Predicting Railway Fare Costs Using Gradient Boosting Regression: A Data-Driven Approach

Abhijith Mallya
Department of Computer Science (AI & ML)
Sahyadri College of Engineering and Management
Mangalore, India
abhijithmallya@gmail.com

Hithesh Shetty
Department of Computer Science (AI & ML)
Sahyadri College of Engineering and Management
Mangalore, India
shettyhithesh22@gmail.com

Manjunath E C
Department of Computer Science (AI & ML)
Sahyadri College of Engineering and Management
Mangalore, India
manjunath.ai@sahyadri.edu.in

***Abstract*—This paper presents a data-driven approach for predicting railway fare costs using a Gradient Boosting Regression model. The model leverages a comprehensive dataset from the Indian Railway Catering and Tourism Corporation (IRCTC), incorporating journey distance, journey time, and diverse service-related attributes. The study explores the intricacies of travel dynamics, aiming to enhance fare pricing strategies within the railway industry. The dataset is preprocessed, and the model is trained to optimize accuracy and adaptability to dynamic travel scenarios. Results demonstrate the model's promising capability in predicting railway fare costs based on key factors.**

I. Introduction

Railway transportation is a crucial component of modern travel systems, and efficient fare pricing is essential for both service providers and passengers [1], [2]. This section introduces the motivation behind predicting railway fare costs and outlines the significance of employing a data-driven approach. The utilization of machine learning models, particularly Gradient Boosting Regression, is discussed as a key methodology in this study [2], [3].

A. Motivation

Effective fare pricing in the railway industry plays a pivotal role in the overall travel experience. Passengers seek transparency and predictability in fare costs, while service providers aim to optimize revenue and resource allocation [4], [5]. The motivation behind this research is to bridge the gap between passenger expectations and industry objectives by developing a robust prediction model. [6], [7]

B. Objective

The primary objective of this study is to develop a data-driven model that accurately predicts railway fare costs [8]. The focus is on leveraging advanced machine learning techniques, specifically Gradient Boosting Regression, to enhance the precision and adaptability of fare predictions. By achieving this objective, the study aims to contribute valuable insights to the field of transportation and fare pricing strategies [9], [10].

II. Literature Review

Previous research in the field of transportation and fare prediction has explored various methodologies [2], [11]. Some studies have employed machine learning techniques, while others have focused on statistical models. However, the use of Gradient Boosting Regression, especially in the context of railway fare prediction, remains relatively unexplored. [1], [7], [9] This section reviews existing literature, highlighting gaps and paving the way for the methodology adopted in this study.

A. Machine Learning in Transportation

Several studies have demonstrated the effectiveness of machine learning models in predicting transportation-related variables. Notably, models such as Support Vector Machines (SVM) and Neural Networks have been applied to various aspects of travel, including fare prediction [7], [12]. However, the specific application of Gradient Boosting Regression in the context of railway fare prediction is an area that warrants further investigation.

B. Statistical Models in Fare Prediction

Traditional statistical models have been widely used in fare prediction studies. Linear regression, time series analysis, and econometric models have been applied to analyze historical fare data and make predictions [13]–[15]. While these models provide a baseline for understanding fare dynamics, their limitations in handling complex, nonlinear relationships motivate the exploration of advanced machine learning models.

III. Methodology

The methodology involves collecting a diverse dataset from IRCTC, encompassing key variables such as journey distance,

travel time, and service-related attributes. This section elaborates on the data collection process, preprocessing steps, and the application of the Gradient Boosting Regression model.

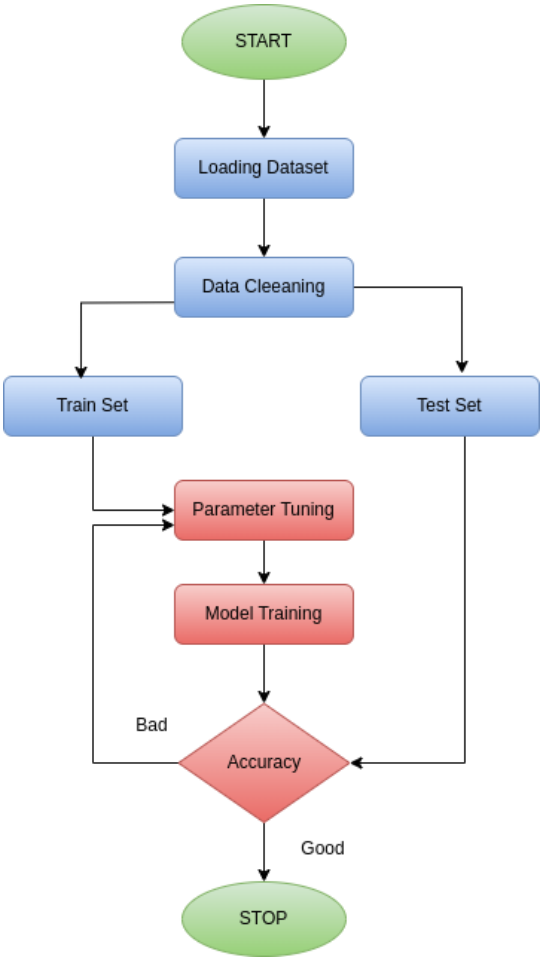

Fig. 1. Illustration of the Data Collection and Preprocessing Process.

A. Data Collection

The dataset used in this study is sourced from the IRCTC, comprising a comprehensive set of features related to railway travel. Key variables include the distance of the journey, the duration of travel, type of service, and historical fare data. The dataset is structured to capture the diverse factors influencing fare costs.

B. Data Preprocessing

To ensure the quality of input data, various preprocessing steps are applied. Outliers are identified and handled using robust statistical methods, and missing values are imputed based on relevant features. The preprocessing phase aims to create a clean and reliable dataset for training the Gradient Boosting Regression model.

C. Gradient Boosting Regression

Gradient Boosting Regression is chosen as the primary predictive model due to its ability to handle complex relationships and adapt to nonlinear patterns in the data. This section provides an overview of the Gradient Boosting algorithm and its application in predicting railway fare costs.

The equation for a single weak learner in Gradient Boosting Regression is given by:

$$F_m(x) = F_{m-1}(x) + \gamma \cdot h_m(x) \tag{1}$$

where:

$F_m(x)$ is the current model prediction at iteration m,
$F_{m-1}(x)$ is the previous model prediction,
γ is the learning rate,
$h_m(x)$ is the weak learner's prediction.

IV. Results and Discussion

The results of the study demonstrate the model's promising capability in predicting railway fare costs. This section presents the quantitative outcomes of the model's performance and delves into a qualitative discussion on the implications of key factors on fare pricing.

A. Decision Tree vs Max Leaf Nodes

Here is a graph comparing the performance of Decision Tree models with different max leaf nodes:

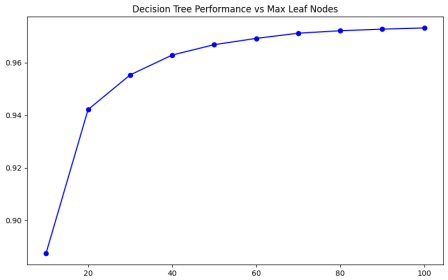

Fig. 2. Decision Tree vs Max Leaf Nodes Performance

B. Quantitative Results

The Gradient Boosting Regression model achieves a high level of accuracy in predicting railway fare costs. Evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) indicate the model's effectiveness in capturing the variability in fare data.

Here is a graph depicting the accuracy variations across different model variations:

C. Key Factors Impacting Fare Pricing

The discussion section analyzes the importance of various factors in influencing fare pricing. Factors such as distance, travel time, service class, and seasonal variations are explored in detail. Understanding the nuanced relationships between these factors is essential for both service providers and passengers.

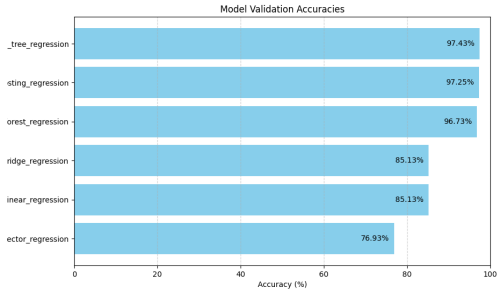

Fig. 3. Accuracy Variation Across Model Variations

D. Implications for the Railway Industry

The study's findings have significant implications for the railway industry. By gaining insights into the factors influencing fare pricing, service providers can optimize their pricing strategies to better align with passenger expectations. This section discusses potential strategies for fare adjustments and dynamic pricing based on the model's predictions.

V. Conclusion and Future Work

In conclusion, this paper presents a robust data-driven approach to predicting railway fare costs. The developed Gradient Boosting Regression model exhibits promising results, providing valuable insights for fare pricing strategies in the railway industry. Future work could explore additional factors, such as passenger demographics and external economic indicators, to further enhance the model's predictive accuracy and applicability.

References

[1] C. Ding, D. Wang, X. Ma, and H. Li, "Predicting short-term subway ridership and prioritizing its influential factors using gradient boosting decision trees," *Sustainability*, vol. 8, no. 11, p. 1100, 2016.

[2] R. Zemel and T. Pitassi, "A gradient-based boosting algorithm for regression problems," *Advances in neural information processing systems*, vol. 13, 2000.

[3] I. Gokasar, A. Karakurt, Y. Kuvvetli, M. Deveci, D. Delen, and D. Pamucar, "Sustainable regional rail system pricing using a machine learning-based optimization approach," *Annals of Operations Research*, pp. 1–28, 2023.

[4] X. Guan, J. Qin, C. Mao, and W. Zhou, "A literature review of railway pricing based on revenue management," *Mathematics*, vol. 11, no. 4, p. 857, 2023.

[5] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[6] C. Li, "A gentle introduction to gradient boosting," *URL: http://www. ccs. neu. edu/home/vip/teach/MLcourse/4_ boosting/slides/gradient_boosting. pdf*, p. 30, 2016.

[7] G. L. Taboada and L. Han, "Exploratory data analysis and data envelopment analysis of urban rail transit," *Electronics*, vol. 9, no. 8, p. 1270, 2020.

[8] S. Guo, C. Chen, J. Wang, Y. Liu, K. Xu, D. Zhang, and D. M. Chiu, "A simple but quantifiable approach to dynamic price prediction in ride-on-demand services leveraging multi-source urban data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, pp. 1–24, 2018.

[9] B. Han, S. Ren, and J. Bao, "Mixed logit model based on improved nonlinear utility functions: a market shares solution method of different railway traffic modes," *Sustainability*, vol. 12, no. 4, p. 1406, 2020.

[10] A. Mowrin, M. Hadiuzzaman, S. Barua, and M. Rahman, "Identifying key factors of commuter train service quality: An empirical analysis for dhaka city," *Malays. J. Civ. Eng*, vol. 31, pp. 23–32, 2019.

[11] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transportation Research Part C: Emerging Technologies*, vol. 58, pp. 308–324, 2015.

[12] P. Prettenhofer and G. Louppe, "Gradient boosted regression trees in scikit-learn," in *PyData 2014*, 2014.

[13] C. Griesbach, B. Säfken, and E. Waldmann, "Gradient boosting for linear mixed models," *The International Journal of Biostatistics*, vol. 17, no. 2, pp. 317–329, 2021.

[14] N. E. Johnson, B. Bonczak, and C. E. Kontokosta, "Using a gradient boosting model to improve the performance of low-cost aerosol monitors in a dense, heterogeneous urban environment," *Atmospheric environment*, vol. 184, pp. 9–16, 2018.

[15] A. Mayr, H. Binder, O. Gefeller, and M. Schmid, "The evolution of boosting algorithms," *Methods of information in medicine*, vol. 53, no. 06, pp. 419–427, 2014.