Airline Passenger Satisfaction: A Comparative Study of Classification and Regression Models

February 2026

1 Introduction

This study analyzes an airline passenger survey dataset (103,904 training and 25,976 test responses) covering 23 demographic, trip, and service rating features. The primary objective is to predict passenger satisfaction (a binary label: *satisfied* vs. *neutral/dissatisfied*).

The analysis is conducted in four stages: (1) binary logistic regression to identify key satisfaction drivers and address confounding variables; (2) multinomial logistic regression to predict booking class from demographics, flight details, and service-related features (excluding satisfaction, which is a post-flight outcome); (3) a comparison of discriminative models (LDA, QDA, Naïve Bayes); and (4) an evaluation of OLS versus Poisson regression for predicting arrival delay in minutes using all available predictors.

2 Data and Preprocessing

Missing values in *Arrival Delay* (0.30%) were imputed with the training median. Categorical variables were one-hot encoded (dropping one dummy per group to avoid multicollinearity), and numeric features were standardised using training set parameters.

The binary target is roughly balanced (57% neutral/dissatisfied vs. 43% satisfied). However, the booking-class variable exhibits class imbalance, with Eco Plus representing only $\approx$7% of the dataset.

3 Exploratory Data Analysis

Initial EDA reveals that Business travellers and Business-class passengers dominate the satisfied category. A correlation heatmap identifies two distinct clusters of positively correlated features: cabin-comfort items (e.g., seat comfort, cleanliness) and digital/ground services (e.g., online boarding, Wi-Fi). Delay variables are strongly collinear with each other but weakly correlated with service scores.

4 Binary Logistic Regression

4.1 Full Model Analysis

A binary logistic regression model was fit to predict satisfaction (Table 1). Online boarding emerged as the strongest positive predictor, while Personal travel and Disloyal customer status were the strongest negative predictors. Flight distance was statistically insignificant ($p = 0.129$).

Table 1: Binary logistic regression — selected coefficients (full model). N = 103,904, McFadden's pseudo-$R^2 = 0.512$.

Variable	**Coef.**	**Std. Err.**	z	p
Online boarding	0.826	0.014	59.86	$< .001$
Type of Travel: Personal	−1.259	0.015	−86.61	$< .001$
Customer Type: disloyal	−0.787	0.012	−68.09	$< .001$
Inflight Wi-Fi service	0.524	0.015	34.46	$< .001$
Arrival Delay in Minutes	−0.341	0.035	−9.69	$< .001$
Departure Delay in Minutes	0.158	0.035	4.50	$< .001$
Class: Eco	−0.366	0.013	−28.74	$< .001$
Flight Distance	−0.017	0.011	−1.52	0.129

The positive coefficient for *Departure Delay* is counterintuitive. It is a suppressor effect caused by extreme collinearity with *Arrival Delay*. The model artificially boosts the satisfaction of flights that departed late but made up time in the air.

4.2 Removing the Confounding Variable

To resolve the collinearity, we dropped *Departure Delay* (Table 2). The model fit (pseudo-R^2) remained unchanged, while the *Arrival Delay* coefficient shrank to its true direct effect (−0.190) with a significantly smaller standard error.

Table 2: Arrival Delay coefficient before and after removing Departure Delay.

	Full model	**Without Departure Delay**
Arrival Delay coef.	−0.341	−0.190
Arrival Delay std. err.	0.035	0.010
McFadden pseudo-R^2	0.5120	0.5119

5 Multinomial Logistic Regression

We applied multinomial logistic regression (with balanced class weights) to predict the passenger's booking class. To keep the model logically sound and deployable (e.g., at booking time), we do *not* use satisfaction as a feature: satisfaction is surveyed after the flight and would not be available when predicting class. Predictors therefore include demographics (gender, customer type, age, type of travel), flight distance, all 14 service ratings, and both departure and arrival delay in minutes.

Table 3: Multinomial logistic regression test-set report ($N = 25{,}976$). Features exclude satisfaction.

Class	Precision	Recall	F1	Support
Business	0.88	0.80	0.84	13,977
Eco	0.79	0.73	0.76	9,929
Eco Plus	0.16	0.34	0.22	2,070

While the model accurately identifies Business and Economy passengers, it fails on Eco Plus (F1 = 0.22). Eco Plus passengers occupy a demographic and service-rating middle ground, lacking clean decision boundaries to separate them from the other two classes.

6 Discriminative Models

6.1 Linear and Quadratic Discriminant Analysis (LDA & QDA)

LDA achieved 87% accuracy. Using Youden's J statistic, we found the optimal classification threshold to be 0.54, which marginally improved precision for the satisfied class without harming overall accuracy.

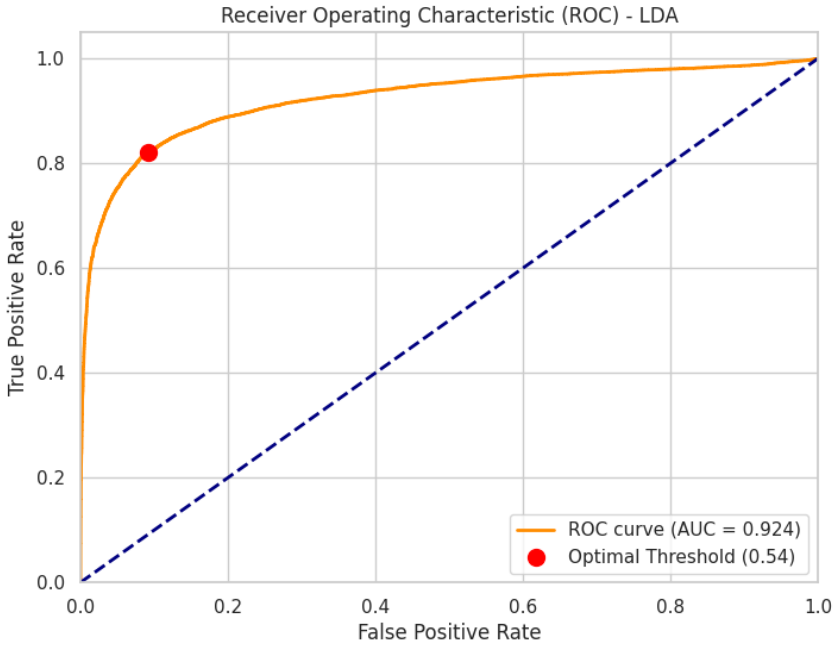

Figure 1: ROC curve confirming the strong predictive performance of the LDA model.

QDA, which estimates a separate covariance matrix per class, achieved slightly lower performance (85% accuracy). The drop suggests that QDA's added flexibility leads to slight overfitting, and LDA's constraint (shared covariance) is actually beneficial for this dataset.

6.2 Naïve Bayes

Gaussian Naïve Bayes achieved 86% accuracy despite violating the core assumption of feature independence. This highlights that Naïve Bayes only requires accurate posterior ranking, not precise probability values; correlated features may distort log-likelihood magnitudes without changing the final predicted class.

7 Model Comparison

Table 4 summarises the binary satisfaction classifiers. LR and LDA performed best (87%), indicating that linear decision boundaries are highly effective for this data. The narrow spread in performance implies that feature engineering and data quality matter more here than the specific algorithm chosen.

Table 4: Binary classification model comparison ($N = 25{,}976$).

Model	**Accuracy**	**Weighted F1**
Logistic Regression (LR)	87%	0.87
LDA (optimal threshold)	87%	0.87
Naïve Bayes (Gaussian)	86%	0.86
QDA (`reg_param` $= 10^{-4}$)	85%	0.85

8 Regression Models: OLS vs. Poisson

Finally, we modeled *Arrival Delay in Minutes* (imputed in preprocessing) using all available predictors: demographics, flight distance, the 14 service ratings, and departure delay in minutes. This yields roughly 21 features (after one-hot encoding of categoricals). Standardisation was applied for numerical stability, especially for the Poisson GLM.

OLS provides interpretable coefficients but can produce negative delay predictions, which are impossible. A Poisson GLM (log link) constrains predictions to be non-negative and is theoretically well suited to non-negative, right-skewed outcomes. Table 5 reports test-set performance. On this dataset, OLS achieved a lower RMSE (11.02 minutes) than the Poisson model (55.44 minutes). The Poisson fit may be hampered by the scale and variance of delay (which is not strictly count-like), whereas OLS adapts to the conditional mean without distributional constraints. The Poisson formulation remains useful when non-negative predictions are required; OLS may produce a small number of negative predicted delays on the test set.

Table 5: Test-set RMSE for Arrival Delay in Minutes ($N = 25{,}976$).

Model	**RMSE (minutes)**
Linear Regression (OLS)	11.02
Poisson Regression (GLM)	55.44

9 Conclusions

- **Key Satisfaction Drivers:** Online boarding quality, travel purpose (Business), and customer loyalty are the strongest predictors of passenger satisfaction.
- **Classification Performance:** Logistic Regression and LDA (87% accuracy) outperform non-linear or independence-assuming models, proving that simple linear boundaries are sufficient for this dataset.
- **Multinomial Limitations:** Eco Plus passengers are difficult to classify due to overlapping traits with both Economy and Business classes.
- **Regression Approach:** For arrival delay in minutes, OLS achieved lower test RMSE (11.02 min) than the Poisson GLM (55.44 min) on this dataset. A Poisson model remains useful when non-negative predictions are required; OLS is more flexible but can yield a few negative predicted delays.