level of 70.25% a approximately. After that, it reaches a threshold accuracy and remains constant no matter how much the node size is increased.

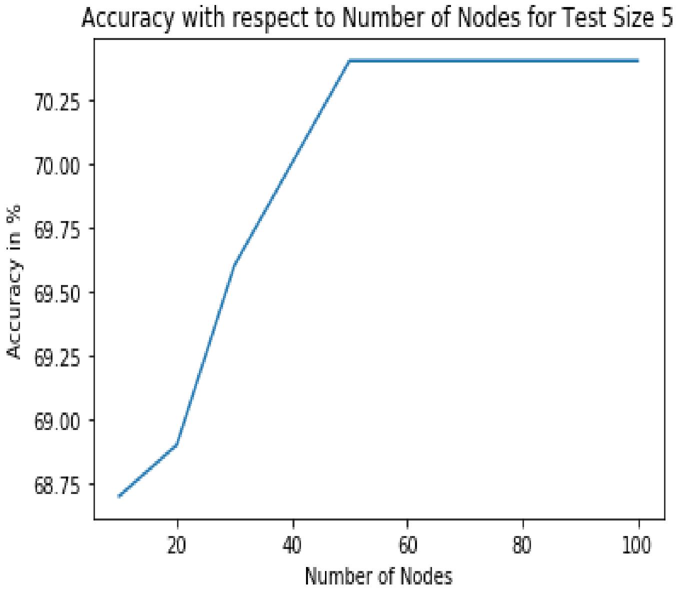


**Figure 6.2.4 Accuracy for different Node Size**

Therefore, from the above graphs, we can deduce that all the accuracies are decreasing with increasing test size. The reason behind the fact that that accuracy is decreasing is the quantity of our data were less which is why the test size with the least value had a higher accuracy, and hence accuracy kept decreasing as test size was increasing.

From the outcomes we can deduce that for our case decision tree has performed better than the other two algorithms e.g. SVM (linear) and Random Forest though the results of SVM (Linear) were closer to the results of Decision Tree, whereas, random forest algorithm outcomes were less than the other two algorithms. The reason behind smaller values of random forest algorithm accuracy is, some of the data we used were categorical rather than numerical like the time of check-in which had day of week and month. So we needed to convert the data into numerical value like we substituted 1 for values of Sunday and 3 for March. So some of the numerical values might not be accurate. Random forest works best with all numerical data. Since we had some categorical values converted to numerical value, the random forest could not perform well with this specific type of dataset, even though it is considered to be more accurate than decision tree.

# Chapter 7

## Conclusion and Future Work

The digital marketing has become extremely competitive in recent years. Companies are looking for ways to improve their strategy to reach potential customers. Consequently using location to find suitable location for live campaigns. Previous check-in from social networks have the potential to predict a crowds next check-in and analyzing the data will give a good site for advertising to the right customers.

This paper talk about finding suitable places for live campaigns, applying different machine learning, in the light of location based network sites using dataset from Foursquare. Some static geographical features are included which can change the movement pattern of people. With Random forest and Decision Tree Algorithm Greenwich Street has the highest score with Barkley Street and Pearl Street the lowest. We achieved 88.25% accuracy in Decision Tree regression model and an accuracy of 88.48% and 70.04% in Support Vector Machine (SVM) and Random Forest respectively.

From the outcomes, we can conclude that decision tree has performed better than the other two algorithms e.g. SVM (linear) and Random Forest though the results of SVM (Linear) were closer to the results of Decision Tree. However, random forest algorithm outcomes were less than the other two algorithms because a part of our dataset had categorical rather than numerical value which has caused some error in the accuracy of random forest.

In future we plan to gather data from mobile phone companies of Bangladesh to get user location with their gender, and then it would be possible to show different types of advertisements in different locations according to user needs. Any product would be promoted to its highest level if advertised to the right customers; this would be really beneficial for business marketing**.**

## Appendix

### Conversion of Geographic Coordinates

The geographic coordinates in the Foursquare dataset are given as longitudes and latitudes. For the predictive model to gauge the distance between two points more accurately, they will need to be converted into 3-dimensional Cartesian coordinates. In this report, we assume Earth to be fully spherical and the conversion formulae are as follows (Wolfram Research, 2017):

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

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

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

R is the earth's radius 6371 km

### Multiclass Classifiers are evaluated:

**Table 6 the types of multiclass classifiers used and their parameters.**

| Item | Description |
|---|---|
| **Decision Tree** | ● The hyper-parameter (minimum number of samples at leaf nodes) is tuned using k-fold cross validation |
| **Random Forest** | ● Average predictions over hundred decision trees |
| **SVM** | ● The hyper-parameter is tuned using k-fold cross validation |

## References

[1] Rahman, M. and Nayeem, M. (2017). Finding suitable places for live campaigns using location-based services Proceedings of the Fourth International ACM Workshop on Managing and Mining Enriched Geo-Spatial Data Article No. 7 Chicago, Illinois, May, 2017

[2] Z. Cheng, J. Caverlee, and K. Lee: Exploring millions of footprints in location sharing services. In Proceedings of the 5th International Conference on Weblogs and Social Media, ICWSM, Barcelona, Catalonia, Spain, July, 2011.

[3] Backstrom, L.; Sun, E.; and Marlow, C. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In WWW '10.

[4] E. J. Zanot:Public Attitudes Toward Advertising: The American Experience. International Journal of Advertising, vol. 3, pages 3-15, 1984.

[5] P. Georgiev, A. Noulas, C. Mascolo: The Call of the Crowd: Event Participation in Location-Based Social Services.In Proceedings of the 8thInternational AAAI Conference on Web logs and Social Media (ICWSM2014).Ann Arbour ,Michigan, USA,June2014.

[6] M. Sklar, B. Shaw, and A. Hogue: Recommending interesting events in real-time withfoursquare check-ins. In Proceedings of the 6th ACM Conference on RecommenderSystems, RecSys, Dublin, Ireland, September, 2012.

[7] S. Porta, V. Latora, F. Wang, S. Rueda, E. Strano, S. Scellato, A. Cardillo, E. Belli, F. CÃărdenas, B. Cormenzana and L. Latora: Street centrality and location of economic activities in barcelona. Urban Studies, vol. 49, no. 7, pages 1471-1488, November, 2011

[8] S. Porta, V. Latora, F. Wang, E. Strano, A. Cardillo, S. Scellato, V. Iacoviello and R. Messora: Street centrality and densities of retail and services in bologna, italy.Environment and Planning B: Planning and Design, vol. 36, no. 3, pages 450âĂŞ465, December, 2009.

[9] D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia and C. Mascolo: Geo-Spotting: Mining Online Location-based Services for Optimal Retail Store Placement. In Proceedings of the 19th ACM SIGKDD International Conference

on Knowledge Discovery and Data Mining, KDD, Chicago, IL, USA, August, 2013

[10] C. Song, Z. Qu, N. Blumm and A.-L. Barabasi: Limits of predictability in human mobility. Science, vol. 327, issue 5968, pages 1018âĂŞ1021, February, 2010

[11] D. Yang, D. Zhang, V. W. Zheng, Z. Yu: Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. IEEE Trans. on Systems, Man, and Cybernetics: Systems, (TSMC), 45(1), 129-142, 2015.

[12] D. Quercia, N. Lathia, F. Calabrese, G. D. Lorenzo and J. Crowcroft: Recommending social events from mobile phone location data. In Proceedings of the 10th IEEE International Conference on Data Mining, Sydney, Australia, December 2010.

[13] D. H. Lee: Pittcult: trust-based cultural event recommender. In Proceedings of the 2nd ACM Conference on Recommender Systems, RecSys, Lausanne, Switzerland, October, 2008.

[14] E. Minkov, B. Charrow, J. Ledlie, S. Teller and T. Jaakkola: Collaborative future event recommendation. In Proceedings of the 19th ACM Conference on Information and Knowledge Management, CIKM, Toronto, Ontario, Canada, October, 2010.

[15] G. Liao, Y. Zhao, S. Xie and P. S. Yu: Latent networks fusion based model for event recommendation in online ephemeral social networks. In Proceedings of the 22nd ACM International Conference on Information and Knowledge Management, CIKM, San Francisco, CA, USA, October, 2013.

[16] Pablo Jensen: Network-based predictions of retail store commercial categories and optimal locations. Physical Review E, 74:035101, September, 2006.

[17] Pablo Jensen: Analyzing the localization of retail stores with complex systems tools. In Proceedings of the 8th International Symposium on Intelligent

Data Analysis: Advances in Intelligent Data Analysis VIII, IDA, pages 10âĂŞ20, Berlin, Heidelberg, 2009.

[18] Z. Cheng, J. Caverlee, K. Kamath and K. Lee: Toward Tra_c-Driven Location-Based Web Search. In Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM, Glasgow, United Kingdom, October, 2011.

[19] M. Ye, D. Shou, W. C. Lee, P. Yin and K. Janowicz: On the semantic annotation of places in location-based social networks. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August, 2011.

[20] Y. Liang, J. Caverlee, Z. Cheng and K. Y. Kamath: How Big is the Crowd? Event and Location Based Population Modeling in Social Media. In Proceedings of the 24th ACM Conference on Hypertext and Social Media (part of ECRC), HT, Paris, France, May, 2013.

[21] S. Hasan and S. Ukkusuri: Urban activity pattern classification using topic models from online geo-location data. Transportation Research Part C, vol. 44, pages 363-381, April, 2014.

[22] C. -H. Lee: Mining spatio-temporal information on microblogging streams using a density-based online clustering method. Expert Systems with Applications, Elsevier, vol. 39, issue 10, pages 9623-9641, February, 2012.

[23] A. Scholkopf, C. Burges, and A. Smola. Advances in Kernel Methods - Support Vector Learning. MIT Press, 1998.

[24] V.N. Vapnik. Statistical Learning Theory. John Wiley and Sons, 1998.

[25] Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[26] Ho, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

[27] Press, William H.; Teukolsky, Saul A.; Vetterling, William T.; Flannery, Brian P. (2007). "Section 16.5. Support Vector Machines". Numerical Recipes: The Art of Scientific Computing. New York: Cambridge University Press.August,,2011

[28] Breiman, Leo (1984). Classification and Regression Trees. Boca Raton: Chapman & Hall/CRC.