Prognostic and risk of failure events using machine learning: An analysis based on onboard aircraft messages

João Francisco Dos Reis Martins Rodrigues
joao.f.r.m.rodrigues@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

October 2019

Abstract

Airline companies have been making a great effort to find ways to optimise maintenance processes in order to maintain high safety standards. The application of data-driven methods to maintenance has been introduced as a breakthrough in aeronautics, due to benefits in cost reduction and safety increase. Over the years, data-driven prognostics has become an important area of study, complementary to the still-dominant strategy in aeronautics, preventive maintenance. These new methods allow maintenance personnel and process engineers to take a proactive instead of a reactive approach to failures, where failures are anticipated and eliminated before they occur. In view of this, there has been a growing concern in the maintenance sector to find indicators or precursors of failures using machine learning and artificial intelligence. This paper reports on the work carried out on the evaluation of the Prognostics and Health Management (PHM) capabilities of the Central Maintenance Computer (CMC) messages. This is achieved by comparing different types of models, varying several properties of the data sets allowing to relate differences in the results to differences in the characteristics of the models. Using a real data set from Portugália Airlines, this study focuses not only on the prediction of the remaining useful life (RUL) of the equipment but also on the prediction of the urgency of an intervention at a given time. The results show that message data associated with the applied machine learning techniques have predictive failure capabilities, aiding the trigger of unplanned maintenance actions.

Keywords: Prognostics, Machine Learning, Predictive Maintenance, Aeronautics, Central Maintenance Computer Messages.

1. Introduction

The world is becoming largely more dependant of machines and systems that are crucial in the present days of the human being's usual daily routine. The level of maintenance required to preserve the functionality and life of an equipment depends largely on how complex the running system is. An aircraft is a result of the continuous interaction of several highly complex systems, that allow the machine to provide the flying capabilities that human could only dream about a century ago. Capable of flying upwards of 40 000 feet, at speeds close to the speed of sound, over its long lifespan of, in some cases, more than 100 000 flight cycles, the commercial aircraft is one of the more complex equipment created and developed by the human being.

Nowadays, the commercial aircraft is a highly instrumented machine. Given the increasing tendency of the amount of data being generated by the global fleet, the major players in the aircraft industry, namely the Original Equipment Manufacturers (OEM), operators and Maintenance, Repair and Overhaul (MRO) companies are trying to use innovative ways to take advantage of the level of instrumentation present in the modern day aircraft, investing mainly in Aircraft Health Monitoring and Predictive maintenance systems. Although health monitoring systems represent an important complement to the maintenance and troubleshooting actions of an aircraft operation, its primary goal is to identify changes that may indicate damage or imminent failure. These systems have some data interpretation capabilities, mainly based on the occurrence rate of certain predefined important failures. However, a significant part of the data interpretation rests on the knowledge of engineers and technicians, being generally the software itself incapable of predicting a failure occurrence. That's where predictive maintenance role lies, as with the significant amounts of data available, the main goal of this technique is to develop data-based or physical degradation based models that may interpret aircraft health monitoring parameters and predict when a failure might occur.

2. Background

2.1 Predictive Maintenance and Prognostics

"Most machine maintenance today is either purely reactive (fixing or replacing equipment after it fails) or blindly proactive (assuming a certain level of performance degradation, with no input from the machinery itself, and servicing equipment on a routine scheduled whether service is actually needed or not" [1]. These scenarios are extremely wasteful considering the level of instrumentation present in the modern days machinery. That is the reason why the maintenance world is adopting and moving forwards using new technologies and adopting the "predict and prevent" maintenance [1].

Predictive maintenance is based on the policy of only applying maintenance actions to the equipment once the magnitude of certain reliability indicators reach a predetermined level and lead to the possible imminent future failure. Using a combination of the available performance and diagnostic data, operation logs, or other available physical or digital data, the predictive maintenance uses a combination of human and technical skills to make decisions about maintenance procedures of certain equipment or systems [1]. Predictive maintenance is essentially "fitting a network of sensors to the aircraft or other equipment to measure condition signals" [2]. Those signals may be then used as condition monitoring variables that may be useful to decide the maintenance intervention to a specific item before it fails. This ability to schedule the intervention before a failure event is the main purpose of predictive maintenance [1].

Prognostics is the term used for the science of making predictions about engineering systems [3]. All the processes that aim to predict the future behaviour of systems are considered a form of prognostics. One of the main goals of prognostics is the estimation of the time at which the system or the components is no longer capable of performing its task. Making use of various indicators of vibration, temperature, lubricant condition, among others, that may be extracted from the highly complex sensor network present in today's most complex machines, such as the aircraft, prognostics aims to correlate the indicators to a possible future failure event, hence reducing the system's unpredictability. It may be stated that the application of prognostics in maintenance results on the predictive maintenance.

This work focuses on the data driven approach to prognostics. It takes advantage of large amounts of data available from historical records, both from normal and faulty operations. No priori knowledge of the process is required as it only develops models from measured data from the process itself. However, general knowledge of the system or process may be useful to interpret results. Prognostics rely on methods that can follow and analyse the trend in data, and forecast the next failure occurrence. Hence, machine learning is a very useful tool for prognostics. It is defined as a sub-field of computer science and artificial intelligence that explores the development of algorithms closely related to linear algebra, probability theory, statistics, and mathematical optimisation that can learn from data and make subsequent predictions. Machine learning allows data analysis otherwise not feasible with more conventional methods. It enables machines to learn by themselves based on provided data with the goal of making predictions [4].

2.2 Related Work

This subsection aims to describe three selected projects ([5], [6], [2]), with objectives that are related to this work. The comparison is focused mainly on the used data types, the general methodologies, the algorithms implemented, and the concluding remarks. The three works aimed the prediction of the health state of a system yet, the methodologies differ, and it is in this subsection presented a brief overview.

Starting with the data types, the authors of [5] used "all the available parameters that could be related to a failure of the system/component under analysis". These included data both from the crash protected and non-protected flight recorders. Basically, the data consisted in raw sensor data from the data recorders placed in the aircraft. Furthermore, the authors considered as failure indicative actions maintenance logs, such as replacements, cleanings and adjustment actions. Instead of using raw sensor data as the previous publication, the authors of [6], used central maintenance system messages as variables. Finally, the last project here presented [2] uses a combination of the last two, analysing the evolution of the results using each data type separately or both in the same analysis.

Focusing first on the project [5], the authors studied the left-hand bleed valve unit. The parameters considered included the " bleed manifold pressure, temperature, the high-pressure compressor speed of the engine where the bleed is taken from (left engine)." These were transformed into 24 statistical features/variables associated with each flight. "To reduce the contribution of the flight profile on the bleed unit behaviour" only the stable cruise flight data from the flights with a minimum of 20 minutes of stable cruise flight phase were considered [5].

The study rested on the classification of the component's health. The authors defined the two classifications on whether the component is within 30

days of a failure event. The definition of failure resulted from the filtering of maintenance logs. Due to the lack of importance of some, the only considered and defined as failures were the ones resultant from a bleed replacement [5]. Post the introduction of the data into a Support Vector Machine (SVM) classifier, the authors decided to define a so-called degradation index that aimed to "smooth the effect" of the misclassifications resultant from the predictions.

Also using a classification machine learning approach based on the SVM algorithm, the authors of the second article [6] stated that the criteria used to identify whether the objectives were or weren't met relied on a Notice Period (NP), i.e., time in advance of a fault occurrence. The overall goal of this project consisted in alerting the user if the failure occurrence is predicted to be between 2 and 12 flight cycles prior to a particular reference time or flight [6].

These first two projects met the desired objectives. In [5], the authors stated that the results were good, especially considering the limited amount of data. No quantitative results were demonstrated in the article, however, the good behaviour of the degradation index confirmed the good results [5]. In the second publication [6], the authors concluded that the results were overall good, as the main goal of developing a model with precision above 50% was achieved [6].

Finally, the author of the project [2] used, unlike the previous, a mix of classical and deep learning algorithms, and compared the results obtained with and without the application of one technique called Kalman filtering. Furthermore, the machine learning algorithms were used to predict the remaining useful life of the equipment and not to predict if it is between a specified time range. Hence, machine learning in question was regressive, unlike in [5] and [6].

The author concluded that "the results confirm the intuition that is easier to extract information from sensory signals" than from message information. Also, the introduction of more advanced deep learning algorithms also contributed to the general decrease in prediction errors, despite the best result not reaching the target of a MAE (Mean Absolute Error) of 10 days. The author stated that "deep learning models present a promising alternative to traditional machine learning models, especially for precision near the potential failure" [2].

2.3 In Summary

The above three mentioned projects are useful to showcase the variety of prognostic approaches already applied in maintenance. the variety of data used to develop the prediction models demonstrates both the high level of instrumentation present in today's aircraft and also the overall uncertainty in the choice of the better data type to predict those failures.

Nothing is perfect, and these models are not an exception. Even if the models would be able to predict with almost perfect precision, the definition of failures used in the three projects, i.e., replacement maintenance logs, do not necessarily represent the end of the life of the equipment. Most replacements are performed, due to safety measures, according to human perception of health monitoring parameters, or even due to complaints. As a result, not all the replacement actions are close to the true end of life of the equipment, and therefore, as the models learn according to those maintenance actions, predictions may fail to identify the real end of life indicators, meaning that the ground truth may fail to be captured.

Prognostics using data science is an emerging field and is attracting the attention of many worldwide companies that seek new and innovative ways to increase safety, efficiency, improving maintenance scheduling, and avoiding all the costs and struggles of unforeseen failure events. The aeronautic sector is one of them, and there is already commercial software available, such as the PROGNOS developed by Air France and KLM [7].

3. Objectives and Methodology

CMC (Central Maintenance Computer) and CAS (Crew Alerting System) messages represent several megabytes of aircraft health reports. The aircraft airworthiness may be dependent on whether a fatal CAS message may or may not be emitted by the avionic system. Some messages represent a AOG (aircraft on ground) risk, and are capable of stopping the otherwise smooth operation on an airline company. Due to the high incidence of maintenance messages which represent the system's behaviour, this paper intends to answer the following industrial hypothesis: Do the maintenance messages have any predictive power over the crew alerting system's failure messages?; Is it possible to predict a future failure event based on the appearance of maintenance messages?.

This work's approach to clarify the predictive capabilities of the message data follows six main steps:

1. Define how the messages' predictive power will be quantified

2. Choose the evaluation measures and the desired results

3. Combine different pre-processing techniques and apply to the data

4. Train and optimise the selected machine learning algorithms

5. Evaluate the models' performance

6. Compare the different results from the different data sets and models built

The messages' predictive power is quantified by the application of five different types of models, varying several properties of the data set, allowing to relate differences in the results and differences in the characteristics of the models. This study applies regression and classification supervised machine learning to each of the different types of analysis. The first intend to predict the remaining useful life of the system, whereas the latter aim to classify the risk of failure as high or low. Besides, a baseline approach, which estimates the failure occurrence solely based on the failure data, is used as a comparison measure.

There are five main steps inherent to the adopted machine learning framework: Feature Construction and Transformation to the tabular form; Train-test data split; Data Scaling; Feature statistics; Machine Learning model construction and evaluation.

The first is explained in the section 4.2. The second consists in dividing the data into training and test data sets. When dealing with models treating multiple independent time-series data, to keep the unseen nature of the test data, the temporal dependencies may be as well divided. To train and test with the highest possible reality, the test and train data have to be divided chronologically [10], to prevent training on data concerning a failure that is also included in the test set. This work allocates 80% of the data into the training data, and the rest into the test data set.

The following steps consist in scaling the data, and in the application of the feature statistics to eliminate one of each pair of highly correlated variables. These concepts are explained in the sections 4.3 and 4.2 respectively.

Finally, the last step in the machine learning framework consists in creating and testing the models. This work uses five different algorithms for the regression problem, and four for the classification, implemented in python's library *sklearn*: From *sklearn.linear_model* - *LinearRegression*; from *sklearn.ensemble* - *RandomForestRegressor* and *RandomForestClassifier*; From *sklearn.neighbors* - *KNeighborsRegressor* and *KNeighborsClassifier*; From *sklearn.svm* - *SVR* and *SVC*; from *sklearn.ensemble* - *GradientBoostingRegressor* and *GradientBoostingClassifier*.

To optimise the validation and performance evaluation, a cross-validation methodology is employed. Due to the variance inherent to the models' results obtained from the numerous ways to split the data into training and test, the cross-validation intends to minimise this effect. The reasoning for the cross-validation application consists, on each iteration, defining one of the K data groups as the test set, and the other as the training set, iteratively running the machine learning process over the K groups of data. The end results come from the average over the K iterations. All the algorithms are trained and tested according to a 5 fold cross-validation method.

This case study analyses the full fault history database from the Portugália Airlines' fleet of 13 aircraft over the last 3 years. In total, the data set is composed with 1.6 million records of CMC messages. The analysis concerns the pneumatic subsystem of the Embraer E190 aircraft.

For the regression methods, the considered error quantifiers are the Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE), defined in the table 1 as:

Table 1: Performance metrics. T_{pred} stands for the predicted remaining useful life, T_{actual} for the observed value, n the number of observations, and i is the observation identifier.

Performance Measures Regression
$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(T_{pred_i} - T_{actual_i})^2}$
$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}\lvert T_{pred_i} - T_{actual_i}\rvert$

For the classification approach, the performance metrics are different. The evaluation is based on the number of true positives, false negatives, and false positives. These concepts are relevant and need to be defined before the evaluating parameters. Notice that the meaning of positive in this context is the cases when there is a "HIGH" risk of failure. The definitions are the following [8]:

- True Positive (TP) : The predicted and the actual result are both positive, or indicate that there is a high risk of failure.
- True Negative (TN) : The predicted and the actual result are both negative, or indicate that there is a low risk of failure.
- False Positive (FP): The predicted result suggests a positive result (high risk) but the actual result is in fact negative (low risk). This is also known as a type I error.
- False Negative (FN) : The predicted result suggests a negative result (low risk) but the actual result is in fact positive (high risk). This is also known as a type II error.

The evaluating metrics for the classification problem are showcased in the table 2

Table 2: Performance metrics for classification.

Performance Measures Classification
$\text{Precision} = \frac{TruePositives}{TruePositive + FalsePositive}$
$\text{Recall} = \frac{TruePositives}{TruePositive + FalseNegatives}$
$\text{Accuracy} = \frac{TruePositives + TrueNegatives}{\text{Total observations}}$
$\text{F1 score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$

In both the classification and regression problems, the main goal is to optimise all the evaluation measures, and therefore obtain the best performance out of the developed models.

4. Data Pre-Processing

The application of machine learning techniques requires a pre-processing phase. Especially for real-world data sets, the original format is often not compatible with most machine learning algorithms. Also, surplus and unreliable data only contribute to the contamination and the weakening of the models. Therefore, the main steps in data pre-processing allow extracting the best possible performance with the available data set.

4.1 Data Filtering and cleaning

This first step consists on filtering and cleaning the data. As the analysis only concerns the pneumatic subsystem, it is included in this step the filtering of the messages, only maintaining the data linked to the ATA (Air Transport Association) 36. Acknowledging that routine pre-flight checks often cause a misleading appearance of messages, and that the aircraft is often incapable of distinguishing maintenance phases from pre-flight phases, the data considered only contemplates messages emitted in the airborne flight phases. All the remaining data was considered meaningful and trustworthy, as unlike other data types, such as raw sensor data, this data results from a manipulative and processing phase performed by the aircraft's internal computers.

4.2 Feature Construction

It was decided to base the analysis on the message's emission frequency. The general reasoning behind this decision was that the more critical messages would increase its frequency when closer to the failure event. Therefore, the main features are set as the sum of emitted messages in a certain period. This is referenced at a variable timestamp, that ranges from the first to the last recorded message, in this work denominated as time reference.

It was also decided to add the aircraft registration as categorical variables. This aimed to measure and compare the influence of the messages' source to the frequency of the messages' emission. A categorical variable differs from the numerical variables as the first contains labelled values. For this purpose, it was used the technique denominated as *one-hot encoding*. This consists in a binary approach, assigning the value of 1 to the column that refers to the actual categorical variable. Another way to integrate categorical variables consist in assigning a number to each variable, from a sequence that ranges from 1 to the n categorical variables needed. This feature is consolidated into one column, reducing the feature space of the machine learning problem, which may or may not improve the results. In this thesis, both methods for scaling are considered and used the one that improves the results.

Apart from the above mentioned, the system's lifetime was also included in the features' set to add temporal sense to the model. The new feature measured, in days, the time difference between the previously recorded failure event and the time reference.

4.2.1 Feature statistics

Whenever two variables are highly correlated between themselves, they may be considered as redundant, as both explain the same variance. Therefore, the standard procedure in this situation is to define a correlation coefficient threshold and disregard one of the two highly correlated variables. The threshold considered in this work is 0.90. Maintaining highly correlated variable pairs would only increase the dimensionality of the problem, without improving the final model. One of the main objectives inherent to machine learning consists in conserving the highest possible variable variance with the least amount of variables possible. This prevents *overfitting*, a common problem in machine learning. It occurs when there is excessive learning of the training data, turning the model biased and not performing well when faced with the test data. On the other hand, when the model fails to train with the training data, the model is *underfitted* [9]. All these pre-processing techniques contribute to avoid both cases.

4.3 Data Scaling

This step aims to maximise the machine learning techniques' performance redefining and scaling the features' values. Usually, due to their different nature or sources, the variables in the variable set have very different scales. Converting the variables into

the same scale is propitious to the overall performance of the machine learning algorithms. Sometimes, more than advisable, it is required by some algorithms implemented in the library used in this project - *sklearn* - that the features all vary in comparable scales. Gradient-based and metric-based algorithms all assume that the data is standardised.

This work applies the *StandardScaler* function available from the *scikit-learn preprocessing* library. The *StandardScaler* re-scales the features so that they end up in a normal distribution with a standard deviation of one and a mean of zero, i.e, $\sigma = 1$ and $\mu = 0$. The values, or z-scores are then given by:

$$z_j = \frac{x - \mu_t}{\sigma_t} \tag{1}$$

On the other hand, the target of the analysis is, in this case, not scaled, either for the classification (would not be possible) or the regressive approaches. On the latter, the author considered that it would not make much sense limiting the range of the label through scaling.

Other pre-processing techniques include feature extraction and feature selection. The first consists of extracting new features based on the existing ones. One very well known technique is the Principal Component Analysis. The second measures how useful the variables may be to the result, introducing the concept of feature importance. Due to the lack of improvement result-wise when applying these techniques, they were disregarded, and not included in the pre-processing set of techniques used.

5. Models

This section aims to present the implementation steps carried out to develop the baseline and the prognostic models.

5.1 Baseline

To compare the results from the more advanced approaches based on machine learning algorithms with a more conventional life data-based methodology, this work applies the Weibull distribution to reliability engineering. It has been used for many years for failure analysis, and it is still used to define maintenance intervals for several preventive maintenance plans. Instead of considering all the already discussed variables referring to the messages' evolution, the distribution aims to fit the life data of the system and extract some useful reliability results to compare with the more advanced methods presented in this paper.

The Weibull distribution may be applied in a variety of forms. This project applies a two-parameter Weibull distribution (shape β and scale η). These parameters are computed by the fitting stage of the process, where, based on a selection of time between failure data, the parameters are estimated to best suit its evolution.

One of the several reliability measurements possible to analyse using the Weibull distribution is the BX life. The BX life refers to the life point in time, that being days or cycles, when less than X% of the population has failed. It may as well be defined as the time when the probability of failure reaches X%.

The BX life may be computed using the Cumulative Density Function (CDF):

$$F(t) = 1 - e^{-(\frac{t}{\eta})^{\beta}} \tag{2}$$

F(t) being the cumulative probability of failure, t the time, and η and β the scale and shape Weibull parameters. Being the BX life the time at which the cumulative probability of failure is X, it may be derived from the equation 2, and it is defined as:

$$BX = \eta[-\ln(1 - X)]^{-\frac{1}{\beta}} \tag{3}$$

Equation 3 is nothing more than equation 2 deriving t (BX) as a function of F(t) (X). Therefore, post the fitting stage, any BX may be calculated with the equation 3.

It was decided, in this paper's behalf, to compare the results from the more advanced machine learning approaches with the BX life evolution considering different X, ranging from 10 to 90 percent, in steps of 10. The value considered for the consequent comparison with the data driven models will be the one that has the least error associated when compared with an unseen subset failure data, hence reducing the uncertainty of only considering a unique value for X. The evaluation of the baseline approach follows a cross-validation method.

The uncertainty associated with only using failure data showcases to what extent is the committed error if the failure event and subsequent maintenance action are expected to occur on a fixed periodic basis defined by the BX value. Therefore the failure based analysis disregards any added information concerning the evolution of messages. Hence, the comparison between the approaches' results helps to conclude about the predictive capabilities of the message data.

5.2 Regressive Models

This section aims to explain all the five regressive model variants analysed in this work. Realising that the methods would have to follow tendencies to extract the best performance out of the machine learning algorithms, it was decided to define the type 1 models considering that the most suitable approach would be to measure the number of emitted

messages since the previous recorded failure event. Therefore, no matter what is the evolution of messages issued per day, the tendency is always positive. Other features included the already mentioned in the section 4.2 messages' source and the period since the previous failure event. In addition, the decision to keep both the categorical variables and the time since last failure feature was supported by the results which showed considerable improvement when these features were included in the model. The remaining useful life is here defined as the period between a certain time reference and the nearest failure event of a certain system. In this type 1 iteration of the problem, the failure event is considered as the appearance of failure indicative CAS messages. For the pneumatic system, the two messages considered were the "BLEED 1 FAIL" and "BLEED 2 FAIL".

The constant search of new data formulations that intend to maximise the machine learning methods' performance led to the redefinition of the data structure of the problem. It is defined as the type 2 analysis, the one that instead of considering the sum of messages from the previous failure event until the time reference, it only counts the messages emitted within 24 hours prior of the reference timestamp, maintaining the rest of the feature set.

In the aeronautical field, the "time" may be measured in several different ways. The first two approaches consisted in predicting the time, in days, left until the next failure event, and the so-called reference time also evolved with time steps of one day. Another approach to this problem consists in considering the evolution metrics as flight cycles or flight legs. One flight leg consists in a performed cycle, where an aircraft performs all the possible flight phases. Based on that, the from this point on denominated as type 3 approach consists in redefining the time measures as number of cycles. This analysis intends to minimise the effects of maintenance periods on the determination of the RUL. Although not very frequent, the day-based approaches were influenced by maintenance periods, as they were unable to distinguish if the aircraft was in fact flying.

Apart from the already mentioned analysis, type 4 and type 5 approaches are in all aspects similar to the type 1 and type 2 respectively, except the definition of failure. To measure the influence of considering one of the many failure definitions, this fourth and fifth analysis considers the failure events as the replacements of the two valves present on each side of the air bleed distribution subsystem. One is the valve at the high-pressure (Valve 1) stage of the engines' compressor, and the other is the low-pressure valve (Valve 2).

5.3 Classification Scheme

Until this point, all the considerations were based on the development of regressive model, that intend to, in one way or the other, estimate the time or number of fights left until the next failure event. Therefore, what the models are aimed to predict are floating point numbers, that have an infinitive range.

To broaden the scope of the analysis, it was decided to try a classification approach. Two out of the three projects discussed in section 2.2 applied classifying approaches to develop prognostic models based on the available data, both with success. Therefore, to complement the five different types of regressive analysis already discussed, this thesis uses a similar structure to those mentioned case studies. The main difference rests on the definition of the label.

Considering the real world operation requirements, it was decided to define the classification approach's target as the system's failure risk. Hence, the system's jeopardy is considered high if the failure is predicted to occur within 20 days of a certain time reference. Otherwise, the risk is considered as low. This attribution of the system's risk is what the models are design to predict. The main objective of the analysis performed in this thesis is to obtain the best possible results from the models. The quantification of what is considered as a good result is discussed in the next section.

6. Results

This section starts by showcasing the results from the baseline approach, followed by the results from the regressive and classification models.

6.1 Baseline Results

Table 3: Minimum values of the error's evolution with the BX life range considered. All the units are in days. MAE stands for Mean Absolute Error and RMSE for Root Mean Squared Error

TYPE		**ERROR/BX MIN**			
		MAE	**BX**	**RMSE**	**BX**
1&2&3		90.4	B40	137.8	B70
4&5	Valve 1	244.2	B30	284.3	B40
	Valve 2	137.8	B50	199.7	B60
	Valve 12	117.6	B50	163.4	B60

The table 3 summarises the results obtained from the baseline approach. It presents the minimum MAE and RMSE and the corresponding BX lifes for the four different failure data analysed in this work. It also presents the cross-reference between

the different failure data and the comparable machine learning models' types. The results are discussed in the section 7.

6.2 Regression Results

The prognostic of the remaining useful life of the pneumatic system is the second analysis to be discussed. Five algorithms are applied to develop all the models. Due to space limitation, this paper only shows the best results from each type of approach.

Table 4: Compilation of the best regressive results. All the units are in days.

TYPE		ERROR	
		MAE	RMSE
1		106.1	143.5
2		108.7	147.2
3		81.7	108.4
4	Valve 1	212.4	244.2
	Valve 2	125.9	159.8
	Valve 12	113.4	151.6
5	Valve 1	204.9	232.1
	Valve 2	131.2	182.9
	Valve 12	108.9	148.3

The best model reaches a minimum MAE of 81.7 days and a minimum RMSE of 108.4 days. Worth mentioning that the best models tend to be derived from the Support Vector Machine algorithm. Also, the models obtained from the 4 and 5 approaches fail to reach the results considering the failure definition as the emission of failure indicative CAS messages.

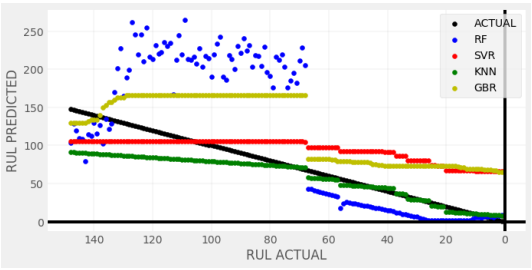

Figure 1: Caparison between the actual and the predicted remaining useful life of a type 1 approach.

Figure 1 presents the plot comparing the evolution of the RUL (Remaining Useful Life) estimation of a selected failure. It's possible to compare how the predicted RUL follow the real evolution. The predictions are extracted from a type 1 approach. All the gathered conclusions from the results presented in this section are discussed in section 7

6.3 Classification Results

This section presents the results from the classification models. These aim to classify the risk of failure inherent to the system at a given moment. Table 5 presents the results from the best models in terms of F1 score, which combines the influence of the precision and recall evaluation measures. The first measures what is the proportion of correct positive predictions compared with all the positive predictions from the models. On the other hand, the recall compares the correct positive predictions with all the samples that should have been predicted as positives. The accuracy computes the percentage of correct predictions between the total number of predictions performed. Notice that the positive results are in this case considered as the prediction of a high risk system.

Table 5: Compilation of the results from the different classification models.

TYPE		EVALUATION MEASURES			
		ACCURACY	PRECISION	RECALL	F1 SCORE
1		0.533	0.153	0.39	0.205
2		0.638	0.179	0.276	0.185
3		0.467	0.199	0.640	0.3
4	Valve 1	0.272	0.044	0.4	0.077
	Valve 2	0.771	0.267	0.348	0.215
	Valve 12	0.695	0.176	0.488	0.235
5	Valve 1	0.399	0.099	0.314	0.209
	Valve 2	0.870	0.294	0.324	0.302
	Valve 12	0.769	0.193	0.337	0.225

The next section discusses and concludes on the shown results.

7. Conclusion

7.1 Achievements

This paper proposed to analyse the potential predictive capabilities of one type of data continuously being emitted by the aircraft - the CMC (Central Maintenance Computer) messages. Several model alternatives were considered, including a non machine learning baseline approach, which only considered failure data and the corresponding time between failures. The result comparison between all the different models allows to extract some relevant conclusions.

Analysing purely the quantification of the errors, the best results from the regressive models reach an MAE (Mean Absolute Error) of 81.7 days and an RMSE (Root Mean Squared Error) of 108.4 days. This would mean if one takes into consideration any prediction in specific, one should expect an error of 81.7 days. In a real-world situation, such high error is prejudicial to the trustworthiness of an autonomous model that would eventually predict the remaining useful life of the equipment. Regarding the RMSE, due to its definition, the results are always equal or greater than the MAE, equality being

the best possible outcome. The RMSE results from this work suggest the presence of high predictive errors, shown by the tendency of the RMSE being substantially higher than the MAE.

Comparing the regressive results with the ones obtained from the baseline analysis allows to extract some relevant conclusions. The undeniable improvement of the regressive results means that a machine learning-based solution would estimate the failures in a more reliable manner than a more conventional preventive maintenance scheduling. Therefore, it proves that such techniques help maintenance move one step forward, allowing to reduce risks and increase safety. Also, this improvement enables to conclude that, associated with more complex methods, the message data assists in providing more accurate predictions regarding the system's remaining useful life.

In view of this, from the figure 1 it is noticeable a well behaved evolution of the RUL's prediction, obtained from the random forest and KNN (K Nearest Neighbors) algorithms, specifically within 60 days of the failure event. This indicates that, in fact, the models were able to extract some information from the learning data that allowed to rightfully predict about the remaining useful life of the system. If the models' data were fully uncorrelated to the failure event, those predictions would not be possible.

The classification results also support this idea. The best overall result in terms of the F1 score was 0.302, from the type 5 approach. However, from the type three approach results a maximum overall recall score of 0.640. This might be considered the most important evaluating measure for a prognostic tool applied to the aeronautics, as it is defined as the ratio between the correct positive scores and the situations that were supposed to be classified as positives. It has an importance that maybe should be more pronounced than the precision score to the final F1 score, as a low recall indicates that there are a significant amount of situations where the model considers the system as healthy, but in fact, it is not. The precision, on the other hand, measures how often the predictions indicated a high failure risk when, in fact, the failure is not imminent. Therefore, a low precision might lead to unnecessary maintenance actions, implying monetary waste. Though, safety being, in this case, a priority, it might be a reasonable assumption to consider the recall as the number one score to maximise.

The results of this thesis indicate that the inherent predictive capabilities of the CMC messages are not enough to develop a fully independent prognostic tool. Despite this, the results show that there are capabilities associated with the CMC messages that enable the improvement of failure prognosis when compared with the more conventional baseline analysis. Also, the classification results reach a recall of 64%, which would not be achievable if the models had no failure predictive capabilities. Therefore, such models may be suitable as a complement to the decision-making process of the unplanned maintenance interventions, and not as the unique decisive factor. From an operational point of view, it is believed that those models may act as alerts to upcoming failure events and that the experienced personnel may also be able to judge the predictions of the models with other indicators, therefore concluding on the truthiness of the prognosis. The results show that the predictions often misjudge the real state of the system, hence the need for the human intervention may not be dismissed. Therefore, a prognostic tool based on CMC messages, although not fully independent, may aid the health management of the fleet, improving the proactiveness to failure events.

7.2 Future Work

To improve quantitatively the results, a future possibility would be to redefine the predictive data. Instead of considering message data that fails to be a promising source to an autonomous prognostic tool, the author suggests considering raw sensor data instead. Raw sensor data represents the behaviour evolution of the system continuously and in greater detail, therefore having greater possibilities in improving the performance of machine learning models, and therefore increasing the chances of developing a fully automated failure prognostic tool. This is supported by the significant improvement result-wise with the introduction of sensor data to the models, discussed in the project [2].

References

[1] Kobbacy, K.A.H. and Murthy, D.P. eds., 2008. Complex system maintenance handbook. Springer Science & Business Media, p57,81.

[2] Baptista, M. (2018). Machine Learning and Deep Learning for Prognostics and Predictive Maintenance of Aeronautical Equipment. Phd. Instituto Superior Técnico, Universidade De Lisboa.

[3] Mishra, M. (2018). Prognostics and Health Management of Engineering Systems for Operation and Maintenance Optimisation. Phd. Luleå University of Technology Luleå, Sweden.

[4] Liu, Y. (2019). Python machine learning by example. 2nd ed. Python Machine Learning By Example: The easiest way to get into machine learning, pp.5-20.

[5] de Pádua Moreira, R. and Nascimento, C.L., 2012, March. Prognostics of aircraft bleed valves using a SVM classification algorithm. In 2012 IEEE Aerospace Conference (pp. 1-8). IEEE.

[6] Nicchiotti, G., 2018, July. Data-Driven Prediction of Unscheduled Maintenance Replacements in a Fleet of Commercial Aircrafts. In PHM Society European Conference (Vol. 4, No. 1).

[7] Kotsiantis, S.B., Kanellopoulos, D. and Pintelas, P.E., 2006. Data preprocessing for supervised leaning. International Journal of Computer Science, 1(2), pp.111-117.

[8] Fortney, K. (2018). Machine Learning — An Error by Any Other Name - Towards Data Science. [online] Towards Data Science. Available at: https://towardsdatascience.com/machine-learning-an-error-by-any-other-name-a7760a702c4d [Accessed 18 Jul. 2019].

[9] Al-Masri, A. (2019). What Are Overfitting and Underfitting in Machine Learning?. [online] Towards Data Science. Available at: https://towardsdatascience.com/what-are-overfitting-and-underfitting-in-machine-learning-a96b30864690 [Accessed 31 Aug. 2019].

[10] Cochrane, C. (2019). Time Series Nested Cross-Validation. [online] Towards Data Science. Available at: https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9 [Accessed 2 Sep. 2019].