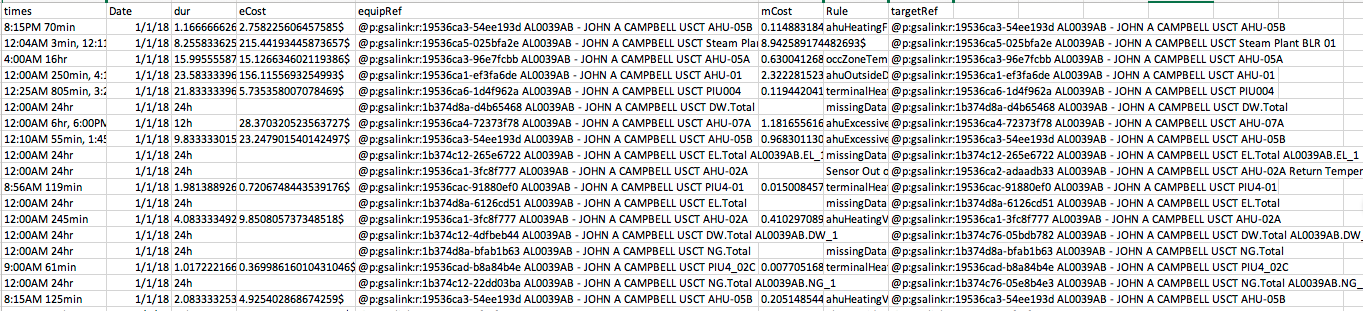
Feb 27

Task statement: checking how accurate eCost, sCost in the skyspark system are for certain spark rules



Is this statement accurate?

If the statement is accurate, what is the definition of eCost, is it more close to the definition of A or B as follows?

1. if I observe some spark rule is not triggered, I anticipate the energy expense of the building being xx amount different than when I observe the spark rule is triggered. In a bit more mathy language, this quantity is the following

The average energy consumption given we observe spark s is ON minus the average energy consumption given we observe spark s is OFF

A bit more mathier is like this

1. if I have two identical buildings at the location, both have the same rule r triggered, I magically fixed the problem in one of them. eCost is the difference in energy expenses between the two from the point when one building has its problem fixed. In a bit more mathy language, this quantity is the following

average energy consumption if we intervene and set the problem indicated by spark s to ON minus the average energy consumption if we intervene and set the problem indicated by spark s we observe spark s is OFF

A bit more mathier is like this

If the eCost is closer to the definition in A, then the key quantity to be estimated is the average building energy consumption given the observed states of a spark. In general, to answer this type of question, we want to

* First collect many variables that we think might make “guessing” the energy consumption easier. For example, outdoor temperature, day of the week, time of the day, season, etc.
* Then use the standard cross-validation technique to evaluate the accuracy of such guesses
* Use the model to estimate the two quantities: and
* Compute and compare with eCost

One subtle thing about the data is that the sparks and eCost are at component level, but the outcome energy consumption is at building level. We consider making the assumption that the effect is proportional to the service area of an equipment.

If the eCost is closer to the definition in B, then we need to first acknowledge the difficulty of this problem, and that our result is going to be very tentative. Let me explain it with the following toy example, following is what we could observe for a building with some spark s on component c

|  |  |  |  |
| --- | --- | --- | --- |
| time | Spark s status of building b component c | Energy of **building b at time t** when spark is OFF | Energy of **building b at time t** when spark is ON |
| 1 | OFF | 30 | m1 |
| 2 | OFF | 30 | m2 |
| 3 | OFF | 30 | m3 |
| 4 | OFF | 30 | m4 |
| 5 | ON | m5 | 60 |
| 6 | ON | m6 | 60 |
| 7 | ON | m7 | 60 |
| 8 | ON | m8 | 60 |

The m1-m8 are the missing data points we cannot observe, because we cannot see a spark being both on and off at the same time for the same building and the same equipment, unless parallel universe exist. The quantity we want if the eCost means definition B is

The missing data could be like this

|  |  |  |
| --- | --- | --- |
| Spark s status of building b component c | Energy of building b at time t when spark is OFF | Energy of building b at time t when spark is ON |
| OFF | 30 | 60 |
| OFF | 30 | 60 |
| OFF | 30 | 60 |
| OFF | 30 | 60 |
| ON | 30 | 60 |
| ON | 30 | 60 |
| ON | 30 | 60 |
| ON | 30 | 60 |

in this case the eCost is

The missing data could be like this

|  |  |  |
| --- | --- | --- |
| Spark s status of building b component c | Energy of building b at time t when spark is OFF | Energy of building b at time t when spark is ON |
| OFF | 30 | 30 |
| OFF | 30 | 30 |
| OFF | 30 | 30 |
| OFF | 30 | 30 |
| ON | 60 | 60 |
| ON | 60 | 60 |
| ON | 60 | 60 |
| ON | 60 | 60 |

in this case the eCost is

There could be many other values the missing data can take on, and we cannot know which one is the right one.

The best way around this issue is to conduct experiments: if we can randomly assign the spark indicated problems to building equipment, we could just observe the effect. If we cannot, we will go with the second-best option: making strong assumptions, and base the analysis based on the strong assumptions.

The problem definition is tricky: what is the target whose effect we are analyzing? My first thought is also, the spark itself, simple. However, if we look at the quantity closer, the effect we wanted might actually be the “problem” indicated by the sparks. In this case