Poistuman tutkiminen logistisella regressiolla

Lauri Suoknuuti

Perustieteiden korkeakoulu

Kandidaatintyö
Espoo 7.7.2020

Vastuuopettaja

Apul.prof. Pauliina Ilmonen

Työn ohjaaja

FM Markus Linnakaari

Aalto-yliopisto
Perustieteiden
korkeakoulu

Aalto-yliopisto
Perustieteiden korkeakoulu

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

Tekijä Lauri Suoknuuti

Työn nimi Poistuman tutkiminen logistisella regressiolla

Koulutusohjelma Teknillinen fysiikka ja matematiikka

Pääaine Matematiikka ja systeemitieteet **Pääaineen koodi** SCI3029

Vastuuopettaja Apul.prof. Pauliina Ilmonen

Työn ohjaaja FM Markus Linnakaari

Päivämäärä 7.7.2020 **Sivumäärä** 25 **Kieli** Suomi

Tiivistelmä

Asiakaspoistumalla tarkoitetaan asiakkaan toimesta tapahtuvaa asioinnin lopettamista yrityksen kanssa. Tässä opinnäytetyössä tutkitaan asiakaspoistumaa logistisen regression avulla. Tämä opinnäytetyö keskittyy asiakaspoistumaan vakuutusyhtiöiden autoliikkeille tarjottavien kampanjatarjousten yhteydessä. Tyypillisesti kampanjatarjouksen päätyttyä vakuutuksen hinta nousee ja suhteellisen suuri osa asiakkaista irtisanoo vakuutuksensa. Tämän työn tavoitteena on rakentaa logistiseen regressioon pohjautuva ennustemalli, jonka avulla uusista asiakkaista pyritään tunnistamaan poistuvat.

Työssä käytetty aineisto koostuu suomalaisen vakuutusyhtiön autoliikkeille tarjottavien kampanjatarjousten asiakasdatasta vuosilta 2017 - 2019. Aineistoa on käytössä noin 30 000:n tapauksen verran. Malliin valitaan 29 selittävää muuttujaa, joista osa on jatkuvia tai kokonaislukuarvoisia ja osa kaksiarvoisia. Mallin vastemuuttujana on kaksiarvoinen asiakkaan poistumaa kuvaava muuttuja.

Työssä muodostetaan logistinen regressiomalli, jonka toimintaa arvioidaan. Mallin ennustekykyä tarkastellaan erilaisten metriikkojen avulla. Työssä havaitaan, että luotu malli kykenee ennustamaan uuden asiakkaan poistuman kohtalaisella tarkkuudella. Lisäksi työssä tunnistetaan ne selittävät muuttujat, jotka vaikuttavat eniten asiakkaan poistuman todennäköisyyteen. Työn lopussa käydään myös läpi, että miten kyseistä mallia voisi vielä jatkokehittää.

Avainsanat logistinen regressio, asiakaspoistuma, luokittelu, vakuutusyhtiö, koneoppiminen

Aalto-yliopisto
Perustieteiden
korkeakoulu

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

Author Lauri Suoknuuti

Title Predicting customer churn using logistic regression

Degree programme Engineering Physics and Mathematics

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

Teacher in charge Asso.prof. Pauliina Ilmonen

Advisor M.Sc. Markus Linnakaari

Date 7.7.2020 **Number of pages** 25 **Language** Finnish

Abstract

Customer churn is said to happen when an existing customer or a client stops doing business with a company. This thesis addresses customer churn using logistic regression. In particular, this thesis focuses on the customer churn associated with an insurance company's discounted car insurances offered to car dealerships. Usually as the discount ends, the price paid for the insurance by the customer increases. This results in a relatively large proportion of customers terminating their insurance after the discount ends. The goal of this thesis is to construct a predictive model based on logistic regression which is able to recognize the customers who are likely to churn.

The data used in this thesis consists of a Finnish insurance company's customer data from 2017 to 2019. The data contains approximately 30 000 instances. 29 explanatory variables are selected to be included in the model, from which some are continuous or ordinal variables and some are binary. The response variable of the model is a binary variable, which represents the churn of a particular customer.

In this thesis, a logistic regression model is formed and its performance is assessed. The model's predictive ability is evaluated using different metrics. The constructed model is observed to be able to predict the churn of a new customer with moderate accuracy. Additionally, the most impactful variables affecting the customer's probability to churn are identified. At the end of the thesis, future prospects for developing the model are discussed.

Keywords logistic regression, customer churn, classification, insurance company, machine learning

Sisältö

1 Johdanto

Asiakaspoistuma aiheuttaa monille yrityksille merkittäviä kustannuksia vuosittain. Kustannuksia aiheutuu, sillä uusien asiakkaiden hankkiminen on tyypillisesti kalliimpaa kuin vanhojen asiakkaiden pitäminen. Poistumalla tarkoitetaan tässä yhteydessä asiakkaan toimesta tapahtuvaa asioinnin lopettamista yrityksen kanssa. Poistumaa vähentämällä ja hallitsemalla yritys voi vähentää sen kuluja huomattavasti.

Poistuman hallinta on yrityksille tärkeä, mutta haastava ongelma. Poistuma on tyypillisesti suhteellisen harvinaista yrityksen koko asiakaskuntaan suhteutettuna. Tämän takia poistumaa käsittelevässä aineistossa poistuvien asiakkaiden osuus on huomattavasti pienempi kuin ei-poistuvien asiakkaiden osuus. Tämä aineiston ominaispiirre vaikeuttaa ongelmaa kuvaavien mallien muodostamista. Poistuma ja sen mallintaminen on runsaasti tutkittu matemaattinen ongelma. Erilaiset koneoppimismallit ovat tyypillinen tapa mallintaa ja pyrkiä ennustamaan poistumaa.

Tämä opinnäytetyö käsittelee poistumaa vakuutusyhtiön autoliikkeille tarjottavien kampanjatarjousten yhteydessä. Kampanjatarjous tarkoittaa vakuutusyhtiön tarjoamaa edullisempaa hintaa vakuutukselle tietyksi määräajaksi. Poistumalla tarkoitetaan tässä tapauksessa vakuutuksen irtisanomista asiakkaan toimesta kampanjatarjouksen päätyttyä. Poistumaa tarkastellaan tässä työssä logistisen regressioanalyysin menetelmin. Työssä perehdytään logistiseen regressioon menetelmänä ja tarkastellaan sen soveltuvuutta asiakaspoistuman ennustamiseen. Työn tavoitteena on rakentaa logistiseen regressioon perustuva koneoppimismalli, jonka avulla pyritään ennustamaan tulevien asiakkaiden poistumaa.

2 Asiakkaiden poistuma

Asiakkaiden poistumalla, englanniksi *customer churn*, tarkoitetaan asiakkaan toimesta tapahtuvaa asioinnin lopettamista yrityksen kanssa (Xie et al., 2009). Asiakkaiden poistuman ymmärtäminen ja rajoittaminen on monille yrityksille hyvin tärkeää, sillä uusien asiakkaiden hankkiminen on yrityksille kalliimpaa kuin olemassaolevien asiakkaiden pitäminen. On arvioitu, että uuden asiakkaan hankkiminen maksaisi yritykselle jopa 5-6 kertaa enemmän, kuin vanhan asiakkaan pitäminen (Bhattacharya, 1998).

Asiakkaiden poistuman ennustaminen on laajalti tutkittu ongelma. Poistuman ennustamiseen on käytetty lukuisia eri koneoppimisen menetelmiä, joista tyypillisiä ovat logistinen regressio, päätöspuut ja neuroverkot (Neslin et al., 2006). Datan perusteella pyritään luomaan malli, joka luokittelee uudet asiakkaat poistuviin ja ei-poistuviin. Luokitteluongelmana poistuman mallintaminen on haastavaa, sillä tämän ilmiön datassa luokat ovat tyypillisesti hyvin epätasapainossa. Tällä siis tarkoitetaan sitä, että poistuvien osuus datasta on huomattavasti pienempi kuin ei-poistuvien osuus. Epätasapainoinen data voi aiheuttaa ongelmia mallin kouluttamisessa. Tyypillinen ongelma on, että malli luokittelee lähes kaikki uudet tapaukset yleisempään luokkaan. Mallia validoitaessa voidaan havaita hyvin korkea ennusteiden tarkkuus, englanniksi *accuracy*, joka kertoo oikeaan luokkaan luokiteltujen suhteen kaikkiin luokitteluihin.

Tämä mittari on kuitenkin tällaisessa tilanteessa harhaanjohtava. Esimerkiksi jos datassa poistuman suhteellinen osuus olisi 0,2%, malli joka luokittelee kaikki tapaukset ei-poistuvien luokkaan saavuttaisi 99,8% tarkkuuden (Kubat et al., 1997). Tällaisesta mallista ei olisi juurikaan hyötyä ennustamisen kannalta. Epätasapainoisen datan perusteella luotuja malleja tuleekin arvioida muilla metriikoilla, kuten positiivisella ennustearvolla ja sensitiivisyydellä, englanniksi *precision* ja *sensitivity/recall* (Powers, 2011).

Epätasapainoisen datan mallintamiseen on kehitetty useita erilaisia ratkaisuja. Suhteellisen yksinkertaisia menetelmiä ovat erilaiset otannat, esimerkiksi yliotanta pienemmästä luokasta (*over-sampling*) tai aliotanta suuremmasta luokasta *(under-sampling*). Molemmat näistä menetelmistä vähentävät aineiston epätasapainoisuutta. Näillä menetelmillä on kuitenkin myös mahdollisia haittavaikutuksia. Aliotanta saattaa jättää käyttämättä mahdollisesti hyödyllisiä yleisemmän luokan esimerkkejä ja täten heikentää mallin luokittelukykyä. Yliotanta taas saattaa aiheuttaa ylisovittamista, sillä yliotannassa yleensä kopioidaan harvinaisemman luokan tapauksia. Täten malli saattaa oppia myös datassa olevan kohinan, mikä heikentää mallin luokittelukykyä. (Weiss, 2004)

Logistinen regressio on suosittu työkalu poistuman mallintamisessa. Yhtenä syynä logistisen regression suosioon on sen yksinkertainen ja hyvin tunnettu toimintaperiaate. Verrattuna niin sanottuihin mustan laatikon malleihin, kuten neuroverkkoihin, logistisen regression parametrit ovat selkeästi tulkittavissa (Burez ja Van den Poel, 2009). Logistisen regressiomallin parametreista on siis helppo nähdä mikä on eri tekijöiden vaikutus luokitteluun ja luokittelua on helpompi perustella. Logistisen regressiomallin on myös näytetty saavuttavan hyviä ja robusteja tuloksia asiakaspoistuman mallinnuksessa (Neslin et al., 2006).

3 Menetelmät

3.1 Logistinen regressio

Regressioanalyysi on tilastollinen työkalu, jonka tavoitteena on selvittää mikä on selitettävän muuttujan eli vasteen ja selittävien muuttujien välinen riippuvuus (Sykes, 1993). Tyypillisin regressioanalyysin muoto on lineaarinen regressio, jossa vasteen ja selittävien muuttujien välille muodostetaan lineaarinen riippuvuussuhde. Lineaarisen regressiomallin oletuksiin kuuluu, että vastemuuttuja on jatkuva. Logistinen regressio eroaa lineaarisesta regressiosta erityisesti vasteen saamien arvojen takia, sillä tavallisessa logistisessa regressiossa vaste on binäärinen eli sille mahdollisia arvoja ovat vain 1 ja 0 (Hosmer Jr et al., 2013). Yleinen käytäntö on, että arvo 1 tarkoittaa tutkittavaan asian tapahtumista ja arvo 0 tarkoittaa, että tutkittavaa asiaa ei tapahdu. Tästä syystä logistista regressiota voidaan käyttää binäärisessä luokittelussa, jossa havainto halutaan luokitella toiseen kahdesta luokasta.

Lineaarinen malli kuvaa ei-satunnaisten ja havaittujen selittävien muuttujien x_{ij} ja satunnaisen ja havaitun vasteen y_i välistä lineaarista riippuvuutta. Lineaarinen

malli on muotoa

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \qquad i = 1, ..., n$$
$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i,$$

jossa β_0 on vakion ei-satunnainen regressiokerroin, β_j on selittävän muuttujan x_{ij} ei-satunnainen regressiokerroin ja ϵ_i on virhetermin satunnainen arvo. Yksi lineaarisen mallin standardioletuksista on systemaattisen virheen puuttuminen eli, että virhetermin ϵ_i odotustarvo $E(\epsilon_i)$ on nolla. Tästä johtuen lineaarisella mallilla voidaan ennustaa vasteen y_i odotusarvoa $E(y_i)$ seuraavalla tavalla

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

(Yan ja Su, 2009) Kuten aiemmin mainittiin, lineaarisen mallin oletuksiin kuuluu, että vastemuuttujan on oltava jatkuva. Joissain tapauksissa vasteen saama arvo halutaan rajata tietylle välille, jolloin tavallista lineaarista mallia ei voida käyttää. Esimerkiksi Bernoullin jakaumaa noudattavan binäärimuuttujan odotusarvo saa arvoja ainoastaan suljetulta väliltä [0,1]. Tällöin voimme käyttää esimerkiksi logistista regressiomallia. Logistinen ja lineaarinen regressio kuuluvat molemmat yleistettyihin lineaarisiin malleihin, englanniksi *generalized linear models* (GLM) (McCullagh, 2018). Yleistetty lineaarinen malli on muotoa

$$g(E(y_i)) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \tag{2}$$

jossa $g(\cdot)$ on niin sanottu linkkifunktio ja merkinnät ovat muuten samat kuin lineaarisessa mallissa 1. Linkkifunktion tarkoituksena on yhdistää lineaarinen prediktori, eli yhtälön 2 oikea puoli, vastemuuttujan y_i odotusarvoon. Logistisen regression tapauksessa vasteen y_i odotusarvo $E(y_i)$ saa arvoja ainoastaan suljetulta väliltä [0,1], joten haluamamme linkkifunktion määrittelyjoukon tulee olla väli [0,1] ja arvojoukon koko reaalilukujen joukko $\mathbb{R}$. Logistisessa regressiossa käytetään niin sanottua *logit*-linkkiä, jota tarkastellaan myöhemmin. Muita eri tarkoituksissa käytettyjä linkkifunktioita ovat esimerkiksi normaalijakauman kertymäfunktio eli *probit*-linkki ja käänteislinkki. (McCullagh, 2018)

Logistisen regression tapauksessa tapahtuman y_i odotusarvoa, annettuna selittävä muuttuja $\mathbf{x}_i$, merkitään usein merkinnällä

$$\pi(\mathbf{x}_i) = E(y_i|\mathbf{x}_i).$$

Logistisessa regressiossa odotusarvo $\pi(\mathbf{x}_i)$ kuvaa ehdollista todennäköisyyttä, jolla vaste y_i, annettuna selittävä muuttuja $\mathbf{x}_i$, saa arvon 1. Toisin sanoen

$$\pi(\mathbf{x}_i) = E(y_i|\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i). \tag{3}$$

Koska logistisessa regressiossa käytetty vastemuuttuja on Bernoulli-jakautunut, vaste saa arvon 0 todennäköisyydellä

$$P(y_i = 0|\mathbf{x}_i) = 1 - \pi(\mathbf{x}_i).$$

Logit-muunnos saadaan ottamalla luonnollinen logaritmi vastasuhteesta, englanniksi *odds*:ista (Hilbe, 2009). Vastasuhteella tarkoitetaan tapahtuman tapahtumisen ja sen tapahtumatta jäämisen suhdetta. Formaalisti siis

$$odds = \frac{P(Y = 1)}{1 - P(Y = 1)}. \tag{4}$$

Kun yhtälöstä 4 otetaan luonnollinen logaritmi ja sijoitetaan todennäköisyyksien tilalle yhtälössä 3 esitetty notaatio, saavutaan seuraavaan muotoon

$$logit(\pi(\mathbf{x}_i)) = \ln\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right). \tag{5}$$

Logit-muunnos voidaan nyt sijoittaa yhtälön 2 yleistettyyn lineaariseen malliin linkkifunktion $g(\cdot)$ tilalle, jolloin malli voidaan kirjoittaa muodossa

$$\ln\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{6}$$

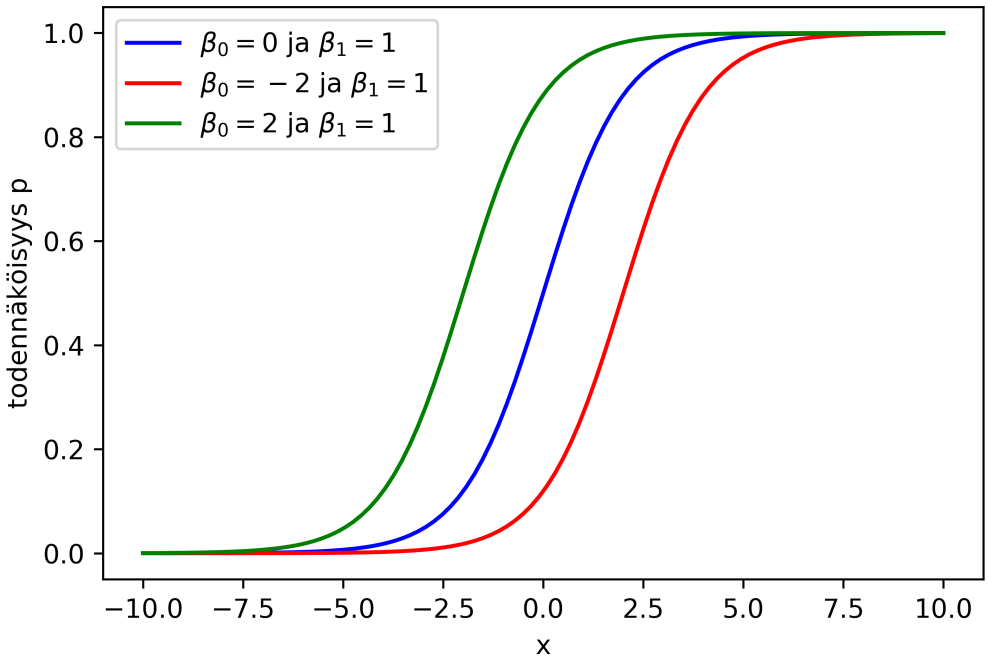
Logit-muunnos on siis lineaarinen selitettävien muuttujien $\mathbf{x}_i$ suhteen. Yhtälöstä 6 on mahdollista ratkaista todennäköisyys $\pi(\mathbf{x}_i)$ seuraavalla tavalla

$$\begin{aligned}
\ln\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right) &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \\
\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)} &= \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}) \\
\pi(\mathbf{x}_i)(1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})) &= \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})) \\
\pi(\mathbf{x}_i) = \frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})} &= \frac{1}{1 + \exp(-(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}))},
\end{aligned} \tag{7}$$

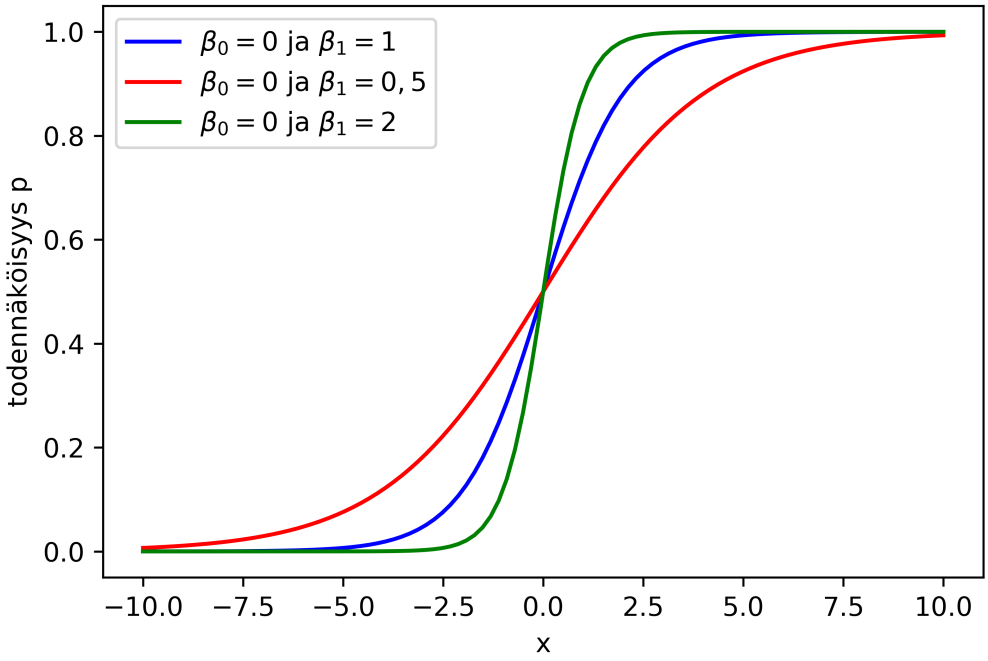
jossa $\exp(x) = e^x$. Lopputuloksena saadaan niin sanottu logistinen funktio, joka yhden selittävän muuttujan tapauksessa yksinkertaistuu muotoon

$$\pi(\mathbf{x}_i) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_i)}}. \tag{8}$$

Kaavan 8 logistinen funktio on kuvattu kuvaajissa 1 ja 2 parametrien β_0 ja β_1 eri arvoilla. Kuten kuvaajasta nähdään, logistinen funktio saa arvoja väliltä [0,1]. Tästä syystä se sopii käytettäväksi todennäköisyyksiä mallintaessa. Logistinen funktio muuttaa sille annetun syötteen todennäköisyydeksi; positiivisista luvuista seuraa suuri todennäköisyys ja negatiivisista luvuista pieni todennäköisyys. Kuvaajista huomataan, että parametrin β_0 muokkaaminen liikuttaa käyrää pitkin x-akselia, ja parametrin β_1 muokkaaminen muuttaa käyrän jyrkkyyttä. Logistinen funktio on niin sanottu sigmoid-funktio, jolla on tyypillinen S-kirjainta muistuttava kuvaaja.



Kuva 1: Logistinen funktio kuvattuna parametrin β_0 eri arvoilla.



Kuva 2: Logistinen funktio kuvattuna parametrin β_1 eri arvoilla.

3.2 Parametrien estimointi

Lineaarisen regression tapauksessa parametrien β_0 ja β_j estimointi tapahtuu tyypillisesti pienimmän neliösumman menetelmällä. Parametrien estimaatit $\hat{\beta}_0$ ja $\hat{\beta}_j$ valitaan siten, että ne minimoivat residuaalien neliösumman. Pienimmän neliösumman menetelmällä parametrien estimaattien arvot on mahdollista ratkaista analyyttisesti. Tosiasiassa pienimmän neliösumman menetelmä tuottaa suurimman uskottavuuden estimaatit parametreille, jos voidaan olettaa mallin residuaalit normaalijakautuneiksi (Kuhn ja Johnson, 2013). Suurimman uskottavuuden estimointi, englanniksi *maximum likelihood estimation* (MLE), on menetelmä, jota voidaan käyttää mallin parametrien estimointiin, jos datan todennäköisyysjakaumasta voidaan tehdä oletuksia. Mallin parametreiksi valitaan ne, joilla uskottavuusfunktio saa maksimiarvonsa. Käytännössä tämä tarkoittaa sitä, että näillä parametreilla otoksen havaitsemisen todennäköisyys on suurinta. Parametrien estimointia kutsutaan koneoppimisessa usein mallin kouluttamiseksi.

Määritellään seuraavaksi logistisessa regressiossa käytetty uskottavuusfunktio. Kuten aiemmin mainittiin, vastemuuttujan y_i oletetaan noudattavan Bernoullin jakaumaa todennäköisyydellä $\pi(\mathbf{x}_i)$, eli $y_i \sim B(\pi(\mathbf{x}_i))$. Täten $P(y_i = 1) = \pi(\mathbf{x}_i)$ ja $P(y_i = 0) = 1 - \pi(\mathbf{x}_i)$. Yksittäiseen havaintoon liittyvä todennäköisyys saadaan Bernoullin jakauman pistetodennäköisyysfunktiosta

$$f(y_i; \pi(\mathbf{x}_i)) = \pi(\mathbf{x}_i)^{y_i}(1 - \pi(\mathbf{x}_i))^{1-y_i}.$$

Logistisessa regressiossa oletetaan myös, että havainnot ovat toisistaan riippumattomia ja identtisesti jakautuneita. Tästä oletuksesta johtuen uskottavuusfunktio l voidaan esittää yksittäisten havaintojen todennäköisyyksien tulona (Dobson ja Barnett, 2018)

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}(1-\pi(\mathbf{x}_i))^{1-y_i}. \tag{9}$$

Matemaattisesti on helpompi käsitellä uskottavuusfunktion logaritmia. Logaritmifunktio on aidosti kasvava, joten logaritminen uskottavuusfunktio saa maksimiarvonsa samassa pisteessä kuin alkuperäinen uskottavuusfunktio. Ottamalla funktiosta 9 logaritmi, saavutaan logaritmiseen uskottavuusfunktioon

$$\begin{aligned} L(\boldsymbol{\beta}) = \ln(l(\boldsymbol{\beta})) &= \ln\left(\prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}(1-\pi(\mathbf{x}_i))^{1-y_i}\right) \\ L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \ln\left(\pi(\mathbf{x}_i)^{y_i}(1-\pi(\mathbf{x}_i))^{1-y_i}\right) \\ L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \ln(\pi(\mathbf{x}_i)) + (1-y_i)\ln(1-\pi(\mathbf{x}_i)) \end{aligned} \tag{10}$$

(Kutner et al., 2005). Sijoittamalla funktioon 10 todennäköisyyden $\pi(\mathbf{x}_i)$ tilalle kaavan 7 logistinen funktio saadaan seuraava muoto

$$\begin{aligned} L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln(&\frac{1}{1+\exp(-(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}))}) \\ &+ (1-y_i)\ln(1-\frac{1}{1+\exp(-(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}))}). \end{aligned} \tag{11}$$

Toiston välttämiseksi ja merkintöjen yksinkertaistamiseksi käytetään notaatiota $z_i = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}$. Uskottavuusfunktio 11 maksimoidaan derivoimalla funktio β_0:n ja β_j:n suhteen ja etsimällä tästä seuraavien derivaattafunktioiden nollakohdat. Derivoidaan funktio ensin β_0:n suhteen

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n}\left(y_i \frac{\partial}{\partial \beta_0}\ln(\frac{1}{1+e^{-z_i}}) + (1-y_i)\frac{\partial}{\partial \beta_0}(1-\frac{1}{1+e^{-z_i}})\right) \\ &= \sum_{i=1}^{n}\left(y_i \frac{e^{-z_i}}{1+e^{-z_i}} + (1-y_i)\left(-1+\frac{e^{-z_i}}{1+e^{-z_i}}\right)\right) \\ &= \sum_{i=1}^{n}\left(y_i - \frac{1}{1+e^{-z_i}}\right) \\ &= \sum_{i=1}^{n}\left(y_i - \pi(\mathbf{x}_i)\right). \end{aligned} \tag{12}$$

Seuraavaksi derivoidaan β_j:n suhteen

$$\begin{aligned}\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n}\left(y_i \frac{\partial}{\partial \beta_j}\ln(\frac{1}{1+e^{-z_i}}) + (1-y_i)\frac{\partial}{\partial \beta_j}(1-\frac{1}{1+e^{-z_i}})\right) \\ &= \sum_{i=1}^{n}\left(y_i x_{ij}\frac{e^{-z_i}}{1+e^{-z_i}} + (1-y_i)x_{ij}\left(-1+\frac{e^{-z_i}}{1+e^{-z_i}}\right)\right) \\ &= \sum_{i=1}^{n} x_{ij}\left(y_i - \frac{1}{1+e^{-z_i}}\right) \\ &= \sum_{i=1}^{n} x_{ij}\left(y_i - \pi(\mathbf{x}_i)\right). \end{aligned} \tag{13}$$

Estimaatit parametreille saadaan asettamalla nämä $p+1$ kappaletta yhtälöitä nollaksi ja ratkaisemalla estimaattien $\hat{\beta}_0$ ja $\hat{\beta}_j$ arvot. Nämä estimaatit maksimoivat uskottavuusfunktion 9. Toisin kuin lineaarisessa regressiossa, nämä yhtälöt eivät ole lineaarisia, eikä niille voida esittää analyyttistä ratkaisua. Yhtälöiden ratkaisuun käytetään erilaisia iteratiivisia menetelmiä, jotka on toteutettu valmiiksi lähes kaikkiin tilastollisiin työkaluihin. Työssä käytetyn scikit-learn kirjaston logistisen regression pakettiin (Pedregosa et al., 2011) on toteutettu useita eri optimointialgoritmeja, kuten Newtonin menetelmä ja gradienttimenetelmä (Battiti, 1992).

Gradienttimenetelmä on iteratiivinen optimointialgoritmi, jolla pyritään minimoimaan tai maksimoimaan jokin funktio. Logistisen regression tapauksessa maksimoitava funktio on kaavan 11 logaritminen uskottavuusfunktio. Maksimoitaessa funktio f gradienttimenetelmällä, aloitetaan pisteestä x_0 ja iteroidaan kaavan

$$x_{k+1} = x_k + \alpha \nabla f(x_k)$$

mukaisesti, kunnes gradientti $\nabla f(x_k)$ on nolla. α on menetelmän askelpituus, joka voidaan asettaa vakioksi tai joissain tapauksissa optimoida jokaisella iteraatioaskeleella erikseen. Iteraatio on päätynyt funktion optimipisteeseen, kun gradientti saa arvon 0. Kaavan 11 uskottavuusfunktio on konkaavi ($-L(\boldsymbol{\beta})$ on konveksi), joten gradienttimenetelmällä saavutettu maksimi on globaali maksimi (Roux et al., 2012).

3.3 Regularisaatio

Ylisovittamisella tarkoitetaan ilmiötä, jossa koneoppimismalli oppii koulutusdatassa olevat ominaisuudet liian tarkasti. Ylisovitettu malli saattaa oppia myös koulutusdatassa olevan kohinan, jolloin mallin ennustekyky uudelle datalle heikkenee (Tetko et al., 1995). Regularisaatio on ylimääräisen tiedon lisäämistä tarkoittava menetelmä, jonka avulla pyritään ehkäisemään ylisovittamista. Regularisaatiota käytettäessä uskottavuusfunktiosta vähennetään regularisaatiotermi $R(\boldsymbol{\beta})$, joka rankaisee mallia parametrien β_j suurista arvoista. Regularisoidun logistisen regression logaritminen uskottavuusfunktio voidaan esittää muodossa

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \ln(\pi(\mathbf{x}_i)) + (1-y_i)\ln(1-\pi(\mathbf{x}_i)) - \lambda R(\boldsymbol{\beta}), \tag{14}$$

joka on kaavan 10 logaritminen uskottavuusfunktio vähennettynä regularisaatiotermillä $R(\boldsymbol{\beta})$. Termi λ on regularisaation voimakkuutta säätelevä vakio, joka on arvoltaan ≥ 0. Suuremmat λ:n arvot lisäävät regularisaation voimakkuutta. Vastaavasti pienemmät λ:n arvot vähentävät regularisaation voimakkuutta. Asettamalla λ:n arvoksi 0, regularisaation vaikutus poistuu ja saadaan kaavan 10 mukainen uskottavuusfunktio.

Eri regularisaatiomenetelmiä on useita, mutta tämän työn kannalta oleellisimpia ovat L1 ja L2 regularisaatiot. L1 ja L2 regularisaatiot eroavat toisistaan regularisaatiotermin $R(\boldsymbol{\beta})$ osalta. L1 regularisaatiossa regularisaatioterminä käytetään parametrivektorin $\boldsymbol{\beta}$ 1-normia, eli $R(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\boldsymbol{\beta}_j|$. L2 regularisaatiossa taas käytetään euklidisen eli 2-normin neliötä, siis $R(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 = \sum_{j=1}^{p} \boldsymbol{\beta}_j^2$. Molemmilla menetelmillä regularisaatiotermi $R(\boldsymbol{\beta})$ saa vain epänegatiivisia arvoja, joten kaavan 14 uskottavuusfunktion arvo pienenee regularisaation seurauksena.

L1 regularisaation on havaittu monissa tilanteissa saattavan osan parametreista β_j nolliksi. Tästä johtuen L1 regularisaatiota voidaan käyttää mallin muuttujien valinnassa. L2 regularisaatiossa tämä käytös ei ole yleistä. L2 regularisaatiota voidaan käyttää ylisovittamisen ehkäisemiseen, ilman että välttämättä halutaan rajoittaa käytettyjen muuttujien määrää.
(Ng, 2004)

4 Tulokset

4.1 Aineisto

Työssä käytetty aineisto koostuu suomalaisen vakuutusyhtiön autoliikkeille tarjottavien kampanjatarjousten asiakasdatasta vuosilta 2017-2019. Kampanjatarjouksella tarkoitetaan autoliikkestä auton ostavalle tarjottavaa autovakuutusta, joka on tietyn ajan edullisempi normaaliin hintaan verrattuna. Tyypillinen määräaika tämänkaltaiselle tarjoukselle on yksi vuosi. Määräajan jälkeen autovakuutuksen hinta tyypillisesti nousee tarjouksen päätyttyä. Tässä yhteydessä poistuma tapahtuu, jos asiakas irtisanoo vakuutuksensa vuoden kuluessa ostohetkestä.

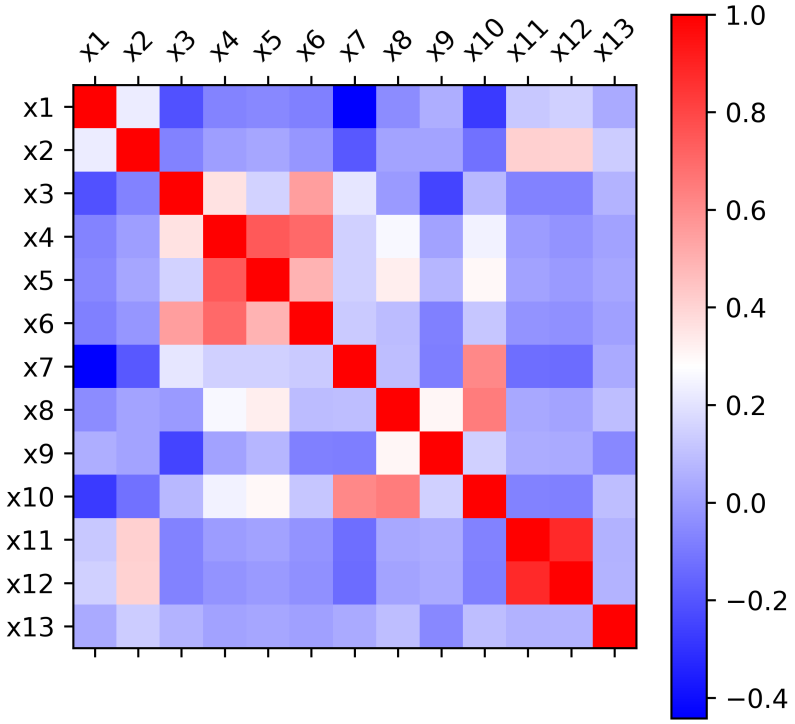
Aineiston selittävinä muuttujina on useita tietoja vakuutuksenottajasta ja vakuutetusta ajoneuvosta. Osa selittävistä muuttujista on kategorisia ja osa on jatkuvia. Aineistossa on myös tietoa vakuutuksenottajan muista vakuutuksista. Aineiston vastemuuttujana on kaksiarvoinen poistumaa kuvaava muuttuja, joka saa arvon 1 jos asiakas on irtisanonut vakuutuksensa vuoden kuluessa ostohetkestä, ja muuten se saa arvon 0. Poistuvien tapausten osuus koko aineistosta on suhteellisen suuri. Täten siis osiossa 2 mainitut epätasapainoisen datan aiheuttamat ongelmat eivät todennäköisesti muodostu mallia rakennettaessa kovin suuriksi.

Malliin valitaan 29 selittävää muuttujaa. Muuttujiin viitataan nimillä $x_1, x_2, ..., x_{29}$. Näistä muuttujat $x_1 - x_{13}$ ovat jatkuvia tai (ordinaalisia) kokonaislukuarvoisia ja loput ovat kaksiarvoisia. Muuttujat $x_{24} - x_{29}$ on muodostettu yhdestä kategorisesta muuttujasta, jolle on 6 mahdollista arvoa. Lisäksi mallin vastemuuttujana on edellä mainittu poistumaa kuvaava kaksiarvoinen muuttuja y. Aineistoa on käytössä noin

30 000:n datapisteen verran.

Kuvassa 3 on kuvattu mallissa käytettyjen jatkuvien ja kokonaislukuarvoisten muuttujien $x_1 - x_{13}$ pareittaiset korrelaatiot. Kuvassa on myös väriasteikko, jonka perusteella kahden muuttujan välisen korrelaation suurutta voidaan arvioida. Kuvasta havaitaan, että korrelaatio on suurinta muuttujien x_{11} ja x_{12} välillä. Pienin korrelaation arvo havaitaan muuttujien x_1 ja x_7 välillä. Samojen muuttujien tilastollisten suureiden arvoja on esitetty taulukossa 1. Taulukosta havaitaan selkeästi, että muuttujien suuruusluokat vaihtelevat runsaasti. Esimerkiksi muuttujan x_{12} keskiarvo on 1,58 ja muuttujan x_4 keskiarvo on 1754,43. Muuttujien suurusluokkien runsas vaihtelu saattaa aiheuttaa ongelmia uskottavuusfunktion 10 maksimoimisessa. Tyypillinen ongelma käytettäessä esimerkiksi osion 3.2 lopussa esiteltyä gradienttimenetelmää on, että gradienttimenetelmä konvergoituu hyvin hitaasti tai ei konvergoidu ollenkaan (Wan, 2019). Muuttujien suuruusluokkien vaihteluun liittyviä ongelmiä pyritään lievittämään skaalaamalla kaikki muuttujat tietyllä tavalla. Työssä käytetyn scikit-learn kirjaston StandardScaler-funktio muuntaa datan siten, että kaikkien muuttujien keskiarvo on 0 ja keskihajonta on 1.

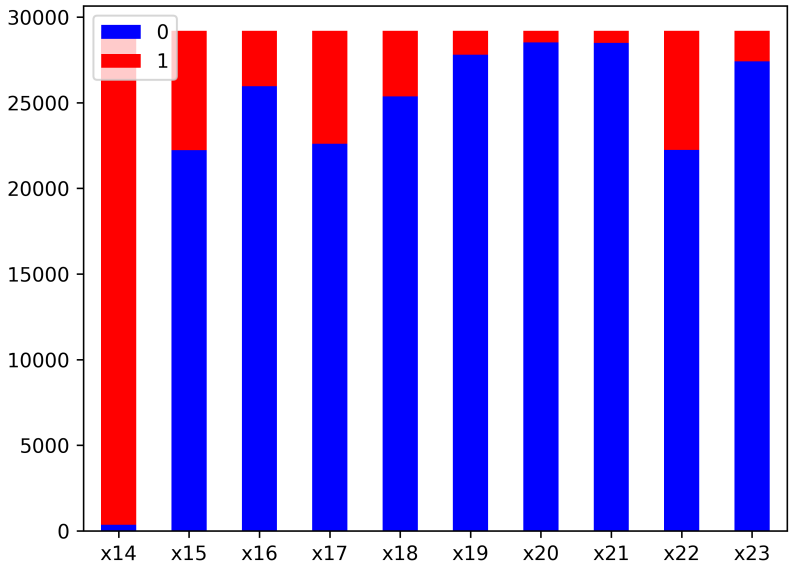
Kuvassa 4 on esitetty mallissa käytettyjen kaksiarvoisten muuttujien jakaumat. Kuvasta havaitaan muuttujan x_{14} jakauman poikkeavan merkittävästi muiden kaksiarvoisten muuttujien jakaumasta. Kuvassa 5 on kuvattuna mallissa käytetyn kategorisen muuttujan jakauma. Kuusiarvoinen kategorinen muuttuja on mallia rakennettaessa muutettu kuudeksi kaksiarvoiseksi muuttujaksi $x_{24} - x_{29}$. Tämä on tyypillinen menettelytapa kategoristen muuttujien kanssa, sillä muuten logistinen regressiomalli ei osaa käsitellä kategorista muuttujaa oikein. Kuvasta 5 havaitaan kategorisen muuttujan koostuvan hyvin suurelta osin muuttujista x_{25} ja x_{26}. Muiden muuttujien esiintymisaste on huomattavasti pienempi.



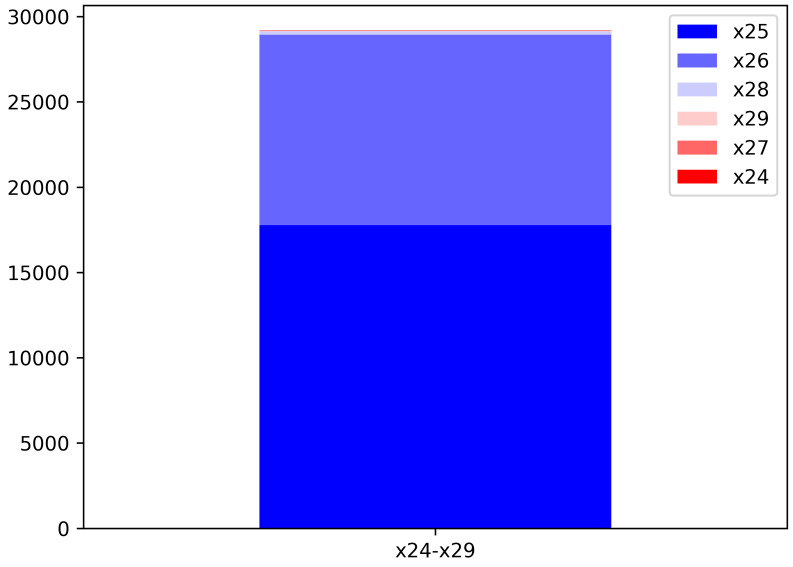
Kuva 3: Mallissa käytettyjen jatkuvien ja kokonaislukuarvoisten muuttujien pareittaiset korrelaatiot.

Taulukko 1: Mallissa käytettyjen jatkuvien ja kokonaislukuarvoisten muuttujien tilastollisten suureiden arvoja.

	keskiarvo	mediaani	keskihajonta	mediaanin absoluuttinen keskipoikkeama	vinous
x_1	44,72	44,00	14,29	11,87	0,34
x_2	2,47	0,00	3,97	3,15	1,49
x_3	5,17	5,00	4,27	3,55	0,61
x_4	1 753,43	1 699,00	609,72	454,28	0,62
x_5	106,90	100,00	37,42	28,31	1,25
x_6	142,01	138,00	42,98	31,17	0,07
x_7	351,35	283,00	200,21	140,67	2,4
x_8	243,90	220,00	110,42	70,75	2,24
x_9	11,36	11,00	1,85	1,34	0,92
x_{10}	569,08	500,00	172,26	129,35	1,7
x_{11}	121,41	0,00	307,91	187,24	4
x_{12}	1,58	0,00	3,33	2,41	2,43
x_{13}	44,86	0,00	137,69	73,82	5,2



Kuva 4: Mallissa käytettyjen kaksiarvoisten muuttujien jakaumat.



Kuva 5: Mallissa käytetyn kategorisen muuttujan jakauma.

4.2 Malli

Mallin rakentaminen aloitetaan muuttujien valinnalla ja datan käsittelyllä. Edellisessä osiossa on esitelty malliin valitut muuttujat. Muuttujat skaalataan edellisessä osiossa mainitun StandardScaler-funktion avulla, jonka jälkeen niiden keskiarvo on 0 ja keskihajonta on 1. Tämän jälkeen data jaetaan satunnaisesti koulutukseen ja validaatioon käytettyihin osiin. Tyypillinen jakosuhde on, että 80% datasta käytetään mallin koulutukseen ja loput 20% käytetään mallin validaatioon ja testaamiseen. Tällöin validaatioon käytetty osio on mallille uutta ja tuntematonta dataa, joten tämän datan avulla voidaan arvioida mallin todellista ennustekykyä. Tätä samaa 80/20 jakosuhdetta käytetään myös tämän työn mallissa.

Seuraavaksi koulutusdataan sovitetaan logistinen regressiomalli scikit-learn kirjaston avulla. Scikit-learnin logistisessa regressiomallissa on muutettavissa useita eri parametreja, joista tämän työn kannalta tärkeimpiä ovat *penalty*, C ja *solver*. Parametrin *penalty* avulla voidaan valita käytetty regularisaatiomenetelmä. Valittavissa on kappaleessa 3.3 esitellyt L1 ja L2 regularisaatiot, sekä näiden yhdistelmä *Elastic Net*-regularisaatio (De Mol et al., 2009). Parametri C määrittää regularisaation voimakkuuden. Parametri C on kappaleessa 3.3 esitellyn λ-parametrin käänteisluku. Parametrilla *solver* voidaan määrittää käytetty ratkaisin uskottavuusfunktio 10 maksimointiin. Valittavissa on useita eri ratkaisimia, joista osa perustuu esimerkiksi kappaleen 3.2 lopussa esiteltyyn gradienttimenetelmään ja osa Newtonin menetelmään.

Mallin koulutuksen jälkeen arvioidaan mallin ennustekykyä. Tämä tapahtuu antamalla koulutetun mallin syötteeksi datan validaatio-osa ja vertaamalla mallin

ennustamia vastemuuttujan arvojen ennusteita $\hat{y}_i$ validaatio-osan todellisiin vasteen arvoihin y_i. Tämän vertailun helpottamiseksi voidaan käyttää työkalua nimeltä luokittelutaulukko, englanniksi *confusion matrix*, joka on esitetty taulukossa 2. Taulukon avulla voidaan laskea mallin ennustekyvyn arvioimiseen käytettyjä metriikoita. Näistä tyypillisimpiä ovat tarkkuus, positiivinen ennustearvo ja sensitiivisyys, englanniksi *accuracy, precision* ja *sensitivity/recall*. Ne lasketaan seuraavilla tavoilla käyttäen luokittelutaulukon merkintöjä:

- Accuracy: $\frac{TP+TN}{\text{ennusteiden määrä}}$
  - Kuinka usein malli luokittelee uuden havainnon oikein
- Precision: $\frac{TP}{TP+FP}$
  - Kuinka usein malli on oikeassa, kun se ennustaa luokan 1
- Recall: $\frac{TP}{TP+FN}$
  - Kuinka usein malli luokittelee luokkaan 1 kuuluvan havainnon oikein.

Taulukko 2: Esimerkki luokittelutaulukosta.

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

Toinen mallin arviointiin käytetty työkalu on niin sanottu erottelukykykäyrä, englanniksi *receiver operating characteristics curve* (ROC curve). Tästä käyrästä mitataan yleensä sen alle jäävän pinta-alan suuruus, englanniksi *area under the curve* (AUC). Erottelukykykäyrä kuvaa mallin sensitiivisyyden ja väärien positiivisten osuuden, englanniksi *false positive raten* (FPR), välistä suhdetta. Sensitiivisyydestä käytetään tässä yhteydessä yleensä nimitystä oikeiden positiivisten osuus, englanniksi *true positive rate* (TPR). Ne määritellään seuraavasti luokittelutaulukon merkintöjä käyttäen:

- TPR: $\frac{TP}{TP+FN}$
- FPR: $\frac{FP}{TN+FP}$.

Nimensä mukaisesti AUC mittaa käyrän alle jäävän pinta-alan suuruutta. Erottelukykykäyrän tapauksessa pinta-alan suuruus voi vaihdella välillä [0,1]. Erottelukykykäyrän AUC:n voidaan tulkita tarkoittavan todennäköisyyttä sille, että malli antaa satunnaisesti valitulle luokkaan 1 kuuluvalle havainnolle suuremman odotusarvon, kuin satunnaisesti valitulle luokan 0 havainnolle. Täten suuret AUC:n arvot kertovat mallin hyvästä erottelukyvystä kahden luokan välillä. AUC:n ollessa 1, malli pystyy erottelemaan havainnot kahteen luokkaan täydellisesti. AUC:n arvolla 0,5 mallin

erottelukyky on olematon, ja samaan tulokseen päästäisiin esimerkiksi kolikkoa heittämällä. AUC:n arvo 0 tarkoittaa sitä, että malli luokittelee luokkaan 0 kuuluvan havainnon aina luokkaan 1 ja vastaavasti luokkaan 1 kuuluvan havainnon aina luokkaan 0.

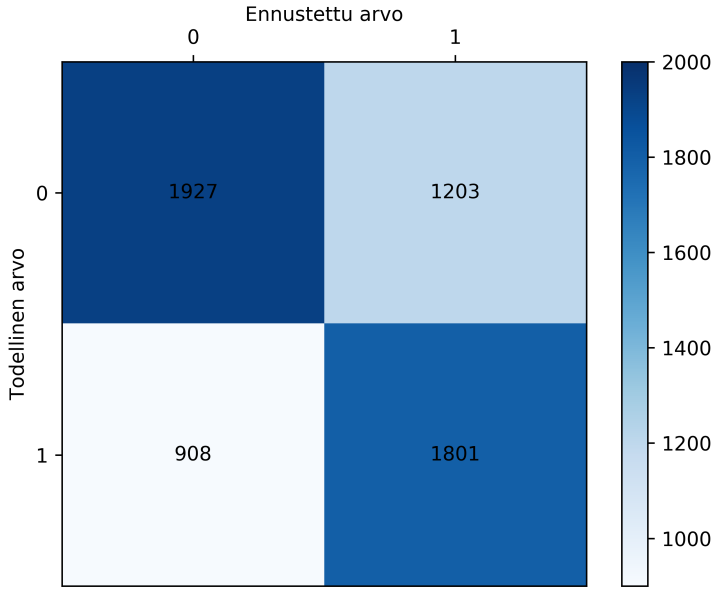
Muodostetaan malli scikit-learnin logistisen regression oletusparametreja käyttäen. Oletusparametreina ovat *penalty*: L2, C: 1,0 ja *solver*: lbfgs. Lbfgs on Newtonin menetelmään perustuva algoritmi, jossa Hessen matriisin sijaan käytetään sen approksimaatiota (Saputro ja Widyaningsih, 2017). Mallin regressiokertoimet ovat liitteissä taulukossa 3. Taulukosta havaitaan regressiokertoimien olevan keskenään samaa suuruusluokkaa. Tämä johtuu aiemmin esitellystä datan skaalaamisesta. Tällä tavoin skaalatun datan perusteella saadut regressiokertoimet ovat vertailukelpoisia, ja itseisarvoltaan suurimmat arvot vaikuttavat luokitteluun eniten. Suurin vaikutus luokitteluun on siis muuttujilla x_2 ja x_8. Koska näiden muuttujien kertoimien arvot ovat negatiiviset, muuttujien x_2 ja x_8 kasvu vähentävää poistuman todennäköisyyttä. Pienin vaikutus luokitteluun on muuttujalla x_{13}, sillä sen regressiokerroin on vain $-0{,}0002$.

Syötetään koulutetulle mallille datan validaatio-osa ja tarkastellaan mallin ennustekykyä. Kuvassa 6 on esitetty mallin perusteella muodostettu luokittelutaulukko. Luokittelutaulukko noudattaa samaa formaattia kuin aiemmin esitelty luokittelutaulukko. Taulukon diagonaalilla on siis onnistuneet luokittelut, ja muut alkiot ovat virheellisiä luokitteluja. Luokittelutaulukon avulla voidaan laskea aiemmin esitellyt mallin ennustekyvystä kertovat metriikat. Niiden arvoiksi saadaan:

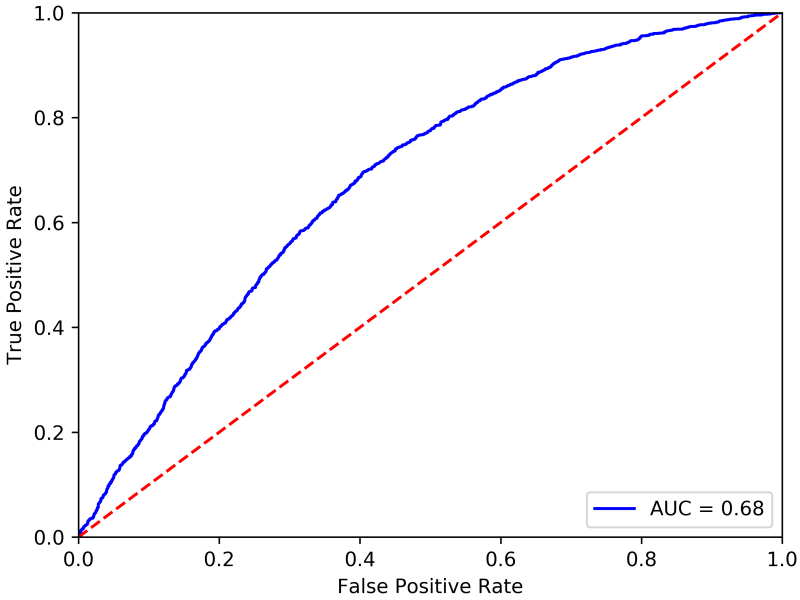
- Accuracy: 0,638
- Precision: 0,600
- Recall: 0,665.

Malli siis luokittelee oikein noin 64% tapauksista. Precision eli positiivinen ennustearvo kertoo, että malli on oikeassa 60% tapauksista, joissa se ennustaa luokan 1 eli asiakkaan poistuman. Recall eli sensitiivisyys taas kertoo, että malli tunnistaa poistuvan asiakkaan n. 67% tarkkuudella. Mallin sensitiivisyys on erityisen tärkeä mallin käyttötarkoitusta ajatellen, sillä poistuvien asiakkaiden tunnistaminen on tälle mallille olennaista. Mallin perusteella laskettujen metriikkojen arvoja voidaan pitää käyttötarkoitus huomioiden kohtuullisina.

Kuvaajassa 7 on kuvattu mallin perusteella muodostettu erottelukykykäyrä. Sininen käyrä kuvaa mallin erottelukykyä eri FPR:n ja TPR:n arvoilla. Kuvaajassa on myös erottelukykykäyrän alle jäävän pinta-alan suurus, jonka arvo on 0,68. Lisäksi punaisella katkoviivalla on merkattu 0,5 AUC:n tasoa. Saatua AUC:n arvoa 0,68 voidaan pitää käyttötarkoitus huomioiden kohtuullisena, mutta todennäköisesti mahdollisuutta parannukseen on vielä olemassa.



Kuva 6: Mallin perusteella muodostettu luokittelutaulukko.



Kuva 7: Mallin perusteella muodostettu erottelukykykäyrä.

5 Yhteenveto

Tässä opinnäytetyössä tutkittiin asiakaspoistumaa logistisen regression avulla. Erityisesti asiakaspoistumaa tutkittiin suomalaisen vakuutusyhtiön kampanjatarjousten yhteydessä. Työn alussa tutustuttiin asiakaspoistumaan yleisesti ja sitä koskevaan aiempaan tutkimukseen. Tämän jälkeen perehdyttiin logistiseen regressioon ja sen keskeisimpiin ominaisuuksiin. Työssä rakennettiin logistiseen regressioon pohjautuva ennustemalli, jonka ennustekykyä ja muita ominaisuuksia arvioitiin.

Työssä käytetty aineisto koostui suomalaisen vakuutusyhtiön autoliikkeille tarjottavien kampanjatarjousten asiakasdatasta. Aineisto oli kerätty vuosilta 2017-2019 ja sitä oli käytössä noin 30 000:n tapauksen verran. Aineistosta valittiin käyttöön 29 selittävää muuttujaa, joista osa oli jatkuvia ja osa kategorisia. Vastemuuttujan aineistossa oli poistumaa kuvaava kaksiarvoinen muuttuja. Kampanjatarjousten keskuudessa asiakkaiden poistuma on suhteellisen yleistä. Poistuman suhteellinen yleisyys helpottaa logistisen regressiomallin muodostamista.

Logistisella regressiomallilla saavutettuja tuloksia arvioitiin työn lopussa. Mallin ennustekykyä tarkasteltiin erilaisia metriikkoja apuna käyttäen. Näistä metriikoista saatujen tulosten perusteella mallin ennustekyvyn arvioitiin olevan kohtalainen. Mallilla saavutettujen ennusteiden kokonaistarkkuus oli noin 64%.

Työssä myös todettiin, että mallin parantaminen olisi vielä todennäköisesti mahdollista. Työssä toteutettu malli muodostettiin käyttämällä scikit-learn ohjelmointikirjaston logistisen regression oletusparametreja. Mallin ennustekykyä voisi vielä todennäköisesti parantaa näitä parametreja muokkaamalla. Voitaisiin esimerkiksi tutkia mallin ennustekykyä regularisaation voimakkuutta säätelevän parametrin C eri arvoilla. Lisäksi voitaisiin vertailla eri regularisaatiomenetelmillä saavutettuja tuloksia. Odotettavissa olisi kuitenkin luultavasti vain maltillisia parannuksia. On todennäköistä, että työssä käytetyllä aineistolla ja menetelmällä ei ole mahdollista saavuttaa merkittävästi parempia tuloksia.

Lisäksi työn asiakaspoistumaa voitaisiin jatkossa mallintaa myös muilla koneoppimismenetelmillä. Eri menetelmiä voitaisiin vertailla keskenään ja valita näistä se, joka suoriutuu parhaiten. Eri malleilla todennäköisesti saavutettaisiin hyvin erilaisia tuloksia. Näistä malleista osa saattaisi sopia tähän ongelmaan paremmin kuin työssä käytetty logistinen regressio.

Viitteet

Roberto Battiti. First-and second-order methods for learning: between steepest descent and newton's method. *Neural computation*, 4(2):141–166, 1992.

CB Bhattacharya. When customers are members: Customer retention in paid membership contexts. *Journal of the academy of marketing science*, 26(1):31–44, 1998.

Jonathan Burez ja Dirk Van den Poel. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, 36(3):4626–4636, 2009.

Christine De Mol, Ernesto De Vito, ja Lorenzo Rosasco. Elastic-net regularization in learning theory. *Journal of Complexity*, 25(2):201–230, 2009.

Annette J Dobson ja Adrian G Barnett. *An introduction to generalized linear models*. CRC press, 2018.

Joseph M Hilbe. *Logistic regression models*. CRC press, 2009.

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

Miroslav Kubat, Stan Matwin, et al. Addressing the curse of imbalanced training sets: one-sided selection. Teoksessa *Icml*, volume 97, pages 179–186. Citeseer, 1997.

Max Kuhn ja Kjell Johnson. *Applied predictive modeling*, volume 26. Springer, 2013.

Michael H Kutner, Christopher J Nachtsheim, John Neter, William Li, et al. *Applied linear statistical models*, volume 5. McGraw-Hill Irwin New York, 2005.

Peter McCullagh. *Generalized linear models*. Routledge, 2018.

Scott A Neslin, Sunil Gupta, Wagner Kamakura, Junxiang Lu, ja Charlotte H Mason. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research*, 43(2):204–211, 2006.

Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. Teoksessa *Proceedings of the twenty-first international conference on Machine learning*, page 78, 2004.

Fabian Pedregosa, Gael. Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, ja Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

David Martin Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2011.

Nicolas L Roux, Mark Schmidt, ja Francis R Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. Teoksessa *Advances in neural information processing systems*, pages 2663–2671, 2012.

Dewi Retno Sari Saputro ja Purnami Widyaningsih. Limited memory broyden-fletcher-goldfarb-shanno (l-bfgs) method for the parameter estimation on geographically weighted ordinal logistic regression model (gwolr). Teoksessa *AIP Conference Proceedings*, volume 1868, page 040009. AIP Publishing LLC, 2017.

Alan O Sykes. An introduction to regression analysis, 1993.

Igor V Tetko, David J Livingstone, ja Alexander I Luik. Neural network studies. 1. comparison of overfitting and overtraining. *Journal of chemical information and computer sciences*, 35(5):826–833, 1995.

Xing Wan. Influence of feature scaling on convergence of gradient iterative algorithm. Teoksessa *Journal of Physics: Conference Series*, volume 1213, page 032021. IOP Publishing, 2019.

Gary M Weiss. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, 6(1):7–19, 2004.

Yaya Xie, Xiu Li, EWT Ngai, ja Weiyun Ying. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3):5445–5449, 2009.

Xin Yan ja Xiaogang Su. *Linear regression analysis: theory and computing*. World Scientific, 2009.

6 Liitteet

Taulukko 3: Mallin regressiokertoimet.

	Kerroin
Vakio	-0,2501
x_1	-0,1139
x_2	-0,3509
x_3	-0,103
x_4	0,0136
x_5	0,0953
x_6	-0,0115
x_7	-0,0315
x_8	-0,3182
x_9	0,1175
x_{10}	0,0308
x_{11}	-0,14
x_{12}	-0,1775
x_{13}	-0,0002
x_{14}	0,1222
x_{15}	-0,035
x_{16}	-0,0663
x_{17}	-0,1189
x_{18}	0,0473
x_{19}	0,0441
x_{20}	0,0215
x_{21}	-0,0516
x_{22}	0,0326
x_{23}	-0,0447
x_{24}	0,0143
x_{25}	-0,0435
x_{26}	0,0394
x_{27}	-0,0151
x_{28}	0,0267
x_{29}	-0,0011