2023-2024 CST2130 Machine Learning Coursework

Data Management and Business Intelligence

General Information

Important Dates:-

● Deadline for Submission:- **22 February 2024(Thursday) 23:59pm.**

● PowerPoint presentation:- **23 February 2024(Friday), 9:00am -12:00 pm.**

You're required to submit your work via the dedicated Unihub assignment link by the specified deadline.

Note that this link will 'timeout' at the submission deadline. If you miss this deadline, your work will not be accepted as an email attachment. Therefore, you are strongly advised to allow plenty of time to upload your work prior to the deadline.

Your submission should comprise a single PDF file. Your PowerPoint presentations should comprise a PowerPoint file.

Problem Description

Santander Cycles (formerly Barclays Cycle Hire) is a public bicycle hire scheme in London. The scheme's bicycles are popularly known as Boris Bikes, after then-Mayor of London, Boris Johnson. The operation of the scheme has been contracted by Transport for London (TfL). The recent success of the scheme has led to its expansion into many areas of London and its rapid growth has led to real challenges in balancing bike-sharing supply with bike-sharing demand.

To provide a possible solution to this problem, bike-sharing usage prediction is critical. To this purpose, the Transport for London (TfL), together with the free meteo provider (https://freemeteo.co.uk/), has released a dataset − named london_bike_data.csv − having the following structure:

id. Identifier of the record.

date. Date in the format 'YYYY-MM-DD'.

hour. Hour of the day, from 00 to 23.

season. A code identifying a season; e.g., spring = 0, summer = 1, etc.

is_weekend. Boolean value, 1 for weekends, 0 otherwise.

is_holiday. Boolean value, 1 for bank holidays, 0 otherwise.

temperature. Weather temperature in celsius degrees.

temperature_feels. Perceived weather temperature in celsius degrees.

humidity. The relative humidity is expressed as a percent.

wind_speed. Wind speed in miles per hour.

weather_code. A code identifying the weather; e.g., sunny = 1, cloudy = 2, etc.

bike_rented. A nominal attribute indicates the number of bikes rented. The classes of this attribute are: "very low" (less than 170 bikes rented), "low" (between 170 and 600 bikes rented), "medium" (between 600 and 1,100 bikes rented), "high" (between 1,100 and 1,900 bikes rented), "very high" (more than 1,900 bikes rented).

You can access to the full dataset in UniHub.

The Challenge

Your goal is to predict the attribute **bike_rented**, indicating the bike sharing usage in each day and in each hour of the day, so as to help TfL to balance bike sharing supply with bike sharing demand. To this end, you need to divide the data into training and testing (using k-fold cross-validation) and build a model using your data **(london_bike_data.csv file).** At the end check the accuracy of your model using Evaluation Metrics and explain your choices in detail. It is essential to provide a comprehensive explanation of the code, detailing the specific purpose of each line of code and its significance.

Your job

You need to work in teams of 3-4 people. Your team need:

- To Submit a PDF submission, which is worth 80 Marks.
- To Prepare a PowerPoint presentation, which is worth 20 Marks.

Your PDF submission

You need to submit a PDF Version of your Jupyter Notebook. download Jupyter notebook as PDF, Go to File->Download->Download as PDF.

Your PowerPoint presentation

You need to prepare a powerpoint presentation of 10-15 minutes covering your work done. You need to present your study in class and you will be evaluated according to:

- Your introduction to the challenge (5 marks)
- Your adopted solution and how you tested its accuracy (5 marks)
- Personal reflection on the issues you faced and how you solved them (5 marks)
- Take away message from this project and how you would improve in the future (5 marks)