

## - Klassifikation mit logistischer Regression -

Beantworten Sie folgende Fragen zum Titanic-Datensatz über Passagiere, die überlebt bzw. nicht überlebt haben (überleben ist wahr bzw. falsch).

1. Beschreiben Sie die Größen, die aus dem Training (*engl. fit*) des Modells in der logistischen Regression als Funktion des Geschlechts resultieren.

```
#Train the model
log_model.fit(X = pd.DataFrame(encoded_sex),
              y = titanic_train["Survived"])
```

### Lösung

Aus dem Training resultieren die Modellparameter für die logistische Regression:

- (a) Regressionskoeffizient  $b$  (*vgl. Vorlesung*) stellt in der linearen Regression das Gewicht der Variablen Geschlecht dar und
  - (b)  $a$  den y-Achsenabschnitt (*engl. Intercept*).
2. Beurteilen Sie die Aussagekraft der logistischen Regression für die Berechnung der Überlebenswahrscheinlichkeit als Funktion des Geschlechts anhand der Ergebnisse der Kreuztabelle:

```
# Make predictions
preds = log_model.predict_proba(X= pd.DataFrame(encoded_sex))
preds = pd.DataFrame(preds)
preds.columns = ["Death_prob", "Survival_prob"]

# Generate table of predictions vs Sex
pd.crosstab(titanic_train["Sex"], preds.loc[:, "Survival_prob"])
```

Survival_prob	0.19314664603456402	0.7342488525817924
Sex		
female	0	314
male	577	0

Kreuztabelle (*engl. crosstab*) für die Überlebenswahrscheinlichkeit als Funktion des Geschlechts

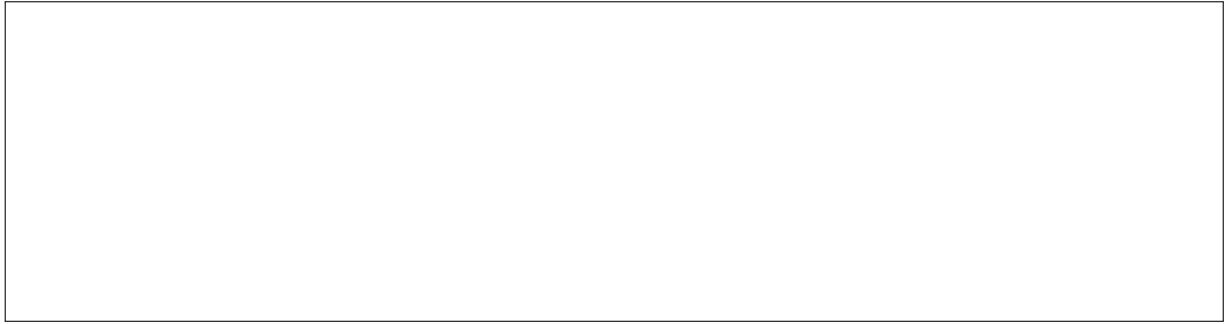
### Lösung

Die Aussagekraft ist gering, da das Modell bei einer überleben ist *wahr* bzw. *falsch* Entscheidung lediglich aussagen würde, dass Frauen überleben und Männer nicht.

3. Interpretieren Sie die Ergebnisse der Wahrheitsmatrix, die aus dem Training eines Modells mit logistischer Regression von insgesamt vier Variablen entstanden ist:

Survived	0	1
row_0		
0	458	89
1	91	253

Wahrheitsmatrix



### Lösung

- Unten rechts: Wahr positiv. Passagiere, die vom Modell als Überlebende vorhergesagt wurden und es tatsächlich waren.
- Oben links: Wahr negativ. Modell sagt kein Überleben voraus, was der Realität entspricht.
- Oben rechts: Falsch negativ: Modell sagt kein Überleben voraus, was nicht der Realität entspricht.
- Unten links: Falsch positiv: Modell sagt Überleben voraus, was nicht der Realität entspricht.