Predictive Modeling of Fatal Outcomes in Commercial Aviation Accidents

Nguyen (Chloe) Pham

DA401 - Fall 2025

Abstract

Commercial aviation remains exceptionally safe, with fatal accidents rare. However, the Boeing 737 MAX tragedies underline the need to analyze multifactorial risk factors. This study applies LASSO logistic regression and Random Forest models to 1,388 NTSB accident records (2008-2024), examining crew, weather, flight phase, and regulatory factors predicting fatality. After addressing missing data via MICE, key risks emerged: FAR Part 135 operations had 32 times higher fatality odds than Part 121 airlines, Unknown/Other flight phases increased odds 38-fold, and Instrument Meteorological Conditions raised risk by 58%. Random Forest identified flight phase and FAR Part as top predictors, but underperformed minority class recall compared to LASSO (0.02 vs. 0.12). Ultimately, LASSO offers better interpretability and balanced performance, supporting targeted safety interventions for Part 135 operations, critical flight phases, and adverse weather protocols.

Introduction

Commercial aviation (scheduled airline services and on-demand air taxi and commuter flights) is widely recognized as one of the safest modes of transportation. The International

Air Transport Association (IATA) reported in its 2024 Safety Report that the global all-accident rate was 1.13 per million flights, equivalent to one accident per 880,000 flights (IATA, 2025). Despite this remarkable safety record, catastrophic accidents can still harm public trust and the reputation of the airlines involved. For instance, the installation of the Maneuvering Characteristics Augmentation System (MCAS) into the Boeing 737 MAX has resulted in the deaths of 346 people and led to a worldwide grounding of the aircraft. This is because MCAS had critical flaws, including reliance on a single angle of attack sensor, repeated uncommanded activations, and excessive control authority (Henricodolfing, 2024). Another significant factor in the accidents was the lack of pilot training, as crews were not adequately informed about MCAS and its recovery procedures before crashes (FAA, 2020; AeroTime, 2022). Subsequently, Boeing faced extensive regulatory scrutiny, severe financial losses, and reputational damage, allowing competitors like Airbus to gain market share (Henricodolfing, 2024).

This example highlights the importance of understanding the conditions that precede fatal aviation accidents in effective risk mitigation and the long-term resilience of the industry. However, aviation safety analysis is complicated by the multifactorial nature of accident causation and the intricate interplay among contributing factors. Research and industry reports consistently attribute fatal accidents to key elements, such as flight crew proficiency, airframe configurations, and the operating environment (Airbus, 2025; IATA, 2025). Hence, traditional safety analysis methods may fall short by examining such factors in isolation, risking the oversight of complex and nonlinear interactions that contribute to accident risk. Moreover, the rarity of fatal accidents creates a statistical challenge that requires the use of advanced analytical approaches capable of extracting reliable patterns from limited datasets while minimizing the risk of overfitting.

Consequently, this study applies modern predictive analytics, including Logistic Regression and Random Forest models, to address these challenges. The central research question of this paper is: Between Logistic Regression and Random Forest, which model offers superior predictive performance for fatal accidents in commercial aviation, and which risk factors are most influential according to the Logistic Regression model? By systematically comparing these methods, the analysis not only evaluates various risk factors and their complex interactions but also provides critical insights for accident prediction. Ultimately, the objective is to highlight multi-factorial risk scenarios and generate actionable knowledge that supports evidence-based risk management and ongoing efforts to reduce fatalities in commercial aviation.

Literature Review

Previous research papers in aviation safety have shown that human factors (flight crew qualifications and age), technical-operational aspects, and environmental factors are important predictors of fatal accidents. These findings provide a strong theoretical framework for applying modern predictive methods, such as Logistic Regression and Random Forest, used in this study.

Particularly, research on commercial aviation safety has long emphasized that accidents emerge from complex, multi-layered socio-technical systems rather than single-point failures. This foundational understanding was introduced by Reason (2000), who used the Swiss Cheese Model (SCM) through theoretical analysis to illustrate how latent organizational failures lead to accidents. His research revealed that incidents typically occur when multiple safety layers, such as organizational, supervisory, environmental, and frontline factors, align to permit hazards to pass through barriers and defenses (Figure.2). Shabani, Jerie, and Shabani (2023) conducted a comprehensive multi-industry review spanning healthcare, aviation, and transportation to assess the application of SCM in identifying system vulnerabilities. Their analysis showed strong empirical and case study support for SCM's core principle that multiple defenses must fail simultaneously for serious accidents to occur. Additionally, their findings revealed that organizational factors such as safety culture, leadership effectiveness, communication systems, and proactive risk management practices play crucial roles in maintaining the integrity and effectiveness of defensive layers. Ultimately, this perspective provides crucial theoretical justification for multivariate analytical approaches and underscores the necessity for predictive models capable of capturing complex, interdependent relationships among risk factors.

Building upon this view, Wiegmann and Shappell (2001) aimed to apply the Human Factors Analysis and Classification System (HFACS) specifically for commercial aviation accidents. Through systematic analysis of accident investigation data (NTSB and the FAA), they classified errors into unsafe acts, preconditions, supervisory issues, and organizational influences. This framework acts as a guide for consistent feature extraction for quantitative modeling. Further validation of the HFACS framework was provided by Ergai et al. (2016), who demonstrated strong evidence for both inter-coder consistency and intra-coder reliability over time. After training 125 safety professionals from various industries to classify 95 real-world causal factors, the study found acceptable reliability levels, particularly after training and when coders repeated classification tasks. These results confirm HFACS as a credible framework for systematically identifying risk factors. Together, these studies provide essential groundwork for selecting structured human and organizational features into quantitative prediction models.

Beyond human and organizational factors, environmental conditions remain a critical area of study. Jarošová et al. (2023) conducted a statistical analysis of aviation events from 2003 to 2022 in the Czech and Slovak Republics. Their study revealed persistent associations between specific weather phenomena (storms, icing conditions, and wind shear) and accident occurrence. This finding is consistent with those of Long and Rupp (2022), who used datasets from the National Transportation Safety Board (NTSB) and the Aviation Safety Reporting System (ASRS) from 2009 to 2018. They identified that turbulence, strong crosswinds, and approaching-phase weather hazards are statistically significant contributors to accidents. Extending this perspective, Storer, Williams, and Joshi (2017) projected increasing environmental risks by analyzing global climate simulations to assess future turbulence exposure patterns. Their modeling projected a significant rise in moderate-to-severe turbulence encounters by 2050–2080, indicating evolving risk profiles that safety systems must address. Collectively, these studies underscore the importance of incorporating weather and environmental variables into predictive accident models.

More recently, research has turned toward predictive modeling of aviation safety outcomes in the US, Canada, and Australia. Omrani et al.(2024) conducted a comparative assessment of machine learning effectiveness for predicting accident severity, applying Artificial Neural Networks (ANN), Decision Trees (DT), and Support Vector Machines (SVM) to aviation datasets. Their analysis concluded that predictive success "depends more on the type and classification of the dataset itself rather than the algorithm choice." Specifically, while SVM performed best (81% accuracy), the effectiveness of any algorithm was heavily constrained by input-data's quality and feature classification. This finding highlights a key challenge in aviation safety prediction: fatal accidents account for only a small portion of the data, leading to class imbalance.

By drawing insights from literature on these key variables, this study contributes to the ongoing discourse on aviation safety improvements. This study, together with Logistic Regression and Random Forest, addresses a critical gap in the literature by handling missing data and class imbalance with Multiple Imputation by Chained Equations (MICE) and Synthetic Minority Over-sampling Technique (SMOTE), respectively. SMOTE addresses class imbalance by generating synthetic examples of the minority class through interpolation between existing minority samples and their nearest neighbors, thereby creating a more balanced training dataset without simply duplicating existing observations (Chawla et al., 2002). Unlike traditional resampling methods, SMOTE has been shown to improve classifier performance on imbalanced datasets while reducing the risk of overfitting that occurs with simple oversampling (He & Garcia, 2009). This integrated approach enables a more nuanced understanding of the complex interactions among human factors, aircraft characteristics, and environmental conditions in predicting fatal accidents, while effectively addressing the inherent challenges of class imbalance that have limited previous predictive modeling efforts. The research aims to provide empirical evidence on the comparative performance of chosen models in aviation safety contexts, while identifying the most statistically significant risk factors that contribute to the occurrence of fatal accidents.

Methods

Data

This study uses the National Transportation Safety Board (NTSB) aircraft accident data available for download for public use at this link. The NTSB dataset comprises nine interconnected datasets that are merged together using unique identifiers. Data aggregation involved left joins of key tables, namely events, aircraft, and injury, to preserve all accident records. Following the merger, records were filtered to include only commercial operations by selecting events with FAR Part 121 and 135 classifications. Variable selection focused on retaining columns that align with four primary analytical categories: crew characteristics, environmental conditions, aircraft systems, and organizational factors. The NTSB data dictionary served as a guide to identify the relevant variables within each category and ensure proper interpretation. All of these steps have been done in SQL.

The dataset is a cross-sectional dataset covering 1,388 aviation accident observations from 2008 to 2024. Each observation represents an individual accident event as the unit of analysis. The dependent variable, "Fatal", is a binary indicator equal to 1 if the accident resulted in at least one fatality and 0 otherwise.

The variable description table (Table 4) provides comprehensive information about each variable included in this analysis, such as variable name, symbol used, and a brief description of what each variable represents. As shown in the table, missing data rates vary considerably across variables and appear to be missing at random. For example, variables like

"Temperature" and "Event Year" have no missing values. In contrast, variables such as "crew_sex" (44.81%) and "crew_age" (33.72%) show substantial missingness. Given these patterns, MICE is well-suited for this analysis because it is specifically designed to handle datasets with different levels and types of missing data. By using MICE, we can obtain more accurate and reliable statistical results (van Buuren & Groothuis-Oudshoorn, 2011).

Summary statistics for numeric variables of interest are presented in Table 5. Several noteworthy patterns emerge from the descriptive analysis. The average crew age is 44.05 years, suggesting a relatively mature and experienced pilot population in commercial aviation. The fatality rate is 8%, indicating that fatal accidents account for a relatively small proportion of all commercial aviation accidents. This low rate suggests a class imbalance issue.

Weather conditions at the time of accidents were generally favorable, with a mean visibility of 9.20 miles and an average wind speed of 10.03 knots. The temporal distribution shows accidents occurring throughout all hours of the day, with a median event time of 1,526 (15:26 PM), suggesting a slight concentration during afternoon hours. Some data quality issues are evident. For example, a maximum crew age of 115 years, and a recorded temperature of 104 °C, along with a surprisingly high median of 43 °C. These values highlight data points that require further examination and cleaning before analysis.

Predictive Models

This study uses a sequential analytical approach that integrates logistic regression with Lasso regularization and Random Forest classification. This combination is purposefully chosen for its complementary strengths in addressing common challenges in aviation safety datasets—specifically, multicollinearity among predictors and the potential for non-linear relationships between predictors and outcomes.

Logistic regression with LASSO enables robust feature selection and model interpretability, especially useful when dealing with highly correlated predictors like in aviation accident datasets (Scarcioffolo, 2024; IJIRT, 2020; Semantic Scholar, 2020). LASSO-based regularization methods have been shown to improve both predictive accuracy and interpretation. This method shrinks less-informative coefficients toward zero, thus preventing overfitting and focusing attention on the most influential risk factors (IJIRT, 2020; Semantic Scholar, 2020; IMSTAT, 2025). The coefficient estimates are directly interpretable with odds ratios, providing clear communication to aviation safety stakeholders and supporting evidence-based policy and operational decisions. (IJIRT, 2020; Semantic Scholar, 2020).

In addition, ensemble tree-based methods like Random Forest are increasingly recognized in the literature for their ability to capture complex, nonlinear interactions and higher-order dependencies among predictors without requiring explicit model specification—a capability especially relevant to aviation, where accident causation often involves interacting technical, human, and environmental factors (IJIRT, 2020; ERAU, 2020; Scholarspace, 2024). Random Forest's variable importance metrics, such as permutation importance and mean decrease in impurity, provide an orthogonal measure to regression-based effect sizes, allowing analysts to cross-validate findings and surface robust data-driven predictors of rare events (Scarcioffolo, 2024; ERAU, 2020).

However, due to the moderately sized sample (N=1,388), it is important to interpret Random Forest results in conjunction with logistic regression coefficients because the stability of variable importance rankings can decrease with smaller datasets (IJIRT, 2020). As such, this study emphasizes concordant predictors across cross-validation folds and methods to ensure statistical robustness and practical interpretability (IJIRT, 2020; ERAU, 2020; Scholarspace, 2024; Semantic Scholar, 2020). Thus, both Logistic Regression and Random Forest balance interpretability and flexibility, enabling accurate predictions while providing explanations relevant to aviation safety stakeholders. (Scarcioffolo, 2024; Semantic Scholar, 2020; IMSTAT, 2025).

Model Specification

All categorical variables are dummy-coded, with reference groups shown in Table 1.

Table 1: Categorical Variables and Their Reference Groups

Variable	**Reference Group**
FAR Part	121
Crew Category	Pilot (PLT)
Crew Sex	Male (M)
Medical Certificate	Class 1 (CL1)
Weather Condition	Visual Meteorological Conditions (VMC)
Phase Group	Standing/Parking

Note. The reference groups are selected based on the most common or standard category within each variable

Hypothesized Equation:

$$\text{Fatal}_i = \beta_0+\beta_1(\text{Human Factors})_i+\beta_2(\text{Environmental Factors})_i+\beta_3(\text{Operational Factors})_i+\varepsilon_i$$

where Fatal_i represents the probability of a fatal outcome in accident i, and ε_i denotes the random error term. Each β coefficient captures the marginal effect of the corresponding factor category on the likelihood of a fatal accident.

Results and Discussion

Before estimating the LASSO logistic regression model, we examined potential multicollinearity among predictor variables to ensure coefficient stability. Pearson correlation coefficients (Figure 3) among continuous predictors revealed generally weak relationships, with all absolute values below 0.33 and substantially below the commonly used threshold of

$|r| > 0.70$ (Krehbiel, 2004). Specifically, only one out of 21 unique variable pairs exhibited moderate correlation ($r = -0.33$ between temperature and wind speed), while all others showed $|r| < 0.20$. Variance Inflation Factors (VIF) (Figure 4) were computed for all predictors, with all values remaining well below the conservative threshold of 5.0, suggesting there is no significant multicollinearity (Hair et al., 2010; O'Brien, 2007). Slightly higher VIFs among "Phase of Flight" variables (2.5-3.0) reflect expected interdependence due to categorical encoding, while VIFs for crew category, weather condition, and FAR Part ranged from 1.8 to 2.2, and environmental variables remained near 1.5. These diagnostics confirm that the predictors are sufficiently independent to conduct LASSO. The initial model estimation is expressed as follows:

Estimated Equation 1:

$$\begin{aligned}\operatorname{logit}\big(P(\text{Fatal}_i = 1)\big) = &- 5.17 - 2.23\,(\text{Phase: Taxi})_i - 0.36\,(\text{Phase: Takeoff})_i \\ &+ 0.42\,(\text{Phase: Initial Climb})_i - 2.56\,(\text{Phase: Descent})_i \\ &+ 2.74\,(\text{Phase: Unknown/Other})_i + 3.37\,(\text{FAR Part 135})_i \\ &- 0.30\,(\text{Crew Category: CPLT})_i + 0.01\,(\text{Crew Age})_i \\ &- 0.05\,(\text{Crew Sex: Female})_i - 0.06\,(\text{Medical Certificate: CL2})_i \\ &+ 0.51\,(\text{Weather: IMC})_i - 0.02\,(\text{Visibility})_i + 0.01\,(\text{Temperature})_i + \varepsilon_i\end{aligned}$$

However, as shown in Figure 5, the number of fatal aircraft accidents fluctuates considerably over time for both Air Carrier and Air Taxi & Commuter operations. This pattern suggests the presence of year-specific factors such as regulatory reforms, macroeconomic conditions, or technological improv ements that may systematically influence accident outcomes. To account for these unobserved temporal effects, year fixed effects were incorporated into the model, resulting in the following specification:

Estimated Equation 2:

$$\begin{aligned}\text{logit}\big(P(\text{Fatal}_i = 1)\big) = &-5.74 - 2.03\,(\text{Phase: Taxi})_i - 0.26\,(\text{Phase: Takeoff})_i \\ &+ 0.45\,(\text{Phase: Initial Climb})_i + 0.02\,(\text{Phase: En Route})_i \\ &- 2.43\,(\text{Phase: Descent})_i + 3.64\,(\text{Phase: Unknown/Other})_i \\ &+ 3.46\,(\text{FAR Part 135})_i + 0.01\,(\text{Crew Age})_i \\ &+ 0.46\,(\text{Weather: IMC})_i - 0.01\,(\text{Visibility})_i + 0.02\,(\text{Temperature})_i + \varepsilon_i\end{aligned}$$

As for class imbalance, the fatal accidents account for only 8% of the dataset. However, we have chosen not to use the Synthetic Minority Over-sampling Technique (SMOTE) to address this class imbalance issue. As this rate represents the true underlying risk structure of commercial aircraft, oversampling would distort actual risk levels (Batuwita & Palade, 2013). Also, SMOTE-generated cases may fail to capture the complex causal dynamics of real fatal events, causing the model to learn from artificial patterns (He & Garcia, 2009). Therefore, we will address class imbalance through LASSO regularization to prevent overfitting and evaluate performance using metrics designed for imbalanced data, namely precision, recall, and F1-score (Chawla et al., 2002). Ultimately, the Logistic Regression model with year fixed effects is selected for comparison with the Random Forest model.

Model Performance Comparison

Table 2: Random Forest Feature Importance

Feature	**Importance (%)**
Phase of Flight	19.04
FAR Part	18.85
Crew Category	11.54
Visibility	8.75
Weather Condition	8.32
Medical Certificate	6.72
Crew Age	6.55
Temperature	4.64
Wind Speed	3.46
Latitude	3.27
Crew Sex	2.77
Event Year	2.71
Event Time	2.60
Longitude	1.70

Note: Feature importance values reflect each variable's contribution to predictive performance.

The Random Forest model confirms ”Phase of Flight” and ”FAR Part” as the two dominant predictors, each contributing approximately 19% to overall model importance (Table 2). These results are consistent with our Logistic Regression model, where both variables also have the largest and most statistically significant coefficients. However, the two models have different performance characteristics when evaluated on classification metrics. As shown in Figure 1, the ROC curves reveal that the Random Forest model achieves superior overall discriminatory power with an AUC of 0.85 compared to the Lasso model's AUC of 0.79.

The Random Forest curve (solid blue line) demonstrates consistently higher true positive rates across most false positive rate thresholds, indicating better separation between fatal and non-fatal accident classes. The optimal operating points marked on each curve, (0.826, 0.791) for Lasso and (0.675, 0.854) for Random Forest, illustrate the trade-off between specificity and sensitivity, with Random Forest achieving higher sensitivity at the cost of reduced specificity.

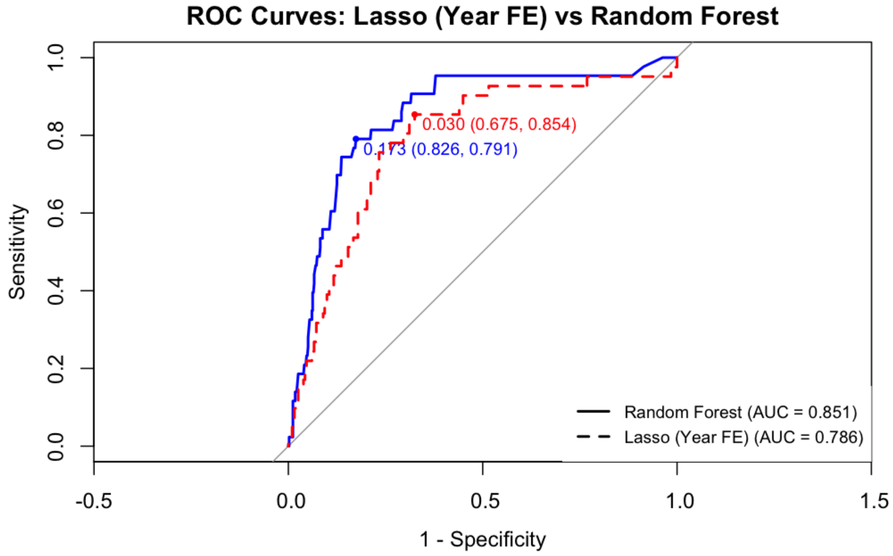

Figure 1: The Random Forest model demonstrates stronger overall discriminatory power (AUC = 0.851) compared to Lasso (AUC = 0.786)

However, the AUC advantage of Random Forest does not equivalent to superior performance for identifying fatal accidents. While both models achieve an identical overall accuracy of 0.91 (Table 3), this metric is misleading in the context of severe class imbalance, where 92% of accidents are non-fatal. The LASSO model demonstrates substantially higher precision (0.28 vs. 0.14), recall (0.12 vs. 0.02), and F1-score (0.17 vs. 0.04) compared to Random Forest. The LASSO model's recall of 0.12 indicates it identifies 12% of actual fatal accidents, whereas Random Forest identifies only 2%, a sixfold difference that has significant practical implications for aviation safety applications. The LASSO model's precision of 0.28 means that when it predicts a fatal accident, it is correct 28% of the time, representing substantial improvement over the 8% base rate and double the precision of Random Forest (0.14).

Table 3: Model Performance Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Lasso (Year FE)	0.91	0.28	0.12	0.17
Random Forest	0.91	0.14	0.02	0.04

The Random Forest model's poor minority class performance despite superior AUC likely stems from the severe class imbalance (92:8 ratio), which causes the ensemble algorithm to optimize primarily for majority class accuracy. The relatively modest sample size (N = 1,388, with only 111 fatal cases) may also be insufficient for Random Forest to learn the complex patterns in fatal accidents. In contrast, the LASSO regularization approach effectively handles class imbalance by preventing overfitting to the majority class while maintaining interpretable coefficients that provide actionable insights for safety practitioners.

In conclusion, the LASSO Logistic Regression with year-fixed effects is the preferred model for aviation safety analysis. Beyond its superior practical performance metrics, it offers interpretable odds ratios that directly inform evidence-based policy decisions and maintain a better balance between precision and recall. The model's ability to identify 12% of fatal accidents indicates strong predictive power. This insight can inform targeted resource allocation and risk mitigation in high-risk contexts, such as FAR Part 135 operations and critical flight phases.

Interpretation of Logistic Regression Results

"Phase Group" is a critical determinant of accident fatality. Relative to the reference category of Standing/Parking, accidents during the Taxi phase demonstrate substantially reduced fatality risk (OR = 0.13), representing an 87% decrease in fatal outcome likelihood.

The Descent phase shows similarly strong protective effects (OR = 0.09), with 91% lower fatal accident odds. Conversely, the Initial Climb phase exhibits elevated risk (OR = 1.57), indicating 57% higher fatality likelihood during this critical flight segment where aircraft operate at low altitude with limited emergency options. Most notably, accidents categorized as Unknown/Other phase demonstrate dramatically higher fatality odds (OR = 38.16), likely reflecting catastrophic structural failures, mid-air collisions, or events of such severity that the flight phase could not be reliably determined during investigation.

Operational context significantly influences accident severity. FAR Part 135 operations (air taxi and commuter flights) exhibit the strongest single-variable effect (OR = 31.88), indicating approximately 32 times greater fatal accident odds compared to Part 121 scheduled airline operations. This substantial disparity reflects systemic differences in aircraft size, equipment redundancy, operational environments, and regulatory oversight. Part 135 operations typically involve smaller aircraft with fewer safety systems, operate in more challenging environments, including remote locations with limited infrastructure, and face different regulatory requirements regarding crew training and maintenance standards. Among human factors, only crew age demonstrates a statistically detectable effect, though the magnitude is modest. Each additional year of crew age corresponds to a 1% increase in fatal accident odds (OR = 1.01). While this effect is substantively small, it may reflect increased exposure to accidents over longer careers, potential age-related declines in reaction time, or cohort effects in training standards. The removal of crew category and medical certificate variables from the final model indicates these factors provide no additional predictive power once operational context and flight phase are controlled.

Environmental conditions contribute significantly to fatality risk. Accidents under Instrument Meteorological Conditions (IMC) show an odds ratio of 1.58, indicating 58% higher fatal accident likelihood compared to Visual Meteorological Conditions (VMC). This finding underscores the compounding effects of adverse weather—poor conditions not only increase accident probability but also worsen outcomes when accidents occur due to delayed emergency response and reduced pilot situational awareness. Finally, visibility and temperature have odds ratios of approximately 1.00, suggesting negligible independent effects on fatality once other operational and environmental factors are controlled.

These results align with prior aviation safety research. The primacy of operational context (FAR Part classification) and flight phase is consistent with Reason's (2000) Swiss Cheese Model, which emphasizes that accidents emerge from multi-layered system failures rather than single factors. The substantial FAR Part 135 effect corroborates Wiegmann and Shappell's (2001) HFACS framework findings that organizational and supervisory factors significantly influence accident outcomes. The persistent weather effects support Jarošová et al. (2023) and Long and Rupp (2022), who identified adverse meteorological conditions as statistically significant contributors to aviation accidents. The modest crew age effect is consistent with human factors literature showing that demographic variables contribute insignificantly to accident severity once operational and environmental contexts are controlled (Ergai et al., 2016).

Conclusion

This study identified critical risk factors associated with fatal outcomes in commercial aviation accidents through systematic comparison of Logistic Regression with LASSO regularization and Random Forest classification. Analyzing 1,388 accident records from the National Transportation Safety Board database (2008-2024), the research examined how human factors, environmental conditions, and operational contexts influence accident severity while addressing methodological challenges of missing data and class imbalance inherent to rare-event prediction.

Three primary risk factors emerged with substantial effects on fatality likelihood. FAR Part 135 operations demonstrate approximately 32 times higher fatal accident odds compared to Part 121 scheduled airlines, reflecting systemic differences in aircraft size, equipment redundancy, operational environments, and regulatory oversight. Phase of flight proved equally critical, with Unknown/Other phases showing 38 times higher fatal accident odds, while taxi and descent phases exhibited strong protective effects. Instrument Meteorological Conditions increased fatality risk by 58% compared to visual conditions, underscoring the compounding effects of adverse weather. Random Forest feature importance analysis corroborated these findings, confirming Phase of Flight and FAR Part as the two most influential predictors.

Model performance evaluation revealed important trade-offs between overall discrimination and practical prediction. While Random Forest achieved superior AUC (0.851 vs. 0.786), it demonstrated substantially inferior minority class performance with a recall of only 0.02 compared to LASSO's 0.12 and F1-scores of 0.04 versus 0.17. The LASSO logistic regression with year fixed effects emerges as the preferred specification, offering superior practical performance and interpretable odds ratios for evidence-based policy decisions. However, modest recall values indicate neither model achieves sufficient predictive power for standalone early warning deployment, reflecting the fundamental challenge of forecasting rare events where fatal accidents constitute only 8% of observations.

The findings yield actionable policy implications. The 32-fold increase in fatal accident odds for Part 135 operations provides strong empirical justification for enhanced regulatory resources, including mandatory advanced safety equipment (terrain awareness systems, weather radar, enhanced ground proximity warning systems), rigorous training requirements, and more frequent safety audits. Policymakers should systematically bridge the regulatory gap between Part 135 and Part 121 standards where feasible, particularly for operations in challenging environments. The pronounced fatality variation across flight phases suggests tailored safety interventions: comprehensive training for initial climb operations, improved emergency protocols when phase classification is compromised, and enhanced automation safeguards during critical segments. The persistent IMC effect reinforces the importance of conservative weather decisions, supporting continued investment in real-time monitoring systems, advanced instrument training, and data-driven go/no-go protocols.

Several limitations suggest future research directions. The modest predictive performance reflects NTSB database constraints lacking detailed information on organizational safety culture, pilot training records, maintenance histories, and crew resource management—factors theory suggests are critical to accident causation. Future research should integrate additional data sources, including FAA safety audits, airline training databases, and maintenance records. The cross-sectional design precludes causal inference; quasi-experimental approaches such as difference-in-differences analyses of regulatory changes would strengthen intervention guidance. Temporal dynamics remain underexplored—examining how risk factors evolve following specific mandates (such as post-737 MAX requirements) would illuminate intervention effectiveness. Developing separate models for Part 121 versus Part 135 operations could reveal context-specific patterns masked by aggregate analysis.

Methodologically, future work should explore techniques designed for extreme class imbalance: cost-sensitive learning algorithms that penalize minority class misclassification, anomaly detection approaches treating fatal accidents as outliers, and ensemble methods combining multiple sampling strategies. Deep learning architectures modeling temporal sequences—such as recurrent neural networks applied to flight data recorder information—may capture dynamic risk patterns that cross-sectional models cannot. International comparative analysis incorporating ICAO global databases would enable identification of best practices from high-performing safety systems and assess whether findings generalize beyond U.S. regulatory contexts.

Commercial aviation remains remarkably safe, with fatal accidents constituting only 8% of this dataset and global rates approaching one per million flights. However, as Boeing 737 MAX tragedies demonstrated, even rare catastrophic events erode public confidence, inflict severe reputational and financial harm, and precipitate fundamental regulatory changes. This study contributes to ongoing safety improvements by providing robust empirical evidence—convergent across distinct analytical methods—identifying high-risk operational contexts warranting prioritized attention. While predictive accuracy remains insufficient for operational deployment, the interpretable findings offer clear guidance for evidence-based resource allocation: enhanced Part 135 oversight, phase-specific risk mitigation, and continued emphasis on weather-related protocols. Ultimately, sustained aviation safety depends on multi-layered defenses consistent with Reason's Swiss Cheese Model, where data-driven identification of where defensive layers most frequently align enables strategic reinforcement of system weaknesses.

References

[1] Batuwita, R., & Palade, V. (2010). Efficient resampling methods for training support vector machines with imbalanced datasets. In Proceedings of the 2010 International Joint Conference on Neural Networks (pp. 1–8). Barcelona.

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

[3] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874.

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

[5] He, H., & Garcia, E.A. (2009). *Learning from Imbalanced Data. Knowledge and Data Engineering, IEEE* , *21*(9), 1263 – 1284.

[6] McKinney, W. (2010). Data structures for statistical computing in Python. In S. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference* (pp. 56–61).

[7] Menard, S. (2002). *Applied logistic regression analysis* (2nd ed.). Sage Publications.

[8] Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, *49*(12), 1373–1379.

[9] Scarcioffolo, A. Classification. Advanced Predictive Methods. Denison University (unpublished manuscript).

[10] Scarcioffolo, A. Tree-Based Methods. Advanced Predictive Methods. Denison University (unpublished manuscript).

[11] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67.

[12] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, *67*(2), 301–320.

Appendix

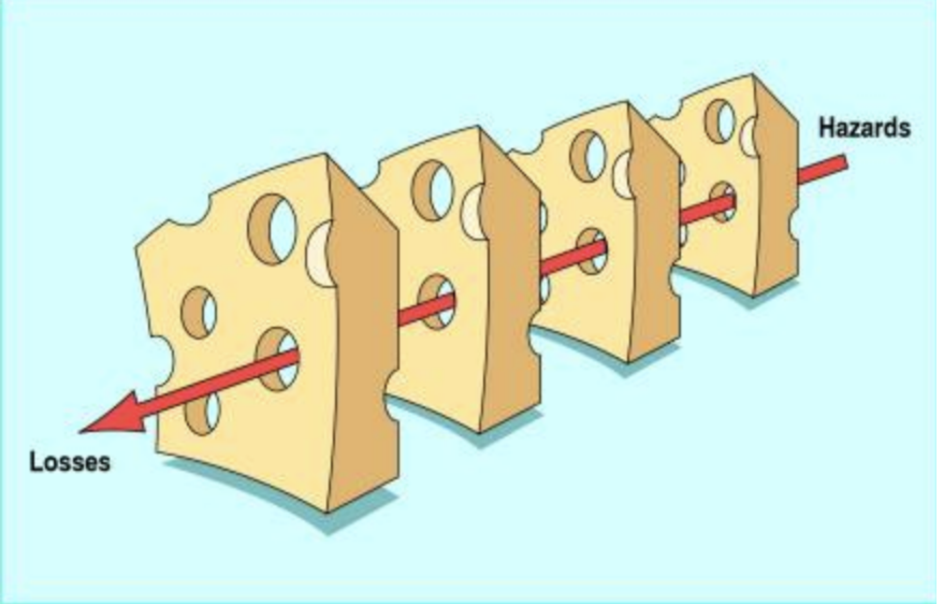

Figure 2: Swiss Cheese Model

Variable	**Symbol**	**% Missing**	**Data Type**	**Description**
Dependent Variable				
Fatal	fatal	0.00	Binary	Indicates whether the accident had at least one fatality (1 = fatal, 0 = non-fatal).

Variable	Symbol	% Missing	Data Type	Description
Predictors				
FAR part	far_part	0.00	Categorical	FAA regulation category (e.g., Part 121, Part 135).
Latitude	latitude	6.99	Numeric	Geographic latitude coordinate of the event.
Longitude	longitude	6.99	Numeric	Geographic longitude coordinate of the event.
Phase of flight	phase_group	1.44	Categorical	Phase of flight when the accident occurred (e.g., takeoff, landing).
Crew category	crew_cat	26.59	Categorical	Crew member classification (e.g., pilot, copilot).
Crew age	crew_age	33.72	Numeric	Age of the primary crew member.
Crew sex	crew_sex	44.81	Categorical	Sex of the crew member (M/F).
Medical certificate	med_certf	32.35	Categorical	Pilot's medical certificate type (Class 1, 2, or 3).
Weather condition	wx_cond	14.19	Categorical	Meteorological conditions during the event (e.g., VMC, IMC).
Wind speed	wind_speed	37.82	Numeric	Wind speed (knots) at the time of the accident.

Variable	Symbol	% Missing	Data Type	Description
Visibility	visibility	29.76	Numeric	Horizontal visibility (miles) during the accident.
Temperature	temperature	0.00	Numeric	Ambient air temperature (°C) at the accident site.
Identifiers				
Event ID	event_ID	0.00	Character	Unique identifier for each accident event.
Event time	event_time	1.01	Numeric	Local time of the accident in 24-hour format.
Event year	event_year	0.00	Year	Year when the accident occurred.
Event state	event_state	12.61	Categorical	US state where the event occurred.
Outcome Variables				
Total fatalities	total_fatal	0.00	Numeric	Total number of fatalities recorded in the event.
Probable cause	cause	15.71	Text	Narrative statement describing the probable cause.

Table 5: Summary Statistics for Numeric Predictors ($N = 1{,}388$)

Variable	**Mean**	**Median**	**SD**	**Min**	**Max**
Crew age (years)	44.05	44.00	13.38	20.00	115.00
Event time (HHMM)	1,316.10	1,526.00	761.69	0.00	2,359.00
Event year	2,015.32	2,015.00	5.02	2,008.00	2,024.00
Fatal (1 = fatal)	0.08	0.00	0.50	0.00	1.00
Latitude (degrees North)	29.00	33.94	20.42	-34.84	70.49
Longitude (degrees East)	-69.07	-81.67	61.17	-170.71	596.88
Phase of flight (coded)	388.74	402.00	136.83	100.00	990.00
Temperature (°C)	38.17	43.00	32.27	-80.00	104.00
Visibility (miles)	9.20	10.00	5.92	0.00	100.00
Wind speed (knots)	10.03	8.00	8.35	0.00	97.00

Note. The table summarizes key numeric variables used in the analysis. All continuous variables are presented with means, medians, standard deviations, and observed ranges.

Table 6: Logistic Regression Coefficients and Odds Ratios

Variable	**Coefficient**	**Odds Ratio**
(Intercept)	-5.17	0.01
Phase: Taxi	-2.23	0.11
Phase: Takeoff	-0.36	0.70
Phase: Initial Climb	0.42	1.52
Phase: Descent	-2.56	0.08
Phase: Unknown/Other	2.74	15.43
FAR Part 135 (Air Taxi & Commuter)	3.37	28.94
Crew Category: CPLT	-0.30	0.74
Crew Age	0.01	1.01
Crew Sex: Female	-0.05	0.95
Medical Certificate: CL2	-0.06	0.94
Weather Condition: IMC	0.51	1.67
Visibility	-0.02	0.98
Temperature	0.01	1.01

Table 7: Logistic Regression with Year Fixed Effects (Chosen Model)

Variable	**Coefficient**	**Odds Ratio**
Intercept	-5.74	0.003
Phase: Taxi	-2.03	0.13
Phase: Takeoff	-0.26	0.77
Phase: Initial Climb	0.45	1.57
Phase: En Route	0.02	1.02
Phase: Descent	-2.43	0.09
Phase: Unknown/Other	3.64	38.16
FAR Part 135 (Air Taxi & Commuter)	3.46	31.88
Crew Age	0.01	1.01
Weather Condition: IMC	0.46	1.58
Visibility	-0.01	0.99
Temperature	0.00	1.00

Note: The model includes year fixed effects (coefficients suppressed). Odds Ratios are exponentiated coefficients from the logistic regression model.

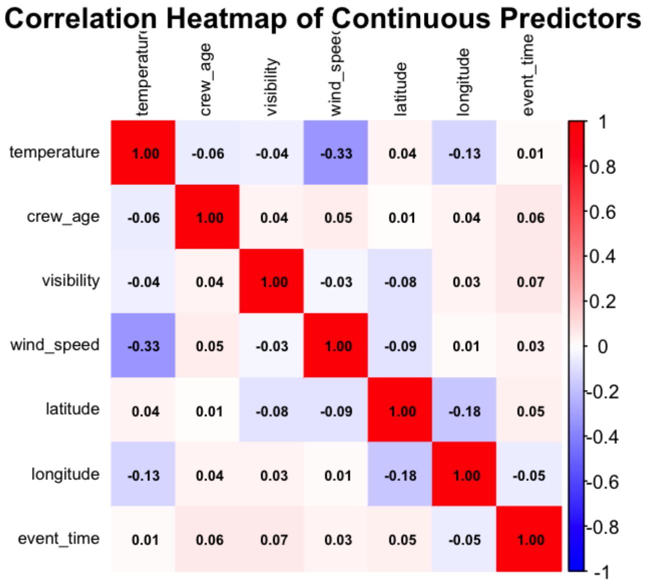

Figure 3: The correlation heatmap demonstrates that most numeric predictors are weakly correlated, with only one variable pair exhibiting a moderate negative correlation

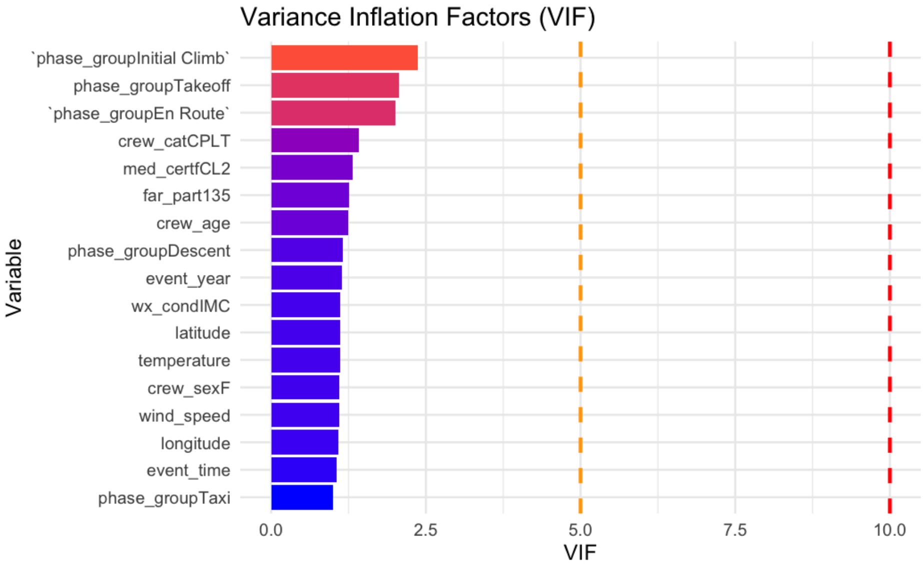

Figure 4: Variance inflation factors indicate low multicollinearity among predictors

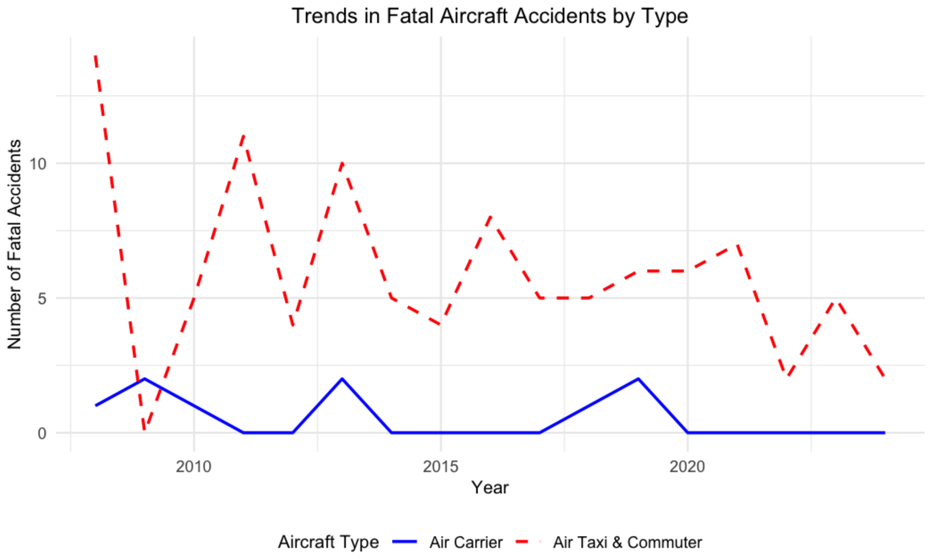

Figure 5: Yearly variations in fatal accident counts highlight the need for year fixed effect