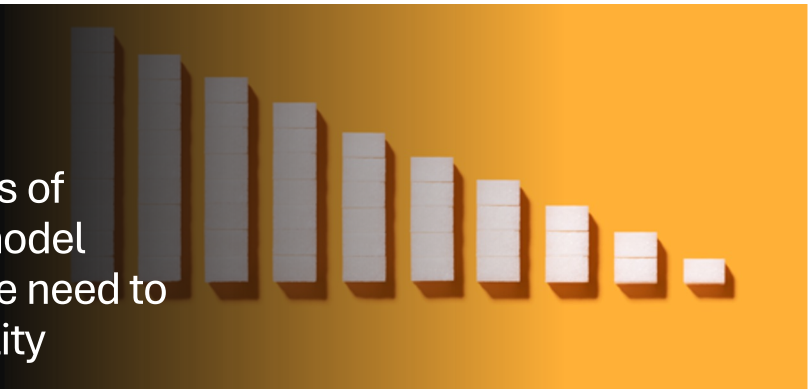
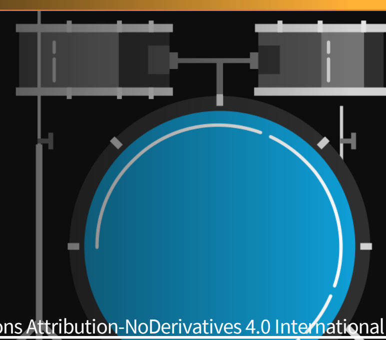


Quantitative analysis of Machine Learning model performance and the need to consider explainability



Vishnu S. Pendyala, Ph.D.

San Jose State University

To cite this presentation: Pendyala, V.S. (2024) "Quantitative analysis of Machine Learning model performance and the need to consider explainability". IEEE Computer Society, Santa Clara Valley Chapter Technical Talk, December 30, 2024

1

Task	Performance Compared to Humans
Text Summarization	Can achieve similar quality
Machine Translation	Near human-quality for some languages
Question Answering on Factual Topics	Can perform well on factual topics with large datasets
Code Generation	Can generate some basic code
Image Captioning	Can generate accurate descriptions of images
Speech Recognition	Achieves high accuracy in controlled environments

THE UNREASONABLE EFFECTIVENESS OF GENERATIVE AI'S MULTIMODAL ABILITIES

2

GPT-4o: Academic Benchmarks

Benchmark	Score	Interpretation
MMLU	88.7	High level of understanding across a wide range of academic subjects, comparable to undergraduates or even a graduates in a general field.
GPQA	53.6	Moderate proficiency in handling complex, nuanced questions, which aligns with the capabilities of an undergraduate.
MATH	76.6	Strong mathematical abilities, akin to a student with an undergraduate degree in mathematics or a related field.
HumanEval	90.2	Excellent programming skills, similar to those of a highly proficient software engineer or computer science graduate.
MGSM	90.5	Exceptional proficiency in solving grade school level math problems across multiple languages.
DROP	83.4	Strong reading comprehension and reasoning abilities, comparable to undergraduates well-prepared for graduate-level work.

Source: https://community.openai.com/t/education-level-interpretation-of-gpt-4os-benchmarks/763947

3

LLMs passed some tough exams too…

CNN Business Markets Tech Media Calculators Videos Watch

ChatGPT passes exams from law and business schools

By Samantha Murphy Kelly, CNN Business
4 minute read · Updated 1:35 PM EST, Thu January 26, 2023

BUSINESS INSIDER

GPT-4 scored in the 90th percentile of the bar exam with a score of 298 out of 400.

BUSINESS INSIDER

ChatGPT passed all three parts of the United States medical licensing examination within a comfortable range.

BUSINESS INSIDER

GPT-4 aced the SAT Reading & Writing section with a score of 710 out of 800, which puts it in the 93rd percentile of test-takers.

4

But with a caveat…

Mirzadeh, et al. "Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models." *arXiv preprint*

"we investigate the fragility of mathematical reasoning in these models and demonstrate that their performance significantly deteriorates as the number of clauses in a question increases. We hypothesize that this decline is due to the fact that current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data"

5

Why care for accuracy?

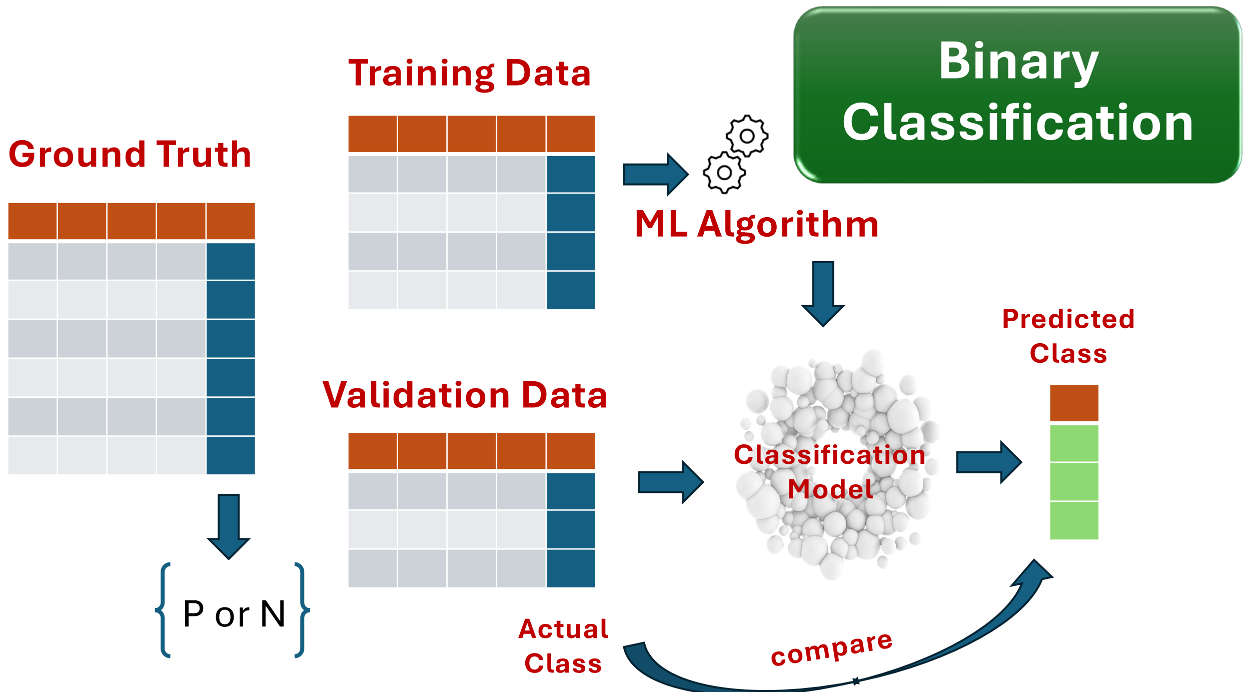
Every decimal % pays!

Source: https://www.datarobot.com/customers/steward-health-care/

"Just a 1% reduction in registered nurses' hours paid per patient day netted $2 million in savings per year, for just eight of the 38 hospitals in Steward's network"

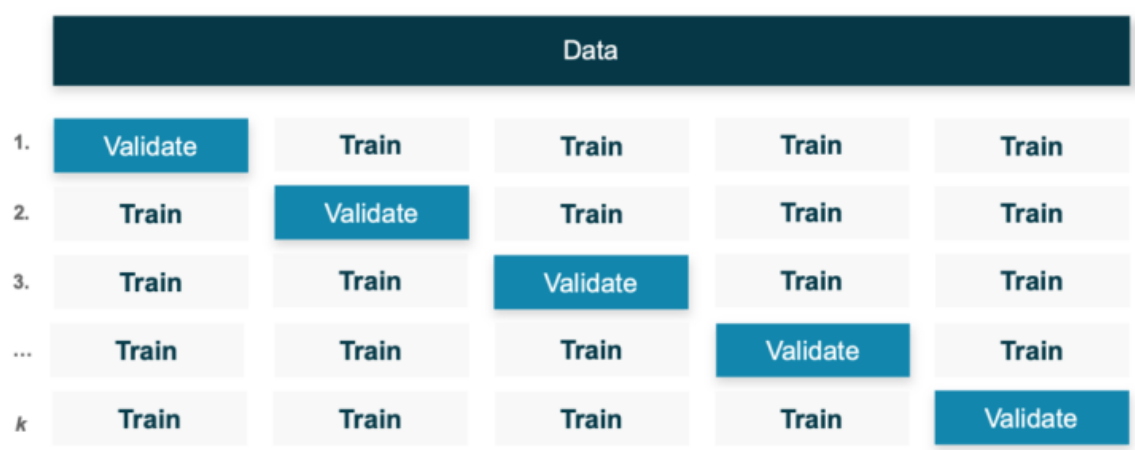
"Reducing patient length of stay by 0.1% results in savings of over $10 million per year"

6

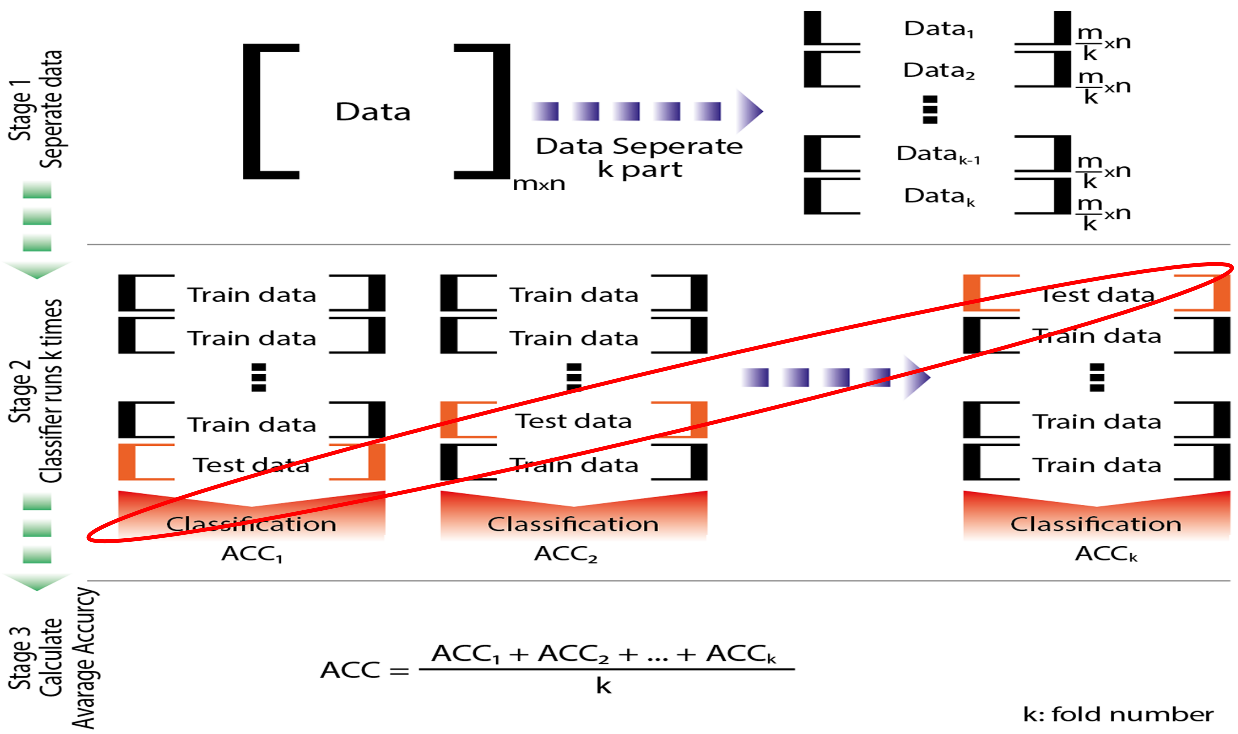


7

Approaches to evaluating Classifiers
k-fold Cross-Validation



8



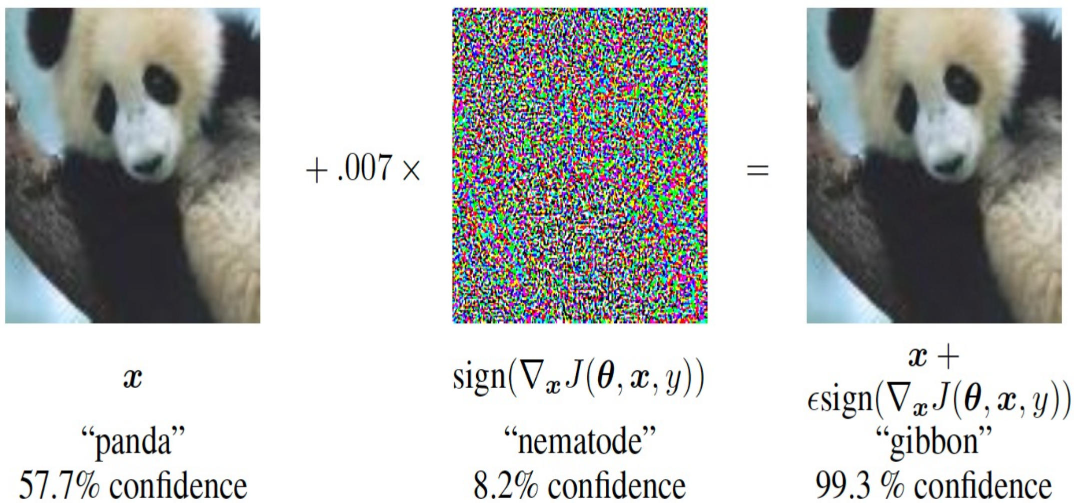
9

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

10

Type II Adversarial attack!



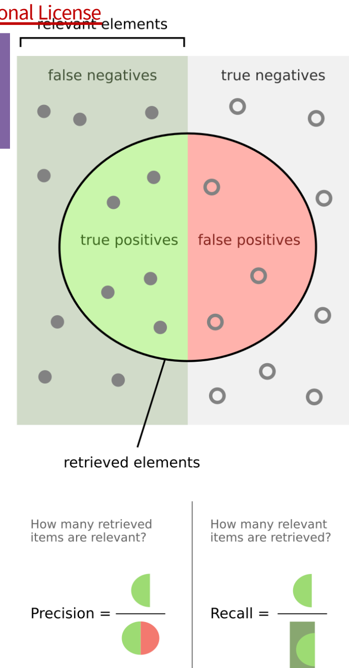
Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

11

A few popular metrics

TP	FP
FN	TN

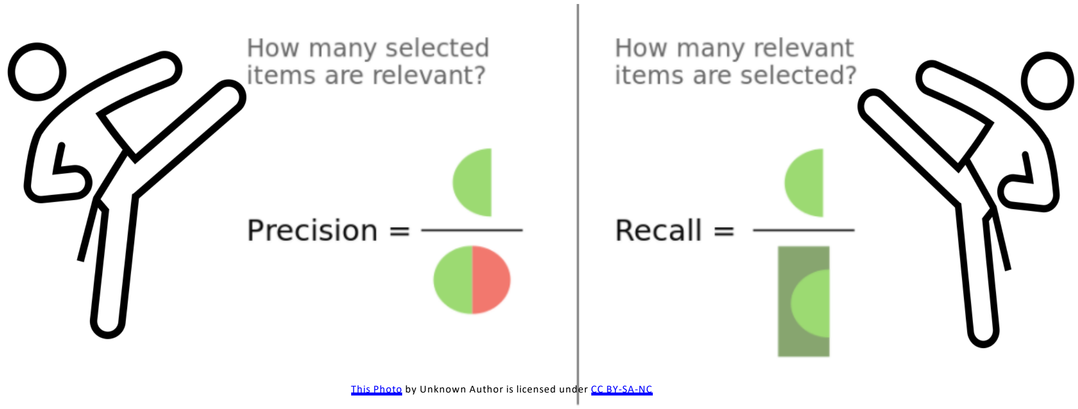
- Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
- Total actual positives, P = TP + FN
- Total actual negatives, N = FP + TN
- Recall or Sensitivity (True Positive Rate) = $\frac{TP}{P}$
- FPR (False Positive Rate) = $\frac{FP}{N}$
- Specificity = 1 – FPR
- Precision = $\frac{TP}{TP+FP}$



12

Precision vs Recall

TP	FP
FN	TN



Precision: Focus on minimizing False Positives

Recall: Focus on minimizing False Negatives

13

Coldmail Spam Filtering

Coldmail is new and is still working on a spam filtering ML model

1M mails in a day, 10,000 of which are spam

Accuracy of a model that classifies every mail as genuine = 990k / 1M = 99%

TP = 0 => Precision = Recall = 0

Accuracy of a model which classifies every mail as spam = 10k / 1M = 1%

Precision = 10k / 1M = 1%; Recall = 10k/10k = 100%

None of accuracy, precision, or recall gives the true picture in all scenarios!

=> Need more metrics

14

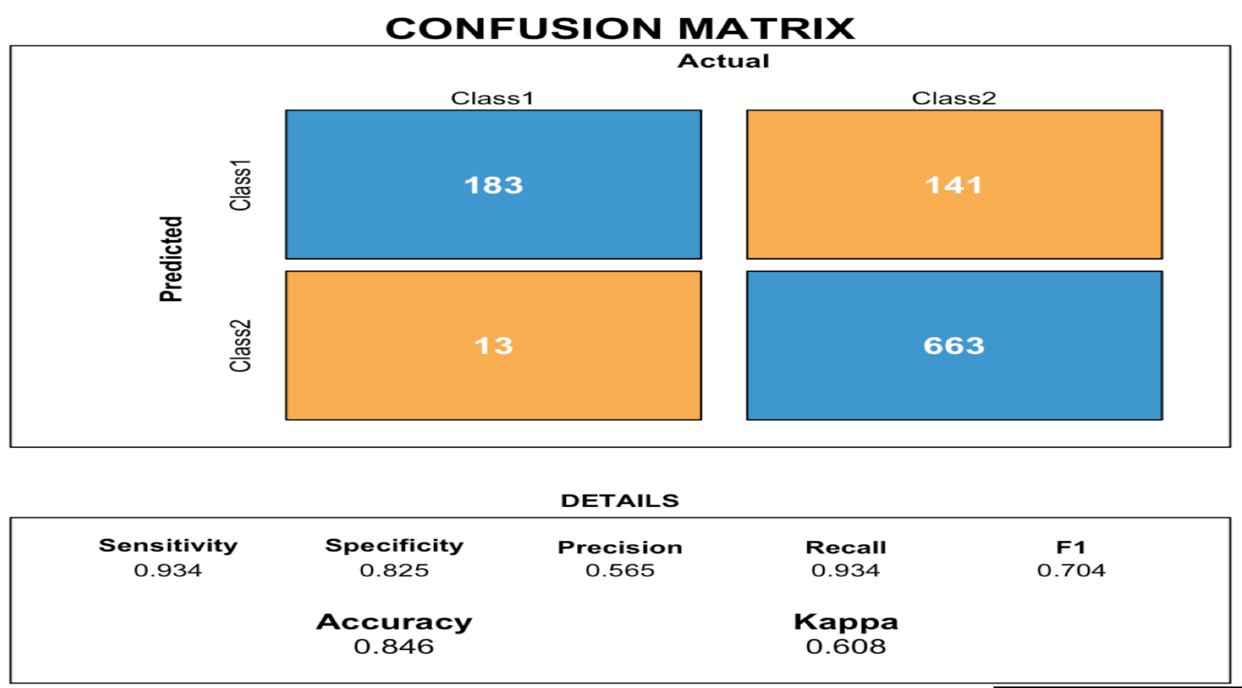
F-measure or F-score or **F_1 score** is the harmonic mean of precision and recall

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2}\right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

F_1 gives equal importance to precision and recall => Domain characteristics are ignored! => Need more metrics!

We therefore adopt a more general form: $F_\beta = \dfrac{(\beta^2 + 1)PR}{\beta^2 P + R}$

15



16

Many metrics and multiple ways to express them

Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR–) $= \frac{FNR}{TNR}$
Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2\ PPV \times TPR}{PPV + TPR} = \frac{2\ TP}{2\ TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Source: Wikipedia

17

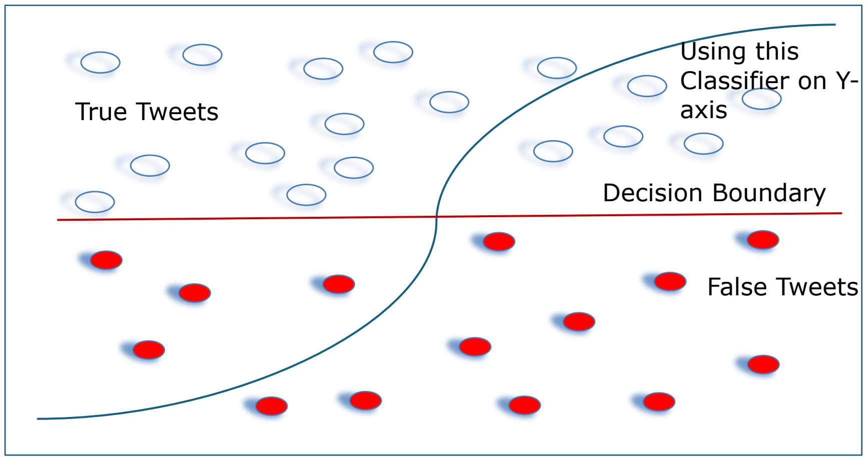
Many metrics (continued)

Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate type II error [c] $= \frac{FN}{P} = 1 - TPR$
False positive rate (FPR), probability of false alarm, fall-out type I error [f] $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$

Source: Wikipedia

18

Prediction Threshold: What happens if you move the decision boundary to extremes?



Legend:

- False Tweets plotted as points in the feature space
- True Tweets plotted as points in the feature space

19

Prediction Threshold

	Predicted 1	Predicted 0
True 1	0	b
True 0	0	d

When threshold > MAX(f(X))

- all cases predicted False (0)
- (b+d) = total
- accuracy = %False = %0's

	Predicted 1	Predicted 0
True 1	a	0
True 0	c	0

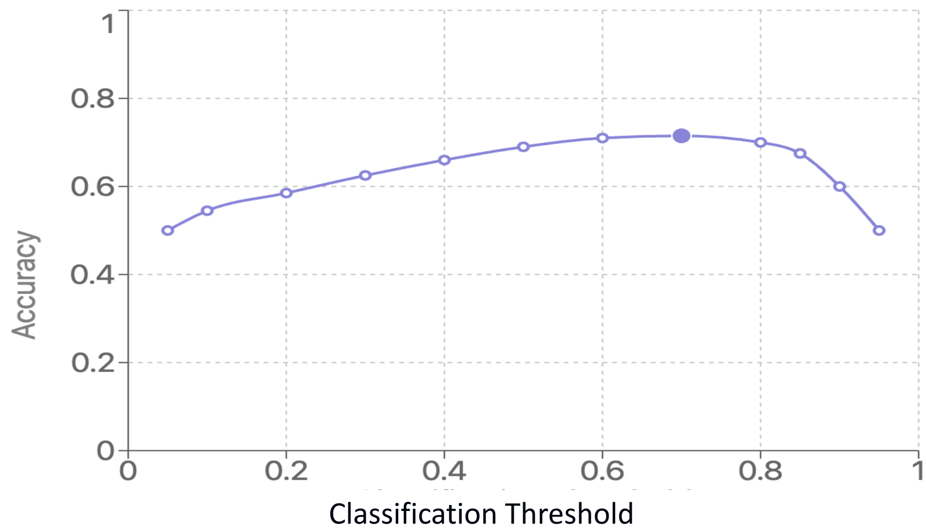
When threshold < MIN(f(X))

- all cases predicted True(1)
- (a+c) = total
- accuracy = %True = %1's

20

How do you choose the Optimal Threshold?

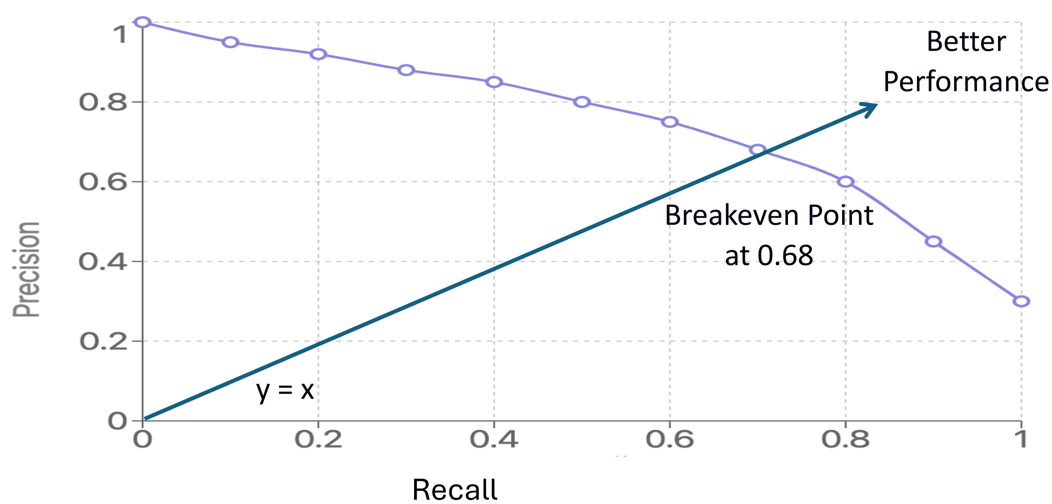
At Optimal Threshold, 0.70,
Accuracy: 0.715
TPR: 0.730,
TNR: 0.700



21

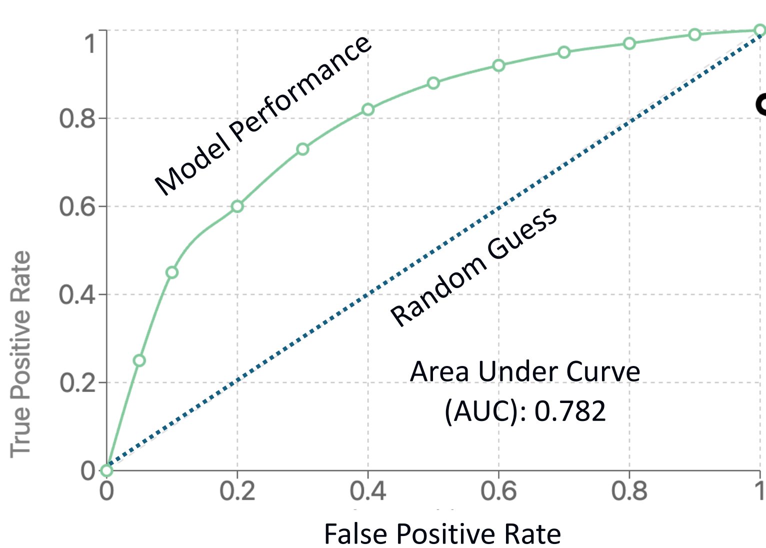
Precision-Recall Curve

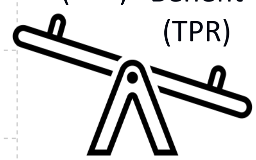
Average Precision (AP): 0.753



22

Receiver Operating Characteristic (ROC) Curve





Thrshld	TPR	FPR
0.95	0.0	0.0
0.9	0.25	0.05
0.85	0.45	0.1
0.8	0.6	0.2
0.7	0.73	0.3
0.6	0.82	0.4
0.5	0.88	0.5
0.4	0.92	0.6
0.3	0.95	0.7
0.2	0.97	0.8
0.1	0.99	0.9
0.05	1.0	1.0

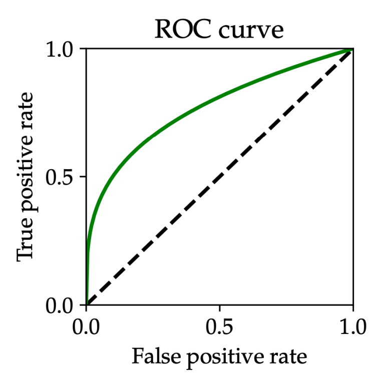
Area Under the ROC Curve

- 1.0: perfect prediction
- 0.5: random prediction
- <0.5: something wrong!

23

Properties of the ideal ROC curve

- The points (0, 0) and (1, 1) are on the ROC curve
- The ROC must lie above the main diagonal
- The ROC curve is concave



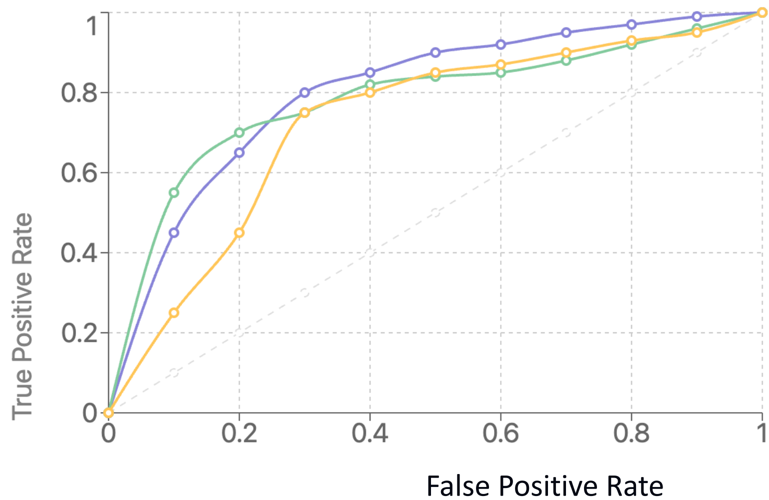
Source: Moritz Hardt, Benjamin Recht, Patterns, predictions, and actions: A story about machine

24

Intersecting ROC Curves

Model A (purple) performs better than Model B (green) and Model C (orange) for most thresholds

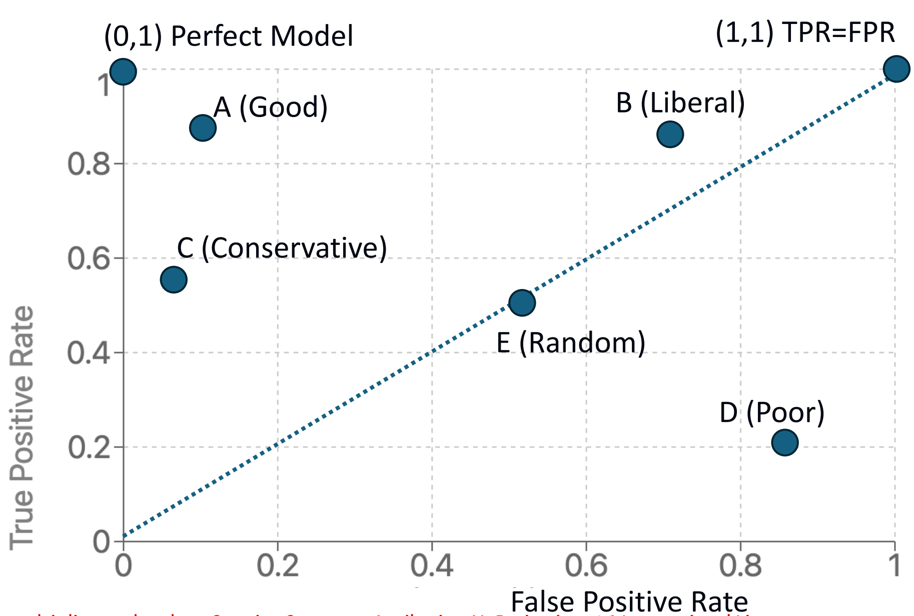
AUC
Model A : 0.798
Model B: 0.777
Model C: 0.725



25

ROC Plot

Performance of Models A, B, C, and D for some threshold



26

Cohen's Kappa Statistic

The Kappa Statistic measures the agreement between the evaluations of actual and predicted values.

It describes agreement achieved beyond chance, as a proportion of that agreement which is possible beyond chance.

27

An Example

- A set of biased coins turn heads 90% of the time. What is the best accuracy of a random system used for prediction?
- (Hint: best case when always predicts the majority class)

  90%

- If the accuracy of a system that you designed is also 90%, how much of that accuracy is attributable to chance?

  100%

=> kappa coefficient = agreement beyond chance = 0

=> The designed system modeled the bias in the coin and could not overcome the bias

28

Interpreting Cohen's Kappa

The value of the Kappa Statistic generally ranges from 0 - 1.00, with larger values indicating better reliability.

- A value of 1 indicates perfect agreement.
- A value of 0 indicates that agreement is no better than chance.

Generally, for ML models, a Kappa > 0.40 is considered satisfactory.

29

Formula for calculating the Kappa Statistic

$$\text{Kappa} = \frac{P_O - P_E}{1 - P_E}$$

where:

P_O = proportion of observed agreements

P_E = proportion of agreements expected by chance

30

intuition for computing the Kappa Score for ML problems

$$Kappa\ score = 1 - \frac{(1 - accuracy)}{(1 - accuracy\ due\ to\ chance)}$$

$$= \frac{(accuracy\ - accuracy\ due\ to\ chance)}{(1 - accuracy\ due\ to\ chance)}$$

31

How to compute P_E?

P_E is the probability of the occurrence of two disjoint events

Predicting +ve by chance and predicting –ve by chance

For predicting +ve by chance, two events must happen:

The prediction must be +ve for the instance and
The actual class must be +ve (like the coin bias)

$$P_+ = probability\ of\ +ve\ prediction * probability\ of\ actual\ classs\ being\ +ve$$

32

Formula for calculating the Kappa Statistic without need for proportions

Predicted	Actual +	Actual -	Total
+	*a*	*b*	p_1
-	*c*	*d*	q_1
Total	p_2	q_2	*N*

$$\kappa = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}$$

33

Why?

$$P_0 = \frac{a+d}{a+b+c+d}$$ Numerator: Observed matches (correct predictions)

$$P_+ = \frac{a+b}{a+b+c+d} * \frac{a+c}{a+b+c+d}$$ Assuming actual and predicted are independent of each other

$$P_- = \frac{b+d}{a+b+c+d} * \frac{c+d}{a+b+c+d}$$

$$P_e = P_+ + P_-$$

Substitute and Simplify

Predicted	Actual +	Actual -
+	*a*	*b*
-	*c*	*d*

34

Formula for calculating the Kappa Statistic

Predicted	Actual +	Actual -	Total
+	*a*	*b*	p_1
-	*c*	*d*	q_1
Total	p_2	q_2	N

$$\kappa = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}$$

Predicted	Actual +	Actual -	Total
+	44	4	48
-	6	46	52
Total	50	50	100

$$\mathrm{K} = \frac{2[44*46 - 4*6]}{48*50+52*50} = 0.8$$

Accuracy = (44+46)/100 = 0.9

35

Makes a difference when the dataset is imbalanced!

Predicted	Actual +	Actual -	Total
+	84	4	88
-	6	6	12
Total	90	10	100

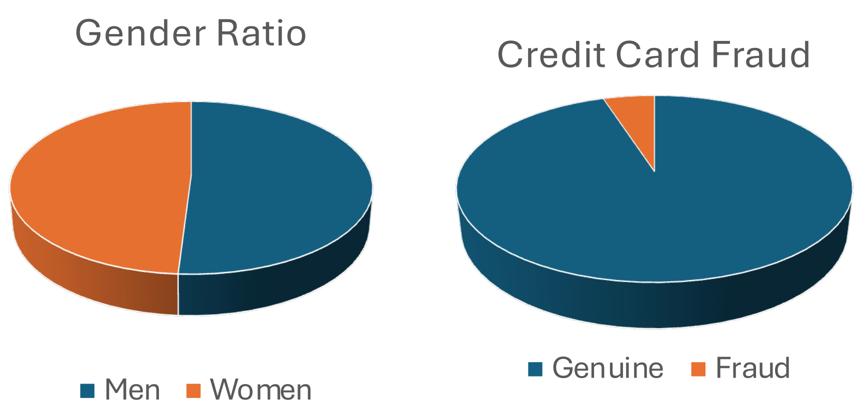
$$\kappa = \frac{2(ad - bc)}{p_1 q_2 + p_2 q_1}$$

$$\mathrm{K} = \frac{2[84*6 - 4*6]}{88*90+12*10} = 0.12$$

Accuracy = (84+6)/100 = 0.9

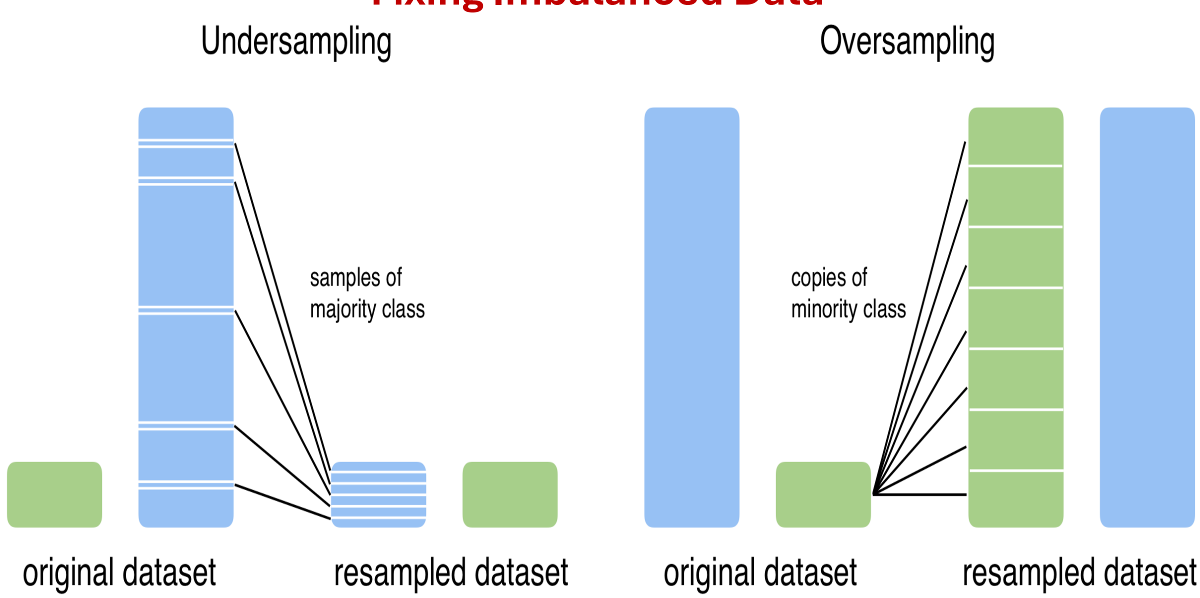
36

Balanced vs Imbalanced data



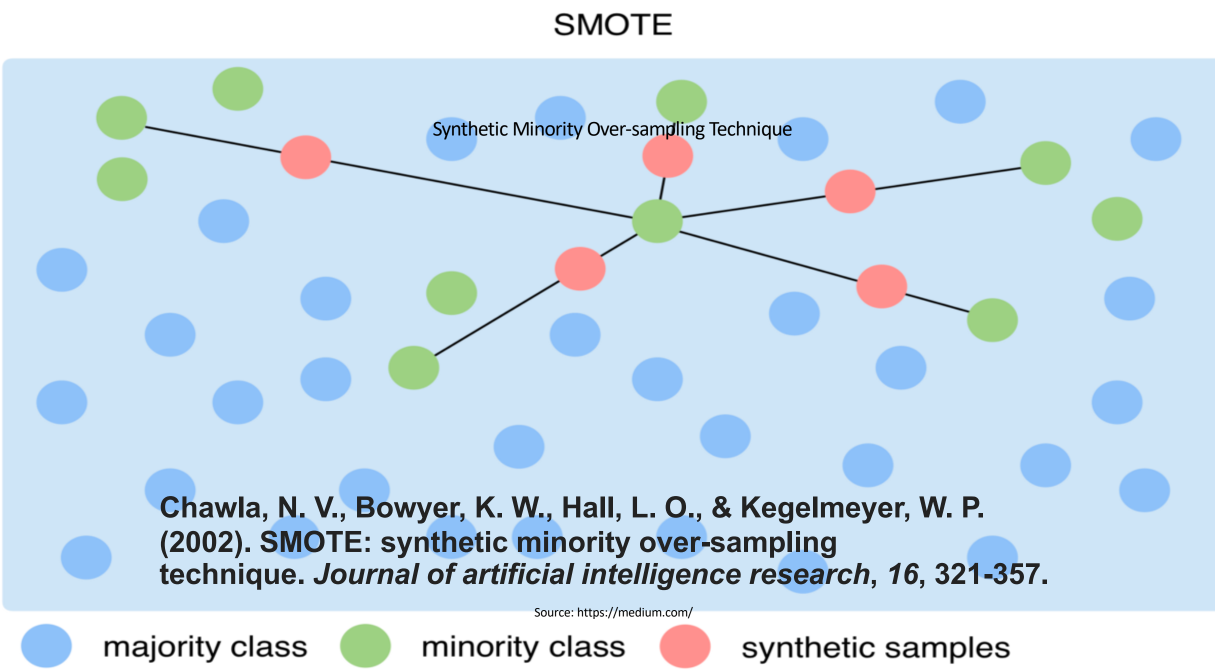
37

Fixing Imbalanced Data



Source: https://medium.com/

38



Synthetic Minority Over-sampling Technique

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Source: https://medium.com/

39

Adaptive Synthetic Sampling (ADASYN)

He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008.

a. Calculate the degree of class imbalance:

$d = \frac{m_s}{m_l}$ where $d \in (0, 1]$

while $d < d_{threshold}$

Calculate the number of synthetic data examples that need to be generated for the minority class:

$G = (m_l - m_s) \times \beta$ where $\beta \in [0, 1]$

40

How do you identify the minority class samples that are difficult to classify?

They are closest to the decision boundary (in the region of disagreement)

Surrounded by samples of the majority class

ADASYN factors this into the choice of minority class points

41

Adaptive Synthetic Sampling (ADASYN)

He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008.

b) For each example $x_i \in$ minorityclass, find K nearest neighbors based on the Euclidean distance in n dimensional space, and calculate the ratio r_i defined as:

$r_i = \Delta_i/K$, $i = 1, \ldots, m_s$ where Δ_i is the number of examples in the K nearest neighbors of x_i that belong to the majority class, therefore $r_i \in [0, 1]$;

c) Normalize r_i according to $\hat{r_i} = \frac{r_i}{\sum_{i=1}^{m_s} r_i}$ so that $\hat{r_i}$ is a density distribution, $\sum \hat{r_i} = 1$

42

Adaptive Synthetic Sampling (ADASYN)

(d) Calculate the number of synthetic data examples that need to be generated for each minority example $\boldsymbol{x}_i$:

$$g_i = \hat{r}_i \times G \quad (4)$$

where G is the total number of synthetic data examples that need to be generated for the minority class as defined in Equation (2).

(e) For each minority class data example $\boldsymbol{x}_i$, generate g_i synthetic data examples according to the following steps:

Do the **Loop** from 1 to g_i:

(i) Randomly choose one minority data example, $\boldsymbol{x}_{zi}$, from the K nearest neighbors for data $\boldsymbol{x}_i$.

(ii) Generate the synthetic data example:

$$\boldsymbol{s}_i = \boldsymbol{x}_i + (\boldsymbol{x}_{zi} - \boldsymbol{x}_i) \times \lambda \quad (5)$$

where $(\boldsymbol{x}_{zi} - \boldsymbol{x}_i)$ is the difference vector in n dimensional spaces, and λ is a random number: $\lambda \in [0,\ 1]$.

End **Loop**

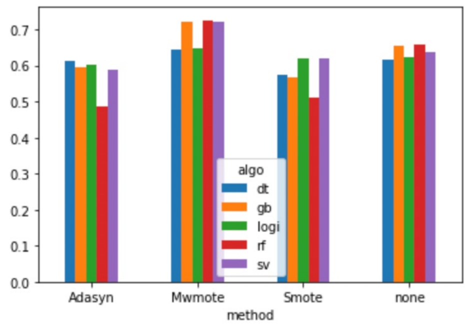
43

Pendyala, Vishnu S., and HyungKyun Kim. "Analyzing and Addressing Data-driven Fairness Issues in Machine Learning Models used for Societal Problems." *2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE)*. IEEE, 2023.

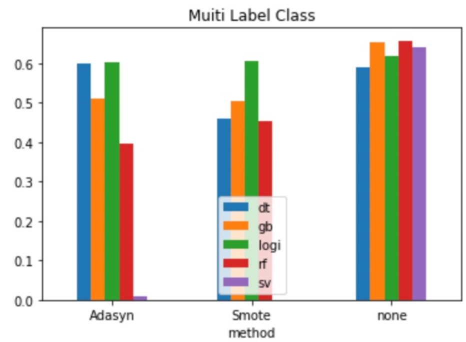
"The experiments also demonstrate that some of the oversampling techniques can degrade the models both in terms of performance and fairness"

44

Performance of ML Algorithms: F1-Score



Histogram of F-1 scores for binary categorization

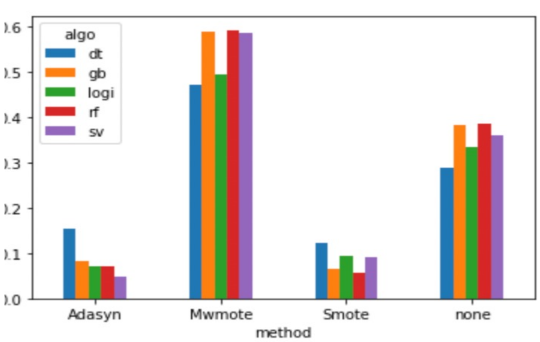


Histogram of F-1 scores for multiple ethnicity categorization

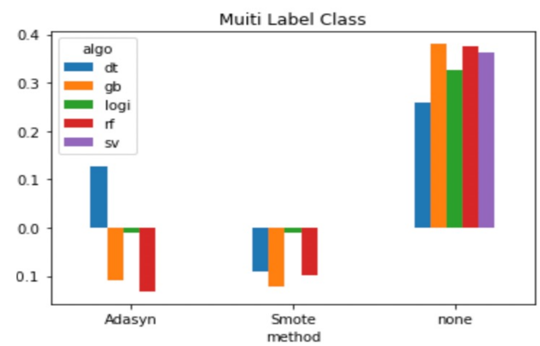
Source: Pendyala, Vishnu S., and HyungKyun Kim. "Analyzing and Addressing Data-driven Fairness Issues in Machine Learning Models used for Societal Problems." *International Conference on Computer, Electrical & Communication Engineering (ICCECE)*. IEEE, 2023.

45

Cohen's Kappa Statistic does it better than F-1



Histogram of Kappa statistic for binary categorization



Histogram of Kappa Statistic for multiple ethnicity categorization

Source: Pendyala, Vishnu S., and HyungKyun Kim. "Analyzing and Addressing Data-driven Fairness Issues in Machine Learning Models used for Societal Problems." *International Conference on Computer, Electrical & Communication Engineering (ICCECE)*. IEEE, 2023.

46

Other Classification Metrics

G-measure: geometric mean of precision and recall

Informedness / Youden's J statistic / Youden's Index = Sensitivity + Specificity – 1

- A value of 1 indicates perfect classification performance
- 0 => performance no better than random chance
- A value below 0 suggests that the model's performance is worse than random chance

Markedness = PPV + NPV – 1

- Positive Predictive Value = TP / (TP + FP)
- Negative Predictive Value = TN / (TN + FN)

MCC combines Informedness and Markedness (next slide)

47

Matthew's correlation coefficient (MCC)

$$\text{MCC} = \frac{(TP.TN)-(FP.FN)}{\sqrt{(TP+FN).(TP+FP).(TN+FN).(TN+FP)}}$$

Range: [-1 (perfect misclassification), +1 (perfect classification)]

Undefined when the whole row or column of a confusion matrix is 0: TP=FP=0 or TN=FN=0, etc

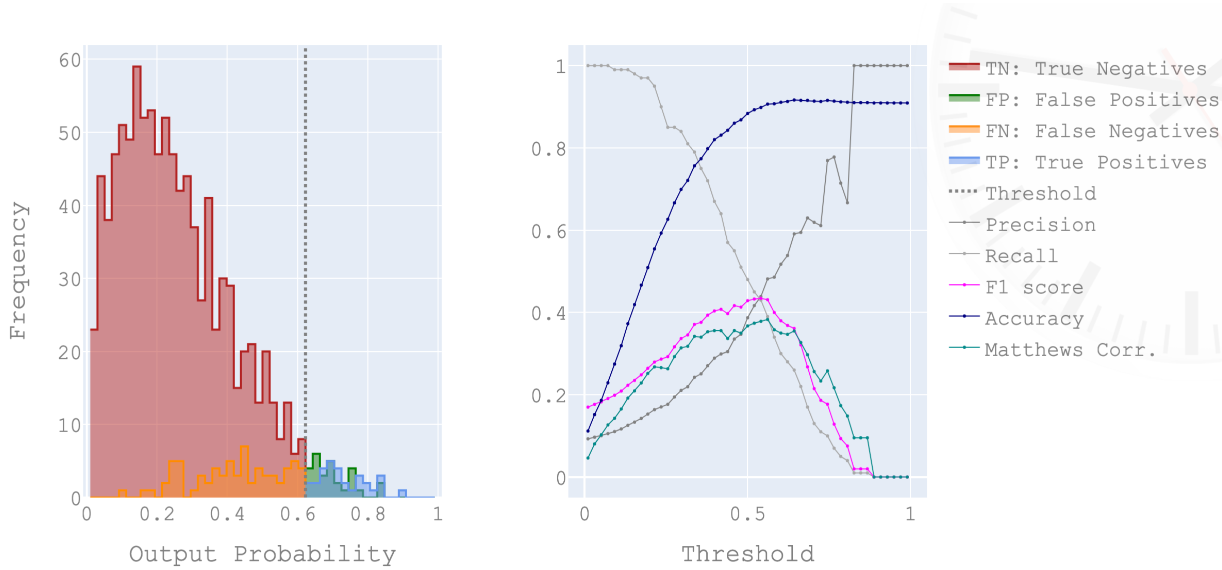
MCC=0 for a coin tossing classifier (perfectly random prediction)

Balanced measure: includes all four elements of the confusion matrix

Often preferred over F1 score

48

Performance of various metrics on imbalanced data



Source: https://felipepenha.github.io/data-science-bits/performance_metrics/Matthews_correlation_unbalanced.html

49

Chicco, D., & Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. BioData Mining, 16(1), 4.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9938573

“we explain why the Matthews correlation coefficient should replace the ROC AUC as standard statistic in all the scientific studies involving a binary classification, in all scientific fields.”

50

Zhu, Q. (2020). On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. Pattern Recognition Letters, 136, 71-80.

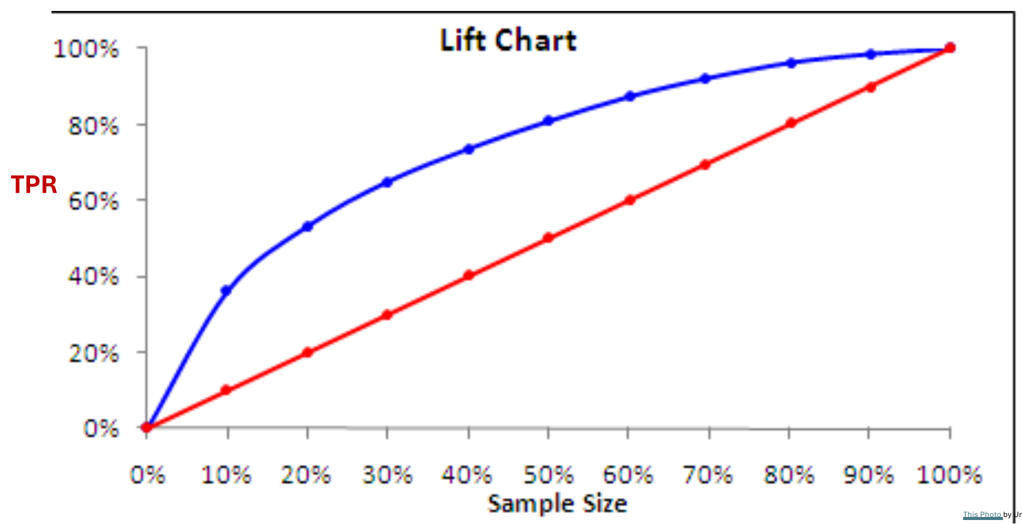
" It has been generally regarded as a balanced measure which can be used even if the classes are of very different sizes. The study of this paper finds that this is not true. MCC deteriorates seriously when the dataset in classification are imbalanced. Experiment results and analysis show that MCC is not suitable for classification accuracy measurement on imbalanced datasets."

51

How well did the model perform in identifying TP?

TP	FP
FN	TN

$$\text{Lift} = \frac{sensitivity\ of\ the\ model}{Ratio\ of\ actual\ positive} = \frac{\text{(TP/(TP+FP)}}{\text{(TP+FN)/(TP+TN+FP+FN)}}$$



52

Regression Metrics

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

53

CASE 1: Evenly distributed errors

ID	Error	|Error|	Error^2
1	2	2	4
2	2	2	4
3	2	2	4
4	2	2	4
5	2	2	4
6	2	2	4
7	2	2	4
8	2	2	4
9	2	2	4
10	2	2	4

CASE 2: Small variance in errors

ID	Error	|Error|	Error^2
1	1	1	1
2	1	1	1
3	1	1	1
4	1	1	1
5	1	1	1
6	3	3	9
7	3	3	9
8	3	3	9
9	3	3	9
10	3	3	9

CASE 3: Large error outlier

ID	Error	|Error|	Error^2
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0	0	0
10	20	20	400

Source: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

MAE	RMSE
2.000	2.000

MAE	RMSE
2.000	2.236

MAE	RMSE
2.000	6.325

54

Regression Metrics: R^2

Total Sum of Squares

$$\text{TSS} = \sum(y_i - \bar{y})^2$$

Inherent variability in *y* before the prediction

Residual Sum of Squares

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

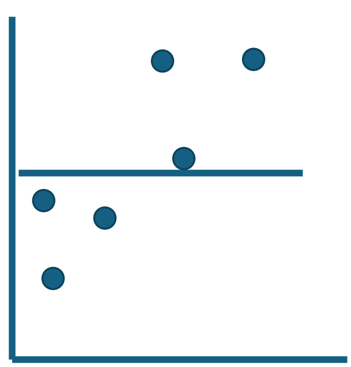
Unexplained variation after applying regression

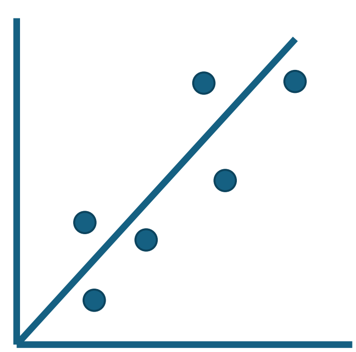
$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

R^2 is the proportion of variance in Y explained using X
R^2 = 0 => not much of the observed variation is explained
=> Model may not be correct or inherent variability is high

55

R^2 measures mean vs regression





56

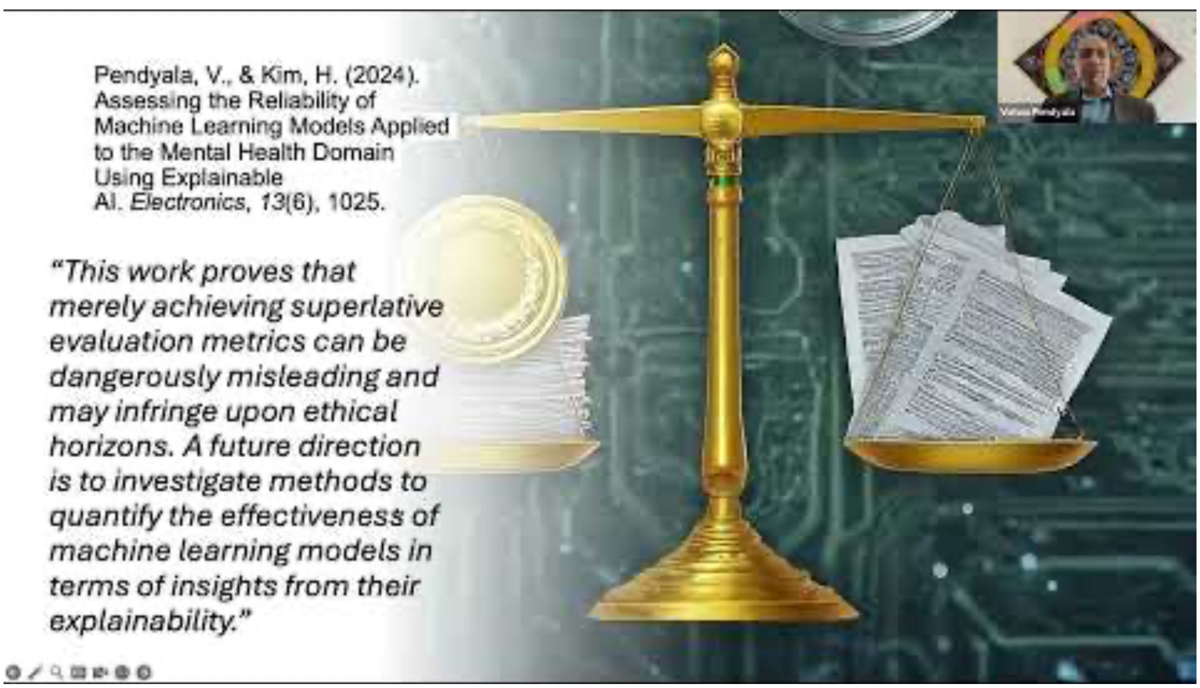
Regression Metrics: F Statistic

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)}$$

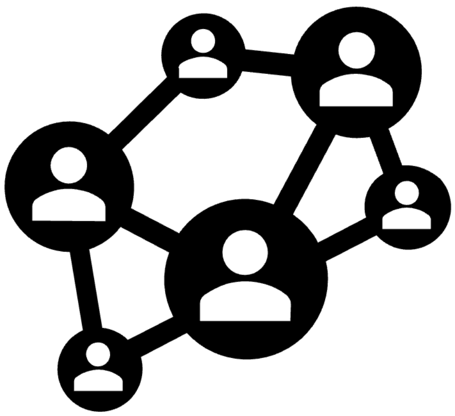
Measures the relationship between X and Y
F=1 => no relationship; otherwise > 1
H0: There is no relationship
Small n requires large F to reject H0
Large n => F slightly > 1 enough to reject H0
Called F because it follows F-distribution when H0 is true and the errors are normally distributed

p is the # of predictors / columns / features

57



58



Stay in touch!

https://twitter.com/VishnuPendyala

https://www.facebook.com/vishnu.pendyala

https://www.instagram.com/vishnupendyala/

https://www.threads.net/@vishnupendyala

https://www.linkedin.com/in/pendyala/

59

Questions?

60