# Upotreba logističke regresije u procjeni kreditnog rizika

Nataša Šarlija

# Zašto logistička regresija?

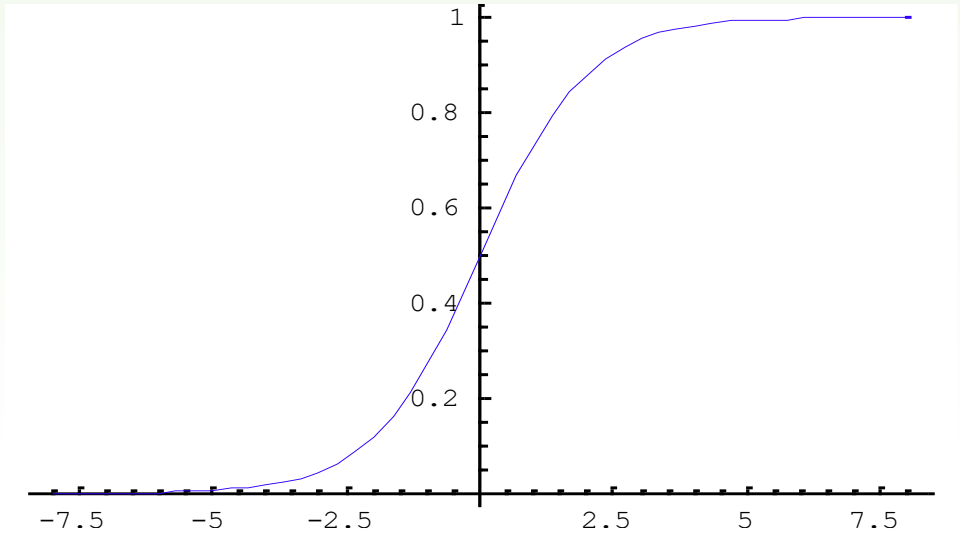
- Jedna od najčešće upotrebljavanih multivarijatnih metoda za klasifikaciju
- Hoće li klijent platiti kredit?
- Hoće li firma rasti?
- Hoće li osoba doživjeti srčani udar?
- Hoće li osoba razviti depresiju?
- Hoće li poruka biti spam?
- Hoće li firma varati pri prijavi poreza?
- Hoće li osoba plaćati alimentaciju?
- Itd.

# Kada upotrijebiti logističku regresiju?

- Kada želimo klasificirati jedinice u 2 (ili više) grupa
- Popularna metoda za klasifikaciju jedinica za dani skup nezavisnih varijabli
- Procjenjuje vjerojatnost da jedinica pripadne određenoj grupi
- Regresijski model gdje je zavisna varijabla dummy/indikator/binarna varijabla

# Logistička regresija

- Neka je $X_1 \ldots X_k$ skup nezavisnih varijabli
- Nelinearna funkcija: $$\frac{1}{1 + e^{-(b_0 + b_1 X_1 + \quad + b_k X_k)}}$$




- Model logističke regresije koristi tu funkciju za procjenu vjerojatnosti da će jedinica pripasti grupi 1

# Logistička regresija

- Ako je p vjerojatnost pripadanja grupi 1, procjenjujemo sljedeći model:

$$\frac{1}{1 + e^{-(b_0 + b_1 X_1 + \quad + b_k X_k)}}$$

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X_1 + \quad + b_k X_k$$

- Postoji linearna veza između log odds (logaritam šanse) i nezavisnih varijabli

# Zašto upotrijebiti logističku regresiju?

- Daje dobre rezultate
- Jednostavna je za primjenu
- Rezultati su razumljivi i nematematičarima
- Pretpostavke za primjenu nije teško zadovoljiti

# Pretpostavke logističke regresije

## ŠTO NE TREBA ZADOVOLJITI

- Ne treba linearna veza između zavisne i nezavisnih varijabli
- Nezavisne varijable i reziduali ne trebaju biti normalni
- Homoskedastičnost nije potrebna
- Nezavisne varijable ne trebaju biti numeričke (*metric*)

## ŠTO TREBA ZADOVOLJITI

- Zavisna varijabla 0/1; 1 – željeni ishod
- Paziti na overfitting i underfitting
- Podaci za svaku jedinicu prikupljeni neovisno
- Model ne smije sadržavati multikolinearnost
- Veliki uzorak (barem 10, a po nekima 30 jedinica po nezavisnoj varijabli)

# Multikolinearnost

- Pravi probleme pri razvoju multivarijatnog logističkog modela
- Povezano sa selekcijom varijabli koje će ući u model – znanje teorije o području koje se istražuje
- Ozbiljno narušava model i tumačenje modela:
  - Procijenjeni regresijski koeficijent jedne varijable ovisi o ostalim varijablama u modelu
  - Preciznost regresijskih koeficijenata se smanjuje uključivanjem novih i novih varijabli
  - Varijabla koja samostalno nije u relaciji sa zavisnom može postati značajna u modelu i obrnuto

# Kako upotrijebiti logističku regresiju?

- Odabrati varijable i postaviti veze između varijabli – potrebno znanje teorije o području koje se istražuje
- Prikupiti odgovarajući skup podataka
- Obratiti pažnju na pretpostavke modela
- Upotrijebiti statistički alat – svi imaju LR, na različite načine uključuju kategorijalne varijable!
- Testirati model

# Procjena defaulta

- Default retail klijenta (zavisna varijabla):
  - 1: ‘loš’ klijent
  - 0: ‘dobar’ klijent
- nezavisne varijable: dob, radni staž, bračno stanje, roba koju kupuje ...
- Cilj: procijeniti default klijenta i otkriti koje varijable imaju utjecaj na default
- glm(losi~bracno, family=binomial(link="logit"), data=retail_odjel za matematiku.csv)

# Interpretacija – numeričke varijable

- Kvalitativno promatramo regresijske koeficijente:
  - b>0: X pozitivno koreliran s pripadnosti grupi 1; s porastom X, povećava se logaritam šanse
  - b<0: X negativno koreliran s pripadnosti grupi 1; s porastom X, smanjuje se logaritam šanse
- dob klijenta
- b= -0.042707
- Postoji negativna veza između dobi i defauta
- S povećanjem dobi smanjuje se vjerojatnost za default

# Interpretacija – numeričke varijable

- Interpretiranje log odds ratio:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1 X_1 + \quad + b_k X_k$$

$$\frac{p}{1-p} = e^{b_0 + b_1 X_1 + \quad + b_k X_k}$$

$$e^{b_1 \Delta}$$

$$\Delta = 1$$

$$e^{b_1}$$

- dob
- b= -0.042707 , $e^b$=0.9581
- Ako se dob poveća za 1, logaritam (kvocijenta) šansi da klijent ode u default smanjuje se za 0.042707
- S jediničnim povećanjem dobi smanjuju se šanse da klijent bude loš (u odnosu na to da bude dobar) za faktor 0.9581

# Interpretacija – kategorijalne varijable

- Kod kategorijalnih varijabli, interpretacija se radi u odnosu na baznu kategoriju
  - b>0: veći je logaritam šanse za tu kategoriju u odnosu na baznu
  - b<0: manji je logaritam šanse za tu kategoriju u odnosu na baznu
- U R-u je, po defaultu, bazna kategorija prva kategorija
- bracno = 1 – samac; 2 – u braku; 3 – udovac/ica; 4 – razvedn/a – BAZA
- Npr. b za samac = 0.585
- Logaritam šansi da bude loš klijent koji je samac u odnosu na razvedenog povećava se za 0.585
- Šanse da bude loš klijent koji je samac povećavaju se u odnosu na klijenta koji je razveden za faktor $e^{0.585}=1.7948$

# Zadatak

- Za projekt treba napraviti 3 scoring modela. Do 19.4. je dovoljno napraviti 1 scoring model na kojem ćemo napraviti testiranje. Isto testiranje ćete ponoviti i za ostale modele.
- Ideje za različite modele: različite kombinacije varijabli, model gdje su sve varijable kategorizirane, model gdje imamo kombinaciju numeričkih i kategorijalnih, pokušati iskoristiti neku selekcijsku proceduru, umjesto originalnih vrijednosti staviti woe vrijednosti, a takve varijable tretirati kao numeričke
- Pri modeliranju obratiti pažnju na multikolinearnost
- Za svaki model napraviti interpretaciju regresijskih koeficijenata – pri tome je VAŽNO da model bude logičan; npr. regresijski koeficijent za dob treba biti negativan, najmanje rizično bračno stanje je ‘u braku’ itd.