ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

Random Forest -A Machine Learning Algorithm to Predict Number of bicycles in Bicycle Rental System

Harish Manghnani[1], Minal Vanage[2], Akshata Naik[3], Shalini Wankhede[4.]

B.E. Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Kondhwa, Pune, India[1].

B.E. Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Kondhwa, Pune, India[2].

B.E. Student, Department of Computer Engineering, Sinhgad Academy of Engineering, Kondhwa, Pune, India[3].

Professor, Department of Computer Engineering,Sinhgad Academy of Engineering, Kondhwa, Pune, India[4].

ABSTRACT: The Bicycle Rental System is recognized by many countries and is a convenient way to travel small distances on the cost environment and personal health. Let it be Red Bike in China or Uber's new venture in San Francisco or OFO in India Bicycle Rental Systems are blooming and successful. But the Company owner needs to keep a count of number of bicycles on each station across city. This is mainly because of customer satisfaction, investment in number of bicycles and the response from clients and users. Thus,Random Forest comes into picture where a Machine Learning algorithm will help predict number of bicycles on a station/s in a Bicycle Rental System.

KEYWORDS: Machine Learning, Random Forest,Prediction,Bicycle Rental System,,R programming RMSLE**.**

I. INTRODUCTION

Bicycle is not only environment friendly but also motivates people for having a reason to exercise. For example, people can grab a bicycle to go to buy daily necessities or for a replacement of evening walk or even to travel small distances. This also helps resolve issues like last mile problem where people would need to walk a certain distance for 15 minutes whereas using a bicycle instead would hardly need 5 minutes.

But today customer satisfaction is given the top most priority. For instance there is a bicycle station but no bicycles are available or there are suppose 100 bicycles on a particular station but daily usage is not more than 30 bicycles then rest 70 are ideal which could be sent to some other station in city where there is shortage of bicycles. Here we can use the Random Forest algorithm to design a machine which can predict number of bicycles on a particular station. This also considers factors like weather, temperature, humidity, weekends/weekdays and peek time hours. Thus this machine would be learning on the basis of historic data and predict for future considering all affecting factors discussed above.

II. RELATED WORK

According to the literature survey conducted we have found various RMSLE values corresponding to various machine learning algorithms. These machine learning algorithms have been experimented on their designed model and then used various matrices to find the RMSLE values. Below table shows that Random Forest shows the least error and thus we have implemented that machine learning algorithm. It is a supervised machine learning algorithm based on decision trees and finds out prediction basis the training model and the testing data provided to the designed model.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

Sr.No	ALGORITHM	RMSLE
1	Random Forest	0.74
2	Decision Tree	0.76
3	Extreme Gradient Boosting	0.75
4	SVM	0.75
5	Linear and Bayesian classification	0.76

Fig.6. Research Analysis of RMSLE

III. OVERVIEW

A. Design Phase-

First we will design a Random Forest algorithm based machine. RMSLE (Root Mean Square Least Error) can be later found out. Later we will be dividing the dataset into training data and testing data. Next, we will be passing the training dataset to train the newly created machine which is nothing but a script in R programming and consider a learning rate to improve the accuracy of model. Finally, we will be passing the testing dataset toget the predictions and find out how accurate the designed model is.

B. Requirements-

The most important requirement will be a historic dataset which has all important factors like weather, atmospheric temperature, seasons, weekday/weekends etc.

For the coding part R programming is used and its libraries like rattle, rpart, rpart.plot and for representation anaconda jupyter notebooks have been used.

IV. ARCHITECTURE

In this architecture the affecting factors to the number of bicycles are windspeed, temperature, humidity, weekdays, weekends and seasons. This architecture basically defines how various factors affect the machine like for example people would not prefer cycling during rains or people would avoid cycling during afternoons. They would rather prefer winter mornings or late Sunday evenings. Similarly, windy weather is more preferable rather than a humid atmosphere. Considering all these factors we will design or in other words program a machine using Random Forest algorithm. This machine will be then trained on the basis of train.csv. finally, as a later input will be test.csv which will be predicting number of bicycles across each date and hour. The final output will be written back to submit.csv file which can be read and analysed. Random Forest is chosen because its one of the most stable machine learning and predicting algorithm where we can get outputs in form of graphs as well as a clear prediction which would be simple for analysis and data interpretation.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

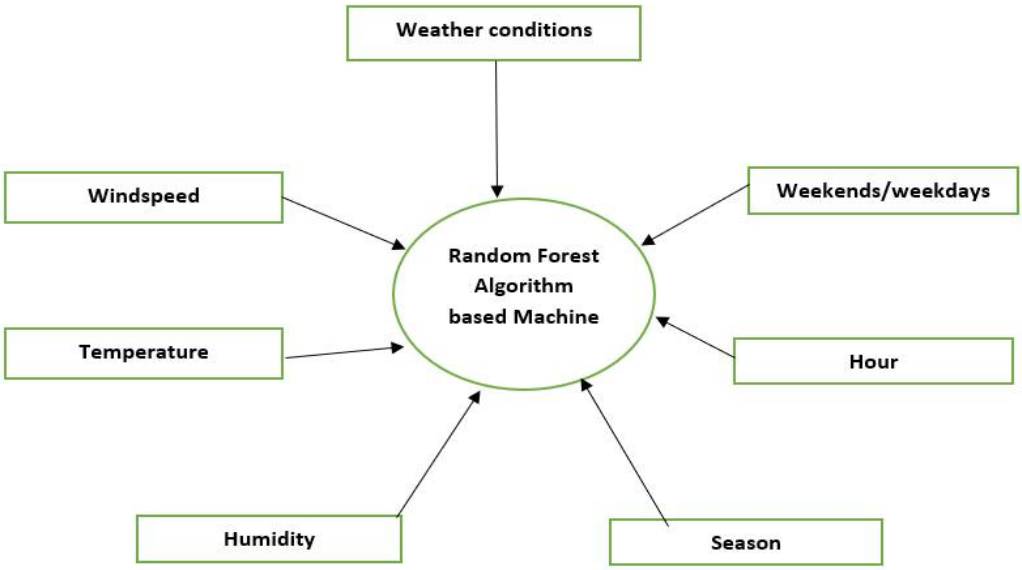

Fig.1. Architecture of the Machine

V. DATASET

The dataset is divided into training and testing dataset. The training will be done using train.csv and the testing will be done using test.csv. The below snapshot is of the test.csv file which has a excel sheet format with date, season, holiday, working day, temperature, atmospheric-temperature, humidity and windspeed. A csv file is a Comma Separated File which is in excel sheet format which has a past historic data base

datetime	season	holiday	workingda	weather	temp	atemp	humidity	windspeed
01-01-2011 00:00	1	0	0	1	9.84	14.395	81	0
01-01-2011 01:00	1	0	0	1	9.02	13.635	80	0
01-01-2011 02:00	1	0	0	1	9.02	13.635	80	0
01-01-2011 03:00	1	0	0	1	9.84	14.395	75	0
01-01-2011 04:00	1	0	0	1	9.84	14.395	75	0
01-01-2011 05:00	1	0	0	2	9.84	12.88	75	6.0032

Fig.2. Snapshot of the train.csv file

VI. DESIGNING OF MACHINE

First we will be factoring the data from numerical using command like data$season=as.factor(data$season). Similarly, we will be factoring all the factors of the training dataset. Later we will produce boxplots of factors like hour v/s number of users using command boxplot(train$registered~train$hour,xlab=”hour”, ylab=”registered users”). This command prints the boxplot of hour v/s registered users. The below boxplot shows basically the number of registered users in a particular hour across the day.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

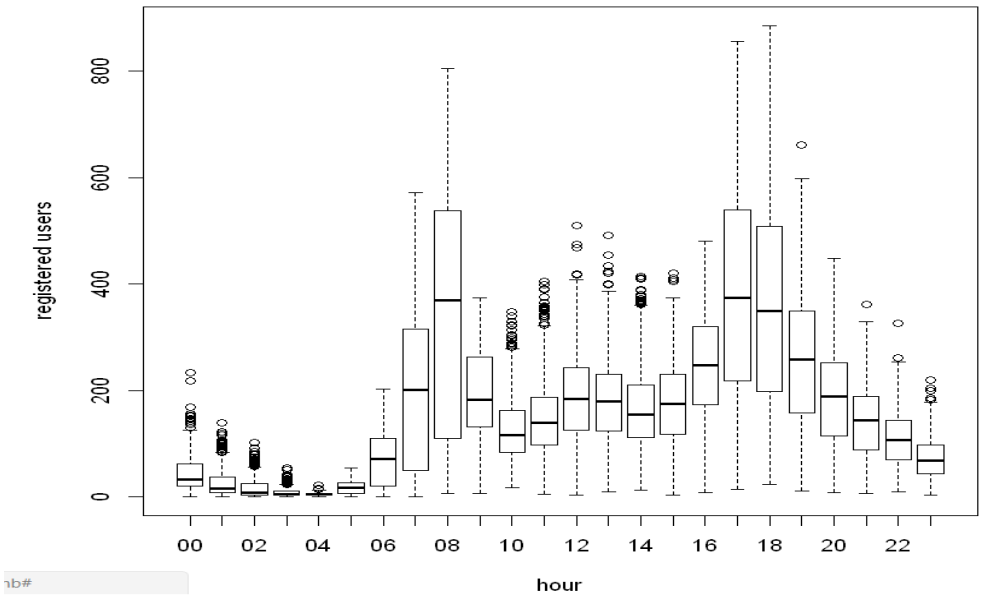

Fig.3. Boxplot of Hour v/s Registered Users

Later we use the commands like
d=rpart(registerd~hour,data=train)
fancyRpartPlot(d). this basically plots the decision tree helps easier to understand the further analysis.

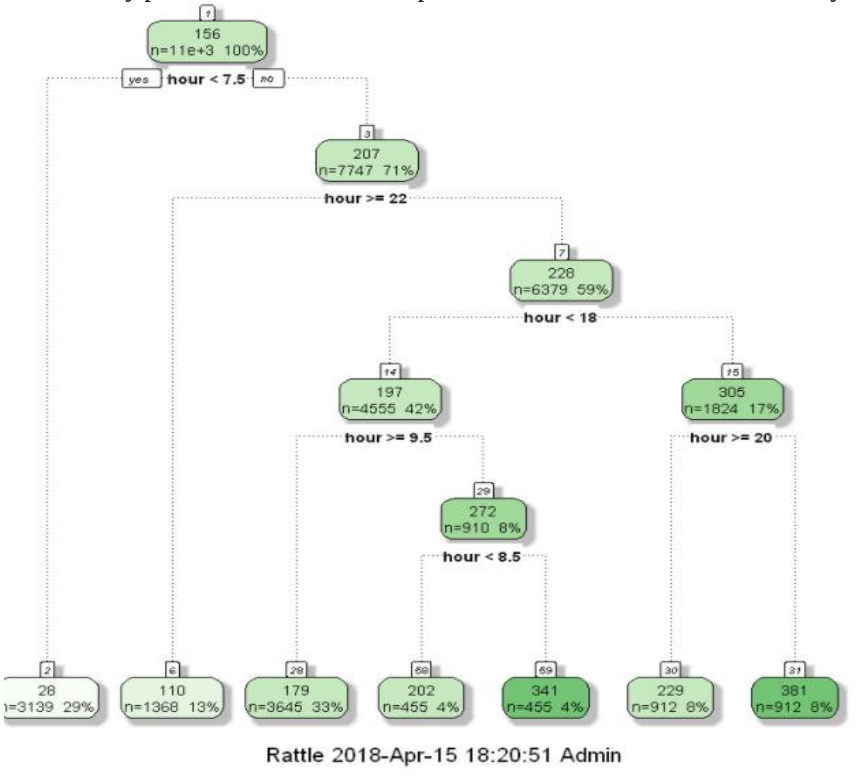

Fig.4. Decision Tree of the data

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

In the final step we pass the testing file to the trained machine(model) by using commands as

```
Set.seed(415)
Fit<- randomForest(factors)
Pred=predict(fit,test)
```

VII. RESULTS

Then finally write it to a submit file and the output will be as shown below a snapshot of submit.csv. the below csv file shows the prediction of a certain day and a particular hour on that day and thus the predicted number of bicycles.

A	B
datetime	count
20-01-2011 01:00	5.472396
20-01-2011 02:00	3.164197
20-01-2011 12:00	83.37935
22-01-2011 13:00	83.3239
22-01-2011 15:00	81.60309
22-01-2011 17:00	80.35441
22-01-2011 22:00	27.29189
24-01-2011 07:00	73.46448
24-01-2011 08:00	175.2444

Fig.5. Final output

VIII. FUTURE SCOPE

The Random Forest algorithm is not only the machine learning algorithm which is widely used but there are various machine learning algorithms like Artificial neural network (ANN), Support Vector Machine (SVM) or the simplest of them all the Linear Regression. The accuracy of the model and how the machine progression is , is found out using RMSLE(Root mean square least error). This error defines how accurate the predicted data is. Below is a table which shows various RMSLE of various machine learning algorithms on the basis of research.

Thus, this not only shows that Random Forest is one of the most stable machine learning algorithm but also flexible and its progression rate is high. The RMSLE functions works as such like
rmsle(actual,predicted). This gives a clear idea that how accurate our prediction is which internally compares the actual number of bicycles and the predicted ones.

For better analysis of data we can also classify data on the basis of machine learning classification algorithms like few of them are, Ordinary least squares, Bernoulli Naïve Bayes, Gradient Boosting Regression trees,Adaboost, Stochastic Gradient Descent etc. These classification algorithms are mainly used to depict the trends in dataset and which factors mainly affect your machine learning model.

Moreover, apart from R programming we can also used Python for machine learning programming. This is because the Python external libraries like NumPy, Scikit-learn, Matplotlib, SciPy and many more allow you to be flexible and provides wide range of options for data analysis.

In addition to not only bicycle rental but this model can also be used to deploy two wheelers like motorcycles, scooters etc as two-wheeler sharing system especially in countries like India where cities like Pune are densely populated with number of two-wheeler.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

IX. CONCLUSION

The Bicycle Rental System is not only popular in many smart cities but also widely accepted due to the benefits it comes along with. Not only environmental factors but also, it's a good choice or lets say a reason to people to exercise. This also motivates people to avoid taking their motorbikes and use a environment-friendly option, instead o wasting time in walking for 201 minutes grab a bicycle for 5 minutes. Thus Bicycle Hailing System is a blooming topic an bicycles availability, its security everything can be taken care of with the help of machine learning.

REFERENCES

[1]. Harish Manghnani, Minal Vanage, Akshata Naik, 'Implementation of Machine Learning to predict number of bicycles in a Bicycle Rental System', IJIRSET 2017.
[2 Gabriel Martins Dias, Boris Bellalta and Simon Oechsner,'].Predicting Occupancy Trends in Barcelona's Bicycle Service Stations Using Open Data',IEEE 2015.
[3]Fei Lin∗, Shihua Wang∗, Jian Jiang∗, Weidi Fan∗, Yong Sun,'.Predicting Public Bicycle Rental Number using Multi-source Data'IEEE 2017.
[4]. Machine Learning or Discrete Choice Models for Car Ownership Demand Estimation and Prediction IEEE 2017.
[5]. Dong Wang , Wei Cao , Jian Li1 Jieping Ye2,'Supply-Demand Prediction for Online Car-hailing Services using Deep Neural Networks 'IEEE 2017.
[6]. Fu-Shiung Hsieh,'Car Pooling based on Trajectories of Drivers and Requirements of Passengers' IEEE 2017