Forecasting the overall activity in the Apache Zookeper Project

Data Science and Big Data Analytics

Group 6: Andrea Bontempelli, Ander Schiavella , Daniel Pruss, Fabio Sortino

Agenda

- Goals 3
- Results 4
- Model Description 5
- Questions 6

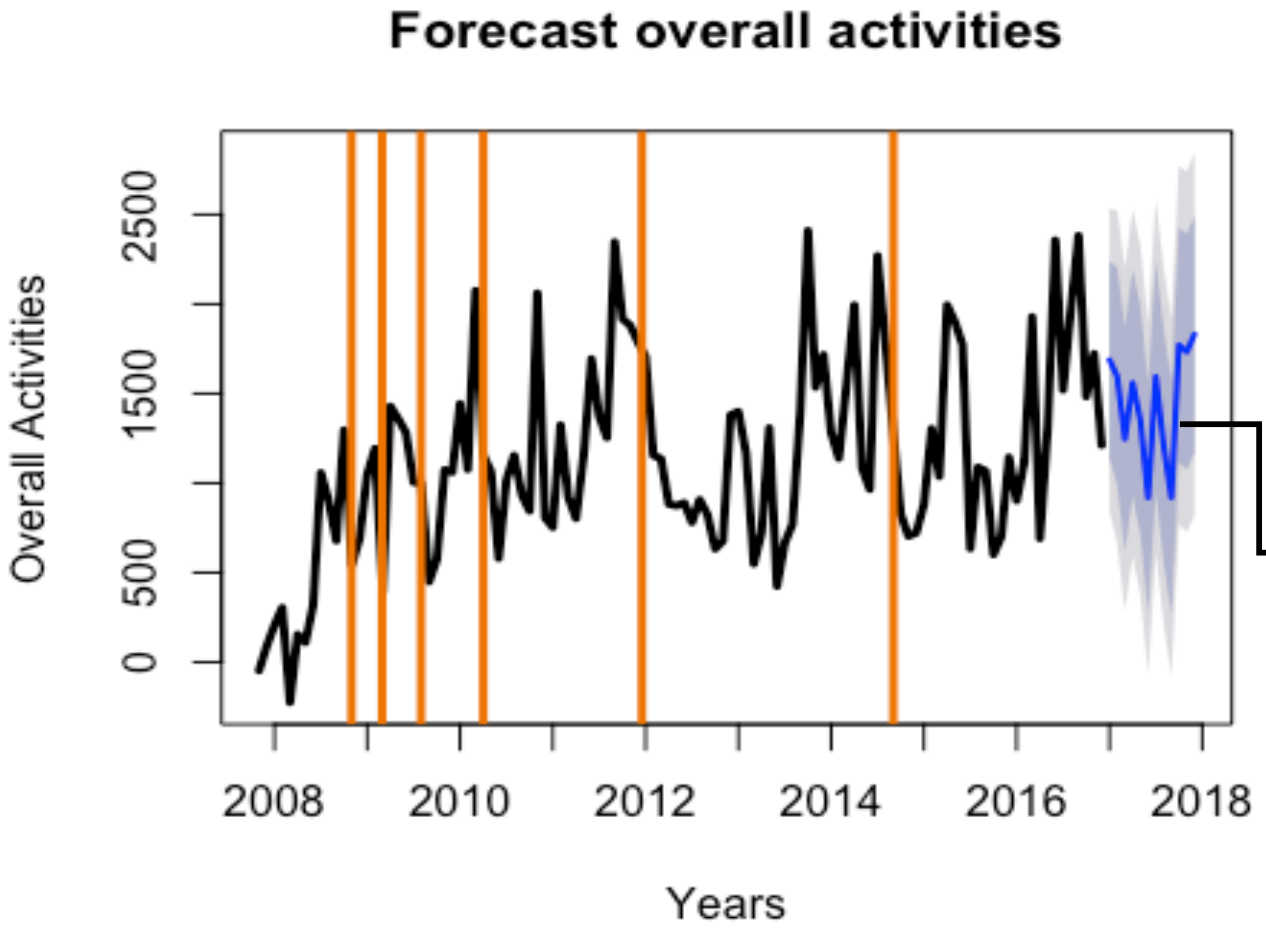
Goals

What we wanted to do	What is the overall activity	How we did that	Our hypothesis
A forecast of the overall activity of the Zookeeper project in 2017	The total amount of operations in the project (commits, messages, events, issues, issue comments, tags)	By analyzing the previous data of the project	A new release is correlated to high overall activity

What is the benefit?

- Measure how active the users are
- Measure when the developers were active
- Forecast the next active period
- Gives information for further decision making

Results



- During the phase-in there were only a few activities
- A new release is related to a peak of activity

The blue line shows the forecast of 2017

Result: In 2018 is a potential new release

Model Description

The tool for the analytics was:


[1]

1. Collect all the dates of the activities
2. Count all the activities for every month
3. Apply the timeseries from November 2007 till December 2016
4. Decompose the timeseries for showing the seasonality and trend
5. Forecast the activity for 2017 with the ARIMA model

Suggestion: This model fits the data but it can be improved e.g. by

- Apply different weights to the parameters (A message has not the same impact as an issue or commit)

Questions

[2]

Sources

[1] R Logo: https://upload.wikimedia.org/wikipedia/commons/thumb/1/1b/R_logo.svg/2000px-R_logo.svg.png Date of access: (26.01.2017)
[2] Questionsign: http://social.eyeforpharma.com/sites/default/files/big20question20marks.png Date of access: (26.01.2017)