FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Classifying Air Traffic Scenarios and Associated Environment Conditions With Respect to Operation Risk

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Business Informatics

by

Markus Bardach, BSc.
Registration Number 1325794

to the Faculty of Informatics

at the TU Wien

Advisor: O. Univ. Prof. Univ. Doz. Dipl.-Ing. Dr. Michael Schrefl
Assistance: Dr. Eduard Gringinger

Vienna, 25th January, 2020

Markus Bardach | Michael Schrefl

Erklärung zur Verfassung der Arbeit

Markus Bardach, BSc.
Dommayergasse 8/15, 1130 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 25. Jänner 2020

Markus Bardach

Acknowledgements

First I would like to thank Professor Schrefl for his expert advice and encouragement throughout this thesis, as well as Dr. Eduard Gringinger for answering my questions whenever needed and for sharing his domain knowledge in the field of air traffic management.

Additionally, I would like to thank everyone who supported me in any form during the time of this thesis, especially my girlfriend Nadine.

Kurzfassung

Das Ziel dieser Diplomarbeit ist es, ein Model zu entwickeln, welches klassische Flugverkehrsszenarien und zugehörige Umweltbedingungen in drei Risikoklassen klassifiziert. Ein Flugverkehrsszenario wird durch den Ort, dem Flughafen an dem das Szenario stattfindet, einer Zeitspanne und dem Typ des Szenarios (Ankunftsszenario oder Abflugszenario) identifiziert und enthält sämtliche Flug- und Flugzeugdaten von ankommenden oder abfliegenden Flügen aus dieser Zeitspanne. Bei Daten über Umweltbedingungen handelt es sich um Wetter- und Notice-to-Airmen-Daten. In dieser Arbeit wird das klassische Flugverkehrsszenario um diese Daten erweitert. Eine Risikoklasse errechnet sich aus den durchschnittlichen Kosten aller Verspätungen der Flugzeuge aus diesem Szenario. Die Forschungsfrage ist, ob so ein Model Szenarien in Risikoklassen klassifizieren kann. Für diese Data-Mining Aufgabe wird der "cross industry standard process for data mining"verwendet, abgekürzt mit CRISP-DM. Dieser besteht aus sechs einzelnen Phasen. Zuerst werden die Flugdaten mit einfachen Abfragen analysiert, um Flughäfen für die Szenarioerstellung auszuwählen. Szenarien müssen für jeden Flughafen einzeln erstellt werden, da verschiedene Attribute auf verschiedenen Flughäfen unterschiedliche Relevanz haben. Basierend darauf wurden die Flughäfen Atlanta in den USA und Wien ausgewählt. Kosten von Verspätungen werden mithilfe der linearen Regressionsfunktion, welche von EUROCONTROL entwickelt wurde, berechnet. Die Regressionsgerade beinhaltet alle taktischen Kosten eines verspäteten Fluges inklusive Folgekosten, die durch die Verspätung ausgelöst werden. Die Kosten berechnen sich basierend auf dem maximalen Startgewicht eines Flugzeuges. Die beiden fertigen Datensätze von Szenarien werden für das Training von einem Random Forest Model und einem Multilayer Neural Network verwendet. Die dabei verwendete Software heißt Rapid Miner. Die beiden Modelle werden mithilfe von Testmetriken verglichen. Für den Vergleich von mehrklassigen Klassifizierungen werden Precision und Recall verwendet. Die Ergebnisse zeigen, dass das Random Forest Modell bessere Werte erreicht als das Multilayer Neural Network. Precision und Recall erreichen bei der Klassifizierung von Risikoklasse 3 Werte über 80%. Diese Klasse beinhaltet Szenarien mit der höchsten durchschnittlichen Verspätung und somit auch mit dem größten Einsparungspotenzial. Diese Klassifizierung kann Fluglotsen helfen, aufkommende Szenarien besser zu evaluieren und entsprechende Maßnahmen zu setzen, um die Verspätung zu verhindern oder zu minimieren. Einige dieser Maßnahmen sind der Tausch von Landeslots, das Öffnen einer weiteren Start- und Landebahn oder eine Änderung der Landebahnkonfiguration.

Abstract

The goal of this thesis is to develop a model to classify air traffic scenarios proper and associated environment conditions into three risk classes. An air traffic scenario proper contains flight data, information about the arriving and departing aircraft and basic data about the airport and runway. It is identified by the airport, a specific time span and the type, describing if it contains only departing or arriving flight data. Environment condition data are meteorological data and notice-to-airmen messages (NOTAMs). The scenario proper enriched by environment condition data is the air traffic scenario that is classified in this thesis. The risk class is calculated based on the average delay cost of all flights in a scenario. The research question is, if a classifier can predict the risk classes of air traffic scenarios. For this data mining task the cross-industry standard process for data mining (CRISP-DM) is used, which consists of six phases. First queries on flight data try to find airports with high capacity and delay. Air traffic scenarios need to be created for individual airports as the relevance of attributes varies locally. The airports of Atlanta and Vienna are selected to create air traffic scenarios. Delay costs are calculated with the linear regression analysis of full tactical delay costs including reactionary costs developed by EUROCONTROL, which is based on the maximum take-off weight of an aircraft. The final datasets for the two airports are then trained with a random forest classifier and a multilayer neural network. The tool used for classification is Rapid Miner. The two classifiers are compared by using the metrics precision and recall. Results show that the random forest classifier outperforms the multilayer neural network. Precision and recall values are analysed with a confusion matrix and reach over 80% for class 3, which includes scenarios with the highest delay and thus with the biggest saving potential. This can help air traffic control to evaluate upcoming scenarios more easily and lets them take actions to try to prevent the delay. Some of these actions can be slot swapping, opening a runway or changing the runway configuration.

Contents

CHAPTER 1

Introduction

1.1 Motivation

The aviation industry has seen an enormous increase in air transportation around the world in recent years. Flying has become more and more affordable and is no longer reserved to a few. The International Air Transport Association (IATA), which is a trade association of the world's airlines estimates that passengers will double from 4 billion air travelers in 2017 to 7.8 billion in 2036 [IAT18]. With this sharp increase in passengers, the amount of delayed flights grows as well. The European Organisation for the Safety of Air Navigation (EUROCONTROL) published an article about the latest statistics on delays in 2018. In the first 231 days of 2018 the total delay reached 14.1 million minutes of en-route delay, which represents just the delay while the aircraft is in the air after take-off and before landing. This is a 123% increase compared to 2017. The two main causes of delay were capacity/staffing issues and weather [EUR18]. These delays can be directly converted into costs. The University of Westminster published a report in cooperation with EUROCONTROL about delay cost reference values. The report states that on average delay costs start at EUR 32 for the first minute and rise to EUR 80.270 for over 300 minutes of delay [oW15, p. 81]. Increasing air traffic has other effects beside economic aspects too, such as environmental pollution and the increasing challenge of planning flights as efficiently as possible while maximising safety and minimising delay.

Aviation has three big stakeholders, which must interact with each other as well as possible to ensure a seamless flight experience for passengers. Those stakeholders are the airlines, airports, and Air Navigation Service Providers (ANSPs). With increasing air traffic, the need for more efficient processes within and between these stakeholders increases. As one of the main causes of delay is the weather, the demand for more adaptable flight planning and air traffic management is necessary as meteorological events cannot be changed, but plans, processes, and workflows can.

1.2 Problem Statement

Every flight scheduled has a flight plan, which includes information about take-off and landing slots as well as the route. This plan is created before the flight and has to be approved by the air traffic control as they coordinate all flights and try to minimise congestion and delay. In addition to flight plans, environment conditions can have influence on a flight. These data includes: notice-to-airmen messages (NOTAMs) that describe runway or airspace closures, Meteorological Terminal Aviation Routine Weather Reports (METAR) that contain basic weather data such as air temperature and information about special weather phenomena from thunderstorms to hail, and airport and runway information such as the number of runways or the length of a runway, which might limit the traffic. Changes in environment conditions are always possible and are one of the main reasons for congestion situations at airports that lead to delay. To reduce the occurrence of such delays air traffic control has to react fast and recognise such changes as early as possible.

The problem is that there is currently no intelligent combination of data sources that can detect upcoming congestions or delays. This means that planning and situational awareness does not adapt fast enough to counteract scenarios with high delay. Therefore the sum of all available information can be expressed into an air traffic scenario. An air traffic scenario proper contains flight data, information about the arriving and departing aircraft and basic data about the airport and runway. It can be identified by the airport, a specific time span, and the type describing if it contains only departing flight data or only arriving flight data. As there is more data available during the time span of the air traffic scenario proper, it can be enriched with environment condition data. These data includes meteorological information and NOTAMs. In this thesis an air traffic scenario is the combination of environment condition data and the air traffic scenario proper.

1.3 Aim of the work

Combining different data sources such as air traffic data, meteorological data and NOTAM data allows to analyse the reasons for the occurrence of delays of aircraft at the departure and the arrival airport. This data can be classified into three risk classes. A risk class is defined by a delay range and each scenario has an average delay per aircraft for the upcoming air traffic at a specific airport in a certain time span. Higher risk classes predict more delay, and thus a more critical situation than low risk classes. If these scenarios are classified into the different risk classes, it can help to identify critical situations earlier and trigger actions to cope with the peaks in flight traffic. The earlier such situations can be detected the better they can be reacted to with the appropriate action according to the risk class. Risk classes comprise the basis for the decision of which processes come in to play. This can influence flight routes or flight planning or require more flexible slot management or slot swapping. Therefore the following questions need to be answered:

- Can a classifier predict pre-defined risk classes of air traffic scenarios based on flight

plan data and environment condition data? What actions can be triggered for the risk classes?

The first question tries to answer if a classifier that was trained on historical flight and environment data can predict the risk class for an air traffic scenario that only includes the planned arrivals, departures and environment information during a fixed time span. If the classifier is able to predict these risk classes to a certain precision the sub-question searches for actions that can influence the planned arrivals, departures or environment conditions to reduce the predicted risk class and minimise delay consequences. An example is that another runway has to be opened as the risk class would be worse if the runway was kept closed for the traffic in the scenario.

1.4 Methodological approach

The methodology of this thesis will be according to the cross-industry standard process for data mining (CRISP-DM). This process defines six phases in how to perform data mining (*business understanding, data understanding, data preparation, modelling, evaluation and deployment*) [CCK+96a]. The research questions described in the previous section are the goal of this process. To answer them, the necessary data has to be collected. For the classification of different scenarios the following data sources will be used: Historic flight data, general information about airports and runways, weather information from Meteorological Terminal Aviation Routine Weather Reports (METAR) data and Notice-to-Airmen messages for the selected airports. These data sources build the basis for the next phases of the data mining process, which is data preparation. In this phase all data is combined into data tuples which represent scenarios at a specific airport. These tuples include information about the number of arrivals or departures grouped by plane size, the delay and weather information regarding wind, visibility, thunderstorms and other weather phenomena. An example of what attributes such a scenario can contain is shown in Table 1.1.

All these tuples will then be grouped into one of three risk classes based on delays of all incoming and outgoing flights in this scenario. The example of Table 1.1 is in risk class 3 because of the average delay of all aircraft in this scenario. A scenario has risk class 1 if the average delay of all aircraft is little. The risk class is an ordinal scale so classes 2 and 3 describe scenarios with more delay. Class 3 is the worst class as it covers all delays over a certain threshold. The exact class ranges will be defined in combination with the delay cost calculation. To perform classification, the data needs to be split into training and test data, which is the last step of the data preparation phase. In the modelling phase a classifier can then, based on classification rules, classify the scenarios into the different classes. In this phase the trained classifiers are compared with test metrics like accuracy, precision or recall. If classifiers do not achieve satisfactory results the adaptations of their parameters or even changes in the data preparation phase might be necessary. The best results then move on to the evaluation phase. This phase compares the results with

Attributes	Data Type	Value Range	Example Value
Scenario Begin	Date	01.01.2016 00:00:00 - 31.12.2017 23:59:59	02.04.2017 14:50:00
Scenario End	Date	01.01.2016 00:00:00 - 31.12.2017 23:59:59	02.04.2017 15:49:00
#Small Aircraft	Integer	0 - 100	10
#Medium Aircraft	Integer	0 - 100	45
#Large Aircraft	Integer	0 - 100	9
Air Temperature	Double	-5 - 100° F	45.00
Dew Point	Double	-5 - 100° F	40.00
Wind Speed	Double	0 - 100 mph	20
Wind Direction	Integer	0 - 360°	120
Visibility	Integer	0 - 10000m	1400
Cloud Ceiling	Integer	0 - 35000 feet	5000
Thunderstorm	Boolean	TRUE/FALSE	FALSE
Rain	Boolean	TRUE/FALSE	TRUE
Snow	Boolean	TRUE/FALSE	FALSE
Open runways	Integer	0 - 5	2
Airspace closed	Boolean	TRUE/FALSE	FALSE
Airport	String	50 characters	Vienna
Risk Class	Integer	1 - 3	3

Table 1.1: Example tuple for an air traffic scenario

the questions asked at the beginning of the process and checks if the process led to a successful result. The last phase, called deployment, takes the results of the evaluation phase into account and discusses actions than can reduce delays or counteract them depending on the risk class.

1.5 Structure of the work

The first part of the work is a literature research, which gives a short description of data mining in general with focus on classification. Additionally, it describes a process to execute data mining tasks, which includes the six phases business understanding, data understanding, data preparation, modelling, evaluation and deployment. Each phase of the data mining process is described in chapter 2 by its tasks and information about the state-of-the-art regarding the phase. Each phase is then given a separate chapter to elaborate the tasks project specific. These chapters apply the description of chapter 2 onto the concrete goal of air traffic scenario classification. The structure of the work is finalised with the critical reflection and contribution in chapter 9 and the future work and summary in chapter 10.

CHAPTER 2

State of the Art

This chapter gives an overview of data mining in general and its methodologies. The method of classification is discussed in more detail as it is applied in this thesis. The data mining process is presented and every phase of the process is described in detail. Every phase outlines its theoretical tasks and the analysis of existing literature in the field of air traffic management as well as general learnings from data mining literature.

2.1 Introduction to Data Mining

Data mining can be defined as

> the study of collecting, cleaning, processing, analysing, and gaining useful insights from data [Agg15, p. 1].

A more striking explanation is the comparison to gold mining, where it is the goal to find a small gold nugget in huge amounts of rocks. With data mining the goal is to find some interesting knowledge in masses of other data [HKP11, p. 5]. However, data mining is an umbrella term which comprises different methodologies to obtain new valuable knowledge from masses of heterogeneous data sources. Even though data mining can be applied in large numbers of different areas all applications are closely connected and can be divided into four "super problems". These are association pattern mining, clustering, classification and outlier detection [Agg15, p. 14]. This can be explained with a simple $n \times d$ data matrix. Each row represents a record n and each column represents an attribute d. The evaluation of the relationship between columns determines the frequency of relationships between values in a particular row, which leads to *association pattern mining*. When one column is of special interest, then the relationship between the selected column and others has to be determined. This can be used to predict the value of this column based on the values of other records. This is referred to as *data classification*. If the focus is

shifted to the relationship between rows, the goal is to find subsets where values in the corresponding columns are related. The division into subsets is called *clustering*. When a row is highly different from other entries, this data point can be seen as an anomaly and is interesting for an *outlier analysis* [Agg15, p. 15]. These four problems are extremely important because they cover an immense range of different scenarios when it comes to data mining applications. The second described super problem, data classification, is the one that is used to predict risk classes of air traffic scenarios. It belongs to the group of supervised learning methods, because of its two-step process. First is the learning step, where a classification algorithm builds a classifier based on a training set. Second is the classification step where the trained classifier predicts categorical class labels of previously unseen data with unknown labels [HKP11, p. 328]. It can be informally defined as follows:

> Given an $n \times d$ training data matrix D, and a class label value in $1...k$ associated with each of the n rows in D, create a training model M, which can be used to predict the class label of a d-dimensional record $\overline{Y} \notin D$ [Agg15, p. 18].

Classification is the problem of the four that can be applied to the three others too. With clustering these data are categorised into k groups, a classification problem categorises a record with an unknown label into a group with a learning model and a training database. This is the difference to unsupervised learning as it does not have a training database to learn from. Furthermore, association pattern mining can be used for classification too. Frequent patterns that contain a class label can provide useful information about correlations for the training model. Classification can also be seen as a specific version of outlier detection. With supervised outlier detection, examples of outliers are available and can be tagged to a so called rare class, while the others belong to the normal class [Agg15, p. 19].

2.2 The Data Mining Process

To perform data mining in a valid and transparent manner a defined process is crucial. All steps from the building of a hypothesis to the results and conclusions should be clear and verifiable. This is the reasons for the cross-industry standard process for data mining (CRISP-DM) which was already defined in 1996 [BP18].

This iterative process was developed to organise business-oriented data mining and is divided into six phases. The first phase is *Business understanding* and gives the possibility to assess the situation and determine the goals of the whole data mining process. It can be seen as hypothesis building. This is followed by *Data understanding* where the data is collected, described, explored and verified. It is important to determine whether all data is available and valid to reach the goal set in the first phase. The next phase is *Data preparation*. Collected data needs to be selected, cleaned and constructed. Due to the

Figure 2.1: CRISP-DM from 1996, [CCK+96b]

fact that almost all data mining processes require data from different sources and formats, data preparation is essential to achieve good and valid results. Phase four is *Modelling* where a model is selected, testing is designed and the model is built and assessed. In this phase the actual data mining is performed. It is followed by *Evaluation*. The results of the previous phase are evaluated and the whole process is reviewed. It is possible that the process owner will have to restart from phase one because the hypothesis from the beginning has to be reformulated or the chosen data sources were not right for the problem. The process is concluded with the phase *Deployment*. The evaluated results have to be deployed, monitored and maintained in the business [CCK+96b]. This was just the description of one cycle. The iterative character of this process specifies that the deployment of data mining results leads to a better business understanding (Phase 1), which then leads to new hypotheses and goals and so on [CCK+96b].

2.3 Business Understanding

The first phase tries to understand the business objective by acquiring domain knowledge with the goal to be able to convert it into a data mining problem definition.

Determine Business Objective
This task elaborates what should be achieved with the data mining process from a business perspective. It ensures that the process finds the right answers to the right question. In the particular case of the master thesis, this task should specify in more detail the aim of the work described in the previous chapter. The objective should be to contribute something new to the state-of-the-art by intelligently combining available data sources of air traffic management to predict risk classes.

Assess Situation
Assessing the situation is about fact-finding of resources, constraints and assumptions in the state-of-the-art, which helps to refine the business objective and to be able to select data mining goals for the process.

In the area of air traffic delay prediction several different findings have been made. When it comes to the prediction of delay, meteorological conditions are of key importance according to the current level of knowledge. For examples paper [ASSRA04] states that weather accounts for nearly 75% of delays due to tight connections and could increase further if no actions are taken. This statement is also supported by the paper [ABE01] which states that reduced ceiling and visibility are the leading contributors to major delays. Similar findings have also been made with different data mining applications. The result of a pattern identification of air traffic flow management in paper [CWLC14] showed that air temperature and wind speed correlates with air traffic delay. Furthermore, a classification done in paper [BMTT16] of flight delays with historic flight information and meteorological data including the two attributes ceiling and visibility highlighted by [ABE01] demonstrate that good results can be achieved with these data sources. This also explains why other similar papers in the field of flight delay classification like [PAYLD18] and [AKW+09] mention that meteorological information would have improved their results and proposed it as ideas for future work.

A huge majority of work about delay prediction focused on the delay of a single flight. Paper [AKW+09] tried to classify flights into the categories "ahead of time", "on time", and "delayed", while paper [BMTT16] used delay thresholds from below 15 minutes to below 90 minutes of delay. This makes the distinctive approach of classifying air traffic scenarios into risk classes new, as the classifier predicts the average delay of a group of planes during the time span of the scenario.

An interesting aspect of flight delay classification was described in paper [AKW+09], that different attributes have locally varying relevance. For example air temperature is more relevant for delay prediction at certain airports than others. The authors took this factor into account by developing their own classifier that was able to slightly outperform traditional classification methods. This fact makes a generalisation of the classifier for different airports difficult and suggests to develop a classifier for a single airport and describe a way to adapt it to others.

When it comes to modelling of classifiers, current knowledge shows that there is not a single best classification method for a certain application, but that different classification methods have strengths and weaknesses according to [HKP11, p. 331]. For example the Random Forest method is robust to outliers, scalable and able to model non-linear decision. These are factors that are important for airport scenarios. The papers [PAYLD18] and [BMTT16] mention in the paragraph about data selection that both used random forest classifiers for their task. In the second paper this method performed best against others like Support Vector Machines, Naive Bayes and Logistic Regression. If computational power allows it, it is always good to try different classifiers and compare the results. Another method is Multilayer Neural Networks which is able to model non-linear and

complex relationships and is able to learn hidden relationships [HKP11, p. 398]. This might be the reason why paper [WLD18] used this method for their delay predictions just with flight data. The papers [FH01] and [KGH15] mention that in case of multi-class classification tasks Multilayer Neural Networks are the classifier to choose.

In conclusion existing literature about classification in the air traffic domain showed the importance of meteorological information when it comes to delay and that the impact of it differs from airport to airport which makes generalisation difficult. Lastly, the dominant classification methods are Random Forest and Multilayer Neural Networks which achieved the best results for flight data sets.

Determine Data Mining Goals
The data mining goal determines the objective in technical terms and what the output should be. Goals can be where the classification has to focus on for a specific reason or in case of a multi-class classification which amount of classes are useful for the business objective.

2.4 Data Understanding

The second phase describes all the data used for the data mining process and its sources in detail. In this phase basic queries are performed that help to understand the data and to get first inputs for the data mining goal.

All data sets necessary for the data mining goal of phase one are analysed by four tasks. **Data collection** includes the location of the data and the methods used to acquire them with a description of any problems that occurred and the appropriate solutions. The second task is **Data description** where the properties of the data set are listed. This includes the format, quantity, description of attributes and any other discoveries. Third is **Data exploration**. This task does basic querying of the data that address the data mining goal by giving the results of simple aggregations or statistical analysis. Last is **Data quality verification** which checks if the data is complete, or contains error or missing values [CCK+96a].

This not only helps to get a better understanding of the collected data, it also helps to make the work reproducible, which is an important aspect in data mining.

2.5 Data Preparation

The third phase prepares the data for the modelling phase. Real data rarely satisfy the quality requirements needed for the data mining process. Data quality is compromised by factors like accuracy, completeness, consistency, timeliness, believability or interpretability. Human or computer errors in data entries can have many causes from purposely submitted wrong data to errors in transmission. Therefore several tasks are necessary in the data preparation phase.

A paper about the impact of data preparation published in the European Journal of Operational Research came to the conclusion that different data preparation schemes have statistically significant impact on the accuracy of different classification methods. Especially neural networks and support vector machines react very sensitively to preparation [CLS06]. This emphasises the importance of this phase and its tasks.

Data Selection

The first task of the data preparation phase decides which data from the sum of collected data sets are selected for the modelling phase. This includes selection of attributes and of records. Attribute selection, which is also called feature selection is an important step as it tries to remove redundant or irrelevant features from the data set with the goal to improve accuracy. Some features might include noisy or missing data that affects the classification process negatively [TAL14]. A comparative study showed how feature selection can increase the accuracy of classification. The study was done on 15 different real datasets comparing Naive Bayes, artificial neural Network as Multilayer Perceptron (MLP) and decision tree classification methods. The results showed that MLP was the most affected classifier and that the accuracy was improved up to 15.55% compared to the data set without feature selection applied. This shows the importance of a good feature selection process [KzI12].

Data Cleaning

Data cleaning raises the data quality to the required level and is done by filling in missing values, smoothing noisy data identifying outliers and resolving inconsistencies. It is not unusual that data has missing entries or errors. Faulty records have to be edited to exclude problems with the modelling afterwards [Agg15].

Data Construction

This task is one of the main ones as the air traffic scenarios are being constructed from the different data sources and the risk classes are defined. The state-of-the-art suggests different methods to divide data into classes. Four basic methods are: equal interval by dividing the range into equal classes, quantile for classes with the same size, natural breaks by using natural groupings inherent in the data and standard deviation by showing how much a record varies from the mean [Inc06].

Data Integration

The last task in this phase is data integration. If the data comes from several sources with different naming conventions causing inconsistencies or redundancies it has to be integrated to a single format. Data reduction with the goal to obtain a smaller dataset which produces still the same analytical result is also part of this task. This can be achieved by dimensionality reduction or numerosity reduction [HKP11, p. 87].

2.6 Modelling

In the modelling phase various modelling techniques are selected. To compare the performance of different classification methods and to check the model's quality and validity a test design is necessary. In the model building task the tool to run the prepared dataset is chosen and the model is built. Lastly the model is assessed if it maps the business objective and data mining goal. Experimenting in this phase leads to going back to the preparation phase when necessary. This is shown in the process flow in section 2.2 in Figure 2.1.

Select Modelling Techniques
It was concluded in the business understanding phase random forest and multilayer neural networks are the two classification methods that lead to the best results with air traffic data sets.

Random forests consists of a large number of individual decision trees. Each decision tree predicts a class for the input data and the class with the most votes becomes the model's prediction. It uses the so-called wisdom of crowds. Low correlation between models can produce ensemble predictions that are more accurate than any of the individual decision trees [Agg15, p. 381]. Let $\hat{C}_b(x)$ be the classification of the b^{th} tree, then the class obtained from the random forest, $\hat{C}_{rf}(x)$, is

$$\hat{C}_{rf}(x) = \text{majority vote}\{\hat{C}_b(x)_1^B\}\,[\text{CKBM16}] \tag{2.1}$$

Each model can be tuned by setting the parameters of the individual classifier. For the Random Forest classifier the important parameters are shown below.

1. Number of trees in the forest: The higher the number the better to learn data. However, adding a lot of trees slows down the process significantly as for each tree a sub-set of examples is selected via bootstrapping.

2. Criterion: This is the criterion on which attributes will be selected for splitting. This can be based on the gain ratio, the accuracy or the least square method.

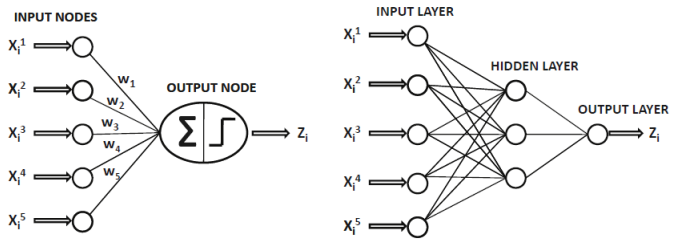
3. Maximal Depth of each tree in the forest: The deeper the tree, the more information it can capture about the data.

4. Minimal gain: The node is split when its gain is greater than the minimal gain. A high value results in fewer splits and smaller trees.

5. Minimum Leaf Size: The minimum number of samples required to be at a leaf node.

6. Minimum Samples Split: Describes the number of samples that are required to split an internal node. By increasing this parameter, each tree becomes more constrained.

Multilayer neural networks have a hidden layer, in addition to the input and output layers and the hidden layer can itself consist of multiple layers. Nodes of one layer feed into nodes of the next layer, which is then called a multilayer feed-forward network. The topology of such a network is automatically determined after the number of layers and nodes is fixed. These networks can use arbitrary functions as input like logistic, sigmoid or hyperbolic tangents. A larger number of nodes and hidden layers provides greater generality, but risks overfitting. These networks are also slow to train and sensitive to noise. However, classification is relatively efficient [Agg15, p. 330].



(a) Perceptron (b) Multilayer

Figure 2.2: Single and multilayer neural networks, [Agg15, p. 328]

The important parameters of the multilayer neural network are:

1. Activation function: An activation function can be defined for the nodes in the hidden layers. Possible functions are: Hyperbolic tangent function (Tanh) which is the same as a scaled and shifted sigmoid, Rectifier Linear Unit that chooses the maximum of (0, x) where x is the input value (Rectifier), Function that chooses the maximum coordinate of the input vector (Maxout) and the Exponential rectifier linear unit function (ExpRectifier).

2. Number of hidden layers: The number of hidden layers and nodes can be defined.

3. Dropout ratios: Defines the fraction of the inputs for the hidden layers to be omitted from training to improve generalisation.

4. Epochs: Sets the amount of iterations for the dataset. A multilayer neural network often needs thousands of epochs to be trained.

5. Learning rate: Controls how much to change the model in response to the estimated error each time the weights of the model are updated.

Generate Test Design

The modelling methods need testing and metrics to make them comparable in order to decide which model outperforms the other and why. The most common way to split the data set into a training and testing set is cross-validation. This is also used in paper [PAYLD18] for flight delays as it divides the training data into k-subsets and runs it k

times while on every run one subset is reserved for testing. Typically k is set to 10 to as it provides a good balance between computational expense and testing. To achieve realistic and comparable results tests should be done on a test set which the classifier has not used for training [HKP11, p. 368].

As the classification is a multi-class problem the important metrics are precision and recall for each of the classes [HKP11, p. 368]. Precision is the percentage of tuples labelled as positive that actually are positive. Recall is the percentage of positive tuples that are labelled as such. The formulas for the two metrics are shown in Figure 2.3. Precision is the ratio of true positive tuples (TP) and the sum of true positive and false positive (FP) tuples. Recall is the ratio between true positive tuples and the sum of true positive and false negative (FN) tuples.

$$precision = \frac{TP}{TP + FP}$$
$$recall = \frac{TP}{TP + FN} = \frac{TP}{P}.$$

Figure 2.3: Formulas for the calculation of Precision and Recall, [HKP11, p. 368]

These metrics can be displayed in a confusion matrix which is a table that indicates the number of tuples that were labelled by the classifier as a certain class together with precision and recall of each class.

Another metric is accuracy. The accuracy is the ratio of number of correct predictions to the total number of input samples. For a multi-class classification accuracy is not the best metric to describe the overall performance. However, related work do often only publish their accuracy values like [AKW+09] or [Din17]. The accuracy metric will be used in chapter 8 in the section of Contribution in comparison to existing work to make comparison between this thesis and related work easier.

Build Model

This task runs the model in the selected modelling tool. For this thesis different data mining tools have been analysed. The paper [WP16] analysed different open source tools. One of the tools is Rapid Miner, which has morphed into a licensed software product with a free and open source community. For educational purposes the license is free. Rapid Miner offers an intuitive UI to create processes and edit data sources before modelling. It contains all models needed for classification tasks and has a prediction explanation implemented that explains the predicted classes for each tuple. In addition processes can easily be sourced out to a Rapid Miner Server which uses Docker as a platform for easy OS-level virtualisation. The two aspects where Rapid Miner is limited is the use of time series analysis and big data processing. In general all analysed tools in the paper [WP16] include the functionality to perform the modelling of air traffic scenarios. Another paper about a comparative study of data mining tools puts Rapid Miner in first place for its

vast amount of functions for analysis and data handling and their parameter optimisation [RB14a]. This made Rapid Miner the tool of choice for this thesis.

Assess Model
The last task of this phase is to interpret the model from a domain knowledge point of view if it meets the business objective and data mining goals. It is likely that the data has to be adapted in the previously mentioned data preparation phase to create a better model. If the model does not produce the expected outcome, contact to domain experts might be necessary to asses the weaknesses of the model and define a way to improve it.

2.7 Evaluation

When the modelling phase produced a model that appears to have high quality from a data analysis point of view. This phase tries to evaluate if the model achieves the business objectives and interprets the results from a domain knowledge view.

Evaluate Results
Results of classification models are analysed by comparing the metrics. For a multi-class classification precision and recall values per class are interpreted if satisfactory results have been achieved. Additionally results might unveil aspects for future research.

Review Process
This task is for quality assurance of the data mining process. The goal is to review every step in order to determine any important factor that might have been left out and that the model was correctly build and only the allowed attributes have been used that make future prediction possible.

2.8 Deployment

The last phase is deployment which is all about applying the gained knowledge in the real world. Depending on the process this phase can be as simple as generating a report or as complex as implementing a repeatable process across a company.

Plan Deployment
This task is about planning the deployment strategy. It is dependent on the context of the process what deployment is necessary in the business. The plan should include all steps and how to perform them.

CHAPTER 3

Business Understanding

This chapter puts the proposed aim of the work of Chapter 1 and the findings of the current knowledge from the previous chapter into concrete objectives and data mining goals. Section 3.2 focuses on the fact-finding that helps to achieve the concrete objectives and data mining goals.

3.1 Determine Business Objective

The objective of the master thesis is to contribute something new to the current knowledge. By combining different available data sources in the field of air traffic management the classification of air traffic scenarios into risk classes should add value to the state-of-the-art. This is done by a new approach based on three aspects.

First, in comparison to delay prediction of single flights, air traffic scenarios of a certain time span containing all flights in the time span are classified which gives the air traffic controller a better situational awareness. Second, the air traffic scenario proper is enriched with environment condition data, which includes not only meteorological data, but also notice-to-airmen messages (NOTAMs). Lastly, delay of flights is derived into cost which results to the fact that delay of a large plane has a higher negative impact than the same delay of a small plane.

3.2 Assess Situation

This task assesses if all resources are available to achieve the three aspects of the approach described in the previous section. For the construction of an air traffic scenario consisting of environment conditions and the air traffic scenario proper as described in the problem statement of Chapter 1, the data sources have to be assessed.

Flight data can be collected from two different sources. The first one is EUROCONTROL, which provides only three months of data for the years 2016 and 2017 combined [EUR19]. This makes it necessary to collect flight data from a second source, which is the US Bureau of Transportation Statistics [oTS19a]. Basic information of airports like the amount of runways on a specific airport can be collected from the Federal Aviation Administration (FAA) [our19]. FAA also provides information about all aircraft [Adm19a]. Environment condition data like meteorological data is available through the Iowa State University, which collects METAR data since 2005 [Uni19] and NOTAMs are also provided by FAA for the past five years [Adm19b].

Delays can be derived into costs with the report of cost reference values created by EUROCONTROL in cooperation with the University of Westminster [oW15]. These costs focus on Europe. However, a similar report that allows the calculation of delay cost based on the size of an aircraft could not be found for the US. The goal of the classification is to set costs for large aircraft and small aircraft into relation. No exact values are evaluated such that the same report can be used for flights in the US.

3.3 Determine Data Mining Goals

The goal is to construct air traffic scenarios from historic flight data and environment condition data for airports with high capacity and delays and categorise each scenario into one of three risk classes based on the sum of the incurred delay cost. A classifier is trained on these data to be able to predict the risk class of an upcoming scenario given only the planned arrivals or departures and the environment conditions. The most important class is risk class 3 as it contains the scenarios with the highest delays. The air traffic controller has to be aware of class 3 scenarios the most and try to reduce the delay with different actions.

The process design of the CRISP-DM is iterative as the evaluation phase can lead back to the business understanding phase if the results did not satisfy the objectives and goals (cf. Figure 2.1). This was also necessary during the thesis. However, the following chapters describe each phase in its final run, when results were achieved that allowed to move on to the deployment phase.

CHAPTER 4

Data Understanding

This phase covers the initial data collection, description and exploration as well as the verification of data quality as described in chapter 2, section 2.4. To achieve the data mining goals determined in the business understanding phase all data that is needed to create air traffic scenarios, is acquired and analysed by the four tasks individually in this phase.

For the basic evaluation and statistical analysis about the raw data Java scripts were used to bring data into a format that can be used by R Studio to create statistical analysis and diagrams.

4.1 Flight Data

Flight Data - US
The United States Department of Transportation gives access to flight data back to 1987 via the website of the Bureau of Transportation Statistics to everyone. [oTS19a]. This website lets users download a report on carrier on-time performance for one month. It includes all domestic flights for carriers that account for at least one percent of scheduled passenger revenues. This means that the data depicts about 77% of US air traffic measured by the amount of passengers. This source is also used in the papers [BMTT16] and [AALB+07] which have been identified in the business understanding phase of the last chapter. For the purpose of this thesis a 24-month period (January 2016 - December 2017) was downloaded. This time period was chosen as the second data source, EUROCONTROL, only makes data from March and June of 2016 and 2017 available.

Each monthly report on carrier on-time performance from the Bureau of Transportation Statistics comes as a csv file and contains information in 28 different attributes per flight. An example of one flight record can be seen in Table 3.1 from 1 January of 2017:

No.	Attribute Name	Type	Values Range	Example
1	FL_Date	Date	2016-01-01 - 2017-12-31	2017-01-01
2	OP_CARRIER	Character	char(2)	UA
3	TAIL_NUM	Character	char(6)	N7728D
4	OP_CARRIER_FL_NUM	Integer	0 - 9999	2429
5	ORIGIN	Character	char(3)	EWR
6	DEST	Character	char(3)	DEN
7	CRS_DEP_TIME	Integer	0000 - 2359	1517
8	DEP_TIME	Integer	0000 - 2359	1512
9	DEP_DELAY	Double	-1000 - 1000	-5.00
10	TAXI_OUT	Double	-1000 - 1000	15.00
11	WHEELS_OFF	Integer	0000 - 2359	1527
12	WHEELS_ON	Integer	0000 - 2359	1712
13	TAXI_IN	Double	-1000 - 1000	10.00
14	CRS_ARR_TIME	Integer	0000 - 2359	1745
15	ARR_TIME	Integer	0000 - 2359	1722
16	ARR_DELAY	Double	-1000 - 1000	-23.00
17	CANCELLED	Double	-1000 - 1000	0.00
18	CANCELLATION_CODE	Character	char(20)	
19	DIVERTED	Boolean	1/0	0
20	CRS_ELAPSED_TIME	Double	-1000 - 1000	268.00
21	ACTUAL_ELAPSED_TIME	Double	-1000 - 1000	250.00
22	AIR_TIME	Double	-1000 - 1000	225.00
23	DISTANCE	Double	0 - 10000	1605.00
24	CARRIER_DELAY	Double	0 - 1000	
25	WEATHER_DELAY	Double	0 - 1000	
26	NAS_DELAY	Double	0 - 1000	
27	SECURITY_DELAY	Double	0 - 1000	
28	LATE_AIRCRAFT_DELAY	Double	0 - 1000	

Table 4.1: Example Record of US Flight Data

This example shows flight UA2429 from Newark, New Jersey to Denver operated by United Airlines, visible in attributes 2, 3, 5 and 6. Properties beginning with the abbreviation CRS, for Certificate of Release to Service, are planned times for departure, elapsed time and arrival. This flight departed 5 minutes early (Attribute 9) and had 18 minutes less air time (attributes 20 + 22) which resulted in a 23-minute early arrival (attribute 16). Attribute 17 shows if the flight was cancelled with the corresponding cancellation code in attribute 18. A diverted flight (attribute 19) is one that has been routed from its original arrival destination to a different destination. The last 5 attributes show the reason for a delay. This reasons can be a delay of the carrier because of problems with the aircraft, delays due to weather, delays due to the National Air System Delay like heavy traffic volume, security delays or late aircraft delays if the aircraft used for the flight had a previous delay. All these delay reasons describe delays at the departure airport.

Figure 3.1. shows the air traffic in the US per month. Each month of the two-year period includes 470,511 flights that departed or arrived at a US airport on average. February was the month with the least flights in both years and July the busiest in 2016 and August the busiest in 2017. The yearly increase of air traffic can be observed in this figure, too. These findings are consistent with other evaluations of the Bureau of Transportation Statistics like passenger numbers per month [oTS19b].

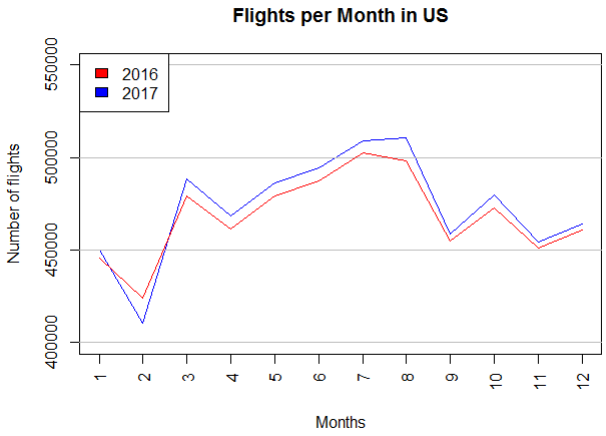


Figure 4.1: Flights per Month in US

Figure 3.2. shows total arrival delays in 2016 and 2017, which peak in July 2017 with over 8 million minutes. This evaluation includes all delayed flights of each month. Flights that have arrived earlier than scheduled have not been included. The amount of delay corresponds to the amount of flights shown in Figure 3.1.

As state-of-the-art of the business understanding phase showed, risk class prediction for air traffic scenarios has to be done for individual airports as attributes have locally varying relevance. To select an airport, a first interesting query is to evaluate where the

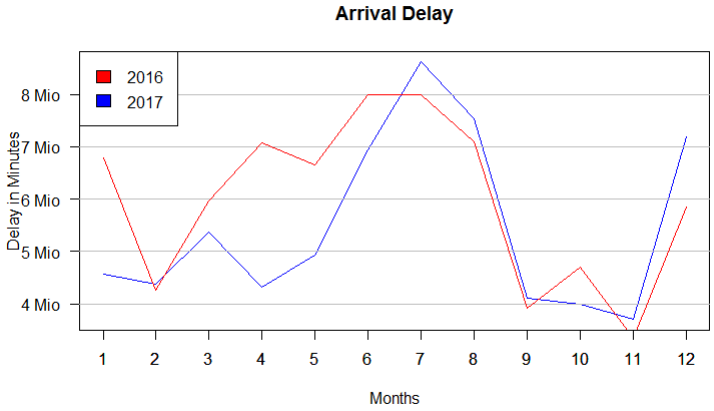


Figure 4.2: Total Arrival Delay per Month in US

biggest delays happen. One might assume that the delay measured in absolute numbers would correspond to airports with the most traffic as probability would suggest the more flights the more chances for delays. However, this simple assumption does not fully apply. Table 4.2 shows a top ten list of airports with the largest total amount of arrival delay. The list is dominated by the airports with the most flights per year. Cells that are filled in grey represent airports that are also in the top ten busiest. However, it can be observed that three of these airports are not among the busiest, and have other reasons for the large amount of delay. Airport 7 in New York is prominently mentioned in the paper [FH01] as one airport that is highly affected by weather. Between the airport of Atlanta and O'Hare, numbers one and two in the table, is quite a large jump of 13.8% in total arrival delay. However, the difference in movements is only 3.5%. It is an interesting fact, which makes Atlanta a suitable candidate for classification as the reasons for the delay might be due to the environment conditions.

Another interesting list is shown in Table 4.3. It shows the average arrival delay per flight. The computation sums up arrival delays greater than zero and creates an average with all flights that arrived at the specific airport. Small airports were discarded in this calculation. This list contradicts the assumption even more as highest delays per flight do not happen at the busiest airports.

No.	Airport	Total Arrival Delay
1	Atlanta (ATL)	3 938 293 min
2	O' Hare (ORD)	3 457 079 min
3	San Francisco (SFO)	3 073 502 min
4	Los Angeles (LAX)	2 997 929 min
5	Dallas (DFW)	2 779 141 min
6	Denver (DEN)	2 392 183 min
7	New York (EWR)	1 956 875 min
8	New York (LGA)	1 732 238 min
9	Las Vegas (LAS)	1 639 679 min
10	Boston (BOS)	1 628 219 min

Table 4.2: Total Arrival Delay for airports in the US in 2016

NR	Airport	Average Arrival Delay per Flight
1	New York (LGA)	17.93 min
2	San Francisco (SFO)	17.83 min
3	New York (EWR)	16.88 min
4	New York (JFK)	16.42 min
5	Miami (MIA)	14.31 min
6	Dallas (DFW)	14.18 min
7	O'Hare (ORD)	14.17 min
8	Los Angeles (LAX)	14.08 min
9	Fort Lauderdale (FLL)	13.47 min
10	Boston (BOS)	13.42 min

Table 4.3: Average arrival delay per flight for airports in the US in 2016

In contrast to Table 4.3, Table 4.4 shows airports with the least average arrival delay per flight. It can be observed that only one of the top ten busiest airports made it into this list. Airports with less than 3000 movements have been discarded here, too. Up to this point it can be said that a busy airport is no indicator for high delays on an average flight, but a flight with the least average delay can be mainly found at airports with miner utilization. These airports are not ideal for the risk class prediction.

The last analysis is a box plot of the average arrival delays of the 50 busiest airports in the years 2016 and 2017. It can be observed that the median increased as well as the amount of upper end outliers, which shows a raised delay overall. So the data chosen shows also that delay increases from year to year, similar to the statements in the problem statement.

The exploration of US historic flight data from the Bureau of Transportation Statistics is summed up by a short description of data quality. In summary the data had high quality and no missing records or any errors except for the five attributes of reason for delay. If

NR	Airport	Average Arrival Delay per Flight
1	Santa Ana (SNA)	7.59 min
2	Honolulu (HNL)	8.39 min
3	Chicago (MDW)	8.49 min
4	Salt Lake City (SLC)	8.90 min
5	Charlotte (CLT)	9.09 min
6	Baltimore/Washington D.C. (BWI)	9.37 min
7	Portland (PDX)	9.44 min
8	San Jose (SJC)	9.46 min
9	Seattle (SEA)	9.71 min
10	Columbus (CMH)	9.83 min

Table 4.4: Least average arrival delay per flight for airports in the US in 2016

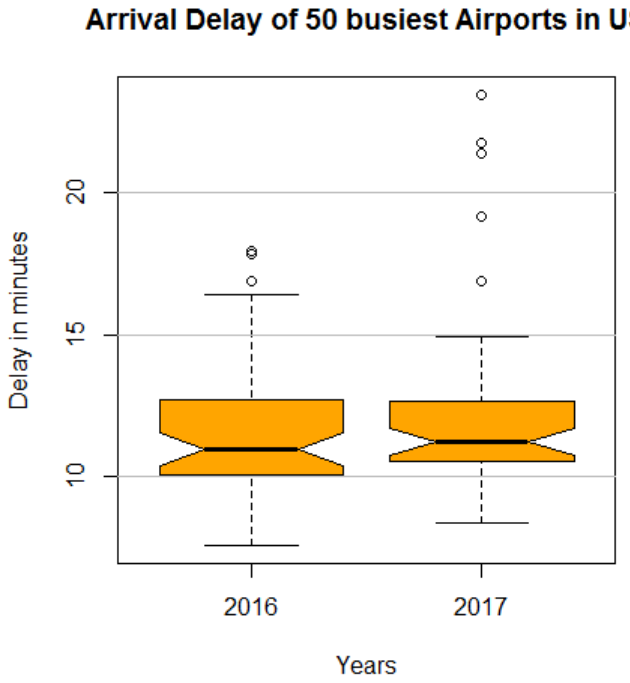


Figure 4.3: Average arrival delay per flight for 2016 and 2017

a flight had a delay these five attributes did not always explain the reason or match with the total amount of delay. The information is given by the airlines voluntarily, which made the five attributes unreliable. In terms of the other 23 attributes each record had complete information. However, there is no possibility to check if the data is complete and no flight is missing.

No.	Attribute Name	Data Type	Values Range	Example Value
1	ECTRL ID	Character	char(9)	197152513
2	ADEP (Depature Airport)	Character	char(4)	LWSK
3	ADES (Destination Airport)	Character	char(4)	LOWW
4	Filed Off Block Time	Date	01-06-2016 00:00:00 - 30-06-2016 23:59:59	01-06-2016 02:30:00
5	Filed Arrival Time	Date	01-06-2016 00:00:00 - 30-06-2016 23:59:59	01-06-2016 03:46:03
6	Actual Off Block Time	Date	01-06-2016 00:00:00 - 30-06-2016 23:59:59	01-06-2016 02:26:00
7	Actual Arrival Time	Date	01-06-2016 00:00:00 - 30-06-2016 23:59:59	01-06-2016 03:47:41
8	AC Type	Character	char(10)	A319
9	AC Operator	Character	char(4)	AUA
10	AC Registration	Character	char(5)	OELDA
11	ICAO Flight Type	Character	char(1)	S
12	STATFOR Market Segment	Character	char(30)	Traditional Scheduled
13	Requested FL	Integer	0 - 1500	280

Table 4.5: Example record for historic flight data from Eurocontrol from June 2016 data set

Flight Data - Europe

The second data source for historical flight data is the EUROCONTROL flight archive for R&D, which is the new portal that replaces the old Demand Data Repository 2 [EUR19]. When the data was needed for this thesis the new flight archive was still in its beta phase and EUROCONTROL granted access as beta user for the purpose of this thesis. During the beta tests only three months of data were available. These three months are June 2016, March 2017 and June 2017. The portal provided information about all flights managed by EUROCONTROL including international flights and transcontinental flights that just flew through European airspace. The dataset provided the following attributes per flight, which can be seen in the example below in Table 4.5. The example shows a flight operated by Austrian Airlines from Skopje in North Macedonia(attribute 2) to Vienna (attribute 3). The flight departed four minutes early (attributes 4 + 6) and arrived with about one minute delay (attributes 5+7), calculated by the difference between filed times and actual times. The AC Type (attribute 8) is the type of aircraft, and AC Registration (attribute 10) is the registration number of the aircraft. The market segment (attribute 12) is based on the aircraft type, the operator and the ICAO flight types and helps in reports and analysis to monitor and track changes in segments through time. The last attribute indicates the requested flight level from the flight plan. In this case the flight level (attribute 13) or altitude of flight was 280.

The US and the EUROCONTROL historic flight data differ in information granularity.

US flight data includes information about taxi times and the exact take-off and landing times, which the European flight data does not include. However, the relevant information of departure delay and arrival delay for the air traffic scenario creation can be calculated from attributes 4 to 7.

The three months of historic flight data included about 842,000 flights on average, split into different segments. It can be observed that 73% of flights are domestic flights in the EU which is similar to the results from US data. To compare airports between EU and US Table 4.6 shows the top ten with the highest and lowest average arrival delay per flight. The list shows airports with at least 6000 flights in the observed month June 2016 which accounts for the 75 busiest airports. This table is an analog representation to the Tables 3.2 and 3.3 from US flight data. The table shows that the top values are similar to the values from US data and that the busiest airports in the EU are also only partly represented in the top ten.

Highest Average Arrival Delay 06/16			Lowest Average Arrival Delay 06/16		
1	London (EGLL)	23.70 min	1	Bergen (ENBR)	2.56 min
2	London (EGKK)	17.46 min	2	Stavanger (ENZV)	3.59 min
3	Istanbul (LTFJ)	17.16 min	3	Warsaw (EPWA)	4.17 min
4	London (EGSS)	14.52 min	4	Budapest (LHBP)	4.45 min
5	London (EGGW)	13.39 min	5	Oslo (ENGM)	4.82 min
6	Amsterdam (EHAM)	13.32 min	6	Bucharest (LROP)	4.93 min
7	Paris (LFPG)	12.81 min	7	Goteborg (ESGG)	5.04 min
8	Frankfurt (EDDF)	12.66 min	8	Athens (LGAV)	5.19 min
9	Paris (LFPO)	12.39 min	9	Moscow (UUDD)	5.45 min
10	Brüssel (EBBR)	11.49 min	10	Kiev (UKBB)	5.67 min

Table 4.6: Highest and Lowest Average Arrival Delay per Flight for June 2016

Even though the airport of Vienna is not part of this list, the increased knowledge due to the geographical closeness makes it worth a closer look. Vienna is on place 11 with 11.44 min of average arrival delay and on place 10 of total arrival delay in Europe.

A comparison in Figure 4.4 shows the delays of the 50 busiest airports from US and Europe each based only on the data of March and June 2017. It shows that airports in Europe have smaller delays per flight. This will have effects on the scenario distribution of the three risk classes if the value ranges of each risk class are the same for European and American airports. All of these findings also coincide with the Comparison of Air Traffic Management-Related Operational Performance Report produced by EUROCONTROL and FAA of 2017 [EUR17].

Similar to the US data, data from EUROCONTROL is of high quality. No missing fields could be found in any record. Data was as complete as available and no other issues were found with data quality.

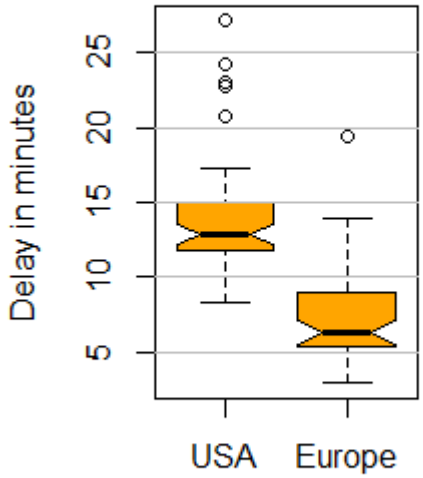


Figure 4.4: Box Plot of Airports in Europe and USA from March and June 2017

4.2 Meteorological Data

Meteorological data were acquired from a website of Iowa State University [Uni19]. This website has been collecting the Meteorological Terminal Aviation Routine Weather Reports (METAR) data from all airports around the world since 2000. It also provides a script written in R to download METAR data for a single airport. For the purpose of this thesis the script was adapted to be able to download weather data for several airports at once. To do so the four-letter ICAO code for all 100 airports (50 U.S. and 50 European airports) was necessary. This code differs from the more common known three letter IATA code which can be found on every ticket or reservation system. Examples are "VIE" for the airport in Vienna or "FRA" for the airport in Frankfurt in Germany. On the technical operations side for air traffic management or planning the ICAO code is more commonly used. The four-letter code is more complex as the code for Vienna is "LOWW" and for Frankfurt "EDDF". Thus the ICAO codes have been added manually to the initial list of all airports. The list in combination with the adapted R script made it possible to download all the METAR data.

The data comes in text files divided per ICAO code where each attribute is separated by a colon. The time period for this data is the same as for flight data from January 2016 until December 2017. The result was 50 text files for the US airports and 50 text files for

the EU airports. Each file of a European airport has 48 records per day as METAR data is published every 30 minutes, adding up to 17,520 records per year. METAR data at US airports was published 24 times per day which leads to 8,760 records per year. Each record consist of 31 properties including information about temperature, wind speed and direction, special weather phenomena like snow or rain and other meteorological information shown in the example in Table 4.7.

Most fields are self-explanatory. All temperatures are given in Fahrenheit, altitudes in feet and speeds in knots. The fields I want to explain in more detail are the following: Wind direction (8) is given in degrees from north which results in this example in a south-south west wind direction as it is 210 degrees. Sky Level Coverage (15-18) is reported by the number of oktas of the sky that is occupied by clouds. In meteorology okta is a measurement for the cloud cover where 0 okta means sky completely clear, 4 oktas means sky half cloudy and 8 oktas means that the sky is completely cloudy. So in this case "FEW" stands for 1-2 oktas, "SCT" means scattered so 3-4 okta and "OVC" means overcast which is 8 oktas. The Present Weather Code (23) says SG for snow grains and BR for Mist. Missing values are indicted by "M", which is either reported missing or it was never received by the sensor. Additionally each record also includes the unprocessed reported observation in METAR format.

Given the flight data and the meteorological data together, a combined analysis is possible. Figure 4.5 shows a scatter plot with the amount of rainy days on the x axis and the average wind speed on the y axis. The data is from 2017 and includes all US airports listed in the data irrespective of size or air traffic. Red dots indicate airports with arrival delays per flight of more than 20 minutes. The horizontal grey line splits the airports

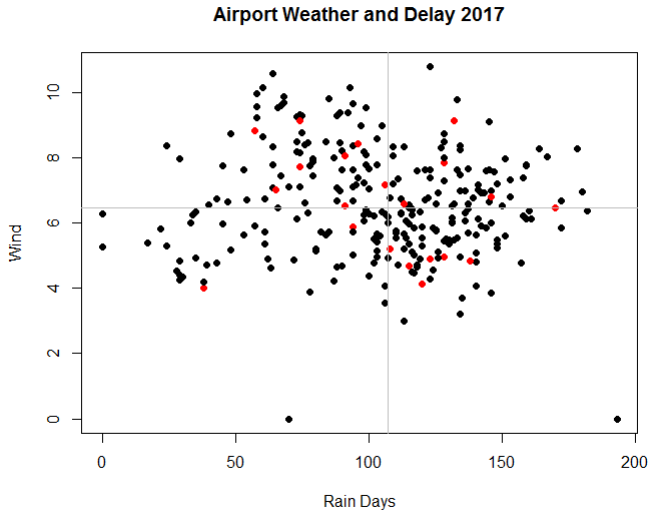


Figure 4.5: Scatter plot of US airports from 2017 based on rainy days and wind speed

in half. 50 percent have more than 6.48 miles per hour of average wind speed and 50

No.	Attribute Name	Data Type	Values Range	Example Value
1	Station	Character	char(4)	EDDF
2	Valid	Date	2016-01-01 00:00 - 2017-12-31 23:59	2017-01-01 00:20
3	Longitude	Character	char(6)	8.5986
4	Latitude	Character	char(6)	50.0464
5	Air Temperature	Double	- 5 - 100 F	24.80 F
6	Dew Point Temperature	Double	- 5 - 100 F	23.00 F
7	Humidity	Double	0 - 100 %	92.74 %
8	Wind Direction	Double	0 - 360	210.00
9	Wind Speed	Double	0 - 100	3.00 knots
10	1 hour precipitation	Double	0 -100	0.00
11	Pressure Altimeter	Double	0 - 100	30.36 inches
12	Sea Level Pressure	Double	0 - 2000	M
13	Visibility in Miles	Double	0 - 10	1.62
14	Wind Gust	Double	0 - 100	M
15	Sky Level Coverage 1	Character	char(3)	FEW
16	Sky Level Coverage 2	Character	char(3)	SCT
17	Sky Level Coverage 3	Character	char(3)	OVC
18	Sky Level Coverage 4	Character	char(3)	M
19	Sky Level 1 Altitude	Double	0 - 10000	100.00
20	Sky Level 2 Altitude	Double	0 - 10000	200.00
21	Sky Level 3 Altitude	Double	0 - 10000	300.00
22	Sky Level 4 Altitude	Double	0 - 10000	M
23	Present Weather Code	Character	char(10)	-SG BR
24	Apparent Temperature	Double	- 5 - 100 F	M
25	Ice Accretion 1h	Double	0 - 100	M
26	Ice Accretion 3h	Double	0 - 100	M
27	Ice Accretion 6h	Double	0 - 100	M
28	Peak Wind Gust	Double	0 - 100	M
29	Peak Wind Gust Direction	Integer	0 - 360	M
30	Peak Wind Gust Time	Date	00.00 - 23.59	20.46
31	Unprocessed METAR	Character	char(100)	EDDF 010020Z 02004KT CAVOK 16/12 Q1021 NOSIG

Table 4.7: Example data record of METAR data

No.	Attribute Name	Data Type	Values Range	Example Value
1	ICAO	Character	char(4)	LOWW
2	type	Character	char(15)	large airport
3	name	Character	char(30)	Vienna International Airport
4	elevation_ft	Integer	0 - 15000	600
5	country	Character	char(2)	AT
6	iata_code	Character	char(3)	VIE
7	amount of runways	Integer	0 - 5	2
8	Runway 1 Length	Integer	0 - 14000	11483 f
9	Runway 1 Width	Integer	0 - 300	148 f
10	Runway 1 Surface	Character	char(3)	ASP
11	Runway 1 Lighted	Boolean	1/0	1
12	Runway 1 Closed	Boolean	1/0	0
13	Runway 1 Heading	Integer	0 - 360	116
14	Runway 2	....	....	....

Table 4.8: Airport and runway data from Vienna International Airport

percent have less. As well the vertical grey line splits the group in half at 107 rainy days. This leads to four groups with the same size. What can be observed is that airports with a large delay of more than 20 minutes appear the most in the group with high wind and/or high rainy days. This supports the conclusion that weather has an impact on arrival delay.

Data quality of raw METAR data was satisfactory, the data included information from every day of the two-year period. The fact that many attributes have already been extracted from the raw METAR message like temperature, wind speed, etc. was very helpful. If a value was missing it can be extracted from the unprocessed METAR record (attribute 31) manually.

4.3 Airport and Runway Data

Airport and runway data was collected from a website called ourairports.com which offers data about all airports in the world and all runways free to download as open data [our19]. The csv file included 55,485 entries including small, medium and large airports as well as heliports, balloonports and closed airports. Other information for every record is location given in longitude and latitude coordinates, elevation in feet, continent and country, municipality, IATA code, ICAO code and local code. The csv file about runways included 41,383 entries with information about ID, length, width, surface and lighted as the most important attributes. In a first step airport and runway data was merged into one csv file. An example record of the merged data is shown in Table 4.8.

Vienna International Airport is a large airport at 600 feet above sea level and has two

No.	Attribute Name	Data Type	Values Range	Example Value
1	Date Completed	Date	2016-Jan-1 - 2019-Dec-31	2016-Feb-29
2	Manufacturer	Character	char(40)	Airbus
3	Model	Character	char(30)	A320-200
4	Physical Class (Engine)	Character	char(30)	Jet
5	# of Engines	Integers	1 - 4	2
6	AAC	Character	char(1)	C
7	ADG	Character	char(3)	III
8	TDG	Character	char(2)	3
9	Approach Speed	Integer	0 - 300	136
10	Wingtip Configuration	Character	char(20)	wingtip fences
11	Wingspan	Double	0 - 300	111,88
12	Length	Double	0 - 200	123,27
13	Tail Height	Double	0 - 100	39,63
14	Wheelbase	Double	0 - 100	41,47
15	Cockpit to Main Gear	Double	0 - 100	50,20
16	MGW	Double	0 - 100	29,36
17	MTOW	Integer	0 - 500 000	171 961
18	Max Ramp Max Taxi	Integer	0 - 500 000	172 842
19	Main Gear Config	Character	char(1)	D
20	ICAO Code	Character	char(10)	A320
21	Wake Category	Character	char(1)	M
22	ATCT Weight Class	Character	char(20)	Large Jet Eqpt
23	Years Manufactured	Character	char(20)	tbd
24	Note	Character	char(50)	tbd
25	Parking Area (WS x Length)	Integer	0 - 100 000	13 791

Table 4.9: Example of the aircraft data: Airbus A320

runways, which are both stored in the last attribute. Each runway is described with 6 more attributes. The first is 11,483 feet long, 148 feet wide, made from asphalt, is lighted and not closed and heads 116 degrees from north.

Data quality of airport and runway data was high. All airports and runways had no missing values.

4.4 Aircraft Data

Aircraft data was downloaded from the aircraft characteristics database of the Federal Aviation Administration (FAA) [Adm19a]. The file is in xls format and includes information about 2,766 different aircraft types. Each aircraft is described with 25 attributes. An example of the Airbus A320 record from the data set is shown in Table 4.9.

No.	Attribute Name	Data Type	Values Range	Example Value
1	Location	Character	char(3)	ATL
2	NOTAM #	Character	char(20)	01/182
3	Class	Character	char(30)	Aerodrome
4	Issue Date (UTC)	Date	01/01/2017 00:00 - 12/31/2017 23:59	01/20/2017 1902
5	Effective Date (UTC)	Date	01/01/2017 00:00 - 12/31/2017 23:59	01/23/2017 0430
6	Cancel Date (UTC)	Date	01/01/2017 00:00 - 12/31/2017 23:59	
7	Expiration Date (UTC)	Date	01/01/2017 00:00 - 12/31/2017 23:59	01/23/2017 1130
8	NOTAM Condition	Character	char(100)	!ATL 01/182 ATL RWY 09R/27L CLSD

Table 4.10: Example a notice-to-airmen message from Atlanta

Even though the data set includes all aircraft needed according to the historic flight data sets from the US and Europe, it had a lot of "to be done" values. If information of a certain aircraft is needed these values cannot be replaced with zero or the average from other records. Missing aircraft information attributes have to be added manually by researching the correct value.

4.5 NOTAMs

Notice to airmen messages have been collected by the Federal Aviation Administration (FAA) for years. They provide an archive search for the last five years on their website [Adm19b]. The website provides no API and displays NOTAMs for one day. So every day needs to be downloaded manually. Therefore this data source was collected lastly, even after the selection process of the data preparation phase in chapter 4. The two offered formats are pdf and xls. 365 xls files have been downloaded and converted to a csv for the airport of Atlanta for 2017 and 91 xls files for the airport of Vienna. Each file consists of all messages that were active on this day. An example of a NOTAM is shown in Table 4.10.

The example NOTAM from the airport of Atlanta describes that the runway 09R/27L is closed (attribute 8) from 23 January 2017 at 04:30 (attribute 5) to 11:30 of the same day (attribute 7). This information was already issued on 20 January at 19:02 (attribute 4). Other NOTAMs can contain information about airspace closures, taxiway closures, defective lightning on taxiways or runways, defective precision approach path indicators or any other important information than can influence the approach to an airport. Data quality of notice-to-airmen messages was high and no missing fields were identified.

CHAPTER 5

Data Preparation

The phase data preparation consists of five steps which prepare the data for the modelling phase. First, all data that is being used in this phase has to be listed including a description of anything that is excluded. This is followed by data cleaning to increase data quality along with the reasons for certain decisions in case of data transformations. Then the data is ready for construction, which adds new attributes created by transforming existing ones or even adds new records. The integration step is the most important one as this merges or combines tables to create new records and values necessary for the modelling phase. Lastly the data needs to be formatted as some tools have specific requirements for input data like order of attributes or that the first field needs to be a unique identifier.

5.1 Data selection

Flight Data
Current knowledge described in the business understanding phase showed that different attribute at airports have a high locally varying relevance. For that reason a classifier will be trained only on single airports. The data understanding phase showed that in the comparison of US airports the airport of Atlanta is a suitable candidate for air traffic scenario classification. It is the largest airport in the US, but the amount of delay could not be explained only by its traffic. From the historic flight dataset of Europe, the airport of Vienna was chosen. Its calculated delay figures in the data understanding phase and its geographical closeness make it a suitable airport for classification.

Experiments with different sizes of air traffic scenario data sets have shown that large data sets are very slow on the provided hardware. This led to the decision to reduce data set sizes. For the airport of Atlanta only data from 2017 is selected and for Vienna all three-months of data. This increases the speed of the classification makes parameter optimisation possible. Details about parameter optimisation are described in chapter 5 Modelling.

Meteorological Data
As for flight data two airports are selected, for meteorological data the same two airports and time span are selected. The result of the selection is the airport of Atlanta for the year 2017 and Vienna for June 2016, March 2017 and June 2017.

Airport and Runway Data
For airport and runway data the airport of Atlanta and the airport of Vienna are selected.

Aircraft Data
For Atlanta a list of all tail numbers is created. This list includes 3,059 distinct numbers. The tail number of an aircraft is similar to a license plate of a car. It is possible to match the number with the aircraft model. The distinct list of tail numbers reduces the list of aircraft to all that have arrived or departed in Atlanta in 2017.

For Vienna a list of all model numbers is created. This list includes 218 distinct model numbers. As the historic flight data included the model number of the aircraft in every record, the list can directly be matched with the aircraft data. Aircraft data was reduced to the aircraft models that have arrived or departed in Vienna in the months of June 2016, March 2017 and June 2017.

NOTAMs
NOTAMs were collected last, after the two airports have been selected as the NOTAM collection had to be done manually for each day needed. That is the reasons why no further selection is necessary. NOTAMs have been collected for Atlanta for 12 months and for Vienna for 3 months.

5.2 Data Cleaning

Flight Data
The cleaning process includes two steps for flight data. The first is about all flights that do not have a valid tail number or model number. A non-valid tail or model number is defined by the fact that no matching aircraft could be found. Reasons for that are typos in the numbers, or that the aircraft is a helicopter or balloon. If no aircraft information is available, the maximum take-off weight is unknown, which is necessary for the delay cost calculation. Therefore these flights are flagged by setting the maximum take-off weight to 9,999,999. This unique value highlights these flights for the scenario creation in the data construction task, where they are processed individually.

The second step is to subtract departure delay of arriving flights from the total delay. If an aircraft is arriving at Atlanta or Vienna the reason for delay at the departure airport is unknown. It might be due to certain environment conditions at Vienna or Atlanta, but this is only one of many possible reasons like security, technical failure, waiting for crew, etc. This departure delay is defined by the time between the filed off block time (attribute 4) and the actual off block time (attribute 7) as shown in Table 3.5 for historic

flight data of Europe. For US flight data the departure delay is stored in dep delay (attribute 9) as shown in Table 3.1.

Meteorological Data
Meteorological data included many missing values for the attributes 5 - 30 shown in Table 4.7. The first reason is that the unprocessed METAR (attribute 31) did not include the information. For example an ice accretion (attributes 25 - 27) is very unlikely in summer. The second reason is that the automatic transformation of the unprocessed METAR to the attributes done by the data provider did not work correctly. In the second case it is possible to fill the missing values with the correct ones from the METAR. However, not all attributes can be extracted from the METAR. Sea level pressure (attribute 12) was dismissed as it was missing in 92.3% of records and cannot be extracted from the unprocessed METAR. This applies also to the apparent temperature (attribute 24), which was missing in 91% of records. Sky Level Coverage 2, 3 and 4 (attributes 16, 17 and 18) were missing in 73.06%, 87.3% and 98.7% as the METAR only included information about Level 1 in 95% of records. Experiments in the modelling phase (cf. Chapter 5 Modelling) have shown that only Sky Level Coverage 1 had an effect on prediction. Therefore these attributes were dismissed in the final construction of air traffic scenarios. This also applies to the Sky Level Altitude 2, 3 and 4 (attributes 20, 21 and 22). Ice accretion 1h, 3h and 6h (attributes 25, 26 and 27) and peak wind gust (attribute 28), peak wind gust direction (attribute 29) and peak wind gust time (attribute 30) were all missing in 99% of records. For the few records that included information about ice accretion and peak wind gust time/direction these environment conditions were expressed in other attributes in the final air traffic scenario (cf. section 4.3 Data Construction). In total no METAR record had to be removed completely, which is important because if no METAR data is available for an air traffic scenario the whole scenario has to be dismissed.

Airport and Runway Data
For airport and runway data no cleaning was needed as it included all necessary values.

Aircraft Data
The needed attributes of the aircraft data are the model (attribute 3), the maximum take of weight (MTOW, attribute 17) and the ICAO code (attribute 20). All attributes are shown in Table 4.9. From the selected aircrafts of the previous step about 35% of aircraft had a missing value for the maximum take of weight. This information has to be added manually by searching each aircraft model on the website skybrary.aero. This website is a library about every aircraft model and includes all facts including the maximum take of weight.

NOTAMs
After the conversion of the data format from xls to csv, which was done in the previous phase, no cleaning was necessary.

5.3 Data construction and integration

In theory the data construction task is about deriving new attributes and the integration task is about combining multiple tables into new records. The creation of air traffic scenarios and calculation of risk classes requires steps from both tasks. To be able to describe these in a comprehensible way the two tasks are combined.

5.3.1 Scenario creation

METAR Data
The base of each scenario is METAR data as it defines the period of time and the meteorological information of a scenario. An air traffic scenario beginning is the time when a METAR record is published and the scenario end is before the next METAR record is published. This sets the time spans to one hour for scenarios at Atlanta airport and to 30 minutes for Vienna airport. From each METAR record the attributes which resulted from the cleaning phase are copied without further transformation to the scenario record:

- Start time and date of scenario: When the METAR information is published. (attribute 2)
- End time and date of scenario: Before the next METAR information is published. (attribute 2)
- Air temperature in Fahrenheit (attribute 5)
- Dew Point Temperature in Fahrenheit (attribute 6)
- Relative humidity in % (attribute 7)
- Wind direction in degrees from north (attribute 8)
- Wind speed in knots (attribute 9)
- Amount of precipitation in the last hour in inches (attribute 10)
- Visibility in miles (attribute 13)
- Wind gust in knots (attribute 14)
- Sky Level Coverage 1 (attribute 15).
- Sky Level Altitude 1 in feet (attribute 19)

The unprocessed METAR (attribute 31) is used to extract further information about special weather phenomena. It includes abbreviations for different phenomena like rain, thunderstorms, shower, hail, snow, fog and wind shear. These seven attributes are added to the scenario as boolean values. They are shown in the air traffic scenario example in Table 5.3.

NOTAMs
Secondly the NOTAM data is added to the scenario. As there is one NOTAM file for each day, the information of the file corresponding to the date of the scenario is added. For the airport of Atlanta ten attributes are added to the air traffic scenarios from NOTAM data. The amount of attributes depends on the amount of runways and relevant taxiways.

- Precision Approach Path Indicator (PAPI): This is a boolean attribute containing the information if the PAPI is in service or out of service.
- Runway 1: This is a boolean attribute containing the information if this runway is open or closed.
- Runway 2: Equal information to Runway 1.
- Runway 3: Equal information to Runway 1.
- Runway 4: Equal information to Runway 1.
- Runway 5: Equal information to Runway 1.
- Taxiway B: This is a boolean attribute containing the information if this taxiway is open or closed.
- Taxiway B1: Equal information to Taxiway B.
- Taxiway B2: Equal information to Taxiway B.
- Taxiway B3: Equal information to Taxiway B.

Experiments in the modelling phase have shown that only taxiways B - B3 have an positive impact on the prediction of risk classes.

For the airport of Vienna NOTAM data only adds 3 attributes. Due to the fact that Vienna has only two runways and no relevant taxiways could be identified for prediction.

- Precision Approach Path Indicator (PAPI): This is a boolean attribute containing the information if the PAPI is in service or out of service.
- Runway 1: This is a boolean attribute containing the information if this runway is open or closed.
- Runway 2: This is a boolean attribute containing the information if this runway is open or closed.

Flight Data
With a sort and merge algorithm all flights that occurred during the time span of the scenario have been identified. The flights are split into three attributes: small aircraft, medium aircraft and large aircraft. The size of an aircraft depends on the maximum take-off weight (MTOW). The range of all MTOWs is split into three equally sized groups. A small aircraft has a weight of up to 70,000 pounds, a medium aircraft of up to 200,000 and a large aircraft has a weight of more than 200,000 pounds. Flights that have been highlighted in the data cleaning task by setting the maximum take-off weight to 9,999,999 get an artificial take-off weight calculated by the average take-off weight of all flights in the particular scenario. Based on this weight the flight is added to one of the three groups. If a scenario consists of more than 5% of highlighted flights (with an unknown take-off weight) the whole scenario is removed from the final dataset. By this rule 9.1% of departure scenarios and 11.7% of arrival scenarios were removed.

In the time span of the scenario the planned arrival and departure distance is calculated, as very close distances are likely to cause congestion. This is done by splitting the time span into five minute blocks and counting how many aircraft depart and arrive during each block. Then the amount is divided by five to get the average per minute of each block. The block with the highest number is added as arrival or departure distance attribute into the scenario. Independently if the particular scenario is about arriving or departing aircraft, the arrival and departure distance attributes are added. The reasons is, that a peak in arrival distance can lead to congestion for departing aircraft as they might have to wait for a free take-off slot on the runway. For Atlanta the arrival distance peaks at 4.8 aircraft per minute. In Vienna the peak is 2.8 aircraft per minute.

Lastly the scenario type is defined, if the flight information is about arriving or departing aircraft. Weather and NOTAM information are equal in both types.

5.3.2 Calculate Risk Class

The last missing attribute to complete an air traffic scenario is the risk class. It is calculated from the cost of all delayed flights of a scenario. To calculate the delay cost of each flight the report of cost reference values is used [oW15, p. 79]. This report was done by EUROCONTROL in cooperation with the University of Westminster. In annex J of the report a regression analysis is proposed that allows to calculate the delay cost for each aircraft based on its maximum take off weight in tons. The linear regression function is of the form

$$y{=}m \cdot x + c \tag{5.1}$$

The function is based on the twelve most common aircraft. Variable x of the linear regression function is the root of the maximum take off weight. There have been two sets of regression coefficients calculated. One for the full tactical costs including reactionary cost for at-gate delay, also known as departure delay and one for en-route delay also

known as arrival delay. Each set includes the coefficients m and c for nine delay categories from below 5 minutes to more than 300 minutes. In Figure 5.1 the coefficients are shown for at-gate delay and in Figure 5.2 the coefficients for en-route delay.

Delay (mins)	5	15	30	60	90	120	180	240	300
r^2	0.961	0.982	0.986	0.987	0.988	0.977	0.970	0.977	0.981
m	12.5	71.6	260	1233	3358	7371	10583	12942	15781
c	-32.9	-178	-663	-3432	-9315	-25015	-38440	-42327	-45932

Figure 5.1: At-gate regression coefficients, [oW15, p. 80]

Delay (mins)	5	15	30	60	90	120	180	240	300
r^2	0.984	0.991	0.994	0.993	0.992	0.981	0.975	0.981	0.986
m	42.9	163	443	1599	3907	8104	11683	14408	17614
c	-147	-524	-1348	-4801	-11367	-27759	-42551	-47809	-52793

Figure 5.2: En-route regression coefficients, [oW15, p. 80]

The delay values of each flight are taken from the historic flight data. For US data the arrival and departure delay of a flight is stored in the attributes dep delay (attribute 9) and arr delay (attribute 16). For data from Europe the delay values have to be calculated. Departure delay is the difference between the filed off block time (attribute 4) and actual off block time (attribute 6). Arrival delay is the difference between the filed arrival time (attribute 5) and the actual arrival time (attribute 7). By knowing the delay the right coefficients can be chosen to calculate delay costs. The maximum take of weight is taken from the aircraft data set. An example calculation for the example of flight data in Table 4.5 would look like following: The arrival delay of this flight was 1.5 minutes and the aircraft model was an A319. The maximum take of weight of this aircraft is 76,500 kilograms. For this calculation the coefficients of the below five minute en-route delay are necessary.

$$y{=}42.9 \cdot \sqrt{76,5} - 147 \tag{5.2}$$

The result of this formula is EUR 228.22 of delay cost for this flight. The costs evaluated in the report are for European flight traffic only. There could be no similar report found for cost evaluation in the US. For the classification of air traffic scenarios it is only important to set different flights and aircraft in a relation to each other. A large aircraft does not have the same cost as a small aircraft and should not have the same impact on the delay class. The real and exact costs for each flight are therefore not important. This is the reason why the linear regression and its coefficients have been used for US delay cost as well.

Region and Year	Min	0.25 Q	Median	Average	0.75 Q	Max
Arr Delay Cost Atlanta	0	31	86	581.08	249	959,523
Dep Delay Cost Atlanta	0	68	368	3424.28	2528.5	413,133
Arr Delay Cost Vienna	0	84	205	307.48	366	24,028
Dep Delay Cost Vienna	0	32	157	442.95	391	23,155

Table 5.1: Quantiles and Average of the Arrival Departure Costs for Atlanta and Vienna

Risk Class	Cost Range	Average Minute Range
1	EUR 0 - 1000	0 - 15 minutes
2	EUR 1001 - 9500	15 - 60 minutes
3	EUR >9500	>60 minutes

Table 5.2: Risk classes for air traffic scenarios

For each scenario the delay costs are divided by the total amount of aircraft to get an average delay cost per aircraft for each scenario. This way each scenario, independently of the amount of aircraft, can be compared with each other. The range of average delay costs now needs to be split into the three risk classes. The quantiles of the value range are shown in Table 5.1.

The high delay costs of up to EUR 959,523 are due to flights that had up to 900 minutes delay by having a planned arrival at 09:05 am and an actual arrival at 11:53 pm. The risk classes are then set to the following three ranges as shown in Table 5.2. The delay in minutes shown in column three is an average over all aircraft. Based on this table every air traffic scenario is put into one of the three risk classes which is the final attribute for each scenario record.

5.4 Data formation

This task is primarily for syntactic modifications to the data that might be required by the modelling. For the chosen data mining tool Rapid Miner no further reformatting was necessary. An example of the finished air traffic scenario data tuple is shown in Table 5.3.

One last analysis of the finished scenario datasets is about class distribution. Figure 5.3 shows how many scenarios are in each of the three delay classes over the two datasets. The bar plots show that the class distribution is highly imbalanced. For Atlanta only 624 scenarios are in risk class 3 and for Vienna only 153.

The class imbalance is solved with undersampling. The classifiers in the modelling phase will have a data set with 647 scenarios for each class for training and testing. Otherwise the classifier will get biased in the training phase towards the most common class.

No.	Attribute Name	Data Type	Value Range	Example Value
1	Airport	Character	char(3)	ATL
2	Scenario Begin	Date	01012017 00:00 - 31122017 23:59	230117 08:52
3	Scenario End	Date	01012017 00:00 - 31122017 23:59	230117 09:51
4	Small Aircraft	Integer	0 - 100	13
5	Medium Aircraft	Integer	0 - 100	51
6	Large Aircraft	Integer	0 - 100	2
7	Arrival Distance	Double	0 - 5	2.6
8	Departure Distance	Double	0 - 5	2.8
9	Scenario Type	Boolean	0/1	0
10	Air Temperature	Double	-5 - 100	48.00 F
11	Dew Point	Double	-5 - 100	46.90 F
12	Humidity	Double	0 - 100	95.95 %
13	Wind Direction	Integer	0 - 360	170
14	Wind Speed	Double	0 - 100	6
15	1h Precipitation	Double	0 - 100	0.12
16	Visibility	Integer	0 - 10000	3000
17	Wind Gust	Double	0 - 100	0.0
18	Sky Level Altitude	Integer	0 - 10000	800
19	Sky Level Coverage	Character	char(3)	BKN
20	Wind Shear	Boolean	YES/NO	NO
21	Thunderstorm	Boolean	YES/NO	NO
22	Rain	Boolean	YES/NO	YES
23	Shower	Boolean	YES/NO	NO
24	Snow	Boolean	YES/NO	NO
25	Hail	Boolean	YES/NO	NO
26	Fog	Boolean	YES/NO	NO
27	P.A.P.I.	Boolean	In service/out of service	out of service
28	Runway 1	Boolean	Open/Closed	Closed
29	Runway 2	Boolean	Open/Closed	Closed
30	Runway 3	Boolean	Open/Closed	Open
31	Runway 4	Boolean	Open/Closed	Open
32	Runway 5	Boolean	Open/Closed	Open
33	Taxiway B	Boolean	Open/Closed	Open
34	Taxiway B1	Boolean	Open/Closed	Open
35	Taxiway B2	Boolean	Open/Closed	Open
36	Taxiway B3	Boolean	Open/Closed	Open
37	# Open Runways	Integer	0 - 5	3
38	Risk Class	Enum	1 / 2 / 3	3

Table 5.3: Example of a finished air traffic scenario tuple

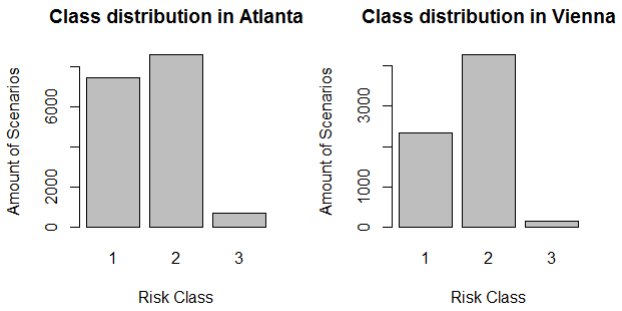


Figure 5.3: Risk Class distribution from the two created scenario datasets

CHAPTER 6

Modelling

The modelling phase is done with help of the tool Rapid Miner version 9.5 as described in section 2.6. In Rapid Miner, training and testing of a classifier is done with a so-called process and these are executed on Rapid Miner Server version 9.5 in a virtual machine with 8 GB RAM. Processes were executed in parallel on the client version too on a laptop with 8GB RAM and an Intel Core i7-4500U with 1.80Ghz. The tool offers to create processes by connecting different operators, similar to workflows. All processes executed for this thesis are described in this chapter.

6.1 Random Forest Classifier

The first classifier that is trained is the random forest classifier. It has been described in section 2.6 including the important parameters. In Rapid Miner two processes are set up to train and test with air traffic scenarios from Atlanta and Vienna. The process is shown in Figure 6.1.

This process starts by retrieving the air traffic data set. The Set Role operator sets the risk class attribute as the attribute that should be classified. Then the Sample operator balances the data set and reduces the records of class 1 and 2 to the amount of class 3. In the case of Atlanta to 624 records, and in the case of Vienna to 153 attributes. Now the data is ready for classifier training. The Optimize Parameter operator varies the parameters of the random forest classifier for every run and consists of a sub-process. In the sub-process is the Cross Validation operator, which splits the dataset into k-folds for training and testing as described in section 2.6. The sub-process is shown in Figure 6.2. For the parameter optimisation the following parameters are varied:

1. Number of trees in the forest: Possible number of trees is set from 0 to 100 divided into 10 steps.

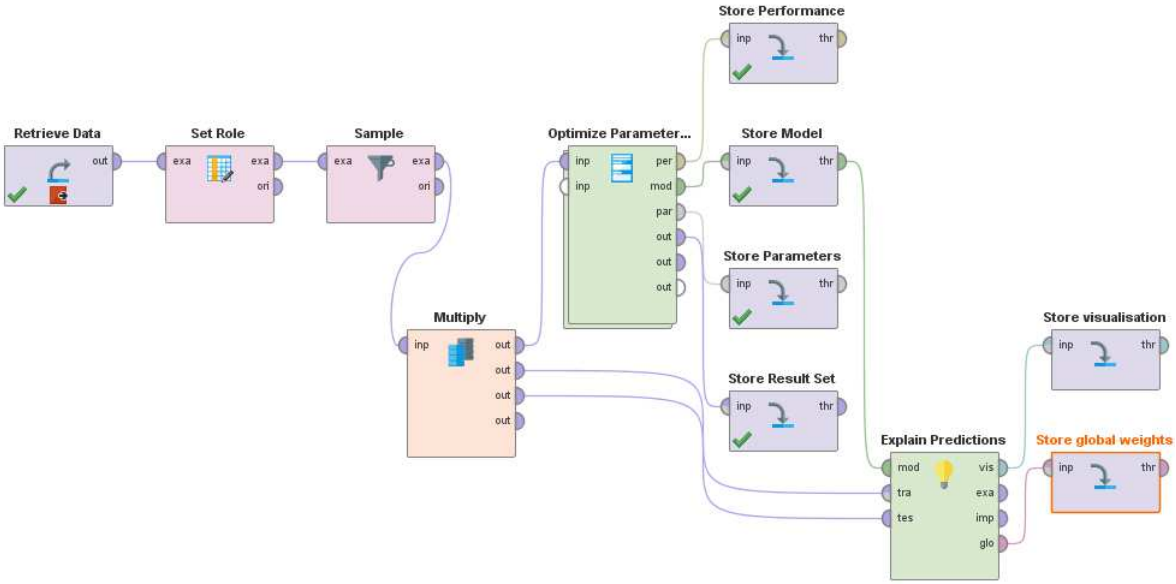


Figure 6.1: Random Forest process in Rapid Miner for air traffic scenario training and testing

2. Maximal Depth of each tree in the forest: Possible depth is set from -1 to 100 divided into 5 steps.

3. Minimum Leaf Size: Possible leaf size is set from 1 to 100 divided into 5 steps.

This parameter optimisation settings lead to 396 possible combinations of parameters and training runs. The running time of all combinations was about 6 hours for the larger dataset on the given hardware. The possibility to try more combinations was limited by the available random access memory.

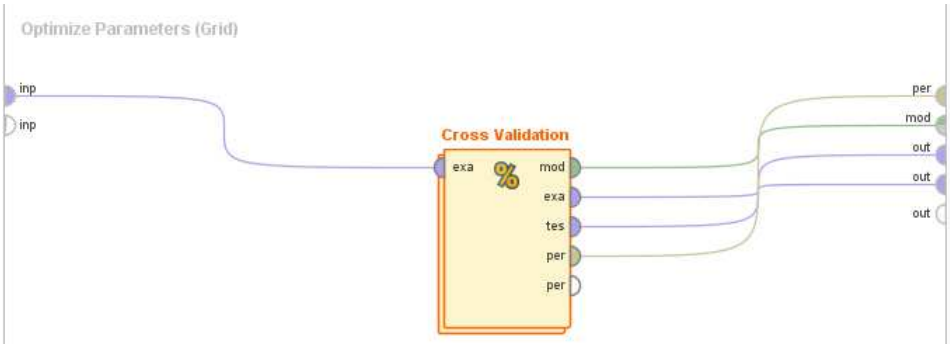


Figure 6.2: Sub-process of the Optimize Parameter operator of the Random Forest process in Rapid Miner

The Cross Validation operator consists of a sub-process itself where the Random Forest classifier is trained, tested and the performance is measured with the metrics accuracy, precision and recall. The metrics have been defined in section 2.6 Generate Test Design. This sub-process is shown in Figure 6.3.

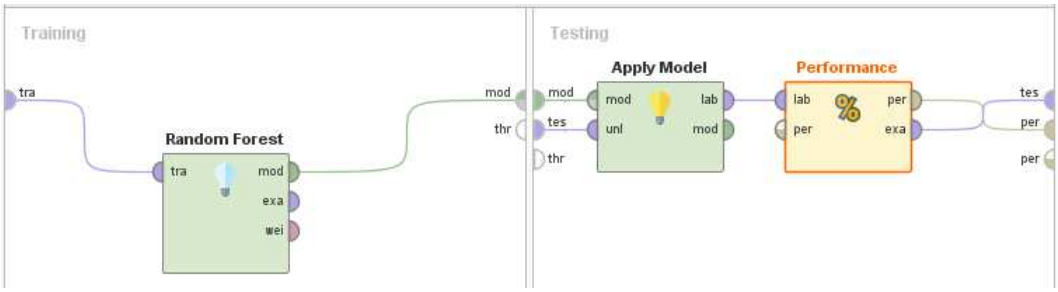


Figure 6.3: Sub-process of the Cross Validation operator of the Random Forest process in Rapid Miner

After the model is finished with training, testing and parameter optimisation the best model moves on to the Explain Predictions operator shown in Figure 6.1. This operator visualizes which attributes were the most important for classification and calculates global weights for every attribute of the air traffic scenario.

Atlanta Dataset

The air traffic scenario dataset for the airport of Atlanta consists of 15,064 scenarios. The Sample operator undersamples the scenarios to 624 per risk class which adds up to 1,872 scenarios for training and testing. A summary of the process for random forest classifier training on the Atlanta dataset is shown in the table below.

Random Forest - Atlanta	
Input Data:	15,064 scenarios from 2017
# Scenarios per Class:	1: 6,344, 2: 8,096, 3: 624
Undersampling:	Delay Class 1 and 2 reduced to the amount of class 3
Oversampling:	None
# Scenarios for Training and Testing	624 scenarios per class
Validation Method	10-fold Cross Validation
Parameter Optimization	
Number of Trees	0 - 100 in 10 steps
Maximal Depth	-1 - 100 in 5 steps
Minimum Leaf Size	1 - 100 in 5 steps

Table 6.1: Process Summary of Random Forest classifier on Atlanta Airport

Vienna Dataset
The air traffic scenario dataset for the airport of Vienna consist of 6,768 scenarios. In this case the dataset has to be undersampled to 153 scenarios per class, which adds up to 459 scenarios for training and testing. A summary of the process for random forest classifier training on the Vienna dataset is shown in the table below.

Random Forest - Vienna	
Input Data:	6,768 scenarios from 2017
# Scenarios per Class:	1: 2,336, 2: 4,277, 3: 153
Undersampling:	Delay Class 1 and 2 to the amount of class 3
Oversampling:	None
# Scenarios for Training and Testing	153 scenarios per class
Validation Method	10-fold Cross Validation
Parameter Optimization	
Number of Trees	0 - 100 in 10 steps
Maximal Depth	-1 - 100 in 5 steps
Minimum Leaf Size	1 - 100 in 5 steps

Table 6.2: Process Summary of Random Forest classifier on Vienna Airport

6.2 Multilayer Neural Network Classifier

The second classifier that is trained is the multilayer neural network classifier. It has been described in section 2.6 including its important parameters. The classifier uses the same process in Rapid Miner described in the previous section and shown in 6.1 and 6.2. Two things need to be changed in the process. First the Optimise Parameter operator has to be adapted to variate the parameters of the multilayer neural network classifier.

1. Activation function: Possible functions are Tanh, Rectifier, Maxout and ExpRectifier.
2. Epochs: Possible values are 0 to 1.797e308 divided into 5 steps.
3. Learning rate: Possible ratio is set from 0 to 1 divided into 10 steps.
4. Hidden Layers: Rapid Miner cannot optimize this parameter automatically. It has to be set before the start of the process. It is set to two hidden layers with 50 nodes each.

This parameter optimisation settings lead to 264 possible combinations of parameters and training runs. The running time of all combinations was about 10 hours for the larger dataset on the given hardware. The possibility of more combinations was limited to the size of the random access memory.

The second change is that the sub-process of the Cross Validation has to be changed as the training is now done with the multilayer neural network classifier. This sub-process is shown in Figure 6.4.

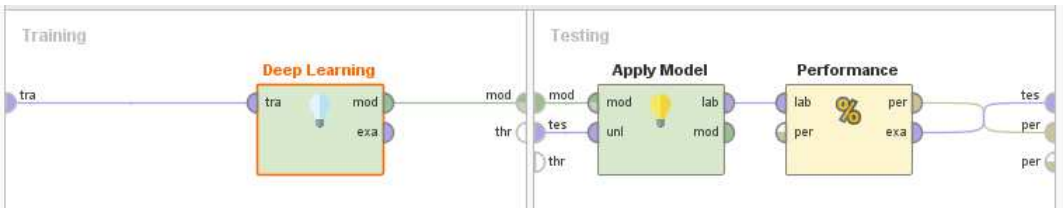


Figure 6.4: Sub-process of the Cross Validation operator of the Multilayer Neural Network process in Rapid Miner

Atlanta Dataset

The same air traffic scenario dataset for the airport of Atlanta was used for the multilayer neural network classifier. A summary of the process for multilayer neural network classifier training on the Atlanta dataset is shown in the table below.

Multilayer Neural Network - Atlanta	
Input Data:	15,064 scenarios from 2017
# Scenarios per Class:	1: 6,344, 2: 8,096, 3: 624
Undersampling:	Delay Class 1 and 2 reduced to the amount of class 3
Oversampling:	None
# Scenarios for Training and Testing	624 scenarios per class
Validation Method	10-fold Cross Validation
Parameter Optimization	
Activation function	Tanh, Rectifier, Maxout and ExpRectifier equals 10 steps
Epochs	0 to 1.797e308 in 5 steps
Learning rate	0 - 1 in 10 steps

Table 6.3: Process Summary of Random Forest classifier on Atlanta Airport

Vienna Dataset

The air traffic scenario dataset for the airport of Vienna consist of 6,768 scenarios. In this case the dataset has to be undersampled to 153 scenarios per class, which adds up to 459 scenarios for training and testing. A summary of the process for random forest classifier training on the Vienna dataset is shown in the table below.

Multilayer Neural Network - Vienna	
Input Data:	6,768 scenarios from 2017
# Scenarios per Class:	1: 2,336, 2: 4,277, 3: 153
Undersampling:	Delay Class 1 and 2 to the amount of class 3
Oversampling:	None
# Scenarios for Training and Testing	153 scenarios per class
Validation Method	10-fold Cross Validation
Parameter Optimization	
Activation function	Tanh, Rectifier, Maxout and ExpRectifier equals 10 steps
Epochs	0 to 1.797e308 in 5 steps
Learning rate	0 - 1 in 10 steps

Table 6.4: Process Summary of Random Forest classifier on Vienna Airport

CHAPTER 7

Evaluation

This chapter evaluates the results of the modelling phase. Results are shown, reviewed, and further steps are determined. The iterative approach of the data mining process (CRISP-DM) led to the fact that the process, including the evaluation phase, was done more than once. Results of earlier runs of the process had to be discarded as precision and recall did not achieve satisfactory results. The reason supposed was that the intial dataset did only include meteorological information as environment condition data and included several airports at once. This lead to the decision to include notice-to-airmen messages and perform classification only for single airports, which substantially improved precision and recall.

For the datasets of Atlanta and Vienna airport the following metrics have been finally achieved:

Atlanta	Class 1		Class 2		Class 3	
	Precision	Recall	Precision	Recall	Precision	Recall
Random Forest	79.56%	86.06%	75.00%	67.79%	81.96%	83.15%
Multilayer Neural Network	74.35%	75.81%	66.01%	64.46%	78.98%	76.61%

Table 7.1: Results of classification on the Atlanta Dataset

Vienna	Class 1		Class 2		Class 3	
	Precision	Recall	Precision	Recall	Precision	Recall
Random Forest	60.69%	69.54%	53.16%	55.63%	64.66%	66.03%
Multilayer Neural Network	54.40%	52.69%	56.27%	56.98%	55.11%	57.12%

Table 7.2: Results of classification on the Vienna Dataset

Atlanta	Scenario Proper Data		Environment Condition Data	
	Attribute Name	Weight	Attribute Name	Weight
1	Medium Aircraft	2.3 %	Open Runways	15.8 %
2	Departure Distance	2.2 %	Sky Level Altitude	7.0 %
3	Large Aircraft	2.0 %	Dew Point	6.3 %
4	Arrival Distance	2.0 %	Snow	5.9 %
5	Small Aircraft	2.0 %	Thunderstorm	5.0 %

Table 7.3: Global weights of the Random Forest classifier for Atlanta

For the airport of Atlanta precision values range from 75.00% to 81.96% and recall values range from 67.79% to 86.06% for the random forest classifier. Risk class 3 achieved the best results. This class includes the air traffic scenarios with the highest delay costs and thus also with the biggest saving potential with over 80% in precision and recall values. Risk class 2 achieved the worst results. The overall best value is recall of class 1 with 86.06%. The same ranking was achieved by the multilayer neural network. However, results metrics of the random forest classifier are better than of the multilayer neural network.

Results show that the classifier is able to predict risk class 3 the best, which means that the chosen environment condition data sources explain the reasons for high delays. The classifier has the most difficulties to predict class 2.

Classification metrics for the dataset of Vienna are worse in comparison to the Atlanta dataset. However, relations are similar. Class 3 achieved the best results, while class 2 achieved the worst. Additionally, the difference between precision and recall of each class is very similar to the differences of the Atlanta dataset. The worse numbers of the precision and recall values are probably due to the reduced size of the data set. For the airport of Vienna only three months of data were available. This leads to the conclusion that a larger dataset for classifier training, of about one year of data, improves the results.

The Explain Predictions operator of Rapid Miner, described and showed in the previous chapter and in Figure 6.1, calculates the weight for classification of each attribute of the air traffic scenario. Table 7.3 and 7.4 shows the five most important attributes of the air traffic scenario proper, as defined in Chapter 1, and the five most important attributes of environment condition data, which were added to the scenario proper.

In Atlanta environment condition data has by far larger impact on the overall classification. The attribute "Open Runways" has the biggest weight with 15.8% and it is also the only attribute of this category that can be influenced by an air traffic controller. Attributes 2 - 5 of environment condition data are meteorological data that cannot be influenced. Attributes from the air traffic scenario proper do not have a high impact. However, all of them can be influenced by the air traffic controller.

In Vienna the distance of weights of attributes between air traffic scenario proper data and environment condition data is less. The attribute "Small Aircraft" has the highest

Vienna	Scenario Proper Data		Environment Condition Data	
	Attribute Name	Weight	Attribute Name	Weight
1	Small Aircraft	25.2 %	Thunderstorm	22.1 %
2	Large Aircraft	7.0 %	Wind Shear	10.6 %
3	Arrival Distance	3.7 %	P.A.P.I	7.0 %
4	Medium Aircraft	3.4 %	Air Temperature	4.9 %
5	Departure Distance	1.9 %	Sky Level Coverage	4.7 %

Table 7.4: Global weights of the Random Forest classifier for Vienna

Random Forest	Parameter 1	Parameter 2	Parameter 3
	Number of Trees	Maximal Depth	Minimal Leaf Size
Atlanta	100	-1	83
Vienna	100	100	67

Table 7.5: Result of the parameter optimisation for the Random Forest classifier

Multilayer Neural Network	Parameter 1	Parameter 2	Parameter 3
	Activation Function	Epochs	Dropout Ratio
Atlanta	tanh	1.797e308	0.5
Vienna	tanh	1.797e308	0.5

Table 7.6: Result of the paramter optimisation for the Multilayer Neural Network

weight with 25.2% and it is also an attribute that can be influenced by the air traffic controller. However, the attribute of environment condition data with the highest weight is "Thunderstorm". The only attribute that can be influenced is the precision approach path indicator (P.A.P.I) with a weight of 7.0%.

Lastly, I want to outline the results of the parameter optimisations for all classifiers. Parameters for the random forest classifier are shown in Table 7.5. A maximal depth of -1 puts no bound on the depth of the trees. Parameters for the multilayer neural network classifier are shown in Table 7.6. The amount of hidden layers for the multilayer neural network classifier could not be optimised automatically with the "Optimise Parameter" operator. It was set to two hidden layers with 50 nodes each as it was the standard setting and paper [WLD18] achieved the best results with two hidden layers for an air traffic dataset.

CHAPTER 8

Deployment

The Deployment phase is the last of the six phases of the data mining process. It describes a deployment strategy for the results evaluated in the previous phase.

The evaluation phase has shown that the classification of air traffic scenarios achieves a precision and recall of up to over 80%. This can help the air traffic controller to be aware of critical situations with high delay earlier and allows to take actions that try to reduce the risk class of a scenario.

While the air traffic controller cannot change the meteorological situation he can take actions in departures and arrivals, as well as in taxiways and runways. One possible action is that he can delay the departure of certain flights on purpose to allow approaching flights a faster landing. This can help to reduce the average delay cost as en-route delay is more expensive than at-gate delay according to the cost reference report used to calculate the risk classes in chapter 4 [oW15]. Similarly the action of slot swapping can reduce average delay cost. If the air traffic controller swaps the landing slots of two flights it can reduce cost if the bigger aircraft can arrive with less delay. A third option is to divert flights to a different airport. While this action would reduce the amount of arriving flights at the classified airport, it is unclear if the total delay cost of this flight really will decrease. The delay cost is just moved from one airport to another and costs are additionally increased as the passengers are arriving at the wrong airport and have to be moved.

The other set of actions involves the taxiways and runways on the classified airport. To reduce average delay cost of a scenario the air traffic controller can decide to open another runway. Of course this action is limited to the total amount of runways and is not possible if all runways are open already. As relevant taxiways on the airport of Atlanta have been identified, which have an effect on classification, the air traffic controller can take action to open another taxiway. A third possible action is to change the runway configuration depending on the meteorological situation. The runway configuration describes from

which compass direction the aircraft approaches the runway. This might reduce the necessary distance between two aircraft which increases possible landings per minute.

Generally the classifier can be used for testing, if a certain action will reduce the delay costs. If the air traffic controller decides to open a runway he can test this scenario with the classifier and decide based on the result.

Lastly, the defined air traffic scenarios from historic data are a good basis to apply associated pattern recognition to draw conclusions from the data. These conclusions can influence flight planning if certain meteorological forecasts are known or a certain runway needs to be closed. However, this is an aspect for future work and is discussed in more detail in chapter 9.2.

CHAPTER 9

Critical Reflection and Contribution

9.1 Critical Reflection

In the scope of this thesis several aspects influenced the work which could be improved if a similar data mining problem is tackled again or the work of this thesis is continued. The first aspect is to invest more time in the first two phases of the data mining process. To define what data is needed to reach the data mining goals and to check if these data is available is an important task. At the start of this thesis EUROCONTROL was about to stop its current data portal. In addition the old portal only allowed to download data from five days per week. This meant that it would have been necessary to collect data for 146 weeks until a two-year period of data was downloaded. The second data source used in this thesis from the US Bureau of Transportation did provide enough data, but not in the same detail as EUROCONTROL did. Information about the type of aircraft or more detailed data about the en-route flight to calculate exactly when delays happened was not included in the US data. It is important to know if all data and attributes are available and how to collect them. However, too much data also has negative impacts as it makes the preparation phase unnecessary complex. This is my second aspect. The amount of work for the data preparation phase can be underestimated, as data cleaning and data construction need a lot of time and thought. So a solid consideration what data and attributes are needed and which can be omitted saves a lot of time in this phase. This consideration also includes a description of how attributes have to be transformed to reach the final dataset for the modelling phase. Data transformation are needed in most cases and are often quite complex as additional data might be necessary. An example in this thesis is the transformation of the tail number of a flight to the aircraft model.

Furthermore data preparation is also computationally very expensive if data records are in the hundred thousands or millions. To execute them on appropriate hardware is of

great advantage due to the chance that multiple repetitions will be necessary as learnings from the modelling phase change the requirements of the dataset. The computation for the scenarios on the given hardware took about 6 to 10 hours depending on the classifier and the size of the dataset. This reduced the possibility to try more combinations of parameters and increased the chance that the best parameters have not been found.

The tool Rapid Miner used in this thesis can be highly recommended as the interface is very intuitive and the way data mining workflows can be developed is very clear and easy. Nevertheless the Rapid Miner Server, where the processes are executed, has high hardware requirements. The recommended specification is quadcore processor with 3Ghz or more, 32GB to 1TB of RAM and enough disk space to store data needed for the processes. The optimal hardware has several advantages. Not only does the process finish faster, the higher speed allows to try more combinations of parameters of each classifier. This increases the possible dataset sizes and parameter variations.

9.2 Contribution

This thesis describes a method to create air traffic scenarios for an airport and to train a classifier that can predict the risk class on each scenario. The objective was to combine different data sources in a new way in comparison to the current knowledge. Instead of predicting delay of single flights, flights were summed to air traffic scenarios containing additional environment condition data like notice-to-airmen messages. Delays were derived into costs to give the air traffic controller a better situational awareness based on delay cost of the upcoming air traffic scenario. The random forest classifier for the airport of Atlanta achieved accuracy values of 86.1% for risk class 1, 67.8% for risk class 2 and 83.0% for risk class three.

As flight delay prediction has mostly been done for single flights in the state-of-the-art, results of this thesis are compared with delay classification of single flights. The paper called "Data Mining For Robust Flight Scheduling" tried to classify flights into the three categories "Ahead of time", "on time" and "delayed" just based on historic flight data as data source [AKW+09]. The results of classification are stated with accuracy values, this is the reason why the accuracy of the risk classes for the Atlanta dataset are stated in the beginning of this section. The highest accuracy achieved in the paper is 45.4%. The authors of that paper adapted their classifier to take attributes of locally varying relevance into account. No environment information is included in that classification and is only proposed as future work idea.

Another paper combined flight and meteorological data to predict the delay of single flights into certain thresholds. By adding weather information from the origin airport and the destination airport the results are up to 79.8% accuracy for the threshold of 90 minutes. Smaller thresholds have a decreased accuracy of 69.1% for the below 15 minute threshold. The dominating classification method in that paper was Random Forest. Different classifiers were applied, but did not achieve better results. [BMTT16].

The work of [RB14b] modelled the US airport network to predict aggregated air traffic delays. That prediction needs information about the entire airport network. The inputs are the delay states of the most influential airports and the global delay state of the National Airspace System. Meteorological information or any other environment condition information is not used. For a delay threshold of 60 minutes that work reaches an accuracy of 81%.

A different research team tried to model flight delays and cancellations based on the expected weather. In addition linear regression and neural network models are compared. They came to the conclusion that the neural network delay model outperforms the linear regression method and that the delay estimation accuracy varies depending on the season. The accuracy is 10% higher in the convective weather season from April to September versus the non-convective season from October to March [SWJK09]. The paper used the same data source for US flight data as was used in this thesis.

Another classification was done by a group of researchers that tried to classify delays of individual origin-destination pairs by including meteorological data at the origin and destination. The goal was to predict if a flight is on-time or delayed. A flight was on time if it had less than 15 minutes of delay. The team compared the following classifiers: random forest, AdaBoost, KNN and decision trees. The result was that the random forest classifier outperformed all others with an accuracy of 80.36% [CKBM16].

One researcher from China tried to predict flight delay by using multiple linear regression. He used historic flight data that also included some meteorological information like wind direction and speed. The goal was to predict if a flight has more than 30 minutes of delay or less. The linear regression model outperformed naive-bayes and C4.5 by achieving 79.1% in accuracy [Din17].

CHAPTER 10

Summary and Future Work

10.1 Summary

The aim of this thesis was to predict one of three risk classes for air traffic scenarios at a certain airport. Those scenarios consist of flight data and environment condition information. The risk classes are calculated based on delay costs of a scenario. Costs are determined by using the linear regression in the cost reference report from EUROCONTROL. The linear regression is based on the maximum take off weight of an aircraft. The data sources are the On-Time performance statistics of the US Bureau of Transportation, Flight Data from EUROCONTROL, METAR data, airport and runway data, aircraft data and notice-to-airmen messages. The analysis of these datasets gives insights into delay frequency and amount in combination with weather information. All data sources are combined to create scenarios of a specific time span, which results in a tuple that includes up to 37 attributes about the airport, air traffic and basic weather and special weather phenomena like thunderstorms or snow. Based on existing literature the random forest classifier and multilayer neural networks are chosen as applicable methods for these datasets. The allocation of the scenario into the specific risk classes for training showed high imbalance in the data as flights with short delays are more common than with long delays. To solve this problem undersampling has been used to achieve balanced classes for the classification training and testing. For modelling the tool Rapid Miner has been used. Results showed that precision and recall values reach more than 80% for classifying risk class 3 which can help the air traffic controller to set actions earlier and try to minimize delay. A smaller training dataset reduces precision and recall, but the relations between the two metrics stay similar. Metrics for risk class 2 are the worst, as the environment conditions of scenarios in risk class 1 and 2 often do not differ. For both datasets the random forest classifier outperformed the multilayer neural network. Based on this classification the air traffic controller has different possibilities to set actions like slot swapping, delay the departure of the flight, open a runway or change the runway

configuration.

10.2 Future work

Research described in this thesis provides a solid basis for an intelligent combination of different air traffic and environment data sources to achieve a support for air traffic management in form of scenario classification. The evaluation phase showed that the result metrics outperformed results of related work. Nevertheless, several ideas for improvement exist. The first one is to provide more data for classifier training. The data that could be collected from EUROCONTROL and the available computational power limited the size of the training data set. Because results show that an increased size of dataset improves results, by providing two or more years of data to the classifier, chances are that the results improve even further.

A second aspect is the fact that time series were not included in this thesis. If an air traffic scenario would have information of past scenarios, this information could increase classification accuracy. This could be information of past air traffic scenarios of the risk class or about the meteorological situation or air traffic. It could be combined by choosing different time spans. In this thesis the time span of a scenario is set to the time when the weather data is updated. However, peaks in traffic do not always happen in between the publication of weather data. By defining new strategies for setting the time span of a scenario, improvements of the result could be possible.

Another aspect is to use the method of scenario creation and risk class calculation as a basis for associated pattern recognition. This can show correlating attributes which might influences flight planning. By knowing the correlations of two attributes flight plans can be adapted in the first place and decrease delay when the aircraft performs the flight.

As evaluation has shown, attributes that have an influence on the risk class classification differ between the two classified airports. Increasing knowledge about the specific airport can help to improve classification. People with domain knowledge of the specific airport can help to find further attributes that have an effect on classification.

List of Figures

List of Tables

Bibliography

[AALB+07] Mohamed Abdel-Aty, Chris Lee, Yuqiong Bai, Xin Li, and Martin Michalak. Detecting periodic patterns of arrival delay. *Journal of Air Transport Management - J AIR TRANSP MANAG*, 13:355–361, Nov 2007.

[ABE01] Shawn Allan, John A. Beesley, and James E. Evans. Analysis of delay causality at newark international airport. *4th USA/Europe Air Traffic Management Research and Development Seminar*, Dec 2001.

[Adm19a] Federal Aviation Administration. Aircraft Characteristics Database. https://www.faa.gov/airports/engineering/aircraft_char_database/, 2019.

[Adm19b] Federal Aviation Administration. NOTAM Search. https://notams.aim.faa.gov/notamSearch/nsapp.html#/, 2019.

[Agg15] Charu C. Aggarwal. *Data Mining - The Textbook*. Springer, 2015.

[AKW+09] Ira Assent, Ralph Krieger, Petra Welter, Jörg Herbers, and Thomas Seidl. *Data Mining For Robust Flight Scheduling*, pages 267–282. Data Mining for Business Applications. Springer, Jan 2009.

[ASSRA04] Khaled F. Abdelghany, Sharmila S. Shah, Sidhartha Raina, and Ahmed F. Abdelghany. A model for projecting flight delays during irregular operation conditions. *Journal of Air Transport Management*, 10(6):385–394, 2004.

[BMTT16] Loris Belcastro, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Using scalable data mining for predicting flight delays. *ACM Trans. Intell. Syst. Technol.*, 8(1), July 2016.

[BP18] Joe Blitzstein and Hanspeter Pfister. Harvard data science course. http://cs109.org/, 2018.

[CCK+96a] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. Crisp-dm 1.0. step-by-step data mining guide. 1996.

[CCK+96b] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rüdiger Wirth. Crisp-dm 1.0: Step-by-step data mining guide, 1996.

[CKBM16] S. Choi, Y. J. Kim, S. Briceno, and D. Mavris. Prediction of weather-induced airline delays based on machine learning algorithms. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pages 1–6, Sep. 2016.

[CLS06] Sven F Crone, Stefan Lessmann, and Robert Stahlbock. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, 173(3):781–800, 2006.

[CWLC14] Leonardo Cruciol, Li Weigang, Leihong Li, and John-Paul Clarke. In-flight cost optimization for air traffic flow management using data mining method on big data. *OPT-i 2014 - 1st International Conference on Engineering and Applied Sciences Optimization, Proceedings*, pages 1491–1498, Jan 2014.

[Din17] Yi Ding. Predicting flight delay based on multiple linear regression. *IOP Conference Series: Earth and Environmental Science*, 81:012198, Aug 2017.

[EUR17] EUROCONTROL. U.s./europe comparison of air traffic management-related operational performance for 2017. https://www.eurocontrol.int/publication/useurope-comparison-air-traffic-management-related-operational-performance-2017, 2017.

[EUR18] EUROCONTROl. Latest on delays. https://www.eurocontrol.int/news/latest-delays, 2018.

[EUR19] EUROCONTROL. Demand data repository web portal. https://www.eurocontrol.int/ddr, 2019.

[FH01] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *European Conference on Machine Learning*, pages 145–156. Springer, 2001.

[HKP11] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann, 2011.

[IAT18] IATA. Traveler numbers reach new heights. https://www.iata.org/pressroom/pr/Pages/2018-09-06-01.aspx, 2018.

[Inc06] ESRI Inc. Standard classification schemes. http://webhelp.esri.com/arcgisdesktop/9.2/index.cfm?TopicName=Standard-classification-schemes, 2006.

[KGH15] Thomas Kopinski, Alexander Gepperth, and Uwe Handmann. A simple technique for improving multi-class classification with neural networks. *ESANN 2015*, Apr 2015.

[KzI12] Esra Karabulut, Selma Özel, and Turgay Ibrikci. Comparative study on the effect of feature selection on classification accuracy. *Procedia Technology*, 1:323 –327, Dec 2012.

[oTS19a] US Bureau of Transportation Statistics. Carrier On-Time Performance. https://www.transtats.bts.gov/DL_SelectFields.asp, 2019.

[oTS19b] US Bureau of Transportation Statistics. Carrier On-Time Performance Statistics. https://www.bts.dot.gov/newsroom/2018-traffic-data-us-airlines-and-foreign-airlines-us-flights, 2019.

[our19] ourairports. Open Data downloads. https://ourairports.com/data/, 2019.

[oW15] University of Westminster. European airline delay cost reference values. https://www.eurocontrol.int/sites/default/files/publication/files/european-airline-delay-cost-reference-values-final-report-4-1.pdf, 2015.

[PAYLD18] Duc-Thinh Pham, Sameer Alam, Su Yi-Lin, and Vu N Duong. A machine learning apporach on past ads-b data to predict planning controller's actions. Technical report, Air Traffic Management Research Institute at Nanyang Technology University, 2018.

[RB14a] Kalpana Rangra and K. L. Bansal. Comparative Study of Data Mining Tools. *Journal of Advanced Research in Computer Science and Software Engineering*, 4(6), Jun 2014.

[RB14b] Juan Rebollo and Hamsa Balakrishnan. Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, 44:231–241, Jul 2014.

[SWJK09] Banavar Sridhar, Yao Wang, Richard Jehlen, and Alexander Klein. Modeling flight delays and cancellartions at the national, regrional and airport levels in the united states. *USA/Europe Air traffic Management Research and Development Seminar*, (8), 2009.

[TAL14] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. In *Data Classification: Algorithms and Applications*, pages 37–64. CRC Press, 2014.

[Uni19] Iowa State University. ASOS-AWOS METAR Data Download. https://mesonet.agron.iastate.edu/request/download.phtml, 2019.

[WLD18] Zhengyi Wang, Man LIANG, and Daniel Delahaye. Automated data-driven prediction on aircraft estimated time of arrival. *Sesar Innovation Days 2018*, pages 1–9, Dec 2018.

[WP16] Hayden Wimmer and Loreen M. Powell. A Comparison of Open Source Tools for Data Science. *Journal of Information Systems Applied Research*, 9(2), Oct 2016.