

## Machine Learning – Regressionsmethoden

<b>Daten</b>	<b>Kreditkartenabrechnung - credit.csv</b>
<b>Inhalte</b>	In diesem Datensatz sind Informationen über 400 Kreditkarteninhaber und ihre demografischen Merkmale wie Alter, Geschlecht, Familienstand, ethnische Herkunft, etc. enthalten.
<b>Informationen</b>	Der Datensatz dient der Kundenstammanalyse.
<b>Fragestellung</b>	Mit welcher Genauigkeit lassen sich aus den persönlichen Kundendaten deren Kreditkarten-Abrechnungen vorhersagen und was sind die stärksten Einflussfaktoren?
<b>Vokabular</b>	<ul style="list-style-type: none"> <li>• Income: Einkommen des Kunden (in USD)</li> <li>• Limit: Verfügbares Kreditkarten-Limit des Kunden</li> <li>• Rating: Bewertung Kreditwürdigkeit</li> <li>• Cards: Anzahl der Kreditkarten des Kunden</li> <li>• Age: Alter (Jahre)</li> <li>• Education: Bildungslevel des Kunden (in Anzahl der Ausbildungsjahre)</li> <li>• Gender: Geschlecht des Kunden (male / female)</li> <li>• Student: Student (True / False)</li> <li>• Married: Verheiratet (True / False)</li> <li>• Ethnicity: Ethnische Herkunft des Kunden</li> <li>• Balance: Kreditkarten-Rechnung des Kunden (monatlicher Mittelwert)</li> </ul>
<b>Quelle</b>	<ul style="list-style-type: none"> <li>• <a href="https://www.statlearning.com/">https://www.statlearning.com/</a></li> </ul>

- 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 12.02. 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/39zCZKFAYZKDQ2z>  
(Passwort: ?solution\_DASC25)

- 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 Veranstaltung am Freitag, den 14.02.25!**