AIRBUS

CORPORATE PROJECT

SARAH AWAD
ALEJANDRO DANUS
FEDERICO HUERTAS
IGNACIO DE PARADA
DIEGO SALAZAR

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In the aviation sector, where operational quality meets important safety standards, the role of maintenance is crucial. Within this context, the aerospace industry's commitment to leveraging advanced technology and data analytics highlights the proactive approach to ensuring flight safety and operational efficiency. Predictive maintenance emerges as a crucial strategy in this environment, offering insight into potential disruptions, enhancing business resilience, and mitigating potentially disastrous events.

Addressing the Core Challenge

Our main initiative to address this issue is the development of an innovative and automated model for fuel leak detection in aircraft. The conventional approach to this critical maintenance aspect is predominantly manual and lacks precision, posing risks of unexpected service interruptions and extended aircraft on ground (AOG) situations. These scenarios can seriously impact fleet scheduling and the operational strength of an airline.

Safety is the top priority in the aerospace industry because it directly impacts people's lives. Every rule and process are carefully crafted to keep passengers and crew safe during every flight. This focus on safety is about more than just following laws; it's a value that influences all areas of aerospace work. Ensuring safety means not only preventing accidents and saving lives but also earning the public's trust. This is why identifying and addressing fuel leakages, with the assistance of innovative machine learning models, is crucial.

Strategic Solutions and Business Implications

Our response to this challenge is the introduction of a machine learning model to quickly identify any potential leaks that may affect the safety of the aircraft, and craft a dashboard to showcase the model's findings.

The primary objective is to harness innovative capabilities to identify leaks swiftly within limited flight intervals, effectively increasing safety standards and reducing AOG risks. The anticipated outcome is a notable enhancement in operational reliability and cost efficiency, with potential savings estimated at up to $10,000 per hour by averting AOG incidents.

Exploratory Data Analysis and Feature Engineering

In order to create our final models, we performed exploratory data analysis to determine which features we deemed important. Through our analysis, we were able to replace NULL values, drop duplicates, and group the data by flight. This gave us a holistic view of each flight for all MSNs.

Next, we decided to continue our preprocessing and feature engineering with only MSN02, because this is the dataset that contains all the information required to proceed. Comparatively, the other MSN flights contained up to 90% NULL values, which would not have been beneficial for our analysis.

Finally, we simulated leaks in our clean MSN02 dataset, randomly assigning fuel leakage flows to the flights. We created leakages in increments of 0.25, 0.75, and 0.9, to test whether our model would be able to distinguish between different flows. We created a binary column to classify those anomalies as leaks, and then concatenated that data together into one dataset for further analysis. This column will be used to compare the recall of our model.

Model Approach

We trained and tested 3 different models, XGBoost Classifier, Logistic Regresison, and Autoencoder. All of them work in very different ways, providing us with different metrics, insights, and results. Finaly, we compared these models to each other, focusing on recall as our metric of success.

DATA IN THE INDUSTRY

Aircraft Industry Overview

The aviation industry is heavily regulated to ensure safety and efficiency, with fuel management being a critical component. Airlines and maintenance providers invest significantly in technology and processes to monitor aircraft fuel systems and identify leaks, which can pose serious safety risks and result in substantial financial losses. The industry is continuously advancing in terms of adopting more sophisticated technologies for fuel monitoring and leakage detection to minimize risks and costs.

Role of Data in the Industry

At the core of aircraft maintenance and safety is data. Data is the driver of modern aircraft maintenance, supporting decision-making processes, enhancing efficiency, and ensuring compliance, all of which contribute to the safer and more economical operation of aircraft.

1. **Enhanced Safety and Reliability:** Data collection and analysis in the aviation sector are vital for identifying trends and potential issues before they escalate into serious problems. By analyzing historical and real-time data, maintenance teams can predict potential failures and address them proactively, significantly enhancing aircraft safety and reliability.

2. **Operational Efficiency:** Data-driven insights allow for more efficient scheduling of maintenance, ensuring that aircraft spend more time in the air and less time on the ground. By optimizing maintenance routines based on data analysis, airlines can improve aircraft utilization and reduce unnecessary maintenance activities.

3. **Cost Reduction:** Analyzing data helps in identifying and rectifying inefficiencies in the maintenance process, which can lead to substantial cost savings. Predictive maintenance, enabled by data analytics, can prevent costly repairs and replacements, reducing the overall operational costs.

Sensor Data

The utilization of data from sensors for fuel leakage detection in the aircraft industry marks a significant advancement in ensuring flight safety and operational efficiency. Sensors strategically placed throughout the aircraft's fuel system collect vital data in real time, including parameters like fuel levels, pressure, temperature, and flow rates. This continuous monitoring enables the immediate identification of any irregularities that could signify a leak, facilitating swift remedial actions. Additionally, integrating predictive analytics into maintenance practices allows for the analysis of accumulated sensor data to identify patterns or anomalies indicative of potential fuel system issues before they manifest as actual leaks. Such predictive capabilities not only enhance the aircraft's safety profile but also optimize maintenance schedules, reducing unnecessary inspections and repairs.

Airbus Skywise

Airbus Skywise, developed with Palantir Technologies, revolutionizes aviation data handling since its 2017 launch. It consolidates aviation data like flight schedules and sensor information on a cloud platform, enhancing data analytics, predictive maintenance, and decision-making:

- **Predictive Maintenance**: Utilizes sensor data and analytics to foresee component failures, shifting maintenance from reactive to preventative, reducing costs, and elevating passenger experience.

- **Health Monitoring**: Offers real-time support for in-flight and maintenance events, enabling swift identification and resolution, minimizing downtime and AOG risks.

- **Fleet Reliability**: Improves analysis and reliability of airline fleets by integrating diverse data, allowing for benchmarking and strategic insights, thereby enhancing operational efficiency and compliance.
  These features emphasize Airbus's dedication to digital advancement, offering airlines powerful tools to harness data for operational excellence.

PROBLEM STATEMENT

Challenge: Aircraft Fuel Leakage

In aviation, fuel leakage represents a significant maintenance concern, often involving extensive time to identify and solve, leading to long periods of Aircraft on Ground (AOG) situations. Fuel leakage in aircraft happens due to breaches in the fuel tanks or the fuel delivery system, causing unexpected fuel loss. These leaks generally stem from structural flaws, the wearing away of seals and adhesives, or could be due to excessive heat and repeated harsh landings.

Identifying a Fuel Leak

The process of identifying a fuel leak is often difficult, relying heavily on manual methods such as visual checks, the use of tracer gases, or pressure assessments carried out after landing during routine maintenance. Currently, fuel leaks are primarily detected either through manual inspection during scheduled or unscheduled maintenance, or by pilots who calculate fuel levels using sensor data. Aircraft sensors typically flag fuel leaks only after they have significantly worsened over time, with a recognized discrepancy threshold of 2,500kg in 20 seconds. Pinpointing the exact leak source can also be painstaking since these leaks usually start off small and may originate from multiple locations.

Routine fuel assessments are carried out before, during, and after flights to spot any inconsistencies in the Fuel on Board, Fuel Used/Fuel Uplifted, and the Initial Fuel on Board, allowing for a certain margin of error.

COSTS OF FUEL LEAKS

Human Cost of Fuel Leaks

The human cost associated with aircraft that do not correctly identify fuel leaks can be significant and multifaceted. Aircraft fuel leaks pose severe safety risks that can lead to catastrophic events, including fires and explosions, which can result in loss of life and severe injuries.

The most serious human cost is the loss of life. Aircraft accidents involving fuel leaks can lead to fatal crashes, resulting in the deaths of passengers, crew members, and, in some cases, individuals on the ground. In addition to loss of life, survivors of incidents involving fuel leaks can suffer from severe burns, respiratory issues due to inhalation of toxic fumes, and other physical injuries, which may require long-term medical treatment.

Monetary Cost of Fuel Leaks

The monetary cost of fuel leaks and maintenance failures on aircraft can be substantial, impacting various aspects of airline operations and associated industries. Here is a breakdown of the potential financial implications:

1. **Direct Repair Costs**: The immediate costs to repair a fuel leak or address a maintenance failure can be significant, especially if parts need to be replaced or specialized labor is required. These costs vary widely depending on the severity of the issue, the type of aircraft, and the urgency of the repairs.

2. **Aircraft Grounding**: When an aircraft is grounded due to a fuel leak or maintenance failure, it is not generating revenue. The cost of aircraft grounding includes lost revenue from canceled flights and the potential need to book passengers on other airlines. There are a multitude of factors which contribute to the cost of an aircraft on ground. An average aircraft lease for an Airbus A320 is around 150,000 USD.

3. **Fuel Wastage**: A fuel leak directly results in the loss of fuel, which is a considerably expense given the high cost of aviation fuel. The cost of lost fuel adds up, especially if the leak goes undetected over time. This poses a serious monetary cost to airlines as the price of jet fuel has dramatically increased in the last 3 years, reaching a record high of 172 USD

per barrel in June 2022. This is especially concerning when considering that the cost of jet fuel accounts for 25-30% of an airline's operating cost.

4. **Compensation and Liability**: In the event of a fuel leak leading to a delay or cancellation, airlines are often obligated to compensate passengers. In severe cases, such as accidents resulting from maintenance failures, the liability costs can be enormous, including settlements and legal fees. In 1999, an air carrier treaty was signed to compensate parties affected by airplane crashes. Presently, the approximate compensation equals 170,000 USD per passenger, if the airline is found to be at fault for the accident.
5. **Regulatory Fines**: Failure to adhere to maintenance protocols or to address known issues can result in regulatory fines from aviation authorities. For example, in 2021, Boeing was fined 6.6. billion USD due to quality and safety oversight lapses such as cutting maintenance checks. Not adhering to the strict safety regulations is not only monetarily costly for Boeing, but also hurts the company's reputation.

Total Cost Estimate

This cost estimate assumes a moderate fuel leak on an Airbus A380, grounding the aircraft at a major international hub with A380 maintenance facilities. The aircraft was scheduled for a long-haul international flight (e.g., London to Singapore) with high passenger occupancy. We estimate approximately 2-3 roundtrip flights would be canceled due to the 3-day repair time. Calculations are based on an 80% passenger load factor and an average ticket price of $1500.

Cost Category	Estimated Cost	Notes
Grounding Cost	$450,000.00	$150,000 per day for 3 days
Labor	$5,000.00	50 hours at $100/hour
Parts	$15,000.00	Moderate fuel leak repair
Overhead	$15,000.00	$5,000 per day for 3 days
Lost Revenue	$1,200,000.00	2 round-trip flights canceled, 80% passenger load, average ticket price $1500
Accommodation/Rebooking	$80,000.00	10% of passengers, 2 nights at average $200/night
Total Cost	$**1,765,000**	

Reputational Cost of Aviation Incidents

For aircraft manufacturers, having a good reputation is very important because it shows they make safe and reliable planes. After a plane incident, the manufacturers reputation and image could be severely affected. When airlines and passengers trust a manufacturer, they are more likely to buy and use their planes, however if that manufacturer has a history of incidents, it is likely that the public will distrust their product. A strong reputation also helps the manufacturer stand out in a competitive market and can lead to more sales and success.

For example, in recent years, Boeing has faced a severe reputational crisis that has significantly impacted its standing in the aerospace industry. Once regarded as the leading airline manufacturer globally, Boeing has now been overshadowed by Airbus. Government data reveals a concerning record: nearly 530 Boeing 737 aircraft have been implicated in various accidents and incidents around the world, accumulating almost 6,000 fatalities. In stark contrast, the Airbus A320 series has experienced 180 accidents, with fatalities numbering around 1,500—substantially fewer than Boeing's record.

Such a reputational shift can significantly affect a company's market perception and financial health. In January 2023, for instance, Boeing experienced a 29% decrease in its 737 Max 9 aircraft deliveries, while Airbus saw significant gains in the same period. Boeing, once celebrated for its safety and engineering excellence, now faces skepticism regarding its reliability and safety standards. Since 2018, this crisis of confidence has led to sizable financial losses for Boeing investors, totaling nearly 87 billion USD. The company's struggles to address manufacturing and safety issues have not only tarnished its reputation but also weakened its competitive position. This situation underscores an important lesson: prioritizing safety is essential for sustaining profitability and trust in the aviation sector.

APPROACH AND MODEL

Exploratory Data Analysis

The data provided consisted of 8 different datasets: MSN 02, MSN 10, MSN 11, MSN 12, MSN 14, MSN 29, MSN 37, and MSN 53. Each dataset was structured in a time series format, taken from the sensors in the engines of the aircraft for each flight. MSN02 was labeled as being the flight test and included 111 features whereas the other datasets included 17 variables. A summary of the variables can be found in the appendix in **figure 1**.

Because we want to use all the data, we are only considering the shared variables between MSN02 and the non MSN02 datasets. Next, we look for null values. We found that Fuel Used for each of the 4 engines is almost 90% NULL, in all but MSN02. Following that, we are plotting the number of NULL values per flight phase. What we found was that flight phase 8 had the most regular data because this is the Cruise part of the flight, which is less sensitive to sensor fluctuations. Next, we started looking at the time element of the data. We plotted the average duration of each flight. Here we can see that some flights have far less duration than others. For example, one flight (V0886) only had data for 37 minutes. Additionally, we found that flight phase 8 had the longest average duration and data among the other phases. We calculated the NULL values for the features FUEL_USED and VALUE_FOB for each aircraft. What we gathered was that MSN02 was the only dataset which did not have majority null values in those features. For those reasons, we are going to focus on exploring MSN02 further.

As a result of our exploratory data analysis, we decided to focus on MSN02, and specifically only on flight phase 8. Additionally, we concluded that the main variables of focus for us will be 'UTC_TIME', 'Flight', 'FLIGHT_PHASE_COUNT’, 'VALUE_FOB', 'FUEL_USED_1', 'FUEL_USED_2', 'FUEL_USED_3','FUEL_USED_4', 'FW_GEO_ALTITUDE'. These variables are important because they are also common between every dataset. **Figure 2**

After analyzing the EDA, we hypothesized that to detect a leak, we will have to compare the expected fuel on board with the actual fuel on board and any deviation from expected fuel on board would indicate a leak. However, we were not given the values for expected fuel on board, so this will have to be calculated using the equation below.

For each flight:

$$EXP\_FOB = VALUE\_FOB_0 - \sum TOTAL\_FUEL\_USED$$

For each UTC_TIME period:

$$LEAKAGE = EXP\_FOB - VALUE\_FOB$$

Data Cleaning and Preprocessing

Following our EDA, we wanted to understand the importance of the variable deltas/difference per UTC_TIME. Knowing the difference per second of our main variables: FW_GEO_ALTITUDE, VALUE_FOB, TOTAL_FUEL_USED, would give us a better understanding of the dataset.

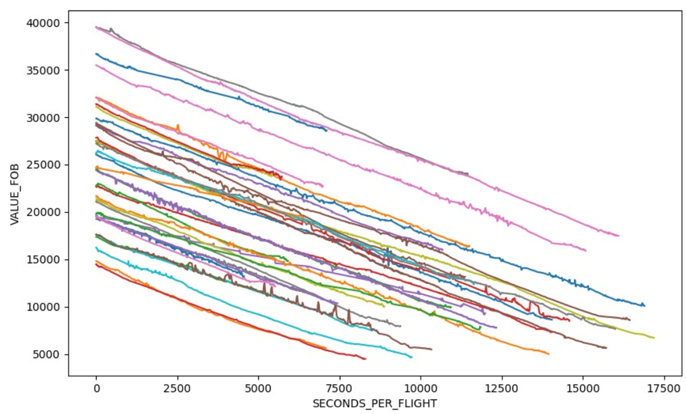
The first thing we did was group by flight and UTC_TIME to identify the sequences of each flight. To accomplish this, we created a column that identifies rows where the difference in UTC_TIME with the previous row is not 1 second. The purpose of this step is to prepare the data for analysis where we know continuous sequences of records at 1-second intervals are important. The code segments the dataset into continuous time blocks which will assist us in analyzing behaviors and measuring distinct time intervals related to our sensor data. We then filled our missing values with 0 since that was the first row in the sequence, this would maintain the integrity of the data.

Additionally, we noticed there were duplicates in the data which we decided to drop. To explain our logic, we will explain each feature we've created below.

Eliminating Noise

Following the creation of our new variables, **Figure 3,** we wanted to plot the flights to get a better understanding of their data structure. After initially plotting all flights in MSN02, we noticed that some flights only had a few minutes' worth of data. This data would only add additional noise to our dataset, so we filtered the dataset to only include flights with more than 1,800 seconds worth of data. This allowed us to maintain the integrity of the data while focusing our analysis on longer flights where fuel data is more relevant.

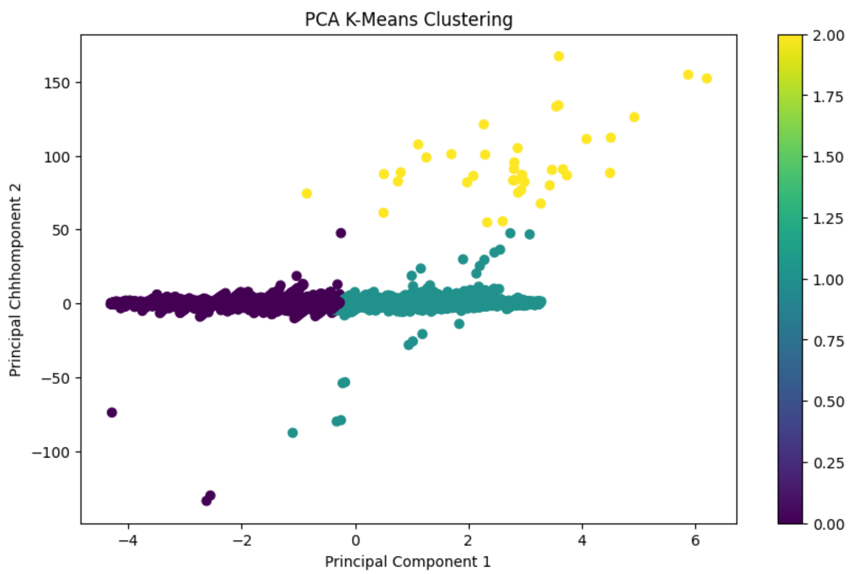
After filtering our dataset to only include the relevant flights, we wanted to understand how VALUE_FOB performed through time. We also noticed that one flight had an immediate drop in its VALUE_FOB. We attributed this to sensor error since the value dropped down to 0 and then went back up. We decided to drop this row from our analysis. We then plotted the continuous values of VALUE_FOB for all the flights and as you can see below, some flights had outliers in VALUE_FOB.

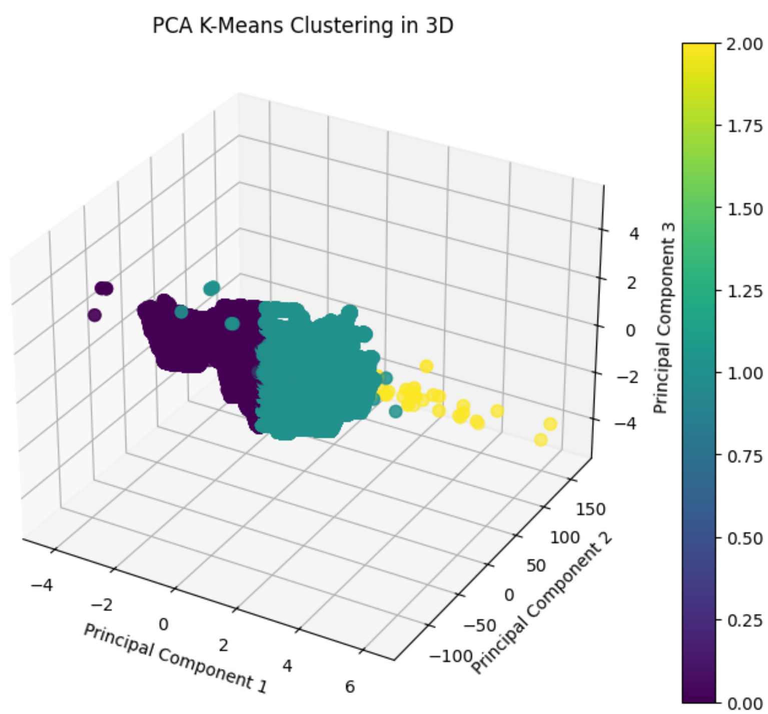


After plotting, we wondered whether we could cluster these flights based on similarity. And second, how can we smooth the values of each flight.

Principal Component Analysis

To cluster our data based on similarity, we performed a PCA analysis with 3 K-Mean clusters and both 2 and 3 dimensions. This also helped us in identifying any anomalies in the dataset. This method provided some very valuable information, we can see that in the 2-D pane, cluster 2 is very dissimilar than cluster 0 and 1. We hypothesized that those data points were outliers and would not be any value add.





To further solidify our hypothesis, we plotted the centroid plot of the clusters formed. In **Figure 4**, we can see that cluster 2 has significantly higher values in the VALUE_FOB and FW_GEO_ALTITUDE features. This could be due to several reasons, such as flights being at higher altitudes with more fuel on board.

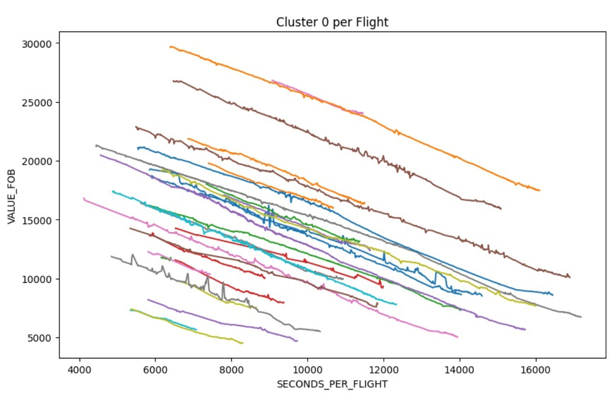
Figure 5 illustrates the loadings of the original features on the first two and three principal components, respectively. In both charts, 'VALUE_FOB' shows a strong positive loading on PC1, whereas 'TOTAL_FUEL_USED' and 'SECONDS_PER_FLIGHT' demonstrate strong negative loadings on PC1, indicating they significantly impact this component. 'VALUE_FOB_DIFF' and 'TOTAL_FUEL_USED_DIFF' have high positive loadings on PC2 across both charts, suggesting they highly influence the variance in PC2. The addition of PC3 in the second chart revealed some new insights: 'FW_GEO_ALTITUDE' and 'TOTAL_SECONDS_PER_FLIGHT' emerge with strong opposing loadings on PC3, which were not visible in the first chart with only two components. This addition of the third component allows the model to account for variations in the data that the first two components alone do not capture, specifically variations associated with the altitude and overall flight duration metrics.

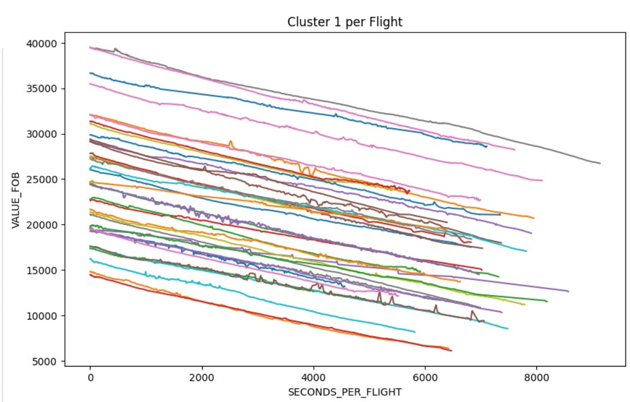
Choosing the right cluster

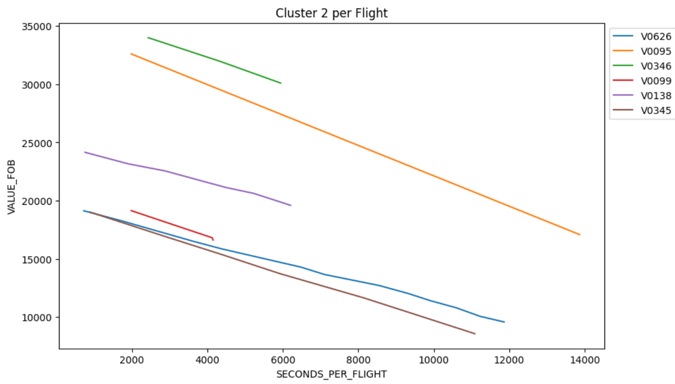
After performng the PCA, we got 3 clusters, each with very different characteristics. In everything but seconds per flight, Cluster 0 and 1 were very similar. However, Cluster 2 was completely different than the others.

Cluster	VALUE_FOB_DIFF	TOTAL_FUEL_USED_DIFF	FW_GEO_ALTITUDE	SECONDS_PER_FLIGHT
0	1.186031	1.154207	13829.427180	10185.637543
1	1.336336	1.314335	16138.462273	3223.196118
2	458.513514	654.962435	3377.791243	5505.000000

To determine which cluster would be more beneficial for further analysis, we plotted cluster 0, 1 and 2. We determined that cluster two does not have enough data and seems to contain mainly outliers.



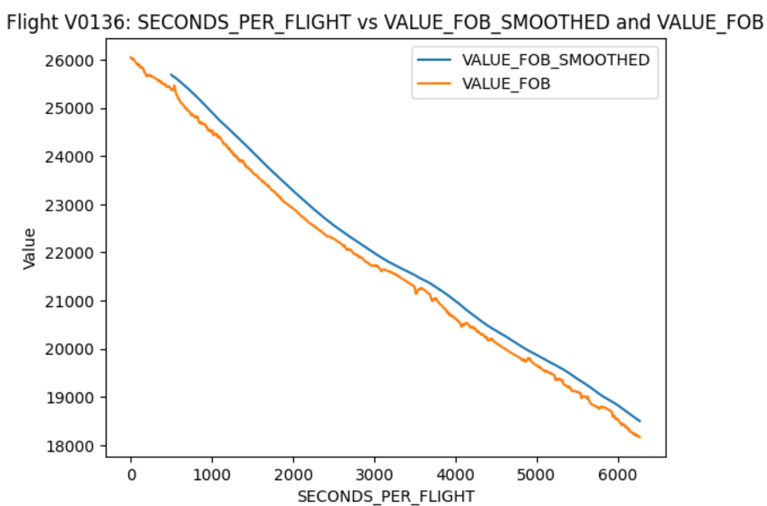




Finally, we observed that cluster 0 had far more variance in VALUE_FOB than cluster 1. Additionally, cluster 1 contained 11 more flights than cluster 0. For those reasons, we continued our analysis with cluster 1.

Cleaning Cluster 1

Despite there being less variance, there were still some anomalies in cluster 1 which we wanted to address. To achieve this, we filtered out the outliers of VALUE_FOB_DIFF, TOTAL_FUEL_USED_DIFF, and FW_GEO_ALTITUDE by calculating the z-scores of each variable and discarding beyond our threshold of 3. Additionally, we performed a moving average calculation on VALUE_FOB using a window size of 500. Our goal is to smooth out the VALUE_FOB data for each flight to help us identify underlying trends in the data and reduce noise and fluctuations.



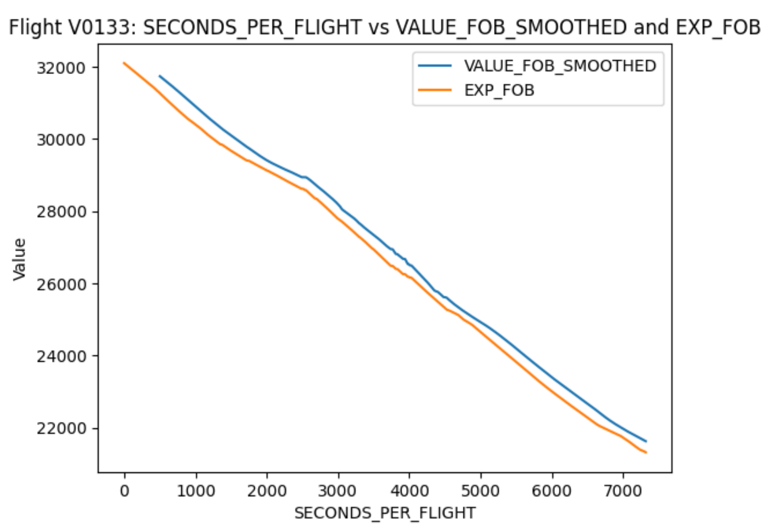
Also, we created another new feature VALUE_FOB_SMOOTHED_DIFF that represents the rate of change in the smoothed VALUE_FOB value from one time point to the next within each flight.

We inverted the sign to make the values more intuitive. This process is useful for detecting trends or patterns in the changes of the 'VALUE_FOB' over time after smoothing out fluctuations. After performing all of these steps, we finally obtain a clean and filtered MSN02 dataset.

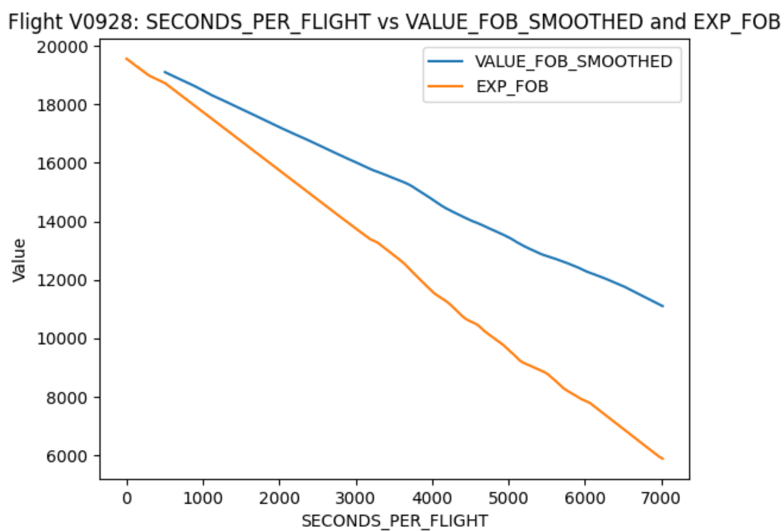
Leak Simulation Approach

Following our analysis, we concluded that our goals were to create a model which would detect and predict leakages. However, because our data was completely standard with no leakages, this meant that we had to simulate leaks.

In order to simulate our leaks, we began by computing the cumulative sum to track the total change in fuel used across consecutive records within each flight. This can reveal trends such as the rate at which fuel is consumed or refilled during the flight. The cumulative sum turns the 'TOTAL_FUEL_USED_DIFF' incremental changes into a total change value from the start of the flight. This is particularly useful for identifying fuel leaks or analyzing fuel usage patterns over the course of each flight. Additionally, we created the 'EXP_FOB' variable to track and estimate the fuel levels throughout the flight, based on the initial fuel amount and the subsequent fuel consumption. This provides a way to monitor fuel usage over time and can be used to detect discrepancies between the expected and actual fuel amounts, which might indicate leaks. We can see that the calculated EXP_FOB variable closely matches the VALUE_FOB_SMOOTHED when there is no known leak.



To achieve this, we separated the MSN02 dataset into 3 random groups of flights, each simulating a different scenario of fuel leakage in the aircraft. Each group represents a different speed of leakage: 0.25, 0.75, and 0.9 liters per second. We specifically chose those rates because anything higher would be too obvious and not as realistic for a minor fuel leak in this scenario. We assigned varied speeds to the variable LEAKEGE_PER_SECOND to better estimate the future state of fuel based on the different leakage conditions, and to identify if our model can predict varied leak speeds. We set the leakage rate to 0 when there is no fuel change and only reflect the leakage when there is an actual drop in the 'VALUE_FOB_SMOOTHED' value. We concatenated each data frame and assigned TOTAL_FUEL_USED_DIFF_LEAK to reflect the sum of the leakage per second and total fuel difference used. Finally, we plotted the difference to visually see the fuel leaks. The gap in VALUE_FOB_SMOOTHED and EXP_FOB was evidently larger the more severe the leak per second was. Also, we created a binary column called LEAKEGE which we assigned a value of 0 for any non-simulated data, and a value of 1 for the simulated fuel leaks. This will provide us with a reference to test the metrics of the models later on.



Finally, we are ready to begin training and testing models with the ultimate goals of predicting and detecting the simulated leaks. We realized that an important variable in our model would be Total Fuel used and the variables that are derived from it, however as previously noted, Total Fuel Used is up to 90% missing in the no-MSN02 datasets. Because of this, we have decided to run two models; one including Fuel used and its derivatives, and one excluding them.

Baseline Models – Logistic Regression and XGBoost Classifier

To kick off our analysis, we created two sets of data: one reflecting the typical conditions of flight and another mimicking potential fuel leak scenarios. The phase was to merge these sets, unifying them under standardized column names for consistency's sake.

The aim here is to use these datasets to refine our machine learning models. Upon merging, we scanned the dataset for any missing entries and removed them to preserve the dataset's integrity. We then produced a more specific version of this dataset, omitting certain columns to tailor it specifically for model training. The 'leakage' variable—our indicator for leaks—allowed us to segment the data into distinct sets for training, validating, and testing. This segmentation is essential to gauge the model's effectiveness on data it has never seen.

The normalization step came next, where we applied the MinMaxScaler to align our features on a uniform scale without skewing their range variances. This step was also applied to the validation and test sets, with the product—a normalized data frame—retaining its structure and labels for further examination.

With our prepared dataset, we tested two machine learning models: the straightforward Logistic Regression and the robust XGBoostClassifier. Logistic Regression served as our baseline model, because of its straightforward approach to binary classification.

We trained both models with the training set, then proceeded to evaluate their predictions using the validation and test sets. The ultimate goal is to develop a predictive model that optimizes recall/sensitivity because it is the metric that captures the percentage of leakage actually identified from all leakage in the dataset. As previously mentioned, we wanted to test the effect of TOTAL FUEL USED on our models to see if we can achieve a high recall.

	MODELS WITHOUT FUEL USED		**MODELS WITH FUEL USED**	
	Logistic Regression	XGBoost	Logistic Regression	XGBoost
Recall	64.6%	41.6%	89%	98%

Overall, both models exhibit significant potential for improvement. Although the models utilizing FUEL USED had significantly higher recall, this would not be beneficial to our overall goal because as stated previously, most of the FUEL USED data from the other MSN datasets were NULL. For this reason, we will focus on improving our model that does not include FUEL USED as it would be a better application for the additional MSN data. In addition to improving those models, we will also implement an autoencoder.

Improvements to Baseline Models

To improve our model, we wanted to see if we could add MSN_02 existing features, as new columns to non-MSN_02 datasets. To get this, what we did is a small feature analysis of the holistic features not shared between the MSN_02 and the non-MSN_02 datasets, to identify variables that are not highly correlated with those of the filtered MSN_02. The goal is to provide more features to this final model that is not included in the other MSN datasets.

We concluded that FUEL_FLOW_AVE (fuel flow average) & PITCH_ANGLE were the most likely features to improve our model’s performance. We proceeded to create predictive models for each of the two variables, both separately and independently. The best metric scores for the final model are achieved by adding only FUEL_FLOW_AVE through the Random Forest Regressor, with a MAPE of 1.41%. Applying this model to the non-MSN_02 datasets allows the creation of the new column FUEL_FLOW_AVE, hopefully improving the final model.

After implementing this feature creation, we ran both Logistic Regression and XGBoost Classifier again to see how our metrics improved. The results were what we had hoped for. The Logistic Regression recall improved from the original 64.6% to 75%.

Alternative Model – Long Short-Term Memory Network Autoencoder

Another approach we took to detecting fuel leaks was using an LSTM autoencoder model on our preprocessed dataset. An autoencoder is beneficial in this scenario because of its ability to reduce the dimensionality of complex data and identify anomalies. They are an unsupervised learning method trained on the "normal" behavior of the data and can reconstruct the input data it is given. When the model identifies a leak or pattern that deviates from the normal, pre-established patterns, the autoencoder struggles to reconstruct the data accurately, flagging it as an anomaly. Since the autoencoder works on unsupervised normal data, we trained and validated it on our clean MSN_02 dataset which does not have our simulated leaks. We added other variables to make sure that the model captures multiple patterns about the normal behavior of flights. Below is a brief description of each of the features that were added to our training data:

- PERCENTAGE_CHANGE_TOTAL_FUEL_USED, PERCENTAGE_CHANGE_FOB, FOB_ERROR, FOB_POLY

To build our model, we used LSTM layers as they are suitable for capturing temporal dependencies in sequence data and we added a L2 regularization and gradient clipping to prevent overfitting and manage gradient stability during training. After defining the architecture of our model, we proceed to scale our data using a 'MinMaxScaler' to transform the data to a range between 0 and 1. Since the model requires the input of data as sequences, we created sequences of our data using a fixed time window of 30 minutes. Once our training data was correctly structured, we trained the Autoencoder for 10 epochs, reaching a final loss of 0.093.

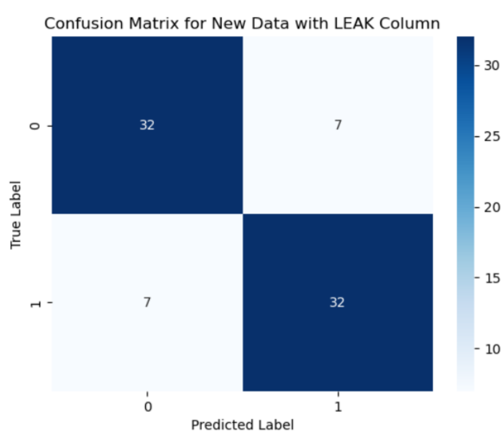
To determine the accuracy and recall of our model, we tested it using our MSN_02 dataset, including the constant leaks and compared the output to our binary LEAKAGE column which we know accurately classifies a leak. The initial purpose of this testing is to see, when trained on normal data, if the autoencoder can distinguish between normal data and simulated leak data. The results obtained for the accuracy, precision, recall and F1 suggested that the model was overfitted since all of them had an equivalent score of 1. These results may be caused by unintentional leakage of information from the test set into the training process or the model is memorizing the data. To prevent having these types of problems, we decided to use a completely different data set to perform the validation of our model. For this purpose, we preprocessed the data included in file

MSN_37, as we know that the flights included in this data is unseen data for the model. After adding artificial leaks to the data and concatenating it with the original values from the file, we structured the data to match the format of our training set and evaluated the model's performance. The results obtained during this process returned an accuracy, precision and recall of 0.82, suggesting that the model has a good capacity to detect both anomalies and normal behaviors in the data.

In summary, while the model is good at identifying normal and anomalous conditions, it still needs further adjustment and experimentation since the model is built based on several assumptions that may require additional analysis to determine the best configuration. These are some of recommendations and further actions that could be applied to tailor our model:

- Evaluate the possibility of incorporating new features for the training process. It's critical to include features that are most indicative of anomalies.
- Increase the size of the training set, particularly with more examples of leaks to manage the imbalance problem in the training set.
- Implement model tuning using different numbers of layers and units within those layers to adjust the model's complexity. This approach can help capture more subtle patterns.
- Incorporate domain knowledge to understand the typical patterns of fuel usage and identify which deviations are most closely related to leaks. This could include as well defining the optimal threshold to classify sequences based on expert feedback.

	Precision	Recall	F1-Score	Support
0	0.82	0.82	0.82	39
1	0.82	0.82	0.82	39
Accuracy			0.82	78
Marco Avg	0.82	0.82	0.82	78
Weighted Avg	0.82	0.82	0.82	78

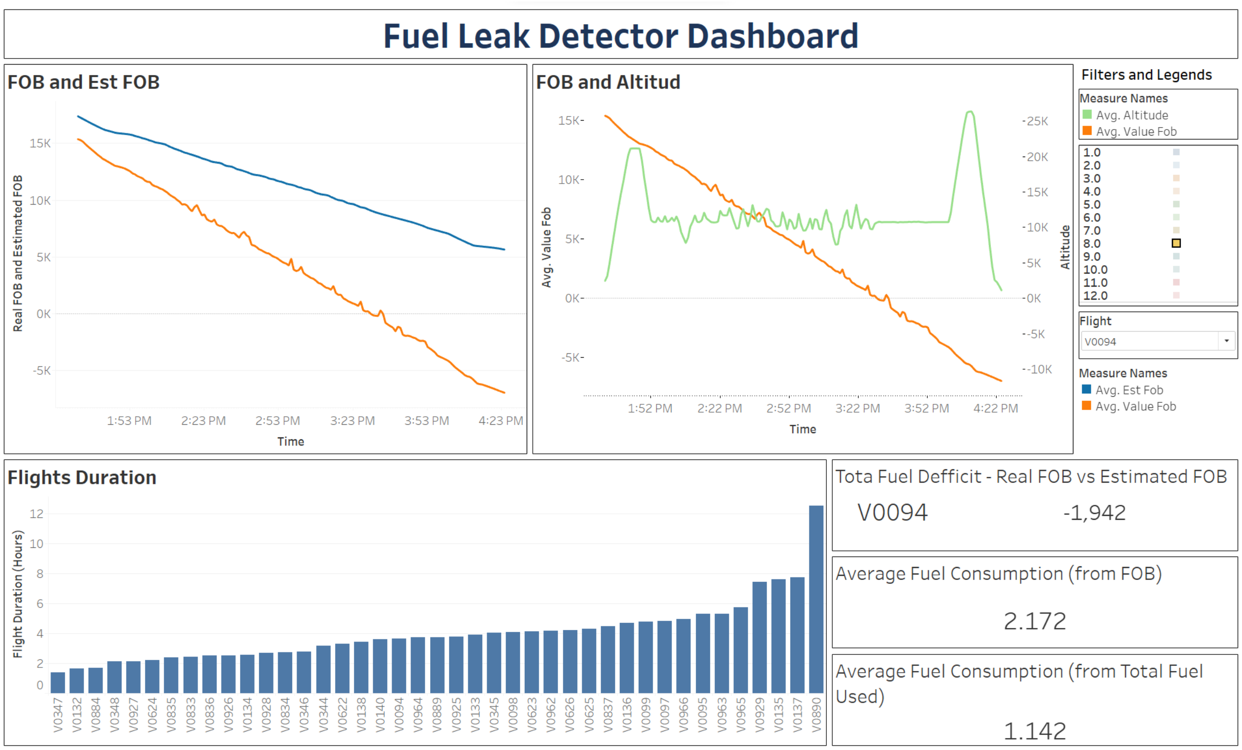


FUEL LEAK DETECTION DASHBOARD

In order to deploy our models, we want to create an interactive dashboard that displays key metrics for each flight and is able to visualize patterns that are indicative of any potential leaks. This dashboard is designed to monitor fuel levels and consumption during flights, helping in the early detection of potential fuel leaks. With this dashboard, Airbus can proactively assign maintenance checks to aircrafts which have been flagged as having a potential leak.

Key features of the dashboard:

- A graph comparing real-time Fuel on Board (FOB) with estimated FOB over time, providing a visual representation of expected versus actual fuel levels.
- A chart correlating FOB with altitude, offering insight into fuel consumption patterns relative to flight altitude changes.
- A histogram displaying the duration of flights, allowing for quick assessment of which flights might require further analysis based on their length.
- Detailed statistics, such as the total fuel deficit (the difference between real and estimated FOB), and average fuel consumption rates based on FOB and total fuel used, equip maintenance crews with precise data to identify and address fuel discrepancies.



CONCLUSIONS

Model Prediction Comparisons

Using all the models we created, we predicted on MSN37 to determine the recall and performance. The results are shown below:

	Models		
Metric	Logistic Reg	Logistic Reg_FU	LSTM
Recall	75.21%	90.02%	82.5%
% leaks	5.56%	3.63%	N/A
No. of unique flights with leakage	35	150	N/A

As we can see here in the comparison of all of our models, the model with the best recall metric is the Logistic Regression models which included FUEL USED. However, as previously stated, this is not realistic given the number of NULL values in FUEL USED in the other MSN datasets. For example, MSN10 has no FUEL USED values.

Also to note, our Logistic Regression model has identified 5.56% of leaks in 35 unique flights. However, Logistic Regression including FUEL USED identified 3.36% of leaks in 150 unique flights. Our observation is that, since we clustered the flights by total seconds per flight, we only receive an average of 3,000 seconds per flight. This leads us to believe that in order to improve this, we need more flights with more duration. Comparatively for anomaly detection, rather than prediction, the LSTM model performed very well, achieving a recall of 82.5%. However because we pre-sequenced the data to feed it into model, it is more difficult to get the exact number of unique flights which had a leak.

The main insight we had is that to utilize our most accurate model, we need consistent and accurate FUEL USED data in all MSN datasets. To achieve this, perhaps some more accurate sensor equipment is required.

IMPROVEMENTS AND LEARNINGS

New form of data measurement

In our analysis, we prioritized examining flight phase 8, where the aircraft is in steady, horizontal flight, which minimizes variables like pitch and noise and enhances the accuracy of our sensor measurements. We decided on this approach due to flight phase 8 including the least null values.

However, in hindsight, it would be more functional to extend our observations to include the other phases as well, as leaks may also be present during those phases. In this scenario, the measurement of fuel could be based on the different weights of the tank rather than sensor data since sensor data might not be as accurate during those phases.

Complex Models

Although our best model was our simplest model, we think that other more complex models might also provide a valuable solution to our problem. Technology and machine learning is constantly evolving, giving the opportunity to try additional, more complex models. Deep learning models might yield better results in our case because they are good at capturing temporal dependencies and long-term patterns in sequence data, which autoencoders may overlook. Additionally, ensemble methods, which combine the predictions from multiple models would be extremely beneficial in improving overall accuracy and recall.

Value Proposition

Implementing a machine learning model for detecting fuel leaks in Airbus planes offers numerous advantages, including enhanced safety through immediate identification and correction of leaks, thereby preventing potential accidents. This innovation not only results in considerable cost savings by averting repairs and optimizing fuel efficiency but also benefits the environment by reducing fuel wastage and lowering emissions. These technological advancements position Airbus as a frontrunner in aviation innovation, fostering trust and support from regulators and customers alike. By setting new industry benchmarks and demonstrating a commitment to safety and sustainability, Airbus enhances its brand reputation, potentially increasing its market share as customers gravitate towards companies prioritizing safe and environmentally friendly.

More extensive preprocessing

The results from our model in identifying fuel leak patterns signals an important move towards preventive maintenance strategies. Although our model demonstrated accuracy, there is still room for improvement. Refining the model, through additional feature engineering, could increase its precision and recall, and significantly increase its utility. This would ensure more reliable detection of fuel leaks in operational settings. Additionally, a change to our preprocessing methods might also benefit the overall accuracy of the model. Improving how we prepare our data is a key step in making our model work better and more accurately. By carefully getting our data ready, we not only make our model more accurate now, but also ensure it remains trustworthy and useful for enhancing aviation safety in the future.

APPENDIX & BIBLIOGRAPH

Figure 1:

A/C and flight data:

- Time, day, month, year → ONLY MSN 02
- UTC date/time → UTC_TIME
- MSN (A/C Name) → MSN
- Flight number → Flight
- Flight phase* → FLIGHT_PHASE_COUNT
- Altitude → FW_GEO_ALTITUDE
- Pitch and roll → ONLY MSN 02

Fuel/Engine system data:

- Engine status (Running or not). → ONLY MSN 02
- Fuel flow (to each engine) → ONLY MSN 02
- Fuel used (by engines; Kg):
    - FUEL_USED_1 → (Engine 1)
    - FUEL_USED_2 → (Engine 2)
    - FUEL_USED_3 → (Engine 3)
    - FUEL_USED_4 → (Engine 4)
- Fuel on board (“FOB” ; Kg) → VALUE_FOB
- Fuel quantity per collector cell and surge tank volume (Kg):
    - VALUE_FUEL_QTY_CT → Centra Tank
    - VALUE_FUEL_QTY_FT1 → Feed Tank 1 (Engine 1)
    - VALUE_FUEL_QTY_FT2 → Feed Tank 2 (Engine 2)
    - VALUE_FUEL_QTY_FT3 → Feed Tank 3 (Engine 3)
    - VALUE_FUEL_QTY_FT4 → Feed Tank 4 (Engine 4)
    - VALUE_FUEL_QTY_LXT → Transfer Tank Left
    - VALUE_FUEL_QTY_RXT → Transfer Tank Right
- Pump status (On/Off, normally/abnormally, immersed/not immersed). → ONLY MSN 02
- Leak detection and leak flow. → ONLY MSN 02
- Fuel transfer mode. → ONLY MSN 02

Figure 2:

	Dataset	FUEL_USED with nulls	FOB with nulls	No null values
0	msn02	45659	4	514663
1	msn10	538649	29670	24
2	msn11	4059234	136220	96431
3	msn12	3116948	30490	90546
4	msn14	4371574	62154	203932
5	msn29	3759784	138784	50138
6	msn37	2920022	72611	243494
7	msn53	2767387	46209	219761

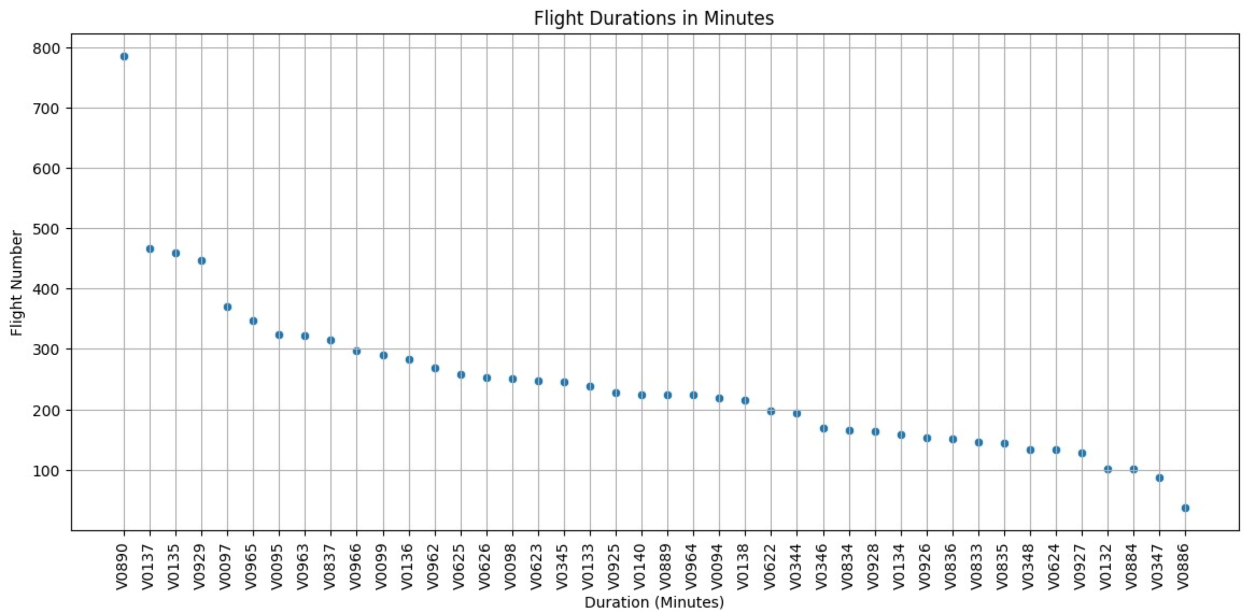


Figure 3:

- **TOTAL_FUEL_USED**: We aggregated the fuel usage from four different metrics (**FUEL_USED_1**, **FUEL_USED_2**, **FUEL_USED_3**, and **FUEL_USED_4**) to formulate a new column, **TOTAL_FUEL_USED**. Subsequently, we removed the initial four fuel usage columns to simplify our dataset.
- **SECONDS_PER_FLIGHT**: We calculated the elapsed time in seconds for each flight by assessing the difference in **UTC_TIME** from the beginning of each flight. This metric assists in evaluating the duration of each flight operation.

- **TOTAL_SECONDS_PER_FLIGHT**: For every flight, we computed the total flight duration in seconds using the **UTC_TIME** by identifying the maximum and minimum time differences. This feature provides the entire duration for each flight.

- **VALUE_FOB_DIFF**: Quantifies the change in FOB value between successive entries within the same flight. We adjusted the sign by multiplying the difference by -1 and filled any null values with zero, offering a consistent and clear dataset for further analysis.

- **TOTAL_FUEL_USED_DIFF**: We determined the variance in **TOTAL_FUEL_USED** between sequential records for each flight and captured this in **TOTAL_FUEL_USED_DIFF**. We then populated any missing values with zero to ensure uniformity across our dataset.

- **VALUE_FOB_SMOOTHED**: Calculate a rolling mean, or moving average, of the 'VALUE_FOB' column within the **cluster_1_data** data frame, grouped by each unique 'Flight' to smooth the data for the 'VALUE_FOB' variable and mitigate the effects of noise and outliers in the data.

- **VALUE_FOB_SMOOTHED_DIFF**: Captures the rate of change in the smoothed fuel on board (FOB) value between consecutive records for each flight. This variable represents the change in fuel amount from one time point to the next but smoothed over time to reduce noise.

- **cumsum_TOTAL_FUEL_USED_DIFF**: Created by computing the cumulative sum of the 'VALUE_FOB_SMOOTHED_DIFF' column within each group of data defined by 'Flight'. The purpose of calculating a cumulative sum is to track the total change over time, which could provide insights into the overall trend of fuel consumption over the course of a flight.

- **EXP_FOB**: This new column represents the expected amount of fuel on board (FOB) for each record within each flight. The purpose of this operation is to estimate how much fuel is expected to be on the aircraft at any given point during the flight, assuming no leak is present. This will help us identify discrepancies that might indicate a leak.

- **LEAKEGE_PER_SECOND (0.25, 0.75, 0.9)**: Assigns values to a new column called 'LEAKAGE_PER_SECOND' in three different data frames (msn_02_leak1, msn_02_leak2, and msn_02_leak3). This new column is intended to quantify the rate of fuel leakage per second in different scenarios that represent different conditions or stages of a leak. This allows us to analyze the impact of different rates of fuel leakage on the aircraft's fuel levels over time.

- **TOTAL_FUEL_USED_DIFF_LEAK**: Takes each entry in the 'TOTAL_FUEL_USED_DIFF' column and adds the corresponding entry from the 'LEAKAGE_PER_SECOND' column. incorporate an estimation of fuel leakage into the overall fuel used calculation. Provides a more complete picture of fuel dynamics during the flight, where simple fuel usage metrics don't capture the full scope of the situation.

- **FUEL_FLOW_AVE:** In parallel, since we do not have any information about TOTAL_FUEL_USED in the non MSN02 data, we tried to create another highly correlated feature from the MSN02 data to then include into the non MSN02 data. This will give us another feature based on the FUEL_FLOW of MSN02.

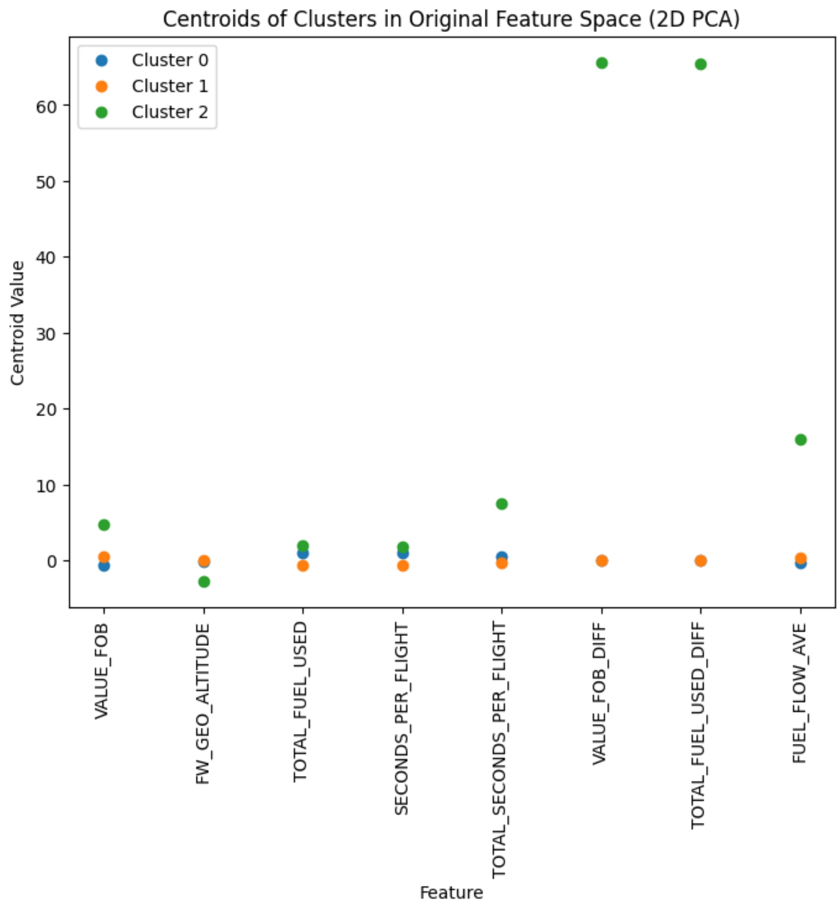
- **PERCENTAGE_CHANGE_TOTAL_FUEL_USED**: Represents the percentage change in the TOTAL_FUEL_USED every second.

- **PERCENTAGE_CHANGE_FOB**: Represents the percentage change in the FOB every second.

- **FOB_ERROR**: Incorporates the difference between the real FOB and the expected FOB calculated using the total fuel consumption.

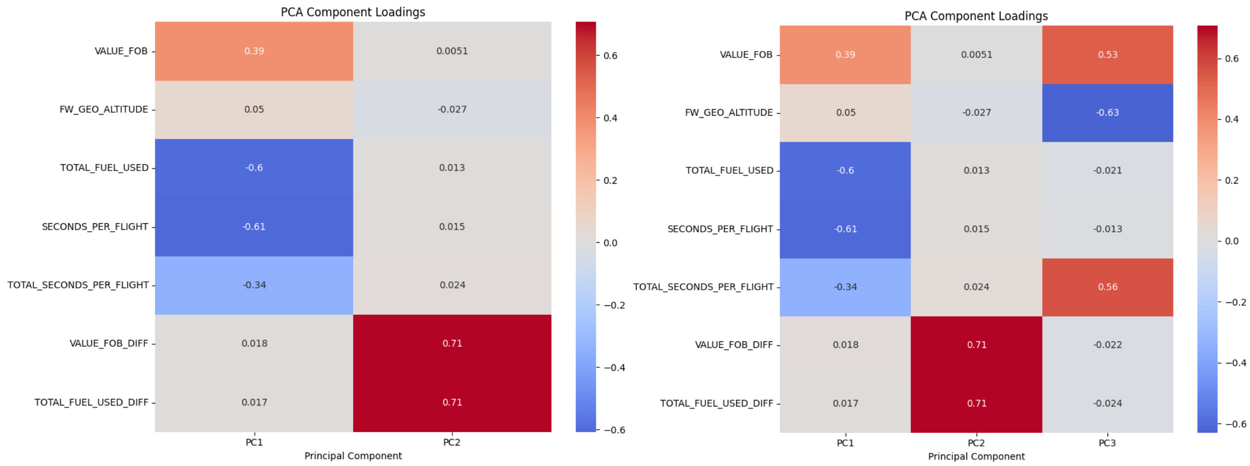
- **FOB_POLY:** This feature is a representation of the FOB value based on a polynomial of degree 5. This variable helps to understand the general behavior of our variable of interest, removing the effects of outliers and high fluctuations.

Figure 4:



Centroids of Clusters in Original Feature Space (2D PCA)
Cluster 0
Cluster 1
Cluster 2
Centroid Value
60
50
40
30
20
10
0
VALUE_FOB
FW_GEO_ALTITUDE
TOTAL_FUEL_USED
SECONDS_PER_FLIGHT
TOTAL_SECONDS_PER_FLIGHT
VALUE_FOB_DIFF
TOTAL_FUEL_USED_DIFF
FUEL_FLOW_AVE
Feature

Figure 5:


PCA Component Loadings
VALUE_FOB
FW_GEO_ALTITUDE
TOTAL_FUEL_USED
SECONDS_PER_FLIGHT
TOTAL_SECONDS_PER_FLIGHT
VALUE_FOB_DIFF
TOTAL_FUEL_USED_DIFF
0.39
0.0051
0.05
-0.027
-0.6
0.013
-0.61
0.015
-0.34
0.024
0.018
0.71
0.017
0.71
PC1
PC2
Principal Component
PCA Component Loadings
0.53
-0.63
-0.021
-0.013
0.56
-0.022
-0.024
PC3
Principal Component

Bibliorgraphy:

Chandler, Charles. "MRO Operation: Finding and Fixing Fuel System Leaks." *Aviation Pros*, Aviation Pros, 24 May 2012, www.aviationpros.com/engines-components/aircraft-airframe-accessories/fuel-systems/article/10618623/mro-operation-finding-and-fixing-fuel-system-leaks.

"Big Data: Racing to Platform Maturity." n.d. https://www.aircraftit.com/articles/big-data-racing-to-platform-maturity/.

"Sky Airlines Testimonial." 2022. Airbus Aircraft. August 11, 2022. https://aircraft.airbus.com/en/services/enhance/skywise/skywise-success-stories.

"Time-of-Flight Sensors - STMicroelectronics." n.d. STMicroelectronics. https://www.st.com/en/imaging-and-photonics-solutions/time-of-flight-sensors.html.

"U.S. Passenger Carrier Delay Costs | Airlines for America." n.d. Airlines for America. https://www.airlines.org/dataset/u-s-passenger-carrier-delay-costs/.

"Fines and Penalties | EASA." n.d. EASA. https://www.easa.europa.eu/en/regulations/fines-and-penalties.

"Aviation and Plane Crash Statistics | Updated 2024." 2024. Panish | Shea | Ravipudi LLP. February 6, 2024. https://www.panish.law/aviation_accident_statistics.html.

Statista. 2023. "Airbus and Boeing: Aircraft Orders 2006-2021." August 23, 2023. https://www.statista.com/statistics/264492/aircraft-orders-from-airbus-and-boeing/.

McCarthy, Niall. 2019. "Infographic: Boeing Sees 51% Year-On-Year Drop in Earnings." Statista Daily Data. October 24, 2019. https://www.statista.com/chart/18806/boeings-net-earnings-by-quarter/.

2023b. "Airbus and Boeing - Key Figures 2021." December 1, 2023. https://www.statista.com/statistics/277047/key-figures-of-airbus-and-boeing/.

"Topic: Airbus and Boeing." 2024. Statista. January 10, 2024. https://www.statista.com/topics/3697/airbus-and-boeing/#topicOverview.

"Claim Compensation if Your Flight's Delayed or Cancelled." n.d. Citizens Advice. https://www.citizensadvice.org.uk/consumer/holiday-cancellations-and-compensation/if-your-flights-delayed-or-cancelled/.