Übungsaufgaben zur Veranstaltung „Modell Validierung“

# Übung 1

Wir nutzen den „Telco Customer Churn“ Datensatz von Kaggle: <https://www.kaggle.com/datasets/blastchar/telco-customer-churn>.

Die csv.Datei liegt bereits im Ordner Übungen.

## Beschreibung aus Kaggle

Each row represents a customer, each column contains customer’s attributes described on the column Metadata.

The data set includes information about:

* Customers who left within the last month – the column is called Churn
* Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
* Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges
* Demographic info about customers – gender, age range, and if they have partners and dependents

## Szenario

Sie sind als Data Scientist in einer Telekommunikationsfirma angestellt. Ihre Vorgesetzte beauftragt Sie, eine gezielte Marketingmaßnahme zu beraten, in der Kunden mit einem hohen Risiko der Vertragskündigung (Churn) adressiert werden sollen. Sie bekommen dazu einen Datensatz zur Verfügung gestellt.

Ihre Vorgesetze legt aus Kosten- und Nachhaltigkeitsgründen besonders viel Wert darauf, nicht zu viele Briefe umsonst zu verschicken, d.h. sie möchte Briefe wenn möglich nur Kunden adressieren, die ohne Brief kündigen würden. Sie möchte aber auch einen möglichst hohen Anteil der Kunden erreichen, die ohne Brief kündigen würden.

## Aufgabe 1

1. Bitte importieren Sie die csv.Datei.
2. Überprüfen Sie die Daten, filtern Sie „störende“ Daten heraus und überprüfen Sie die relevanten Daten.
3. Splitten sie die Daten zufällig in 80% Training und 20% Test. An dieser Stelle nutzen wir nur eine Zweiteilung zwischen Train und Test, ohne eine getrenntes Validierungsset.
4. Trainieren Sie ein Klassifikationsbaum zur Klassifikation der Variable „Churn“ im Training Set. Bitte nutzen Sie die default-Hyperparameter.

## Aufgabe 2

1. Testen Sie die Güte Ihres Modells anhand der visuellen ROC und der AUC im Test-Set.
2. Reichen die Informationen aus der ROC aus, um Entscheidungen über die praktische Anwendung Ihres Modells nach den Wünschen Ihrer Vorgesetzten zu spezifizieren? Bitte begründen Sie Ihre Antwort.
3. Bitte vergleichen Sie verschiedene Wahrscheinlichkeits-Cut-Offs und wählen Sie einen optimalen Cut-Off, um Ihre Marketingmaßnahmen zu steuern. Bitte begründen Sie, anhand welcher Performance Metriken Sie Ihr Modell spezifiziert haben.