

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 screenshot shows a Stack Overflow question page. The title "How exactly do decision trees split the input region?" is highlighted with a red border. The question is asked 24 days ago, modified 24 days ago, and viewed 176 times. The question text is: "Let's assume I have a rectangular input region with 3 points belonging to class 0 on the left and 1 point belonging to class 1 on the right. Let's assume these points are near the ends of the rectangle. In this scenario, how will the decision tree make the split? Will it split it right down the middle to offer 50% region to both sides, or will class 0 get 75% (3/4) of the region, leaving 25% (1/4) to class 1? Thanks in advance!". The question is tagged with "decision-trees". The page includes a sidebar with navigation links (Home, Questions, Tags, Users, Companies, Unanswered, TEAMS), an advertisement for Google, and a section titled "The Overflow Blog" with links to articles about GenAI features and ML education. The bottom of the page shows the user's name "Stella" and the time "asked Feb 12 at 13:03".

StackExchange Search on Data Science... Log in Sign up

How exactly do decision trees split the input region?

Asked 24 days ago Modified 24 days ago Viewed 176 times

← Ads by Google Send feedback Why this ad? ▷

2

This is more of a stupid question!

Let's assume I have a rectangular input region with 3 points belonging to class 0 on the left and 1 point belonging to class 1 on the right. Let's assume these points are near the ends of the rectangle.

In this scenario, how will the decision tree make the split? Will it split it right down the middle to offer 50% region to both sides, or will class 0 get 75% (3/4) of the region, leaving 25% (1/4) to class 1?

Thanks in advance!

decision-trees

Share Improve this question Follow

asked Feb 12 at 13:03 Stella

← Ads by Google Send feedback Why this ad? ▷

The Overflow Blog

- Building GenAI features in practice with Intuit Mailchimp
- A leading ML educator on what you need to know about LLMs

Featured on Meta

- Our partnership with Google and commitment to socially responsible AI
- Shifting the data dump schedule: A proposal
- On AI-generated answers

Free Create a free Team

The Problem

❶ Question:

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

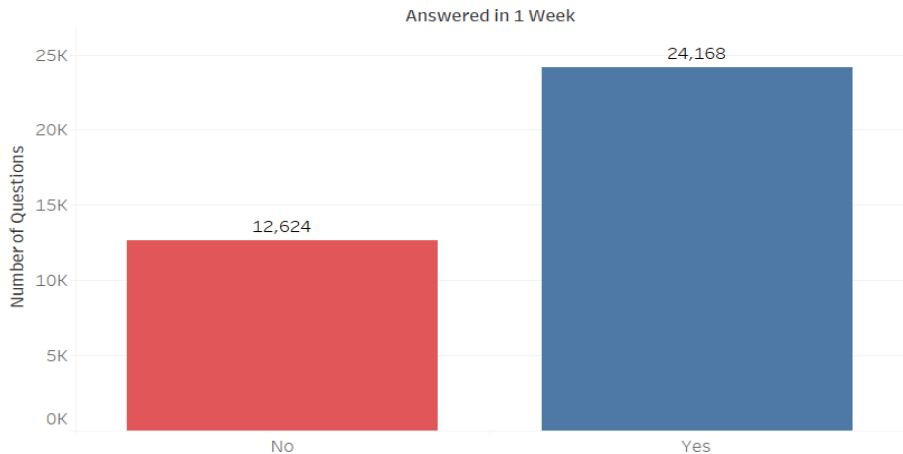
❷ Who should care?

Users who want to optimize their questions.

The Data

| | PostType | id | 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 |

66% of Questions are Answered in 1 Week



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)

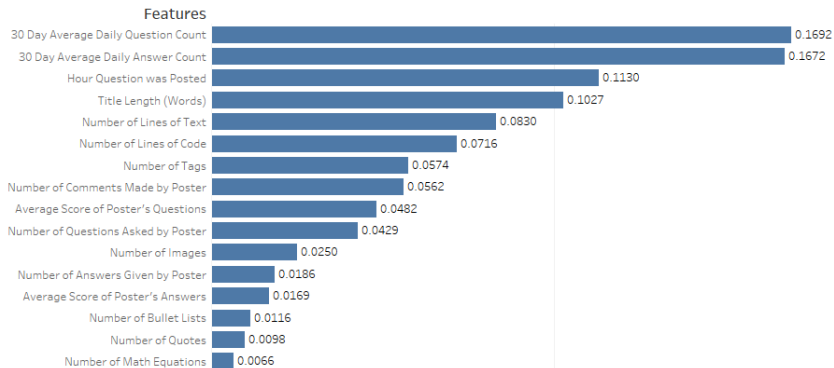
Model Performance

| | | 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

Stack Exchange Activity is the Biggest Factor Which Decides if Questions Get Answered

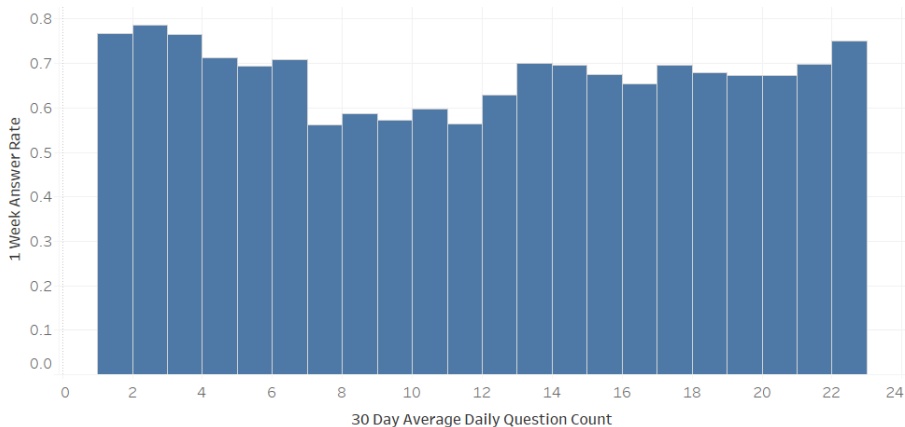


Importance

Answer Prediction (Random Forest)

Interpretation

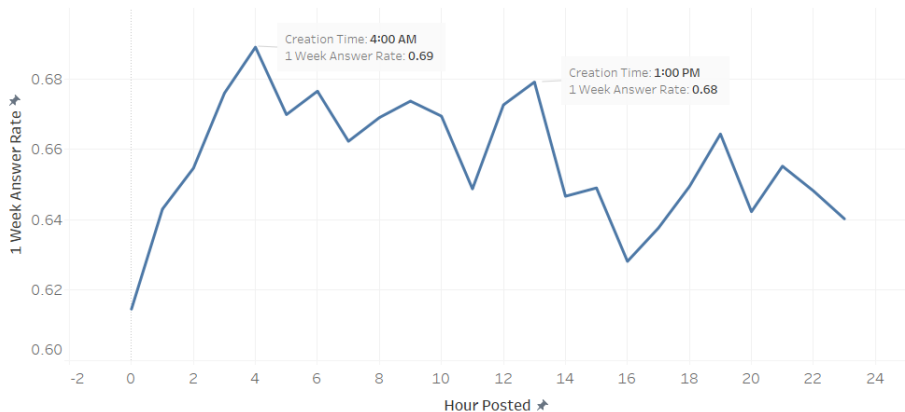
Answer Rate is Highest When Daily Question Activity is Extreme



Answer Prediction (Random Forest)

Interpretation

4:00 AM-1:00 PM is the Optimal Time to Post Questions



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)

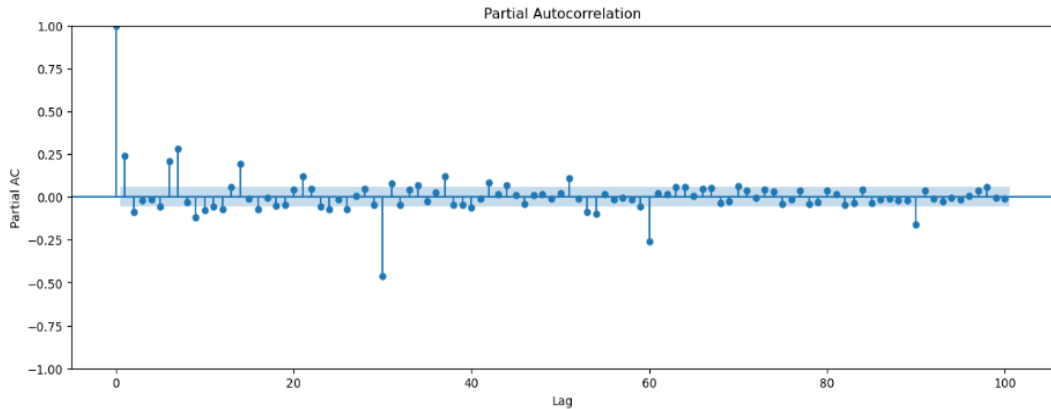
EDA and Hyperparameter Selection

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

Average Daily Post Numbers (Actual vs. Predicted)



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 is on the decline. This could be due to widespread use of ChatGPT.