Summary of "Signals Matter: Understanding Popularity and Impact of Users"

The paper is an attempt to analyze the impact features such as the possession of badges, reputation scores, the time the user has been using the site etc have on the overall popularity and 'impact score' of the user in the community.

Stackoverflow, in summary, has following rewarding system:
- A user must first be active and prove himself in 3 most important activities: **asking, answering and editing** - none of which require any reputation at all.
- The primary way to gain **reputation** is by posting good questions and useful answers: points are awarded when questions are upvoted, answers are accepted, suggested edits are approved etc.
- Badges can be classified into questions badges, answer badges, moderation badges etc
- It is once that the user is comfortable in the above 3 activities that he gains special privileges: the activation of which depends on the badges and reputation and may allow the user to perform tasks like : create a new tag, delete questions, flag a post, create a topic wiki etc.

Based on a survey conducted, the paper considers the following **criteria** as a measure of the competency of users:
Popularity: total number of distinct views on their profile page.
Impact Score: sum total views on questions, and answers with non-zero scores that have either been accepted, or are in the top 3 answers etc.
RIghtfully, the paper neglects 'inactive/throwaway accounts' from the used dataset.
The paper tries to predict the popularity and impact score of the user using _Gradient Tree Boosting Regression model_ on 3 sets of features:
CM: Age; Number of questions,answers,upvotes,downvotes
RM: features of CM + reputation_score
BM: features of CM + reputation_score + number of badges of diff types.

The model is validated using the R-squared metric.

Results

For popularity

Model	R-squared score	Most useful feature
CM	0.911	Age on the site
RM	0.939	Reputation score
BM	0.957	Age on the site

For impact score

Model	R-squared score	Most useful feature
CM	0.685	Age on the site
RM	0.767	Reputation score
BM	0.858	Number of Necromancer Badges

Conclusions from the results
- RM performs better than CM for both criteria as CM depends **_solely on features like 'age of site' (misleading as a user registered a long time ago may not have any expertise at all) , number of questions asked (a user can keep asking low-moderate quality questions despite not being helpful to the community)._** So, RM which has he additional dependence on reputation score (useful feature giving importance to accepted answers etc which, at least in theory, seem indicative of user's competence in a field).
- Reputation score is a good indicator of popularity (0.313).
- BM performs better than RM and CM for both criteria. This is as **badges which reward quality as opposed to 'reputation score' which can be increased by quantitative accomplishments like providing low-medium quality answers(with 1-2 upvotes) for a large number of questions.** Eg: **User A :** provides answers to 15 low level questions with 2 upvotes each. **User B:** answers a much difficult question and receives 29 upvotes. By number of upvotes, user A would have a better rep score even though the odds are that User B is more knowledgeable in his/her domain.
- Not all badges are good predictors. This is as **some badges are representative of effort rather than skill**. Eg: _badges awarded to beginners for motivation, badges awarded for flagging duplicates, correcting grammatical errors, improving appearance of a question, moderating etc._ **Answer badges were found to be much better**

features than 'Question badges' and other badges as good questions can be asked by even intermediately skilled users or even beginners, but providing good answers to questions (while in competition with other skilled users) is a better indicator of area-level knowledge.
- Most highly ranked users had joined the site early (representative of the fact that it is likely to take time to build impactful knowledge in CS).

Possible reasons for inconsistencies in the results
- Many people use the site **without signing in**. Often users do not upvote an answer even if they found it useful or do not have the option to do so (since they are not signed in).
- **Views are tracked per IP address per rolling time window**. If one comes back to the question in, say, 15 minutes we will see the counter go up by 1 again.
- The rewarding system is not perfect : If User A provides a quick correct answer to a question like https://stackoverflow.com/questions/11828270/how-do-i-exit-the-vim-editor with large number of views , he may have a very high impact score even though he may not be a regular impactful contributor.
- **The bounty system of stackoverflow** where a highly skilled user with a theoretically high reputation has an actual low reputation score as he had set bounties on questions to encourage better discussions on that particular questions and hence, loses his own reputation score.

Future work/Improvements
- Just like the scores of different badges were studied, I would like to divide **reputation score** also into 3-4 categories like 'points gained due to answering', 'points due to questions', points due to edits etc.
- For StackOverflow, **time of joining the site** was found to be a crucial factor. I would like to repeat the same exercise with non-theoretical communities like **Science fiction StackExchange (https://scifi.stackexchange.com/)** . Here**, I expect 'time of joining' etc to be a non-essential factor** as even new users may be excellent contributors with questions related to a particular tag (here, possibly a TV series recently released).
- I would like to study the **correlation in the badges/reputation score obtained by the user** while participating in discussions related to different tags. A user with high contribution wrt to the tag 'pandas' is likely to have a high contribution wrt tag 'dataframe'.