3. óra

Esettanulmány: Hitelbírálati probléma

Három kategóriába tesszük a hitelt igénylőket.

1. Elutasítottak
2. Szürke sáv
3. Elfogadottak

Prediktív analitika

A célja, hogy az összes többi (bemeneti, vagy független) változó segítségével (egy adott példán belül) előre tudjam jelezni a célváltozót (függő változót).

Modelt tanítom a múltbeli adatokkal, ami alapján ő majd megbecsüli a későbbi célváltozókat.

- Lehetséges több célváltozót is megjósolni
- Egy speciális eset, amikor ezek hierarchikus viszonyban állnak egymással

Adatmodell

- oszlopok, v változók
- sorok, v entitások v példák

Adattípusok és elemzési módszerek

- kategória: osztályozás
- numerikus: regresszió

De ezek nem annyira egyértelműek

Adatelőkészítés

Sokféle adatot nem egyértelmű, hogy hogyan rögzítünk és használunk.

Credit scoring vizsgálati módok

- Application credit scoring: Az ügyfél által, a hitelkérelemmel kapcsolatban megadott adatokat tudjuk róla
  - Rengeted változó
  - Ezek egy sorban jelennek meg az adatmodellben
- Viselkedési: Az ügyfél korábbi viselkedéséről van adatunk
  - Itt egy ügyfélről több sorban is találunk adatokat

Célváltozó

Nem egyértelmű, hogy ki a jó vagy rossz adós.

- Basel II alapján, Európában a rossz adós az aki akár csak egyszer is 3*30 napon meghaladóan tartozott.
- Nyilvánosan csak erről van lista
- Menet közben a bankoknak folyamatosan becsülniük kell a bedőlt hiteleket és tartalékot kell kialakítaniuk ezzel kapcsolatban.

Tanulási adat minta

- Nem feltétlenül kell az egész adathalmazra nézve reprezentatív mintát venni, de pl sokkal inkább a jelenre és jövőre nézve.
- Hosszú futamidőnél nincs sok adat, ezért viszonylag friss adatokból kell dolgozni
- Három hónapnál fiatalabb ügyletek viszont nem kerülnek be (mert ott még nem lehet tudni, hogy bedőlt-e vagy sem)

Konkrét gyakorlat

2 dimenzióban néz:

- Késedelmesek aránya
- Hitel élettartalam

Ez a függvény telítődik egy idő után és emiatt a telítődési utáni pont utáni élettartamú hiteleket érdemes behozni a tanuló adathalmazba.

Modellezés

- Egyelőre osztályozási problémaként tekintünk erre.
- Logisztikus regresszió (ez alapvetően egy osztályozási model)
  - Feltételezzük, hogy az X és a célváltozó között van összefüggés
  - Mindegyik bemeneti változóhoz számol egy együtthatót (koefficiens)

$$X = \sum \omega_i * BV_i$$

- Az értéktartomány egyelőre $-\infty/\infty$, de ezt normálnunk kell a [0,1] tartományban

$$P = \frac{e^x}{1 + e^x} \rightarrow [0, 1]$$

- Az attribútomok csak számszerű adatokat kezelnek
  - Ez nem foglalkozik azzal, hogy a változóknak mi az értelme
  - 'Dummyzás': kategória változókat át kell alakítani valamilyen numerikus változóvá