

★ Member-only story

Probability Calibration



Tapan Kumar Patro · Following

Published in DataDrivenInvestor · 6 min read · Feb 12



12



Building a model with risk appetite, don't miss this step!



Hi Guys Tapan here, I am back with a blog about **Probability calibration**. If you never heard of it or if you have an idea about this, this blog definitely can give you simpler yet informative details about this topic. So let's get started.

Introduction to Probability calibration



Photo by [Anne Nygård](#) on [Unsplash](#)

The term “probability calibration” describes how accurately a machine learning model predicts probabilities. It assesses how closely the probabilities predicted match the outcomes. simple r8 !

Consider a model that indicates the likelihood that an event will occur, such as the likelihood that it will rain tomorrow. Tomorrow's rain is predicted to have a probability of 0.8 by a properly calibrated model, but it actually rains 80% of the time. On the other hand, a model that is not calibrated might

forecast a 0.8 probability of rain tomorrow, but only 30% of the time it actually does.

Probability calibration put simply, is a method of determining whether the probabilities predicted by a machine learning model match the outcomes of actual experiments. Predictions from a model that has been properly calibrated are more accurate and trustworthy.

In simple terms, you can put it as the truthfulness of a model.

The problem of uncalibrated probabilities

Uncalibrated probabilities describe a situation in which a machine learning model's predictions of probabilities do not accurately reflect the actual probabilities of the events they are projecting. This may result in the following issues:

Overconfidence: The model may make predictions with excessively high probabilities, a sign that it is overconfident in those predictions. This may cause decisions based on the predictions to be made incorrectly.

Underconfidence: The model may make predictions with too low of probabilities, which shows a lack of confidence in those predictions. This may lead to missed opportunities or bad choices.

Inaccurate risk assessment: The model's forecasts might not accurately reflect the risks connected to different events, which can result in inaccurate risk assessments and poorly informed choices.

Performance inconsistency: Depending on the particular data the model is using, performance can vary significantly.

Bias: If the training data is biased, uncalibrated probabilities may produce biased predictions. Discrimination and unfair results may result from this.

Calibration metrics

A machine learning model's probabilities predictions are evaluated using probability calibration metrics. Several frequently employed probabilistic calibration metrics are:

A calibration curve compares the observed percentages of successful outcomes to the typically predicted probabilities for each bin. The calibration curve of a well-calibrated model will be relatively close to the line of perfect calibration, which represents the ideal situation where the predicted probabilities exactly match the outcomes that were observed. The **Brier Score** is a measurement of the mean squared error between the probabilities that were predicted and the outcomes that actually occurred. Better calibration is indicated by a lower Brier Score.

Reliability Diagram: The distribution of the predicted probabilities is depicted visually by a reliability diagram, which is a representation of the calibration curve.

Logarithmic Loss: A measure of the average negative log-likelihood of the predicted probabilities is called logarithmic loss. In multiclass classification problems where the predicted probabilities must range from 0 to 1, it is a frequently employed metric.

AUC curve: The overall effectiveness of a binary classification model is measured by the AUC-ROC, which disregards the probabilities' calibration.

Expected Calibration Error: The ECE is a measurement of the discrepancy between the observed frequencies for each bin and the average predicted probabilities. It offers a scalar value that represents the total amount of calibration error.

Techniques for improving calibration

A machine learning model's calibration can be improved using a variety of methods, including:

By fitting a logistic regression model to the expected probabilities and the actual results, Platt Scaling is a post-processing technique that modifies a model's predicted probabilities.

Isotonic Regression is a non-parametric technique for calibrating predicted probabilities that modifies the predictions to make them monotonic with respect to the actual results.

With temperature scaling, which is a straightforward post-processing method for neural networks, the predicted probabilities are modified by multiplying them by a temperature hyperparameter.

Ensemble Methods: By combining the predictions of various base models, ensemble methods, such as bagging and boosting, can be used to enhance the calibration of a model.

Re-sampling: Class distributions in the training data can be balanced using re-sampling techniques like over- and under-sampling to enhance model calibration.

Cross-validation: This technique divides the data into multiple folds, trains the model on each fold, and then evaluates the model's performance on each fold. A model's predictions' stability and dependability can be evaluated using cross-validation.

Regularization: To avoid overfitting and enhance the generalization of a model, which can lead to improved calibration, regularization techniques like L1 and L2 regularization can be used.

Example of probability calibration in practice

Wherever a risk factor is associated with the model, probability calibration is needed.

The area of medical diagnosis serves as a practical illustration of probability calibration.

Based on a patient's symptoms, medical history, and test results, a machine-learning model may be trained to forecast the likelihood that the patient has a particular disease. A doctor can then use the predicted probabilities to make a diagnosis and choose the best course of treatment for the patient. It is crucial to calibrate the model so that the predicted probabilities accurately reflect the likelihood that the patient will have the disease in order to ensure that the probabilities are reliable. By assessing the model's calibration using probability calibration metrics like the Brier Score or the Calibration Curve, this can be accomplished.

There are other examples like in the fintech industry where a model identifies a member or a transaction as fraudulent it can be useful, or in the retail industry where the model is predicting customer churn calibration can be helpful.

Limitations and challenges



Photo by [Sylwia Bartyzel](#) on [Unsplash](#)

Probability calibration has a number of restrictions and difficulties:

Model complexity: Because deep neural networks have many parameters and non-linear relationships between inputs and outputs, complex machine learning models like these can be challenging to calibrate.

Data quality: The calibration of the model can be significantly impacted by the caliber of the data used to train the model. A model that is not calibrated can be produced by data that is unbalanced, noisy, or missing values.

Lack of ground truth: It can be challenging to obtain accurate and dependable ground truth labels for the data in many real-world applications.

Because of this, assessing the model's calibration and fixing any calibration errors may be challenging.

Overfitting: When a model fits the training data too closely, it is said to be overfitting. This can lead to an uncalibrated model and poor generalization performance.

Small sample size: Small sample sizes can lead to high variance in a model's predictions, making it challenging to precisely determine the model's calibration. **Limited evaluation metrics:** The evaluation metrics for evaluating the calibration of a machine learning model are limited, and they might not be appropriate for all applications.



Buy me a coffee

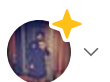
What Now?

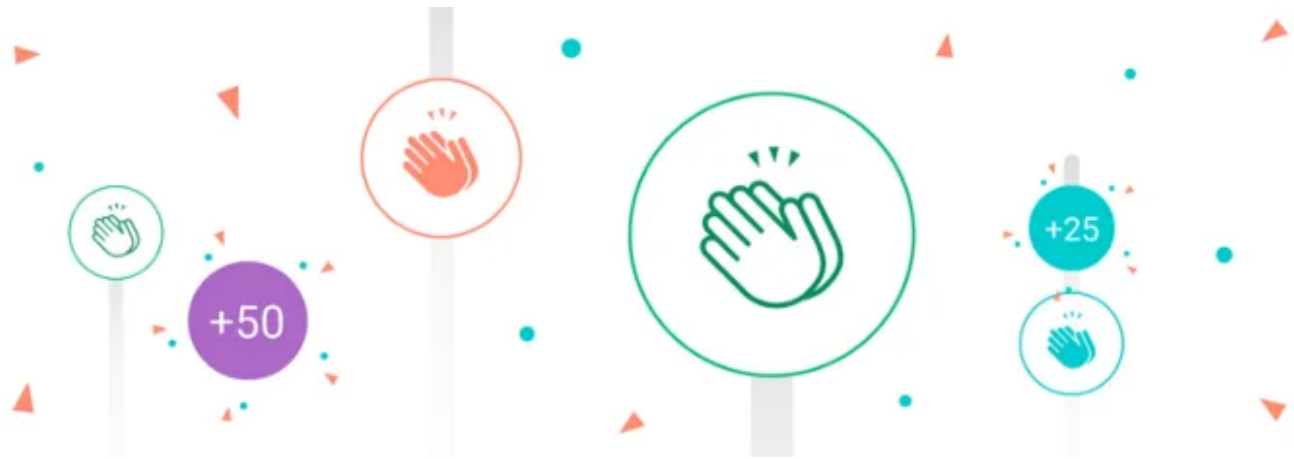
Open in app ↗



Search Medium

Write





If you like the article please make sure to give a clap. Please follow me for more projects and articles on my Github and my medium profile.

Lets Connect!

<https://www.linkedin.com/in/tapankpatro/>

Tapan Kumar Patro - Medium

Read writing from Tapan Kumar Patro on Medium. 📖 Machine learning | 🤖 Deep Learning | 👁 Computer vision | 🧠 Natural...

tapanpatro.medium.com

Subscribe to DDIntel [Here](#).

Visit our website here: <https://www.datadriveninvestor.com>

Join our network here: <https://datadriveninvestor.com/collaborate>

Statistics

Data Science

Risk Management

Machine Learning

More from the list: "ML"

Curated by Himanshu Birla



Kyosuke... in Towards Dat...

Probability Calibration for Imbalanced Dataset

★ · 8 min read · Oct 20, 2019



Mattia Ci... in Analytics Vi...

How Probability Calibration Works

★ · 6 min read · May 28, 2020



Jason Yo...

Why Calibrat the Series on

7 min read · Oc

[View list](#)

Written by Tapan Kumar Patro

451 Followers · Writer for DataDrivenInvestor

Following



📚 Machine learning | 🧠 Deep Learning | 📺 Computer vision | 🗣️ Natural Language processing | 🎧 Audio Data | 📄 End to End Software Development | ✍️

More from Tapan Kumar Patro and DataDrivenInvestor



Tapan Kumar Patro in Analytics Vidhya

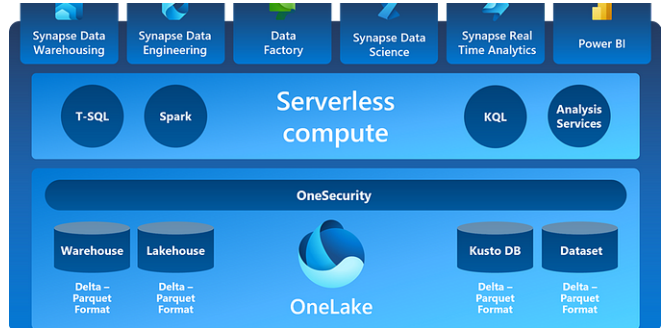
Random Cut Forest

The things you should know about this unsupervised machine learning algorithm.

4 min read · Feb 10, 2021



57



Gabe Araujo, M.Sc. in DataDrivenInvestor

Microsoft Fabric vs. Power BI: What's the Difference?

Microsoft Fabric vs. Power BI: Architecture, Capabilities, Data Governance, and Use Cases

★ · 10 min read · Aug 8



305

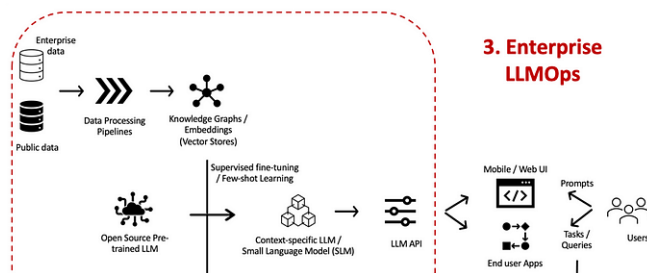


1



2. Enterprise Apps in LLM App Store

Provider



3. Enterprise LLMops



Debmalya Biswas in DataDrivenInvestor

Generative AI—LLMOps Architecture Patterns

Deploying Large Language Models (LLMs) in the Enterprise

★ · 4 min read · Jun 25



Tapan Kumar Patro in Nerd For Tech

Deep Learning | Keras vs Pytorch | CNN, RNN, GAN, Autoencoder

Complex Data need complex architecture to understand and find insight, Deep learning...

★ · 8 min read · Dec 5, 2022



580



2



...



198



2



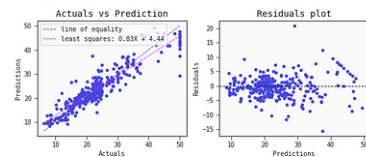
...

See all from Tapan Kumar Patro

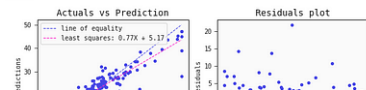
See all from DataDrivenInvestor

Recommended from Medium

Training Metrics



Test



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



...



407

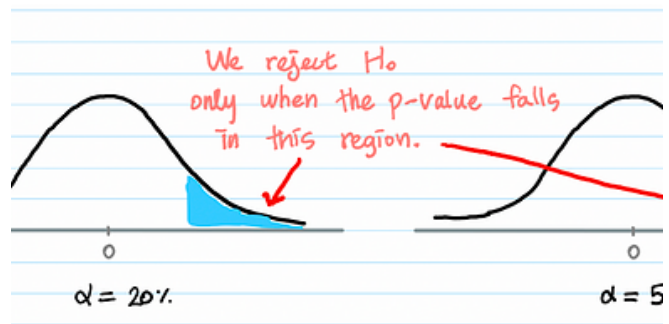


3



...

Lists

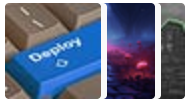


Ms Aerin in IntuitionMath

Chi Square Test—Intuition, Examples, and Step-by-Step...

The best way to see if two variables are related.

🌟 · 15 min read · Feb 13



Predictive Modeling w/ Python

20 stories · 473 saves



Practical Guides to Machine Learning

10 stories · 544 saves



Natural Language Processing


689 stories · 304 saves



New_Reading_List

174 stories · 143 saves



 Juan Broglio

Gamma Regression vs Linear Regression (in Python)

General Linear Models and Gamma Regression

4 min read · Aug 9



 Biman Chakraborty

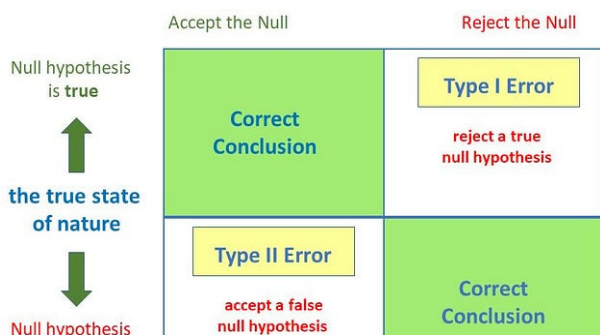
Two-Sample t Tests, Power, Effect Size and Sample Size Calculator i...


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

13 min read · Apr 30



8



 misun_song

Hypothesis Test and P-Value

Discerning Real Differences with P-values



 Shubham Pandey

Spotting the Unusual: A Guide to Outlier Detection (Part-1)

What is an outlier?

7 min read · Sep 24

 139 

4 min read · Aug 26

 1 

See more recommendations