AEROJET ROCKETDYNE

Assessing Predictive Capability and Contribution for Binary Classification Models

Mindy Hotchkiss

Defense and Aerospace Test and Analysis Workshop (DATAWorks) https://dataworks.testscience.org/

April 25 – 27, 2023

Institute for Defense Analysis, Alexandria, VA

Abstract

Classification models for binary outcomes are in widespread use across a variety of industries. Results are commonly summarized in a misclassification table, also known as an error or confusion matrix, which indicates correct vs incorrect predictions for different circumstances. Models are developed to minimize both false positive and false negative errors, but the optimization process to train/obtain the model fit necessarily results in cost-benefit trades. However, how to obtain an objective assessment of the performance of a given model in terms of predictive capability or benefit is less well understood, due to both the rich plethora of options described in literature as well as the largely overlooked influence of noise factors, specifically class imbalance. Many popular measures are susceptible to effects due to underlying differences in how the data are allocated by condition, which cannot be easily corrected.

This talk considers the wide landscape of possibilities from a statistical robustness perspective. Results are shown from sensitivity analyses for a variety of different conditions for several popular metrics and issues are highlighted, highlighting potential concerns with respect to machine learning or ML-enabled systems. Recommendations are provided to correct for imbalance effects, as well as how to conduct a simple statistical comparison that will detangle the beneficial effects of the model itself from those of imbalance. Results are generalizable across model type.

Approved for Public Release / This document contains no ITAR or EAR controlled technical information.

Overview

- Introduction & Terminology
- Motivating Example
- Issues & Concerns
- Implications for Machine Learning
- Candidate Metrics
- Sensitivity Study Results
- Alternative Approaches
- Summary & Recommendations

Approved for Public Release / This document contains no ITAR or EAR controlled technical information.

What are Classification Models?

- A mathematical function f used to predict outcome **Y**

$$f(\mathrm{X}) \Rightarrow \hat{\mathbf{Y}}$$

from inputs **X**

where **Y** is *NOT* continuous and takes on only a limited discrete set of values

- Binary: 2 levels, denoting any dichotomy
- Ordinal: 3+ classifications with natural sequencing
- Nominal / Categorical: unordered categories or classifications

- Standard regression models not applicable
  - Need a link function to convert to probability scale
- Results are summarized in an error matrix
  - Also designated contingency table, truth table, confusion matrix, misclassification table…

Raw Data Table

X1	…	Xk	Actual Y	Predicted Y
0.19	…	0.59	0	0
0.03	…	0.27	0	1
0.62	…	0.87	1	0
0.75	…	0.58	1	1
⋮	⋮	⋮	⋮	⋮

Binary Classification

		Model Prediction	
		No	Yes
Actual Condition	No	# True Negatives	# False Positives
	Yes	# False Negatives	# True Positives

Models must balance error types:

- **False Positive** – model is over-sensitive, detects too often
- **False Negative** – model is under-sensitive, fails to detect

How are Classification Models Assessed?

From this table, many measures can be calculated:

Binary Classification

		Model Prediction: No	Model Prediction: Yes	Row Totals
Actual Condition	No	**TN** (# True Negatives)	**FP** (# False Positives)	**AN** (Total # Actual Negatives)
	Yes	**FN** (# False Negatives)	**TP** (# True Positives)	**AP** (Total # Actual Positives)
Col Totals		**PN** (Total # Predicted Negatives)	**PP** (Total # Predicted Positives)	**N** (Total)

Prevalence $= \frac{AP}{N}$

* Class Imbalance *

False Positive Rate (FPR = 1- TNR)

$$= \frac{FP}{AN} = \frac{FP}{FP + TN}$$

Fall-Out
Probability of False Alarm

True Negative Rate (TNR)

$$= \frac{TN}{AN} = \frac{TN}{FP + TN}$$

Specificity
Selectivity

True Positive Rate (TPR)

$$= \frac{TP}{AP} = \frac{TP}{TP + FN}$$

Recall
Sensitivity
Hit Rate
Power
Probability of Detection (POD)

False Negative Rate (FNR = 1- TPR)

$$= \frac{FN}{AP} = \frac{FN}{TP + FN}$$

Miss Rate

Negative Predictive Value (NPV)

NPV $= \frac{TN}{PN} = \frac{TN}{FN + TN}$

Positive Predictive Value (PPV)

$= \frac{TP}{PP} = \frac{TP}{TP + FP}$ *Precision*

False Omission Rate (FOR = 1- NPV)

$$= \frac{FN}{PN} = \frac{FN}{FN + TN}$$

False Discovery Rate (FDR = 1- PPV)

$$= \frac{FP}{PP} = \frac{FP}{TP + FP}$$

These are individual **intermediate calculations**, incomplete measures for **model characterization or performance assessment**

How are Classification Models Assessed?

(an incomplete list)

		Model Prediction: No	Model Prediction: Yes	Row Totals
Actual Condition	No	TN (# True Negatives)	FP (# False Positives)	AN (Total # Actual Negatives)
	Yes	FN (# False Negatives)	TP (# True Positives)	AP (Total # Actual Positives)
Prevalence $= \frac{AP}{N}$	Col Totals	PN (Total # Predicted Negatives)	PP (Total # Predicted Positives)	N (Total)

False Positive Rate (FPR = 1- TNR) $= \frac{FP}{AN} = \frac{FP}{FP+TN}$ — *Fall-Out, Probability of False Alarm*

True Negative Rate (TNR) $= \frac{TN}{AN} = \frac{TN}{FP+TN}$ — *Specificity, Selectivity*

True Positive Rate (TPR) $= \frac{TP}{AP} = \frac{TP}{TP+FN}$ — *Recall, Sensitivity, Hit Rate, Power, Probability of Detection (POD)*

False Negative Rate (FNR = 1- TPR) $= \frac{FN}{AP} = \frac{FN}{TP+FN}$ — *Miss Rate*

Functions of Row Proportions: Recall (Sensitivity) & Specificity

- Balanced Accuracy (BA) Arithmetic mean of Recall (Sensitivity) & Specificity
- Bookmaker Informedness (BMI), Youden's J
  = Recall + Specificity -1
  = 2*BA -1 (Scaled BA)

Functions of All Row & Column Proportions: Recall(Sensitivity), Specificity, NPV, & Precision

- Diagnostic Odds Ratio (DOR, OR), theta, cross-product ratio, Log DOR
- Yules Q, Y are scaled versions of theta
- Matthews Correlation Coefficient (MCC), Pearson's phi coefficient, also = Yules Y

Functions of Column Proportions: NPV & Precision

- Markedness (Mk) Δp
  = PPV(Precision) + NPV -1

Calculated analogously to BMI & YJ except relative to columns instead of rows. Equivalently, could be scaled.

Functions of Mixed Row & Column Proportions: Recall (Sensitivity) & Precision

- Area Under the Precision vs Recall Curve (AUC)
- Gini Coefficient = 2*AUC -1 (Scaled AUC)
- F-Score (F-Measure)
  F_1: Harmonic Mean of Precision & Recall
  F_β: F_1 weighted to accommodate detection error trades
- Fowlkes-Mallow Index: Geometric Mean of Precision & Recall

***** Omit True Negatives from all calculations*****

Functions of Neither Row nor Column Proportions

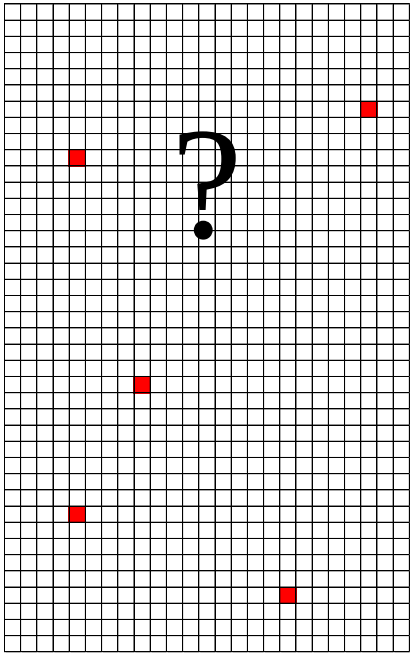
- Overall Accuracy

$$OA = \frac{TN+TP}{N}$$

Not all metrics are useful measures of model performance

Rare Event Prediction

Example: 1000 predictions



True incidence rate: 5/1000

- Objective is to successfully predict the outcome for each instance
- A high success rate can easily be achieved
  … *without doing any actual modeling*
- Predict Y = No <u>for all</u> instances (“null” or “naïve” model)

		Model Prediction		
		No	Yes	Total
Actual Condition	No	995	0	995
	Yes	5	0	5
	Total	1000	0	1000

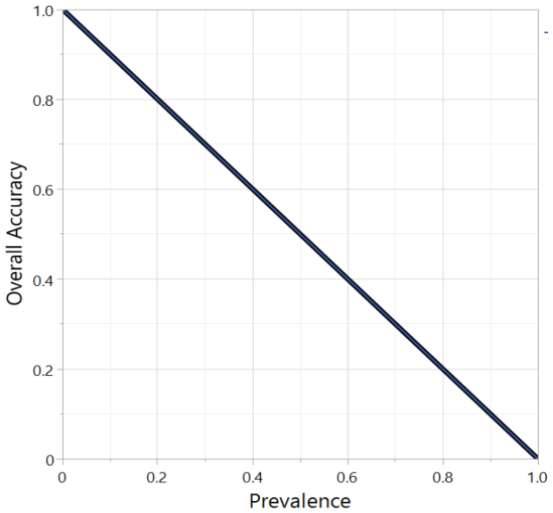
→ Overall Accuracy = $\frac{995+0}{1000}$ = **99.5%**

But this assessment is clearly meaningless – this “model” provides no information

Rare Event Prediction Implications

- If there is no value-added contribution from the model,

Overall Accuracy = 1 – Prevalence



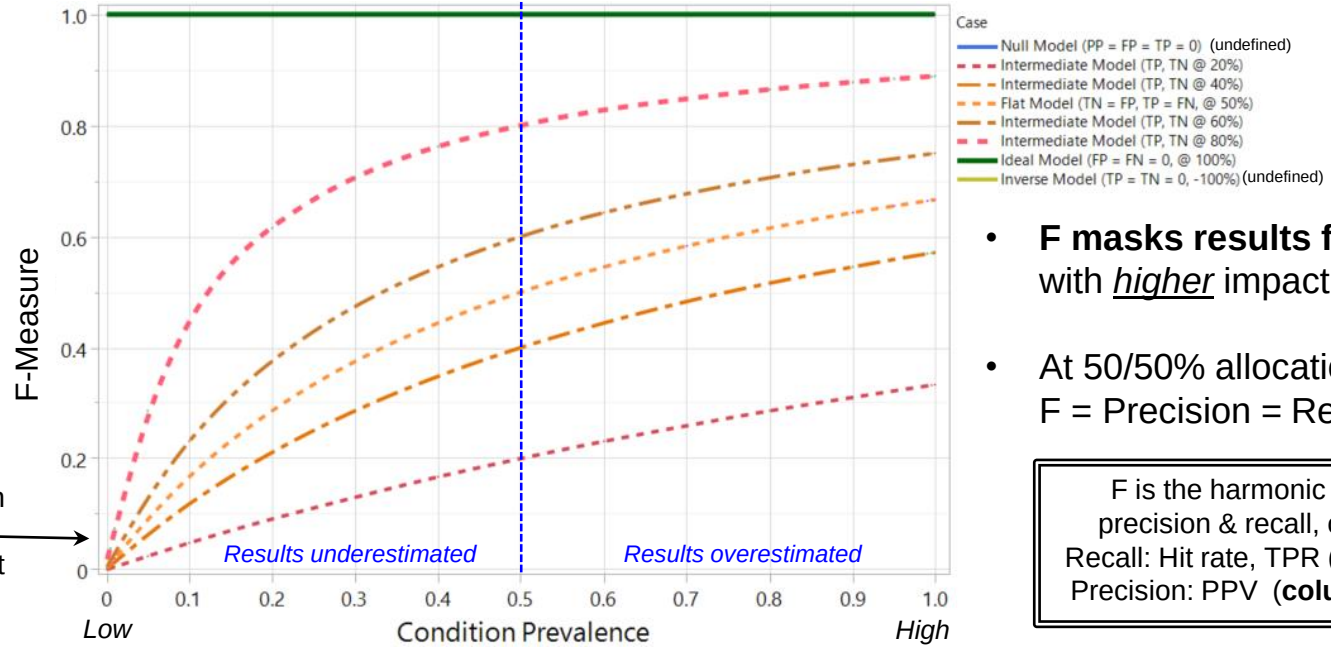
Class Imbalance Proportion

- If model effects are present, OA provides no way to disentangle its contribution from imbalance effects
- A good model will have high OA, but high OA itself is NOT an indicator of a good model
  - .. *necessary* but not *sufficient*
- OA is a *trailing* indicator, not a *leading* indicator

Overall Accuracy provides no insight into actual model performance

Class Imbalance Affects Other Metrics

- Sensitivity study results: F-measure also shows variable dependencies



Nonlinearities become more extreme for certain cases due to asymmetry in what data are excluded.

- **F masks results for rarer events**, with *higher* impact on *better* results
- At 50/50% allocation, F = Precision = Recall

F is the harmonic mean of precision & recall, omits TN.
Recall: Hit rate, TPR (**row** metric)
Precision: PPV (**column** metric)

Undesirable properties generalizable to similarly derived functions

Robustness

- Statistical measures are tools with their own properties, that need to meet certain suitability criteria
- Tools that are more “robust” will be more broadly useful, less affected by violations in assumptions and usage conditions, but still providing a correct assessment
- Model assessment metrics should provide the SAME answer ***regardless of changes in “noise” factors that are unimportant***
- Class imbalance is unimportant, not integral to any analysis, and it should NOT drive conclusions

Implications for Machine Learning

- Class model assessment metrics in widespread use today have flaws that implicate results in majority of cases
- ML applications involving classification models should also have used a model assessment metric in development and during implementation if embedded into a larger more integrated system
- Down-stream decision-making based on these class model assessments assume that the metrics provide a fair and independent assessment of model performance when they do not
- Assessments of actual value contributed by classification models developed are highly likely to have been misrepresented
- Potential downstream effects include poor repeatability, increased inefficiencies, and loss of system credibility, as well as lost opportunities

Incorporating more robust class model metrics and assessments will provide greater chance of system success

Directionality affects Metric Robustness

Cross-Directional

		Model Prediction		
		No	Yes	Total
Actual Condition	No	True Negative (TN)	False Positive (FP)	Actual Negative (AN)
	Yes	False Negative (FN)	True Positive (TP)	Actual Positive (AP)
	Total	Predicted Negative (PN)	Predicted Positive (PP)	Total (N)

Metrics calculated relative to one column and one row, with some element omitted.

Unidirectional

		Model Prediction		
		No	Yes	Total
Actual Condition	No	True Negative (TN)	False Positive (FP)	Actual Negative (AN)
	Yes	False Negative (FN)	True Positive (TP)	Actual Positive (AP)
	Total	Predicted Negative (PN)	Predicted Positive (PP)	Total (N)

Metrics calculated relative to Columns

Unidirectional

		Model Prediction		
		No	Yes	Total
Actual Condition	No	True Negative (TN)	False Positive (FP)	Actual Negative (AN)
	Yes	False Negative (FN)	True Positive (TP)	Actual Positive (AP)
	Total	Predicted Negative (PN)	Predicted Positive (PP)	Total (N)

Metrics calculated relative to Rows

Bidirectional

		Model Prediction		
		No	Yes	Total
Actual Condition	No	True Negative (TN)	False Positive (FP)	Actual Negative (AN)
	Yes	False Negative (FN)	True Positive (TP)	Actual Positive (AP)
	Total	Predicted Negative (PN)	Predicted Positive (PP)	Total (N)

Metric calculated interchangeably relative to both rows AND columns

Balance is important: preferred metrics are in bidirectional family

Fully Balanced Functions

- **Correlation: Phi, Matthews Correlation Coefficient (MCC)**
  - Range [-1,1], can be calculated on raw 0-1 data using standard Pearson correlation function
  - Binary tabular formulation (one version):

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

Binary Case Layout

		Model Prediction: No	Model Prediction: Yes	Total
Actual Condition	No	True Negative (TN)	False Positive (FP)	Actual Negative (AN)
	Yes	False Negative (FN)	True Positive (TP)	Actual Positive (AP)
	Total	Predicted Negative (PN)	Predicted Positive (PP)	Total (N)

Metrics calculated using both rows AND columns

Functions of All Row & Column Proportions: Recall(Sensitivity), Specificity, NPV, & Precision

- Diagnostic Odds Ratio (DOR, OR), theta, cross-product ratio, Log DOR
- Yules Q, Y are scaled versions of theta
- Matthews Correlation Coefficient (MCC), Pearson's phi coefficient, also = Yules Y

- **(Diagnostic) Odds Ratio (DOR, OR), theta (θ), cross-product ratio**
  - Range [0, inf), 1 if no effect, undefined when FP, FN = 0
  - Can be written equivalently in terms of *either* row or columns

$$\mathrm{OR} = \frac{LR+}{LR-} = \frac{\frac{TPR}{FPR}}{\frac{FNR}{TNR}} = \frac{\frac{sensititivity}{1-specificity}}{\frac{1-sensitivity}{specificity}} = \frac{\frac{TP/_{AP}}{FP/_{AP}}}{\frac{TN/_{AN}}{FN/_{AN}}} = \frac{TP/_{FP}}{FN/_{TN}} = \frac{TP \cdot TN}{FP \cdot FN} = \frac{TP/_{FN}}{FP/_{TN}} = \frac{\frac{TP/_{PP}}{FN/_{PN}}}{\frac{FP/_{PP}}{TN/_{PN}}} = \frac{\frac{PPV}{FOR}}{\frac{FDR}{NPV}} = \frac{\frac{PPV}{1-NPV}}{\frac{1-PPV}{NPV}} = \theta$$

$$= \frac{sensitivity \cdot specificity}{(1-sensitivity) \cdot (1-specificity)} = \frac{PPV \cdot NPV}{(1-PPV) \cdot (1-NPV)}$$

  - **Variants Yule's Q, Y** *(coefficient of colligation)*
    - Functions of OR, converted to [-1,1] scale

$$Q = \frac{(\theta-1)}{(\theta+1)};\ Y = \frac{(\sqrt{\theta}-1)}{(\sqrt{\theta}+1)}$$

*undefined when θ = -1, NA since θ E [0, inf)

Yule's Y conversion
= Delta Rate
= TNR-FPR

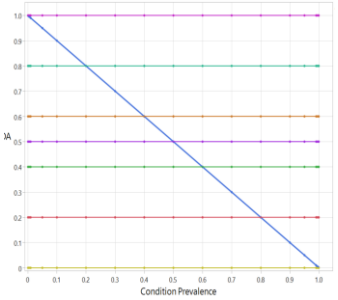
Sensitivity Study Comparisons & Results

AEROJET ROCKETDYNE

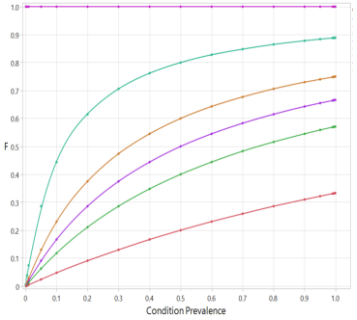
Case
- Null Model (PP = FP = TP = 0)
- Intermediate Model (TP, TN @ 20%)
- Intermediate Model (TP, TN @ 40%)
- Flat Model (TN = FP, TP = FN, @ 50%)
- Intermediate Model (TP, TN @ 60%)
- Intermediate Model (TP, TN @ 80%)
- Ideal Model (FP = FN = 0, @ 100%)
- Inverse Model (TP = TN = 0, -100%)

*All cases defined to have specific OA

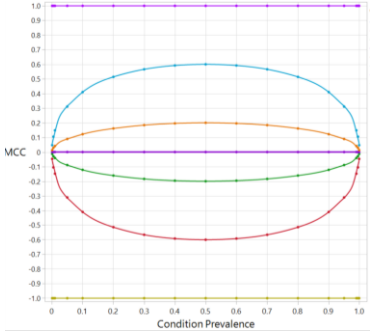
Overall Accuracy



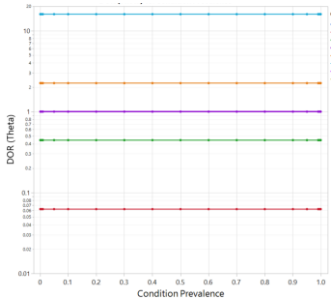
F-measure



MCC



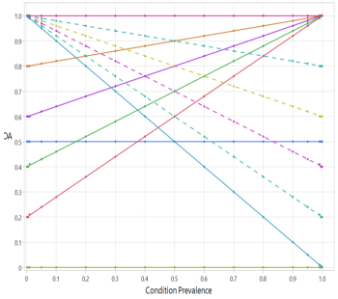
DOR (logscale)

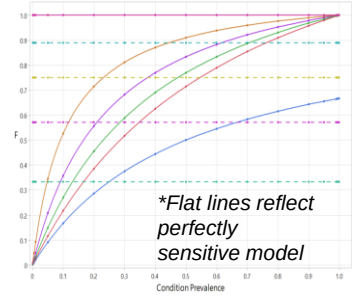


*Flat lines reflect good behavior

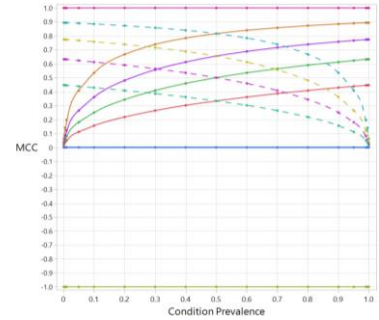
Case
- Flat Model (TN = FP, TP = FN, @ 50%)
- Highly Sensitive Model (FN = 0, TN @ 20%)
- Highly Sensitive Model (FN = 0, TN @ 40%)
- Highly Sensitive Model (FN = 0, TN @ 60%)
- Highly Sensitive Model (FN = 0, TN @ 80%)
- Highly Specific Model (FP = 0, TP @ 20%)
- Highly Specific Model (FP = 0, TP @ 40%)
- Highly Specific Model (FP = 0, TP @ 60%)
- Highly Specific Model (FP = 0, TP @ 80%)
- Ideal Model (FP = FN = 0, @ 100%)
- Inverse Model (TP = TN = 0, -100%)
- Null Model (PP = FP = TP = 0)

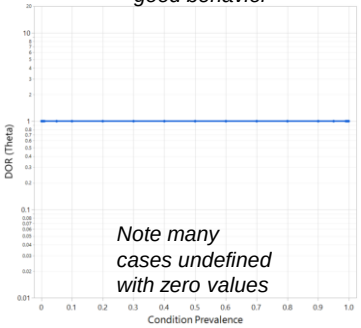
* Incorporated cases with unbalanced error prediction rates





*Flat lines reflect perfectly sensitive model





Note many cases undefined with zero values

Metrics calculated for hypothetical cases, varying class imbalance

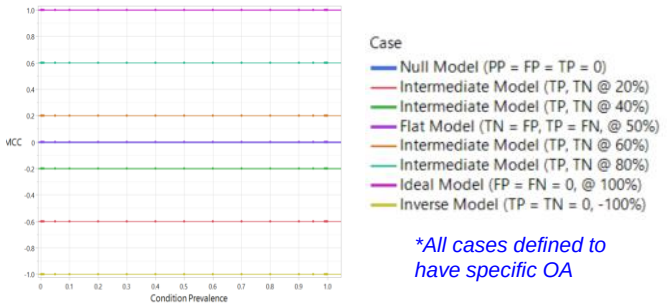
Modified MCC

- Alternative calculation approaches can improve MCC
  - Scale every cell element to be out of 100%, calculate as a function of proportions
  - Or divide by AP^2*AN^2 instead of standard denominator

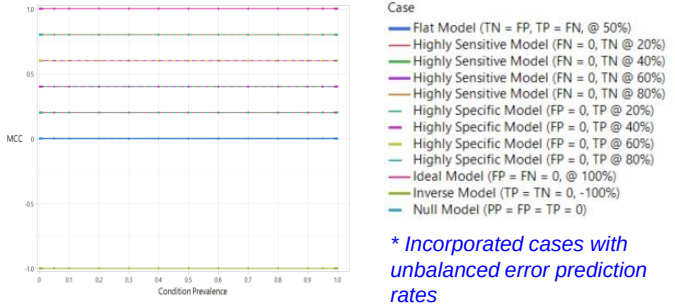
$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$$

$$MCC^{*mod} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)^2 \cdot (TN+FP)^2}} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{AP^2 \cdot AN^2}}$$

Modified MCC/Delta Rates



*All cases defined to have specific OA



* Incorporated cases with unbalanced error prediction rates

Adjustments stabilize MCC over full range of prevalence, equivalent to difference between True Positive to False Negative Rates

What is a "Good" Model?

- Adds value over the null model, provides some improved capability to differentiate between classes

Worthless Model – No Value Added

Model predicts same proportion for hit rate and false positive rate

		Model Prediction			*Row Proportion*
		No	Yes	Total	
Actual Condition	No	810	90	900	10%
	Yes	90	10	100	10%
	Total	900	100	1000	

FPR
TPR
$\Delta = 0$

Ideal Model – Provides Perfect Solution

Model predicts different proportions hit rate >>> false positive rate

		Model Prediction			*Row Proportion*
		No	Yes	Total	
Actual Condition	No	900	0	900	0%
	Yes	0	100	100	100%
	Total	900	100	1000	

FPR
TPR
$\Delta = 100\%$

FPR = False alarm rate, False Positive Rate
TPR = Recall, sensitivity, Probability of Detection (POD), hit rate

Focus on maximizing the difference between True & False Positive Rates

Importance can be assessed with standard statistical t-test of 2 proportions

Summary and Recommendations

- Many popular metrics used to assess classifier performance are affected by class imbalance, so impacts are widespread but unrecognized

- Be skeptical of claims made about classification model performance, especially if given in terms of “accuracy”

- A “good” model is one that adds value, provides some improved capability to differentiate between classes – and effects can be quantified

- Modifications can be made to stabilize metrics over full range of possible prevalence values, but model quality can be best quantified & assessed for statistical significance through the comparison of the true positive to true negative rates

Questions?

AEROJET ROCKETDYNE