International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 2507–2512

Feature based Visual Analysis of Indian Railways Station data with a statistical approach for Train delay prediction

Dr. V.Asha[1], Heena Gupta[2], Dr. B. Meenakshi Sundaram[3]
[1]*Department of MCA, New Horizon College of Engineering, Bangalore.*
[2]*Department of Computer Science, Mount Carmel College, Bangalore.*
[3]*Department of MCA, New Horizon College of Engineering, Bangalore.*
[1]*asha.gurudath@gmail.com*, [2]*heenag2248@gmail.com*, [3]*bmsundaram@gmail.com*

Abstract

Indian railways are an integral mode of transport for any Indian. The delay in train travel adds to problems sometimes. To predict the delays efficiently, the right preprocessed data is needed about the infrastructure, train travel and other subjects of concern about the railways. The extraction of correct set of features is also imperative. The paper thus aims to extract and thus understand the right set of features affecting the rail infrastructure by using data visualization techniques.

Keywords: *Indian Railways, Visualization, Preprocessing*

1. Introduction

Transport is an important part of any country's economy; the major part is from railways. Rail transport is one of the important and economical transports used for travelling long distances. In India almost all rail operations are handled by a state-owned organization, Indian Railways. Indian rail transport network is fourth largest rail network in the world carrying millions of passengers, and this number of passengers keeps increasing. People travel from one state to another in search of work, to look for better job opportunities or to meet family in various states. This leads to crowding of trains due to huge demand and lesser supply. If train delays can be appropriately predicted it will lead to hassle free travel among people. The train infrastructure data needs to be rightly analyzed and preprocessed so that addition of trains and infrastructure are optimized.

1.1 Objectives:

i) The aim of the paper is to analyze the rail infrastructure data and extract the right set of features.

ii) The various data visualization techniques will also help to understand how the features affect the railway data.

2. Literature Survey

Work related to prediction and using prediction algorithms for various applications has been an active research topic. Various papers discuss different schemes, ensemble techniques or new approaches to data preprocessing to enhance the accuracy in prediction and to minimize the loss.

Sai Nageshware et al studied the representation of Indian Railways which is one of the biggest networks in the world using hypergraphs. Hypergraphs better capture qualitative information than simple graphs. [1] They discuss several latent properties. Diego Arenas et al [2] apply a new mathematical model to solve real time train timetabling problem for German railways. Francesco Cormam at al [3] focus on unavailability of a track due to blockage and hence distributed approaches are presented to manage effectively larger networks. The paper [4] presents a framework for statistical analysis of large amount of data for various features. Many global diagnostics were created to assess.

ISSN: 2233-7857 IJFGCN

The paper [5] uses Artificial Intelligence rule based decision making with automatic feedback for performing statistical analysis for network data visualization example. The paper [6] discusses a new approach for performing statistical analysis for big data. Xu, Xiaoming et al [7] discus the time–space status of trains in the railway system, the status may be one of three categories, including dwelling at a station, waiting at a station and traveling on a segment. A check algorithm is particularly proposed to guarantee the feasibility of transition. Shakibayifar, Masoud et al [8] discus the best possible infrastructure plan for scheduling new train services. It also determines the best stop locations in "An integrated train scheduling and infrastructure development model in railway networks"

3. Technical Details

To perform train delay prediction the preprocessing of data is to be done so that the machine learning algorithm performs efficiently. So we perform various data query operations to get the right visualization. These visualization techniques will help us to understand each feature's role for train delay prediction.

3.1 Different Plots:

The different plots generated are:

- Relational Plot:

This is one of the easiest and one of the most frequently used chart. It shows relationship among various chart elements.

- Histogram:

This is a chart which shows distribution of data as per various bins.

- Distribution Plot:

This is a chart which shows distribution of data with more options. It is similar to a histogram.

- Rug Plot:

This is a chart which shows distribution of data by placing a dash for each occurrence.

- Kernel Density Estimate Plot:

This is a plot which shows distribution of data using a continuous probability density curve.

4. Experiment

4.1. Dataset

Open Government Data (OGD) Platform India is a platform for supporting Open Data initiative of Government of India. The portal is intended to be used by Government of India Ministries to publish datasets, documents, services, tools and applications collected by them for public use. Uploading of such data tend to increase transparency in the functioning of Government and also open avenues for many more innovative uses of Government Data. Various datasets related to Indian railways are available. The infrastructure data considered for this study has 186124 records.

The fields are:

- Train No : This gives a unique train ID
- Train Name: This is the unique train name
- SEQ: This indicates a sequence number for the stations for each train.
- Station Name: This indicates the station where the train will arrive to.
- Arrival Time: This gives the arrival time from that station.

ISSN: 2233-7857 IJFGCN

- Departure Time: This gives the departure time from that station.
- Source Station: This gives the source station name.
- Destination Station: This gives the destination station name.

4.2. Data Preprocessing

The redundant columns or columns conveying redundant information were deleted. The columns were renamed to indicate right meaning. Basic transformations were also applied to encode data in appropriate way.

Various grouping functions were well applied to know the trains with maximum number of stops and distance. This will help us to understand the frequency of stops by trains and the distance that is being covered by them. For our study, the top ten trains' data with maximum number of stops and ten trains' data with maximum distance were chosen for study.

4.3. Visualization

The data has several fields of interest. The visualization is drawn for these fields using few sample records.

- Relational Plot:

The figure 1 shows the relational plot with stations and distance.

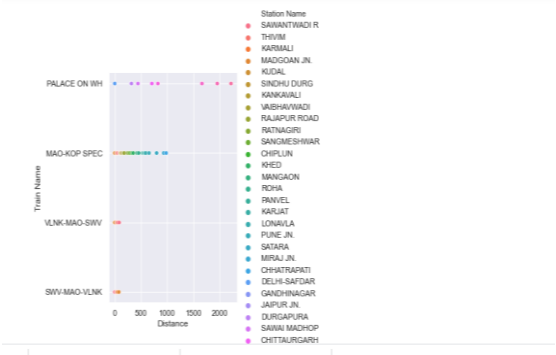

Figure 1. Relational Plot

- Histogram:

The histogram to understand the number of stops for a sample set of trains was generated. The figure 2 shows the histogram in this case.

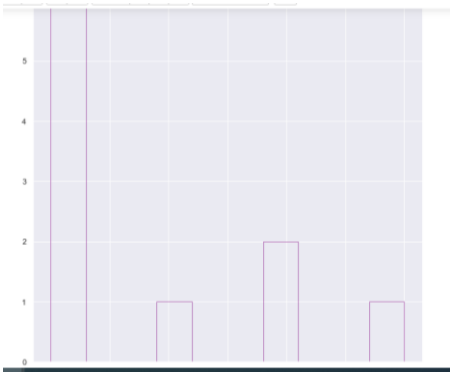

Figure 2. Histogram

ISSN: 2233-7857 IJFGCN

- Distribution Plot:

The distance plot will help to understand the distance covered by the sample set of trains. The figure 3 shows the distribution plot with respect to distance covered by a train.

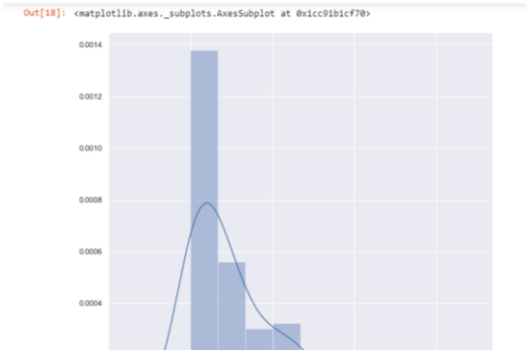

Figure 3. Distribution Plot

- Rug plot:

The figure 4 shows the rug plot and each occurrence of the distance is marked with a dash. This helps to understand the various distances that the trains cover.

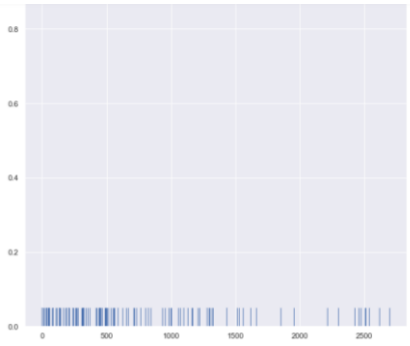

Figure 4. Rug plot

- Kernel Density Estimate Plot:

The figure 5 shows the Kernel Density Estimate plot. This rightly shows the probability density curve for the distance covered.

ISSN: 2233-7857 IJFGCN

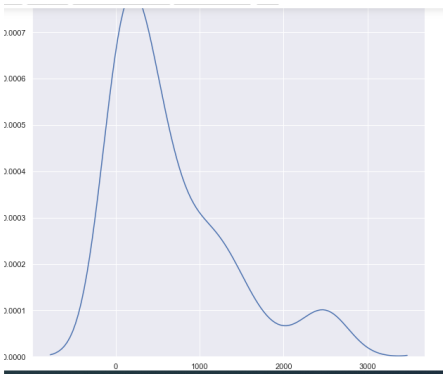

Figure 5. KDE plot

5. Results and Discussion

Using the above visualization techniques and charts, the important features of the dataset were studied. The dataset rightly shows how various features affect the train movement.

6. Conclusion

In conclusion, this study shows how the dataset was used to study useful insights about the railway infrastructure to perform train delay prediction. This will help in easy and timely scheduling of trains to help the general public, at large.

References

[1] S. N. Satchidanand, S. K. Jain, A. Maurya and B. Ravindran, "Studying Indian Railways Network using hypergraphs," 2014 Sixth International Conference on Communication Systems and Networks (COMSNETS), Bangalore, (**2014**), pp. 1-6.

[2] Diego Arenas et al, “Solving the Train Timetabling Problem, a mathematical model and a genetic algorithm solution approach”, 6th International Conference on Railway Operations Modelling and Analysis (RailTokyo2015), Tokyo, Japan, (**2015**) March.

[3] Francesco Corman, A. D'Ariano, I. A. Hansen, D. Pacciarelli and M. Pranzo, "Dispatching trains during seriously disrupted traffic situations," 2011 International Conference on Networking, Sensing and Control, Delft, (**2011**), pp. 323-328.

[4] J. C. Bennett et al., "Feature-Based Statistical Analysis of Combustion Simulation Data," in IEEE Transactions on Visualization and Computer Graphics, vol. 17, no. 12, (**2011**) December, pp. 1822-1831.

[5] D. H. Wong and K. Chai, "Statistical analysis learning approach: The use of artificial intelligence in network data visualization system," 2010 International Conference on Computer Applications and Industrial Electronics, Kuala Lumpur, (**2010**), pp. 206-210.

[6] K. Wang, J. Xu, J. Woodring and H. Shen, "Statistical Super Resolution for Data Analysis and Visualization of Large Scale Cosmological Simulations," 2019 IEEE Pacific Visualization Symposium (PacificVis), Bangkok, Thailand, (**2019**), pp. 303-312.

[7] Xu, Xiaoming & Li, Ke-Ping & Yang, Lixing & Gao, Ziyou, “An efficient train scheduling algorithm on a single-track railway system”, Journal of Scheduling, (**2018**).

ISSN: 2233-7857 IJFGCN

[8] Shakibayifar, Masoud & HassanNayebi, Erfan & Mirzahossein, Hamid & zohrabnia, Shaghayegh & Shahabi, Ali, "An integrated train scheduling and infrastructure development model in railway networks", Scientia Iranica, vol 24, (**2017**).

ISSN: 2233-7857 IJFGCN