Evaluation metrics for classification models

Why do we need evaluation metrics for classification models?

- In machine learning on broad aspect, we have a problem statement to address, then we fetch data related to it, perform analysis and feature engineering. Then use the data to train the model which we finally use to do the prediction.
- Thus, the output of trained machine learning model is consumed by end users for making their decisions.
- In our cybersecurity use case, the impact of the models becomes extremely critical because of the nature of outcome helps to make important decisions about benign and malicious events or type of malicious events.
- In order to use machine learning models in real world scenario, we need to address the fundamental questions such as: -
    1. Why should the end user trust the trained model?
    2. How does our model perform relative to the other models trained by others?
- To address the above fundamental questions, we need to define the governance and framework of evaluation of models which help us understand the given model's performance and also compare them on reliable and useful metrics with other models, which finally allows the end users to make decisions on determining the quality of output produced by the given model and describe the same in detail.
- In terms of building structure for evaluation of classification models, we need to perform seven major tasks: -

    1. List the metrics that can be used for the use case.

    2. Define each metric in detail and explain its benefits and limitations (if any).

    3. Document the evaluation results of all previous models observed from literature survey.

    4. Compute the performance of our model based on each metric defined in task 2.

    5. Quantitatively document the comparison of performance of our model with previously trained models observed in literature survey.

    6. Describe the performance of our model with respect to previously trained model using the data documented in task 6. We need to compute the gap between performance of our model with respect to other models for all available metrics.

    7. Derive the inferences based on task 5 and task 6, explain reason for the same. If our model performs better than previously trained models, we need to explain the reasons for achieving better results. Similarly, if our model performs worse than previously trained models, we need to identify the gaps that we need to work on to reach that performance.

- Robust documentation of the above tasks will enable us define the performance of our models which will provide clarity about its application and also convey the same to end users.
- Additionally, in real world scenarios, the evaluate and decisions to adopt machine learning solutions are taken by different stakeholders. Thus, the specific details in evaluation metrics along with relevant context and research will build the ability of our project to articulate well for different audiences.

Task 1: List of metrics for evaluation of classification models (both binary and multi-class)

1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall
5. F1-Score
6. ROC curve
7. AUC score
8. Specificity
9. Balanced accuracy
10. Matthews Correlation Coefficient (MCC)
11. Logarithmic Loss
12. Binary Crossentropy
13. Categorical Crossentropy
14. Concordance and Discordance
15. Somers-D Statistic
16. Gini coefficient
17. Type 1 error
18. Type 2 error
19. Negative predictive value
20. False discovery rate
21. Cohen kappa metric
22. Precision – Recall curve
23. Brier score

Task 2: Definition and details about each metric: -

1. Confusion matrix: -
   - It is used to consolidate data to measure and evaluate performance of classification model.
   - In binary classification, we have 2X2 matrix.
   - In multi-class classification, we have matrix of size same as number of classes in the target feature.

   - Representation of Confusion Matrix for Binary classifier: -

		Actual values	
		Positive	**Negative**
Predicted values	**Positive**	True Positive	False Positive
	Negative	False Negative	True Negative

   - Representation of Confusion Matrix for Multi-class classifier: -
     Assuming 3 classes: Class 1, Class 2, Class 3.

		Actual values		
		Class 1	**Class 2**	**Class 3**
Predicted values	**Class 1**	Cell 1	Cell 2	Cell 3
	Class 2	Cell 4	Cell 5	Cell 6
	Class 3	Cell 7	Cell 8	Cell 9

     - True Positive: {Cell 1}, {Cell 5}, {Cell 9}
     - False Positive: {Cell 2 + Cell 3}, {Cell 4 + Cell 6}, {Cell 7 + Cell 8}
     - True Negative: {Cell 5 + Cell 6 + Cell 8 + Cell 9}, {Cell 1 + Cell 3 + Cell 7 + Cell 9}, {Cell 1 + Cell 2 + Cell 4 + Cell 5}
     - False Negative: {Cell 4 + Cell 7}, {Cell 2 + Cell 8}, {Cell 3 + Cell 6}

   - True Positive: The number of records model correctly predicts the positive outcome.
   - True Negative: The number of records model correctly predicts the negative outcome.
   - False Positive: The number of records model incorrectly predicts the positive outcome.
   - False Negative: The number of records model incorrectly predicts the negative outcome.
   - Examples of above four components in the domain of cyber security: -
     - True Positive: The model correctly predicts a malicious event as an attack.
     - True Negative: The model correctly predicts a normal event as normal.
     - False Positive: The model incorrectly predicts a normal event as an attack.
     - False Negative: The model incorrectly predicts an attack event as normal.

- Confusion matrix is useful in Binary classification due to compact nature of the structure and complex in multi-class classification due to a greater number of classes to be incorporated in the matrix its dimensions will have higher order and interpreting the results will become difficult.
- Confusion matrix can also give incorrect or misleading representation of a model's performance in the datasets having an imbalanced nature of target classes. This is because if the target class is heavily skewed in one direction, and thus the model may showcase high accuracy by predicting everything in the favour of dominant class while failing to detect the rare class.
- For example: In a network dataset, we have 1000 events, 995 are benign and 5 are malicious. Thus, the dataset is highly imbalanced and skewed against malicious events. Now if the classification model predicts everything as benign then the confusion matrix may portray the model has high accuracy but in reality, it failed to correctly detect malicious events which was more critical for evaluating performance of the classification model.

2. Accuracy: -
    - It gives overall correctness of the model.
    - Accuracy = (True Positive + True Negative)/(True Positive + True Negative + False Positive + False Negative)
    - It helps us understand how often the model predicts correctly.
    - It is useful in scenarios when the dataset is balanced that is the target feature has balanced representation of all classes.
    - It fails to justify false negative in imbalanced dataset.

3. Precision: -
    - It gives accuracy of positive predictions.
    - Precision = True Positive/(True Positive + False Positive)
    - It emphasizes on positive predictions made by the model.
    - For example: A model makes predictions about 10 events: -
        - 5 predictions were correct and 5 were incorrect. Then precision=5/10 = 0.5 – (Scenario 3.1)
        - 8 predictions were correct and 2 were incorrect. Then precision=8/10 = 0.8 – (Scenario 3.2)
        - 2 predictions were correct and 8 were incorrect. Then precision=2/10 = 0.2 – (Scenario 3.3)
    - Given the above 3 scenarios: -
        - Ranking in terms of performance: - Scenario 3.2 > Scenario 3.1 > Scenario 3.3
        - Scenario 3.1 is as good as tossing a fair coin with each side having an equal probability. Thus, it does not give any useful result.
        - Scenario 3.3 indicates the model fails to identify the characteristics of the class it was trying to predict.
    - It fails to address False Negatives while measuring the performance of the model.

- It is better than accuracy while working on imbalanced datasets since it demands minimization of False Positives to have higher score.

4. Recall: -
    - It is also called Sensitivity.
    - It gives model's ability to find all positive cases.
    - Recall = (True Positive)/(True Positive + False Negative)
    - In our cybersecurity use-case, if we have 10 malicious events, how many events were successfully classified as malicious by the model will give the value of recall.
    - Thus, if a model has high recall, it means it mostly identifies malicious events correctly.
    - Taking the above example, out of 10 malicious events: -
        - 8 are classified as malicious and 2 are classified as benign. Then recall=8/10 = 0.8 (Scenario 4.1)
        - 5 are classified as malicious and 5 are classified as benign. Then recall=5/10 = 0.5 (Scenario 4.2)
        - 2 are classified as malicious and 8 are classified as benign. Then recall=2/10 = 0.2 (Scenario 4.3)
    - Given the above 3 scenarios: -
        - Ranking in terms of performance: Scenario 4.1 > Scenario 4.2 > Scenario 4.3
        - Scenario 4.2 is similar to tossing fair coin with each side having equal probability of occurrence.
        - Scenario 4.3 indicates the model fails to identify the malicious events which is more critical than failing to identify the benign events.
    - It works well on imbalanced datasets.

5. F1-Score: -
    -
6. ROC curve
7. AUC score
8. Specificity
9. Balanced accuracy
10. Matthews Correlation Coefficient (MCC)
11. Logarithmic loss
12. Binary Crossentropy
13. Categorical Crossentropy
14. Concordance and Discordance
15. Somers-D Statistic
16. Gini coefficient
17. Type 1 error
18. Type 2 error
19. Negative predictive value
20. False discovery rate
21. Cohen kappa metric

22. Precision – Recall curve
23. Brier score

Task 3: Evaluation results of models in literature survey: -

Sources for evaluation metrics: -

https://www.linkedin.com/pulse/mastering-model-evaluation-comprehensive-guide-accuracy-bin-liao

https://www.linkedin.com/advice/3/what-advantages-disadvantages-using-accuracy#:~:text=Accuracy%20is%20easy%20to%20calculate%20and%20understand%2C%20and,make%20it%20misleading%20or%20inappropriate%20for%20some%20situations.

https://www.coursera.org/articles/what-is-a-confusion-matrix

https://www.kdnuggets.com/2020/04/performance-evaluation-metrics-classification.html

https://www.explorium.ai/blog/machine-learning/top-10-evaluation-metrics-for-classification-models/

https://www.appsilon.com/post/machine-learning-evaluation-metrics-classification

https://community.alteryx.com/t5/Data-Science/Metric-Matters-Part-1-Evaluating-Classification-Models/ba-p/719190

https://www.deepchecks.com/a-guide-to-evaluation-metrics-for-classification-models/

https://www.evidentlyai.com/classification-metrics/multi-class-metrics

https://cloud.google.com/vertex-ai/docs/video-data/classification/evaluate-model

https://www.mlwhiz.com/p/eval_metrics

https://www.datacamp.com/blog/classification-machine-learning

https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/

https://www.machinelearningplus.com/machine-learning/evaluation-metrics-classification-models-r/

https://neptune.ai/blog/evaluation-metrics-binary-classification

https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html

https://cs229.stanford.edu/lectures-spring2022/cs229-evaluation_metrics_slides.pdf

https://www.kdnuggets.com/understanding-classification-metrics-your-guide-to-assessing-model-accuracy

https://www.appliedaicourse.com/blog/evaluation-metrics-in-machine-learning/

https://colab.research.google.com/github/datacommonsorg/api-python/blob/master/notebooks/intro_data_science/Classification_and_Model_Evaluation.ipynb

https://keylabs.ai/blog/overview-of-evaluation-metrics-for-classification-models/

https://thesai.org/Publications/ViewPaper?Volume=12&Issue=6&Code=ijacsa&SerialNo=70

https://deepai.org/machine-learning-glossary-and-terms/evaluation-metrics

https://www.youtube.com/watch?v=5vqk6HnITko&list=PLiteiKUvOPTelAIiquq5-VKFguUJpv1BT

https://www.youtube.com/watch?v=LbX4X71-TFI

https://www.youtube.com/watch?v=lt1YxJ_8Jzs

https://www.youtube.com/watch?v=7ZfcoEao_5c

https://www.youtube.com/watch?v=PeYQIyOyKB8

https://www.youtube.com/watch?v=BEeXPgfhCsA

https://www.youtube.com/watch?v=wpQiEHYkBys

https://www.youtube.com/watch?v=sgv1Q46Tdmw