User Churn in Focused Question Answering Sites: Characterizations and Prediction

Jagat Pudipeddi
Stony Brook University

Leman Akoglu
Stony Brook University

Hanghang Tong
City College of New York

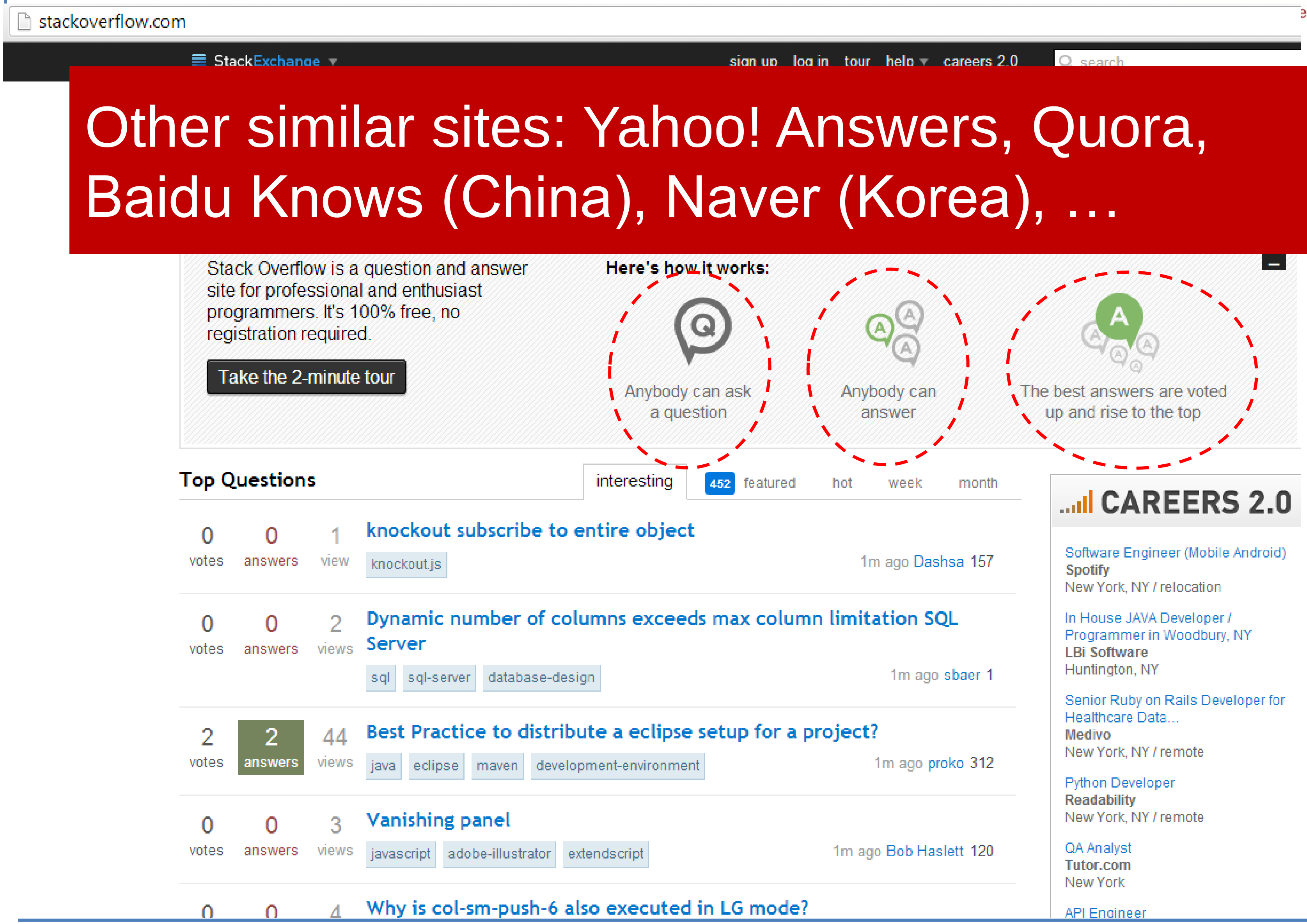
WWW WebScience
Seoul, Korea
April 7-10, 2014

Stony Brook University
Computer Science

CU
NY

StackOverflow

Other similar sites: Yahoo! Answers, Quora, Baidu Knows (China), Naver (Korea), …



stackoverflow.com

Stack Overflow is a question and answer site for professional and enthusiast programmers. It's 100% free, no registration required.

Take the 2-minute tour

Here's how it works:

Anybody can ask a question

Anybody can answer

The best answers are voted up and rise to the top

Top Questions

0 votes 0 answers 1 view **knockout subscribe to entire object**
knockout.js 1m ago Dashsa 157

0 votes 0 answers 2 views **Dynamic number of columns exceeds max column limitation SQL Server**
sql sql-server database-design 1m ago sbaer 1

2 votes 2 answers 44 views **Best Practice to distribute a eclipse setup for a project?**
java eclipse maven development-environment 1m ago proko 312

0 votes 0 answers 3 views **Vanishing panel**
javascript adobe-illustrator extendscript 1m ago Bob Haslett 120

0 0 4 **Why is col-sm-push-6 also executed in LG mode?**

CAREERS 2.0

Software Engineer (Mobile Android)
Spotify
New York, NY / relocation

In House JAVA Developer / Programmer in Woodbury, NY
LBi Software
Huntington, NY

Senior Ruby on Rails Developer for Healthcare Data...
Medivo
New York, NY / remote

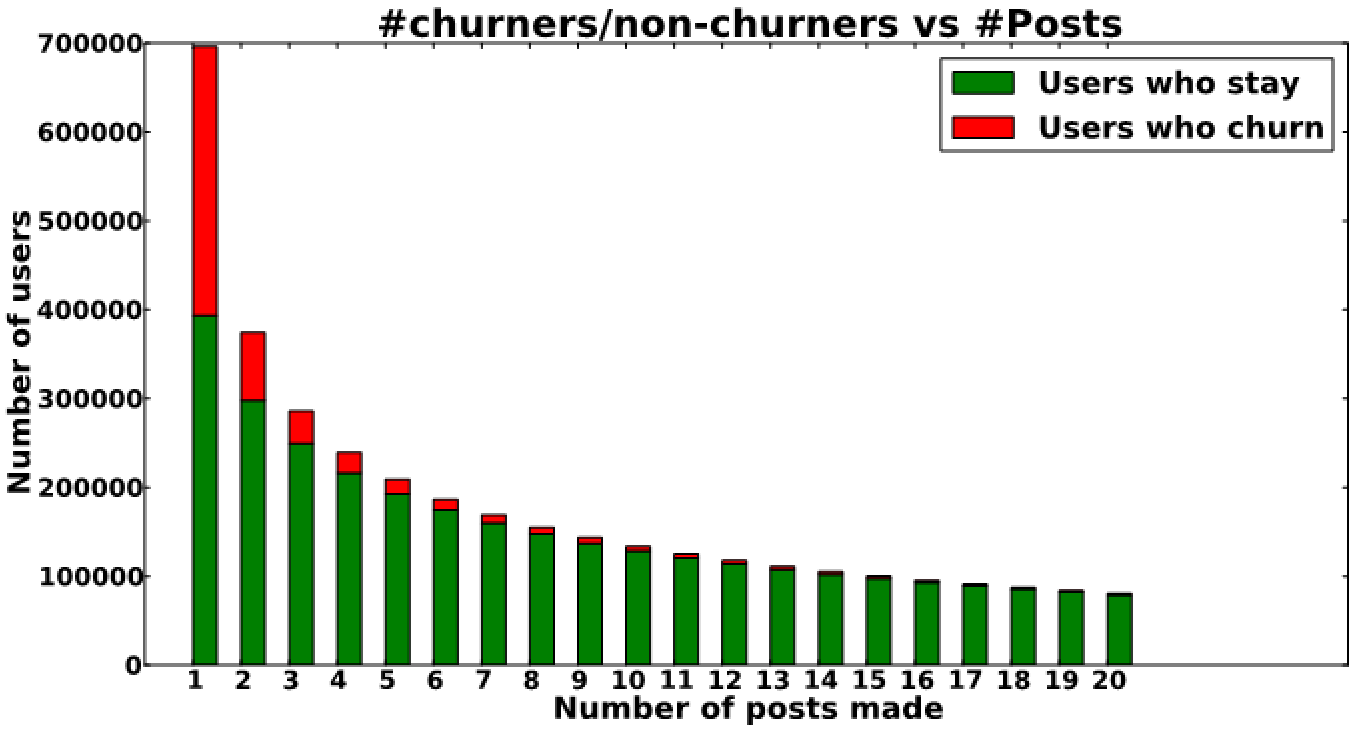
Python Developer
Readability
New York, NY / remote

QA Analyst
Tutor.com
New York

API Engineer

User Churn

- User churn is a problem: large fraction of users churn after only a few posts



Research questions

Characterization

- What are intrinsic factors / signals that make a new user (newbie) leave after a few posts?
- What makes a prolific user (veteran) leave after a certain number of posts?
- Are there common factors across two user groups (i.e., newbie vs. veteran)?

Prediction

- How well can we predict if a user is likely to churn using evidential features?

2 Prediction Tasks

Task 1.

Given the first k posts (questions and answers) of a user,

$$1 \leq k \leq 5 \text{ and } 16 \leq k \leq 20$$

Task 2.

Given the first T days of site activity of a user,

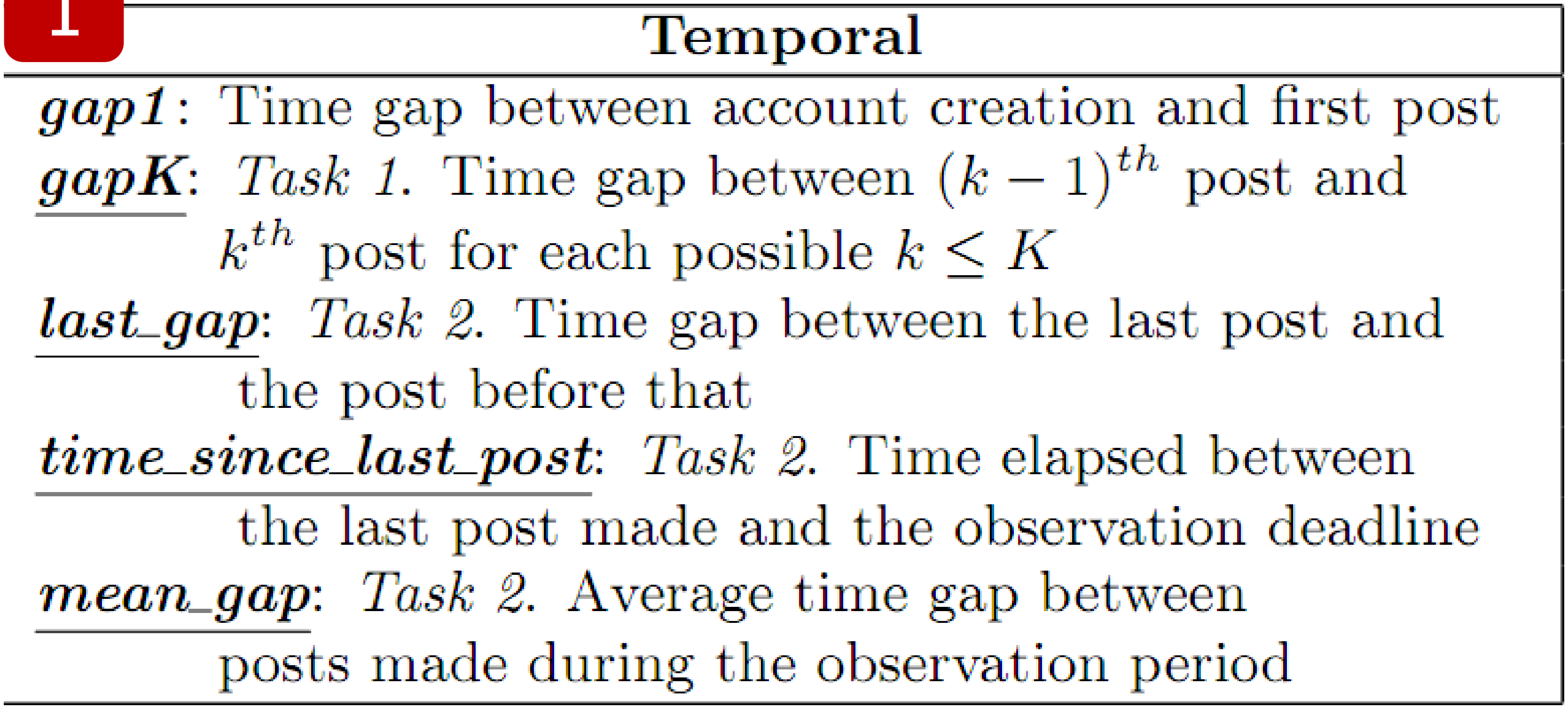
$$T = \{7, 15, 30\} \text{ days}$$

Predict how likely it is that the user will churn (i.e., will have no activity for the next 6 months).

Feature Extraction

- We find and organize 9 groups of features likely associated with churn

1

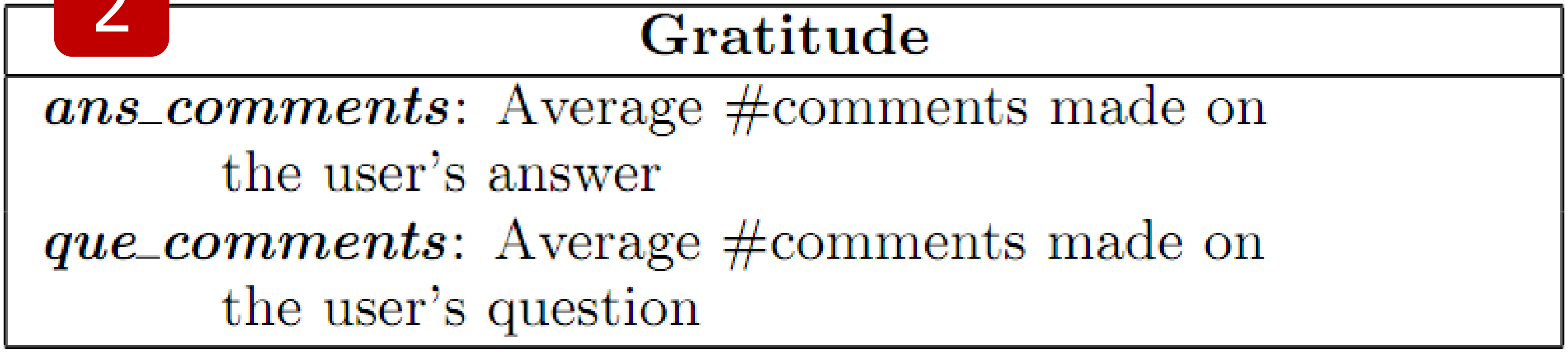


Temporal
gap1: Time gap between account creation and first post
gapK: *Task 1.* Time gap between $(k-1)^{th}$ post and k^{th} post for each possible $k \leq K$
last_gap: *Task 2.* Time gap between the last post and the post before that
time_since_last_post: *Task 2.* Time elapsed between the last post made and the observation deadline
mean_gap: *Task 2.* Average time gap between posts made during the observation period

Feature Extraction

- We find and organize 9 groups of features likely associated with churn

2

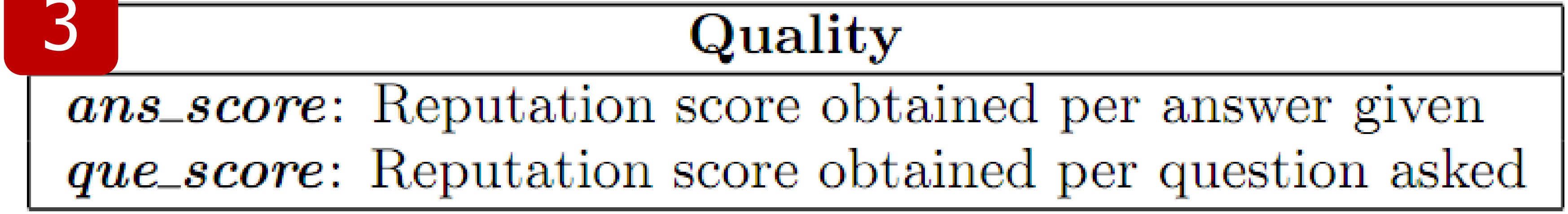


Gratitude
ans_comments: Average #comments made on the user’s answer
que_comments: Average #comments made on the user’s question

Feature Extraction

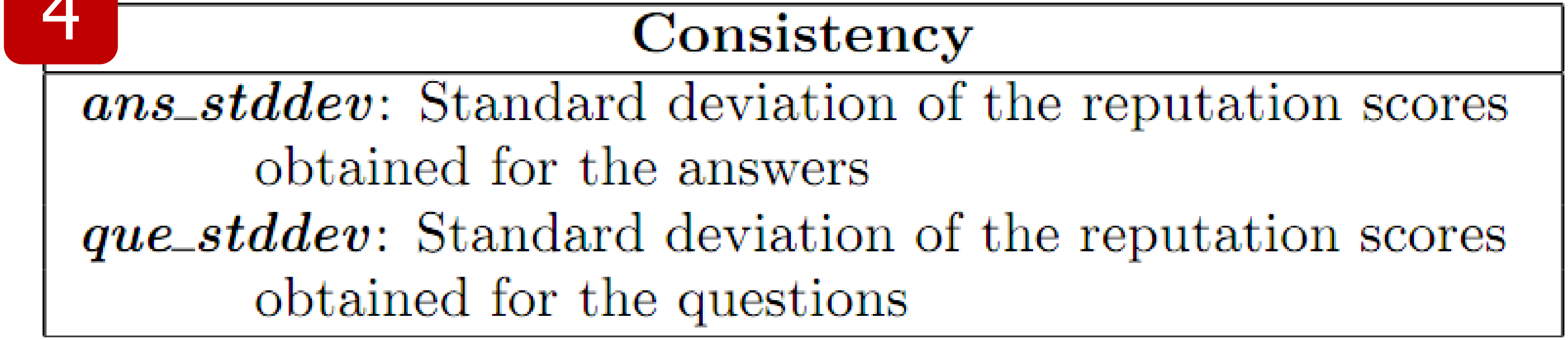
- We find and organize 9 groups of features likely associated with churn

3



Quality
ans_score: Reputation score obtained per answer given
que_score: Reputation score obtained per question asked

4



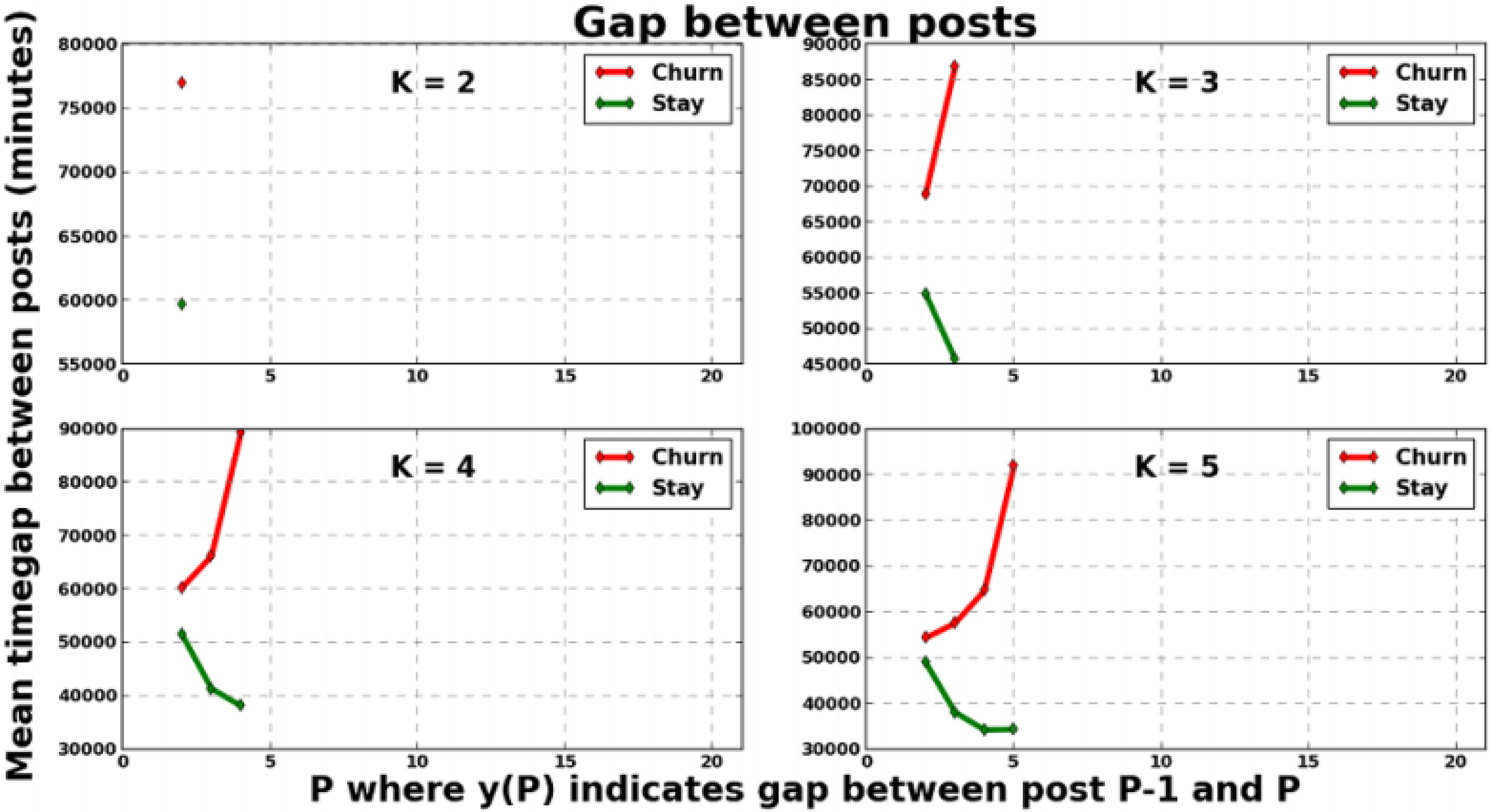
Consistency
ans_stddev: Standard deviation of the reputation scores obtained for the answers
que_stddev: Standard deviation of the reputation scores obtained for the questions

Feature Extraction

- We find and organize 9 groups of features likely associated with churn
  - Temporal
  - Gratitude
  - Quality
  - Consistency
  - Frequency
  - Speed
  - Content
  - Competitiveness
  - Knowledge Level

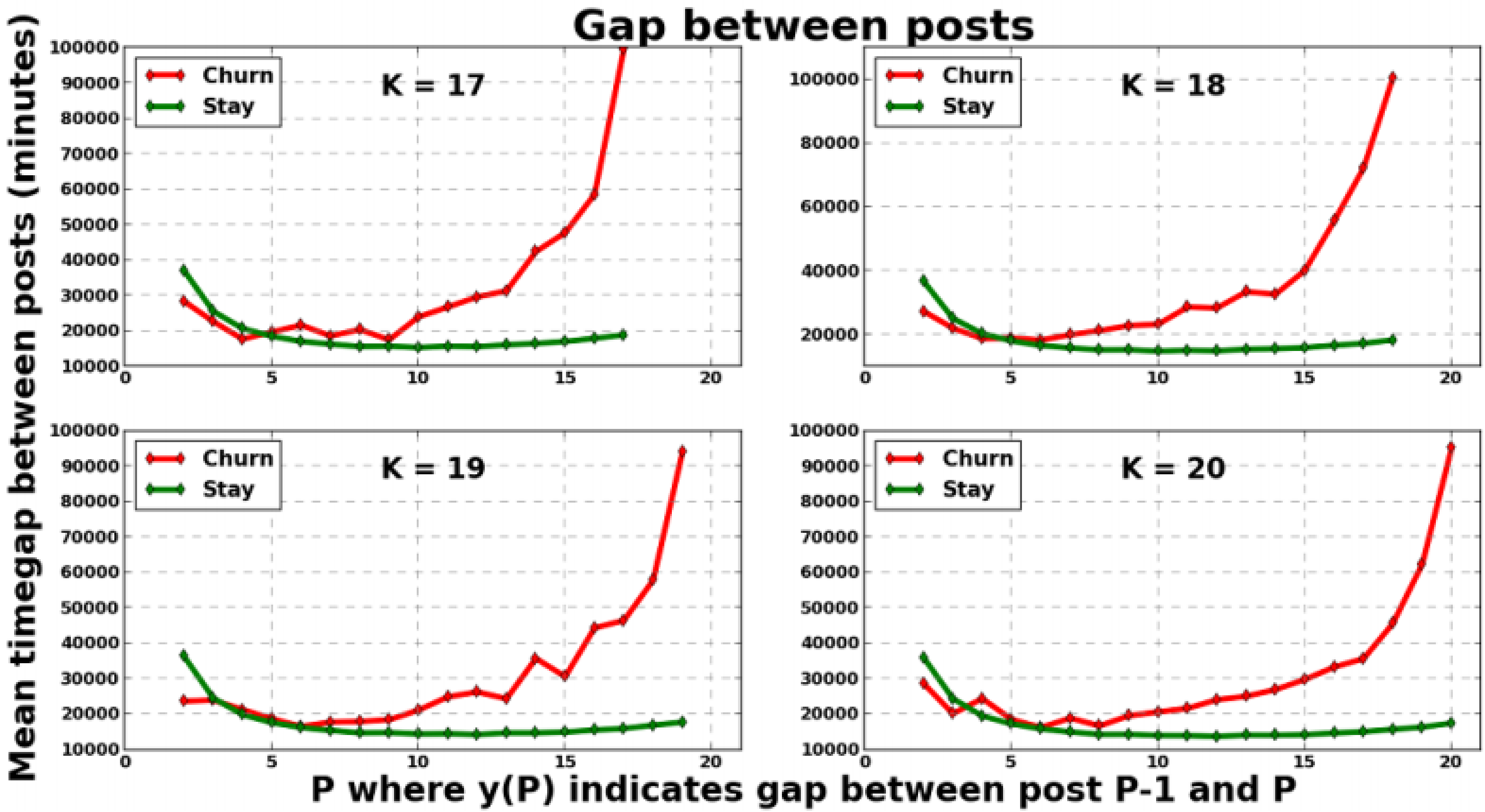
Feature Analysis

- Most significant signal: temporal gaps



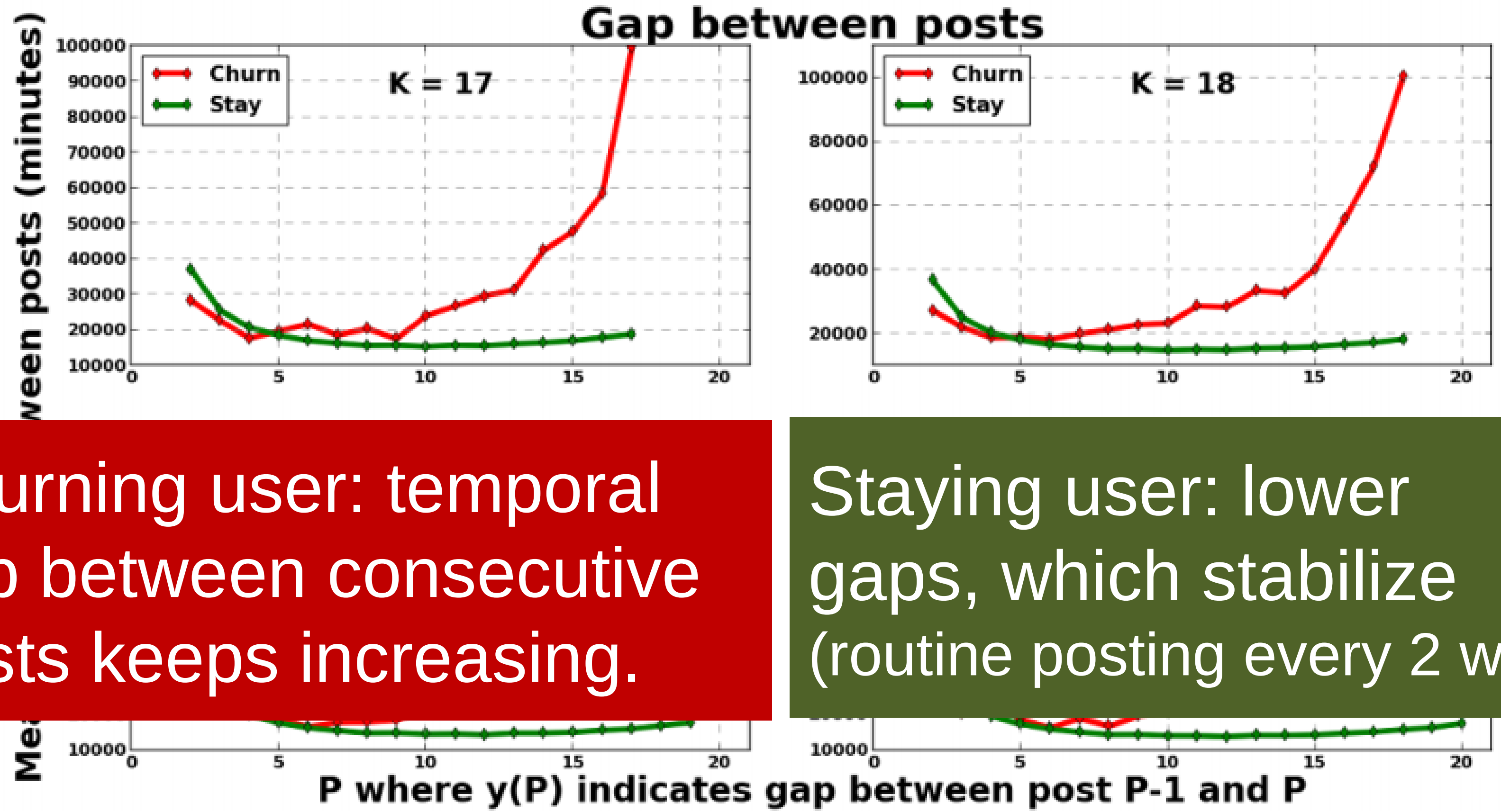
Feature Analysis

- Most significant signal: temporal gaps



Feature Analysis

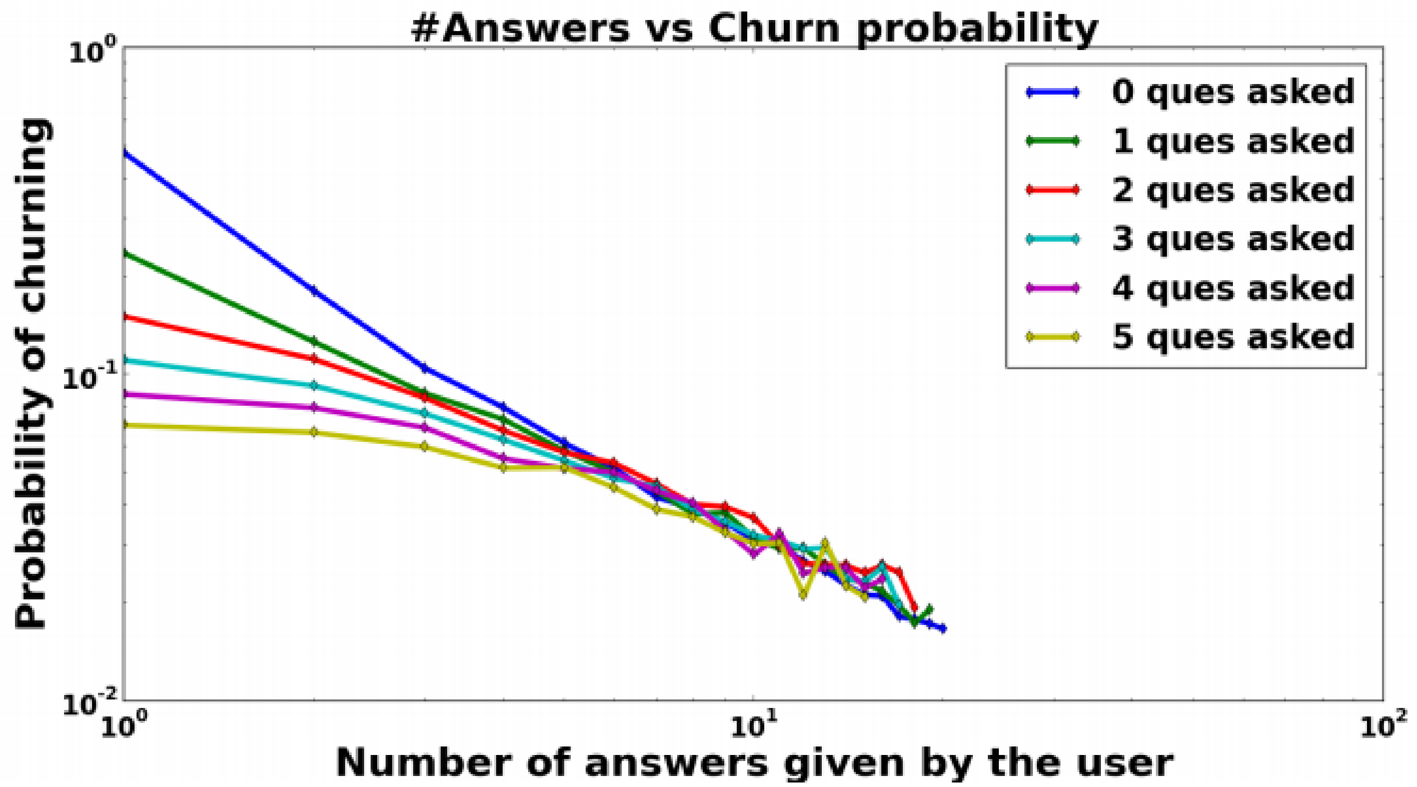
- Most significant signal: temporal gaps



Churning user: temporal gap between consecutive posts keeps increasing.

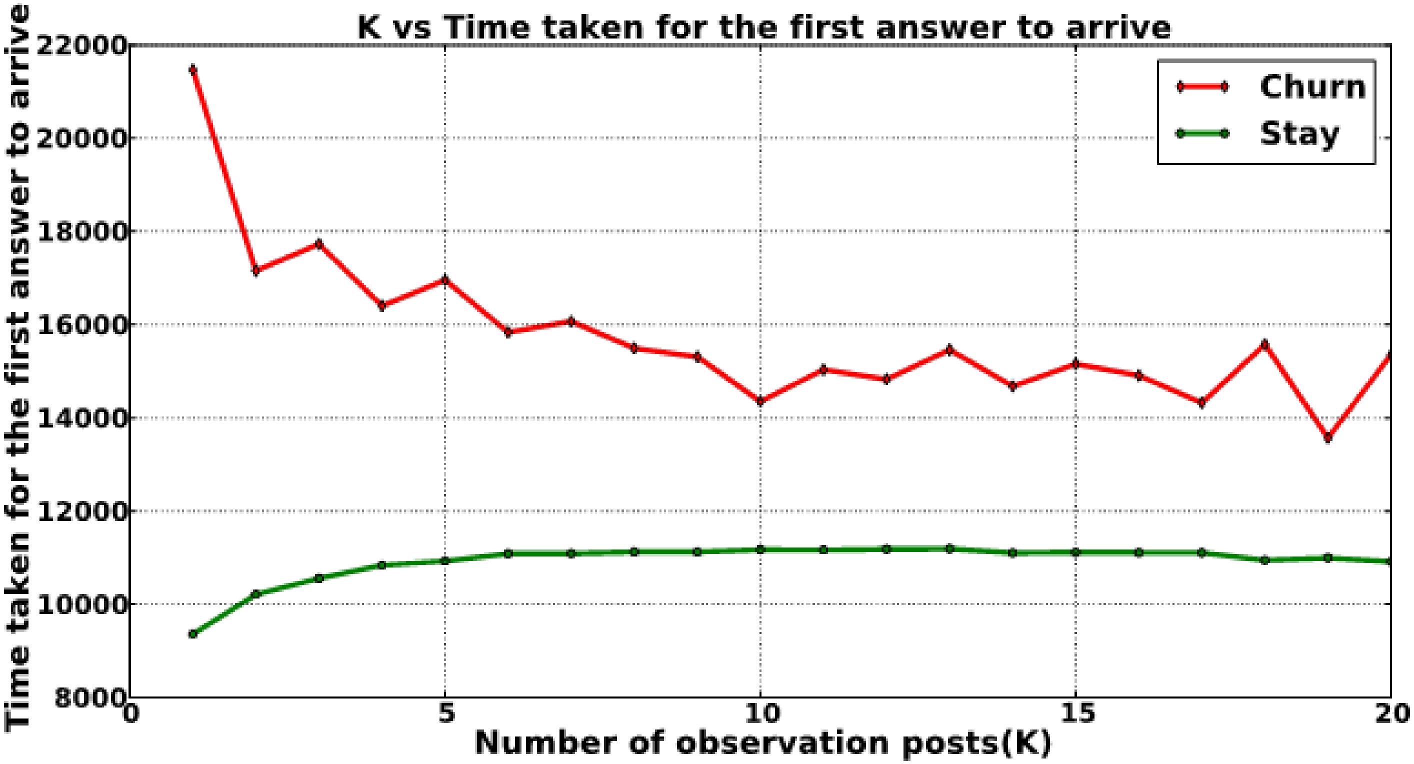
Staying user: lower gaps, which stabilize (routine posting every 2 wks)

Feature Analysis



The more answers by a user, the lower probability of churn; even lower if more questions asked alongside.

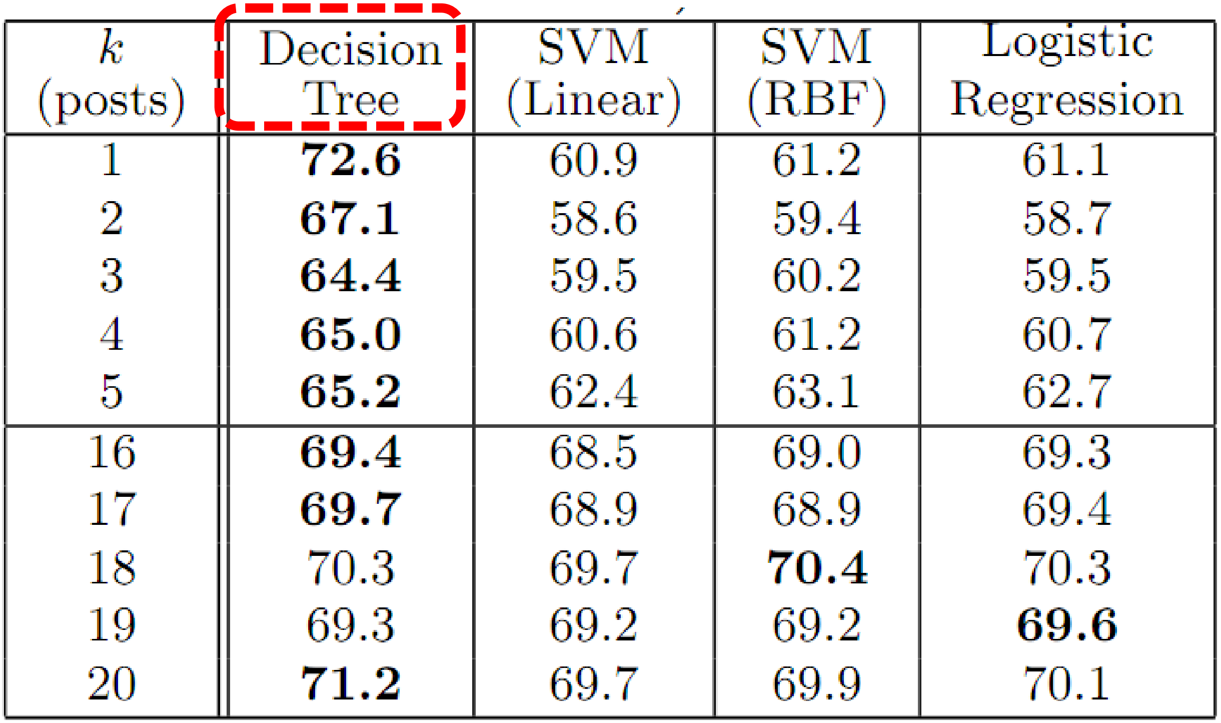
Feature Analysis



The more the time taken to receive an answer, the lesser satisfaction level, more chances of churn.

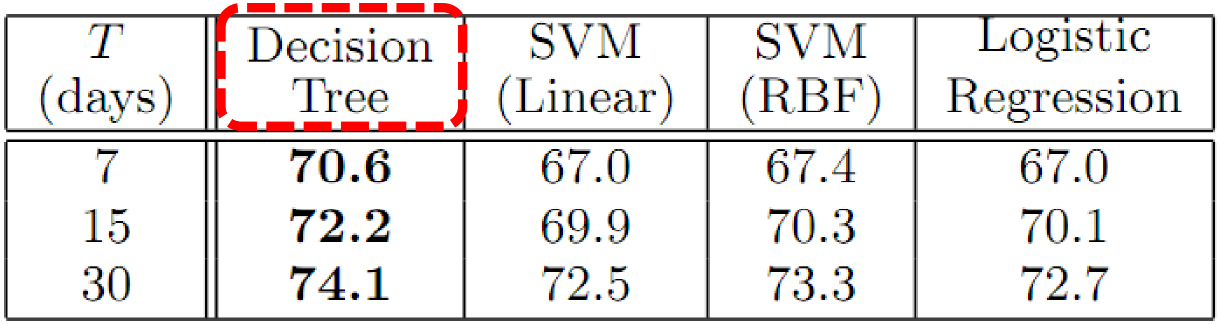
Prediction results

Task 1



k (posts)	Decision Tree	SVM (Linear)	SVM (RBF)	Logistic Regression
1	**72.6**	60.9	61.2	61.1
2	**67.1**	58.6	59.4	58.7
3	**64.4**	59.5	60.2	59.5
4	**65.0**	60.6	61.2	60.7
5	**65.2**	62.4	63.1	62.7
16	**69.4**	68.5	69.0	69.3
17	**69.7**	68.9	68.9	69.4
18	70.3	69.7	**70.4**	70.3
19	69.3	69.2	69.2	**69.6**
20	**71.2**	69.7	69.9	70.1

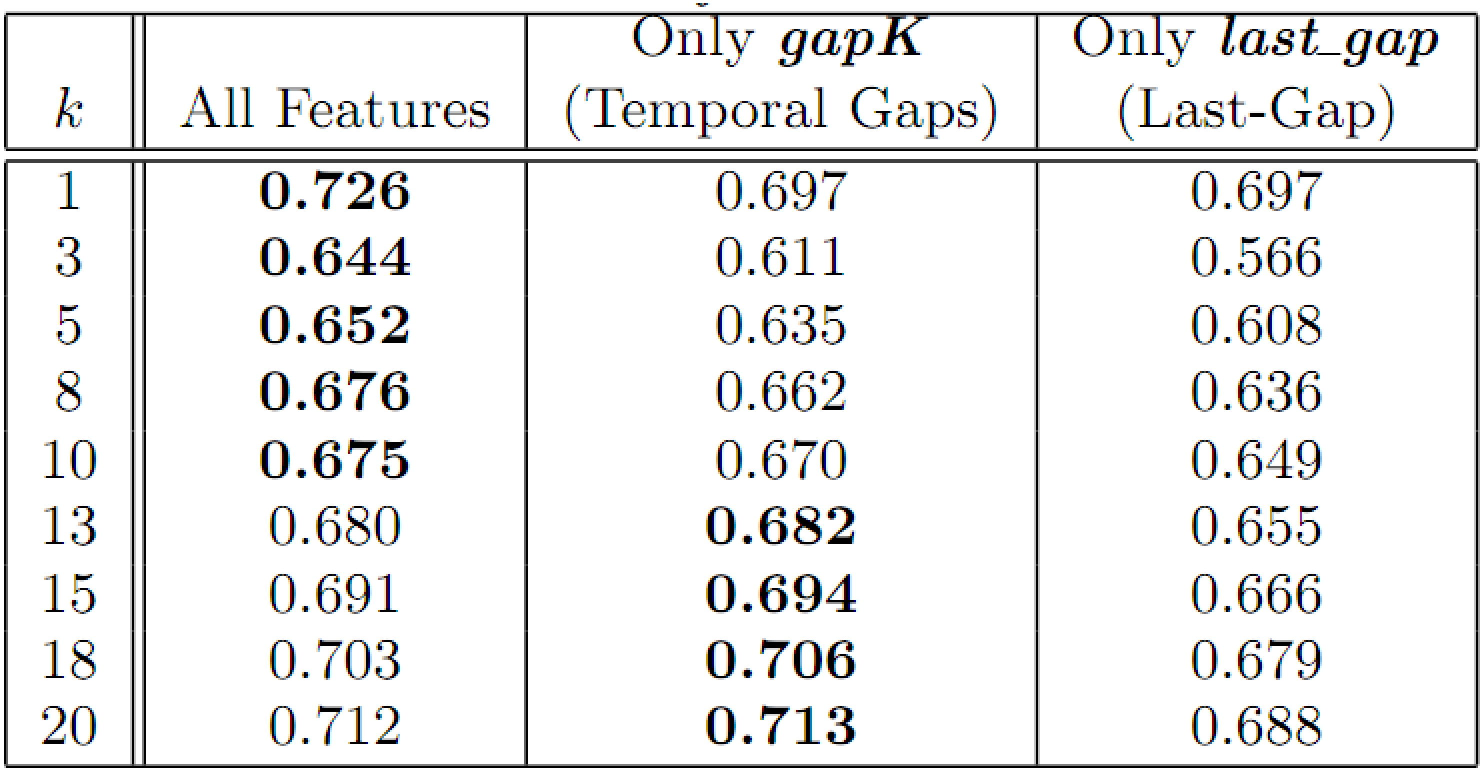
Task 2



T (days)	Decision Tree	SVM (Linear)	SVM (RBF)	Logistic Regression
7	**70.6**	67.0	67.4	67.0
15	**72.2**	69.9	70.3	70.1
30	**74.1**	72.5	73.3	72.7

Prediction analysis

- Recall most significant signal: temporal gaps



k	All Features	Only *gapK* (Temporal Gaps)	Only *last_gap* (Last-Gap)
1	**0.726**	0.697	0.697
3	**0.644**	0.611	0.566
5	**0.652**	0.635	0.608
8	**0.676**	0.662	0.636
10	**0.675**	0.670	0.649
13	0.680	**0.682**	0.655
15	0.691	**0.694**	0.666
18	0.703	**0.706**	0.679
20	0.712	**0.713**	0.688

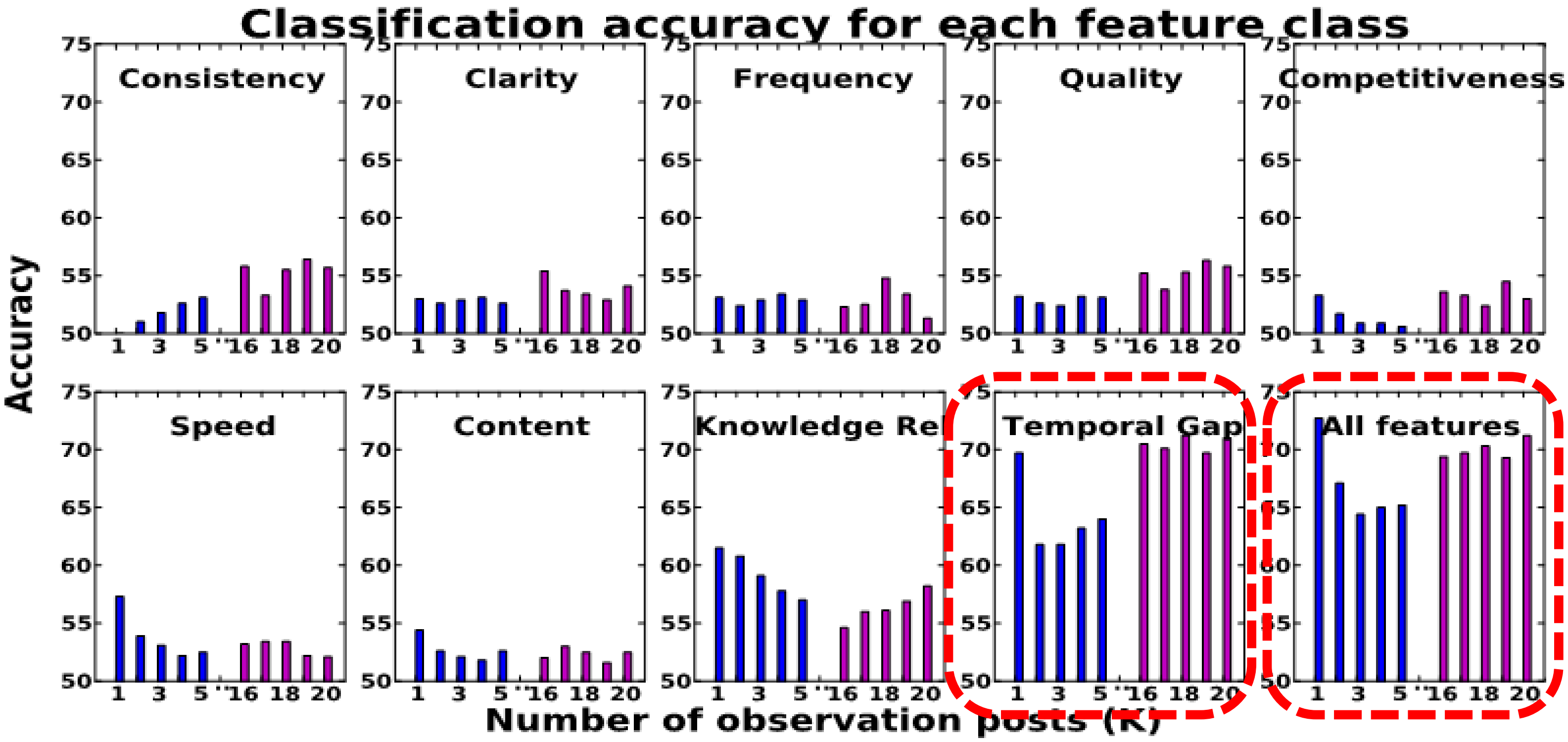
≥ 30 features

K-1 features

1 (!) feature

Prediction analysis

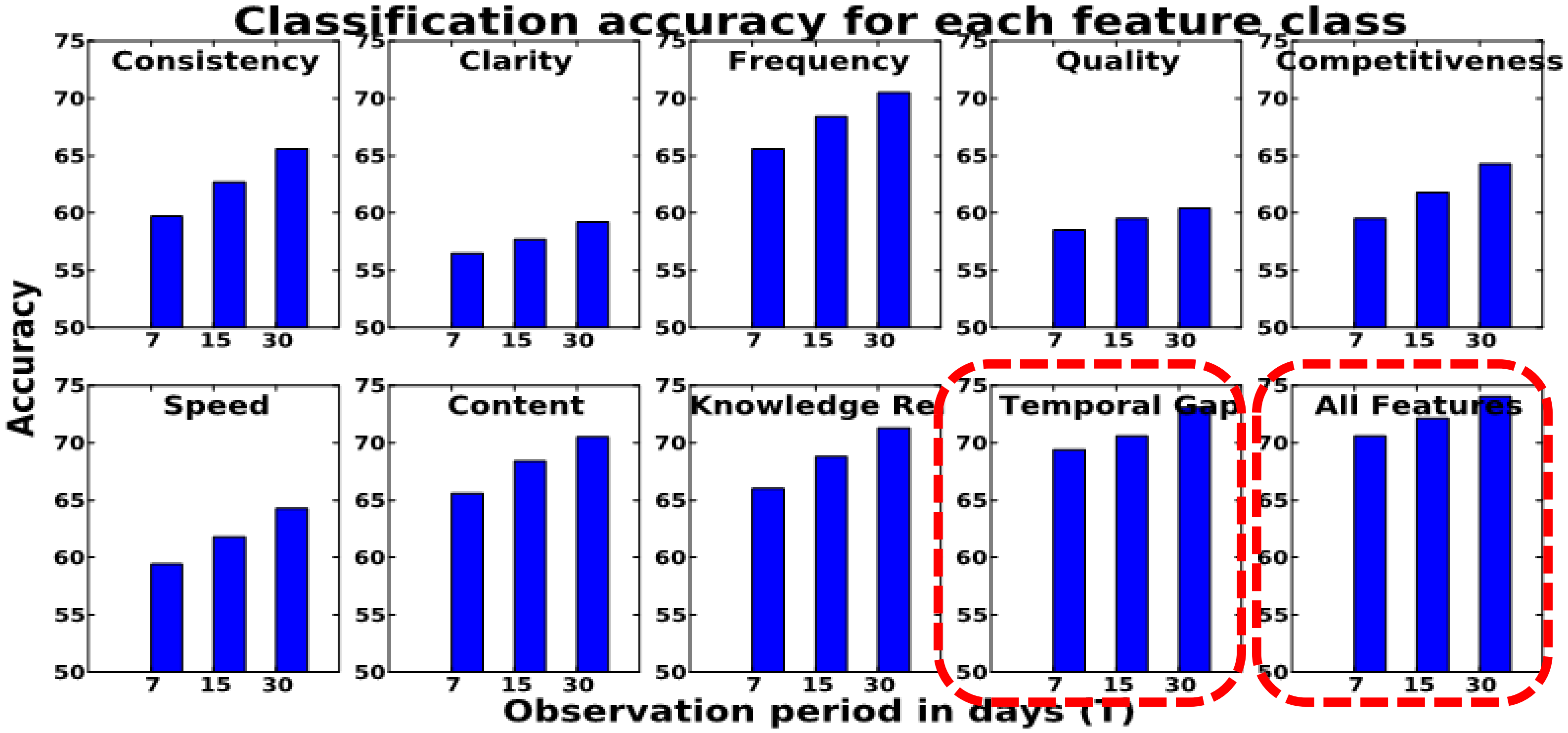
- Recall most significant signal: temporal gaps



Churn prediction accuracy with features from each category in isolation, for varying K (Task 1)

Prediction analysis

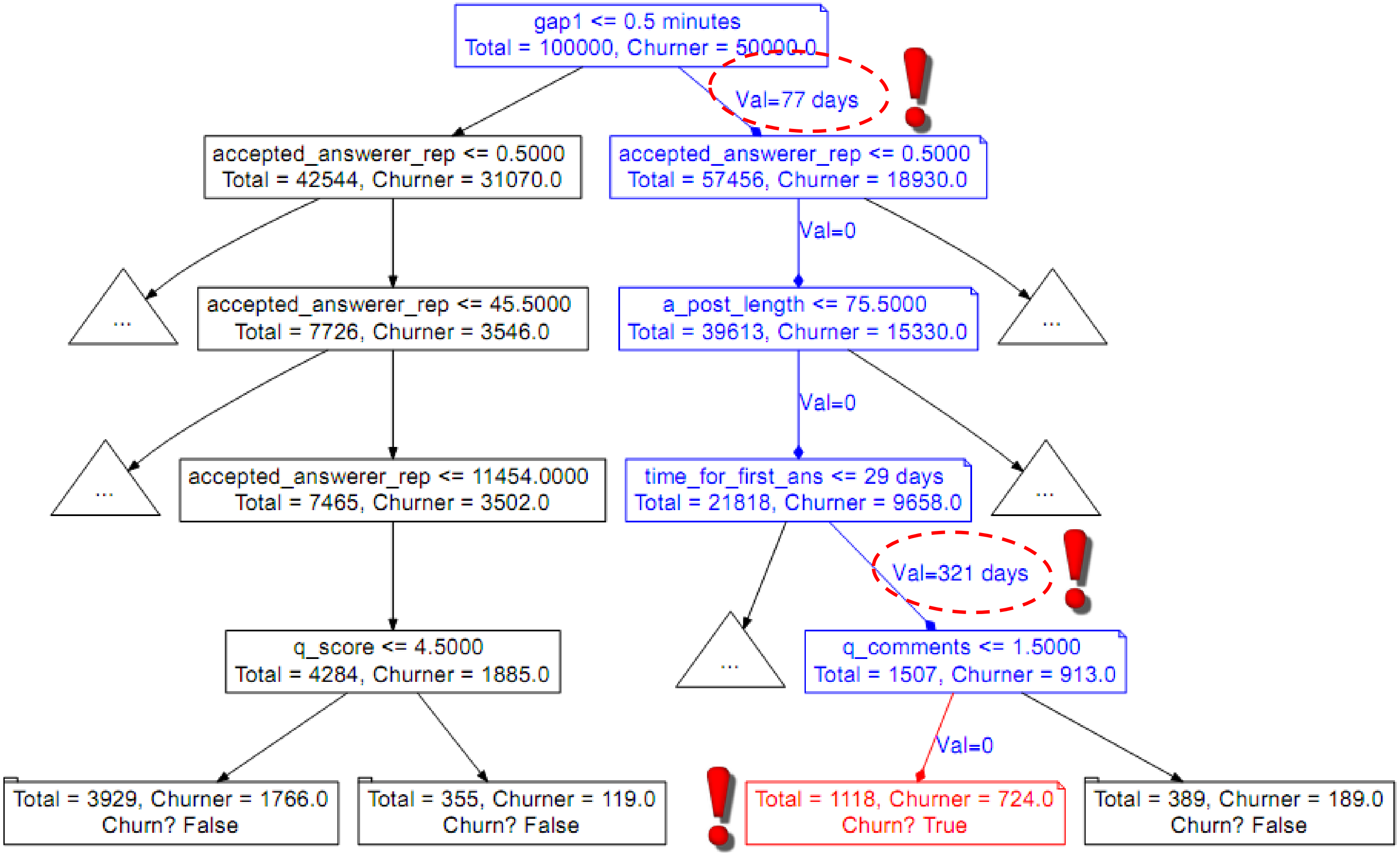
- Recall most significant signal: temporal gaps



Churn prediction accuracy with features from each category in isolation, for varying T (Task 2)

Use-Case: churn analysis of a user

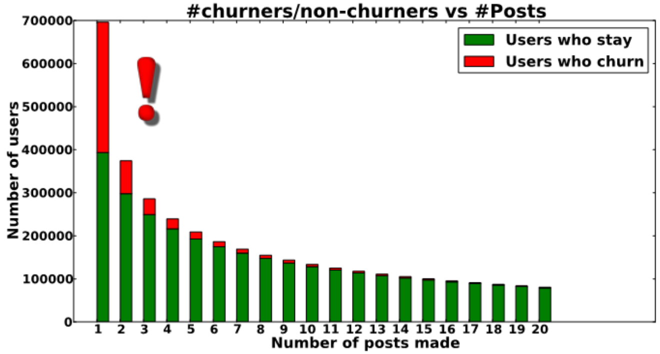
- Learned models (trees) help characterizing:

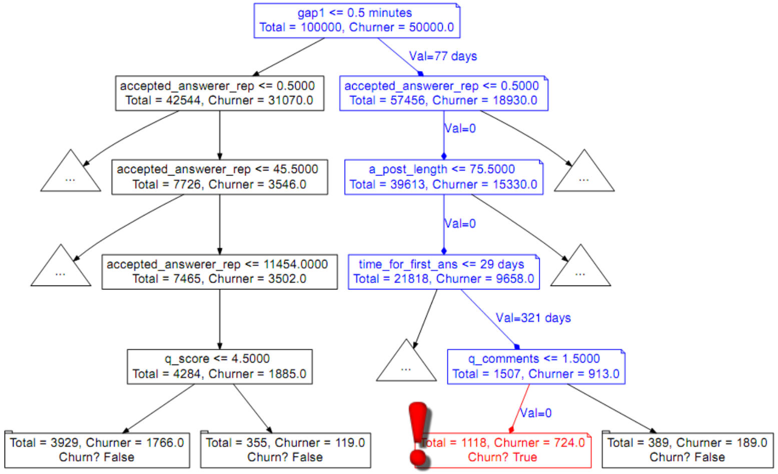

gap1 <= 0.5 minutes
Total = 100000, Churner = 50000.0
Val=77 days
accepted_answerer_rep <= 0.5000
Total = 42544, Churner = 31070.0
accepted_answerer_rep <= 0.5000
Total = 57456, Churner = 18930.0
Val=0
...
accepted_answerer_rep <= 45.5000
Total = 7726, Churner = 3546.0
a_post_length <= 75.5000
Total = 39613, Churner = 15330.0
...
Val=0
...
accepted_answerer_rep <= 11454.0000
Total = 7465, Churner = 3502.0
time_for_first_ans <= 29 days
Total = 21818, Churner = 9658.0
...
Val=321 days
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

Summary

- Study of user churn in Q&A sites
- Associated/potential factors
- 9 groups of features
- Best signal: trend in gap change (growth!)
- Prediction & characterizing by decision trees

Thank you!





gap1 <= 0.5 minutes
Total = 100000, Churner = 50000.0

Val=77 days

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 <= 29 days
Total = 21818, Churner = 9658.0

Val=321 days

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