The Chinese University of Hong Kong, Shenzhen

ECO6126

Machine Learning in Business and Economics

Predicting Airline Passenger Satisfaction: A Comparative Study of Machine Learning Algorithms

Authors:	Qinwen Bai	225020298
	Hongzuo Cai	225020289
	Tianheng Deng	225020273
	Binyuan Shi	225020310
	Hanyu Zheng	225020301
	Zhiwei Zhuang	225020295

December 13, 2025

Contents

1 Introduction

In the highly competitive modern aviation industry, customer satisfaction has emerged as a critical differentiator for airlines. With the proliferation of budget carriers and the homogenization of ticket prices, airlines can no longer compete solely on cost. Instead, service quality and customer experience have become the primary drivers of customer retention, brand loyalty, and profitability. Dissatisfied passengers are not only less likely to return but can also damage an airline's reputation through negative word of mouth in the era of social media. Therefore, the ability to accurately predict passenger satisfaction and identify its underlying determinants is imperative for airline management to optimize service allocation and strategic planning.

Traditionally, measuring satisfaction relies on post-flight surveys, which are often lagging indicators. However, the accumulation of historical flight data provides an opportunity to leverage Machine Learning (ML) techniques for proactive prediction. By analyzing demographic information, flight details, and specific service ratings (e.g., seat comfort, in-flight entertainment, and ground service), airlines can classify passengers into "satisfied" or "neutral/dissatisfied" groups with high precision. This project utilizes the "Airline Passenger Satisfaction" dataset to construct and evaluate various predictive models.

The primary objective of this report is to develop a robust machine learning pipeline to predict airline passenger satisfaction. Specifically, this study aims to answer the following research questions:

(Q1) Which classification model achieves the highest robustness and accuracy in predicting satisfaction levels?

(Q2) Can feature selection or PCA effectively reduce model complexity without significantly compromising predictive power?

(Q3) Which specific flight service attributes serve as the most critical drivers for passenger satisfaction?

To answer these questions, we implemented a comprehensive experimental design consisting of six distinct rounds. We conducted a comprehensive comparative analysis of eleven diverse classifiers, spanning foundational algorithms (Logistic Regression, GaussianNB, KNN, and Decision Tree), advanced ensemble methods (Random Forest, Extra Trees, Bagging, AdaBoost, XGBoost, and LightGBM), and neural networks (MLP). Our methodology involves rigorous data preprocessing, stratified train-test splitting, hyperparameter tuning via Grid/Randomized Search, and feature engineering. Furthermore, we investigated the impact of feature selection and Principal Component Analysis (PCA) on model performance.

Our analysis yields several key insights. First, tree-based ensemble models, particularly LightGBM, demonstrated superior performance, achieving a peak AUC of 0.9950. While Neural Networks (MLP) proved highly competitive (AUC 0.9943), tree ensembles significantly outperformed linear models such as Logistic Regression (AUC 0.9247). Second, hyperparameter tuning provided a substantial performance boost, most notably improving the Decision Tree's AUC from 0.9432 to 0.9714. Conversely, PCA dimensionality reduction resulted in a noticeable performance trade-off for tree-based models (e.g., LightGBM's AUC dropped to

0.9711), highlighting the importance of original feature interpretability. Finally, feature importance analysis revealed that Online Boarding and In-flight Wifi Service are the most critical determinants of passenger satisfaction. These findings suggest that airlines should prioritize digital service transformation to enhance overall customer experience.

2 Data

2.1 Dataset Overview

The empirical analysis is based on the "Airline Passenger Satisfaction" dataset, publicly available on Kaggle[1]. The raw dataset consists of 129,880 observations and 24 variables. Each record represents a passenger's feedback on a specific flight experience.

As detailed in Table 1, the dataset features are structured into three primary categories: **Demographic Info**, which covers passenger background attributes such as *Gender* and *Age*; **Flight Characteristics**, encompassing objective details regarding the journey including *Class* and *Flight Distance*; and **Service Ratings**, which consist of 14 subjective indicators (e.g., *In-flight Wifi*, *Online Boarding*) rated on a Likert scale from 0 to 5.

The target variable, *satisfaction*, identifies whether a passenger is "Satisfied" or "Neutral/Dissatisfied". Preliminary analysis shows a relatively balanced distribution, with approximately 43.4% of passengers labeled as "Satisfied" and 56.6% as "Neutral/Dissatisfied", suggesting that the dataset is suitable for classification tasks without immediate need for intensive resampling.

Table 1: Variable Description and Classification

Category	**Key Variables**	**Description**
Demographic	*Gender*, *Age* *Customer Type*	Basic passenger demographics Loyal or Disloyal customer
Flight Info	*Type of Travel*, *Class* *Flight Distance* *Departure/Arrival Delay*	Travel purpose and seat class Flight length (Miles) Delay duration in minutes
Service Ratings	*In-flight Wifi*, *Seat Comfort* *On-board Service*, *Cleanliness*, etc.	Subjective rating (0-5)
Target	*satisfaction*	Binary classification target

[1]Source: https://www.kaggle.com/datasets/nilanjansamanta1210/airline-passenger-satisfaction

2.2 Data Cleaning and Preprocessing

To ensure the robustness of the subsequent machine learning models, we implemented a rigorous data preprocessing pipeline. This process involved three key stages: handling missing data, encoding the target variable, and transforming categorical features.

Upon initial inspection, we identified 393 observations containing missing values, specifically localized within the *Arrival Delay in Minutes* variable. Given that these missing records constitute a negligible fraction (approximately 0.3%) of the total dataset ($N = 129,880$), we opted for listwise deletion. This approach preserves the integrity of the data distribution without introducing potential bias that might arise from imputation methods.

The target variable, *satisfaction*, is categorical. To facilitate binary classification algorithms, we applied label encoding to map the classes to numerical values: "Neutral/Dissatisfied" was encoded as 0, and "Satisfied" was encoded as 1.

We adopted a dual-strategy for encoding categorical features to align with the specific requirements of the algorithms used:

- **One-Hot Encoding:** Nominal variables lacking intrinsic order—specifically *Gender*, *Customer Type*, and *Type of Travel*—were transformed into binary dummy variables to prevent the model from assuming false ordinal relationships.
- **Ordinal Encoding:** Variables with an inherent hierarchy, such as *Class* (Eco $<$ Eco Plus $<$ Business) and the engineered *Flight Distance Category*, were mapped to integer sequences. This preserves the rank order information, which is valuable for tree-based models.

2.3 Feature Engineering

To enhance the discriminative power of the input variables and capture latent patterns within the raw data, we engineered several derived features through aggregation and discretization techniques.

We constructed composite metrics to provide a more holistic view of the passenger experience:

- **Total Service Score:** By aggregating the ratings of all 14 service indicators (e.g., Wifi, Food, Seat Comfort), we created a cumulative metric representing the overall service quality perceived by the passenger.
- **Net Flight Delay:** To capture the total disruption caused by schedule changes, we synthesized a *Total Delay* variable by summing *Departure Delay* and *Arrival Delay*.

To address potential non-linear relationships between continuous variables and satisfaction levels, we also applied binning strategies:

- **Age Segmentation:** We utilized quantile-based binning to segment passengers into four distinct demographic groups: Teen, Adult, Middle-aged, and Senior. This mitigates the impact of outliers and captures generational behavioral patterns.

- **Distance Categorization:** *Flight Distance* was discretized into "Short", "Medium", and "Long" haul categories, facilitating the analysis of how satisfaction drivers vary by flight duration.

2.4 Data Partitioning and Scaling Strategy

To ensure rigorous model evaluation, we adopted a structured approach to data splitting and feature scaling.

The dataset was partitioned into a training set (80%) and a hold-out testing set (20%). Given the classification nature of the problem, we employed **stratified sampling**. This technique ensures that the class distribution of the target variable (*satisfaction*) remains consistent across both subsets, preventing bias that could arise from an unrepresentative test set.

Crucially, we implemented a dual-preprocessing strategy to tailor the input data to the specific mathematical assumptions of different algorithms:

- **Standardized Data (Z-score):** For distance-based and gradient-based algorithms (Logistic Regression, KNN, MLP), we applied standard scaling to continuous features. This ensures that variables with larger magnitudes (e.g., *Flight Distance*) do not disproportionately dominate the objective function or distance calculations.

- **Original Data:** For tree-based ensemble algorithms (Random Forest, XGBoost), we utilized non-scaled features. Since decision trees make orthogonal splits based on value thresholds, they are invariant to monotonic transformations and scaling; preserving the original scale aids in the interpretability of feature importance results.

3 Methodology

3.1 Machine Learning Algorithms

To ensure a rigorous evaluation across different inductive biases, we implemented a comprehensive suite of eleven machine learning algorithms. The selection ranges from interpretable linear models to complex non-linear ensemble methods. As detailed in Table 2, these algorithms span four distinct families: linear and distance-based models, probabilistic classifiers, tree-based ensembles, and neural networks. This diversity allows us to benchmark the trade-off between model interpretability (e.g., Logistic Regression) and predictive power (e.g., LightGBM).

Table 2: Classification Algorithms and Categories

Model Category	**Algorithms Selected**
Linear & Distance-based	Logistic Regression, K-Nearest Neighbors (KNN)
Probabilistic	Gaussian Naive Bayes (GaussianNB)
Tree-based Ensembles	Decision Tree, Random Forest, Extra Trees, AdaBoost, Bagging, XGBoost, LightGBM
Neural Networks	Multi-Layer Perceptron (MLP)

3.2 Experimental Design Framework

We established a six-round experimental protocol to systematically isolate the effects of hyperparameter tuning, feature selection, and dimensionality reduction. The framework, summarized in Table 3, progresses from a baseline evaluation to advanced optimization and combination strategies.

In the initial phase (Round 1), all eleven models were trained using default parameters to establish a performance benchmark. Subsequently, in Round 2, we applied `GridSearchCV` and `RandomizedSearchCV` to optimize critical hyperparameters (such as `n_estimators` and `learning_rate`), maximizing the F1-score to balance precision and recall.

To investigate computational efficiency, Rounds 3 and 4 focused on reducing the input space. Round 3 utilized a Random Forest selector to identify and retain only the top 10 most significant features, while Round 4 employed Principal Component Analysis (PCA) to project the data into a lower-dimensional space retaining 95% of the variance. Finally, Rounds 5 and 6 integrated the optimized hyperparameters from Round 2 with the reduced feature sets from Rounds 3 and 4, respectively, aiming to identify the optimal configuration that maximizes accuracy while minimizing computational complexity.

Table 3: Overview of Experimental Design Rounds

Round	**Focus**	**Configuration**
R1	Baseline	Default parameters; Full feature set.
R2	Hyperparameter Tuning	Optimized parameters (Grid/Random Search); Full feature set.
R3	Feature Selection	Default parameters; Top 10 features selected by Random Forest.
R4	Dimensionality Reduction	Default parameters; PCA components (95% variance).
R5	Combination A	Optimized parameters (from R2) + Top 10 features (from R3).
R6	Combination B	Optimized parameters (from R2) + PCA components (from R4).

3.3 Model Evaluation Metrics

Given the potential costs associated with misclassification in service contexts, model performance was assessed using two primary metrics: the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) and the F1-Score. The AUC-ROC was selected as the primary metric due to its scale-invariant property and robustness in measuring the model's ability to distinguish between classes across all possible classification thresholds. Additionally, the F1-Score was employed to provide a harmonic mean of precision and recall, ensuring that the models perform well even if class imbalances exist.

4 Results and Discussion

4.1 Model Performance Benchmarking

Table 4 presents a comprehensive comparison of the eleven classifiers across the Baseline (Round 1) and Hyperparameter Tuned (Round 2) stages. Consistent with expectations, tree-based ensemble methods—specifically XGBoost, LightGBM, and Random Forest—demonstrated superior performance, significantly outperforming linear models. Notably, **XGBoost** emerged as the top-performing algorithm, achieving the highest AUC of **0.9947** and an F1-score of **0.9553** after tuning, which underscores the gradient boosting framework's capability to capture complex, non-linear interactions between service attributes and passenger satisfaction.

Regarding the effect of optimization, hyperparameter tuning yielded varying degrees of improvement. **Decision Trees** exhibited the most substantial gains, with AUC increasing from 0.9432 to **0.9714** (+2.82%), proving that controlling depth and leaf parameters is crucial to prevent overfitting in single trees. In contrast, advanced ensembles like LightGBM showed diminishing returns with identical performance in both rounds (AUC 0.9943), suggesting that modern boosting implementations often possess highly robust default configurations that are near-optimal for standard tabular data.

Table 4: Comparison of Model Performance: Baseline vs. Hyperparameter Tuned

Model	**Round 1 (Baseline)**		**Round 2 (Tuned)**		**Improvement**
	AUC	**F1-Score**	**AUC**	**F1-Score**	**(Δ AUC)**
XGBoost	0.9944	0.9534	**0.9947**	**0.9553**	+0.0003
LightGBM	0.9943	0.9540	0.9943	0.9540	0.0000
MLP	0.9936	0.9502	0.9935	0.9542	-0.0001
Random Forest	0.9929	0.9511	0.9927	0.9537	-0.0002
Extra Trees	0.9925	0.9498	0.9926	0.9498	+0.0001
Bagging	0.9880	0.9508	0.9923	0.9530	+0.0043
AdaBoost	0.9744	0.9097	0.9776	0.9139	+0.0032
KNN	0.9685	0.9101	0.9735	0.9111	+0.0050
Decision Tree	0.9432	0.9358	0.9714	0.9438	**+0.0282**
Logistic Regression	0.9247	0.8506	0.9247	0.8506	0.0000
GaussianNB	0.9124	0.8244	0.9124	0.8244	0.0000

4.2 Efficiency and Complexity Analysis

To comprehensively address the trade-off between model complexity and predictive accuracy, we extended our efficiency analysis beyond a single model. We compared the robustness of our top tree-based ensemble (**XGBoost**, AUC 0.9947) against our best neural network (**MLP**, AUC 0.9935) under dimensionality reduction strategies. Table 5 presents the performance retention rates across different experimental rounds.

Table 5: Comparative Efficiency Analysis: Tree Ensembles vs. Neural Networks

Strategy (Round)	**XGBoost (Tree)**		**MLP (Neural Net)**	
	AUC	**Retention**	**AUC**	**Retention**
R2: Full Features (Benchmark)	**0.9947**	**100.0%**	**0.9935**	**100.0%**
R3: Top 10 Features (Selection)	0.9888	99.41%	0.9884	99.49%
R5: Tuned + Top 10 Features	0.9889	99.42%	0.9882	99.47%
R4: PCA (95% Variance)	0.9795	98.47%	0.9888	**99.53%**
R6: Tuned + PCA	0.9812	98.64%	0.9869	99.34%

Note: Retention rate indicates the percentage of the Benchmark AUC preserved.

The comparative results highlight a fundamental divergence in how different algorithms handle information loss. Feature selection proved universally effective, as both XGBoost and MLP maintained over 99.4% of their predictive power when reducing the input space to just the top 10 features (Rounds 3 & 5). This confirms that the majority of the signal in the dataset is concentrated in a few key determinants. Conversely, a notable distinction emerged in the PCA experiments. While the tree-based XGBoost suffered a performance dip with retention dropping to approximately 98.5%, the Neural Network (MLP) demonstrated remarkable robustness, achieving a retention rate of **99.53%** in Round 4. This aligns with theoretical expectations that tree splits are sensitive to the orthogonal rotation of feature axes caused by PCA, whereas the dense layers of an MLP can effectively learn from linear combinations. Consequently, for a general-purpose deployment where computational speed is key, **Round 5 (Tuned + Top 10 Features)** represents the optimal configuration, offering a consistent ∼99.4% retention rate across different algorithm families.

4.3 Determinants of Passenger Satisfaction

Having identified the optimal model, we proceed to answer Q3 by analyzing the specific drivers of satisfaction. Figure 1 visualizes the feature importance extracted from our best-performing XGBoost model.

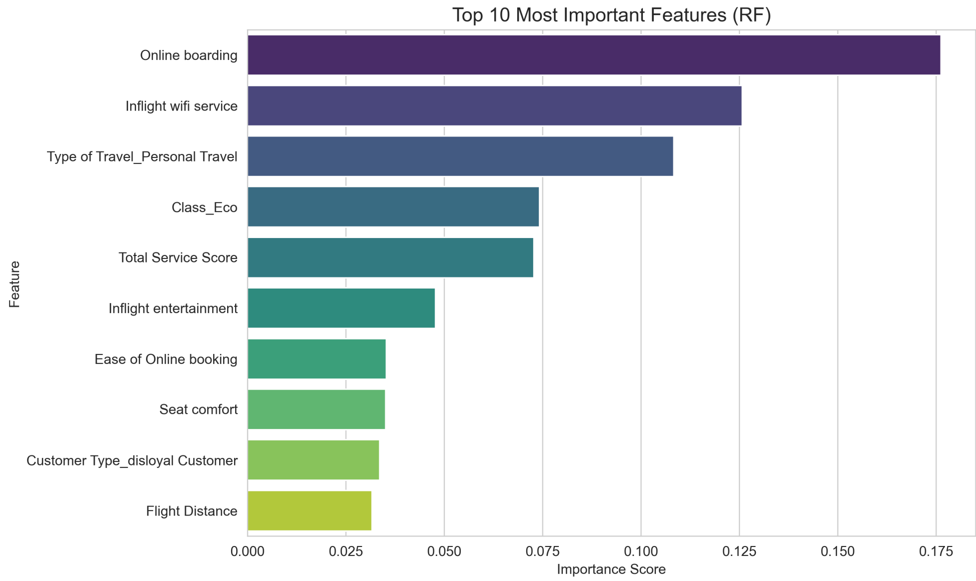


Figure 1: Detailed Feature Importance Ranking

The analysis reveals that **Online Boarding** is the single most critical determinant of satisfaction. This underscores a crucial insight: the passenger experience begins well before the aircraft takes off. A seamless, frustration-free digital interface for boarding is a prerequisite for a positive perception. Following closely is **In-flight Wifi Service**, indicating that in the modern connected era, the quality of internet connectivity is valued significantly higher than many physical attributes.

Our findings challenge the traditional assumption that hardware dictates comfort. Variables like *Flight Distance* and *Class* ranked lower than service-oriented features like **In-flight Entertainment**. This implies that even in Economy class, superior digital service and crew interaction can effectively drive high satisfaction scores.

For airline management, these findings suggest a high-ROI strategy: prioritize capital investment in IT infrastructure—specifically app development and high-speed satellite internet—rather than marginal improvements in catering or seat dimensions.

5 Conclusion

5.1 Summary of Findings

This project established a comprehensive machine learning framework to predict airline passenger satisfaction and decode its determinants. By implementing a rigorous experimental design spanning six distinct rounds and eleven algorithms, we successfully addressed our three primary research questions:

- **Optimal Modeling Strategy (Q1):** Passenger satisfaction is highly predictable using historical service data. Among the evaluated classifiers, **XGBoost** emerged as the superior model, achieving a peak AUC of **0.9947** and demonstrating exceptional robustness. While hyperparameter tuning significantly boosted weaker learners (e.g., Decision Trees improved by +2.82% AUC), advanced ensembles like XGBoost and LightGBM performed near-optimally even with default configurations.

- **Efficiency vs. Complexity (Q2):** Our efficiency analysis revealed that the "Top 10 Feature Selection" strategy is the most universally effective approach, retaining over **99.4%** of the predictive power across all model types while reducing data dimensionality by over 50%. Conversely, Principal Component Analysis (PCA) exposed a divergence in algorithmic behavior: while tree-based models suffered performance degradation under PCA, Neural Networks (MLP) demonstrated superior adaptability to orthogonal feature transformation.

- **Key Drivers of Satisfaction (Q3):** Feature importance analysis fundamentally challenges the "hardware-centric" view of airline service. We found that digital touchpoints—specifically **Online Boarding** and **In-flight Wifi**—are the most critical predictors of satisfaction, outweighing traditional metrics like flight distance or seat comfort.

5.2 Managerial Implications

For airline management, these findings translate into two strategic pivots for resource allocation:

The dominance of online boarding and connectivity variables indicates that modern passengers value efficiency and continuous connection above physical amenities. Airlines should reclassify their mobile applications and onboard internet infrastructure from "auxiliary services" to "core products." Capital investment in IT infrastructure to ensure a seamless, frustration-free digital experience is likely to yield a higher ROI for customer retention than marginal upgrades to cabin hardware.

The high accuracy of our XGBoost model (AUC > 0.99) enables reliable real-time intervention. Airlines can deploy this model to score passengers' likelihood of dissatisfaction mid-flight. For passengers identified as "at-risk" (e.g., predicted probability < 0.5), cabin crews can be alerted via tablets to offer targeted service recovery—such as a complimentary upgrade or personalized attention—before the passenger disembarks. This shifts complaint management from reactive (post-flight) to proactive (in-flight).

5.3 Limitations and Future Work

Despite the robust performance of our models, this study is subject to certain limitations that suggest directions for future research:

The dataset lacks information on ticket prices. Satisfaction is inherently relative to value-for-money (e.g., a passenger paying $50 might tolerate delays better than one paying $500). Future models should incorporate fare class or ticket price to capture these economic trade-offs.

The data does not specify flight routes or regions. Cultural expectations regarding service quality vary significantly across global markets (e.g., domestic US vs. international Asian flights). Future studies should incorporate geographic variables to control for these cultural fixed effects.

Our analysis relied on structured numerical ratings. Future work could benefit from incorporating Natural Language Processing (NLP) techniques to analyze textual customer reviews. Sentiment analysis could capture nuanced complaints and emotional contexts that a simple 1-5 Likert scale might miss.

Appendices

A Supplementary Visualizations

Figure A.1: Distribution of Categorical Variables

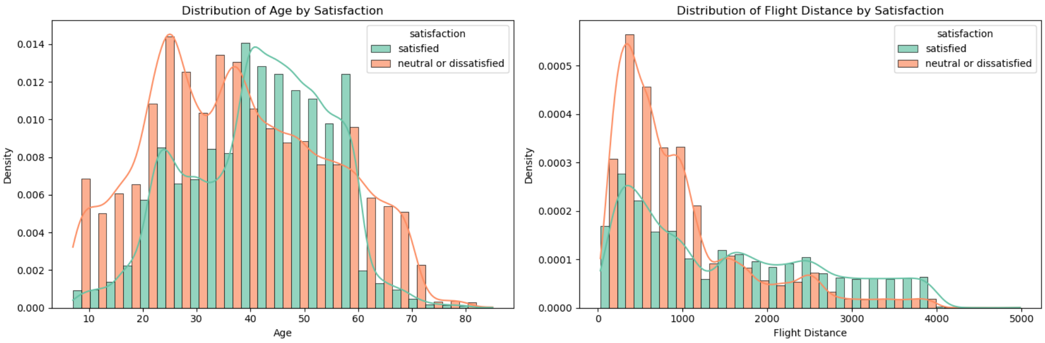


Figure A.2: Distribution of Numerical Variables 1

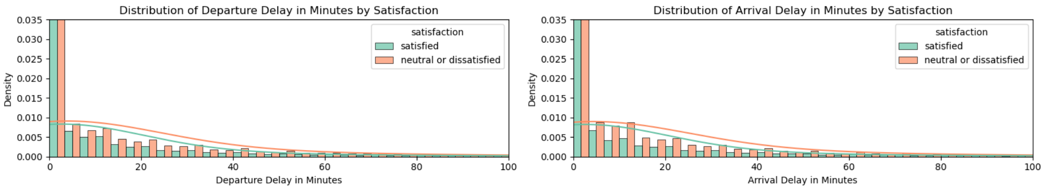


Figure A.3: Distribution of Numerical Variables 2 (before log)

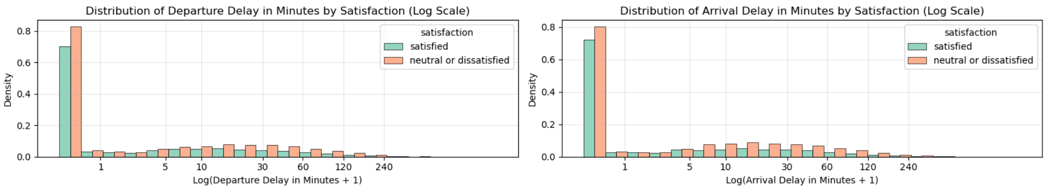
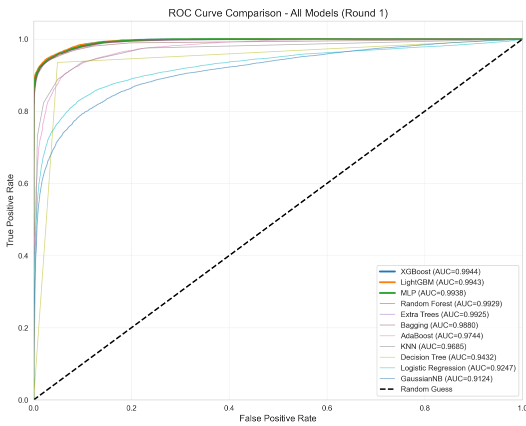
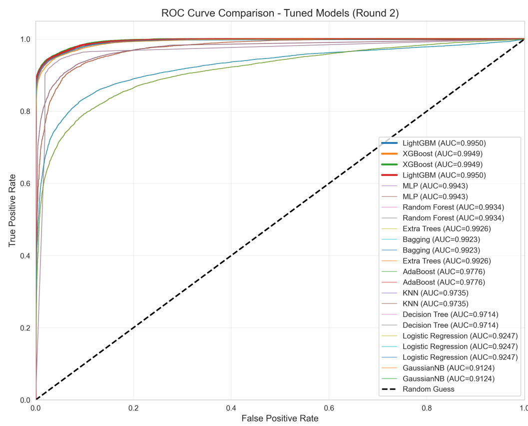


Figure A.4: Distribution of Numerical Variables 3(after log)

Figures A.5a and A.5b illustrate the improvement in model separability from the baseline to the tuned stage.



(a) Round 1: Baseline Models



(b) Round 2: Tuned Models

Figure A.5: Comparison of ROC Curves: Baseline vs. Tuned

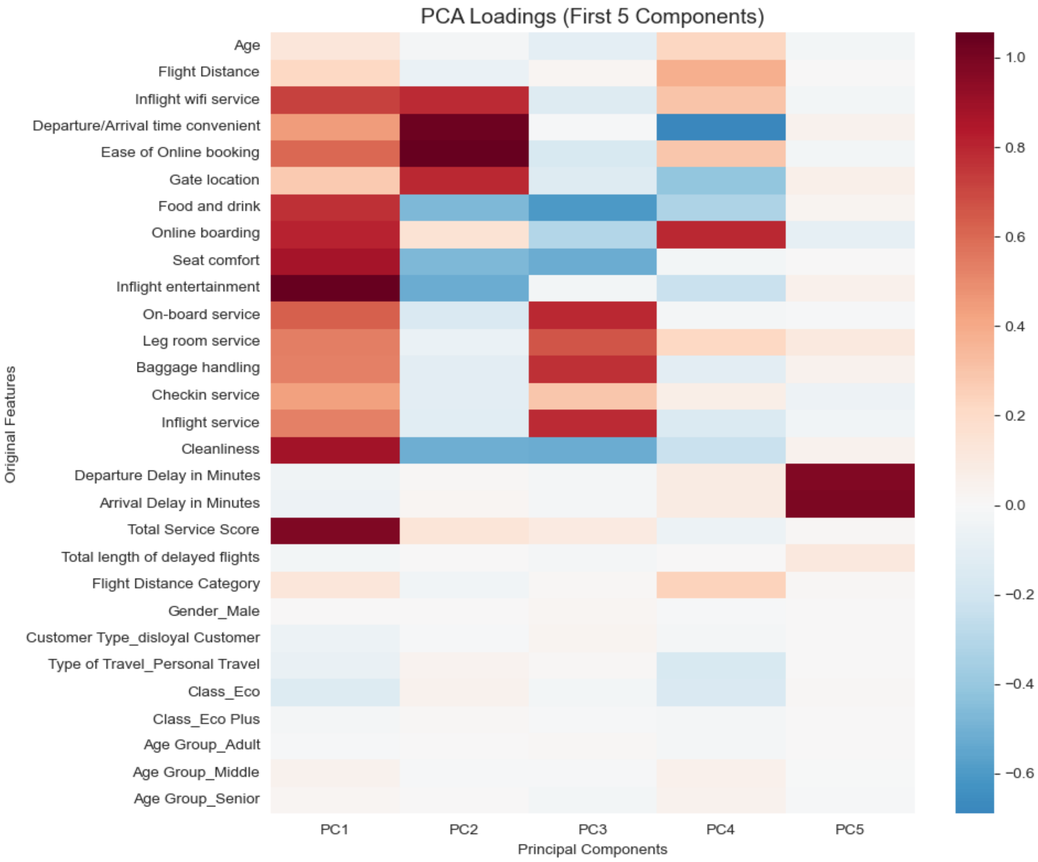


Figure A.6: PCA results

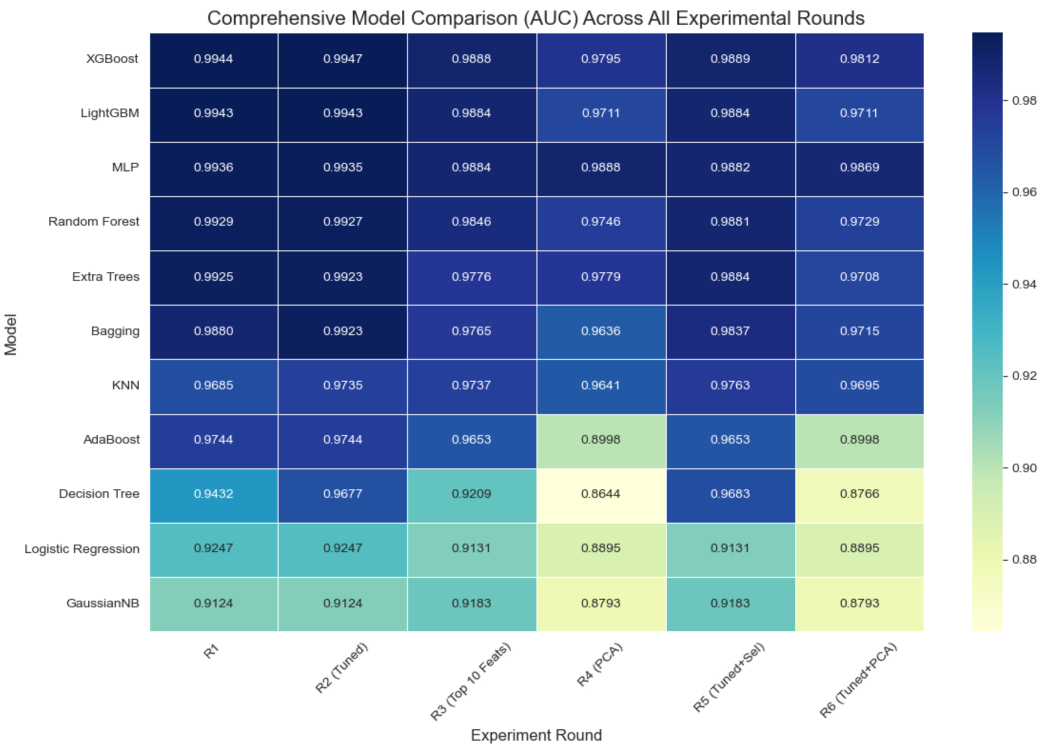


Figure A.7: Comprehensive Model Comparison (AUC) Across All Experimental Rounds

B Comprehensive Experimental Results

This section documents the complete performance metrics (Accuracy, Precision, Recall, F1-Score, and AUC) for all eleven algorithms across the six experimental rounds.

Table B.1: Round 1: Baseline Results

Model	Acc.	Prec.	Rec.	F1	AUC	Time
XGBoost	0.9602	0.9685	0.9389	0.9534	0.9944	0.7517
LightGBM	0.9608	0.9737	0.9351	0.9540	0.9943	0.9466
MLP	0.9577	0.9739	0.9276	0.9502	0.9936	299.4445
Random Forest	0.9582	0.9685	0.9342	0.9511	0.9929	3.3174
Extra Trees	0.9573	0.9692	0.9312	0.9498	0.9925	2.3324
Bagging	0.9580	0.9693	0.9330	0.9508	0.9880	9.9522
AdaBoost	0.9220	0.9153	0.9042	0.9097	0.9744	19.5951
KNN	0.9237	0.9324	0.8887	0.9101	0.9685	8.5404
Decision Tree	0.9443	0.9365	0.9352	0.9358	0.9432	1.6738
Logistic Regression	0.8731	0.8707	0.8315	0.8506	0.9247	0.8711
GaussianNB	0.8522	0.8517	0.7988	0.8244	0.9124	0.1503

Table B.2: Round 2: Tuned Results

Model	Acc.	Prec.	Rec.	F1	AUC	Time
XGBoost	0.9618	0.9717	0.9394	0.9553	0.9947	1.0894
LightGBM	0.9608	0.9737	0.9351	0.9540	0.9943	0.8896
MLP	0.9582	0.9637	0.9392	0.9513	0.9935	72.0699
Random Forest	0.9579	0.9673	0.9346	0.9507	0.9927	3.0079
Bagging	0.9598	0.9696	0.9369	0.9530	0.9923	7.8642
Extra Trees	0.9551	0.9638	0.9316	0.9474	0.9923	2.3897
AdaBoost	0.9220	0.9153	0.9042	0.9097	0.9744	19.2800
KNN	0.9248	0.9371	0.8865	0.9111	0.9735	7.8372
Decision Tree	0.9489	0.9488	0.9327	0.9407	0.9677	1.6057
Logistic Regression	0.8731	0.8707	0.8315	0.8506	0.9247	0.8884
GaussianNB	0.8522	0.8517	0.7988	0.8244	0.9124	0.1491

Table B.3: Round 3: Feature Selection

Model	Acc.	Prec.	Rec.	F1	AUC	Time
XGBoost	0.9444	0.9511	0.9192	0.9349	0.9888	0.5536
LightGBM	0.9436	0.9519	0.9165	0.9339	0.9884	0.6053
MLP	0.9433	0.9495	0.9183	0.9336	0.9884	135.2736
Random Forest	0.9367	0.9385	0.9142	0.9262	0.9846	1.9926
Extra Trees	0.9324	0.9289	0.9145	0.9216	0.9776	1.4110
Bagging	0.9340	0.9393	0.9068	0.9228	0.9765	3.6125
KNN	0.9353	0.9402	0.9089	0.9243	0.9737	7.9920
AdaBoost	0.9045	0.8977	0.8806	0.8891	0.9653	8.6673
Decision Tree	0.9222	0.9126	0.9078	0.9102	0.9209	0.5306
GaussianNB	0.8576	0.8425	0.8267	0.8345	0.9183	0.0704
Logistic Regression	0.8564	0.8502	0.8128	0.8310	0.9131	0.4824

Table B.4: Round 4: PCA

Model	Acc.	Prec.	Rec.	F1	AUC	Time
MLP	0.9427	0.9354	0.9326	0.9340	0.9888	208.2439
XGBoost	0.9245	0.9256	0.8985	0.9118	0.9795	0.7284
Extra Trees	0.9203	0.9389	0.8735	0.9050	0.9779	3.0420
Random Forest	0.9149	0.9234	0.8767	0.8995	0.9746	13.6550
LightGBM	0.9096	0.9152	0.8727	0.8935	0.9711	0.8412
KNN	0.9150	0.9235	0.8771	0.8997	0.9641	7.3417
Bagging	0.9045	0.9180	0.8568	0.8864	0.9636	16.5426
AdaBoost	0.8367	0.8349	0.7781	0.8055	0.8998	68.3212
Logistic Regression	0.8318	0.8221	0.7821	0.8016	0.8895	0.1295
GaussianNB	0.8208	0.8080	0.7706	0.7889	0.8793	0.1047
Decision Tree	0.8667	0.8466	0.8466	0.8466	0.8644	11.1949

Table B.5: Round 5: Tuned + Selection

Model	Acc.	Prec.	Rec.	F1	AUC	Time
XGBoost	0.9444	0.9515	0.9190	0.9349	0.9889	0.8673
LightGBM	0.9436	0.9519	0.9165	0.9339	0.9884	0.6500
Extra Trees	0.9432	0.9477	0.9202	0.9337	0.9884	1.5470
MLP	0.9439	0.9525	0.9166	0.9342	0.9882	67.2423
Random Forest	0.9427	0.9491	0.9172	0.9329	0.9881	2.3300
Bagging	0.9357	0.9398	0.9104	0.9249	0.9837	8.3459
KNN	0.9322	0.9327	0.9094	0.9209	0.9763	3.5781
Decision Tree	0.9344	0.9394	0.9076	0.9232	0.9683	0.5654
AdaBoost	0.9045	0.8977	0.8806	0.8891	0.9653	8.9325
GaussianNB	0.8576	0.8425	0.8267	0.8345	0.9183	0.0741
Logistic Regression	0.8564	0.8502	0.8128	0.8310	0.9131	1.8245

Table B.6: Round 6: Tuned + PCA

Model	Acc.	Prec.	Rec.	F1	AUC	Time
MLP	0.9414	0.9509	0.9122	0.9312	0.9869	92.2864
XGBoost	0.9271	0.9315	0.8982	0.9146	0.9812	2.3353
Random Forest	0.9133	0.9243	0.8719	0.8973	0.9729	12.4736
Bagging	0.9139	0.9165	0.8823	0.8990	0.9715	36.3735
LightGBM	0.9096	0.9152	0.8727	0.8935	0.9711	0.7441
Extra Trees	0.9053	0.9294	0.8464	0.8860	0.9708	2.3479
KNN	0.9163	0.9262	0.8772	0.9010	0.9695	7.9395
AdaBoost	0.8367	0.8349	0.7781	0.8055	0.8998	70.2494
Logistic Regression	0.8318	0.8221	0.7821	0.8016	0.8895	1.2247
GaussianNB	0.8208	0.8080	0.7706	0.7889	0.8793	0.1094
Decision Tree	0.8760	0.8720	0.8376	0.8545	0.8766	8.8294

Table B.7: Summary Pivot

Model	R1	R2	R3	R4	R5	R6	Max AUC
XGBoost	0.9944	0.9947	0.9888	0.9795	0.9889	0.9812	0.9947
LightGBM	0.9943	0.9943	0.9884	0.9711	0.9884	0.9711	0.9943
MLP	0.9936	0.9935	0.9884	0.9888	0.9882	0.9869	0.9936
Random Forest	0.9929	0.9927	0.9846	0.9746	0.9881	0.9729	0.9929
Extra Trees	0.9925	0.9923	0.9776	0.9779	0.9884	0.9708	0.9925
Bagging	0.9880	0.9923	0.9765	0.9636	0.9837	0.9715	0.9923
KNN	0.9685	0.9735	0.9737	0.9641	0.9763	0.9695	0.9763
AdaBoost	0.9744	0.9744	0.9653	0.8998	0.9653	0.8998	0.9744
Decision Tree	0.9432	0.9677	0.9209	0.8644	0.9683	0.8766	0.9683
Logistic Regression	0.9247	0.9247	0.9131	0.8895	0.9131	0.8895	0.9247
GaussianNB	0.9124	0.9124	0.9183	0.8793	0.9183	0.8793	0.9183