POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Forecasting of low frequency high severity events: A commercial aviation safety study for airline insurance

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Pietro Spina**

Student ID: 968902
Advisor: Prof. Piercesare Secchi
Co-advisors: Olivier Marre, Alberto Candida De Matteo
Academic Year: 2022-23

Abstract

This thesis delves into the realm of commercial aviation accidents, seeking to enhance our understanding of their causes and the potential predictive power of advanced data analysis techniques. The research is bifurcated in two distinct phases each leveraging unique analytical methodologies.

The first phase consists in the employment of the Agresti-Coull confidence intervals to conduct an in-depth analysis of accident rates across a multitude of flights, considering a wide array of features. This comprehensive approach provides insights into the characteristic trends of accident occurrences, offering a nuanced perspective on risk factors within commercial aviation.

In the second phase, we shift our focus towards predictive modelling by implementing a gradient boosting binary safety classifier for flights. Leveraging machine learning techniques this classifier harnesses the power of data-driven insights to predict the likelihood of aviation accidents. By training on historical aviation data , the model becomes a valuable tool for evaluating risks, ultimately contributing to enhanced aviation safety insurance procedures.

This study underscores the importance of employing diverse analytical approaches to gain a comprehensive understanding of commercial aviation accidents. By combining traditional statistical approaches with cutting-edge machine learning techniques, we not only identify critical risk factors but also develop a predictive tool to proactively address safety concerns, ultimately fostering a more aware aviation industry.

Keywords: aviation safety, accident rate analysis, rare events classification, gradient boosting, CatBoost.

Abstract in lingua italiana

Questo progetto di tesi si addentra nel campo degli incidenti dell'aviazione commerciale, cercando di migliorare la comprensione delle loro cause e del potenziale predittivo di tecniche avanzate di analisi dei dati. Il lavoro è suddiviso in due fasi distinte, ognuna delle quali fa leva su metodologie analitiche specifiche.

La prima fase consiste nell'impiego degli intervalli di confidenza di Agresti-Coull per condurre un'analisi approfondita in merito ai tassi di incidenti su una moltitudine di voli, considerando un'ampia gamma di caratteristiche. Questo tipo di approccio consente di evidenziare le principali tendenze relative all'occorrenza di incidenti aerei, offrendo una prospettiva sui fattori di rischio nell'ambito dell'aviazione commerciale. dell'aviazione commerciale.

Nella seconda fase, la ricerca si concentra sulla modellazione predittiva implementando un classificatore binario, tramite gradient boosting, per valutare la sicurezza dei voli. Grazie a tecniche di machine learning, questo classificatore sfrutta la potenzialità di un approccio basato su grandi quantità di dati per prevedere la probabilità di incidenti aerei. Basandosi su dati storici dell'aviazione, il modello diventa uno strumento prezioso per la valutazione del rischio, contribuendo in ultima analisi a migliorare le procedure assicurative nell'ambito dell'aviazione.

Questo lavoro evidenzia l'importanza di utilizzare diversi approcci analitici per ottenere un quadro completo degli incidenti nell'aviazione commerciale. Combinando approcci statistici tradizionali con tecniche di machine learning all'avanguardia, non solo identifica i fattori di rischio più critici, ma permette anche di sviluppare uno strumento predittivo per la sicurezza aerea, favorendo la crescità di un'industria areonautica più consapevole.

Parole chiave: sicurezza aerea, analisi dei tassi di incidenti, classificazione di eventi rari, gardient boosting, CatBoost.

Contents

Introduction

The commercial aviation industry has revolutionized global travel and connectivity, becoming an essential part of our society. While air travel is generally considered the safest mode of transportation, commercial aviation accidents continue to happen and are a cause of concern for both industry stakeholders and the public. These accidents, thought statically extremely rare, have highly significant and far-reaching consequences, affecting the safety of the passengers, air industry professionals, as well as causing immense financial losses for airlines, aircraft manufacturers and insurance companies involved.
The importance of studying commercial aviation accidents lies in the need of continuously enhancing safety standards and obtain a broader outlook on this phenomenon. Each accident, whether large scale disasters or minor accidents, provides valuable insights into potential weaknesses in aircraft design, insights regarding the safety of specific flight routes, and airlines.
Furthermore as air travel keeps growing and becoming an essential part of our day to day life the technological advancements in this field have been substantial over the last few decades, thus significantly reducing the occurrence of commercial aviation accidents, making the pattern identification process for this phenomena increasingly difficult and requiring extreme precision and granularity when it comes to data collection.

This thesis aims to investigate commercial aviation accidents, analyzing their causes. By examining past accidents and their implications, this study seeks to identify trends and recurring patterns, shedding light on potential risk factors that may persist in the current aviation safety framework. The finding of this research could be instrumental in helping aviation professionals in their work as an additional tool to be paired with their industry experience.

There is a plethora of studies conducted to analyze this issues, from the usage of the implementations of artificial neural networks and fault tree models ([18]), the more sophisticated use of Bayesian neural networks ([11]) or the implementations of ensemble machine learning and deep learning models ([14]) to less specific, but more broad, analyses of commercial aviation accident occurrence rates ([15]). The first ones although interesting often rely on extremely specific features for their analyses (like cabin pressure and other indicators registered on the aircraft before an accident [14, 18]), while the latter ([15]) offers a broader view on the commercial aviation accidents phenomenon but lacking the in depth multivariate feature analysis offered by the others.

In opposition to the previous developed studies this thesis aims at analyzing the phenomenon of commercial aviation accidents from a broad perspective on the flights, while trying to keep an in depth multivariate approach (in the later stages) to better identify patterns and trends.

In this thesis the following research questions will be addressed:

- How do commercial aviation accident rates vary among the different characteristics of the flights, and how can we evaluate these variations?
- Is it possible to evaluate the risk of future flights to turn into an accident and can modern Machine Learning techniques help us in this task?

To address these concerns the following tools will be implemented:

- Evaluation of commercial aviation accident rates through the use of specific confidence intervals, such as the Agresti-Coull confidence interval, that are more suitable to handle small proportions than the more common counter part options (see [1, 4]).
- Implement a classification model for the flights through the use of Gradient Boosting Machines (GBMs) for binary classification (see [12]), in order to have a multivariate point of view on the commercial aviation accidents phenomenon and to make predictions and evaluate the risk of each flight.

For the latter point the Python library CatBoost will be used due to the high efficiency at handling categorical features, and the tools it provides for results interpretation see ([16, 19]), when compared with its competitors (XGBoost and LightGBM).

This thesis structure is laid out as follows.

First a complete overview of the available data will be made, explaining its features and the cleaning process adopted in order to obtain a satisfactory outcome fit for our future analyses (Chapter 1). Then a chapter conducting a brief analysis of the available data is presented, in order to understand the nature of commercial aviation accident and how to handle their future analysis (Chapter 2).

Subsequently we present available theoretical options to deal with confidence interval es-

timation for binomial proportions, and select the Agresti-Coull interval as the best fit for our analysis (Chapter 3). Based on the conclusions of the previous chapter we then move on into an in depth analysis of the commercial aviation accident rates, with the objective of identifying interesting trends among flights characteristics (Chapter 4).
We then move on with a theoretical chapter explaining the inner workings of GBMs and how they can be used for rare events classification, focusing on specific model choices that will be implemented later on (Chapter 5). Subsequently we proceed with the implementation, optimization and results interpretation of flight safety classifier based on a gradient boosting approach, in order to gain deeper insights on flights behaviour (Chapter 6).
Finally we conclude this study with an overview of the obtained results, discussing its limitations and possible further developments (Chapter 7).

This study was carried out during the author's internship at the R&D department of *elseco*, an independent multi-line managing general agent and Lloyd's coverholder based in Dubai specialized in space, aviation and energy insurance. The objective of this thesis is to provide a comprehensive overview of the commercial aviation safety phenomenon, for the aviation underwriting team, and to develop a novel forecasting tool for their future underwriting purposes.

1 | Dataset

In this study two different data sources will be considered, one for the flights and one for the accidents, that will be subsequently combined in a single dataset to achieve our aviation accident analysis goal. Overall, 4 years of data (2019-2022) are considered due to lack of availability of previous years flights data, although this is not to be considered a drawback given the extremely high amount of data at our disposal.
One of the main goals of this study is to build the most comprehensive and accurate data in order to have a clear overview of aviation accidents and their behaviour. Due to the lack of a comprehensive dataset containing flights and accidents, one of the major challenges of this study has been to come up with an effective matching procedure between flights and accidents

1.1. Flights Data

The focus of this section is to give a comprehensive explanation of the flights data considered in this study, highlighting the features of interest from a more general view of a flight (operator, region, schedule, ecc.) down to its specifics (delay, METAR data, ecc.).

1.1.1. OAG Data

OAG (see [13]) is data provider for commercial aviation flights data, this data is privately owned thus no data source will be provided.
Each OAG dataset contains all the recorded information of commercial flights in a given year, in particular we have the following information at our disposal for each flight:

- Flight ID (OAG specific)
- Scheduled departure and arrival times (UTC format)
- Scheduled departure and arrival airports
- Registration number of the aircraft
- Airline operator ID

- Type of service offered for the flight
- Minutes of delay at departure and arrival
- Departure and arrival METAR ids (see 1.1.2)

It is of crucial importance to notice how ids for airline operators and aircraft allows us to retrieve additional information about this features:

- Operating region/country
- Operator fleet size
- Aircraft age
- Aircraft manufacturer
- Aircraft type (i.e. the model)

1.1.2. METAR Data

METARs (METerological Aerodrome Reports [3]) is a form of reporting weather information used by pilots and meteorologists for weather forecasting specifically in the aviation sector.
In this study METAR data will be used to access weather information for each flight at take off and landing locations and times.
METARs provide an extensive collection of weather variables; in this study the focus will be restricted to the following:

- METAR id
- TMPF - Temperature $[°F]$
- RELH - Relative humidity
- SKNT - Wind speed $[kn]$
- VSBY - Visibility $[mi]$

These variables were selected after experts opinions on the most influential weather factors in aviation accidents. It is important to highlight that the impact of weather conditions on aviation accidents has reduced significantly over the last decades, due to the prominent technological advancements that have been made in this field.

1.2. Aviation Safety Network Data

The Aviation Safety Network (ASN, see [2]) is the most reliable publicly available database of aviation accidents (both for commercial and general aviation), for the collection of this data an online data scraper has been built retrieving all accidents data available in the considered time period (2019 - 2022). ASN provides various information about the accidents; following is a list of the considered features:

- Date and time of the accident (local time)
- Registration number of the aircraft
- Phase of the flight in which the accident took place
- Geographical location of the accident (name)
- Scheduled departure and arrival airports of the flight
- Damage sustained by the aircraft
- Number of fatalities

1.2.1. Data cleaning

From the rather extensive data provided by ASN only commercial aviation flights were selected, since general aviation is not considered in this study, significantly reducing the accidents at our disposal.

Furthermore, geocoding (through the use of a specific Python library [9]) of the accidents locations has been applied in order to retrieve with relative accuracy the latitude and longitude coordinates of the accidents, which were then used to identify the correct timezone of each accident, allowing for the conversion of accidents date and time to the UTC time zone for an accurate matching with the flights.

Overall, 1614 accidents were identified, distributed over the considered time period as shown in Table 1.1.

	2019	**2020**	**2021**	**2022**
Accidents	499	233	404	478

Table 1.1: Accident frequency per year

The lower accident occurrence for the year 2020 can be easily explained by the COVID-19 pandemic, in fact compared to 2019 during 2020 the flights decreased by almost 50%.

1.3. Final Dataset

In this section the matching procedure to uniquely match an ASN accident to an OAG flight will be presented and then the quality of the obtained data will be analyzed.

1.3.1. Accident-Flight matching procedure

The final step in the data processing phase is to match the reported accidents from ASN to the flights data provided by OAG. The matching procedure was conducted as follows:

1. For each available aircraft registration number among the ASN accidents extract all the recorded flights available.
2. Among the available flights select all the flights with scheduled departure date or scheduled arrival date within a neighborhood, of length of 2 days, of the accident date.
3. Check if the departure airport recorded for the accident matches with the scheduled departure airport provided by OAG.
4. Check if the arrival airport recorded for the accident matches with the scheduled arrival airport provided by OAG.

For our matching purposes step 1 and 2 presented above are necessary for an accident and a flight to be considered a match. Notice that step 3 and 4 are not conducted in parallel, since in case the accident occurs during the take off or initial climb phases ASN might not report the arrival airport of the accident.
After this initial matching procedure we are sometimes left with accidents matched to multiple flights, to fix this this issue and finally achieve unique accident-flights matches a control procedure has been put in place that operates as follows:

1. Check that the time of the accident is exactly between the scheduled departure time and the arrival departure time. When this step fails we proceed with step 2.
2. If the damage sustained by the aircraft is of category *substantial* or *major* we select the last available flight among the candidate matches following the logical reasoning that in these cases the considered aircraft will not fly again soon. When this step fails we proceed with step 3.

3. Make use of the flight phase in which the accident occurred to select the most likely flight candidate as follows:
   (a) Take off/Initial climb: Select the candidate flight that has scheduled departure time closer to the recorded accident time in ASN.
   (b) En route: Select the candidate flight that minimizes the sum of time difference between scheduled departure/arrival and the accident time.
   (c) Approach/Landing: Select the candidate flight that has scheduled arrival time closer to the recorded accident time in ASN.

After this lengthy and computational intense, but necessary, procedure all of the matched flights have been reduced to a unique one to one match with the accidents.

1.3.2. Data quality

Finally let us go through a brief summary of results achieved in terms of data quality and the final cleaning procedures.
Over the considered time period we have the following quality of flight-accident matched data as shown in Table 1.2, where the last column represents the percentage of uniquely matched accidents among the selected candidates:

Year	No. Flights	No. Accidents	% of matched accidents
2019	31,549,494	396	80.59%
2020	16,246,688	222	66.67%
2021	19,248,202	381	63.67%
2022	24,482,124	400	73.37%

Table 1.2: Matched data quality

Overall, we have satisfactory matching quality of the data (with a total of 1399 accidents uniquely matched to a flight) considering that around 30% of the OAG flights lack an aircraft registration number, thus not making them eligible for the matching procedure.
Finally for future analyses only accidents with at least *minimal* damage caused to the aircraft will be considered, effectively eliminating all the accidents with *none* or *unknown* damage due to their lack of interest for our purposes.
It is also worth noticing how moving forward we consider only flights associated with a carefully selected list of aircraft types (see Table A.1 in Appendix A), compiled with the

help of aviation experts, in order to remove aircraft types mainly associated with cargo flights and also get rid of outdated models that are almost out of use.

The deep data granularity achieved for the flights will allow for an agile computation of the accident rates among different features (operators, countries, aircraft manufacturers, aircraft types, ecc.), so that we can swiftly compare accident rates belonging to different features sub classes and gain a deep understanding of their behavior.

2 | Preliminary analysis

In this chapter we will take a first look at the commercial aviation accident phenomenon, starting from a brief look into the ASN accidents and their characteristics, and then the rate of accidents across the 4 years in analysis (2019 - 2022) will be considered, to gain a better understanding and select the appropriate tools to deal with the issue at hand.

2.1. ASN accidents

After the previously described data cleaning measures adopted (see 1.2.1) and having gone through the appropriate flights-accidents matching procedure (see 1.3.1), we can now take a close look at the accidents and their behaviour focusing in particular on the *damage* sustained by the aircraft and *phase of the flight* in which the accident occurred.

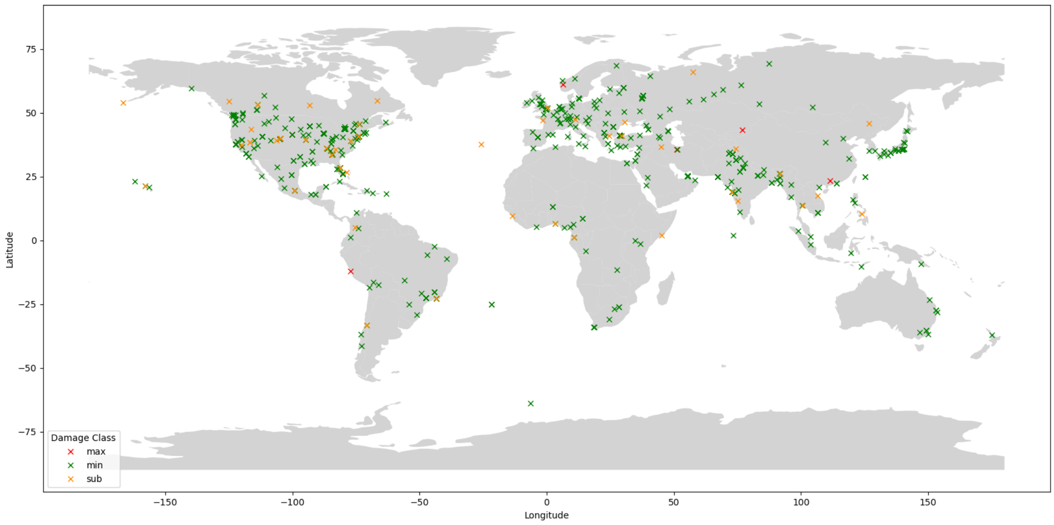


Figure 2.1: ASN accidents classified by aircraft damage in 3 classes: minimal, substantial and major

By looking at Figure 2.1 we notice that the majority of commercial aviation accidents exhibit *minimal* damage to the aircraft (approximately 89.9%), while the *substantial* (approximately 8.8%) and *major* (approximately 1.3%) damage represent an extremely small fraction of the accidents making them rare events among commercial aviation accidents which are rare events themselves. This is something to keep in my mind in our future analyses, since by nature minimal damage are more random and might be dependent from external factors not at our disposal (i.e. human error).

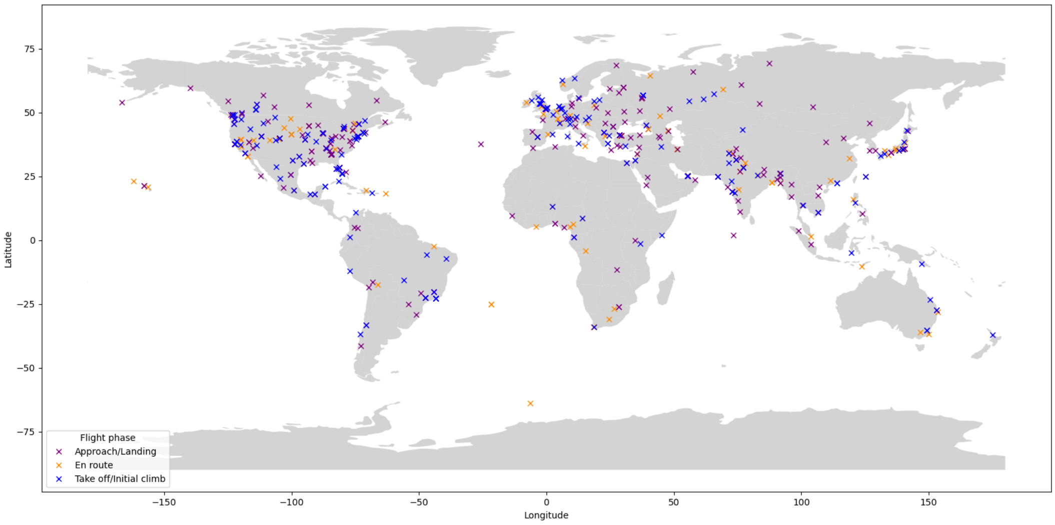


Figure 2.2: ASN accidents classified by flight phase in 3 classes: Take off/Initial climb, En route and Approach/Landing

By visual inspection of Figure 2.2 most accidents, as expected, happen when the plane is either in the *Take off/Initial climb* or the *Approach/Landing* phases (approximately 80.3% combined), this feature is crucial in achieving accurate flight-accident matches (see 1.3.1) since the *En route* accidents are the hardest to appropriately match. Furthermore, this is something to keep in mind moving forward when arrival and departure METAR data will be considered.

Finally let us shed a light on ASN accidents' seasonality in order to identify any particular monthly trends.

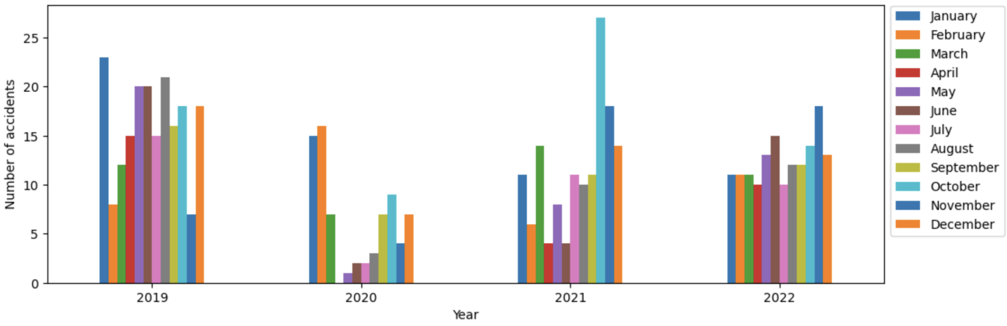


Figure 2.3: ASN accidents monthly distribution

The results depicted in Figure 2.3 are to be considered cautiously. First of all we notice the significant impact of the COVID-19 pandemic that caused a substantial reduction of the flights (and consequently of the accidents) starting from March 2020. Furthermore, the rest of the seasonality trends can be explained by the high seasons and low seasons of the aviation industry (peaks in December and during the summer months), paired with the randomness of the phenomenon.

2.2. Commercial aviation accident rates

First we proceed by removing all flights with matching registrations and all accidents with either *none* or *unknown* aircraft damage, to minimize accident reporting bias, due to different accident reporting procedure laws belonging to different countries or regions; furthermore the accident rates will be represented as percentage for visualization purposes.

In the following Table 2.1 a preliminary look into accident rates is presented:

Year	No. Flights	No. Accidents	Accident Rate (%)
2019	29,063,825	233	0.00080168
2020	14,340,483	82	0.00057181
2021	15,865,361	153	0.00096437
2022	18,489,566	181	0.00097893
TOT	77,759,235	649	0.00083463

Table 2.1: Commercial aviation accidents yearly rates

It is immediately clear the extreme rarity of such events even when considering low damages to the aircraft. We also notice a significant decrease in flights after *2019*, this is easily explained by the COVID-19 pandemic which hit the commercial aviation industry particularly hard. The same reasoning can be adopted for the extremely low accident rate recorded in 2020. Since we have to keep in mind that ASN reporting procedure is executed mainly by volunteers, it is safe to assume that its reporting is quality mainly focused on more important events losing on accuracy. Moving forward we will have to be aware at results and trends relative to 2020 (and partially also 2021) due to this issue.

The next logical step in the commercial aviation accident rates analysis would be to categorize the flights, and subsequently the accidents, according to the different categorical features at our disposal (region, country, age of the aircraft, aircraft type, delay, ecc.) in order to spot interesting patterns among them, and try to evaluate their influence on flight safety by evaluating the accident rates.

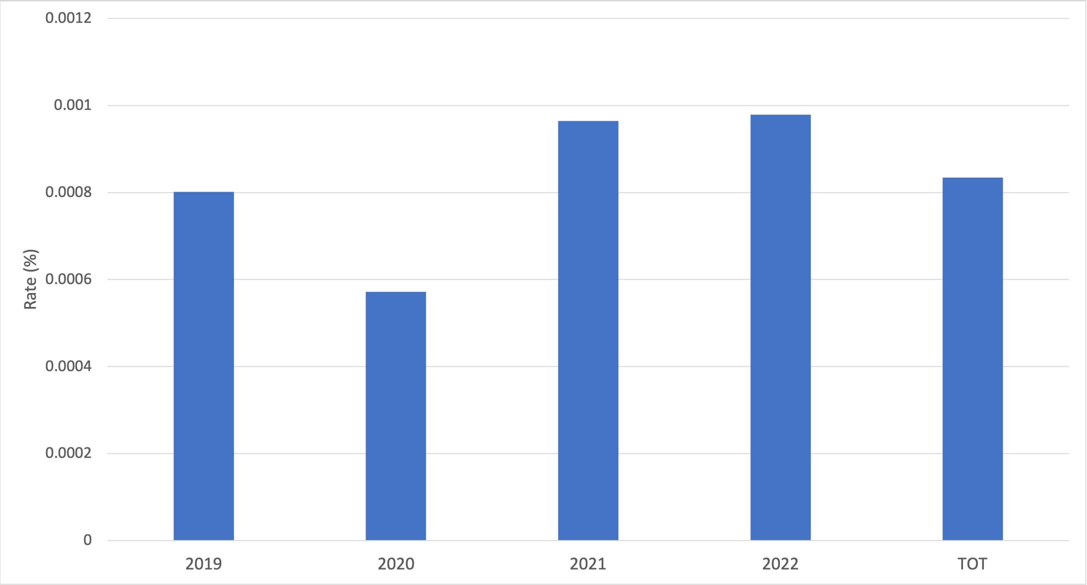


Figure 2.4: Commercial aviation accidents yearly rates.

Furthermore, by looking at Figure 2.4 we can put in perspective the differences among these accident rates, it is immediately noticeable how the considered rates show considerable variability and thus the need of a tool to evaluate such statistical differences becomes apparent, in particular we will look into binomial proportions confidence intervals (see Chapter 3).

3 | Binomial proportion confidence intervals

In the world of statistics a binomial proportion confidence interval is an interval for the estimation of the probability of successes computed from the result of a series of Bernoulli trials. In other words, a binomial proportion confidence interval acts as an estimate of the probability of success p, given that the number of trials n and the number of successes n_{S} are known.

There are several available formulas for the estimation of binomial confidence intervals and all of them share one common trait: they rely on the assumption of a binomial distribution. In general a binomial distribution is considered when an experiment is repeated a known amount of times, where each trial admits only two possible outcomes (success or failure) making the success probability (and analogously the failure probability) equal for each trial, furthermore assuming statistical independence among the trials.

3.1. Normal approximation interval

The most common example of a binomial proportion confidence interval is the *normal approximation interval* or *Wald interval*. This interval relies on approximating the distribution of a binomial variable, n_S, with a normal distribution. This kind of approximation relies on the Central Limit Theorem, which estimates the distribution of $\hat{p}$, i.e. of binomial distributed variable divided by its parameter n, as n grows; this estimate becomes extremely unreliable when the sample size is small (not of our concern for the application) or when the success probability is close to 0 or 1 (this is the main concern for us since commercial aviation accident are extremely rare events).

Using the normal approximation we obtain the following estimate for the success probability p:

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{3.1}$$

or equivalently,

$$\frac{n_{\text{S}}}{n} \pm \frac{z_{\frac{\alpha}{2}}}{n\sqrt{n}}\sqrt{n_{\text{S}}n_{\text{F}}} \tag{3.2}$$

where $\hat{p} = \frac{n_{\text{S}}}{n}$ is the estimate proportion of successes in a Bernoulli process, measured with n trials with an outcome of n_{S} successes and $n_{\text{F}} = 1 - n_{\text{S}}$ failures, and $z_{\frac{\alpha}{2}} = \Phi^{\text{-1}}\left(1 - \frac{\alpha}{2}\right)$ represents the $1 - \frac{\alpha}{2}$ quantile of a standard normal distribution which corresponds to the interval confidence level α (yielding $z_{\frac{\alpha}{2}} = 1.96$ in case of $\alpha = 0.95$).

As previously stated the main weakness of the *normal approximation interval* is its unreliability when dealing with small proportions close to 0 which makes it unfit for our purposes, thus the need of introducing a stronger interval fit to deal with this issues arises.

3.2. Agresti-Coull interval

The *Agresti-Coull interval* (see [1]) is a non-parametric interval for the estimate of binomial proportions, meaning that it does not make any assumptions on the underlying distributions of the data. This particular feature makes it a more robust choice than parametric intervals, such as the *Wald interval*, when the assumptions of the parametric interval are not met.

The *Agresti-Coull* interval is constructed in the following way.

Given n_{S} successes in n trials, define

$$\tilde{n} = n + z^2_{\frac{\alpha}{2}} \tag{3.3}$$

and

$$\tilde{p} = \frac{1}{\tilde{n}}\left(n_{\text{S}} + \frac{z^2_{\frac{\alpha}{2}}}{2}\right) \tag{3.4}$$

Then the *Agresti-Coull confidence interval* for p is obtained as follows

$$\tilde{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\tilde{p}}{\tilde{n}}(1-\tilde{p})} \tag{3.5}$$

3.3. Conclusion

It is proven by L. D. Brown and T. T. Cai (see [4]) that the *Agresti-Coull interval* is more fit than the *normal approximation interval* to deal with close to 0 proportions interval estimation providing a wider coverage.

Thus moving forward, for our interval estimation purposes of the commercial aviation accident rates, the *Agresti-Coull interval* will be considered due to its previously shown benefits when dealing with small proportions. The latter also provides an asymmetric coverage for $\hat{p}$ (symmetric for $\tilde{p}$) which is larger for values above p, when it's close to 0, which conveniently fits our interval estimation goals since it provides a larger coverage in case of higher accident rates.

4 | In depth commercial accident rates analysis

In this chapter we proceed with the explorative analysis of an extensive selection of different flight's features on the commercial aviation accident rates, starting from more generic airline operator features, such as the *operating region* and the *fleet size*, to then get into the aircraft specifics and finally look into the flight exclusive features, such as the *delay* and the *flight duration* (for an in depth overview of the considered features in the study see Section 1.1.1).

To conduct such analysis of the accident rates we will make use of the *Agresti-Coull interval* introduced in Section 3.2, due to the previously explained robustness of this approach when dealing with proportions close to 0. All the intervals represented in this chapter are evaluated at a confidence level of 95% ($\alpha = 0.95$) and considered as one-at-the-time confidence intervals for each individual rate.

Recall that, as previously mentioned (see Section 2.2), the accident rates will be represented as a percentage for better visualization purposes.

4.1. Airline operators

In this section we will analyze all the aspects that identifying a flight with its respective airline operator brings to the table. Starting from the more broad view of regional categorization of operators, to then move into a deeper look of each individual airline shining a light on the most active ones and finally focusing on the trend provided by the *fleet size* of the operators when compared with their respective accident rates.

4.1.1. Operating region

Now the regional commercial aviation accident rates will be considered. It is important to remark that the flights are classified by *operating region*, that is the region of residence of the airline operator of each flight; in most cases this coincides with the *departure region* and the *arrival region*, since the vast majority of commercial flights are not intercontinental, but that is not always the case.

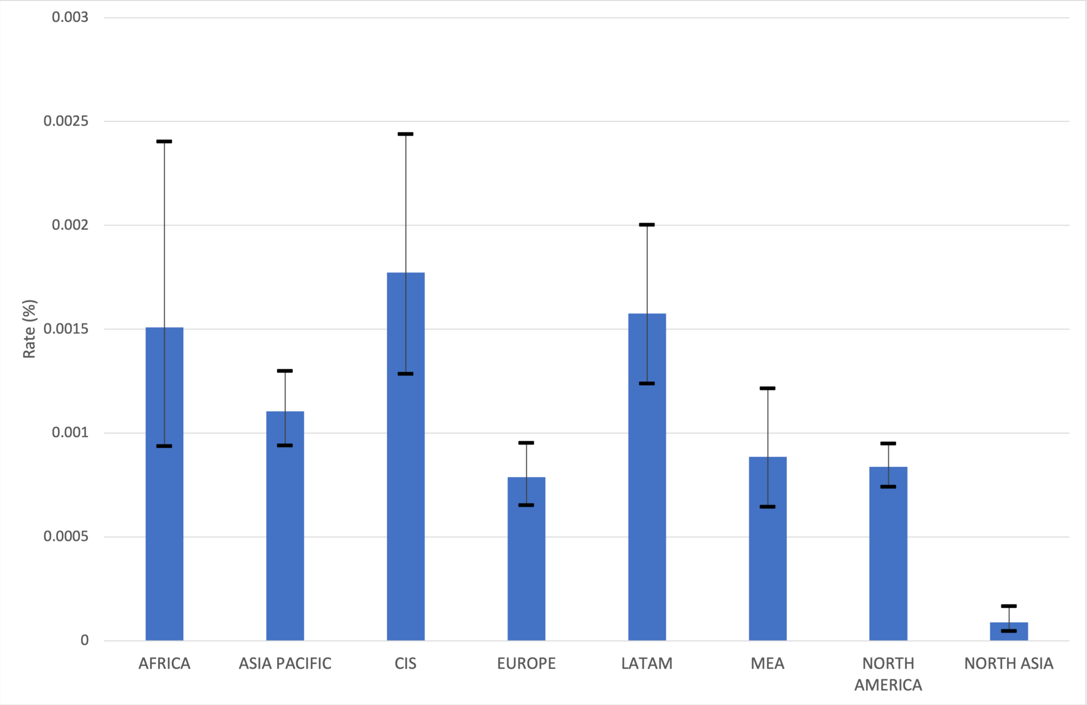


Figure 4.1: Operating region accident rates.

Figure 4.1 depicts an interesting image that is, for the most part, consistent with aviation experts opinions.

Africa, *CIS* (i.e. *Commonwealth of Independent States*, a regional organization of 12 countries that were formerly part of the Soviet Union) and *LATAM* (i.e. *Latin America*) all show the highest, and comparable, accident rates among the considered regions and their higher confidence interval amplitudes are mainly due to the lower air traffic when compared to the other regions. This view is consistent with experts' opinions which consider *Africa* as the most dangerous region shortly followed by *LATAM*, it is interesting how the behaviour of the *CIS* accident rate is higher than expected.

Asia Pacific seems to be in a category of its own which stands in between the previously mentioned higher rate category and the lower one which will be shortly discussed, this might due to the fact that this region includes, by far, the most heterogeneous group of countries in terms of wealth and thus flight safety procedures.

Moving on we get to the third group in our analysis which is made up from the wealthiest regions in the world: *Europe*, *MEA* (i.e. Middle East and North Africa) and *North America*. All of the just mentioned regions show very similar rates and a significantly smaller confidence interval amplitude mainly due to their higher air traffic. This results are uniform with the expert's opinions which consider these regions among the safest in the world.

Finally the most interesting case is by far the one of *North Asia* (i.e. China) which operates completely in a category of its own, displaying an extremely low accident rate as well as the smallest confidence interval among all the regions. Although this might surprise a naive reader, and lead to think this is due to a lack of accident reporting by the region's air traffic institutions, *North Asia* is unanimously considered the safest region to fly in by experts as well.

4.1.2. Airlines

In our study we consider an extensive list of 396 commercial airline operators from industry giants to less active and only local airlines.

For visualization purposes we will represent only the top 50 operators ranked by total amount of flights in the considered time period, which represent over 60% of the world wide air traffic and around 57% of the considered accidents.

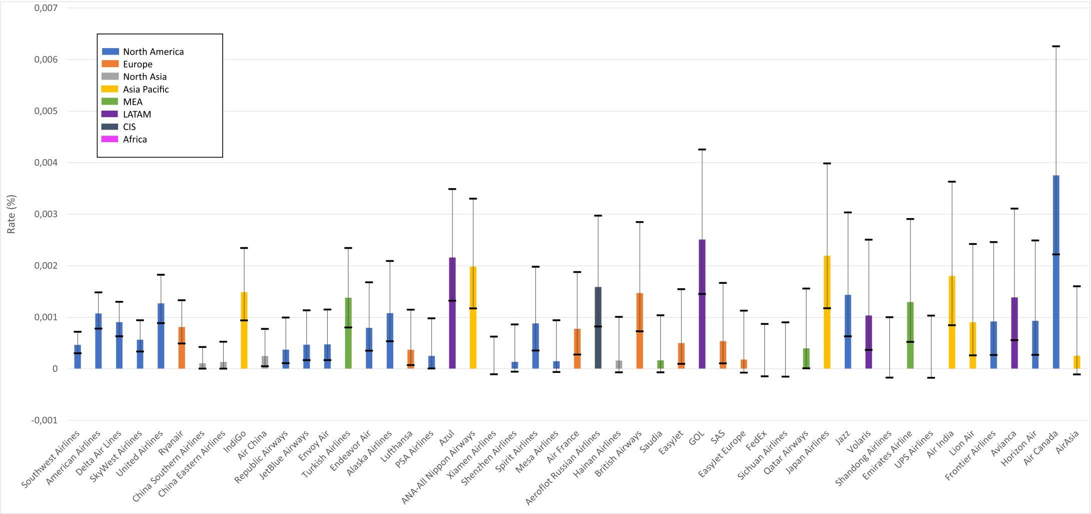


Figure 4.2: Airline operators accident rates.

Although the results shown in Figure 4.2 are of difficult interpretation, due to high amplitude and fluctuation of the confidence intervals caused by the large number of classes introduced, some interesting observations can still be made.

First of all we notice low accident rates for the top Chinese airline operators such as *China Southern Airlines*, *China Eastern Airlines*, *Air China*, *Xiamen Airlines*, *Shenzhen Airlines*, ecc. when compared to other region's top operators. This result is consistent with the already discussed regional trends in the previous section.

Furthermore, we observe how almost all of the high accident rate airline operators belong to higher rated regions, for example *Azul* operates in LATAM, while *ANA-All Nippon Airways*, *GOL*, *Japan Airlines* and *Air India* all operate in Asia Pacific. The only noticeable outlier from the regional trends is represented by *Air Canada* which operates in *North America.*

Finally, as the last step in the airline operators analysis, it is interesting to focus on the influence of the *fleet size* of airlines, that is the number of aircraft employed by an airline, on the accident rates since it is commonly believed, by aviation experts, that wealthier (and thus bigger) airlines usually are more reliable than local ones.

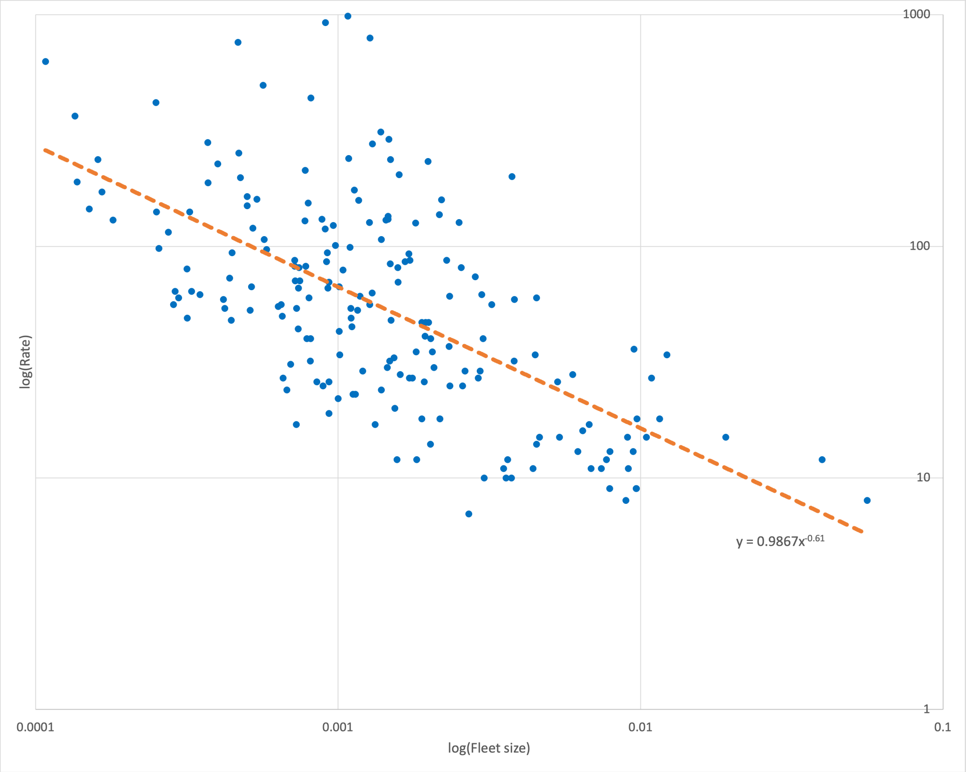


Figure 4.3: Airline operators fleet size accident rates trend.

As we can see in the log-log plot represented in Figure 4.3, a clearly decreasing trend in the accident rates is noticeable as the fleet size increases (for visualization purposes only the positive rates operators are represented). Furthermore, it is also noticeable how the spread between observation tightens as the fleet size increases, as expected.

Moreover, a power trend line is represented which gives the following relation between the two variables:

$$log(rate) = 0.9867 \cdot log(fleet\,size)^{-0.61} \tag{4.1}$$

yielding,

$$rate \approx fleet\,size^{-0.602} \tag{4.2}$$

which is to be considered carefully given the fact that:

1. null rate observations are not being considered
2. the trend line seems to be heavily leveraged by the presence of outlying observations with extremely high *fleet size*

4.2. Aircraft specifics

Moving further in our study we will consider the types of aircraft considered and their characteristics from the manufacturer, the different aircraft types and body types and finally a bin classification of the years each aircraft has been active.
Recall that, for our data cleaning purposes, we decided to consider only a specific list of aircraft types in order to remove out of use aircraft types not relevant to our analysis (see Table A.1 in Appendix A).

4.2.1. Manufacturer

Now the subset of available aircraft manufacturers at our disposal is to be considered, overall we consider *8* different manufacturers ranked in order of their respective aircraft usage.
It is worth underlining the fact that *Airbus* and *Boeing*, the two biggest manufacturers in terms of number of flights employing their aircraft, add up to over 78% of the world wide air traffic and over 72% of the active aircraft with almost equal shares in both categories.

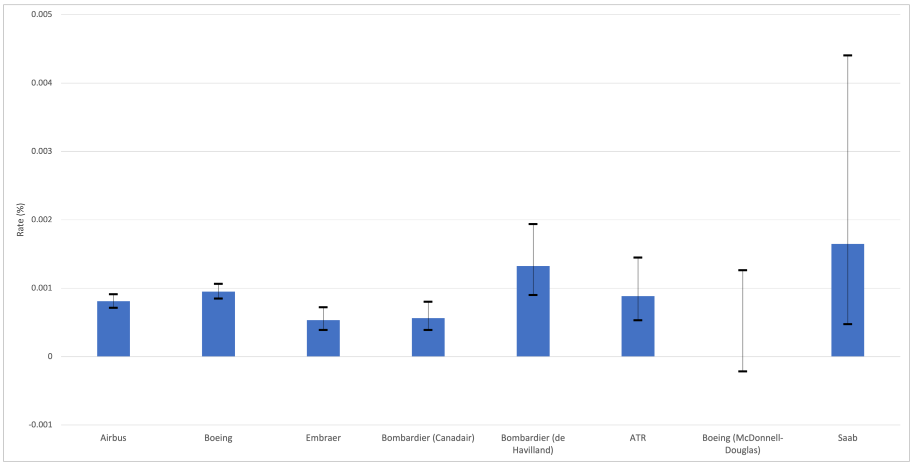


Figure 4.4: Aircraft manufacturers accident rates.

By visual inspection of Figure 4.4, it is interesting to notice how *Airbus* displays slightly better performance when compared to the other industry giant *Boeing*, this is worth noticing, although the difference seems to be minor, because the two manufacturers look extremely similar on paper: both produce a wide and complete range of different aircraft, covering all different needs from domestic to intercontinental flights, (see A.1) and also have almost the same numbers in terms of market share.

Then we observe lower rates for *Embraer* and *Bombardier (Canadair)* and higher rates for *Bombardier (de Havilland)* and *ATR*. These manufactures, unlike *Airbus* and *Boeing*, are more specific about the kinds of aircraft they produce focusing on selected body types (the first two on *narrow bodies* and *regional jets*, while the latter two on *wide bodies* and *turboprops* respectively), their difference in rates is most likely due to this fact and will be shortly explored.

Finally it is hard to draw conclusions on the last two manufacturers, *Boeing (Mc Donnell-Douglas)* and *Saab*, since collectively they represent less than 1% of the worldwide air traffic, and of course this issue is reflected by the amplitude of their confidence intervals, especially for *Saab* which, although displays the highest accident rate among all the manufacturers, it also has an extremely large confidence interval making it hard to evaluate when compared to the others.

4.2.2. Aircraft type

For a broader view on the influence of an aircraft on the accident rates an overview of the aircraft types will be now presented.

In total 33 aircraft types are considered, among the previously listed manufacturers (see Table A.1), ranked in descending order when it comes to their employments (i.e. number of flights).

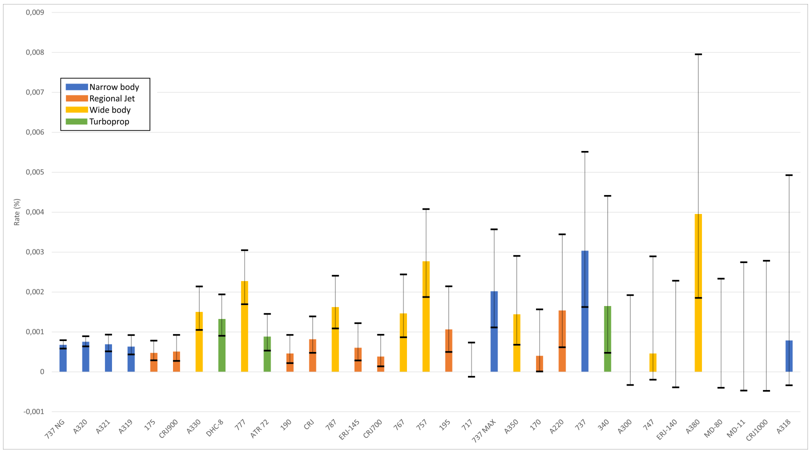


Figure 4.5: Aircraft types accident rates.

The results depicted in Figure 4.5 show a very similar behaviour among the most commonly employed aircraft types (*737 NG*, *A320*, *A321*, *A319*, *175* and *CRJ900*), which all are either *narrow body* or *regional jets* generally used for short distance flights (we will analyze aircraft body types shortly).
Moving further down the list, the estimation of the influence of different types becomes harder, due to a fast increase of the amplitude of the confidence intervals.
One interesting aspect is the high accident rate of the *737 MAX* type when compared to other *narrow body* aircraft.

The need of introducing a new perspective on the aircraft types by classifying them by *body type* becomes apparent. Moreover, we consider 4 different *body types* as follows:

- Narrow body - Mainly employed in short and medium length flights
- Regional jet - Exclusively employed for short distance flights
- Turboprop - Much like regional jets are employed for short distance flights (they have different propulsion system)
- Wide body - Mainly employed for long flights (e.g. intercontinental flights)

For an extensive list of all the aircraft type and their respective body type, as always, refer to Table A.1

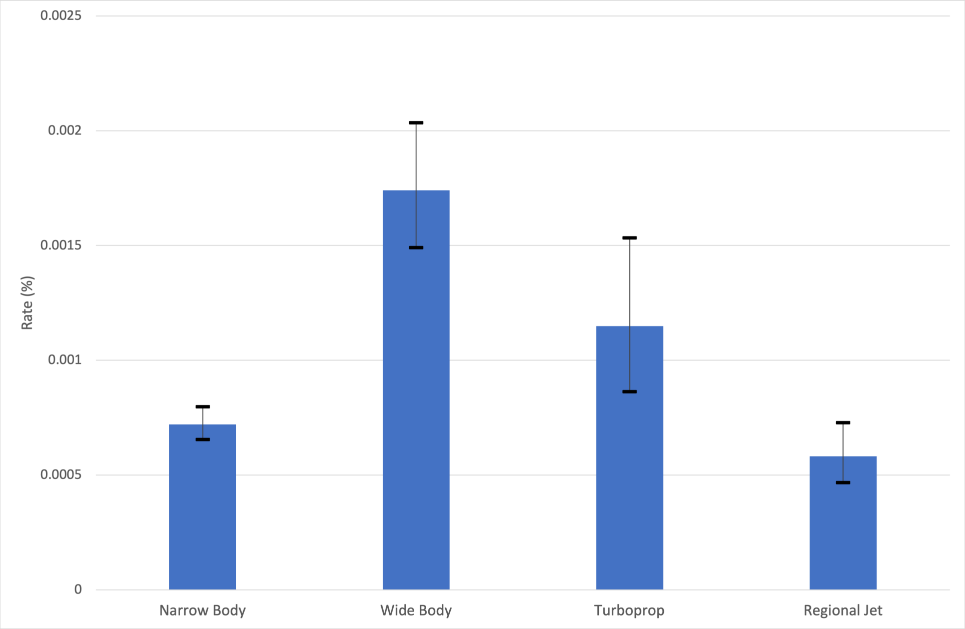


Figure 4.6: Aircraft accident rates classified by body types.

The results shown in Figure 4.6 are interesting and of easier interpretation compared to the aircraft type results.
First of all we observe how *narrow bodies* and *regional jets* have significantly the lowest rates, and are comparable among each other. *Turboprop* display an higher accident rate accompanied by also a larger confidence interval, and finally the most surprising result is the extremely high accident rate displayed by *wide bodies*. This last result is not consistent with the wide spread opinion of industry experts that generally *turboprops* are the worst performing body type, this conviction might be led by the fact that the latter are often used for cargo flights which are inherently more dangerous that commercial flights, but not considered in our analysis.
Moreover, this interesting result leads us to think that the *flight duration* might play a bigger role in the evaluation of accident rates than one might think and will be shortly investigated (see Section 4.3).

4.2.3. Aircraft age

Finally we conclude by shifting our attention towards the role played by the aircraft *age* on the accident rates' trend.
Moreover, the aircraft at our disposal have been grouped up according to their age in 9 different bins of equal size of 3 years apart form the first bin, *1-6* years, and the last bin, *28+* years.

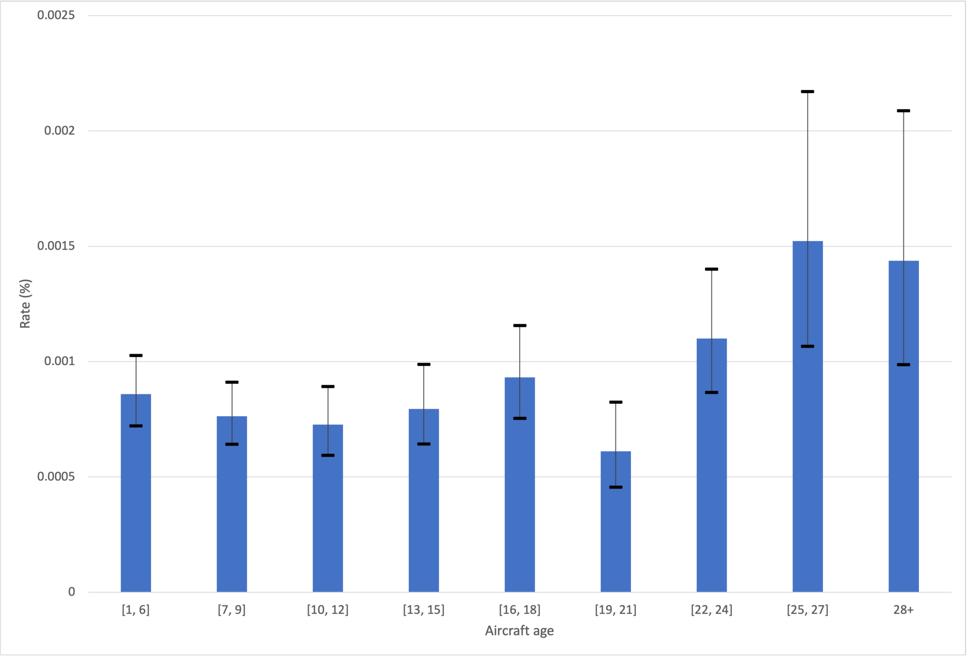


Figure 4.7: Aircraft accident rates categorized by age.

Overall, for the first 18 years of an aircraft lifespan the rates seem to be comparable both when it comes to their value and to the amplitude of the confidence intervals, thus displaying no apparent increasing safety issues in this period, as one might expect.
Furthermore, we have a peculiar drop of the accident rate in the *19-21* bin, this is an unusual behaviour given the fact that in the remaining final 3 bins we observe an increase of the accident rates as expected, although the confidence intervals grow significantly in size due to the lack of still active aircraft in that age range.
Moreover, the *age* of the aircraft doesn't seem to play the crucial role that a naive individual might expect before a closer inspection,.

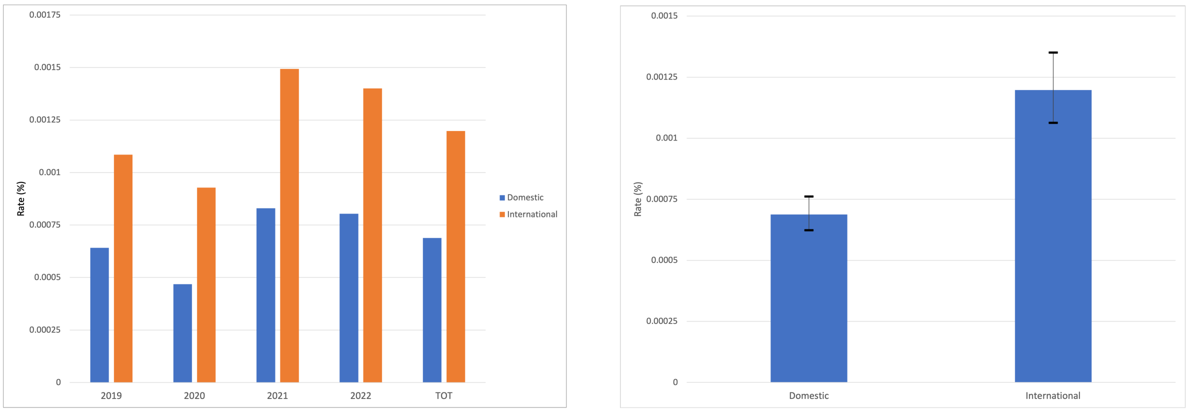
4.3. Flight specifics

In this section we will focus on the flight specific variables and their effect on commercial aviation accident rates. Starting from a broad classification of flights according to their *nature* (i.e. domestic or international), to then look into the *flight duration*, and finally the *flight delay* will be considered.

4.3.1. Nature

Now we compare *domestic* and *international* flights and how they might influence accident rates, but first let us introduce the following assumption:

- A flight is considered *domestic* when its departure and arrival airports are located in the same country, on the other hand when this condition isn't met it is considered *international*



(a) Yearly rates.

(b) Total rates with confidence intervals.

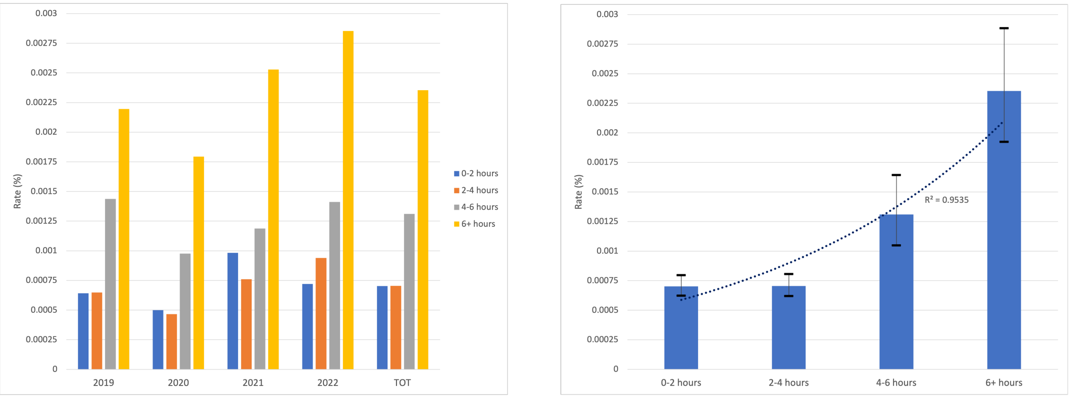
Figure 4.8: Accident rates classified by nature.

The results represented in Figure 4.8a depict a clear disparity between *domestic* and *international* flights with the latter ones clearly showing more accident predisposition.
Moreover, by taking in consideration Figure 4.8b we notice how the previous observation is also true when we include the confidence intervals; in fact the two intervals are relatively small and show no overlapping.

This significant trend might be due to different reasons. First of all it could be linked to the *flight duration* since, for obvious reasons, *international* flights in the majority of the cases display an higher flight time. Furthermore, according to industry experts, this trend could also be connected to the issues that flying to another country might introduce, for example, language barriers, different safety procedures and fatigued pilots due to long hours flights, just to mention a few.

4.3.2. Duration

Moving forward, in the deep dive into flight specific features, now we shine a light on the influence of flight duration on the accident rates.
A classification of the flights based on 2 hours bins will be considered with the exception of the last bin, for which all flights with a flight time over 6 hours are grouped together.



(a) Yearly rates. (b) Total rates with confidence intervals.

Figure 4.9: Accident rates classified by flight duration.

As we can see in Figure 4.9a the trend is consistent among the 4 considered years, in particular there seem to be no noticeable differences between flights with below 4 hours duration, and then a significant increase in the accident rates in the other two classes, with all the available years hitting their respective peaks when considering *6+ hours* flights.
Furthermore, by taking in consideration the confidence intervals shown in Figure 4.9b we

solidify the observations made above. Indeed, the classes *0-2 hours* and *2-4 hours* share not only almost identical rates, but also completely overlapping (and of small amplitude) confidence intervals, confirming the assumption of no statistical difference of the accident rates among the two classes.

Moreover, the other remaining classes, *4-6 hours* and *6+ hours*, show a distinct increase in their rates and, despite having significantly bigger confidence intervals, no overlapping among said intervals, confirming a significant statistical difference between them and with the first two classes.

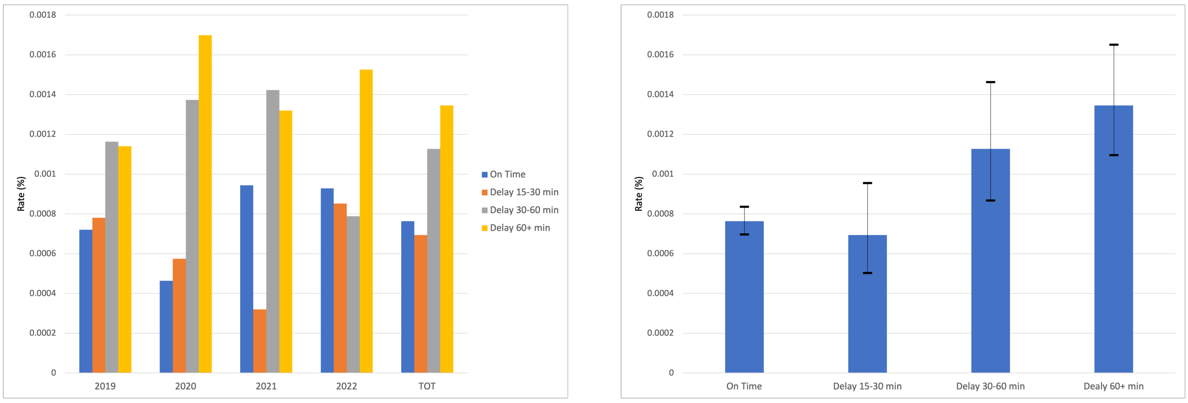
Finally an exponential trend line is represented which fits the class rates extremely well with $R^2 = 0.9535$. Although this result looks impressive it is to be taken into account carefully since only 4 points (the 4 different classes) are used to identify the trend.

The results obtained for the *flight duration* feature are very crucial since they are most likely the main reason of previously observed trend for the features *nature* and the aircraft *body type* (in Section 4.2), in particular when it comes to the extremely high accident rate observed for *wide body* aircraft, which are used almost exclusively for long distance flights.

4.3.3. Flight delay

As the final step, in this flight specific analysis of the commercial aviation accident rates, we shift our attention towards *flight delays.* The overall delay is calculated as the sum of the departure and landing delays.

At first we start by classifying the flights into 4 different classes based on their delay as follows: *On time*, which includes also includes flights that have negative overall delay (i.e. ahead of schedule), *15-30 minutes* delay, *30-60 minutes* delay and *60+ minutes* delay.



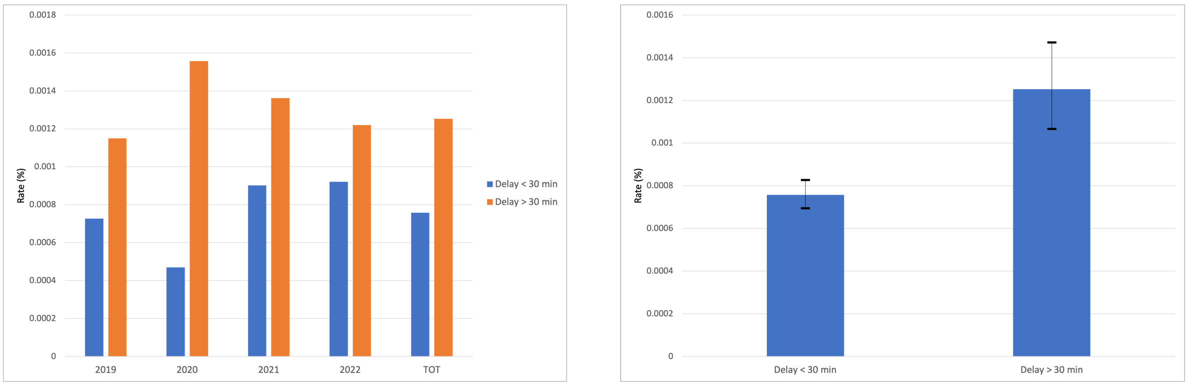
(a) Yearly rates.

(b) Total rates with confidence intervals.

Figure 4.10: Accident rates classified by overall flight delay among 4 classes.

As we can see from Figure 4.10, there doesn't seem to be consistent results among the 4 different delay classes. In particular by looking at Figure 4.10b we notice how the classes *On time* and *15-30 minutes* seem to be very similar due to significant overlapping of the confidence interval, the same reasoning stands even stronger for the classes *30-60 minutes* and *60+ minutes*.
Moreover, this conclusions seem to be true across all the considered years, see Figure 4.10a, with the odd exception of 2021. This considerations lead us to rethink the *flight delay* classification process.

Let us now consider only 2 classes, based on the previous observations we are led to classify *flight delays* as follows: *delay < 30 minutes*, effectively unifying the previous *on time* and *15-30 minutes*, classes and *delay > 30 minutes*, effectively unifying the previous *30-60 minutes* and *60+ minutes*.



(a) Yearly rates.

(b) Total rates with confidence intervals.

Figure 4.11: Accident rates classified by overall flight delay among 2 classes.

As we can see now in Figure 4.11a we have a much more consistent distribution among the different years; only 2020 shows a more pronounced difference than the other years.
Moreover, also in Figure 4.11b we get confirmation that with this new classification we have much better separation among the classes, indeed we have significantly different rates among them and no overlapping between the confidence intervals.

The much higher accident rate when it comes to significantly delayed flights (i.e. *delay > 30 minutes*), when compared to lower delayed flights (i.e. *delay < 30 minutes*), might be due to a collection of different reasons from the stress caused to the pilots arising from time pressure, to airport organizational issues caused by sudden schedule changes.

4.4. Results

Overall, the in depth analysis of the behaviour of commercial accident rates conducted in this chapter has led to some interesting and eye opening conclusions when it comes to understand the main causes of aviation accidents.
Among them we confirmed experts opinions when it comes to considering certain world regions as more dangerous than others, aircraft body type having an influence on accidents and the role played by flight delays.
On the other hand some surprising trends were also spotted such as the extremely significant increasing trend linked to the duration of flights, the not so significant trend linked to different aircraft manufacturers and finally, the most surprising of all, the fact that the age of an aircraft does not seem to play as big of a role as previously expected.

Although this analysis has provided interesting insights when it comes to the issue at hand, there are some limitations. The major one is given by the fact that this approach allows us to look at the influence of the available features one at time, without allowing us to evaluate correlation, interaction and importance among these features and most of all it doesn't allow for methodical feature selection process.
For these reasons, moving on in our analyses, we will consider a machine learning classification approach for flights in order to find the main features that constitute an accident and tackle the shortcomings of the analysis conducted in this chapter.

5 | Gradient boosting machines

There is a plethora of different approaches to tackle the problem of commercial aviation accidents classification, from the usage of the implementations of artificial neural networks and fault tree models ([18]), the more sophisticated use of Bayesian neural networks ([11]) to the implementations of ensemble machine learning and deep learning models ([14]). Although very interesting these previously proposed solutions often rely on extremely specific features for their analyses (i.e. cabin pressure values monitored during the flights) and are rarely fit to handle properly large amounts of data which is of course one of our primary concerns (see Chapter 1).

For these reasons we decided to consider a gradient boosting approach for our classification purposes.

Generally the most common approach for data-driven modeling is to build a single strong predictive model. An alternative is to consider an ensemble of weaker models for a specific learning task. The ensemble approach relies on considering predictions of a large scale of weaker models synthesized in a strong ensemble prediction.

Other ensemble techniques, like random forest, rely on a simple average of the weak models considered in the ensemble, instead boosting methods are based on different procedure: at each iteration a new weak model (base-learner) is trained with respect to the overall error of the ensemble learnt up to that point.

In gradient boosting machines (GBMs) the learning phase sequentially fits new weak models in order to improve the prediction on the target variable, this is done by introducing the new base-learners so that they are maximally correlated with the negative gradient of the loss function of the ensemble. There is an extensive collection of loss functions studied in literature to tackle a variety of different problems from regression, to classification and more. Furthermore there are also different base-learners to choose from when fitting a GBMs with decision trees being the most popular option.

In this chapter a complete overview of GBMs' methodology will be presented with particular attention on how they can be used for classification purposes, for a more in depth and complete theoretical framework of GBMs see [7, 12].

5.1. Methodology

We start by considering the function estimation problem in the common supervised learning setting.
Considering a dataset $(x, y)_{i=1}^{N}$, where $x = (x_1, ..., x_d)$ is the collection of input variables and y refers to the associated response variable, the goal is to extract the functional dependence f between x and y. To achieve this we rely on an estimate $\hat{f}(x)$ such that a specific loss function $\Psi(y, f)$ is minimized:

$$\begin{gathered}\hat{f}(x) = y, \\ \hat{f}(x) = \underset{f(x)}{\arg\min} \Psi(y, f(x))\end{gathered} \tag{5.1}$$

The optimization problem can be written in terms of the conditional expectation as follows:

$$\hat{f}(x) = \underset{f(x)}{\arg\min}\ \mathbb{E}_x[\mathbb{E}_y[\Psi(y, f(x))]|x] \tag{5.2}$$

where $\mathbb{E}_y[\Psi(y, f(x))]$ is the expected y loss, and $\mathbb{E}_x[\mathbb{E}_y[\Psi(y, f(x))]|x]$ is the expectation over the whole dataset.
The decision of the loss function for the problem comes from the distribution of the response variable y, in particular in the case of a binary response, $y \in \{0, 1\}$, we could opt for the log loss function, that is the binary version of the cross-entropy function.

5.1.1. Gradient descent

In this section an overview of the gradient descent method is presented as described by *A. Natekin* and *A. Knoll* (see [12]).
In the case of conventional machine learning models to tackle the function estimating problem we restrict the search to a parametric collection of functions $f(x, \theta)$, effectively changing the optimization problem to the following:

$$\begin{gathered}\hat{f}(x) = f(x, \hat{\theta}), \\ \hat{\theta} = \underset{\theta}{\arg\min}\ \mathbb{E}_x[\mathbb{E}_y[\Psi(y, f(x, \theta))]|x]\end{gathered} \tag{5.3}$$

Since generally the closed form solutions of 5.3 are not available we adopt an iterative numerical approximation procedure. Consider M iterations, and rewrite the parameter estimate in incremental form:

$$\hat{\theta} = \sum_{i=1}^{M} \hat{\theta}_i \tag{5.4}$$

The most utilized parameter estimation technique is gradient descent where, having considered the dataset $(x, y)_{i=1}^{N}$, we want to minimize $J(\theta)$, an empirical loss function, over the data:

$$J(\theta) = \sum_{i=1}^{N} \Psi(y_i, f(x_i, \hat{\theta})) \tag{5.5}$$

The main idea of gradient descent optimization is to continuously improve the estimate along the direction of the gradient of the loss function $\nabla J(\theta)$.
The gradient descent optimization is conducted as follows:

1. Initialize the parameter estimate $\hat{\theta}_0$
   At each iteration t:
2. Compute a parameter estimate $\hat{\theta}^t$ from all the preceding iterations:

$$\hat{\theta}^t = \sum_{i=0}^{t-1} \hat{\theta}_i \tag{5.6}$$

3. Compute $\nabla J(\theta)$ and evaluate it given the ensemble parameter estimates:

$$\nabla J(\theta) = \left[\frac{\partial J(\theta)}{\partial J(\theta_i)} \right]_{\theta = \hat{\theta}^t} \tag{5.7}$$

4. Compute the new incremental parameter estimate, with step-size ρ:

$$\hat{\theta}_t \leftarrow \hat{\theta}_{t-1} - \rho \nabla J(\theta) \tag{5.8}$$

5. update the ensemble estimate by adding the new estimate $\hat{\theta}_t$

5.1.2. Gradient boosting

In this section an overview of the gradient boosting algorithm is presented as described by *A. Natekin* and *A. Knoll* (see [12]).

The main difference between conventional machine learning procedures, such as the one presented above, and boosting methods is that in latter ones the optimization procedure is carried out in the functional space.
This means that we directly parameterize the function estimate in an incremental way:

$$\hat{f}(x) = \hat{f}^M(x) = \sum_{i=0}^{M} \hat{f}_i(x) \tag{5.9}$$

where M is the number of iterations considered, $\hat{f}_0$ is the initial function estimate and $\{\hat{f}_i\}_{i=1}^{M}$ are function increments also known as "boosts".
Now we introduce the parameterized "base-learner" functions $h(x, \theta)$ in order to differentiate them from the overall ensemble estimate $\hat{f}(x)$. There is a plethora of different base-learners to choose from according to ones specific modelling needs, the most common choice are decision trees (we will go through the considered options in Section 5.2.2).
Now we can finally detail the so called "greedy stagewise" approach with the base-learners as function increments (introduced by Friedman, see [7]). In this case the optimal step-size ρ has to be specified at each iteration t, thus defining the optimization procedure at each iteration as follows:

$$\begin{gathered} \hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t), \\ (\rho_t, \theta_t) = \arg\min_{\rho,\theta} \sum_{i=1}^{N} \Psi(y_i, \hat{f}_{t-1}) + \rho h(x_i, \theta), \end{gathered} \tag{5.10}$$

Problems arise in practical implementations of the procedure, indeed having specified the loss function $\Psi(y, f)$ and the base-learner the solution of parameter estimation can be difficult to obtain.
To deal with this issue *Friedman* ([7]) proposed to consider a new function $h(x, \theta_t)$ to be the most parallel to the negative gradient $\{g_t(x_i)\}_{i=1}^{N}$ along the considered data:

$$g_t(x) = \mathbb{E}_y \left[\frac{\partial \Psi(y, f(x))}{\partial f(x)} | x \right]_{f(x)=\hat{f}^{t-1}(x)} \tag{5.11}$$

This allows us to simply select the new function increment such that it is the most correlated with $-g_t(x)$, which means that we can circumnavigate a potentially arduous optimization task with a least-squares minimization problem:

$$(\rho_t, \theta_t) = \arg\min_{\rho,\theta} \sum_{i=1}^{N} \big(- g_t(x_i) + \rho h(x_i, \theta)\big)^2 \tag{5.12}$$

Finally we can summarize the Gradient Boost algorithm proposed by *Friedman* (see [7]):

1. Initialize $\hat{f}_0$ as a constant
   For each iteration t:

2. Compute the negative gradient $g_t(x)$ as shown in 5.11

3. Fit a new base-learner function $h(x, \theta_t)$ such that is the most parallel to the negative gradient.

4. Compute the optimal step-size ρ_t:

$$\rho_t = \arg\min_{\rho} \sum_{i=1}^{N} \Psi\big[y_i, \hat{f}_{t-1}(x_i) + \rho h(x_i, \theta_t)\big] \tag{5.13}$$

5. Update the function estimate as follows:

$$\hat{f}_t \leftarrow \hat{f}_{t-1} + \rho_t h(x, \theta_t) \tag{5.14}$$

Furthermore, for our future implementation of gradient boosting we will use the Python library CatBoost ([19]) which makes use of a specific GBM algorithm that allows for a smooth handling of categorical features without the need of one-hot encoding, that is the conversion of categorical features to numerical ones (for more information on CatBoost refer to [6, 8, 16]).

5.2. Model design

The design choices, for a specific GBM built for particular task, consist in the selection of the functional parameters, that is the loss function $\Psi(y, f)$ and the base-learner $h(x, \theta)$. Namely, one has to first specify the objective of the optimization, i.e. the loss function, and then select the functional form used to achieve the solution, i.e. the base-learner.

In this section we will go through the most popular options available when it comes to the selection of said functional parameters. Furthermore, the main focus will be on classification tasks and on the selection of options that will be implemented later on.

5.2.1. Loss function

Depending on the learning task at hand there is an extensive selection of loss functions $\Psi(y, f)$, with different options available depending on the response variable. When considering a categorical response variable the most popular options are the quantile loss function and the cross-entropy loss function.
Moving forward we will consider the case of a binary categorical response variable, $y \in \{0, 1\}$, and the associated cross-entropy loss function which in this case is also known as the log loss function.

Log loss function

The log loss function is a widely used performance metric in machine learning and optimization tasks when it comes to dealing with binary response variables.
The goal of the log loss function is to evaluate the discrepancy between the true label y and the predicted probability of the positive class (i.e. class 1) $\hat{y}$:

$$\Psi(y, \hat{y})_{log} = -[y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})] \tag{5.15}$$

The value of the log loss ranges from 0 to infinity with lower values representing better model performances.
When evaluated over a dataset of sample size N the log loss is calculated as follows:

$$\Psi(y, \hat{y})_{log} = -\sum_{i=1}^{N} [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)] \tag{5.16}$$

When considering highly imbalanced data, which is of our concern since we are dealing with commercial aviation accidents (see Chapter 2), it is common practice to consider a weighted version of the log loss function in order to encourage the model to focus on predicting correctly samples belonging to the minority class. For a binary classification problem the weighted log loss function is computed as follows:

$$\Psi(y, \hat{y})_{weighted\,log} = -\sum_{i=1}^{N} [\omega_0 y_i \log(\hat{y}_i) + \omega_1 (1 - y_i) \log(1 - \hat{y}_i)] \tag{5.17}$$

where ω_0 is the weight assigned to the negative class (i.e. class 0) and ω_1 is the weight assigned to the positive class (i.e. class 1).

5.2.2. Base-learners

There is a diverse selection of base-learners in the literature to choose from when considering a specific GBM design.

The base-learners of popular use can be classified in three different categories: linear models, smooth models and decision trees.

Although linear models, such as linear regression and ridge regression, and smooth models, such as p-splines, can be effective as base-learners they have some drawbacks when it comes to dealing with large datasets. For these reasons decision trees are the most common choice when it comes to popular GBM algorithms (such as XGBoost, LightGBM and CatBoost) and we will consider them as base-learners moving forward.

Decision trees

Decision tree models are a computationally friendly approach at capturing interactions between variables in GBM models.

The concept behind a decision tree is to partition the input variables space in sub-areas following a tree rule system. Each tree split corresponds to a if-then decision over a specific variable. Due to this structure interactions between variables are naturally modelled. Usually the number of splits is a parameter (called *interaction depth*) set by the user; one might think that the deeper the tree the more accurate the results will be, but it has been proven that, in an ensemble model framework such as GBMs, complex trees ($interaction\ depth > 20$) do not provide significant improvements compared to compact ones ($interaction\ depth \approx 5$).

The most important characteristic of decision trees is that, by construction, a single decision tree will always output a constant value function.

5.2.3. Early stopping

The biggest concern when designing any machine learning model from data is the final model generalization effectiveness. If we overlook proper applications of the learning process overfitting issues are not uncommon. These problems are the same for GBMs.

Indeed, it is common to encounter a situation where base-learners are added until the data is completely overfitted. In statistics this concept is commonly known as the bias-variance trade-off, which refers to the fact that it is difficult to simultaneously minimize the bias, that is the difference between the expected value of a model's prediction and the true value of the target variable, and the variance, that is the amount of variation in the model's predictions.

To tackle these issues different regularization approaches were considered, in particular we will focus on early stopping since it will be the regularization technique used during the later modelling stages.

Early stopping stems from practical considerations, where after a certain number of iterations the loss function evaluated on the test set starts increasing, while the same on the training set keeps decreasing effectively confirming that overfitting is taking place. To solve this issue we reduce the ensemble to the number of trees corresponding to the test set minima on the loss curve. To achieve this result an iteration threshold k is set, and the learning procedure is stopped if, after k consecutive iterations, there are no improvements on the minimization of the test set curve.

5.3. Model interpretation

When it comes to practical applications it is of paramount importance to be able to interpret the model results. When using additive GBM models, that is with linear models as base-learners, the results can be trivially explained since the additive components correspond to the marginal dependence plots by design.

This approach is of no use when one uses a GBM with high interaction depth decision trees as base-learners, and despite the simple structure of a decision tree, when it comes to an ensemble of thousands of them the results interpretation becomes a challenging task.

A collection of different tools has been developed to tackle interpretation issues in decision tree based GBMs. In this section we will describe the later on implemented tools for GBM interpretation.

5.3.1. SHAP values

Shapley values are a model agnostic interpretation tool that comes from cooperative game theory, recently their employment as machine learning explanation tools as become prevalent due to their versatility (see [10]).

This method requires the training of the model on every feature subset $S \subseteq F$, where F is the complete set of features. In order to do this we effectively train two models, $f_{S\cup\{i\}}$ with the considered feature present and f_S with the feature withheld. Successively the predictions from the two models are compared on a specific input $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$, where x_S represents the feature values in S. The latter difference is then computed for all subsets $S \subseteq F \setminus \{i\}$, and the Shapley values are a weighted average of said differences:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \tag{5.18}$$

Once the Shapley values are computed for all the features they can effectively be used as feature importance measure.

Moving forward for our model interpretation purposes we will consider SHAP (SHapley Additve exPlanation) values (see [10, 17]).
SHAP values are the Shapley values of a conditional expectation function of the original model under analysis. SHAP values assign to each feature the modification in the prediction of the expected model when conditioning with respect to that feature. They allow us to explain how to get from the base expected value $\mathbb{E}[f(x)]$ to the model output $f(x)$. For a more in depth theoretical framework of the inner workings and properties of SHAP values refer to Scott M. Lundberg and Su-In Lee's paper: *A Unified Approach to Interpreting Model Predictions* ([10]).

6 | Flights safety classification model

In this chapter an implementation of a safety GBM tree-based classification model for the OAG flights will be presented (see Section 1.1.1), with the aim of identifying patterns in the available descriptive features to gain a deeper understanding of the characteristics of commercial aviation accidents. Furthermore, also METAR data (see Section 1.1.2) at departure and arrival will be taken into account for each flight.
For the modelling and analyses conducted in this chapter only the flights relative to 2019 will be considered, this is due to two main reasons: the first, and more obvious, reason is to avoid extremely high computational burdens, the second reason is because, as previously discussed in Chapter 2, 2019 data is not affected by the COVID-19 pandemic, therefore it allows to consider more data compared to the previous years and also will reflect more accurately the trends of future years.

Among the available flight features (presented in Chapter 1, or introduced in Chapter 4), the following will be considered in the modelling:

- Operating region - OPERATING_REGION
- Airline operator - OPERATOR
- Aircraft body type - BODY_TYPE
- Aircraft type - TYPE
- Aircraft age - AGE
- Flight nature - NATURE (see Section 4.3.1)
- Flight delay class - DELAY_CLASS (see Section 4.3.3)
- Flight duration - DURATION

Additionally we will also considered the following METAR features (introduced in Section 1.1.2), for each flight, both at departure and arrival (with prefix *DEP* and *ARR* respectively):

- Temperature $[°F]$ - DEP_TMPF and ARR_TMPF
- Relative humidity - DEP_RELH and ARR_RELH
- Wind speed $[kn]$ - DEP_SKNT and ARR_SKNT
- Visibility $[mi]$ - DEP_VSBY and ARR_VSBY

Finally the target variable in this modelling chapter is a binary variable, $y \in \{0, 1\}$, where the positive class represents the occurrence of an aviation accident for each flight. Before tackling the model implementation we should address the data processing measures taken into account to deal with the high class imbalance at hand.

6.1. Data processing

As previously discussed in Chapter 2, the main issue to deal with is the extreme class imbalance among our data. There are several ways to tackle this problem from to data-processing methods, such as oversampling of the minority class (i.e. accidents) or undersampling of the majority class (i.e. non-accidental flights); to considering alternative loss functions, such as the weighted log loss function (see 5.17).

It's pretty clear, from initial attempts at fitting a classification model, that due to this extreme imbalance we cannot rely on only one of this methods to obtain somewhat decent results. Indeed, if we were to perform a total undersampling of the non-accidental flights, we would reduce the dataset to an extremely small fraction of itself, loosing any possible identifiable pattern in the flights. On the other hand if we were to perform a total oversampling of the accidents, the dataset size would augment to an extent that would cause serious computational issues. Finally also considering a weighted loss function on the original data would cause serious misclassification issues assigning an extremely high weight to the accident class.
For the above reasons moving forward we decided to adopt a hybrid method among the proposed solutions. Indeed, both undersampling for the majority class and oversampling for the minority class will be performed to a controlled extent, in order to maintain the characteristic imbalanced nature of the phenomenon and tame the above listed problems. Furthermore, in the model training stage the weighted log loss function will be considered. Moreover, only flight samples with available departure and arrival METAR data are taken into consideration.

6.1.1. Data oversampling and undersampling

The main issue when it comes to sampling techniques on the considered flight data lays in the fact that we are dealing with both numerical and categorical features, therefore significantly restricting the options at our disposal. First of all we perform a train-test split of our dataset so that 80% of the available flights will be used for the training phase, while the remaining 20% for the testing phase. It is important to highlight that all of the following sampling strategies described will be implemented only on the training data, effectively maintaining the structure of the test data untouched.

To reduce the size of the safe flights a random undersampling procedure has been adopted. That is random samples are selected from the majority class, obtaining a new dataset for the latter corresponding to 50% of its original size.

When it comes to generating synthetic data for the aviation accidents a finer technique as been implemented: SMOTE-NC (Synthetic Minority Over-sampling Technique for Nominal and Continuous features, see [5]). SMOTE-NC extends the concept of SMOTE, a commonly used synthetic data generation technique, to deal with both categorical and numerical variables. Continuous features are numeric and can be interpolated to generate new values. When it comes to continuous features SMOTE-NC interpolates the continuous features of the target instance and its chosen neighbor (or neighbours) to create synthetic samples. The interpolation process considers the feature values of the target instance and the neighbors, and it generates new values within the range of the chosen feature, by multiplying the difference between the target instance and the neighbors by a random number between 0 and 1 and adding it to the target instance's feature value. This generates a new instance that lies on the line segment connecting the target instance and the nearest neighbor in the feature space.

On the other hand, for each categorical feature SMOTE-NC selects a target instance and its k-nearest neighbors, based on a suitable distance metric for categorical features. For each categorical feature, it selects a value from either the seed instance or the nearest neighbor with a certain probability. The probabilities are often determined by the ratio of the number of this feature's values in the nearest neighbor and the seed instance.

This procedure allows us to create synthetic accident data samples (for a more in depth theoretical overview of SMOTE-NC see [5]). We apply SMOTE-NC to the training data effectively increasing the count of accident data by 10 times its original size.

In the following Table 6.1 we present the structure of the new training data obtained after the METAR data empty values removal (performed on the whole dataset) and the above discussed sampling techniques (performed only on the training data).

Data	No. Safe flights	No. Accidents	Accident Rate (%)
Original	12,158,646	110	0.000905
Resampled	6,080,368	1100	0.018091

Table 6.1: Training data resampling to mitigate class imbalance

As we can see from the results in the table, we are able to significantly improve the representation of accidents in the training data while still maintaining the intrinsic imbalanced nature of the phenomenon.
Now that we have completed the data processing stage we can finally move to the GBM aviation safety classification modelling part.

6.2. GBM classification model

Now that all the pieces are in the correct place we can finally proceed with the modelling stage. We consider a gradient boosting binary classifier for the available flights to discern patterns between safe flights and aviation accidents, taking into account all the available features mentioned at the beginning of this chapter. To do so a CatBoost binary classifier model is considered, that is a gradient boosting ensemble model with decision trees as base-learners, with the following optimized parameters:

Parameter	Value
Loss function	Log loss
Class weights	Balanced
Learning rate	0.01
Max tree depth	6
No. Iterations	4000
Early Stopping iterations	150

Table 6.2: Modelling parameters of the full GBM classifier model

As we can see, from the selected parameters shown in Table 6.2, we are opting for low learning rate with a high number of total iterations ($M = 4000$) in order to be able to capture the most nuances possible from the model. Moreover, we opt for a log loss function with balanced class weights, that is of the form 5.17, to account for the class imbalance. The balanced class weight calculation for a class $i \in \{0, 1\}$ has the following formula:

$$\omega_i = \frac{1}{\frac{n_i}{N}} \tag{6.1}$$

where n_i represents the sample size of class i, while N represents the full sample size. Therefore, by substitution of the results in Table 6.1, we obtain $\omega_0 \simeq 1$ and $\omega_1 \simeq 5528$. This high discrepancy among the weights will favor the correct classification of accidental flights instead of safe ones.

Furthermore, to avoid any possible overfitting problems we apply an early stopping procedure to the model (see 5.2.3), with threshold on the iterations of $k = 150$.

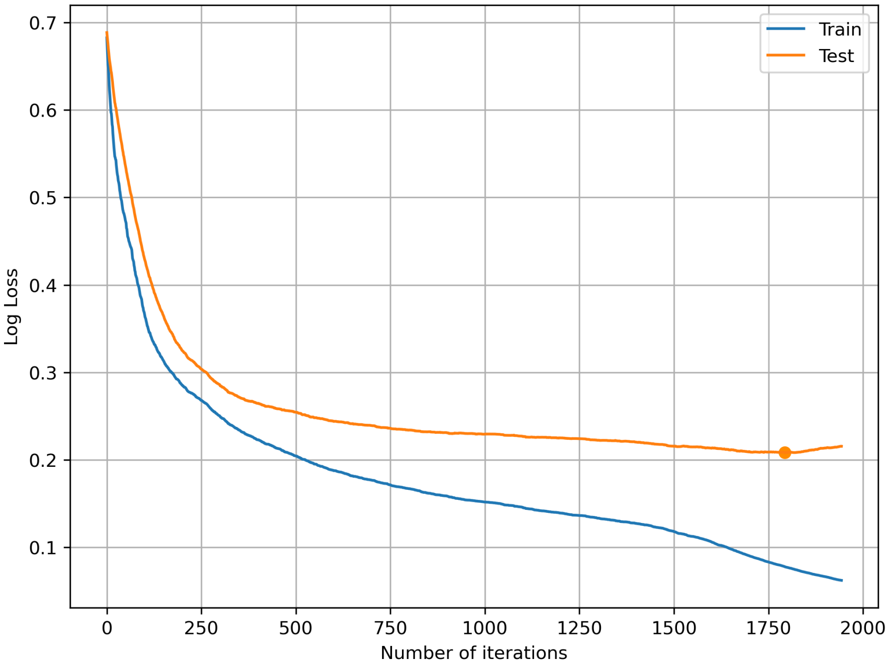


Figure 6.1: Full GBM model log loss evaluation for both training and testing.

By visual inspection of Figure 6.1 we notice how the full 4000 iterations are not needed. Indeed, we reach a minimum for the testing log loss, with value 0.2083 at iteration $M_{opt} = 1793$ (highlighted on the testing loss curve), after which the model starts overfitting the training data; indeed there is a noticeable increase in the values of the testing log loss, and the training process is truncated by early stopping before the 2000^{th} iteration.

Overall, both the learning and testing curves display regular decreasing trends without any spike or major oscillatory behaviour, furthermore reaching good low levels of the respective log losses: 0.0778 for the learning and 0.2083 for the testing.

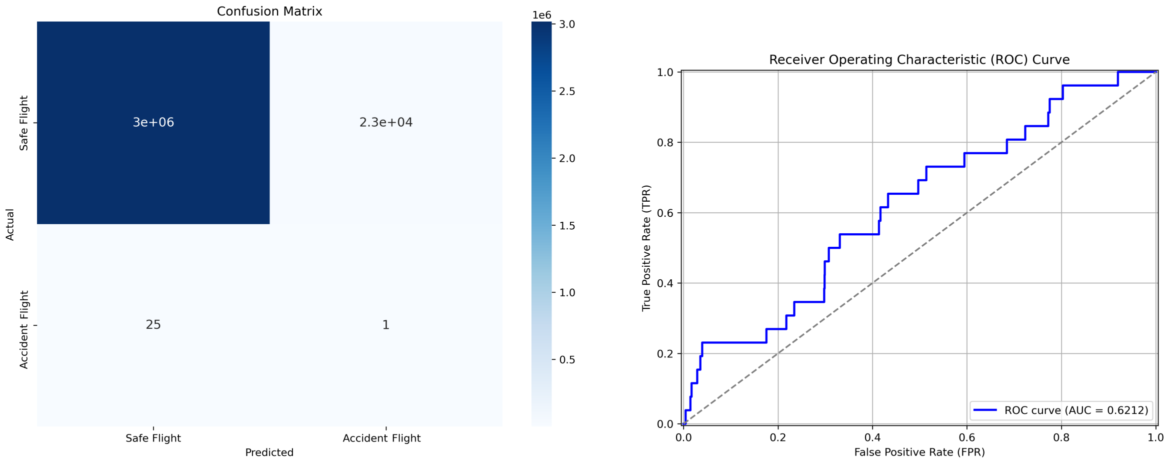
6.2.1. Performance evaluation

Now we are interested in analyzing the model's performance. To do so we will rely on two specific metrics: the accuracy and the recall. Moreover, our attention will be mainly towards trying to keep a decent level of accuracy with a recall above 0.5, this approach has been adopted since, for our purposes, we are more interested in correctly classifying the minority class samples (i.e. the accidents). Furthermore, also visualization tools such as the confusion matrix and the receiver operating characteristic (ROC) curve, and its area under curve (AUC) as an additional metric, are considered to visualize the performance of the classifier.

Initially we consider the model with default value for the classification threshold $\epsilon = 0.5$ and we have the following results on the testing set:

Metric	Value
Accuracy	0.992395
Recall	0.038461
AUC	0.621156

Table 6.3: Full GBM classifier evaluation metrics with classification threshold $\epsilon = 0.5$.



(a) Confusion matrix.

(b) ROC curve.

Figure 6.2: Full GBM classifier visualization tools with classification threshold $\epsilon = 0.5$.

As we can see from Table 6.3 we have an extremely high model accuracy but we sacrifice a lot when it comes to the recall. This is also reflected by the results displayed in the confusion matrix in Figure 6.2a. Indeed, we correctly classify only one accident among the 26 available in the testing set.
From the above results its clear that the default threshold $\epsilon = 0.5$ is way too conservative when it comes to our main concern which is to identify patterns in accidents.

For this reason we opt to select an optimal classification threshold, in order to improve the recall, by maximizing the difference between the true positive rate and the false positive rate on the ROC curve (see Figure 6.2b). By doing so the following optimal threshold for the model is obtained: $\epsilon_{opt} = 0.040783$.
Adopting this new threshold for classification we observe the following results, evaluated on the testing set.

Metric	Value
Accuracy	0.566950
Recall	0.653846
AUC	0.621156

Table 6.4: Full GBM classifier evaluation metrics with classification threshold ϵ_{opt}.

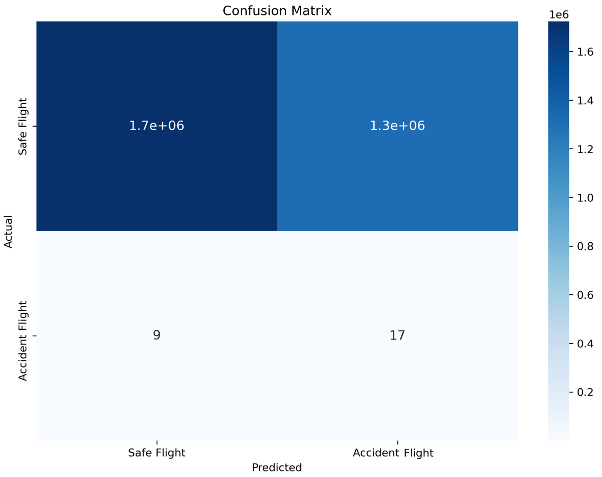


Figure 6.3: Full GBM classifier confusion matrix with classification threshold ϵ_{opt}.

As we can see from the metrics displayed in Table 6.4, we are now able to achieve a much higher recall while still maintaining an above average accuracy. This also clear from the confusion matrix displayed in Figure 6.3, indeed even if we lose classification accuracy when it comes to the safe flights we gain a lot the accident classification accuracy; this trade-off is to be expected when dealing with such rare events.

Now that we have maximized the classification potential of the GBM classifier, when considering all the available features, the only logical step forward is to perform a feature selection process to improve the model classification performance.

6.3. Features selection

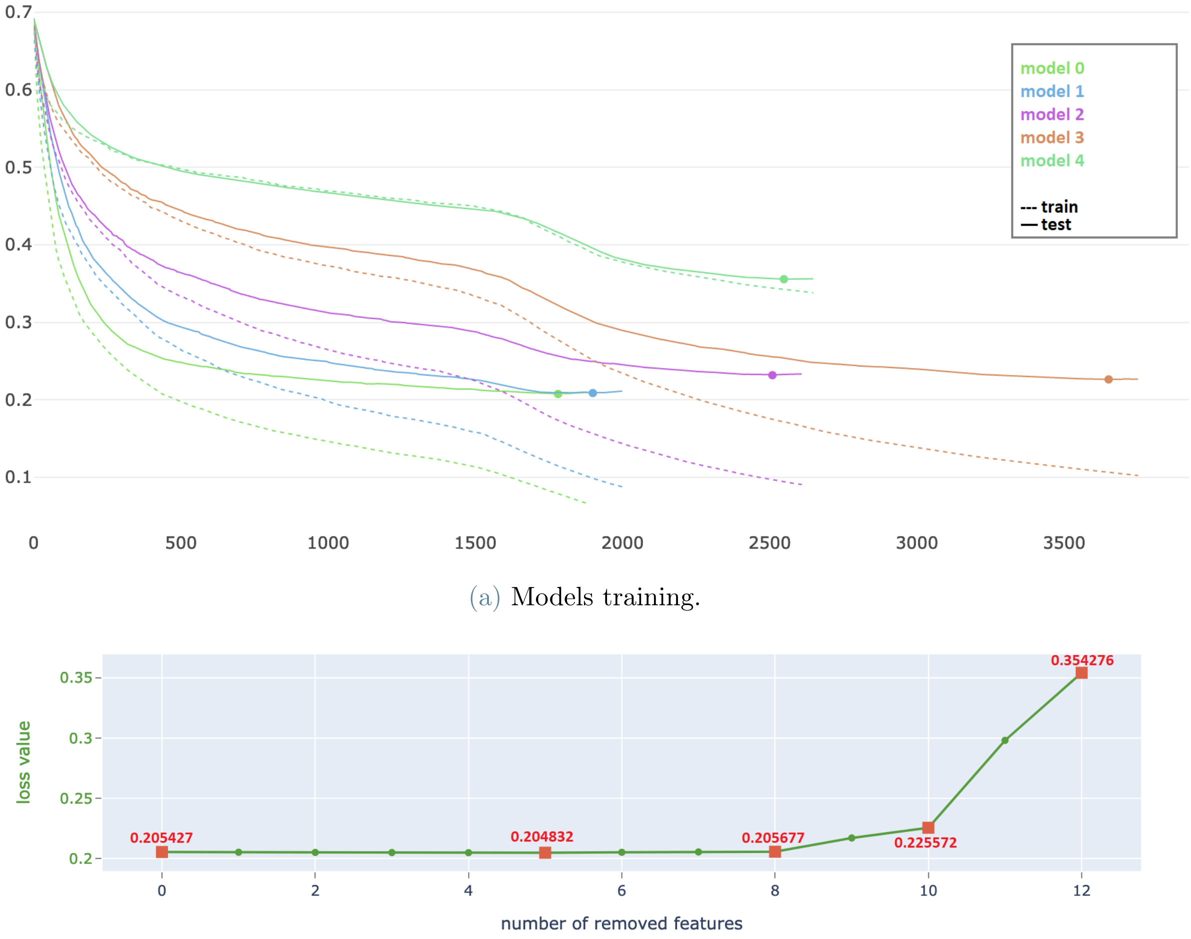
In order to retrieve the best subset of features to achieve higher model performance the built-in feature selection tool of CatBoost is considered, evaluating the change in the loss function (in our case the weighted log loss function, see 5.17) caused by each feature's removal.

CatBoost uses a technique called permutation importance to select features for the model. Permutation importance works by randomly permuting the values of a feature and observing the change in the model's accuracy. If we have a significant decrease in the model's accuracy when the values are permuted than that feature is to be considered important. For a detailed explanation of the inner workings on CatBoost and permutation importance see [6, 8, 16].

In our case after the evaluation of the features' importance we proceed by removing them one-at-the-time the least important features, and evaluating the reduced models accuracy at each step through the evaluation of the log loss function on the testing set.

For the feature selection process the same parameters as the ones set for the training of the full GBM model have been set for consistency reasons (see Table 6.2). Furthermore, for computational reasons the training of the reduced models is performed at 5 steps, that are the steps at which individual reduced model is evaluated, selected considering the features' permutation importance. The final set of features evaluated to conclude the feature selection process will consist of 4 features, from the starting 16 considered in the full model. In particular we will evaluate the model accuracy at 16 features (Model 0, that is the full model presented in Section 6.2), 11 features (Model 1), 8 features (Model 2), 6 features (Model 3) and 4 features (Model 4).

Furthermore, for computational speed issues, a one-hot encoding parameter for the categorical features has been enabled in the training of the models which might yield a slight reduction in terms of loss function performance and overfitting problems, due to CatBoost characteristic of being optimized for raw categorical features, reason for which once the reduced model is selected it will be retrained again on its own.



(a) Models training.

(b) Minimum loss function value at each step.

Figure 6.4: Feature selection with 4 step model training and evaluation.

As we can see from visual inspection of Figure 6.4a we achieve similar results for model 0 and model 1, while for the rest of the trained models the performance significantly decreases as we keep removing features.

Moreover, in Figure 6.4b we plot the minimum values achieved at each step of the training, and we can see that we have slight improvement in the loss value when considering model 1. Indeed, this model achieves a minimum loss value of 0.204832 removing 5 features in the following order:

1. Operating region (OPERATING_REGION)
2. Flight delay class (DELAY_CLASS)
3. Flight duration (DURATION)

4. Arrival relative humidity (ARR_RELH)

5. Departure visibility (DEP_VSBY)

This result is rather interesting since, from our previous analyses of the commercial aviation accident rates (see Chapter 4), features such as *Operating region* and *Fight duration* seemed to have a strong influence on the accident rates.

6.4. GBM reduced classification model

In this section we proceed with training and evaluation of the reduced GBM classifier identified in the previous section, that is removing the identified 5 features (Operating region, Flight delay class, Flight duration, Arrival RELH and Departure VSBY) from the original pool of features presented at the beginning of the chapter, thus considering 11 features (Airline operator, Aircraft body type, Aircraft type, Aircraft age, Flight nature, Arrival TMPF, Departure TMPF, Departure RELH, Arrival SKNT, Departure SKNT and Arrival VSBY). Furthermore, we will compare the performance of the reduced model with the full model to identify the best one for our purposes.

For the training of the GBM reduced classifier we consider a CatBoost classifier with the following optimized parameters:

Parameter	**Value**
Loss function	Log loss
Class weights	Balanced
Learning rate	0.01
Max tree depth	6
No. Iterations	5000
Early Stopping iterations	150

Table 6.5: Modelling parameters of the reduced GBM classifier model

As we can see from the selected parameters in Table 6.5, once again we opt for a low learning rate with high number of maximum iteration ($M = 5000$), increased with respect to the full GBM model since we expect more iterations might be needed in order to reach convergence. And just as before we opt for a log loss function with balanced weights (see 5.17 and 6.1).

Finally, once again, to avoid overfitting issues we apply early stopping with $k = 150$ iterations threshold.

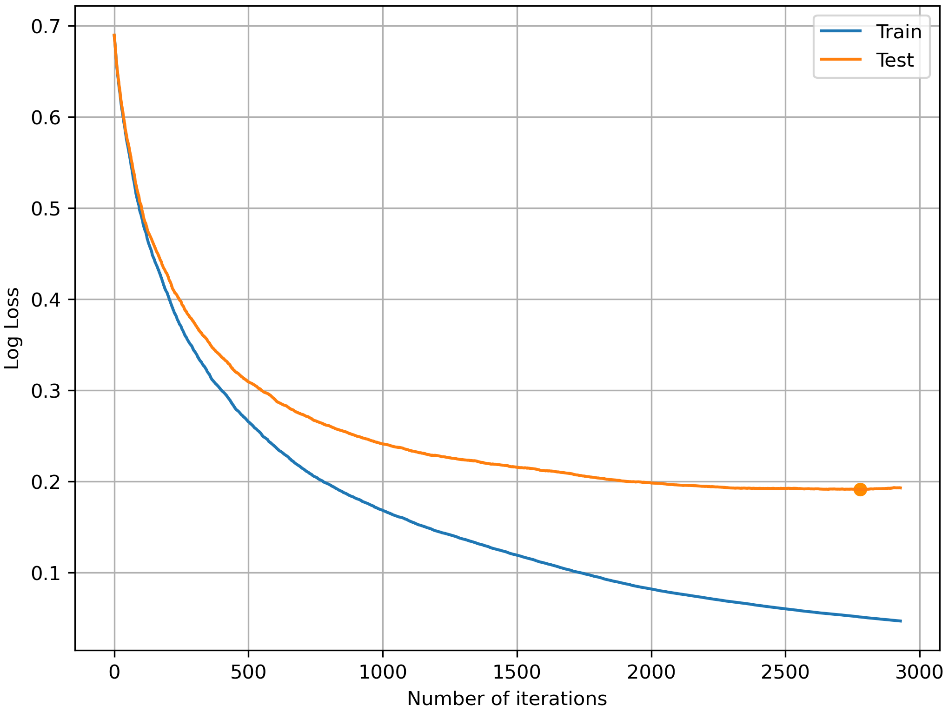


Figure 6.5: Reduced GBM model log loss evaluation for both training and testing.

By visual inspection of Figure 6.5 we notice how the full 5000 iterations are not needed. Indeed, we reach a minimum for the testing log loss, with value 0.1913 at iteration $M_{opt} = 2778$ (highlighted on the testing loss curve), after which the model starts overfitting the training data; indeed there is a noticeable increase in the values of the testing log loss, and the training process is truncated by early stopping before the 3000^{th} iteration.

Overall, both the learning and testing curves display regular decreasing trends without any spike or major oscillatory behaviour, furthermore reaching good low levels of the respective log losses: 0.0515 for the learning and 0.1913 for the testing.

We can already see from the loss function evaluation that we achieve better results when compared to the full model (see Figure 6.1).

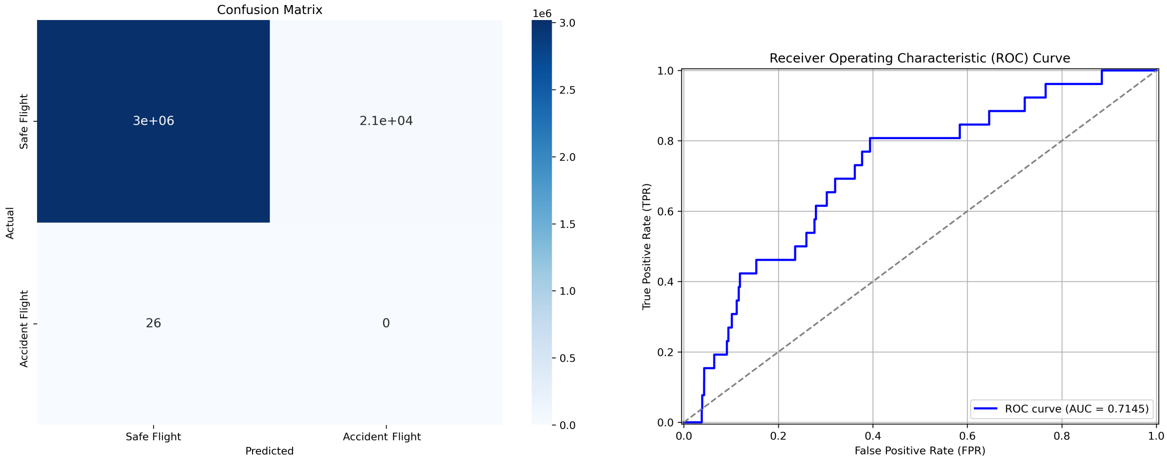
6.4.1. Performance evaluation

Now we proceed with a performance evaluation of the reduced model. Once again we rely on evaluation metrics such as accuracy and recall with an analogous approach on trying to keep decent levels of accuracy while maximizing the recall. Furthermore, to evaluate the classifier's performance also the area under curve (AUC) of the receiver operating characteristic (ROC) curve is considered.

Once again initially the model is evaluated with default value for the classification threshold $\epsilon = 0.5$ yielding the following results on the testing set:

Metric	**Value**
Accuracy	0.992944
Recall	0.0
AUC	0.714485

Table 6.6: Reduced GBM classifier evaluation metrics with classification threshold $\epsilon = 0.5$.



(a) Confusion matrix.

(b) ROC curve.

Figure 6.6: Reduced GBM classifier visualization tools with classification threshold $\epsilon = 0.5$.

As we can see, from the results displayed in Table 6.6, we achieve an extremely high level of accuracy but we are not able to correctly capture any accident, indeed this is portrayed both by the null recall and in the confusion matrix in Figure 6.6a.
Nonetheless we observe an high level of the $AUC = 0.714485$ of the ROC curve in Figure 6.6b. This leads us to further optimize the classification threshold, since the default one of $\epsilon = 0.5$ is clearly too conservative for our purposes.

Once again, as previously done with the full model, we opt to select an optimal classification threshold that maximizes the difference between the true positive rate and the

false positive rate on the ROC curve (see Figure 6.6b). This procedure leads to obtain the following optimal classification threshold: $\epsilon_{opt} = 0.025418$.
Adopting this new improved threshold we observe the following results on the reduced model, again evaluated on the testing set.

Metric	Value
Accuracy	0.605911
Recall	0.807692
AUC	0.714485

Table 6.7: Reduced GBM classifier evaluation metrics with classification threshold ϵ_{opt}.

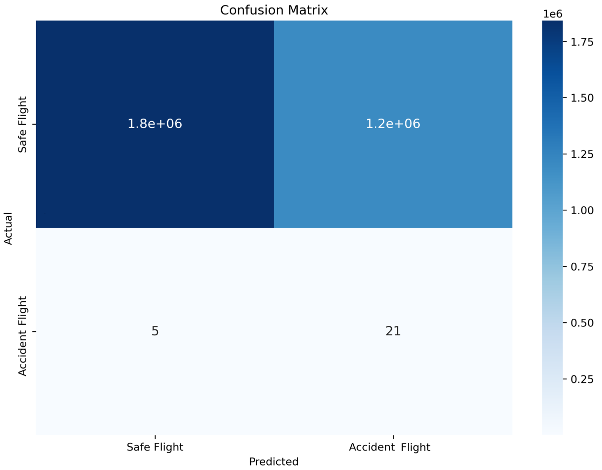


Figure 6.7: Reduced GBM classifier confusion matrix with classification threshold ϵ_{opt}.

As we can see from the metrics displayed in Table 6.7, we are able to achieve a very satisfactory value of recall, above 80%, while maintaining an overall decent level of accuracy. This is also displayed in the confusion matrix in Figure 6.7, indeed even if we lose classification accuracy when it comes to the safe flights we gain a lot when it comes to accidents, correctly classifying 21 instances over the 26 available in the testing set.
The only concern when it comes to the reduced model is the fact that we observe null true positive rate for very small threshold values on the ROC curve (see Figure 6.6b). Indeed, this indicates that the reduced model might struggle to correctly classify instances with very small probability of belonging to the positive class.

Before moving forward in our analysis we will present a comparison between the two presented models considering their respective optimal classification thresholds, highlighting pros and cons of each one considering our accident analysis purposes.

6.4.2. Model selection

We can now compare the evaluation metric of the full and reduced models (presented in the Sections 6.2 and 6.4 respectively), with respect to their relative optimal thresholds (0.040783 and 0.025418 respectively), to better understand which one to consider to move forward in our analysis.

Metric	Full Model	Reduced Model
Min. Test loss value	0.2083	0.1913
Accuracy	0.5669	0.6059
Recall	0.6538	0.8077
AUC	0.6212	0.7145

Table 6.8: Comparison of the evaluation metrics of the full and reduced models with optimal thresholds.

As we can see form the results displayed in Table 6.8, the reduced model outperforms the full model in every metric. In particular it is interesting to see how the reduced model has significantly improved recall and AUC compared to the full model, and how it is able to achieve better results with only a marginal decrease in the loss value and furthermore by maintaining an higher level of accuracy. This means that the reduced model, not only is better at identifying accidents by a significant margin, but also at classifying safe flights at the same time (this can also be noticed by visual comparison of the full model's confusion matrix, see Figure 6.2a, with the reduced model's counterpart, see Figure 6.6a).

If we compare the ROC curves of the two proposed models (see Figure 6.2b and Figure 6.6b), we notice how the reduced model's ROC curve has null true positive rate for very small values of the classification threshold, unlike the full model. This, in general, could be a concern since it means that the reduced model might have problems at correctly classifying instances with low probability of belonging to the positive class. However, this concern is trumped by the overall increase in performance under every metric of the reduced model when compared to the full model. Indeed, given the critical nature of flight safety a high recall is extremely valuable, and the relatively high accuracy suggests that the model is performing well overall. Moreover, the choice of the optimal threshold, $\epsilon_{opt} = 0.025418$, seems to be offering good trade-off between correctly identifying accidents and maintaining a reasonable level of accuracy, which aligns with the priorities and requirements of flight safety.

For the extensive list of reasons mentioned above, moving forward for our flight classification and analysis purposes, the reduced model presented and evaluated will be considered over the full model initially developed.

6.5. Results interpretation

In this section the focus will be shifted on the interpretation of the results obtained with the reduced model introduced in Section 6.4.

Moreover, this interpretation will be conducted through the use of the SHAP Python library (see [17]), which allows for useful graphical tools to represent the Shapley values (see Section 5.3.1) for each considered feature. Furthermore, a global overview of the Shapley values for the model will be initially presented, to try and explain the global effect of the features on the model's output; later on a local interpretation by comparison of similar flights and their Shapley values will be provided, to further understand what might be the causes that separate a safe flight from an accident.

6.5.1. Global interpretation

Let us now focus on the global effect of different features on the reduced model output, moreover on their associated Shapley values computed on the training set so that we can try to understand how the model learns to classify flights.

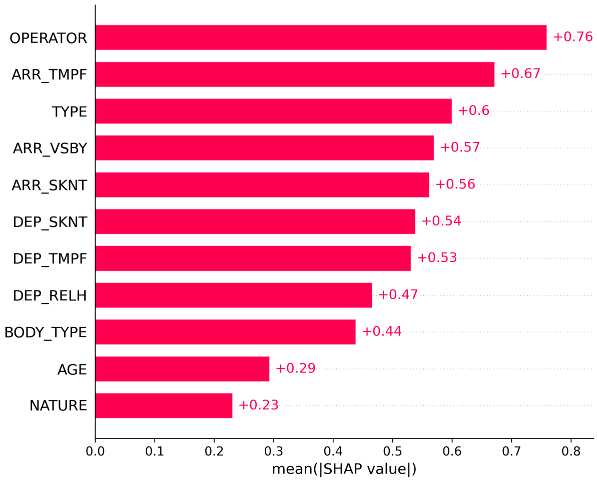


Figure 6.8: Bar plot of the mean Shapley values associated to each feature

Figure 6.8 shows interesting results. In particular it is interesting to notice how the features *Operator* and *Type* seem to be the most influential features, among the general flight ones, when it comes to classifying flights' safety; while the *aircraft age* and the flight's *nature* play a marginal role.
Furthermore, by looking at the METAR features (see Section 1.1.2) we notice a striking trend. That is the fact that the *arrival* METAR features are all consistently more crucial than the *departure* ones. This could be explained by the fact that it is common knowledge that most of the accidents happen upon landing (for a visual representation see Figure 2.2).

Now we shift the focus on the influence of numerical variables, in particular the METAR features.

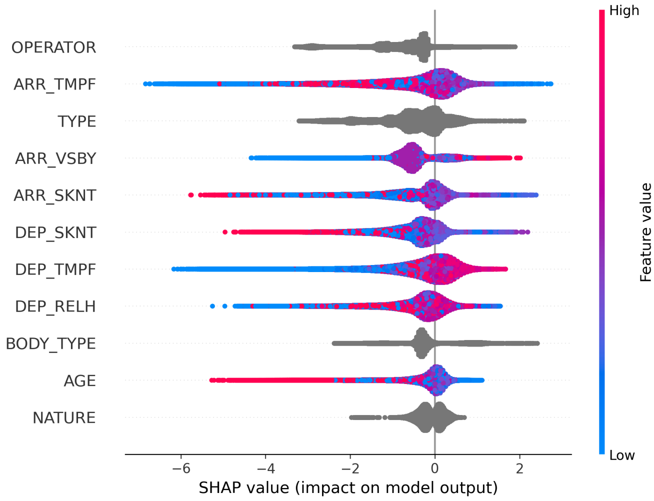


Figure 6.9: Beeswarm plot of the Shapley values of each flight feature

The beeswarm plot represented in Figure 6.9 is to be interpreted in the following way: for each feature all the flights are represented, color coded depending on the value of the feature. Furthermore, the positive Shapley values are associated with a positive prediction (i.e. an accident), while the negative Shapley values are associated with a negative prediction (i.e a safe flight).
While no particular pattern is spotted for the arrival *temperature*, with both low values of the feature being present in a positive and negative prediction, we notice a clearly increasing trend for its departure counterpart. Indeed, the departure *temperature* shows a compelling increasing trend, with high values for the feature being associated with a positive classification.
Furthermore, it is surprising to see how both the arrival and departure *wind speed* seems to be to be negatively correlated with positive predictions, indeed for both these features

high values of the features are mostly towards predicting a flight as safe. This could be due to erroneous measurements or data reporting.

An extremely surprising result is given by the arrival *visibility*, where we notice a clear separation among the feature values with low towards a negative prediction and high values towards a positive one. One might think that this is due to different units of measurements but this is not actually the case, indeed all visibility data has been converted in miles [mi]; this is more likely due to different reporting procedures when it comes to visibility across different regions; in particular with *North America*, which is by far the region with more commercial aviation flights, being associated with extremely high *visibility* values (see Figure A.1 in Appendix A for a more detailed explanation of this phenomenon).

Finally another surprising result is represented by the aircraft *age*, indeed we have low values shifting the predictions towards a positive outcome, this is likely due to high values of the age being very rare and thus their representation among the accidents might be overshadowed due to the data processing procedures explained in Section 6.1, particularly the removal of flights with missing METAR data and the oversampling.

Now we would like to focus on the airline *operators*, moreover on producing a ranking system for them based on the mean of Shapley values associated to the feature of all the available flights considered in the training. This procedure will allow us to compare operators' performances among each other.

In Table 6.9 a ranking of the airline *operators* with positive mean of the Shapley values is shown:

Airline operator	**Mean shap value**
Pakistan International Airlines	0.841733
Onur Air	0.482366
GoAir	0.475167
Alliance Air	0.365707
TUI UK	0.332119
Hawaiian Airlines	0.211060
BoA	0.188303
ANA-All Nippon Airways	0.007476

Table 6.9: Ranking of the most dangerous airline operators based on the mean of their Shapley values.

The results displayed above are very much aligned with the previously conducted airlines operators rates analysis in Chapter 4. In particular *Pakistan International Airlines* is among the worst performing airlines across the 4 years, counting 8 accidents with only 73619 recorded flights. The same can be said also for the other airlines in the list which all display an above average accident rate. Moreover, it is interesting to see how also *ANA-All Nippon Airways* is part of this list confirming the results analyzed previously in Figure 4.2, where it was identified as one of the worst performing airlines operating in *Asia Pacific*.
To complete this analysis of the airline *operators* it is interesting to also take a look at the best performing ones based on the mean of their Shapley values.

Airline operator	**Mean shap value**
Ryanair	-2.858941
EasyJet	-2.521200
China Airlines	-2.383004
Korean Air	-2.177598
Pegasus	-1.897567
Etihad Airways Airlines	-1.682134
Vistara	-1.425271
Air China	-1.332490

Table 6.10: Ranking of the safest airline operators based on the mean of their Shapley values.

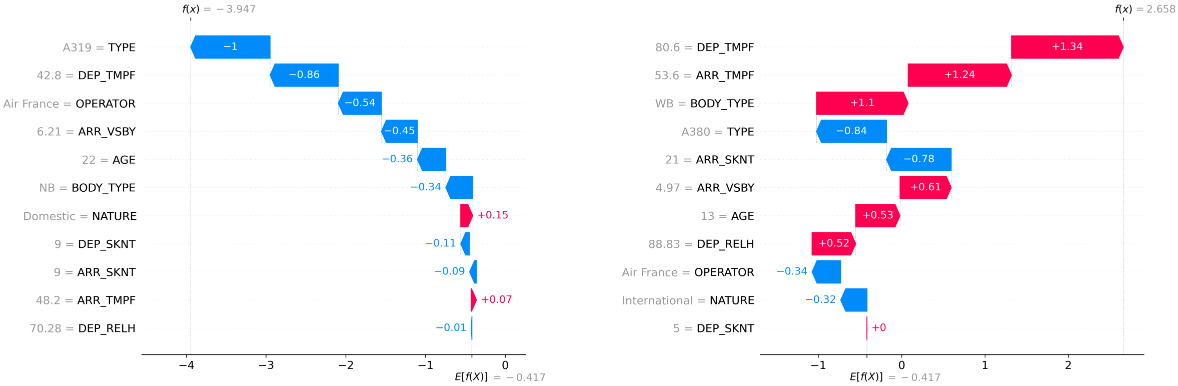
All of the results displayed in Table 6.10 are in line with commercial aviation accident rates analysis, with all the above listed operators having a below average accident rate. In particular it is interesting to see how airlines such as *Ryanair*, *EasyJet* ans *Air China* are among the safest in the top 50 operators represented in Figure 4.2. It is important to remark that the accident rate analysis was conducted over 4 years of data, while the GBM classification models were trained and tested only over 2019 data, reason for which we might find some discrepancies among the results.

6.5.2. Local interpretation

One of the main advantages of working with Shapley values as an interpretation tool, is that they allow us to understand the features' influence among each individual sample.

This approach is extremely useful for comparing similar flights, for example flights belonging to the same airline or carried out through the same aircraft type, yielding different classification outcomes (i.e. safe flight or accident), granting us an in depth look at the micro differences among two (or more) samples and how they affect the classification process.

Now we can compare two flights with different outcome belonging to the same operator. In this case *Air France* was selected, due to its variety in terms of flights since it handles both domestic and international flights with a variety of aircraft types when it comes to its fleet. Overall, in both cases, we expect a negative influence of the common airline operator towards a negative classification, since Air France displays an above average performance in accident rates terms.



(a) Safe flight.

(b) Accident flight.

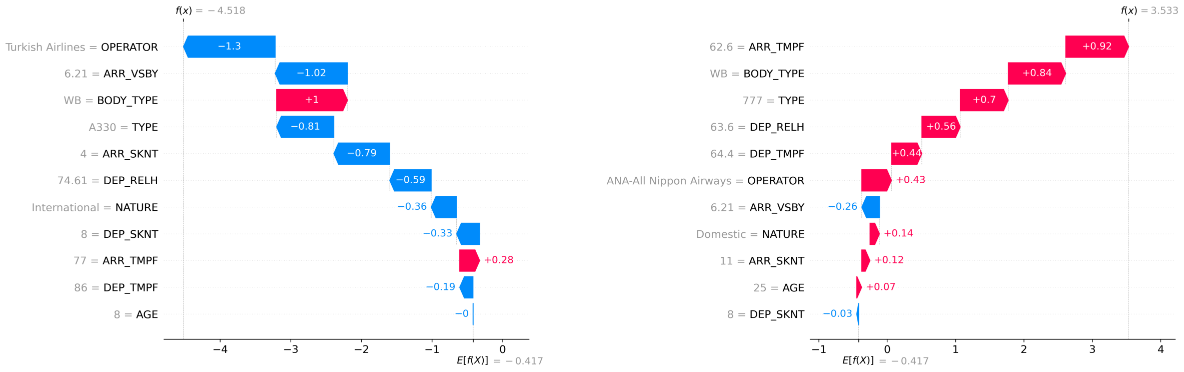
Figure 6.10: Shapley values waterfall plots comparison of a safe flight with an accidental flight from Air France.

From visual inspection of Figure 6.10 we notice that the main difference among the two flights is represented by the type of aircraft employed. Indeed, the safe flight is carried out through an *A319* narrow body aircraft while the accidental flight employs an *A380* wide body. It is interesting to see the influence of the *body type* on the outcome, that is that the narrow body has a negative influence on the classification of the safe flight shifting it towards a negative outcome in striking contrast with the wide body category in the accidental flight, which is among the most influential features in shifting its classification towards a positive outcome. This result is in line with the outcome of the accident rate analysis previously conducted and displayed in Figure 4.6, which shows wide bodies aircraft as the riskiest body type.

Moreover, going deeper in the aircraft types differences, we notice in both cases a negative contribution for both the *A319* and *A380* types. While the first result is very

much in line with our previous analysis, represented in Figure 4.5, that shows the *A319* as one of the best performing aircraft types on the market, the same cannot be said of the *A380*. Indeed, from the previous analysis the latter is one of the worst performing aircraft types. This mismatch among the model's result and the commercial aviation accident rate analysis can be explained by the oversampling techniques adopted (see Section 6.1) in combination with the lack of data for this specific aircraft type. Indeed, we count only 239 active aircraft when it come to the *A380* type (compared to the 1243 of the *A319*) making one of the least employed aircraft types among our list, thus resulting in an unlikely generation of synthesized accidents for this aircraft type (especially when considering only one year of data).
Finally it is interesting to see the influence of METAR data on the two different outcomes. First of all we notice how, for the accidental flight, an higher departure temperature has the most influence on the positive classification, while a lower value has the opposite impact on the safe flight. Moreover, we notice how also the arrival visibility plays a crucial role in the classification outcome. Indeed, in the safer flight we have a visibility value of $6.21mi$ (i.e. $10km$), which is the European standard for optimal conditions, shifting the prediction towards a safer outcome; while a lower value for the accidental flight shifts the prediction in the opposite direction.

Now as a final comparison among flights we are interested in comparing two flights carried out by wide body aircraft, to identify safer types among this category and also investigate the influence of different airlines.



(a) Safe flight.

(b) Accident flight.

Figure 6.11: Shapley values waterfall plots comparison of a safe flight with an accidental flight of wide body aircraft.

The results depicted in Figure 6.11 allow us to extract interesting conclusions. First of all, we can see how the influence of different airline operators plays a crucial role in the classification outcome. Indeed, we notice how for the safe flight the operator *Turkish Airlines* represent the major factor towards a positive classification, while for the accidental one *ANA-All Nippon Airways* has the opposite effect. This result is very much in line with the analysis conducted in Section 4.1.2, which shows *ANA-All Nippon Airways* as one of the riskiest operators form the *Asia Pacific* region; furthermore *Turkish Airlines* displays a good performance in terms of accident rates.

Moreover it is compelling the difference among the two aircraft types, even when they belong to the same body type category, representing an *A330* as a safer option compared with the *777*. This result is also reflected by the previous analysis (see Figure 4.5), although we would not expect such a striking difference. Furthermore, it is interesting to see how the aircraft age has almost no influence on the classification, indeed in this case we have a significantly older aircraft in the accidental flight but with an almost null (but still positive) associated Shapley value, while the aircraft age of the safe flight has no influence on the classification.

When it comes to the available METAR data the results are of arduous interpretation. In both cases we notice a negative contribution to the prediction given by the same standard value of the arrival visibility in good weather conditions (as for the safe flights of the previous case in Figure 6.10a). When it comes to the other METAR features the interpretation is not as straightforward, indeed among the most influential factors towards a positive outcome in the accidental flight we have the arrival temperature and the departure relative humidity, which don't seem to display a particular pattern towards classification outcomes (see Figure 6.9).

Overall, the results interpretation provided in the final part of this chapter yielded compelling results and new tools to tackle the commercial aviation safety problem, while shining a light on different shortcomings, especially when it comes to the interpretation of METAR features and under represented categories among the categorical features.

In general we are satisfied with the interpretation of the results provided by the classification model, considering this is one of the main issues when dealing with machine learning models compared to a classical additive model alternative approach.

7 | Conclusions

This study provided a comprehensive analysis of the commercial aviation safety phenomenon, concluding significant results.
Starting from a custom developed matching procedure to uniquely identify accidents to flight data, to then move on to more intricate analyses. This included a more classical analysis approach of commercial aviation accident rates, allowing us to identify the most interesting trends among a plethora of different features, through the employment of the *Agresti-Coull* binomial proportion interval to best evaluate the issue at hand.

Moving on, a binary classification model for flight safety was implemented through the use of gradient boosting machines, thus introducing deeper insights to the problem thanks to a modern machine learning approach. This allowed us to introduce weather aviation data (i.e. METAR) into the picture, to gain an insight into the safety of specific flights and, chiefly, to develop a forecasting tool specifically tailored for commercial aviation flights. The data processing, model training, parameter tuning and feature selection were all interesting and challenging phases of this study, that granted us with a well performing and finely tuned prediction model.
Furthermore, the interpretation of the model's findings through the exploitation of the Shapley values, associated to the model's learning phase, yielded compelling results. Among those the high dependence of the commercial aviation accidents phenomenon from the flights' airline operators allowed us to extract a safety exposure ranking system, for the operators, based on their associated Shapley values. Moreover, the conducted comparison of the Shapley values associated to specific flights, provides an extremely detailed analysis tool for flight evaluation; allowing us to compare similar flights, based on a specific set of features, to investigate the reasoning behind their classification and gain insights into the model learning process.

Overall, between the results obtained through the commercial aviation accident rates analysis and the development of the GBM flight safety classification model, not only we were able to depict a complete picture about commercial aviation safety, but also to

provide different evaluation and prediction tools that could be key for future aviation underwriting purposes.
Thus we were able to achieve the goals set out at the beginning of the study; notwithstanding, the intrinsic extremely rare nature of commercial aviation accidents, which are among the rarest events, thus making their analysis extremely arduous.

7.1. Limitations of the study

During the unfolding of this study we encountered a series of different limitations towards our analysis goals.
First and foremost, the quality of the available data. Indeed not having a unique source of data for both flights and recorded accidents posed a big limit to our further analyses, in spite of the the efficient flight-accident developed algorithm. Indeed, we recall how around 30% of the OAG flights had missing registration number, making them completely unfit for accident matching to begin with; on top of that having to rely on ASN, an open source online database, for accidents data could cause some issues due to reporting biases form different countries.

Furthermore, the intrinsic rare nature of commercial aviation accidents posed a limitation in itself, forcing us to resort to a restricted family of models for our forecasting purposes, and requiring beforehand heavy data processing procedures which might restrict the model's generalization capabilities, given the rarity of the phenomenon. This issue led us to consider a machine learning classification approach, through the implementation of a GBM binary classifier, which, although yielded satisfactory results, made the interpretation of the results a complicated task. Indeed, the necessary generation of synthetic data for the accidents through SMOTE-NC, can cause biased results when it comes to evaluating rare instances for the considered categorical features, especially since we considered only one year of data.

7.2. Further developments

When it comes to possible improvements and further developments to this study the possibilities are virtually endless.
Moreover, the extension of the GBM flight safety classifier to a larger pool of data, for example the available time period of 4 years, could produce higher results and significantly improve the model's generalization capabilities towards forecasting future flights' safety. The same can be said of the considered features, indeed the pool of considered

features could be increased including features like the departure and arrival airports, specific countries instead of the broader geographical regions considered and so on.

Moreover, a plethora of model optimization can be done moving forward. From the development of a custom loss function that accounts better for extreme level of class imbalance at hand. To the development of a multiclass classification model, based on the level of damages sustained by the aircraft involved in the accidents, or even based on flight's phase in which the accident occurred. In particular the latter one could allow for a smarter approach towards the available weather (i.e. METAR) data, since a separation between departure and arrival weather is provided. Furthermore, a different feature selection approach, such as one based on Shapley values, could be considered. This approach is generally more accurate than the loss gain approach implemented in this study, but it was avoided due to its high computational burden.

Overall, each possible further development would pose its own set of challenges, especially when it comes to the multiclass classification alternatives previously proposed, since it would make synthetic accidents data generation an even more delicate and crucial process.

Bibliography

[1] A. Agresti and B. A. Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998. ISSN 00031305. URL `http://www.jstor.org/stable/2685469`.

[2] AviationSafetyNetwork. ASN aviation safety wikibase, 2023. URL `https://aviation-safety.net/wikibase/`.

[3] AviationWeatherCenter. METereological Areodrome Reports (METARs) - AWC, 2023. URL `https://www.aviationweather.gov/metar`.

[4] L. D. Brown, T. T. Cai, and A. DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–117, 2001. ISSN 08834237. URL `http://www.jstor.org/stable/2676784`.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[6] A. V. Dorogush, V. Ershov, and A. Gulin. Catboost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*, 2018.

[7] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[8] J. T. Hancock and T. M. Khoshgoftaar. Catboost for big data: an interdisciplinary review. *Journal of big data*, 7(1):1–45, 2020.

[9] E. Kostya. GeoPy GitHub repository, 2022. URL `https://github.com/geopy/geopy`.

[10] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[11] J. Luxhoj and D. Coit. Modeling low probability/high consequence events: an avi-

ation safety risk model. In *RAMS '06. Annual Reliability and Maintainability Symposium, 2006.*, pages 215–221, 2006. doi: 10.1109/RAMS.2006.1677377.

[12] A. Natekin and A. Knoll. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, 2013. ISSN 1662-5218. doi: 10.3389/fnbot.2013.00021. URL `https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021`.

[13] OAG. OAG flights database, 2023. URL `https://oag.com`.

[14] A. Omar Alkhamisi and R. Mehmood. An ensemble machine and deep learning model for risk prediction in aviation systems. In *2020 6th Conference on Data Science and Machine Learning Applications (CDMA)*, pages 54–59, 2020. doi: 10.1109/CDMA47397.2020.00015.

[15] C. V. Oster, J. S. Strong, and C. K. Zorn. Analyzing aviation safety: Problems, challenges, opportunities. *Research in Transportation Economics*, 43(1):148–164, 2013. ISSN 0739-8859. doi: https://doi.org/10.1016/j.retrec.2012.12.001. URL `https://www.sciencedirect.com/science/article/pii/S0739885912002053`. The Economics of Transportation Safety.

[16] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin. Catboost: unbiased boosting with categorical features. *Advances in neural information processing systems*, 31, 2018.

[17] S. V. Sundararajan, S. M. Lundberg, and Y. Lee. Shap: A unified approach to explaining black boxes, 2017. URL `https://github.com/slundberg/shap`.

[18] X. Yan, H. Wang, Q. Fu, and J. Zhao. Civil aviation safety risk assessment for rare events. In *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, pages 898–906, 2019. doi: 10.1109/ISKE47853.2019.9170344.

[19] Yandex. CatBoost - open-source gradient boosting library, 2022. URL `https://catboost.ai`.

A | Additional content

Here we present the full list of aircraft *types* considered in our analyses, organized by their respective manufacturers with additional information regarding the body type:

Manufacturer	Aircraft type	Body type
Airbus	A220	Regional Jet
Airbus	A300	Wide Body
Airbus	A318	Narrow Body
Airbus	A319	Narrow Body
Airbus	A320	Narrow Body
Airbus	A321	Narrow Body
Airbus	A330	Wide Body
Airbus	A350	Wide Body
Airbus	A380	Wide Body
ATR	ATR 72	Turboprop
Boeing	717	Narrow Body
Boeing	737	Narrow Body
Boeing	737 MAX	Narrow Body
Boeing	737 NG	Narrow Body
Boeing	747	Wide Body
Boeing	757	Wide Body
Boeing	767	Wide Body
Boeing	777	Wide Body
Boeing	787	Wide Body
Boeing (McDonnel-Douglass)	MD-11	Wide Body
Boeing (McDonnel-Douglass)	MD-80	Wide Body
Bombardier (Canadair)	CRJ	Regional Jet
Bombardier (Canadair)	CRJ700	Regional Jet
Bombardier (Canadair)	CRJ900	Regional Jet

Bombardier (Canadair)	CRJ1000	Regional Jet
Bombardier (de Havilland)	DHC-8	Turboprop
Embraer	170	Regional
Embraer	175	Regional Jet
Embraer	190	Regional Jet
Embraer	ERJ-140	Narrow Body
Embraer	ERJ-145	Regional Jet
Saab	340	Turboprop

Table A.1: List of aircraft type considered for the analyses

Here we present a comparison between the arrival *visibility* distributions of the *Europe* and *North America* operating regions. Highlighting the fact that it is very likely that the two regions share different reporting standards when it comes to this METAR feature.

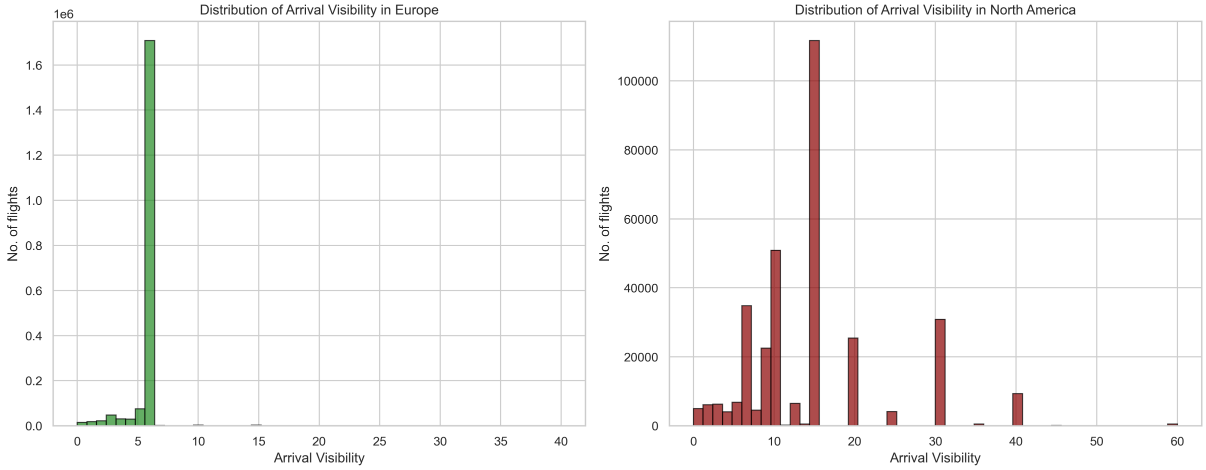


Figure A.1: Comparison of the distributions of the arrival visibility among *Europe* and *North America*

Indeed, by visual inspection of Figure A.1, we notice how *Europe* seems to have a cut-off point at 6.21mi (which corresponds to 10km) to represent clear visibility. On the other end *North America* doesn't display this pattern, instead has a more homogeneous distribution with no apparent cut-off point.

In this instance only *Europe* and *North America* have been compared, this is because *North America* is the only region displaying this peculiar behaviour while all the other regions seem to follow the standard employed by *Europe*, with the same cut-off point when it comes to optimal visibility conditions.

List of Figures

List of Tables

Acknowledgements