OPEN ACCESS

Volume: 11

Special Issue: 1

Month: July

Year: 2023

E-ISSN: 2582-0397

P-ISSN: 2321-788X

Impact Factor: 3.025

Received: 07.05.2023

Accepted:18.06.2023

Published: 01.07.2023

Citation:
Subburaj, T., and RS Dhanushree. "Real-Time Passenger Train Delay Prediction Using Machine Learning: A Case Study with Amtrak Passenger Train Routes." *Shanlax International Journal of Arts, Science and Humanities*, vol. 11, no. S1, 2023, pp. 77–82.

DOI:
https://doi.org/10.34293/sijash.v11iS1-July.6319

Real-Time Passenger Train Delay Prediction Using Machine Learning: A Case Study with Amtrak Passenger Train Routes

Dr. T. Subburaj
Department of Masters of Computer Applications
Rajarajeswari College of Engineering

Dhanushree R.S
Department of Masters of Computer Applications
Rajarajeswari College of Engineering

Abstract

The traveller train delays have a substantial impact on users' decision to use rail transit. Using methods of machine learning, this paper provides real-time passenger train delay prediction (PTDP) models. The influence on PTPD models employing Real-time based Data-frame Structure (RT-DFS) and Real-time with Historical based Data-frame Structure (RWH-DFS) is looked at in this essay. The outcomes prove that PTDP models that combine MLP and RWH-DFS outperform all other models. The result of external variables such as historical delay profiles at the destination (HDPD), ridership and population, day of the week, geography, and weather data on real-time PTPD models is also examined and explored. This system's ability to improve the precision of anticipating train arrival delay time is critical for airport improvement.

Transportation effectiveness.In our procedure, we must use to increase the accuracy of dataset as input. Following that, we must incorporate using machine learning like logistic regression and random forest. The experimental findings reveal that each algorithm's accuracy and error values are different. The model has a high forecast accuracy and can accurately follow the trends of several delay indicators.

Keywords: Train Delay Forecast, Machine Learning, Logisting Regression Random Forecast.

Introduction

Transport networks are crucial elements of infrastructure that have grown significantly in many nations throughout the world. Rail transport systems have advanced tremendously, including the ability to provide long-distance transit services. Between 2013 and 2016, the overall distance travelled by rail in Sweden grew by 8%. 1Ridership on state-supported lines climbed by more than 10% in the United States (U.S.), making it the fastest growing component of Amtrak's services. Ridership and revenue on long-distance routes climbed by 6.2% and 7.3%, respectively, in fiscal year 2018. Maintaining competition and attracting new riders requires a good on-time performance. Poor on-time performance can have an influence on passenger trust and happiness, and it may lead to a move to alternative means of transportation.

Lower train punctuality and customer satisfaction are caused by service interruption. Accidents, issues with train operation, malfunctioning or broken equipment, normal maintenance, construction, passenger boarding or alighting, and even harsh weather can cause major service interruptions.

3Rail service interruptions with instant effect on the scheduled timetable and eventually result in train delays. Significant railway delays might finally result in service interruption or cancellation. Furthermore, train delays might have a severe impact on connecting trains and passengers' itineraries or activities. Thus, delay estimations or projections can assist train operators in developing better plans to more efficiently manage, reschedule, or alter the itinerary of the current and subsequent trains, in addition to alert passengers in advance that they can modify their travel plans in time. Using or referring to historical average delay is insufficient for estimating future train delay since passenger trains can be impacted by a variety of factors such as ridership, cumulative delay from previous trains, or weather conditions.

Several methodologies and strategies have been used to make and model prediction of passenger train delays (PTDP). To forecast train behaviour, a fuzzy Petri net (FPN) model is used. delays. The primary aim is to forecast or analyse the delayed train based on the delayed train dataset.

- To put the various categorization algorithms into action.
- MachineLearningAlgorithms can assist us in creating realistic visual representations of train delays.
- To enhance the overall performance of classification algorithms.

Literature Survey

Current train[1] delay prediction systems do not make use of cutting-edge tools and techniques for processing and extracting relevant and actionable information from the massive amounts of historical train movement data acquired by railway information systems. Instead, they rely on static rules developed by railway infrastructure specialists based on basic univariate statistics. (TDPS) for large-scale railway networks that makes use of cutting-edge big data technologies, learning algorithms, and statistical tools. The proposal was compared against the most recent state-of-the-art TDPSs. Results using real-world the data from Italian railway network reveal that our concept outperforms current state-of-the-art TDPSs.

FCLL-Net [3]model, which combines a whole-network neural architecture (FCNN) and two long short-term memory (LSTM) components to capture operational interactions, is presented in this research. FCLL-Net's performance is analysed using information from two high-speed railway lines in China. The results demonstrate that FCLL-Net outperforms the frequently used state-of-the-art models by more than 9.4% on both lines, regarding the specified absolute and relative measures. Furthermore, the sensitivity Analysis demonstrates that relationships between train operations and weather-related characteristics are important to account in delay prediction models.

The hybrid technique[4] combining extraordinary learning device (ELM) and particle swarm optimisation (PSO) is presented in this work to gauge arrival of train delays, which may then be utilised for subsequent delay management and timetable optimisation. First, nine factors linked with arrival (such as buffer time, train number, and station code) are selected and assessed using an additional trees classifier. Following that, an ELM with one a covert layer is created to forecast delays in train arrival using the previously given characteristics as input features. Furthermore, the PSO method is chosen to optimise the ELM's hyper parameter over Bayesian optimisation and genetic algorithms, alleviating the arduousness of hand regulating. Finally, a case study is conducted to validate the benefit of the suggested paradigm.

Preventing train delays[5] is a difficult challenge for railway networks all around the world. Due to the enormous number of passengers and the prior system's inadequate updating, the situation in

India is significantly worse than in other developing countries. According to a study in the Times of India (TOI), a daily newspaper, around 25.3 million passengers travelled by rail in 2006, which climbed dramatically year on year to 80 million in 2018. Use a ML model to forecast the arrival time of the train(s) in minutes before beginning the travel on a valid date. To estimate delay, we integrated past train delay data and meteorological data in this work. We employ In the suggested model, four different machine learning techniques are used. (linear regression, gradient descent, and decision trees).

Existing System

Currently, passenger train delays have a substantial impact on travellers' decision to choose rail transportation as their means of transportation. In this, following algorithms are suggested: article for real-time passenger train delay prediction (PTDP) models: random forest (RF), gradient boosting machine (GBM), and multi-layer perceptron (MLP). The influence on PTPD models employing Real-time based Data-frame Structure (RT-DFS) and Real-time with Historical based Data-frame Structure (RWH-DFS) is looked at in this essay. 4The outcomes prove that PTDP models that combine MLP and RWH-DFS outperform all other models. The result of external factors on real-time PTPD models, such as historical delay profiles at the destination (HDPD), ridership, population, day of the week, geography, and weather information.

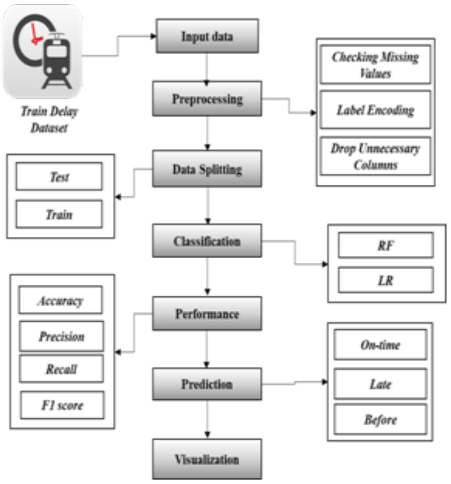

Figure 1 Proposed Architecture

Proposed System

Train delays are a key issue in the aviation industry. 5The train delay dataset must use suggested system. Following that, we must carry out the pre-processing stage. In this phase, we must implement the handling of missing values to avoid incorrect prediction and the label encoder to make it machine readable. 6Following that, we must apply several classification algorithms, such as Random Forest and logistic Regression, to analyse or forecast the train delay. Finally, the experimental findings reveal that the f1 score, accuracy, precision, recall, and are all higher than expected. Then we can properly estimate whether the train will arrive on schedule, early, or late.

Implementation

Input Data

- Data selection refers to the process of selecting data for anticipating railway delays.

- The time series dataset is utilised in this system to anticipate train delays.
- The dataset including train information such as arrival time, departure time, status, so forth.• In Python, we must read the dataset using Panda packages.
- Our dataset includes file extension of '.csv'.

Preprocessing of Data

- Data pre-processing is the removal of undesirable data from a dataset.
- Pre-processing data transformation techniques are working to turn the dataset into a machine-learning-friendly structure.
- Missing data removal: This method replaces null values such as missing values and Nan values with 0.
- Categorical data encoding: Categorical data is defined as variables with a limited number of label values.

Data Splitting

- Necessary is data for ML in order for learning to occur.
- Test data are also necessary to assess the algorithm's performance and determine its effectiveness needed for training.
- In our process, we considered 70% of the 30% of the dataset will serve as training data, and the rest will serve as testing data.
- Splitting accessible data into two parts is called data splitting. portions, usually for cross-validator purposes.
- One a portion from data is utilized to create a predictive model, and the remaining portion using in assess model's performance.

Classification

- We must apply algorithms like RF and LR in our process.
- Random forest outperforms bagging because it decorrelates the trees by splitting on a random selection of features.
- This implies that at each split of the tree, the model considers just a small portion of the model's features rather than all of the model's characteristics.
- A type of Logistic-Regression-Machine-Learning classification approach that predicts the likelihood of specific classes based on various dependent variables.

Results

Input Data

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

Figure 2 Input Data

```
-----------------------------------------------------
Machine Learning ----> Random Forest
-----------------------------------------------------

1. Accuracy : 89.47368421052632 %

2. Classification Report

              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       0.94      0.94      0.94        54
           2       0.00      0.00      0.00         1

    accuracy                           0.89        57
   macro avg       0.31      0.31      0.31        57
weighted avg       0.89      0.89      0.89        57
```

Figure 3 ML Random Forest

```
-----------------------------------------------------
Machine Learning ----> Logistic Regression
-----------------------------------------------------

1. Accuracy : 91.22807017543859 %

2. Classification Report

              precision    recall  f1-score   support

           1       0.96      0.95      0.95        55
           2       0.00      0.00      0.00         2

    accuracy                           0.91        57
   macro avg       0.48      0.47      0.48        57
weighted avg       0.93      0.91      0.92        57
```

Figure 4 Logistic Regression

Conclusion

We infer that the repository from which the input dataset was retrieved of datasets. We created many categorization methods such as Logistic-Regression and Random-Forest.Finally, the results demonstrate that various performance indicators, such as accuracy, precision, recall, and f1 score, are significantly improved. 6The train delay is then forecasted or analysed and visualised. In the future, we hope to combine the two types of machine learning. It is feasible to give enhancements or modifications to the suggested clustering and classification algorithms in the future to obtain even higher performance. Aside from the tried-and-true combination of data mining techniques, additional combinations and clustering algorithms can be utilised to increase detection accuracy.

References

1. S. Derrible, Sustainable Urban Engineering. MIT Press, Cambridge, MA, USA, 2019.
2. R. Nilsson and K. Henning are the authors of this work. Machine Learning Predictions of Train Delays. 2018. [Online]. http://urn. kb.se/resolve?urn=urn:nbn:se:kth:diva-230224 (Accessed: 27 July 2019).
3. Amtrak, Washington, DC, USA, Five Year Service Line Plans FY20-24, 2019.
4. “Influencing factors on train punctuality—Results from some Norwegian studies,” N. O. E. Olsson and H. Haugland. Transportation Policy, vol. 11, no. 4, Oct. 2004, pp. 387-397, doi: 10.1016/j.tranpol.2004.07.001.

5. W. Peetawan and K. Suthiwartnarueput, "Identifying Factors Contributing to a Rail Infrastructure Development Project's Success Logistics Platform: A Thailand Case Study," 39, No. 2, Kasetsart Journal of Social Sciences, p.320-327, doi: 10.1016/j.kjss.2018.05.002.
6. P. Wang and Q. Zhang, "Train late assessment and forecasting based on big data fusion," Transp. Safety Environ., vol. 1, no. 1, July 2019, pp. 79-88, doi: 10.1093/tse/tdy001.
7. Oneto, L. and colleagues, "Train Using big data analytics, describe delay prediction systems. 11th volume of Big Data Research, p. 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 the IEEE Proceedings Bus. Eng. Ind. Appl. Colloquium (BEIAC), April 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, April 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," Transportation Research Record, 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," 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," IEEE Int. Conf. Data Sci. Adv. Anal. (DSAA), Oct. 2016, pp. 458-467, doi: 10.1109/DSAA.2016.57.