Hochschule Bochum
Bochum University
of Applied Sciences

Fachbereich Wirtschaft
Prof. Dr. Christian Bockermann

Data Science

Sommersemester 2021

Übungsblatt 4

Aufgabe 1 (Daten Exploration)

Im Verzeichnis `Kurse/DataScience1/data/` finden Sie den Datensatz `telco-churn.csv`. Der Datensatz enthält Kunden eines Telekommunikationsunternehmens und deren Vertragseigenschaften (Geschlecht, Telefon: ja/nein, Internet: ja/nein,...).
Die Spalte `Churn` gibt an, ob der Kunde im letzten Monat seinen Vertrag gekündigt hat, oder nicht. Das Thema *Churn Prediction* ist ein typischer Anwendungsfall in vertragsbasierten Geschäftsmodellen.

1. Laden Sie den Datensatz in einen DataFrame `churn`.
   Welche Spalten hat der Datensatz? Welchen Typ haben die Spalten?

2. Wenden Sie den folgenden Befehl auf dem DataFrame an:

   ```
   churn.replace({ "PhoneService": { 'Yes': 1, 'No': 0 } })
   ```

   Wie verändert dies den Datensatz? Was hat das für Vorteile?

3. Wie hoch ist die Churn-Rate? Bei welchem Geschlecht liegt die Churn-Rate höher?

4. Wie hoch ist der Anteil an weiblichen Kunden, die einen DSL-Anschluss besitzen? Welche Anschluss-Arten gibt es noch? Wie ist deren Verteilung?

Aufgabe 2 (Vorhersage-Fehler)

In dieser Aufgabe schauen wir uns nochmal den *Churn Prediction* Datensatz an.
In Pandas können wir eine Spalte mit gleichen Werten für jede Zeile erzeugen, indem wir eine Zahl der Spalte zuweisen, z.B.:

```
df = pd.read_csv(..)

df['x1'] = 42      # Spalte 'x1' ist jetzt immer 42
```

Die Aufgabe im *Churn Prediction* Datensatz ist die Vorhersage der Spalte `churn`. Im folgenden wollen wir uns mit dem Vorhersage-Fehler eines Modells auf diesem Datensatz beschäftigen.

1. Angenommen, wir haben ein Modell, dass immer den Wert 1 vorhersagt.
   Erzeugen Sie im DataFrame eine Spalte `y_hat`, die nur aus 1en besteht.

   Zählen Sie, wie häufig in diesem Datensatz die Spalte `churn` und `y_hat` übereinstimmen. Wie groß ist die relative Häufigkeit der Zeilen, in denen diese beiden Spalten *nicht* übereinstimmen?

2. Wie groß ist der relative Vorhersage-Fehler bei den männlichen bzw. bei den weiblichen Kunden?

3. Schreiben Sie eine Funktion `rel_error(s1, s2)`, die als Parameter zwei Series-Objekte (Spalten) bekommt und berechnet, wie hoch der Anteil der Zeilen ist, in denen `s1` und `s2` den gleichen Wert haben.

Aufgabe 3 * (Modellfehler berechnen)

In der Vorlesung wurde u.a. der Klassifier `Zufall` aus dem `datascience` Modul vorgestellt. Die Klasse `Zufall` sammelt in der Trainingsphase alle Werte der Label-Spalte ein und berechnet bei der Vorhersage jeweils einen zufälligen Wert aus diesen eingesammelten Werten als Vorhersageergebnis.

1. Benutzen Sie das Pandas Modul um den *Iris* Datensatz einzulesen. Erstellen Sie aus dem Datensatz einen DataFrame `X`, der die Merkmalsspalten enthält (alle ausser `species`) und eine Series `y`, die die `species` Spalte enthält.

2. Erzeugen Sie ein neues Modell `m` der Klasse `Zufall` und trainieren Sie es auf `X` und `y`. Nutzen Sie das Modell `m` um auf dem DataFrame `X` die Vorhersagen zu berechnen und speichern Sie diese in der Variable `y_hat`.

3. Definieren Sie eine Funktion `errors`, die zwei Series Objekte als Eingabe bekommt und die Anzahl der Stellen berechnet, an denen die eingegebenen Series Objekte nicht gleich sind.

4. Die Klasse `Zufall` hat zusätzlich die Methode `describe()`, die einen Überblick über die Verteilung der Klassen aus der Trainingsphase enthält. Schauen Sie sich die Ausgabe von `describe()` an.

   Welche durchschnittliche Fehlerrate erwarten Sie, wenn Sie das Modell auf dem Trainingsdatensatz `X, y` auswerten? (Trainingsfehler)

5. Definieren Sie eine Funktion `rel_errors`, die aus zwei Series Objekten den durchschnittlichen Fehler berechnet. Benutzen Sie als Fehlerfunktion dazu die Funktion `errors`, die Sie zuvor definiert haben.