Aalto-yliopisto
Perustieteiden
korkeakoulu

Poistuman tutkiminen logistisella regressiolla (valmiin työn esittely)

Lauri Suoknuuti

10.08.2020

Ohjaaja: FM *Markus Linnakaari*
Valvoja: Apul.prof. *Pauliina Ilmonen*

Systeemianalyysin
laboratorio

Tausta

Asiakkaiden poistuma engl. *customer churn*

- Asiakkaan toimesta tapahtuva asioinnin lopettaminen yrityksen kanssa
- Uusien asiakkaiden hankkiminen tyypillisesti kalliimpaa kuin vanhojen asiakkaiden pitäminen
- Asiakaspoistuman ennustaminen laajalti tutkittu ongelma
- Voidaan määritellä binäärisenä luokitteluongelmana

Menetelmät
Logistinen regressio

- Regressioanalyysin tavoitteena on selvittää vasteen ja selittävien muuttujien välinen riippuvuus
- Logistinen regressio kuuluu yleistettyihin lineaarisiin malleihin
- Logistisessa regressiomallissa vastemuuttuja on tyypillisesti binäärinen
- Malli antaa vastemuuttujan odotusarvon, joka tulkitaan todennäköisyytenä sille, että vastemuuttuja saa arvon 1

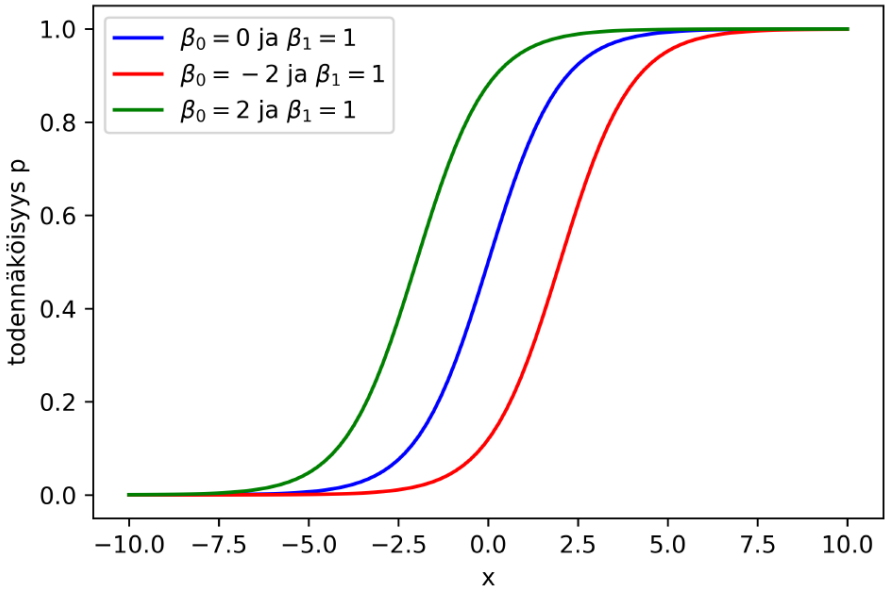
Menetelmät
Logistinen regressio

$$\pi(\mathbf{x}_i) = \frac{1}{1 + \exp(-(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))} \qquad 1)$$

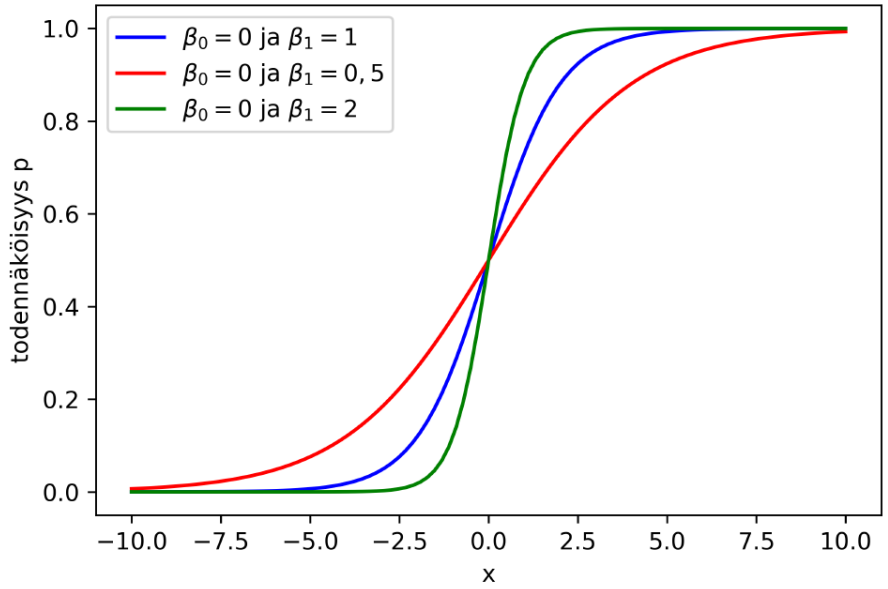
$$\pi(\mathbf{x}_i) = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_i))} \qquad 2)$$

1) Malli, jossa on p selittävää muuttujaa
2) Malli, jossa on yksi selittävä muuttuja

Menetelmät
Logistinen regressio



Kuva 1: Logistinen funktio kuvattuna parametrin β_0 eri arvoilla.

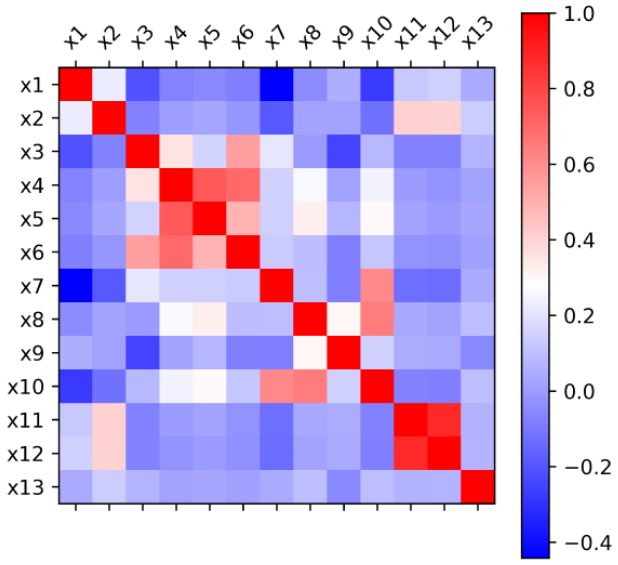


Kuva 2: Logistinen funktio kuvattuna parametrin β_1 eri arvoilla.

Tulokset
Aineisto

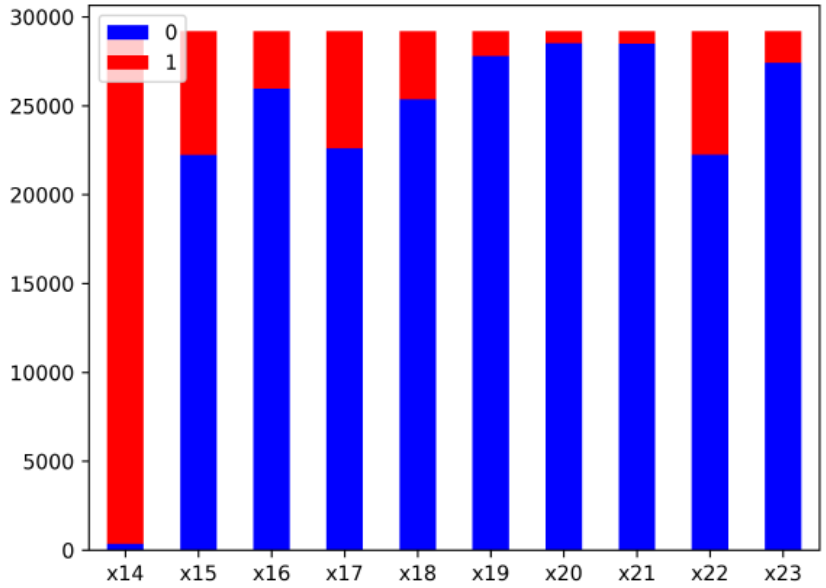
- Fennian autoliikkeille tarjottavien kampanjatarjousten asiakasdataa vuosilta 2017-2019
- 29 selittävää muuttujaa, jotka sisältävät tietoja vakuutuksenottajasta ja vakuutetusta ajoneuvosta
- Vasteena kaksiarvoinen muuttuja, joka kuvaa asiakkaan poistumaa
  - Vaste saa arvon 1, jos asiakas irtisanoo vakuutuksensa vuoden kuluessa
- Selittäviin muuttujiin viitataan nimillä x1, x2, …, x29

Tulokset
Aineisto



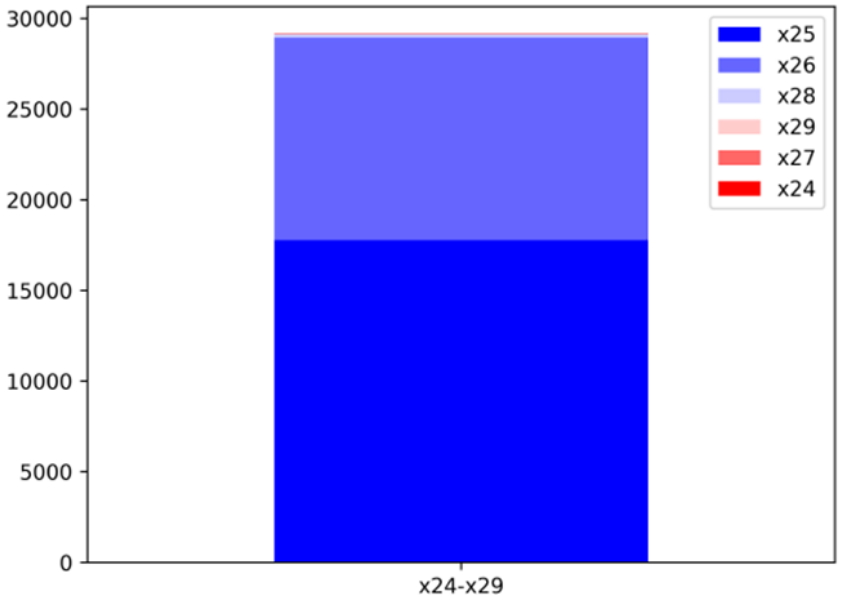
Kuva 3: Mallissa käytettyjen jatkuvien ja kokonaislukuarvoisten muuttujien pareittaiset korrelaatiot.

Tulokset
Aineisto



Kuva 4: Mallissa käytettyjen kaksiarvoisten muuttujien jakaumat.

Tulokset
Aineisto



Kuva 5: Mallissa käytetyn kategorisen muuttujan jakauma.

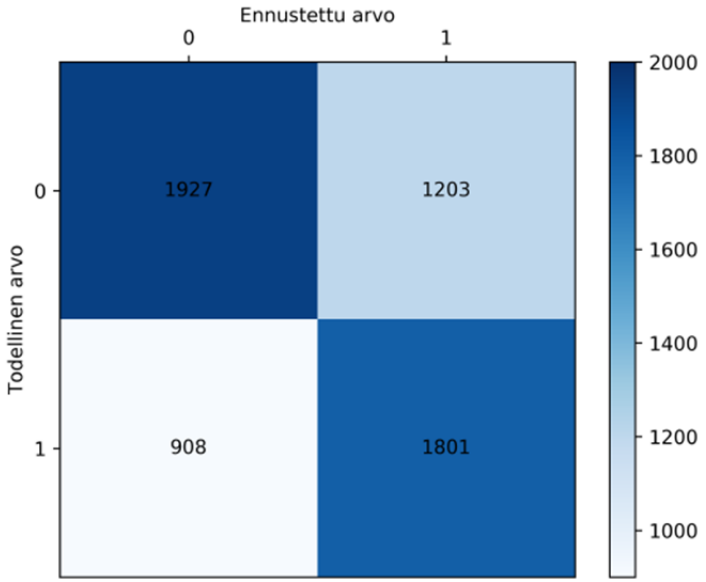
Tulokset
Malli

- Mallissa käytetään edellä esiteltyjä muuttujia
- Data jaetaan satunnaisesti koulutukseen käytettyyn osaan (80%) ja validaatioon käytettyyn osaan (20%)
- Koulutusdataan sovitetaan logistinen regressiomalli käyttäen scikit-learn ohjelmointikirjaston oletusparametreja
- Mallin ennustekykyä arvioidaan datan validaatio-osan avulla

Tulokset
Regressiokertoimet

- Regressiokertoimet ovat keskenään vertailukelpoisia, johtuen datan skaalaamisesta
- Itseisarvoltaan suurimmat regressiokertoimet vaikuttavat luokitteluun eniten
- Suurin vaikutus luokitteluun on muuttujilla x2: -0,3509 ja x8: -0,3182
  - Kertoimien arvot negatiiviset => näiden muuttujien kasvu vähentää poistuman todennäköisyyttä

Tulokset
Luokittelutaulukko



Kuva 6: Mallin perusteella muodostettu luokittelutaulukko.

Accuracy: 0,638
Precision: 0,600
Recall: 0,665

Tulokset
Yhteenveto

- Työn luokitteluongelma on tunnetusti haastava
- Mallilla saavutettua ennusteiden tarkkuutta voidaan pitää kohtalaisena
- Mallia voitaisiin yrittää parantaa muokkaamalla logistisen regression parametreja
- Voitaisiin myös tutkia muiden koneoppimismenetelmien soveltuvuutta työn asiakaspoistuman ennustamiseen