Identifying Unmaintained Projects in GitHub

Jailton Coelho, Marco Tulio Valente
Federal University of Minas Gerais, Brazil
{jailtoncoelho,mtov}@dcc.ufmg.br

Luciana L. Silva
Federal Institute of Minas Gerais, Brazil
luciana.lourdes.silva@ifmg.edu.br

Emad Shihab
Concordia University, Canada
eshihab@encs.concordia.ca

ABSTRACT

Background: Open source software has an increasing importance in modern software development. However, there is also a growing concern on the sustainability of such projects, which are usually managed by a small number of developers, frequently working as volunteers. Aims: In this paper, we propose an approach to identify GitHub projects that are not actively maintained. Our goal is to alert users about the risks of using these projects and possibly motivate other developers to assume the maintenance of the projects. Method: We train machine learning models to identify unmaintained or sparsely maintained projects, based on a set of features about project activity (commits, forks, issues, etc.). We empirically validate the model with the best performance with the principal developers of 127 GitHub projects. Results: The proposed machine learning approach has a precision of 80%, based on the feedback of real open source developers; and a recall of 96%. We also show that our approach can be used to assess the risks of projects becoming unmaintained. Conclusions: The model proposed in this paper can be used by open source users and developers to identify GitHub projects that are not actively maintained anymore.

KEYWORDS

Unmaintained Projects, Prediction Models, Machine Learning, Open Source Software, GitHub

1 INTRODUCTION

Open source projects have an increasing relevance in modern software development [12]. For example, many critical software is currently available under open source licenses, including operating systems, compilers, databases, and web servers. Similarly, it is common nowadays to depend on open source libraries and frameworks when building and evolving proprietary software. For example, in a recent survey—conducted by Black Run’s consulting firm—78% of the over 1,300 companies surveyed acknowledged the use of open source software [1]. Concretely, Instagram—the popular photo-sharing social network—is currently implemented using more than 20 open source libraries [2]. Furthermore, the emergence of worldwide code sharing platforms—GitHub is the most well-known example—is contributing to transform open source development in a competitive market. Indeed, in a recent survey with open source maintainers we found that the most common reason for the failure of open source projects is the appearance of a stronger competitor in GitHub [9].

However, GitHub does not include clear and aggregated data about project status, in terms of maintenance activity. Users can access historical data about commits or global project metrics, like number of stars, forks, and watchers. However, based on the values of these metrics, they should judge themselves whether a project is being actively maintained (and therefore if it is worth to use it or not). Therefore, in this paper we propose and evaluate a machine learning approach to identify unmaintained (or sparsely maintained projects) on GitHub. Our goal is to provide a simple and effective mechanism to alert users about the risks of depending on a GitHub project. This information can also contribute to attract new maintainers to a project. For example, users of libraries facing the risks of continuation can be motivated to assume their maintenance.

Previous work in this area relies on the last commit activity to classify projects as unmaintained or in similar status. For example, Khondhu et al. use an one-year inactivity threshold to classify dormant projects on SourceForge [16]. The same threshold is used in works by Mens et al. [25], Izquierdo et al. [15], and in our previous work about the motivations for the failure of open source projects [9]. However, in this paper, we do not use this definition when investigating unmaintained projects due to three reasons. First, because defining a threshold to characterize unmaintained projects is not trivial. For example, in the mentioned works, this decision is arbitrary and it is not empirically validated. Second, our intention is to detect unmaintained projects as soon as possible; preferably, without having to wait for one year of inactivity. Third, our definition of unmaintained projects does not assume a complete absence of commits during a given period; instead, a project is considered unmaintained even when sporadic and few commits happen in a given time interval. Stated otherwise, by our definition, unmaintained projects do not necessarily need to be dead, deprecated or archived.

In this paper, we first train ten machine learning models to identify unmaintained projects, using as features standard metrics provided by GitHub about a project’s maintenance activity, e.g., number of commits, forks, issues, and pull requests. Then, we select the model with the best performance and validate it by means of a survey with the owners of projects classified as unmaintained and also with a set of deprecated GitHub projects. Particularly, we ask three research questions about properties of this model:

RQ1: What is the precision according to GitHub developers? The intention is to check precision in the field, according to the feedback provided by the principal developers of popular GitHub projects.

RQ2: What is the recall when identifying unmaintained projects? Usually, recall is more difficult to compute in the field, because it requires the identification of all unmaintained projects in GitHub. To circumvent this problem, we compute recall considering only

1 https://www.slideshare.net/blackduckssoftware/2015-future-of-open-source-survey-results
2 https://www.instagram.com/about/legal/libraries/
we remove 331 projects with null size, measured in lines of code.

Section 7 concludes the paper and outlines further work.

2 MACHINE LEARNING MODEL

In this section, we describe our machine learning approach to identify projects that are no longer under maintenance.

2.1 Experimental Design

Dataset. We start with a dataset containing the top-10,000 most starred projects on GitHub (in November, 2017). Stars—GitHub’s equivalent for likes in other social networks—is a common proxy for the popularity of GitHub projects [5]. Then, we follow three strategies, in order, to discard projects from this initial selection. First, we remove 2,810 repositories that have less than two years from the first to the last commit (because we need historical data to compute the features used by the prediction models). Second, we remove 331 projects with null size, measured in lines of code (typically, these projects are implemented in non-programming languages, like CSS, HTML, etc). Finally, we remove 74 non-software projects, which are identified by searching for the following topics: books and awesome-lists. We end up with a list of 6,785 projects.

Next, we label these projects in two groups: active and unmaintained. The active (or under maintenance) group is composed by 754 projects that have at least one release in the last month, including well-known projects, like facebook/react, d3/d3, and nodejs/node. Thus, we assume that projects with recent releases are active (under maintenance). By contrast, the unmaintained group is composed by 248 projects, including 104 projects that were explicitly declared by their principal developers as unmaintained in our previous work [9] and 144 archived projects. Archiving is a recent feature provided by GitHub that allows developers to explicitly move their projects to a read-only state. In this state, users cannot create issues, pull requests, or comments, but can still fork or star the projects.

Features. Our hypothesis is that a machine learning classifier can identify unmaintained projects by considering features about (1) projects, including number of forks, issues, pull requests, and commits; (2) contributors, including number of new and distinct contributors (the rationale is that maintenance activity might increase by attracting new developers); (3) project owners, including number of projects he/she owns and total number of commits in GitHub (the rationale is that maintenance might be affected when project owners have many projects on GitHub). In total, we consider 13 features, as described in Table 1. However, these features are not extracted considering the whole history of a project, but considering only the last \(n\) months, counting from the last commit; moreover, we collect each feature in intervals of \(m\) months. The goal is to derive temporal series of feature values, which can be used by a machine learning algorithm to infer trends in the project evolution, e.g., an increasing number of opened issues or a decreasing number of commits.

---

Table 1: Features used to identify unmaintained projects.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project</td>
<td>Forks</td>
<td>Number of forks created by developers</td>
</tr>
<tr>
<td></td>
<td>Open issues</td>
<td>Number of issues opened by developers</td>
</tr>
<tr>
<td></td>
<td>Closed issues</td>
<td>Number of issues closed by developers</td>
</tr>
<tr>
<td></td>
<td>Open pull requests</td>
<td>Number of pull requests opened by developers</td>
</tr>
<tr>
<td></td>
<td>Closed pull requests</td>
<td>Number of pull requests closed by the project developers</td>
</tr>
<tr>
<td></td>
<td>Merged pull requests</td>
<td>Number of pull requests merged by the project developers</td>
</tr>
<tr>
<td></td>
<td>Commits</td>
<td>Number of commits performed by developers</td>
</tr>
<tr>
<td></td>
<td>Max days without commits</td>
<td>Maximum number of consecutive days without commits</td>
</tr>
<tr>
<td></td>
<td>Max contributions by developer</td>
<td>Number of commits of the developer with the highest number of commits</td>
</tr>
<tr>
<td>Contributor</td>
<td>New contributors</td>
<td>Number of contributors who made their first commit in the considered period</td>
</tr>
<tr>
<td></td>
<td>Distinct contributors</td>
<td>Number of distinct contributors that committed in the considered period</td>
</tr>
<tr>
<td>Owner</td>
<td>Projects created by the owner</td>
<td>Number of projects created by a given owner</td>
</tr>
<tr>
<td></td>
<td>Number of commits of the owner</td>
<td>Number of commits performed by a given owner</td>
</tr>
</tbody>
</table>

---

\(^3\)READMEs are the first file a visitor is presented to when visiting a GitHub repository.\n
\(^4\)GitHub topics allow tagging a repository with keywords.
assuming that \( n \) is 24 months and that \( m \) is 3 months. In this case, for each feature, we collect 8 data points, i.e., feature values.

![Figure 1: Feature collection during 24 months in 3-month intervals.](image1)

We experiment with different combinations of \( n \) and \( m \); each one is called a scenario, in this paper. Table 2 describes the total number of data points extracted for each scenario. This number ranges from 13 data points per feature (scenario with features extracted in intervals of 6 months during 6 months) to 104 data points per feature (scenario with features extracted in intervals of 3 months during 24 months, as in Figure 1).

Table 2: Scenarios used to collect features and train the machine learning models (length and intervals are in months; data points is the total number of data points collected for each scenario).

<table>
<thead>
<tr>
<th>Scenario</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td>6</td>
<td>12</td>
<td>12</td>
<td>18</td>
<td>18</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Intervals</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>12</td>
<td>12</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Data points</td>
<td>26</td>
<td>13</td>
<td>52</td>
<td>26</td>
<td>13</td>
<td>78</td>
<td>39</td>
<td>104</td>
<td>52</td>
<td>26</td>
</tr>
</tbody>
</table>

Correlation Analysis. As usual in machine learning experiments, we remove correlated features, following the process described by Bao et al. [3]. To this purpose, we use a clustering analysis—as implemented in a R package named \texttt{Hmisc}—to derive a hierarchical clustering of correlations among data points (extracted for the features in each scenario). For sub-hierarchies with correlations larger than 0.7, we select only one data point for inclusion in the respective machine learning model, as common in other works [3, 38]. For example, Figure 2 shows the final hierarchical clustering for the scenario with 24 months, considering a 3-month interval (scenario 8). The analysis in this scenario checks correlations among 104 data points (13 features \( \times \) 8 data points per feature). As a result, 78 data points are removed due to correlations with other points and therefore do not appear in the dendogram presented in Figure 2. Finally, Table 3 shows the total number and percentage of data points removed in each scenario, after correlation analysis. As we can see, the percentage of removed points is relevant, ranging from 43% (scenario 7) to 75% (scenario 8).

![Figure 2: Correlation analysis for the 104 data points collected for the features in scenario 8 (24 months, 3-month interval). 78 data points (75%) are removed in this case, due to correlations with other data points, and therefore do no appear in this final clustering.](image2)

Machine Learning Classifier. We use the data points extracted in each scenario to train and test models for predicting whether a project is unmaintained. In other words, we train and test ten machine learning models, one for each scenario. After that, we select the best model/scenario to continue with the paper. Particularly, we use the Random Forest algorithm [6] to train the models because it has several advantages, such as robustness to noise and outliers [38]. In addition, it is adopted in many other software engineering works [2, 14, 26, 29, 31, 35]. We compare the result of Random Forest with two baselines: baseline #1 (all projects are predicted as unmaintained) and baseline #2 (random predictions). We use the Random Forest implementation provided by \texttt{randomForest}'s R package\footnote{https://cran.r-project.org/web/packages/randomForest/} and 5-fold stratified cross validation to evaluate the models effectiveness. In 5-fold cross validation, we randomly divide the dataset into five folds, where four folds are used to train a classifier and the remaining fold is used to test its performance. Specifically, stratified cross validation is a variant, where each fold has approximately the same proportion of each class [6]. We perform 100 rounds of experiments and report average results.

Evaluation Metrics. When evaluating the projects in the test fold, each project has four possible outcomes: (1) it is truly classified as unmaintained (True Positive); (2) it is classified as unmaintained but it is actually an active project (False Positive); (3) it is classified as an active project but it is actually an unmaintained one (False Negative); and (4) it is truly classified as an active project (True Negative). Considering these possible outcomes, we use six metrics to evaluate the performance of a classifier: precision, recall, F-measure, accuracy, AUC (Area Under Curve), and Kappa, which are commonly adopted in machine learning studies [11, 18, 22, 37, 38]. Precision and recall measure the correctness and completeness of the classifier, respectively. F-measure is the harmonic mean of precision and recall. Accuracy measures how many projects are classified
Table 4: Prediction results (mean of 100 iterations, using 5-cross validation); best results are in bold.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>0.5 Year</th>
<th>1 Year</th>
<th>1.5 Years</th>
<th>2 Years</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3 months</td>
<td>6 months</td>
<td>12 months</td>
<td>3 months</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.90</td>
<td>0.91</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td>Precision</td>
<td>0.83</td>
<td>0.87</td>
<td>0.87</td>
<td>0.84</td>
</tr>
<tr>
<td>Recall</td>
<td>0.78</td>
<td>0.74</td>
<td>0.77</td>
<td>0.75</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.80</td>
<td>0.79</td>
<td>0.81</td>
<td>0.79</td>
</tr>
<tr>
<td>Kappa</td>
<td>0.74</td>
<td>0.76</td>
<td>0.76</td>
<td>0.73</td>
</tr>
<tr>
<td>AUC</td>
<td>0.86</td>
<td>0.85</td>
<td>0.86</td>
<td>0.85</td>
</tr>
</tbody>
</table>

2.2 Experimental Results

Table 6: Top-5 most relevant features, by Mean Decrease Accuracy (MDA).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Period</th>
<th>MDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commits</td>
<td>T_{22,24}</td>
<td>38.5</td>
</tr>
<tr>
<td>Max days without commits</td>
<td>T_{22,24}</td>
<td>28.6</td>
</tr>
<tr>
<td>Max days without commits</td>
<td>T_{10,12}</td>
<td>21.9</td>
</tr>
<tr>
<td>Max contributions by developer</td>
<td>T_{16,18}</td>
<td>21.1</td>
</tr>
<tr>
<td>Closed issues</td>
<td>T_{1,3}</td>
<td>18.0</td>
</tr>
</tbody>
</table>

3 EMPIRICAL VALIDATION

In this section, we validate the proposed machine learning model by means of a survey with the owners of projects classified as unmaintained and also with a set of deprecated GitHub projects. Overall, our goal is to strengthen the confidence on the practical value of the model proposed in this work. Particularly, we provide answers to three research questions about this model:

RQ1: What is the precision according to GitHub developers?
RQ2: What is the recall when identifying deprecated projects?
RQ3: How early the model identifies unmaintained projects?

3.1 Methodology

RQ1: To answer RQ1, we conduct a survey with GitHub developers. To select the participants, we first apply the proposed machine learning model in all projects from our dataset that were not used in the model’s construction, totaling 5,783 projects (6,785 – 1,002 projects). Then, we select 2,856 projects classified as unmaintained by the proposed model. From this sample, we remove 264 projects whose developers were recently contacted in our previous surveys [9, 10]. We make this decision to do not interrupt again these developers, with new e-mails and questions. Finally, we remove 2,270 projects whose owners do not have a public e-mail address on GitHub. As a result, we obtain a list of 323 survey participants (2,856 – 2,270 – 264). However, before e-mailing these participants, the first author inspected the main page of each project on GitHub, to check whether it includes mentions to the project status, in terms of maintenance. We found 21 projects whose documentation states they are no longer maintained, by means of messages like this one:
This project is deprecated. It will not receive any future updates or bug fixes. If you are using it, please migrate to another solution.

Therefore, we do not send mails to the project owners, in such cases; and automatically consider these 21 projects as unmaintained.

Survey Period: The survey was performed in the first two weeks of May, 2018. It is important to highlight that the machine learning model was constructed using data collected on November, 2017. Therefore, the unmaintained predictions evaluated in the survey refer to this date. We wait five months to ask the developers about the status of their projects because it usually takes sometime until developers actually accept the unmaintained condition of their projects. In other words, this section is based on predictions performed and available on November, 2017. However, these predictions are validated five months later, on May, 2018.

Survey Pilot and Questions: Initially, we perform a pilot survey with 75 projects (∼25%), randomly selected among the remaining 302 projects (323 – 21 projects). We e-mail the principal developers of these projects with a single open question, asking them to confirm (or not) if their projects are unmaintained. We received 23 answers, which corresponds to a response ratio of 30.6%. Then, two authors of this paper analyzed together the received answers to derive a list of recurrent themes. As a result, the following three common maintainability status were identified:7

(1) **My project is under maintenance and new features are under implementation** (6 answers): As an example, we can mention the following answer:

> [Project-Name] is still maintained. I maintain the infrastructure side of the project myself (e.g. make sure it’s compatible with the latest Ruby version, coordinate PRs and issues, mailing list, etc.) while community provides features that are still missing. One such big feature is being developed as we speak and will be the highlight of the next release. (P57)

(2) **My project is finished and I only plan to fix important bugs** (13 answers): As an example, we mention the following answers:

> It’s just complete, at least for now. I still fix bugs on the rare occasion they are reported. (P10)

> I view it as basically “done”. I don’t think it needs any new features for the foreseeable future, and I don’t see any changes as urgent unless someone discovers a major security vulnerability or something. I will probably continue to make changes to it a couple times per year, but mostly bug fixes. (P68)

(3) **My project is deprecated and I do not plan to implement new features or fix bugs** (4 answers): As an example, we can mention the following answer:

> The project is unmaintained and I’ll archive it. (P74)

After the pilot study, we proceed with the survey, by e-mailing the remaining 227 projects. However, instead of asking an open question—as in the pilot—we provide an objective question to the survey participants, about the maintainability status of their projects. In this objective question, we ask the participants to select one (out of three) status identified in the pilot study, plus an other option. This fourth option also includes a text form for the participants detailing their answers, if desired. Essentially, we change to an objective question format to make answering the survey easier, but without limiting the respondents’ freedom to provide a different answer from the listed ones. In this final survey, we received 89 answers, representing a response ratio of 39.2%. When considering both phases (pilot and final survey), we sent 302 e-mails, received 112 answers, representing an overall response ratio of 37.1%. After adding the 21 projects that document their maintainability status, this empirical validation is based on 133 projects.

**RQ2:** To answer this second question, we construct a ground truth with projects that are no longer being maintained. First, we build a script to download the README (the main page of GitHub’s projects) and search for a list of sentences that are commonly used to declare the unmaintained state of GitHub projects. This list is reused from our previous work [9], where we study the motivations for the failure of open source projects. It includes 32 sentences; in Table 7, we show a partial list, with 17 sentences.

<table>
<thead>
<tr>
<th>Sentences documenting unmaintained projects</th>
</tr>
</thead>
<tbody>
<tr>
<td>no longer under development, no longer supported or updated, deprecation notice, dead project, deprecated, unmaintained, no longer being actively maintained, not maintained anymore, not under active development, no longer supported is not supported, is not more supported, no longer supported, new features should be expected, isn’t maintained anymore, no longer be shipping any updates, don’t want to maintain</td>
</tr>
</tbody>
</table>

We searched (in May, 2018) for these sentences in the README of 5,783 projects, which represent all 6,785 projects selected for this work minus 1,002 projects used in Section 2. In 451 READMEs (7.8%) we found the mentioned sentences. Then, the first author of this paper carefully inspected each README, to confirm the sentences indeed refer to the project’s status, in terms of maintenance. In the case of 112 projects (∼25%), he confirmed this fact. Therefore, these projects are considered as a ground truth for this investigation.8

**RQ3:** To answer this third research question, we rely on projects whose unmaintained status, as predicted by the proposed model, is confirmed by the participants of the survey conducted in RQ1. Then, we compute the number of days between November, 30, 2018 (when the machine learning model proposed in this paper was built) and the last commit of the mentioned projects. For projects where this interval is less than one year, there is a gain, compared with the strategy adopted in previous work [9, 15, 16, 25] that requires one year of commit inactivity to identify unmaintained projects.

### 3.2 Results

**RQ1: Precision according to GitHub developers**

Before presenting the precision results, Figure 3 shows the survey results, including answers retrieved from the project’s documentation, answers received in the pilot and answers received in the final

---

7Project names are omitted, to preserve the respondent’s anonymity; survey participants are identified by means of labels, in the format Pxx, where xx is an integer.

8Usually, the unconfirmed cases refer to the deprecation of specific elements, e.g., methods or classes.
survey. As we can see, the most common status, with 54 answers (42%) refers to finished projects, i.e., cases where maintainers see their projects as feature-completed and only plan to resume the maintenance work if relevant bugs are reported.\footnote{In our previous work\cite{9}, we also identified finished or completed open source projects. However, we argued these projects do not contradict Lehman’s Laws about software evolution\cite{21}, because they usually deal with a very stable or controlled environment (whereas Lehman’s Laws focus on E-type systems, where E stands for evolutionary). In fact, we manually classified the application domain of the 54 projects considered by their developers as finished in our survey. 41 (76%) are libraries and frameworks, 10 (18%) are software tools, and only 3 (6%) are applications.} We also received 41 answers (32%) mentioning the projects are deprecated and no further maintenance is planned, including the implementation of new features and bug fixes. Finally, we received 18 answers in the other option. In this case, four participants did not describe their answer or provide unclear answers; furthermore, one participant mentioned his project is in a limbo state:

> The status of [Project-Name] fits into a special category. Some of the tools it’s based on are either deprecated or not powerful enough for the goal of the project. This is part of the reason what’s keeping the project from being “done”. I would call this status limbo. (P24)

Seven participants answered their projects are stalled, as in this answer:

> It is under going a rewrite... but has been stalled based on my own priorities (P33)

To compute precision, we consider as true positive answers related to the following status: finished (54 answers), deprecated (41 answers), stalled (7 answers), and limbo (1 answer). The remaining answers are interpreted as false negatives, including answers mentioning that new features are being implemented (20 answers) and the answers associated to the fourth option (other option), but without including a description or with an unclear description (4 answers). By following this criteria, we received 103 true positive answers and 26 false negative ones, which results in a precision of 80%.

By validating the proposed model with 20 GitHub developers, we achieve a precision of 80%, which is a result very close to the one obtained when building the model (86%).

RQ2: Recall considering deprecated projects

The proposed machine learning model classifies 108 (out of 112) projects of the constructed ground truth as unmaintained, which represents a recall of 96%. This value is significantly greater than the one computed when testing the model in Section 2. Probably, this difference is explained by the fact that only projects that are completely unmaintained expose this situation in their READMEs. Therefore, it is easier to detect and identify this condition.

By validating the proposed model with projects that declare themselves as unmaintained, we achieve a recall of 96%.

RQ3: How early can we detect unmaintained projects?

77 (out of 103) projects classified as true positives by the surveyed developers have commits after November, 2016. Therefore, these projects would not be classified as unmaintained using the strategy followed in the literature, which requires one year of commit inactivity. In other words, in November, 2017, the proposed model classified 77 projects as unmaintained, despite the existence of recent commits, with less than one year. Figure 4 shows a violin plot with the age of such commits, considering the date of November, 2017. The first, second, and third quartiles are 35, 81, and 195. Interestingly, for two projects the last commit occurred exactly on November, 30, 2018. Despite this fact, the proposed model classified these projects as unmaintained in the same date. If we relied on the standard threshold of one year without commits, these projects would have had to wait one year to receive this classification.

75% of the studied projects are classified as unmaintained despite having recent commits, performed in the last year.

4 LEVEL OF MAINTENANCE ACTIVITY

In this section, we define a metric to express the level of maintenance activity of GitHub projects, i.e., a metric that reveals how often a project is being maintained. The goal is to alert users about projects that although classified as under maintenance by the proposed model are indeed close to an unmaintained status.

4.1 Definition

The proposed machine learning model—generated by Random Forest—consists of multiple decision trees built randomly. Each tree in the ensemble determines a prediction to a target instance and the most voted class is considered as the final output. One possible prediction type of the Random Forest is the matrix of class probabilities. This matrix represents the proportion of the trees’ votes. For example, projects predicted as under maintenance have probability...
Figure 6: Comparing number of commits, issues, pull request, and forks over time of ten projects with maximal LMA (green lines) and ten projects with the lowest LMA in our dataset (red lines). Metrics are collected in intervals of 3 months (x-axis).

Figure 5: Level of maintenance activity (LMA).

Figure 7 shows scatterplots correlating LMA and number of stars, contributors, core contributors, and size (in LOC) of projects classified as under maintenance. To identify core contributors, we use the most common heuristic described in the literature: core contributors are the ones responsible together for at least 80% of the commits in a project [17, 27, 33]. To measure the size of the

4.2 Results

Figure 5 shows the LMA values for each project predicted as under maintenance (2,927 projects, after excluding the projects used to train and test the proposed model, in Section 2). The first, second, and third quartiles are 48, 82, and 97, respectively. In other words, most studied projects are under constant maintenance (median 82). Indeed, 171 projects (5.8%) have a maximal LMA, equal to 100. This list includes well-known and popular projects such as twbs/bootstrap, meteor/meteor, rails/rails, webpack/webpack, and elastic/elasticsearch.

Figure 6 compares a random sample of 10 projects with LMA equals to 100 (actively maintained, therefore) with ten projects with the lowest LMA (0 ≤ LMA ≤ 0.4). These projects are compared using number of commits (Figure 6a), number of issues (Figure 6b), number of pull requests (Figure 6c), and number of forks (Figure 6d), in the last 24 months. Each line represents the project’s metric values. The figures reveal major differences among the projects, regarding these metrics. Usually, the projects with high LMA present high values for the four considered metrics (commits, issues, pull requests, and forks), when compared with projects with low LMA. In other words, the figures suggest that LMA plays an aggregator role of maintenance activity over time.

The equation simply converts the probabilities \( p \) computed by Random Forest to a range from 0 to 100; LMA equals to 0 means the project is very close to an unmaintained classification (since \( p = 0.5 \)); and LMA equals to 100 denotes a project that is actively maintained (since \( p = 1.0 \)).

This equation simply converts the probabilities \( p \) computed by Random Forest to a range from 0 to 100; LMA equals to 0 means the project is very close to an unmaintained project; by contrast, \( p = 1.0 \) means the project is actively maintained. Using these probabilities, we define the level of maintenance activity (LMA) of a GitHub project as follows:

\[
LMA = 2 \times (p - 0.5) \times 100
\]

\( p \) ranging from 0.5 to 1.0. If \( p = 0.5 \), the project is very similar to an unmaintained project; by contrast, \( p = 1.0 \) means the project is actively maintained. Using these probabilities, we define the level of maintenance activity (LMA) of a GitHub project as follows:
Figure 7: Correlating LMA with (a) stars, (b) contributors, (c) core contributors, and (d) size. Spearman’s $\rho$ is also presented.

4.3 Validation with False Negative Projects

In Section 3, we found four projects that although declared by their developers as *unmaintained* are predicted by the proposed machine learning model as *under maintenance*. Therefore, these projects are considered false negatives, when computing recall. Two of such projects has a very low LMA: *nicklockwood/iRate* (LMA = 2) and *gorangajic/react-icons* (LMA = 12). Therefore, although predicted as *under maintenance*, this project is similar to projects classified as *unmaintained*, as suggested by its low LMA. A second project has an intermediate LMA value: *spotify/HubFramework* (LMA = 39.2). Finally, one project *Homebrew/homebrew-php* has a high LMA value (LMA = 99.2). However, this project was migrated to another repository, when facing continuous maintenance. In other words, in this case, the GitHub repository was deprecated, but not the project; therefore, *Homebrew/homebrew-php* is a false, false negative (or a true negative).

5 THREATS TO VALIDITY

The threats to validity of this work are described as follows:

**External Validity:** Our work examines open source projects on GitHub. We recognize that there are popular projects in other platforms (e.g., Bitbucket, SourceForge, and GitLab) or projects that have their own version control installations. Thus, our findings may not generalize to other open source or commercial systems. A second threat relates to the features we have considered. There might be additional features that could be more help improve the the prediction of unmaintained projects, however, given our high prediction performance, we feel confident that our features are effective. Also, some of the features we use may not be available in other projects, however, most of our features are available in most code control repositories/ecosystems. In the future, we intend to investigate additional projects and consider more features.

**Internal Validity:** The first threat relates to the selection of the survey participants. We surveyed the project owner, in the case of repositories owned by individuals, or the developer with the highest number of commits, in the case of repositories owned by organizations. We believe that developers who replied to our survey are the most relevant given their level of activity in the project. The themes of the surveys were defined and organized by the authors of
the paper. As with any human activity, the derived themes may be subject to bias and different researchers might reach different observations. However, to mitigate this threat, a first choice of themes was conducted in parallel by the first two authors of this paper. Also, they attended several meetings during a whole week to improve the initial selected themes. A third threat relates to the parameters used to perform our experiment. We set the number of trees to 100 to train our classifier. To attenuate the bias of our results, we run 5-fold cross-validation and use the average performance for 100 rounds. A forth threat is related to how the accuracy of our machine learning approach was evaluated. We relied on developer replies about their project to evaluate the performance of our machine learning classifier. In some cases, the developer replies (or developers who did not reply) may impact our results. That said, our survey had a response rate of 37.1%, which is very high for a software engineering study, giving us confidence in the reported performance results.

Construct Validity: A first threat relates to the definition of active projects. We consider as active projects those with at least one release in the last year (Section 2). We acknowledge a threat in the definition of the time frame. To mitigate this threat, the first paper’s author inspected each selected project to look for deprecated projects (21 projects declare they are no longer being maintained) and we conduct a survey with 112 developers to confirm our findings. A second threat is related to the projects we studied. Our dataset is composed of the most starred projects (and additional filtering). Although the starred projects may not be representative of all open source projects, we did carefully select such projects to ensure that our study is conducted on real (and not toy) projects.

6 RELATED WORK

Machine Learning. Recently, the application of machine learning in software engineering contexts has gained much attention. Several researchers have used machine learning to accurately predict defects (e.g., [29]), improve issue integration (e.g., [1]), enhance software maintenance (e.g., [13]), and examine developer turnover (e.g., [3]). For example, Gousios et al. [13] investigate the use of machine learning to predict whether a pull request will be merged. They extract 12 features organized into three dimensions: pull request, project, and developer. They conduct their study using six algorithms (Logistic Regression, Naive Bayes, Decision Trees, AdaBoost with Decision Trees, and Random Forest). Bao et al. [3] build a model to predict developer turnover, i.e., whether a developer will leave the company after a period of time. They collect several features based on developers monthly report from two companies. The authors evaluate the performance of five classifiers (KNN, Naive Bayes, SVM, Decision Trees, and Random Forest). In both studies, Random Forest outperforms the results of other algorithms. In another study, Martin et al. [24] train a Bayesian model to support app developers on causal impact analysis of releases. They mine time-series data about Google Play app over a period of 12 months and survey developers of significant releases to check their results. Tian et al. [38] use Random Forest to predict whether an app will be high-rated. They extract 28 factors from eight dimensions, such as app size and library quality. Their findings show that external factors (e.g., number of promotional images) are the most influential factors. Our study also uses machine learning techniques, however, our main goal is to detect projects that are not going to be actively maintained. Moreover, our study extracts project, contributor and owner features that we input to the machine learning models.

Open source projects maintainability. In previous work [9], we survey maintainers of 104 failed open source projects to understand the rationale for such failures. Their findings revealed that projects fail due to reasons associated with project properties (e.g., low maintainability), project team (e.g., lack of time of the main contributor), and to environment reasons (e.g., project was usurped by a competitor). Later, we report results of a survey with 52 developers who recently became core contributors on popular GitHub projects [10]. Our results show the developer’s motivations to assume an important role in FLOSS projects (e.g., to improve the projects because they use them), the project characteristics (e.g., a friendly community), and the obstacles they faced (e.g., lack of time of the project leaders).

Also related is the work by Yamashita et al. [39], which adapts two population migration metrics in the context of open source projects. Their analysis enables the detection of projects that may become obsolete. Khondhu et al. [16] report that more than 10,000 projects are inactive on SourceForge. They use the maintainability index (MI) [28] to compare the maintainability between inactive projects and projects with different statuses (active and dormant). Their results reveal that the majority of inactive systems are abandoned with a similar or increased maintainability, when compared to their initial status. Nonetheless, there are critical concerns on using MI as a predictor of maintainability [4]. Eghbal [12] reports risks and challenges to maintain modern open source projects. She argues that open source plays a key role in the digital infrastructure of our society today. Opposed to physical infrastructure (e.g., bridges and roads), open source projects still lack a reliable and sustainable source of funding. Other recent research on open source has focused on the organization of successful open source projects [27] and on how to attract and retain contributors [8, 20, 30, 34, 40]. Our work enhances the aforementioned work by contributing factors and proposing the use of machine learning to accurately identify projects that are not going to be actively maintained.

7 CONCLUSION

In this paper, we proposed a machine learning model to identify unmaintained GitHub projects. By our definition, this status includes three types of projects: finished projects, deprecated projects, and stalled projects. We validated the proposed model with the principal developers of 127 projects, achieving a precision of 80% (RQ1). Then, we used the model with 112 deprecated projects—as explicitly mentioned in their GitHub page. In this case, we achieved a recall of 96% (RQ3). We also showed that the proposed model can identify unmaintained projects early, without having to wait for an one year of inactivity, as commonly proposed in the literature (RQ3). Finally, we defined a metric, called Level of Maintenance Activity (LMA), to assess the risks of projects become unmaintained. We provided evidence on the applicability of this metric, by investigating its usage in 2,927 projects classified as under maintenance, in our dataset.

Due to its high accuracy (precision= 80% and recall= 96%), the model proposed in this paper can be used by developers to check
the maintenance status of an open source project, before deciding to use it. This information has a key value, since there is a growing concern on the sustainability of modern open source projects [12].

As future work, we intend to implement a tool to provide information about the maintenance status and the level of maintenance activity of open source projects. We also intend to evaluate the accuracy of other machine learning results, besides Random Forest.

The dataset used in this paper is available at: https://goo.gl/K8pZDV [we plan to move it to a permanent repository, in case of acceptance]. Due to privacy concerns, we are omitting the name of the unmaintained projects, as classified by the surveyed developers.

ACKNOWLEDGMENTS

Our research is supported by CAPES, FAPEMIG, and CNPq. We would also like to thank the 112 GitHub developers who kindly answered our survey.

REFERENCES


