aerospace

Article

Study of the Impact of Traffic Flows on the ATC Actions

Guillermo Gutiérrez Teuler [1,*], Rosa María Arnaldo Valdés [1], Victor Fernando Gómez Comendador [1], Patricia María López de Frutos [2] and Rubén Rodríguez Rodríguez [2]

[1] Aerospace Systems, Air Transport and Airports Department (SATAA), School of Aeronautical and Space Engineering (ETSIAE), Polytechnic University of Madrid (UPM), 28040 Madrid, Spain
[2] Reference Centre for Research, Development and ATM Innovation (CRIDA), 28040 Madrid, Spain
* Correspondence: guillermo.gutierrez.teuler@alumnos.upm.es

Abstract: It has always been a topic of great interest in air transport management to be able to estimate controller workload. So far, research has not had the opportunity to make use of real data on the controller's actions. We have enough data to be able to use machine learning methods. The aim of this work is to predict the controller's actions to know his workload. Several machine learning models were tested to try different combinations of features and the selected algorithms and two models were finally chosen. The predictions provided by the models are good enough to be used when a first approximation of the workload in a sector is to be obtained. Finally, explainability techniques were employed to discover the patterns found by the AI in the machine learning models. Thanks to these techniques, we can build a profile of the critical flights that increase the workload the most.

Keywords: air traffic management; artificial intelligence; air traffic control; mental workload; machine learning; explainable AI

Citation: Gutiérrez Teuler, G.; Arnaldo Valdés, R.M.; Gómez Comendador, V.F.; López de Frutos, P.M.; Rodríguez Rodríguez, R. Study of the Impact of Traffic Flows on the ATC Actions. *Aerospace* **2022**, *9*, 467. https://doi.org/10.3390/aerospace9080467

Academic Editor: Alexei Sharpanskykh

Received: 7 July 2022
Accepted: 18 August 2022
Published: 22 August 2022

1. Introduction

1.1. Background and Related Work

Air traffic controllers' mental workload is one of the most historically researched topics in air transport management. The complexity and the number of a controller's tasks may increase the mental workload and decrease the operator's performance. This is critical because it has a significant impact on the safety of the Air Traffic Management (ATM) system. In the last decade, there has been renewed interest in this topic due to the need to study the impact that automation has had on the behaviour of controllers [1,2]. To assess this impact, new mental workload models have been developed [3]. The objective of these models is to analyse how the reallocation of automation functions between human and system has an impact on related cognitive processes and mental workload [4].

Workload measurement methods can be grouped into three categories [5]: subjective measure [6], physiological measure and performance measure. There has been a great development of physiological measurement technologies [7]. This type of measurement aims to associate physiological changes with workload levels and has several advantages: the ability to provide real-time data, more accurate reporting of mental workload, standardisation and being able to compare between different studies [8]. Our study falls into the category of performance measure because of the nature of the data we have gathered. We have operational data from ENAIRE, Spain's Air Navigation Service Provider (ANSP). Performance measures are not as accurate as physiological ones, but positive relations between measures of mental workload and task difficulty are usually found in empirical research [9].

This study is not based on assessing test subjects as they perform planned tasks through physiological measurements. It focuses on trying to predict the actions that the controller will perform given a known traffic situation. This is possible thanks to data recorded during operations in various sectors of Spanish airspace and provided by CRIDA.

The data include information about traffic, sector flows and the actions taken by controllers. These data are used to train machine learning models. Thanks to the use of artificial intelligence, these models can make predictions and discover patterns in the given data.

There has been already research with a similar aim. Theoretical models have been developed to try to predict the mental workload of en-route controllers [10]. An investigation is carried out to find how traffic factors, airspace factors and operational constraints interact with task demands and controller mental workload. Then, the theoretical model provides a framework for studying possible ATCOs control strategies.

Machine learning had been used to try to predict controller actions, detect conflicts between aircraft, and resolve these conflicts [11]. We are more interested in the first problem addressed, prediction of controller actions. This problem is identified as a supervised learning problem and the Random Forest algorithm is used to solve it. Controllers usually work in pairs, the main responsibility of the planner controller is to coordinate with the neighbouring sectors while the executive controller is responsible for issuing clearances and instructions to aircraft that are in their sector. The main differences with our work are that the predicted actions are those of the planning controller and not those of the executive controller, and the lack of real data to work with. Duc-Thinh Pham overcomes this problem by analysing aircraft trajectories into and out of sectors. From the extraction of patterns from the trajectories, he infers the macro actions (simpler definitions of actions) that the controllers have performed.

It is fundamental to understand how models make predictions, and therefore an important part of this work is the application of explainability methods to the trained models. These methods are included in the concept of eXplainable AI (XAI). XAI has gained prominence recently due to the need to increase trust in the new AI-based ATM decision-support systems (DSS) that are being developed. There is a gap between research projects and their practical implementation due to different policy and practical challenges. To help overcome these challenges, frameworks have been developed to build trust and enable end-user feedback [12]. The aim is that these frameworks help to advance AI within ATM and reduce the gap between research and implementation of this technology.

It has also been proposed to include explanations generated by the explainability methods in the interfaces used by controllers to increase users' confidence in new AI-based systems [13]. A model is built that predicts the probability of an aircraft having a weather-related accident or incident. An XGBoost model is used to generate the predictions. Two modules are used to obtain the explanations, SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). This research is one of the most similar to our work, as the choice of the ML model and the modulus of explainability is the same.

1.2. Focus and Structure of the Document

As we have seen throughout the introduction, there are certain areas that have not been covered very well. Although there is previous work on machine learning and the mental workload of controllers, data from real operations has often not been used. Thanks to ENAIRE (The Air Navigation Service Provider (ANSP) of Spain) we have had the opportunity to use real operational data to train the machine learning models. We have also proposed to predict the controller's actions from the air traffic situation. We believe it is a novel approach, as it brings together two fields of research that are usually separated. In addition, being able to predict the actions controllers will take in a sector will allow to estimate future workload which can be an excellent metric for air traffic management.

The structure of this work is as follows. First, we will start with a description of the data used in the study. It will then go on to define the methods used to process the data. That is, the machine learning models and the explainability techniques. In the second part of the paper, the results are presented. It begins by showing the findings of the exploratory data analysis. Then the new variables that have been created are defined. Finally, the process followed to train the machine learning models and the results they produce are described. The paper ends with the conclusions drawn and the proposed future lines of research.

2. Materials and Methods

2.1. Data

The data have been obtained thanks to CRIDA and the information comes from four sectors. Countries' airspaces are divided in sectors and assigned to controllers in order to break down the provision of air traffic services into tasks with a manageable workload Three different sets of data are provided: information on the characteristic flows of the sectors, traffic data for flights crossing the sectors under study and controller actions data, hereafter referred to as ATOM data. After cleaning and processing the data, a total of 55,000 flights between the four sectors are available. All data are records taken from actual flights that occurred throughout 2018.

2.1.1. Flow Data

It gives information on the characteristic flows of the sectors. The geometry of the flow is defined by the reporting points through which it passes. The objective is to locate the characteristic flow through which each flight has crossed the sector. If an exact match is not found, each flight is assigned the characteristic flow that most closely matches its route. Flights that do not have a clearly defined characteristic flow are an important factor to take into account. Flows of traffic which are used very infrequently or traffic that is not following a defined flow can cause the most complexity and increase workload considerably due to their irregularity. The end result is that each flight has a new column with its assigned characteristic flow. In this way, we collect the flight path information in a way that is easy for machine learning algorithms to read.

2.1.2. Traffic Data

It contains information about the flights that have crossed the sector. For each flight, its callsign, departure and destination airport and the reporting points it passed through when crossing the sector are recorded. In addition, on entering and leaving the sector, the time, flight level and trend (climb, descent or cruise) are recorded too. This dataset is the main one and contains most of the variables that machine learning models will try to make predictions with. The information from the rest of the datasets is processed and modified to be added to this one.

2.1.3. Atom Data

ATOM data are real-time recordings of the actions taken by the controllers. This is a CRIDA project that seeks to gather more information on the actions of controllers. Each action is manually recorded along with the time it occurs. The type of action and the callsign of the aircraft to which it is directed are noted. This information is cross-referenced with traffic data. This way we know the actions that each flight that has crossed the sector has received.

2.2. Machine Learning

We have chosen to use machine learning because it is ideal for complex problems for which there is no direct solution. There are several types of machine learning systems [14] depending on whether or not they are trained with human supervision, whether or not they can learn incrementally on the fly, and whether they work by simply comparing new data points with known data points, or whether they detect patterns in the training data and build a predictive model.

For the problem addressed in this study, the machine learning system needs to be of the following types:

1. Supervised learning: In supervised learning, the training data that are fed into the algorithm include the desired solutions, called labels. In this case, the label

will be related to the workload that a flight causes to the controller. This type of model has been chosen because the feature to be predicted is known in advance. A typical supervised learning task is classification. Flights are divided into two groups according to the actions received, and the model learns to classify a flight into one or the other. Another typical task is to predict a target numerical value, given a set of features called predictors. This type of task is called regression. In this case, the model tries to predict the number of actions a flight has received from the controller given a set of features. Supervised learning has proven to work well for solving both classification and regression problems in the ATM field [15,16].

2. Batch learning: In batch learning, the system does not learn incrementally, it is trained using all available data. First the system is trained, then it goes into production and runs without learning more; it only applies what it has learned. This is called offline learning. In order for a batch learning system to learn new data, a new version of the system has to be trained from scratch on the complete dataset (not only the new data, but also the old data). This is the type of system chosen, since it is based on initial data and there will not be a regular arrival of new data. In the context of a real use of the models studied in the work, an online system would be necessary. The system would continue to learn and adapt continuously with the new data generated every day.
3. Model-based learning: The aim is to generalise from a set of examples to build a model. This model is then used to make predictions. Other works have recently used this approach in the field of ATM, using machine learning to develop new mathematical models to determine air traffic complexity [17].

2.3. Machine Learning Algorithms

All the algorithms used in this study are batch-supervised, model-based machine learning algorithms. Two algorithms have been used in this work: Random Forest and XGBoost. Both of which are decision trees algorithms.

There are other interesting machine learning algorithms which have not been used in this work, such us neural networks [18] and support vector machine (SVM) [19]. These algorithms have been used successfully in the ATM field [20,21].

Neural networks are considered deep learning algorithms. Deep Learning is a type of machine learning whose algorithms are inspired by the human brain, mimicking the way biological neurons signal each other. They have not been chosen in this work because they usually decimate the interpretability of the features to the point where they become meaningless. Instead, we wanted to focus on explainability, as it is more relevant for practical use in the ATM field.

SVM are well suited for the classification of complex datasets but because of the tabular nature of our data, Random Forest was way more accessible for creating the ML models. Early exploratory data analysis showed that the data were not very spare and were easy to classify. The best approach was not clear, and SVM were discarded.

2.3.1. Decision Trees

Decision trees are a series of sequential steps designed to answer a question and provide probabilities, costs or other consequences of making a particular decision. Decision trees have two advantages: they are easy to understand and provide a clear view to guide decision-making progress. However, this simplicity comes with some serious disadvantages, such as overfitting, bias error and variance error.

Overfitting occurs for many reasons, such as the presence of noise and lack of representative instances. Overfitting is possible despite having a large tree. Bias error occurs when too many constraints are imposed on the objective functions. For example, restricting the result with a restrictive function or by a simple binary algorithm will often result in bias. Error variance refers to how much an outcome will change as a function of changes in the training set. Decision trees have high variance, which means that small changes in the training data can cause large changes in the outcome.

This algorithm has not been used in this study because it is too simple. It has been explained in this section because it is the basis on which the other two algorithms that have been used are built.

2.3.2. Random Forest

As noted above, decision trees are fraught with problems. If there were a way to generate numerous trees, by averaging their solutions, then one would probably get an answer very close to the true answer. Thus, the random forest is a collection of decision trees with a single aggregate outcome. Random forests are often considered the most accurate learning algorithm [22].

Random forests reduce the variance of decision trees by using different samples for training, specifying subsets of random features, and constructing and combining small trees.

A single decision tree is a weak predictor, but is relatively quick to construct. A larger number of trees provides a more robust model and avoids overfitting. However, the more trees you have, the slower the process. Each tree in the forest has to be generated, processed and analysed. Also, the more features you have, the slower the process; reducing the feature set can drastically speed up the process. The disadvantage compared to decision trees is that random forests are more complicated to interpret. This algorithm has been used with good results to predict air traffic delays [23].

2.3.3. XGBoost

Like random forests, gradient boosting is a set of decision trees [24,25]. The two main differences are:

1. Random forests build each tree independently, while gradient boosting builds one tree at a time. This additive model works incrementally, introducing a weak learner to improve the deficiencies of existing weak learners;
2. Random forests combine the results at the end of the process by averaging, while gradient boosting combines the results throughout the process.

If the parameters are carefully tuned, gradient boosting can result in better performance than random forests [26]. However, gradient boosting may not be a good option where there is a lot of noise, as it can lead to over-tuning. They are also often more difficult to tune than random forests.

Random forests and XGBoost excel in different areas. Random forests work well for multi-class object detection and bioinformatics, which tends to have a lot of statistical noise. XGBoost works well when there is unbalanced data, such as in real-time risk assessment. This is precisely the scenario here in the study, an unbalanced dataset, which is why XGBoost works best in our case. In the vast majority of flights, there are few controller actions. The flights that are of interest are those that receive numerous actions and increase the workload. But they are a small part of the flights studied, so that is why it is an unbalanced dataset.

This algorithm has also been used successfully in ATM research. It has been used to evaluate air traffic complexity in the operation of the air traffic system [27]. This is crucial for air traffic safety and air traffic controller deployment. It has also been used to forecast air traffic flow at Hong Kong International Airport, obtaining good short-term predictions [28].

2.4. Explainability Methods

These methods have been used to better understand how the developed models make the predictions. Recently, there has been a lot of interest in explaining the outputs of machine learning algorithms. To address these needs, new software has been developed [29].

In this work, explainability methods have been used to analyse the patterns found by the models. The aim is to better understand how traffic affects the workload of the controllers. Once the models have been trained, different evaluation methods have been used to study their performance and draw conclusions about the results. ELI5 has been used to find out which variables are most important in the performance of the model. SHAP has been used to see what patterns the machine has learned and how it calculates the predictions. Confusion matrices have been used in the evaluation of all classification algorithms.

2.4.1. ELI5

This library can examine a classification or regression model in two ways: Inspect the model parameters and try to find out how the model works globally, or inspect an individual prediction of a model and try to find out why the model makes the decision it does. For some classifiers, inspection and debugging is easy, for others it is difficult. ELI5 aims to handle not only simple cases, but even for simple cases to have a unified API (Application Programming Interface). In this way, ELI5 supports multiple models of all types and allows a common evaluation method for all of them.

2.4.2. SHAP (Shapley Additive Explanations)

This algorithm is a brilliant way to reverse engineer the results of any predictive algorithm [30]. SHAP values are used whenever you have a complex model (such as gradient boosting or random forest) and you want to understand what decisions the model is making. Predictive models answer how much. SHAP answers the why.

SHAP values are based on Shapley values, a concept from game theory. But game theory is based on two things: a game and players. The game is to reproduce the outcome of the model, and the players are the characteristics included in the model. What SHAP does is to quantify the contribution that each characteristic makes to the prediction made by the model. Shapley's values are based on the idea that the outcome of each possible combination (or coalition) of players should be considered in determining the importance of a single player. This is a fundamental characteristic of SHAP values: the sum of the SHAP values for each trait of a given observation yields the difference between the model prediction and the null model.

2.4.3. Confusion Matrix

In the field of machine learning, and specifically in the statistical classification problem, a confusion matrix is a specific table design that allows visualizing the performance of an algorithm, usually in supervised learning (in unsupervised learning it is often called a matching matrix). Each row of the matrix represents instances of an actual class, while each column represents instances of a predicted class, or vice versa. The name comes from the fact that it makes it easy to see if the system confuses two classes, i.e., if it tends to mislabel one as the other.

If the actual classification set is compared to the predicted classification set, there are 4 different results that could occur in any particular column.

First, if the actual classification is positive and the predicted classification is positive, this is called a true positive result because the positive sample was correctly identified by the classifier;

Second, if the actual classification is positive and the predicted classification is negative, this is a false negative result because the positive sample was incorrectly identified by the classifier as negative;

Third, if the actual classification is negative and the predicted classification is positive, this is a false positive result because the classifier incorrectly identifies the negative sample as positive;

Fourth, if the actual classification is negative and the predicted classification is negative, this is a true negative result because the classifier correctly identifies the negative sample.

3. Results

3.1. Exploratory Data Analysis

Before starting to apply machine learning to the data, the raw data were analysed to extract as much information as possible. Obtaining this information is useful as it allows for a better understanding of the problem and the ability to design better predictive models.

3.1.1. Flight Levels

By analysing the flight levels at which aircraft exit and enter the sectors, we have discovered that each sector has a distinctive profile (see Figure 1). The vast majority of flights in all sectors tend to be cruise flights. They cross the sector entering and leaving on the same flight level. But each sector has some differences that distinguish it from the others.

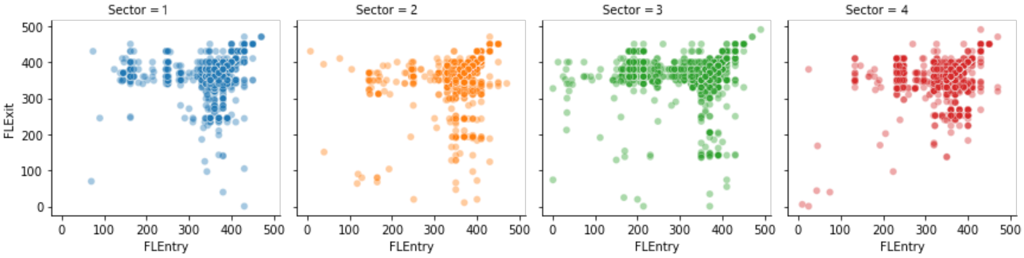

Figure 1. Entry and exit flight levels according to each sector.

Sector 1 is distinguished from the rest by the fact that almost all of its flights are cruise flights. Most aircraft cross the sector entering and exiting between FL 300 and FL 400.

Sector 2 is notorious for having a significant flow of descending flights. These flights typically enter at cruising altitudes and depart between FL 100 and FL 300.

Sector 3 is the opposite; it has a flow of ascending flights. These flights enter between FL 50 and FL 300 and usually leave the sector at cruising altitudes.

Sector 4 is a mixture of the above. It has both ascending and descending flows, but these are smaller. These flights are typically between FL 250 and FL 300, closer to typical cruising altitudes.

The trends explained above are more clearly seen in one of the new variables created, FLDelta. See Section 3.2.1.

We analyse the entry and exit flight levels according to the flight trend (cruise or evolution) in the Figure 2. Almost all cruise flights enter and leave the sectors between FL 340 and FL 450. Evolution flights usually enter and leave via two bands. The first coincides with the cruise flight slot. The second is between FL 180 and FL 250.

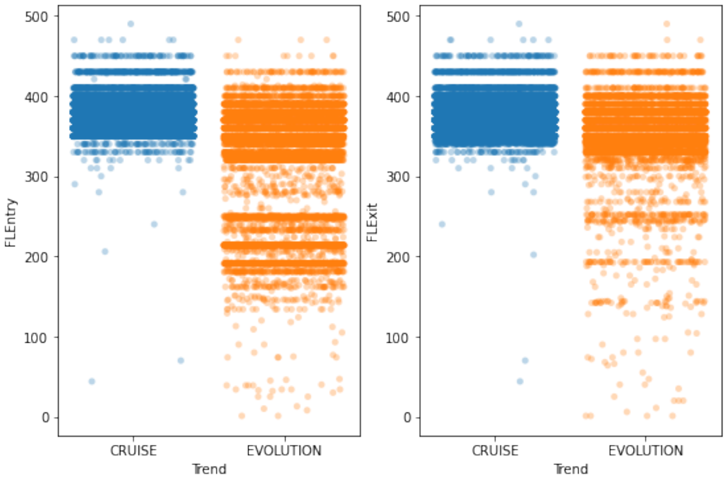

Figure 2. Representation of the entry and exit flight levels depending on the flight trend.

3.1.2. Events

Events are what we call the number of actions taken by controllers on flights. There are different possible actions with variable complexity, but events are just the number of total actions taken. Each action contributes the same to the sum, despite having different complexity. All the possible actions that a controller can take and that can be recorded are shown in Table 1.

Table 1. IDs of each possible ATCo action.

Action ID	Description
CTE	Confirmation of contact. Aircraft enters with radar coverage
CTET	Confirmation of contact. Aircraft enters without radar coverage
CTET31	Physical entry into the sector of a shared-flow flight or flight of interest
CS	Flight transfer
CTE32	Physical exit of the sector of a shared-flow flight or flight of interest
Ac5	Instruction (suggestion in VFR) to change level or ROC/ROD
Ac6	Horizontal speed variation required by the aircraft
Ac7	Approach clearance
Ac8	Direct route clearance to a point to shorten a flight plan
Ac9	Provide relevant information or information on VFR intentions
Ac10	Provide traffic information
Ac11	Flight rule change
Ac12	SSR transponder code change
Ac13	STAR Assignment / Entry or Exit Point Confirmation Free Route
S2	Course instruction (VFR course suggestion) for separation or sequence
S3	Diversions caused by storm areas
X1	Vector guidance instruction (VFR suggestion) for sequence or procedure
A1	Instruction (VFR suggestion) to change level for sequence or separation
A2	Horizontal speed settings for separation or sequence
A3	Direct route authorisation for separation or sequencing.
A4	Stand-by instruction
A6	Separation via non-approval of request
H1	Entry of an aircraft into a holding area
Co1	Coordination with non-operational offices and units
Co2	Receiving and transmitting
Co3	Internal coordination
Co5	Verbal issuing /receiving of estimates and exchange of general information
Y1	Creation of a new flight plan
Y2	Modification of a flight plan
Y3	Gather necessary information without verbal coordination.
Mo1	Monitoring of specific situations
Mo2	Search for interactions with other traffic

Each flight receives on average 2.26 controller actions. Most flights receive only one or two actions. Flights that receive six or more actions are considered to be outliers; they are quite rare.

The number of actions received varies depending on whether the flights are cruise or evolution flights (see Figure 3). Almost 80% of cruise flights receive only one or two actions. In contrast, most evolution flights (those entering and exiting at different flight levels) receive more than two actions.

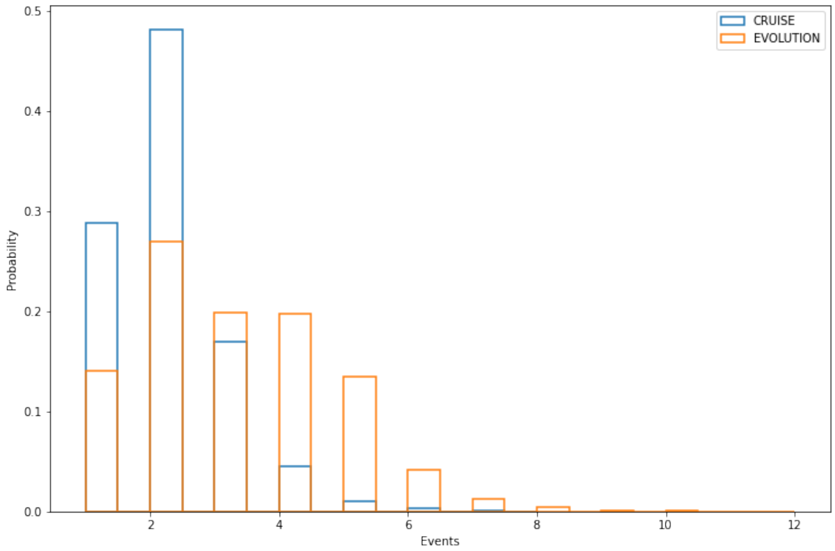

Figure 3. Distributions of the number of actions for each trend.

3.1.3. ATCo Actions

We analysed all actions registered by the ATOM programme. In total there are 140,000 actions spread over the four sectors. Figure 4 shows the number of actions taken in each sector classified by the action type. It is noteworthy that Sector 4 has the second-highest number of actions, while it has the lowest number of flights. Its average number of events is 2.5, higher than the average of all the sectors combined. There are many types of actions, but they can be divided into two broad categories.

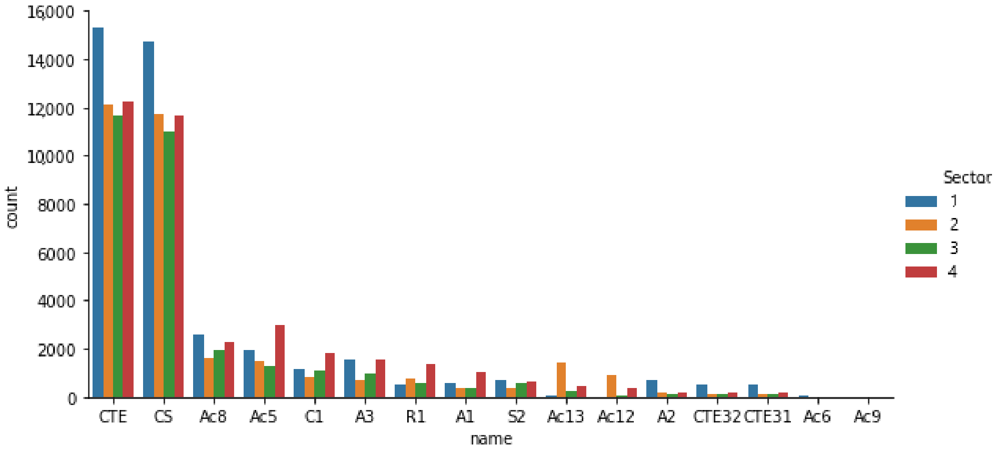

Figure 4. Actions taken by controllers according to sector.

Confirmation actions (CTE and CS, see Figure 4) comprise contact confirmations when an aircraft enters the sector and flight transfers when exiting. They constitute the bulk of the actions, approximately 100,000 of the total number. They involve little workload for the controller, and at most there can be two for the same flight. This is why they are the least important for studying the controller's workload.

The actions that we are going to call complex are the rest. Complex actions are usually changes of course (Ac8 and A3) or altitude (Ac5, A1, C1 and R1) by controllers. They involve a change in the trajectory of the aircraft. These actions do have an impact on the workload and are the focus of this study. Critical flights will be those with a high number of this type of action.

3.2. Feature Engineering

After the exploratory data analysis, we tried to create new variables from the raw data we had. The objective of feature engineering is simplifying and speeding up data transformations while also enhancing model accuracy. It achieves that by creating new variables that are not in the training set. At the same time, a prototype of a predictor model was made using random forest to measure the impact of the variables created on the effectiveness of the model. The aim was to obtain new variables that would make it easier for artificial intelligence to interpret the data to obtain more accurate models. Two variables were found that significantly improved the accuracy (how many times the ML model was correct overall) of the models.

3.2.1. FLDelta

It is obtained by making the difference between the exit flight level and the entry flight level. The formula is as follows: $\Delta_{FL} = FL_f - FL_0$. Thus, we have directly how much each flight has climbed or descended during its transit through the sector. This makes it easier for us and the AI to interpret flight level changes.

Figure 5 provides further information. Of the 55,000 flights, 40,000 are cruise flights ($FLDelta = 0$). Even among evolution flights, small changes are most common. Large changes in flight level ($FLDelta > 100$) are rare. Although Sector 3 has a significant number of flights that climb near 200 FLDelta.

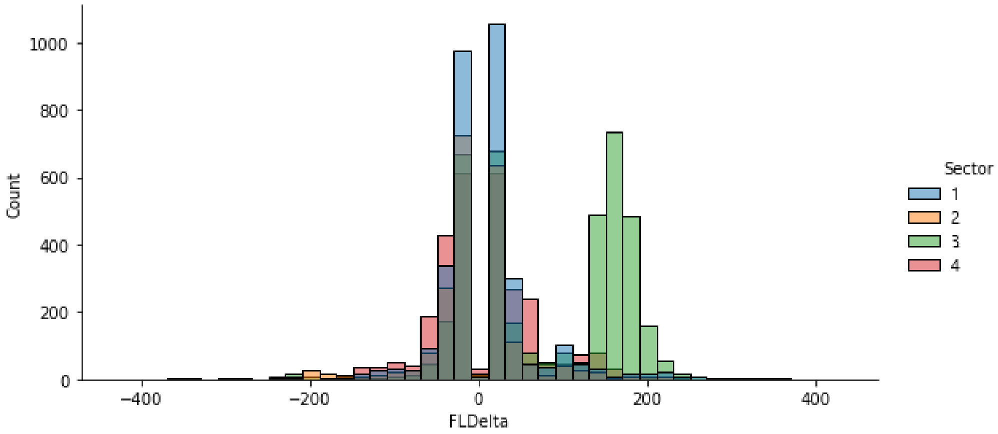

Figure 5. FLDelta of evolution flights by sector.

3.2.2. TDelta

It is the duration of the flight in the sector in seconds and is obtained from the sector entry and exit time. The formula is as follows: $\Delta_T = T_f - T_0$. Figure 6 is a boxplot of TDelta and shows the following information. The average flight duration in the sector is around 600 s. There are flights with extreme values up to 3500 s, but values above 1500 s are considered outliers.

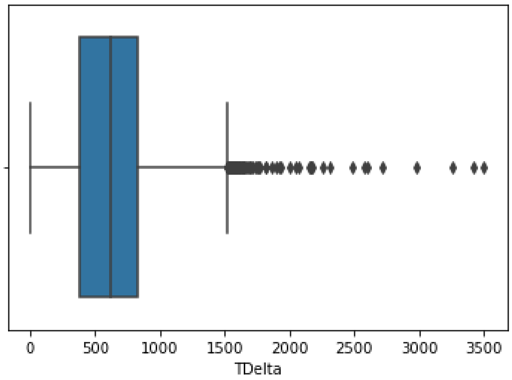

Figure 6. TDelta boxplot.

Each sector has its own characteristic layout, as shown in Figure 7. Sector 3 is characterised by longer flights, while Sector 2 is characterised by shorter ones.

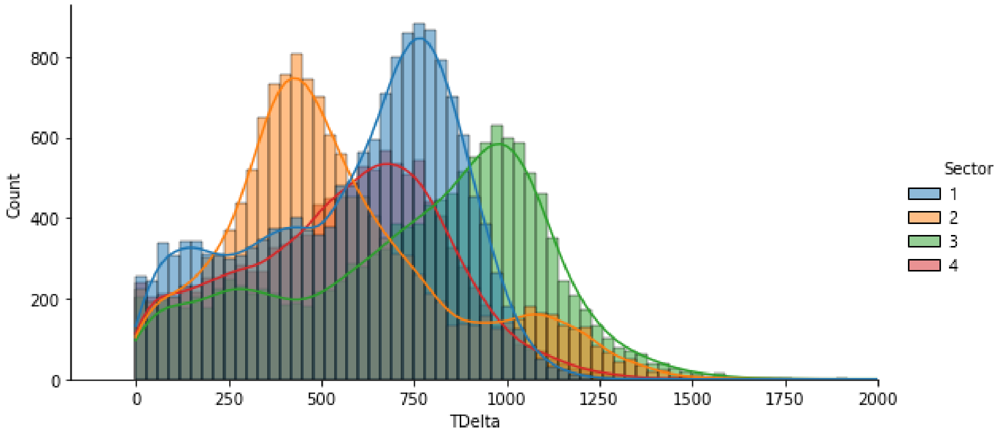

Figure 7. TDelta by sector.

3.3. Machine Learning Models

Five models were developed to test different combinations of variables and algorithms in an iterative process. All five of them are tested in Sector 1, and the results obtained for this sector are compared. Of all these models, two were finally chosen. In all models we choose to split randomly the data into training at testing datasets. The training set contains 80% of the data and the test set the remaining 20%.

Model 1: It is the most basic model. Its purpose is to serve as a basis for comparing the other models. Its algorithm is RandomForestRegressor and uses flight levels, flight trend and characteristic flows as variables. It predicts the number of actions a flight will receive. It has a relative error of 42%, which is quite high.

Model 2: All variables available in the data are added, including the newly created variables FLDelta and TDelta. This brings the relative error down to 34%, which is a big improvement. This means that the new variables that have been created are useful for the model to make better predictions. The flight duration in the sector and the change in altitude are related to the controller’s actions.

Model 3: Confirmation events, which have the lowest workload, are removed. The aim is to focus only on complex events and try to predict them. The number of data is reduced too much as many flights run out of events received. When a flight has not received any event when crossing the data, it does not appear in the final dataset. This is why the relative error increases at 39%. This model is discarded as it performs worse than the previous one.

Model 4: Unlike the previous model, all flights are retained. Instead of removing the confirmation events, these are still counted, although they are worth 0. Thus, if a flight has only confirmation events, it will appear in the final dataset but with 0 events. A slight improvement is obtained in the Mean Absolute Error (MAE), so it outperforms model 3. The relative error of this model tends to infinity because of the large number of zeros, due to the fact that many flights have 0 events. This does not allow this model to be compared with the relative errors of the rest.

Model 5: In this model, the algorithm is changed to a RandomForestClassifier. This is a classification model. First, the flights are divided into two types: those with 0 events are low load, the rest are high load. The low load are the flights that only have confirmation events, and the high load are the flights that have some complex events. The metric to measure this model is accuracy, the number of correct predictions. Unlike the others, this model does not predict a numerical result. The accuracy is 73%, which is a good performance that seems to indicate that the classifier approach is interesting for the problem addressed.

Eventually, the models chosen are 2 and 5. We have chosen model 2 as the best regression model of the four tested. Model 5 was chosen because of the models that focus on high load flights, the classification model performs best. In addition, in these last two models, the RandomForest algorithms have been replaced by their XGBoost equivalents. This change has been made because of the ease of training and tuning the models and the saving of computational time. The performance (relative error and accuracy) of the models improved, but not in any significant way.

3.4. Results of Selected Models

Once the two best models have been chosen, they are applied to the four sectors under study. The results obtained are a combination of the usual metrics of the models and the analysis with the explanatory methods. Explanatory methods make it possible to inspect the parameters of the model and to try to find out how the model works globally, and to inspect an individual prediction of a model and try to find out why the model makes the decision it does.

3.4.1. Regression Model

Each sector has been trained separately, so there are actually four different models. Each model is trained separately, as there may be patterns specific to each sector that do not appear in the others. Also, the flows are different for each sector, flow 1 in Sector 1 has nothing to do with flow 1 in Sector 2. Table 2 shows the errors obtained for each sector. The newly calculated sectors give better results than Sector 1, which is the one used in the study.

Table 2. Regression model results.

Sector	1	2	3	4
MAE	0.62302	0.57554	0.59438	0.66385
RMSE	0.85607	0.80965	0.85798	0.86883
MAPE	0.33598	0.29480	0.30877	0.33227
Average of events	2.1529	2.3415	2.1268	2.4967

The sector with the best predictions is Sector 2, which has a lower relative error (MAPE) than Sector 30. Regarding Sector 3, predictions also improve. In Sector 4, the MAE and RMSE worsen, but the relative error is lower, as it is the sector with the highest average number of events per flight.

Figure 8 shows the variables that are most important and influential for the model. The most important variable is TDelta, followed by the trend. The next most important variable is FLExit. The model gives more importance to the outbound flight level than to the inbound flight level. Flows have less importance because their overall impact is smaller, but they are nevertheless very influential. The output of the flights they affect changes a lot. FLDelta is not one of the best predictors.

Figure 9 shows that the longer the flight stays in the sector, the higher the number of estimated actions. In addition, if the flights are ascending or descending, the slope becomes steeper. This means that the most critical flights are those that spend the longest time in the sector, especially if they are ascending or descending flights.

Figure 10 shows that the higher the absolute value of FLDelta, the higher the number of actions. Flights are more critical the larger the climb or descent.

Figures 11 and 12 show that the number of estimated actions increases significantly when exiting or entering the sector below 350. The most critical flights are ascending or descending flights entering or leaving the sector at lower altitudes.

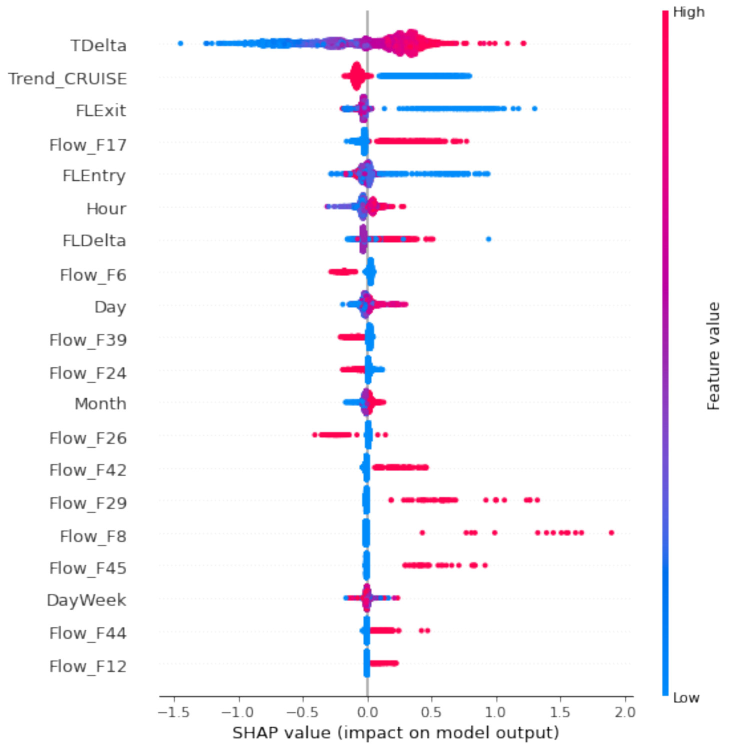

Figure 8. SHAP values of the regression model for Sector 1.

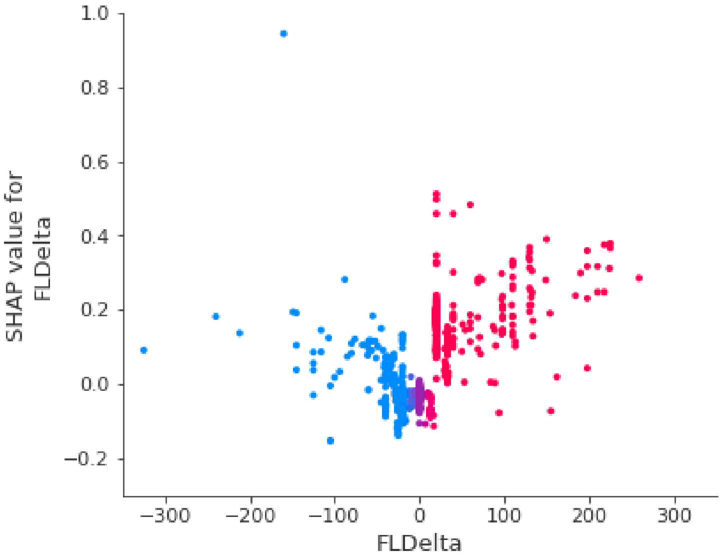

Figure 9. SHAP values (regression) of FLDelta for Sector 1.

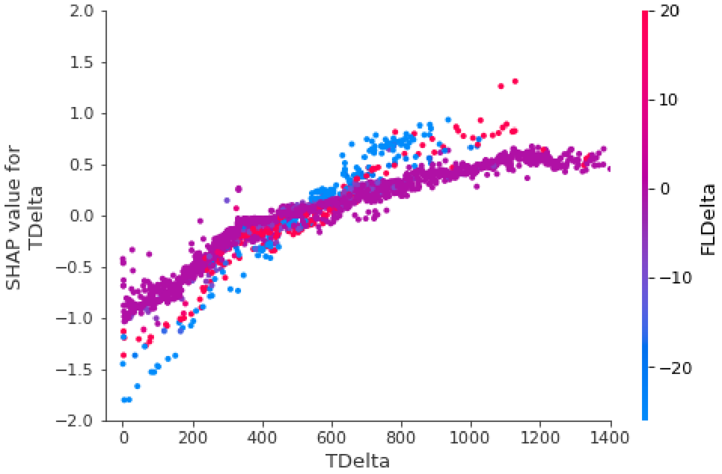

Figure 10. SHAP values (regression) of TDelta according to FLDelta for Sector 2.

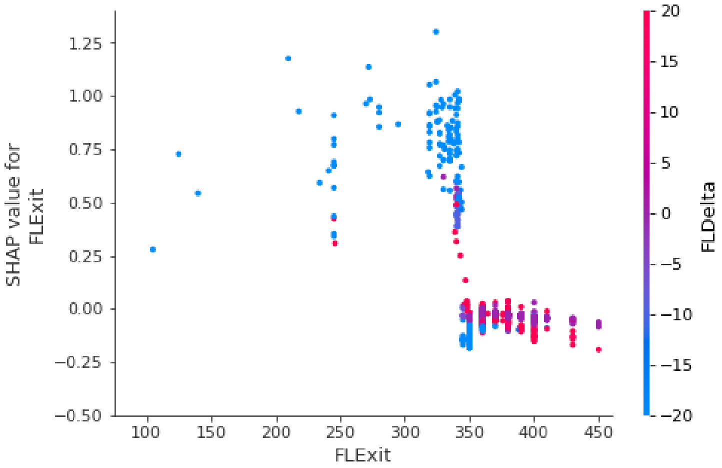

Figure 11. SHAP values (regression) of FLExit according to FLDelta for Sector 1.

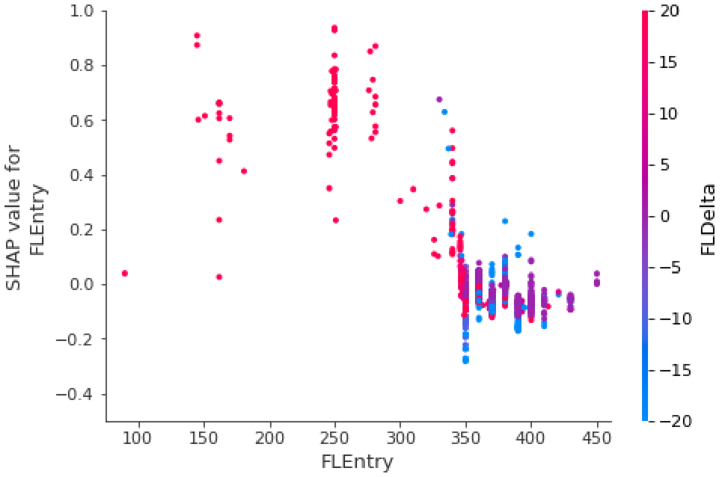

Figure 12. SHAP values (regression) of FLEntry according to FLDelta for Sector 1.

3.4.2. Classification Model

As in the regression model, a separate model has been trained for each sector.

Table 3 shows the results of the model. The metric against which they are compared is accuracy. This is the percentage of flights that have been correctly assigned to their category. The accuracy value in all sectors is good because it exceeds 70%. It varies slightly depending on the sector, with the worst accuracy in Sector 1 at 73% and the best in Sector 2 at 75%.

Table 3. Classification model results.

Sector	1	2	3	4
Accuracy	0.73070	0.75323	0.74863	0.73114
Average of events	0.56594	0.63311	0.52803	0.870353

The confusion matrices indicating which predictions were correct in each category can be seen in Figure 13. The sum of the percentages of the main diagonal gives the accuracy of each sector, as it is the sum of the correct data. The advantage is that they can now be looked at separately, in addition to the failed predictions. There are 4 possible situations:

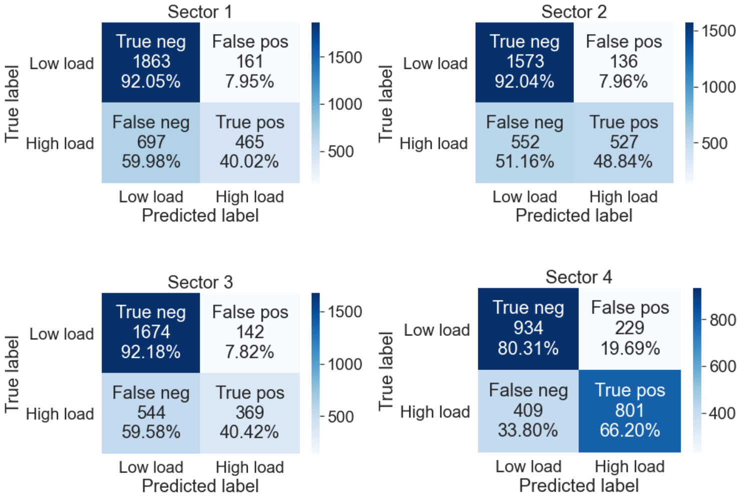

Figure 13. Confusion matrices for each sector.

The flight is low load and the model predicts it correctly. A high number is interesting because the model manages to detect many low load flights correctly. By identifying them in advance, a decision can be made to allocate fewer resources to these flights and leave them reserved for more demanding flights. In all four sectors, it is the largest set. The vast majority of low load flights are correctly identified. In the first three the accuracy is above 90% and in the fourth at 80%.

The flight is low load and the model predicts that it is high load. This set-up is problematic, as it results in a misuse of resources. Low load flights are treated as if they were to be high load. This group is the smallest of the four in all sectors. In Sector 1, 2 and 3, only 8% of low load flights are misidentified. In Sector 4, this number reaches 20%.

The flight is high load and the model predicts that it is low load. This set is the most critical, as it is believed that it is a flight that will not increase the workload much, when in fact it will. It is the most important set because if it is too large, it could lead to an underestimation of the workload and the sector may have to be regulated. Unfortunately, more than 50% of the high load flights are misclassified in the first three sectors. In contrast, in Sector 4, this number drops to 33%, making it the sector where the model works best.

The flight is high load and the model predicts that it is high load. The higher this number is, the more efficiently resources are allocated to the flights that require them most. More resources can be allocated in several ways. A sector can be split into more sectors, allocating controller teams to each new one. Another way is not to change the sector but to increase the rotation of controllers so they can have longer rest periods. Of particular note is Sector 4 where 66% of high load flights are correctly identified.

It is worth highlighting the fact that the model improves considerably in Sector 4. This is the sector with the highest number of critical flights and the highest average number of events. It is very interesting to note that the model works best in the most problematic sectors where it is most needed.

The interest in analysing the SHAP values now is to point out the differences with respect to the regression model. Some differences will be found because in this model the best variables that serve to differentiate high and low load flights appear. What each model evaluates is slightly different.

Figure 14 shows that flows have gained importance (F6) and are still very influential (F29) with respect to the regression model. In contrast, TDelta has lost importance. It seems that geometry has become more important in distinguishing between the two types of flights.

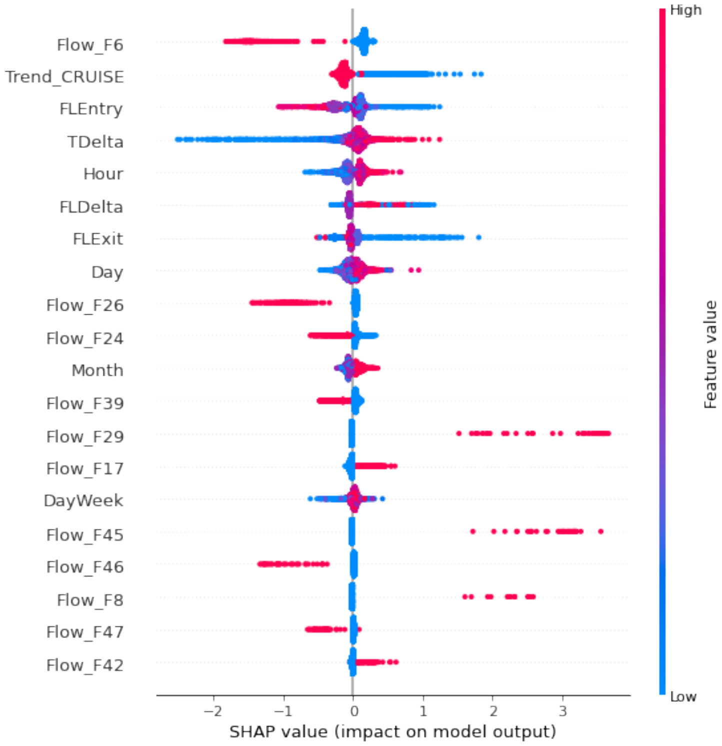

Figure 14. SHAP values of the classification model for Sector 1.

Figure 15 shows that TDelta is less important than in the regression model. As for the slopes, the slope of the evolutionary flight remains steeper.

Figure 15 shows that FLDelta has a greater impact than in the regression model. The conclusion is that this variable is better related to more complex actions.

Figures 16 and 17 show that there are no significant changes with regard to entry and exit flight levels. The importance of the two variables increases slightly. The probability of a flight being a high load flight is high if it enters or departs at a flight level lower than 350.

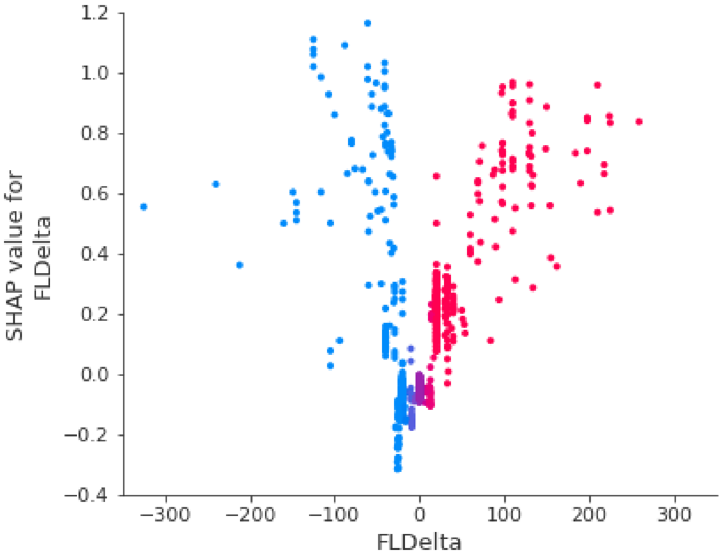

Figure 15. SHAP values (classification) of FLDelta for Sector 1.

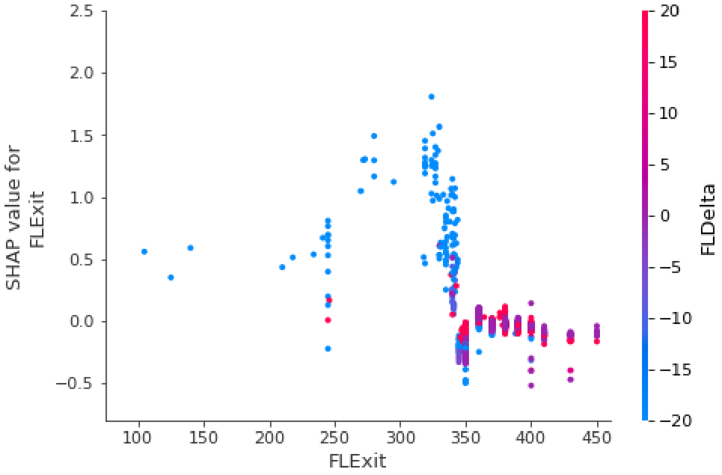

Figure 16. SHAP values (classification) of FLExit according to FLDelta for Sector 1.

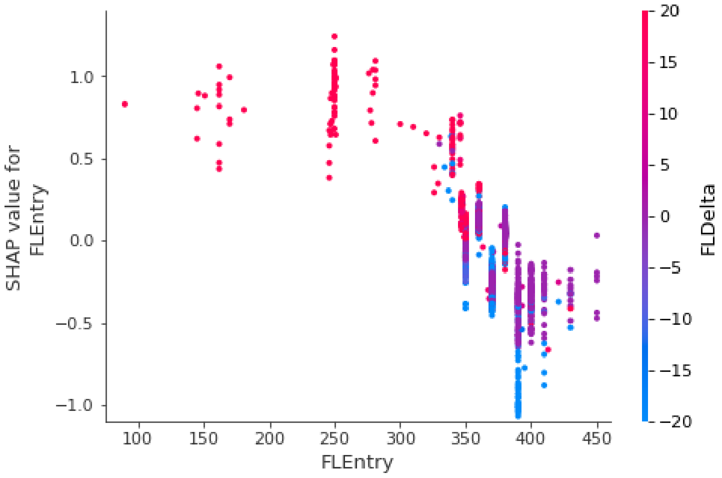

Figure 17. SHAP values (classification) of FLEntry according to FLDelta for Sector 1.

3.4.3. General Approach

Instead of training one model per sector, a single model is trained by pooling data from all sectors. The aim is to see how important local patterns are. Traffic flows are a variable that only makes sense within sectors. It cannot be used in this model. This implies a drop in performance as the flows give information on the trajectory that the aircraft has followed. It is expected to compensate for the loss of performance by increasing the amount of data available for training. This happens by pooling data from all sectors. This method is applied to the two models chosen: the regression model and the classification model.

Tables 4 and 5 show the results obtained by applying the general approach to the two chosen models.

Table 4. Results of the general regression model.

Sector	1	2	3	4	General
MAE	0.62302	0.57554	0.59438	0.66385	**0.67875**
RMSE	0.85607	0.80965	0.85798	0.86883	**0.93156**
MAPE	0.33598	0.29480	0.30877	0.33227	**0.36525**
Average of events	2.1529	2.3415	2.1268	2.4967	**2.2680**

Table 5. Results of the general classification model.

Sector	1	2	3	4	General
Accuracy	0.73070	0.75323	0.74863	0.73114	**0.71043**
Average of events	0.56594	0.63311	0.52803	0.870353	**0.63872**

The performance, being of the same order of magnitude, is the worst seen so far. The estimates provided by the general model are slightly less accurate.

It seems that the model has been able to learn a large part of the patterns, as its results are close to those of the other models. But the lack of the geometric information of the characteristic flows seems to penalise it. Local patterns and flows allow for fine-tuning the models to further improve their performance.

4. Conclusions and Future Lines of Research

4.1. Conclusions

This article has sought to gain a better understanding of the relationship between air traffic and the actions of air traffic controllers. The actions of air traffic controllers are an approximation of their workload. Multiple insights have been gained through exploratory data analysis. Some of the most important variables found are the trend of flights and their entry and exit levels. Other important variables that were not initially in the data have been found, thanks to feature engineering. These variables are the time delta and the flight level delta when crossing the sector.

The other major goal of this paper was to use machine learning to create models that were able to predict the actions of controllers in a sector. The models that have been obtained are a good first step. Their accuracy does not allow them to be used for day-to-day operations management. But they could be useful to obtain first approximations when estimating the workload of a new sector. In addition, once the models have been trained, explanatory techniques have been used to gain more insight into the relationship between traffic and workload. In this way, it is possible to know which are the most critical flights according to the artificial intelligence.

The predictions of the number of actions of the regression model have errors close to 30 percent. They are not accurate, but are useful as a first approximation to the complexity of a flight. The model has picked up certain patterns that, when they occur, estimate a higher number of actions. The flights that the model considers most critical are flights with a high sector stay, evolution flights (especially descending flights), flights with a departure flight level of less than 350 and flights with high flight level deltas.

The flight type predictions of the classification model are good, with predictions as high as 75 per cent. Although it also has weaknesses, the model has more difficulty in correctly predicting high load flights. In this model, the characteristic flow is more important and the flight level delta is more important, while the flight duration in the sector is less important.

In both the regression model and the classification model, characteristic flows of great interest have been located. In some sectors, there are flows that are good predictors and that are closely related to the most critical flights. The geographical dimension is critical.

It is important to highlight that exploratory data analysis has been essential for designing the models, as it allows us to understand the problem and the relationships between variables.

The two new variables created have been the action that has improved the models the most. It is worth investing time in thinking about new variables to extend the information available in the models. TDelta correlates well with the number of actions, and FLDelta with the presence of more complex actions.

Although the results of this work are not as good as one might wish, they mark a promising path. In the near future, when AI algorithms and the quality and quantity of data continue to improve, we are confident that appropriate models will be developed for use in industry.

4.2. Future Lines of Research

The research work of this project can be extended in the following directions in the future:

Using unsupervised learning techniques. The aim of the work has been to develop models to be able to predict metrics related to the workload of a flight for a controller. This has meant that supervised machine learning systems have been chosen. It would be interesting to analyse the data by applying unsupervised learning systems. Some unsupervised learning techniques that would be interesting to apply are clustering, visualisation and dimensionality reduction and association rule learning. The goal would be to use these techniques to obtain more information from the data and to better understand the problem. A more in-depth understanding of the patterns that exist between flights and workload would lead to better predictive models;

Improving the quality of the data, especially the geometrical part. During this work, many problems have been encountered in the assignment of flights to a characteristic flow of the sector. Sometimes the definition of these flows was inconsistent. There were also problems in assigning each flight to a flow. For almost half of the flights, the exact flow was not found, and an approximation had to be assigned. In the machine learning models developed, it has been found that the flow of a flight is significant in predicting its workload. Improving these data could lead to a great improvement in the predictions made by the models;

Getting data from sectors other than those already studied. In the search to improve the machine learning model, it is important to adapt it to different contexts to make it as general as possible. Even if a model is capable of making excellent predictions in one sector, it is not as useful if it cannot make them in other sectors while maintaining its performance.

Author Contributions: Conceptualization, G.G.T. and V.F.G.C.; methodology, G.G.T. and R.M.A.V.; software, G.G.T.; formal analysis, G.G.T. and R.M.A.V.; investigation, G.G.T.; resources, R.R.R.; writing—original draft preparation, G.G.T.; writing—review and editing, R.M.A.V.; visualization, G.G.T.; supervision, P.M.L.d.F.; project administration, P.M.L.d.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data were provided by ENAIRE (Spain's ANSP) and is not publicly available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, Y.; Hu, R.; Lin, S.; Schultz, M.; Delahaye, D. The Impact of Automation on Air Traffic Controller's Behaviors. *Aerospace* **2021**, *8*, 260. [CrossRef]
2. Metzger, U.; Parasuraman, R. Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. In *Decision Making in Aviation*; Routledge: Oxfordshire, UK, 2017; pp. 345–360.
3. Rodríguez, S.; Sánchez, L.; López, P.; Cañas, J.J. Pupillometry to Assess Air Traffic Controller Workload through the Mental Workload Model. In Proceedings of the 5th International Conference on Application and Theory of Automation in Command and Control Systems (ATACCS '15), Toulouse, France, 30 September–2 October 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 95–104. [CrossRef]
4. Suarez, N.; López, P.; Puntero, E.; Rodriguez, S. Quantifying air traffic controller mental workload. In Proceedings of the Fourth SESAR Innovation Days, Madrid, Spain, 25–27 November 2014.
5. Hilburn, B. Cognitive complexity in air traffic control: A literature review. *EEC Note* **2004**, *4*, 1–80.
6. Lamoureux, T. The influence of aircraft proximity data on the subjective mental workload of controllers in the air traffic control task. *Ergonomics* **1999**, *42*, 1482–1491. [CrossRef] [PubMed]
7. Aricò, P.; Borghini, G.; Di Flumeri, G.; Colosimo, A.; Pozzi, S.; Babiloni, F. A passive brain–computer interface application for the mental workload assessment on professional air traffic controllers during realistic air traffic control tasks. *Prog. Brain Res.* **2016**, *228*, 295–328. [CrossRef] [PubMed]
8. Tao, D.; Tan, H.; Wang, H.; Zhang, X.; Qu, X.; Zhang, T. A Systematic Review of Physiological Measures of Mental Workload. *Int. J. Environ. Res. Public Health* **2019**, *16*, 2716. [CrossRef] [PubMed]
9. Pagnotta, M.; Jacobs, D.M.; de Frutos, P.L.; Rodríguez, R.; Ibáñez-Gijón, J.; Travieso, D. Task difficulty and physiological measures of mental workload in air traffic control: A scoping review. *Ergonomics* **2021**, *65*, 1–24. [CrossRef] [PubMed]
10. Loft, S.; Sanderson, P.; Neal, A.; Mooij, M. Modeling and Predicting Mental Workload in En Route Air Traffic Control: Critical Review and Broader Implications. *Hum. Factors* **2007**, *49*, 376–399. [CrossRef] [PubMed]
11. Pham, D.T. Machine Learning-Based Flight Trajectories Prediction and Air Traffic Conflict Resolution Advisory. Ph.D. Thesis, PSL Research University, Paris, France, 2019.
12. Sanchez Hernandez, C.; Ayo, S.; Panagiotakopoulos, D. An Explainable Artificial Intelligence (xAI) Framework for Improving Trust in Automated ATM Tools; An Explainable Artificial Intelligence (xAI) Framework for Improving Trust in Automated ATM Tools. In Proceedings of the 2021 IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), Portsmouth, VA, USA, 3–7 October 2021. [CrossRef]
13. Xie, Y.; Pongsakornsathien, N.; Gardi, A.; Sabatini, R. Explanation of Machine-Learning Solutions in Air-Traffic Management. *Aerospace* **2021**, *8*, 224. [CrossRef]
14. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2019; p. 851.
15. Bosson, C.S.; Nikoleris, T. Supervised learning applied to air traffic trajectory classification. In Proceedings of the AIAA Information Systems-AIAA Infotech at Aerospace, Kissimmee, FL, USA, 8–12 January 2018. [CrossRef]
16. Pham, D.T.; Alam, S.; Duong, V. An Air Traffic Controller Action Extraction-Prediction Model Using Machine Learning Approach *Complexity* **2020**, *2020*, 1659103. [CrossRef]
17. Antulov-Fantulin, B. Air Traffic Complexity Model Based on Air Traffic Controller Tasks. Ph.D. Thesis, Faculty of Transport and Traffic Sciences, University of Zagreb, Zagreb, Croatia, 2020.
18. Chandra, R. Competition and collaboration in cooperative coevolution of Elman recurrent neural networks for time-series prediction. *IEEE Trans. Neural Netw. Learn. Syst.* **2015**, *26*, 3123–3136. [CrossRef] [PubMed]
19. Liao, C.H.; Wen, C.H.P. SVM-based dynamic voltage prediction for online thermally constrained task scheduling in 3-D multicore processors. *IEEE Embed. Syst. Lett.* **2017**, *10*, 49–52. [CrossRef]
20. Gui, G.; Zhou, Z.; Wang, J.; Liu, F.; Sun, J. Machine Learning Aided Air Traffic Flow Analysis Based on Aviation Big Data. *IEEE Trans. Veh. Technol.* **2020**, *69*, 4817–4826. [CrossRef]
21. Le Fablec, Y.; Alliot, J.M. Using Neural Networks to Predict Aircraft Trajectories. In Proceedings of the IC-AI, Las Vegas, NV, USA, 28 June–1 July 1999; pp. 524–529.
22. Liaw, A.; Wiener, M. Classification and Regression by RandomForest. *Forest* **2001**, *23*, 18.
23. Rebollo, J.J.; Balakrishnan, H. Characterization and prediction of air traffic delays. *Transp. Res. Part Emerg. Technol.* **2014**, *44*, 231–241. [CrossRef]
24. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K. Xgboost: Extreme Gradient Boosting; R Package Version 0.4-2. 2015. Available online: https://cran.r-project.org/ (accessed on 1 July 2022).

25. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
26. Li, P. Robust logitboost and adaptive base class (abc) logitboost. *arXiv* **2012**, arXiv:1203.3491.
27. Zhang, M.; Xie, H.; Ge, J.; Zhang, D. Air traffic complexity evaluation with novel complexity features and mRMR-XGBoost. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *638*, 012036. [CrossRef]
28. Hon, K.k. Artificial intelligence prediction of air traffic flow rate at the Hong Kong International Airport. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *865*, 012051. [CrossRef]
29. Arya, V.; Bellamy, R.K.; Chen, P.Y.; Dhurandhar, A.; Hind, M.; Hoffman, S.C.; Houde, S.; Liao, Q.V.; Luss, R.; Mojsilović, A.; et al. AI Explainability 360 Toolkit. *ACM Int. Conf. Proc. Ser.* **2020**, *20*, 376–379. [CrossRef]
30. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.