The Role of Machine Learning in Predicting Real-Time Passenger Train Delay Using Algorithms

Saragadam Sridhar[1]
Associate Professor.
Miracle Educational Society Group of Institutions

Amara Venkatesh
MCA Student.
Miracle Educational Society Group of Institutions

ABSTRACT

Passenger train delay significantly influences riders' decision to choose rail transport as their mode choice. This article proposes real-time passenger train delay prediction (PTDP) models using machine learning techniques. In this article, the impact on the PTPD models using Real-time based Data-frame Structure (RT-DFS) and Real-time with Historical based Data-frame Structure (RWH-DFS) is investigated. The results show that PTDP models using MLP with RWH-DFS outperformed all other models. The influence of the external variables such as historical delay profiles at the destination (HDPD), ridership, and population, day of the week, geography, and weather information on the real-time PTPD models are also further analysed and discussed. This system is, to improve the accuracy of predicting train arrival delay time is of great significance for improving airport transportation efficiency.In our process, we have to take the input as time series dataset. After that, we have to implement the machine learning algorithms such as logistic regression and random forest. The experimental results shows that the accuracy and error values for each algorithm. The model has good prediction accuracy and can track the trends of multiple delay indicators well.

INTRODUCTION

Transport systems are critical pieces of infrastructure and they have substantially increased in size in many countries worldwide. This includes rail transport systems that have evolved significantly, including to provide long-distance travel services. In Sweden, the total distance travelled by trains increased by 8% between 2013 and 2016. In the United States (U.S.), ridership on state supported routes increased by more than 10%, making it the fastest growing segment of Amtrak's services. On long-distance routes, both ridership and revenue increased in fiscal year 2018 by 6.2% and 7.3%, respectively. To sustain its competitiveness and attract more riders, ensuring a high on-time performance is critical. Poor on-time performance can impact passenger trust and their satisfaction, and it may result in a shift to other modes of transport, especially private vehicles and air transport.

Service disruption is a root cause of lower rail punctuality and customer satisfaction. Major Service disruptions result from various conditions or factors such as accidents, problems in train operation, malfunctioning or damaged equipment, routine maintenance, construction, passenger boarding or alighting, and even extreme weather conditions.

Rail service disruptions directly affect scheduled timetable and inevitably cause train delay. Significant train delay can eventually lead to service loss or even cancellation.

In addition, train delay can also negatively affect connecting trains and passengers' journeys or activities. Thus, delay estimations or predictions can help train operators develop better plans to manage, reschedule, or adjust the timetable of the current and consecutive trains more effectively, as well as to inform passengers in advance so they themselves can adjust their travel plans in time. Using or referring to historical average delay is insufficient to estimate future train delay as passenger train can potentially be affected by different factors such as ridership, accumulated delay from prior trains, or weather conditions.

Passenger train delay prediction (PTDP) has been made and modeled in several ways using a variety of approaches and techniques. A fuzzy Petri net (FPN) model to estimate train delay of the Belgrade rail service (the train primary delays were simulated by a fuzzy Petri net module. It used dispatch simulation software to simulate traffic volume and estimate train delays on single and double track rail lines. Wang and Work indicated that although simulation methods can be used to

estimate complex train operations, they require tremendous effort to configure parameters such as dispatching rules as well as calibrating the models for complex train systems.

Objectives:

The main objective is,

- To forecast or to analyze the train delay based on the train delay dataset.
- To implement the different classification algorithms.
- *Machine Learning Algorithms* can help us plot accurate visual representations of such *train delay.*
- *To enhance the overall classification algorithms performance.*

LITERATURE SURVEY

Title: Train Delay Prediction Systems: A Big Data Analytics Perspective Year: 2019 Author: Luca Oneto a,∗, Emanuele Fumeo a, Giorgio Clerico a, Renzo Canepa b, Federico Papa c,Carlo Dambra c, Nadia Mazzino c, Davide Anguita

Methodology: Current train delay prediction systems do not take advantage of state-of-the-art tools and techniques for handling and extracting useful and actionable information from the large amount of historical train movement's data collected by the railway information systems. Instead, they rely on static rules built by experts of the railway infrastructure based on classical univariate statistic. The purpose of this paper is to build a data-driven Train Delay Prediction System (TDPS) for large-scale railway networks which exploits the most recent big data technologies, learning algorithms, and statistical tools. Proposal has been compared with the current state-of-the-art TDPSs. Results on real world data coming from the Italian railway network show that our proposal is able to improve over the current state-of-the-art TDPSs.

Consequently, with reasonably small modifications, we are able to take advantage of a simple deep architecture by exploiting. Simulations have been performed for all the trains included in the dataset adopting an online-approach that updates predictive models every day, in order to take advantage of new information as soon as it becomes available

- Prediction is poor.

Title: Train Time Delay Prediction for High-Speed Train Dispatching Based on Spatio-Temporal Graph Convolutional Network

Year: 2020

Author: Bo Zhang, Dandan Ma

Methodology:

In this paper, we don't try to predict the specific delay time of one train, but predict the collective cumulative effect of train delay over a certain period, which is represented by the total number of arrival delays in one station. We propose a deep learning framework, train spatio-temporal graph convolutional network (TSTGCN), to predict the collective cumulative effect of train delay in one station for train dispatching and emergency plans. The proposed model is mainly composed of the recent, daily and weekly components.

The above methods have advantages in traffic flow prediction, it is not suitable for train delay prediction in high-speed railway network because they only establish the relationship between nodes through graph structure and ignore the influence of distance. Training time is high.

Title: Modeling train operation as sequences: A study of delay prediction with operation and weather data Year: 2020 Author: Ping Huang a,b, Chao Wen a,b,c*, Liping Fu b , Javad Lessan b , Chaozhe Jiang a *, Qiyuan Peng a , and Xinyue Xu

This paper presents a carefully designed train delay prediction model, called FCLL-Net, which combines a fully-connected neural network (FCNN) and two long short-term memory (LSTM) components, to capture operational interactions. The performance of FCLL-Net is tested using data from two high speed railway lines in China. The results show that FCLL-Net has significantly improved prediction performance, over 9.4% on both lines, in terms of the selected absolute and relative metrics compared to the commonly used state-of-the-art models. Additionally, the sensitivity analysis demonstrates that interactions of train operations and weather-related features are of great significance to consider in delay prediction models.

The primary advantages of the present work include: 1) the ability of the proposed model to feed operational and non-operational factors into corresponding units to efficiently recognize their respective influences, and 2) the use of adjacent trains as a group to predict individual train delays, which can be regarded as sequences, to uncover cumulative interactions within the train operation data. The big disadvantage is Error rate is high.

Title: Prediction of Train Arrival Delay Using Hybrid ELM-PSO Approach

Year: 2011 Author: Xu Bao,1Yanqiu Li,2Jianmin Li,2Rui Shi,2and Xin Ding

In this study, a hybrid method combining extreme learning machine (ELM) and particle swarm optimization (PSO) is proposed to

forecast train arrival delays that can be used for later delay management and timetable optimization. First, nine characteristics (e.g., buffer time, the train number, and station code) associated with train arrival delays are chosen and analysed using extra trees classifier. Next, an ELM with one hidden layer is developed to predict train arrival delays by considering these characteristics mentioned before as input features. Furthermore, the PSO algorithm is chosen to optimize the hyper parameter of the ELM compared to Bayesian optimization and genetic algorithm solving the arduousness problem of manual regulating. Finally, a case is studied to confirm the advantage of the proposed model.

- ELM has the advantages of small computation, good generalization, and fast convergence.
- PSO algorithm is a random and parallel optimization algorithm, which has the advantages of fast convergence speed and simple algorithm. The big disadvantage is Training time is high

3.5 Title: Train Delay Estimation in Indian Railways by Including Weather Factors through Machine Learning Techniques

Year: 2019 **Author**: Mohd Arshad1,* and Muqeem Ahmed2

Methodology:

Railway systems all over the world face an uphill task in preventing train delays. Categorically in India, the situation is far worse than other developing countries due to the high number of passengers and poor update of the previous system. As per a report in Times of India (TOI), a daily newspaper, around 25.3 million people used to travel by train in 2006 which drastically increased year on year to 80 million in 2018. : Deploy Machine learning model to predict the delay in arrival of train(s) in minutes, before starting the journey on a valid date. In this paper we combined previous train delay data and weather data to predict delay. In the proposed model, we use 4 different machine learning methods (Linear regression, Gradient Boosting Regression, Decision Tree and Random Forest) which have been compared with different settings to find the most accurate method. Linear Regression gives 90.01% accuracy, while Gradient Boosting Regressor measure 91.68% and the most accurate configuration of decision tree give 93.71% accuracy. When the researcher implemented the ensemble method, Random forest regression, the researcher achieved 95.36% accuracy.

- The model accuracy and training time may be improved over met heuristic methods

such as genetic algorithms. Prediction is not accurate.

Proposed System

Train delay is a major problem in the aviation sector. In proposed system, we have to use the train delay dataset. After that, we have to implement the pre-processing step. In this step, we have to implement the handling missing values for avoid wrong prediction and label encoder for machine readable. After that,

we have to implement the different classification algorithms such as Random forest and logistic Regression for analysing or forecasting the train delay. Finally, the experimental results shows that the accuracy, precision, recall and f1 score. Then, we can predict the train is arrived (on-time or before or late) effectively.

High accuracy is performed for both supervised.It also display the visual graphs. (i.e) comparison graph.

It is efficient for large number of dataset.

The process is implemented with removing unwanted data.

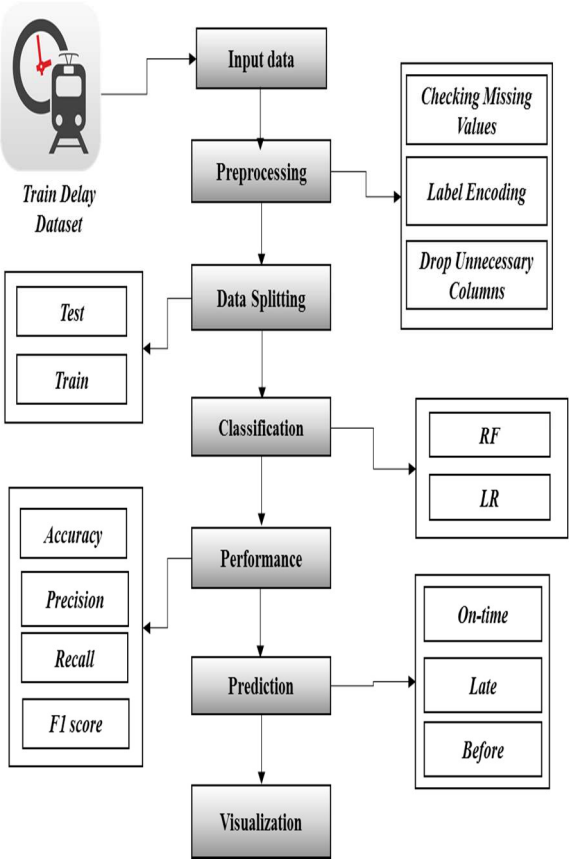


Fig: System Architecture

Algorithms

INPUT DATA:

The data selection is the process of selecting the data for forecasting the train delay.

In this system, the time series dataset is used for predicting the train delay.

The dataset which contains the information about the train such as arrival time, starting time, status and so on.

In python, we have to read the dataset by using the panda's packages.

Our dataset, is in the form of '.csv' file extension.

4.2.2: PREPROCESSING:

Data pre-processing is the process of removing the unwanted data from the dataset.

Pre-processing data transformation operations are used to transform the dataset into a structure suitable for machine learning.

Missing data removal: In this process, the null values such as missing values and Nan values are replaced by 0.

Encoding Categorical data: That categorical data is defined as variables with a finite set of label values.

4.2.3: DATA SPLITTING:

- During the machine learning process, data are needed so that learning can take place.
- In addition to the data required for training, test data are needed to evaluate the performance of the algorithm in order to see how well it works.

In our process, we considered 70% of the dataset to be the training data and the remaining 30% to be the testing data.

Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes.

One Portion of the data is used to develop a predictive model and the other to evaluate the model's performance.

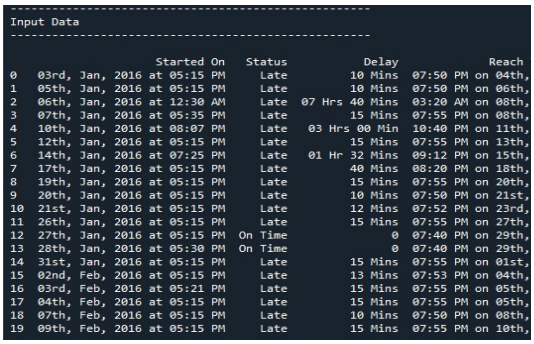
CLASSIFICATION:

In our process, we have to implement the machine learning algorithm such as RF and LR.

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

Random forest improves on bagging because it decorrelates the trees with the introduction of splitting on a random subset of features.

- This means that at each split of the tree, the model considers only a small subset of features rather than all of the features of the model.
- **Logistic regression** is a Machine Learning classification algorithm that is used to predict the probability of certain classes based on some



Input Data

	Started On	Status	Delay	Reach
0	03rd, Jan, 2016 at 05:15 PM	Late	10 Mins	07:50 PM on 04th,
1	05th, Jan, 2016 at 05:15 PM	Late	10 Mins	07:50 PM on 06th,
2	06th, Jan, 2016 at 12:30 AM	Late	07 Hrs 40 Mins	03:20 AM on 08th,
3	07th, Jan, 2016 at 05:35 PM	Late	15 Mins	07:55 PM on 08th,
4	10th, Jan, 2016 at 08:07 PM	Late	03 Hrs 00 Min	10:40 PM on 11th,
5	12th, Jan, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 13th,
6	14th, Jan, 2016 at 07:25 PM	Late	01 Hr 32 Mins	09:12 PM on 15th,
7	17th, Jan, 2016 at 05:15 PM	Late	40 Mins	08:20 PM on 18th,
8	19th, Jan, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 20th,
9	20th, Jan, 2016 at 05:15 PM	Late	10 Mins	07:50 PM on 21st,
10	21st, Jan, 2016 at 05:15 PM	Late	12 Mins	07:52 PM on 23rd,
11	26th, Jan, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 27th,
12	27th, Jan, 2016 at 05:15 PM	On Time	0	07:40 PM on 29th,
13	28th, Jan, 2016 at 05:30 PM	On Time	0	07:40 PM on 29th,
14	31st, Jan, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 01st,
15	02nd, Feb, 2016 at 05:15 PM	Late	13 Mins	07:53 PM on 04th,
16	03rd, Feb, 2016 at 05:21 PM	Late	15 Mins	07:55 PM on 05th,
17	04th, Feb, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 05th,
18	07th, Feb, 2016 at 05:15 PM	Late	10 Mins	07:50 PM on 08th,
19	09th, Feb, 2016 at 05:15 PM	Late	15 Mins	07:55 PM on 10th,

dependent variables.

- In short, the logistic regression model computes a sum of the input features (in most cases, there is a bias term), and calculates the logistic of the result.

PERFORMANCE METRICS:

- The Final Result will get generated based on the overall classification and prediction. The performance of this proposed approach is evaluated using some measures like,

- **Accuracy**

Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

AC= (TP+TN)/ (TP+TN+FP+FN)

- **Precision**

Precision is defined as the number of true positives divided by the number of true positives plus the number of false positives.

Precision=TP/ (TP+FP)

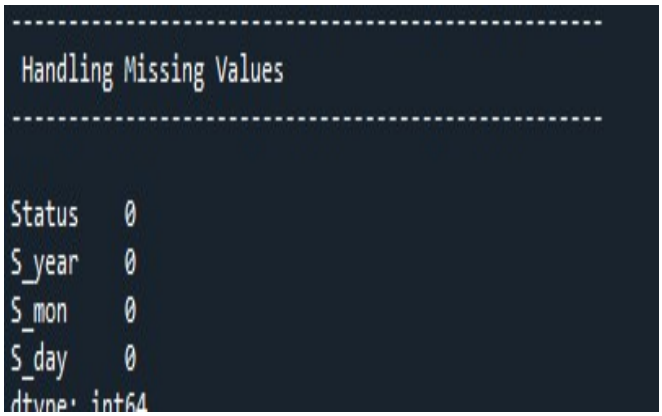
- **Recall**

Recall is the number of correct results divided by the number of results that should have been returned. In binary classification, recall is called sensitivity. It can be viewed as the probability that a relevant document is retrieved by the query.

Recall=TP/ (TP+FN)

RESULTS

Data pre-processing is the process of removing the unwanted data from the dataset. That most machine learning algorithms require numerical input and output variables. Encoding Categorical data: That categorical data is defined as variables with a finite set of label values.



Handling Missing Values

```
Status    0
S_year    0
S_mon     0
S_day     0
dtype: int64
```

Fig: Checking Missing Values

Missing data removal, Missing data removal: In this process, the null values such as missing values and Nan values are replaced by 0.

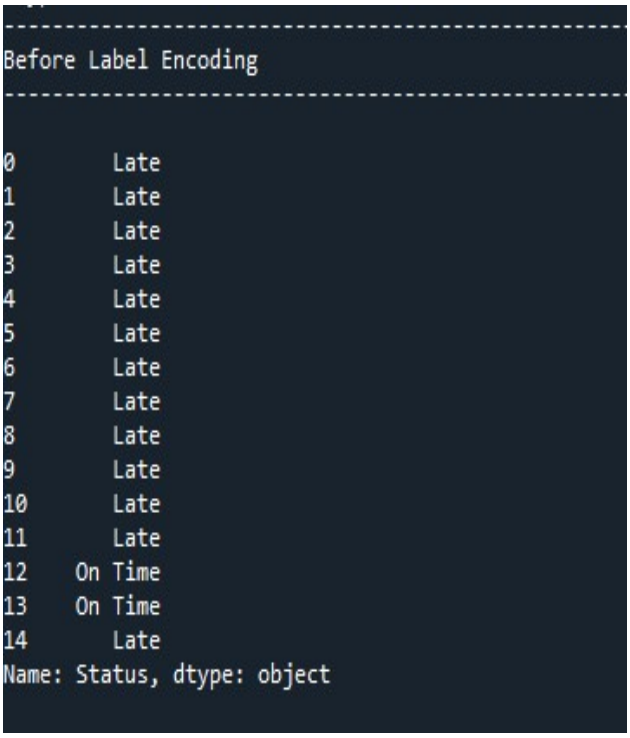


fig Label Encoding

Missing and duplicate values were removed and data was cleaned of any abnormalities.

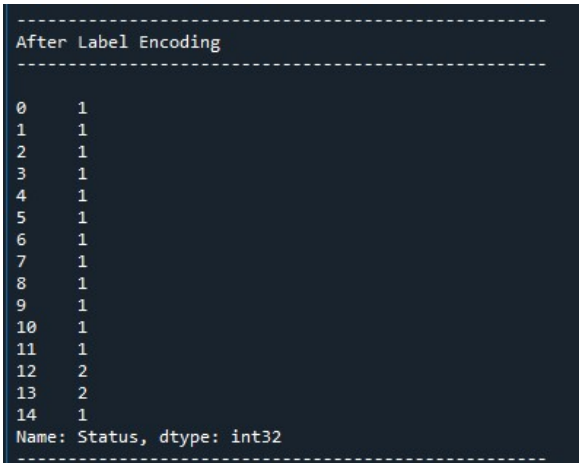


Fig Label Conding

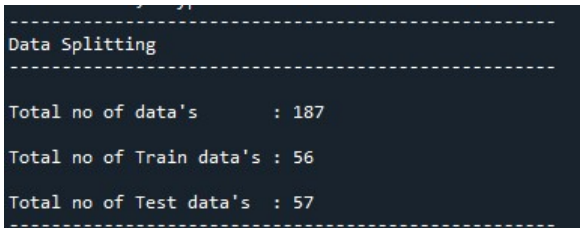


Fig: Data Splitting

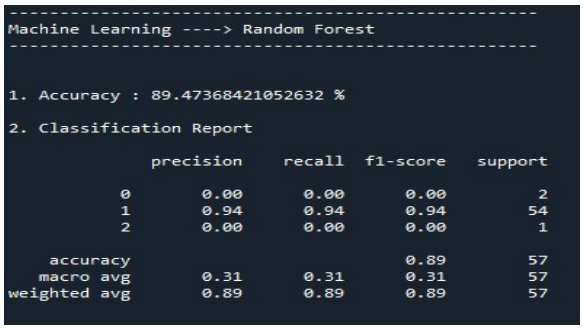


Fig: Classification

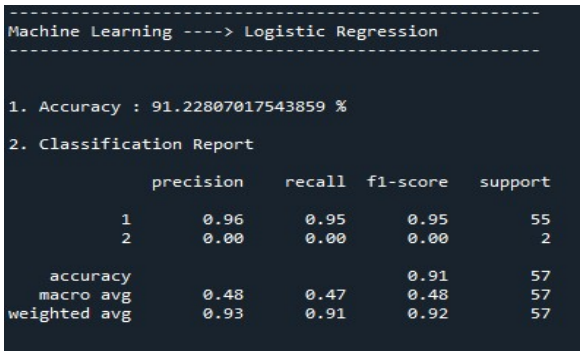
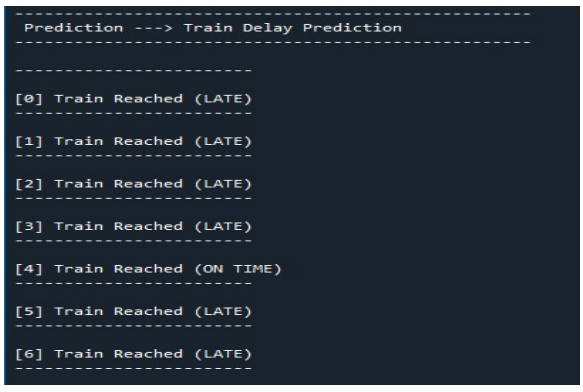


Fig: classification



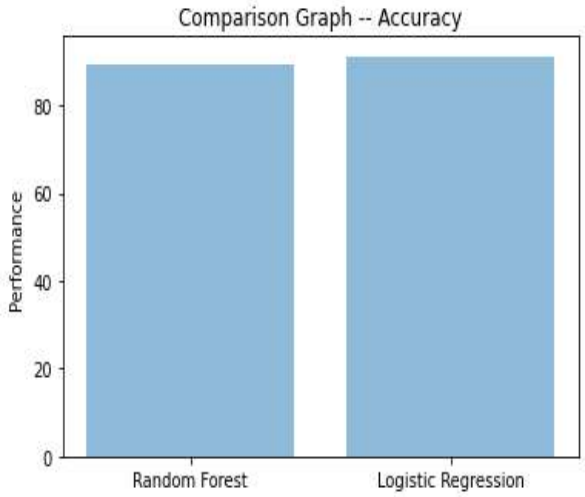
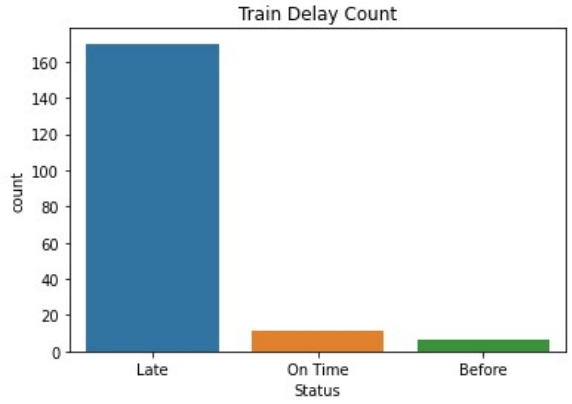


Fig: performance metrics



Conclusion

We conclude that, the input dataset was taken from dataset repository. We are developed the different classification algorithms such as logistic regression and random forest.

Finally, the result shows that some performance metrics such as Accuracy, precision, recall and f1 score. Then, we are forecast or analysed the train delay and visualization. In the future, we should like to hybrid the two different machine learning. In future, it is possible to provide extensions or modifications to the proposed clustering and classification algorithms to achieve further increased performance.

Apart from the experimented combination of data mining techniques, further combinations and other clustering algorithms can be used to improve the detection accuracy.

REFERENCES

[1] S. Derrible, Urban Engineering for Sustainability. Cambridge, MA, USA: MIT Press, 2019.

[2] R. Nilsson and K. Henning. Predictions of Train Delays Using Machine Learning. 2018. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-230224 (Accessed: Jul. 27, 2019).

[3] Amtrak Five Year Service Line Plans FY20-24, Amtrak, Washington, DC, USA, 2019.

[4] N. O. E. Olsson and H. Haugland, "Influencing factors on train punctuality—Results from some Norwegian studies," Transp. Policy, vol. 11, no. 4, pp. 387–397, Oct. 2004, doi: 0.1016/j.tranpol.2004.07.001.

[5] W. Peetawan and K. Suthiwartnarueput, "Identifying factors affecting the success of rail infrastructure development projects contributing to a logistics platform: A Thailand case study," Kasetsart J. Social Sci., vol. 39, no. 2, pp. 320–327, 2018, doi: 10.1016/j.kjss.2018.05.002.

[6] P. Wang and Q. Zhang, “Train delay analysis and prediction based on big data fusion,” Transp. Safety Environ., vol. 1, no. 1, pp. 79–88, Jul. 2019, doi: 10.1093/tse/tdy001.

[7] L. Oneto et al., “Train delay prediction systems: A big data analytics perspective,” Big Data Res., vol. 11, pp. 54–64, Mar. 2018, doi: 10.1016/j.bdr.2017.05.002.

[8] Z. Alwadood, A. Shuib, and N. A. Hamid, “Rail passenger service delays: An overview,” in Proc. IEEE Bus. Eng. Ind. Appl. Colloquium (BEIAC), Apr. 2012, pp. 449–454, doi: 10.1109/BEIAC.2012.6226102.

[9] S. Milinkovic, M. Markovi ´ c, S. Veskovi ´ c, M. Ivi ´ c, and ´ N. Pavlovic, “A fuzzy Petri net model to estimate train ´ delays,” Simulat. Model. Pract. Theory, vol. 33, pp. 144–157, Apr. 2013.

[10] B. W. Schlake, C. P. L. Barkan, and J. R. Edwards, “Train delay and economic impact of in-service failures of railroad rolling stock,” Transport. Res. Rec., vol. 2261, no. 1, pp. 124–133, Jan. 2011, doi: 10.3141/2261-14.

[11] R. Wang and D. B. Work, “Data driven approaches for passenger train delay estimation,” in Proc. IEEE 18th Int. Conf. Intell. Transport. Syst., Sep. 2015, pp. 535–540, doi: 10.1109/ITSC.2015.94.

[12] L. Oneto et al., “Advanced analytics for train delay prediction systems by including exogenous weather data,” in Proc. IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA), Oct. 2016, pp. 458–467, doi: 10.1109/DSAA.2016.57.

[13] A. Gal, A. Mandelbaum, F. Schnitzler, A. Senderovich, and M. Weidlich, “Traveling time prediction in scheduled transportation with journey segments,” Inf. Syst., vol. 64, pp. 266–280, Mar. 2017, doi: 10.1016/j.is.2015.12.001.

[14] A. Estes, M. O. Ball, and D. Lovell, “Predicting performance of ground delay programs,” presented at the 12th USA/Europe air traffic management R&D seminar, Seattle, WA, USA, 2017.

[15] R. Gaurav and B. Srivastava, “Estimating train delays in a large rail network using a zero shot Markov model,” Jun. 2018, arXiv:1806.02825.

[16] R. Nair et al., “An ensemble prediction model for train delays,” Transport. Res. C, Emerg. Technol., vol. 104, pp. 196–209, Jul. 2019, doi: 10.1016/j.trc.2019.04.026.

[17] P. Taleongpong, S. Hu, Z. Jiang, C. Wu, S. Popo-Ola, and K. Han, “Machine learning techniques to predict reactionary delays and other associated key performance indicators on British railway network,” J. Intell. Transport. Syst., vol. 26, no. 3, pp. 311–329, Dec. 2020, doi: 10.1080/15472450.2020.1858822.

[18] C. M. Zappi, Amtrak Host Railroad

Report Card 2019, Amtrak, Washington, DC, USA, Jan. 2020.

[19] "Amtrak Host Railroad Report Card 2020." Amtrak. Apr. 2021. [Online]. Available: https://www.amtrak.com/content/dam/projects/ dotcom/english/public/documents/corporate/ HostRailroadReports/ Amtrak-2020-Host-Railroad-Report-Card-FAQs.pdf (Accessed: Aug. 17, 2021).

[20] P. Lapamonpinyo, S. Derrible, and F. Corman. "A Python Tool and Database of Amtrak Departure and Arrival Times with Weather Information." Oct. 19, 2021. [Online]. Available: https://engrxiv.org/index.php/engrxiv/preprint/view/