LU03d - Kaggle.com: Begrifflichkeiten

Einleitung

Um die Titanic-Challenge erfolgreich zu meistern, ist ein tiefes Verständnis dieser Kernbegriffe unerlässlich. Nachfolgend finden Sie eine kleine Auswahl.

1. Random Forest

Stellen Sie sich vor, Sie fragen nicht nur einen Experten um Rat, sondern ein ganzes Komitee. Ein Random Forest (zu Deutsch: Zufallswald) besteht aus einer Vielzahl von einzelnen Entscheidungsbäumen, die unabhängig voneinander trainiert werden.

- **Das Prinzip:** Jeder Baum im Wald trifft eine eigene Entscheidung (z. B. „Passagier überlebt“ oder „Passagier stirbt“). Das Endergebnis ist die Mehrheitsentscheidung aller Bäume.
- **Der Vorteil:** Da die Bäume auf unterschiedlichen Datenstichproben basieren, korrigieren sie gegenseitig ihre Fehler. Dies macht das Modell extrem robust gegen Rauschen in den Daten.

2. Mean (Arithmetisches Mittel)

Der Mean ist der klassische Durchschnittswert. Sie berechnen ihn, indem Sie alle Werte einer Spalte addieren und durch die Anzahl der Werte teilen.

- **Anwendung:** Er eignet sich gut für gleichmässig verteilte Daten.
- **Das Problem:** Er ist sehr anfällig für Ausreisser. Wenn ein einziger Passagier ein extrem teures Ticket für 500 Dollar gekauft hat, während alle anderen 10 Dollar zahlten, zieht dieser eine Wert den Durchschnitt stark nach oben und verzerrt das Bild.

3. Median (Zentralwert)

Der Median ist der Wert, der genau in der Mitte einer sortierten Datenreihe liegt. 50 Prozent der Werte sind kleiner oder gleich, 50 Prozent sind grösser oder gleich dem Median.

- **Anwendung:** Bei der Titanic-Challenge ist der Median oft besser als der Mean geeignet, um fehlende Altersangaben zu füllen.
- **Der Vorteil:** Er ist „robust“. Der oben genannte Passagier mit dem 500-Dollar-Ticket lässt den Median völlig unbeeindruckt. Er repräsentiert die „typische“ Mitte viel besser als der Durchschnitt.

4. Confusion Matrix (Konfusionsmatrix)

Die Confusion Matrix ist das ultimative Werkzeug, um die Leistung Ihres Klassifikationsmodells zu bewerten. Sie zeigt Ihnen nicht nur, wie oft Sie recht hatten, sondern auch, welche Art von Fehlern Sie machen.

Sie ist typischerweise als Tabelle aufgebaut:

	Vorhersage: Überlebt	**Vorhersage: Gestorben**
Realität: Überlebt	**True Positive (TP)**	False Negative (FN)
Realität: Gestorben	False Positive (FP)	**True Negative (TN)**

- **True Positive (TP):** Sie haben korrekt vorhergesagt, dass jemand überlebt.
- **True Negative (TN):** Sie haben korrekt vorhergesagt, dass jemand stirbt.
- **False Positive (FP):** Sie sagten „Überlebt“, aber die Person ist gestorben (ein „falscher Alarm“).
- **False Negative (FN):** Sie sagten „Gestorben“, aber die Person hat überlebt (ein „übersehener“ Überlebender).

5. Der Modus (Modalwert)

Während Mean und Median für numerische Werte (wie Alter oder Ticketpreis) gedacht sind, ist der Modus der Wert, der in einer Datenreihe am häufigsten vorkommt.

- **Warum er für Sie wichtig ist:** Stellen Sie sich vor, Sie analysieren die Spalte Embarked (Einstiegshafen). Sie können aus den Buchstaben „S“, „C“ und „Q“ keinen Durchschnitt (Mean) berechnen. Wenn dort Werte fehlen, schauen Sie, welcher Hafen am häufigsten vorkommt (bei der Titanic war dies „S“ für Southampton) und füllen die Lücken mit diesem Modus.
- **Anwendung:** Er ist das Standardwerkzeug zur Imputation (Auffüllen) von fehlenden Werten bei kategorialen Variablen (Text-Spalten).

6. Die Standardabweichung (Standard Deviation)

Dieser Wert beschreibt, wie stark die Daten um den Mittelwert (Mean) streuen.

- **Praxisbezug:** Eine hohe Standardabweichung beim Ticketpreis (Fare) verrät Ihnen, dass die Preise sehr weit auseinandergingen – von extrem günstigen Tickets bis hin zu Luxussuiten. Eine niedrige Standardabweichung beim Alter würde bedeuten, dass die meisten Passagiere einer sehr ähnlichen Altersgruppe angehörten.
- **Nutzen:** Sie hilft Ihnen zu entscheiden, ob Sie die Daten skalieren müssen, damit extreme Streuungen Ihr Modell nicht verwirren.