Luottoriskien arviointi logistisella regressiolla

Olivia Antikainen

Perustieteiden korkeakoulu

Kandidaatintyö
Espoo 28.2.2022

Vastuuopettaja

Prof. Ahti Salo

Työn ohjaaja

Prof. Ahti Salo

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto-yliopisto
Perustieteiden korkeakoulu

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Tekniikan kandidaatintyön tiivistelmä

Tekijä Olivia Antikainen

Työn nimi Luottoriskien arviointi logistisella regressiolla

Koulutusohjelma Teknillinen fysiikka ja matematiikka

Pääaine Matematiikka ja systeemitieteet **Pääaineen koodi** SCI3029

Vastuuopettaja Prof. Ahti Salo

Työn ohjaaja Prof. Ahti Salo

Päivämäärä 28.2.2022 **Sivumäärä** 25 **Kieli** Suomi

Tiivistelmä

Luottoriski tarkoittaa lainanottajan maksukyvyttömyyttä eli sitä, että lainanottaja ei maksa otettua lainaa takaisin. Luottoriskiä arvioivat mallit ennustavat lainanottajan maksukyvyttömyyden todennäköisyyttä. Tässä kandidaatintyössä arvioidaan luottokorttiyhtiön asiakkaiden luottoriskiä logistisella regressiolla. Asiakkaan maksukyvyttömyyden todennäköisyyden ennustamiseksi rakennetaan logistinen regressiomalli sen määrittämiseksi, mistä tekijöistä asiakkaan maksukyvyttömyyden todennäköisyys riippuu. Rakennettavan mallin data-aineistona on taiwanilaisen luottokorttiyhtiön asiakkaiden luottotietoja ja muita lähtötietoja. Tulosten valossa asiakkaan maksukyvyttömyyden todennäköisyys kasvaa eniten asiakkaan koulutustason myötä, kun taas aikaisemmin maksettujen laskujen suuruus pienentää sitä. Työn lopussa mallin tarkkuutta arvioidaan erilaisilla menetelmillä. Näiden valossa mallin ennustekyvyn havaitaan olevan kohtalainen.

Avainsanat luottoriski, maksukyvyttömyys, logistinen regressio, luottokortti, mallin arviointi, luokitteluongelma

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the bachelor's thesis

Author Olivia Antikainen

Title Evaluating credit risk using logistic regression

Degree programme Engineering Physics and Mathematics

Major Mathematics and Systems Sciences **Code of major** SCI3029

Teacher in charge Prof. Ahti Salo

Advisor Prof. Ahti Salo

Date 28.2.2022 **Number of pages** 25 **Language** Finnish

Abstract

A credit risk is risk of default on a debt, which means that the borrower is not able to repay a loan or make the required payments. Credit scoring is a statistical method for predicting the probability of a default. In this thesis, the goal is to evaluate the credit risk of credit card clients with logistic regression. Logistic regression is used to build a model that predicts the probability that the customer will default. From the model we can examine which factors affect the probability of default. The model is estimated using data from Taiwanese credit card clients past payments and other information. Based on the results, education increases the probability of default the most and the quantity of past payments decreases the probability of default the most. Finally, the performance of the model is evaluated with different methods, which indicate the model has decent diagnostic accuracy.

Keywords credit risk, default, logistic regression, credit card, model evaluation, classification problem

Sisällys

1 Johdanto

Tässä kandidaatintyössä tutkitaan luottokorttiyhtiön asiakkaiden luottoriskiä, ja siihen vaikuttavia tekijöitä. Luottokorttiyhtiön luottoriskillä tarkoitetaan sitä, että asiakas ei kykene maksamaan luottokortin käyttöön liittyviä laskuja ajallaan. Luottoriskiä arvioivat mallit ennustavat asiakkaiden maksukyvyttömyyden todennäköisyyttä. Siten lainanantajat voivat hyödyntää näitä malleja lainahakemusten riskien arvioinnissa. Luottoriskien arvioinnista on tullut lainanantajien tärkeimpiä työkaluja, sillä onnistunut riskien arviointi parantaa lainanantajien tehokkuutta ja tuottavuutta (Weber, 2012). Luottoriskien arvionnin laiminlyönti voi johtaa pahimmassa tapauksessa lainanantajien konkursseihin.

Luottoriskin arviointiin ja mallintamiseen liittyy monia haasteita. Tekoälyä hyödyntävät kehittyneet menetelmät kykenevät ennustamaan maksukyvyttömyyttä tehokkaasti, mutta ne saattavat tahottamasti tuottaa harhaisia malleja (Commission et al., 2018). Tällä tarkoitetaan sitä, että mallin algoritmi saattaa olla puolueellinen, eli se saattaa tahattomasti syrjiä tiettyjä ihmisryhmiä. Harhaisuus voi johtua käytettävästä menetelmestä, tai käytettävän data-aineiston muuttujista. Oikeiden muuttujien valinta on näin ollen yksi merkittävimmistä haasteista luottoriskien arvioinnissa. Yleinen ongelma luottoriskien arvioinnissa on myös data-aineiston epätasapaino (Li et al., 2019), joka johtuu maksukyvyttömien asiakkaiden pienestä osuudesta koko datassa. Aineiston epätasapaino aiheuttaa ongelmia mallin muodostamisessa.

Työn tavoitteena on arvioida luottokorttiasiakkaiden luottoriskiä. Luottoriskiä arvioidaan kehittämällä asiakkaan maksukyvyttömyyden todennäköisyyttä ennustava logistinen regressiomalli. Logistinen regressio on suosittu työkalu luottoriskin arviointiin silloin, kun asiakkaiden luottoriskiä lähestytään binäärisenä luokitteluongelmana (Thomas, 2009). Tämä tarkoittaa sitä, että asiakkaat pyritään luokittelemaan kahteen luokkaan maksukyvyttömyyden perusteella. Mallin parametrit estimoidaan käyttäen data-aineistona taiwanilaisten luottokorttiyhtiön asiakkaiden luottotietoja ja muita lähtötietoja. Logistisen regression avulla selvitetään, mitkä tekijät kasvattavat asiakkaan maksukyvyttömyys riskiä ja kuinka paljon. Lisäksi rakennetun mallin toimivuutta eli sen erottelukykyä arvioidaan.

Työn rakenne on seuraava. Ensimmäisenä käsitellään myös aikaisempaa tutkimusta luottoriskien arvioinnista ja mallintamisesta. Kirjallisuuskatsauksessa tutustutaan luottoriskien arvioinnin ja mallintamisen haasteisiin sekä mahdollisuuksiin, etenkin kasvavan datan ja uusien menetelmien näkökulmasta. Tämän jälkeen perehdytään logistiseen regressioon, jota käytetään luottoriskien mallintamiseen. Työssä käsitellään myös mallin arvioinnin työkaluja ja toimintaa. Lopulta työssä kehitetään maksukyvyttömyyttä ennustava malli, jonka tuloksia sekä ennustekykyä arvioidaan.

2 Kirjallisuuskatsaus

2.1 Luottoriskien arviointi

Luottoriskillä tarkoitetaan riskiä siitä, että lainanottaja ei maksa otettua lainaa takaisin. Luottoriskiä voidaan arvioida yksityishenkilöille luottopisteytyksen (credit sco-

ring) avulla ja yrityksille käyttämällä luottoluokitusta (credit rating) (Group, 2019). Luottopisteytyksen sekä luottoluokituksen avulla voidaan tilastollisen analyysin avulla tutkia lainanhakijan luottokelpoisuutta, joka vaikuttaa lainan hyväksymiseen ja lainan ehtoihin. Tässä työssä keskitytään luottopisteytykseen. Luottopisteytyksellä voidaan arvioida millä todennäköisyydellä kuluttaja on kykenemätön maksamaan lainaa takaisin.

Luottoriskin arvioiminen on yksi tärkeimmistä työkaluista lainanantajille, kuten pankeille ja luottokorttiyhtiöille, jotka tarjoavat lainoja yksityishenkilöille. Antolainaus on keskeinen osa pankin liiketoimintaa, joten lainojen luottojärjestelyt ovat tärkeimpiä toimia pankin menestyksen kannalta (Weber, 2012). Monet rahoituslaitokset ovat kärsineet rahoitusongelmista tai pahimmassa tapauksessa joutuneet konkurssiin puutteellisten luottojärjestelyiden takia. Luottojärjestelyiden tarkoituksena on estää tulevia tappioita, jotka johtuvat asiakkaiden maksukyvyttömyydestä (Kvesić ja Dukić, 2012). Usein menestyvät rahoituslaitokset ovat niitä, jotka pystyvät arvioimaan asiakkaiden luottoriskiä parhaiten. Luottopisteytystä käyttämällä rahoituslaitokset voivat arvioida haettujen ja myönnettyjen lainojen riskiä, ja näin muokata lainojen ehtoja riskien mukaan tai evätä laina. Luottoriskin arviointi luottopisteytyksen avulla parantaa näin ollen rahoituslaitosten tehokkuutta lainojen myöntämisessä sekä parantaa rahoituslaitosten tuottavuutta (Group, 2019). Luottopisteytystä voidaan käyttää myös lainan myöntämisen jälkeen esimerkiksi tulevaisuuden maksuhäiriöiden ja luottoriskien arvioimiseen (Group, 2019).

Tässä työssä keskitytään luottoriskin arvioimiseen luottokorttimarkkinoilla. Luottokortit ovat kehittyviä luottolimiittejä (line of credit), jonka takia niiden riskien hallinta eroaa suuresti muista pankin myöntämistä lainoista (Butaru et al., 2016). Luottokorttien käyttöä voidaan monitoroida jatkuvasti, sillä luottokorttiyhtiöt saavat helposti kerättyä data-aineistoa luottokorttien toiminnasta ja niiden haltijoista. Tarvittaessa luottokorttiyhtiöt voivat puuttua luottokorttien käyttöön, esimerkiksi jäädyttämällä kortit tai muuttamalla korttien ehtoja, jos data-aineiston pohjalta tehdyn luottoriskin arvioinnin jälkeen kortin haltijan maksukyvyttömyys vaikuttaa todennäköiseltä (Butaru et al., 2016). Näin voidaan välttyä ylimääräisiltä tappioilta, esimerkiksi vähentämällä todennäköisesti maksukyvyttömien asiakkaiden määrää. Lisäksi luottokorttiyhtiöt voivat hyödyntää luottoriskien arvioimista asettaessaan uusien asiakkaiden korttien luottorajoja sekä hyväksyessään luottokorttihakemuksia.

Luottoriskin arvioiminen on kehittynyt digitalisaation myötä, koska saatavilla on enemmän data-aineistoja luottoriskien arvioimiseksi (Group, 2019). Laajempien data-aineistojen ja uusien menetelmien avulla, esimerkiksi tekoälyä hyödyntämällä, on mahdollista arvioida kuluttajien luottokelpoisuutta täsmällisemmin. Kasvavan datan määrän ja erilaisten menetelmien yhdistäminen on mahdollistanut lainojen myöntämisen ihmisille, jotka eivät aikaisemmin olleet luokiteltuja lainojen saamiselle. Tämä kasvattaa taloudellista tasa-arvoa. Kaikilla kuluttajilla ei ole luottotietoja tai ne ovat vähäiset. Perinteiset lainaehdot perustuvat menneisiin luottotietoihin, joten rahoituslaitokset, jotka hyödyntävät monipuolisempaa dataa, kykenevät tarjoamaan tasapuolisemmin lainoja useammille ihmisille (Aitken, 2017). Monipuoliseen dataan sisältyy esimerkiksi kuluttajan matkapuhelimen käytön tiedot ja koulutus. Esimerkiksi tämän työn logistisessa regressiossa voidaan halutessa valita tarkasteltavaksi laaja

valikoima muuttujia, jotka voivat parantaa mallin tarkkuutta ja tasapuolisuutta.

Vaikka laajemmat data-aineistot ja uudet menetelmät edistävät luottoriskien arvioimisessa tasapuolisuutta lainan saamisessa, saattavat ne myös aiheuttaa tietämättään puolueellisuutta (Commission et al., 2018). Tällä tarkoitetaan sitä, että mallin algoritmi saattaa olla puolueellinen, ja näin ollen malli saattaa tietämättä syrjiä joitain ihmisryhmiä. Lisäksi data voi sisältää muuttujia, jotka arvioivat henkilön etnisyyttä, mikä saattaa mallissa aiheuttaa syrjintää. Oikeiden muuttujien valinta on yksi keskeisistä haasteista luottoriskien arvioimisessa. Muita haasteita luottoriskin arvioimisessa esiintyy esimerkiksi datan keruun ja käytön yhteydessä, jotka saattavat aiheuttaa tietoturvariskejä.

2.2 Mallintaminen

Luottoriskimallit ennustavat todennäköisyyttä, jolla lainan hakija tai jo lainan saanut henkilö ei maksa lainaa takaisin. Malleja rakennetaan asiakasdatan avulla, esimerkiksi käyttämällä asiakkaiden luottokorttien käyttötietoja. Asiakasdataa mallintamalla pyritään ennustamaan tulevien ja jo olemassa olevien asiakkaiden luottoriskiä. Yksityishenkilön luottoriskin arvioimiseen voidaan käyttää monia eri menetelmiä, jotka kehittyvät perinteisistä tilastollisista menetelmistä kohti kehittyneempiä tekoälyä hyödyntäviä tilastollisia menetelmiä (Group, 2019). Perinteisiä menetelmiä ovat esimerkiksi lineaarinen regressio, diskriminanttianalyysi ja logistinen regressio. Kehittyneisiin menetelmiin kuuluvat tekoälyä hyödyntävät menetelmät, kuten päätöspuut ja neuroverkostot, joiden avulla voidaan mallintaa erittäin monimutkaisiakin yhtälöitä (Abdou ja Pointon, 2011). Perinteisten ja kehittyneiden menetelmien avulla saatavilla olevaa dataa käytetään muodostamaan luottoriskimalleja, joita voidaan käyttää luottojärjestelmissä. Mallien avulla voidaan luokitella yksilöt sen mukaan, tuleeko heille maksuhäiriö vai ei, eli pystyvätkö he maksamaan lainan sopimuksen mukaiset laskut ajallaan.

Suosituin menetelmä yksityishenkilöiden luottoriskin mallintamiseen on logistinen regressio (Thomas, 2009). Logistista regressiota on helppo tulkita, validoida, kalibroida ja kehittää. Menetelmässä käytettyjä parametreja on helppo tulkita, eli eri parametrien merkitys mallin tekemään luokitteluun on helppo nähdä. Näin ollen logistisia regressiotuloksia voidaan hyödyntää tulevien asiakkaiden luokittelun lisäksi esimerkiksi luottokorttimarkkinoinnin kohderyhmän valintaan, kun luotettavien asiakkaiden määrää halutaan lisätä. Kehittyneemmät menetelmät ovat osoittaneet parempaa tarkkuutta, mutta koska ne ovat vaikeita tulkita logistinen regressio on pysynyt suosiossaan luottoriskiä arvioitaessa (Dong et al., 2010).

Luottoriskin mallintamisessa ilmenee monia haasteita, kuten epätasapainoisen datan ja variaation vaikutus menetelmästä saatuun tulokseen (Li et al., 2019). Epätasapainoinen data tarkoittaa, että toinen luokista esiintyy datassa huomattavasti vähemmän kuin toinen luokista. Epätasapaino saattaa aiheuttaa ongelmia koulutusdatan tasapainottamisessa, joka saattaa aiheuttaa muodostettuihin malleihin harhaisuutta. Voimakkaasti epätasapainoinen data on yleistä luottoriskimallinnuksessa (Li et al., 2019), minkä takia sen huomioon ottaminen on tärkeää esimerkiksi mallin hyvyyttä arvioitaessa. Mallin hyvyyden arviointiin käytetään usein mallin

tarkkuutta (accuracy) kuvaavaa mittaria, joka kertoo oikeiden tehtyjen ennusteiden määrän suhteessa kaikkiin mallin tekemiin ennusteisiin. Voimakkaasti epätasapainoisesta datasta muodostetut mallit voivat tällöin saada korkean tarkkuuden, jos ne luokittelevat kaikki tapaukset luokkaan, josta on enemmän dataa (Akosa, 2017). Tällaiset mallit ovat kuitenkin hyödyttömiä ennustamisessa. Tämän takia mallia tulee tutkia myös muiden mittareiden, kuten sensitiivisyyden (sensitivity) ja positiivisen ennustearvon (precision) avulla, jotka arvioivat mallia myös epätasapainoisen datan tapauksessa. Mallin sensitiivisyys kertoo positiivisten ennusteiden tarkkuudesta eli suhteessa vähemmän esiintyvän luokkaan kuuluvien ennusteiden tarkkuudesta (Akosa, 2017). Positiivinen ennustearvo lasketaan oikein positiiviseksi luokiteltujen suhde kaikkiin positiiviseksi luokiteltuihin (Akosa, 2017).

Mallin epätasapainoa voidaan korjata esimerkiksi perinteisillä tavoilla hyödyntämällä aliotantaa, yliotantaa tai niiden yhdistelmää. Aliotanta poistaa suuremman luokan tapauksia datasta, kun taas yliotanta lisää pienemmän luokan tapauksia datasta toiston avulla. Näihin otantamenetelmiin liittyy kuitenkin monia haittoja, esimerkiksi yliotannan aiheuttama datan ylisovittaminen ja aliotannan mahdollisesti tärkeiden datojen poistaminen. Otantamenetelmien lisäksi epätasapainoa voidaan korjata painottamalla uudelleen logistisessa regressiossa käytettävää uskottavuusfunktiota (Li et al., 2019).

3 Menetelmät ja aineisto

3.1 Logistinen regressio

Työssä tehtävä malli toteutetaan logistisen regression avulla. Logistinen regressio on hyvä työkalu luottoriskin arvioimiseen, kun luottoriskiä lähestytään binäärisenä luokitteluongelmana. Tässä työssä tämä tarkoittaa sitä, että luottokorttiyhtiön asiakkaat luokitellaan maksukyvyttömyyden mukaan. Regressioanalyysillä tarkoitetaan tilastollista menetelmää, jolla voidaan tutkia selitettävän muuttujan riippuvuutta selittäviin muuttujiin (Hilbe, 2009). Regressioanalyysin muotoja on monia, ja ne eroavat toisistaan muuttujien ominaisuuksien, määrän sekä riippuvuuden mukaan. Regressionanalyysin muoto valitaan sen mukaan, mikä menetelmä tuottaa parhaimman mallin. Menetelmän valinnassa täytyy ottaa huomioon myös selitettävän muuttujan ominaisuudet.

Regressioanalyyseistä tyypillisin on lineaarinen regressio, jonka avulla voidaan tutkia selitettävän muuttujan lineaarista riippuvutta selittäviin muuttujiin (Hilbe, 2009). Selitettävän muuttujan täytyy lineaarisen regression oletuksien mukaan olla jatkuva. Tämän takia lineaarista regressiota ei voida käyttää binääriseen luokitteluun. Binäärisissä luokitteluongelmissa voidaan tällöin käyttää tavallista logistista regressiota, jonka selitettävä muuttuja on binäärinen (Hilbe, 2009). Binäärisen muuttujan mahdollisia arvoja ovat 0 ja 1. Tyyppilisesti arvo 0 tarkoittaa, että tutkittava asia ei toteudu ja arvo 1 tarkoittaa asian toteutumista. Logistinen regressio on yksi regressioanalyysin muotoja.

Lineaarinen regressio ja logistinen regressio kuuluvat yleistettyihin lineaarisiin malleihin (McCullagh ja Nelder, 1989). Yleistetyt lineaariset mallit ovat laajennus

tavallisesta lineaarisesta mallista. Tavallinen lineaarinen malli on muotoa

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \tag{1}$$

jossa y_i on havaittu selitettävä muuttuja, β_0 on vakioselittäjän ei-satunnainen regressiokerroin, β_j on selittävän muuttujan x_{ij} ei-satunnainen regressiokerroin ja ϵ_i mallin satunnainen virhetermi. Virhetermin odotusarvo on nolla lineaarisen mallin oletuksien mukaan, jolloin selitettävän muuttujan odotusarvoa voidaan ennustaa kaavoilla

$$\begin{aligned} E(y_i) &= E(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i) \\ &= \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + E(\epsilon_i) \\ &= \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j. \end{aligned} \tag{2}$$

Yhtälö 2 on lineaarisen regression lauseke, jossa regressiokerroin β_j kertoo, kuinka paljon selitettävän muuttujan odotusarvo $E(y_i)$ muuttuu, kun selittävä muuttuja x_{ij} kasvaa yhden yksikön (Weisberg, 2014). Yleistetyssä lineaarisessa mallissa rakennetaan lineaarinen malli selitettävän muuttujan odotusarvon muunnokselle (Madsen ja Thyregod, 2010). Yleistetty lineaarinen malli on tavallisen lineaarisen mallin laajennus. Sen lauseke on muotoa

$$g(E(y_i)) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, \tag{3}$$

jossa linkkifunktio g määrittää funktion rakenteen ja yhdistää selitettävän muuttujan odotusarvon lineaarisesti selitettäviin muuttujiin (McCullagh ja Nelder, 1989).

Regressiomenetelmä valitaan data-aineistolle sopivaksi, ja näin ollen linkkifunktion valintaan vaikuttaa selitettävän muuttujan tyyppi ja jakauma. Lineaarisen regression tapauksessa linkkifunktio on $g(E(y_i)) = E(y_i)$.

Logistisessa regressiossa linkkifunktio on logit-muunnos $g(E(y_i)) = ln(E(y_i)/(1 - E(y_i))$ (Hilbe, 2009). Selitettävän muuttujan odotusarvoa merkitään usein logistisen regression tapauksessa $p(x_i) = E(y_i|x_i)$, jossa $p(x_i) = P(y_i = 1|x_i)$. Näin ollen logistisen regression tapauksessa odotusarvo kuvaa ehdollista todennäköisyyttä, jolla selitettävä muuttuja saa arvon 1, selittävällä muuttujalla x_i. Logistisessa regressiossa selitettävä muuttuja on Bernoulli-jakautunut, jolloin selitettävä muuttuja saa arvon 0 todennäköisyydellä $P(y_i = 0|x_i) = 1 - p(x_i)$. Vastasuhde (odds) saadaan laskettua tapahtuman toteutumisen ja sen tapahtumatta jäämisen suhteena. Lopullinen linkkifunktio, eli logit-muunnos, saadaan vastasuhteen luonnollisesta logaritmista. Logistisen regressiolauseke saadaan tällöin sijoittamalla logit-muunnos linkkifunktion paikalle lausekkeessa 3. Näin ollen lausekkeeksi saadaan

$$logit(p(x_i)) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, \tag{4}$$

missä logit-muunnos on muotoa

$$logit(p(x_i)) = ln(\frac{p(x_i)}{1 - p(x_i))}), \tag{5}$$

missä

$$\frac{p(x_i)}{1 - p(x_i))} = \frac{P(y_i = 1|x_i)}{1 - P(y_i = 1|x_i)}. \tag{6}$$

Lausekkeen 4 regressiokerroin β_j kertoo, kuinka paljon tapahtuman $y_i = 1$ log-vastasuhde (log-odds) muuttuu, kun selittävä muuttuja x_{ij} kasvaa yhden yksikön. Muutos vastasuhteessa voidaan tällöin laskea lausekkeen e^{β_j} avulla. Regressiokertoimen etumerkki kertoo selittävän muuttujan vaikutuksesta tutkittavan tapahtuman toteutumisen todennäköisyyteen $p(x_i)$. Positiivinen arvo kertoo muuttujan kasvattavan todennäköisyyttä $p(x_i)$, kun taas negatiivinen arvo vähentää todennäköisyyttä $p(x_i)$. Todennäköisyyden $p(x_i)$ arvo voidaan ratkaista lausekkeesta (4)

$$\begin{aligned} ln(\frac{p(x_i)}{1 - p(x_i)}) &= \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \\ \Leftrightarrow \frac{p(x_i)}{1 - p(x_i)} &= e^{(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)} \\ \Leftrightarrow p(x_i) &= \frac{e^{(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)}}{1 + e^{(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)}} \\ \Leftrightarrow p(x_i) &= \frac{1}{1 + e^{(-(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j))}} \end{aligned} \tag{7}$$

Merkitään jatkossa $b = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$ lausekkeiden yksinkertaistamiseksi. Lauseke (7) on logistinen funktio, joka muuttaa sille annetut syötteet todennäköisyydeksi. Näiden todennäköisyyksien avulla logistista regressiota voidaan käyttää tapausten luokitteluun.

3.2 Mallin sovittaminen

Mallin sovittamisella pyritään löytämään aineistolle paras mahdollinen sovite estimoimalla parametrit β_0 ja β_j (Hosmer Jr et al., 2013). Tavallisen lineaarisen mallin sovitus tehdään usein pienimmän neliösumman menetelmällä. Pienimmän neliösumman menetelmässä parametrien estimaatit $\hat{\beta}_0$ ja $\hat{\beta}_j$ valitaan niin, että ne ne minimoivat virhetermien neliösumman. Kyseistä menetelmää ei voida käyttää logistiseen regressioon, sillä selitettävän muuttujan ollessa binäärinen parametrien estimaattoreilla ei ole menetelmään tarvittavia oletuksia (Hosmer Jr et al., 2013). Logistisessa regressiossa mallin sovittamiseen käytetään suurimman uskottavuuden menetelmää, joka on yleistys pienimmän neliösumman menetelmästä. Suurimman uskottavuuden menetelmässä parametrit estimoidaan siten, että todennäköisyys havaitulle datalle on mahdollisimman suuri. Menetelmässä käytetään uskottavuusfunktiota, joka

kuvaa havaittujen datapisteiden yhteistä todennäköisyyttä tuntemattomien parametrien funktiona. Toisin sanoin, suurimman uskottavuuden menetelmä maksimoi uskottavuusfuktion mallin parametrien β_0 ja β_j suhteen.

Logistisen regression selitettävä muuttuja on Bernoulli-jakautunut $y_i = B(p(x_i))$, jossa $p(x_i)$ kertoo tuloksen 1 todennäköisyyden. Täten selitettävä muuttuja saa arvon 1 todennäköisyydellä $p(x_i)$ ja arvon 0 todennäköisyydellä 1-$p(x_i)$. Logaritmisessa regressiossa käytettävässä uskottavuusfunktiossa käytetään Bernoulli-jakauman tiheysfunktiota, joka kertoo yksittäisen havainnon todennäköisyyden. Bernoullin-jakauman tiheysfunktio on muotoa

$$f(y_i|p(x_i)) = (1 - p(x_i))^{1-y_i} p(x_i)^{y_i}. \tag{8}$$

Kun havainnot voidaan olettaa toisistaan riippumattomattomiksi ja identtisesti jakautuneiksi, uskottavuusfunktioksi saadaan

$$\begin{aligned} L(\beta) &= f(y_1|\beta) \cdots f(y_n|\beta) \\ L(\beta) &= \prod_{i=1}^{n} f(y_i|\beta). \end{aligned} \tag{9}$$

Logaritmisen regression oletuksiin kuuluu, että havainnot ovat identtisesti jakautuneita ja riippumattomia, joten logistisessa regressiossa käytettävä uskottavuusfunktio on saadaan sijoittamalla yhtälöön (9) Bernoullin-jakauman tiheysfunktio yhtälöstä (8) (Hosmer Jr et al., 2013)

$$L(\beta) = \prod_{i=1}^{n} (1 - p(x_i))^{1-y_i} p(x_i)^{y_i}. \tag{10}$$

Uskottavuusfunktion maksimointi on usein helpompaa logaritmisen muutoksen avulla eli tarkastelemalla logaritmista uskottavuusfunktiota $l(\beta) = ln(L(\beta))$. Logaritmi on aidosti kasvava, joten logaritminen uskottavuusfunktio $l(\beta)$ saavuttaa maksiminsa samassa pisteessä kuin uskottavuusfunktio $L(\beta)$. Näin ollen lausekkeen (9) uskottavuusfunktiota vastaa logaritminen uskottavuusfunktio

$$\begin{aligned} l(\beta) &= ln(\prod_{i=1}^{n} (1 - p(x_i))^{1-y_i} p(x_i)^{y_i}) \\ l(\beta) &= \sum_{i=1}^{n} (1 - y_i) ln(1 - p(x_i)) + y_i ln(p(x_i)). \end{aligned} \tag{11}$$

Sijoittamalla todennäköisyyden $p(x_i)$ lauseke (7) lausekkeeseen (11) saadaan logaritmiseksi uskottavuusfunktioksi:

$$l(\beta) = \sum_{i=1}^{n} (1 - y_i) ln(1 - (\frac{1}{1 + e^{-b}})) + y_i ln(\frac{1}{1 + e^{-b}}), \tag{12}$$

jossa $b = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$

Uskottavuusfunktio maksimoidaan derivoimalla logaritmista uskottavuusfunktiota erikseen parametrien β_0 ja β_j suhteen ja ratkaisemalla näiden derivaattafunktioiden nollakohdat eli

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta_0} &= \sum_{i=1}^{n}(1-y_i)\frac{\partial}{\partial \beta_0} ln(1-(\frac{1}{1+e^{-b}})) + y_i \frac{\partial}{\partial \beta_0} ln(\frac{1}{1+e^{-b}}) \\
&= \sum_{i=1}^{n}(1-y_i)(-1+(\frac{e^{-b}}{1+e^{-b}})) + y_i(\frac{e^{-b}}{1+e^{-b}}) \\
&= \sum_{i=1}^{n}(y_i - \frac{1}{1+e^{-b}}) \\
&= \sum_{i=1}^{n}(y_i - p(x_i)),
\end{aligned} \tag{13}$$

$$\begin{aligned}
\frac{\partial l(\beta)}{\partial \beta_j} &= \sum_{i=1}^{n}(1-y_i)\frac{\partial}{\partial \beta_j} ln(1-(\frac{1}{1+e^{-b}})) + y_i \frac{\partial}{\partial \beta_j} ln(\frac{1}{1+e^{-b}}) \\
&= \sum_{i=1}^{n}(1-y_i)x_{ij}(-1+(\frac{e^{-b}}{1+e^{-b}})) + y_i x_{ij}(\frac{e^{-b}}{1+e^{-b}}) \\
&= \sum_{i=1}^{n} x_{ij}(y_i - \frac{1}{1+e^{-b}}) \\
&= \sum_{i=1}^{n} x_{ij}(y_i - p(x_i)).
\end{aligned} \tag{14}$$

Näin ollen uskottavuusfunktion maksimoivat parametrit saadaan ratkaisemalla yhtälöt

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - p(x_i)) = 0, \\
\sum_{i=1}^{n} x_{ij}(y_i - p(x_i)) = 0.
\end{aligned} \tag{15}$$

Yhtälöt (15) eivät ole lineaarisia, joten niille ei voida esittää analyyttistä ratkaisua (Hosmer Jr et al., 2013). Näin ollen parametrit ratkaistaan numeerisesti. Yhtälöiden ratkaisuun käytetään iteratiivisia menetelmiä, jotka ovat useimmissa ohjelmistoissa valmiiksi asennettuina. Tässä työssä käytetään RStudiota, joka käyttää iteratiivisena menetelmänä Fisherin pisteytystä (Fisher's scoring). Fisherin pisteytys on Newtonin menetelmän muoto. Menetelmässä malli sovitetaan samalla tavalla kuin iteratiivisessa painotetussa pienimmän neliösumman menetelmässä (IRLS) (Schworer ja Hovey, 2004). Painotetun pienemmän neliösumman lauseke on muotoa

$$\sum (y_i - \hat{y})^2 w_i, \tag{16}$$

jossa y_i on havaittu datapiste, $\hat{y}$ on ennustettu arvo ja w_i on paino. Fisherin pisteytyksessä lauseketta (16) käytetään iteraatiossa, eli lauseketta iteroidaan kunnes päästään optimaaliseen tulokseen. Fisherin pisteytys on ohjelmistossa oletuksena, mutta RStudiossa on myös mahdollista käyttää muita iteratiivisa mentelmiä.

3.3 Mallin arviointi

Mallin sovittamisen jälkeen sen hyvyyttä täytyy arvioida. Hyvyyden arviointi on tärkeä osa malli rakentamista, sillä sen avulla voidaan valita paras mahdollinen malli tutkimalla niiden ennustuskykyä. Tässä työssä arviointiin käytetään luokittelutaulukkoa (confusion matrix) ja erottelukykykäyrää (ROC, receiver operating characteristic). Arvioinnissa data-aineisto jaetaan satunnaisesti koulutus- ja testi-osaksi. Tämän jälkeen koulutus-osalla koulutetulle mallille annetaan syötteeksi testi-osa ja sen antamia ennusteiden arvoja verrataan testi-osan todellisiin selitettävän muuttujan arvoihin. Arvoja voidaan helposti vertailla luokittelutaulukon avulla, jota yleisesti käytetään mallin hyvyyden visualisointiin. Luokittelutaulukossa käytetään yleensä luokittelun kynnysarvona 0.50. Tämä tarkoittaa että tapaukset, jotka ovat logistisen funktion mukaan yli tai tasan 50 prosentin todennäköisyydellä luokkaan 1 kuuluvia, luokitellaan luokkaan 1. Luokittelutaululukon rivit vastaavat todellisia arvoja, kun taas sarakkeet vastaavat ennustettuja arvoja, taulukon 1 mukaisesti:

Taulukko 1: Luokittelutaulukko

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

Taulukossa 1 käytetty merkintä TP tarkoittaa oikeaa positiivista (true positive) ja TN tarkoittaa oikeaa negatiivista (true negative). Oikean positiivisen tapauksessa malli ennustaa oikein positiivisen lopputuloksen (arvon 1) ja oikean negatiivisen tapauksessa se ennustaa negatiiviset tapaukset (arvon 0) oikein. Väärää positiivista, eli FP (false positive), kutsutaan myös tyypin I virheeksi. Tässä tapauksessa malli luokittelee negatiiviset tapaukset virheellisesti positiivisiksi. Väärää negatiivista, eli FN (false negative), kutsutaan tyypin II virheeksi, eli malli luokittelee positiiviset tapaukset virheellisesti negatiivisiksi.

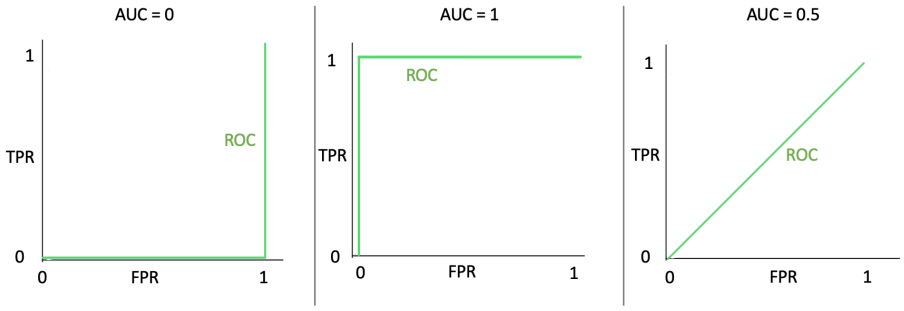
Luokittelutaulukon merkintöjen avulla voidaan laskea erilaisia mittareita, joiden avulla voidaan arvioida mallin ennustuskykyä. Nämä mittarit ovat kappaleessa 2.2 mainitut tarkkuus, sensitiivisyys ja positiivinen ennustearvo. Tarkkuus kertoo mallin oikein luokiteltujen tapausten määrän suhteessa kaikkiin tehtyihin ennusteisiin. Näin ollen tarkkuus kertoo, kuinka usein malli luokittelee havainnon oikeaan luokkaan. Sensitiivisyys kertoo, kuinka usein malli luokittelee luokkaan 1 kuuluvan havainnoin oikein. Positiivinen ennustearvo nimensä mukaisesti kertoo, kuinka usein mallin harvemman esiintyvän luokan ennustukset ovat oikein. Näin ollen mittareille saadaan seuraavat kaavat:

$$\begin{aligned} \text{Tarkkuus} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{Sensitiivisyys} &= \frac{TP}{TP + FN} \\ \text{Positiivinen ennustearvo} &= \frac{TP}{TP + FP} \end{aligned} \tag{17}$$

Näiden mittarien lisäksi mallia on hyvä arvioida muilla tavoilla, esimerkiksi erottelukykykäyrän, eli ROC-käyrän avulla, joka voidaan muodostaa luokittelutaulukon merkintöjen avulla. ROC-käyrä on kaksiulotteinen kaavio, jossa pystyakselilla on mallin oikeiden posiitivisten osuus (TPR) ja vaaka-akselilla sen väärien positiivisten osuus (FPR) kullakin mahdollisella kynnysarvolla laskettuna (Fawcett, 2006). Oikeiden positiivisten osuus on mallin sensitiivisyys. Väärien positiivisten osuus lasketaan väärin positiivisesti luokiteltujen tapahtumien ja todellisten negatiivisten tapahtumien kokonaismäärän suhteena

$$FPR = \frac{FP}{FP + TN}. \tag{18}$$

Ensimmäiseksi kynnysarvoksi valitaan arvo, joka luokittelee kaikki tapaukset luokkaan 1, jonka jälkeen lasketaan mallin FPR ja TPR. Seuraavaksi kynnysarvoksi valitaan arvo, joka luokittelee kaikki paitsi logistisen funktion ensimmäisen tapauksen luokkaan 1, ja tämän jälkeen lasketaan jälleen FPR ja TPR. Kynnysarvon vaihtamista jatketaan valitsemalla aina yksi enemmän luokkaan 0 luokiteltavaksi, ja laskemalla FPR ja TPR (Fawcett, 2006). Tarkempi ROC-käyrä saataisiin laskemalla FPR ja TPR arvot jokaisen mahdollisen kynnysarvon kohdalla. Lopulta näin saadaan muodostettua ROC-käyrä. Näin ollen ROC-käyrä yhdistää monien luokittelutaulukkojen tiedot yhteen helposti tulkittavaan muotoon. Kuvassa 1 on esitetty esimerkkejä erilaisista ROC-käyristä.



Kuva 1: Esimerkkejä ROC-käyristä

Mitä parempi malli on kyseessä, sitä lähempänä ROC-käyrä kulkee vasenta yläkulmaa. Mallilta halutaan mahdollisimman korkea TPR ja alhainen FPR, jotta se luokittelee positiivisen luokkaan, eli luokkaan 1, kuuluvat tapaukset oikein. Näin ollen ROC-käyrän kulkiessa vasenta laitaa kuvan 1 keskimmäisen käyrän mukaisesti, erottelukyky on täydellinen. ROC-käyrän ollessa oikeanpuoleisin kuvan 1 käyristä erottelukyky on olematon. Tällöin malli on arvoton, ja se erottelee luokat satunnaisesti toisistaan. Jos mallin diagnoosikriteeri on valittu väärinpäin ROC-käyrä kulkee läheltä oikeaa alakulmaa, kuvan 1 vasemmanpuoleisimman käyrän mukaan. Tällöin malli luokittelee luokan 0 tapaukset luokkaan 1, ja luokan 1 tapaukset luokkaan 0.

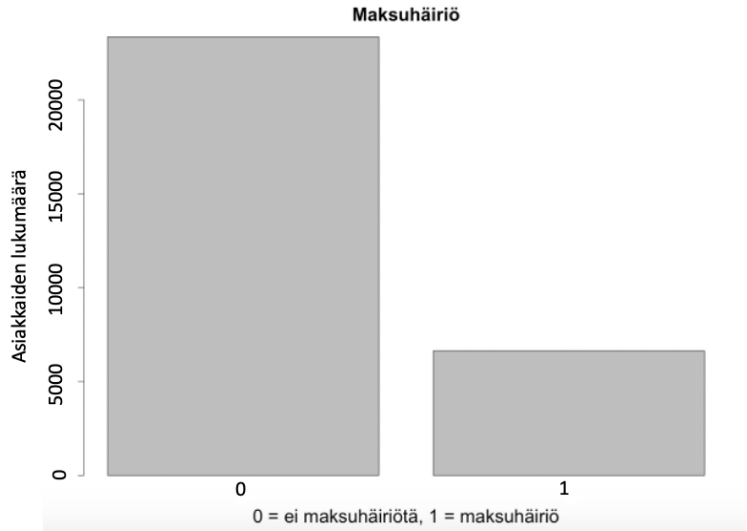
ROC-käyrää voidaan käyttää optimaalisen kynnysarvon löytämiseen. Tämän lisäksi ROC-käyrää voidaan käyttää sen alle jäävän pinta-alan suuruuden laskemiseksi,

eli AUC:in (area under the curve) määrittämiseksi. AUC kertoo todennäköisyyden sille, että malli antaa suuremman odotusarvon luokkaan 1 kuuluvalle havainnolle, kuin luokkaan 0 kuuluvalle havainnolle (Cortes ja Mohri, 2003). Näin ollen mitä korkeampi AUC-arvo, sitä parempi erottelukyky on. AUC voi saada arvoja väliltä [0,1]. Näin ollen kun ROC-käyrä on kuvan 1 vasemmanpuoleisimman käyrän mukainen, käyrän alle jäävä pinta-ala saa arvon 0, mikä tarkoittaa, että mallin diagnoosikriteeri on valittu väärinpäin. Kun taas AUC saa arvon 1, malli erottelee havainnot kahteen luokkaan täydellisesti, tällöin käyrä kulkee vasenta laitaa kuvan 1 keskimmäisen käyrän mukaisesti. Arvon ollessa väliltä]0.5,1[malli pystyy todennäköisesti erottelemaan havainnot kahteen luokkaan oikein. Arvo 0.5 tarkoittaa, että malli ei pysty erottamaan 0 luokkaan kuuluvia havaintoja 1 luokaan kuuluvista havainnoista, ja tällöin ROC-käyrä on kuvan 1 oikeanpuoleisimman käyrän mukainen. AUC-arvon avulla voidaan arvioida eri luokittelumallien erottelukykyä, ja se antaa usein laajemman kuvan mallin suorituskyvystä, sillä se on laskettu monien kynnysarvojen avulla, toisin kuin luokittelutaulukosta lasketut edellä mainitut mittarit.

4 Tulokset

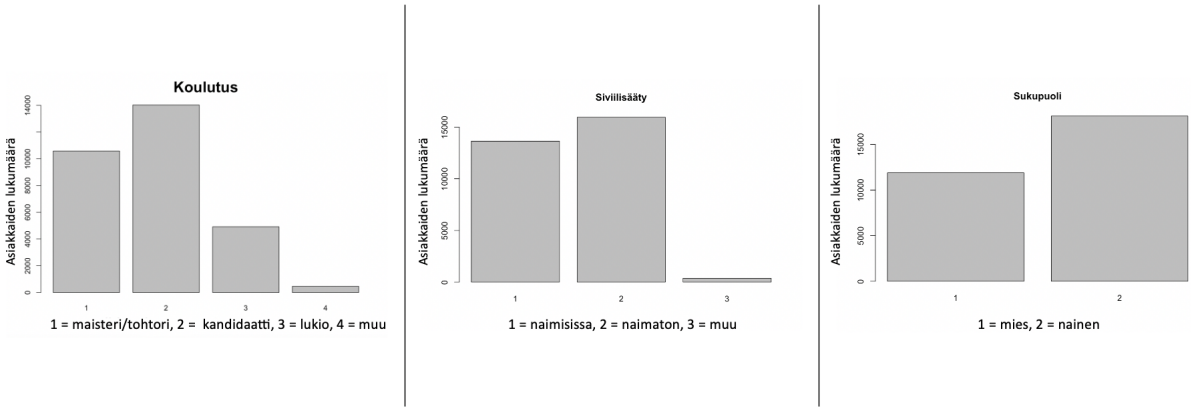
4.1 Aineisto

Mallin toteuttamiseen käytetään taiwanilaisen luottokorttiyhtiön asiakasdataa huhtikuusta syyskuuhun vuodelta 2005. Aineistossa on yhteensä 30 000 datapistettä. Data-aineistossa on yhteensä 24 muuttujaa, jotka ovat muodoltaan jatkuvia, ordinaalisia, kategorisia sekä binäärisiä. Selitettäväksi muuttujaksi ainestosta valitaan binäärinen asiakkaan maksukyvyttömyyttä (default) kuvaava muuttuja, joka saa arvon 1 maksuhäiriön tapahtuessa ja arvon 0 muuten. Maksuhäiriöllä tässä työssä tarkoitetaan sitä, että asiakas ei pysty maksamaan luottokorttimaksua seuraavana kuukautena. Kuvasta 2 voidaan nähdä maksukyvyttömien osuus data-aineistossa:



Kuva 2: Muuttujan maksuhäiriö jakauma

Kuvasta 2 nähdään, että maksukyvyttömien osuus data-aineistossa on suhteellisen suuri. Tämä tarkoittaa, että kappaleessa 2.2 mainitut epätasapainoisen datan ongelmat eivät todennäköisesti tule esille mallia rakennettaessa, joten ne voidaan sivuuttaa tässä työssä. Data-aineistossa kategorisia muuttujia ovat sukupuoli, koulutus sekä siviilisääty. Kategoriset muuttujat muutetaan binäärisiksi muuttujiksi, sillä logistinen regressio ei osaa käsitellä kategorisia muuttujia. Kategoristen muuttujien muuttamisen jälkeen muuttujia on yhteensä 30. Muuttujien muuttaminen tapahtuu muuttamalla jokainen kategorinen muuttuja yhtä moneksi binääriseksi muuttujaksi, kuin muuttujassa on kategorioita. Kategoristen muuttujien jakaumat on esitetty kuvassa 3.



Kuva 3: Kategoristen muuttujien jakaumat

Kuvan 3 jakaumista voidaan nähdä, että sukupuoli on jakautunut suhteellisen tasaisesti, mutta naisten määrä on kuitenkin suurempi. Siviilisääty on jakautunut suhteellisen tasaisesti naimisissa olevien ja naimattomien välillä, muiden osuus on todella pieni. Koulutuksen jakaumasta voidaan nähdä, että suurin osa asiakkaista on korkeakoulutettuja. Pienempi osuus asiakkaista on lukiolaisia, ja hyvin harvat ovat muista koulutuksista. Kategoriset muuttujat vaikuttavat olevan suhteellisen tasaisesti jakautuneita, joten mallin harhaisuus ei vaikuta tulevan suureksi ongemaksi mallia toteuttaessa.

Muut muuttujat ovat jatkuvia tai ordinaalisia. Muuttujat kertovat asiakkaan iästä, annetusta luoton määrästä, aikaisempien maksujen määrästä ja ajasta, sekä luottokortilta käytetystä määrästä. Taulukossa 2 on esitetty jatkuvien ja ordinaalisten muuttujien tilastollisia suureita. Muuttujia merkitään selkeyden takia nimillä x_i, jossa i=1,2,...,20.

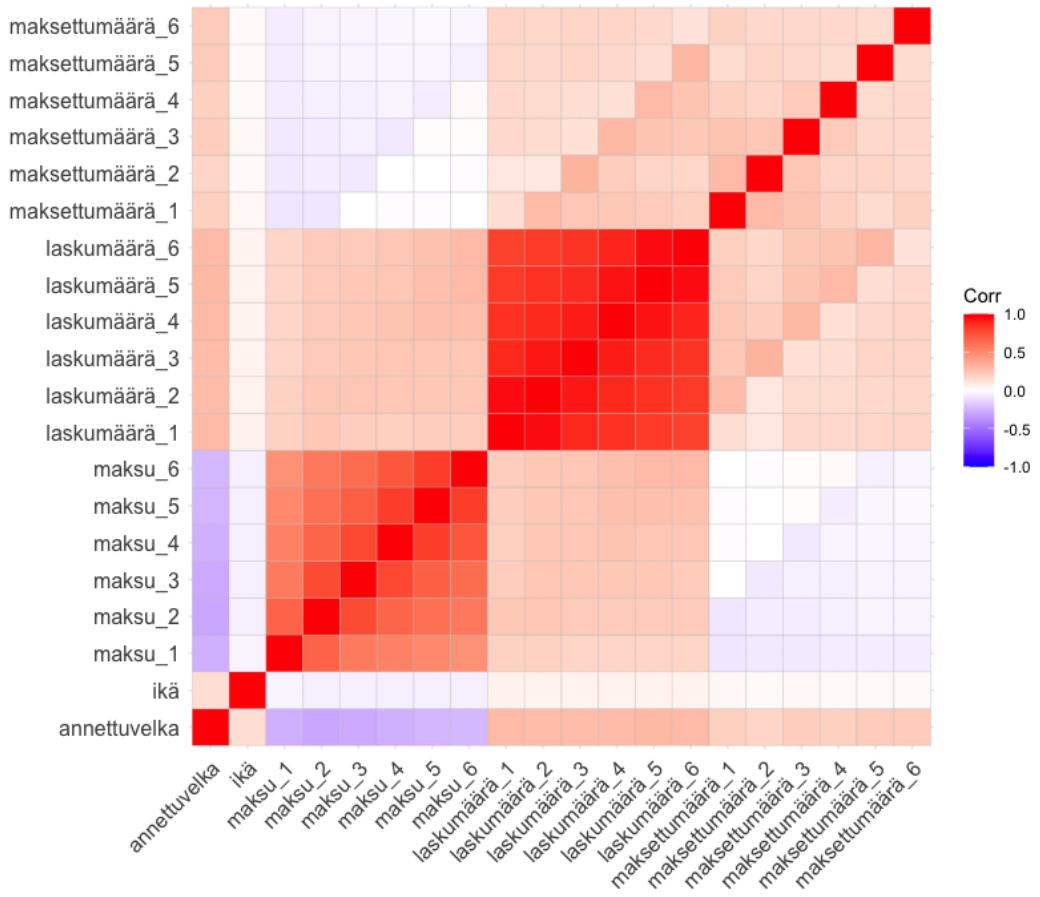
Taulukosta 2 nähdään, että suuruusluokat vaihtelevat suuresti muuttujien välillä. Esimerkiksi muuttujien $x_3 - x_8$ keskiarvot ovat lähellä nollaa, kun taas muuttujien $x_9 - x_{14}$ keskiarvot ovat kymmenentuhannen luokassa. Muuttujien suuruusluokan vaihtelu saattaa aiheuttaa vaikeuksia uskottavuusfunktion (11) maksimoimisessa ja ongelmia mallin kouluttamisessa. Suuruusluokan vaihtelun aiheuttamat ongelmat voidaan välttää muuttujien skaalauksella. Muuttujat skaalataan niin, että niiden keskiarvoksi tulee 0 ja keskihajonnaksi 1. Muuttujien skaalaus auttaa valittua ite-

	Keskiarvo	Mediaani	Keskihajonta
x_1	167484,30	140000,00	129747,70
x_2	35,49	34,00	9,22
x_3	-0,02	0,00	1,12
x_4	-0,13	0,00	1,20
x_5	-0,17	0,00	1,20
x_6	-0,22	0,00	1,17
x_7	-0,27	0,00	1,13
x_8	-0,29	0,00	1,15
x_9	51223,33	22381,50	73635,86
x_{10}	49179,08	21200,00	71173,77
x_{11}	47013,15	20088,50	69349,39
x_{12}	43262,95	19052,00	64332,86
x_{13}	40311,40	18104,50	60797,16
x_{14}	38871,76	17071,00	59554,11
x_{15}	5663,58	2100,00	16563,28
x_{16}	5921,16	2009,00	23040,87
x_{17}	5225,68	1800,00	17606,96
x_{18}	4826,08	1500,00	15666,16
x_{19}	4799,39	1500,00	15278,31
x_{20}	5215,50	1500,00	17777,47

Taulukko 2: Jatkuvien muuttujien tilastollisten suureiden arvoja

raatiomenetelmää konvergoitumaan nopeammin. Lisäksi skaalaus tekee lopullisen mallin regressiokertoimet vertailukelpoisiksi (Hilbe, 2009). Työssä skaalaus tehdään RStudio-ohjelmiston scale-funktion avulla.

Kuvassa 4 on esitetty korrelaatiomatriisi, josta voidaan nähdä työssä käytettävien jatkuvien muuttujien pareittaiset korrelaatiot. Korrelaatiomatriisissa punainen väri tarkoittaa voimakasta positiivista korrelaatiota ja sininen väri voimakasta negatiivista korrelaatiota. Valkoinen väri tarkoittaa, että muuttujien välinen korrelaatio on nolla. Muuttujien perässä olevat numerot kuvastavat vuoden 2005 kuukausia. Kuvan 4 matriisista voidaan nähdä, että $laskumäärä_{(1-6)}$ nimiset muuttujat korreloivat keskenään erittäin voimakkaasti. Myös $maksu_{(1-6)}$ nimiset muuttujat korreloivat keskenään suhteellisen voimakkaasti. Heikoin korrelaatio on muuttujalla ikä muiden jatkuvien muuttujien kanssa, sekä $maksettumäärä_{(1-6)}$ korralaatio muuttujien $maksu_{(1-6)}$ välillä on heikkoa. Logistisen regression yksi oletuksista on selittävien muuttujien korreloimattomuus (Kumari, 2008). Voimakas korrelaatio selittävien muuttujien välillä kasvattaa regressiokertoimien keskivirhettä, jolloin niistä tulee epävakaita. Tällöin regressiokertoimien tilastollinen merkittävyys saattaa muuttua, ja niiden suuruus sekä vaikutuksen suunta saattaa vaihdella helposti datan muuttuessa. Tämä tekee mallista epäluotettavan ja huonontaa sen ennustekykyä. Näin ollen mallia kehittäessä selittävien muuttujien korrelaatiota tulisi vähentää poistamalla voimakkaasti korreloivia muuttujia, muuttamalla muuttujien muotoa tai lisäämällä data-aineistoa (Kumari, 2008).



Kuva 4: Jatkuvien muuttujien korrelaatiomatriisi

4.2 Malli

Aineiston läpikäymisen jälkeen malliin valitaan yhteensä 25 muuttujaa. Muuttujat $laskumäärä_{(2-6)}$ poistettiin mallista selittävien muuttujien välisen korrelaation vähentämiseksi. Tämän lisäksi sovitettavasta datasta poistetaan muuttujat $sukupuoli_2$, $avioliitto_2$, $avioliitto_3$ ja $koulutus_4$, sillä ne ovat riippuvaisia muiden samassa luokassa olevien kategorioiden kanssa. Muuttujien valinnan lisäksi ennen mallin sovittamista tehdään kappaleessa 4.1 mainittu data-aineiston skaalaus, sekä muu datan käsittely. Datan käsittelyyn sisältyy kategorisissa muuttujissa muihin kuin määriteltyihin luokkiin kuuluvien arvojen siirtäminen määritettyihin luokkiin. Data-aineiston käsittelyn ja muuttujien valinnan jälkeen data-aineisto jaetaan kappaleessa 3.3 mainittuun koulutus- ja testi-osaan. Tässä työssä 70% data-aineistosta käytetään koulutus-osioon ja loput 30% on testi-osiota. Koulutus-osaa käytetään mallin kouluttamiseen, ja testi-osaa saadun mallin arviointiin.

Malli toteutaan RStudio-ohjelmiston glm-funktion avulla. Työssä hyödynnettävää glm-funktiota voidaan käyttää erilaisiin yleistettyjen lineaaristen mallien sovittamiseen. Funktion parametreiksi tarvitaan selitettävä muuttuja, selittävät muuttujat, linkkifunktio sekä käytettävä data-aineisto. Tässä työssä selitettävänä muuttujana on

maksukyvyttömyys ja selittävinä muuttujina loput 20 muuttujaa. Linkkifunktiona on osiossa 3.1 esitelty logit-muunnos ja data-aineistona käytetään valittua koulutus-osaa. Iteraatiomenetelmänä käytetään funktion vakiona olevaa Fisherin pisteytystä.

Sovittamisen tulokset, eli regressiokertoimien arvot ja niiden luottamusvälit ovat esitetty taulukossa 3. Taulukkoon on merkitty vihreällä värillä muuttuja, jolla on maksukyvyttömyyden todennäköisyyteen suurin vaikutus ja punaisella muuttuja, jolla on pienin vaikutus todennäköisyyteen. Data-aineiston skaalauksen avulla regressiokertoimet ovat kaikki samaa suuruusluokkaa ja näin ne ovat keskenään vertailukelpoisia. Näin ollen itseisarvoltaan suurin arvo vaikuttaa maksukyvyttömyyden todennäköisyyteen eniten, ja itseisarvoltaan pienin arvo vaikuttaa vähiten. Regressiokertoimen etumerkki kertoo vaikutuksen suunnasta. Negatiivinen arvo pienentää maksukyvyttömyyden todennäköisyyttä, kun taas positiivinen arvo kasvattaa maksukyvyttömyyden todennäköisyyttä.

	Regressiokerroin	Luottamusvälin alaraja	Luottamusvälin yläraja
vakio	-2,52940	-2,96401	-2,13633
annettuvelka	-0,08119	-0,12900	-0,03378
ikä	0,04291	0,00293	0,08275
$maksu_1$	0,63349	0,58647	0,68057
$maksu_2$	0,10792	0,05096	0,16470
$maksu_3$	0,07034	0,00663	0,13369
$maksu_4$	0,06650	-0,00179	0,13448
$maksu_5$	0,03350	-0,03799	0,10478
$maksu_6$	0,02681	-0,03272	0,08613
$laskumäärä_1$	-0,10429	-0,15054	-0,05849
$maksettumäärä_1$	-0,19856	-0,28661	-0,11998
$maksettumäärä_2$	-0,16711	-0,26750	-0,07906
$maksettumäärä_3$	-0,07339	-0,14482	-0,01115
$maksettumäärä_4$	-0,06797	-0,13174	-0,01175
$maksettumäärä_5$	-0,02504	-0,07909	0,02323
$maksettumäärä_6$	-0,07747	-0,14027	-0,02036
$sukupuoli_1$	0,12928	0,05733	0,20111
$avioliitto_1$	0,19736	0,11733	0,27743
$koulutus_1$	0,99185	0,59648	1,42828
$koulutus_2$	0,90936	0,51606	1,34405
$koulutus_3$	0,88894	0,48916	1,32888

Taulukko 3: Mallin regressiokertoimet ja luottamusvälit

Taulukossa 3 vihreällä merkitty muuttuja $koulutus_1$ on itseisarvoltaan suurin, ja näin ollen se vaikuttaa eniten maksukyvyttömyyden todennäköisyyteen. Muuttuja $koulutus_1$ tarkoittaa sitä, että asiakas on koulutukseltaan maisteri tai tohtori. Muuttujan arvo on positiivinen, joten asiakkaan ollessa maisteri tai tohtori, maksukyvyttömyyden todennäköisyys kasvaa. Myös muuttujat $koulutus_2$ ja $koulutus_3$ vaikuttavat voimakkaasti maksukyvyttömyyden todennäköisyyteen, joten korkea koulutustaso kasvattaa maksukyvyttömyyden riskiä. Data-aineisto sisältää kuitenkin vain pienen määrän asiakkaita joilla ei ole lukio-, kandidaatti- tai maisteritason koulutusta, mikä saattaa aiheuttaa mallin harhaisuutta. Muuttujalla $maksu_1$ on myös suhteellisen suuri vaikutus maksukyvyttömyyden todennäköisyyden kasvuun. Muuttuja $maksu_1$ kertoo kuinka myöhässä syyskuun aikana tulleet luottokorttimaksut ovat maksettu. Maksut myöhässä maksanut asiakas maksaa laskut todennäköisesti myöhässä myös tulevaisuudessa, mikä kasvattaa maksukyvyttömyyden riskiä. Taulukossa 3 punai-

sella on merkitty muuttuja $maksettumäärä_5$, jolla on pienin itseisarvo. Siten sen vaikutus on pienin maksukyvyttömyyden todennäköisyyteen. Maksukyvyttömyyden todennäköisyyttä pienentää muuttujat $laskumäärä_1$, sekä kaikki $maksettumäärä$ muuttujat. Muuttuja $laskumäärä$ kertoo, kuinka paljon luottokortilta on käytetty rahaa kuukauden aikana. Muuttuja $maksettumäärä$ kertoo, kuinka paljon asiakas on maksanut kuukauden aikana luottokorttilaskuja. Parempituloisille ja luotettavammille asiakkaille on usein asetettu suuremmat luottorajat, jolloin he voivat käyttää luottokorttiaan enemmän, mikä kasvattaa $laskumäärä$ muuttujan arvoa. Hyvätuloiset ja luotettaviksi määritellyt asiakkaat ovat myös todennäköisesti kykeneviä maksamaan luottokorttilaskut, mikä selittää muuttujan pienentävän vaikutuksen maksukyvyttömyyden todennäköisyyteen. On myös todennäköisempää, että aikaisemmin paljon luottokorttilaskuja maksanut asiakas maksaa laskut myös tulevaisuudessa, mikä selittää $maksettumäärä$ muuttujan maksukyvyttömyyden riskiä pienentävän vaikutuksen.

Taulukossa 3 on esitetty myös regressiokertoimien luottamusvälien alarajat ja ylärajat 95% luottamustasolla. Luottamusväli kertoo välin, jolla regressiokertoimen estimaatti on 95%:n luottamuksella. Näin ollen luottamusvälejä voidaan käyttää estimaattien luotettavuuden tarkasteluun. Leveämpi luottamusväli viittaa epävakaaseen estimaattiin, kun taas kapea luottamusväli kertoo luotettavasta estimaatista. Taulukosta 3 voidaan nähdä, että luottamusvälit ovat suhteellisen kapeita. Muuttujien $maksu_4$, $maksu_5$, $maksu_6$ ja $maksettumäärä_5$ luottamusvälin alaraja on negatiivinen ja yläraja on positiivinen. Edellä mainitut muuttujat eivät ole sovitetun mallin mukaan tilastollisesti merkittäviä 5%:n merkitsevyystasolla. Tämä tarkoittaa, että kyseisillä muuttujilla ei ole tilastollisesti merkittävää vaikutusta asiakkaan maksukyvyttömyyteen. Muut taulukossa 3 esitetyt muuttujat ovat mallin mukaan tilastollisesti merkittäviä, joten niillä on tilastollisesti merkittävä vaikutus asiakkaan maksukyvyttömyyteen.

Mallin sovittamisen jälkeen mallin ennustekykyä voidaan arvioida osiossa 3.3 mainituilla menetelmillä. Malille muodostetaan luokittelutaulukko 4. Taulukosta 4 voidaan laskea mallin tarkkuus, sensitiivisyys ja positiivinen ennustearvo kaavojen 19 avulla.

Taulukko 4: Mallin luokittelutaulukko

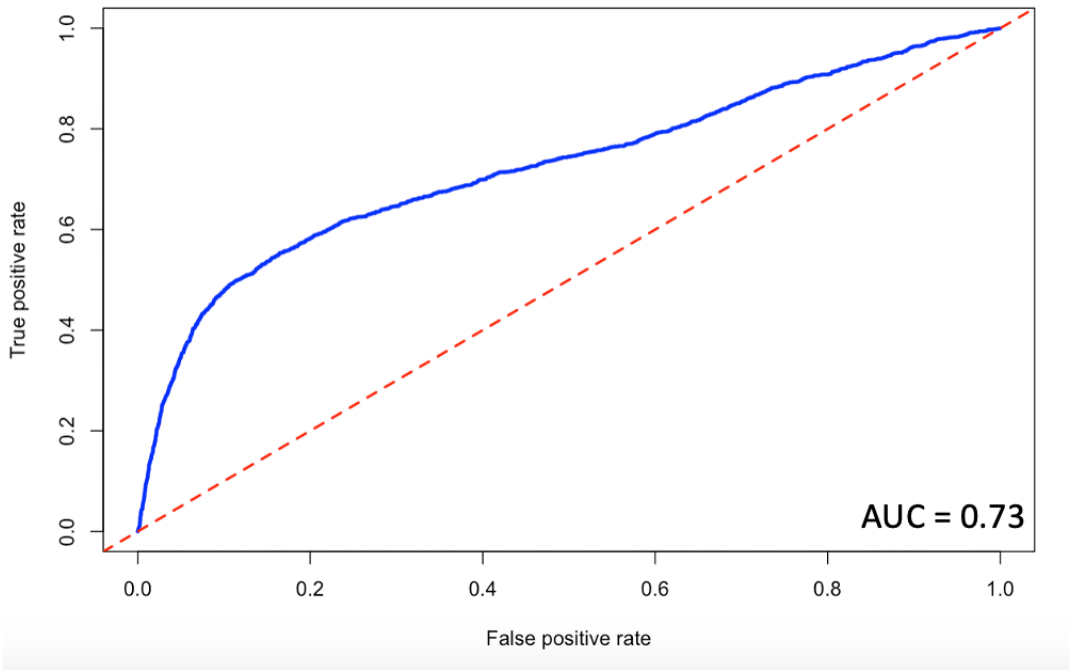
	Ennustettu 0	Ennustettu 1
Todellinen 0	6798	198
Todellinen 1	1501	503

$$\begin{aligned} \text{Tarkkuus} &= 0,811 \\ \text{Sensitiivisyys} &= 0,251 \\ \text{Positiivinen ennustearvo} &= 0,718 \end{aligned} \tag{19}$$

Mallin tarkkuus kertoo, että se luokittelee noin 81% tapauksista oikein. Sensitiivisyys kertoo, että malli tunnistaa maksukyvyttömän asiakkaan 25% todennäköisyydellä. Mallin positiivinen ennustearvo kertoo, että se luokittelee asiakkaan

maksukyvyttömäksi oikein noin 72% tarkkuudella. Näin ollen mallin tarkkuus ja positiivinen ennustearvot ovat hyviä, mutta sensitiivisyys on kohtuullisen alhainen. Sensitiivisyyttä voitaisiin nostaa pienentämällä luokittelutaulukon kynnysarvoa, jolloin tapauksia luokiteltaisiin enemmän maksukyvyttömyyden luokkaan.

Mallia arvioidaan myös osiossa 3.3 läpikäydyn ROC-käyrän avulla. Mallin ROC-käyrä on esitetty kuvassa 5 sinisellä viivalla. Punainen katkoviiva merkitsee AUC=0.5 tasoa. Kuten kuvasta 5 voidaan nähdä, ROC-käyrän alle jäävä pinta-ala on 0.73. Tämä AUC-arvo on hyväksyttävää tasoa. Näin ollen mallin ennustekyky on kohtuullisen hyvä.



Kuva 5: Mallin ROC-käyrä ja AUC

5 Yhteenveto

Työn tavoitteena oli arvioida luottoriskiä logistisen regression avulla. Tämä toteuttiin logistisella regressiomallilla, jonka avulla tutkittiin luottokorttiyhtiön asiakkaiden maksukyvyttömyystodennäköisyyttä. Työ aloitettiin luottoriskiin liittyvän aikaisemman tutkimuksen läpikäymisellä, jossa luottoriskin teoriaan ja mallintamiseen syvennyttiin. Lisäksi aikaisemmassa tutkimuksessa tutustuttiin luottoriskin mallintamiseen liittyviin haasteisiin ja etuihin. Tämän jälkeen esitettiin työssä käytettävä menetelmä eli logistinen regressio. Logistisen regression teorian jälkeen käsiteltiin mallin arviointia. Ennen mallin sovittamista käytössä olevaa data-aineistoa tutkittiin ja tarvittaessa muokattiin. Data-aineistosta rakennettin logististen regression avulla maksukyvyttömyyttä ennustava malli, jonka jälkeen mallin antamia ennusteita tulkittiin ja sen ennustekykyä arvioitiin.

Mallin tulosten perusteella asiakkaan maksukyvyttömyystodennälöisyys riippuu eniten asiakkaan koulutuksesta. Suurin vaikutus todennäköisyyteen oli muuttujalla, joka kertoi, oliko asiakas koulutukseltaan maisteri tai tohtori. Korkealla koulutustasolla oli täten maksukyvyttömyyden todennäköisyyttä kasvattava vaikutus. Mallin mukaan maksukyvyttömyyden todennäköisyyttä pienensi kuukauden aikana luottokortilta käytetyn rahan määrä sekä asiakkaan maksettujen luottokorttilaskujen suuruus. Mallin sovittamisen jälkeen sen ennustekykyä arvioitiin luokittelutaulukon ja erottelukykykäyrän avulla, joiden mukaan ennustekyky oli kohtalainen.

Aineistona työssä käytettiin taiwanilaisen luottokorttiyhtiön asiakkaiden dataa. Mallista saatuja tuloksia voitaisiin myös jossain määrin verrata muiden maiden asiakkaisiin, sillä luottokorttiyhtiöt ovat usein kansainvälisiä, eli niiden toimintaperiaatteet ovat samanlaisia maailmanlaajuisesti. Tuloksia voitaisiin täten hyödyntää myös suomalaisten luottokorttiyhtiöiden asiakkaiden luottoriskin arviointiin.

Työssä käytettyssä aineistossa selitettävän muuttujan, eli maksukyvyttömyyden, osuus oli suhteellisen suuri, mutta silti huomattavasti pienempi kuin toisen luokan osuus. Näin ollen data-aineiston epätasapaino saattoi aiheuttaa mallissa ongelmia. Näitä ongelmia voitaisiin jatkossa välttää korjaamalla data-aineiston epätasapainoa hyödyntämällä esimerkiksi ali- tai yliotantaa. Tämä voisi parantaa mallin ennustekykyä. Lisäksi logistisella regresiolla voitaisiin päästä parempiin tuloksiin, jos data-aineistoa ja erilaisia muuttujia olisi saatavilla enemmän. Maksukyvyttömyyden todennäköisyyttä voitaisiin tarkemmin ennustaa erilaisilla koneoppimenetelmillä, esimerkiksi neuroverkostoilla ja päätöspuilla. Toisaalta tekoälymenetelmien tulkittavuus olisi vaikeampaa.

Rakennetussa mallissa ei keskitytty muuttujien aiheuttamaan eri ryhmien tasapuoliseen kohteluun. Esimerkiksi muuttujat sukupuoli ja avioliitto voivat aiheuttaa luokittelussa joidenkin ihmisryhmien tahatonta syrjintää. Oikeiden muuttujien valintaa ja sen haasteita voitaisiin tutkia pidemmälle. Tällöin mallin ennustekykyä voitaisiin tutkia erilaisilla muuttujilla, mikä mahdollisesti voisi parantaa mallin ennustekykyä ja samalla myös mallin tasapuolisuutta.

Viitteet

Hussein A Abdou ja John Pointon. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management*, 18(2-3):59–88, 2011.

Rob Aitken. All data is credit data: Constituting the unbanked. *Competition & Change*, 21(4):274–300, 2017.

Josephine Akosa. Predictive accuracy: A misleading performance measure for highly imbalanced data. Teoksessa *Proceedings of the SAS Global Forum*, volume 12, 2017.

Florentin Butaru, Qingqing Chen, Brian Clark, Sanmay Das, Andrew W Lo, ja Akhtar Siddique. Risk and risk management in the credit card industry. *Journal of Banking & Finance*, 72:218–239, 2016.

EU Commission et al. Communication from the commission to the European parliament, the European council, the council, the European economic and social committee and the committee of the regions: Artificial intelligence for Europe. *A Clean Planet for All. A European strategic long-term vision for a prosperous, modern, competitive and climate neutral economy. Brussels*, 28:2018, 2018.

Corinna Cortes ja Mehryar Mohri. AUC optimization vs. error rate minimization. *Advances in Neural Information Processing Systems*, 16:313–320, 2003.

Gang Dong, Kin Keung Lai, ja Jerome Yen. Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science*, 1(1):2463–2468, 2010.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8): 861–874, 2006.

The World Bank Group. Credit scoring approaches guidelines. *World Bank*, 2019.

Joseph M Hilbe. *Logistic Regression Models*. Chapman and hall/CRC, 2009.

David W Hosmer Jr, Stanley Lemeshow, ja Rodney X Sturdivant. *Applied Logistic Regression*, volume 398. John Wiley & Sons, 2013.

Shantha Kumari. Multicollinearity: Estimation and elimination. *Journal of Contemporary research in Management*, 3(1):87–95, 2008.

Ljiljanka Kvesić ja Gordana Dukić. Risk management and business credit scoring. Teoksessa *Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces*. IEEE, 2012.

Yazhe Li, Tony Bellotti, ja Niall Adams. Issues using logistic regression with class imbalance, with a case study from credit risk modelling. *Foundations of Data Science*, 1(4):389, 2019.

Henrik Madsen ja Poul Thyregod. *Introduction to General and Generalized Linear Models*. CRC Press, 2010.

Peter McCullagh ja John A Nelder. *Generalized Linear Models II*. Chapman and Hall, 1989.

Andrew Schworer ja Peter Hovey. *Newton-Raphson versus Fisher scoring algorithms in calculating maximum likelihood estimates*. Undergraduate Mathematics Day, Electronic Proceedings, 2004.

Lyn C Thomas. *Consumer Credit Models: Pricing, Profit and Portfolios*. OUP Oxford, 2009.

Olaf Weber. Environmental credit risk management in banks and financial service institutions. *Business Strategy and the Environment*, 21(4):248–263, 2012.

Sanford Weisberg. *Applied Linear Regression*. Wiley, 2014.