University JEAN MONNET

Data Mining and Knowledge Discovery

2022/2023

Issue analytics for project management

Searching for main issues and impacts in budget and time estimation.

Authors:
Anasco Loor CESAR WASHINGTON

April 2023

Contents

1 Introduction

Jira is a project management tool developed by Atlassian, its beginnings were in 2002 with the objective of monitoring the inconveniences that arose in the development of software. Today it offers a series of packages (Jira Service Desk, Jira Core, Jira Software) whose objective is based on optimizing the productive level (productivity) of all the members of the work team, additionally it is possible to use Jira as a management system task management.[1]

This document studies the public Atlassian issue tracker data set. Where we have issue related to real projects being done in the jira workspace. Issues are related to requirements users find for and during software development process.

2 Content

2.1 Problem Understanding

Priority and time estimation is a crucial part of software project management, and is this crucial activity that is usually done by human estimation. Requirements are assigned a timeline base of reporters (software developers) expertise.

Cost and schedules are assigned via this way as well, which in several cases can end in overrun for projects budget and schedule delays. The objective of this study is to predict the priorities assigned to requirements in previous projects, create a decision tree based on the issues attributes such as votes, projects type, company, etc.

2.2 Data Understanding

The current explained data structure was downloaded via the Atlassian api used with python. Given that the dataset has 491 columns fore every row, the principal ones that have the more information about the dataset are the following:

- Summary: Issue description.
- Issue.id: row unique id.

- Issue.Type: issue type referring to bugs or issue suggestions.
- Status: last issue status.
- Project.key: project id
- Project.name: project name. Created: issue creation date.
- Priority: requirement impact.
- Severity: requirement importance.
- Company: Company name
- Resolution: date of issue resolution.
- Resolved: resolved status
- Date response: date of the first response.
- Environment: description to issue nature.
- Story points: issue story points as sub requirements.

2.3 Data Preparation

The following techniques where applied to the data set:

- Null registries or NA rows where deleted.
- Columns where selected until obtaining 14 left columns.
- Numeric values to every column left for plotting and the model implementation.

2.4 Modeling

The prediction target is the priority that the issues are being assigned, taking into account the severity, issue type, resolution, etc. The model is based on building a decision tree via the "rpart()" function. The following decision tree was generated as showed in the Figure 1.

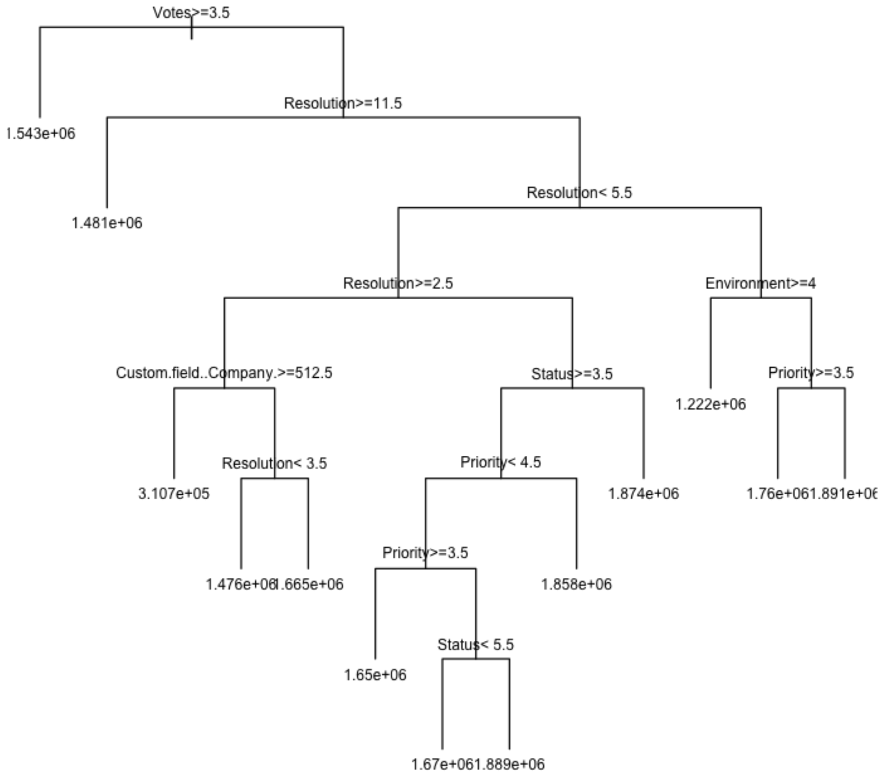
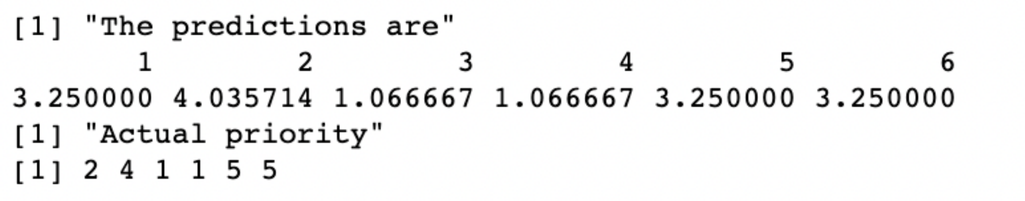


Figure 1: Decision tree generated

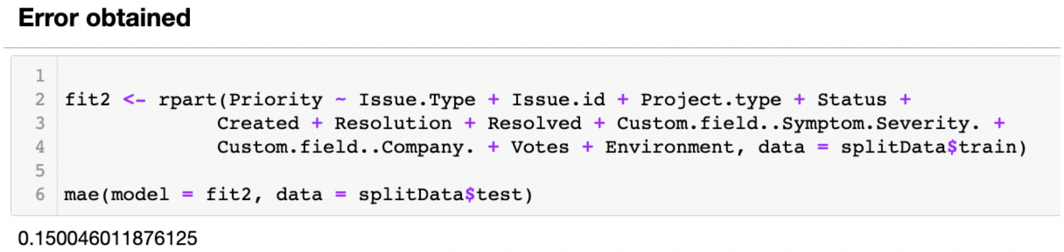
3 Evaluation

Quality of the model was evaluated by the measure of the predictive accuracy. Therefore, a train and a test dataset where generated from the original set of information. The prediction obtained error is as showed in Figure 2 and 3:



```
[1] "The predictions are"
       1        2        3        4        5        6 
3.250000 4.035714 1.066667 1.066667 3.250000 3.250000 
[1] "Actual priority"
[1] 2 4 1 1 5 5
```

Figure 2: Predictions bade by decision tree



Error obtained

```
fit2 <- rpart(Priority ~ Issue.Type + Issue.id + Project.type + Status +
             Created + Resolution + Resolved + Custom.field..Symptom.Severity. +
             Custom.field..Company. + Votes + Environment, data = splitData$train)

mae(model = fit2, data = splitData$test)
```

```
0.150046011876125
```

Figure 3: Error percentage obtained

4 Deployment

In conclusion, using the 49000 project issues obtained from the Atlassian public dataset, there was a model proposed for predicting issue priority for improving the effort management that will be assigned to each issue in every project. There error obtained was 13 percent for the final test set, which may imply and overt=fitting of the model. But that could be improve with the selection of different features for the model creation.

The proposed approach is straight forward and directly makes decision for issue priority based on a large learning dataset. Which outperforms human expertise for effort estimation. Next steps should involve expanding the study for more projects, not only for software development.

References

[1] Patrick Li. *Jira 8 Essentials: Effective issue management and project tracking with the latest Jira features.* Packt Publishing Ltd, 2019.