PERFORMANCE METRICS USE CASES

For each project below, describe which measure(s) are best, and why. Also, give an example of a measure which would be horrible to use, and why. List any assumptions you are making, about the dataset, problem, or business priorities that were involved in the project.

a) The fraud department at a bank wanted to predict which transactions were fraudulent. The training dataset had 100K credit card transactions, of which 97K are legit and 3K are fraud.
b) A hospital wanted to predict whether a MRI scan contained cancer.
c) An IT team wanted to filter spam from email inboxes.
d) A sports analytics department wants to predict which team will win the match.
e) A city government wanted to build a system to monitor Twitter to see if any local residents were tweeting about emergencies that needed quick response from the police department. They don't trust Twitter that much; they only want to send police in true emergencies.
f) *[Describe one more project, whereby the best measure is one that you have not yet listed in parts a-e above.]*

Solution

a) It is given that the dataset used in fraud detection project is highly imbalanced. Therefore, it is important that we use a classification measure like Matthews Correlation Coefficient (MCC) to accommodate this imbalance. In addition, MCC being a balanced measure considers all four values in a confusion matrix while measuring the correlation between observed and predicted classifications.

For such a project, accuracy would not be an ideal performance measure because it might give us high accuracy values (TP and TN), especially because our dataset is highly imbalanced, but also a high number of misclassified cases (FN) which we might not be able to catch. When detecting fraudulent transactions, we want to be careful while labelling a transaction fraud or not, which means we need to use a metric that considers all the four values in a balanced way, like MCC and unlike accuracy.

b) For this project it is important to keep in mind that the dataset might be highly imbalanced because not many patients are diagnosed with cancer and detecting if an MRI scan has cancer or not can be difficult with just any metric. Given that we do not want to misdiagnose a patient we want a metric with high precision, that is if they have cancer, we want to detect it correctly, and if they do not have cancer we want to be as sure as possible with high recall. Hence, I believe F-1 score, the weighted average between precision and recall, will be the ideal performance metric for this project.

A horrible measure for this project would be accuracy because we do not have symmetric split between actual negatives and actual positives. Which means our model might give high accuracy in detecting TP and TN but might also have a lot of misdiagnosed patients under FN, which is not desired when detecting cancer.

c) When filtering out spam emails it is important to remember the cost of misclassification. When a non-spam is classified as spam there is high cost such as losing important company information. Similarly, let's assume the user relies completely on the IT spam email detection model and does not double check before opening an email. There is a risk of user opening a spam email that is misclassified as non-spam. This is also a high cost for the company because they must figure out if there is information leakage, virus, etc. Therefore, I believe ROC curve is the best performance metric for this project as it takes true positive rate and false negative rate into consideration and tells us when a label is negative how often does it predict incorrectly. In other words, how many times are we going to misclassify and lets us determine how well our model is performing.

A metric I think will be horrible to use for this project is precision because it costs company more to miss a non-spam email than falsely labelling a spam one as not, assuming the user can delete the email using his/her judgment.

d) Since target label is binary, to identify which team will win the match the best performance metric would be accuracy. Accuracy will tell us the number of times our model predicted 'wins' or 'losses' correctly and allows us to understand a team's 'odds' of winning.

A metric that would not be ideal for a project like this is log loss. This is because it measures the accuracy of the classifier and output probability for each class rather than telling us if a team will win or lose.

e) Since the city government does not trust Twitter much there is a high chance that they will overlook many tweets, and this may cost them to miss out on actual emergencies. Therefore, for this project F-1 score would be ideal metric because it considers false positives and false negatives as crucial values.

Assuming that there are not many people that tweet about emergencies and just call 911, this dataset is probably very imbalanced. Which means using accuracy is not ideal for this project. Model might be very good at predicting true values but does not give us information on false positives; and we do not want to miss out any actual emergencies assuming our model is good because of high accuracy.

f) A grocery store wants to build a model that predicts if a customer will churn or not. For this project I think recall will be the best performance measure because it is important to know which customers are leaving and try to retain them.

Accuracy will be worst metric, especially when not many customers churn, and the dataset is imbalance. Model might predict the TP and TN correctly and give high accuracy, but it will not give high importance to FN and might mislead the interpretation.