Intelligente Datenanalyse

Prüfung: Kreditwürdigkeit

Datum: 13. Juni 2016

Gerrit Gruben
Prof. Tobias Scheffer

Die Projektaufgabe ist Teil der Prüfung *Intelligente Datenanalyse*. Jede Aufgabe soll durch einen Studenten selbstständig bearbeitet und die Lösung innerhalb der mündlichen Prüfung vorgestellt werden. Ein Ausdruck des Python-Programmcodes und der Ergebnisse in Form eines Diagramms, Tabelle o.ä. werden vorausgesetzt; die Art der Präsentation der Ergebnisse ist dem Studenten freigestellt.

Problemstellung

Eine Bank möchte die Kreditwürdigkeit ihrer Kunden vorhersagen. Basierend auf den Kundendaten, der Kredithistorie usw. soll ein Kunde in kreditwürdig bzw. kreditunwürdig eingestuft werden. Dabei ist es für die Bank fünfmal so „teuer“ einen kreditunwürdigen Kunden als kreditwürdig einzuschätzen wie umgekehrt. Hinzukommt, dass für einige Attribute nur ein Teil der Daten gegeben sind. Für 1.000 repräsentativ ausgewählte Kunden ist die Kreditwürdigkeit bekannt. Von diesen Kunden wurden die folgenden Kundendaten erfasst. (Attribute für welche nicht alle Werte bekannt sind, sind mit dem Zusatz „incomplete“ markiert.)

- Status of existing checking account
  - A11: ... < 0 EUR
  - A12: $0 \leq$ ... < 200 EUR
  - A13: ... ≥ 200 EUR / salary assignments for at least 1 year
  - A14: no checking account
- Duration in month
- Credit history
  - A30: no credits taken/ all credits paid back duly
  - A31: all credits at this bank paid back duly
  - A32: existing credits paid back duly till now
  - A33: delay in paying off in the past
  - A34: critical account/ other credits existing (not at this bank)
- Purpose (incomplete)
  - A40: car (new)
  - A41: car (used)
  - A42: furniture/equipment
  - A43: radio/television
  - A44: domestic appliances
  - A45: repairs

  - A46: education
  - A47: (vacation - does not exist?)
  - A48: retraining
  - A49: business
  - A410: others
- Credit amount
- Savings account/bonds
  - A61: ... $<$ 100 EUR
  - A62: 100 $\leq$ ... $<$ 500 EUR
  - A63: 500 $\leq$ ... $<$ 1000 EUR
  - A64: .. $\geq$ 1000 EUR
  - A65: unknown/ no savings account
- Present employment since (incomplete)
  - A71: unemployed
  - A72: ... $<$ 1 year
  - A73: 1 $\leq$ ... $<$ 4 years
  - A74: 4 $\leq$ ... $<$ 7 years
  - A75: .. $\geq$ 7 years
- Installment rate in percentage of disposable income
- Personal status and sex
  - A91: male: divorced/separated
  - A92: female: divorced/separated/married
  - A93: male: single
  - A94: male: married/widowed
  - A95: female: single
- Other debtors / guarantors
  - A101: none
  - A102: co-applicant
  - A103: guarantor
- Present residence since
- Property
  - A121: real estate
  - A122: if not A121: building society savings agreement/ life insurance
  - A123: if not A121/A122: car or other, not in attribute 6
  - A124: unknown / no property
- Age in years
- Other installment plans

  - A141: bank
  - A142: stores
  - A143: none

- Housing
  - A151: rent
  - A152: own
  - A153: for free

- Number of existing credits at this bank

- Job (incomplete)
  - A171: unemployed/ unskilled - non-resident
  - A172: unskilled - resident
  - A173: skilled employee / official
  - A174: management/ self-employed/
  - A175: highly qualified employee/ officer

- Number of people being liable to provide maintenance for

- Telephone
  - A191: none
  - A192: yes, registered under the customers name

- Foreign worker (incomplete)
  - A201: yes
  - A202: no

- Creditworthy
  - 1: yes
  - 2: no

Sie wurden damit beauftragt ein Vorhersagemodell zu entwickeln welches die Kreditwürdigkeit zukünftiger Kunden bewertet. Es ist davon auszugehen, dass auch bei diesen Kunden ein Teil der Werte für die 4 Attribute „Purpose“, „Present employment since“, „Job“ und „Foreign worker“ fehlen.

Aufgabe

Lesen sie die Daten in Python ein und führen sie eine Datenvorverarbeitung durch. Ersetzen sie dabei fehlende Werte mittels linearer Regression bzw. Klassifikation. Identifizieren sie ein geeignetes Verfahren zum Lösen des Vorhersageproblems, implementieren sie dieses in Python und trainieren sie ein Modell. Evaluieren sie das Modell bzgl. des von der Bank vorgegebenen Kostenmodells. Begründen und dokumentieren sie kurz alle durchgeführten Schritte.