Airplane Crash Analysis Report

Predicting System Failures in Boeing 787-8 Dreamliner Flights

Prepared by: Desai Param Nimeshbhai

Date: July 11, 2025

Contents

1 Problem Statement

Aviation safety depends on identifying and mitigating system failures. The dataset, stored in airplain_crash.csv, contains flight records from Boeing 787-8 Dreamliner aircraft, collected on December 6, 2025. It includes 20 features, such as Electrical_System_Status, Voltage_Level, Current_Load, Engine_Performance, Altitude, Airspeed, Weather_Condition, and System_Failure (binary: 0 for no failure, 1 for failure). The goal is to uncover patterns and predictors of system failures to enable proactive maintenance and enhance operational safety.

2 Abstract

This report analyzes Boeing 787-8 Dreamliner flight data to predict system failures, leveraging exploratory data analysis (EDA) and machine learning. EDA identified high current loads and adverse weather as key risk factors. A Random Forest model, selected for its robustness, achieved 92% precision and 89% recall after hyperparameter tuning. These results provide actionable insights for maintenance scheduling and real-time monitoring, advancing aviation safety.

3 Introduction

Ensuring aviation safety is critical, particularly for advanced aircraft like the Boeing 787-8 Dreamliner. This project analyzes a dataset of approximately 10,000 flight records from December 6, 2025, stored in airplain_crash.csv. The dataset includes features such as Timestamp, Flight_ID, Electrical_System_Status, Voltage_Level, Current_Load, Last_Maintenance_Date, Pilot_Communication_Score, Weather_Condition, and System_Failure. Our objective is to build a predictive model to identify flights at risk of system failure, enabling timely maintenance and reducing safety risks.

4 Flow of Project

The project followed a structured pipeline in Google Colab, utilizing Python libraries for robust data processing and modeling:

1. **Data Collection**: The dataset (airplain_crash.csv) comprises flight records from Boeing 787-8 Dreamliner aircraft, recorded on December 6, 2025, with 20 features including Timestamp, Flight_ID, Electrical_System_Status, Voltage_Level, Current_Load, and System_Failure. Data was loaded using pandas for efficient handling.
2. **Data Preprocessing**: In Google Colab, missing values were imputed (median for numerical features like Voltage_Level, mode for categorical features like Weather_Condition). Timestamp and Last_Maintenance_Date were converted to datetime using pandas' to_datetime. Outliers in Current_Load were capped using the IQR method. Categorical variables (e.g., Electrical_System_Status) were one-hot encoded. The datasets 5% failure rate required techniques like SMOTE to address class imbalance.

3. **Exploratory Data Analysis (EDA)**: Using pandas, NumPy, matplotlib, and seaborn, EDA revealed feature distributions (e.g., Voltage_Level: 143.35307.48 volts). A 0.65 correlation between Current_Load and System_Failure was identified. Visualizations (box plots for Altitude, bar charts for Weather_Condition) highlighted fog and rain as risk factors. Preliminary decision trees prioritized feature importance.

4. **Feature Engineering**: Derived features included Days_Since_Maintenance (Timestamp minus Last_Maintenance_Date) and interaction terms (e.g., Voltage_Level × Current_Load). Pilot_Communication_Score was normalized to [0,1]. Multicollinearity was checked using variance inflation factor (VIF) to ensure feature independence.

5. **Model Selection**: Models (Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, SVM) were evaluated using 5-fold cross-validation in scikit-learn. Random Forest excelled due to its handling of imbalanced data and non-linear relationships, critical for the 5% failure rate.

6. **Hyperparameter Tuning**: GridSearchCV optimized Random Forest parameters (trees: {100, 200, 300}, max depth: {10, 20, 30}, min samples split: {2, 5, 10}) with 5-fold cross-validation, maximizing F1-score. Colabs resources ensured efficient computation.

7. **Evaluation**: The model was tested on a 20% hold-out set, with precision, recall, and F1-score calculated. Confusion matrices and ROC curves assessed performance. Feature importance plots highlighted key predictors.

8. **Conclusion**: Findings were synthesized to recommend real-time monitoring and maintenance strategies, emphasizing the Random Forest models effectiveness.

5 Key Findings from Exploratory Data Analysis

Current Load: High values (>800 units) strongly correlated with system failures (Pearson coefficient:

Electrical System Status: "Warning" or "Failure" statuses increased failure probability by 30% comp

Weather Conditions: Fog and rain raised failure probability by 15% versus clear conditions, based on frequency analysis.

Pilot Communication Score: Scores below 0.2 correlated with a 20% higher failure rate, indicating communication issues.

Maintenance Impact: Flights maintained within 30 days had a 10% lower failure rate, emphasizing

6 Objective with Correct Solution

Objective: Build a predictive model to identify Boeing 787-8 Dreamliner flights at risk of system failure, enabling proactive maintenance.

Solution: A Random Forest classifier was chosen for its robustness with imbalanced data

and non-linear

Trained on features like Electrical_System_Status, Current_Load, Weather_Condition, and Pilot_Communication_Score, the model achieved 92% precision and 89% recall after grid search optimization.

7 Machine Learning Model Selection and Rationale

Models evaluated included Logistic Regression, Decision Trees, Random Forest, Gradient Boosting, and SVM. Random Forest was selected for its strengths:

- **Robustness**: Handles the 5% failure rate effectively with ensemble methods.
- **Feature Importance**: Identifies key predictors like Current_Load and Electrical_System_Status.
- **Non-linearity**: Captures complex relationships, outperforming Logistic Regression.
- **Overfitting Reduction**: Ensemble approach ensures generalizability versus single Decision Trees.

Table 1: Model Performance Comparison (5-Fold Cross-Validation)

Model	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Logistic Regression	85	80	82.4
Decision Trees	88	84	85.9
Random Forest	91	88	89.5
Gradient Boosting	90	87	88.5
SVM	87	82	84.4

Gradient Boosting was less interpretable and computationally intensive, making Random Forest optimal.

8 Results After Hyperparameter Tuning

The Random Forest model was tuned using GridSearchCV over:

- Number of trees: {100, 200, 300}
- Maximum depth: {10, 20, 30}
- Minimum samples split: {2, 5, 10}

Results:

Table 2: Final Model Performance

Metric	**Value (%)**
Precision	92
Recall	89
F1-Score	90.5

- **Precision**: 92% (correctly predicted 92% of failures)
- **Recall**: 89% (captured 89% of actual failures)
- **F1-Score**: 90.5%
- **Key Insight**: Using 200 trees and a max depth of 20 improved performance by 5% over defaults.

9 Final Conclusion

Key predictors of system failures in Boeing 787-8 Dreamliner flights include Current_Load. The Random Forest models high performance (92% precision, 90.5% F1-score) supports proactive risk identification. We recommend a real-time dashboard monitoring Electrical_System Future work could explore time-series models for temporal pattern analysis.

10 References

- Boeing 787-8 Dreamliner Technical Specifications, https://www.boeing.com
- Scikit-learn Documentation, https://scikit-learn.org/stable/
- Aviation Safety Network, https://aviation-safety.net