CSP7001 – Introduction to Machine Learning
Lecturer: Giovanni Di Liberto
Academic Year 2022-2023
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 be working on a larger portion of the dataset, starting from "2018 Q3". It is important that all members of the groups 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 (e.g., Friday afternoon). 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 impact of COVID-19 on the city-bikes usage as they are planning to optimise the city-bike system. Dublin City Council had originally structured the city-bike network based on the forecasts of bike usage up to 2030. However, they think that the usage may not match the initial prediction because of the impact of the pandemic on our mobility. FUTURE-DATA decided to investigate this rapidly and by formulating multiple scenarios that should be considered by the City Council. To do so, they assigned the task to *k* small teams or machine learning and smart and sustainable cities experts. You are part of one of these *k* teams.

Task: The company proposed 3 goals. Pick 2 out of the following 3 tasks (two tasks correctly carried out will give you full points; an extra task will not give any extra points).

1. To assess the impact of the pandemic on the city-bike usage;
2. To estimate how the city-bike usage would have been without the pandemic (e.g., 2020);
3. To predict the city-bike usage for 2022 in both the pandemic and no-pandemic scenarios. Use both qualitative and quantitative comparisons.

The manager suggested focussing on two or three strategically placed bike stations (of course, you are free to do more than that, if you like). Make sure that the data for that bike station is available on all the datasets. 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).

Suggestion: The original features tell us about bike and bike stand availability. However, that is a different concept from "bike usage". The optimal approach involves deriving a new feature quantifying the "bike usage" in a given station. Other ways of tackling the tasks are also accepted, but remember to justify your choices and to plot clear, compact figures.

See next page:

Submission:

- A brief report (~2 pages including figures, max 3 pages) with text answering the three 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).
- 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: 21 April 2023.
- Late submission penalty: 5% * number of days (e.g., 2 days late -> 10% penalty)