ISYE 7406 Description of the Course Project

The Project Description is long. I appreciate your patience in reading the entire Description carefully and thoroughly.

As a part of the assigned work for this course, you are required to participate in a team and conduct a course project. The purpose of this project is to encourage you to explore an avenue related to, but not limited to, the material in the class. To perform well in the course project, you may need to read additional literature and broaden your knowledge base. The project will enable you to gain hands-on experience about developing data science-enabled fault and anomaly anticipation (or prediction).

Introduction

In this course project, you are given the data from five wind turbines on a wind farm belonging to the fourth largest wind energy operator, EDP Renewables (referred to as EDP hereinafter). You are asked to develop an early warning system to predict or anticipate the occurrence of failures or faults on the wind turbines. There are two-year worth of data to be used in this project, of which 20 months are used as the training set and 4 months are used as the test set. At the beginning of the semester, almost all data are made available in Canvas, except the actual failure events in the 4-month test period, which are withheld and will be used to test the quality of your method.

There are five turbines, labeled as T01, T06, T07, T09, T11. The failure and fault events that happened during the two full years on the five turbines are registered with five subsystems: gearbox, generator, generator bearing, transformer, and hydraulic group. Modern engineering systems are reasonably robust so that there are not that many failures. The number of failures that happened in the training period, i.e., the first 20 months, are 23 events—this translates to fewer than one fault per turbine per category in the 20-month period. If the failure event frequency does not change in the test period, one can expect 4-5 failures in that 4-month period (the exact number of failures in the test period is withheld).

In the wind turbine maintenance (or generally all maintenances), it is important to foretell the happening of a failure event, so that an early warning can be issued and countermeasures be taken to either prevent the catastrophe or prepare early enough for the inevitability. A utility function is used by EDP to weigh costs and benefits, while factoring in how early ahead a failure event is predicted. The specific formular is as follows:

$$\text{Total saving} = \sum_{\#TP}\left\{(R-M)\cdot\frac{\text{Lead Time}}{\text{Allowable Time Window}}\right\} - \#FN\cdot R - \#FP\cdot I,$$

where

- *R*, *M*, *I* are the costs associated with replacement, maintenance (i.e., repair), and inspection, respectively. Their values differ for different subsystems, and the specific euro amount is shown in Table 1.

Table 1. Cost values specified by EDP for use in their utility function (in euro)

Subsystems	***R***	***M***	***I***
Gearbox	100,000	20,000	5,000
Generator	60,000	15,000	5,000
Generator bearings	30,000	12,500	4,500
Transformer	50,000	3,500	1,500
Hydraulic	20,000	3,000	2,000

- *#TP*, *#FN*, and *#FP* are the number of true positives (correct detections), false negatives (missed detections), and false positives (false alarms) of a detection/prediction method, respectively.
- Lead time is in the unit of days, measuring how early a method anticipates a fault event ahead of its actual occurrence.
- Allowable Time Window is also in the unit of days, which sets the earliest possible time that a fault prediction could be taken seriously. If a Lead Time is greater than the Allowable Time Window, then the detection is no longer considered a true positive, but treated as a false positive. In this course project, the Allowable Time Window is set to be 60 days.

How are *#TP*, *#FN*, and *#FP* decided?

When a method is applied to the test data, the method should issue a number of alarm warnings. Each warning is characterized by three parameters: the day of issuance, which turbine, and which subsystem. These warnings are compared with the withheld true events and they are deemed a true positive if

1. The alarm's turbine/subsystem matches with the true event's turbine/subsystem,
2. The day of issuance is within 60 days of the true event (i.e., Day Issuance – Day Event ≤ 60).

Otherwise, the alarm warning is considered as a false positive. For each true event, if there is no alarm issued within 60 days prior to its occurrence, then there is a false negative.

In the case where a method issues multiple alarms within 60 days of an event, the first (earliest) alarm will be treated as the alarm and all others will be discarded. It is to your team's benefit to make sure that all final alarms you submitted are more than 60 days apart from each other.

Ranking and winner of the course project competition

I would like to run this course project as a competition. All teams will be ranked by their method's performance and the winning team will receive a free lunch from me (we will find a nice restaurant in mid-town).

The total saving for the five turbines, as measured on the test period data, will be used to rank the teams in the class. Whoever attains the highest saving is the winner of the course project competition. If all team's total savings are negative, then it means that the failure warning system

does not save but cost money. Then, whichever team costs the least is the winner.

You do not have to use the same method for all five subsystems or all five turbines. You have the discretion to choose however you want to build your method for issuing early warnings. Your team's performance will be judged by the same criterion regardless of your strategy.

You can also choose to work on a single or a subset of the subsystems, rather than all five of them. Let us say that you have an effective method that only works for gearbox and it predicts correctly the happening of gearbox failures. Since your method does not work for other subsystems and will lead to false negatives for the failures associated with other subsystems in the test data. The costs associated with these false negatives will be computed accordingly and factored in when determining the total savings.

One more piece of information given here is that in the specific test dataset we use for this course project, there are only gearbox faults and hydraulic faults (the number of faults, total or individual category, is withheld). Other subsystems do not experience a fault/failure event in the 4-month test period.

Datasets and variable descriptions

(**Please use the datasets ONLY for your ISYE 7406 course project. Do NOT share or repost in any fashion or anywhere else any of the files in the *Dataset Files* Folder. Please delete all local copies of the dataset files from your devices when the Spring 2024 Semester is over, but no later than May 3, 2024. Failure to observe this rule is considered a violation of the Georgia Tech Honor Code.**)

There are five data files to be uploaded in the Datasets and Descriptions module block on Canvas. They are:

1. wind-farm-1-signals-training.csv: This is the SCADA signals measured on five turbines in the 20-month training period (SCADA stands for Supervised Control And Data Acquisition).
2. wind-farm-1-signals-testing.csv: This is the SCADA signals measured on five turbines in the 4-month test period.
3. wind-farm-1-metmast-training.csv: This is the weather data measured on the meteorological mast (referred to as met mast) on the wind farm in the 20-month training period. Because there is a single met mast, the weather data is common for all five turbines.
4. wind-farm-1-metmast-testing.csv: This is the weather data measured on the meteorological mast (referred to as met mast) on the wind farm in the 4-month test period.
5. htw-failures-training.csv: This is the list of fault and failure events in the 20-month training period.

In the same module block, there are two description/documentation files:

A. Wind Farm - Signals Variables.pdf, which explains the variable headers in the SCADA signal data files.
B. WindTurbine_Datasheet.pdf, which provides some basic information about the wind

turbines dealt with in this course project.

Project logistics and evaluation

Teaming: A team should include 2-5 students, i.e., no fewer than 2 students and no more than 5 students.

Report and Presentation: Each team should submit one written report and make a set of Power Point slides for project presentation.

Format of your report: You need to start with a single-page executive summary (no more than one page) to state what you have done and the insights you gained from doing the project, that is, *anything* that you feel you have a better understanding because of doing this project. The main text of the report should clearly present your approaches, justification, results, and conclusion. You can briefly recap the problem but there is no need for any lengthy repetition of the problem described in this project description.

There is no page limit for the final report. Nevertheless, keep in mind that the grade is given upon the quality instead of the length of your report. Also, please **do NOT include** any code or pseudo-code in your report.

When you have formatting and writing style questions, you can consult the publications in *IISE Transactions* for guidance. This is not to ask you to write a paper to be submitted to *IISE Transactions*. Instead, it is to give you an idea how professionals in our field write a formal technical report (a published paper is a technical report by itself), i.e., how they lay out their problem, argue their cases, and reach their conclusions. Papers in the *Transactions* also show you how to present your figures, how to present your tables, and how to cite a source and list the references.

Please make sure to write your report in a single column format.

Format of your Power Point slides: Your set of slides will start with a title slide, which includes the title of your project and the team members. The total number of slides, including the title slide, MUST be equal to, or fewer than, **ten** (**10**). Inclusion of each EXTRA slide will receive one point reduction in the project score, until the project score reaches zero. Suppose that you have 30 slides. It gives you – 20 points. Even if your team has done everything else perfectly, gaining +25 points, your project score will still be five (5).

Timeline:

Formation of team: The list of students in your team is due on **Feb 15, 2024, by 5PM**. Please email your team formation (names and GTID of your team members) to TA's email address, which will be announced in Canvas.

Submission of your failure early warnings: by **5PM on April 15**. This submission will be used to determine the ranking of the teams in the competition and select the winner. You can change your results in your report or presentation. But those changes

made after your submission do not affect your team's ranking. You can submit multiple times but the last submission before the deadline is used as your team's true submission. If after submitting a new result, you regret and believe an earlier submission would be better, you will have to resubmit the earlier submission AGAIN before the submission deadline.

The total saving and ranking of all teams will be released after last team's presentation on April 23.

Project presentation: April 16, 18, and 23.

Report/Slides submission:

- The written report and the slides of your presentation are due by **5PM on April 24**. One submission per team, including a written report and a file of slides.
- **Electronic submission is required**, so the electronic time stamp will be used to determine whether the report is submitted on time. The report and slide files should be submitted in GT Canvas.
- **You must submit both report and presentation in PDF files**. Canvas will not take other file formats.
- A late submission will be penalized 0.5 points for the first hour it is late and an additional 0.5 points per hour it is late. The time is counted starting from 5 PM on the project due date. This means that if your submission happens after 5PM and before 6PM, your project score will be deducted 0.5 points, and if it is after 6PM and before 7PM, your project score will be deducted 1 point, and so on. A complete submission is defined as submitting ALL required files in the required format. Anything other than a complete submission is considered a "late submission."

Grading: A total of 25 points are allocated to this project, where 5 points are based on the performance of your method and 20 points are for the report and presentation slides combined. Your method's performance score is decided through a comparison with the best team's method and the benchmark method.

The benchmark method used here is to assume no detection at all. So there are no true or false positives but only false negatives. The total saving is calculated based on all false negatives (the resulting value is guaranteed to be negative, i.e., a cost and no saving).

Denote by TS_0 the total saving (sign included) using the benchmark method, by TS_{best} the highest total saving value obtained among all ISYE 7406 teams, and by TS the total saving of your team's method. Assuming $TS_{best} > TS_0$. Then your team's score for the performance part, out of the 5 points, is

$$\text{performance score } = \frac{TS - TS_0}{TS_{best} - TS_0} \times 5 \, .$$

Obviously, the winning team, whose $TS = TS_{best}$, receives the full 5 points for the performance part. The lowest possible performance score is zero, even if the above formula returns a negative

value. In the case when TS_{best} is equal to or smaller than TS_0, i.e., when all teams perform no better than the benchmark method, then the performance score of all teams will be assigned zero.

Regarding the report and presentation slides, emphasis will be given to the sophistication of your methodology as demonstrated in your reasoning/arguments and results (10 points). Please note that *sophistication* of your methodology does not equate *complexity* of a method. If you believe a simple method will do the best for your problem, it is Okay but you still need to provide convincing arguments and good test results to support your claim. In addition to the sophistication consideration, other considerations for grading include the organization and clarity of your report and presentation (5 points for the report and 5 points for presentation slides).

Mandatory point deductions:

-1	per extra slide beyond 10;
-1	executive summary longer than one page;
-1	report not in a single column format;
-0.5	per hour the report/presentation/outcome is late.