Practical Work: Data Exploration and Preparation – Spring 2022

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

Load the following raw data (car-crashes.csv) in an orange workflow.

The file contains 16 features and around 8k rows, it represents the car crashes in the city of San Francisco between 2016 and 2020.

Here are the columns descriptions :

- **Lat** : Latitude of the incident
- **Lng** : Longitude of the incident
- **Bump**, **Crossing**, **Give_Way,Junction**, **NoExit**, **RailWay**, **Roundabout**, **Stop, Amenity, Side** : The characteristics of the location where the incident has taken place, several can be true at the same time. Side is the side of the street.
- **State:** the state from which this dataset is coming from
- **Distance:** the distance of the traffic jam provoked by an accident
- **Severity**: An indicator representing the impact on the traffic. Values can range from 1 to 4, the highest value translates the highest impact on traffic.
- **Timestamp**: the moment when the incident has occurred.

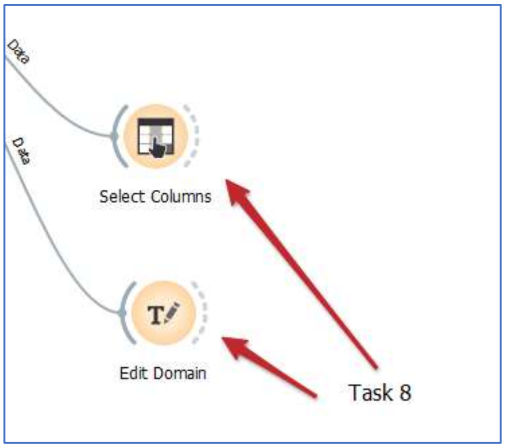
Severity is your target variable for that exercise. The goal **is not to adjust** the algorithm parameters to train your model, but to improve your performance **only** with the data preparation.

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**

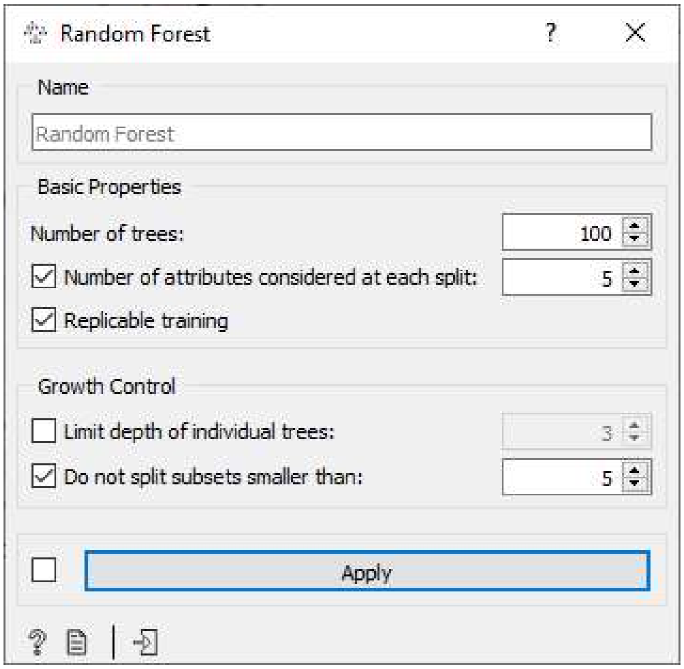
In the following tasks those marked as (Mandatory) are mandatory for the exam

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

Load data in a dataframe, explore data

A. Put some widgets to analyze features type, quality, distribution the features vs the target feature.
B. Analyze consistency (geographical consistency for instance) of data.
C. Put 5 interesting visualizations in your widget, put a comment for each visualization.

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



Random Forest

Name

Random Forest

Basic Properties

Number of trees: 100

☑ Number of attributes considered at each split: 5

☑ Replicable training

Growth Control

□ Limit depth of individual trees: 3

☑ Do not split subsets smaller than: 5

□ Apply

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.

Other ideas to try… (16pts)

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*

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*

Date feature processing

Add a column named “weekday” (using a feature constructor, add a categorical variable and set the possible values to 0,1,2,3,4,5,6) and initialize this column with the week day number, available thanks to this code:

```
import datetime
import pytz
from pytz import timezone

timezone = timezone('US/Pacific')
for data in in_data:
    timestamp: str = str(data['timestamp'])
    date = timezone.localize(datetime.datetime.strptime(timestamp, '%Y-%m-%d %H:%M:%S'))
    weekday: int = date.weekday()
    data['weekday'] = weekday

out_data = in_data
```

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

Task 7: Dealing with balance between classes – *2 points*

🎯 *Rebalancing techniques*

Compare the precision of the model and the distribution of the target variable. What can you propose to improve the precision for minor classes?

Check if your solution has a positive impact on the model performance.

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

🎯 *Dataframe join, duplicates handling*

Add weather based external information to the dataset (provided file: weather-sfcsv.csv)

Task 9: Use a python library to build new features - *3 points*

🎯 *Call external library to process data*

Encode the daylight information thanks to the astral library, compare the incident date with the sunset and sunrise information given by the library. You must add a new column beforehand (like isDay or isNight), using a feature constructor, save the information under whatever format you like.

Hints:

- You can reuse the code provided at task 6 by making required modifications, beware to localize the time zone for effective transformation.
- To install astral, just go in orange Options window > add-ons > Add more… > type "astral" and "OK" (resolved version should be 2.2)
- More information on astral : https://astral.readthedocs.io/en/latest/index.html

Task 10: Bonus exercise – *4 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.

Hints: think about geographical segmentation, other events like federal or national vacations, local cultural events, re-encoding time ranges to identify rush hours… be innovative!