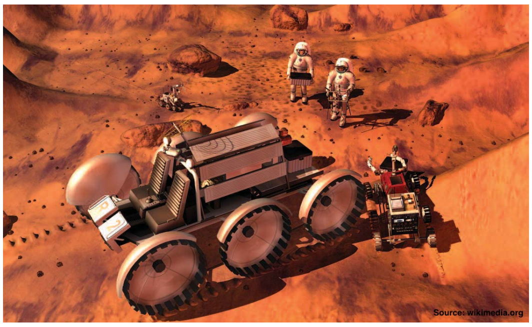
rapidminer

Data Science Competitions



Farming on “Mars”

September 12 – October 13, 2017

Welcome to the 2nd RapidMiner Data Science Challenge! We are very excited to bring this opportunity to our entire user community with more prizes and more fun! Please read this document carefully before beginning your race to the top of the leaderboard…

Disclaimer

This RapidMiner Data Science Challenge is based on real, earth-based data and its goal is to find an improved solution for a real-world problem. In order to make this challenge more interesting and allow for a better visualization of the problem, a fictional story about “Mars” has been created, which puts the provided data into a practical context. For reasons of confidentiality, all attribute names have been anonymized. *By participating in this challenge, you agree to only use the provided data in context of this challenge and not to re-distribute it to third parties.*

Background

One of the major challenges of the human colonization of “Mars” is the introduction of Earth-independent food production facilities, i.e. farming. A key element to farming on “Mars” will be the fertilization of available soil, which in its current state is not farmable due to a lack of nutrients.

In order to address this, an experimental setup has been created under “Martian” environmental conditions to produce bio-fertilizer made from algae and measure the usable yield after each production run. This yield varies based on the exact quantities of certain base nutrients and *the optional addition of one of two possible additional nutrients,* α or β*, inserted into the bio-fertilizer at some time t during the production run*. In addition, the yield of usable bio-fertilizer grown in identical conditions still varies per production run due to random events outside the controlled environment.

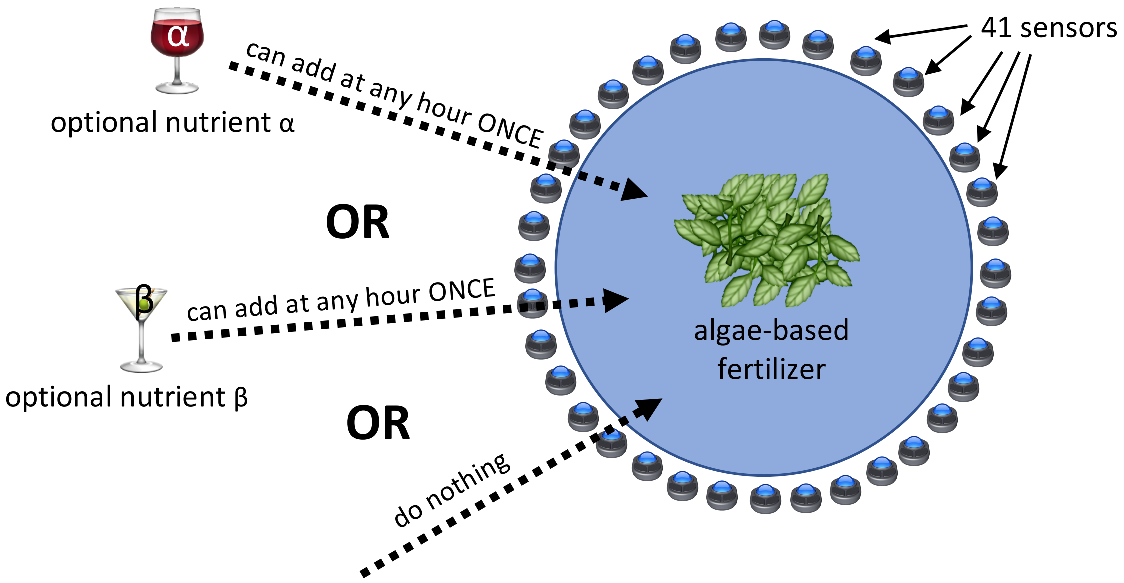
Although most production runs yield usable bio-fertilizer, some production runs may produce none due to random factors. Worse, *if the wrong additional nutrient, α or β, is inserted into the production run, there will be a net negative output of the production run*, irrespective of when it was inserted into the process (the output will be unusable and the input nutrients spoiled by the process). It is unknown which additional nutrient α or β will boost the yield of any production prior to the end of the process; only after the algae is harvested is it determined whether or not α or β would have benefitted. The insertion of either nutrient α or nutrient β (but not both) can be done at any time during the 36-hour production run. The insertion of either nutrient α or β into the production run is irreversible.

Hence production runs can have the following possible outcomes:

α or β inserted during the production run	**Nutrient that would have boosted yield as determined at the end of the production run, and its respective "Process Type"**	**Outcome**
neither	α –> Type A	baseline yield
neither	β –> Type B	baseline yield
α	α –> Type A	yield is increased vs baseline
α	β –> Type B	fertilizer spoiled
β	α –> Type A	fertilizer spoiled
β	β –> Type B	yield is increased vs baseline

Currently there is a net positive yield of fertilizer for large batches of production runs due to existing process controls without the insertion of either α or β. The configuration of these process controls has been done with the help of data from 41 sensors that take measurements of certain process values during each hour of a production run. See the attached Excel spreadsheet "training set - run 37 – annotated.xlsx" for an annotated version of a production run.

The current challenge faced by scientists lies in the *correct and timely identification of whether or not an optional nutrient,* α or β*, should be added while sensor data is received during a production run, and when*. Again, absolute certainty on which nutrient, α or β, would have boosted the yield is only available after the production run is completed, with no further possibilities to add the optional nutrient.



Scoring

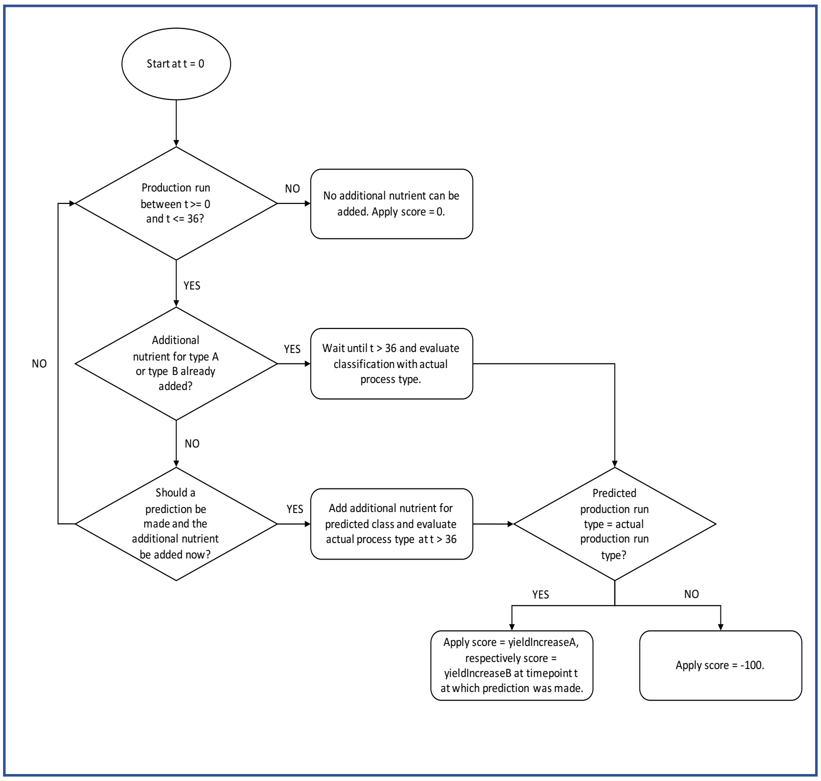
Incorrect predictions receive a score of –100. If the process type was correctly predicted, the score that is applied corresponds with the values of attributes *yieldIncreaseA* for correct predictions of process type A, respectively *yieldIncreaseB* for correct predictions of process type B at the respective point in time *t* in which the prediction has been made. The values for *yieldIncreaseA* and *yieldIncreaseB* are available for all values of *t*.

Those production runs for which no prediction on process type was made during the process run receive a score of 0, as no additional output could be created without the addition of optional nutrients α or β.

α **or** β **inserted during the production run**	**Nutrient that would have boosted yield as determined at the end of the production run, and its respective "Process Type"**	**Outcome**	**Production Run Score**
neither	α –> Type A	baseline yield	0
neither	β –> Type B	baseline yield	0
α	α –> Type A	yield is increased vs baseline	value of "yieldIncreaseA" at insertion time *t*
α	β –> Type B	fertilizer spoiled	-100
β	α –> Type A	fertilizer spoiled	-100
β	β –> Type B	yield is increased vs baseline	value of "yieldIncreaseB" at insertion time *t*

The figure below illustrates a flowchart of the classification and ultimate scoring of one single process run. Once the process run has started, there needs to be a regular evaluation at each point in time *t* (where $t \in \mathbb{Z}^+$, $0 \leq t \leq 36$), checking if the additional nutrients α (for processes of type A) or β (for processes of type B) can still be added.

The conditions for adding the optional nutrients are (1) that the process is still running ($t \geq 0$ & $t \leq 36$) and (2) that no optional nutrients (either α or β) have been added earlier in the production run.


Start at t = 0
Production run between t >= 0 and t <= 36?
NO
No additional nutrient can be added. Apply score = 0.
YES
Additional nutrient for type A or type B already added?
YES
Wait until t > 36 and evaluate classification with actual process type.
NO
NO
Should a prediction be made and the additional nutrient be added now?
YES
Add additional nutrient for predicted class and evaluate actual process type at t > 36
Predicted production run type = actual production run type?
YES
NO
Apply score = yieldIncreaseA, respectively score = yieldIncreaseB at timepoint t at which prediction was made.
Apply score = -100.

Once it is established that those two conditions are fulfilled and that therefore additional nutrients can be added to the process run at the respective point in time *t*, a decision needs to be made on the classification of the process run. If the decision is to not make a prediction on process type (i.e. classification) due to low confidence levels of the prediction or low predicted scoring values, the classification and scoring loops back to the previous step, which will evaluate the two required conditions for adding nutrients at the next point in time *t + 1*.

After all 178 production runs are completed, a cumulative score is determined:

Production Run	Correct Classification	Predicted Classification	Predicted Insertion Hour	Score
1	A	B	t=23	-100
…	…	…	…	
…	…	…	…	
50	B	B	t=17	yieldIncreaseB at t=17
…	…	…	…	…
…	…	…	…	…
90	A	A	t=29	yieldIncreaseA at t=29
…	…	…	…	…
…	…	…	…	…
178	A	none	n/a	0
			CUMULATIVE SCORE	[≥ 1000]

Figure 3: Cumulative Scoring illustrated for all 178 testing production runs

Current Conditions

The above described process controls (i.e. the current model) and the net process output are at a non-optimal level. The graph below illustrates the current cumulative score from a batch of 178 production runs: 1000. These 178 production runs also represent the data of the test set (see Data section for details) on which models built in this competition will be scored.

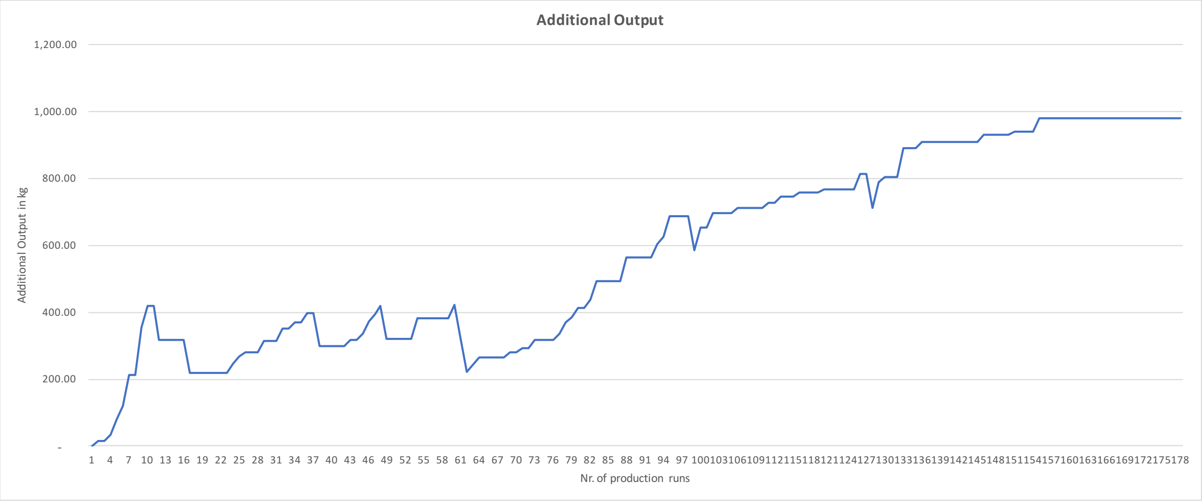


Figure 2: Additional output (in kg) for production runs in test set

Challenge

The goal of the challenge is to build a model that will classify which additional nutrient, α or β, and at what time *t*, will be most likely to boost yield during a production run. The metric to be optimized is the <u>cumulative score value of the same 178 production runs in the test set</u>; the baseline example above has a cumulative score value of 1000.

Model building needs to be done on the provided training set (see Data section for details), whereas the test set will be only used for scoring purposes. Each classification needs not only to include a predicted label (A or B) for the production run indicating which additional nutrient (α or β) was beneficial to the yield of that production run, but also the time *t* at which the additional nutrient was inserted. It is crucial that models only use data that has been available at the point in time in which the prediction is made, e.g. a prediction made at t = 17 can only use data from t = 0 to t = 17.

Data

The data consists of 1653 unique production runs. Each production run consists of the following attributes:

- 1 unique ID attribute, which acts as an identifier for the production run

- 1 label attribute, indicating the correct classification of the two possible process types, A or B as determined at the end of the production run
- 2 numeric scoring attributes, yieldIncreaseA and yieldIncreaseB, indicating the scores for a correct prediction of an A or B process type, respectively.
- 41 regular numeric attributes, sensor1 to sensor41, which represent values from sensors that were taken during the production run.

Whereas the Label attribute is obtained after completion of the production runs, all other attributes are available at each time point from t = 1 to t = 36. Furthermore, at t = 0 data are provided for the following attributes: Id, yieldIncreaseA, yieldIncreaseB and attribute41. Despite the sensors working reliably, there can be instances in which some data are missing due to malfunction of a sensor. This also includes possible missing values for yieldIncreaseA and yieldIncreaseB.

The data are split in two sets: the first contains 1475 production runs and represents the *training set*. The second set, consisting of 178 production runs is the *testing set* and will be used to score the models. Each set will be posted as a zip file containing one Excel spreadsheet per production run.

SUBMISSION AND EVALUATION

All submissions in this competition need to be posted in this thread with the entire XML of the process and the score. This includes the finished models, as well as the entire training process and all pre-processing steps. Processes that use scripts (e.g. within the Execute Script operator) or other third-party elements need to include reasonable commenting on that element. Participants are welcome to submit multiple entries; only the *last* entry will be reviewed for each participant prior to the submission deadline.
All submissions will be evaluated within 72 hours of submission and confirmed within this thread. Next to the scoring value on the test set which will be the basis for ranking submissions, it will be checked that trained models only use data available in the training set and that models only use data that has been available at the respective point in time in which the prediction has been made (see Task section for details).

The deadline for submissions is October 13, 2017 at 23:59:59 UTC.

RAPIDMINER SERVER INSTANCE

In order to increase the efficiency of model training and to demonstrate RapidMiner's powerful parallel processing capabilities with its new SaaS on Amazon AWS EC2 , *RapidMiner has agreed to provide a free Server EC2 instance for all participants for the duration of this competition*. This server instance can be used by any participant free of charge, as often as desired, for the duration of the competition as long as all use is restricted to this competition only. Participants wishing to use this server must send a private message (PM) to the RapidMiner Senior Community Manager, Scott Genzer (handle: @sgenzer) on the RapidMiner User Community to register and obtain the relevant connection details.

Winner and Prizes

The winner of the competition will be selected based on the highest aggregate score value of the 178 testing production runs ≥ 1000, after applying the test dataset to the submitted models. All submissions will be validated by RapidMiner and the competition's sponsor within 72 hours after their submission. The winners of this RapidMiner Data Science Challenge will be announced by October 17, 2017 in the competition's thread.

RapidMiner and the competition sponsor will award the following prices to the winners:

1st place: US$1000
2nd place: US$250
3rd place: US$100

PLUS all participants who submit a valid entry in the thread prior to the deadline will be eligible to win one or more amazing RapidMiner "swag" items. Supplies are limited and will be awarded on a first come-first served basis.

Restrictions

All participants of the RapidMiner Data Science Competitions must be registered users in good standing of the RapidMiner User Community and age 18 or older at the time of entry. Employees, directors, consultants, and any other persons affiliated with RapidMiner, Inc. are not eligible to participate in this competition.

GOOD LUCK AND MAY THE BEST MODEL WIN!!