Practical Work: Data Exploration and Preparation – Fall 2023

This exam will evaluate your skills in preparing datasets to train your models with appropriate data.

The skills checked are the following:

* Your ability to load data from different sources.
* Your ability to analyze raw data:
  + Basic analysis, distribution of features, labels
  + Deal with temporal data
  + Geographical data
* To choose correct transformations and extract values from existing features in the dataset.
* To add external values to enrich the dataset.

# Practical work: 4 hours

## Main task

The zip file “data.zip” contains your primary data for this practical work. It is composed of 3 files:

* train.csv
* test.csv
* external\_data.csv

train and test files are containing the same kind of data, they have the same structure. The data is about number of air passengers, recorded daily. The structure is as follows:

* The date of departure (**DateOfDeparture**)
* The departure airport (**Departure**)
* The arrival airport (**Arrival**)
* the mean and standard deviation of the number of weeks of the reservations made before the departure date, respectively: **WeeksToDeparture** and **std\_wtd**
* **log\_PAX** which is related to the number of passengers (the actual number were changed for privacy reasons)

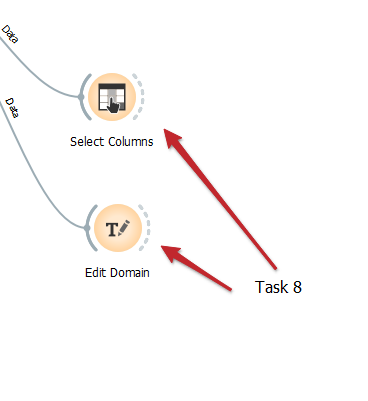
In this exercise we want to predict the variable **log\_PAX**

## Deliverables

For the practical exam, **you should submit an ows file** (Orange Worfklow file) containing the “mandatory tasks” (see next section). Once the mandatory tasks are completed, you can proceed with the other ones.

**Very important remark**

For each task, annotate your workflow with the name of the task, as in the following screenshot:

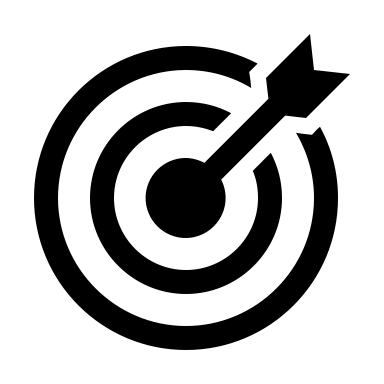


If it is not possible to easily understand what exactly you have tried, then you will not be graded.

# Mandatory steps (10 pts)

Make sure you have read the previous **important remark**

In the following tasks those marked as (Mandatory) are mandatory for the exam

****Task 1 (Mandatory)**: explore the dataset – *3 points*

*Load data in a data frame, explore data*

Load the train.csv file and:

1. Perform the following analysis: features type, quality, distribution the features vs the target feature. Pay attention to the encoding of the variables. (mention as Task1.A)
2. Analyze availability and consistency of data (mention as Task1.B)
3. Put minimum 3 interesting visualizations. In your widget, put a comment for each visualization. You can use distribution, correlation, scatter plots, etc (Task1.C)

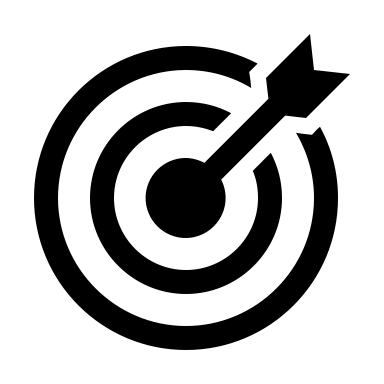
Hint: seasonality can be interesting at 2 levels: year based and month based, you can even refine to daily base to check the underlying model driving the passengers behaviour.

**Task 2 (Mandatory)**: Create a baseline model that trains on all the variables (use a random forest, configured like the following screenshot) – *1 point*

Une image contenant texte

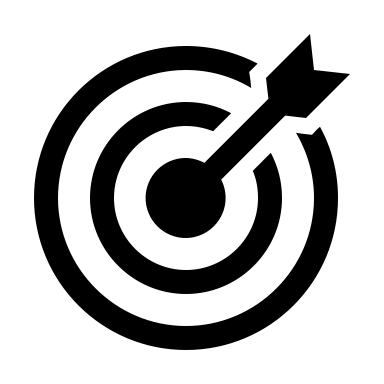
Description générée automatiquement

Make sure you correctly configure the split between test and train data in the “test and score” widget.

****Task 3 (Mandatory):** Feature selection – *2 points*

*Business understanding and feature selection*

Refine the variables selected to train the model. You can select them after a correlation or feature contribution analysis. Also think well about what the available data at the time are when the system makes the prediction

****Task 4 (Mandatory)**: Feature construction – *4 points*

*Feature construction*

Transform the features to add information to your dataset, taken in account for the training phase. For instance, transform the date so you can include the date and time aspects in your features. Other attributes of your choice can be processed as well, to deal, for instance, with missing values or date handling.

# Other ideas to try… (15pts)

You can try the following tasks to improve your model score (and your grade for this exam!)

**Task 5**: Enrich the dataset with external xml data – *3 points*

|  |  |
| --- | --- |
| *Mille avec un remplissage uni* | *XML data to dataframe, dataframe join* |

Enrich this dataset with some external information (provided in holidays.xml), encode the fact it is a regular day or a holiday day

*Hint* : create an other data table that receives those information, and merge “append-style” (left join from the main dataset) them on the date and time aspects, then use a feature constructor that transforms the right side into :

* If value is missing then : set “**regular**”
* If value is present : replace the value by “**holiday**”

Check if it has a positive impact on the model performance.

**Task 6**: enrich the dataset with data already present but not used yet *– 2 points*

|  |  |
| --- | --- |
| *Mille avec un remplissage uni* | *Date feature processing* |

The goal is to add a Weekday column to check if it has an influence.

|  |
| --- |
|  |

Observe the distribution of departure across weekdays, then re-encode the newly created variable “Weekday” with the appropriate kind of variable and check if it has a positive impact on the model performance,

**Task 7**: Adding more data – *1 point*

|  |  |
| --- | --- |
| *Mille avec un remplissage uni* | *Dataframe concatenation* |

Load now the test.csv, append it to the primary dataframe (got from train.csv), to do that, you can use the “concatenate” widget. Then observe the test and score performance.

Note: appending the test data can seem weird, but actually orange will split it again into test and train at model building phase. It allows us to benefit from the transformation.

**Task 8**: Add external data source in csv format – *2 points*

|  |  |
| --- | --- |
| *Mille avec un remplissage uni* | *Dataframe join, duplicates handling* |

Add weather based external information to the dataset (provided file: **external\_data.csv**). Observe again what feature could be interesting. Keep only the ones which are interesting.

**Task 9**: merging the IATA information - *2 points*

|  |  |
| --- | --- |
| *Mille avec un remplissage uni* | *Call external library to process data* |

Merge with the data at this link:

[airport-codes](https://data.opendatasoft.com/explore/dataset/airports-code%40public/table/?flg=fr-fr&refine.country_name=United+States)

Do again a feature observation and selection to retain only the valuable ones.

**Task 10**: Bonus exercise – *6 points*

Propose **two** data enrichments / transformation from this dataset that could help in improving the model performance. If you don’t have time to implement them, put a text note in your workflow to explain what you could try. Name them “Task10.A” and “Task10.B)

*Hints:* think about geographical facts, other events like federal or national vacations, local cultural events… be innovative!