Analyzing the factors affecting the delay of train delays using Data mining

Jyoti Verma , Assistant Professor, ROFEL BBA & BCA College, Vapi, India

moksh.mj.mv@gmail.com

Tanvi Rana , Assistant Professor, ROFEL BBA & BCA College, Vapi, India proftanvi@gmail.com

Abstract - People always prefer safe , fast and cheap mode of transportation and such mode of transportation is none other than the railways. Also the passengers travelling by any means of transportation always prefer the mode of transport to be on time. Maximum trains are available for the transportation and also majority of times the trains are on time but under some unforeseen or unavoidable circumstances the trains may get delayed. The paper here is a review paper which has studied certain papers to find and predict the reason of the delay in train. The study has used the Neural Network, KNN,Decision tree, Decision tree with Adaboot for making the delay predictions.

Key words: Adaboost, Classification, KNN, Multivariate regression, Machine Learning, SVM, Classification

I. INTRODUCTION

People use the technology for making their life easy and this gives the contribution to structured and unstructured data this data has paved the way to the new horizons in Data mining tries to find the link between totally unknown and hidden data from the warehouse of the data given to it. Finding the unexpected result from data can be termed as Data mining. Discovery of the hidden information in the data the data is trained based on specific data mining algorithms like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

In 2017-18, Indian Railways had the worst punctuality performance in three years. 30 per cent trains ran late in 2017-18, according to official data. From April 2017-March 2018, the punctuality of mail and express trains was 71.39 per cent, down from 76.69 per cent in 2016-17, which is a deterioration of 5.30 percent. (The Economics Times, 2018) But as per the current study in the beginning of this year the train are 100% punctual.

This paper ties to present the review of various papers involved in the prediction of the reasons behind the delay and the accuracy of various algorithms.

The rest of the paper is organized as section 2 is literature review of research papers, section 3 includes the algorithms used in research paper summary, section 4 is the discussion of the methods and finally the section 5 is conclusion.

II. LITERATURE REVIEW

The authors (Weiwei Mou, 2019) have tried to develop a predictive model for predicting the train delay. They have compared the Long Short Term Delay(LSTD) with the random forest and Artificial Neural Network. The data of the railways for the study was taken from the dutch railways. They have identified the dependent and independent variables based on the passed records. Major factors cover the climate condition but the infrastructure is not included.

The data of the French railways is considered for the study. The data has related to the infrastructure related issue has been considered as the factors responsible for the train delay. Infrastructure issues like feature of track, engineering structure, level crossing, the kind of goods transported through train. The data for 2015 January to December 2019 are used for the study. The authors have tried to visualize the prediction better understanding. Only two models have been implemented for the prediction. Random forest and SVM are the models used. The authors (Lbazri Sara, August 2020) have suggested that more algorithms can be implemented to get more accurate result.

(Heena Gupta, March 2021) have done the research on the Indian railways. The main cause for the delay which have been considered is the speed restriction imposed for the safety of the passengers. Machine learning algorithms like KNN, Logistic regression, SVM, Decision tree and random forest are used to test the efficiency of the study. The speed constraint can be made more precise by dividing speed delay to the delay caused by the track and other train delay which impose the delay of the train. The data of the southern Indian railways is used for the study. The Random forest has shown the best result in the delay predictions.

The (Ramashish Gaurav, 2018) have considered the data of Mughalsaria Indian railway station from the period of March 2016 to February 2018. The delay prediction has been done from the actual arrival time of the train. The predictions are made using the N order Marckov Late

 DOI : 10.35291/2454-9150.2021.0538

Minutes Prediction Framework. The inline station are the factors because of which the delay has been caused.

In this paper the authors (Lokaiah Pullagura, 2019) have considered the reasons like fog, traffic, signal failure, accidents, derailments for the train delays. The authors have assumed that if a train is delayed is delayed at previous station than it will be delayed at subsequent stations as well. The data is collected from the etrain.info/in website for the study of the stations between vijaywada to Chennai for the period of January 2018 to March 2018. The major cause for the delay is the infrastructure which causes the accident or derailments which delays the train in the schedule. Decision tree with Adaboost and Recurent Machine Learning algorithms are used for the predictions.

The authors (Mohd Arshad, 2019) have tried to predict the delays in train caused by the poor signal and the lack of track availability. They have used 3 different machine learning methods Multivariate regression, Neural Network, and Random Forest for the accurate findings. Weather data play an important factor in prediction of train delay in India because weather in India is differing from one state to another state, unlike from other countries. The paper can be foundation for the other researches.

Here the data set of the Iranian Railways is considered for the study. The data of the delay and the reasons from start of 2005 till the end of 2009 have been collected. The main algorithm used is the Neural network for getting the result of the predictions. Neural network Numeric ,binary data. Decision tree and logistic regression both work on the numeric data. The factors which affect the delay are Delay at the origin, Incidence with another passenger train, Unscheduled waiting time at overtaking points, Engine breakdown, other train's engine break down, Wagon breakdowns, infrastructure faults such as track and signal failure. (Masoud Yaghini, 2012)

(Suporn Pongnumkul)historical data of arrival time or departure time from the last station to predict the delays. It uses k-NN and moving average . The number of stop stations was found as one of the factors that caused an increase in train delay and it also contributed to the prediction error. The arrival time of the historical data is used instead of the current and online data for the prediction so in future the study can be applied to the real time data.

III. METHODS

Decision tree

A decision tree is a tree-like graph with nodes representing the place where we pick an attribute and ask a question; edges represent the answers the to the question; and the leaves represent the actual output or class label. They are used in non-linear decision making with simple linear decision surface. (htt1)

Decision tree with AdaBoost

The algorithm is majorly used in the classification and regression tasks. Adaboost algorithm uses the weak classifiers and using them it creates a strong and high performance classifier for the output generation. AdaBoost algorithm includes several classifiers and classifiers have different classification rates on datasets. During the training process, weak classifier which has the least error is used. Weights of wrongly classified samples are increased to be classified more accurate by next classifier (M. Barstuğan, 2014)

The main difference between Adaboost and bagging methods (including Random Forests) is that, at the end of the process, when all the classifiers built during the iterations will be asked to vote for the target of a new observation, there will be trees with a heavier vote than others. Those are the trees that performed the best during all the iterations (so, they showed very few misclassifications) (Alto, 2020)

Random Forest

In case of regression it handles continuous variables and in case of classification it can handle categorial variables. It works by collecting various models like bagging and adaboost. Bagging maximum voted value is selected as result and multiple weak learners are used to create a strong learner. (SRUTHI94, 2021)

The random forest algorithm is an extension of the bagging method as it utilizes both bagging and feature randomness to create an uncorrelated forest of decision trees (Educator, 2020)

Linear Regression

It is used to find the relationship between dependent and independent variable by generating a line that best fits the equation

Y= A *X + B Where, Y = Dependent Variable, X = independent variable, A = slope, B= intersect (Mohd Arshad, 2019)

Multivariate Regression

Regression used when there are more than one variable to predict a value. The mathematical function/hypothesis of a Multivariate regression is of the form:

$$y = \beta_0 + \beta_1.x_1 + \ldots + \beta_n.x_n$$

where, n represents the number of independent variables, β0~ βn represent the coefficients and x1~xn, are the independent variable (Raj, 2020)

Neural network

The machine learning branch Deep learning uses the concept of the simulated neural networks known as the

 DOI : 10.35291/2454-9150.2021.0538

Artificial Neural Network (ANN) which mimics the functionality of the human brain to make the decisions. Artificial neural networks (ANNs) are comprised of a node layers, containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. Otherwise, no data is passed along to the next layer of the network. (Education, 2020). The ANN can be classified based on their use as the Convolutional neural networks (CNNs), Recurrent neural networks (RNNs) . Deep learning and the neural network are used interchangeably but the difference lies between the number of the layers used the neural network uses 3 layers whereas the deep learning included more than 3 layers by increasing the number of the layers the complexity of the decision making can be increased.

Long Short Term Memory Network

Long Short Term Memory Network is an advanced version of the RNN. The RNN uses the past information to predict the current situation but the drawback of the same is that it is not able to keep up with the long term dependency and hence the Long Short Term Memory Network came into existence to overcome the vanishing long term dependency. It consists of 3 parts these three parts of an LSTM cell are known as gates. The first part is called **Forget gate, the** second part is known as **the Input gate** and the last one is **the Output gate**. (SAXENA, 2021)

K nearest neighbors

It works on the comparison of the tuple value with the training set. N attributes are used to show the training set. When given an unknown tuple, a ***k*-nearest-neighbours classifier** searches the pattern space for the *k* training tuples that are closest to the unknown tuple. These *k* training tuples are the *k* "nearest neighbours" of the unknown tuple. (Jiawei Han, 2012). Euclidean distance is used to measure the distance between the values.

IV. FINDINGS

The factors which affect the smooth functioning as understood from the papers are wether, infrastructure, delay a the origin, speed restriction on the train. The number of the halts a train has is also one of the major factors affecting the train delay as if the train gets delayed at an intermediate station than that delay can subsequently affect the halting time in other coming halts. The weather condition cannot be controlled but various precautionary measures can be taken so that delay time can be decreased. Infrastructure related issues can be also managed and the delay time can be controlled. The research papers have been written based on the Iranian, Thailand, French, Dutch and Indian Railway data. Much little work has been done on Indian Railways.

V. DISCUSSION

The machine learning algorithms are used to make the predictions and most of the authors have used the historical data for making the prediction. to make the study worthy and accurate the real time data can be used to make the predictions. Also specific algorithms are used to implement the study but new algorithms can also be tried to implement the accurateness of the work.

VI. CONCLUSION & FUTURE WORK

It is expected that the trains be on time so that the passengers don't face inconvenience and they can reach their destination on time. This study can help the trains predict the delay time in advance and accordingly manage any other ways to reach the destination on time. Algorithms most commonly used for the predictions are the KNN ,ANN, Decision tree, Multivariate Logistic regression. Among these algorithms the NN has proved out to be the best in the case of accuracy. Incase of the KNN algorithm the value of the k is best suited for the value of 16 and 32 not above. And the factors which are responsible for the delay of the rain are the infrastructure and the climatic condition. The climate factors cannot be controlled but the infrastructure related issues can be resolved and the delay can be minimized. Other issues like chain pulling , riots can also be considered

REFERENCES

[1] (n.d.). Retrieved from https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial/

[2] Alto, V. (2020, january 11). Retrieved from https://towardsdatascience.com/understanding-adaboost-for-decision-tree-ff8f07d2851

[3] Education, I. C. (2020, 8 17). Retrieved from https://www.ibm.com: https://www.ibm.com/cloud/learn/neural-networks

[4] Educator, I. c. (2020, 12). Retrieved from https://www.ibm.com/cloud/learn/random-forest

[5] Heena Gupta, V. S. (March 2021). Train Delay Prediction System in India Using Machine Learning Techniques. *International Advanced Research Journal in Science, Engineering and Technology*, 139-145.

[6] Jiawei Han, M. K. (2012). *Data Mining Concetpts and Techniques.* United States of America: Morgan Kaufmann Publishers.

[7] Lbazri Sara, O. s. (August 2020). predict French train delays using visualization and machine learning techniques. *The International Workshop on Artificial Intelligence and Smart City Applications (IWAISCA)* (pp. 700-705). Leuven, Belgium: www.elsevier.com.

[8] Lokaiah Pullagura, J. K. (2019). Train Delay Prediction using Machine Learning. *International Journal of Engineering and Advanced Technology (IJEAT)*, 1312-1315.

[9] M. Barstuğan, R. (2014). Comparison of Decision Tree and SVM Based AdaBoost Algorithms on Biomedical Benchmark Datasets. *MBEC.*

 DOI : 10.35291/2454-9150.2021.0538

[10] Masoud Yaghini, M. M. (2012). Railway passenger train delay prediction via neural network model. *JOURNAL OF ADVANCED TRANSPORTATION*, 355-368.

[11] Mohd Arshad, M. A. (2019). Prediction of Train Delay in Indian Railways through Machine Learning Techniques. *International Journal of Computer Sciences and Engineering*, 405-411.

[12] Raj, A. (2020, 5 23). Retrieved from https://towardsdatascience.com/: https://towardsdatascience.com/applied-multivariate-regression-faef8ddbf807

[13] Ramashish Gaurav, B. S. (2018). Estimating Train Delays in a Large Rail Network Using a Zero Shot Markov Model.

[14] SAXENA, S. (2021, 3 16). Retrieved from https://www.analyticsvidhya.com: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/

[15] SRUTHI94. (2021, 5). Retrieved from https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

[16] Suporn Pongnumkul, T. P. (n.d.). Improving Arrival Time Prediction of Thailand's Passenger Trains Using Historical Travel Times.

[17] *The Economics Times*. (2018, August 9). Retrieved from The Economics Times Website: https://economictimes.indiatimes.com

[18] Weiwei Mou, Z. C. (2019). Predictive Model of Train Delays in a Railway System. *8th International Conference on Railway Operations Modelling and Analysis*, (pp. 913-929). RailNorrkroping.

 DOI : 10.35291/2454-9150.2021.0538