Simulating and Analysing delay in Indian Railways

Rahul Pradhan
Ashutosh Kumar
Mayank Kumar
Bhakti Sharma
GLA University, Mathura

Abstract. Delay is one of the serious issue all over the world in railways, however it becomes more important to country like India having the top four biggest railway network all over the world and where the majority of the population depends on railways for making their journey’s in different parts of our country. Categorically most of the train in Indian railway not able to run on its schedule time which in turn affects the passenger to reach their destination timely. So we have make a small step in order to learn and analyze these problems and its factor we have used a real world dataset from internet and perform simulation on the data using regression to predict the total delay take place while planning a particular journey from a provided date by a passenger and our approach accuracy is promising.

1. Introduction
Big Data along with its latest technology plays an efficient and a vital role in performing and analyzing the data collected from different sources having high volume and velocity helps in gaining insights for making strategic decisions that helps business activity to boom by verifying and removing the faulty area. When we talks about Delay in railways which is one of the major problem throughout the World [1] takes place due to various reasons causes the trains to be stopped at a particular point for almost 5 to 6 hours or more depends on factor which causes these delays, which cost an effective loss of time to the railways as well as for the passengers who use to travel through the train at that time [3]. Therefore, it becomes more important when we talks about Indian Railway System as it is one of the vast railway network of 66,687Km in the world ranking as the fourth largest one with 11,122 locomotives [3] and for every day almost about 23 million of population travels along different parts of the country [1] and about 8 billion passenger travels in a year. According to paper [2] delays can be due to various causes: disruptions in the operations flow, accidents, malfunctioning or damaged equipment, construction work, repair work, and severe weather conditions like snow and ice, floods, and landslides, to name just a few. Although trains should respect a fixed schedule called Nominal Timetable (NT), Train Delays (TDs) occur daily and can negatively affect railway operations, causing service disruptions and losses in the worst cases[12,13]. According to paper [11] this rail transit system also used this key concept of big data to get advantage of analyzing the sophisticated processing of large volumes of data. This is the key
concept to all the professionals in transportation to understand the role of analytics in their respective domain sectors which help them to boost their sector revenue.

So in a country like India where most of the trains unable to run on its schedule time due to poor communication and less number of tracks will affect the passenger not to reach their destinations on a correct time and, this happens to large population of our country which depends on this sector for planning their journey of short or long distances based on their requirement [1]. Therefore, it becomes important that there is a system which in advance predict the accurate delay that might takes place for a passenger planning its journey on a particular date from a particular station, so that the passenger plan its journey accordingly [1] and also helps in gaining the faith of passenger for a making their journey by train which in turn increases the revenue in this sector.

So, this proposed work helps us in learning and understanding of how to build a prediction system as this is not the first research taken in this area many of other researcher also used to provide well defined system for predicting the delays based on the data they collected from different resources [4]. We were going to use a machine learning model on the data set we have collected and before that we do some exploratory analysis which in turn provide a best view for us to apply which model is best for our dataset based on that we can predict the accurate delay for a journey which in sights help the passenger to plan accordingly.

Figure 1: - Information regarding Indian Railways System

2. Related Work
The table shows the research work done by other researcher in predicting delay systems in railways and what are the models and approach they had taken to define their model more accurately.
| Year | Author                  | Title                                                                 | Approach/ Model/ Basic Idea                                                                 | Dataset used                                                                 | Pros                                                                                      | Cons                                                                                       |
|------|------------------------|----------------------------------------------------------------------|-------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------|
| 2010 | [5] A. Hansen et. al. | Online train delay recognition and running time prediction          | Offline Statistical analysis [16]                                                        | Netherlands historical train data, Dutch railway corridor Rotterdam           | Piecewise linear function may maximize the accuracy of prediction in case of large dataset. | Not suitable for large datasets.                                                             |
| 2012 | [3] Masoud Yaghini et. al. | Railway passenger train delay prediction via neural network model | Artificial Neural Network                                                                | Historical train dataset of Iranian Railways                                   | To evaluate the quality of result decision tree and multinomial logistic regression are used. Approximately 90% accuracy is used. | Can be improved through meta heuristic methods such as genetic algorithm or hybrid algorithm. |
| 2013 | [6] Kecman Pavie et. al. | Train Delay evolution as a stochastic process                      | Timed event graph with dynamic arc weights using recursive depth first search algorithm [16] | Netherlands train data                                                        | Produced accurate estimates for train traffic and route conflicts within 30 min.[17]       | Adding some more factors like railways assets condition, whether data may increase the efficiency. |
| 2014 | [9] Suporn Pongnum Kul et. al. | Improving arrival time prediction of Thailand’s passenger trains using historical travel times | k-Nearest Neighbors-based algorithm                                                        | Historical data of Thailand local passenger train                              | Algorithm based on k-NN improves the prediction error by 23% on average( when k=16)            | Missing most important factor i.e. weather info.                                             |
| 2014 | [8] Jia Hu, Bernd Noche et. al. | Application of Artificial neuron network in analysis of railway delays | Multilayered perception, GA-BPNN model                                                     | Implementing GA in BPNN improves prediction performance.                      | Small dataset, less parameter and the runtime of the GA-BPNN is more.                     |                                                                                              |
| Year | Reference | Methodology | Dataset | Accuracy | Improvement |
|------|-----------|-------------|---------|----------|-------------|
| 2015 | [10] Yaghini, Masoud et al. | Passenger train delay classification | Decision tree and multinomial logistic regression | Iranian Railways | Good accuracy. Can be improved through metaheuristic methods such as genetic algorithm or hybrid algorithm |
| 2016 | [7] Robert Nilsson et al. | Prediction of train delays using machine learning | Decision tree, Decision tree with Ada boosting, Neural Network | Local traffic of Stockholm using Trafikverket’s API | Average error reported 3 minutes in case of neural network. Other methodologies may give better prediction. |
| 2018 | [2] L. Oneto et al. | Train Delay prediction systems: A big data analytics perspective | Kernel method, Extreme learning machines, Ensemble methods | The Italian Case, Italian train historical data and weather data | Addition of weather info increases the accuracy by 5-10%. Railways assets conditions need to be added. |
| 2018 | [4] R. Gaura et al. | Estimating train delay in a large rail network using a zero-shot Markov model | N-Markov Model, N-OMLMPF algorithm, Random forest regressor (RFR) model | Indian railways historical data. | Used to predict late minutes at an inline station. Some factors like railways assets conditions could increase the efficiency. |
| 2019 | Ping Huang et al. [13] | A deep learning approach for multi-attribute data: A study of train delay prediction in railway system. | 3-dimensional convolutional neural network (3D-CNN), Long short-term memory recurrent neural network (LS TM), Fully connected neural network (FCNN) | Chinese railway networks and Netherlands railway networks. | Used to capture a sequence of train operations and make use of historical data to capture the cumulative interactions between trains and stations.[19] Train heterogeneities such as time and space differences in timetables are not involved. |
### 2019
| Authors          | Section Title                                                                 | Methodology                                                                 | Description                                                                                   | Result                                                                 |
|------------------|-------------------------------------------------------------------------------|----------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|----------------------------------------------------------------------|
| Yuxiang Yang et. al. [12] | Statistical delay distribution analysis on high-speed railway trains. | Kolmogrov-Smirnov test (K-S) | Train operational database of China Railway corporation. | Has great practical application value as most of the distributed models fitted accurately fit the impact of HSR disturbance of train delay time. |
|                   |                                                                                |                              |                                                                                                 | CSSF model doesn’t passed the K-S test as the data scale is not adequate for the precise calculation. |

### 2020
| Authors           | Section Title                                                                 | Methodology                                                                 | Description                                                                                   | Result                                                                 |
|------------------|-------------------------------------------------------------------------------|----------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|----------------------------------------------------------------------|
| Patgiri et. al.   | Empirical study on airline delay analysis and prediction                      | Logistic regression with L2 regularization, Gaussian Naïve Bayes, K-nearest neighbors, Decision tree classifiers and Random forest model. | American statistical association and national digital forecast database.                     | Showed that the flight can be scheduled, organized and analyzed in a much better way. |
|                   |                                                                                |                              |                                                                                                 | Work can be carried out with other machine learning models for better accuracy. |
| Esmaeilzadeh et. al. [15] | flight departure delay prediction and analysis using machine learning | SVM and K-fold cross-validation model | New York airlines data | Various explanatory variables are analyze that might produce an impact on delay variable. |
|                   |                                                                                |                              |                                                                                                 | More effective TMI implementation can help to reduce overall delay. |

### 3. Proposed Methodology
Big Data Analytics is the use of data analytics tool on a very large quantity of data to get some information or some pattern to make future prediction for a business or organization. In this project, we followed the following steps to perform analysis on a dataset of trains to predict delays in arrival of trains:

1. Collection of dataset
2. Feature Selection, Filtering and Weighing
3. Learning Multiple Regression Model
4. Evaluating the model
5. Predict the Model and deploy it
Figure 2: Workflow in predicting Train Delays

At first it was important to find the factors which were responsible for train delays in the past to predict about the delays in future. One such factor is weather and climatic conditions. The other factors are railway traffic which we did not included in this prediction. The next step was to collect a dataset. So we collected a dataset of trains that ran in 2016 from the Indian railways site. A small sample of dataset we used –

The first dataset is as follows –

Table 1. Data set of trains from Delhi to Mumbai with temperatures variation and Delay

| sday | stime | eday | MUMH | MUML | DELH | DELL | delay |
|------|-------|------|------|------|------|------|-------|
| 2    | 1060  | 3    | 32   | 23   | 18   | 10   | 15    |
| 3    | 1060  | 4    | 31   | 22   | 20   | 20   | 11    |
| 4    | 1060  | 5    | 32   | 23   | 23   | 12   | 15    |
| 5    | 1060  | 6    | 31   | 23   | 21   | 12   | 0     |
| 6    | 1060  | 7    | 32   | 23   | 20   | 14   | 14    |
| 7    | 1060  | 8    | 31   | 23   | 15   | 14   | 145   |

The attributes of this dataset are sday for starting day, stime for starting time, eday for end day and as we have taken dataset of train from Mumbai to Delhi so MUMH, MUML,DELH, DELL, are highest and lowest temperature of Mumbai and Delhi respectively and delay is the delay in train time in minutes(show in Table 1).
The second dataset is of merged data –

Table 2. Data set of variations in Temperature.

| date | tempLo | tempHi | tempLow | tempHi |
|------|--------|--------|---------|--------|
| 60   | 19     | 33     | 11      | 24     |
| 61   | 19     | 33     | 11      | 24     |
| 62   | 19     | 33     | 11      | 24     |
| 63   | 19     | 34     | 11      | 24     |
| 64   | 21     | 35     | 15      | 26     |
| 65   | 21     | 35     | 16      | 27     |

The second dataset is of merged data containing attributes as date, tempLo, tempHi, tempLow and tempHi (show in Table 2).

Figure 3. Prediction of train delay shows the stations and weather information.

In the working code in Python, at first a function is called in which two strings are passed - the date and time of the train to predict the delay. The function converts the date into day and month and by the number of the month adds the days from the month and all the previous month and returns the days. After that the dataset merged data is made from the date passed and necessary calculations are done. After that the first dataset above is used and the attributes are converted into an array removing delay and delay is stored in another variable and then linear regression is performed on the two variables. And at last we predict the delay using linear regression model.

The next step was to filter out data which is useful for the prediction from the large dataset using feature selection, filtering and weighting. As it is important to work only on meaningful data from the
dataset as the irrelevant data decreases the accuracy in the prediction. A model cannot be based on irrelevant features so for finding important features feature selection is performed.

Now we have to find a model using which we will be able to make the prediction for our project. After trying certain amount of models we decided to stick to Multiple Regression for predicting delay times.

Now we have to find a model using which we will be able to make the prediction for our project. The tools we used for performing exploratory analysis are done by using R-language. A language and environment for statistical computing and graphics. At first we plotted a graph that discuss the relation exist between various attributes which in result affect the dependent variable that is delay attribute, we have also shown it by numerical value.

3.1 A small introduction of Multiple Regression[5]:

A Multiple regression is used when there is need to find the relationship between more than two variables by fitting these variables data into model similar to linear model. Here one variable is called a dependent variable and the rest other variables is called an independent variable.

A multiple regression line has similar equation that a line has -

\[ F(x_1, x_2, x_3, \ldots, x_n) = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n \]

here \( x_1, x_2, x_3, \ldots, x_n \) are the independent variable, \( F(x_1, x_2, x_3, \ldots, x_n) \) is the dependent variable, and the intercept is \( w_0 \), and \( w_1, w_2, \ldots, w_n \) are the slopes.

At last we used the Multiple Regression model for predicting delay times. The training data had data of all the trains that ran in 2016. The constraint on input is that we have only included booking dates of March, 2017.

4. Experimental Setup And Result Analysis

During performing our proposed work we have decided to use real world data. For this we have managed to create our own dataset from net using python library. This dataset includes some delay information of a train running from Delhi Nizamudin to Mumbai and also scraps a data related to weather condition on that particular area on the respective journey date and all these data’s are of 2016 year. After collecting the dataset we are going to merge the data and make our final dataset after analyzing what are the various attributes that might help in contributing the delay. After that there is need to use some tools which perform this prediction for this we have choose the pycharm software, we loaded our data set
after that we perform our machine learning model on that historical dataset. And finally we have completed our prediction part, then we decided to present this result on a GUI. For making GUI we can use python tkinter library which helps in making GUI for showing the result in decorative way to the viewers. Tkinter is standard library in python used for creating Graphical User Interface for desktop applications. First we import the module, after that we create top-level windowing object that contains our entire GUI application and set up all our GUI components and their functionality which connects to our application code and at the end we just enter the main event loop using mainloop function[20]. The result came to be a very significant and the accurate prediction done by our model. We have used a multiple regression model for our project which shows the weather factors is one of the most important contributing factors to the railways delay.

The [3,17] model in the study can be applied to other stations although similar data must be collected. With respect to the expansion direction of the model, the current model does not consider an excessive number of infrastructure factors. With respect to further model expansion, it is possible to consider additional train delay influence factors and extract increasingly accurate feature variables to obtain better prediction results.

**Figure 5:** Relationship between the datasets of Table 1 using graphical representation

Graphical representation of correlation among the attributes as the S DAY and E DAY are closely related also there is strong relation between M UML and M UMH, similarly between D ELH and D ELL and all this significantly affect the delay attributes.
Figure 6: Relationship between the datasets of Table 1 using numerical value.

Correlation among the various attributes that are used in this model are shown using the graph and also with the help of numerical value as the sday and eday have a correlation of 0.98299844 which is close to the value 1 hence this attributes are closely related to each other and similarly, the other attributes and its correlation between itself and with others are listed in the above diagrams.

Figure 7: Calculated p-value and Adjusted R-Squared

Here this regression model the stime and the DELH parameter provide strong relation with delay attribute. We have Adjusted R-squared value 0.2721 which means if the new attribute added in the dataset it affect the dependent variable delay significantly by the variance of 0.2721.

A lower p-value is interpreted as meaning that there is a stronger relationship between variables, the probability under a specified regression model [21] that a statistical summary of the data would be equal to or more extreme than its observed delay value and since we have a very less p-value which is less than \(2.2e^{-16}\) closely to 0. Therefore this model predicts the delay more accurately.
5. Conclusion
In India train delay is the major problem and Government of India try to solve this by introducing new routes analyzing the delay factors from the past data. Train delay is depending on the train movements and weather conditions. Weather conditions also affect the temperature and we have dataset which have temperature data. The main purpose of this paper is finding delay of train time. In this paper we have used temperature records along with historical train status record which used to predict delays and also marks which variable in a dataset is most contributing factor in train delay. We have put forward key train delay stations from Delhi to Mumbai and determine the total delay done by a train on a particular journey on a particular date. The model provide the accurate results but not the exact delay, since we have used limited variable and records to predict the delay. This can be improved further by adding some more independent variable and the most recent delay records along with other climatic conditions provide the accurate prediction.

References
[1] Arshad M and Ahmed M 2019 Prediction of train delay in Indian railways through machine learning techniques Int. j. comput. sci. eng. 7 405–11
[2] Oneto L, Fumeo E, Clerico G, Canepa R, Papa F, Dambra C, Mazzino N and Anguita D 2018 Train delay prediction systems: A big data analytics perspective Big Data Res. 11 54–64
[3] Yaghini M, Khoshraftar M M and Seyedabadi M 2013 Railway passenger train delay prediction via neural network model: RAILWAY PASSENGER TRAIN DELAY PREDICTION J. Adv. Transp. 47 355–68
[4] Gaurav R and Srivastava B 2018 Estimating train delays in a large rail network using a zero shot Markov model 2018 21st International Conference on Intelligent Transportation Systems (ITSC) (IEEE)
[5] Hansen I A, Goverde R M P and van der Meer D J 2010 Online train delay recognition and running time prediction 13th International IEEE Conference on Intelligent Transportation Systems (IEEE)
[6] Kecman P, Corman F and Meng L Train delay evolution as a stochastic process Diva-portal.se
[7] Nilsson R and Henning K 2018 Predictions of train delays using machine learning (Dissertation DIVA)
[8] Hu J and Noche B 2016 Application of artificial neuron network in analysis of railway delays Open J. Soc. Sci. 04 59–68
[9] Pongnumkul S, Pechprasarn T, Kunaseth N and Chaipah K 2014 Improving arrival time prediction of Thailand’s passenger trains using historical travel times 2014 11th International Joint Conference on Computer Science and Software Engineering (JC SSE) (IEEE)
[10] Yaghini M, Sanai M S and Sadrabady H A 2013 Passenger train delay classification Int. j. appl. metaheuristic comput. 4 21–31
[11] Thaduri A, Galar D and Kumar U 2015 Railway assets: A potential domain for big data analytics Procedia Comput. Sci. 53 457–67
[12] Yang Y, Huang P, Peng Q, Li J and Wen C 2019 Statistical delay distribution analysis on high-speed railway trains J. Mod. Transp. 27 188–97
[13] Huang P, Wen C, Fu L, Peng Q and Tang Y 2020 A deep learning approach for multi-attribute data: A study of train delay prediction in railway systems Inf. Sci. (Ny) 516 234–53
[14] Patgiri R, Hussain S and Nongmeikapam A 2020 Empirical study on Airline Delay Analysis and prediction arXiv [cs.LG]
[15] Esmailzadeh E and Mokhtarimousavi S 2020 Machine learning approach for flight departure delay prediction and analysis Transp. Res. Rec. 2674 145–59
[16] Wen C, Huang P, Li Z, Lessan J, Fu L, Jiang C and Xu X 2019 Train dispatching management with data-driven approaches: A comprehensive review and appraisal IEEE Access 7 114547–71
[17] Kecman P and Goverde R M P 2013 An online railway traffic prediction
A predictive model of train delays on a railway line [18] Wen C, Mou W, Huang P and Li Z 2020 J. Forecast. 39 470–88

Modeling train operation as sequences: A study of delay prediction with operation and weather data [19] Huang P, Wen C, Fu L, Lessan J, Jiang C, Peng Q and Xu X 2020 Transp. Res. Part E: Logist. Trans. Rev. 141 102022

Core Python Programming [20] Chun W J 2007 (Philadelphia, PA, USA: Prentice Hall)

P-Values and Statistical Significance [21] Mcleod S 2019 Simplpsychology.org

Train delay analysis and prediction based on big data fusion [22] Wang P and Zhang Q-P 2019 Transportation Safety and Environment 1 79–88