***Title:***

To what extent can a Solver-based software reproduce the rating system in Formula-2?

***Subtitle:***

How accurately and precisely can a Solver-based software create the rating system of Formula-2, based on the volume of data fed provided?

***Author:***

Dániel Barnabás Bradali

***Institution of Studies:***

Óbudai Egyetem

***Abstract:***

 The article debates, whether it is possible to recreate the official (in long term unchanged) rating system fully, just by using an optimizer (Solver) and the data recorded from the previous years of Formula-2 (F2). To clarify, the usage of the previously mentioned software means, that after all the data was gathered and used as an input, the Solver should be able to find the official patterns (score system parameters), and even errors in the input (this will be discussed later), with which it may be able to recreate the so-called hidden rating system. The beauty of this method is that neither the author, nor the reader should be familiar with the proper working of the program, because its output can be directly checked by the developers/users and its “goodness”, can be determined (since the official rating system is available to anyone, who is interested) or even without the official rules of the scoring system – based only the knowledge: there is given a very simple rules system.

***Keywords:*** similarity analysis, staircase function, index of homogeneity as dynamic Goodness-term, approximation, impact of one single record,

***Hypothesis:***

 The aim of this article is to shed lights on how the volume and content of the data influence the preciseness of the program exploring hidden rules for scoring. It is expected, that if more data is available, then the accuracy of the robot will also increase, but how big are the steps, can a single record be ground-breaking, if so, then which one is the important (next) record, are the records weigh the same in every phase of the exploration (is every record has the same impact on the output, or are there any, that worth “more”)? Is there any room for improvement on how the program works, or is this the best way to approach the problem? All these questions will be discussed and evaluated throughout the article.

***Discussion:***

 In this part of the essay, the data bank will be discussed. Keep, that in mind, that most of the representation of it, will be partial, since it is too big to be included in the document. Therefore, the link for the complete data bank is given here (<https://miau.my-x.hu/miau/284/f2_2/f2_2.xlsx>), and all the required tables will be mentioned.


…(part of the whole dataset)…

Figure 1: Formula 2 official data from 2017

 As it can be seen in the first figure, the official data gathered looks like this, and can be found in the document (link above) on “Sheet 1”, with all the other data of year from 2018 till 2021. To understand this figure, the meaning of each column will be discussed. Firstly, pos in the first column is the aggregated position of each driver of the year, the one after this is the name of driver. Moving to the following columns, they represent the location of the race, and what was the position the driver had finished with, it is important to note, that the number 99 is a substitute for the fact, that a driver has not finished the given race, it is not part of the pointing system (this way, the solver can work with the data, and figure out, that the number 99 should not be taken into account, when trying to recreate the rating system). The last column is the aggregate points earned by each driver in the given year.


Figure 2: Formula 2 official data converted for the solver

 Continuing with the second figure, which can be found in the document (link above) as table 2017\_Y0, that only contains, the positions for each driver and each race. The solver (<https://miau.my-x.hu/myx-free/coco/index.html>) can only process numbers, so all the columns with names was left out, and an additional column was added to the end. It is worth to mention, that the method is shown and explained with the year 2017, but the document linked, shows every data for every year since Formula 2 exists.

 Moving on with the data, that the solver produced after the raw data was fed to it. Since, the volume of data is enormous, it is pointless to include any figure of it, but in the linked document above, 2017\_becsles table everything is included.

***Evaluation and Reflection:***

 To evaluate, it is clear, that the solver could not reproduce the original rating system but could estimate what it might be. The original rating system states, that only the first 10 driver should be awarded with points, but the solver gave points to the 20th driver, even though, most of the points were given to the first 5-6 drivers. The problem with the experiment is straight-forward, since Formula 2 began in 2017 there is not enough data for the solver to work with, or at least not work precisely enough to reproduce the official rating system, which raises the question: How could the solver work more efficiently, giving a satisfying answer to the original hypothesis? (c.f. <https://miau.my-x.hu/miau/284/kodtores.zip>)

 Firstly, by manipulating the original data by replacing the values from “not attended” to 99 (can be seen in the first figure), the working of the solver might be affected, therefore a better solution could be worked out.

 Secondly, not only the data volume should be bigger, but the quality of it could be also better, by this it is meant, that for instance, if the drivers would perform similarly in each of the races (the first driver wins most of the race and so on), it would be much easier to the solver to figure out, how much points to give out.

 Finally, combining the first two mentioned argument, it is clear, that the answer for the question is, increasing the volume and the quality of data. As it can be seen in the linked document, there is an aggregate solution, which contains all the raw data combined and fed to the solver. From this, it is clear, that the product of the solver is more precise (c.f. the ratio of the given points by the solver is higher in the higher ranks than before and/or the standard deviation for the potential ranking levels – so called stairs – is less, then before in the different cases with single years, etc.) since it only gives points to the first 14-15 driver and the ratios are better as well.

 Summing up, this experiment has shown, how useful and diverse the solver can be, although there are downsides of it as well, for example it can not be predicted, how much data is enough for the experiment to be successful, not to mention, that not every data (c.f. driver’s characteristic for a year) is equally worthy for the solver. That being said, further investigation can be done, aiming to answer the simple question, how much and which records do we need to achieve a solver, which gives the desired alternative rating system?

 It is important to say, that the solver does not know, that a system is existing behind the driver’s characteristics. I would be much easier to solve the problem, where the solver has further restrictions like each point for a ranking level should be the same in case of each attribute. In this less complex case, the question is only: for which ranking number what kind of point-value should be expected? The same question can be answered by solver, if the solver does also know that only the first n position can be leading to scores about zero.

 It is also a challenge to derive whether a scoring system might be seen as a constant system over all the years. If the aggregated points (Y) can not be re-calculated based on a solver solution, then different suspicions can be formulated like:

* The published/processed data have errors.
* The system of the scoring is not constant for the proved years at all.

***Target audience:***

The article is recommended to anyone, who is interested in the artificial intelligence, and how this kind of software, can solve problems, many believe to be restricted to humans, because of intuition. Furthermore, sport enthusiasts can enjoy this article too, since it can bring the analysis of racing sports to a new level, meaning more and more everyday-problems can be solved in the future, with programs like this, giving them opportunity to further improve and be suitable for more complicated problems (c.f. <https://miau.my-x.hu/miau/284/Formula2.docx>).

***Outer References:***

 My co-worker Baticz Levente worked on a similar project, dealing with the alternative rating system of Formula-2, raising questions as, who is best pilot in the season and discussing, whether someone achieving the same point as other pilots necessarily means, that they are equally good drivers? The topic of his work is not directly connected to mine, but all the measurements, were carried out together, therefore, the data provided is shared. His work can be read here: <https://miau.my-x.hu/miau/284/Formula2.docx>