CSP7001 – Introduction to Machine Learning
Lecturer: Giovanni Di Liberto
Academic Year 2023-2024
Trinity College Dublin

Group project

General instructions: Download the *Dublinbikes* dataset at https://data.gov.ie/dataset/dublinbikes-api. This is the same dataset that we used in the tutorials. In this project, you will have the chance to work on a larger portion of the dataset, starting from "2018 Q3", at your discretion. It is important that all members of a group get familiar with the core skills required to complete this project (e.g., basic coding, visualisation, supervised learning). Nevertheless, you are free to organise your work as you find most productive. For example, one member of the group could focus on data visualisation in this project, while another member could focus on supervised learning. Also, feel free to ask questions about the project in class/lab. Please avoid talking with other groups about the project though. There are many possible answers to each question, so interaction between groups would bias your solution and make this much less interesting for you.

Scenario: You are working for FUTURE-DATA, a local company specialised in data science. Dublin City Council hired your company to study the optimisation of the Dublin bike grid. Dublin City Council plans to modify the stations to improve the user experience. For example, adding new stations or even reducing the bike stands for inactive stations. FUTURE-DATA decided to investigate this by formulating multiple scenarios that should be considered by the City Council. To do so, they assigned the task to *k* small teams of machine learning and smart and sustainable cities experts. You are part of one of these *k* teams. Please register your team through Blackboard i.e., agree on a team number with your team members and register on the corresponding team on the online portal.

Task: The company proposed 2 goals. The two tasks correctly carried out will give you full points. The two tasks have the same value.

1. To identify at least 5 bike stations that could be removed or substantially reduced in size, and to identify areas where additional bike stations could be useful for increasing the user experience;
2. To predict the time-course for the new bike stations and for the modified stations; to identify a metric for comparing the new city bike infrastructure with the old one, making a case that your solution will actually improve the user experience.

A few tips: The manager suggested focussing on a subset of the city for simplicity (of course, you are free to do more than that, if you like). Make sure that the data for that bike station is available if you plan on combining datasets from different quartiles. Missing or bad data-points can be a problem. So, identifying stations with good data will make your life easier (but feel free to make your life more complicated if you like the challenge).

See next page:

Submission:

- A brief report (maximum 4 pages including figures – additional pages will not be considered), minimum font size 11pt. The report should answer the two questions above, and figures (e.g., bar plots) or tables showing the results and data supporting your considerations. Remember: You are planning to present this report to your manages AND to Dublin City Council. As such, figures must be easy to understand (e.g., large font size, brief but meaningful axis labels, include a short caption describing each figure, lines and datapoints must be clearly visible).
- Your Python scripts. Please write them well, with comments so that I can understand what you did. I will use the scripts to double-check your results where necessary.
- Submission deadline: 19th April 2024.
- Late submission penalty: 10% * number of days