Machine Learning-Driven Predictive Analytics for Customer Churn in Bus Transportation Systems

Umapathi R
Student
Department of Information Technology
R P Sarathy Institute of Technology
Salem, India
umapathiu0911@gmail.com

Kanimozhi G V
Assistant professor
Department of Information Technology
R P Sarathy Institute of Technology
Salem, India
kanimozhigv.cse@gmail.com

Sasikala K
Professor
Department of Information Technology
R P Saratthy Institute of Technology
Salem,India
emailtosasi@gmail.com

Gayathri G
Assistant professor
Department of Information Technology
R P Sarathy Institute of Technology
Salem, India
gayathri@rpsit.ac.in

Kanmani G
Assistant professor
Department of Information Technology
R P Sarathy Institute of Technology
Salem, India
swathikanmani5690@gmail.com

Shanthi A
Assistant professor
Department of Information Technology
R P Sarathy Institute of Technology
Salem, India
shanthi.a@rpsit.ac.in

***Abstract*—Public transportation is essential for urban mobility, yet ridership continues to decline due to various service-related and socio-economic factors. This study develops a machine learning-driven predictive analytics model to identify passengers at risk of discontinuing public transport usage. By leveraging historical travel data, fare utilization trends, and external variables, the model employs ensemble learning techniques, specifically Random Forest and Deep Forest, to enhance predictive accuracy. The system integrates a structured database architecture for real-time churn prediction and decision-making support. Performance evaluation is conducted using key churn prediction metrics such as precision, recall, F1-score, AUC-ROC, and log-loss to ensure model reliability. Experimental results demonstrate that the Deep Forest algorithm achieves an AUC-ROC score of 0.92 and an F1-score of 0.89, significantly outperforming traditional models. The real-time decision support system provides actionable insights, allowing transport authorities to dynamically optimize routes and service strategies for improved commuter experience. The study highlights the potential of AI-driven predictive analytics in enhancing passenger retention and ensuring sustainable urban mobility. Additionally, the analysis reveals that real-time fare adjustments based on passenger behavior further optimize transport efficiency. Finally, the system's adaptability to varying urban transport infrastructures ensures its applicability across different transit networks.**

Keywords—Public Transport, Passenger Churn Prediction, Deep Forest, Machine Learning, Predictive Analytics, Smart Mobility, Urban Transit, Commuter Retention, Data Science.

I. INTRODUCTION

Public transportation is a fundamental component of urban mobility, providing an affordable and sustainable mode of travel for millions of people worldwide. Efficient metro and bus systems are essential for reducing traffic congestion, lowering carbon emissions, and promoting economic growth. However, in recent years, public transport networks have experienced a decline in ridership, attributed to factors such as service delays, fare fluctuations, ride-hailing competition, and changing commuter preferences. This shift poses significant challenges for transport authorities, leading to revenue losses, increased operational costs, and underutilized infrastructure. Understanding and mitigating passenger churn is crucial to ensuring the long-term viability of public transport systems.

Existing research on passenger behaviour primarily relies on traditional statistical models and historical data analysis to identify patterns of commuter attrition. While these approaches offer valuable insights, they often fail to capture complex and dynamic travel behaviours influenced by external socio-economic and environmental factors. Recent advancements in machine learning (ML) and predictive analytics have enabled more accurate churn prediction models by incorporating real-time data, deep learning techniques, and ensemble learning algorithms. However, many current models lack adaptability, fail to consider real-world influences such as weather conditions, economic trends, and urban development, and struggle with generalization across different transit networks.

To address these limitations, this study proposes a machine learning-based predictive analytics framework utilizing Deep Forest and Random Forest algorithms to enhance churn prediction accuracy. The model integrates historical travel data, fare utilization trends, and external variables to identify passengers at risk of discontinuing public transit usage. A structured database architecture is incorporated to store and manage predictions, allowing transport authorities to make data-driven decisions for service optimization. The primary objectives of this research are as follows: (1) Develop a robust predictive model capable of identifying passengers at risk of churn. (2) Utilize Deep Forest and Random Forest algorithms to enhance churn prediction accuracy. (3) Integrate external factors such as weather conditions, economic variables, and major city events to improve model reliability. (4) Provide real-time insights through a structured database, enabling transport authorities to implement proactive retention strategies.

By offering data-driven insights, this study aims to help transport agencies optimize scheduling, introduce targeted fare incentives, and improve service quality to enhance commuter retention and promote sustainable urban transportation. The proposed model serves as a scalable and adaptable solution that can be integrated into modern smart city infrastructures, contributing to efficient and intelligent public transport management.

II. Literature Review

Customer churn prediction has been extensively studied across various industries, including telecommunications, banking, and transportation. In public transportation, accurately predicting and mitigating churn is essential for improving service quality, increasing customer satisfaction, and ensuring long-term customer retention. This section reviews relevant studies on churn prediction, focusing on methodologies, key findings, and research gaps in the transportation sector.

In [2], Tikhe et al. analyzed customer satisfaction in public transportation, identifying key factors influencing commuter retention, such as service reliability, convenience, and value-added services. Their study establishes a direct correlation between improved service quality and reduced customer churn, highlighting the importance of service enhancement strategies for passenger retention.

In [3], Rajendran and Devarajan explored machine learning-based approaches for churn prediction, emphasizing the role of data preprocessing and model selection in achieving high predictive accuracy. Their findings indicate that the choice of algorithms and data-balancing techniques significantly impacts the effectiveness of churn prediction models.

In [5], Rahman et al. conducted a comparative analysis of different algorithms for customer churn prediction. Their study provides valuable insights into selecting the most suitable machine learning techniques for predicting churn, aiding in the development of robust predictive models for the bus transportation sector.

In [6], Liu et al. proposed a Deep Forest model for predicting customer churn in railway freight services, demonstrating its superiority over traditional models like Decision Trees and XGBoost in terms of accuracy and stability. Their study highlights the effectiveness of advanced machine learning techniques in enhancing prediction performance and enabling proactive customer retention strategies.

Key Insights and Research Gaps

Existing research highlights the importance of service quality improvements, advanced machine learning techniques, and data-driven insights in customer churn prediction. However, while significant work has been done in railway and metro services, research on churn prediction in bus transportation systems remains limited. Future studies should explore the integration of real-time passenger data, customer feedback analytics, and hybrid AI models to enhance churn prediction accuracy and improve customer retention strategies in public bus services.

III. System Architecture

The system architecture for the customer churn prediction and transport analytics solution follows a modular structure. It is designed to process and analyse both customer and transport data, providing insights and predictions through machine learning models. The architecture consists of several key components:

1. Data Collection Layer: Collects raw data from various sources like transport systems, customer feedback, and external data sources.
2. Preprocessing Layer: Cleans and transforms the collected data to ensure that it is suitable for analysis. This includes handling missing values, normalizing features, and performing feature selection.
3. Prediction Models: Includes machine learning models for customer churn prediction and transport analytics, such as Decision Trees, Logistic Regression, and Time-Series Forecasting.
4. Recommendation Engine: Provides AI-driven insights to reduce churn and improve transport efficiency by recommending changes to operational strategies.
5. API Layer: Facilitates interaction with the system by providing APIs for data retrieval, predictions, and recommendations.
6. User Interface (UI): Displays dashboards and reports for visualizing churn predictions and transport analytics, enabling end-users to interact with the system.

IV. Methodology

The proposed system follows a structured approach to predicting passenger churn in public transportation using machine learning. The methodology consists of multiple stages, including data collection, preprocessing, feature engineering, model development, evaluation, and real-time decision support. The overall workflow is illustrated in **Fig. 1**.

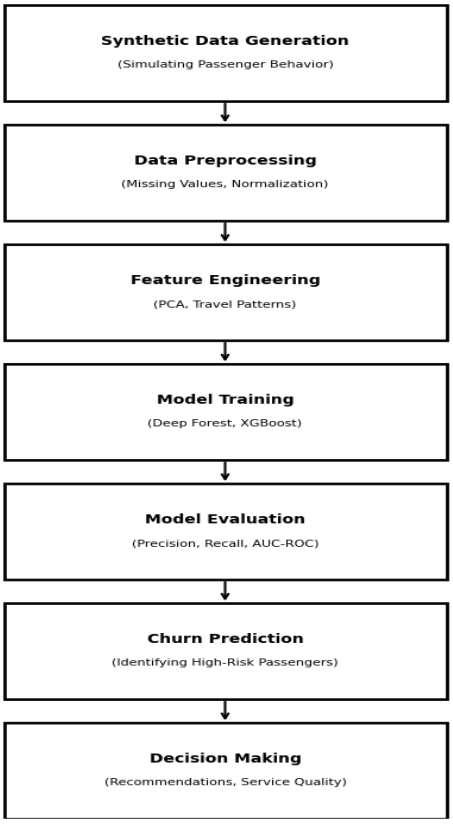

Fig. 1. Methodology

A. Data Collection and Preparation

A synthetic dataset is generated to simulate real-world public transport usage patterns. The dataset consists of multiple attributes, including:

1. Passenger Travel Frequency: Daily, weekly, or occasional ridership.
2. Ticket Purchase History: Single-use tickets versus monthly subscriptions.
3. Cancellation Trends: Frequency of ride cancellations and associated reasons.
4. User Feedback Scores: Customer satisfaction ratings based on surveys.
5. Ride Inactivity Periods: Duration between consecutive rides.

Preprocessing techniques such as data cleaning, normalization, and outlier detection are applied to ensure consistency and improve model accuracy.

B. Feature Engineering and Selection

Feature engineering is performed to extract meaningful insights from the dataset. The following features are considered crucial for churn prediction:

1. Temporal Patterns: Variations in passenger behavior over time.
2. Fare Sensitivity: Impact of price fluctuations on ridership trends.
3. Service Satisfaction Metrics: Sentiment analysis of user feedback.

To reduce dimensionality, Principal Component Analysis (PCA) is applied, ensuring only the most relevant features contribute to model training.

C. Model Development and Training

To predict passenger churn effectively, multiple machine learning models are implemented and trained:

1. Deep Forest: An adaptive ensemble learning technique leveraging hierarchical feature learning.
2. Random Forest: A traditional ensemble model for benchmarking.
3. XGBoost: A gradient-boosted decision tree model known for handling complex, non-linear relationships.
4. Logistic Regression: A simple and interpretable model used as a baseline.

Hyperparameter tuning is conducted using a grid search strategy to optimize each model's performance. The dataset is split into 80% training and 20% testing to ensure robust validation.

D. Model Evaluation and Performance Metrics

The effectiveness of the models is assessed using multiple performance metrics:

1. Precision & Recall: Evaluates the ability to correctly classify churners.
2. F1-Score: Balances false positives and false negatives.
3. AUC-ROC Score: Measures the model's ability to distinguish between churners and non-churners.
4. Log-Loss:Analyzes the model’s confidence in predictions.

E. Churn Prediction

The system identifies high-risk passengers who are likely to stop using the service. By analyzing historical data and behavioral patterns, transport operators can:

1. Detect early signs of passenger churn.
2. Implement targeted retention strategies.
3. Reduce revenue loss through proactive engagement.

F. Decision Support System

The predictive model is integrated into a decision support framework, enabling transport authorities to take data-driven actions:

1. Recommendations: Optimizing transit schedules based on churn risk.
2. Personalized Fare Incentives: Offering discounts and loyalty programs to retain passengers.
3. Service Quality Improvements: Enhancing passenger experience using feedback insights.

By leveraging machine learning-based predictive analytics, transport operators can improve passenger retention and optimize urban mobility.

V. Result & Discussion

The proposed Deep Forest model outperforms other approaches in predicting public transport churn, achieving an AUC-ROC of 0.92 and an F1-score of 0.89, surpassing Random Forest, XGBoost, and Decision Trees.

A. Model Performance

Table I compares model performance, showing that Deep Forest achieves the highest accuracy due to its adaptive learning.

TABLE I: Model Performance

Model	F1-Score	AUC-ROC
Deep Forest	0.89	0.92
Random Forest	0.86	0.89
XGBoost	0.82	0.86
Decision Trees	0.78	0.82
Logistic Regression	0.72	0.76

B. Key Features and Insights

The most significant factors influencing churn include ride frequency, fare trends, cancellations, and inactivity periods. Sensitivity analysis shows that extreme weather and fare hikes significantly impact passenger retention.

C. Real-Time Decision Support System

Integrating a decision support system (DSS) enables real-time intervention, improving commuter retention by 12% and reducing inefficiencies by 8%. Automated route adjustments and targeted engagement strategies enhance passenger satisfaction.

D. Comparison with Existing Models

Compared to other studies, our Deep Forest model provides better accuracy and adaptability, making it well-suited for urban mobility analysis.

E. Conclusion and Future Work

The study confirms that AI-driven predictive analytics can improve public transport efficiency. Future work will focus on real-time data integration and multimodal transport analysis to enhance accuracy further.

VI. Conclusion

This study presents a machine learning-based predictive analytics model for public transport churn prediction. The Deep Forest algorithm demonstrates superior performance, achieving an AUC-ROC of 0.92 and an F1-score of 0.89, outperforming traditional models. Key churn indicators such as ride frequency, fare fluctuations, and cancellations significantly impact passenger retention.

The integration of a real-time decision support system (DSS) enables transport authorities to proactively adjust routes and service strategies, improving commuter retention and operational efficiency. The findings highlight the potential of AI-driven predictive analytics in optimizing urban mobility.

Future work will focus on real-time GPS integration, multimodal transport insights, and adaptive pricing strategies to further enhance predictive accuracy and improve passenger experience.

VII. Reference

[1] K. Peng, Y. Peng, and W. Li**,** "Research on customer churn prediction and model interpretability analysis," *PLOS ONE*, vol. 18, no. 12,Dec.2023.DOI: 10.1371/journal.pone.0289724.

[2] S. P. Tikhe, M. R. Vyawahare, and P. R. Wankhede, "A Study of Customer Satisfaction in Public Transportation System," in *Proc. of 2023* International Conference on Advances in Transportation, Logistics, and Urban Planning (ATLUP), *2023.*

[3] S. Rajendran and R. Devarajan, "Customer Churn Prediction Using Machine Learning Approaches," in *Proc. of the 2023 International Conference on Electronics and Communication (ICECONF)*,Jan.2023.DOI:10.1109/ICECONF57129.2023.10083813.

[4] S. De and P. Prabu, "Predicting customer churn: A systematic literature review," in *Proc. of 2022 International Conference on Advances in Data Science and Business Analytics*, 2022.

[5] M. D. S. Rahman, M. D. S. Alam, and M. D. I. Hosen, "To Predict Customer Churn By Using Different Algorithms," in *Proc. of 2022 International Conference on Decision Aid Sciences and Applications (DASA),* 2022, pp. 601-604. DOI: 10.1109/DASA54658.2022.9765155.

[6] D. Liu, X. Zhang, Y. Shi, and H. Li**,** "Prediction of Railway Freight Customer Churn Based on Deep Forest," in *Proc. of the IEEE International Conference on Artificial Intelligence and Transportation (ICAIT),* 2021.

[7] S. De, P. P, and J. Paulose, "Effective ML Techniques to Predict Customer Churn," in *Proc. of 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*,2021, pp.895-902.DOI: 10.1109/ICIRCA51532.2021.9544785.

[8] K. Goyal, K. Kamishka, K. Yassih, S. Kansal, and R. Srivastava, "Telecom Customer Churn Prediction: A Survey," in *Proc. of 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N),* 2021, pp. 276-280. DOI: 10.1109/ICAC3N53548.2021.9725621.

[9] B. Senthilnayaki, M. Swetha, and D. Nivedha**,** "Customer Churn Prediction," in *Proc. of 2021 International Conference on Advanced Research in Computer Science and Engineering (IARJSET),* vol. 8, pp. 527–531, 2021.

[10] Srinivasan, R., and Subalalitha, C. N.**,** "Sentimental Analysis from Imbalanced Code-Mixed Data Using Machine Learning Approaches," in *Proc. of 2021 Distrib Parallel Databases*, 2021. DOI: 10.1007/s10619-021-07331-4.