THE UNIVERSITY OF AUCKLAND
BUSINESS SCHOOL

INFOSYS 750- Final Project

Evolution of Open Source Software Projects

Open source projects are software artefacts, which are developed and maintained by software developers and volunteers. Generally, the source code of these projects are available online and end-users can freely use them under the constraints defined on project license type. GitHub is the largest open source repository with millions of projects. You have been provided with a panel dataset of projects hosted on GitHub. It contains two years (8 quarters) of data for each project. You can investigate how projects change over time by analysing these datasets.

Academic Report:

In recent years, software companies have started collaborating with and/or sponsoring open source software projects. Before forging any partnership, software companies try to understand OSS projects' progress and how they change over time.

Mr. Richard Thompson, a software architect at Mega Software who often collaborates with or sponsors OSS projects, needs your help in understanding evolution in some of the OSS projects. Mr. Thompson, would like any insights about these projects (see the dataset).

By applying longitudinal data models, you need to provide an insight about the projects evolution. Clearly define your research questions and test your hypotheses using this dataset. Students are strongly recommended to refer to research papers to follow the writing style and structure.

The final report should be have following structure-

1. Introduction (**5 points**)
2. Research questions and Hypotheses (**10 points**)
3. Definition of main variables, Visual Exploration (**5 points**)
4. Data cleaning and preparation (any modification, transformation and sampling techniques that you have applied) (**5 points**)
5. Methodology (**10 points**)
6. Results an discussions (**12.5 points**)
7. References (**2.5 points**)

Every figure or table in the report should be explained. Students should not just dump output from R script file in the report.

Everyone needs to submit a project report. All group members are expected to contribute equally. For each group, R script file would be same for every group member, however, report should be written individually.

Length of the report: 2000 words (minimum), excluding references and script file.

If there are any problems in your group, please notify the instructor immediately. You can google to get familiar with the Git terminology to understand the dataset better. Bring your questions to lectures and tutorials.

Dataset details:

Variable	Definition
PrjID	A unique id number for each project
Period	Year and month when data was collected. (data is collected quarterly)
Time	A sequence for time of observations
SatrtDate	Beginning of data collection for this period
EndDate	End of data collection for this period
Forks	Total number of times a project is forked
Members	Total number of members
Commits	Total number of coding activities (commits)
Issues	Total number of problem/bugs raised or requests for new features
Watchers	Total number of people interested in project (number of users who have this project on their watchlist)
PullReq	Total number of code changes request for review.
CommitCmnt	Total number of discussion/comments on commits
PullReqCmnt	Total number of discussion/comments on pull requests.
PR Issue Cmnt	Total number of comments on issues related to Pull Requests
IssueCmnt	Total number of discussion/comments on issues.
Committers	Total number of users who committed on this project.
MemCommitters	Total number of project members who committed on this project.
PRClosedCnt	Total number of closed pull requests (merged/rejected)
IssueClosedCnt	Total number of closed issues
PRClosedTime	Average time spent on closing pull requests (minutes).
IssueClosedTime	Average time spent on closing issues (minutes).
Health	The health indicator of a project scaled (0-100). 100 means best.
License	License of the project
ContribFile	The joining script and contribution guideline
OwnerFollower	Number of followers of the project Owner
AvgFollower	Average number of followers of project team members
OwnerType	Type of the owner of the project (organization, user, etc.)