Short Explainer: Accuracy, Confusion Matrix, and AUC (Area Under the Curve)

Data Science Team

Andrew Wheeler, PhD

andrew.wheeler@hms.com

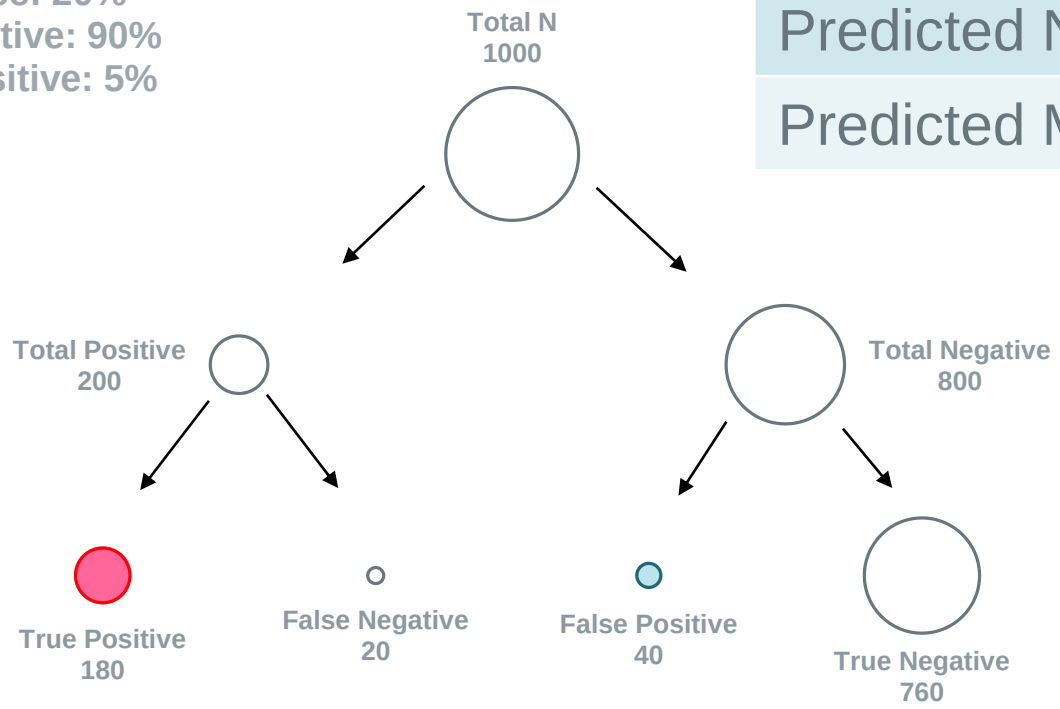
Confusion Matrix

	Actual False	Actual True
Predicted False	True Negative (TN)	False Negative (FN)
Predicted True	False Positive (FP)	True Positive (TP)

- Hypothetical Example:
- Score 1,000 claims for probability it is an overpayment:
  - **True Positive Rate** of model is 90% (proportion of actual overpayments we capture)
  - **False Positive Rate** is 5% (proportion of cases that aren't overpayments our model incorrectly flags as overpayments)
  - **Prevalence** of match is 20% (overall proportion of overpayments, positive mix %)
  - **Accuracy** is the proportion of cases we predict correctly, (TP + TN)/Cases

Hypothetical Example

Prevalence: 20%
True Positive: 90%
False Positive: 5%



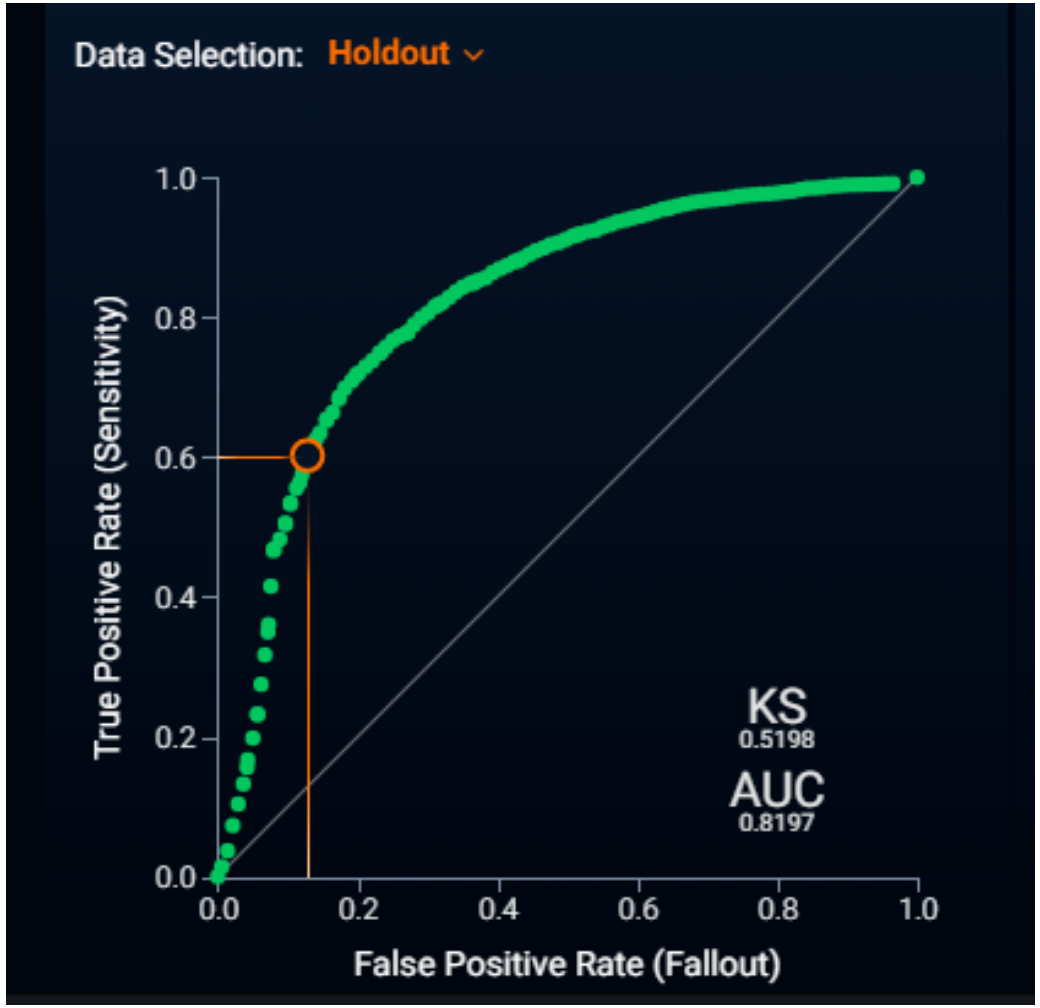
	No Match	Match
Predicted No Match	760	20
Predicted Match	40	180

Accuracy is 94%:

(760 + 180)/1000

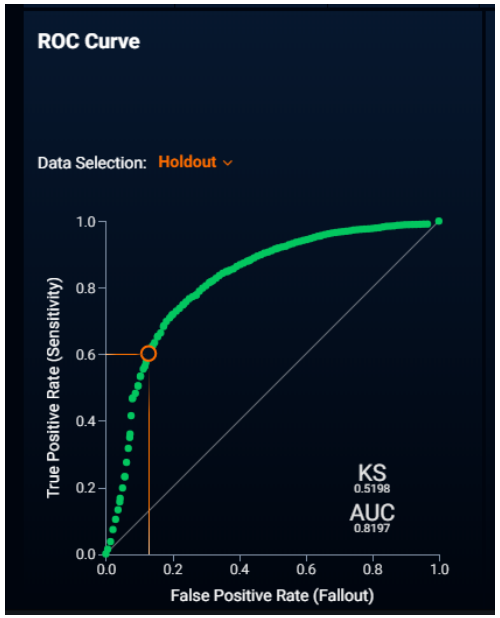
Area Under The Curve (AUC)

- If you flag more claims, you will capture more true positives, but will increase false positives
- ROC curves show this trade-off
- AUC is the area under the curve.
  - 1 is perfect
  - will get 0.5 with random guessing
- Is AUC = 0.82 good enough? Depends on costs/benefits of false positives/true positives
- For cases with extremely low positive mix % (e.g. 5% positive), there might be many more *false positives* than *true positives. In this case accuracy is NOT a good metric.*



Translating AUC to Confusion Tables

- Where to set the threshold depends on costs of false positives and benefits of true positives.
- Tradeoff: different threshold settings yield different accuracy and error rates, for the model with same AUC

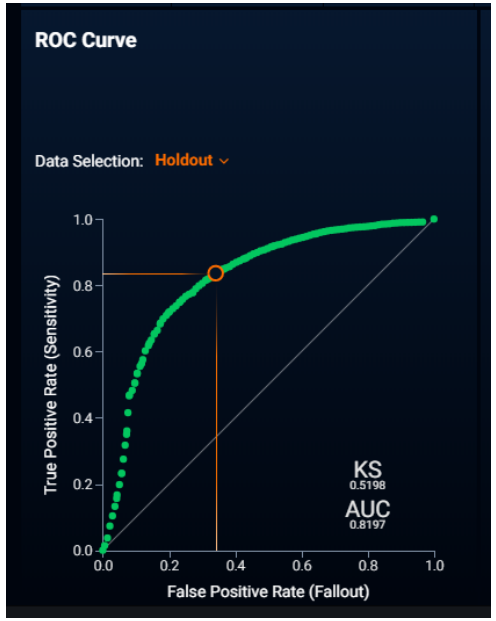


Low Thresh.	Actual False	Actual True
Predicted False	29,349 (TN)	1,962 (FN)
Predicted True	4,318 (FP)	2,969 (TP)

Lower False Positives (13%) and True Positives (60%), Accuracy is 84%

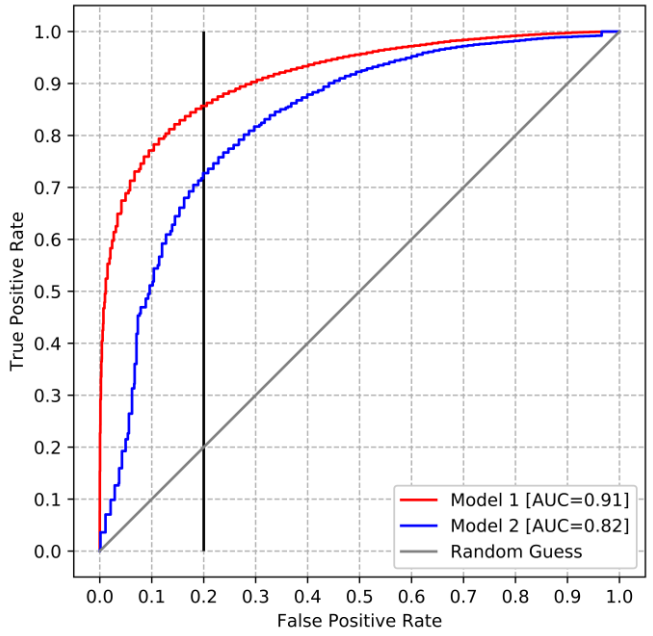
Higher False Positives (34%) and True Positives (84%), Accuracy is 68%

High Thresh.	Actual False	Actual True
Predicted False	22,193 (TN)	807 (FN)
Predicted True	11,474 (FP)	4,124 (TP)



Comparing AUC for Different Models

- Larger AUC values will capture more *true positives* for a given *false positive rate* if the line *is above* the alternative in a ROC chart.
- In most cases, a higher AUC yields higher true positive rates and accuracy given the same false positive rates



Better Model 1 (Red), False Positives (20%) and True Positives (86%), Accuracy 80%

Low Thresh.	Actual False	Actual True
Predicted False	106,943 (TN)	2,823 (FN)
Predicted True	27,727 (FP)	16,900 (TP)

Worse Model 2 (Blue), False Positives (20%) and True Positives (72%), Accuracy 79%

High Thresh.	Actual False	Actual True
Predicted False	107,898 (TN)	5,572 (FN)
Predicted True	26,772 (FP)	14,151 (TP)

Takeaways

- *Accuracy*, *AUC*, and *true positives rates* are all metrics to measure how precise machine learning models are.
- Accuracy is the total number of correct predictions from the model (both True Positive and True Negatives)
- AUC is a measure of the trade-off in True Positives and False Positives. Generally speaking high AUC leads to high accuracy
- Selecting thresholds to separate positives from negatives should be based on business needs -> it is a trade off