Machine Learning – Regressionsmethoden

Daten	**Kreditkartenabrechnung - credit.csv**
Inhalte	In diesem Datensatz sind Informationen über 400 Kreditkarteninhaber und ihre demografischen Merkmale wie Alter, Geschlecht, Familienstand, ethnische Herkunft, etc. enthalten.
Informationen	Der Datensatz dient der Kundenstammanalyse.
Fragestellung	Mit welcher Genauigkeit lassen sich aus den persönlichen Kundendaten deren Kreditkarten-Abrechnungen vorhersagen und was sind die stärksten Einflussfaktoren?
Vokabular	• Income: Einkommen des Kunden (in USD) • Limit: Verfügbares Kreditkarten-Limit des Kunden • Rating: Bewertung Kreditwürdigkeit • Cards: Anzahl der Kreditkarten des Kunden • Age: Alter (Jahre) • Education: Bildungslevel des Kunden (in Anzahl der Ausbildungsjahre) • Gender: Geschlecht des Kunden (male / female) • Student: Student (True / False) • Married: Verheiratet (True / False) • Ethnicity: Ethnische Herkunft des Kunden • Balance: Kreditkarten-Rechnung des Kunden (monatlicher Mittelwert)
Quelle	• https://www.statlearning.com/

1) Nutzen Sie das Script der EDA-Analyse aus ÜA2 als Ausgangspunkt für die sich jetzt anschließende statistische Modellierung. Untersuchen Sie die allgemeine Fragestellung: *Mit welcher Genauigkeit lassen sich aus den persönlichen Kundendaten die Kreditkarten-Abrechnungen vorhersagen und was sind die stärksten Einflussfaktoren?*

2) Führen Sie eine einfache lineare Regression durch! Zielvariable ist die Kreditkarten-Balance, Prädiktor des Kreditkartenlimit. Welche Güte erreichen Sie mit diesem einfachen Modell (welches Bestimmtheitsmaß R^2)?

3) Entwickeln Sie ein Modell für die Vorhersage der Kreditkarten-Balance unter Verwendung der multiplen linearen Regression. Welche Modellgüte erreichen Sie hier unter Einbeziehung mehrerer Prädiktoren in die Vorhersage?

4) Welche Merkmale haben in Ihrem Modell signifikanten Einfluss auf die Kreditkartenabrechnung? Fassen Sie ihre Ergebnisse kurz zusammen!

5) Warum ist es so wichtig, die Daten der Kunden ohne Kreditkartenumsatz aus der Modellierung auszuschließen?

- **Abgabe bis Donnerstag 28.11. 16:00 Uhr zwei Folien pro Gruppe (pdf!!), im Namen des Files bitte Gruppennummer bzw. Namen der Verfasser + Nummer der Übung verwenden (z.B. GruppeXX_uebung-a3.pdf)**

- **in Verzeichnis Upload_DASC hochladen: :** https://nc.ufz.de/s/ai2zHScBEic8S8r
  (Passwort: !DASC_ba_2024)

- Folie 1: Grafik zu Aufgaben 2 (einfache lineare Regression)
  Folie 2: Grafik auswählen aus Aufgabe 3 + 4 (multiple lineare Regression), kurze Zusammenfassung der Ergebnisse aus der Analyse
- **5-Minuten Pitch: Jede Gruppe präsentiert kurz ihre Ergebnisse (5 min maximal) zu Beginn der nächsten Veranstaltung am 29.11.24!**