Advances in Artificial Intelligence and Machine Learning; Research 3 (1) 711-730 Received 14-01-2023; Accepted 15-02-2023; Published 22-02-2023

Parametric PDF for Goodness of Fit

Natan Katz natan.katz@gmail.com
Kehilat Wrsaw 15B Tel Aviv

Uri Utai uri.itai@gmail.com
Kehilat Wrsaw 15B Tel Aviv

Corresponding Author: Natan Katz

Abstract

The methods for the goodness of fit in classification problems require a prior threshold for determining the confusion matrix. Nonetheless, this fixed threshold removes information that the model's curves present and may be beneficial for further studies such as risk evaluation and stability analysis. We present a different framework that allows us to perform this study using a parametric PDF.

Keywords: Beta Distribution, KL-divergence, Porbiality metrics, Density function, ROC curves, Interval of confidence

1. INTRODUCTION

Machine learning (ML) projects have become a leading tool in enormous domains of the computer industry. Their rule is far beyond computational aspects. Indeed, they are a focal point in designing analytical business decisions. The commercial usage of these models raises new challenges which are not often covered under traditional ML research. The latter often assumes that :

- The data in the database represents well the global data distribution.
- Training methodology aligns with the model's **Key Performance Indicators (KPI)**.
- There are no production-driven drawbacks.

kPI's are the indicators that data scientists focus on when they develop models. Commonly these indicators are determined upon the model's **False Reject (FR)** and **True Reject (TR)**. Unfortunately, none of these assumptions hold in real-world models. In addition, there are factors that are uniquely cardinal for commercial tasks such as complexity and stability (e.g. **"what is the efficient way to set a threshold to have both good and stable performance"**). In the academy, researchers focus mainly on abstract KPIs such as accuracy and precision. We use these KPIs for other scaling indicators such as Creamer's V, F1-score, AUC [1], and Matthew correlation coefficient (MCC) [1–4]. These indicators require a prior threshold for using them. Thus they all act as *discrete signals* . In the following sections, we discuss the derived drawbacks of *discrete signals* and suggest solutions.

Citation: Natan Katz and Uri Utai. Parametric PDF for Goodness of Fit. Advances in Artificial Intelligence and Machine Learning 2023;3(1):47.

2. MOTIVATION AND RELATED WORKS

The motivation for this work came from a hands-on project. However, researchers study the sensitivity of models for a threshold setting [5–7]. The problem occurs since the performances of an ML model are determined not only by the quality of the training process but by the threshold setting, as well. When we use a model with its fixed threshold, we omit many of the models' strengths:

- It tells nearly nothing about additional KPIs (e.g. when we set a threshold upon accuracy needs, we totally ignore precision or recall).
- Stability - We cant potentially estimate degradation when the model's input traffic is changed

We wish to estimate models in a wider fashion that allows us to have a broad knowledge of its strength. Studying continuous functions is a known approach in speech [8], and ML [9–11]. In such a manner, we suggest replacing the traditional performance manners that rely on tools such as confusion matrix and using a functional approach.

3. DISCRETE SIGNALS

In this section, we discuss the disadvantages of discrete signals. To do so, we need to review the typical inference process.

3.1 Inference Overview

Consider a well trained model **M** and an evaluation set **Dtest** one can easily deduce from 1 that the confusion matrix fully determines the model's evaluation. It leads to the following definition.

Definition 1 (Discrete signal) *Let M be the confusion matrix and $\mathbb{R}$ the real numbers (We take the entire real line since we wish to avoid prior constraints on F). Consider the function*

$$F\colon M \to \mathbb{R}$$

*If F is monotone for each entry of M, then F is a **Discrete signal**.*
*If F does not depend on M then it is called **Continuous signals**.* We note that the domain on the Discrete signal can be every nonempty subset of the entries of M.

The output of a classification model is a probabilities vector [12, 13]. We use these vectors to calculate our **FR** and **TR** curves. For classifying the data, we set a threshold. This threshold determines the confusion matrix. This matrix is the domain of the discrete signals [1]. Most of the common goodness of fit KPIs are discrete signals, nonetheless, these signals may suffer from three essential disadvantages:

- Unstable concerning the threshold

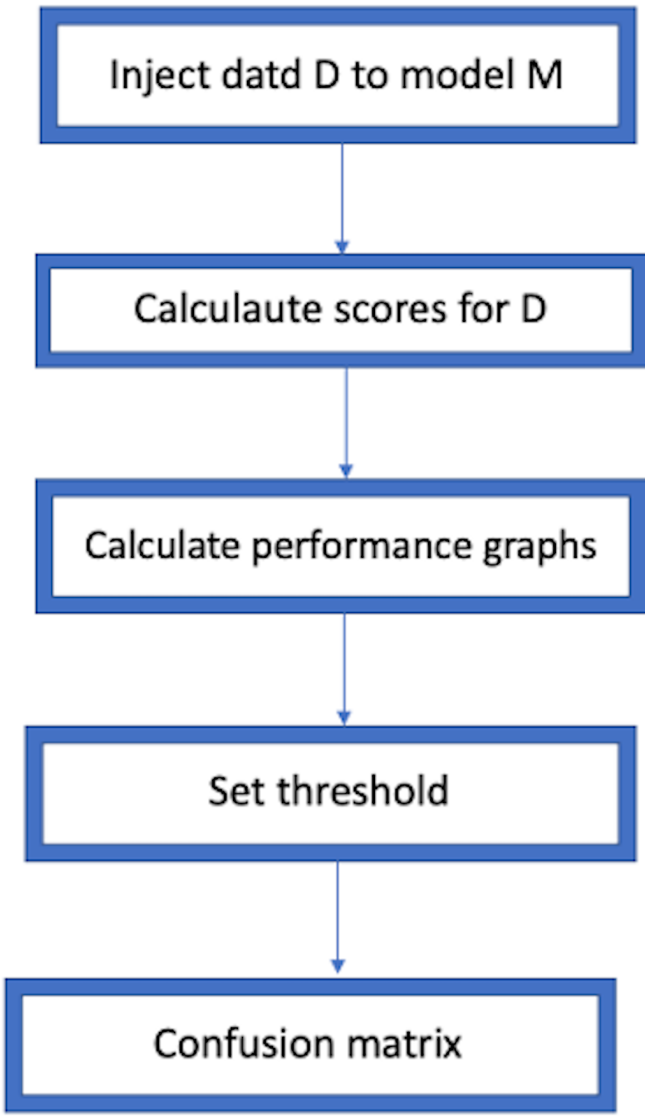


Figure 1: Generic Inference Process

- Difficult for risk calculations
- Absence of good mathematical toolbox

In the following subsections, we discuss these disadvantages.

3.2 Instability

Model's performances have a substantial capital impact. Therefore it is crucial to evaluate our indicators accurately. Setting a fixed threshold on the model graphs may provide two caveats:

- Typical graphs suffer from steep slopes concerning the thresholds
- Real-world statistics do not always identical to the distribution of the evaluation test

Academically, these phenomena are seldom studied. Nonetheless, different distributions and steep slopes often indicate instability. Thus, we find these caveats cardinal in the commercial world.

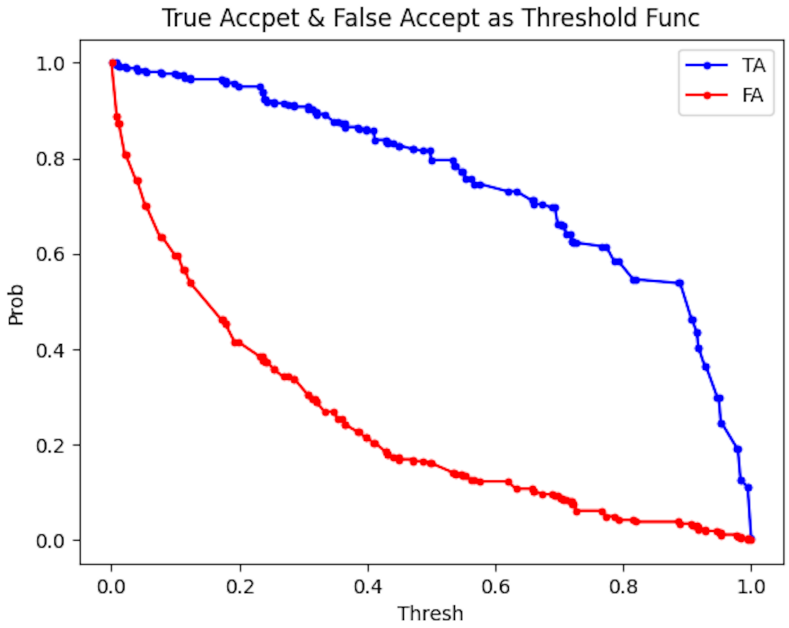


Figure 2: True Accept and FA graphs.

3.3 Risk Estimation

A cardinal tool in classical statistics is risk estimation. Whether a statistician is a Bayesian and uses **credible interval** [14, 15], or a frequentist that uses **confidence interval** [16, 17], this tool is essential. When we study distribution parameters, the ideal outcome consists of the parameters and a confidence measurement based on the distribution family. When we set a threshold or perform statistics such as maximum, we truncate our statistical information and collapse it to a single number. We can estimate the risk based on the threshold settings. However, the latter depends on our model, which leads to a non-coherent process. In contrast to the academy, the model's risk estimation is crucial in the commercial world.

3.4 Lack of Mathematical Toolbox

The final disadvantage of discrete signals is motivated by dynamical systems. Since we define discrete signals on the confusion matrix, which is a fixed matrix, we cannot define open sets. We can study neither infinitesimal perturbations nor stability analysis. These two are cardinal for the model's pre-deployment tests.

3.5 Predict Proba

The data scientists among the readers may wonder **"What about predict proba ?"**,[18]. Indeed, predict proba is not a discrete signal since it doesn't use a confusion matrix. However, it merely provides a scores histogram and has no canonical form. Therefore we can have no generic methodology to study its stability or evaluate its risk. Nevertheless, one can consider the discussion in the following sections as **"Methods for continuous approximation of proba"**

3.6 So What Can We Do?

We over-viewed the main drawbacks of discrete signals. **Can we provide a remedy?** If we search for common manners of these drawbacks, it is clear that a more "continuous" framework can be beneficial. Thus defining **PDFs** on models' curves can assist in this study.

4. CONTINUOUS SIGNALS- PARAMETRIC PDF

4.1 Motivation

We discussed the drawbacks of discrete signals. Nonetheless, models output continuous signals: their scores' curves. If we replace the common analysis that studies a confusion matrix with an analysis of these curves, we may overcome some of the drawbacks:

- Curves allows you to calculate different order derivatives which indicate stability status
- It allows to use of metrics such as the L_P or probabilistic such as Jensen-Shannon or Kullback-Leibler [19, 20].
- it allows to obtain the behavior of common indicator upon perturbation
- Using distribution family manners, it can evaluate risk using the interval of confidence

We can cleverly choose a distribution family that handles most of the discrete signals' drawbacks using its parameters. It preserves the probabilistic nature of ML models.

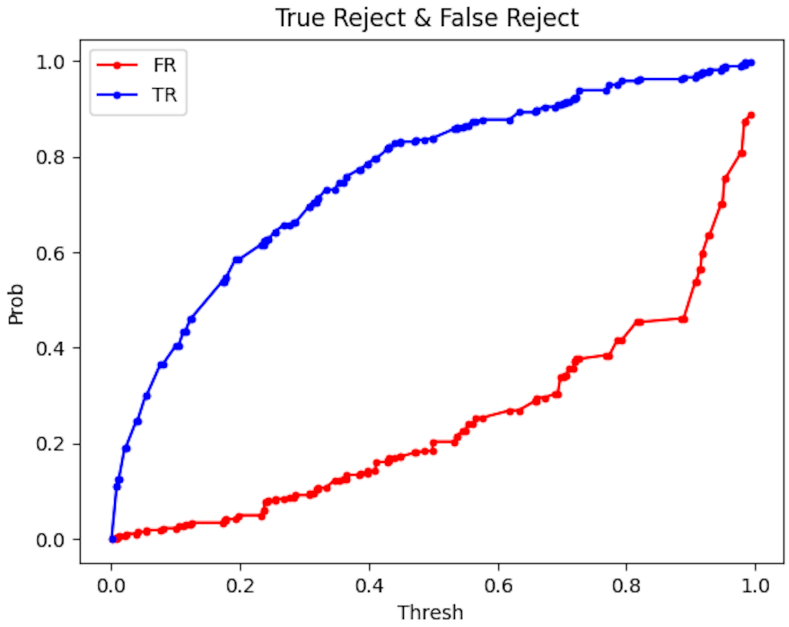


Figure 3: TR (blue) and FR (red)

4.2 Kullback-Leibler Metric

The Kullback-Leibler metric **KL** is commonly used in applied probability. It can measure the distance between two discrete distributions or continuous densities. Although it suffers from the absence of some metric manners (e.g. it is not symmetric), it allows handling with every type of probability function. In some cases such as in the Gaussian or Beta (which we will discuss in the next section), it provides an analytic closed form that simplifies our calculations. Clearly, there are additional methods to estimate the distance between two densities such as **total variation** and **Wasserstein distance**.

4.3 Parametric Distributions

Consider a standard binary classification problem. We train a model using a deep learning architecture or a classical tool such as logistic regression. In the inference, the model outputs a vector of probabilities of length 2 (number of classes). 3 presents a typical scenario. We will give a mathematical definition that probably most of the readers are familiar with:

Definition 2 *Cumulative A function **F** is said to be **a Cumulative Distribution Function** (CDF) if it satisfies the following:*

- *Non decreasing*
- *Right continuous*

- $\lim_{x \to -\infty} \boldsymbol{F}(x) = 0$
- $\lim_{x \to \infty} \boldsymbol{F}(x) = 1$

Definition 3 *Density We say that a function* ***P*** *is a* ***density function*** *if it is a derivative of a CDF.*
In 3, we can see that **FR** and **TR** satisfy the required. If we have an explicit form of the function, we can derive this function, evaluate risk and perform stability analysis. Moreover, we can calculate error areas analytically, as appears in 4.

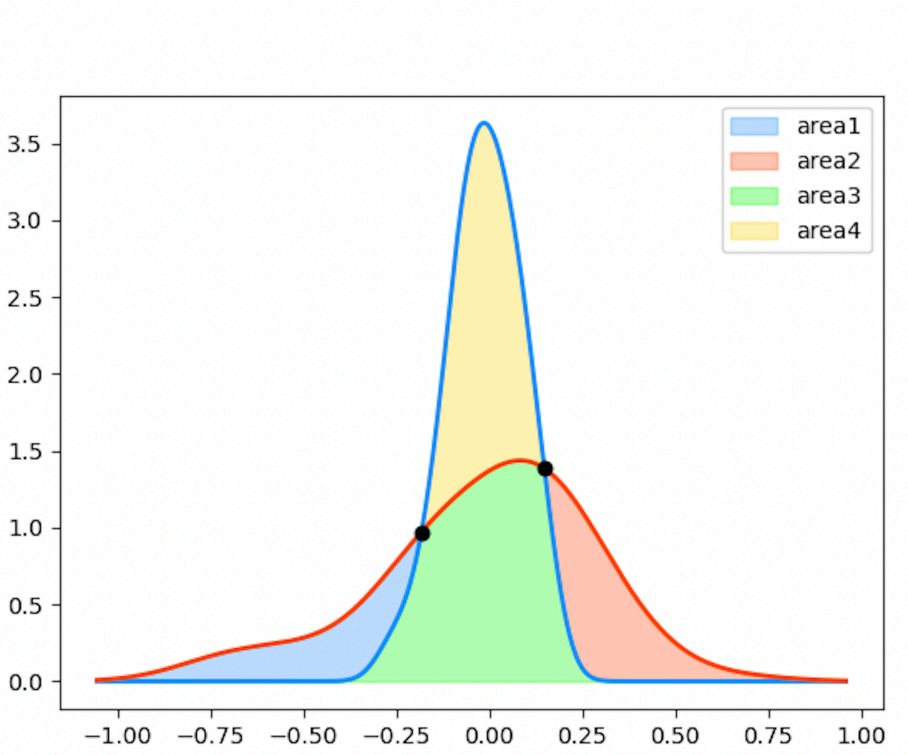


Figure 4: Area 3 represents the intersection between PDFs

4.4 Beta as a Case Study

Consider a binary classification problem. The model detects whether an input is an element in class **"1"** and provides the probability for this event. We wish to model the FR and TR using a sound distribution family. A natural choice is **Beta** function [15, 21, 22].

4.4.1 Beta's properties

We will describe Beta's main properties:

- Beta's support is on $[0, 1]$ interval. Moreover, it is strictly great in the interior of the support.
- The Beta distribution is infinitely continuous.
- The distribution has two positive parameters α and β.

We denote by μ the mean of a random variable and by σ the standard deviation. A random variable X with Beta distribution satisfies the following:

$$\mu[X] = \frac{\alpha}{\alpha + \beta} \tag{1}$$

$$Var[X] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)} \tag{2}$$

We can revert the formula [23]:

$$\alpha = \left(\frac{1 - \mu[X]}{\sigma[X]^2} - \frac{1}{\mu[X]} \right) \mu[X]^2 \tag{3}$$

$$\beta = \left(\frac{1}{\mu[X]} - 1 \right) \alpha \tag{4}$$

A typical shape of Beta appears in 5.

We complete this section by presenting the KL closed form formula of Beta [24]. Let positive number α_1,β_1,α_2,β_2 We have Γ and Ψ functions (in some books,Ψ appears as **digamma** or **polygamma** or order 0).

$$\begin{aligned} KL[(B(\alpha_1, \beta_1)||B(\alpha_2, \beta_2)] = \ln \frac{B(\alpha_2, \beta_2)}{B(\alpha_1, \beta_1)} + (\alpha_1 - \alpha_2)\Psi(\alpha_1) + \\ (\beta_1 - \beta_2)\Psi(\beta_1) + (\alpha_2 - \alpha_1 + \beta_2 - \beta_1)\Psi(\alpha_1 + \beta_1) \end{aligned} \tag{5}$$

4.4.2 Example

In this section, we compare common indicators with the performances of a KL divergence between FR and TR during a model training of a binary classification problem. We follow three indicators

- Accuracy
- MCC
- KL distance between FR and TR

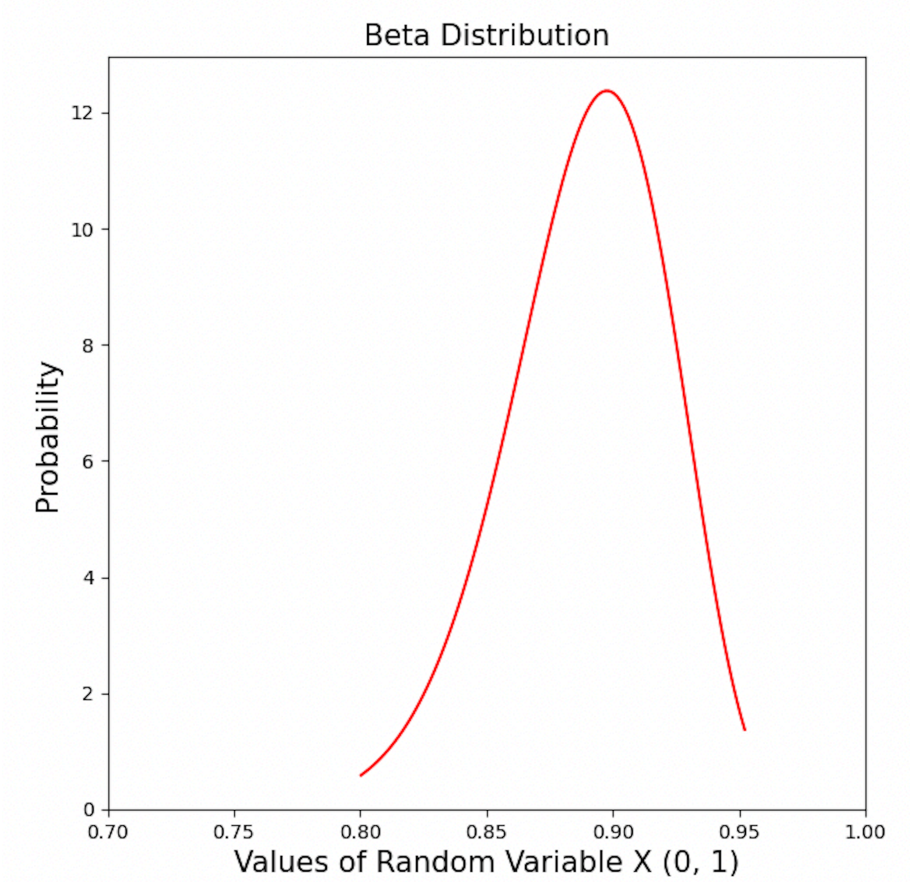


Figure 5: Beta's PDF

One can see the obtained graphs om in FIGURE 6, and the code we used [25]. The number in the graphs' headers represents the number of epochs. We can see that as this number increases, the gaps between the function increases. More importantly, we see that the KL increases with the accuracy and MCC, which gives an optimistic perspective on our hypothesis.

4.5 More Examples

We presented the Beta distribution study as it fits classification models since its support is the interval $[0, 1]$. However, we wish to present some examples that present different scenarios::

4.5.1 Non improving model

Here we will show a scenario in which the continuous approach can detect a wider range of processes than common KPIs: The accuracy and MCC don't change were as the KL metric does. We can relate it to some that take place during training but do not necessarily improve accuracy (e.g. precision or recall).

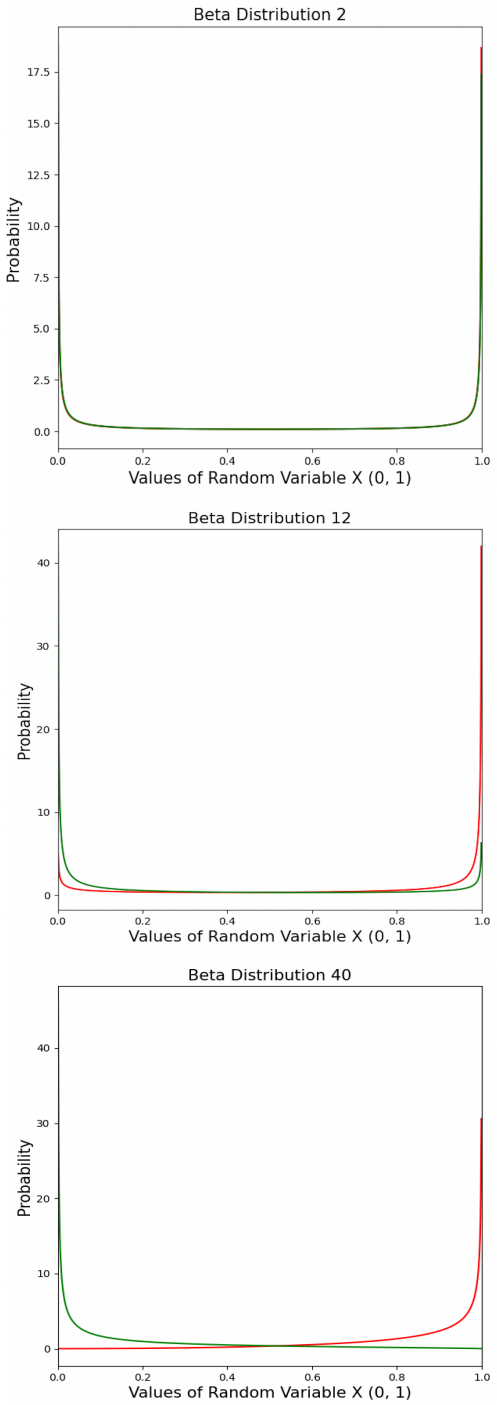


Figure 6: Comparison of our measurements using the tuple (accuracy, KL, MCC), As we move down, the models are better trained: Top= (0.82,0.04, 0.66), Middle =(0.89,1.06, 0.78), Lowest= (0.91,3.9,0.82).

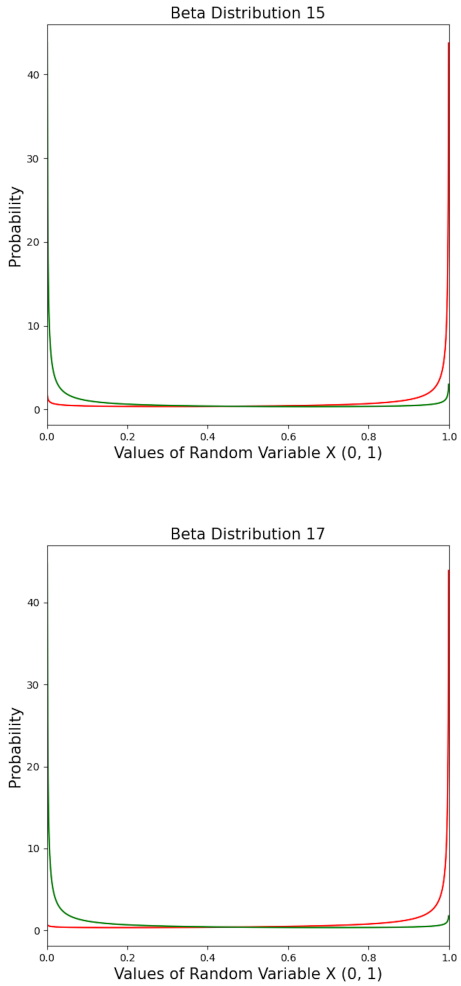


Figure 7: Comparison of our measurements using the tuple (accuracy, KL, MCC). Top= (0.9466,1.73,0.8933) Lowest= (0.9466,2.02.0.8933).

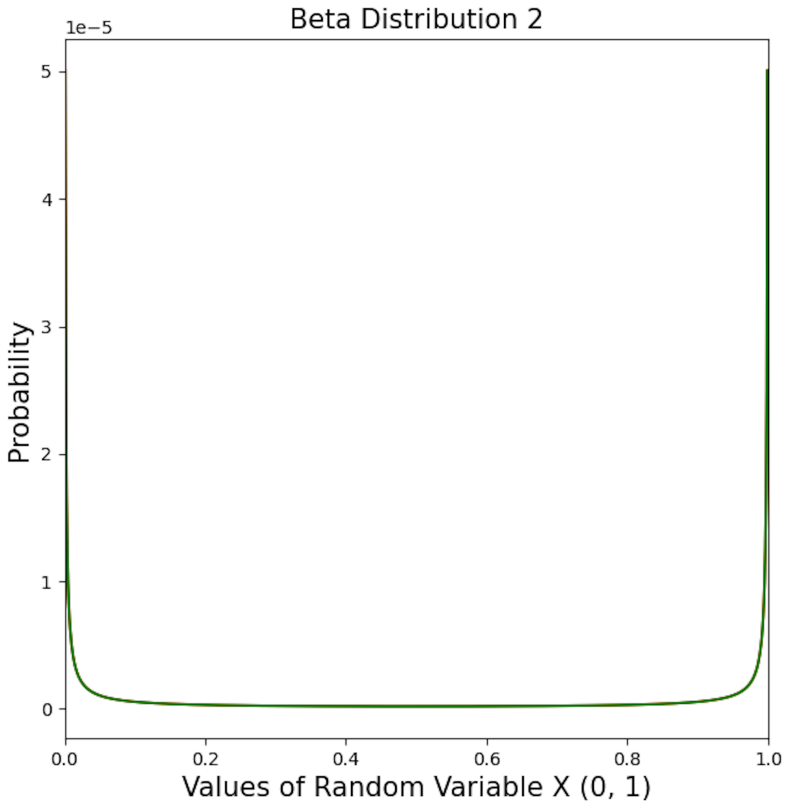


Figure 8: Random Model Behavior

FIGURE 7, presents the described above: we can see the gaps between the distributions, particularly near the edges. and we see them increase. However, accuracy remains fixed.

4.5.2 The random model

Here we discuss a random model. It is binary with an accuracy of 0.49 When we ran the KL divergence formula we got a coded warning and an overflow. But when we tested the reason we realized that the measured hyperparameters of the beta distributions were similar namely, no separation exists at all (which we expect from a random model).
It can be seen in FIGURE 8, that the model is random since only a single distribution is visible.

4.5.3 A cyber example

In the following example, I used ”real-world” data. this is a multi-class classification problem from the cyber domain. In order to modify it to a binary problem, I aggregated several classes together

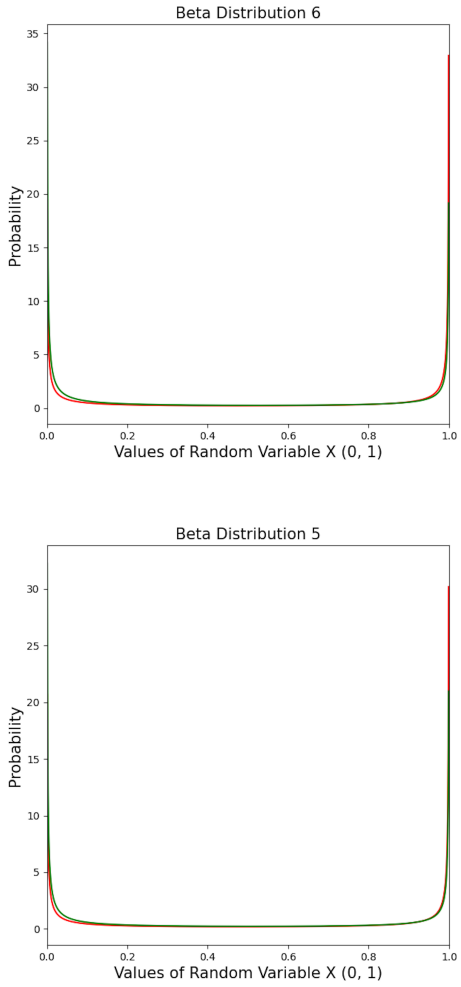


Figure 9: Multi class cyber1 (accuracy, KL, MCC). Top= (0.87,0.37,0.74) Lowest= (0.85,0.29,.0.71).

on expert knowledge suggestions. I used the same format as in the previous sections to present the results in figures 9 , 10 and 11 . Due to obvious reasons, I cannot elaborate more.

Another example using the same data but with a different aggregation approach Here is the last cyber example.

We can see that in our cyber example, our continuous approach works well.

4.6 Goodness of Fit - Summary

We proposed the continuous signal approach and discussed its theoretical improvements for the discrete signals as a goodness-of-fit method. We presented several examples from both known

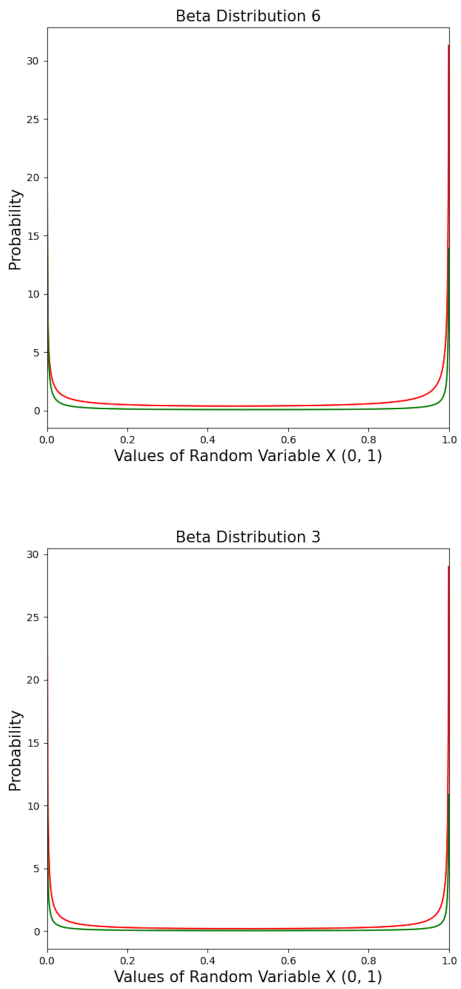


Figure 10: Multi class cyber1 (accuracy, KL, MCC). Top= (0.93,0.937,0.768) Lowest= (0.92,0.702,.0.75).

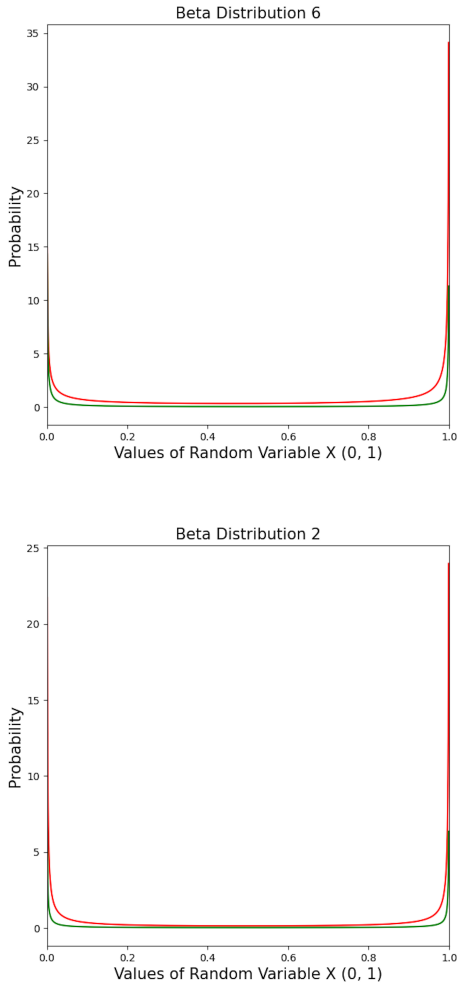


Figure 11: Multi class cyber1 (accuracy, KL, MCC). Top= (0.955,1.18,0.82) Lowest= (0.93,0.88,.0.80).

common data and commercial real-world data to estimate this approach for binary classification problems. For modeling the curves, we used Beta distribution and KL divergence. We deduced that this approach can identify when a model doesn't work (the random model) and detect improvement in a wider range than accuracy itself. In the next section, we will study another approach for using continuous signals.

5. Training

In the previous sections, we tested the idea that the separation between TR and FR graphs can be a goodness of fit indicator. We have seen some examples of how this hypothesis works well. It leads to a further question: Can we use this approach during training by adding a regulation term? We begin the discussion by presenting an intuition for using this method,**(It is intuition and not a proof!**).

Definition 4 (Left epsilon-Beta) *Consider a Beta distribution and a positive small* ϵ*. A* ***Left*** ϵ***-Beta function*** *is a Beta distribution where*

$$\frac{\beta}{\alpha} < \epsilon \tag{6}$$

The right Beta function is defined by the reciprocal (see FIGURE 12).

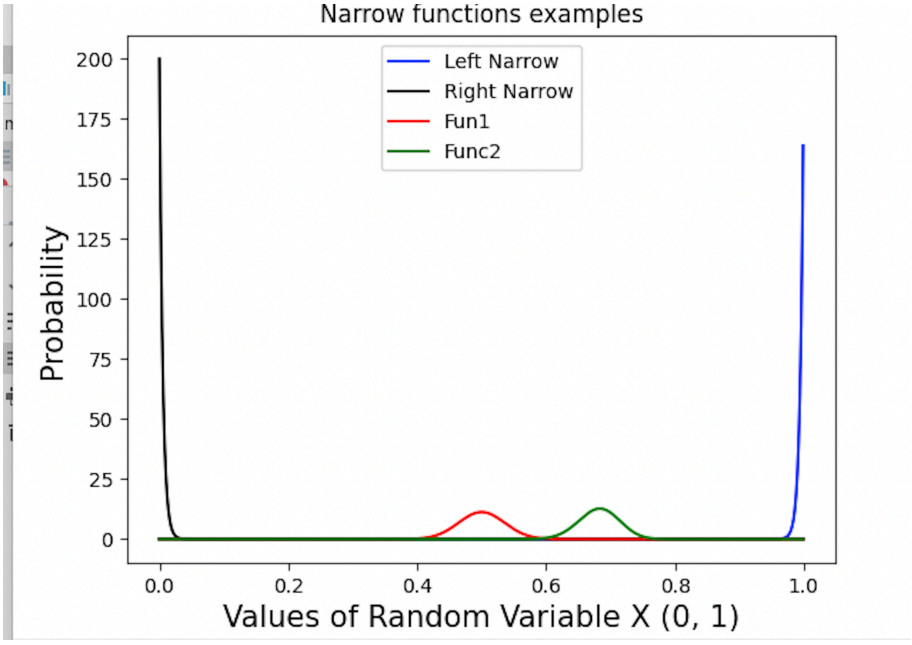


Figure 12: Epsilon Beta functions

We aim to maximize the distance between two Beta functions **P** and **Q**. Consider a metric **d** that satisfies the triangle inequality (KL and J-S do not always do). Let **R, L** right and left ϵ-Betas. The following inequalities hold:

$$\mathbf{d(R,L)} \geq \max[\mathbf{d(P,Q)}] \tag{7}$$

$$\begin{aligned} \mathbf{d(R,L)} \leq \min[\mathbf{d(L,P)} + \mathbf{d(P,Q)} + \mathbf{d(Q,R)}] \leq \\ \min[\mathbf{d(L,P)} + \mathbf{d(Q,R)}] + \max(\mathbf{d(P,Q)}) \end{aligned} \tag{8}$$

The LHS is constant. Combining with the upper inequality, we obtain that for some cases maximizing **d(P, Q)** is equivalent to minimizing the other terms of the RHS,

5.1 Training Example

We present an xgboost model training: [26]: one model is a vanilla xgboost, and the other uses a new regulation term: the gradient of KL divergence [25]. We compared the models using three indicators:

- Accuracy
- Precision
- MCC

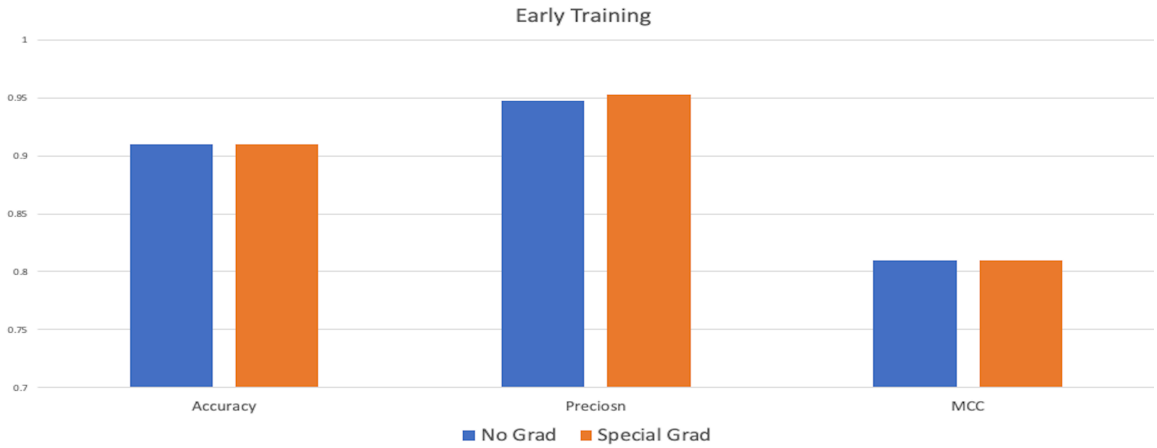
The results are in FIGURE 13.

Considering these results, we can't declare a clear winner. However, the "regulated model" shows an advantage in all KPIs compared to the vanilla model. It hints that this approach is not far-fetched and requires further study.

6. SUMMARY AND FUTURE WORK

We described the approaches for evaluating the goodness of fit of ML models and discussed some of their inherent failures. We presented new notions: **discrete signals** and **continuous signals** that allowed us to develop a different methodology to overcome these failures. We suggested that parametric PDFs can act as continuous signals and that by using these, we can evaluate the model's risk and analyze its stability. We tested this approach for both the goodness of fit purposes and as a training regulation function. The results are promising, but it is evident that further massive research is required:

- Test on various databases
- Test on different methodologies such as DL
- Generalize binary problems to multi-classes by replacing Beta to Dirichlet
- Test **Isotonic Regression** [27], which is extremely common in regression problems
- Test on various distributions such as Gamma



and during training:

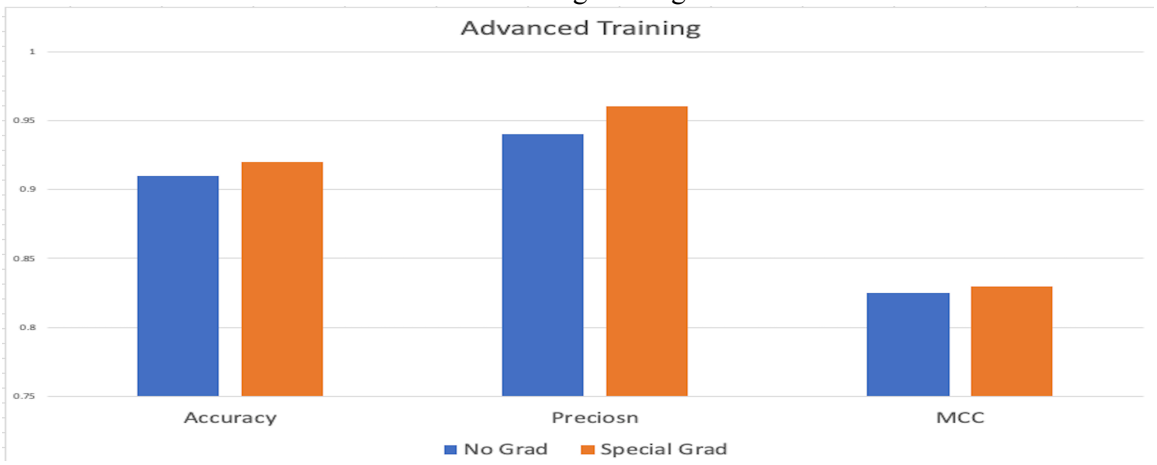


Figure 13: Comparison of KPIs between vanilla model and regulation term.

These are all plausible tools for improving the offered approach and enhancing its usage. Finally, we believe such frameworks will enhance the usage of classical statistics and dynamical system tools. These tools are mandatory in deploying prediction models, particularly in the commercial world.

7. ACKNOWLEDGEMENT

I wish to thank Erez Israel, Irinia Shalem, and Ofek Dadush from Checkpoint for presenting the severity of the problem, and Konstantin Gedalin for his fruitful ”out of the box” ideas.

References

[1] Itai U, Katz N. Goodness of Fit Metrics for Multi-Class Predictor.2022. arXiv preprint: arxiv.org/pdf/2208.05651.pdf

[2] Jurman G, Riccadonna S, Furlanello C. A Comparison of MCC and CEN Error Measures in MultiClass Prediction. Public Library of Science San Francisco, USA.PLOS ONE.2012;7:e41882.

[3] Chicco D, Jurman G. The Advantages of the Matthews Correlation Coefficient (MCC) Over F1 Score and Accuracy in Binary Classification Evaluation. BMC genomics.2020;21:1-3.

[4] Anderson TW, Darling DA. A Test of Goodness of Fit. J Am Stat Assoc. 1954;49:765-769.

[5] Charitos T, van der Gaag LC. Sensitivity Analysis for Threshold Decision Making With Dynamic Networks. 2012. arXiv preprint: https://arxiv.org/pdf/1206.6818.pdf

[6] https://towardsdatascience.com/classification-models-and-thresholds-97821aa5760f

[7] https://towardsdatascience.com/calculating-and-setting-thresholds-to-optimise-logistic-regression-performance-c77e6d112d7e

[8] Rabiner L. Fundamentals of Speech Recognition, PTR Prentice Hall. 1993.

[9] Rokach L, Maimon O. Data Mining With Decision Tree; Series in Machine Perception And Artificial Intelligence. World Scientific. 2014;89:81:61-62.

[10] https://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15381-s06/www/DTs2.pdf

[11] https://www.ise.bgu.ac.il/faculty/liorr/hbchap9.pdf

[12] *https://pytorch.org/*,

[13] https://scikit-learn.org/stable/tutorial/index.html

[14] http://www2.stat.duke.edu/ rcs46/lecturesModernBayes/601-module3-ebayes/lecture5-more-bayes.pdf

[15] `http://varianceexplained.org/r/credible_intervals_baseball/`

[16] https://www.redjournal.org/article/S0360-3016(21)03256-9/fulltext

[17] https://datatab.net/tutorial/confidence-interval

[18] https://scikit-learn.org/stable/modules/calibration.html

[19] https://en.wikipedia.org/wiki/Jensenshannondivergence

[20] Kullback S, Leibler RA. On Information and Sufficiency. The annals of mathematical statistics. 1951;22:79-86.

[21] https://web.stanford.edu/class/archive/cs/cs109/cs109.1166/pdfs

[22] https://stephens999.github.io/fiveMinuteStats/beta.html

[23] https://stats.stackexchange.com/questions/12232/calculating-the-parameters-of-a-beta-distribution-using-the-mean-and-variance

[24] https://en.wikipedia.org/wiki/Betadistribution

[25] https://github.com/natank1/Betapaper

[26] https://xgboost.readthedocs.io/en/stable/

[27] https://en.wikipedia.org/wiki/Isotonicregression