A Machine Learning Based Bike Recommendation System Catering To User's Travel Needs

Ananta Kumar Das[1], Amogh Manoj Joshi[2], Subhasish Dhal[3]
[1]*Department of Software Engineering*
International Institute of Information Technology, Bangalore, India
Email: ananta.kumar@iiitb.org
[2]*Department of Electronics and Telecommunication Engineering*
Vivekanand Education Society's Institute of Technology, Mumbai
Email: joshiamogh9@gmail.com
[3]*Department of Computer Science and Engineering*
Indian Institute of Information Technology Guwahati, India
Email: subhasis@iiitg.ac.in

***Abstract*—Bike sharing system has emerged as a popular transportation system. Each ride accumulates data which includes bike id, trip information, station information, etc. Prevailing weather conditions also play a major role in bike sharing systems. This paper attempts to infer from the public bike sharing dataset and weather data to help the riders in selecting a suitable bike for their travel demands. For this, we propose a novel bike recommendation system using machine learning techniques to cluster bikes with similar behavioural patterns and predict the best cluster for user's travel needs. A strategy is further proposed to identify actively used bikes in predicted cluster. We evaluate the proposed method using real world bike sharing dataset from Divvy Bike System at Chicago, North America. Results show that the method effectively infers bike trip patterns along with the weather conditions to suggest the best cluster of bikes suitable for every individual trip.**

***Index Terms*—Machine Learning, Bike Sharing system, Weather data, K-means clustering, Neural Network**

I. Introduction

In the recent years, Bike Sharing Systems have gained a lot of popularity. Bike Sharing Systems (BSS) are used by a wide range of users from corporate professionals to daily cyclists owing to the cheaper and faster alternative that they provide in facilitating short-distance trips among various stations in the city. BSS has been proven a better alternative as compared to public transport services for several factors such as alleviating city traffic, reducing pollution and providing health benefits to the users in the form of exercise while riding the bikes, etc. As a result of this growing popularity, BSS have been deployed in more than 2000 cities all across the world [1]. According to a report by the National Association of City Transportation Officials (NACTO), 35 million bike-share trips took place within the United States in 2017 across 100 bike sharing systems [2]. BSS users go to the desired start station, rent a bike and return the bike at the destination. While renting, the users select any one bike out of the bikes available at the station while renting the bike. Many a times, the users feel uncomfortable with the bike for possible reasons ranging from rigidity while pedalling due to lack of usage of that bike, to uncertainty if that bike is appropriate for travelling the desired distance in desired time, in the prevailing weather conditions such as rain, windy storms, humidity conditions, etc. Such unsatisfied users return the rented bike to the start station immediately and select another bike. BSS companies discard such trips less with duration less than one minute from their database considering those events as unsatisfied or mock trips. This issue wastes user's time while returning and changing bikes, and also adds to the unsatisfactory feedback of the users, thus proving detrimental for the future of bike sharing company. Also, if a user selects a bike and realizes certain issues with the bike after riding a certain distance away from the station, he/she has no option but to either continue with that bike or return that bike at another nearby station and rent another bike to continue his/her journey. Thus, there is a lot of uncertainty from the user's perspective for which bike is best according to the desired distance and desired time duration of travel, taking in consideration the prevailing weather conditions at the time of rental. Having said that, choosing a bike should be the user's decision as the user may consider some other factors as well while choosing a bike like physical condition of the bike, proper functioning of brakes etc. Therefore, directly allocating a particular bike to the user is not an optimal solution. By analysing the bike sharing trip network, similarity between a group of bikes can be observed based on the trip patterns. Due to the large size of trip records in the dataset, its not manually possible to observe and gain insights about

978-1-7281-6916-3/20/$31.00 ©2020 IEEE

the bikes with similar trip patterns. Machine learning clustering techniques effectively analyse patterns in the data and form clusters with similar features, thus helping us in feature extraction and pattern analysis. In this paper, we propose a novel bike recommendation system based on user's travel demands like desired trip duration and desired trip distance along with some weather factors into consideration, further allowing the user to choose any bike among them according to his/her choice.

The noteworthy contributions of this paper can be summarized as :

- Performs a fine grained analysis of the trip patterns for each bike and clusters the bikes with similar behavioural patterns using K-means clustering approach.
- Proposes a novel bike recommendation system which predicts the best cluster of bikes based on user's travel demands and prevailing weather conditions at the time of rental, thus solving user's uncertainty issue.
- Presents a machine learning based approach for training the proposed system by using several ensemble machine learning methods and perform a comprehensive evaluation of the proposed methods based on evaluation metrics.
- Proposes a bike usage optimization strategy on top of bike recommendation system, by analysing the active and passive bikes per station and recommending top 3 active bikes, thus improving the bike flow network.

The paper is organized as follows. Section I gives the introductory part and the need of developing a bike recommendation system based on user's travel needs. Section II gives an overview of the existing works proposed in the bike sharing domain. Section III describes the framework of the proposed system. Section IV discusses the experimental setup for the training approaches while Section V discusses the results. Finally, section VI concludes the paper.

II. Related Works

The growth of public bike sharing companies around the world and some of their open-sourced data has facilitated a significant amount of studies being published in this domain in the recent years. Various problems have been addressed by the researchers with a variety of approaches to improve the service quality of bike sharing systems. Several studies have been summarized in this section. A lot of works have been performed in calculating demand per station by predicting the number of bikes at each station. In [3], Huang et. al developed a bimodal poisson algorithm for predicting the number of bikes at each station. In [4], Yao et. al used temporal and spatial features associated with each station to calculate the demand. Almanna et. al used mini-batch gradient descent for linear regression and locally weighted regression(LWR) for predicting bike counts in [5]. In [6], Li et. al proposed a multi-categorical probabilistic approach for short-term demand prediction. Huang et. al predicted hourly demand prediction for top stations using clustering approach in [7]. Station analysis is an important research area when studying bike sharing systems. In [8], Tomaras et. al used a demand forecasting model to identify stations as high or low demand and developed relocation strategy which maximised station utility. Zhang et. al proposed a novel visual method for analysing bike sharing systems and planning station dock capacity in [9]. Some studies have also been conducted for bicycle allocation for repositioning [10]− [12]. In [13], Chen et. al conducted an empirical study for inferring bike flow patterns from trip data. A greedy trip planning algorithm was used by Li. et al for large scale trip planning in [14]. In [15], Zhou et. al used a context aware flow prediction model to infer bike flows. Ljubenkov et. al used several machine learning approaches for predicting flow attributes with a bike rebalancing strategy [16]. In [17], Guo et. al clustered stations based on their functional profiles and validated the predictions by comparing with the data of points of interest and station names, stating that the study was useful for city planning purposes. To summarize, lot of works have been done in predicting station demand, analysing bike trip flows and repositioning bikes to the stations in demand. Thus, recommending bikes to users based on their travel demands and helping them to choose a bike is a largely unsolved problem. None of the above mentioned works focused on user inputs or developing a system for user's travel needs. We cluster bikes based on their behavioural patterns inferred from trip records. Further, we use this clustered data to propose a novel bike recommendation system. On top of this, we use a bike usage optimisation strategy while recommending bikes, which improves the usage of active bikes and maintains the flow of bikes. The proposed system solves multiple issues along with prioritizing user's travel demands, and helps improve the service quality of bike sharing systems, having achieved promising results.

III. Methodology

In this section, we brief over the data used in this study, the clustering technique used and the methodology followed in the proposed bike recommendation system. Fig. 1 describes the work flow followed in this study.

A. Data Acquisition

The latest Divvy Bike Sharing Data as of 2019 was used in this study [18]. This data contained 10,13,659 trip records from 600 stations containing 5989 bikes. It contained the following features:

- Trip ID
- Bike ID
- Rental Date and Time
- Return Date and Time
- Start Station ID along with its coordinates
- End Station ID along with its coordinates
- Trip Duration

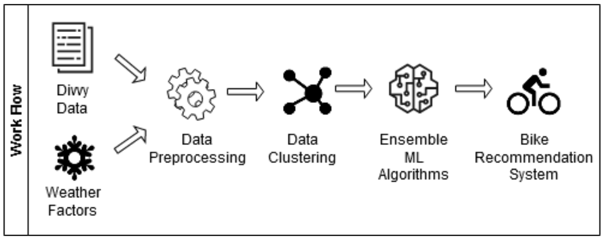

Fig. 1. Work Flow

We wanted to analyse the bike behaviour based on trip distance and trip duration. As the weather conditions also play a major role from the user's perspective while choosing a bike, we also wanted to analyse bike behaviour in accordance with the weather factors. Weather data was not available in Divvy Bike System dataset. So, weather data of Chicago was collected from city of Chicago database [19]. The weather dataset was merged with Divvy dataset based on the year, month, day and hour of rental. This newly customized data contained weather attributes like Temperature, Humidity, Rain Intensity and Wind Speed.

B. Bike Clustering

The proposed bike recommendation system aims to help the user to choose a bike by recommending top 3 bikes from the rent station. However, predicting the best bike based on user's inputs is not an appropriate solution as the predicted bike may not be available at that station. To have a robust bike recommendation system, it is optimal if the system predicts the best cluster of bikes for the given inputs as each cluster is fundamentally, a set of bikes similar in behavioural and trip patterns. We wanted to cluster bikes based on the trip patterns along with the weather data. Two most important attributes associated with every trip are trip duration and trip distance. Trip duration was available in Divvy Data. Due to absence of trip distance data in the divvy database, the minimum road distance between the start and end stations for every trip was found out from their coordinates. Though this distance wasn't accurate, it was sufficient for gaining insights about the behavioural patterns of each bike. To analyse the behavioural patterns of each bike, we took the mean trip distance, mean trip duration and the mean weather factors at which the bikes were used throughout the dataset. We used K-means clustering approach for clustering bikes with similar patterns. To decide the number of clusters for the data, we used the Kelbow method which indicates the best number of clusters by calculating the distortion score by fitting the model with a range of values for K. The second dip in the plotted graph in Fig. 2 indicates the best number of clusters for the input data. We grouped the bikes into 3 clusters by Kmeans clustering. The clusters along with number of bikes in each cluster have been described in Table 1. We further explored these clusters by analysing the trip records of bikes belonging to respective clusters. Fig. 3 plots the hourly distribution of bike trips for respective clusters. The next section clearly explains the functioning of the proposed bike recommendation system.

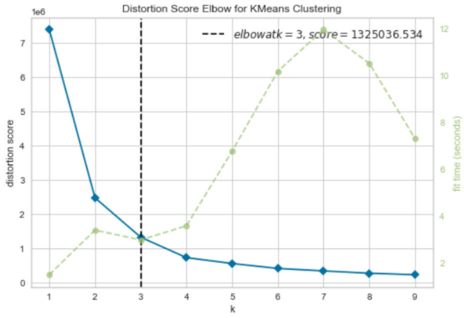

Fig. 2. Kelbow Visualization Results

TABLE I
Clustered Bikes

Cluster	Bikes
0	3370
1	1942
2	677

C. Proposed System

The motivation behind proposing this system is to help the users in choosing a bike best fit for their travel. The proposed bike recommendation system accepts user's travel needs like desired trip duration and trip distance and the prevailing weather conditions as inputs and recommends bikes accordingly. Based on these inputs, the system predicts the best bike cluster for the trip. We further propose a bike usage optimization strategy, which recommends the top 3 active bikes among the available bikes at the time of rent. From the trip records, we calculated the number of check-ins and check-outs of each bike id per station. Thus for each station, we got the list of bikes along with their number of uses by adding the no. of

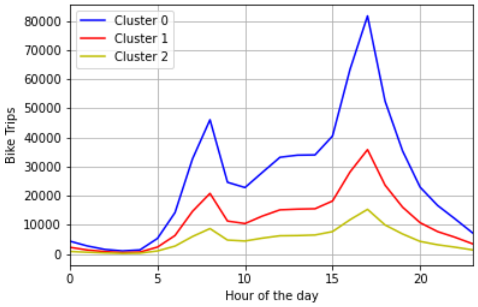

Fig. 3. Hourly distribution of trips for clusters

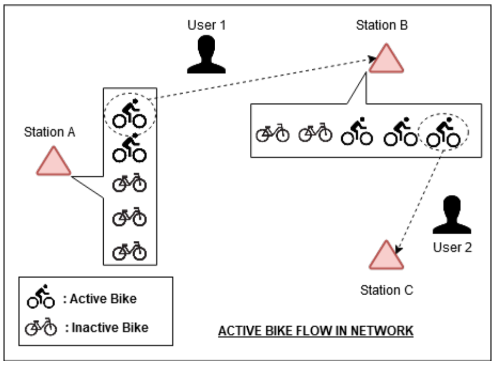

Fig. 4. Visualization of flow of active bikes in the network

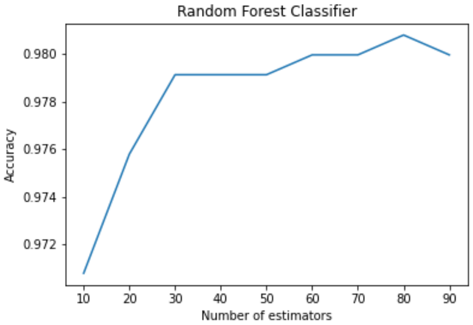

Fig. 5. Random Forest: Accuracy vs Estimators

check-ins and check-outs for that station. After predicting the cluster, the system finds the number of bikes belonging to that cluster in that station. Then, using the bike usage optimization strategy, the system recommends top 3 most used bikes. We have set the recommendation number to 3, but it can be modified based on bike sharing company's needs. This strategy increases the use of active bikes in the bike flow network as that improves the chances of another user choosing that bike for his/her journey and this effectively improves bike circulation.

The robustness of this system is explained below:

1) The system uses a probabilistic distribution to predict the best cluster out of the three. If no bike belonging to the predicted cluster is available at the start station, the system selects the next best cluster and finds the number of bikes belonging to that cluster. Thus, even in such an extreme case, user gets a list of recommended bikes.
2) If there are more than 3 bikes belonging to the predicted cluster in that station, then the system recommends top 3 bikes. If the bikes are less than 3, then the system recommends all those bikes.
3) This effectively reduces the problem of uncertainty whether the selected bike is best fit for user's travel needs. Thus, this system recommends 3 bikes, giving user a choice instead of allocating a particular bike. This further helps the user consider some other factors like physical condition of bike, condition of tyres etc. as all the recommended bikes are best fit for user's travel needs for the prevailing weather conditions.

The training approaches followed are described in the next section.

IV. Experimental Setup

The newly curated dataset was ready after extracting the bike features from the original trip records. This dataset contained mean of features from trip records for every bike id. The cluster in which the bike belonged to was added as another attribute in this dataset. Since, bike id is not the input to the system, it was then dropped from the data. Table II describes the inputs and output for the system. In this study, major libraries used were keras, pandas, numpy and scikit-learn. 20+ hours of experiments were performed by using the proposed training methods by changing their parameters to observe which parameter yielded us best results. Since this problem was a classification problem, the clusters were converted into categorical variables using one hot label encoding. We used three machine learning approaches for classifying the bikes into one of 3 clusters. We used a neural network approach, a random forest classifier model and KNN classifier model. The training approaches are explained in the next section. We designed a small neural network model consisting of 4 hidden layers and an ouput layer of 3 neurons with softmax activation function. The neural network model was trained for 300 epochs with a learning rate of 0.0001. A random forest classifier was the second training approach followed. Number of estimators is an important parameter governing the performance when using random forest classifier. Fig. 5 plots the classification accuracy for different number of estimators for random forest model. We got the best results for estimators=80. The third training method used was a K-Nearest Neighbours(KNN) classifier. Fig. 6 describes the classification accuracy for different number of neighbours. For KNN classifier, we got the best results for n=23. Table III describes the parameters used in the training approaches followed. Experimental results from all training approaches are described in the next section.

TABLE II
System Features

Type	Attributes
Input Features	Trip Distance Trip Duration Temperature Humidity Rain Intensity Wind Speed
Output	Cluster

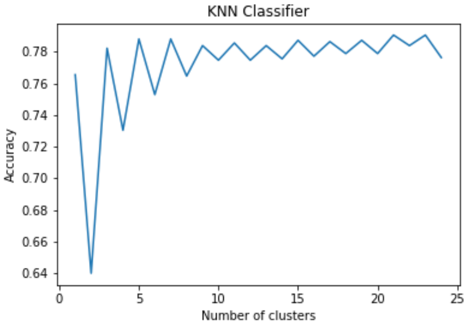

Fig. 6. KNN: Accuracy vs No. of Neighbours

TABLE III
Parameters used

Training Method	Parameter	Value
Neural Network Approach	Hidden Layers Optimizer Learning Rate Activation Epochs	4 Adam 0.0001 Softmax 300
Random Forest Classifier	Estimators	80
KNN Classifier	Neighbours(n)	23

V. Experimental Results

In this study, we proposed a novel bike recommendation system. Further, we used three machine learning methods as training approaches. The modified dataset contained 5989 records belonging to three clusters. For the neural network approach, the dataset was initially split into Train and Test set at 80:20 ratio. Then 20% of Train set was further split into Validation set. The final data split was:

- Train: 3832 records
- Validation: 959 records
- Test: 1198 records

For the other two methods, dataset was split into Train and Test in the ratio 80:20. Random forest classifier gave us an accuracy of 98.08% whereas KNN classifier resulted in an accuracy of 79.04%. By using the proposed neural network model, we got training accuracy of 96.32%, validation accuracy of 96.98% and testing accuracy of 96.49%. Fig. 7 plots the accuracy graphs and the confusion matrix for the classifications in the test set is plotted in Fig. 8. Table IV describes training, validation and testing results obtained using neural network approach and Table V summarizes other classification metrics like F1 Score, AUC Score, Recall and Precision. Table VI compares the results obtained by the three methods.

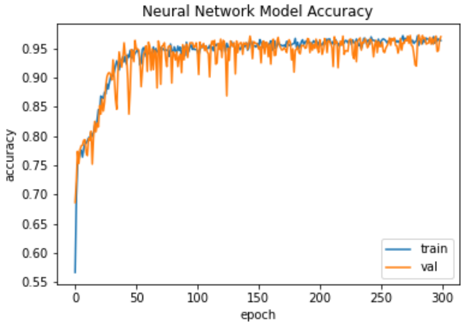

Fig. 7. Accuracy Graph using Neural Network approach

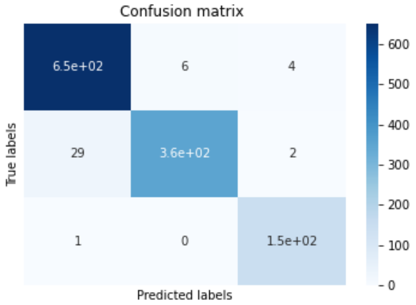

Fig. 8. Confusion Matrix for Neural Network Model classifications

TABLE IV
Results using Neural Network Approach

Phase	Accuracy	Loss
Training	96.32	0.0845
Validation	96.98	0.0764
Testing	96.49	0.0790

TABLE V
Evaluation Metrics

Metrics	Values
F1 Score	96.47
AUC Score	99.75
Recall	96.60
Precision	96.68

TABLE VI
Comparison of Results

Training Method	Accuracy	Loss
Neural Network	96.49	0.0790
Random Forest Classifier	98.08	-
KNN Classifier	79.04	-

VI. Conclusion

Bike sharing systems have gained a lot of popularity in the recent years. Users choose any one of the available bikes at the start station and begin their journey. But many times, users feel uncomfortable with the chosen bike due to many factors ranging from some physical issues with the bike to uncertainty if the chosen bike is best fit for the desired travel duration and travel distance in the weather conditions at the time of renting the bike. In such cases, users come back to the start station immediately, return the bike and choose another bike. This wastes the user's time and adds to the negative feedback of the user, proving detrimental for the bike company's reputation. At present, there is no system to help the user select a bike based on his/her travel demands. To this end, our proposed bike recommendation system solves this issue to a great extent, by recommending top 3 bikes by taking the user's travel needs and prevailing weather conditions into consideration, thus helping the user in choosing a bike. Also, the bike usage optimization strategy used in the system improves the circulation of active bikes and maintains the bike flow between different stations. The near accurate predictions of the training approaches followed prove the robustness of our system, thus making it fit for real time deployment.

References

[1] "The Bike-sharing Blog", http://bikesharing.blogspot.com/, May 2018.

[2] Marshall, Aarian, "Americans Are Falling in Love With Bike Share", May 2018.

[3] F. Huang, S. Qiao, J. Peng and B. Guo,"A Bimodal Gaussian Inhomogeneous Poisson Algorithm for Bike Number Prediction in a Bike-Sharing System," IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, August 2019, vol. 20, no. 8, pp. 2848-2857.

[4] X. Yao, X. Shen, T. He and S.H. Son,"Demand Estimation of Public Bike-Sharing System Based on Temporal and Spatial Correlation," 4th International Conference on Big Data Computing and Communications, August 2018, pp. 60-65.

[5] M.H. Almanna, M. Elhenawy, F. Guo and H.A. Rakha,"Incremental Learning Models of Bike Counts at Bike Sharing Systems," 2018 21st International Conference on Intelligent Transportation Systems (ITSC) Maui, Hawaii, USA, November 4-7, 2018, pp. 3712-3717.

[6] D. Li and Y. Zhao, "A Multi-Categorical Probabilistic Approach for Short-Term Bike Sharing Usage Prediction, " IEEE Access Digital Object Identifier, June 2019, 10.1109/ACCESS.2019.2923766.

[7] J. Huang, X. Wang and H. Sun, "Central Station Based Demand Prediction in a Bike Sharing System," IEEE International Conference on Mobile Data Management (MDM), 10-13 June 2019, 10.1109/MDM.2019.00-38.

[8] D. Tomaras, I. Boutsis and V. Kalogeraki, "Modeling and Predicting Bike Demand in Large City Situations," IEEE International Conference on Pervasive Computing and Communications (PerCom), March 2018, 10.1109/PERCOM.2018.8444588, pp. 1-10.

[9] J. Zhang and Y. Pang, "Planning Station Capacity and Bike Rebalance Based on Visual Analytics of Taxi and Bike-Sharing Data," International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, October 2018, 10.1109/CyberC.2018.00061, pp. 305-3054.

[10] I.L. Wang and C.W. Wang, "Analyzing Bike Repositioning Strategies based on Simulations for Public Bike Sharing Systems," Second IIAI International Conference on Advanced Applied Informatics, Aug 2018, 10.1109/IIAI-AAI.2013.9, pp. 306-311.

[11] X. Yao, X. Shen, L. Wang and T. He, "Hybrid Bicycle Allocation for Usage Load Balancing and Lifetime Optimization in Bike-Sharing Systems," IEEE 18th International Conference on Mobile Data Management, Daejeon, 2017, 10.1109/MDM.2017.24, pp. 112-117.

[12] H. Xu and J. Ying, "An improved GRASP for the bike-sharing rebalancing problem," International Conference on Smart Grid and Electrical Automation, Changsha, 2017, 10.1109/ICSGEA.2017.117, pp. 324-328.

[13] L. Chen and J. Jakubowicz, "Inferring Bike Trip Patterns from Bike Sharing System Open Data," IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, 10.1109/BigData.2015.7364115, pp. 2898-2900.

[14] Z. Li, J. Zhang, J. Gan, P. Lu and F. Lin, "Large-Scale Trip Planning for Bike-Sharing Systems," IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems, Orlando, FL, 2017, 10.1109/MASS.2017.36, pp. 328-332.

[15] Y. Zhou and Y. Huang, "Context Aware Flow Prediction of Bike Sharing Systems" IEEE International Conference on Big Data (Big Data),Seattle, WA, USA, 2018, 10.1109/BigData.2018.8621918, pp. 2393-2402.

[16] D. Ljubenkov, F. Kon and C. Ratti, "Optimizing Bike Sharing System Flows Using Graph Mining, Convolutional and Recurrent Neural," European Technology and Engineering Management Summit (E-TEMS), 10.1109/E-TEMS46250.2020.9111707.

[17] Y. Guo, X. Shen, Q. Ge and L. Wangi, "Station Function Discovery: Exploring Trip Records in Urban Public Bike-Sharing System," IEEE Access (Volume: 6), 10.1109ACCESS.2018.2878857, pp. 71060 - 71068.

[18] Divvy Data: https://www.divvybikes.com/system-data

[19] Weather Dataset: https://data.cityofchicago.org/Parks-Recreation/Beach-Weather-Stations-Automated-Sensors/k7hf-8y75