# Prediction Of User Churn In Q&A Sites: A Case Study Of Stack Overflow

Jagat Sastry Pudipeddi
Stony Brook University
Department of Computer Science
jpudipeddi@cs.stonybrook.edu

Leman Akoglu
Stony Brook University
Department of Computer Science
leman@cs.stonybrook.edu

## ABSTRACT

Question and Answer (Q &A) sites form an excellent repository of crowdsourced knowledge, provided by those who ask questions and those who answer those questions. To make sure that these sites are self-sustainable, it is important for the site owners to make sure that new users, most of whom ask questions don't stop asking, while also ensuring that site veterans, who provide most of the answers continue to do so. In this study, we make use of data from stackoverfow.com to to find significant factors that cause users to leave in the early stage and those that cause veterans to leave in the later stage. We find that in both the cases, gap between subsequent posts are the most significant in determining the diminishing interest of users, and other factors like answering speed, reputation of those who answer their questions and number of answers received by the user come close. We use the features extracted to perform classification of users to identify those who are about to leave, and obtain accuracies that range from 64% to 72% as the number of control posts are changed. We then study these features by fixing the observation time frame instead of the number of posts. We conclude by explaining how this study and the insights provided by the classifiers can be used by the owners of such sites to try and stop the users who are about to leave.

## 1. INTRODUCTION

To get answers to questions which haven't been documented yet, or are difficult to find in an existing article, users make use of Q&A sites for refuge. It is the online equivalent of a room full of experts where users can walk in and ask questions which other users might have answers to. Users are motivated to provide answers for reasons beyond monetary benifits, since almost none of the popular Q&A sites pay users to answer questions. Instead, they have various form of virtual rewards and reputation system, to keep the users motivated. Since the site's popularity is based on the breadth of the questions and answers provided, it is important for the site to make sure that users who post questions look at the site as reliable and also to motivate those who answer to continue doing so. In this paper, we use Stack Overflow's data to perform our analysis. As shown in[ Figure 1], the number of users who post a particular number of posts follow a power law distribution. In fact, 94% of the users have less than 21 posts. Moreover, as shown in [Figure 2], the more a users posts, the more his tendency to provide answers, and more is the probability that the post get accepted by the user who asked the question. Hence, for the website to have higher quality answers, it is important that users who post continue to do so. We study two sets of users, those who make 1-5 posts as those representing the users, who need to be nurtured into experts, and those who make 15-20 posts as those experts who provide answers. Since both the type of posts are equally important for the site, we focus our study on the factors that cause either newbies(post 1-5), or the experts (15-20 posts) to leave. However, every now and then we study how the discriminative power of a feature varies across posts. As far as we are aware of, this is the first study of user churn in Q&A sites that takes both these classes of users into account. We also provide a model based on decision tree which can be used by such site admins in predicting if a user is about to leave and to easily find the reasons for that.

## 2. Q&A SITES AND DATA DETAILS

***** here we will discuss what q&a sites look like, what do they offer, how easy they are to use, etc. We will give more details for StackOverflow, present plots such as distributions, and finally give data statistics, counts etc. ***** StackOverflow is a popular Q&A site for programmers. As of May 13, 2013 the site reports 5.5 million visits/day.The data we use for our research is based on the activity on the site from July 31, 2008 to July 31, 2012 - a period of four years. Since it is not possible to know if a user has left the site forever, we consider a user having left the site if he does not post anything for more than six months. That comprises less than 13Since it doesn't make sense to study those users who haven't posted anything, we exclude those users from the control set.

Some of the details that might help in gaining some pespective are provided in table 1

**Table 1: StackOverflow data details**

| Feature | Number |
|---|---|
| Questions | 3.4m |
| Answers | 6.8m |
| Comments | 13.2m |

## 3. RESEARCH QUESTIONS

***** introduce the 2 types of questions we are interested in (and of course why they matter): (1) predicting prolific vs.unproductive users, (2) churn (who leaves, who stays) ***** We are interested in finding the intrinsic factors and signals that cause a user to stop posting. Some of the extrinsic factors like the user losing interest in online activities manifest in the form of signals like increase in inter-post gap. However, others like job loss or end of college studies cause the users to leave abruptly, and these are factors that are hard to account for. Explicitly, the questions we are interested in answering through this study are

1. What are the intrinsic factors and signals that make a new user leave after a particular post?
2. What makes a user who has been posting for a very long time leave after a particular post?
3. Are the causes common? If not, how do the analyzed features vary across these two classes?
4. Can we extend this study to make predictions for any user at any given time?

I categorize these questions into two different task buckets. 1, 2, and 3 are covered by Task 1 and 4, which is more of a generalization task, is covered by Task 2.

## 4. METHODOLOGY

For the first task where we fix the number of posts made, our control set consists of those users who leave after having made K (1-5,16-20) posts and those who stay back after each of these posts. We extracted 21 fixed and 20 variable features (gapK, described in the list below),with each feature belonging to one of eight different categories. All the features and the corresponding categories are listed in table 2.

The number of users who

For the second task where we fix the observation time, the temporal features are collapsed into a few other features, since , as we will see below, only a few inter-post gaps are significant for the prediction task. The control set for this task consists of those users who leave either after 7 days or 15 days or 30 days after their account creation.

In subsequent sections, wherever I say ”K”, I refer to the fixed number of observation posts in task 1 and wherever I use ”T”, I refer to the fixed observation time in task 2. Features are referred to by the short names given in the table.

Since the number of users who churn reduces drastically relative to the users who stay, we undersample the users who stay before performing classification.

## 5. ANALYSIS AND RESULTS

The probability of a user who answers a lot reduces with time. ItâĂŹs even lower if he asks more questions alongside. However, for large number of answers, a user who does not ask any question at all tends to stay. Note that this can simply be attributed to the plot under this which shows that as a user engages with the site, his chances of staying with the site increases as well. This can be attributed to some sort of expertise the user might have, which motivates him to stay and answer a lot, while not being required to ask any question.

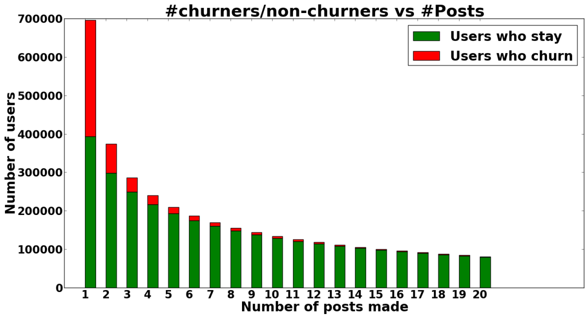

**Figure 1: Notice how the number of users, as well as the churn/stay ratio reduces as the number of posts increases**

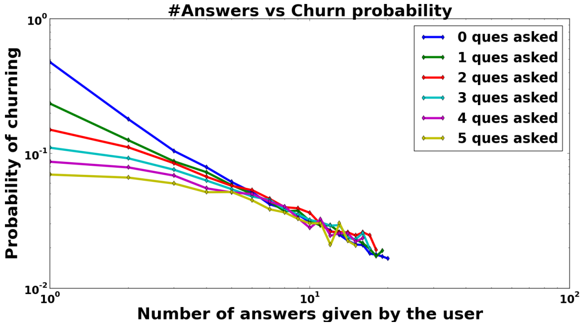

**Figure 2: The probability of churning for a user who answers a lot reduces with time. It is even lower if he asks more questions alongside.**

## 6. EXPERIMENTS

## 7. RELATED WORK

## 8. CONCLUSION

In this study, we explored the factors that are correlated with users leaving after the first few posts(<=5) and those that are correlated with them leaving after a large number of posts (16-20). We found that the gaps between posts, especially between the latest two posts, are the most significant factors in determining whether a user is about to leave the site or not. A quick look at the plots representing gaps between posts reveal that the gap keeps increasing till the user leaves the site, while successive gaps for users who stay are stable.

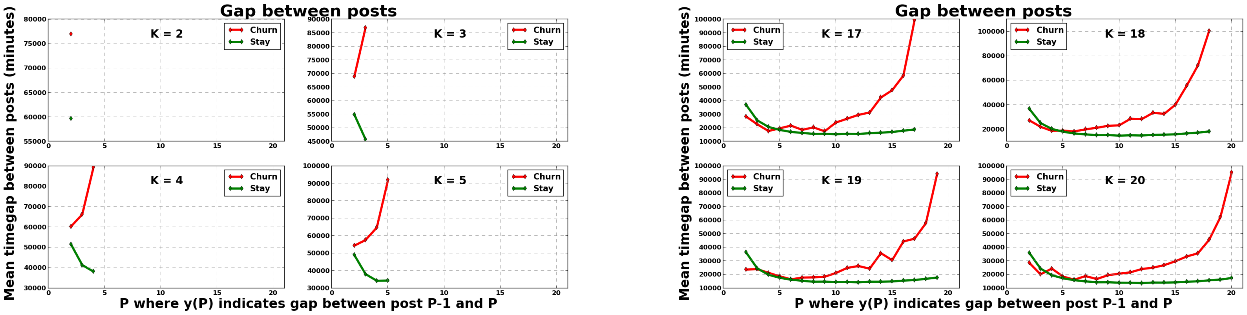

**Figure 3: For a user who churns, gap between every posts keeps increasing. Gaps for those who stay, remain relatively stable. Users who stay for a long time tend to post at least once every two weeks.**

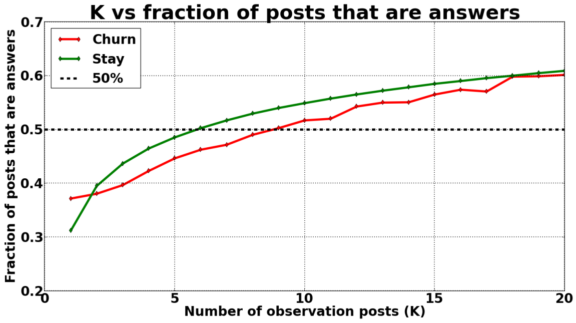

**Figure 4: Users who stay tend to provide more answers compared to those who churn at a given K.**

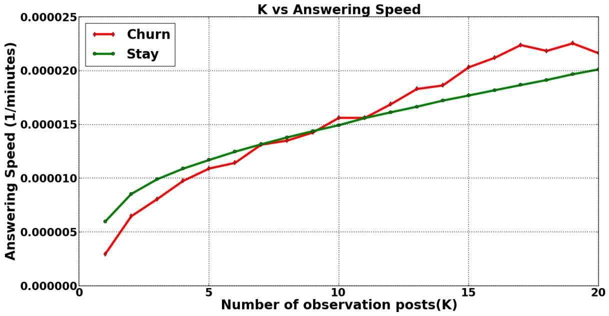

**Figure 5: More the time taken for the user to give an answer, more the chance of churning. After K=11, higher answering speed correlates with higher chances of churning. This can be explained by the corresponding decrease in the weighted rank of answers.**

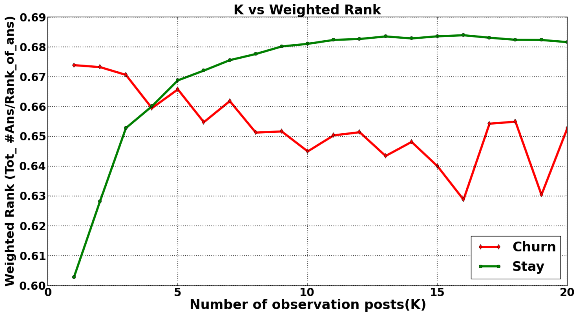

**Figure 6: Weighted rank acts as a measure of gratification. This also explains the correlation between higher answering speed and churn probability at higher K.**

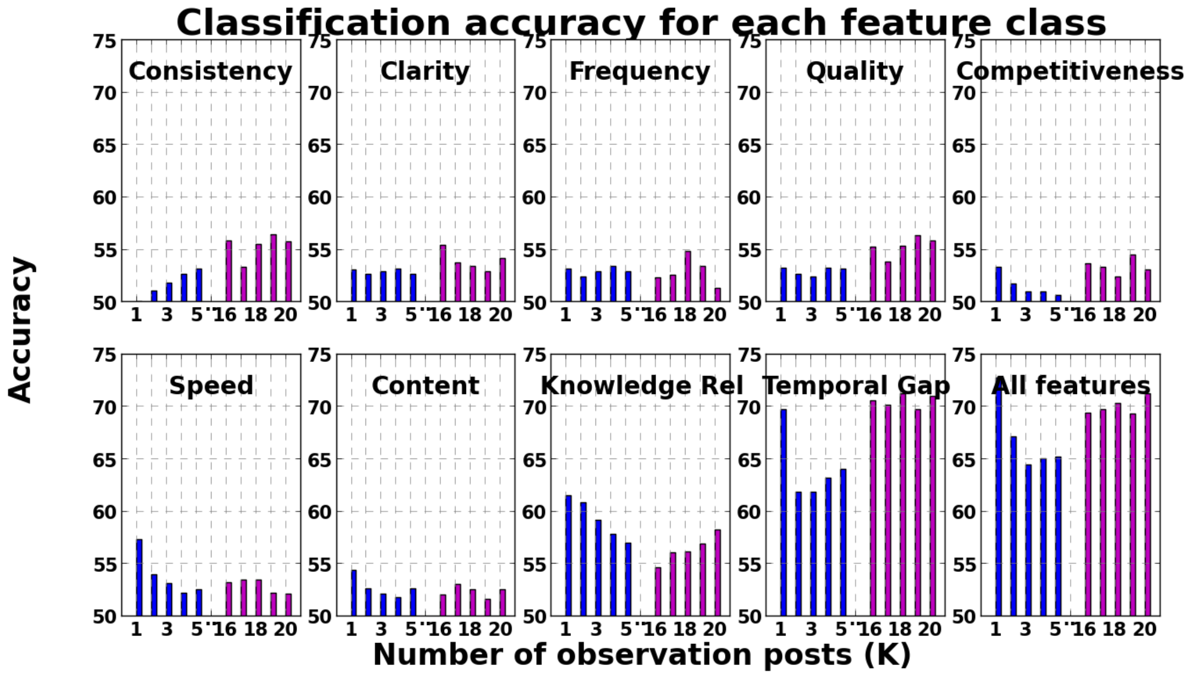

Figure 9: Decision Tree classifier accuracy when features from each feature class are used in isolation, as K varies

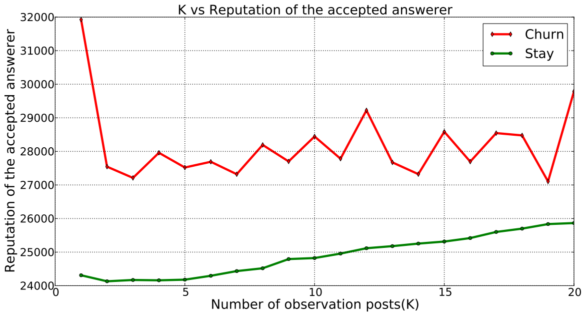

Figure 7: Higher reputation of those whose answer correlates with the questioner churning. This is because uses who stay tend to answer more and answers constitute most of the reputation points.

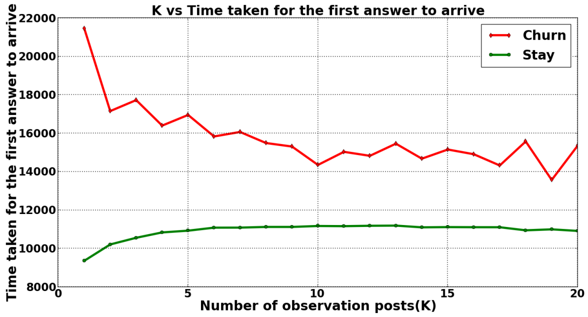

Figure 8: More the time taken for a user to receive an answer, lesser the satisfaction level and more the chances of churning.

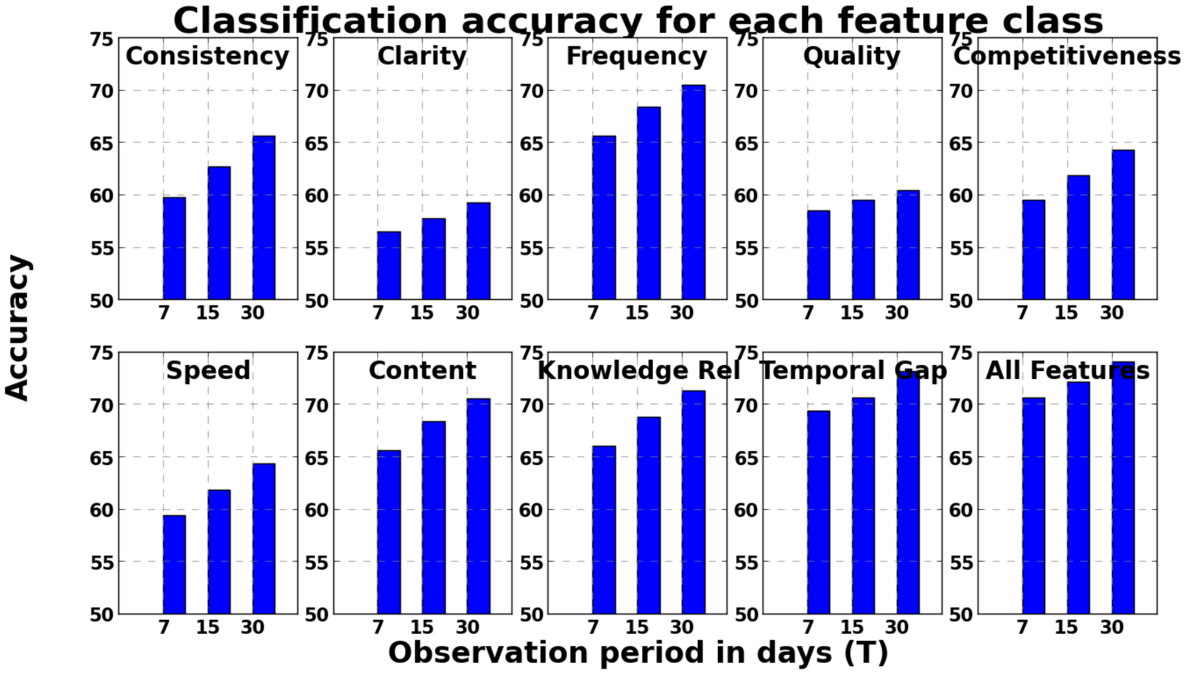

**Figure 10: Decision Tree classifier accuracy when features from each feature class are used in isolation as T varies**

**Table 2: Features Extracted**

| **Temporal Features** |
|---|
| ***gap1***: Time gap between account creation and first post. |
| ***gapK***: *Task 1.* Time gap between (k-1)th post and kth post for each possible $k \leq K$. |
| ***last_gap***: *Task 2.* Time gap between the last post and the post before that. |
| ***time_since_last_post***: *Task 2.*Elapsed between the last post made and the observation deadline. |
| ***mean_gap***: *Task 2.* Mean of all the time gaps between posts for all the posts made during the observation period. |
| **Frequency** |
| ***num_answers***: Number of answers |
| ***num_questions***: Number of questions |
| ***ans_que_ratio***: Ratio of #answers to #questions |
| ***num_posts***: *Task 2.* Number of posts |
| **Quality** |
| ***ans_score***: Reputation score obtained per answer given |
| ***que_score***: Reputation score obtained per question asked |
| **Consistency** |
| ***ans_stddev***: Standard deviation of the reputation scores obtained for the answers |
| ***que_stddev***: Standard deviation of the reputation scores obtained for the questions |
| **Speed** |
| ***answering_speed***: Inverse of the time gap between someone posting a question and the user answering it. |
| **Gratitude** |
| ***ans_comments***: Number of comments made on the user's answer |
| ***que_comments***: Number of comments made on the user's question |
| **Competitiveness** |
| ***relative_rank_pos***: Total number of answers for a question divided by Rank of user's answer |
| **Content** |
| ***ans_length***: Length of an answer |
| ***que_length***: Length of a question |
| **Knowledge Relevance** |
| ***accepted_answerer_rep***: Mean reputation of the user whose answer was accepted |
| ***max_rep_answerer***: Mean reputation of the user who had the maximum reputation among all those who answered a question |
| ***num_que_answered***: Number of questions posted by the user that got answered |
| ***time_for_first_ans***: Time taken for the arrival of the first answer to a question. |
| ***rep_questioner***: Mean reputation of the user whose question was answered. |
| ***rep_answerers***: Mean reputation of the users who answered the question. |
| ***rep_co_answerers***: Mean reputation of the users who answered the same question as the control user |
| ***num_answers_recvd***: Mean number of answers received for every question the user posts |

**Table 3: List of all the features classes, sorted by their computational complexity. Underlined features represent those that are used in one task but not the other.**

| K=1 | K=5 | K=16 | K=20 |
|---|---|---|---|
| gap1 | gap5 | gap16 | gap20 |
| acc_answerer_rep | gap4 | gap15 | gap19 |
| answering_speed | gap1 | gap14 | gap17 |
| ans_post_length | acc_answerer_rep | que_post_length | gap18 |
| time_for_first_ans | rep_questioner | num_answers_recvd | gap2 |
| rep_answerers | num_answers_recvd | gap12 | rep_questioner |
| que_post_length | gap3 | gap11 | rep_answerer |
| ans_score | gap2 | accepted_answerer_rep | answering_speed |
| rep_questioner | max_rep_answerer | gap8 | que_post_length |
| max_rep_answerer | que_post_length | num_questions | num_questions |

**Table 4: *Task 1.* Top 10 important features based on sum of information gain weighted by the number of samples split[CITE: http://bus.utk.edu/stat/datamining/Decision Trees for Predictive Modeling (Neville).pdf ]** . For every K, the latest time gaps are the most important.

| T=7 | T=15 | T=30 |
|---|---|---|
| time_since_last_post | time_since_last_post | time_since_last_post |
| accepted_answerer_rep | accepted_answerer_rep | accepted_answerer_rep |
| num_posts | num_posts | num_posts |
| num_answers_recvd | ans_que_ratio | ans_que_ratio |
| ans_score | que_score | que_score |
| ans_que_ratio | num_answers_recvd | rep_questioner |
| que_score | ans_score | num_answers_recvd |
| rep_questioner | rep_questioner | ans_score |
| gap1 | first_post_gap | first_post_gap |
| ans_comments | num_answers | mean_gap |

**Table 5: *Task 2.* Top 10 important features based on sum of information gain weighted by the number of samples split**

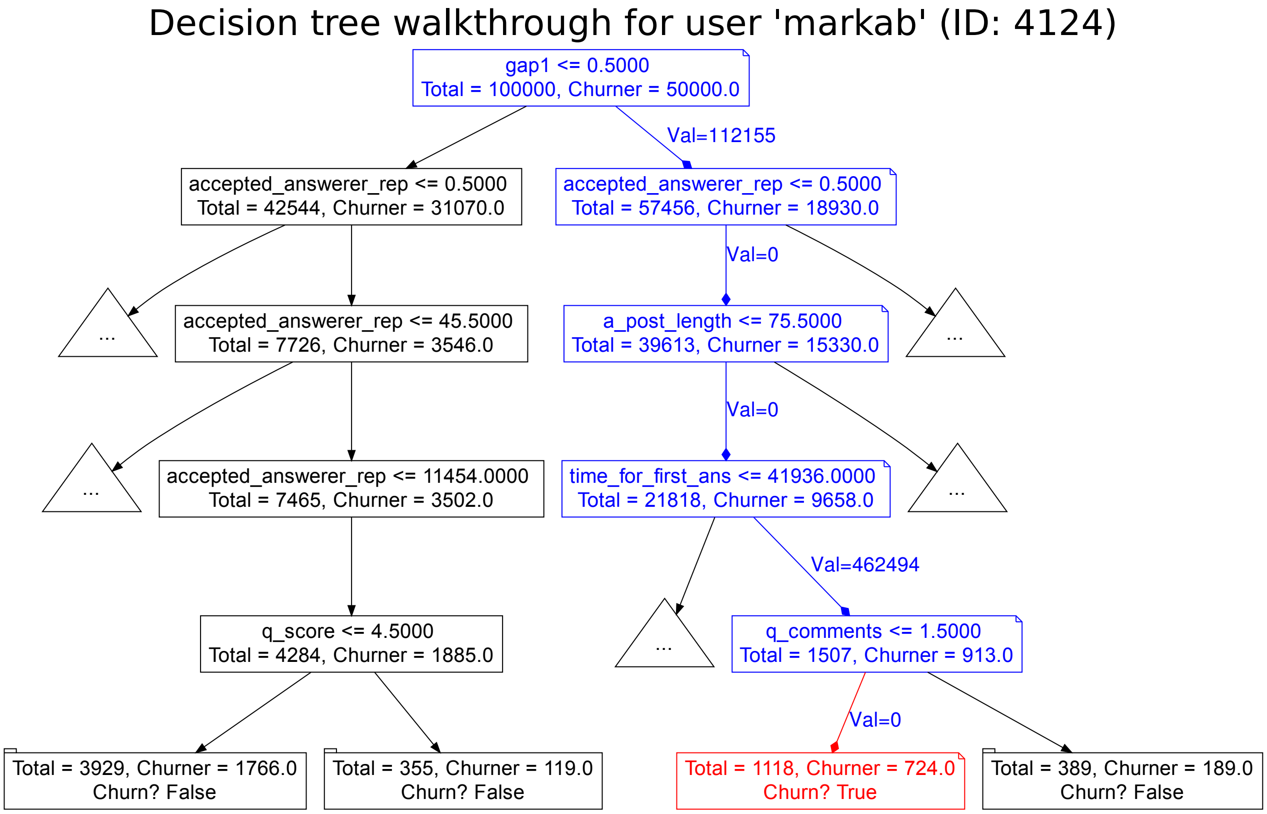
Decision tree walkthrough for user 'markab' (ID: 4124)
gap1 <= 0.5000
Total = 100000, Churner = 50000.0
Val=112155
accepted_answerer_rep <= 0.5000
Total = 42544, Churner = 31070.0
accepted_answerer_rep <= 0.5000
Total = 57456, Churner = 18930.0
Val=0
...
accepted_answerer_rep <= 45.5000
Total = 7726, Churner = 3546.0
a_post_length <= 75.5000
Total = 39613, Churner = 15330.0
...
Val=0
...
accepted_answerer_rep <= 11454.0000
Total = 7465, Churner = 3502.0
time_for_first_ans <= 41936.0000
Total = 21818, Churner = 9658.0
...
Val=462494
q_score <= 4.5000
Total = 4284, Churner = 1885.0
...
q_comments <= 1.5000
Total = 1507, Churner = 913.0
Val=0
Total = 3929, Churner = 1766.0
Churn? False
Total = 355, Churner = 119.0
Churn? False
Total = 1118, Churner = 724.0
Churn? True
Total = 389, Churner = 189.0
Churn? False

**Figure 11: Simple example of how the learnt decision tree model predicts if a certain user will churn or not after one post. (Truncted and restricted to depth 5.)**