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	8/24/2
5	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	14	FALSE	8/24/2
6	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13	FALSE	8/24/2
7	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	14	FALSE	8/24/2
8	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13	FALSE	8/24/2
9	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	1001	TRUE	8/24/2
10	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	63027	FALSE	8/24/2
11	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	2	FALSE	8/24/2
12	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	7	FALSE	8/24/2
13	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	18	FALSE	8/24/2
14	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	13902	FALSE	8/24/2
15	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	14001	FALSE	8/24/2
16	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	15001	FALSE	8/24/2
17	Park024	4	GGS	BBC	B	8894	PL	178	yes	8/25/2016 13:53	4	8	FALSE	8/24/2

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)