Data Science

Übungsblatt 4

Aufgabe 1 (Wiederholung Pandas DataFrames)

Unter der URL

`https://data.hsbo.de/bitcoin.csv`

haben wir Daten zum Kursverlauf der Kryptowährung Bitcoin bereitgestellt. Laden Sie diese Datei mit Hilfe von Pandas in einen DataFrame und untersuchen Sie den Datensatz. Folgende Fragen sind dabei von Interesse:

- Welche Spalten gibt es? Welchen Datentyp haben die Spalten?
- Berechnen Sie den minimalen, maximalen und durchschnittlichen Schlusskurs über den gesamten Datensatz.
- Berechnen Sie eine neue Spalte Mittelwert, die das arithmetische Mittel aus Höchst- und Tiefpunkt je Tag enthält.

Aufgabe 2 (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 3 (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.