Practical assignment: flight delay prediction

Data Mining

1 Introduction

The goal of this practical assignment is to get hands-on experience with the core data mining concepts that have been covered in the lectures of this course. In this second assignment we focus on practical experience with a second data mining tool, Rapidminer.

The assignment consists of three parts. First, in Section 3, we will take a look at the data. In Sections 4 and 5, we ask you to create a classification and regression model for flight delay prediction in Rapidminer.

1.1 Practical matters

The assignment needs to be handed in individually. Your final submission should consist of two parts:

1. The files corresponding to the Rapidminer process that you created for classification.
2. The files corresponding to the Rapidminer process that you created for regression.

The Rapidminer process file (.rmp) should have the following naming scheme: `sXXXXXXX-classification.rmp` and `sXXXXXXX-regression.rmp`, where the X's mark your student number. Since the files will be graded automatically you should **not** deviate from this format. Collect the files of your submission in a single ZIP file and submit this file by uploading it to Brightspace before the deadline which is indicated on the Assignment page.

2 The Rapidminer toolkit

This assignment assumes that you have access to the Rapidminer toolkit (`www.rapidminer.com`) with an Educational License. If you do not have such a license yet, request one[1]. Then, download the tool.

[1] `https://rapidminer.com/educational-program/`

2.1 Rapidminer tutorial

When starting Rapidminer, you are prompted with a series of tutorials. If this is not the case, you can access them by clicking `Help` → `Tutorials`. In case you haven't done this yet, make sure that you follow and understand the eight tutorials in the `Get started` chapter to get familiar with the most common Rapidminer operators. We will use these operators while working with the flight delay data.

3 The flights dataset

The problem that we will consider in this assignment is that of *using data to learn more about flight delays*, which might help airlines to improve their schedules in the future, which in turn might result in reduced flight delays, reduced fuel usage and costs, and reduced CO^2 emissions.

To this end, we will use real-world flight data collected from the Federal Aviation Administration (FAA). This data concerns historical aircraft movements in the United States, including information about departures (date, time, and place) and arrivals. Also included are the registration number of the aircraft, the airline, and the departure and arrival delays. In addition, we have access to a (separate) dataset with information on the weather at the relevant airports, which was measured hourly.

For this assignment you will use the following two datasets:

- `flights_sample.csv`: A set of flights that have been executed in the United States in 2016 and 2017. These flights have been randomly sampled from a larger database of flights.
- `airport_weather.csv`: The weather for each airport in the filtered flights dataset. Once every hour, the temperature, rain, wind, visibility, etc., were measured.

You can download the datasets from from the Brightspace page.

3.1 Knowledge discovery

We will first take a look at the `flights_sample.csv` dataset. In Rapidminer, use the `Import Data` button and select the `csv` file. In the next screen, `Specify your data format`, you should not have to change any settings, since Rapidminer is smart enough to detect the examples. A preview of the data is shown. Each row represents a flight, or *example*. The columns are the characteristics, or *attributes* of the flights. Take a look at the attributes and try to understand them. You might need to look some of them up. Understanding the data helps to determine later which attributes are important and which ones are not.

On the next page, `Format your columns`, we can specify the data types of the attributes by clicking the ▼ and selecting `Change type`. Again, Rapidminer is pretty smart, but made

some mistakes. Find out for each of the seven data types that are used what they are used for and change the attributes that are wrong. It is important to note that while some attributes might look numerical, they are actually binominal or polynominal! Also, Rapidminer does not seem to understand our notation of time and selecting `time` as data type will show errors. Leave those data types as `polynominal`.

Go to the next screen and select a location to save the dataset. The `Local Repository` is fine. Click `Finish` to import the dataset. After the import, Rapidminer shows the dataset again. In the `Statistics` tab on the left, you can see some properties of the data, grouped by attribute.

3.2 Target values

Now that you have a feel for what type of data we are dealing with, we can talk about the goal of this assignment. In this assignment we will create models to predict departure delay of the flights in the dataset. More specifically, the two *supervised learning* paradigms that we will involve are *classification* and *regression*. Our goal is to create two types of models that can answer the following questions:

1. *Classification*: Can we predict whether a flight will depart with a delay of 15 minutes or more (yes or no)?
2. *Regression*: Can we predict the departure delay, in minutes?

For the classification problem, we use the attribute `dep_delay_15`. It has a value of 1 when a flight has a departure delay of 15 minutes or more, and 0 when the departure was reasonably on time (< 15 minutes). The attribute used for regression is the exact departure delay, as given by the `dep_delay` column.

3.3 Weather data

An important factor that strongly influences air traffic is the weather. We therefore introduce another dataset, `airport_weather.csv`, which contains data on the weather at each airport in our flight datasets. Import the dataset in the same way that you imported the flight data. Try to find out what each attribute describes and think about the data types you use for the attributes. Also, add a `Select Attributes` node so we can filter the weather attributes.

To obtain the weather at the departure airport for each flight, we need to merge the datasets. Combining datasets in such a way is called a *join*. The Rapidminer `join` operator can do this for us. Add this operator to the process and connect the `Select Attributes` of the two datasets. Take a look at `join`'s parameters. There are four `join types`. Read the documentation or look up online what each join type does and select the one we need. In the `key attributes` properties, we define the attributes that we want to match on. This is important, because we do not want to add the weather of each hour and each airport to each

flight. Think about which attributes you want to match. Hint: there are five attributes to match on. You may need to change the `Select Attributes` configurations.

Spoiler: one of the attributes that we want to match on is the departure hour of the flight. This is easy to do in the weather dataset, since there is an attribute `hour`. However, we do not have such an attribute in the flights datasets. There is an attribute `dep_time_planned` though, which contains the hour and minute that the plane was supposed to take off. This is currently stored as text and the type is `polynomial`. This is a complex attribute that is hard to use for modelling and we therefore need to perform *attribute extraction* to get its desired components. To split the hour from these timestamps and convert them to numbers, we can use the `Split` and `Parse Numbers` operators. Find out how they work and add them to your process.

3.4 Attributes & Train/Test set

You can refine the attributes you want to use by adding another `Select Attributes` operator before splitting the data. Introducing new attributes by combining existing ones can be a good way of increasing a models performance, e.g. constructing a feature which can detect certain weather types which are extremely disruptive for flights. Find out if you can introduce and refine some attributes in such a way that the performance of the model increases, compared to the base attributes.

We are trying to create a model which can predict **future** departure delays, this needs to be taking into account when creating a train and test set. Therefore, it is important to only use data points from the year `2017` as your test set. Connect the output of the `join` operator to the `Filter Examples` operator to generate a training and test set which meet this requirement.

Please note that you are **not** allowed to use any features in your model which are not available before departure time, e.g. you cannot use the feature 'actual arrival time' to predict the departure delay. The model is supposed to be able to predict the delay of a flight, **before** the flight has actually taken place.

4 Classification

It is now time to create a classification model. Use the knowledge of the Rapidminer tutorials and the problems described above (training/test data, joining the weather dataset and attribute selection) to create your classification model. When creating the model, and trying out different types of classification techniques, try out at least some sort of `Decision Tree` and one other classifier type. Although you should try several models, your submission should include only the finally selected model. Also include a relevant evaluation node at the end of the process. Make sure this evaluation node is informative for the type of model you are

creating.

5 Regression

In the previous section, we created a Rapidminer process to create a classification model. Now you are asked to create a regression model that predicts the `dep_delay` attribute. When creating this model, try three types of models, and experiment with other sets of attributes. Although you should try several models, your submission should include only the finally selected model. Also include a relevant evaluation node at the end of the process, make sure this evaluation node is informative for the type of model you are creating.