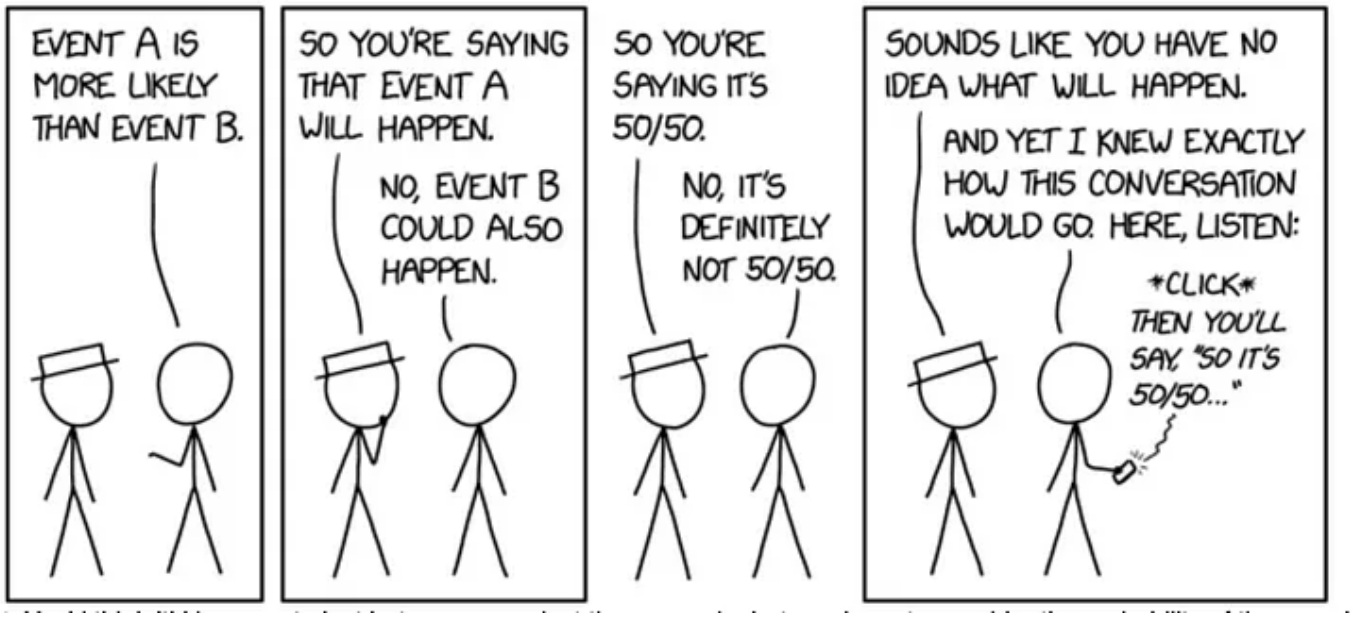
Probability calibration : why it matters ?

Jaideep Ray · Follow

Published in Better ML · 3 min read · Jul 10, 2021

Why returning probability instead of prediction from your classification model is a good idea ?



https://www.explainxkcd.com/wiki/index.php/2370:_Prediction

1. **It is easier to reason around probabilities.**

A typical exercise in evaluating classifiers is going through a sample of false positives and false negatives and understanding classifier characteristics. Probabilities are easier to reason than values. As you keep on iterating on the classifier, this is a major win.

2. Picking thresholds

In typical binary classification scenario we often use classifier outputs with thresholds,
if pred > threshold, the predicted class is X else Y

If predictions are not probabilities, you have to reason through threshold values every time you update classifier.

3. Updating models in multi stage ranking.

In production we often see chain or cascade of classifiers (model outputs are used as input for other models) working together.

Even though this is not a great practice, this pattern finds its way into many important products. We discuss this multi-stage-ranking. It is important that each stage in multi stage ranking is emitting probabilities and not just predictions.

Let's say we have a 2 stage system, classifier A feeds to classifier B. Every iteration on classifier A changes its output distribution somewhat. If the outputs are not calibrated then it is hard to estimate the impact on classifier B and every iteration on A would require proper A/B experimentation with entire system retraining.

OTOH, with calibrated probabilities we can estimate the impact and retrain if the distributions change too much.

4. How to check if your classifier outputs are calibrated ?

Calibration plots

On the y axis plot fraction of positives (true label) and on the x axis plot bins that group examples by predicted score.

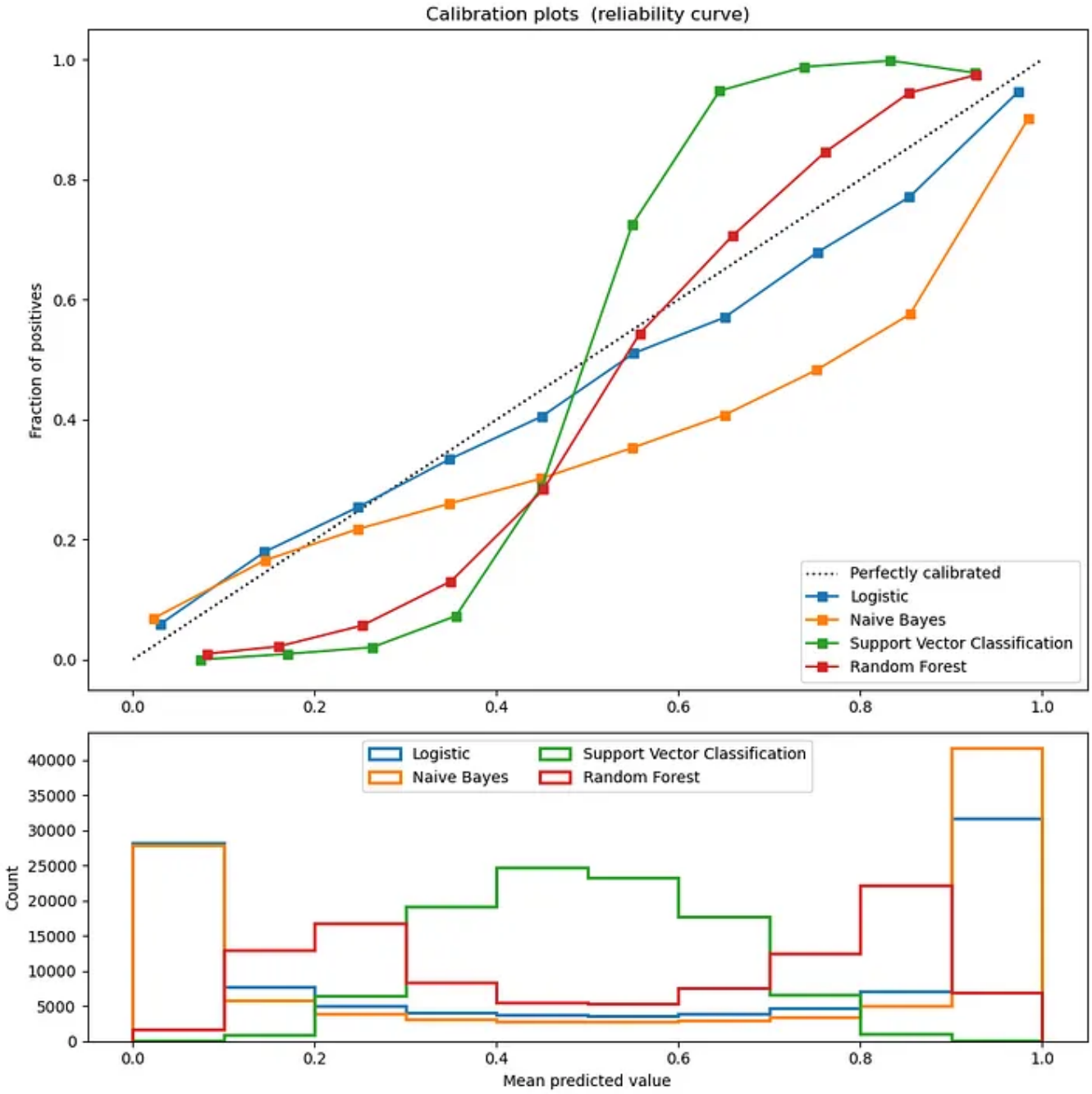
So, the y axis point for the interval [0.0–0.2] signifies :

n = number of examples for which prediction falls in [0.0–0.2] range.

n_p = number of examples for which true label is 1.

y_axis value = n_p / n

The diagonal represents perfectly calibrated scores. In this diagram, Logistic regression is better calibrated than other classifiers such as Random Forest.



Source : comparison of calibration curves

See point (6) for metrics for calibration.

5. How to calibrate your classifier ?

- **Collect calibration dataset.** This should be a held out dataset matching production traffic. Split this is calibration_train & calibration_test.
- Use platt scaling or isotonic regression to train a calibration your classifier score.
- Make sure inference stack can stitch the calibrated model appropriately.

Before calibration :

y = model(input)

6. Metrics for calibration:

Brier score loss : The Brier score measures the mean squared difference between the predicted probability and the actual outcome. The Brier score always takes on a value between zero and one, since this is the largest possible difference between a predicted probability (which must be between zero and one) and the actual outcome (which can take on values of only 0 and 1). The lower, the better.

Recap :

- Output well calibrated probabilities from classifier wherever possible.

Machine Learning | Machine Learning Ai | Automl

Written by Jaideep Ray

Follow

225 Followers · Editor for Better ML

Engineer | ML Platforms | Model lifecycle| https://www.linkedin.com/in/jaideepray/

More from Jaideep Ray and Better ML

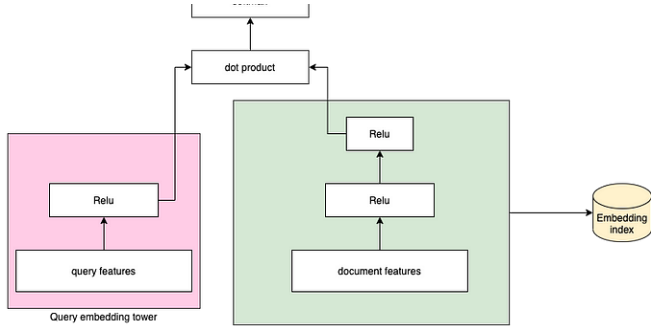
Jaideep Ray in Better ML

Where did all my memory go ?

LLM finetuning version.

4 min read · May 28

11



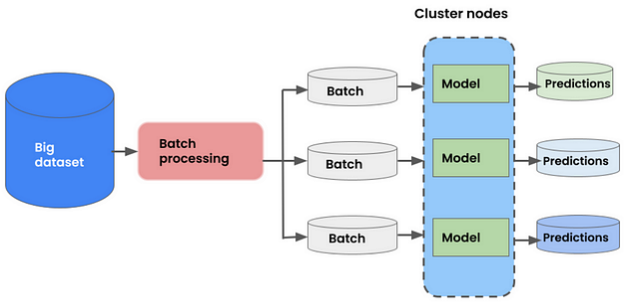
Jaideep Ray in Better ML

Embedding learning for retrieval

TLDR :

5 min read · Jul 13, 2021

44



Jaideep Ray in Better ML

Offline Batch Inference for large models

Batching up for profit!

$$t_{math} > t_{mem}$$

$$ops/BW_{math} > bytes/BW_{mem}$$

$$ops/bytes > BW_{math}/BW_{mem}$$

Jaideep Ray in Better ML

Arithmetic Intensity : Understand Op limits—Memory or Compute

Why ?

3 min read · Aug 13

52

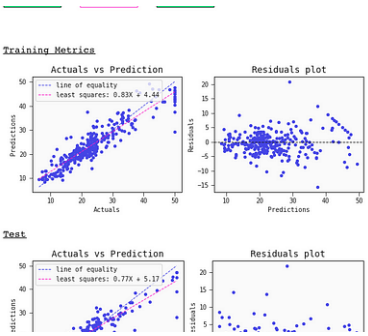
3 min read · Dec 31, 2021

14

See all from Jaideep Ray

See all from Better ML

Recommended from Medium



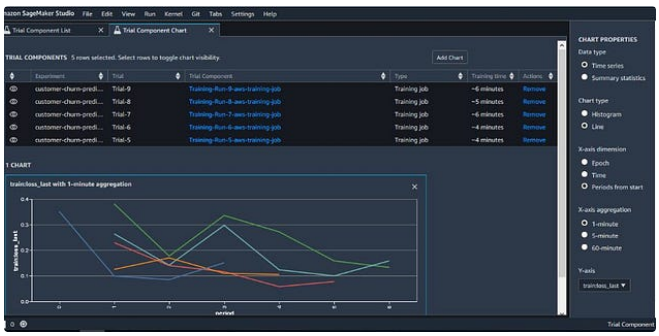
Casper Skern Wilstrup

Symbolic Regression: a Simple and Friendly Introduction

Symbolic Regression is like a treasure hunt for the perfect mathematical equation to...

3 min read · May 5

18 1



Yugank Aman

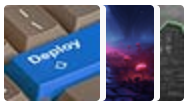
Top MLOps Tools to Manage Machine Learning Lifecycle

Businesses continue transforming their operations to increase productivity and...

10 min read · May 30

94

Lists



Predictive Modeling w/ Python

20 stories · 473 saves

Practical Guides to Machine Learning

10 stories · 544 saves

Natural Language Processing

689 stories · 304 saves

The New Chatbots: ChatGPT, Bard, and Beyond

13 stories · 137 saves

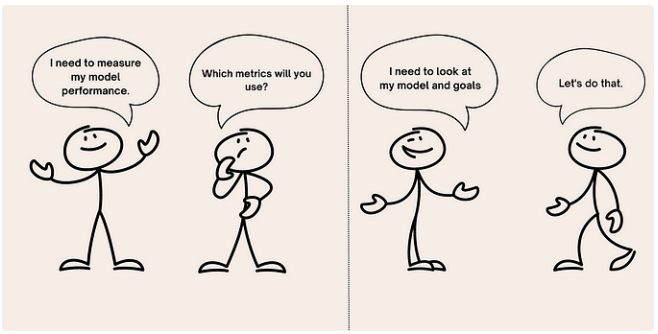
Chandra Prakash Bathula

Machine Learning Concept 68: Platt's Scaling

Platt's Scaling:

3 min read · Apr 13

5



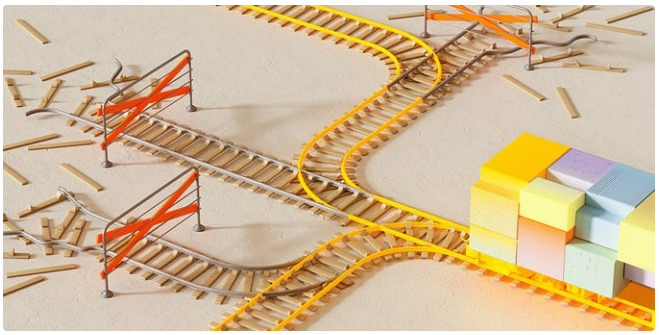
Hima

9. Model Metrics That Matter / AI Product Management

Howdy friends! So far we have discussed how to choose the right problem to solve using A...

8 min read · Jul 20

Biman Chakraborty



Andrew Blance in Better Programming

Two-Sample t Tests, Power, Effect Size and Sample Size Calculator i...

I was going over my daily dose of coffee while redaing the newspaper in the morning. A...

13 min read · Apr 30

8

MLOps and MLflops

How to understand MLOps architectures and diagrams

9 min read · Apr 3

496 3

See more recommendations