ECO-PEDALS - The Public Bicycle Sharing Management and Maintenance System

Mayank Jain[1], Shreya Bishnoi[2], Kritika Dutta[3], Shreya Mittal[4], Sanjay Kumar Sonker[5]

[1-4]Student, Department of Computer Science Engineering MIET, Meerut, UP, India
[5]Assistant Professor, Department of Computer Science Engineering MIET, Meerut, UP, India

---

Abstract - *The Objective of our proposed idea is to make the existing "Bicycle Sharing System" smart and efficient by assisting the management with indications representing the demand of bicycles on a particular day considering some crucial factors (weather and holidays).It will help the management to cope up with fluctuating demands. Moreover, it will also highlight the bicycles which require maintenance so that the operations of the system remain smooth and cost effective for the management and ensure that good working cycles are always available. Thus, our proposed system will add two extra features to the prevailing system adding value to it. It will eliminate the loop holes of existing system and will take it two steps ahead which will result in attaining higher customer satisfaction by ensuring availability of the bicycle at the station all the time. Many Bicycle sharing systems are already successfully operating in many countries including India and our effort is to make it more reliable and promising. We can categorize the functionalities of this system into two:*

1. The Management

2. The User (Rider)

Our proposed solution is to make the working of management better and easy while the functionalities of the user remain the same. Management is provided with a website to control its operations and maintain and view the record of users. Now, after changes proposed by us, there will be two add-on Options on that website –

1. Demand

2. Maintenance.

The workings of these are explained further in the thesis.

Keywords: Analysis, Bicycle sharing system, classification, demand, machine learning, prediction, regression.

1. Introduction:

1.1 About Public bicycle sharing system

Public Bicycle sharing [PBS] system is not a new concept; in fact, it is already in execution in many countries [1] like US, Germany, Spain, India etc. In India, cities like Bhopal, Mumbai, Mysore, Ahmedabad and many more are successfully running this system (Namma cycle system in Bengaluru, Cycle chalao in Pune are few examples). In the rush of today's modern lifestyle, problems like traffic jam, health issues, and pollution have hiked a lot so, in effort to control this, the Public Bicycle Sharing System was recognized as a neat and most effective solution. It was launched in many countries round the world. In India it was first launched in Mysore in 2009(India).Bicycles were considered as the best solution to all these problems as it is a sustainable, low maintenance form of commute which is not only cost effective but also promotes health benefits (energization, increase in stamina, fat reduction, leads to healthy heart, stress reduction ,etc.),it is not at all dependent on fossil fuels and can bring a major impact in controlling pollution and making roads traffic free .So, Entrepreneurs world-wide recognized the potential of bicycles as future changing component and implemented it successfully in many parts of the world.(china, Italy ,United States, Germany, Spain being the top 5 leading countries).

1.2 Existing system:

The implementation of existing system includes creation of several stations across the city at considerable distance with bicycles available at each station. A user can pick up a bicycle from any station after payment and take a ride and drop the cycle at any station close to the destination of the user. An app or website is provided to the user from where the user can view the availability of the bicycle and can view the list of all the stations in that city after creating an account. Tracking system is also implemented for real time cycle tracking for security and customer assistance.

The Website accepts two logins-**Management** and **User**. Their functions are as follows:

Management: it can view the no. of cycles at a particular station, can view the details and payments made by customers and track their current location during the ride.

User: User can sign up by creating account and login using username and password after which he/she can view the no. of cycles available at a station and location of all stations in the city. He can also make payment.

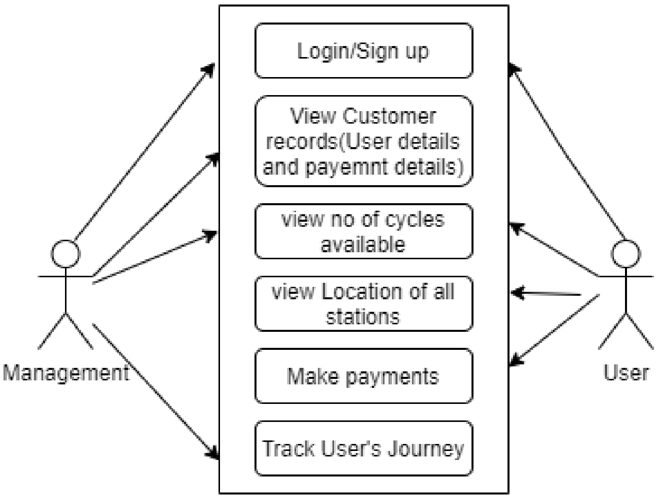
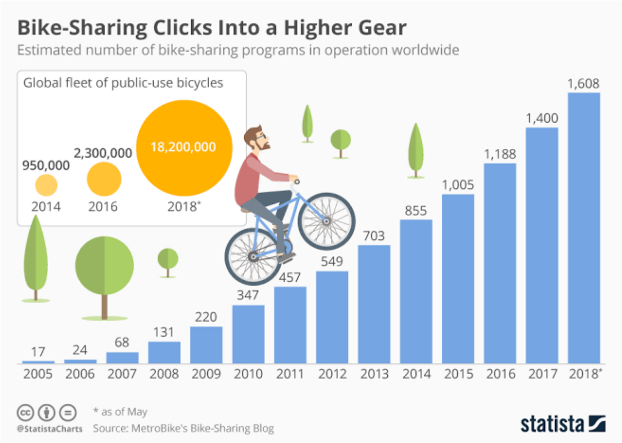

Login/Sign up
View Customer records(User details and payemnt details)
view no of cycles available
view Location of all stations
Make payments
Track User's Journey
Management
User

Fig 1.2.1: USE CASE OF EXISTING SYSTEM

Initially there were many obstacles faced like consideration of cycle as poor peoples' ride, no proper system etc. but as a result of awareness campaigns and change in the mind set of people, this system became more popular.

According To MetroBike's bike sharing blog of statistic as of 3rd June, 2018,[2] the increase in use of PBS is shown below shown in fig.1.2.2:


Bike-Sharing Clicks Into a Higher Gear
Estimated number of bike-sharing programs in operation worldwide

* as of May
Source: MetroBike's Bike-Sharing Blog

Fig 1.2.2 showing the increase in PBS worldwide

Fig 1.2.2 shows that public bicycles sharing system has a promising future and is a business worth investing.

1.3 Proposed system:

Our proposed system is regarding addition of new functionalities on management side while keeping the functions of the user same.

1. There are frequent fluctuations in the demand of bicycles due to several factors which can't be evaluated by management accurately thus, affecting the revenue. Sometimes the demand is high whereas the bicycles available at station are less and vice versa. Demand of the bicycles depend on some factors like weather conditions(inversely proportional) and holidays(directly proportional).By predicting the demand of bicycles on a station with the help of these factors, we can help the management to maintain required amount of cycles so that the customer never faces problem of non-availability. For this, a day will be divided into slots and demand for a particular slot will be predicted and shown to the management for each slot.

2. The second modification will be regarding the identification of cycles which require maintenance and not used by customers due to bad condition. The reason may be an old body, bad appearance etc. We can easily recognize such cycles by considering factors like-*How frequently the cycle is chosen by user*, and *the distance covered by that cycle* (A cycle which covered a particular large distance may need maintenance).

So, the use case of management after modifications will be:

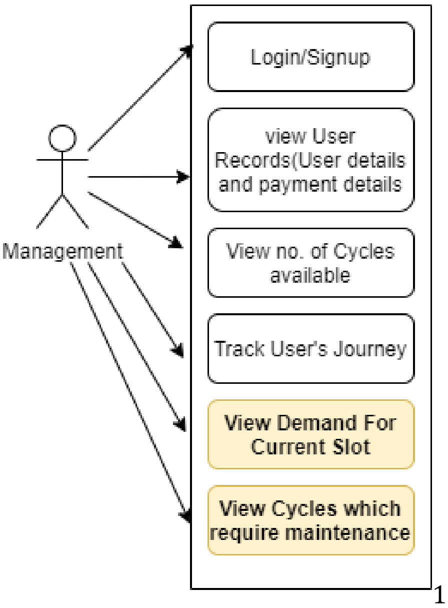

Login/Signup
view User Records(User details and payment details
View no. of Cycles available
Track User's Journey
View Demand For Current Slot
View Cycles which require maintenance
Management
1

Fig: 1.3.1: USE CASE OF MODIFIED SYSTEM (MANAGENENT SIDE).

2. Description and Implementation of Modules:

We are using NEW YORK CITI BIKE data set (2014 for the month of January) for analysis which has 30041 records in total. [3]

We have created a website for Management showing the 2 add on features which we have proposed. The predictions and identification of cycles done in modules is done by using machine learning using python.

Our proposed system has two modules which are described below:

2.1 Module 1: Predicting demand of the Bicycles:

Description: In this module, we have used supervised learning approach of machine learning to predict the no. of cycles (demand) on the current slot at the station. A day is divided into 4 quarters/slots, each of 6 hours. The factors considered for prediction are-Weather conditions, weekends and Holiday. We have provided the mentioned data set as training data which is 80% with labeled output to prepare the model and used 20% data to test the model.

The prediction will be shown to the management through the website under DEMAND option.

Multivariate Regression Algorithm:

In this module, we have used multiple linear regression algorithm to determine the demand of the bicycles that are needed in a particular slot at the station.

As the name suggests, multivariate regression algorithm has one dependent variable, which is to be calculated and has multiple independent variables (on which the dependent variable depends).

We have performed this analysis on the data that is provided by Citi Bike Ride data [3] which is a bicycle sharing management system in New York. The schema of raw data is shown in Table 2.1.1.

Table 2.1.1- Schema of raw data.

Tripduration	Int64
Starttime	Object
Stoptime	Object
Start station id	Int64
Start station name	Object
Start station latitude	Float64
Start station longitude	Float64
End station id	Int64
End station name	Object
End station latitude	Float64
End station longitude	Float64
Bikeid	Int64
Usertype	Object
Birth year	Object
Gender	Int64

The first step that is required in any machine learning algorithm is to perform data analysis and to understand the structure and behavior of the data. We have used Python NumPy and Pandas library for this. After that, the next thing is to clean the data and remove extra features that are not required in our model. After preprocessing the data, our schema of data is shown in Table 2.1.2.

Table- 2.1.2- Schema of data after preprocessing.

Start_station_id	Int64
Start_station_name	Object
Start_Date	Object
Start_time_slot	Int64
Holiday_parameter	Int64
Number_of_bicycles	Int64

The next step was to select the dependent and independent variables by understanding the trend of the data. In this, the basic requirement is to calculate the correlation between the different variables of the data given. Let two variables x and y, the correlation between them is C that is:

$$C = \frac{\sigma_{xy}}{\sigma_x * \sigma_y}$$

Here σ_{xy} is the covariance between the two and σ_x and σ_y are the standard deviation of the variables x and y. The one variable which have largest correlation with other variables is chosen as dependent variable. Python Seaborn Heatmap is used to find out the relation between variables, that is:

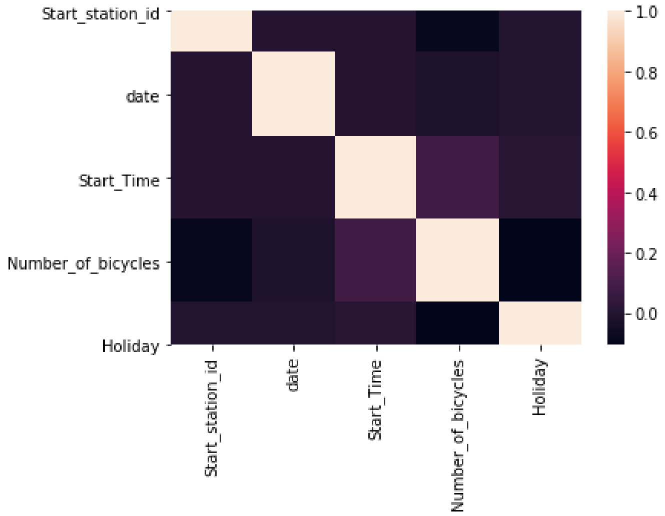


Fig: 2.1.1- Heatmap of various attributes.

And in our case, it is the 'number of bicycles'.

To determine the independent variables, we checked the scatterplots between the different attributes and dependent variable.

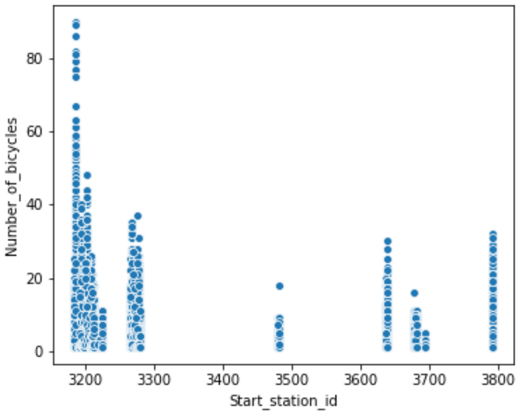


Fig: 2.1.2- Scatterplot between station id and number of bicycles.

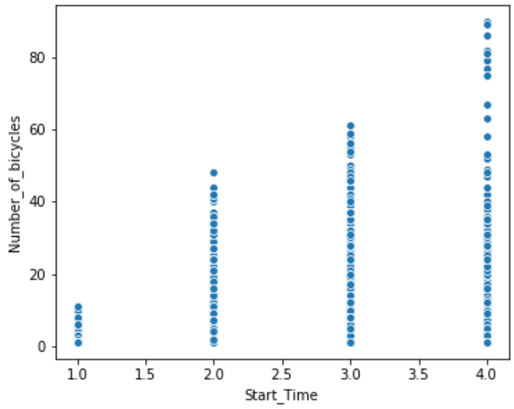


Fig: 2.1.3- Scatterplot between start time and number of bicycles.

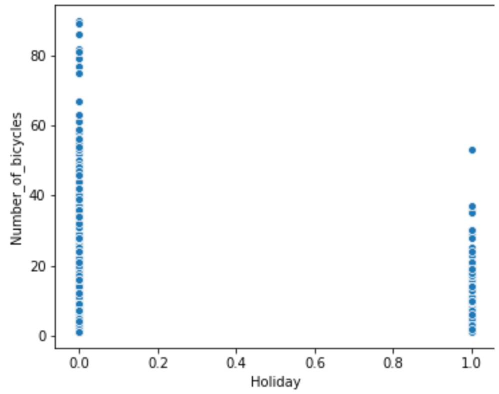


Fig: 2.1.4- Scatterplot between holiday and number of bicycles.

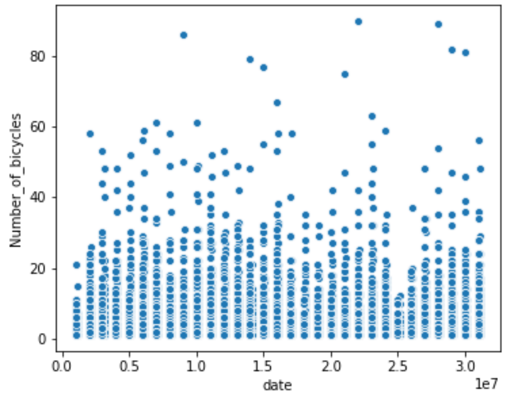


Fig: 2.1.5- Scatterplot between date and number of bicycles.

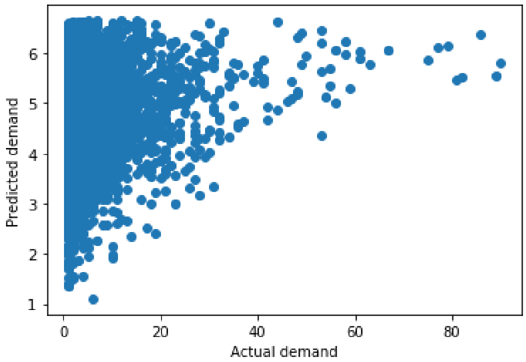
As there is no specific relationship between the variables, Start_station_id, date, Start_Time and Holiday are used as independent variables.

Further multiple linear regression is used on the above selected variables and tries to fit a linear equation to the data. The equation for multiple linear regression used is:

$$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + \ldots + m_n x_n + c$$

Here x_1 to x_n be the independent variables, m_1 to m_n be the respective slopes and c is a constant.

Using the above procedure, we have predicted the demand of the bicycles. Along with the above mathematical process, one important step is to divide the data into train and test datasets. Train data will first train the model according to the variables selected that how they are related and how prediction can be done. Test data is used to test the model that is created at the end. Graph between the actual and predicted data is:



4

Fig: 2.1.6- Plot between actual demand and predicted demand.

There are errors in the graph and the accuracy can be seen by calculating the value of errors in the predicted demand. We have used Mean squared error (MSE) to calculate the value of error by using:

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}(predicted\_value\ - true\_value)^2$$

Here, n is the total number of predictions done which is 4479 in our case. Using the above procedure, we have got the value of error approximately 59 which shows that our predictions are not exact but near to accurate result.

2.2 Module 2: IDENTIFICATION OF CYCLES WHICH REQUIRE MAINTENEACE

DESCRIPTION: In this module we have determined the bicycles that need maintenance. Bicycle id is used as a unique feature of each bicycle. The maintenance will be dependent on factors like how frequently the bicycle is chosen by the customer and how much distance is covered by the bicycle until the last maintenance. We have used our data and performed Naïve Bayes Classification on it to find out those bicycles which require maintenance.

Naïve Bayes Classification:

In this module we have used Naïve Bayes Classification to classify those bicycles which require maintenance without taking the round trips to every station. Naïve Bayes Classifier provides the class as output to which the combination of attribute belongs to. The output of this algorithm will be the class to which a bicycle belongs depending on its features. There will be two classes as output-'YES' and 'NO', which describe whether a bicycle requires maintenance or not.

We have performed this analysis on the data that is provided by Citi Bike Ride data which is a bicycle sharing management system in New York. We have also used the monthly reports of Citi Bike Ride [4]. The schema of raw data is shown in table 2.2.1.

Table:2.2.1- Schema of raw data.

Tripduration	Int64
Starttime	Object
Stoptime	Object
Start station id	Int64
Start station name	Object
Start station latitude	Float64
Start station longitude	Float64
End station id	Int64
End station name	Object
End station latitude	Float64
End station longitude	Float64
Bikeid	Int64
Usertype	Object
Birth year	Object
Gender	Int64

The first step of any machine learning algorithm is to analyze and understand data. We have used Python NumPy and Pandas libraries for this step. The very next step is to preprocess data which includes data cleaning, data transformation, data reduction and many more. Extra attributes are removed like user type, birth year, etc. and some new attributes are added by normalizing the remaining attributes. Such as, we have found the distance between the different stations using the latitude and longitude of start station and end station and aggregation is performed using 'pandas Sql' to determine the distance covered by the bicycle till now. A new attribute is added named 'Distance_covered_in_km' which tells the distance covered by bicycle. The formula used to calculate the distance between two points using latitude and longitude is:

$$distance = 6371.01 * \cos^{-1}[\sin(latitude_1) * \sin(latitude_2) + \cos(latitude_1) * \cos(latitude_2) * \cos(longitude_1 - longitude_2)]$$

After preprocessing, schema of the data is shown in table 2.2.2.

Table:2.2.2- Schema of data after preprocessing.

Bike_id	Int64
Total_number_of_trips	Int64
Total_distance_covered	Int64
Class	String

The attribute 'class' can have two entries either 'YES' when the bicycle require maintenance or 'NO' when they do not require.

Further Naïve Bayes Classifier [5] is used on the above variables. Bayes' Theorem is used to predict the probabilities of different classes. The formula of Bayes' theorem is:

$$P\left(\frac{A}{B}\right)=\frac{P\left(\frac{B}{A}\right).P(A)}{P(B)}$$

Here, A&B are events,

P (A/B) =Probability of A given B is true,

P (B/A) = Probability of B given A is true,

P (A) = Independent Probability of A,

P (B) = Independent Probability of B.

We have used 'scikit learn' library of python. Using the above procedure, we have predicted the bicycles which require maintenance. Along with the above mathematical process, one important step is to divide the data into train and test datasets. Train data will first train the model according to the variables selected that how they are related and how prediction can be done. Test data is used to test the model that is created at the end. We have used grid to show the predicted probabilities of the various classes in our model:

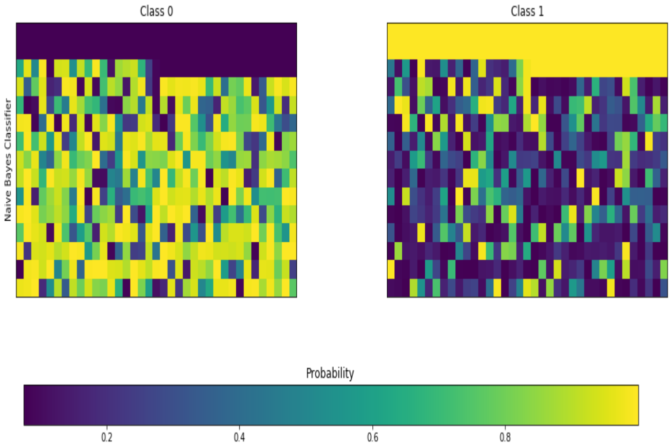


Fig: 2.2.1: Grid showing probabilities of different classes.

The accuracy of the model is an important factor which determines the success of the model. So we have measured the accuracy of the model using accuracy matrix which is the fraction of samples predicted correctly having formula: [6]

$$Accuracy_{Fraction\ predicted\ correctly} = \frac{TP + TN}{TP + TN + FP + FN}$$

Here TP= True Positive

TN=True Negative

FP= False Positive

FN=False Negative

We have obtained the accuracy score using scikit-learn, which takes the actual labels and the predicted labels as input. Our value of accuracy is 0.774 which shows that our model is quite good.

Conclusion: As we can easily analyze the growing market and scope of Public bicycle sharing system, we can predict that it is going to flourish in upcoming years. Therefore, continuous improvements and innovations must be made to make it more successful and customer oriented. With the effort we have made, not only the organization of stations will be improved but it will also enhance the revenues and will prove to be a more reliable system. Currently, the managers can not accurately predict the demand and their manual predictions are not up to mark, our solution will provide them with reliable predictions based on previous scenarios. It will help them to effectively manage the cycles in station-both in amount of cycles and regarding maintenance as well. This will result in longer life of cycles and the factors which earlier used to create unpredictable situations will now be converted into a utility. Thus, manual analysis and calculations will be converted into automated predictions thus, making changes in the system for sure.

3. References

1. Napoli, B. S. (2013, Dec 7). "Bike share boom: 7 cities doing it right". Retrieved from http://www.bikesharingnapoli.it/best-practices/

2. Ritcher, F. (2018, July 3). “Bike-Sharing Clicks Into Higher Gear”. Retrieved from https://www.statista.com/chart/14542/bike-sharing-programs-worldwide/

3. Motivate. (2017). "System Data". Retrieved from https://www.citibikenyc.com/system-data

4. Motivate. (n.d.). Citi Bike Monthly Operating Reports. Retrieved from https://www.citibikenyc.com/system-data/operating-reports

5. Gandhi, R. (2018, May 5). Naive Bayes Classifier. Retrieved from https://towardsdatascience.com/naive-bayes-classifier-81d512f50a7c

6. "Accuracy, Precision, Recall or F1?". (n.d.). Retrieved from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9