KI Summer School

Übung 2
Logistische Regression

Hochschule Karlsruhe
Technik und Wirtschaft
UNIVERSITY OF APPLIED SCIENCES

Prof. Dr. Patrick Baier

Titanic Datensatz

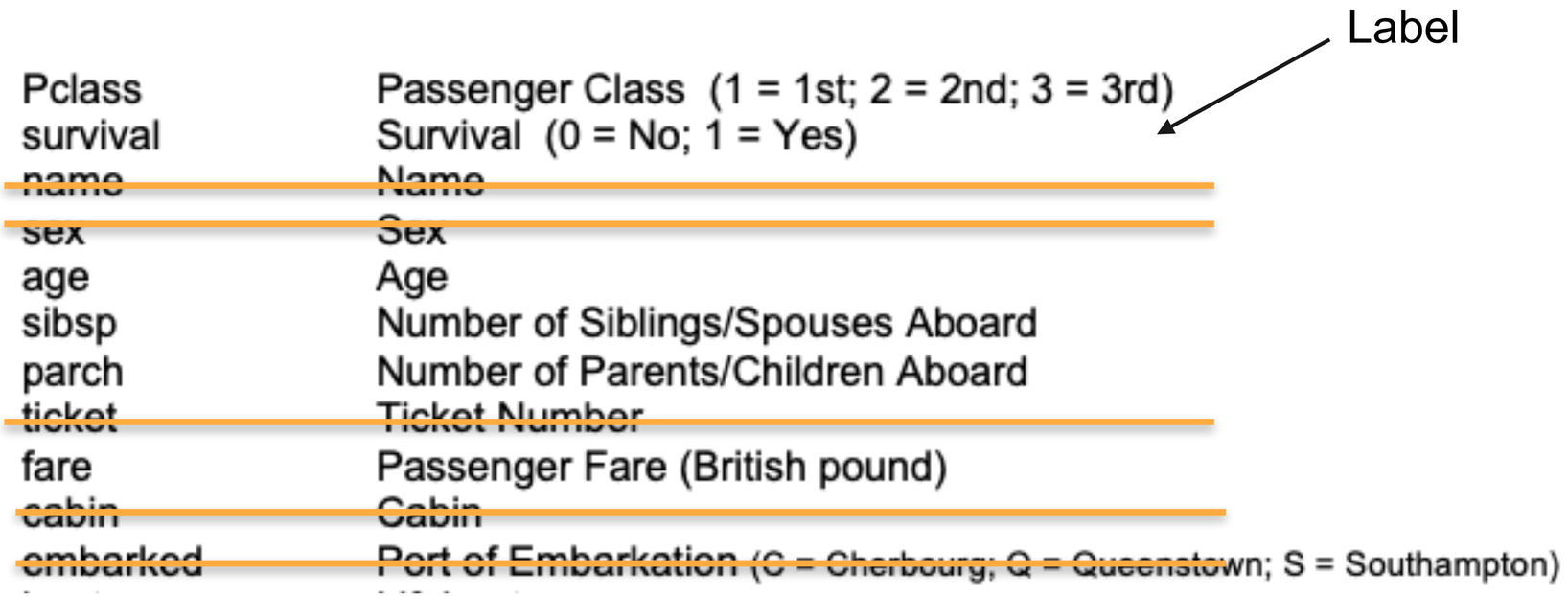
Vorhersageproblem:

- Gegeben: Verschiedene Attribute zu den Passagieren auf der Titanic.
- Frage: Hat der Passagier das Unglück überlebt?

Pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
survival	Survival (0 = No; 1 = Yes)
name	Name
sex	Sex
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare (British pound)
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

Titanic Datensatz

Da wir noch nicht den Umgang mit kategorischen Features eingeführt haben, benutzen wir nur die folgenden Attribute:



Aufgabe 1

1. Laden Sie den Datensatz aus `titanic.csv` in einen Pandas DataFrame.
2. Erstellen Sie einen neuen DataFrame, der nur die folgenden Spalten enthält: "Survived", "Pclass", "Age", "Fare", „SibSp" und „Parch"
3. Entfernen Sie die leere Feldern aus dem DataFrame (d.h. Felder mit `NaN`-Werte) indem Sie die Methode `dropna()` auf dem DataFrame aufrufen. Zählen Sie wie viele Zeilen der DataFrame vor und nach dem Aufruf dieser Methode hat.

Aufgabe 2

1. Unterteilen Sie den aus Aufgabe 1 entstandenen DataFrame mit Hilfe der Methode `train_test_split` in Trainings- und Testdaten.
2. Trainieren Sie eine logistische Regression auf den Trainingsdaten mit „survival“ als Label.
3. Machen sie mit dem trainierten Modell Vorhersagen auf den Testdaten und berechnen Sie:
    1. Accuracy
    2. Precision
    3. Recall.

Aufgabe 3

1. Fügen Sie nun das Geschlecht als weitere Feature hinzu und führen Sie die Schritte aus Aufgabe 1 und 2 erneut aus. Beachten Sie dabei, dass das Geschlecht kein numerischer Wert ist, d.h. Sie müssen daraus ein numerisches Feature erstellen.
   Beispiel: Feature „isFemale“ hat den Wert 1 wenn Sex==„female“, sonst 0.
2. In wie weit verbessert sich die Accuracy durch das neue Feature?
3. Versuchen Sie nun auf ähnliche Weise die Spalte „Embarked“ als Feature zu nutzen.

Sex
male
female
female
male

isFemale
false
true
true
false

Hinweis: Python wandelt Booleans automatisch in Nummern um:

isFemale
0
1
1
0

Bonus:

1. Welches sind die wichtigsten Features des Modells für die Vorhersage?
2. Wie verändert sich die Accuracy wenn Sie das unwichtigste Feature weglassen?

Fragen?