Nedostajuće vrednosti atributa

- Do sada smo razmatrali samo podatke gde za svaku opservaciju u skupu podataka imamo sve vrednosti atributa
- U realnim situacijama smo često suočeni sa nedostajućim vrednostima
- Npr. u problemu dodele kredita

Starost	Pol	Godišnja zarada	Trenutni dug	Poseduje nekretninu	Rok za otplatu	Odobriti kredit (y)
23	M	30 000	15 000	Da	3 godine	Da
40	Ž	40 000	?	Da	5 godina	Da
30	M	35 000	0	Ne	3 godine	Da
21	Ž	12 000	0	Ne	3 godine	Ne
25	M	35 000	30 000	?	5 godina	Ne

Nedostajuće vrednosti atributa

- Problem treniranja modela:primeri na kojima obučavamo model sadrže nedostajuće vrednosti
- Problem predikcije: primeri za koje želimo da predvidimo vrednost ciljne varijable sadrže nedostajuće vrednosti
- Moguće strategije:
  - Pročišćavanje podataka
  - Procena nedostajućih vrednosti
  - Modifikacija obučavajućeg algoritma da bude robustan na nedostajuće vrednosti

Pročišćavanje podataka

- Ukloniti/preskočiti opservacije sa nedostajućim vrednostima

Starost	Pol	Godišnja zarada	Trenutni dug	Poseduje nekretninu	Rok za otplatu	Odobriti kredit (y)
23	M	30 000	15 000	Da	3 godine	Da
40	Ž	40 000	?	Da	5 godina	Da
30	M	35 000	0	Ne	3 godine	Da
21	Ž	12 000	0	Ne	3 godine	Ne
25	M	35 000	30 000	?	5 godina	Ne

Starost	Pol	Godišnja zarada	Trenutni dug	Poseduje nekretninu	Rok za otplatu	Odobriti kredit (y)
23	M	30 000	15 000	Da	3 godine	Da
30	M	35 000	0	Ne	3 godine	Da
21	Ž	12 000	0	Ne	3 godine	Ne

Pročišćavanje podataka

- Ukloniti/preskočiti obeležja sa nedostajućim vrednostima

Starost	Pol	Godišnja zarada	Trenutni dug	Poseduje nekretninu	Rok za otplatu	Odobriti kredit (y)
23	M	30 000	15 000	Da	3 godine	Da
40	Ž	40 000	?	Da	5 godina	Da
30	M	35 000	?	Ne	3 godine	Da
21	Ž	12 000	0	?	3 godine	Ne
25	M	35 000	?	?	5 godina	Ne

Starost	Pol	Godišnja zarada	Rok za otplatu	Odobriti kredit (y)
23	M	30 000	3 godine	Da
40	Ž	40 000	5 godina	Da
30	M	35 000	3 godine	Da
21	Ž	12 000	3 godine	Ne
25	M	35 000	5 godina	Ne

Pročišćavanje podataka

- Prva strategija: ukloniti opservacije sa nedostajućim vrednostima
  - Primenjivati samo kada imamo mali broj takvih opservacija
- Druga strategija: ukloniti obeležje za koje mnogim opservacijama nedostaje vrednost
  - Primenjivati samo ako ima malo takvih obeležja

Pročišćavanje podataka

- Prednosti:
  - Jednostavno za razumevanje i implementaciju
  - Može se primeniti na bilo koji model
- Nedostaci:
  - Uklanjanje opservacija/obeležja može da rezultuje odbacivanjem informacija važnih za rešavanje problema
  - Nije jasno definisano kada ukloniti opservaciju a kada obeležje
  - Ove strategije ne pomažu kada treba da damo predikciju za opservaciju sa nedostajućim vrednostima

Procena nedostajućih vrednosti

- Nedostajuću vrednost zamenjujemo našom procenom te vrednosti
- Kategoričko obeležje možemo zameniti medijanom poznatih vrednosti tog obeležja
- Numeričko obeležje možemo zameniti srednjom vrednošću ili medijanom poznatih vrednosti tog obeležja

Starost	Pol	Godišnja zarada	Trenutni dug	Poseduje nekretninu	Rok za otplatu	Odobriti kredit (y)
23	M	30 000	15 000	Da	3 godine	Da
40	Ž	40 000	?	Da	5 godina	Da
30	M	35 000	0	Da	3 godine	Da
21	Ž	12 000	0	Ne	3 godine	Ne
25	M	35 000	30 000	?	5 godina	Ne

Starost	Pol	Godišnja zarada	Trenutni dug	Poseduje nekretninu	Rok za otplatu	Odobriti kredit (y)
23	M	30 000	15 000	Da	3 godine	Da
40	Ž	40 000	22.500	Da	5 godina	Da
30	M	35 000	0	Da	3 godine	Da
21	Ž	12 000	0	Ne	3 godine	Ne
25	M	35 000	30 000	Da	5 godina	Ne

Procena nedostajućih vrednosti

- Postoje i naprednije tehnike:
  - Ocena nedostajuće vrednosti regresionom metodom
  - K-NN imputacija
  - Expectation-maximization (videćemo kasnije)

Procena nedostajućih vrednosti

- Prednosti:
  - Jednostavno za razumevanje i implementaciju
  - Može da se primeni na bilo koji model
  - Može da se koristi i prilikom predikcije

- Nedostaci:
  - Može uzrokovati sistematskom greškom
  - Može rezultovati dodavanjem sistematskog odstupanja
    - Npr. da je u Washington- u ilegalno da se objavi starost aplikanta. Za sve takve aplikante ćemo starost zameniti prosečnom starošću ispitanika – ispašće da svi iz Washington-a imaju npr. 40 godina starosti

Modifikacija obučavajućeg algoritma

- Obučavajući algoritam se modifikuje kako bi rukovao eksplicitno sa nedostajučim podacima

- Prednosti:
  - Može se koristiti i prilikom obučavanja modela i prilikom predikcije
  - Rezultuje tačnijim predikcijama

- Mane:
  - Zahteva modifikaciju obučavajućeg algoritma što može biti veoma kompleksno
  - Za određene algoritme je lako, npr. za Decision trees