ROC-Kurven, Klassifikationskurven und AUC

Für Evaluation von binären Klassifikationsmodellen

Wofür steht ROC und AUC?

Receiver Operating Characteristic (ROC) Kurven.

Area Under the Curve (AUC).

Was versteht man unter der Evaluation (von Modellen)? Anders gefragt: Welche Aufgabe erfüllt ROC und AUC?

Die Definition der "Evaluation" im Zusammenhang mit ROC (Receiver Operating Characteristic) bezieht sich auf die Bewertung der Leistung eines Klassifikationsmodells, insbesondere in Bezug auf binäre oder multiklassen Klassifikationsprobleme. Bei der ROC-Evaluation werden die folgenden Aspekte bewertet:

Bewertung der Modellleistung: Die ROC-Analyse bewertet die Fähigkeit des Modells, zwischen verschiedenen Klassen oder Zuständen zu unterscheiden. Sie ermöglicht die Beurteilung, wie gut das Modell in der Lage ist, echte positive Ergebnisse von falsch positiven Ergebnissen zu unterscheiden.

Messung von Zielen und Ergebnissen: ROC-Kurven und AUC (Area Under the Curve) sind Maßnahmen, um zu überprüfen, ob ein Modell die gewünschten Ziele erreicht. Sie helfen bei der Bewertung, ob das Modell in der Lage ist, die Klassen richtig zu klassifizieren.

Bewertung von Modellleistung bei verschiedenen Schwellenwerten: Die ROC-Analyse erlaubt es, die Modellleistung bei verschiedenen Schwellenwerten für die Klassifikation zu bewerten. Dies ist besonders wichtig, um den Schwellenwert anzupassen und die gewünschten Kompromisse zwischen Sensitivität und Spezifität zu erzielen.

Bewertung von Modellen im Multiklassen-Kontext: Wie bereits erwähnt, können ROC-Kurven und AUC auch auf Multiklassen-Klassifikationsprobleme erweitert werden, wobei sie häufig die "Eins gegen Alle"-Technik verwenden, um die Leistung jeder Klasse zu bewerten.

In diesem Sinne bezieht sich die "Evaluation" im Kontext der ROC-Analyse auf die Beurteilung der Leistung eines Klassifikationsmodells und die Bestimmung, wie gut es die Aufgabe der Klassifikation in Bezug auf Sensitivität und Spezifität bewältigt. Dies ist ein entscheidender Schritt, um festzustellen, wie effektiv ein Modell in der Praxis ist und ob es die gesteckten Ziele erreicht.

Frage: Wofür werden ROC-Kurven und AUC benutzt/gebraucht?

ROC-Kurven und AUC sind in erster Linie für die **Evaluierung** von **binären Klassifikationsmodellen** relevant. Das bedeutet, dass sie hauptsächlich verwendet werden, wenn Sie ein Modell bewerten

möchten, das zwischen zwei Klassen oder Zuständen unterscheidet, wie beispielsweise Ja/Nein, Krank/gesund, Spam/Nicht-Spam usw.

> Aufgabe: **Evaluation**.

> Auf welche Modelle anwendbar: **binären Klassifikationsmodellen**

Welche binären Klassifikationsmodelle gibt es, die mit ROC und AUC bewertet werden können?

Diese Metriken sind für eine breite Palette von binären Klassifikationsmodellen nützlich, einschließlich, aber nicht beschränkt auf:

Logistische Regression: Ein einfaches statistisches Modell für binäre Klassifikation.

Support Vector Machines (SVM): Ein Modell, das eine marginale Trennfläche zwischen den Klassen zu finden versucht.

Random Forest + Entscheidungsbäume: Ein Ensemble-Modell, das aus einer Kombination von Entscheidungsbäumen besteht.

K-nearest Neighbors (K-NN): Ein Modell, das auf der Klassifikation basierend auf den Klassen der K-Nächsten-Nachbarn eines Datenpunkts beruht. Nicht parametrisches Modell.

Naive Bayes: Ein probabilistisches Modell, das auf dem Bayes'schen Theorem beruht.

Neuronale Netzwerke: Tief lernende Modelle können ebenfalls für binäre Klassifikation verwendet werden.

Wie bereits erwähnt, ist die AUC-ROC-Kurve nur für binäre Klassifikationsprobleme geeignet. Dennoch können wir sie auf Multiklassen-Klassifikationsprobleme mit der sogenannten "Eins gegen Alle"-Technik erweitern. Das bedeutet, wenn wir zum Beispiel drei Klassen haben, nämlich 0, 1 und 2, wird die ROC-Kurve für Klasse 0 erstellt, indem wir Klasse 0 gegen alle anderen Klassen (1 und 2) klassifizieren.

> Klasse 1 ist 0 und alle anderen Klassen werden zusammengefasst als andere Klasse.

Ab hier geht es um die Umsetzung von ROC & AUC

Modell Validierung: Performance Metriken von (binären) Klassifikationsmodellen.

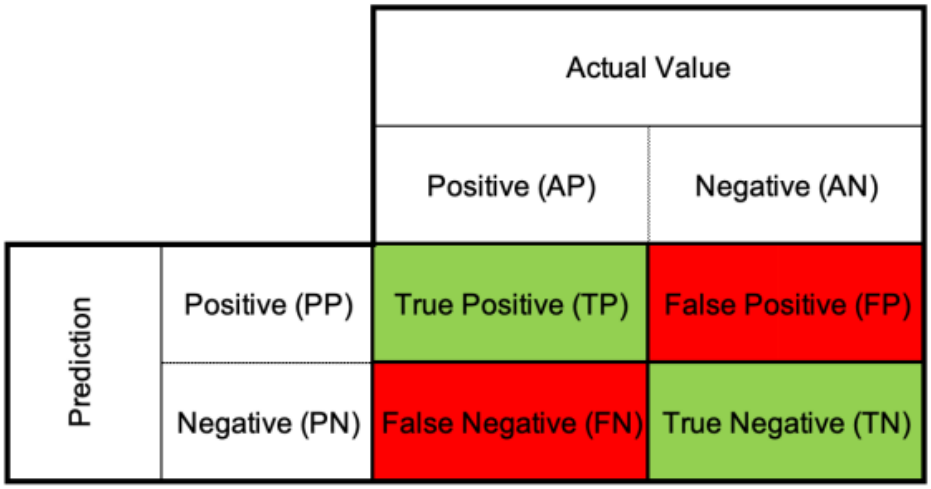


Abbildung 1: Konfusionsmatrix. Auf der linken Seite (Prediction. Y-Achse, Senkrecht) steht, was die Vorhersage für die einzelnen Fälle war. Auf der X-Achse (Actual Value, Waagerechte) steht, was der eigentliche Wert gewesen ist.

Bei binären Klassifikationsmodellen sind Kontingenztafeln (auch Kreuztabellen genannt) wichtig.

Definition: Kontingenztafeln (auch: Kontingenztabellen oder Kreuztabellen) sind Tabellen, die die absoluten oder relativen Häufigkeiten (Häufigkeitstabellen) von Kombinationen bestimmter Merkmalsausprägungen enthalten. Eine tabellarische Darstellung (2x2 im binären Fall) der Häufigkeiten der Beobachtungen nach ihrem vorhergesagten und tatsächlichen Klassifikationsstatus.

> Im Englischen **Confusion Matrix / Konfusionsmatrix** genannt.

> Siehe Abbildung 1.

Auf der linken Seite (Prediction. Y-Achse, Senkrecht) steht, was die Vorhersage für die einzelnen Fälle war. Auf der X-Achse (Actual Value, Waagerechte) steht, was der eigentliche Wert gewesen ist. Die Diagonale gibt die Anzahl korrekter Vorhersagen an. Um Actual Values auftragen zu können und damit man Kontingenztafeln erstellen kann, muss die Klassifikationsmodelle auf eine Reihe von Fällen anwenden, bei denen man bereits im Nachhinein weiß, welcher Klasse die jeweiligen Objekte angehören.

> Man braucht Label/ Response-Variablen = Kontingenztafeln werden nur bei supervidierten Lernmodellen genutzt.

TP (**True Positive**): Das Modell hat korrekt vorhergesagt, dass etwas positiv ist, und es ist tatsächlich positiv.

TN (**True Negative**): Das Modell hat korrekt vorhergesagt, dass etwas negativ ist, und es ist tatsächlich negativ.

FP (**False Positive**): Das Modell hat fälschlicherweise vorhergesagt, dass etwas positiv ist, obwohl es in Wirklichkeit negativ ist. Dies wird auch als "Type I Error" bezeichnet. Positiv vorhergesagt, war aber in wirklichkeit negativ.

FN (**False Negative**): Das Modell hat fälschlicherweise vorhergesagt, dass etwas negativ ist, obwohl es tatsächlich positiv ist. Dies wird auch als "Type II Error" bezeichnet. Negativ vorhergesagt, war aber in Wirklichkeit positiv.

Diese Begriffe sind grundlegend für die Berechnung von Leistungsmetriken in der Klassifikation, wie z.B. Genauigkeit, Präzision, Recall und die ROC-Kurve.

Die Gütemaße zur Beurteilung eines Klassifikators beruhen im Wesentlichen auf der Konfusionsmatrix

> Die meisten **Performance-Metrikern für Klassifizierungsprobleme nutzen die Kontingenztafel als ein Fundament** -> Sind davon abgeleitet.

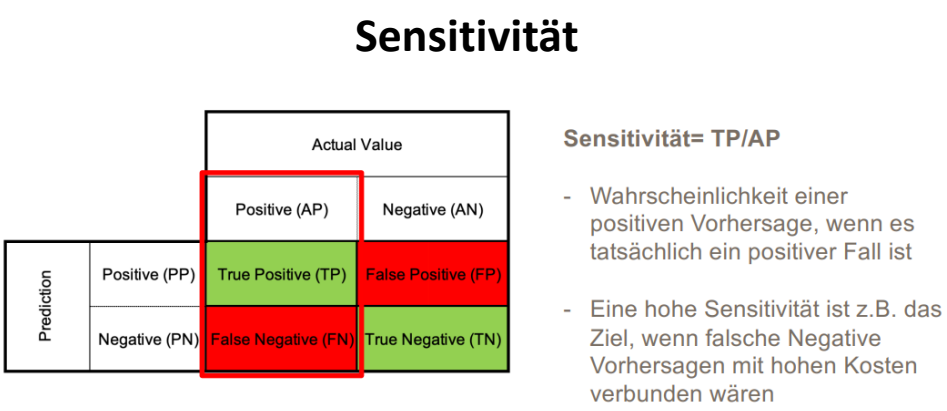


Abbildung 2: Visualisierung der Sensitivität mit Definition und Formel.

Definiere Sensitivität, was bildet es ab, wie wird es ausgerechnet und welchen Zahlenwert kann es annehmen? Wie wird Sensitivität sonst genannt?

Einer der am häufigsten verwendeten Metriken für die Bewertung von (binäre) Klassifizierungsmodellen ist die **Sensitivität**. Die Sensitivität gibt wieder, wie hoch die Wahrscheinlichkeit einer richtigen positiven Vorhersage (TP) ist, wenn der Fall tatsächlich positiv (AP) ist.

Formel: $Sensitivity = \frac{True\ Positive}{Actual\ Positive} = \frac{True\ Positive}{True\ Positive + False\ Negative}$

Die Sensitivität lässt sich berechnen, indem man alle True Positive Fälle durch alle Positive Fälle teilt. Die Sensitivität kann einen Wert zwischen **0 und 1** annehmen und gibt damit eine Wahrscheinlichkeit (%) an. Interpretationsbeispiel: Eine Sensitivity von 0,7 bedeutet, dass 70 % aller wirklich positiven Ergebnisse auch als positiv vorhergesagt wurden. Mit anderen Worten: 70 % aller positiven Ergebnisse wurde richtig vorhergesagt und 30 % wurden als negativ erkannt, obwohl Sie in Wirklichkeit positiv sind.

Wie wird Sensitivität sonst genannt?

Die Sensitivität wird sonst **True-Positive Rate (TPR)/ Richtig-Positiv-Rate (RPR) und Recall** genannt.

Frage: Wann ist Sensitivität wichtig?

Man erkennt also, dass eine hohe Sensitivität ein Ziel ist, **wenn falsch negative Aussagen (FN) mit hohen Kosten verbunden sind**. Wenn Sie zum Beispiel eine schlimme Krankheit (z.B. Krebs) identifizieren wollen, dann ist es sehr wichtig, dass Sie alle Patienten identifizieren, die die untersuchte Krankheit wirklich haben. Wenn Sie viele Falsch-negative Aussagen haben, dann werden die Patienten fälschlicherweise als gesund eingestuft und kriegen dann möglicherweise keine weitere Behandlung, was im schlimmsten Fall ihr Leben kosten würde. Deshalb ist die Maximierung der Sensitivität sehr wichtig.

Wird Sensitivität nur in Machine Learning gebraucht?

Nein. Die klassische Diagnostikverfahren nutzen Sensitivität.

Spezifität

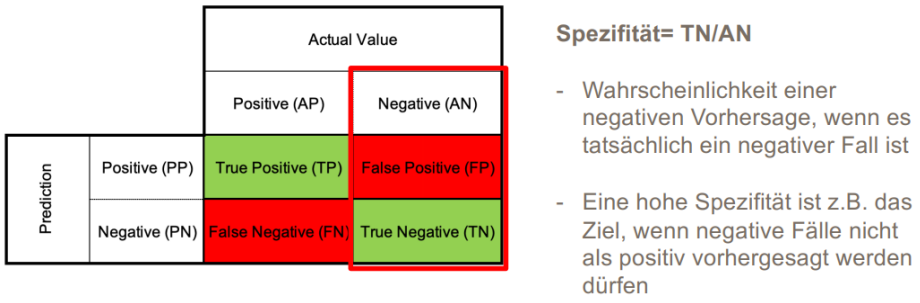


Abbildung 3: Visualisierung der Spezifität mit Definition und Formel.

Definiere Spezifität, was bildet es ab, wie wird es ausgerechnet und welchen Zahlenwert kann es annehmen?

Das Äquivalent für die Negativen Ergebnisse zur Sensitivität ist die Spezifität. Die Sensitivität gibt wieder, wie hoch die Wahrscheinlichkeit einer richtigen negativen Vorhersage (TN) ist, wenn der Fall tatsächlich positiv (AN) ist.

Formel: $Spezifität = \frac{True\ Negatives}{Actual\ Negatives} = \frac{True\ Negative}{True\ Negative + False\ Positives}$

Die Sensitivität lässt sich berechnen, indem man alle True Negative Fälle durch alle Negative Fälle teilt. Die Sensitivität kann einen Wert zwischen **0 und 1** annehmen und gibt damit eine Wahrscheinlichkeit (%) an. Interpretationsbeispiel: Eine Spezifität von 0,7 bedeutet, dass 70 % aller wirklich negativen Ergebnisse auch als negativ vorhergesagt wurden. Mit anderen Worten: 70 % aller negativen Ergebnisse wurde richtig vorhergesagt und 30 % wurden als positiv erkannt, obwohl Sie in Wirklichkeit negativ sind.

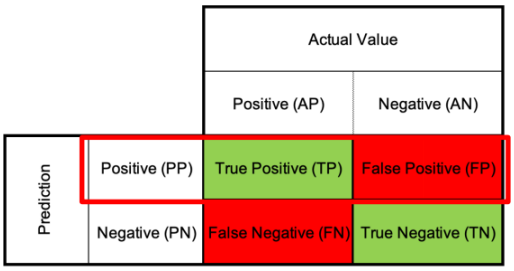
Wie wird Spezifität sonst genannt?

Die Sensitivität wird sonst **True-Negative Rate (TNR)/ Richtig-Negative-Rate (RNR)** genannt.

Frage: Wann ist Spezifität wichtig?

Man erkennt also, dass eine hohe Spezifität ein Ziel ist, **wenn negative Fälle (AN) nicht als positiv (FP) vorhergesagt werden dürfen**. Ein konkretes Beispiel, in dem Spezifität von großer Bedeutung ist, betrifft Drogentests im Arbeitsumfeld. Arbeitgeber führen Drogentests durch, um sicherzustellen, dass ihre Mitarbeiter keine illegalen Drogen konsumieren. In diesem Kontext ist es wichtig, dass negative Testergebnisse (wenn der Mitarbeiter keine Drogen genommen hat) nicht fälschlicherweise als positiv eingestuft werden. Eine hohe Spezifität gewährleistet, dass nur diejenigen Mitarbeiter als positiv getestet werden, die tatsächlich Drogen genommen haben (True Positives), und dass Mitarbeiter, die keine Drogen konsumiert haben, korrekt als negativ getestet werden (True Negatives). Eine niedrige Spezifität könnte dazu führen, dass unschuldige Mitarbeiter fälschlicherweise als drogenpositiv eingestuft werden (False Positives), was schwerwiegende Konsequenzen für ihre Karriere haben könnte. Daher ist in diesem Fall die Maximierung der Spezifität von entscheidender Bedeutung, um genaue und gerechte Testergebnisse sicherzustellen.

Präzision



Präzision= TP/PP

- Wahrscheinlichkeit eines positiven Falles, wenn die Vorhersage positiv war
- Eine hohe Präzision ist z.B. das Ziel, wenn Fehlalarme vermieden werden sollen

Abbildung 4: Visualisierung der Präzision mit Definition und Formel.

Definiere Präzision, was bildet es ab, wie wird es ausgerechnet und welchen Zahlenwert kann es annehmen?

Die Präzision lässt sich berechnen, indem man alle True Positive Fälle durch alle Predicted Positives (PP) teilt. Die Präzision kann einen Wert zwischen **0 und 1** annehmen und gibt damit eine Wahrscheinlichkeit (%) an. Ein Wert, der nicht so oft berichtet wird, ist die Präzision. Präzision ist vor allem klinisch sehr relevant, auch bei der Bewertung von Vorhersagemodellen. Es ist die Wahrscheinlichkeit eines positiven Falles (TP), wenn die Vorhersage positiv war (PP). In diesem Fall schauen wir auf folgendes: Wenn wir eine positive Vorhersage haben, wie wahrscheinlich ist es, dass die Vorhersage auch wirklich positiv (TP) ist.

Formel: $Präzision = \frac{True\ Positives\ (TP)}{Positive\ \Pr ediction\ (PP)} = \frac{True\ Positives\ (TP)}{True\ Positives + False\ Positives\ (FP)}$

Die Präzision beantwortet die Frage: Wenn das Modell oder der Test eine positive Vorhersage macht, wie wahrscheinlich ist es, dass diese Vorhersage tatsächlich korrekt ist? Eine hohe Präzision bedeutet, dass die positiven Vorhersagen des Modells zuverlässig und genau sind, während eine niedrige Präzision darauf hinweist, dass viele der positiven Vorhersagen falsch sind. In klinischen Anwendungen, wie bei medizinischen Tests oder Diagnosen, ist die Präzision von großer Bedeutung. Ein hoher Präzisionswert zeigt an, dass das diagnostische Verfahren in der Lage ist, tatsächlich kranke Patienten korrekt zu identifizieren, ohne zu viele gesunde Patienten fälschlicherweise als krank zu diagnostizieren. Dies reduziert unnötige Sorgen und Behandlungen für gesunde Patienten und verbessert die Qualität der medizinischen Versorgung. Eine Präzisionswert von 0,7 (oder 70%) bedeutet, dass, wenn ein Vorhersagemodell oder ein Test eine positive Vorhersage macht, es mit einer Wahrscheinlichkeit von 70% korrekt ist. Mit anderen Worten, 70% der positiven Vorhersagen sind tatsächlich korrekt, während die verbleibenden 30% falsch sind.

Wie wird Präzision sonst genannt?

Mediziner kennen die Präzision als den positiv-Prädiktiven Wert. Ein anderer Name für Präzision ist Relevanz.

Frage: Wann ist Präzision wichtig?

Eine hohe Präzision ist das Ziel, wenn Fehlalarme vermieden werden sollen. Mit anderen Worten: Präzision ist das Ziel, wenn **Falsch-Positive Fälle** so selten wie möglich gehalten oder ganz vermieden werden sollen.

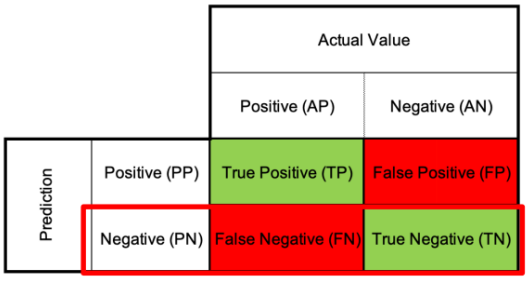
Angenommen, es gibt einen medizinischen Test zur Früherkennung von Brustkrebs, der bei Frauen durchgeführt wird. Dieser Test gibt entweder ein positives oder ein negatives Ergebnis aus.

Hohe Präzision: Wenn der Test eine hohe Präzision aufweist, bedeutet dies, dass er in der Regel sehr zuverlässig ist. Wenn er positiv ist, können Sie mit hoher Sicherheit davon ausgehen, dass Sie tatsächlich Brustkrebs haben. Dies reduziert unnötige Sorgen und Behandlungen. Eine hohe Präzision bedeutet in diesem Fall, dass Fehlalarme vermieden werden. Nur diejenigen Frauen, bei denen der Test positiv ist und tatsächlich an Brustkrebs erkrankt sind, erhalten eine Behandlung.

Niedrige Präzision: Wenn der Test eine niedrige Präzision hat, bedeutet dies, dass er häufig falsche positive Ergebnisse liefert. Frauen, die den Test bestehen, könnten fälschlicherweise denken, dass sie Brustkrebs haben, und sich unnötig Sorgen machen. Dies führt zu Fehlalarmen und möglicherweise zu unnötigen Untersuchungen und Behandlungen.

In diesem Beispiel ist eine hohe Präzision für den Brustkrebstest von entscheidender Bedeutung, da sie dazu beiträgt, Fehlalarme zu vermeiden und sicherzustellen, dass die positiven Testergebnisse tatsächlich auf das Vorliegen von Brustkrebs hinweisen. Dies ist wichtig, um die körperliche und emotionale Belastung der Patientinnen zu minimieren und die Ressourcen im Gesundheitswesen effizient zu nutzen.

Negativ Prädiktor Wert



Negativ Prädiktiver Wert= TN/PN

- Wahrscheinlichkeit eines negativen Falles, wenn die Vorhersage negativ war

Abbildung 5: Visualisierung der Präzision mit Definition und Formel.

Definiere Negativ Prädiktor Wert, was bildet es ab, wie wird es ausgerechnet und welchen Zahlenwert kann es annehmen?

Das Äquivalent für die Negativen Ergebnisse zur Präzision ist der Negativ Prädiktor Wert. Es ist die Wahrscheinlichkeit eines negativen Falles, wenn die Vorhersage negativ war. Der NPW beantwortet die Frage: Wenn ein Test oder Modell ein negatives Ergebnis vorhersagt, wie wahrscheinlich ist es, dass dieses Ergebnis korrekt ist?

Formel:

$$Negativ\ Prädiktor\ Wert\ = \frac{True\ Negative\ (TN)}{\mathrm{Pr}\ edicted\ Negatives\ (PN)} = \frac{True\ Negative\ (TN)}{True\ Negative + False\ Negative\ (FN)}$$

Den Negativ Prädiktor Wert errechnet man, indem man die wahren negativen Werte (TN) durch die vorhergesagten negativen Werte (PN) teilt. Der negativ Prädiktor Wert kann einen Wert zwischen 0 und 1 annehmen und gibt damit eine Wahrscheinlichkeit (%) an. Der NPV gibt die Wahrscheinlichkeit an, dass ein negatives Testergebnis korrekt ist, was insbesondere in diagnostischen und medizinischen Anwendungen von großer Bedeutung ist. Ein hoher NPV bedeutet, dass ein negatives Testergebnis zuverlässig auf das Fehlen des betrachteten Zustands hinweist, während ein niedriger NPV darauf hinweist, dass negative Ergebnisse weniger vertrauenswürdig sind und Fehlalarme wahrscheinlicher sind. Ein Wert von 0,7 für den Negativ Prädiktor Wert (Negative Predictive Value, NPV) bedeutet, dass, wenn ein Test oder ein Modell ein negatives Ergebnis vorhersagt, die Wahrscheinlichkeit, dass dieses Ergebnis korrekt ist, 70% beträgt. Mit anderen Worten, in 70% der Fälle, in denen der Test ein negatives Ergebnis liefert, ist dies tatsächlich korrekt und deutet darauf hin, dass der betrachtete Zustand oder das Ereignis wahrscheinlich nicht vorliegt.

Frage: Wann ist Negativ Prädiktor Wert wichtig?

Der NPW ist wichtig, wenn man **False Negatives (FN)** vermeiden muss. Ein False Negative tritt auf, wenn ein Test oder ein Modell ein negatives Ergebnis vorhersagt, obwohl die Realität tatsächlich positiv ist.

Beispiel: Drogentests bei Flughafensicherheit

Angenommen, bei der Sicherheitskontrolle an einem Flughafen wird ein Drogentest durchgeführt, um das Vorhandensein illegaler Drogen bei Passagieren zu identifizieren. Der Negative Prädiktor Wert (NPV) spielt in diesem Szenario eine entscheidende Rolle aus folgenden Gründen:

Vermeidung von Fehlern: In der Flughafensicherheit ist es von entscheidender Bedeutung, Drogenkuriere oder gefährliche Personen zuverlässig zu identifizieren. Ein False Negative, bei dem ein Drogentest eine Person fälschlicherweise als "negativ" einstuft, obwohl sie tatsächlich Drogen bei sich trägt, kann schwerwiegende Sicherheitsprobleme verursachen.

Hoher NPV: Ein hoher NPV des Drogentests bedeutet, dass ein negatives Testergebnis zuverlässig darauf hinweist, dass die Person wahrscheinlich keine Drogen bei sich hat. Ein hoher NPV trägt dazu bei, False Negatives zu minimieren und die Sicherheit am Flughafen zu gewährleisten.

Prävalenz

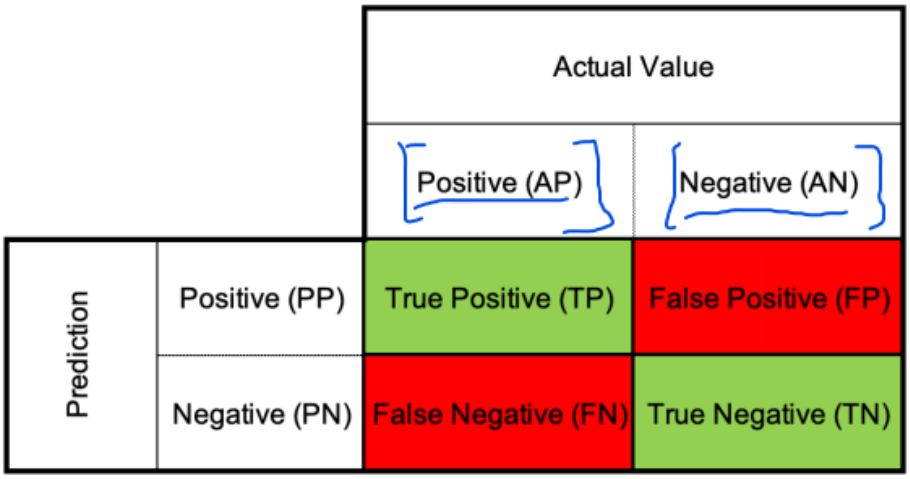


Abbildung 6: Visualisierung der Prävalenz.

Prävalenz, was bildet es ab, wie wird es ausgerechnet und welchen Zahlenwert kann es annehmen?

Die Prävalenz gibt nicht die Qualität des Modells wieder. Stattdessen sagt es einem, wie häufig die positiven Ergebnisse (AP) im Datensatz vorkommen. In vielen Klassifikationsproblemen (wie binärer Klassifikation) ist die Prävalenz relevant, um zu verstehen, wie häufig das positive Ereignis oder die positive Klasse auftritt.

Formel: $Prävalenz = \frac{Actual\ Positives\ (AP)}{Actual\ Positives\ (AP) + Actual\ Negatives\ (AN)}$

Die Prävalent berechnet man, indem man die tatsächlich positiven Fälle durch alle Fälle teilt. Prävalenz kann einen Wert zwischen 0 und 1 annehmen und gibt damit eine Wahrscheinlichkeit (%) an.

Frage: Wann ist Prävalenz wichtig?

Stellen Sie sich eine ländliche Gemeinde mit 500 Einwohnern vor. In dieser Gemeinde gibt es eine Initiative, um den Zugang zum Internet zu fördern und die digitale Bildung zu verbessern. Um die Initiative effektiv zu gestalten, ist es wichtig, die Prävalenz der Internetnutzung in der Gemeinde zu verstehen.

In einer Umfrage stellt die Gemeindeverwaltung fest, dass von den 500 Einwohnern:

300 Personen das Internet regelmäßig nutzen, um online zu arbeiten, zu lernen oder sich zu unterhalten.

200 Personen haben keinen Zugang zum Internet und nutzen es daher nicht.

Die Prävalenz der Internetnutzung in dieser ländlichen Gemeinde kann wie folgt berechnet werden:

Prävalenz = 300/500 = 0,6 = 60 %

Die Prävalenz der Internetnutzung beträgt in diesem Fall 0,6 oder 60%. Das bedeutet, dass in dieser Gemeinde 60% der Einwohner das Internet nutzen. Dies ist eine wichtige Information für die Planung der Initiative zur Förderung der Internetnutzung.

Die Gemeindeverwaltung könnte nun gezielt Ressourcen und Programme entwickeln, um den verbleibenden 40% der Einwohner den Zugang zum Internet zu erleichtern und die digitale Bildung zu fördern.

Genauigkeit (Accuracy)

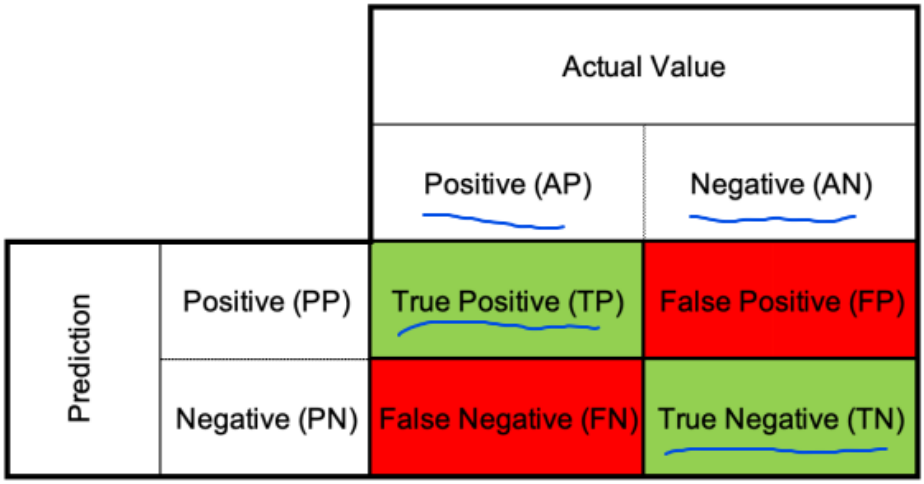


Abbildung 7: Visualisierung der Genauigkeit.

Genauigkeit, was bildet es ab, wie wird es ausgerechnet und welchen Zahlenwert kann es annehmen?

Die Genauigkeit ist die Wahrscheinlichkeit der richtig vorhergesagten Fälle (TP und TN) im Vergleich zu allen Fällen. Die Genauigkeit misst, wie gut ein Modell oder ein Algorithmus in der Lage ist, die richtigen Vorhersagen (True Positives und True Negatives) im Vergleich zu allen Vorhersagen (sowohl positiv als auch negativ) zu treffen. Sie gibt an, welcher Prozentsatz der Vorhersagen korrekt ist.

Formel: $Genauigkeit = \frac{True\ Positives\ (TP) + True\ Negatives\ (TN)}{Actual\ Positives\ (AP) + Actual\ Negatives\ (AN)}$

Die Genauigkeit kann Werte zwischen 0 und 1 annehmen, oder sie wird oft in Prozent ausgedrückt. Ein Genauigkeitswert von 1 oder 100% bedeutet, dass alle Vorhersagen korrekt sind, während ein Wert von 0 oder 0% anzeigt, dass keine Vorhersagen korrekt sind. Ein Wert von 70% in Bezug auf die Genauigkeit bedeutet, dass das Modell oder der Algorithmus in 70% der Fälle die richtigen Vorhersagen getroffen hat. Mit anderen Worten, von den insgesamt durchgeführten Vorhersagen waren 70% der Vorhersagen korrekt.Dies zeigt an, dass das Modell in der Lage ist, die Mehrheit der Fälle richtig zu klassifizieren, was grundsätzlich positiv ist. Ein Genauigkeitswert von 70% bedeutet jedoch auch, dass in 30% der Fälle die Vorhersagen falsch waren.

Achtung: **Genauigkeit ist abhängig von der Verteilung von positiven (AP) und negativen Fällen (AN)**! Wenn die Verteilung ungleich ist, kann die Genauigkeit leicht täuschen. Beispielsweise, wenn in einem Datensatz 95 % der Fälle zur Klasse N gehören und nur 5 % zur Klasse P, könnte ein Klassifikator, der immer N vorhersagt, eine Genauigkeit von 95 % haben. Doch dies wäre in den meisten Anwendungen nicht sinnvoll, da er die Fälle in Klasse P vollständig ignoriert.

Frage: Was ist balanced Accuracy?

Die ausgewogene Genauigkeit (balanced accuracy) ist eine Metrik, die speziell entwickelt wurde, um mit unbalancierten Datensätzen umzugehen. Unbalancierte Datensätze sind Datensätze, in denen eine Klasse (z. B. die positive Klasse) erheblich seltener oder häufiger vorkommt als die andere Klasse (z. B. die negative Klasse). Bei solchen Datensätzen kann die normale Genauigkeitsberechnung zu irreführenden oder verzerrten Ergebnissen führen.

Formel: $Balancierte\ Genauigkeit = \frac{Sensitivität + Spezifität}{2}$

Die ausgewogene Genauigkeit behebt dieses Problem, indem sie die Genauigkeit für jede Klasse getrennt berechnet und dann den Durchschnitt dieser Genauigkeiten nimmt. Dadurch wird sichergestellt, dass die Genauigkeit unabhängig von der Prävalenz der Klassen korrekt bewertet wird.

F1- Score

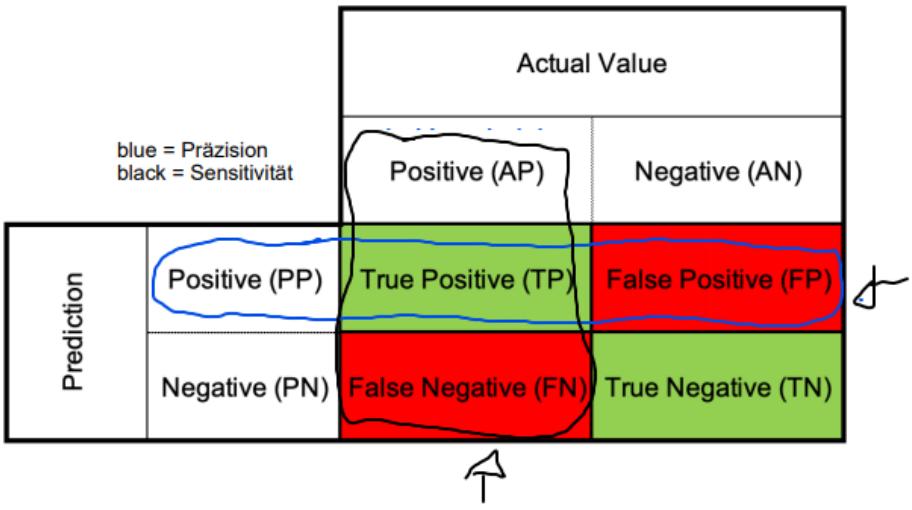


Abbildung 8: Visualisierung des F1-Scores.

F1-Score, was bildet es ab, wie wird es ausgerechnet und welchen Zahlenwert kann es annehmen?

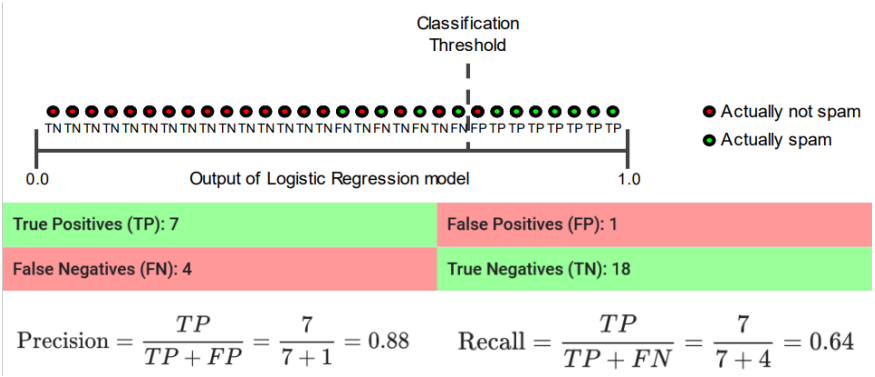
Der F1-Score ist eine Metrik zur Bewertung der Leistung eines Klassifikationsmodells und wird oft verwendet, wenn sowohl Präzision als auch Recall (Sensitivität) von Bedeutung sind. Der F1-Score ist das harmonische Mittel zwischen Präzision und Recall. Der F1-Score berücksichtigt sowohl Präzision als auch Recall und gibt eine einzige Metrik, die die Balance zwischen diesen beiden Leistungskennzahlen darstellt.

Formel: $F1 = \frac{Präzision \cdot Sensitivität}{Präzision + Sensitivität} \cdot 2$

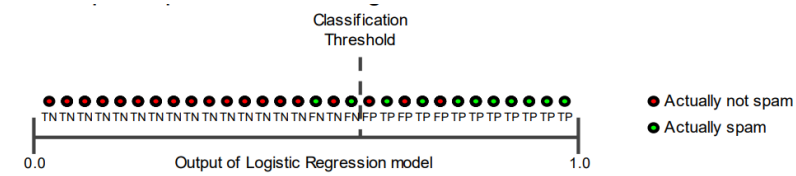
Der F1-Score bewertet, wie gut ein Klassifikationsmodell sowohl genaue Vorhersagen für positive Fälle macht als auch keine positiven Fälle übersieht. Er hilft dabei, die Balance zwischen "korrekten positiven Vorhersagen" (Präzision) und "verpassten positiven Fällen" (Sensitivität) zu beurteilen. Der F1-Score kann Werte zwischen 0 und 1 annehmen. Ein hoher F1-Score (nahe 1) zeigt an, dass das Modell sowohl Präzision als auch Recall in Balance hat, was auf eine gute Gesamtleistung hinweist. Ein niedriger F1-Score (nahe 0) zeigt an, dass das Modell entweder viele falsch positive Vorhersagen macht oder viele positive Fälle übersieht. Ein F1-Score von 0,7 zeigt an, dass Ihr Klassifikationsmodell eine angemessene Balance zwischen Präzision (genaue Vorhersagen der positiven Fälle) und Recall (Fähigkeit, alle positiven Fälle zu erkennen) hat. Dies bedeutet, dass das Modell dazu neigt, positive Fälle genau zu identifizieren, ohne zu viele davon zu übersehen und ohne zu viele falsch positive Vorhersagen zu machen.

Welchen Zweck erfüllt der F1-Score?

Da der F1-Score sowohl Präzision als auch Sensitivität berücksichtigt, ist es bei nicht balancierten Datensätzen ein besseres Maß für die Genauigkeit.



Anderer Threshold:



True Positives (TP): 9	False Positives (FP): 3
False Negatives (FN): 2	True Negatives (TN): 16

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{9}{9+3} = 0.75 \quad \text{Recall} = \frac{TP}{TP+FN} = \frac{9}{9+2} = 0.82$$

Präzision und Sensitivität (eigentlich alle behandelten Metriken) sind vom Cut-Off abhängig, was in den zwei Abbildungen demonstriert wird. Der F1-Score ist ein kombiniertes Maß aus Präzision und Sensitivität.

ROC-Kurven

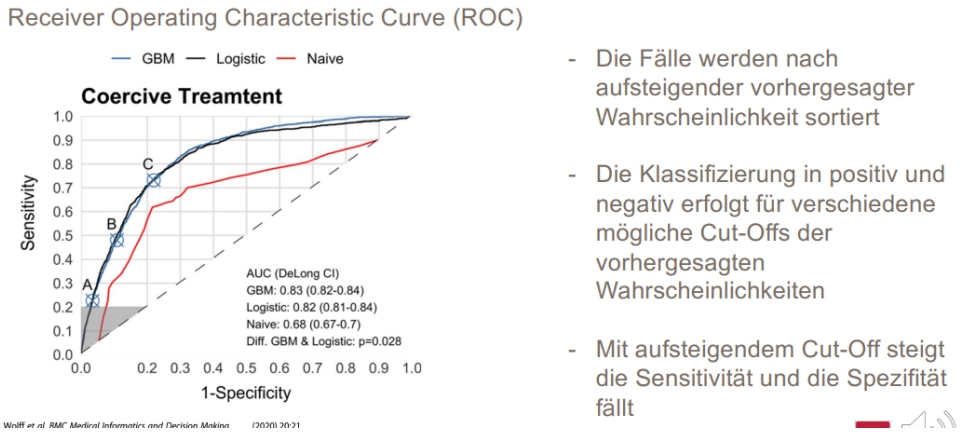


Abbildung 9: Visualisierung der ROC-Kurve.

Die Receiver Operating Characteristic (ROC)-Kurve ist ein wichtiges Werkzeug in der Bewertung und Visualisierung der Leistung von binären Klassifikationsmodellen, insbesondere in Machine Learning und diagnostischen Tests. Die Hauptaufgabe einer ROC-Kurve besteht darin, die Fähigkeit eines Modells zu bewerten, zwischen den beiden Klassen (normalerweise als "positiv" und "negativ" bezeichnet) zu unterscheiden, indem sie die Trade-offs zwischen Sensitivität und Spezifität darstellt. Die ROC-Kurve zeigt, wie sich die Sensitivität und die Spezifität eines Modells bei verschiedenen Schwellenwerten (Thresholds) für die Vorhersage ändern. Sie ermöglicht eine visuelle Bewertung der Fähigkeit des Modells, die richtige Balance zwischen der Identifizierung von positiven Fällen und der Vermeidung von Fehlalarmen zu finden.

Wozu ROC-Kurven?

Die bisherigen Metriken, wie Sensitivität, Spezifität und andere, basieren auf einer Tabelle, die die Übereinstimmung zwischen den vorhergesagten Werten eines Modells und den tatsächlichen Werten zeigt. Sie sagen uns jedoch nicht direkt, welche Klasse vorhergesagt wird. Stattdessen liefern sie die Wahrscheinlichkeit eines positiven Ergebnisses, was die Wahrscheinlichkeit für tatsächliche positive Fälle (AP - Actual Positives) darstellt.

Um jedoch die Klasse vorherzusagen, müssen wir die Wahrscheinlichkeiten für Sensitivität und Spezifität in Abhängigkeit von verschiedenen Schwellenwerten auf den X- und Y-Achsen auftragen. Indem wir dies tun, können wir feststellen, wie sich die Sensitivität und Spezifität ändern, wenn wir den Schwellenwert für die Wahrscheinlichkeit anpassen. Dies bedeutet, dass wir einen Cut-Off-Wert auswählen können, der festlegt, ob die Klasse als positiv oder negativ eingestuft wird.

Wenn die Schwellenwerte niedrig sind, neigen wir dazu, mehr Fälle als positiv zu klassifizieren, was die Sensitivität erhöht. Dies bedeutet, dass wir eine bessere Fähigkeit haben, tatsächliche positive Fälle zu

identifizieren. Aber dies kann auch zu mehr falsch positiven Vorhersagen führen, was die Spezifität verringert.

Auf der anderen Seite, wenn die Schwellenwerte hoch sind, klassifizieren wir nur Fälle mit sehr hohen Wahrscheinlichkeiten als positiv. Dies kann die Spezifität erhöhen, da die Wahrscheinlichkeit geringer ist, dass negative Fälle fälschlicherweise als positiv eingestuft werden. Allerdings kann dies die Sensitivität verringern, da einige tatsächlich positive Fälle möglicherweise übersehen werden.

> Bsp Cut-Off = 0,5. D.h. alle Fälle mit einer vorhergesagten Wahrscheinlichkeit höher 0,5 sind positiv und darunter sind Sie alle negativ.

Jetzt zur Antwort: Die ROC-Kurve, oder Receiver Operating Characteristic-Kurve, wird verwendet, um die Beziehung zwischen Sensitivität und Spezifität darzustellen. Sie bietet einen visuellen Rahmen, um zu verstehen, wie sich die Sensitivität und die Spezifität eines Klassifikationsmodells in Abhängigkeit von verschiedenen Schwellenwerten für die Vorhersagewahrscheinlichkeit verändern. Dies ermöglicht es, die Leistung des Modells in Bezug auf die Unterscheidung zwischen den beiden Klassen zu bewerten und den optimalen Schwellenwert auszuwählen. Mit steigendem Schwellenwert steigt die Sensitivität, was bedeutet, dass das Modell besser darin ist, tatsächlich positive Fälle zu erkennen. Gleichzeitig fällt die Spezifität, da die Wahrscheinlichkeit sinkt, dass negative Fälle fälschlicherweise als positiv klassifiziert werden. Die ROC-Kurve ist ein nützliches Werkzeug, um die Kompromisse zwischen Sensitivität und Spezifität zu visualisieren und zu verstehen.

Wie wir gesehen haben, sind wir *gezwungen zwischen Sensitivität und Spezifität abzuwägen*. Es gilt: **je mehr positive Ergebnisse erfasst werden, desto mehr negative Ergebnisse werden fälschlicherweise als positiv klassifiziert (mehr false positives)**. Wir nutzen die ROC-Kurve, um den Konflikt zwischen Sensitivität und Spezifität abzubilden.

> genutzt, um die Güte eines Klassifizierungsmodells zu bewerten.

Wann sollte ROC genutzt werden und welche Alternative hat man?

Wenn die Daten balanciert sind. Ansonsten Precision and Recall Plot nutzen.

Wie verhält sich Cut-Off-Wert zu Sensitivität und Spezifität?

Mit aufsteigendem Cut-Off steigt die Sensitivität und die Spezifität fällt (Abbildung 10).

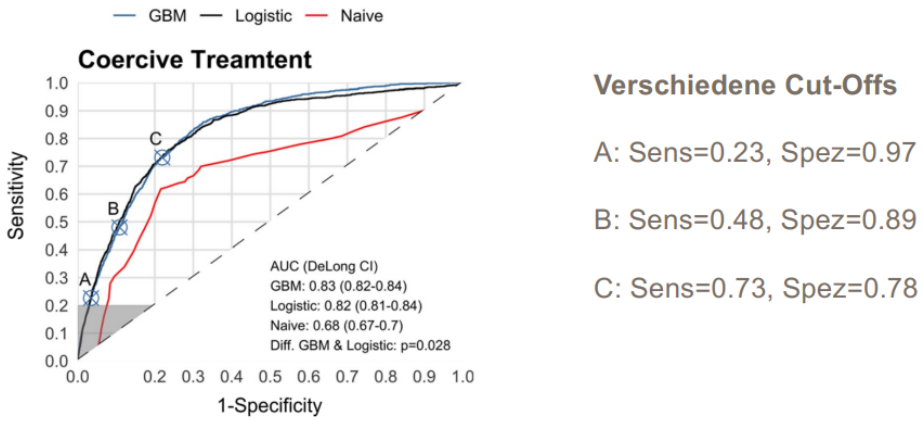


Abbildung 10: Visualisierung der Cut-Offs in ROC.

Abbildung 10 zeigt, dass mit steigendem Cut-Off die Sen. steigt, während Spez. Fällt. Der Cut-Off A hat eine Sens von 23 % (Modell hat 23% der wirklich positiven Werte richtig als positiv erkannt) und eine Spezifität von 97 %, weil es nur sehr wenige Fälle als positiv bewertet hat (wegen dem niedrigen Threshold).

Wie generiert man eine ROC-Kurve?

Man trägt die Fälle nach aufsteigender Wahrscheinlichkeit auf. Dann setzt man Cutoffs und trägt das Verhältnis aus Spezifität und Sensitivität auf. Das sieht man in der Abbildung anhand der Kurve.

Auf der Y-Achse wird die Sensitivität aufgetragen.

Für die X-Achse hat man zwei Varianten:

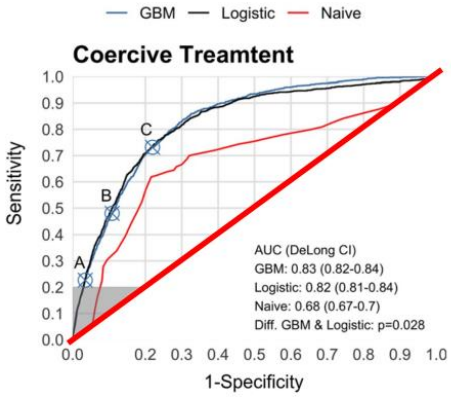
1. Spezifität auf der X-Achse aufgetragen mit den Wert 1 auf der linken Seite und 0 auf der rechten Seite (absteigend).
2. 1- Spezifität auf der X-Achse aufgetragen mit dem Wert 0 auf der linken Seite und 1 auf der rechten Seite (aufsteigend).

Wie sieht das Modell aus, wenn man viele Feature-Variablen hat und wie, wenn man wenige hat?

Viele Feature-Variablen: Wenn ein Modell mit vielen Feature-Variablen arbeitet, erzeugt es oft eine glattere und kontinuierlichere ROC-Kurve, wie sie in der dunklen Linie in Abbildung 9 dargestellt ist. Das bedeutet, dass das Modell eine größere Anzahl von unterschiedlichen Wahrscheinlichkeiten für die Vorhersage generiert, was zu einer kontinuierlichen Variation von Sensitivität und Spezifität führt.

Wenige Features: Im Gegensatz dazu, wenn ein Modell nur wenige unterschiedliche Wahrscheinlichkeiten vorhersagt, weil es nur wenige Feature-Variablen als Eingabe verwendet, führt dies oft zu einer "kantigeren" ROC-Kurve. Dies wird in der roten Linie in Abbildung 9 deutlich. Bei wenigen Features sind die möglichen Kombinationen und Variationen begrenzt, was zu einer diskreteren Darstellung von Sensitivität und Spezifität führt.

Was bedeutet ein Wert von 0,5 in der ROC-Kurve?



- Rein zufällige Klassifizierung

Abbildung 11: ROC Diagonale Linie.

Im Kontext von ROC-Kurven repräsentiert ein Wert von 0,5 die sogenannte Zufallslinie oder diagonale Linie, die von der linken unteren Ecke zur rechten oberen Ecke des Diagramms verläuft. Diese Linie zeigt an, wie das Modell abschneiden würde, wenn es rein zufällige Vorhersagen treffen würde, ohne jegliche Unterscheidungsfähigkeit zwischen den Klassen.

Ein ROC-AUC-Wert (Area Under the Curve) von 0,5 bedeutet, dass die Leistung des Modells nicht besser ist als der Zufall. Mit anderen Worten, das Modell ist nicht in der Lage, zwischen den beiden Klassen besser zu unterscheiden als ein Münzwurf oder eine reine Vermutung. Dies ist in der Regel ein Hinweis darauf, dass das Modell unbrauchbar ist oder dass es erhebliche Probleme bei der Vorhersage gibt.

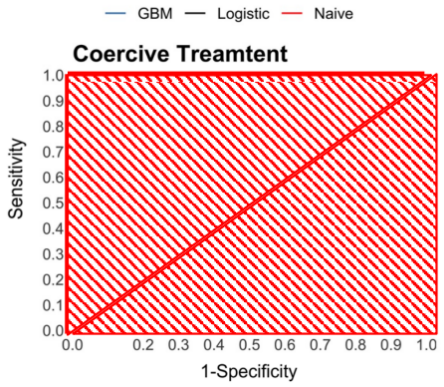
In der Praxis strebt man an, einen ROC-AUC-Wert deutlich über 0,5 zu erzielen, um sicherzustellen, dass das Modell tatsächlich eine sinnvolle und nützliche Unterscheidung zwischen den Klassen trifft. Ein ROC-AUC-Wert von 0,5 ist ein Indikator für schlechte Modellleistung, während Werte über 0,5 auf eine bessere Leistung hinweisen, wobei höhere Werte normalerweise auf bessere Diskriminierungsfähigkeit hindeuten.

> 0,5 oder darunter ist schlecht.

> nahe an 1 = besser

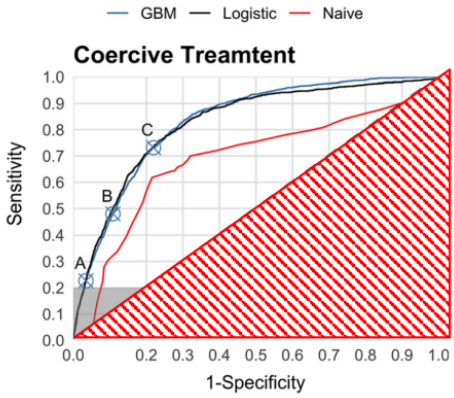
> 1 wäre Perfekt, aber wird man nie sehen.

Was ist die AUC?



Area Under the Curve (AUC)

- 1= Perfekt



Area Under the Curve (AUC)

- 0,5= Zufällige Vorhersage

Das ist eine Metrik, die in einer Zahl wiedergibt, wie gut das Modell gemessen an der ROC ist. Die AUC gibt die Fläche unterhalb der Kurve wieder.

1 wäre der beste Wert, aber das gibt es in der Praxis nicht. **Ein äußerst effektiver Klassifikator hat eine ROC-Kurve, die sich der oberen linken Ecke annähert**.

0,5 wäre der reine Zufall.

Interpretation: Gibt die Wahrscheinlichkeit an, dass ein zufällig ausgewählter, tatsächlich positiver Fall eine höhere vorhergesagte Wahrscheinlichkeit hat, als ein zufällig ausgewählter, tatsächlich negativer Fall

$$
\text{If}\begin{cases} ROC = 0.5 & \text{This suggests no discrimination, so we might as well flip a coin.} \\ 0.5 < ROC < 0.7 & \text{We consider this poor discrimination, not much better than a coin toss.} \\ 0.7 \leq ROC < 0.8 & \text{We consider this acceptable discrimination.} \\ 0.8 \leq ROC < 0.9 & \text{We consider this excellent discrimination.} \\ ROC \geq 0.9 & \text{We consider this outstanding discrimination.} \end{cases}
$$

Was sind die Vorteile und Nachteile der AUC und ROC?

+ **Gesamtbewertung der Performance in einer Kennzahl**: Die AUC und ROC fassen die Leistung eines Klassifikationsmodells in einer einzigen Zahl (AUC) oder einem grafischen Diagramm (ROC) zusammen. Dies erleichtert die schnelle Beurteilung der Modellleistung, ohne dass eine Vielzahl von Metriken berücksichtigt werden muss.

+ **Häufige Verwendung ermöglicht Benchmarking**: Kann mein Modell mit anderen Modellen anhand ihrer Kennzahl vergleichen.

Nachteile:

- **AUC hat relativ wenig klinische Bedeutung**

- **ROC deckt auch klinisch sinnlose Bereiche ab** (siehe zB grauer Bereich)

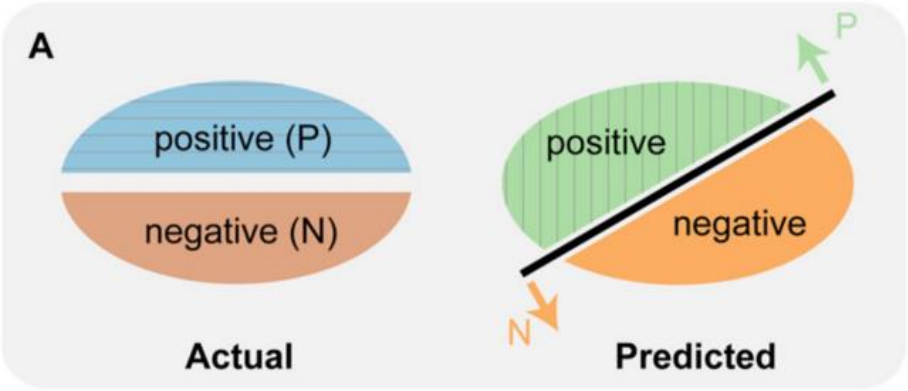
- **verschiedene Modelle können gleiche AUC haben, sich jedoch in den relevanten Bereichen unterscheiden.**

- **ROC hat Probleme bei unbalanced Data**.

ROCs Probleme mit unbalanced DATA und die Alternative

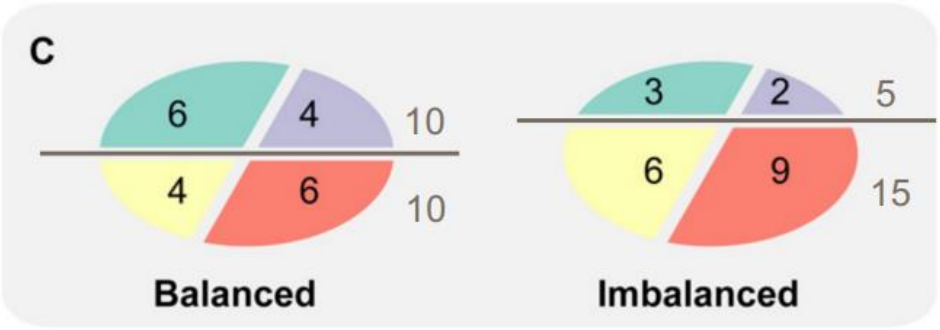
Unbalanced DATA

ROC funktioniert nicht, wenn man es auf unbalanced Data anwenden will.



In der Abbildung sieht man links die tatsächliche Aufteilung der Fälle in positiv und negativ. Da ein Vorhersage-Modell nie perfekt sein kann, weicht die tatsächliche Aufteilung von der Vorhersage ab. Das sieht man rechts. Man sieht, wenn man ausgewogene Daten vorliegen hat, dann ist die Prävalenz 50 % (50 % aller Fälle sind tatsächlich positiv (AP)).

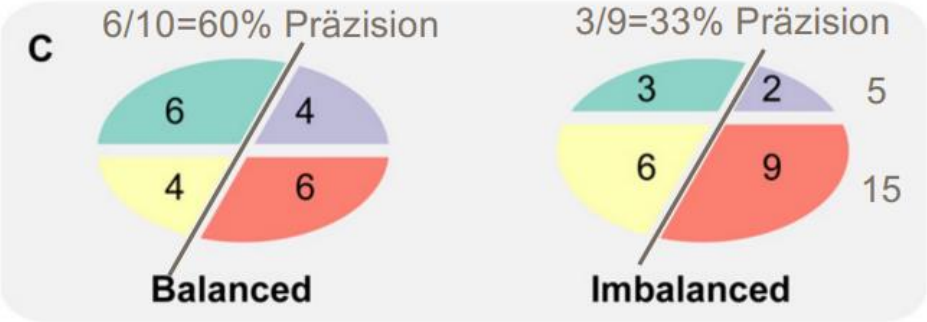
Frage: Was sind unbalanced Datensätze?



Das waren Beispiele für balancierte Datensätze. Unausgeglichene Datensätze (unbalanced datasets) sind Datensätze, in denen das Verhältnis zwischen den Klassen oder den Zielvariablen stark unausgeglichen ist. Das bedeutet, dass eine Klasse (normalerweise die Minderheitsklasse) in der Stichprobe erheblich seltener vorkommt als die andere Klasse (normalerweise die Mehrheitsklasse).

> In Wirklichkeit sind nicht balancierte Datensätze der Normalfall, da die interessanten Ergebnisse (z.B. Krankheiten) selten sind.

> Unbalanced Datensätze haben eine Prävalenz ungleich 50 %.



Präzision ändert sich zwischen Balancierten und unbalancierten Datensätzen.

Welche Metriken variieren zwischen balanced und unbalanced Datensätzen?

Table 2. Example of basic evaluation measures on a balanced and on an imbalanced dataset.

Measure	Balanced	Imbalanced
ACC	0.6	0.6
ERR	0.4	0.4
SN (TPR, REC)	0.6	0.6
SP	0.6	0.6
FPR	0.4	0.4
PREC (PPV)	0.6	0.33
MCC	0.2	0.17
$F_{0.5}$	0.6	0.37
F_1	0.6	0.43
F_2	0.6	0.52

Metriken, die gleichbleiben:

- Genauigkeit (Accuracy)
- Errorrate (1-Accuracy)
- Spezifität
- Sensitivität / Recall

Metriken, die sich zwischen Balanced und Unbalanced ändern:

- Präzision
- Matthews-Korrelationskoeffizient (MCC)
- F1-Score.

Frage: Kann ich anhand der ROC-Kurve sagen, ob Daten balanciert oder unbalanced sind?

Das heißt, die AUC kann nicht sagen, ob Daten balanciert oder unbalanced sind, weil sich die Sensitivität und die Spezifität (beide auf den Achsen der ROC-Kurve aufgetragen) nicht verändert! Die ROC kann also klinisch relevante Unterschiede in der Peformance bei unbalancierten Daten nicht wiedergeben.

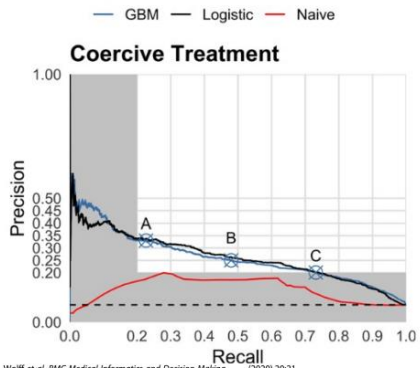
Wieso brauchen wir die anderen Metriken? Wieso reicht Genauigkeit nicht aus?

Wenn Sie einen Klassifikator erstellen, der etwas mit einer Genauigkeit von 99 % voraussagt, was sagt Ihnen das? 99 % klingt gut, berücksichtigt aber nicht die fehlende Balance zwischen den Kategorien nicht. Dies bedeutet, dass Genauigkeit allein nicht ausreicht, um die Leistung eines Klassifikators angemessen zu bewerten. Genauigkeit berücksichtigt nicht, wie gut der Klassifikator mit verschiedenen Kategorien umgeht, insbesondere wenn die Datensätze ungleich verteilt sind.

Ein Beispiel dafür ist, wenn wir einen medizinischen Test betrachten, der darauf abzielt, eine seltene Krankheit zu erkennen, von der nur 1 % der Bevölkerung betroffen ist. Selbst wenn der Test in 99 % der Fälle richtig ist, könnte er alle Fälle der Krankheit verpassen und dennoch eine hohe Genauigkeit aufweisen, da die meisten Menschen die Krankheit nicht haben (Erinnerung: True Negatives stehen bei Genauigkeit im Zähler. Wenn nur 1 % der Fälle positiv sind, dann kann ich auch alles als negativ vorhersagen und trotzdem eine hohe Genauigkeit erreichen, weil der True Negative Wert wegen der großen Unausgewogenheit des Datensatzes sehr groß sein wird). Aus diesem Grund brauchen wir andere Metriken wie Sensitivität und Präzision für die Bewertung unseres Klassifikators.

Welche Alternative zur ROC und AUC hat man, die mit unbalanced Data umgehen kann?

ROC Alternative bei unbalanced Data: Precision and Recall Plot



Präzision= TP/TP+FP

- Wahrscheinlichkeit eines positiven Falles, wenn die Vorhersage positiv war
- Eine hohe Präzision ist z.B. das Ziel, wenn Fehlalarme vermieden werden sollen

Recall= Sensitivität

Alternative zu ROC: **Precision- Recall-Plot**.

Anders als ROC trägt der Precision-Recall-Plot die **Sensitivität gegen die Präzision** auf, nicht Sensitivität gegen die Spezifität.

Achtung! In ROC-Kurven steht oft “Sensitivität” in der Abbildung, während Precision-Recall-plots “Recall” nutzen. Das ist aber DASSELBE!

Wie interpretiert man eine RPR?

Idealer Punkt:

Der ideale Punkt auf der PR-Kurve ist oben rechts, wo sowohl Präzision als auch Recall maximal sind. In diesem Fall werden alle positiven Fälle korrekt erkannt, und es gibt keine falsch positiven Vorhersagen.

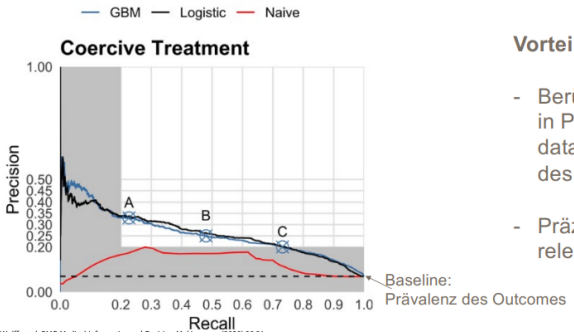
Schlechtester Punkt:

Der schlechteste Punkt auf der PR-Kurve ist unten links, wo sowohl Präzision als auch Recall minimal sind. Dies bedeutet, dass das Modell keine positiven Fälle erkennt und alle Vorhersagen falsch positiv sind.

Eine Precision von 0,5 spricht für no Skill. Ein Perfekter Klassifikator hätte Precision = 1 und Recall = 1.

Was sind die Vorteile der PRP?

ROC Alternative bei unbalanced Data: Precision and Recall Plot



Vorteile PRP:

- Berücksichtigt die Unterschiede in Präzision durch unbalanced data (AUC Baseline= Prävalenz des Outcomes)
- Präzision klinisch in vielen Fällen relevanter als Spezifität

> Zeige Immer ROC und RPR!

		Predicted condition		Sources: [9][10][11][12][13][14][15][16] view · talk · edit	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) = TPR + TNR − 1	Prevalence threshold (PT) = $\frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power = $\frac{TP}{P}$ = 1 − FNR	False negative rate (FNR), miss rate = $\frac{FN}{P}$ = 1 − TPR
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection	False positive rate (FPR), probability of false alarm, fall-out = $\frac{FP}{N}$ = 1 − TNR	True negative rate (TNR), specificity (SPC), selectivity = $\frac{TN}{N}$ = 1 − FPR
	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 = $\frac{2PPV \times TPR}{PPV+TPR} = \frac{2TP}{2TP+FP+FN}$	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 = $\frac{TP}{TP+FN+FP}$