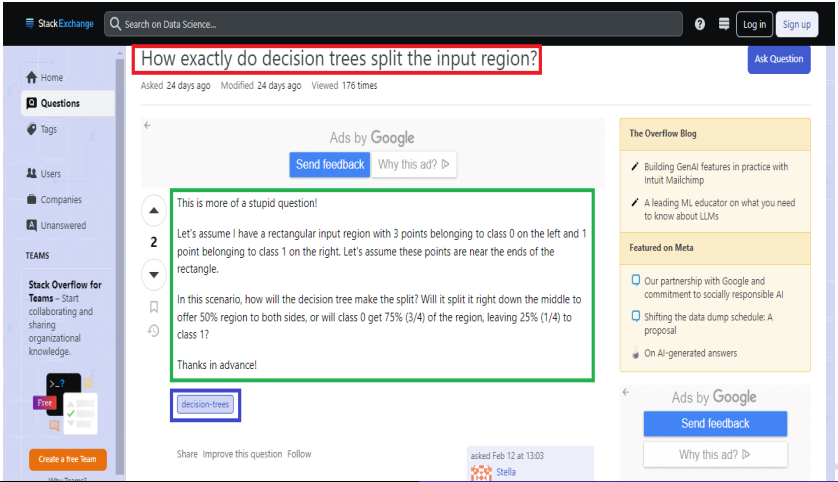
Stacking Your Odds on the Stack Exchange

Ethan Alwaise

April 12, 2024

The Stack Exchange

A network of Q&A websites for various fields



The Problem

1. Question:

   Can we predict if a given question will be answered within a week?

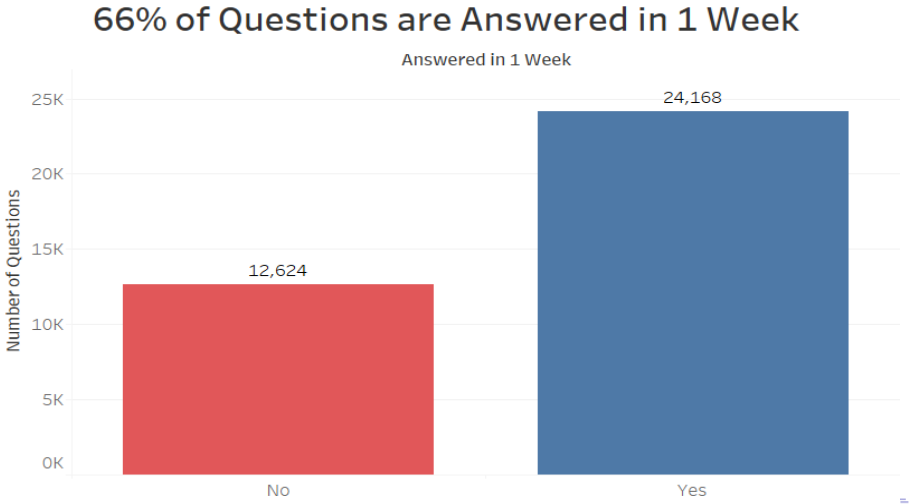
2. Who should care?

   Users who want to optimize their questions.

The Data

	PostTypeId	CreationDate	Score	ViewCount	Body	LastActivityDate	Title	Tags	AnswerCount	CommentCount	LastEditDate
0	1	2014-05-13T23:58:30.457	9	959.0	<p>I've always been interested in machine lear...	2014-05-14T00:36:31.077	How can I do simple machine learning without h...	<machine-learning>	1.0	1	None
1	1	2014-05-14T00:11:06.457	4	503.0	<p>As a researcher and instructor, I'm looking...	2014-05-16T13:45:00.237	What open-source books (or other materials) pr...	<education> <open-source>	3.0	4	2014-05-16T13:45:00.237
2	2	2014-05-14T00:36:31.077	5	NaN	<p>Not sure if this fits the scope of this SE,...	2014-05-14T00:36:31.077	None	None	NaN	0	None
3	2	2014-05-14T00:53:43.273	13	NaN	<p>One book that's freely available is "The El...	2014-05-14T00:53:43.273	None	None	NaN	1	None
4	1	2014-05-14T01:25:59.677	26	1925.0	<p>I am sure data science as will be discussed...	2020-08-16T13:01:33.543	Is Data Science the Same as Data Mining?	<data-mining> <definitions>	4.0	1	2014-06-17T16:17:20.473

EDA


66% of Questions are Answered in 1 Week
Answered in 1 Week
Number of Questions
25K
20K
15K
10K
5K
0K
12,624
24,168
No
Yes

Answer Prediction (Random Forest)

Feature Engineering and Data Processing

Steps:

- Count daily post numbers at question time.
- Collect question poster data (number of prior questions, etc.)
- Count math equations, lines of code, etc. in questions.
- Create dummy variables for question tags.

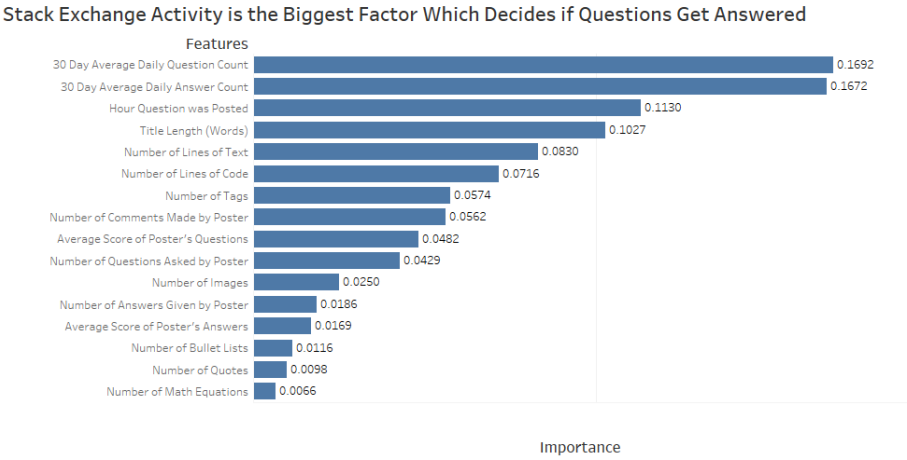
Answer Prediction (Random Forest)

Model Performance

		Precision	Recall	F1-score
TRAIN	0 (Unanswered)	1.00	1.00	1.00
	1 (Answered)	1.00	1.00	1.00
	Accuracy			1.00
TEST	0 (Unanswered)	0.60	0.29	0.39
	1 (Answered)	0.71	0.90	0.79
	Accuracy			0.69

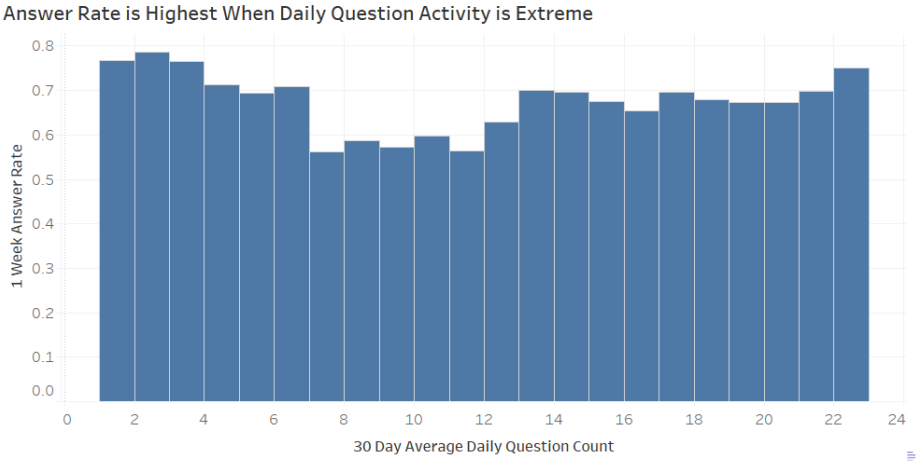
Answer Prediction (Random Forest)

Interpretation



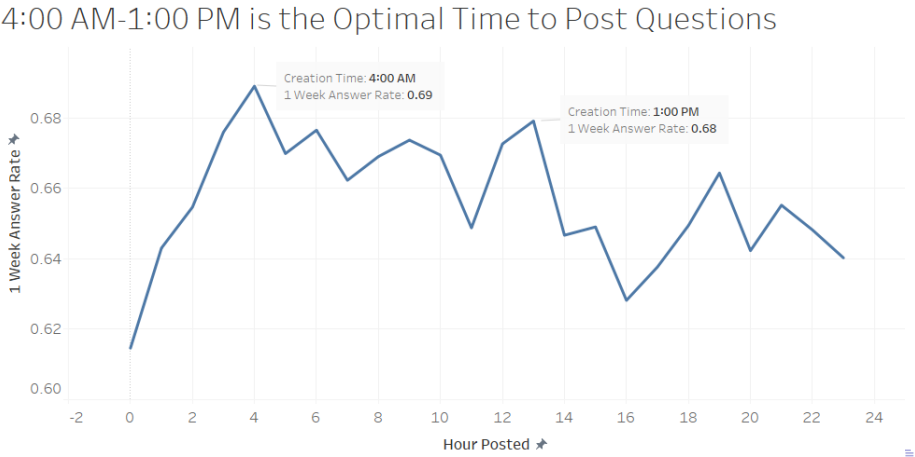
Answer Prediction (Random Forest)

Interpretation



Answer Prediction (Random Forest)

Interpretation



A Secondary Question

Post Activity Prediction

Since response rate depends on post activity, let's investigate another question:

Can we predict the number of daily posts on a given future day?

Daily Post Activity Prediction (SARIMA)

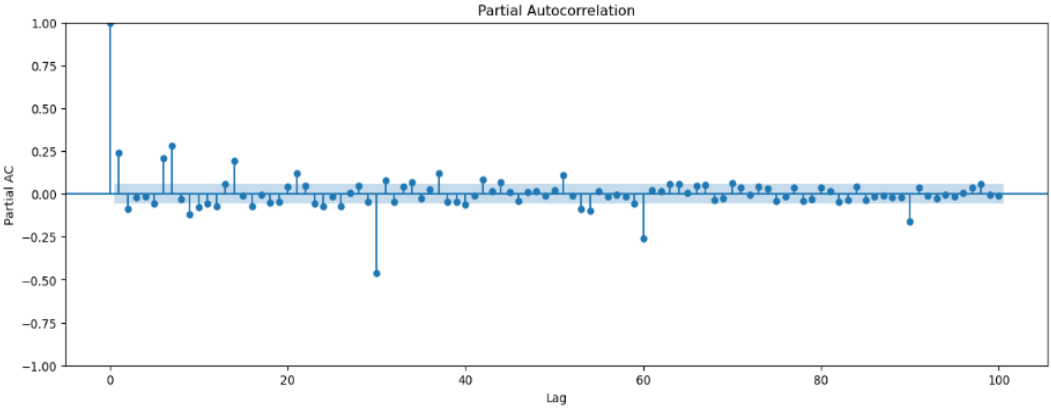
EDA and Hyperparameter Selection

Modeling steps:

- Count number of posts per day and observe trends.
- Determine order of differencing using unit root testing.
- Look at partial autocorrelations to determine autoregressive order.

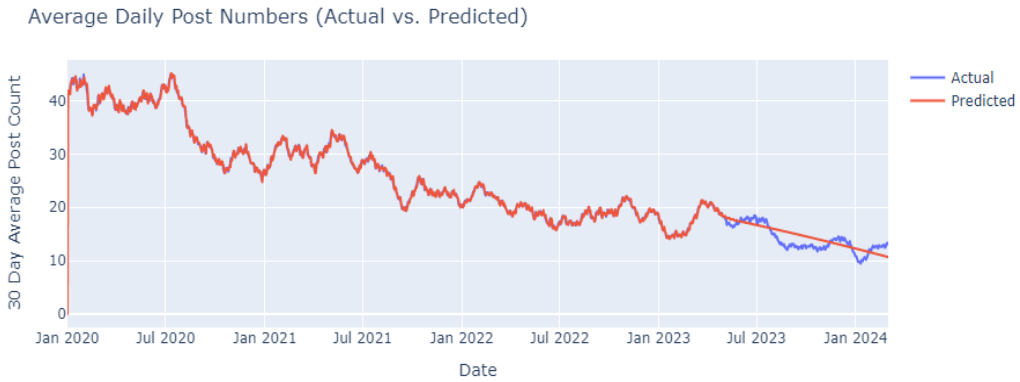
Daily Post Activity Prediction (SARIMA)

PACF Plot



Daily Post Activity Prediction (SARIMA)

Model Predictions



Daily Post Activity Prediction (SARIMA)

Model Performance

	MAPE	RMSE
TRAIN	1.0%	1.23
TEST	10.0%	1.53

Conclusions

Although predicting if a question will be answered is difficult, it appears that Stack Exchange activity is the primary factor in question response odds. The time of posting as well as the length of the title and body also seem important.

Unfortunately, the Stack Exchange on the decline. This could be due to widespread use of ChatGPT.