Luchtvaartfeiten.nl AviationFacts.eu

CRISP-DM METHODOLOGY

A structured approach for data mining projects

Introduction

The aviation industry continues to increasingly generate and store data. It is expected that the global fleet will generate 98 million terabytes on an annual basis in 2026[1] (**figure 1**). The effective use and analysis of the vast amounts of data, through data mining, could help to optimise processes in a variety of businesses in aviation. A data mining project aims to analyse historical data to generate new, valuable knowledge for an organisation[2]. To structurally approach a data mining project, the Cross-Industry Standard Process for Data Mining (CRISP-DM) was developed. This process model is used to structurally guide the data mining process. This paper will further elaborate on the process of data mining and will discuss how the CRISP-DM methodology could be used to effectively perform data mining projects.

The use of data mining to overcome challenges in aviation

The aviation industry is by nature a rapidly changing environment with growing figures of air transportations every year. Naturally, this growth is followed by an increasing demand in MRO services[3]. MRO organisations are required to keep improving their processes to remain competitive and be able to respond to the increasing demand of reliable and predictable turnaround times. However, aircraft maintenance operations are mainly characterized by their unpredictability regarding process times and material requirements. In today's operations, unpredictability is taken into account by implementing large buffers in terms of time, manpower, and material. These buffers result into relatively inefficient and expensive processes.

To overcome the challenge of unpredictable maintenance processes, data-mining comes into play. Data mining is a process of analysing large sets of data in order to discover any meaningful, possibly unexpected, correlations and patterns in data[2]. In the aviation business, the data could be categorised into three different sources: maintenance data, data from the Flight Data Recorder (FDR), and external data (weather, airport data, etc.). Ideally, data mining these sources would contribute to improve the ability of predicting failures in maintenance processes and, therefore, help to better anticipate on turnaround times and material requirements[4].

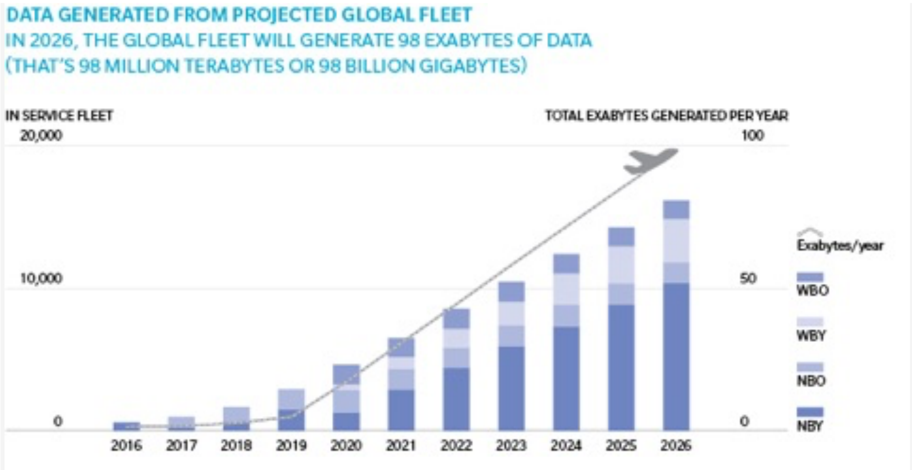


Figure 1: Projected data generated by global fleet

The need for a standard process model

Most practitioners of data mining can admit the complexity of a data mining process. The process requires various tools and different people and the success of such projects highly depends on a mix of proper tools and good analytical skills. Furthermore, a structured methodology and an effective project management is strongly advised. A standard process model could help to understand and manage the complexity of data mining processes[5].

Any data mining process starts off with the original data and has its main objective to generate valuable knowledge out of this data (**figure 2**). However, there are many steps in between. Initially, a selection of the original data should be made. The selection of data is based on the desired objectives of the data mining project. Hereafter, the data needs to be pre-processed and transformed in order to fit these into the data mining tool or algorithm. The transformed data is modelled to recognise any potential patterns. The obtained results need to be interpreted in order to actually acquire new valuable knowledge.

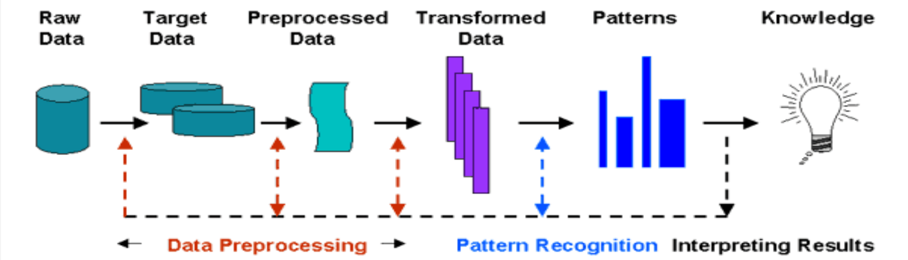
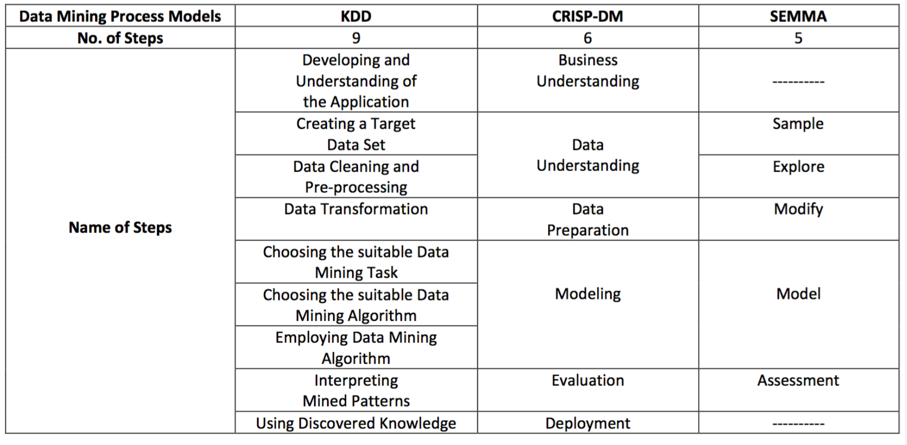


Figure 2: Data mining process

Today's data mining projects particularly make use of three well-known process models: KDD (Knowledge Discovery Databases), CRISP-DM, and SEMMA (Sample, Explore, Modify, Model and Access).
These models divide the data mining process into several phases (**figure 3**). Out of these three, CRISP-DM has proven to be the most commonly preferred model[6]. Both CRISP-DM and SEMMA could be viewed as an implementation of the KDD model[7]. However, CRISP-DM is considered being more complete than SEMMA as SEMMA was specifically designed for the SAS enterprise. Therefore, not designed to be applicable in a broader range of projects[8].



Data Mining Process Models	**KDD**	**CRISP-DM**	**SEMMA**
No. of Steps	9	6	5
Name of Steps	Developing and Understanding of the Application	Business Understanding	----------
	Creating a Target Data Set	Data Understanding	Sample
	Data Cleaning and Pre-processing		Explore
	Data Transformation	Data Preparation	Modify
	Choosing the suitable Data Mining Task	Modeling	Model
	Choosing the suitable Data Mining Algorithm		
	Employing Data Mining Algorithm		
	Interpreting Mined Patterns	Evaluation	Assessment
	Using Discovered Knowledge	Deployment	----------

Figure 3: Data mining process models

CRISP-DM Methodology

CRISP-DM was conceived in 1996 by three "veterans" in the early days of the data mining market: Daimler Chrysler, SPSS and NCR4. It provides a structured approach together with guidelines to help the user to execute a data mining project. The CRISP-DM methodology is based on an iterative nature and consists of six phases[9] (**figure 4**).

1. Business understanding

The initial phase aims to obtain a clear understanding of the desired objectives from a business perspective. Consequently, the business goal needs to be converted to a data mining goal problem definition, which should result in a project plan.

2. Data understanding

In the data understanding phase, the researcher needs to become familiar with the available data. The initial data is collected from the project resources and further examined to identify the main characteristics of the data. The data requires further exploration to address the specific data mining questions. Lastly, the data needs to be assessed on its quality.

3. Data preparation

The data preparation phase involves all activities carried out to construct the final set of data from the raw initial data of the previous phase. The data needs to be cleaned in order to correct for any inaccurate records. The set of data might still require a few adjustments, such as: derivation of attributes by certain calculations and the generation of completely new records. On top of that, the data is integrated whereby information of multiple tables is combined. Ultimately, the data might require any formatting transformations in order to properly feed the data to the modelling tools.

4. Modelling

Once the data is pre-processed, it could be modelled by selecting one or multiple specific modelling technique(s) related to the data mining goal. Before the model is actually built, a procedure needs to be created to test the quality and validity of the model. Thereafter, the modelling tool could start running on the prepared set of data to generate one or more models. The data mining engineer assesses the success of models according to the test procedure.

5. Evaluation

In the modelling phase, the quality of the models should be assessed from a technical data-mining perspective. The evaluation phase assesses whether the business objectives were reached. Once the business needs are satisfied, the entire process will be reviewed in order to identify any aspects which were overlooked. Lastly, the researcher(s) should make a decision on how to proceed with the obtained results.

6. Deployment

The deployment phase considers the results of the evaluation to determine a strategy for deployment within a certain company. Once the results of the project will be used extensively, it is important that the enterprise is fully aware of the required actions to take in order to actually use the models. This phase results in a final report and presentation of the obtained results. The last step includes the final review of the entire project to identify any points of improvement.

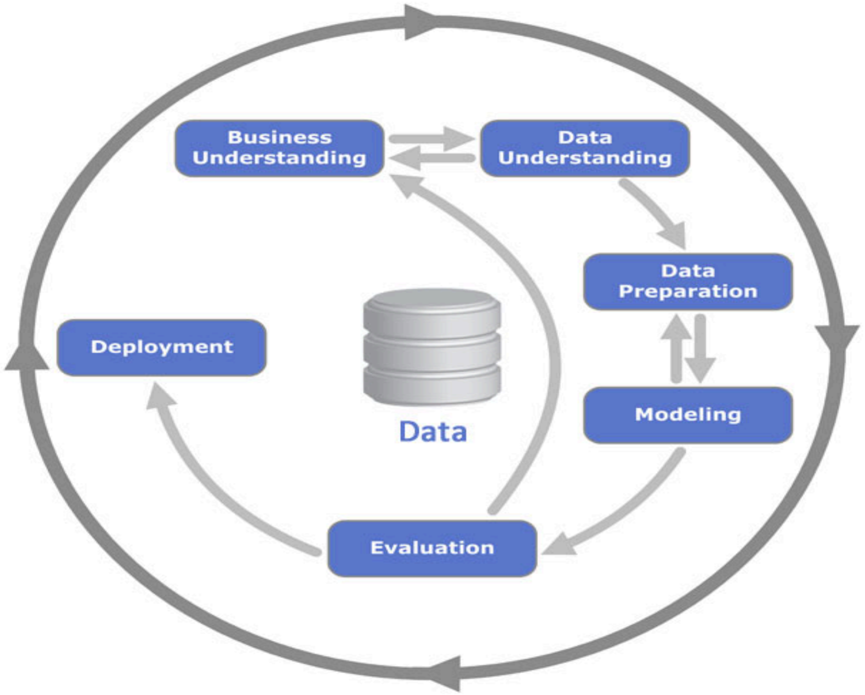


Figure 4: CRISP-DM methodology

CRISP-DM Application on fog prediction (case study)

The CRISP-DM methodology could be applied on a variety of cases in the aviation industry. A data mining project, conducted in 2011, aimed to create a data mining model which could predict the short-term fog near airports using historical data[10]. Unpredictable fog near airports could cause many ongoing problems and could negatively influence airport operations. Improving fog forecasting allows airports to adjust their operation in advance, which could lead to cost savings and saver operations. The used historical data consists of several Meteorological Terminal Air Reports (METAR) and satellite images. The data from both the METARs and the satellite images was extracted and integrated in a singular set of data in order to be able to model the data. The models were developed using 90% of the available amount of data (so-called training data) and tested with the remaining 10% of the data. Hereafter, the model has been evaluated to assess whether the tool could be used in practice. The evaluation showed that using the CRISP-DM approach resulted in a data mining model which is comparable with existing fog prediction methods. The model will continue to develop by integrating more sources of data in future projects. Thereby, increasing the accuracy of the fog prediction model.

Figure 5: Foggy conditions decrease airport capacity

CRISP-DM Application case studies in MRO

In 2016 the Amsterdam University of Applied Sciences (AUAS) initiated a two-year research project: “Data Mining in MRO”. Together with numerous partners in the aviation industry, the project aims to identify how data mining could be used in order to optimise maintenance processes for MRO SMEs[11]. The cooperative companies have already provided the AUAS with numerous case studies which are executed by graduation students of the Aviation Academy. After a literature study, the CRISP-DM methodology was chosen to acquire a structured approach of the data mining activities. The project already achieved several results, which are considered to be promising by both the University and the cooperative companies perspective. A few examples of the obtained results from the case studies are described below.

Investigation of drop in fleet availability

One of the case studies, executed within the AUAS project, investigated the causes of a drop in fleet availability during high season[12]. It has been tried to correlate Air Transport Association (ATA) chapters to the drop in fleet availability. The study used historical data such as weather data, flight data and unscheduled ground time events. These data had to be cleaned and integrated. Thereafter, this data was modelled using a descriptive analysis. The results of the analysis identified a correlation between the performance drop and ATA subchapters related to tyres, brakes and cabin air quality. The results will be used in order to anticipate on the influence of these ATA subchapters on the operation of the aircraft. Thereby resulting into an increase in aircraft uptime and a decrease in part costs.

Optimal moment to change aircraft tyres

A second case study executed within the AUAS project investigated the optimal moment to change aircraft tyres[12]. The study used historical data such as aircraft weight, braking action, runway length and temperature. These data had to be cleaned and integrated into a single dataset. Hereafter, a predictive model was created which could be used to determine the optimal moment to change aircraft tyres. Using this model will lead to increased aircraft availability and decreased maintenance costs.

Prediction of component failure

A third case study investigated how the failure of aircraft components could be predicted using external historical data sources such as maintenance data and weather data[13]. The data have been split into different flight phases after which it has been modelled. The model already showed flight anomalies before the component actually failed. Therefore, using the model will contribute to better anticipation on future component failures.

Optimal moment of engine replacement

A fourth case study investigated the prediction of the optimal engine replacement moment[12]. The study found that the Life Limiting Parts (LLP) and the Exhaust Gas Temperature (EGT) determine the optimal engine replacement moment. The study has been conducted using historical data such as fuel consumption, oil pressure and oil consumption. These data had to be selected per engine type, where after the data had to be cleaned. Thereafter, the data had to be modelled which resulted in an Engine Health Monitoring model. This model is capable of forecasting the optimal engine replacement moment. Using the model will lead to an increase in aircraft availability as well as a decrease in maintenance costs.

Maintenance planning optimisation

Another example is the case study which investigated how to increase aircraft availability with an improved planning of maintenance tasks[14]. The study used historical maintenance data which had to be cleaned and integrated in order to be able to model it. The modelling consisted of a visualisation between planned and actual performed maintenance. In addition, forecasting algorithms were created which were based on actual duration of task cards (which describe which maintenance tasks must be performed). The use of these algorithms would lead to a more efficient maintenance planning.

Discussion

As mentioned previously, the CRISP-DM methodology has proven its practicality in several case studies for different MRO SMEs. In order to assess the usefulness of the CRISP-DM methodology, a study was conducted which aimed to examine the experiences of the graduation students regarding their execution of the data mining activities[15]. By interviewing the graduation students, the researcher discovered that not all phase-related activities were executed because they were not applicable for the project. However, the students did address the use of CRISP-DM to be a useful guideline for a structured data mining project. The model did contribute to an efficient project planning and help to effectively communicate the project's progress.

During the case studies, some difficulties were encountered within the phases of the CRISP-DM methodology. One of these difficulties is the fact that several databases contain a large amount of data, which is often found to be improperly documented[10]. These issues affect the outcomes of the specific data mining tasks. Furthermore, the CRISP-DM methodology did expose some critical aspects, which could be improved in order to optimise the usefulness of the data. One of these issues is the fact that companies do not see the advantages which come along with their gathered data from the past. Another issue is the fact that the case studies revealed a gap between the data that a company should store (according to their business model) and the data that is actually stored[11].

Nevertheless, the outcomes of the case studies, using the CRISP-DM methodology, did identify valuable insights and practical results. The obtained results will be used in future case studies to further develop the knowledge regarding data mining in MRO SMEs.

How could companies benefit from the use of CRISP-DM?

As the aviation industry continues to increasingly gather and store data, effective use of these data could provide many benefits for a variety of businesses. By analysing historical and ongoing data, companies could generate valuable knowledge, which could contribute to the optimisation of processes. However, the process of data mining is considered to be a process with high complexity and a structured process model is necessary to successfully execute such projects. The most commonly preferred process model, CRISP-DM, already has proven to be a successful guideline in a variety of cases. The methodology was used in a variety of case studies for MRO SMEs. These case studies actually identified relevant patterns and created predictive models from the data sources, which could be used to optimise maintenance processes. Thereby, potentially reducing maintenance costs and increasing aircraft availability.

However, as CRISP-DM is a relatively general approach, the model was not always entirely applicable to any project. Therefore, certain activities, described by the model, were unnecessary to be carried out along the project. Overall, the use of CRISP-DM did add value to these projects in terms of project planning, communication and documentation.

Glossary

- ***Cross-Industry Standard Process for Data Mining (CRISP-DM)****:* A structured process model used for data mining projects.
- ***Maintenance Repair and Overhaul (MRO)****:* A company specialized in performing maintenance on aircraft and their components.
- ***Small to Medium-sized Enterprises (SME):*** A company with less than a specified number of employees, annual sales, assets, or any combination of these.

References

1 Maire, S., & Spafford , C. (2017, June 16). The Data Science Revolution That's Transforming Aviation. Retrieved from Forbes: https://www.forbes.com/sites/oliverwyman/2017/06/16/the-data-science-revolution-transforming-aviation/#61771e367f6c

2 The Economic Times . (2018). Definition of 'Data Mining' . Retrieved from The Economic Times : https://economictimes.indiatimes.com/definition/data-mining

3 Aviation Maintenance . (2017, July 28). EUROPEAN MRO: Challenges Mount But Opportunities Remain . Retrieved from Aviation Maintenance : https://www.avm-mag.com/european-mro-challenges-mount-opportunities-remain/

4 Centre for Applied Research Technology. (2017, February 16). DATA MINING IN MRO. Retrieved from CENTRE OF APPLIED RESEARCH TECHNOLOGY: http://www.amsterdamuas.com/car-technology/shared-content/projects/projects-general/data-mining-in-mro.html

5 Wirth, R., & Hipp, J. (n.d.). CRISP-DM: Towards a Standard Process Model for Data Mining. Retrieved from Semantic Scholar: https://pdfs.semanticscholar.org/48b9/293cfd4297f855867ca278f7069abc6a9c24.pdf

6 Piatetsky, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. Retrieved from KDnuggets: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html)

7 Azevedo, A., & Santos, M. F. (2008). KDD, semma and CRISP-DM: A parallel overview.

8 Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA) . Gujrat: Innovative Space of Scientific Research Journals.

9 Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). CRISP-DM 1.0.

10 Bednár, P., Albert , F., Babič, F., Paralič, J., & Bartók, J. (2011). Design and implementation of local data mining model for short-term fog prediction at the airport. International Symposium on Applied Machine Intelligence and Informatics.

11 Borst, M., Broodbakker, J., Pelt, M., & de Boer, R. J. (2017). RAAK MKB Data Mining in MRO.

12 Pelt, M. (2017). Data Mining in MRO process optimisation . Amsterdam, Netherlands .

13 Brienen, S. v. (2017). Data potentials: Scheduling unplanned maintenance of legacy aircraft.

14 Killaars, M. (2017). Predictive Maintenance in MRO.

15 Doolhoff, S. (2017). Data mining in aviation MRO.

Image references

Front page: https://www.wearefinn.com/topics/posts/data-collaboration-for-mro-why-sharing-is-caring-in-the-aviation-industry/

1 https://www.forbes.com/sites/oliverwyman/2017/06/16/the-data-science-revolution-transforming-aviation/#61771e367f6c

2 https://www.loginworks.com/data-mining-services-various-type/

3 2014 Innovative Space of Scientific Research Journals http://www.ijisr.issr-journals.org/

4 http://dwgeek.com/9-laws-data-mining.html/

5 https://pxhere.com/en/photo/1117514

Dutch Summary

De luchtvaartindustrie genereert data in toenemende mate. Het effectief gebruik van deze data kan het optimaliseren van processen bevorderen. Het analyseren van (historische) data om nieuwe, waardevolle kennis te verkrijgen, wordt ook wel 'Data Mining' genoemd. Om data mining processen effectief uit te kunnen voeren zijn er verschillende procesmodellen ontwikkeld. Het meest gebruikte model is de CRISP-DM methode (Cross-Industry Standard Process for Data Mining).
CRISP-DM beschrijft het proces van data mining aan de hand van zes fasen: Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation en Deployment. Deze methode is al in verschillende studies gebruikt, voornamelijk met het doel om de voorspelbaarheid van onderhoudsprocessen te verbeteren. Aan de hand van deze resultaten kunnen onderhoudsbedrijven beter anticiperen op de verwachte onderhoudstijd. Het gebruik van deze methode heeft vele voordelen met zich meegebracht. Met name in de planning, communicatie en documentatie van resultaten, zorgt deze methode voor een gestructureerde aanpak van data mining projecten.

This is a Luchtvaartfeiten.nl / AviationFacts.eu publication.

Authors: M.P. Bloothoofd, A. Francken, R. Graas
Editorial staff: R.J. de Boer PhD Msc, G. Boosten MSc & G.J.S. Vlaming MSc

Copying texts is allowed. Please cite:
'AviationFacts.eu (2016), *CRISP-DM Methodology* Fact sheet,
www.AviationFacts.eu'

AviationFacts.eu is an initiative by the Aviation Academy at the Amsterdam University of Applied Sciences (AUAS). Students and teachers share knowledge with politicians and the general public to ensure discussions are based on facts.

January 2018