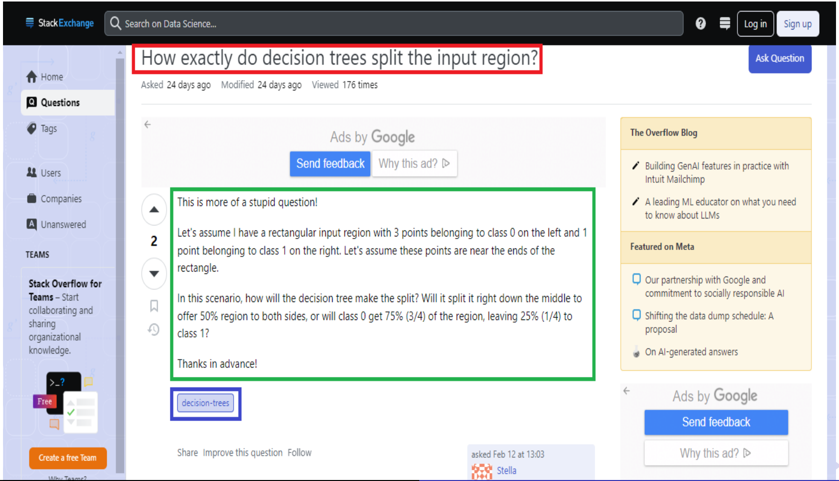
BrainStation Capstone Sprint 2

Ethan Alwaise

April 2, 2024

The Stack Exchange

A network of Q&A websites for various fields



How exactly do decision trees split the input region?

Asked 24 days ago Modified 24 days ago Viewed 176 times

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!

asked Feb 12 at 13:03
Stella

Stack Overflow for Teams – Start collaborating and sharing organizational knowledge.

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

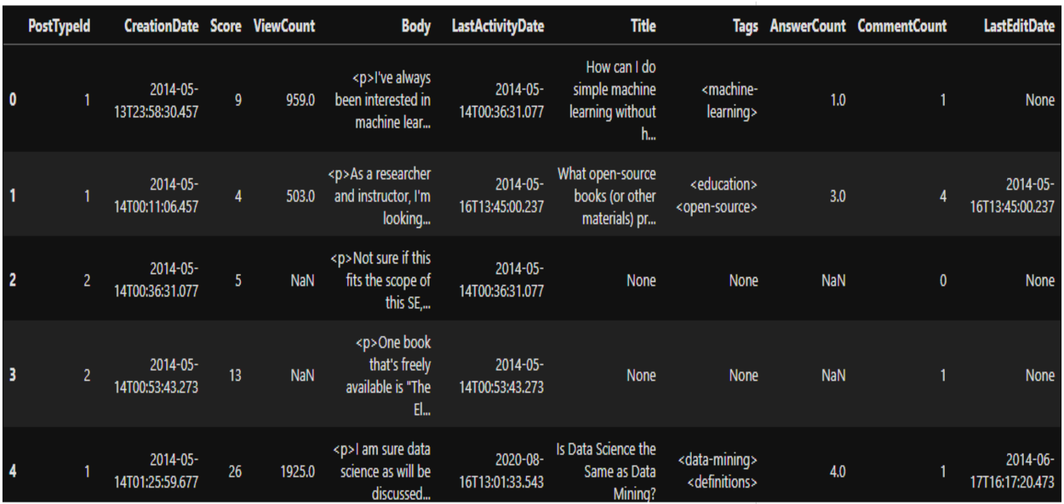
The Problem

1. Questions:
    - What factors affect the likelihood that a question is answered?
    - What patterns can we observe in posting activity over time?
2. Who should care?
    - Users who want to optimize their questions.
    - Stakeholders in the Stack Exchange.

The Problem

- Can we predict if a given question will be answered within a week?
- Can we predict the number of posts made on a given day?

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

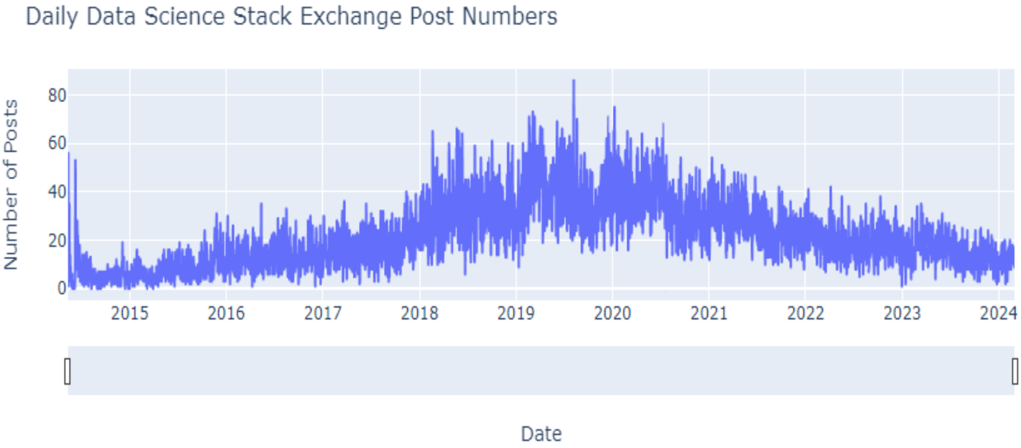
EDA (Answer Prediction)

Class balance:

$\sim$ 65% of questions are answered within 7 days.

EDA (Daily Post Activity)

We focus on 2020 onward since we observe two opposite trends:



Answer Prediction (Logistic Regression)

Feature Engineering and Data Processing

Steps:

- Count math equations, lines of code, etc. in questions.
- Create dummy variables for question tags.
- Vectorize the text (bag of words).

Answer Prediction (Logistic Regression)

Model Performance

		Precision	Recall	F1-score
TRAIN	0 (Unanswered)	0.71	0.39	0.50
	1 (Answered)	0.74	0.92	0.82
	Accuracy			0.74
TEST	0 (Unanswered)	0.49	0.19	0.27
	1 (Answered)	0.67	0.90	0.77
	Accuracy			0.65

Answer Prediction (Logistic Regression)

Next Steps

- Improve tokenization (include non-alphabetic characters)
- More sophisticated text processing.
- Try dimensionality reduction and hyperparameter optimization.
- Experiment with other types of models (random forest, neural nets, etc.).

Daily Post Activity Prediction (SARIMA)

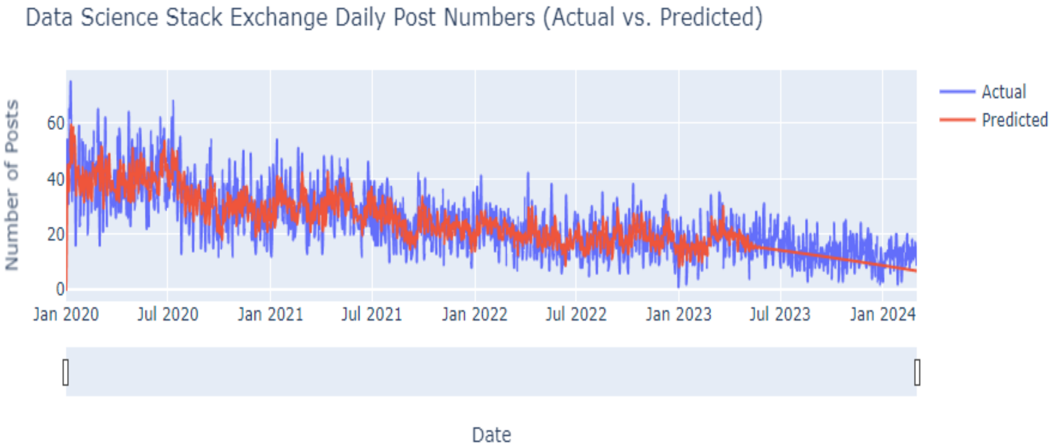
EDA and Hyperparameter Selection

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)

Model Predictions



Daily Post Activity Prediction (SARIMA)

Model Performance

	MAPE	RMSE
TRAIN	30.0%	7.96
TEST	40.0%	5.72

Daily Post Activity Prediction (SARIMA)

Next Steps

- Tune hyperparameters.
- Incorporate seasonality.
- Experiment with rolling averages / monthly numbers.