STA6704 Data Mining Methodology II

Spring 2017 Midterm

The midterm is the Siemens 2017 Wind Analytics Contest. Please see the attached files for the problem statement, official rules, general requirements, required materials, and report format, etc. Students can submit as individual or on teams not larger than 5 people.

DAEDLINEs:
All registrations must be received by February 28, 2017. Final contest projects must be submitted in full no later than March 12.

Submission:
On or before the due date, contestant should email the required materials to **BOTH daoji.li@ucf.edu and jenzelmanski@yahoo.com** with subject line "Siemens 2017 Wind Analytics Contest" Include your registration ID and contact information in the email. Results file should be in this format:

Results file: Separate Cover page with team identification information, Code files and report with the file name being ID_x where "x" is your registration ID. All files should be in one compressed ZIP file.

Siemens 2017 Wind Analytics Contest

Siemens Wind Power Inc. and the Data Mining program in the Department of Statistics at University of Central Florida are pleased to announce a Data Analytics contest using real life (but disguised) data from wind turbines.

Background

During operation, wind turbines automatically generate event information, warnings, and faults; some of which then cause the turbine to shut down and require intervention before restart. Wind turbines are maintained by technicians, who visit the turbines when some action is needed. Visits are always performed by teams of at least two, and travel time between the maintenance building and turbines can be as much as 30 minutes or more.

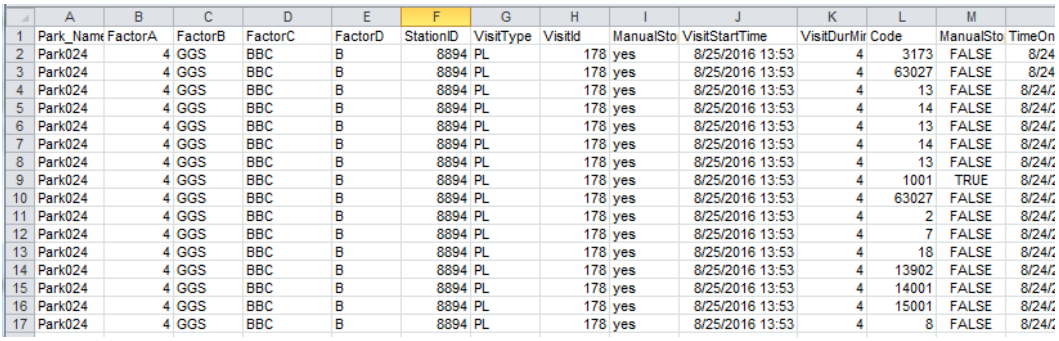
A manual switch is located at the base of every turbine, and upon arriving at the turbine, the mandatory procedure is that the technician switches this from remote to local operation. This signal has been utilized to determine that the turbine was visited, and to track the Park operations with regards to the frequency with which they personally visit a turbine.

Problem overview

Data on visits made to turbines by technicians has been compiled. A subset of this visit information is provided for this contest.

The main Visit Data File contains park, turbine, visitID, and visit start date/time data, as well as some park-specific coded attributes or characteristics (Factors A through D) which may or may not have any correlation with the Visits.

Also provided is historical turbine-specific data from the information log: numerical codes of events, warnings, and faults for each specific turbine organized by the unique Visit ID. The list of codes experienced, including the time on and off of each code, is provided for this timeframe relative to the visit start: history going back in time from 12:01AM the day ***prior*** to the visit, up until midnight on the day ***of*** the visit. (Note: For event codes, there is only a time-on associated: events do not have a Time Off.)



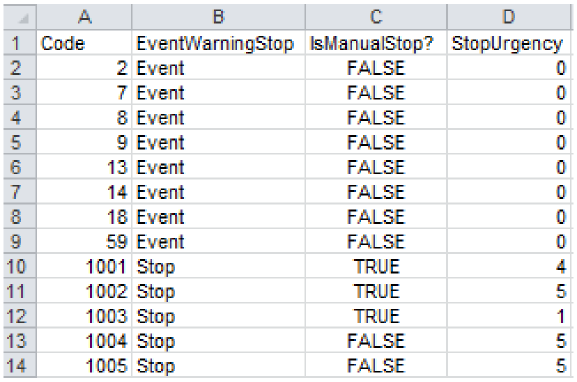
	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	Park_Name	FactorA	FactorB	FactorC	FactorD	StationID	VisitType	VisitId	ManualSto	VisitStartTime	VisitDurMir	Code	ManualSto	TimeOn
2	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	3173	FALSE	8/24
3	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	63027	FALSE	8/24
4	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13	FALSE	8/24/
5	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	14	FALSE	8/24/
6	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13	FALSE	8/24/
7	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	14	FALSE	8/24/
8	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13	FALSE	8/24/
9	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	1001	TRUE	8/24/
10	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	63027	FALSE	8/24/
11	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	2	FALSE	8/24/
12	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	7	FALSE	8/24/
13	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	18	FALSE	8/24/
14	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13902	FALSE	8/24/
15	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	14001	FALSE	8/24/
16	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	15001	FALSE	8/24/
17	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	8	FALSE	8/24/

The primary goals: to ascertain all possible information and intelligence, associations and interactions, from what the turbine *itself* can tell us through its automated event and fault logs, about why the technicians visited the turbines; to search for and categorize patterns exhibited prior to visit start, the time during the visit, and/or the time after the visit is complete, from the Information Log.

The intent is to look for common themes, patterns, sequences, and/or instances of the various turbine codes in order to identify a cause or situation associated with technician visit. (It is understood and expected that students’ responses and findings will need to be interpreted by Siemens technical personnel, who will provide meaning through proprietary controls and software knowledge.)

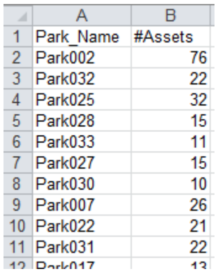
There are status signals included in the Information Log which indicate something (normal) happened; so not every code indicates something unexpected or an error state.

Provided in an additional data table is a list of the information codes, classified as Event (information only), a Warning (oddity, but safe operation of turbine continues), or a stop (abnormality detected which causes the turbine to shut down); and whether or not the code is a manually-initiated stop (by human intervention); and a StopUrgency code (0 means no stop; 1 means the slowest, gentlest equipment shutdown, while 6 means an immediate stop of the turbine.) Descriptions of the codes are not provided; this data table is provided in case any correlations or patterns exist only within a particular subgroup of codes (i.e. events vs. warnings vs. stops).



	A	B	C	D
1	Code	EventWarningStop	IsManualStop?	StopUrgency
2	2	Event	FALSE	0
3	7	Event	FALSE	0
4	8	Event	FALSE	0
5	9	Event	FALSE	0
6	13	Event	FALSE	0
7	14	Event	FALSE	0
8	18	Event	FALSE	0
9	59	Event	FALSE	0
10	1001	Stop	TRUE	4
11	1002	Stop	TRUE	5
12	1003	Stop	TRUE	1
13	1004	Stop	FALSE	5
14	1005	Stop	FALSE	5

Also provided is a data table listing Park_Names and the Number of Assets (StationIDs) located at each park, for use in determining the extent or percent of Park Assets visited in a single day:



	A	B
1	Park_Name	#Assets
2	Park002	76
3	Park032	22
4	Park025	32
5	Park028	15
6	Park033	11
7	Park027	15
8	Park030	10
9	Park007	26
10	Park022	21
11	Park031	22
12	Park017	13

The Code (1020) which indicates a Visit may occur multiple times within a single VisitID’s information code history. You can tell which Visit Code is the one corresponding to the start of the official visit by comparing the VisitStartTime to the TimeOn(Alarms).

Contest tasks

1. Investigate the codes leading up to the visit for patterns in seeing either particular codes or a specific pattern of code sequence. Ideas: look at ***between/among different sites***, ***different groups of turbine types, seasonality indications, time of day*** the visit occurs (within vs outside expected working day hours).

   Generate common sequences or paths among the all alarm sequences. Can visits be segmented according to these common paths?

2. Categorize / cluster analysis; correlations: investigate commonalities by various factors: time of day of visit start, quantities of codes associated with the visit, what percent of total turbines at the same Park were visited on the same day, similarity of type or pattern of associated codes, length of the visit (very short, <5 minutes vs 25 minutes), and/or other various provided Factors.

> What variables are significant indicators of visit duration or visit segmentation (from task 1 above)?

3. Consider not just patterns and occurrence of codes, but also the ***timing of the codes*** relative to each other. Your analysis should distinguish between a code sequence spread out over many hours vs. just a few minutes. Remember, some codes occur on every turbine on a periodic basis as a matter of normal operation; so the occurrence of a code may, or may not, be relevant to the visit.

4. Consider breaking the analysis into codes and patterns occurring ***prior*** to the Visit code, separately from those occurring ***on*** or ***after*** the Visit code starts. Classifying visits in groups or clusters with similar "before visit" behavior vs. "after visit" behavior can provide valuable insight. There may be codes during the visit which indicate similar work was performed, even if/when no pattern is apparent in the codes leading up to the visit, and vice versa. Be able to capture a similarity, even if it only exists in the pre-visit codes or in the during-visit (or after-visit) codes.

5. Take special note of ***many turbines at a Park being visited on the same day***. It may be useful to specifically seek out visits occurring on the same day at a site, as this could be an indication of a common task or cause. On the other hand, there may be a pattern or oddity noted in investigating only "one-off" visits (where only one turbine experienced a visit at a Park in a single day) vs. the occurrence of multiple visits to various turbines at a Park in a day.

> Not every instance of a local/remote switch signal input might actually be caused by a physical turning of the switch at the turbine. It is possible that some sort of electrical glitch could cause the switch input signal to latch in, when a technician is not actually activating the input. Absence of additional turbine codes in the hours proceeding or during the visit could indicate this. A second clue to this situation would be observing many visit starts occurring simultaneously (or near-simultaneously) across a significant portion of turbines at any specific Park (it's unlikely that a limited # of teams would be at every turbine at once and/or putting the turbine into local mode simultaneously). A third clue to this situation could be observing abnormal Visit Start Times outside of what is normally seen at a Park -- generally before 6AM or after 8PM. Siemens is particularly interested in identifying any code "signature" which indicates visits could fall into this category of "potential electrical glitches".

Dataset details

Students will be provided with one main large dataset containing information codes and timing associated with each recorded Visit. Field descriptions (column headers) are as follows:

Park_Name – designation of the windfarm

FactorA, FactorB, FactorC, FactorD – various disguised characteristics of the parks

StationID – unique wind turbine identifier

VisitType: indication (by technicians) as to the category of visit

VisitID – unique identifier for each visit

Manual Stop during Visit – flag which indicates true/false if was there were any manual stop codes registered among the provided alarm history for each visit

VisitStartTime – the date and time stamp of the "time on" of the Visit code (one unique timestamp per VisitID)

VisitDurMinutes: length of visit in minutes as documented in the Visit History Table (calculation defined as the VisitCode 1020's (TimeOff – TimeOn); but if more than one 1020 occurred the same day, then the calculation is (the last 1020's TimeOff – the first 1020's TimeOn). Due to a data validation algorithm, you will notice that *not every* 1020 will be interpreted as a visit with a VisitID.)

Code – a number representing the actual alarm, event, or fault code from the turbine's historical information log

ManualStop – True/False for each code; indicates if that code is initiated by command of a person (either locally at the turbine or remotely through the software). (This field is also seen in the Code Listing data table (mentioned below) and was included in the main dataset for convenience.)

TimeOff – when the code cleared (date/time format)

TimeOn – when the code came on (date/time format)

In addition, two additional data tables are provided:

> 2 – Park Size listing each Park_Name and the Total # of Turbines (StationIDs) located at that Park
>
> 3 – Code Listing identifying each code as Event, Warning, Fault; if it's a manually-initiated stop; and a Stop Urgency code (0 = it's not a stop; 1 = slowest, gentlest stop sequence, up to 6 = the fastest most urgent stop of the turbine)

2017 Wind Analytics Contest
Sponsored by Siemens Wind Power Inc.
OFFICIAL RULES
NO PURCHASE NECESSARY

HOW TO ENTER: Click “Register Now” button and fill out the registration form at the following URL:

http://sciences.ucf.edu/statistics/dms/siemens-2017-wind-analytics-contest-registration-form/

Once registered, a registration ID will be assigned to each contestant by the system. The contestant will receive an email with the registration ID and a link to download the datasets.

DEADLINES:
All registrations must be received by February 28, 2017. Final contest projects, as outlined below, must be submitted in full no later than March 12. Winners will be announced no later than March 20, 2017.

ELIGIBILITY: The contest is open to current Full-time and Part-time UCF students. **Students can submit as individual or on teams not larger than 5 people**. Employees (including immediate family members and/or those living in the same household of each) of Siemens Wind Power Inc., their advertising, promotion and production agencies, the affiliated companies of each, and the immediate family members of each are not eligible.

PRIZE: Three (3) winning teams (first place, second place, and third place) will be recognized with $500, $300, and $200 cash prizes, respectively.

GENERAL REQUIREMENTS:
• Entries must be the original work of the submitter and/or his/her team; must be suitable for publication; and must not infringe third- party rights. .
• Participants may only use the variables provided in the original data or transformed variables as a result of their work.
• Teams must include their assigned team number (issued at registration) in the header or footer of all materials uploaded for judging.
• The required cover page should be the only place where it is possible to identify the University, Department(s), and/or the individuals involved in the submission.

REQUIRED MATERIALS:
1. **A separate cover page** that references the department, team members, supporting faculty member, primary contact, team number assigned at registration and date submitted. The cover page will be removed for judging.
2. **Project report** that includes the following sections: executive summary, data analysis, modeling approach, results and conclusions (see judging criteria in the next section). Please do not include any school/member identifying information on the report other than the team number.
3. Include any/all programming code and/or flows to be examined by the judges.
4. Provide instructions for running the code and/or flows, including estimated running time with brief description of hardware environment, neatly organized code and/or flow running order, and clear labels for the order of execution.
5. **Project codes** can be in SAS, R, Python, C, C++ or MATLAB. Code files type should be compatible with the software used (example R file should be in “*.R”, Python file should be in “*.py”)

REPORT FORMAT:
All submissions should follow this suggested (but not required) format:
1. Margins: 1” from all sides
2. Typeface: Times New Roman size 12
3. Spacing: 1.5
4. Paragraph style: Block style
5. Captions: Captions to be placed under tables or figures, numbered sequentially across the document \

6. Document size: maximum 30,000 words or 10 pages (no limit on charts, graphs, visualizations which are instrumental in conveying results)
7. Page numbering: Page number to be placed at the bottom of each page (e.g., Page 1 of x)
8. File type: PDF format

HOW TO SUBMIT RESULTS:
On or before the due date, contestant should email the required materials to jenzelmanski@yahoo.com with subject line “Siemens 2017 Wind Analytics Contest” Include your registration ID and contact information in the email. Results file should be in this format:

Results file: Separate Cover page with team identification information, Code files and report with the file name being ID_x where “x” is your registration ID. All files should be in one compressed ZIP file.

JUDGING:
All entries will be judged based on the following criteria:
• 75% of the total points will be awarded based on the content in the Project Report. Major sections should include:

- Executive Summary: – 5%
- Exploratory Data Analysis – 20%
- Modeling Approach: rationale for modeling approach(es), selection, and evaluation process; Creativity in partitioning / slicing of data to obtain insights as described in Contest Tasks – 30%
- Results and Conclusions: clarity of interpretations, visualizations, and implications – 20%

• 25% of the points will be awarded based on the ability for the judges to replicate the analysis using the team’s code on a test dataset

Judging will be based on the application of the data, the method(s) used to reach conclusions, the number of significant relationships or correlations unearthed, and a solid explanation of the findings.

Important Dates:

Contest start date: February 2, 2017 (data files will be available on this date)
Registration deadline: February 28, 2017
Final submissions due: On or before 11pm ET March 12, 2017
Notification of top three winners: March 20, 2017
Award ceremony: March 22, 2017

NOTIFICATION: Winner will be notified on March 22, 2017.

GENERAL: All federal, state and local laws and regulations apply. By accepting prize, winner consents to Sponsor’s use of their name and likeness without additional compensation, unless prohibited by law. By entering, you release and hold harmless Sponsor, its parent, subsidiaries, affiliates, employees and agents from any and all liability or any injuries, loss or damage arising from or in connection with participation in this promotion or acceptance/use of the prize. By entering, participants release Siemens Wind Power Inc., its affiliates, directors, officers, employees and agents from any and all liability with respect to all aspects of the contest.

Except where prohibited, acceptance of the prize constitutes winner’s consent to the use of his/her name, likeness and biographical information for advertising and promotional purposes, without limitation and without additional compensation.

Submission of any entry constitutes the entrant’s irrevocable, non-exclusive license to Siemens Wind Power Inc., and their agents to publish, use, adapt, edit and/or modify such entry in any way, in any and all media, without limitation, for use in association with advertising, promotion, archiving and review, including, without limitation, screenshots and selected portions of the winner’s entry, all without additional compensation.

Sponsor is not responsible for phone, technical, network, electronic, computer hardware or software failures of any kind, misdirected, incomplete, garbled or delayed transmissions. Sponsor will not be responsible for incorrect or inaccurate entry information, whether caused by entrants or by any of the equipment or programming associated with or utilized in the contest.

Entries must be the original work of the entrant; must be suitable for publication; and must not infringe third-party rights. All entries become the property of the sponsor. By entering, entrants acknowledge compliance with these official rules including all eligibility requirements. In the event of non-compliance with these requirements, the selected entrant may be disqualified and an alternate winner selected, at Sponsor’s discretion. Sponsor reserves the right to suspend, cancel, or modify this promotion if fraud or any other causes beyond its control destroys the integrity of the promotion, as determined by Sponsor’s sole discretion. If the promotion is cancelled, unawarded prizes may be returned to Sponsor or may be awarded by random drawing from eligible entries, to the extent a fair random drawing can be conducted, at Sponsor’s discretion.

SPONSOR: The sponsor of this promotion is Siemens Wind Power Inc.

For any further questions about the dataset, contact “jenzelmanski@yahoo.com”