Machine Learning-Based Passenger Impact Forecasting for Flight Cancellations Using Operational and Route Analytics

Dr Munidhana Lakshmi[1], Md Asma[2], Dr Humera Anjum[2], Ms Vibhavari[2]

[1]Associate Professor, [2]Assistant Professor,[1,2]Department of Computer Science and Engineering

[1,2]Lords Institute of Engineering and Technology (UGC-Autonomous), Himayath sagar, Hyderabad, 500091, Telangana, India.

To Cite this Article

Dr Munidhana Lakshmi, Md Asma, Dr Humera Anjum, Ms Vibhavari, "Machine Learning-Based Passenger Impact Forecasting for Flight Cancellations Using Operational and Route Analytics", Journal of Science Engineering Technology and Management Science, Vol. 01, Issue 01, September 2024, pp: 364-371, DOI: http://doi.org/10.64771/jsetms.2024.v01.i01.pp364-371

Submitted: 08-08-2026 *Accepted: 19-09-2026* *Published: 25-09-2026*

Abstract

Flight cancellations significantly disrupt airline operations and cause passenger inconvenience, financial losses, and scheduling challenges. Accurate forecasting of the number of passengers affected by flight cancellations is essential for improving operational planning and supporting effective disruption management. This paper proposes a machine learning-based Passenger Impact Forecasting System (PIFS) that estimates the number of passengers affected by flight cancellations using historical flight data, operational attributes, and route analytics. The proposed system considers features such as scheduled departure and arrival times, origin and destination locations, route types, cancellation reasons, refund status, passenger counts, and date-related attributes. The data preprocessing pipeline incorporates time-based feature extraction, categorical encoding, and numerical feature scaling to prepare the dataset for predictive modelling. Four regression approaches, namely Poisson Regressor, Ridge Regressor, Light Gradient Boosting Machine (LightGBM) Regressor, and Stacking Regressor, are implemented and evaluated using Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). The Stacking Regressor combines Huber and Tweedie regressors as base learners with Ridge Cross-Validation (RidgeCV) as the meta-learner to improve prediction performance. Experimental results demonstrate that the Stacking Regressor achieves the best reported performance, obtaining an R^2 score of 0.953, outperforming the other regression models. Furthermore, the system integrates dataset management, model evaluation, passenger impact prediction, and visualization through a graphical user interface. Joblib-based model persistence enables the reuse of trained models and preprocessing components for predicting passenger impact on new flight datasets. The proposed approach supports data-driven airline disruption management, passenger impact assessment, and operational decision-making.

Keywords: Passenger Impact Forecasting, Flight Cancellation Analytics, Airline Operations Management, Machine Learning, Regression Analysis, Stacking Ensemble Learning, Huber Regressor, Tweedie Regressor, LightGBM, Route Analytics, Passenger Impact Prediction.

This is an open access article under the creative commons license https://creativecommons.org/licenses/by-nc-nd/4.0/

CC BY-NC-ND 4.0

1. Introduction

Flight delays and cancellations are unpleasant events in the flight transportation system that may be caused by several factors such as adverse weather conditions, inefficiency in planning, plane

breakdowns, and unexpected events [1]. The occurrence of these events in one flight may indirectly affect a chain of other flights [2]. On the other hand, flight cancellations and delays, in addition to imposing financial losses on airlines, cause passengers' dissatisfaction [3]. In such a situation, advance notice of flight cancellation is very important. In case of information about the possibility of flight cancellation, airline managers can prevent more losses by planning alternative strategies [4]. Also, having a flight cancellation prediction system can be useful in improving flight scheduling [5]. These advantages have caused many researchers to deal with the issue of flight cancellation or delay prediction based on flight data in recent years.

Most of the previous studies in this field have used machine learning (ML) techniques to solve the problem, and the goal of all of them was to achieve a more accurate prediction of flight delay cancellation.

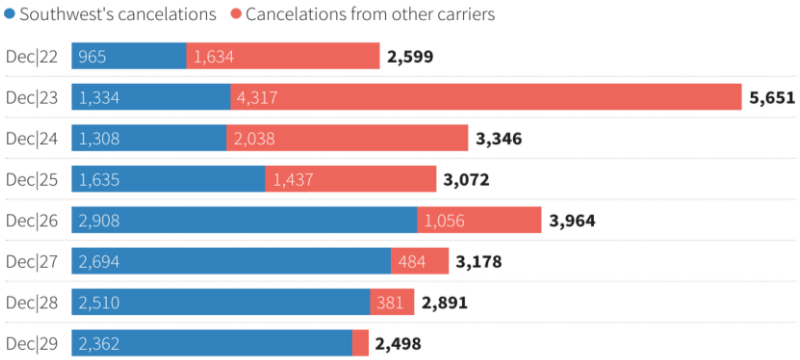

Figure 1: Southwest airlines cancellations.

However, some specific challenges with flight data have made the applicability of these models in real conditions questionable, as shown in Figure 1 The first challenge in this field is the complexity of the problem of flight cancellation/delay prediction. Because this event can be caused by various direct (e.g., aircraft features) and indirect (e.g., delay in previous flights) factors [6].

2. Literature Survey

Lambelho et al. [7] researched a method for predicting flight cancellations or delays using a combination of feature selection techniques and machine learning models. They applied recursive feature elimination to reduce the dataset to 14 features and utilized LGBM, Random Forest (RF), and Multilayer Perceptron (MLP) models. Their results demonstrated the superiority of the LightGBM model over the other approaches. Shu [8] examined the use of machine learning techniques for predicting flight cancellations or delays. However, this study did not include any feature selection or dimensionality reduction strategy, making it difficult to compare its reported accuracy with other works.

Bandyopadhyay et al. [9] developed a deep learning-based approach to predict flight cancellations during the COVID-19 pandemic. Their model used 19 indicators and combined two models into an ensemble framework. Although the model achieved high accuracy, its complexity raised concerns regarding scalability for large datasets. Mohammed et al. [10] investigated the impact of data imbalance in flight cancellation prediction, noting that canceled flights represent less than 2% of total flights. They applied various sampling techniques and evaluated them using multiple machine learning models. Their findings indicated that combining random sampling with an MLP model improved prediction accuracy. Hendrickx et al. [11] explored the issue of data imbalance using an equilibrium-based sampling strategy. Their approach iteratively determined an optimal balance between classes and applied RF and MLP models for prediction, concluding that MLP achieved better performance.

Ayadin and Akcayol [12] proposed a deep learning-based method for predicting flight delay and cancellation using 10 features. They also employed Random Forest models for comparison, and the results showed that RF outperformed deep learning models in terms of overall accuracy. Gong et al. [13] developed a customized recommender system to provide alternative solutions for canceled flights. Their approach used an XGBoost model to process customer and flight data, aiming to maintain customer loyalty by offering personalized recommendations. Taylor et al. [14] designed a decision support system for flight cancellation recommendations. Their system used an RF model to process weather and flight features, while also considering forecast uncertainty and potential losses due to cancellations.

Imran et al. [15] studied flight delay prediction using various machine learning models, including Logistic Regression (LR), Decision Tree (DT), RF, and Gradient Boosting Regression (GBR). Their results showed that the RF model achieved the highest accuracy. Li et al. [16] proposed a hybrid model combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to capture spatial and temporal features. These features were then classified using an RF model to predict flight delays. Wu et al. [17] introduced a Graph Convolutional Network (GCN)-based model for flight delay forecasting. Their model effectively captured both temporal and spatial relationships by modelling them separately.

3. Proposed System

The proposed system transforms flight cancellation management from a manual, reactive process into a data-driven, automated, and predictive solution. It uses historical flight data, operational details, and route analytics to accurately forecast the number of passengers affected by cancellations. Key features such as departure and arrival times, route type, cancellation reasons, refund status, day of the week, and passenger count are collected and stored in a centralized database. The data undergoes preprocessing including feature extraction, encoding of categorical variables, and scaling of numeric features to prepare it for predictive modelling. Multiple ML regression algorithms PR, LGBM Regressor, RR, and SR are trained and evaluated using metrics like MAE, MSE, RMSE, and R^2 to ensure high prediction accuracy as demonstrate in Figure 2.

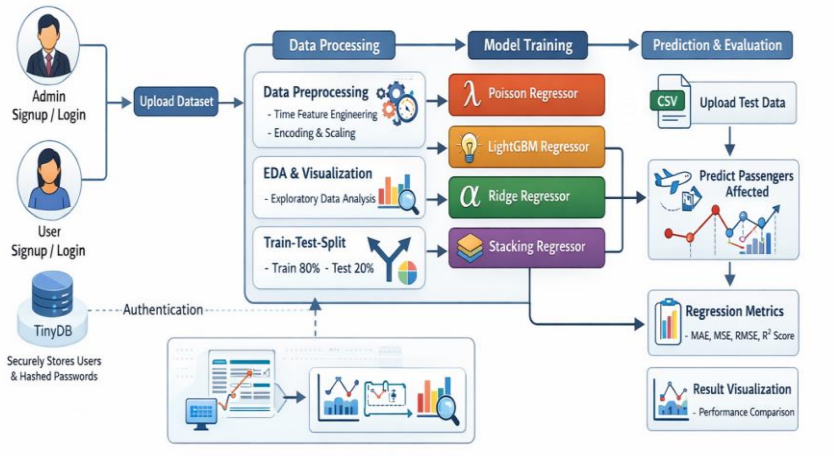

Figure 2: Proposed system architecture of airline cancellations using operational and route analytics.

User Management Module: This module handles signup, login, and role-based access for Admin and User. User credentials are securely stored in a database with hashed passwords, ensuring secure and controlled access.

Data Upload and Storage Module: Allows uploading of training and test datasets and stores them in a centralized database. This ensures data is organized, accessible, and ready for processing and analysis.

Data Preprocessing Module: Prepares the data for modelling by performing feature extraction, encoding of categorical variables, and scaling of numeric features. It ensures consistency between training and test datasets.

Exploratory Data Analysis (EDA) Module: Generates visualizations to identify patterns and relationships in the data. It helps in understanding cancellation trends, passenger impact, and key influencing factors.

Model Training and Evaluation Module: Implements multiple Machine Learning regression algorithms, including PR, LGBM, Ridge, and SRs. Models are evaluated using MAE, MSE, RMSE, and R^2 to select the most accurate predictor.

Prediction Module: Processes new or test datasets to forecast the number of passengers affected by cancellations. Predictions are stored in the database for future reference and analysis.

Visualization and Reporting Module: Produces plots for actual vs predicted values and comparative performance of models. This helps airline staff make data-driven decisions and track trends effectively.

Graphical User Interface (GUI) Module: Provides a user-friendly interface for Admin and User operations. It allows easy access to functionalities like upload, preprocess, train, predict, visualize, and logout.

3.1 SR Model

The SR is an ensemble learning technique that combines multiple regression models to improve predictive performance. Instead of relying on a single model, stacking leverages the strengths of different base regressors and uses a meta-model to learn how to best combine their predictions. In this project, the SR uses Huber Regressor and Tweedie Regressor as base learners, with Ridge Regression as the final estimator. Stacking is a model-agnostic ensemble technique that operates in two levels to improve prediction accuracy. At the base level, multiple individual regression models are trained independently on the original dataset to learn patterns from the features. In this research, the base models include Huber Regressor and Tweedie Regressor. The Huber Regressor is a robust linear regression model that is less sensitive to outliers, combining squared loss for small errors with linear loss for large errors, making it suitable for datasets with irregular passenger impact values.

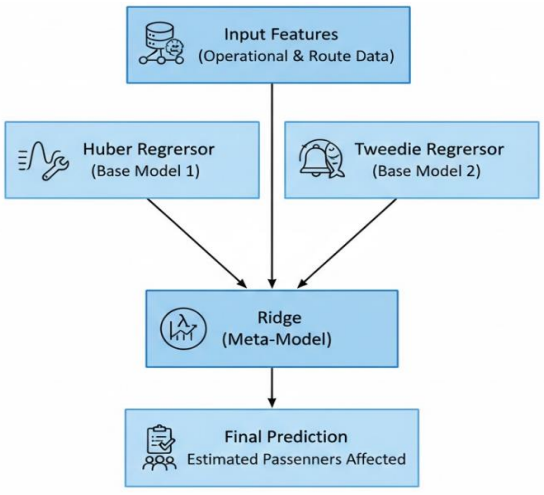

Figure 3: Internal working of SR Model.

The Tweedie Regressor, on the other hand, is a generalized linear model capable of handling continuous, count, and mixed distribution targets, and can effectively model non-negative, skewed data, such as passenger counts with high variability. At the meta level, a final estimator, Ridge Regression, takes the predictions of the base models as input and learns how to optimally combine them, as shown in Figure 3. This two-layer approach reduces bias and variance, leveraging the complementary strengths of the base models to produce a more accurate and reliable forecast of the estimated number of passengers affected by flight cancellations.

4. Result Description

This section presents the experimental results obtained from the implementation of the PIFS for airline cancellations. The results demonstrate the effectiveness of the graphical user interface, data preprocessing pipeline, and machine learning models in predicting the number of passengers affected. Each figure highlights a specific module of the system, from authentication and data handling to model performance evaluation and prediction output. The analysis confirms the practical applicability and reliability of the proposed system in real-world airline operations.

	origin_city	origin_airport_code	destination_city	destination_airport_code	scheduled_departure_time	scheduled_arrival_time	cancellation_reason	route_
0	Hyderabad	HYD	Delhi	DEL	23:30	02:00	Crew Shortage	Internat
1	Bengaluru	BLR	Kolkata	CCU	05:00	07:30	Crew Shortage	Internat
2	Bengaluru	BLR	Jaipur	JAI	17:00	18:30	Crew Shortage	Internat
3	Ahmedabad	AMD	Bengaluru	BLR	06:15	07:15	Regulatory Constraints	Dom
4	Hyderabad	HYD	Mumbai	BOM	04:00	05:00	Crew Shortage	Dom
...	...	...	...	...	...	...	...	
53743	Jaipur	JAI	Kolkata	CCU	03:30	05:00	Operational Issues	Internat
53744	Chennai	MAA	Jaipur	JAI	03:45	04:45	Crew Shortage	Internat
53745	Jaipur	JAI	Mumbai	BOM	11:00	12:30	Operational Issues	Dom
53746	Ahmedabad	AMD	Bengaluru	BLR	06:30	07:30	Crew Shortage	Internat
53747	Pune	PNQ	Jaipur	JAI	08:45	10:00	Operational Issues	Internat

53748 rows × 13 columns

Figure 4: Uploading dataset of flight cancellation.

Figure 4 illustrates the dataset loading stage of the PIFS. The system reads the airline cancellation dataset using pd.read_csv(), which imports the CSV file into a structured DataFrame. This dataset contains attributes such as scheduled departure and arrival times, origin and destination details, route type, cancellation reason, refund status, day, and weekday. The successful loading of the dataset is confirmed by displaying its shape and preview, ensuring that the data is ready for preprocessing and further analysis.

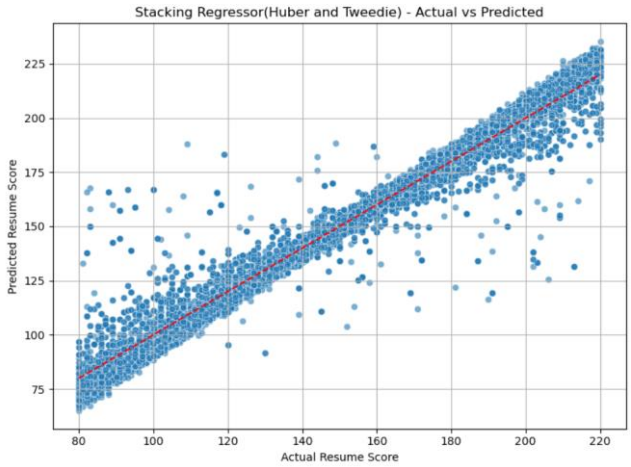

Figure 5: Performance metrics of SR model

Figure 5 presents the performance of the SR model, which combines multiple base learners using StackingRegressor(). The integration of Huber Regressor and Tweedie Regressor with RidgeCV as the meta-learner results in improved predictive accuracy. The evaluation metrics show the lowest error values and highest R^2 score among all models. The actual vs predicted plot demonstrates close alignment along the diagonal line, confirming superior model performance.

scheduled_departure_time	scheduled_arrival_time	cancellation_reason	route_type	refund_status	day	month	weekday	predicted_passengers_affected
19:00	21:00	Regulatory Constraints	International	Pending	5	12	4	194.014066
11:45	12:15	Operational Issues	International	Refunded	21	12	6	167.135691
18:15	20:15	Regulatory Constraints	Domestic	Pending	29	12	0	108.195741
14:30	17:15	Operational Issues	Domestic	Pending	17	12	2	166.322838
18:15	21:45	Crew Shortage	International	Refunded	12	12	4	133.377782
04:00	06:45	Crew Shortage	International	Rescheduled	30	12	1	208.734837
05:30	08:45	Regulatory Constraints	International	Pending	18	12	3	131.722391
01:45	02:15	Regulatory Constraints	Domestic	Refunded	15	12	0	189.591790
15:30	18:30	Operational Issues	Domestic	Rescheduled	2	12	1	205.861837
00:15	02:00	Crew Shortage	Domestic	Rescheduled	28	12	6	101.696513
09:00	12:45	Operational Issues	International	Refunded	24	12	2	68.022860
15:00	17:00	Crew Shortage	Domestic	Pending	28	12	6	91.648640
19:45	21:00	Operational Issues	Domestic	Rescheduled	6	12	5	102.535197
10:45	13:30	Operational Issues	International	Rescheduled	13	12	5	90.969532
05:45	07:00	Regulatory Constraints	International	Rescheduled	11	12	3	176.291284
10:30	13:45	Crew Shortage	International	Rescheduled	7	12	6	99.696326
01:30	03:15	Regulatory Constraints	Domestic	Refunded	28	12	6	73.424229
17:15	18:45	Crew Shortage	Domestic	Rescheduled	23	12	1	76.734341
22:00	00:45	Operational Issues	Domestic	Rescheduled	29	12	0	188.586170
19:45	21:00	Regulatory Constraints	Domestic	Rescheduled	13	12	5	212.484237

Figure 6: Prediction on sample test data.

Figure 6 represents the final prediction output of the system. New input data is loaded using pd.read_csv() and processed using the saved preprocessing pipeline through joblib.load(). The trained SR model generates predictions using model.predict(), and the output is stored in the column predicted_passengers_affected. The final dataset displays predicted passenger impact values for each record, completing the forecasting process.

Table. 1: Performance comparison for the PR, LightGBM Regressor, RR, and SR models.

Algorithms Name	**MAE**	**RMSE**	**R2**
PR Model	0.028	0.025	0.001
LGBM Model	0.032	0.028	0.21
RR Model	0.028	0.025	0.001
SR Model	0.190	0.114	0.953

Table 1 presents the performance comparison of the regression models implemented in the PIFS, including PR, LightGBM Regressor, RR, and SR. The evaluation metrics include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 score, which quantify prediction accuracy and model reliability. PR and RRs show low MAE and RMSE values but very low R^2 scores, indicating limited explanatory power for the variability in passenger impact. LightGBM improves R^2 moderately to 0.21, capturing some nonlinear relationships in the dataset. The SR, which combines Huber and Tweedie base learners with Ridge as the meta-learner, outperforms all models, achieving a high R^2

score of 0.953 while maintaining reasonable error rates. This demonstrates the ensemble model's effectiveness in accurately forecasting the number of passengers affected by flight cancellations.

5. Conclusion

The PIFS successfully predicts the number of passengers affected by flight cancellations by utilizing historical flight data, operational attributes, and route-based features. The system incorporates data preprocessing, EDA, and multiple regression models including PR, RR, LGBM, and SR to ensure accurate forecasting. Among all models, the SR demonstrates the best performance with the highest prediction accuracy and an R^2 score of 0.953, indicating its effectiveness in capturing complex patterns within the data. The implementation follows a structured workflow that includes dataset loading, preprocessing, model training, evaluation, and prediction generation on new data. The use of joblib enables efficient storage and reuse of trained models and preprocessing components. By automating the prediction process, the system eliminates the limitations of manual estimation and provides consistent and reliable results. Overall, the proposed system delivers a robust and scalable solution for passenger impact forecasting, supporting improved operational planning and effective decision-making in airline management during flight disruptions.

References

[1] Gössling, S., Neger, C., Steiger, R. & Bell, R. Weather, Climate Change, and Transport: a Review1–20 (Natural Hazards, 2023).

[2] Carvalho, L. et al. On the relevance of data science for flight delay research: a systematic review. Transp. Reviews. 41 (4), 499–528 (2021).

[3] Zachariah, R. A., Sharma, S. & Kumar, V. Systematic review of passenger demand forecasting in aviation industry. Multimedia Tools Appl., 1–37. (2023).

[4] Afonso, F. et al. Strategies towards a more sustainable aviation: A systematic review. Prog. Aerosp. Sci. 137, 100878 (2023).

[5] Zhou, L., Liang, Z., Chou, C. A. & Chaovalitwongse, W. A. Airline planning and scheduling: models and solution methodologies. Front. Eng. Manage. 7 (1), 1–26 (2020).

[6] Wang, F., Bi, J., Xie, D. & Zhao, X. Flight delay forecasting and analysis of direct and indirect factors. IET Intel. Transport Syst. 16 (7), 890–907 (2022).

[7] Lambelho, M., Mitici, M., Pickup, S. & Marsden, A. Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions. J. Air Transp. Manage. 82, 101737 (2020).

[8] Shu, Z. Analysis of flight delay and cancellation prediction based on machine learning models. In 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI) (pp. 260–267). IEEE. (2021), December.

[9] Bandyopadhyay, S. K., Goyel, V. & DUTTA, S. Prediction of Air Flight Cancellation during COVID-19 Using Deep Learning Methods (ScienceOpen Preprints, 2020).

[10] Mohammed, Z., Asghar, M. & Kanwal, N. Analyzing the impact of COVID-19 on flight cancellation using machine learning and deep learning algorithms for a highly unbalanced dataset. In 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET) (pp. 1–6). IEEE. (2021).

[11] Hendrickx, R., Zoutendijk, M., Mitici, M. & Schäfer, J. Considering Airport Planners' Preferences and Imbalanced Datasets when Predicting Flight Delays and Cancellations. In 2021

IEEE/AIAA 40th Digital Avionics Systems Conference (DASC) (pp. 1–10). IEEE. (2021), October.

[12] Ayadin, A. & Akcayol, M. A. Deep Learning Based Forecasting Of Cancellation, Delay And Orientation On Flights. In 2021 1st International Conference On Informatics And Computer Science (p. 79). (2021).

[13] Gong, Z., Wang, H., Nie, Q., Zhang, Z. & Xiao, Q. The personalized recommendation for OTA flight cancellation and change services during the pandemic. J. Revenue Pricing Manag. 22 (2), 157–165 (2023).

[14] Taylor, C., Tien, S. L., Vargo, E. & Wanke, C. Strategic flight cancellation under ground delay program uncertainty. J. Air Transp. 29 (1), 5–15 (2021).

[15] Imran, S. H., Anjum, A., Ananthasiri, C. H. & Viraja, D. U. Build an machine learning model for prediction flight delays with error calculation. Int. J. Manage. Res. Reviews. 13 (2), 12–17 (2023).

[16] Li, Q., Guan, X. & Liu, J. A CNN-LSTM framework for flight delay prediction. Expert Syst. Appl. 227, 120287 (2023).

[17] Wu, Y., Yang, H., Lin, Y. & Liu, H. Spatiotemporal Propagation Learning for Network-Wide Flight Delay Prediction (IEEE Transactions on Knowledge and Data Engineering, 2023).