Stack Overflow Analysis

Gian Cercena, Prithaj Nath
Social Computing Systems

November 21, 2024

Abstract

We analyze the Stack Overflow userbase, defining what it means to be a super-user, and how they are set apart from a regular user. Succinctly, super-users are a very small subset of users that are important to the survival of an online social network (OSN) due to the large amount of content they produce, and how they become core users to a platform, both of which drive activity. Logistic regression and XGBoost classifier models were created to predict whether or not someone will become a super-user at certain intervals since their account creation. Findings show that it is possible to predict whether or not someone will become a super-user, something that increases with more user history. Accuracy rates start at 62.5% seven days after registration, and increase to 85% two years out using easily attainable data. Additionally, we look at the rate at which super-users churn, or depart from an OSN. With the inclusion of more data and development of more complex models, this level of accuracy can be increased, opening a new path for online social network to grow their dedicated userbase.

1 Introduction

Stack Overflow is a widely used resource for many [1]. From beginner programmers asking questions to advanced users delving into the minutiae of their fields, Stack Overflow hosts a wide variety of programming-related content. It is most often seen by developers when looking up a question due to the nature of the website; a user will ask a question, and others will pose responses to it, with each response being voted on by the community and one eventually being selected by the original poster as the one that has solved their problem.

Often, when looking at answers to a question, it will be by a user that has accumulated a high reputation (the amount of points gained by other users voting on their responses) as well as many badges, which tend to indicate outstanding posts or comments. These users are individuals who spend an outsized amount of time on Stack Overflow when compared with the average user, as the typical visitor to the website likely doesn't even have an account, as all Q&As are available for free. Those who post frequently, and are consistently active in an online space are known as super-users.

Super-users are noted to be active minorities [2], and crucial actors, opinion leaders, or active users [3]. Given the wide variety of online social networks (OSN), including their number of active users, content reach, and user participation style, what defines a super-user tends to be unique to a given OSN. That being said, super-users can most generally be defined as those who regularly participate in an OSN, and to a higher degree than the median user, whether it be making posts, writing comments, or even moderating groups or forums.

In Stack Overflow, a super-user can be defined as someone with a high in-degree, a high reputation, and the acquisition of rare badges [4]. For our analysis, we will focus on users in the top percentiles of these metrics. These super-users are often seen as critical, whether it be to them being well-known across a community, someone who drives traffic to the OSN, or are helpful and answer many questions [5], such as our case with Stack Overflow.

We intend to define the concept of a super-user within the Stack Overflow community, identify key behavioral trends that distinguish super-users from regular users, and develop predictive models that can accurately guess whether or not a given user will become—or is already—a super-user.

1.1 Purpose

Super-users are users who contribute disproportionately to the site despite being a small minority. However, the number of questions answered by super-users has been declining over the years, shown in Figure 1. This research aims to investigate the trajectory and behaviors of users that lead them to become Stack Overflow super-users and their implications for user retention across all social computing networks. Understanding what causes people to become attached to a community and website such as Stack Overflow could be a valuable resource in understanding user retention, and the factors contributing to participation across all social networks.

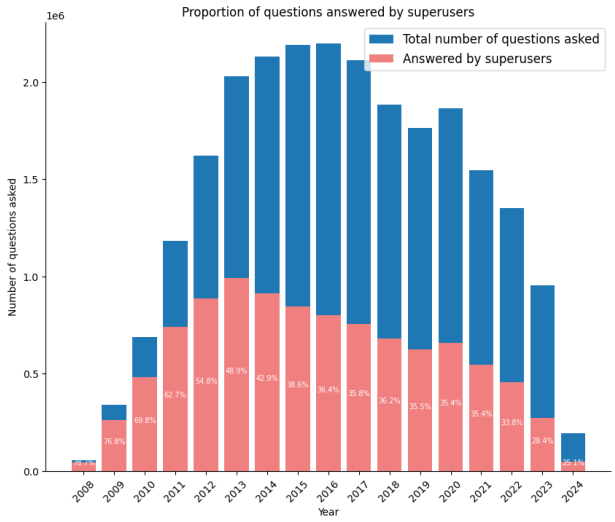

Figure 1: *The proportion of questions answered by super-users on Stack Overflow has been declining over the years.*

2 Previous Works

Understanding how users interact in OSNs has been a key point of research, specifically targeting active users, often called super-users. Below, we review relevant literature in this subject, and shed light on what has been done previously and how our work builds off of it.

We first start by defining super-user. By noting previous works, super users are typically semantically described as "leaders", "figureheads", or "mentors" [2, 3]. They are generally very active, and have some vested interest in a community, whether it be through interest as a hobby, or through professional engagement. These individuals often take on significant roles in these communities, such as moderator or admin roles, and disproportionally contribute to content and discussion on these sites, something commonly noted as the Pareto principle [6].

Other works have looked at the concept of churn, or the process of users ceasing their participation on an OSN. This issue is extremely pertinent for these OSNs as part of their usefulness relies on others participating and user-genearated content. Wilson et al. [7] looked at the interactions that users may have across online social networks, showing that there are certain patterns that are correlated with churn.

One potential reason for super-user churn is that the areas in which they have expertise in may no longer be popular on websites. Barua et al. [8] analyzed topic trends, showing swings of popularity in subjects, which could lead to decreased activity in certain topics. With less attention to a topic, super-users might churn as they may feel like their expertise is either not needed, or as positively welcomed. Akin to that, Fu et al. [9] found that users who focus on specific areas or concentrations tend to have higher reputations, suggesting that these individual super-users are densely connected around a limited set of topics. Over time, if fewer questions appear in an area that a super-user frequents, they might spend less time on the platform, which would follow with our reasoning.

The falloff of interaction with websites due to this was also seen by Asaduzzaman et al. [10]. The authors explored the increasing amount of unanswered questions (something we have also historically noted when using Stack Overflow). Their findings line up with a previous analysis of our data, which showed that many super-users no longer tend to answer new questions on the site but answer old questions instead. We hypothesize that these users are just updating obsolete responses to old questions and are only interested in answering questions from their domain of expertise, which we suspect is no longer popular on the site.

Kabir et al. [11] examine the impact of ChatGPT on Stack Overflow, finding that responses from ChatGPT are preferred, even if they include misinformation, as the well-done presentation of the information by ChatGPT gets the users to overlook these mistakes. This could be one of the many contributing factors to super-user churn in the platform since they could feel like the effort they put into more accurate posts could go underappreciated.

Lastly on churn, we see Adaji and Vassileva [4], who developed models to predict the churn of these experts or super-users, found that they were able to accurately predict churn based off of the current activity a user had compared to previous points in their history. By noting various attributes such as time between posts, reputation scores for recent answers, and the number of badges received, high levels of predictive accuracy were achieved.

Zhang et al. [12] looked at obsolete answers, or in other words, answers that have been provided such a long time ago that they are no longer helpful, either due to software updates or obsolete technology. This can be remedied, though, as users can edit previous answers by others to keep them up to date. This is useful for very often-frequented posts, such as those that may appear within some of the first results on a search engine. This also brings us back to our initial hypothesis that states super-users, over time, tend to answer fewer new questions and just update old answers to old questions due to a variety of reasons, which we wish to explore in our paper.

Bachschi et al. [13] hit on one of the critical points of our project, noting when users stop asking questions, and start answering them instead. This is a key insight needed for our analysis of user evolution, especially novice users, who we wish to understand the progression of to super-users. Their findings show a detailed breakdown of users and what causes or prevents them from answering questions or making posts. There is also the potential to expand upon some of the limitations they stated, such as taking into account the effect of having a highly up-voted post on future postings.

3 Threats to Validity

There are a few threats to validity that are important to note. First is that our models, we include columns denoting the presence of optionally filled profile fields, which are presented in the form of a boolean. It is important to note that these values are determined as of the export of this dataset, and do not have a history associated with them, meaning we have to assume their presence or absence across the entire dataset based on the date the data was exported by Stack Exchange.

Another is the issue of class imbalance. When talking about super-users, or any extremely small minority group, there is the issue of class imbalance. Due to how we later define super-users, they are necessarily an extremely small part of the entire userbase. This is can be an issue for model accuracy reporting, and because of so, we utilize balanced accuracy, which is the average accuracy across all groups, equalizing that factor. As for other data, such as posts or comments, since super-users are partially defined as those who are extremely active, and well-known in the community, their posts and comments far exceed normal users, meaning that there is less of an imbalance to see there.

Finally, for the super-user vs. regular user prediction portion of this project, it assumes that each record of the activities later described are of equal value. For example, it does not differentiate between a well-written out post that has taken a lot of effort and a post that was accidentally sent and contains a single letter. This level of effort that users can put into activities can be used as another, likely highly predictive measurement, but is something we do not take into account for the sake of this project.

4 Ethics

The Stack Overflow data dump is released publicly [14], and accounts do not have the possibility to be privated. Therefore, while IDs can be used to track back to a given user, they will be anonymized for any reporting in this paper or subsequent work. Due to this, and the fact that the data is released by Stack Overflow itself through its CC BY-SA 4.0 license, it is safe to say that this work falls within ethical guidelines. This is especially so since we work with the data as aggregates, not singling out individual users, increasing their anonymity.

Since users are anonymized, and worked as aggregates, this project should have no adverse effects on the users. Additionally, later mentioned sampling techniques are used to avoid bias throughout this project.

This work is intended to analyze and seek insights from user behaviors and patterns while using Stack Overflow. Actions can be taken given the results we have found, but it still is imperative that they are not taken and used maliciously, or to target a group of users in a harmful way.

5 Methodology

This section outlines the way in which we acquired the data for this project, including any preprocessing needed to use it within our analysis.

5.1 Data Collection

The data we utilized is the Stack Exchange Data Dump [14]. This data, hosted publicly under Stack Exchange's CC BY-SA 4.0 license, is a trove of information relating to all Stack Exchange boards. For the purpose of this project, we look solely at the data pertaining to Stack Overflow.

The data is divided into several large XML files such as `Badges.xml`, `Comments.xml`, `Posts.xml`, and `Users.xml`. Due to the large size of the data (certain files were larger than 100 GB) we took steps to create a preprocessing pipeline to convert it to a more usable format for this project.

5.2 Data Preprocessing

The original data, being in an XML format was not ideal for our process of analysis. XML, being a format where an arbitrary number of key value pairs

can placed into any entry, could not be easily converted to a tabular format, especially due to the fact that we did not have access to machines that could fit all of the data into memory. Each file had a first pass where all entries were looked over, and all possible value names were recorded. These value names then consisted of the columns in a CSV that we then converted the XML to, in batches. The CSV file format was chosen due to its uncompressed nature since we had to continuously append to it due to memory constraints.

It is important to note that due to the abundance of written and formatting text within various posts and comments throughout our data, we wanted to ensure that the value that separated our data was never accidentally escaped. This lead us to choose the ASCII character 23 (End of transmission block), a non-printing and legacy character as our separator. This distinguishes our "CSV" files from typically comma separated files. This character did not show up once in our data, giving us confidence it would work as a separating character.

After converting the original XML files to CSVs, some of the data was transferred to BigQuery using a Rust tool called `dbcrossbar` [15] for faster processing and quick analysis via a SQL interface.

5.3 Sampling Techniques

Samples for initially analysis were done via purely random sampling. When moving onto the creation of models and detecting super-users as seen in Section 6.2, since the data is large, we employed certain sampling techniques to reduce bias throughout the model creation. Specifically, what we used was stratified sampling done where users were grouped into bins based on their creation year and month (e.g. all accounts created in January 2010 were in one bin).

The sampling technique was chosen to ensure that we get a properly unbiased set of new users. This is due to the fact that, for example, many accounts were registered during the initial COVID-19 shutdown. Taking a normal random sample would favorably take users from that part of the userbase. By either setting a percentage or a max value to be taken from each of these bins, we can ensure a more generalized group of accounts.

Additionally, since the User IDs are public, and can be traced back to Stack Overflow, any IDs reported in this project will be anonymized.

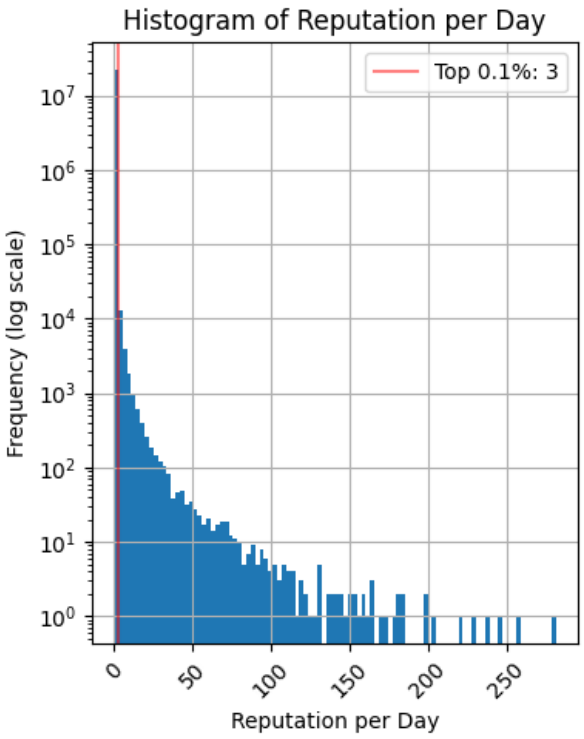

Figure 2: *Average reputation per day for each user with an account age older than 30 days.*

6 Analysis

6.1 What is a Stack Overflow Super-User?

The discussion of super-users presumes a category of non-super-users though, which then defining the separation between the two become a problem. At what point does someone "become" a super-user?

As seen in Figure 2 we can see an extremely strong, heavy right-tailed skew for reputation per day per user. To belong in the top 0.1% of users, one must have, on average, a gain of 3 reputation per day. One person upvoting either a question, post, or answer gives a user +10 reputation, meaning that these top users receive at least one upvote every $\sim$3.3 days. This metric gives us 22,076 super-users, and 22,053,173 non-super-users across all of our Stack Overflow data.

Since Stack Exchange doesn't publish active user counts frequently, by taking advantage of their Data Explorer, we can view roughly how many active registered users there are on Stack Overflow [16]. Pre-2023, values of registered users that made at least one per month were above 100,000 and reached up to 180,000. Post-2023, we see a dramatic decrease in users that have made posts, falling to values around 70,000, and as low as 45,000. This decreased traffic

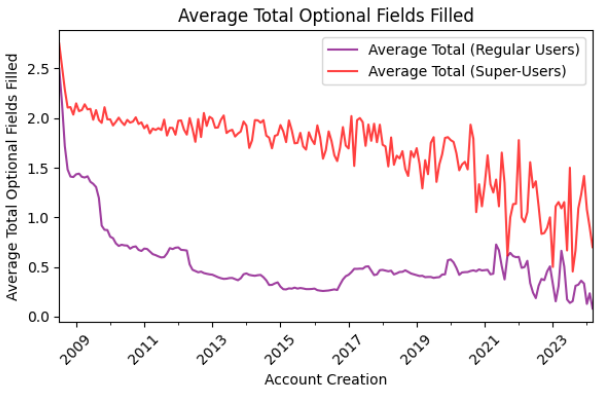

Figure 3: *The average amount of optional fields filled (About, Location, Website) for each user group by their account creation date.*

has been noted to be due to two primary reasons, LLM chatbots, and a modification to how Stack Overflow monitors their usage. While LLMs seem to have a large correlation to this decrease in usage [17], Stack Exchange asserts that this issue is not that damaging, and believe it is inaccurate [18]. This information is important to note as the following calculations are based on rough estimates and should not be considered precise but rather a general representation of the underlying trends.

With the above information, assuming an average of 100,000 unique registered users per month [16] and a monthly visitor count of 200 million [19], we can assume that on average 0.05% (100k/200 million = 0.0005) of people who view a post has the ability to upvote or downvote it. One can presume that roughly for every 1 upvote, there is at least 2,000 ($1/0.0005 = 2000$) others who have viewed it. This calculation is likely a slight underestimate too, considering that not every registered Stack Overflow user will vote on a post, and they are more likely to see a post than a non-registered individual.

What this means is that a super-user, as defined by our logic, will be an account that receives on average one upvote (or $\sim$10 reputation), and around 2,000 views on their posts every 3 days. These are the metrics that put a user into the top 0.1% of the Stack Overflow userbase.

6.2 Super vs. Regular

There are a few distinguishing characteristics between a typical super-user and regular user. One key metric to look at are the optional fields on a user's profile. Each user is, by choice, able to fill out 3 fields to flesh out their profile: About, Location, and Website. The About field is a short summary where a user typically describes themselves, Location is another field that allows the user to show where they are from, and Website is a place to put a URL, usually to a personal webpage or a LinkedIn account.

Since these fields are initially empty, and left up to the user to fill in, it can be presumed that a more dedicated Stack Overflow user (e.g. a super-user) would have them filled in. This can be seen within Figure 3 where generally, for accounts created before 2018, super-users tended to have at least two of the fields filled, where regular users tended to have less than 0.5 filled on average.

A large spike in optional fields filled can be seen towards the beginning, likely due to the fact that Stack Overflow was initially opened up to a smaller group of people that followed one of the creator's blog [20]. These people, being a smaller community (before Stack Overflow became as big as it has) might have felt more comfortable sharing this information, or due to a different registration procedure, they could've been more heavily prompted to give input on these fields. But seeing this overall discrepancy towards what could be considered "additional optional effort" shows that those who participate more on Stack Overflow generally like to keep a more complete profile.

The downward trend for super-users as of the past few years (see Figure 1) is likely due to the fact that the accounts are younger. The optional fields are not static upon account creation, and as someone uses the site more, there will be more chances for them to fill in the information.

As accounts mature, we see a roughly similar growth of Answers, Comments, and Questions, as seen in Figure 4. This would follow given as each are similar in nature—a text post made by the user. While they vary in counts, their pattern of increase over time can be seen sharply rising during the first few weeks, and then slowing to a sublinear rate. This is anecdotally due to users who create accounts to have one specific issue or question solved, which then afterwards they promptly stop interacting with the website.

The mean and standard deviation for the Comment count is the highest compared to all other activities. This is likely due to the fact that Comments tend to be shorter, and require less decorum and effort to post, as they tend to be follow ups to specific Question or Answer posts. A clearer look at the large discrepancy of the comment standard deviation can be seen in Figure 7. Since, a super-user is defined as someone in the top 0.1% of users, this would put them at 3.719 standard deviations (SD), meaning the line would be exaggerated far more. For example, at

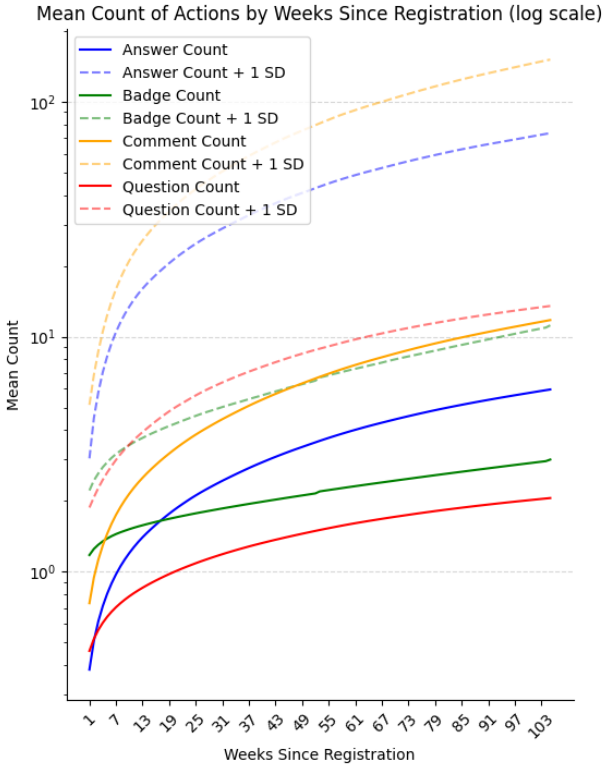

Figure 4: *Mean and mean + 1 SD of actions taken by users for 2 years after their registration. Shown with the y-axis log scaled.*

the end of year two, one could expect a super-user to have approximately 530 comments, where a typical user would have just over 10.

Badges are seen following a different rate when compared with the other activities. This is because users cannot "directly" receive Badges, but instead have to qualify for them in certain ways. This could be in the amount of Answers they have given, having a post reach over a certain milestone of upvotes, or being a registered user on Stack Overflow for a certain amount of time. This last one can be clearly seen if looking closely at the graph at the one or two year mark, where the Badge Count line (green) in Figure 4 increases sharply by a small amount.

6.3 Predicting Super-Users

Being able to quick assess whether or not someone will become a super-user would be a great asset to have. Comprehending the underlying patterns that give rise to those who use and contribute to your platform the most can lead to a mutual benefit. The platform can notice, and either lend resources, additional information, or help to these users, and in turn, these users can help give back to the platform, helping it mature and grow. In order to quickly find these users, we utilize user activity data in order to predict whether or not someone will become a super user.

We have two datasets, one that includes all activity information, which contains the amount of answers, badges, comments, and questions a given user has posted since time t, where t is the number of days since their account registration, and a full dataset that includes that activity, as well as boolean values for whether or not the 3 optional profile fields (About, Location, and Website) are filled out. Both datasets have the full inclusion of super-users, and a stratified 10% sampling of regular users, based on the creation year and month of their account. This was done in order to get a more representative sample of users across the data. This information was gathered for numerous values of t, ranging from $t = 7$ to $t = 735$ at intervals of 7, or one week to just over two years. Both a logistic regression model and a XGBoost classifier [21] were created at each value of t for each dataset and evaluated on whether or not someone was defined as a super-user. The results of this model can be seen in Figure 5.

XGBoost performed considerably better than logistic regression. Being a gradient boosted library often touted for its flexible and efficient performance on tabular data, it does at points upwards of 5% better on the balanced accuracy metric. Balanced accuracy was chosen due to class imbalance, since super-users, as previously defined, are only 0.1% of the Stack Overflow userbase. If normal accuracy was used, and the models predicted non-super-users each time, accuracy would be about 99.9%. With accuracies Starting at around 60% from 7 days since registration, and reaching up to 85% 2 years in, we can certainly say there is some predictive power in these collected variables.

The full models, which included the optional values, did do quite better than the models that did not have access to that data increasing accuracies by up to 2.5% for XGBoost and 1.4% for logistic regression during lower values of t as seen in Figure 8. The amount of extra information gained from these variables decreased as t increased, as the activity counts gave larger insights into the user. This can be seen by looking at the coefficients in Figure 9. Almost all coefficients start off as positive. The sole negative coefficient is for Comments. This can be due circumstances where an individual makes an account solely to answer one question, during which, they post many comments in order to figure out what is going on. Regardless, each coefficient is positive at one point, meaning that each contributes towards a prediction being for a super-user. The values change drastically towards the first few weeks, which after-

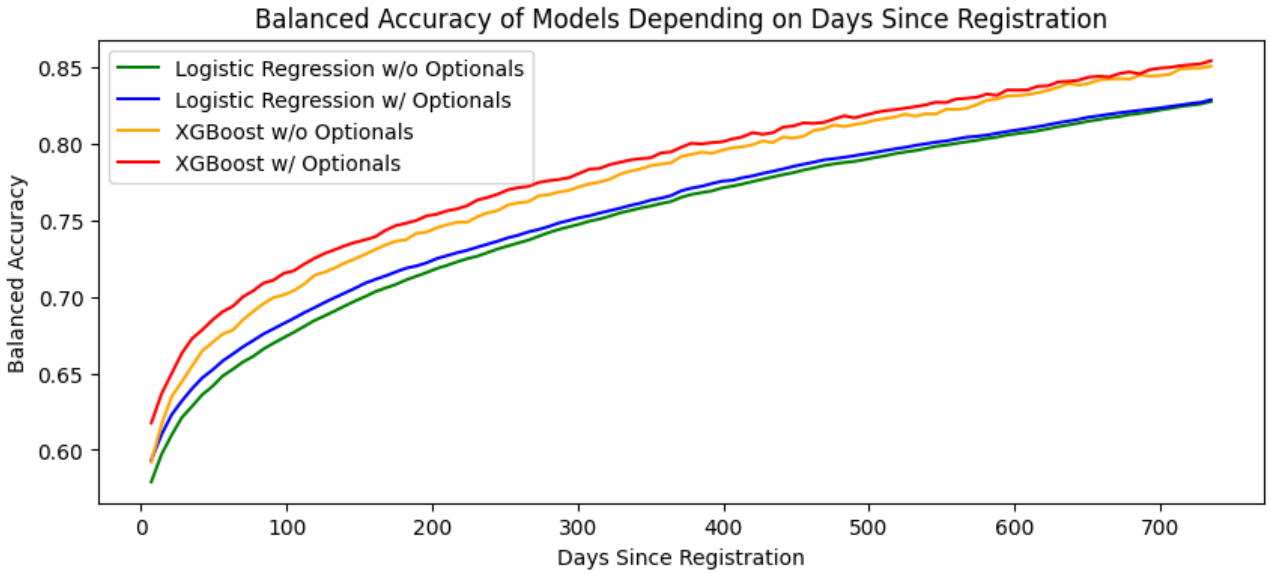

Figure 5: *Logistic and XGBoost model balanced accuracies across values of t for predicting super-users.*

wards tend to stabilize.

All 3 of the optional fields sit for most of the graph as the strongest coefficients, meaning that these values contribute a large amount to each prediction. Specifically it is seen that Location is the largest coefficient by a factor of 2 (see Figure 9. This could follow given that it is likely that someone would only include their location on a Stack Overflow account (which can be easily kept anonymized) if they wanted to have some level of professional connection to it, purposefully "doxxing" themselves (revealing personal private information online). By associating a Stack Overflow account with your Location, and therefore likely your real name, it is conceivable that an individual would either be more serious about contributing to Stack Overflow, or keeping good appearances if others were to stumble onto their account. This also goes for the About and Website sections, though, at least for the latter, having a personal website (whether it is listed or not) is probably less frequent across all users.

6.4 Analyzing Super-User Churn

Super-users contribute disproportionately to Stack Overflow. Hence, churn from these users can significantly affect the longevity and sustainability of the site. We investigate the extent to which super-users are in the process of churning or have already churned. Churn could be defined in many ways, such as decline in edits or site visits, but for our analysis we defined churned as decline in answers contributed to questions. To be more precise, we calculated the rolling six-month average of the number of answers posted by super-users and saw that it follows an initial upward trend until it reaches a peak and then declines over time. We modeled this post-peak decline in activity using exponential decay

$$y = a \cdot e^{-kt}$$

where y is the activity levels (rolling six-month average of number of answers posted) since peak, a is the peak activity level, k is the decay rate and t is the time since peak activity level.

We define the decay rate, k, as the *rate of churn*. We used the inter-quartile range (where q_k denotes quartile k) to categorize super-users into the following categories based on their churn rates.

Churn rate	Churn category
$k < 0$	no churn
$0 < k < q_1 - 1.5 * \text{IQR}$	slow churn
$q_1 - 1.5 * \text{IQR} <= k < q_3 + 1.5 * \text{IQR}$	fast churn
$k >= q_3 + 1.5 * \text{IQR}$	churned

Based on this definition, we found that around 5% of super-users have already churned, and around 95% of them are in the fast churn category. The number of super-users who have not churned makes up less than 1% of the total number of super-users. These findings show that Stack Overflow's most influential contributors are at a very high risk of leaving the site, which can severely affect the longevity and sustainability of the site. More resources need to be allocated for outreach programs to target super-users in the fast churn category.

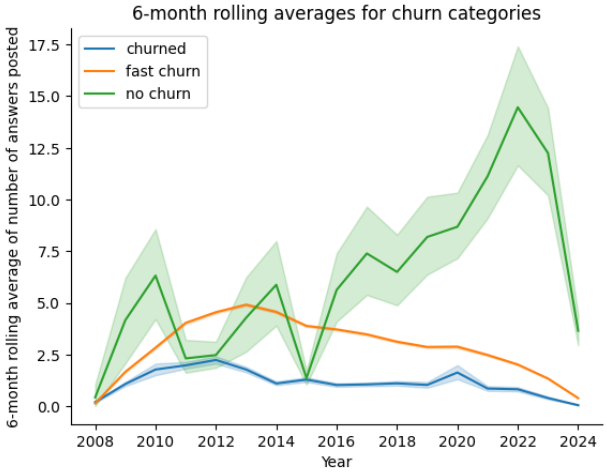

Figure 6: *Non-churning super-users show very distinct trajectories from churned and churning super-users.*

7 Discussion

This analysis shows that it is feasible to distinguish super-users from regular users on Stack Overflow. Being able to develop predictive models in forecasting a user's evolution into a super-user is useful as a tool that can be used by OSNs in order to drive more activity and keep users interacting on their site. Our findings, through high predictive model accuracies, show that these super-users stand out from regular uses in notable ways, those that are easy to pick up in an automated fashion.

The results we have obtained align themselves closely with previous works that looked that the role of super-users in OSNs [2, 3]. Specifically, the contribution of activity and having optional profile fields filled out being predictors of super-users back up other works that had similar findings [4]. This consistency between works shows that there is indeed something that correlates super-users to these metrics.

Finding that these optional profile fields such as "Location" and "About" are strong predictors of super-user behavior shows that there still are other avenues to build a stronger community in Stack Overflow. By using Location, for example, Stack Overflow can reach out to users for local events that may be happening, either sponsored, or attended by Stack Overflow in some capacity. Since those who include those fields tend to be super-users, which are more likely to be an account publically connected to a specific professional in the field, this could be a good targeting technique.

Additionally, when looking at the analysis of super-user churn, we see several insights into the sustainability of super-users on the platform. The exponential decay model is used to great extent in discovering insights related to the super-user churn. Decisions from Stack Overflow can be made in order to target users within the "fast churn" category, either reaching out to them, or recognizing and assisting with common issues.

We can see super-users churning already when we look at Figure 1, which raises important issues for the longevity of the platform. With the wide-spread adoption of LLMs, and other competitive resources, Stack Overflow might need to change its strategy as it continues into the near future. Reaching out to super-users would appear to be a strong choice to make given the societal want for answers by real professional humans, not LLMs, something we have shown Stack Overflow to—though shrinking—have a large quantity of.

7.1 Future Work

While our predictive models show good results, the access and time we had for this project limited their complexity, and therefore the upper limit on how accurate they could be. With access to more resources and data, Stack Overflow can take similar steps to build models, and expand upon them, reaching higher prediction rates.

Utilizing real-time, or more fine grain data can help improve the performance of these models. Additionally, incorporating sentiment analysis on the text of the posts or comments, noting their relative positivity, or other factors such as sophistication or length, could also contribute. Additionally, expending this project to other forums within the Stack Exchange ecosystem besides Stack Overflow could prove to be useful, especially when comparing the forums to each other.

Our findings also showed that a significant portion of super-users are at risk of churning in the future. Doing a deep dive into the possible reasons that could contribute to super-user churn will be really helpful in creating outreach programs to incentivize these users to keep engaging with the site.

8 Conclusion

Overall, this project displays the importance of super-users within an OSN. Being able to foster a strong core of users can help drive engagement and traffic to an OSN. This then feeds back to users as they additionally have additional resources they are able to

take advantage of, which if they are all concentrated in one location, could drive those other users to participate more. By using predictive models to reliably identify these super-users, OSNs can help support them, continuing the production of high quality user-generated content.

References

[1] *Stack Overflow*. 2024. URL: https://stackoverflow.com/.

[2] T. Graham and S. Wright. "Analysing 'Super-Participation' in Online Third Spaces". In: *Analyzing Social Media Data and Web Networks*. Ed. by M. Cantijoch, R. Gibson, and S. Ward. London: Palgrave Macmillan, 2014. DOI: 10.1057/9781137276773_8.

[3] F. Zhang, S. Li, and Z. Yu. "The super user selection for building a sustainable online social network marketing community". In: *Multimedia Tools and Applications* 78 (2019), pp. 14777–14798. DOI: 10.1007/s11042-018-6829-0.

[4] I. Adaji and J. Vassileva. "Predicting Churn of Expert Respondents in Social Networks Using Data Mining Techniques: A Case Study of Stack Overflow". In: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. Miami, FL, USA, 2015, pp. 182–189. DOI: 10.1109/ICMLA.2015.120.

[5] M. A. Stelzner. *Social Media Marketing Industry Report*. Available at: http://www.socialmediaexaminer.com/socialmediamarketing-industry-report-2010/. 2010.

[6] Vilfredo Pareto. *Cours d'Économie Politique*. French. Lausanne, Switzerland: F. Rouge, 1896.

[7] Christo Wilson et al. "Beyond Social Graphs: User Interactions in Online Social Networks and their Implications". In: *ACM Transactions on the Web* 6.4 (Nov. 2012), 17:1–17:31. DOI: 10.1145/2382616.2382620.

[8] A. Barua, S. W. Thomas, and A. E. Hassan. "What are developers talking about? An analysis of topics and trends in Stack Overflow". In: *Empirical Software Engineering* 19.3 (2014), pp. 619–654. DOI: 10.1007/s10664-012-9231-y.

[9] C. Fu et al. "Patterns of interest change in Stack Overflow". In: *Scientific Reports* 12 (2022), p. 11466. DOI: 10.1038/s41598-022-15724-3.

[10] M. Asaduzzaman et al. "Answering questions about unanswered questions of Stack Overflow". In: *2013 10th Working Conference on Mining Software Repositories (MSR)*. San Francisco, CA, USA, 2013, pp. 97–100. DOI: 10.1109/MSR.2013.6624015.

[11] Samia Kabir et al. "Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions". In: *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. New York, NY, USA: Association for Computing Machinery, 2024, 935:1–935:17. DOI: 10.1145/3613904.3642596.

[12] H. Zhang et al. "An Empirical Study of Obsolete Answers on Stack Overflow". In: *IEEE Transactions on Software Engineering* 47.4 (Apr. 2021), pp. 850–862. DOI: 10.1109/TSE.2019.2906315.

[13] T. Bachschi et al. "From asking to answering: Getting more involved on Stack Overflow". In: *arXiv preprint* (2020). eprint: 2010.04025.

[14] Stack Exchange. *Stack Exchange Data Dump (August 2024) [Data set]*. Retrieved October 10, 2024, from Internet Archive. 2024. URL: https://archive.org/details/stackexchange.

[15] Faraday. *dbcrossbar*. 2024. URL: https://www.dbcrossbar.org/.

[16] Stack Exchange. *Active Users Per Month*. Query generated by Gian Cercena executed on the Stack Exchange Data Explorer. 2024. URL: https://data.stackexchange.com/stackoverflow/query/1873573/active-users-per-month (visited on 11/17/2024).

[17] Gordon Burtch, Dokyun Lee, and Zhichen Chen. "The consequences of generative AI for online knowledge communities". In: *Scientific Reports* 14.1 (May 2024), p. 10413. ISSN: 2045-2322. DOI: 10.1038/s41598-024-61221-0. URL: https://doi.org/10.1038/s41598-024-61221-0.

[18] Des Darilek. *Insights into Stack Overflow's Traffic*. Accessed: 2024-11-18. Aug. 2023. URL: https://stackoverflow.blog/2023/08/08/insights-into-stack-overflows-traffic/.

[19] Semrush. *Stack Overflow Website Traffic, Ranking, Analytics [October 2024]*. Accessed: 2024-11-18. 2024. URL: https://www.semrush.com/website/stackoverflow.com/overview/.

[20] Jeff Atwood. *Introducing Stackoverflow.com*. Accessed: 2024-11-19. 2008. URL: https://blog.codinghorror.com/introducing-stackoverflow-com/.

[21] Tianqi Chen and Carlos Guestrin. "XGBoost: A Scalable Tree Boosting System". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. ACM, Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. URL: http://dx.doi.org/10.1145/2939672.2939785.

9 Appendix

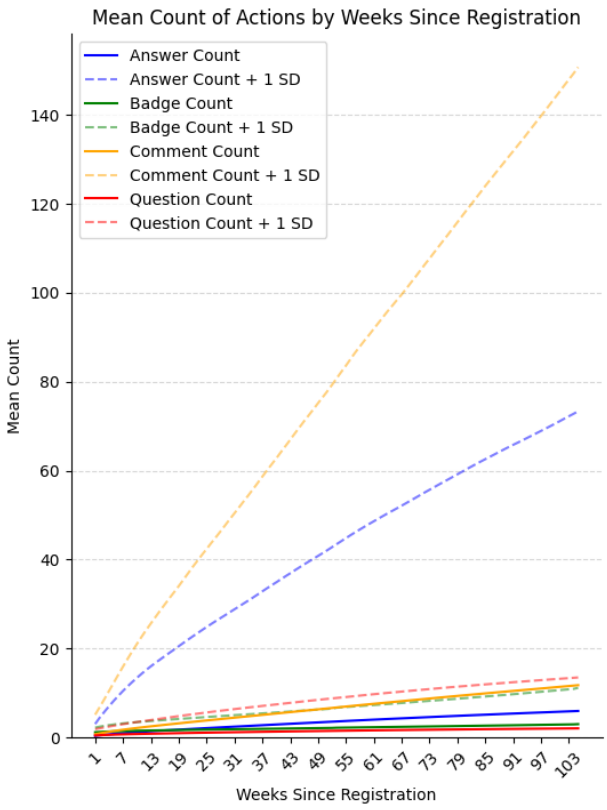

Figure 7: *Mean count plot, without the log y-scale.*

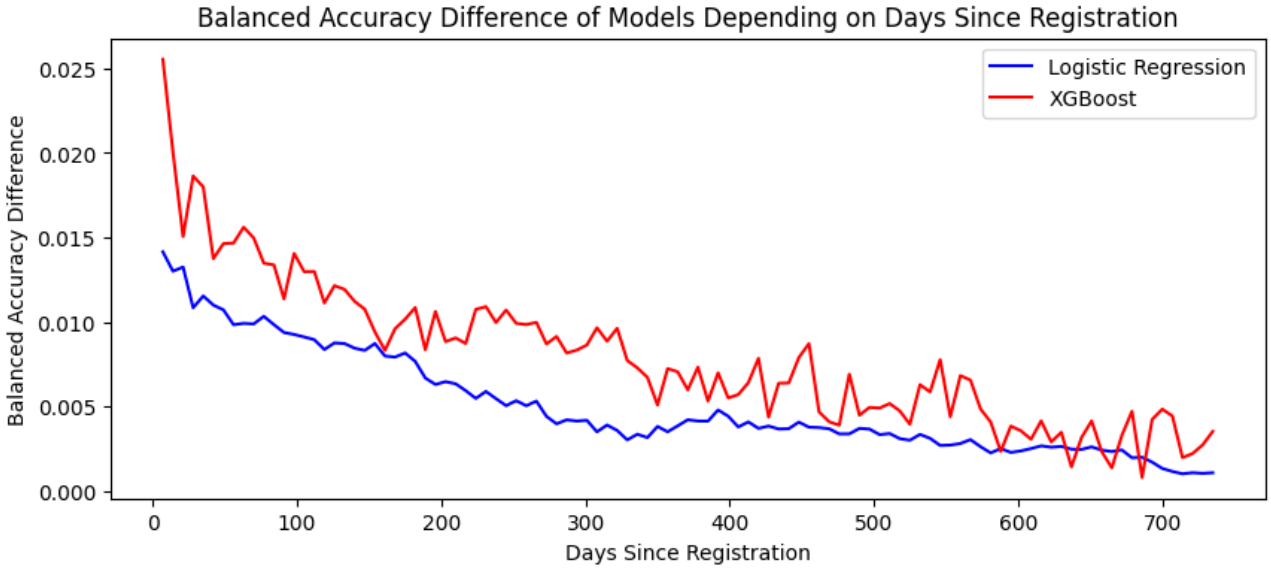

Figure 8: *The difference between using the full dataset and the activity dataset for each model (positive values mean the full dataset did better).*

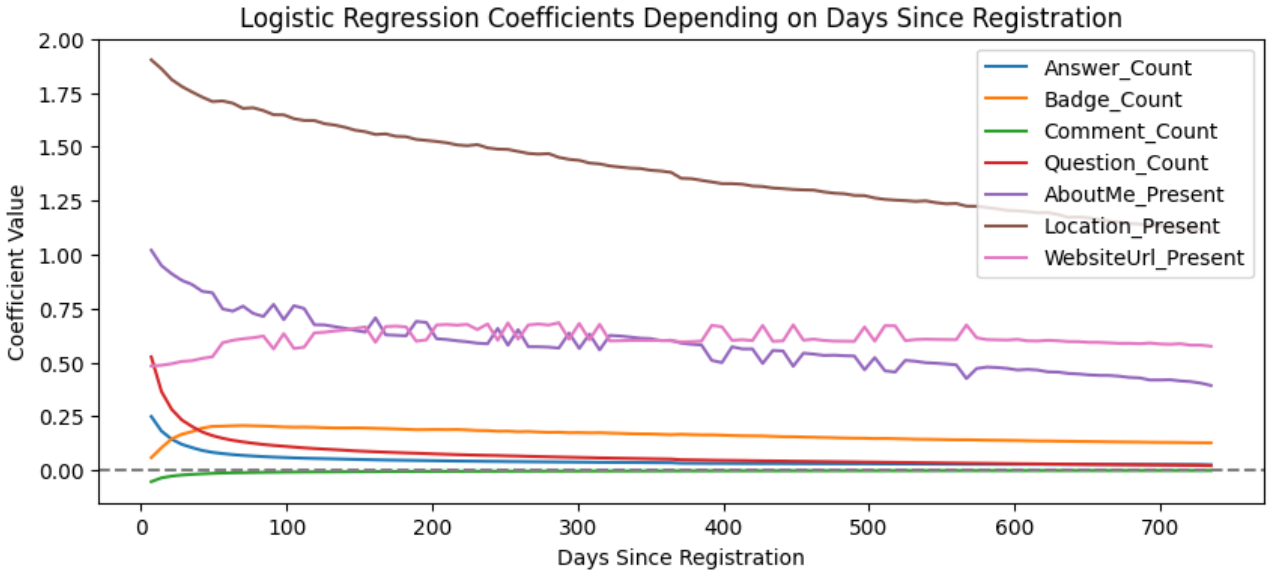

Figure 9: *Logistic model coefficients for predicting whether a user will become a super-user or not at different intervals of t.*

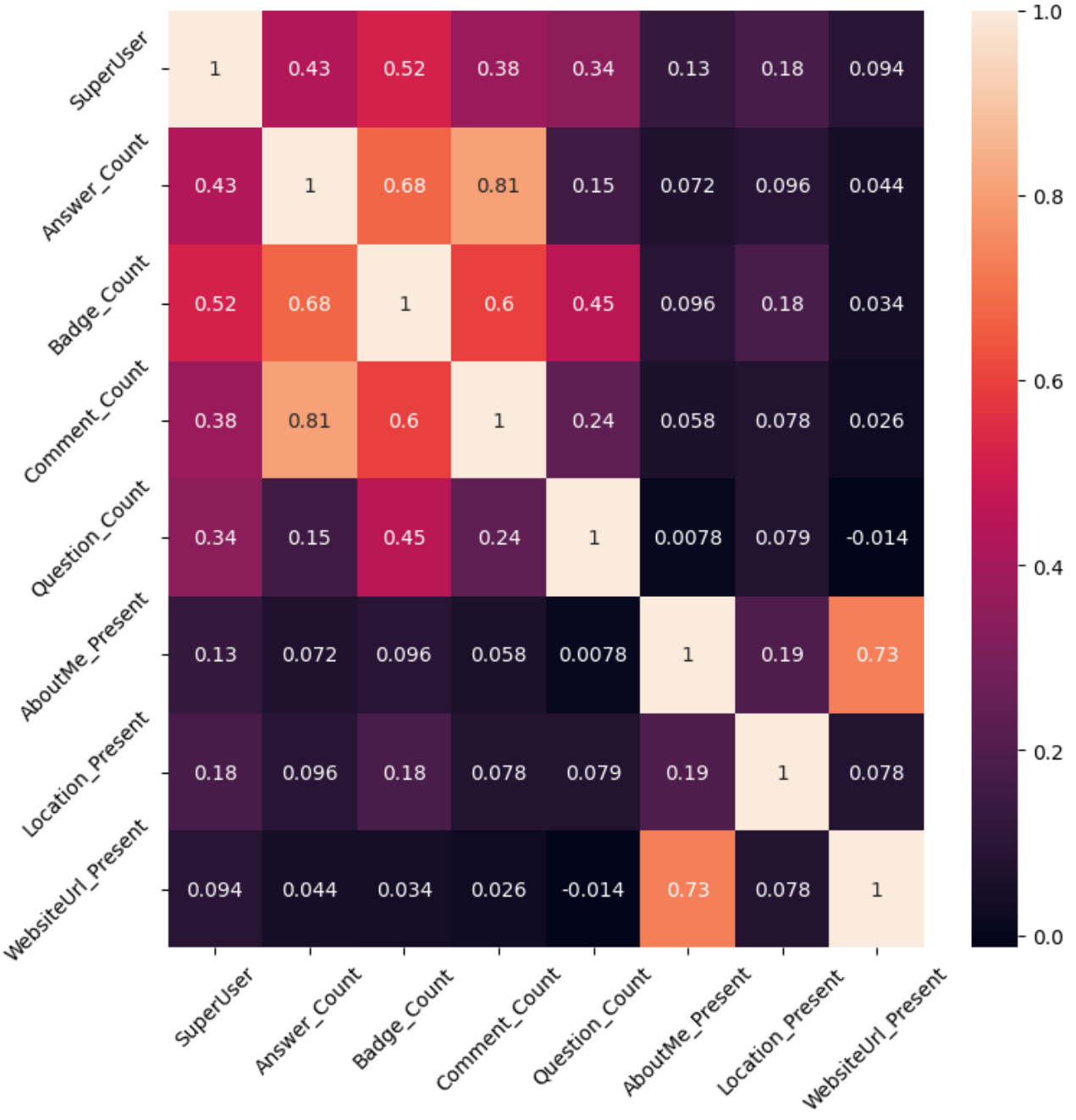

Figure 10: *Correlation table for all variables in the full dataset.*