TU DORTMUND
Fakultät Statistik
Prof. Dr. Jörg Rahnenführer
M.Sc. Marieke Stolte
Dr. Franziska Kappenberg
Dr. Julia Duda

Wintersemester 2024/2025

Projekt 5 - Logistische Regression

Am 5. November 2024 fand in den Vereinigten Staaten von Amerika die Wahl der Delegierten für das Electoral College statt. Das Electoral College wählt den Präsidenten der Vereinigten Staaten. Am 17. Dezember 2024 wählte das am 5. November 2024 gewählte Electoral College den demokratischen Präsidenschaftskandidaten Donald Trump zum 47. Präsidenten der USA. Der Wahlsieger wurde am 6. Januar 2025 offiziell verkündet und wird am 20. Januar ins Amt eingeführt.

Die Datei *US_election_2024.csv* enthält die folgenden Variablen:

- *State*: Bundesstaat der USA (+ District of Columbia)
- *Leading_Candidate*: Gewinner bei der Wahl des Electoral College am 5. November 2024 (Harris oder Trump) (The Associated Press, 2024)
- *Total_Area*: Gesamtfläche in Square Miles (U.S. Census Bureau, 2010)
- *Population*: Geschätzte Bevölkerungszahl (U.S. Census Bureau, 2020)
- *Population_Density*: Bevölkerungsdichte, mittlere Einwohnerzahl pro Square Mile im Jahr 2020 (U.S. Census Bureau, 2020)
- *Median_Age*: Medianes Alter in Jahren im Jahr 2023 (U.S. Census Bureau, U.S. Department of Commerce, 2023a)
- *Birth_Rate*: Geburtenrate, Anzahl Frauen im Alter zwischen 15 und 50 Jahren mit Geburten in den vergangenen 12 Monaten im Jahr 2023 (U.S. Census Bureau, U.S. Department of Commerce, 2023b)
- *HDI*: Index der menschlichen Entwicklung (Human Development Index) im Jahr 2022 (Global Data Lab, 2022)
- *Unemployment_Rate*: Arbeitslosenrate (U.S. Census Bureau, U.S. Department of Commerce, 2023c)
- *Health_Insurance_Coverage*: Anteil der Bevölkerung mit Krankenversicherung (U.S. Census Bureau, U.S. Department of Commerce, 2023c)
- *Median_Rent*: Mediane Miete in US-Dollar (U.S. Census Bureau, U.S. Department of Commerce, 2023d)

Aufgabenstellung

1. Modellieren Sie die Zielvariable *Leading_Candidate* anhand aller anderen gegebenen Variablen, gegebenenfalls geeignet transformiert. Verwenden Sie hierzu die logistische Regression.
2. Nutzen Sie ein geeignetes Variablenselektionsverfahren, um ein möglichst „gutes“ reduziertes Modell zu erhalten. Interpretieren Sie das resultierende Modell hinsichtlich der Koeffizientenschätzer unter Berücksichtigung der zugehörigen Konfidenzintervalle.
3. Bewerten Sie beide Modelle mittels Grafiken zur *Receiver Operating Characteristic* (ROC) und mit der *Area under the Curve* (AUC) jeweils unter Verwendung von Kreuzvalidierung. Vergleichen Sie die beiden Modelle hinsichtlich dieser Gütekriterien.

Referenzen für die Daten

- CDC/National Center for Health Statistics (2020): *Life Expectancy at Birth by State*. URL: `https://www.cdc.gov/nchs/pressroom/sosmap/life_expectancy/life_expectancy.htm` (besucht am 12.11.2024)
- Global Data Lab (2022): *Subnational HDI (v8.1)*. URL: `https://globaldatalab.org/shdi/table/shdi/USA/` (besucht am 12.11.2024)
- The Associated Press (2024): *2024 Presidential Election Results*. URL: `https://apnews.com/projects/election-results-2024/?office=P` (besucht am 12.11.2024)
- U.S. Census Bureau (2010): *State Area Measurements and Internal Point Coordinates*. URL: `https://www.census.gov/geographies/reference-files/2010/geo/state-area.html` (besucht am 12.11.2024)
- U.S. Census Bureau (2020): *Historical Population Density Data*. URL: `https://www.census.gov/data/tables/time-series/dec/density-data-text.html` (besucht am 12.11.2024)
- U.S. Census Bureau, U.S. Department of Commerce (2023a): *Age and Sex*. URL: `https://data.census.gov/table/ACSST1Y2023.S0101?g=010XX00US,$0400000` (besucht am 12.11.2024)
- U.S. Census Bureau, U.S. Department of Commerce (2023b): *Fertility*. URL: `https://data.census.gov/table/ACSST1Y2023.S1301?g=010XX00US,$0400000` (besucht am 12.11.2024)
- U.S. Census Bureau, U.S. Department of Commerce (2023c): *Selected Economic Characteristics*. URL: `https://data.census.gov/table/ACSDP1Y2023.DP03?g=010XX00US,$0400000` (besucht am 12.11.2024)
- U.S. Census Bureau, U.S. Department of Commerce (2023d): *Selected Housing Characteristics*. URL: `https://data.census.gov/table/ACSDP1Y2023.DP04?g=010XX00US,$0400000` (besucht am 12.11.2024)

Literaturempfehlungen

- Fahrmeir, L., Kneib, T., Lang, S. (2009): *Regression: Modelle, Methoden und Anwendungen*, 2. Auflage, Springer, Berlin Heidelberg.
- Agresti, A. (2007): *An Introduction to Categorical Data Analysis*, 2. Auflage, Wiley, New York.
- Tutz, G. (2012): *Regression for Categorical Data*, 1. Auflage, Cambridge University Press, New York.
- Hosmer, D. W., Lemeshow, S., Sturdivant, R. (2013): *Applied Logisitic Regression*, 3. Auflage, John Wiley & Sons, Inc., New Jersey.
- Dunn, P. K., Smyth, G. K. (2018): *Generalized Linear Models With Examples in R*, 1. Auflage, Springer, New York.
- Behnke, J. (2015): *Logistische Regressionsanalyse: Eine Einführung*, 1. Auflage, Springer, Wiesbaden.

Abgabe

Abgabe des Berichts und des zugehörigen (lauffähigen und kommentierten) Programmcodes bis Donnerstag, den 30.01.2025, 10:00 Uhr, im Moodle.