Case: Apache Avro

INTRODUCTION

Apache-AVRO is a software environment for the objects serialization. A community of developers (https://issues.apache.org/jira/browse/AVRO), helps to identify bugs and improvements for the software, typically the signalman (who open a bug) is called Reporter.

Those bugs are then assigned to a member of the community (defined as Assignee, who may be the same as the Reporter).

The bugs can be in the following statuses:

1. **Open**: the issue or the improvement has been reported and it's ready to be handled by the Assignee;
2. **In progress**: the Assigne is working on the issue;
3. **Patch**: the solution of the issue has been identified and the patch has been made available;
4. **Resolved**: the Assignee reports that he solved the issue;
5. **Reopened**: the confirmation of resolution is not given;
6. **Closed**: the alert has been closed for the acceptance of resolution or a different motivation.

However, closing causes can be:

- **Cannot reproduce**: the user made a script with a specific issue that cannot be replicated by developers to do bugfixing.
- **Duplicated**: another alert for the same problem has been pointed out (identical, that include it or that is included in it). In this case, one of them is closed and the other follow a standard course.
- **Fixed**: problem solved.
- **Incomplete**: issue's description incomplete.
- **Invalid**: a fake problem was pointed out. In this case, an explanation of the error is given to the user and the issue is closed.
- **Not a problem**: reporting errors pointed out by the same reporter or problems already solved somewhere else.
- **Won't fix**: the issue is closed but a complete fix could not be made or an alternative way to bypass the problem was found.

The different types of closing causes could be roughly decreased to two principal ones:

1. closed because solved (**Fixed**) or
2. closed because deleted (**all others**).

Sometimes, the issue's assignment is immediate, but sometimes it's very time consuming (there are issues not assigned for years). There are also issues that are closed without being assigned: most of times they are "Cannot reproduce", "Duplicated", "Incomplete", "Invalid", "Not a problem", "Won't fix" closings (often identified by the same Reporter). Sometimes there are issues closed as "Fixed" without being assigned: they are generally quickly closed issues by the Reporter. As an initial approximation, the

"not assigned but closed" issues may be regarded as equivalent to that where the Reporter corresponds to the Assignee.

THE PROBLEM

The file AVRO-ISSUES.CSV include a subset of the information related to the open issues in almost 5 years. The complete dataset, in JSON format, can be found enclosed with this document. It is required:

1. To develop a data science model with the data in the CSV file to predict the required time to resolve an alert, based on its characteristics (for example "Issue Type", "Priority", how long has the alert been opened, Reporter type and/or Assignee type based on a previous profile, and every other information that could be relevant);
2. Implement the proposed model using the data and extract the predicted resolution time for 3 interesting cases;
3. Identify at least another information, available on the site of the community, that could be useful to improve the quality of the model;
4. (optional) All the Information on the website is included in the JSON file mentioned above. Extract from the file the information identified at the point 3 and insert it in the CSV.

DELIVERY

- The answers to the questions should be reported in a **short paper**, no longer than 10 pages, where the candidate will have to explain the choice of the software used, the model, the algorithm, the results about the three chosen cases (in addition attach the **new CSV** if it had been modified with respect to what it was provided point 4);
- The **code** which implement the model and all the analysis;
- Moreover, the candidate will have to prepare a **presentation**, no longer than 15 minutes, to show his own "data interpretation", persuading the committee about the effectiveness of the developed analysis and model.

The paper, code and presentation will have to be delivered no later than 1 week after receipt at the following e-mail addresses:

silvia.stellato@enel.com
rachel.koons@enel.com