Influencing factors and Short/Long-term prediction of availability for the Dublin Bike Scheme

Nidhin George[1], Joseph Timoney[1] and Thoa Pham[2]

[1]Department of Computer Science, Maynooth University, Maynooth, Co. Kildare, Ireland
[2]College of Business, Dublin Institute of Technology, Dublin, Ireland

Summary

The popularity of bike sharing in Dublin city has been demonstrated by its growing subscriber numbers. This paper presents continued work on Dublin bike data analysis. The focus now is on different prediction strategies. These will help users plan their trip by telling if there are bikes or empty docks at a specific station nearby. It also assists management with effective rebalancing of inventory between stations. This analysis takes into account the usage history, weather data and holiday data. It investigates regression and classification predictive modelling. A comparison of these models is performed to decide on the most accurate model.

KEYWORDS: public bicycle systems, prediction of bike availability, regression models, classification models

1. Introduction

Dublin's public bike sharing system was established in 2009 with 450 bicycles in 40 stations. Since then, it has grown to 110 stations with 42,000 annual subscribers (Dublinbikes, 2019). Previous work (Pham Thi et al, 2017), and (Timoney at al, 2018) offered some analyses on discovering usage patterns across the stations and introduced a preliminary investigation into prediction. However, what might be key influencers such as the weather and holiday periods were not accounted for. This paper provides a more complete analysis and also examines the creation of prediction models using different machine learning techniques. Predictions in the short-term and the long-term are now considered. They require Regression and Classification models respectively. Predictions from the most accurate of the models could thus be used to operate Dublin Bikes more efficiently for both users and managers. The next sections will explain the data collection, analysis, modeling, and results.

2. Data capture

The data was received from the JCDecaux API every 10 minutes between 12-09-2016 to 26-04-2017. It included the stations, time of update, the number of available bikes, and the number of available docks. To determine other influential factors, additional information such as holiday data, season and weather data was gathered. The dates for Irish national holidays and observances for 2016 and 2017 were obtained from the website (TimeandDate.com, 2019). The weather data for Dublin was collected from Weather Buoy Network data hosted by Marine.ie (Irish Marine institute, 2019). Only windspeed and air-temperature values were used from this dataset. Pre-processing involved data cleaning and joining the datasets from the different sources. Pre-processing and feature engineering was done in R using dplyr and using numpy in python. Table 1 lists the feature set used for analysis.

Table 1 Analysis features

Feature	Meaning
Time	Minute value converted to half hour interval.
Weekday	String value denoting the day of the week.
Prev_period_diff	Difference in the number of bikes available compared to the previous row for a specific station.
Check_in	Cumulative number of bikes that were checked in compared to the previous time period.
Check_out	Cumulative number of bikes that were checked out compared to the previous time period.
av_ind	Categorical value to indicate the level of availability of bikes in a station based on percentage.
Holiday	Categorical value to indicate whether the day is a holiday.
Prev_bike_num	Number of bikes available from the previous time frame for the same station.
Prevweek_bikenum	Number of bikes available at the station during the previous week at the same time.

3. Holiday, Season and Weather analyses

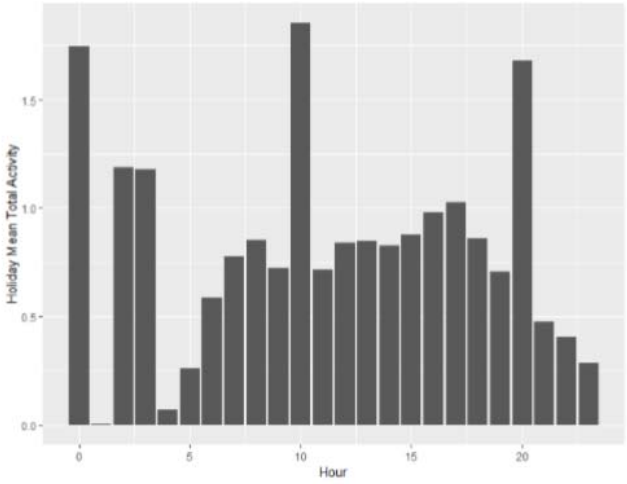

Figure 1 Activity during Holidays

The effect of holiday times on bike usage activity was first examined. Figures 1 and 2 provide plots for averaged activity on holidays and non-holidays respectively. The most significant difference is the holiday-time activity in Figure 1 around 20:00 compared to other times in the evening or night, and in Figure 2 for non-holiday time there are notable peaks around the rush hour working times in the morning and evening.

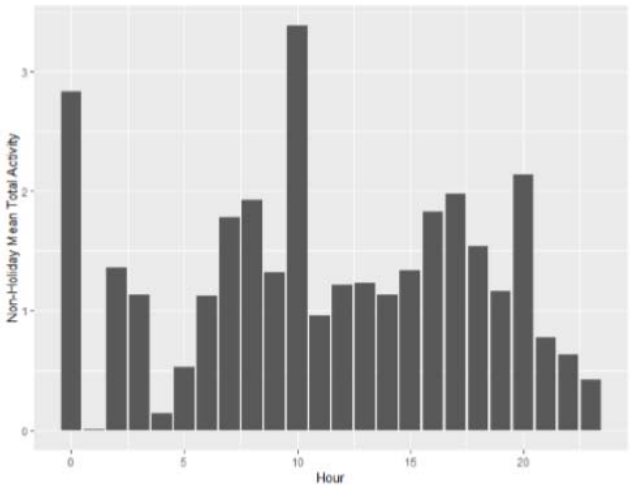

Figure 2 Activity during Non-Holidays

From a seasonal perspective, in Figure 3 noticeable differences can be seen between average usage values. Winter and Spring have an almost similar amount of usage, but the activity was consistently higher during the Autumn months for each day of the week.

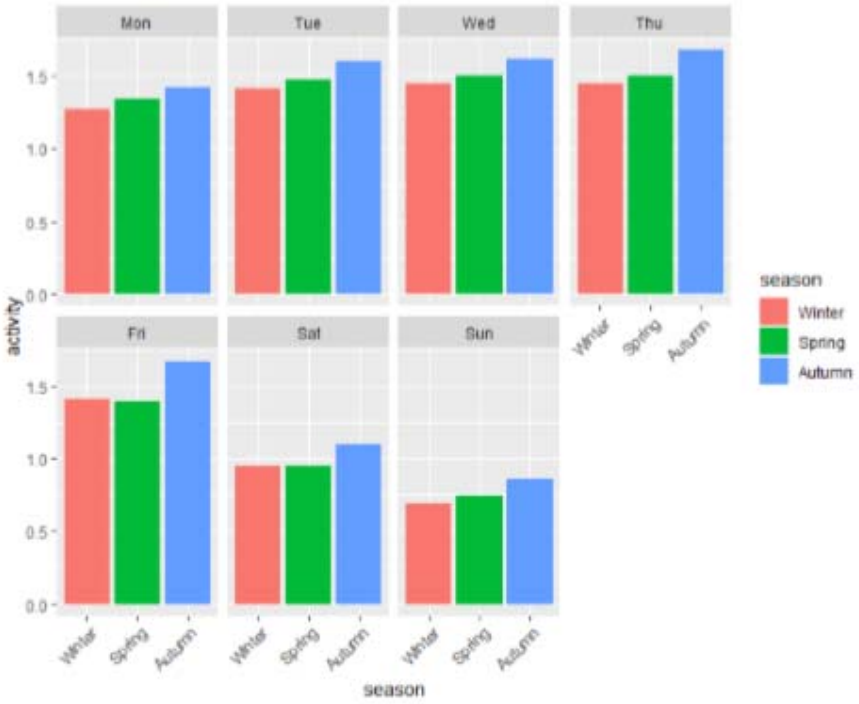

Figure 3 Activity across different seasons

The importance of air-temperature and wind speed were also explored. In Figure 4, an increase in temperature has a positive effect on the number of bikes being rented whereas the user numbers appear to be somewhat higher when the wind is blowing at a moderate speed. Usage is smaller on average at times of high wind speeds and low air temperatures.

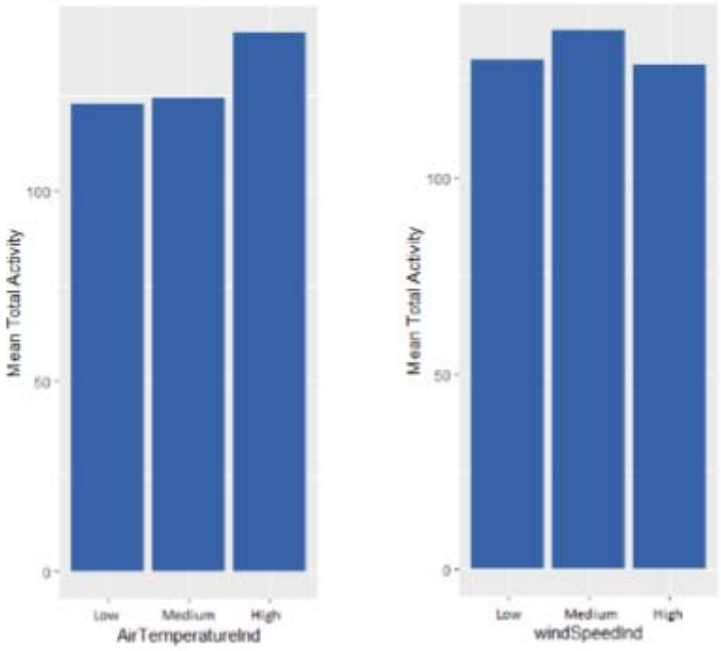

Figure 4 Activity during different weather

4. Short-term prediction analysis and evaluation

The short-term prediction can be useful for those who wish to know about the availability of bikes in nearby stations for the immediate future. Regression models were built using the techniques of Boosted Linear Regression, Linear Regression with Stepwise Selection, Gradient Boosting machine, Random Forest, and LSTM. The models were based on the features in Table 1, but excluded data on week, the station location, the cluster the station belongs to, and the weather data. Figure 5 shows examples of the actual (in blue) and predicted (in red) values for the mean number of bikes available for fifteen different stations that were determined using the Random Forest model. In almost all cases the prediction results are very good with just some errors in the model for the King Street North and Western way stations.

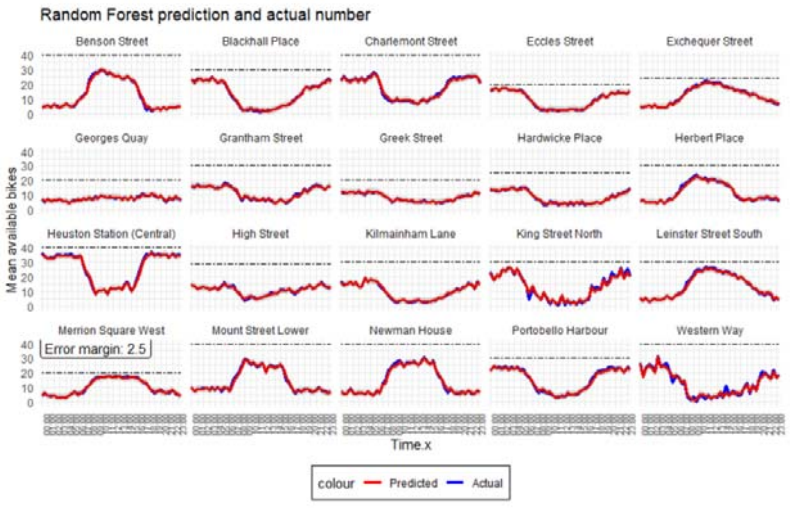

Figure 5. Analysis result with Random Forest model

From the prediction results for each technique, the Root Mean Square (RMS) and Root Mean Square Logarithmic Error (RMSLE) of each model were computed. Table 2 shows the errors of each.

Table 2. The regression prediction result comparison

Model	RMSE	RMSLE
Boosted Linear Regression	3.33	0.61
Linear Regression with Stepwise Selection	2.99	0.42
Gradient Boosting machine	2.87	0.41
Random Forest	2.49	0.39
LSTM	2.68	0.41

Accuracy: From Table 2, the Random Forest model is most accurate.. The prev_bike_num predictor was found to have the highest importance in the model. Among the models, those based on linear regression gave the least accurate results which indicated a lack of linearity in the dataset.

Computational Time: The Random Forest algorithm took the most amount of time in training the model (12 hours) whereas all others took considerably less time (less than or equal to an hour) in producing a model after training and validation.

5. Long-term prediction analysis and evaluation

The long-term prediction categorises stations into low and high bike availability using classification models. The features are similar to the short-term predictors but exclude data on the number of bikes available from the previous time frame for the same station. Additional data included was the availability numbers at the same times for a specific station over the four previous weeks. Table 3 presents the accuracy of each model and the Matthews Correlation Coefficient (MCC). The MCC illustrates the effectiveness of a binary classifier and varies from +1 to -1.

Table 3. The classification prediction result comparison

Model	Accuracy	MCC
Gradient Boosting Machine	80.12%	0.59
Linear Discriminant Analysis	87.47%	0.56
Quadratic Discriminant Analysis	69.69%	0.37
Random Forest	88.43%	0.76
K-Nearest Neighbours	77.83%	0.54
LSTM	87.02%	0.73

Accuracy: The Random Forest gave the most accurate classification of stations into low and high categories of inventory availability with 88.26% accuracy and a Matthews Correlation Coefficient (MCC) value of 0.76. The four predictors representing the bike numbers from previous 4 weeks showed the most importance in the random forest classification model.

Computational Time: Random Forest and KNN took the highest amount of time (multiple hours) for training compared to the other algorithms which completed the process within an hour. However, the overall time taken for classification models was considerably less compared to that of regression model training.

6. Conclusion

An ongoing analysis was made for data related to the Dublin bike scheme. A new investigation was made into the influential factors of season, holiday period, and weather. This was followed by the development of predictive models. Regression models were used to provide short term predictions whereas Classification models were applied to long term prediction. Overall, the Random Forest model gave the most accurate results for both. The result of this analysis can easily be incorporated in mobile or web applications to give predictions to users about source stations with sufficient inventory or with free docking stands at the destination. These models would also help management create load balancing strategies.

References

Dublinbikes, 2019. Dublin Bikes - How does it work? URL: http://www.dublinbikes.ie/Howdoes-it-work.

T.T. Pham Thi, J Timoney, S Ravichandran, P Mooney, A Winstanley (2017). *Bike Renting Data Analysis: The Case of Dublin City*, GISRUK, 2017

J Timoney, C Amaral, TTP Thi, A. Winstanley (2018). *A continuation on the data analysis for the Dublin Bike rental scheme*, GISRUK, 2018

TimeAndDate, https://www.timeanddate.com/, 2019

Irish Marine Institute, 2019. Weather Buoy Network Real Time Data. http://data.marine.ie/Dataset/Details/20972.

Biographies

Nidhin George has recently completed a MSc in Data Analytics at the Computer Science Department in Maynooth University. Joe Timoney is a lecturer in the Computer Science Department, Maynooth University. Thanh Thoa Pham Thi is a lecturer at the school of Management in Dublin Institute of Technology. All are interested in developing spatial analysis tools for helping users of public bike schemes.