
# Prediction of Dublin Bike Sharing System

*A thesis submitted in fulfilment of the requirements for the degree*

*of*

***Master of Science (Data Analytics)***

*by*

**Poojitha Vishukumar**

**(Student ID: 18231013)**

*in the*

*Discipline of Information Technology*

Supervised by

**Dr. Edward Curry**

*National University of Ireland, Galway*

*August 2019*

# *Acknowledgement*

I would like to thank Dr. Enda Howley for allocating me this project and giving me the opportunity to work on a growing and applied area of research. I wish to acknowledge the support of my enthusiastic supervisor Dr. Edward Curry whose dedicated support ,expertise and timely feedback helped me understand the direction of the project and without whom this thesis would not have been possible. I also would like to express my deepest gratitude to God almighty. Lastly, I would like to thank my family and friends, who have been there for me over the course, providing advice and encouragement along the way. I hereby, certify that this material, which I now submit for assessment on the programme of study leading to the award of Masters in Science in Computer Science (Data Analytics) is entirely my own work, that I have exercised care to ensure the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

**Signed:**

**Dated: August 28, 2019**

# *Abstract*

Bike sharing systems are rapidly being introduced in European cities for daily mobility leading to sustainable transportation. These systems add to city's public transport system providing fast and easy services to the public. The bike sharing is increasing dramatically due to its unique feature of being convenient, low-cost and environmentally friendly, giving rise to problems to both users and operators. Amongst the European countries, Dublin city is considered in this thesis for the study and research. In this city, there are two types of bikes sharing systems, docked and dockless bike sharing systems. Dockless are newly being introduced but docked bikes are increasing tremendously since few years. In docked based bike sharing system, the bikes are rented by user by signing up to the services, checking out and returning the bike at the dock. When the user signs up and checks out, the basic information such as user's information, station check-in, check-out, bike usage duration and many more is stored.

Nowadays, the primary issue of bike sharing is the uneven distribution of bicycles caused by the ever-changing usage and supply. The uneven distribution of bikes can be due to station area, timing, population, weather, events and many more. These external factors affect the availability of the bikes and gives rise to necessary efficient model for prediction of bike's availability. Efficient prediction model will help in future prediction of availability of bikes with respect to the factors helping the users to use the bikes also, guiding the operators for redistribution of bikes accordingly.

This thesis aims at art of bike sharing systems in Dublin city proposing spatio-temporal bike mobility prediction model based on historical bike sharing data and real data on a per - station with sub-hour granularity. With the time series multivariate prediction model, advancement in re-balancing of bikes could be seen paving the way for rapid deployment and adoption of bike sharing system across the world. To analyse the bike, weather and event data and build the prediction model, deep learning models such as regression model and recurrent neural network model are studied and implemented. In particular, Autoregressive Integrated Moving Average model (ARIMA) and Long Short-Term Memory model (LSTM) have been implemented to predict the availability of the bikes at the given timestamp. The root mean square error is calculated and depicted on trend curve deviation graph for both the models where LSTM proves to work best with defined large series data whereas ARIMA shows best results with comparatively less series data.

# Table of Contents

# List of Figures:

# List of Tables:

# *Abbreviations:*

| | |
|---|---|
| **CRISP-DM** | **C**ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining |
| **ARIMA** | **A**uto **R**egressive **I**ntegrated **M**oving **A**verage |
| **ETL** | **E**xtract, **T**ransform, **L**oad |
| **WEKA** | **W**aikato **E**nvironment for **K**nowledge **A**nalysis |
| **LSTM** | **L**ong **S**hort-**T**erm **M**emory |
| **RNN** | **R**ecurrent **N**eural **N**etwork |
| **RMSE** | **R**oot **M**ean **S**quare **E**rror |
| **DNN** | **D**eep **N**eural **N**etwork |
| **AR** | **A**uto **R**egression |
| **MA** | **M**oving **A**verage |
| **ARMA** | **A**uto **R**egressive **M**oving **A**verage |
| **ARIMA** | **A**uto **R**egressive **I**ntegrated **M**oving **A**verage |
| **ACF** | **A**uto **C**orrelation **F**unction |
| **AIC** | **A**kaike **I**nformation **C**ritera |
| **BSS** | **B**ike **S**haring **S**ystem |
| **DCC** | **D**ublin **C**ity **C**ouncil |
| **SBRP** | **S**tatic **B**ike **R**epositioning **P**roblem |
| **DBRP** | **D**ynamic **B**ike **R**epositioning **P**roblem |

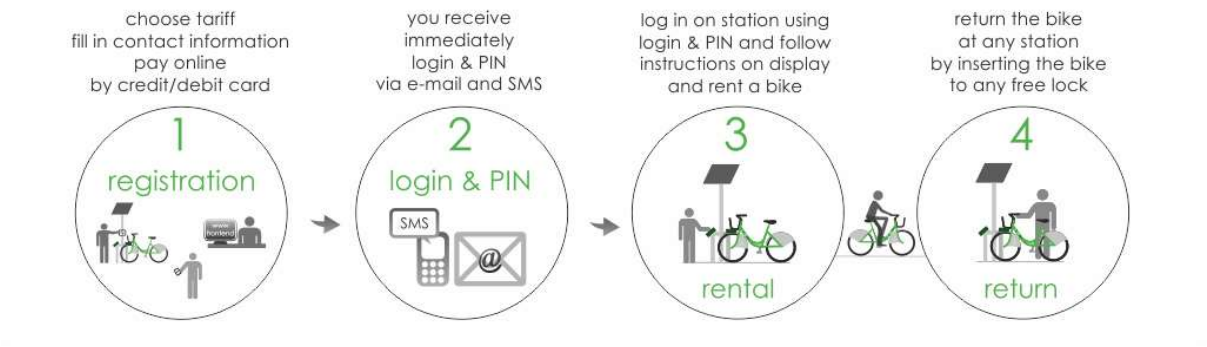
# 1 Introduction

## 1.1 General Overview

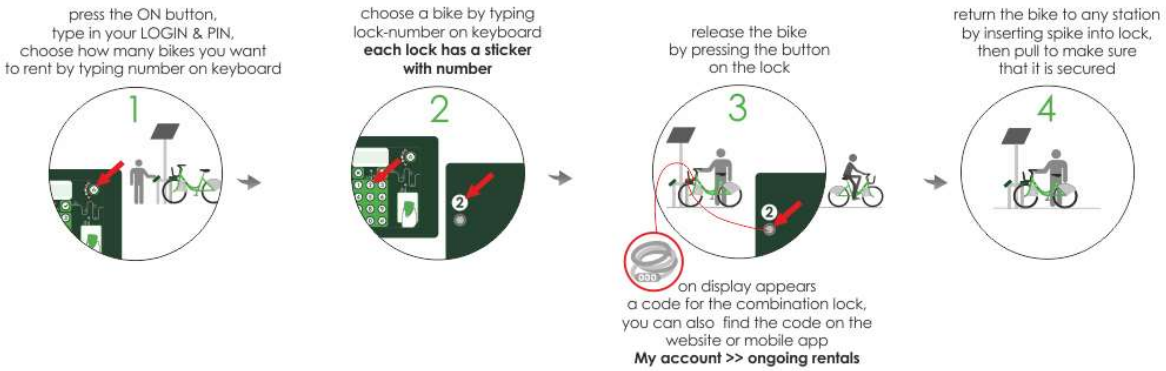
Bike sharing systems is a new generation of traditional bike rentals where user is able to easily rent, return and take membership of bikes from a particular place to another automatically without any human resources. According to recent statistics, there are about over 500 bike sharing systems around the world with more than 500 thousand bikes. (Dua, D. and Graff, C.,2019). The BSS program is growing tremendously due to their important role in traffic, environment and health issues. It can be used seamlessly for individual trips consisting of multiple transportation modes which is important component of a modern, sustainable and efficient multimodal transportation network. In general, Distributed bike sharing system is of two types, docked based and dock-less based bike sharing system. In docked based BSS, the bikes are rented and returned from and to the docking stations.

Typically, a rental station includes one terminal stand and several bikes shown in Fig. 1.1. A terminal works on IoT(Internet of Things), where a device is attached to the terminal which communicates with the electronic locker attached to the bikes. (Raviv, T., Tzur, M. and Forma, I. A., 2013) Initially, user needs to register online to the services by entering the basic contact details which proceeds to the payment. After the payment is done, an electronic message is received with PIN number which is used to login at the terminal for rental of bikes as shown in fig 1.2. Internally, when the user rents a bike, signal is sent to the terminal that the bike with respective locker has been vacated. A user can return the bike to destination stand when there is a vacant locker as shown in fig 1.3. All the rental and return transactions are recorded and reported in real time by JCDecaux of Dublin City Council which is the central control facility for Dublin city. Thus, this information of state of system, in terms of number of available bikes and number of vacant lockers available for each station is updated by DCC to the BSS operators and these operators make the information available online for the users.

*1.1 Dublin Docking Terminal*
*(En.wikipedia.org, 2019)*




*1.2 Bike sharing working*
*(Anon, 2019)*




*1.3 Working of Rental and Return of bikes*
*(Anon, 2019)*

Apart from real- world application of bike sharing systems for the accessibility of the users and redistributors, the data generated by these systems can be used for further researches. Antithetical to other transport services such as bus or subway or luas, the total duration of travel, the arrival time and the departure time and position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city (Fanaee-T, Hadi & Gama, João.,2013).

## 1.2 Problem Statement

Bike sharing has significantly been enhancing the urban mobility as a sustainable mode of transportation, but it has few limitations due to the effects of fluctuating spatial and temporal demand. According to previous researches, (Chen et al., 2016, Li et al., 2015, Lin, 2018, Zhou, 2015), it is very common for bike sharing systems with fixed number of terminals that some terminals can be completely empty with no bikes to check out during peak hours of usability while others are full precluding bikes from being returned at those locations. For dock-less bike sharing systems, there is lot of flexibility with the parking system but it even possesses more challenges with respect to ensuring bike availability at some places and also, prevent surplus bikes from blocking sidewalks and parking areas. For both types of BSSs, accurate bike-sharing demand predictions can help operators to generate optimal routes and schedules for rebalancing to improve the efficiency of BSSs.

Docked bike sharing service is widely being used in Dublin bringing much of convenience for the short trips, reducing greenhouse gas emission and congestions but the management of the bikes due to its fluctuating spatial and temporal demand , limiting the capacity usage of bikes and docking stations has increased the concern. Empty or saturated bike stations at certain places at certain time causes problems to the commuters. In order to address this issue, bikes are manually rebalanced by the trucks with an arbitrary timetable. Due to the unbalanced distribution in the number of bikes at docking stations, these resources cannot be fully utilized (Wang and Kim, 2018).This problem of predicting the frequency usage of bike and manually loading of bikes accordingly can be effectively dealt by predicting the availability of the bikes with respect to time at respective stations.

## 1.3 Core Research Questions

The core goal of this research is to answer the questions related to rebalancing problem. For instance, what if everyone gets off from the Dublin bus at the major stop like City Centre and then they all grab the bikes from the nearest station, those bikes will be depleted before long resulting in unequal riding patterns. (Dell'Amico, M., Hadjicostantinou, E., Iori, M. and Novellani, S., 2014).

The major challenging questions that have been focused on and is studied are:

- Is it possible to classify bike sharing stations according to the bike availability?
- What if the station is empty when the user wants to rent a bike?
- What if the station is saturated when the user wants to return the bike?
- How to build a complete ETL framework to analyse the data and report results?
- Can one-commodity pickup-and-delivery capacitated vehicle routing problem be solved?

## 1.4 Existing Approaches

For both types of Bike sharing systems, accurate bike-sharing demand predictions can help operators to generate optimal routes and schedules for rebalancing to improve the efficiency of BSSs. The affordability and availability of bikes influences the variability of users. The ability to predict the number of users can allow the sponsored businesses and governments that oversee these systems to manage them in a more efficient and cost-effective manner.

The problem of repositioning operations which consists of removing and reloading of bikes from overloaded stations to shortage stations using dedicated fleet of trucks where the decisions regarding the routes the truck needs to follow and the number of bikes that needs to be removed or loaded at each stations at each visit of truck is solved by dividing the problem into two categories, Static bike repositioning problem(SBRP) and dynamic bike repositioning problem (DBRP). The repositioning activities carried at the night when the rate of usage of bikes is negligible falls under SBRP and the operations carried during day when the bike systems usage is rapidly changing is considered as DBRP. The Static problem is solved by estimating the demand pattern on past demand data on similar days. Solution for SBRP fails to work efficiently during the unpredictable events such as traffic slowdown, vehicle breakdown, heavy rain, and many more unfortunate events( Raviv, T., Tzur, M. and Forma, I. A., 2013).

The demand faced by the BSSs is also sensitive towards the partial predictable effects such as weather conditions, public events in the city and many more. This kind of situations affecting the usability of the bikes requires dynamic repositioning and which is dealt by statistical time series model which analyses the historical data, the traffic data, arrival and departure rates on hourly, daily and weekly basis. To make the dynamic approach to predict the demand pattern more efficiently, the current traffic trends and weather data is considered along with the historical data.

## 1.5 Proposed Approaches

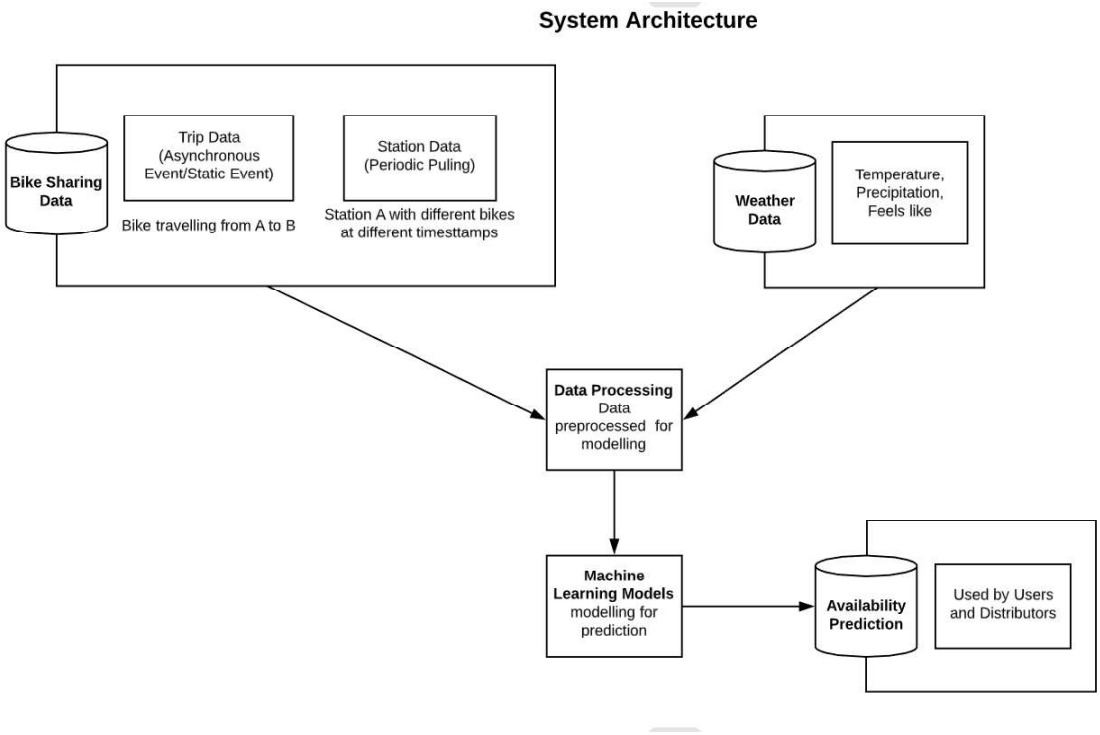
The use of Internet of Things technologies having embedded sensors to obtain the real-time data has allowed the cities to provide a better service, where the data is used to study, plan and manage the bike sharing docking station and adapt to the need of the commuters. As discussed in the previous sections, the efficient way of rebalancing scheme is by predicting the demand of bikes in different locations covered by different services and selecting the way to relocate bikes in order to satisfy user's demand.

Rather than following the conventional static method for rebalancing, this thesis proposes a dynamic rebalancing schemes which aims at providing user's satisfaction by minimizing probability of system failures. (Chiariotti, F., Pielli, C., Zanella, A. and Zorzi, M., 2018). There are many other factors that alters the usability of the bike sharing patterns, such as time of day, population of the city, location of the station, nearby amenities, weather and climatic changes, week of the day, nearby events and bank holidays and many more. Different machine learning time series regression models help to predict the availability of bikes on basis of multivariate attributes. In this, time series model, a sequence is taken at successfully equally spaced points at each hour of time. There are lot of work done with univariate dataset wherein just the availability of the bikes is considered against the time interval. To make the work closer to real life problems, multivariate data set has been chosen. In this thesis, multivariate time analysis consists of simultaneous multiple time series that deals with dependent data such as time stamp which includes hour of the day, weekend of not, season of the year, public holiday or not, temperature, weather and feels like attributes.

The goal of the thesis to use deep learning models in Keras such as Multivariate Linear Regression and Long short-term memory model which effectively predicts the number of bikes shared which are available at any given time period. It can be effectively be done by using available information of bike shared on hourly basis with weather, day of the week, status of

the day and events. The focus on behavioural pattern of the stations depending on the space and usage can be done to categorize them into highly used to least used stations during the period of time.

## 1.6 System Architecture




*1.4 System Architecture*

Figure 1.4 describes the system architecture of the data flow from data collection to data transformation and data usage. The bike data and weather data is collected through APIs which include trip data of bike from station A to station B and Station data with its respective number of bikes and the temperature, precipitation and feels like data respectively. The data is further processed and used in the machine learning models for prediction and evaluation of the model. The predicted data is stored and used by the commuters and the distributers in the end.

## 1.7 Research Methodology

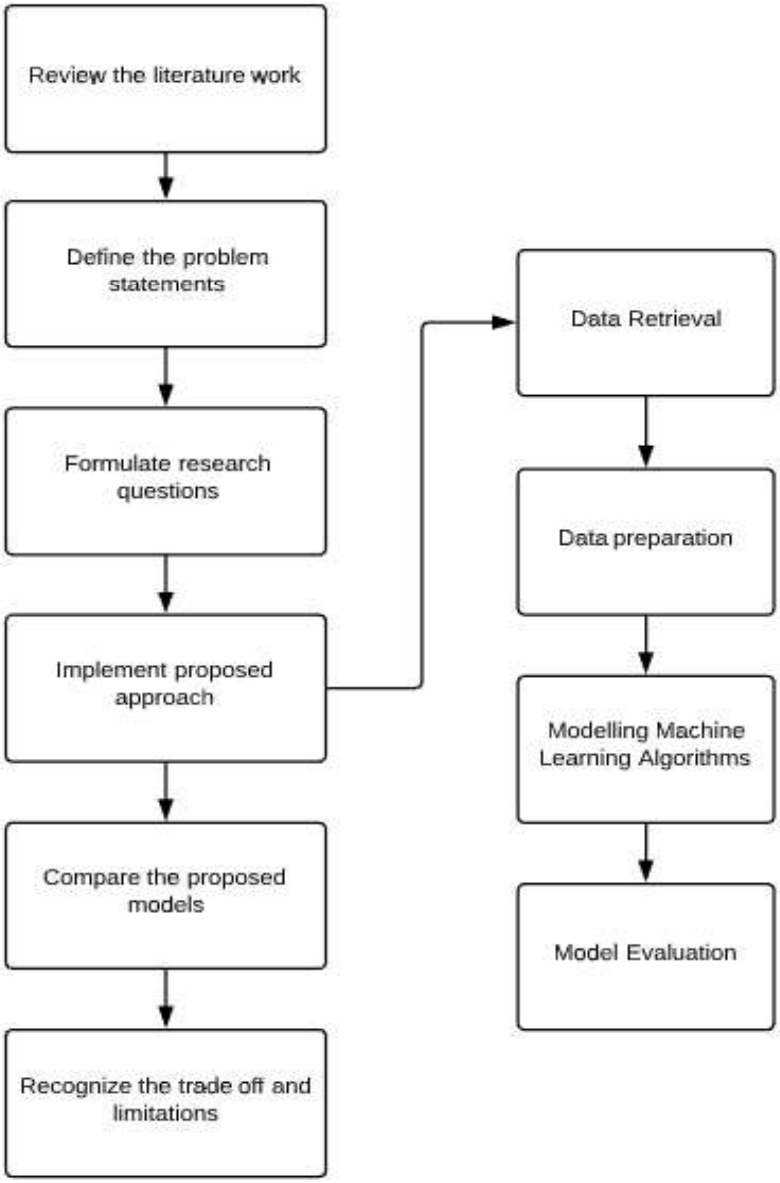



*Figure 1.5 Methodology Flowchart*

The methodology followed in this thesis is shown in the above fig 1.6 and is explained in the following steps:

- Initially, review of literature work related to Bike sharing systems around the world and their challenges were studied.
- The problems that exists which were gathered from different researches were formulated.
- Research questions were formulated that is planned to study and answer through the model developed in the thesis.
- Two machine learning timeseries models (LSTM and ARIMA) have been developed to solve the defined problem of Bike sharing which follows CRISP-DM steps. The detailed pipeline is explained in Chapter 4.
- After each model's evaluation, its performance is checked through comparing respective root mean square errors.
- Limitations of the proposed models is recognised, and future work is proposed.

## 1.8 Contributions

The contributions of this work are manifold:

- The main challenge and contribution were to collect the historical and real time bike data which resulted in extracting the real time Dublin bike data from JCDecaux API managed by Dublin City Council. A python script was run for a month to retrieve data. Similarly, weather data has been collected using pwyapi library in python.
- The data was collected, cleaned and processed for modelling and has been made publicly available as open source for future use.
- Machine learning timeseries models have been developed based on multivariate attributes to analyse and predict the forecast. Previous researches include study and implementation on single feature but this thesis deals with multiple features.
- According to the previous researches and study, it is observed that there is no work done on prediction of bikes considering multiple factors like weather, event, day of the week for the city Dublin. This thesis contributes by working on these factors by modelling multivariate timeseries model and comparing their performances.

- The various factors affecting the availability of the bikes such as temperature, wind speed, precipitation, events, day of the week is considered and is visualized based on time, day and station in Tableau.

## 1.9 Thesis Outline

The rest of the thesis is organised as follows:

- *Chapter 2: State of Art* – This chapter elaborates about the timeseries related work and reviews the previous study done on the bike sharing problems. It discusses about the Regression, Long Short-Term Memory and Auto Regressive Integrated Moving Average model work done on bike sharing systems and their limitations with the takeaway points which has been implemented in the thesis, discussed later in chapter 4.
- *Chapter 3*: *Dataset Description* – This chapter includes brief description about the method of dataset collected, explains its attributes and the use as features for future feature engineering procedures explained in Chapter 4.
- *Chapter 4*: *Method and Implementation* – This chapter explains the in-depth description about the methodology of the thesis which was mentioned in section 1.6. It provides the complete knowledge about the CRISP-DM pipeline which includes elaborative explanation of all the steps from data retrieval to model evaluation along with comprehensive study of LSTM and ARIMA model.
- *Chapter 5*: *Results and discussions* - This chapter discusses the results of the implementation of the model, comparison of their performances and discussions about the better model for the bike series dataset.
- *Chapter 6*: *Conclusion and future work* – This chapter concludes the thesis with listing the advantages and limitations of the work and discusses the potential impact along with future work.
- *Chapter 7*: *References* – This includes all the referenced sources cited in Harvard style.