International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Train Delay Analysis Using Logistic Regression Approach

Harnoor Huda[1], Aaradhya Chaple[2], Ayush Bhutada[3], Kanak Mishra[4], Shreyas Marudwar[5], Dr. Pradip Selokar[6]

[1,2,3,4,5,6]Shri Ramdeobaba College of Engineering and Management, Nagpur, India

ABSTRACT

Train transportation plays a crucial role in modern society, facilitating the movement of goods and people efficiently. One of the primary challenges faced in railway operations is the occurrence of train delays, which can result from various factors, such as weather conditions, infrastructure issues, or operational constraints. Timely detection and prediction of train delays are essential for ensuring a smooth and reliable rail transportation system. This study explores the use of logistic regression as a method for determining train delays. Logistic regression, traditionally employed for binary classification tasks, is adapted to model the likelihood of train delays based on a combination of relevant features and historical data. The research begins with a comprehensive collection of historical train data and past delays. This data is preprocessed and used to train the logistic regression model. The model is evaluated and fine-tuned to optimize its predictive performance, with metrics like accuracy, precision, and recall considered to assess its effectiveness. The logistic regression model's output provides a probability score for the likelihood of a train delay, which can be used to prioritize and allocate resources effectively. This predictive approach enables railway operators to make informed decisions and implement strategies to prevent or minimize delays, ultimately leading to improved rail transportation efficiency and customer satisfaction.

INTRODUCTION

Train transportation is a vital component of modern infrastructure, facilitating the efficient movement of both people and goods. The reliability and punctuality of train services are crucial to ensuring that transportation networks operate seamlessly. However, train delays continue to challenge the rail industry, arising from various factors, such as inclement weather, infrastructure limitations, and operational constraints. To proactively manage these delays, the application of data-driven methods has become increasingly essential.

This research investigates the use of logistic regression as a powerful tool for predicting and managing train delays, with a specific focus on its merits in comparison to other machine learning algorithms. Logistic regression, traditionally utilized for binary classification, can be adapted to assess the likelihood of train delays. This model's unique characteristics offer several advantages:

Interpretability: Logistic regression provides straightforward and interpretable results. It offers insights into which features contribute to the prediction of train delays, enabling railway operators to understand the factors that most significantly impact their scheduling.

Efficiency: Logistic regression is computationally efficient, making it suitable for real-time applications. It can deliver predictions quickly, allowing for timely decision-making and resource allocation.
Probabilistic Output: Logistic regression produces a probability score, indicating the likelihood of a train delay. This probability information enables operators to prioritize and allocate resources effectively, focusing their efforts where they are most needed.
Applicability to Binary Classification: For tasks such as classifying whether a train will be delayed or not, logistic regression is well-suited due to its inherent binary classification nature. It excels when predicting binary outcomes.
While logistic regression offers several advantages, it is important to recognize that other machine learning algorithms, such as decision trees, random forests, support vector machines, and neural networks, have their strengths and applications in the field of train delay prediction. These alternatives may excel when dealing with complex, non-linear relationships within the data or regression tasks.
This study seeks to assess the effectiveness of logistic regression in train delay prediction and compare its performance against other machine learning algorithms. By doing so, it aims to provide a comprehensive understanding of the strengths and limitations of logistic regression and to determine the most suitable algorithm for different scenarios. The findings will contribute to the enhancement of rail operations, ensuring more reliable and efficient train services for passengers and the broader transportation industry.

Existing System

In existing, 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 the following machine learning techniques: random forest (RF), gradient boosting machine (GBM), and multi-layer perceptron (MLP). 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, population, day of the week, geography, and weather information on the real-time PTPD models.
The system exhibits several disadvantages that significantly impact its overall performance. Firstly, it demonstrates low accuracy, posing a substantial limitation to its reliability and effectiveness. This issue can compromise the system's ability to deliver precise and dependable results, hindering its practical utility. Secondly, the system proves to be inefficient when handling a large volume of data. The challenges associated with scalability raise concerns about its suitability for applications that involve extensive datasets, potentially leading to sluggish performance and compromised efficiency. Additionally, the presence of theoretical limits further constrains the system's capabilities, indicating that certain inherent constraints or boundaries may prevent it from achieving optimal outcomes in certain scenarios. These collective drawbacks underscore the need for improvements and optimizations to enhance the system's overall efficacy and broaden its potential applications.

LITERATURE SURVEY

Comparisons:

Serial Number	References	Advantages	Disadvantages
1	[3] 2021	PSO algorithm is a random and parallel optimization algorithm, which has the advantages of fast convergence speed and simple algorithm.	Training time is high
2	[4] 2019	Consequently, with reasonably small modifications, we are able to take advantage of a simple deep architecture by exploiting.	Prediction is poor.
3	[5] 2019	The model accuracy and training time may be improved over met heuristic methods such as genetic algorithms.	Prediction is not accurate.
4	[6] 2020	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
5	[7] 2020	The ability of the proposed model to feed operational and non-operational factors into corresponding units to efficiently recognize their respective influences	Error rate is high.

The article [8]proposes real-time passenger train delay prediction (PTDP) models using the following machine learning techniques: random forest (RF), gradient boosting machine (GBM), and multi-layer perceptron (MLP). 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 aim of [9] 2019 paper was to present the prediction of Train delay in Indian Railways through machine learning techniques to achieve higher accuracy. In the proposed model, 3 different machine learning methods (Multivariate regression, Neural Network, and Random Forest) were used which were compared with different settings to find the most accurate method. To compare different methods, training time and accuracy of the method over the test data set was compared. In the paper [2] 2019 there was a combination of previous train delay data and weather data to predict delay. In the proposed model, 4 different machine learning methods (Linear regression, Gradient Boosting Regression, Decision Tree and Random Forest) were used which were compared with different settings to find the most accurate method. In the paper [1] 2021 researchers 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.

Proposed System

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

that, we must implement different classification algorithms such as Random Forest and logistic Regression for analyzing or forecasting the train delay. Finally, the experimental results show accuracy, precision, recall and f1 score. Then, we can predict the train will arrive (on-time or before or late) effectively.

The system boasts several notable advantages. Firstly, it excels in achieving high accuracy, particularly in the realm of supervised learning. Additionally, the system is equipped with the capability to generate visual representations, such as comparison graphs, providing users with insightful and easily interpretable results. Its efficiency is particularly pronounced when handling large datasets, ensuring that the processing of substantial amounts of data is both seamless and effective. Moreover, the implementation of the process includes a feature for removing unwanted data, contributing to the system's overall robustness and the production of more refined outcomes. These combined strengths make the system a versatile and powerful tool in the realm of data analysis and machine learning.

Logistic Regression

Logistic regression is a statistical method used for binary classification, which means it's commonly employed to predict the probability of an instance belonging to one of two classes. It is a type of regression analysis that is well-suited for situations where the dependent variable is categorical and represents two classes, such as 0 and 1, Yes and No, True and False, etc.

In logistic regression, the output is transformed using the logistic function (also known as the sigmoid function) to ensure that the predicted values fall between 0 and 1. The logistic function, denoted as "σ(z)", maps any real-valued number "z" to a value in the range [0, 1]. The formula for the logistic function is:

σ(z) = 1 / (1 + e^(-z))

Here, "e" is the base of the natural logarithm.

The logistic regression model computes a weighted sum of the input features and applies the logistic function to it. The mathematical expression for logistic regression can be represented as follows:

p(Y=1|X) = σ(β0 + β1*X1* + *β2*X2 + ... + βn*Xn)

Where:

p(Y=1|X) is the probability that the dependent variable Y is equal to 1 given the independent variable X.

σ() is the logistic (sigmoid) function.

β0, β1, β2, ..., βn are the coefficients of the model, which are estimated during the training process.

X1, X2, ..., Xn are the independent variables or features.

During training, the model learns the values of the coefficients (β) by minimizing a suitable cost function, typically the logistic loss or cross-entropy loss. Once trained, the model can be used to predict the probability of a new data point belonging to one of the two classes.

Logistic regression is widely used in various fields, including medicine, social sciences, finance, and machine learning, where binary classification is a common task. It can be extended to handle multiclass classification using techniques like one-vs-all (OvA) or SoftMax regression.

METHODOLOGY

A logistic regression strategy is used in the rail delay analysis project to predict and identify factors driving train delays. The project's objectives are first explicitly established, laying the groundwork for succeeding processes.

Data collection is an important phase that includes historical train delay records as well as various aspects such as weather conditions, time-related factors, maintenance schedules, and other variables that may affect train operations. Following that, data preprocessing ensures that the dataset is clean and ready for analysis by doing tasks such as addressing missing values, encoding categorical variables, and scaling numerical characteristics.

Following that, exploratory data analysis (EDA) delves into the dataset to uncover insights through visualizations, trend analysis, and anomaly detection. The discovery of significant predictors based on domain expertise, statistical testing, and feature importance analysis is critical in feature selection.

The dataset is then divided into training and testing sets, with 80% usually going to training. For model training, logistic regression is used, using tools such as scikit-learn and fine-tuning hyperparameters as needed. To determine predictive performance, the model is evaluated using the testing set, examining parameters such as accuracy, precision, recall, F1 score, and visualizing the ROC curve.

Understanding the logistic regression model's coefficients, identifying relevant predictors, and utilizing odds ratios to quantify their impact on the likelihood of train delays are all part of the interpretation process. The model may be fine-tuned depending on evaluation results, taking regularization approaches into account, and modifying feature selection.

Validation techniques such as cross-validation verify that the model is generalizable. Data sources, preprocessing methods, model parameters, and evaluation measures must all be documented thoroughly. Deployment may also include incorporating the model into a system for real-time predictions.

The final phases are continuous monitoring and improvement, which ensure that the model remains relevant and effective as new data becomes available. This comprehensive methodology provides a step-by-step guide, with an emphasis on adaptation based on the features of the dataset and project goals.

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 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)

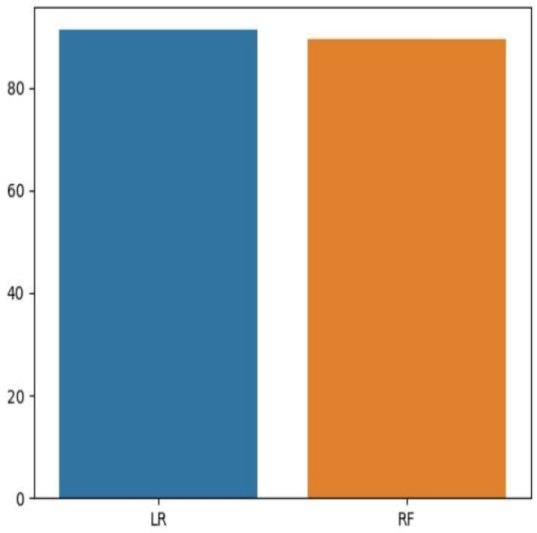

Figure showing accuracy of logistic regression and random forest

CONCLUSION

In this study, we utilized a dataset sourced from a repository to address the pervasive issue of train delays through the implementation of various classification algorithms, including logistic regression and random forest. The input dataset, drawn from a dataset repository, served as the foundation for the development of these algorithms. Our analysis involved evaluating the performance of the models using key metrics such as accuracy, precision, recall, and F1 score. The results gleaned from these metrics provided crucial insights into the effectiveness of the logistic regression and random forest models. Notably, the study concluded with a forecast and analysis of train delays, leveraging visualization techniques to enhance the interpretability of the findings. Through this comprehensive approach, our research not only contributes to the prediction of train delays but also sheds light on the advantages and performance nuances of logistic regression and random forest algorithms in this context, ultimately paving the way for more informed decision-making in the realm of transportation logistics.

REFERENCES

1. Pradhan, Rahul & Kumar, Ashutosh & Kumar, Mayank & Sharma, Bhakti. (2021). Simulating and Analysing delay in Indian Railways. IOP Conference Series: Materials Science and Engineering.

2. Arshad, Mohd & Ahmed, Muqeem. (2019). Train Delay Estimation in Indian Railways by Including Weather Factors Through Machine Learning Techniques. Recent Advances in Computer Science and Communications.
3. Bao, Xu & Li, Yanqiu & Li, Jianmin & Shi, Rui & Ding, Xin. (2021). Prediction of Train Arrival Delay Using Hybrid ELM-PSO Approach. Journal of Advanced Transportation. 2021.
4. Oneto, Luca & Fumeo, Emanuele & Clerico, Giorgio & Canepa, Renzo & Papa, Federico & Dambra, Carlo & Mazzino, Nadia & Anguita, Davide. (2017). Train Delay Prediction Systems: A Big Data Analytics Perspective. Big Data Research.
5. Arshad, Mohd & Ahmed, Muqeem. (2019). Train Delay Estimation in Indian Railways by Including Weather Factors Through Machine Learning Techniques. Recent Advances in Computer Science and Communications.
6. Zhang, Dalin & Peng, Yunjuan & Zhang, Yumei & Wu, Daohua & Wang, Hongwei & Zhang, Hailong. (2021). Train Time Delay Prediction for High-Speed Train Dispatching Based on Spatio-Temporal Graph Convolutional Network. IEEE Transactions on Intelligent Transportation Systems. PP
7. Huang, Ping & Wen, C. & Fu, Liping & Lessan, Javad & Jiang, Chaozhe & Peng, Qiyuan & Xu, Xinyue. (2020). Modeling train operation as sequences: A study of delay prediction with operation and weather data. Transportation Research Part E: Logistics and Transportation Review.
8. P. Lapamonpinyo, S. Derrible and F. Corman, "Real-Time Passenger Train Delay Prediction Using Machine Learning: A Case Study With Amtrak Passenger Train Routes," in IEEE Open Journal of Intelligent Transportation Systems, vol. 3, pp. 539-550, 2022.
9. Arshad, Mohd & Ahmed, Muqeem. (2019). Prediction of Train Delay in Indian Railways through Machine Learning Techniques. International Journal of Computer Sciences and Engineering.