IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 4, April 2015.
www.ijiset.com
ISSN 2348 – 7968

Bike Share Demand Prediction using RandomForests

Akshay Patil[1], Karishma Musale[2] and BVANSS Prabhakar Rao[3]

[1] School of Computer Science and Engineering, VIT University,
Chennai Campus, Tamil Nadu - 600127, India.

[2]Department of Computer Engineering, SVIT Chincholi,
Nashik, Maharashtra - 422102, India.

[3]School of Computer Science and Engineering, VIT University,
Chennai Campus, Tamil Nadu, 600127, India.

***Abstract*—Bike Sharing System is an emerging mode of transport in the world and most of the developing countries are on the path of following the western model of Bike Sharing Systems.In India some entrepreneurs have tried to setup a bike share system and have failed in the past as they have failed to use data analytics properly. There is a possibility that bike stations can be full or empty when a traveler comes to the station. Thus to predict the use of such system can be helpful for the users to plan their travels and also for the entrepreneurs to setup the system properly. This paper presents different ways to predict the number of bikes that can be rented in such a system, for case study purpose we have used a public data set. The predictions are made for every hour of a day.**

Keywords—bike sharing system, randomforests,prediction

I. INTRODUCTION

Bike sharing systems allow the users to take one way bi-cycle trips over short distances.Generally these systems are operated via automated kiosks to save manpower and reduce waiting time for the users. Bike Sharing System ensures that pollution is reduced as with use of bicycles there is reduction in use of motor vehicles which leads to reduction in emission of pollutants in the air. This practice of Bike Sharing Systems is common in Western Countries while the same is not seen yet in countries like India.In India most of the bike sharing systems could not achieve their maximum potential as data analysis was not used properly. The advantages of this system is that we can have public bike stations without any human involvement. Even local Chennai Municipal Corporation has invited biddings for a new bicycle sharing system in Chennai which will be likely be operational by 2016.

II. PROBLEM DEFINITION

Generally in bicycle sharing systems it is very important that the administrators should know how many cycles will be needed in each bicycle station, knowing this count enables them the arrange proper number of cycles at the stations and decide whether a particular station needs to have extra number of bicycle stands.So in this research work we study various prediction algorithms i.e. randomforests, decision trees, gradient boosting machines. This research work focuses on which algorithm can work better for the real world problem of bicycle sharing demand prediction.

III. RESEARCH SCOPE

Today in India, the number of vehicles is increasing day by day, with this number of increase in vehicles there is also increase in the CO2 emission in the atmosphere. To overcome this problem there is need of Bicycle sharing systems in India. With the emergence of bicycle sharing systems people will be encouraged to use bicycles for short distance travel. Due to this
traffic will be reduced and there will be less emission of harmful gasses in the atmosphere. Along with this there are also health benefits for the people using bicycles. People can stay healthy, a research has proved that cycling is a good exercise to prevent heart diseases. Cycling builds stamina, it increases cardio vascular fitness, burns calories and reduces stress.

Table 1. : Pollution Comparison in Indian Cities

City	Pollution load in metric tons per day	Average distance in km covered in 1 hour by car in traffic.	Average distance covered in 1hour by bicycle in traffic.
Delhi	421	15	12
Mumbai	189	12	11
Chennai	177	11	13
Kolkata	137	10	15
Bengaluru	207	18	15
Hyderabad	163	21	16

Source: Auto Fuel Policy Report 2013.

As seen in the above table, the pollution load in metro cities in India is quite high, this can be reduced by introduction of a healthier and greener way of transport in India. Cycle sharing systems will be useful to reduce this pollution. Also there is problem of traffic as people tend to use cars for traveling short distances. The distance travelled in peak traffic by car and a bicycle is nearly same. Hence we can say that India needs a bicycle sharing program to reduce the pollution and traffic.

ISSN 2348 – 7968

IV. Literature Review

A. *Existing Bike Sharing Systems*

Bike sharing is an emerging industry and it is very popular in western countries, while people have tried to start the same in India, we will look into some of the stats regarding how many people use bike sharing systems. According to Wikipedia by August 2014 only 600 cities in the world had bike sharing systems and most of them were in western countries with a fleet of about 500000 bicycles with them. There is a sharp increase in NextBike,Cogo BikeShare are some of the leading Bike Sharing systems that are currently in operation in the world.

While considering Indian perspective in the Bike Share industry, India has not yet adapted the application of this emerging industry. Currently there are a few bike share systems in India and still are running on test basis, some of them are:

i)NammaCycles:

This was started in Bengaluru in August 2012, as a IISC project and is still working efficiently. This system has 5 cycle stations and about 50 bicycles with them. Limitation with Namma Cycles is that it is practiced in a closed area i.e. a Campus area and Government is also participating in this initiative.

Fig.1: Bicycle Sharing Station

ii)CycleChalao

This project was started in Mumbai, and had 25 stations with a fleet of 300 bicycles, unfortunately this project was closed due to many reasons like no co-operation from the government, lack of use of analytics, less seed capital, failure to implement the model on a large scale etc.

B. *Current use of Analytics in Bike Share Systems.*

The current use of analytics in Bike Share Industry is encouraging, most of the companies are having their own analytics divisions. Before starting any new Bike Station the companies analyze how much the station will be useful and will it generate appropriate revenue and whether setting up the station is feasible or not. With enhancements in analytic techniques in current era , with some simple surveys Bike Share companies can forecast the use of the system and then plan accordingly.

C. *Methodologies used for Prediction.*

The existing methodologies for predictions are regression, decision trees, randomforests, svm, neural networks etc. This research work allows to have insight of performance of various prediction algorithms and walkthrough the whole process of prediction.

V. Methodology

Most the successful bike sharing systems use analytics to a great extent. The system architecture is as follows:

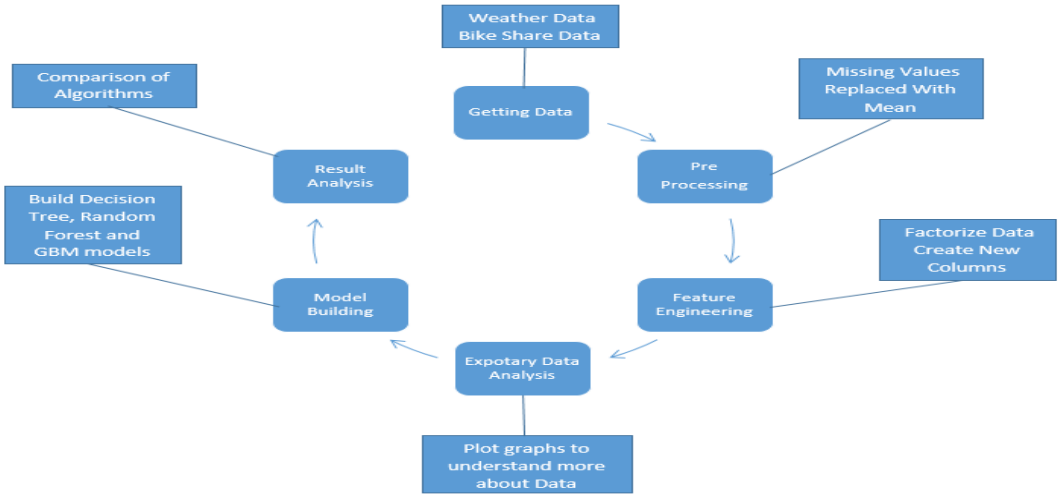

Fig.2 : System Architecture for Prediction Models

ISSN 2348 – 7968

A. *Getting the Data*

Getting data is an important factor in building a predictive model, in most of real time situations we cannot have the luxury of having fully structured data every time. In this research work, a public dataset provided by Capital Bike Share on the UCI Repository is used for model construction.

The data has following attributes and are explained in below table:

Table 2: Data Description

Features/Labels	Values
DateTime	hourly date + timestamp
Season	1 = spring, 2 = summer, 3 = fall, 4 = winter
Holiday	whether the day is considered a holiday
Working Day	whether the day is neither a weekend nor holiday
Weather	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
Temperature	temperature in Celsius
ATemp	"feels like" temperature in Celsius
Humidity	relative humidity
Windspeed	wind speed
Casual	number of non-registered user rentals initiated
Registered	number of registered user rentals initiated
Count	number of total rentals

B. *Pre-Processing*

There is a need of data pre-processing because the data may be incomplete or inconsistent or noisy. There are many ways to deal with un-processed data viz:

i)**Data Cleaning**: By this term we mean to fill the missing values in data, identifying and removing outliers in data, smoothningdata.

ii)**Data Transformation**: In this stage operations like normalization and aggrigation are performed.

iii)**Data Reduction**: In this stage the data set is modified such that the results produced by the model are almost the same but un neccesary values in dataset are removed.

iv)**Data Integration**: In this stage data is merged from different sources if needed , again redundancies are removed too.

C. *Feature Engineering:*

It is a process in which analysts use domain knowledge about the data and to create new features in the data set in a way such that the new features help in improving the model accuracy. There is no definite path for feature engineering , but it depends on the skills of the analyst and type of data. Feature engineering needs to be done on both training and testing data and is very important part of building a good prediction model.

In this dataset , we do following feature engineering:

i) Convert the data-time attribute in proper format and we separate day, month , year and hour into separate columns so that it is easy to perform operations on the data.

ii)Divide temprature, humidity and windspeed variables into categories. Doing so we can better accuracy in the model.

iii)Create dummy variables for season attribute , here season variable is broken down into 3 binary variables i.e spring, summer and winter.

iv) Normalize all the continuos variables in data set.

D. *Exploratory Data Analysis*

Exploratory data analysis is an statistical way of understanding the data which is usually done in a visual way.The graphs plotted in explotary data analysis are for better understanding of data to the analyst. For the current data set exploratory data analysis is done as follows:

i)Since we have to predict the number of bikes that will be rented, the best way to begin is with the variable to predict, "count". We can stratify the "count" distribution as boxplots for the categorical variables, and draw the "count" and numeric variables in another plot.

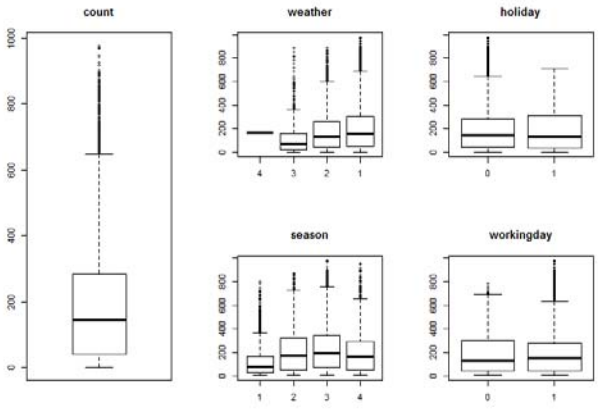

Fig.3: Exploratory Data Analysis for count variable in dataset.

ISSN 2348 – 7968

As seen in the boxplots above, count distribution is larger plot. The median count is about 150 units, and there are many outlier counts above 600. The range of counts is in between 0 to 1000units. When the weather is extreme the count i.e number of rented bicycles is less and otherwise its median count increases. There is not much difference other than the outliers.Median value increases in season summer and fall.

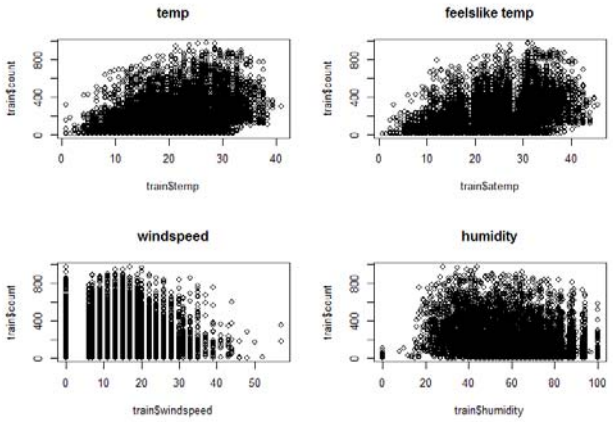

Fig 4. Exploratory Data Analysis for other variables.

Let's shift to the numeric variables. By taking a look at the distributions for the numeric variables we can conclude that the count is high when temperature is in between 25 and 30. Also the count is high when the windspeed is low. The count does not vary much for the humidity values.

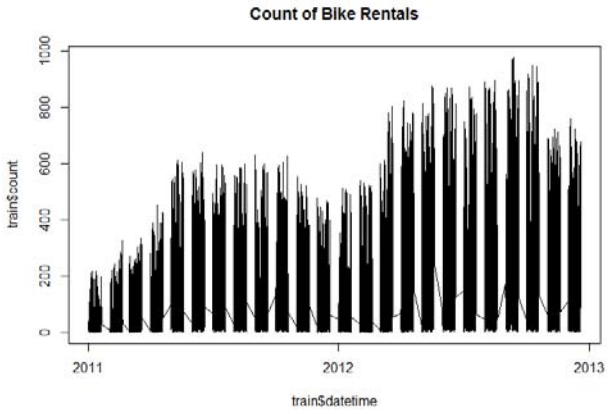

Fig.5: Number of Bicycles rented in each year.

Now from the plot of date-time vs. count we can conclude that the Capital Bike Share program became more popular as the years went by and we can see that the count increases in summer of each year.

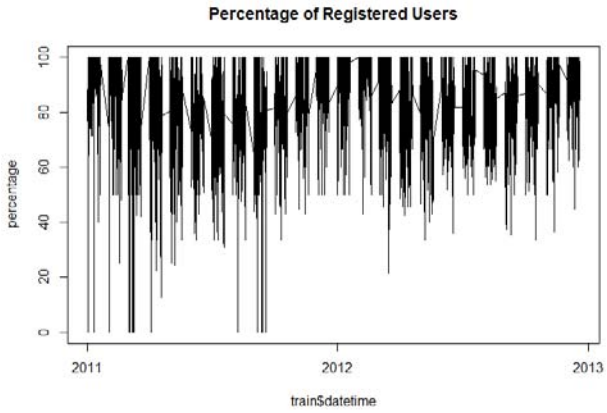

Fig.7: Registered users for Captital Bike Share System.

Now we can have the pot of percentage of registered and casual users, from this we can determine that number of casual users started decreasing as months passed by and mostly the users of Capital Bike Share are registered users.

E. Model Building

Hence as the data is understood properly, randomforest predictive model can be built for this data to predict the count variable.

i)Randomforest Model

```
fit <- randomForest(count ~ season + atemp + humidity +
                    workingday + hour + dayofweek + year,
                    data = df.train, ntree = 20000,
                    mtry = 6, importance = TRUE)
```

ii)Enhancing model using TuneRF

TuneRF is a function by which we can determine the optimal parameters that are needed to be passed to the randomforest algorithm.

```
bestmtry <- tuneRF(df.train_small, df.train$count,
                   ntreeTry = 100, stepFactor = 1.5, improve = 0.01,
                   trace = TRUE, plot = TRUE, dobest = TRUE)
```

VI. Advantages & Limitations

A. Advantages

i) The problem of empty bicycle stations i.e.when a user goes to a bicycle station and finds no bikes can be solved by already predicting how many bikes will be needed on a particular day.

ISSN 2348 – 7968

ii)The problem of full bicycle stations i.e. when a user goes to bicycle station and finds that there is no place to return the bicycle can be solved as with the predictions as if it known how many bikes will be needed each day then the bicycle redistribution can be planned accordingly.

B. Limitations

i) There are always some outliers in predictions, the system may fail in such case. Outliers can be like if a large group of people decide to travel from a same location via the bicycle sharing system on a random day then the predictions may fail for that day.
ii) Along with this the problem of thefts and misuse of the systems cannot be ignored. The administrators need to think of various innovative ways to keep the bicycles safe.
iii)The issue of safety of bicycle riders comes into consideration, bicycles should always be rode with a helmet on for the protection of the rider.

VII. Results Comparison

For comparison of the results given by various algorithms Root Mean Squared Logarithmic Error (RMSLE) is determined for the counts predicted. Is is given by:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(p_i+1)-\log(a_i+1))^2}$$

Where:

- *n* is the number of hours in the test set
- *pi* is your predicted count
- *ai* is the actual count
- log(*x*) is the natural logarithm

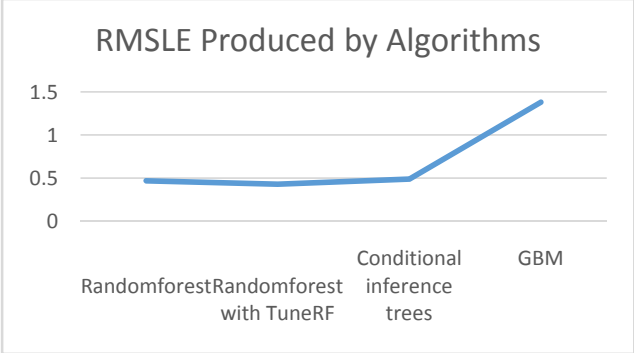

Fig.8: Comparison of RMSLE produced by Algorithms.

From the above mentioned results we can see that randomforest algorithm when combined with TuneRF function gives us better results.

VI. CONCLUSION

Bicycle sharing systems can be the new boom in India, with use of various prediction models the ease of operations will be increased.The four algorithms are applied on the bikeshare dataset for predicting the count of bicycles that will be rented per hour. We got some good results and accuracy with random forest and by using TuneRF function with the original randomforest algorithm.The accuracy and performance has been compared between the models using Root Mean Squared Logarithmic Error (RMSLE). If these systems include the use of analytics the probability of building a successful system will increase.

References

[1] K. Gebhert, R. Roland ,"Impact ofWeather on Capital Bike Share Trips,"presented at the 92nd Annual Meeting of the Transportation Research Board 2013.

[2] J. Yoon, Fabio Pinelli, "Cityride: a predictive bike sharing journey advisor"2012 IEEE 13th International Conference on Mobile Data Management.

[3] R. Giot, R.Cherrier"Predicting Bike Share Demand upto One hour ahead" 2013 IEEE 9th International Confrence on Data Management, France.

[4] I.Frade, A.Ribbero , "Bicycle Sharing Systems Demand" unpublished.

[5] T. Rui, Lin Li Hua"Quantitative Research on Vehicle Exhausts Pollution in the City"Published in 2012.

[6] Y. Zhang, Z Huang "Performance evaluation of bike sharing system in Wuchang area of Wuhan, China**" ,** 6th China Planning Conference (IACP), 2012

[7] Kaggle: Bike Share Demand ,"*https://www.**kaggle**.com/c/**bike-sharing**-demand"*

[8] Using Gradient Boosted Trees to Predict Bike Share Demand, "http://blog.dato.com/using-gradient-boosted-trees-to-predict-bike-sharing-demand"

[9] M. Meriem ,J. Usmail "A comparative study of predictive algorithms for time series forecasting ", 3rd International Conference on Information Science and Technology , 2014.

[10] C. Ramirez , N. Naggapan , "Studying the impact of evolution in R libraries on software engineering research " , 1st International Workshop on Software Analytics, 2015.

[11] D.J. Newman A. Asuncion. UCI machine learning repository, 2007.

[12] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques",second edition Morgan Kaufmann publisher.

[13] Ya Su, Xinbo Gao, Xuelong Li,and Dacheng Tao. "Multivariate Multilinear Regression",IEEE transactions on systems, man and cybernetics-Part B: cybernetics, vol 42.No.42 .

[14] Decision Trees for Business Intelligence and Data Mining: Using SAS Enterprise Miner.

[15] Simon Bernard, Laurent Heutte, Sebastian Adam. "On the selection of decision trees in Random Forests". International Joint Conference on Neural Networks IEEE, Jun 2009, France.

[16] Myungsook Klassen," Learning microarray cancer datasets by random forests and support vector machines",IEEE, 2010.

[17] Mohammed S. Alam and Son T. Vuong, "Random Forest Classification for Detecting Android Malware", 2013 IEEE International Conference

ISSN 2348 – 7968

on Green Computing and Communications and IEEE Internet of Things and IEEE Cyber, Physical and Social Computing.

[18] Vrushali Y Kulkarni and Dr Pradeep K Sinha, " Random Forest Classifiers :A Survey and Future Research Directions", International Journal of Advanced Computing, ISSN:2051-0845, Vol.36, Issue.1.

[19] Yasser Ganjisaffar, Rich Caruana, Cristina Videira Lopes, "Bagging Gradient-Boosted Trees for High Precision, Low Variance Ranking Models", ACM SIGIR'11, July 24–28, 2011, Beijing, China.

[20] Chun-Xia Zhang, Jiang-She Zhang, Gai-Ying Zhang, " An efficient modified boosting method for solving classification problems", Science Direct, An efficient modified boosting method for solving classification problems.

[21] Chun-Xia Zhang, Jiang-She Zhang, "A local boosting algorithm for solving classification problems", Science Direct, Computational Statistics & Data Analysis 52 (2008) 1928 – 1941.

[22] Uyen Nguyen Thi Van, and Tae Choong Chung, "An Efficient Decision Tree Construction for Large Datasets", IEEE journal 2008.

[23] Christopher M. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics).Springer-Verlag New York, Inc., Secaucus, NJ, USA,2006.

[24] Thangaparvathi.B and Anandhavalli.D, "An Improved Algorithm of Decision Tree for Classifying Large Data Set Based on RainForest Framework",IEEE journal 2010.