



Aalto-yliopisto
Perustieteiden
korkeakoulu

Luottoriskin arvioiminen logistisen regression avulla (valmiin työn esittely)

Olivia Antikainen

21.1.2022

Ohjaaja: Prof. *Ahti Salo*

Valvoja: Prof. *Ahti Salo*

Työn saa tallentaa ja julkistaa Aalto-yliopiston avoimilla verkkosivuilla. Muilta osin kaikki oikeudet pidätetään.

Tausta

- Luottoriski tarkoittaa riskiä, joka liitetään lainanottajan maksukyvyttömyyteen
 - Mallit ennustavat maksukyvyttömyyttä todennäköisyyksinä
- Luottoriskien arviointi on yksi lainanantajien tärkeimmistä työkaluista
 - Haettujen lainahakemusten riskien arviointi → parantaa tehokkuutta ja tuottavuutta
- Luottoriskien arvioiminen on kehittynyt
 - Enemmän dataa ja uusia menetelmiä → edut ja haasteet
- Luottokorttien luottoriskin arvioiminen
 - Luottokorttien käyttöä seurataan jatkuvasti, luottokorttiyhtiöt pystyvät tarvittaessa puuttumaan → tuottavuus kasvaa

Työn tavoite

- Arvioidaan luottokorttiasiakkaiden luottoriskiä
 - Maksukyvyttömyyden todennäköisyys
- Kehitetään malli logistisen regression avulla hyödyntäen dataa luottokorttiyhtiön asiakkaiden luotto- ja muista lähtötiedoista
- Selvitetään mitkä tekijät vaikuttavat riskin todennäköisyyden kasvuun ja miten

Menetelmät

- Logistinen regressio
 - Regressioanalyysin muoto, tutkitaan selitettävän muuttujan riippuvuutta selittäviin muuttujiin
 - Selitettävä muuttuja on binäärinen
 - Kuuluu yleistettyihin lineaarisiin malleihin:

$$g(E(y_i)) = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j,$$

josta saadaan logistinen regressiomalli:

$$\ln\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j \quad \longrightarrow \quad p(x_i) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^p x_{ij}\beta_j)}}$$

, jossa $p(x_i) = P(y_i=1 \mid x_i)$

Menetelmät

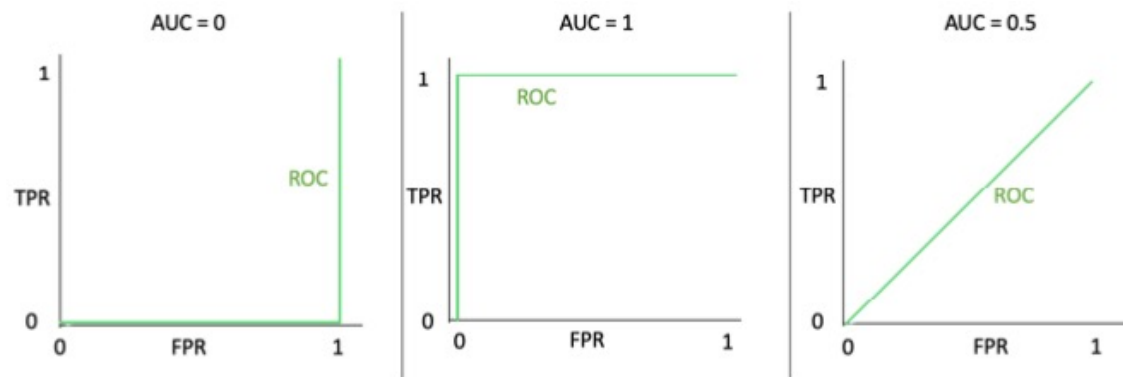
- Luottoriski binäärisenä luokitteluongelmana
- Selitettävänä muuttujana maksukyvyttömyys
- Selittävinä muuttujina ovat maksuhistoria (määrä ja aika), luottokortin käyttötiedot, luoton määrä sekä kortin haltijan ikä, sukupuoli, koulutus ja siviilisääty

Menetelmät

- Mallin arviointi:
 - Luokittelutaulukko (Accuracy, Precision, Sensitivity)

	Ennustettu 0	Ennustettu 1
Todellinen 0	TN	FP
Todellinen 1	FN	TP

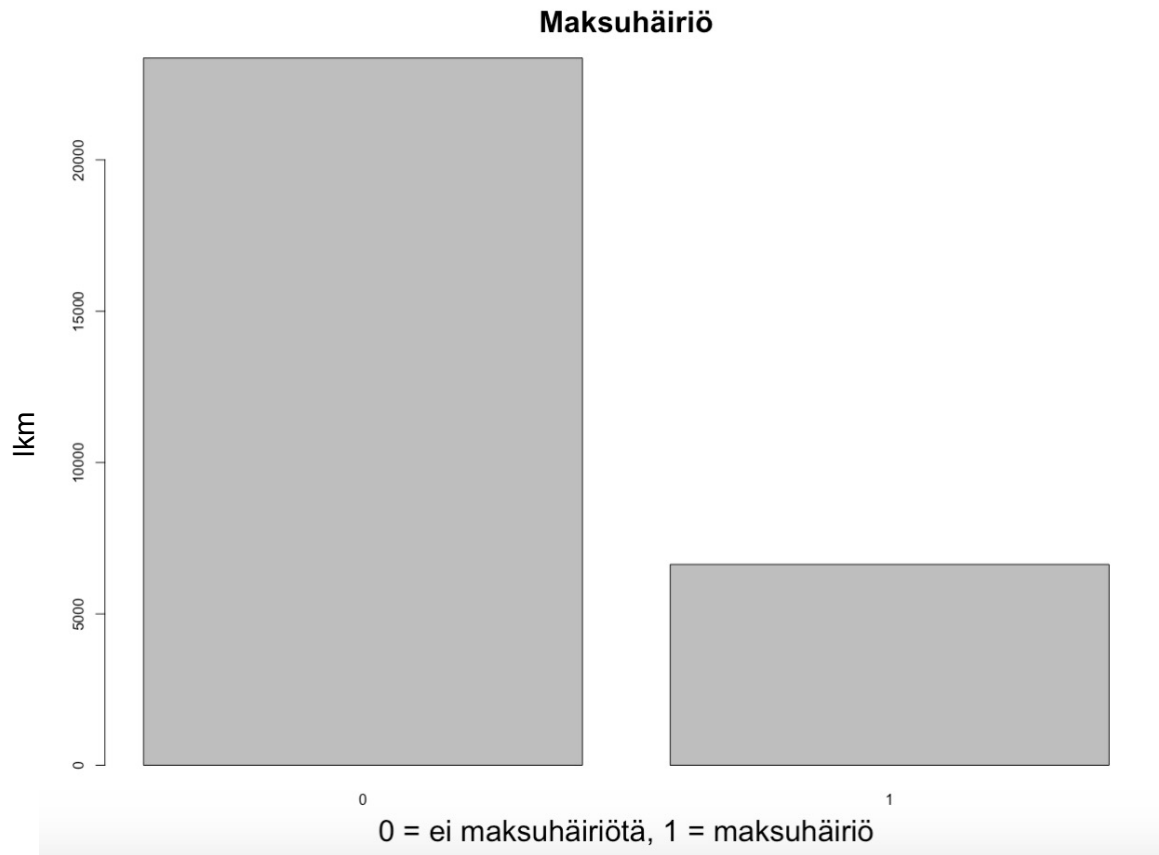
- ROC-käyrä (receiver operating characteristic) → AUC (area under the curve)



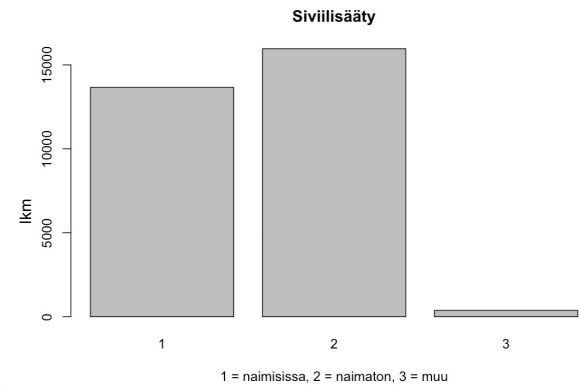
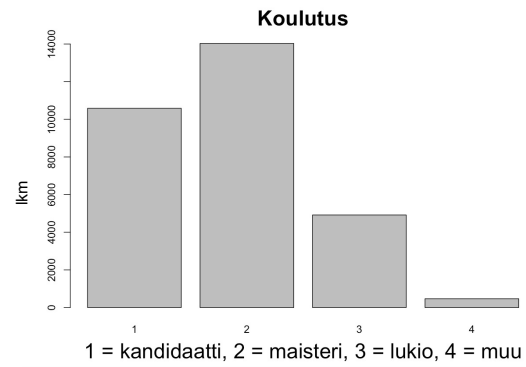
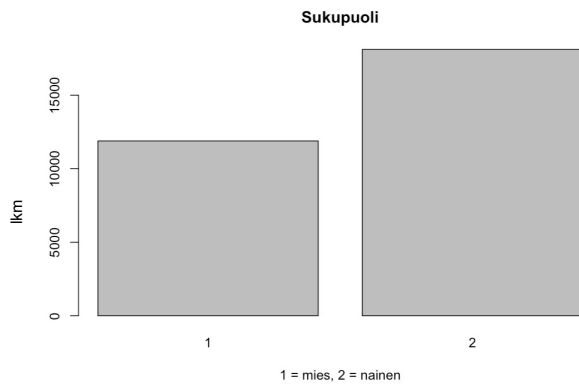
Aineisto

- Taiwanilaisen luottokorttiyhtiön asiakasdataa huh-
tikuusta syyskuuhun vuodelta 2005
- 24 muuttujaa, jotka ovat muodoltaan jatkuvia, ordi-
naalisia, kategorisia sekä binäärisiä

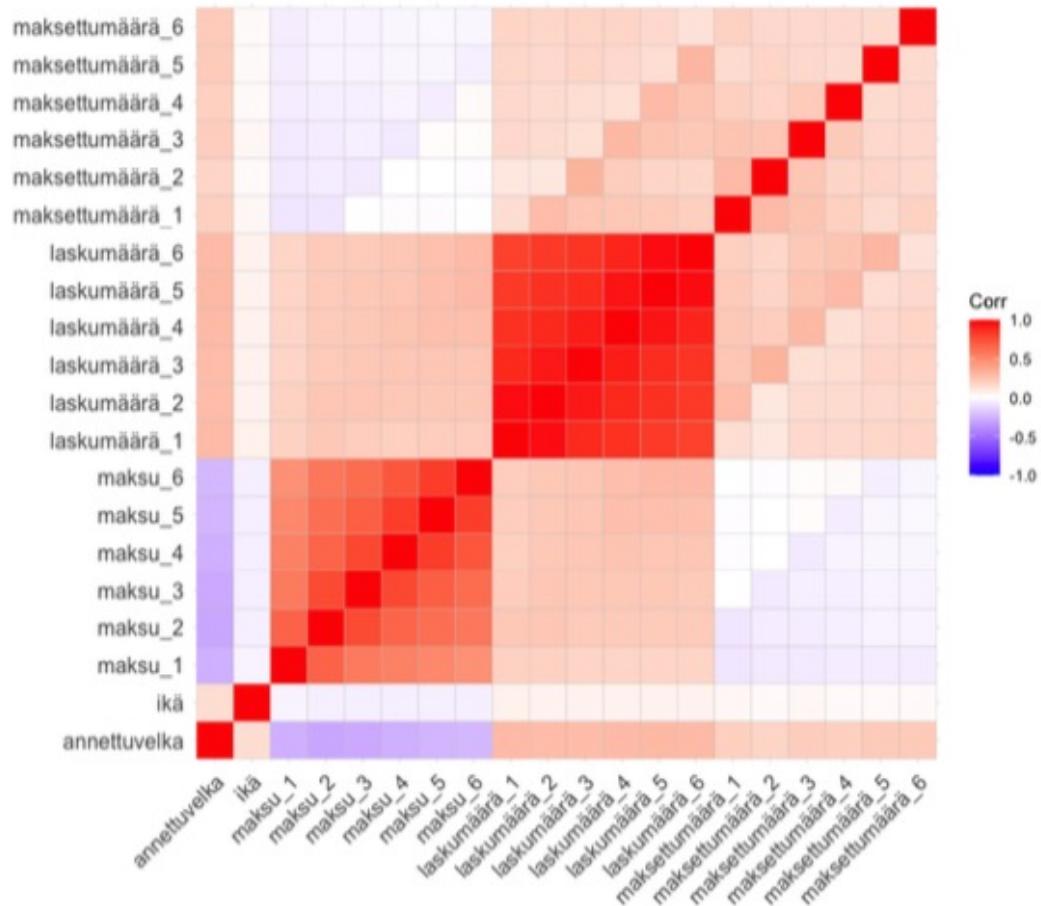
Aineisto



Aineisto



Aineisto



Malli

- Ennen mallin sovittamista tehdään muuttujien valinta ja jatkuvien muuttujien skaalaus
- Malli toteutetaan RStudio-ohjelmiston glm-funktion avulla.
- Data-aineisto jaetaan koulutus- ja testiosaan (70% ja 30%)
 - Koulutusosaa käytetään glm-funktion data-aineistona ja testiosaa mallin arvioinnissa

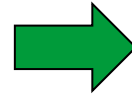
Tulokset

- Itseisarvoltaan suurin arvo vaikuttaa maksukyvyttömyyden todennäköisyyteen eniten, itseisarvoltaan pienin arvo vaikuttaa vähiten.
- Etumerkki kertoo vaikutuksen suunnasta → negatiivinen arvo pienentää maksukyvyttömyyden todennäköisyyttä, positiivinen arvo kasvattaa

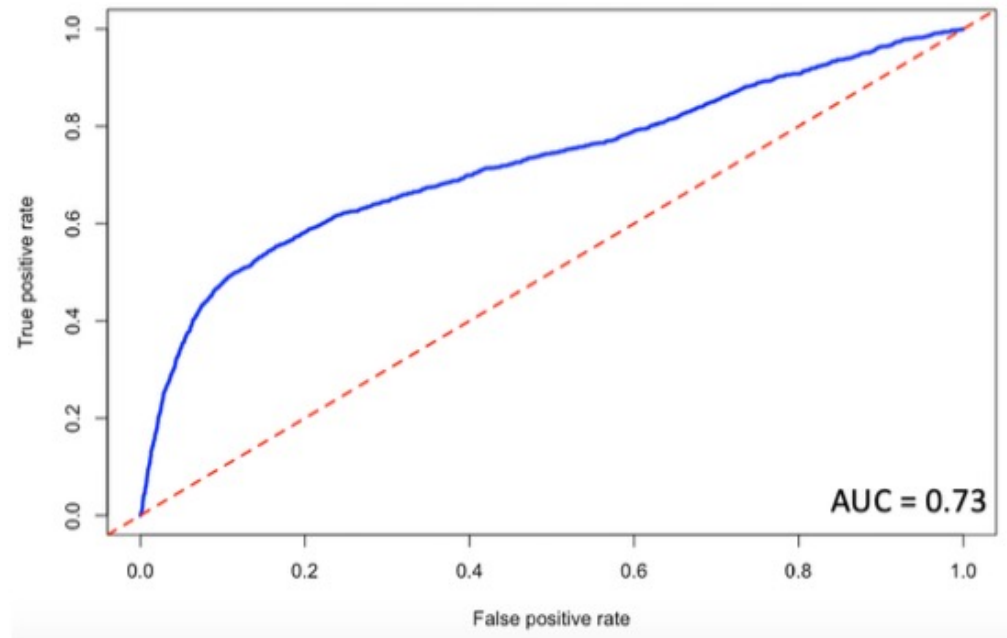
	Regressiokerroin
vakio	-2,52940
annettuvelka	-0,08119
ikä	0,04291
maksu ₁	0,63349
maksu ₂	0,10792
maksu ₃	0,07034
maksu ₄	0,06650
maksu ₅	0,03350
maksu ₆	0,02681
laskumäärä ₁	-0,10429
maksettumäärä ₁	-0,19856
maksettumäärä ₂	-0,16711
maksettumäärä ₃	-0,07339
maksettumäärä ₄	-0,06797
maksettumäärä ₅	-0,02504
maksettumäärä ₆	-0,07747
sukupuoli ₁	0,12928
avioliitto ₁	0,19736
koulutus ₁	0,99185
koulutus ₂	0,90936
koulutus ₃	0,88894

Tulokset

	Ennustettu 0	Ennustettu 1
Todellinen 0	6798	198
Todellinen 1	1501	503



Tarkkuus = 0,811
Sensitiivisyys = 0,251
Positiivinen ennustearvo = 0,718



Yhteenveto

- Koulutuksella on maksukyvyttömyyden todennäköisyyttä kasvattava vaikutus ja asiakkaan aikaisemmin maksettujen laskujen suuruudella pienentävä vaikutus
- Mallin ennustekyky on kohtalainen
- Mallin suorituskyky voisi parantua aineiston epätasapainon korjaamisella sekä data-aineiston ja muuttujien suuremmalla määrällä
- Luottokorttiyhtiöt ovat usein kansainvälisiä → mallia ja sen tuloksia voitaisiin hyödyntää myös suomalaisten luottokorttiyhtiön asiakkaiden luottoriskin arviointiin.

Tärkeimmät tietolähteet ja data-aineisto

- Data-aineisto: University of California, Irvine (UCI) / default of credit card clients
- Tietolähteet:
 - World Bank (2019) 'Credit Scoring Approaches Guidelines', The World Bank Group.
 - Hilbe, J. M. (2009) 'Logistic Regression Models' , CRC Press.
 - McCullagh, P., & Nelder, J. A. (1989) 'Generalized linear models II', Chapman & Hall/CRC.
 - Hosmer, D.W., & Lemeshow, S., & Sturdivant, R. X. (2013) 'Applied logistic regression', Wiley.