# Teollisuusyritysten konkurssien ennustaminen logistisella regressiolla

Juho Lahti

**Perustieteiden korkeakoulu**

Kandidaatintyö
Espoo 04.06.2024

**Vastuuopettaja**

Prof. Ahti Salo

**Työn ohjaaja**

DI Leevi Olander

Aalto-yliopisto
Perustieteiden
korkeakoulu

**Aalto-yliopisto**
**Perustieteiden korkeakoulu**

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

**Tekijä** Juho Lahti

**Työn nimi** Teollisuusyritysten konkurssien ennustaminen logistisella regressiolla

**Koulutusohjelma** Teknistieteellinen kandidaattiohjelma

**Pääaine** Matematiikka ja systeemitieteet **Pääaineen koodi** SCI3029

**Vastuuopettaja** Prof. Ahti Salo

**Työn ohjaaja** DI Leevi Olander

**Päivämäärä** 04.06.2024 **Sivumäärä** 34 **Kieli** Suomi

**Tiivistelmä**

Yrityskonkurssilla tarkoitetaan menettelyä, jossa maksukyvyttömän yrityksen ulosmitattava omaisuus myydään ja jaetaan velkojien kesken. Tämän työn tavoitteena on luoda yritysten tilinpäätöstietoihin tukeutuva logistinen regressiomalli konkurssiriskien arvioimiseksi. Mallin avulla konkurssiriskiin vaikuttavat tekijät voidaan tunnistaa ja todennäköiset konkurssiyritykset erotella aktiivisten yritysten joukosta.

Työssä käytetty aineisto koostui Euroopan Unionin alueella sijaitsevista teollisista pienistä ja keskisuurista yrityksistä ja niiden tilinpäätöstiedoista. Aineistosta 62 oli konkurssiyrityksiä ja noin 42 000 aktiivisia yrityksiä. Aikaisempaan eri toimialan yrityskonkursseja ennustavaan malliin pohjautuen muuttujia valikoitui malliin 15, joiden pohjalta kehitettiin kaksi erilaista mallia konkurssiriskin ennustamiseksi.

Konkurssiriskiä eniten kasvattavaksi tekijäksi osoittautui käyttöpääoman kasvu taseen loppusummaan nähden. Toisaalta konkurssiriskiä eniten vähentävä tekijä oli kokonaispääoman tuottoprosentin kasvu. Työssä rakennettujen regressiomallien sensitiivisyydet yritysten erottelussa jäivät heikolle tasolle. Aiemmin kirjallisuudessa esitetty toiselle toimialalle rakennettu malli soveltuu heikosti teollisten yritysten konkurssien luokitteluun. Työn lopussa kuvataan ennustamisessa käytetyn mallin kehittämismahdollisuuksia.

**Avainsanat** konkurssi, teollisuusala, logistinen regressio, ennustaminen, koneoppiminen

**Aalto-yliopisto**
**Perustieteiden korkeakoulu**

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

**Author** Juho Lahti

**Title** Predicting industrial company bankruptcies with logistic regression

**Degree programme** Bachelor's Programme in Science and Technology

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

**Teacher in charge** Prof. Ahti Salo

**Advisor** M.Sc. (Tech.) Leevi Olander

**Date** 04.06.2024 **Number of pages** 34 **Language** Finnish

**Abstract**

Corporate bankruptcy refers to the procedure in which the assets of an insolvent company are sold off and distributed among its creditors. The aim of this thesis is to build a logistic regression model based on the financial statements of companies to assess bankruptcy risks. Using this model, factors affecting bankruptcy risk can be identified, and potential bankruptcies can be distinguished from operational companies.

The data used in this thesis consisted of financial statements for industrial small and medium-sized enterprises located in the European Union. The data contained 62 bankrupt companies, and approximately 42,000 operational companies. Based on a previous model predicting bankruptcies in a different industry, 15 variables were selected for the model. Based on the variables two different models were developed for predicting bankruptcy risk.

The factor that increased bankruptcy risk most was found to be the growth of working capital relative to total assets on the balance sheet. Conversely, the factor that most effectively reduced the risk of bankruptcy was the increase in the return on total capital. The sensitivities of the logistic regression models for classification were found to be weak. A model previously proposed in the literature for another industry does not fit well for classifying bankruptcies in industrial companies. Finally, the thesis also examines the possibilities for developing the model further.

**Keywords** bankruptcy, industry sector, logistic regression, prediction, machine learning

# Sisällys

# 1 Johdanto

Yrityksen konkurssilla tarkoitetaan tilaa, jossa yritys on kykenemätön suoriutumaan sille asetetuista maksuvaateista. Konkurssissa realisoidaan yrityksen jäljellä oleva omaisuus ja jaetaan yrityksen lopulliset varat velkojien kesken. Konkurssia voi hakea yritys itse tai velkojat yhdessä, mutta yhteistä näille on yritystoiminnan loppuminen. Konkurssista aiheutuu yleensä suuria tappioita kaikille yrityksen sidosryhmille, kun velkojat ja rahoittajat menettävät omaisuuttaan ja sijoituksiaan, yhteiskunta verotuloja ja yrittäjä menettää mahdollisesti henkilökohtaisia varojaan. Konkurssin ennustaminen onkin tärkeää lainanantajien, sijoittajien ja yrityksen johdon näkökulmasta, jotta konkurssin ennusmerkit voidaan tunnistaa ajoissa ja yrityksen toimintaa kehittää konkurssin välttämiseksi (Laitinen, 1990).

Konkurssien ennustaminen ja niiden luokittelu kasvattivat suosiotaan 1900-luvulla erityisesti 30-luvun laman aikaan (Bellovary et al., 2007). Konkurssia on pyritty aluksi ennustamaan yhden tai useamman muuttujan lineaarisilla malleilla, joiden tulkitseminen on edellyttänyt kriittisen kynnysarvon määrittämistä aineistosta (Laitinen, 1990). Myöhemmin konkurssitutkimuksessa otettiin käyttöön logistinen regressiomalli, joka soveltuu binääriseen luokitteluongelmaan. Yritykset voidaan karkeasti luokitella aktiivisiin yrityksiin ja konkurssiyrityksiin, jonka takia logistista regressiota voidaan konkurssien ennustuksessa hyödyntää (Bellovary et al., 2007). Muuttujina malleissa käytetään yritysten julkisesti saatavilla olevia tilinpäätöstietoja.

Konkurssien ennustaminen on haastavaa yritys- ja toimialakohtaisten erojen vuoksi. Eri toimialat reagoivat talouden suhdanteisiin eri tavoin, joten yleispätevää mallia on hankala luoda. Eri ajanjaksot ja data-aineisto voivat johtaa harhaanjohtaviin malleihin, joiden ennustetarkkuudet vaihtelevat suuresti eri aineistoilla (Grice ja Dugan, 2001).

Tässä työssä tarkastellaan teollisuusyritysten konkurssiriskiä logistisella regressiolla. Tavoitteena on rakentaa yritysten tilinpäätöstietoihin tukeutuva malli ja selvittää, miten konkurssiriskiä kasvattavat tekijät vaikuttavat konkurssin todennäköisyyteen. Aineistona työssä käytetään EU:n teollisuusalan pienten ja keskisuurten, eli pk-

yritysten tilinpäätöstietoja (Moody's, 2024). Edelleen työn tarkoituksena on selvittää, miten aikaisemmin kehitetty EU:n rakennusalan yrityskonkursseja ennustava malli soveltuu EU:n teollisuusyritysten konkurssien ennustamiseen.

Työn rakenne on seuraavanlainen. Luvussa 2 perehdytään aikaisempaan konkurssitutkimukseen sekä logistisen regressioanalyysin käyttöön konkurssien ennustuksessa. Kirjallisuuskatsauksessa tarkastellaan aikaisemmin kehitettyjä malleja, niissä käytettyjä muuttujia sekä tarkastellaan konkurssien ennustukseen liittyviä haasteita. Luvussa 3 tutkitaan tarkemmin logistista regressioanalyysia ja sen soveltamista aineiston luokitteluun. Luvussa 4 kehitetään yrityskonkursseja ennustava malli, jonka tarkkuutta ja ennustekykyä arvioidaan.

# 2 Kirjallisuuskatsaus

## 2.1 Yrityskonkurssien ennustaminen

Yrityskonkurssilla tarkoitetaan tilaa, jossa yritys ei kykene suoriutumaan sille asetetuista maksusitoumuksista, jolloin yrityksen jäljellä oleva varallisuus joudutaan ulosmittaamaan ja jakamaan velkojien kesken. Maksuvaikeuksiin ajautumassa olevan yrityksen johdon tulee tunnistaa mahdollinen kriisitilanne ja pyrkiä toiminnallaan estämään sen jatkuminen. Yrityksen eri sidosryhmät pyrkivät osaltaan arvioimaan yrityksen luottokelpoisuutta, mikä voi vaikuttaa lainatarpeen ja lainaehtojen määrittämiseen. Yrityksen johdolla on tässä tilanteessa etu, että se pääsee käsiksi yrityksen sisäisiin tietoihin ja pystyy seuraamaan tarkasti tilanteen kehittymistä. Ulkoiset sidosryhmät pyrkivät tilinpäätöstietojen avulla ja tilastollisen analyysin keinoin arvioimaan yrityksen luotettavuutta (Laitinen, 1990).

Yrityksen tilanteen ja sen mahdollisen taloudellisen ahdingon ennustaminen kiinnostaa sekä yrityksen johtoa että sidosryhmiä. Lainanantajien ja sijoittajien näkökulmasta on tärkeä ennustaa maksuvalmiuskriisin todennäköisyyttä, sillä se määrittelee sijoituksiin liittyvien maksujen jakautumisen. Jo varhaisissa konkurssitutkimuksissa on painotettu sitä, ettei päätarkoituksena ole ennustaa konkurssin todennäköisyyttä vaan ennemmin arvioida siitä aiheutuvia tappioita (Beaver et al., 2011).

Yrityksen tilinpäätöstietoihin perustuvassa analyysissä tunnuslukuja analysoimalla pyritään selvittämään ja tunnistamaan mahdolliset konkurssiin tai maksuvalmiuskriisiin ajautuvat yritykset. Tilinpäätöksestä saatavien tunnuslukujen käyttö perustuu oletukseen, että käytettyjen tunnuslukujen jakaumat eroavat aktiivisten- ja konkurssiyritysten kesken johdonmukaisesti ja näitä eroavaisuuksia voidaan hyödyntää konkursseja ennustaessa (Laitinen, 1990). Konkurssiennustuksessa yhden tai monen tunnusluvun pohjalta pyritään luomaan malli, jota voidaan käyttää yritysten luokittelussa aktiivisiksi- ja konkurssiyrityksiksi tietyn tarkastelujakson aikana.

Yrityskonkursseja on pyritty ennustamaan jo monen vuosikymmenen ajan ja ennustamisessa käytetyt mallit ja keinot ovat vaihdelleet suuresti (Bellovary et al., 2007). Varhaisimmat tutkimukset tehtiin 1930-luvulla, jolloin pystyttiin osoittamaan, että konkurssiyritysten tunnusluvut olivat huonompia verrattuna vastaaviin aktiivisiin yrityksiin. Beaver (1966) oli yksi varhaisimpia ja tunnetuimpia konkurssitutkimuksen kehittäjiä, joka hyödynsi tunnuslukuanalyysiä konkurssiennustuksen osalta. Aineistossa aktiivisten- ja konkurssiyritysten tilinpäätösten tunnuslukuja vertailtiin vastinparimenettelyllä, jolloin eri kokoluokkien ja toimialojen yritykset olivat tasavertaisesti edustettuina. Yritysten tunnuslukujen keskiarvoja tarkasteltiin viiden vuoden ajalta ennen konkurssia. Luokittelu lineaarisen mallin avulla toteutettiin optimoimalla kynnysarvot siten, että luokitteluvirheet saatiin minimoitua. Konkurssitutkimuksessa virhetyyppi I tarkoittaa konkurssiyrityksen luokittelua aktiiviseksi maksukykyiseksi yritykseksi ja virhetyyppi II aktiivisen yrityksen luokittelua konkurssiyritykseksi. Virhetyypin I aiheuttamat kustannukset voivat olla jopa 7-46-kertaa suuremmat kuin virhetyypin II aiheuttamat kustannukset, joita tutkimuksessa pyrittiin minimoimaan (Laitinen, 1990).

Logistista regressiota konkurssitutkimuksessa hyödynsi myöhemmin Ohlson (1980), jonka tutkimus oli toteutettu otantatutkimuksena, kun aikaisemmat usean muuttujan mallit toteutettiin vastinparimenettelynä. Logistinen regressiomalli erosi aikaisemmista kriittiseen arvoon perustuvista lineaarisista ja diskriminanttianalyysimalleista, sillä se soveltuu paremmin binäärisen luokittelun tekemiseen, jossa konkurssiyrityksiä (1) ja aktiivisia yrityksiä (0) erotellaan toisistaan. Logistisen regressioanalyysin

etuna aikaisempiin tutkimuksiin oli sen yksinkertaisuus ja yleistettävyys myös muille aineistoille, kuin mitä mallin kehittämisessä oli käytetty (Ohlson, 1980).

## 2.2 Konkurssitodennäköisyyden mallintaminen

Tässä työssä keskitytään konkurssitodennäköisyyden ennustamiseen logistisella regressiolla. Logistisen regressiomallin edut konkurssiennustuksessa ovat yksinkertaisuus ja helppo tulkittavuus. Mallin parametreista saadaan määritettyä suoraan todennäköisyys konkurssin tapahtumiselle ennalta määritetyssä aikaikkunassa (Ohlson, 1980). Lineaarinen malli sekä myöhemmin käytetty diskriminanttimalli edellyttävät vertailuarvojen määrittämistä konkurssitodennäköisyyden arvioinnissa (Beaver, 1966). Molemmat mallit perustuvat tiukempiin oletuksiin tilinpäätöksen tunnuslukujen jakaumasta (Ohlson, 1980).

Haasteet konkurssien ennustamisessa ja mallintamisessa liittyvät yritys- ja toimialakohtaisiin eroihin. Konkurssitutkimuksessa tilinpäätöstietojen pohjalta tehdään vahvoja oletuksia yritystoiminnan jatkumisesta samanlaisena tai hyvin vastaavana myös tulevaisuudessa. Ajan saatossa yritysten toimintamallit voivat kuitenkin muuttua rajusti ja yksittäiset suuret muutokset tunnusluvuissa voivat vääristää tutkimustulosta merkittävästi. Yritysjohto voi toiminnallaan pyrkiä parantamaan tai vääristämään tilinpäätöksen tunnuslukuja tavoitteenaan antaa yritystoiminnasta todellisuutta parempi ja positiivisempi kuva (Laitinen, 1990).

Logistisessa regressioanalyysissä käytettävillä muuttujilla eli tunnuslukujen valinnalla on myös merkitystä ennusteen tarkkuudessa ja soveltuvuudessa eri aineistolle. Muuttujien valinta pohjautuu usein empiiriseen tutkimukseen eli valinta toteutetaan sen perusteella, että lopullinen malli toimii mahdollisimman hyvin tutkimusaineistossa (Laitinen, 1990). Muuttujien valinnassa käytetään esivalintaa kategorisoimalla eri muuttujia, jotta tutkittavista yrityksistä saadaan mahdollisimman kattava kuva. Mallin lopulliset muuttujat valitaan testauksen perusteella esimerkiksi poistovalinnalla siten, että muuttujia tiputetaan yksi kerrallaan pois, kunnes jäljelle jää vain mallin kannalta merkittäviä ja toistensa kanssa mahdollisimman vähän korreloivia muuttujia jäljelle (Lehtinen, 2016).

Mallit, jotka ovat kehitetty muuttujien empiirisen valinnan perusteella, voivat toimia hyvin niiden kehitystyössä käytetylle aineistolle. Kuitenkin niiden tarkkuus ja ennustekyky saattavat heiketä muilla toimialoilla tai eri ajanjaksoilla (Grice ja Dugan, 2001). Tunnuslukujen suhteuttaminen toimialan keskiarvoon tuottaakin usein parempia tuloksia (Platt ja Platt, 1990). Lisäksi näiden tulosten pohjalta kehitetty malli säilyttää ennustetarkkuutensa paremmin eri tutkimusaineistoilla. Tässä työssä ennustetaan teollisuusyritysten konkursseja. Muuttujien esivalinta perustuu tutkimukseen, jossa käsiteltiin EU:n rakennusalan yritysten konkurssiennustusta (Lappalainen, 2021).

Konkurssitutkimuksissa on usein ongelmana epätasapainoinen data, sillä keskivertoyrityksellä on vain noin 2 %:n todennäköisyys konkurssille satunnaisesti valitulla ajanjaksolla (Moody's Analytics, 2003). Tämän tiedon perusteella malli, joka luokittelee jokaisen yrityksen aktiiviseksi ja maksukykyiseksi olisi siis oikeassa 98 %:n tarkkuudella, mutta sen hyöty ennustuksessa olisi mitätön. Myös toimialakohtaisessa tutkimuksessa havaittiin, että mallin parametriarvot voivat olla vääristyneitä muuttujien osalta (Platt ja Platt, 2002). Konkurssitutkimuksen päämääränä on kuitenkin ensisijaisesti estää konkurssin aiheuttamia suuria tappioita, mihin päästään tarkastelemalla konkurssin todennäköisyyttä (Beaver et al., 2011). Voimakkaasti epätasapainoisen aineiston pohjalta kehitetyn mallin arvioinnissa käytetään tarkkuuden lisäksi muita mittareita, kuten sensitiivisyyttä (recall) ja kykyä erotella positiiviset arvot oikein (precision) (Bekkar et al., 2013). Mallin sensitiivisyys kuvaa positiivisten ennusteiden tarkkuutta ja positiivisten arvojen erottelu oikein kuvaa kaksiarvoisen luokittelun erottelukykyä (Bekkar et al., 2013).

# 3 Menetelmät

## 3.1 Logistinen regressio

Regressioanalyysi on tilastollinen menetelmä, jonka avulla pyritään selittämään ja ennustamaan yhden tai useamman muuttujan vaikutusta toiseen muuttujaan eli vastemuuttujaan (Hilbe, 2009). Monista regressioanalyysimalleista valitaan sopi-

va muuttujien ominaisuuksien ja riippuvuuksien mukaan, jotta menetelmä tuottaa parhaan mahdollisen lopputuloksen. Tässä työssä rakennetaan logistinen regressiomalli, joka soveltuu binääriseen luokitteluongelmaan (Hosmer Jr et al., 2013). Tilinpäätöksen tunnuslukuihin pohjautuvassa konkurssitutkimuksessa yrityksiä pyritään luokittelemaan kahteen luokkaan, konkurssiin ajautuviin yrityksiin ja aktiivisiin maksukykyisiin yrityksiin.

Tyypillinen regressiomalli on lineaarinen, jossa vastemuuttujan ja yhden tai useamman selittävän muuttujan välinen riippuvuus on lineaarinen. Lineaarisen riippuvuuden yksi oletuksia on, että vastemuuttujan arvot ovat jatkuvia. Kategorisen vastemuuttujan tapauksessa monet lineaarisen regression oletuksista eivät päde, jonka vuoksi logistinen regressio soveltuu tarkoitukseen paremmin (Menard, 2002). Tyypillisesti logistisessa regressiossa arvo 1 kuvaa tutkittavan asian tapahtumista ja arvo 0 tarkoittaa, ettei tutkittavaa asiaa tapahdu. Konkurssitutkimuksessa ja tässä työssä konkurssiin ajautuvia yrityksiä mallinnetaan arvolla 1 ja aktiivisia yrityksiä arvolla 0.

Logistinen ja lineaarinen regressiomalli ovat yleistyksiä tavallisesta lineaarisesta mallista (McCullagh, 2019). Tavallinen lineaarinen malli on muotoa

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \quad i = 1, ..., n \\ y_i &= \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \end{aligned} \tag{1}$$

jossa $n$ on havaintojen lukumäärä, $y_i$ on selitettävä vastemuuttuja, $\beta_0$ ei-satunnainen regressiokerroin, $p$ selittävien muuttujien lukumäärä, $\beta_j$ selittävän ei-satunnaisen muuttujan regressiokerroin, $x_{ij}$ selittävä muuttuja ja $\epsilon_i$ mallin satunnainen virhetermi. Lineaarisessa mallissa standardioletuksena on, että virhetermin $\epsilon_i$ odotusarvo on

nolla. Selitettävän vastemuuttujan $y_i$ odotusarvo voidaan siis esittää muodossa

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

Yhtälössä (2) esitetty lineaarinen regressiomalli kertoo, miten vastemuuttujan odotusarvo muuttuu suhteessa selittävän muuttujan $x_{ij}$ arvoihin (Montgomery et al., 2021). Yleistetyssä lineaarisessa mallissa pyritään luomaan malli, joka sopii tilanteisiin, joihin tavallinen lineaarinen malli ei sovi. Tavallisessa lineaarisessa mallissa oletetaan, että vastemuuttuja on jatkuva, mutta joskus tarvitaan mallia, joka rajaa muuttujat tietylle välille. Logistisen regression tapauksessa binäärisen vastemuuttujan odotusarvo noudattaa Bernoullin jakaumaa, jonka arvot voivat olla vain arvojoukossa {0,1} (Hosmer Jr et al., 2013). Sekä logistinen että lineaarinen regressiomalli ovat mallista (2) johdettuja yleistettyjä lineaarisia malleja, joiden lauseke on muotoa

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

jossa $g(\cdot)$ on linkkifunktio, joka määrittää yhteyden selittävien muuttujien ja vastemuuttujan välille muuntaen vasteen odotusarvon siten, että se riippuu lineaarisesti selittävistä muuttujista (McCullagh, 2019). Linkkifunktio muuttaa yleistä mallia siten, että sillä voidaan käsitellä erilaisia vastemuuttujia ja niiden jakaumia, kuten Bernoullin jakaumaa. Logistisen regression tapauksessa linkkifunktion määrittelyjoukkona on odotusarvoa $E(y_i)$ vastaava väli [0,1] ja se voi saada arvonsa väliltä $(-\infty, \infty)$ riippuen $x_{ij}$:stä. Näin ollen linkkifunktion $g$ arvojoukkona on koko reaalilukujen joukko $\mathbb{R}$ (Hosmer Jr et al., 2013). Linkkifunktiona logistisessa regressiossa käytetään niin kutsuttua logit-linkkiä (McCullagh, 2019).

Logistisessa regressiossa tapahtuman $y_i$ odotusarvo merkitään usein seuraavasti

$$\pi(\boldsymbol{x}) = E(y_i|\boldsymbol{x}), \tag{4}$$

jossa $\boldsymbol{x} = (x_1, x_2, ..., x_p)$ ja $\pi(\boldsymbol{x})$ kuvastaa muuttujista $\boldsymbol{x}$ riippuvan tapahtuman $P(y_i = 1|\boldsymbol{x})$ todennäköisyyttä. Odotusarvo $\pi(\boldsymbol{x})$ kuvaa siten ehdollista todennäköisyyttä, jolla vastemuuttuja $y_i$ saa arvon 1 (Hosmer Jr et al., 2013). Logistisessa regressiossa binäärinen vastemuuttuja noudattaa Bernoullin jakaumaa, jolloin vastatapahtuman $P(y_i = 0|\boldsymbol{x})$ todennäköisyys voidaan laskea kaavalla

$$P(y_i = 0|\boldsymbol{x}) = 1 - \pi(\boldsymbol{x}). \tag{5}$$

Logistiseen regressioon liittyy olennaisesti vastasuhde (odds), joka kuvaa tapahtuman toteutumisen ja toteutumatta jäämisen todennäköisyyksien suhdetta eli

$$\text{odds} = \frac{P(y_i = 1|\boldsymbol{x})}{P(y_i = 0|\boldsymbol{x})} = \frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}. \tag{6}$$

Logistisessa regressiossa käytettävä linkkifunktio eli logit-muunnos saadaan ottamalla luonnollinen logaritmi vastasuhteesta (Hilbe, 2009). Ottamalla luonnollinen logaritmi yhtälöstä (6) ja sijoittamalla tulos odotusarvon paikalle kaavaan (4) saadaan

$$\text{logit}(\pi(\boldsymbol{x})) = \ln\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right). \tag{7}$$

Tämä logit-muunnos voidaan nyt sijoittaa yleisen lineaarisen mallin linkkifunktion $g$ paikalle yhtälöön (3), jolloin lopullinen malli on muotoa

$$\ln\left(\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}. \tag{8}$$

Yhtälön (8) logit-muunnos on lineaarinen parametrien $\beta_0$ ja $\beta_j$ suhteen. Muutokset regressiokertoimissa kuvaavat, miten selittävän muuttujan $x_{ij}$ muutokset vaikuttavat tutkittavan tapahtuman todennäköisyyteen (Hosmer Jr et al., 2013). Todennäköisyys $\pi(\boldsymbol{x})$ saadaan ratkaistua kaavasta (8) seuraavasti

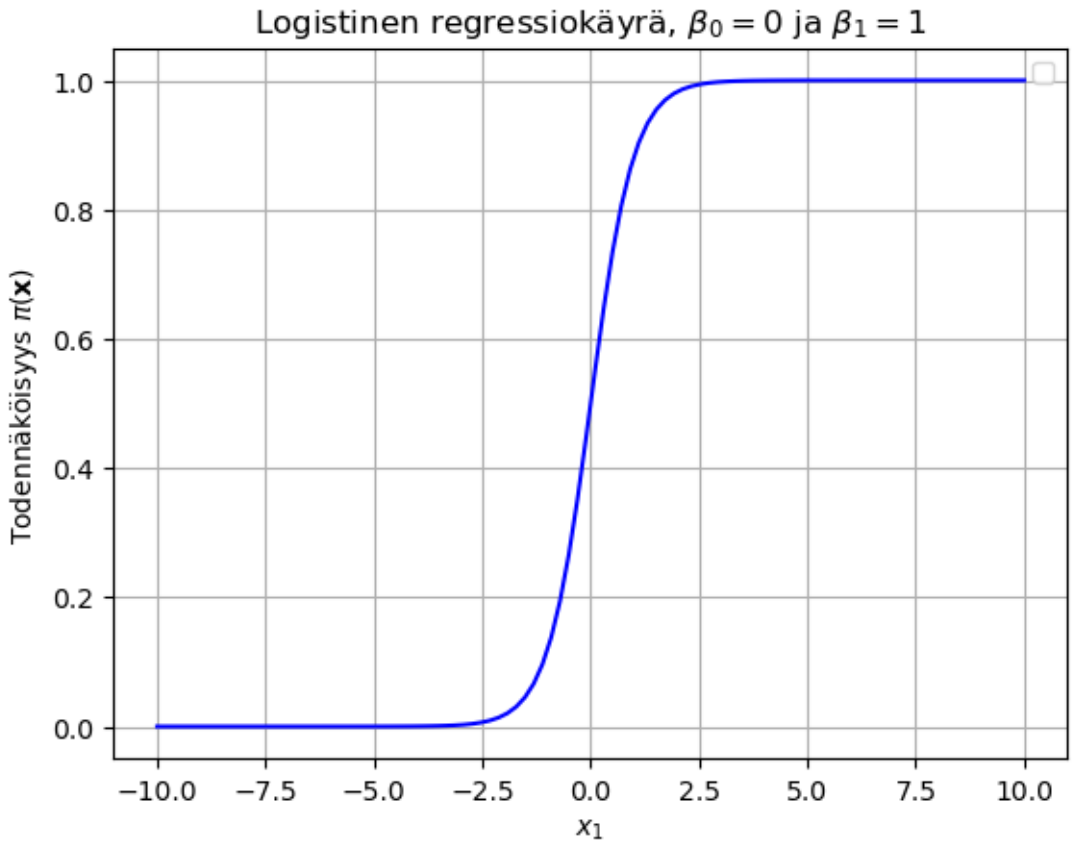
$$
\begin{aligned}
\ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) &= \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} \\
\Leftrightarrow \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} &= \mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}} \\
\Leftrightarrow \pi(\boldsymbol{x})(1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}) &= \mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}} \\
\Leftrightarrow \pi(\boldsymbol{x}) = \frac{\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}}{1+\mathrm{e}^{\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}}} &= \frac{1}{1+\mathrm{e}^{-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}}}.
\end{aligned}
\tag{9}
$$

Yksinkertaisuuden vuoksi merkitään jatkossa $g_i = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}$. Verrattuna lineaarisen mallin rajoittamattomiin arvoihin, kaavan (9) logistinen funktio muuttaa sille annetut syötteet todennäköisyyksiksi välillä [0,1]. Funktiosta saatavia todennäköisyyksiä käytetään vastemuuttujan luokittelussa tapahtuviin ja ei-tapahtuviin luokkiin.

Yhden muuttujan mallissa kaavan (9) logistinen funktio yksinkertaistuu muotoon

$$
\pi(\boldsymbol{x}) = \frac{1}{1+\mathrm{e}^{-\beta_0-\beta_1 x_1}}, \tag{10}
$$

jossa $\boldsymbol{x} = x_1$. Kuvaajassa 1 on kuvattu yksinkertainen yhden muuttujan logistinen funktio parametreilla $\beta_0 = 0$ ja $\beta_1 = 1$.




Kuva 1: Logistisen funktion kuvaaja.

## 3.2 Parametrien estimointi

Logistisen regressiomallin sovittamiseksi aineistolle kaavan (9) mukaisesti parametrit $\beta_0$ ja $\beta_j$ on estimoitava. Tavallisessa lineaarisessa regressiossa parametriestimaattien $\hat{\beta}_0$ ja $\hat{\beta}_j$ arvot määritetään yleensä pienimmän neliösumman menetelmällä, jonka tarkoituksena on minimoida neliöllisten poikkeamien summa havaittujen ja ennustettujen arvojen välillä. Binäärimuuttujilla pienimmän neliösumman menetelmä aiheuttaa ongelmia estimaattien jakaumien suhteen (Hosmer Jr et al., 2013). Binäärimuuttujaan ei voida soveltaa vastaavia lineaarisen mallin olettamuksia, kuten vastemuuttujan lineaarisuutta ja normaalijakautuneisuutta, jonka vuoksi logistisen regressiomallin parametrien estimoinnissa käytetään suurimman uskottavuuden menetelmää (Menard, 2002). Suurimman uskottavuuden menetelmässä estimointi

perustuu havaintoaineistoa vastaavaan todennäköisyysjakaumaan. Tavoitteena on löytää ne parametrien arvot, jotka maksimoivat havaitun aineiston todennäköisyyden.

Uskottavuusfunktiona logistisessa regressiossa käytetään Bernoullin jakauman pistetodennäköisyysfunktiota, joka kertoo yksittäiseen tapahtumaan liittyvän todennäköisyyden. Bernoullin jakauman pistetodennäköisyysfunktio on muotoa

$$f(y_i|\pi(\boldsymbol{x})) = \pi(\boldsymbol{x})^{y_i}(1-\pi(\boldsymbol{x}))^{1-y_i}. \tag{11}$$

Logistisessa regressiossa yksittäiset havainnot voidaan olettaa toisistaan riippumattomiksi ja identtisesti jakautuneiksi, jolloin uskottavuusfunktio saadaan yksittäisten havaintojen todennäköisyyksien tulona (Hosmer Jr et al., 2013)

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x})^{y_i}(1-\pi(\boldsymbol{x}))^{1-y_i}, \tag{12}$$

jossa $\boldsymbol{\beta}$ sisältää parametrit $\beta_0$ ja $\beta_j$.

Suurimman uskottavuuden periaatteen mukaan $\boldsymbol{\beta}$:n estimaattina käytetään parametria, joka maksimoi yhtälön (12). Matemaattisesti on helpompi käsitellä yhtälöä (12) logaritmisen muunnoksen avulla. Yhtälöstä (12) johdettu logaritminen uskottavuusfunktio on muotoa

$$\begin{aligned} L(\boldsymbol{\beta}) = \ln(l(\boldsymbol{\beta})) &= \ln(\prod_{i=1}^{n} \pi(\boldsymbol{x})^{y_i}(1-\pi(\boldsymbol{x}))^{1-y_i}) \\ &= \sum_{i=1}^{n} [y_i \ln(\pi(\boldsymbol{x})) + (1-y_i)\ln(1-\pi(\boldsymbol{x})).] \end{aligned} \tag{13}$$

Kun sijoitetaan todennäköisyyden $\pi(\boldsymbol{x})$ lauseke kaavaan (13), logaritminen uskottavuusfunktio voidaan esittää muodossa

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[ y_i \ln\left(\frac{1}{1+\mathrm{e}^{-g_i}}\right) + (1-y_i)\ln\left(1-\frac{1}{1+\mathrm{e}^{-g_i}}\right)\right], \tag{14}$$

jossa $g_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$. Uskottavuusfunktio (14) maksimoidaan derivoimalla funktio parametrien $\beta_0$ sekä $\beta_j$ suhteen ja ratkaisemalla derivaattafunktioiden nollakohdat.

Derivaattafunktio $\beta_0$:n suhteen on

$$\begin{aligned}\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n}\left[y_i \frac{\partial}{\partial \beta_0}\ln\left(\frac{1}{1+\mathrm{e}^{-g_i}}\right) + (1-y_i)\frac{\partial}{\partial \beta_0}\ln\left(1-\frac{1}{1+\mathrm{e}^{-g_i}}\right)\right] \\ &= \sum_{i=1}^{n}\left[y_i\left(\frac{\mathrm{e}^{-g_i}}{1+\mathrm{e}^{-g_i}}\right) + (1-y_i)\left(-1+\frac{\mathrm{e}^{-g_i}}{1+\mathrm{e}^{-g_i}}\right)\right] \\ &= \sum_{i=1}^{n}\left[y_i - \frac{1}{1+\mathrm{e}^{-g_i}}\right] \\ &= \sum_{i=1}^{n}\left[y_i - \pi(\boldsymbol{x})\right].\end{aligned} \tag{15}$$

Derivaattafunktio $\beta_j$:n suhteen on

$$\begin{aligned}\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n}\left[y_i \frac{\partial}{\partial \beta_j}\ln\left(\frac{1}{1+\mathrm{e}^{-g_i}}\right) + (1-y_i)\frac{\partial}{\partial \beta_j}\ln\left(1-\frac{1}{1+\mathrm{e}^{-g_i}}\right)\right] \\ &= \sum_{i=1}^{n}\left[y_i x_{ij}\left(\frac{\mathrm{e}^{-g_i}}{1+\mathrm{e}^{-g_i}}\right) + x_{ij}(1-y_i)\left(-1+\frac{\mathrm{e}^{-g_i}}{1+\mathrm{e}^{-g_i}}\right)\right] \\ &= \sum_{i=1}^{n} x_{ij}\left(y_i - \frac{1}{1+\mathrm{e}^{-g_i}}\right) \\ &= \sum_{i=1}^{n} x_{ij}\left(y_i - \pi(\boldsymbol{x})\right).\end{aligned} \tag{16}$$

Uskottavuusfunktion maksimoivat parametrit saadaan siis ratkaisemalla yhtälöt

$$\begin{aligned}\sum_{i=1}^{n}\left[y_i - \pi(\boldsymbol{x})\right] &= 0, \\ \sum_{i=1}^{n} x_{ij}\left(y_i - \pi(\boldsymbol{x})\right) &= 0,\end{aligned} \tag{17}$$

jossa $j = 1, 2, ..., p$.

Tavallisessa lineaarisessa regressiossa uskottavuusfunktion maksimoivat yhtälöt (17) ovat lineaarisia ja siten helposti ratkaistavissa. Logistisen regression tapauksessa yhtälöt (17) ovat epälineaariset ja niiden ratkaisussa käytetään erilaisia iteratiivisia menetelmiä (Hosmer Jr et al., 2013). Tässä työssä käytetään scikit-learn-kirjastoa, joka tarjoaa eri optimointialgoritmeja kuten Newtonin menetelmän ja gradienttimenetelmän (Hackeling, 2017).

## 3.3 Mallin arviointimenetelmät

Logistisen regressiomallin arvionnissa hyödynnetään luokittelutaulukkoa (confusion matrix) sekä erottelukykykäyrää (receiver operating characteristic curve = ROC-curve). Luokittelutaulukko kuvaa luokittelijan eli mallin suorituskykyä testidataan nähden. Luokittelutaulukon rivit edustavat todellisia luokkia ja sarakkeet ennustettuja luokkia (Bekkar et al., 2013). Esimerkki luokittelutaulukosta on taulukossa 1.

Taulukossa 1 merkintä TN (true negative) ilmaisee, kuinka monta tapausta on ennustettu oikein luokkaan 0. FP (false positive) ilmaisee, kuinka monta tapausta on ennustettu väärin luokkaan 1, vaikka ne kuuluisivatkin luokkaan 0. FN (false negative) ilmaisee, kuinka monta tapausta on ennustettu väärin luokkaan 0, vaikka ne kuuluisivatkin luokkaan 1. TP (true positive) ilmaisee, kuinka monta tapausta on ennustettu oikein luokkaan 1 (Fawcett, 2006).

Taulukko 1: Esimerkki luokittelutaulukosta.

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

Luokittelutaulukon avulla voidaan laskea useita mallin suorituskykyyn liittyviä metriikoita, kuten tarkkuus (accuracy), sensitiivisyys (recall) sekä positiivinen ennustearvo (precision). Tarkkuus kuvaa oikein luokiteltujen tapausten osuutta kaikista tapauksista ja lasketaan kaavalla

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

Sensitiivisyys kuvaa oikein positiivisesti luokiteltujen tapausten osuutta kaikista todellisista positiivisista tapauksista. Se lasketaan kaavalla

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

Positiivinen ennustearvo kuvaa oikein luokiteltujen tapausten osuutta kaikista posi-

tiivisesti, eli luokkaan 1, luokitelluista tapauksista ja lasketaan kaavalla

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

Eri metriikoiden lisäksi kaksiarvoiseen luokitteluongelmaan rakennettua mallia voidaan arvioida erottelukykykäyrän (ROC-curve) avulla. Erottelukykykäyrä on graafinen esitys mallin suorituskyvystä eri kynnysarvoilla. Erottelukykykäyrää varten mallin ennusteet luokitellaan eri kynnysarvoilla positiivisiksi ja negatiivisiksi tuloksiksi. Kummallekin luokalle lasketaan TPR (true positive rate) ja FPR (false positive rate) arvot, jotka voidaan määrittää luokittelutaulukon avulla (Fawcett, 2006). TPR lasketaan kaavalla

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

ja FPR lasketaan kaavalla

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}.$$

Erottelukykykäyrä esittää graafisesti TPR:n ja FPR:n välisen suhteen. Sekä TPR että FPR edustavat osuuksia, joten niiden arvot ovat väliltä [0,1]. Erottelukykykäyrän alle jäävästä pinta-alasta käytetään nimitystä AUC (area under curve). AUC kuvaa mallin suorituskyvyn keskimääräistä arvoa kaikilla mahdollisilla kynnysarvoilla. Mitä suurempi AUC on, sitä parempi mallin luokittelukyky on. Ihanteellinen AUC:n arvo olisi 1 (Fawcett, 2006). Täysin satunnaisen luokittelijan AUC-arvo olisi 0,5; eli erottelukykykäyrä olisi täysin suora. Tällöin mallista ei ole luokittelun kannalta hyötyä (Hosmer Jr et al., 2013).

# 4 Tulokset

## 4.1 Aineisto

Tutkimuksessa käytetty aineisto on kerätty Bureau Van Dijk:n Orbis-tietokannasta. Aineisto kattaa Euroopan Unionin alueella toimivat teollisuusyritykset. Toimia-

lakohtaisessa rajauksessa on käytetty Executive Office of the President Office of Management and Budget (2017) mukaista NAICS luokitusta, joka luokittelee teollisuusyrityksiin tuotteita valmistavat ja jalostavat yritykset. Teollisuusyritykset ovat jaettu edelleen omiin alaluokkiin riippuen tuotannossa vaadittavista resursseista tai tuotantolaitteista. Aineisto rajattiin kattamaan kaikki erilliset alaluokat.

Yrityskoon osalta aineisto rajattiin pk-yrityksiin. Määritelmän mukaan näitä ovat yritykset, joiden henkilöstömäärä on alle 250 ja vuotuinen liikevaihto alle 50 miljoonaa euroa tai tase enintään 43 miljoonaa euroa (Euroopan komissio, 2015). Aineistosta rajattiin pois mikroyritykset, joiden henkilöstömäärä on alle 10. Mikroyritykset eivät ole vastaavalla tavalla kirjanpitovelvollisia kuin pk-yritykset, mikä voi vaikuttaa tilinpäätöstietojen saatavuuteen ja luotettavuuteen (Leppiniemi, 2017).

Lopullinen aineisto kattaa konkurssiyritykset, jotka ovat hakeutuneet konkurssiin vuosina 2020-2021. Kahden vuoden tarkastelujaksoon päädyttiin, jotta aineistosta saadaan tarpeeksi kattava. Myös Lappalainen (2021) ja Lehtinen (2016) ovat hyödyntäneet samaa kahden vuoden aikaväliä konkurssitarkastelussa. Lisäksi yrityksillä piti olla vähintään kolme tilikautta ennen konkurssia. Näin ollen yrityksestä on saatavilla tarpeeksi tietoja ja konkurssitutkimuksen luotettavuus kolmen vuoden tarkastelujaksolla on vielä edelleen hyvä (Deakin, 1972).

Tutkittavassa aineistossa on yhteensä 104 konkurssiyritystä ja 156 759 aktiivista maksukykyistä yritystä. Monen yrityksen osalta tilinpäätöstiedot ovat kuitenkin joko vajavaiset tai puutteelliset, jolloin lopullinen aineisto sisältää 62 konkurssi- ja 41813 aktiivista yritystä sekä näiden tilinpäätöstiedot vuosilta 2017-2019. Kaikki tilinpäätöstietoja koskevat arvot ja niistä johdetut tunnusluvut on ilmoitettu tuhansissa euroissa.

Muuttujien esivalinta pohjautuu konkurssitutkimukseen EU:n rakennusalan yrityksillä (Lappalainen, 2021). Esivalinta kattaa 17 tunnuslukua eri kategorioista, jotka ovat listattuna tarkemmin taulukossa 2. Lopullisessa mallissaan Lappalainen (2021) käytti viittä eri tunnuslukua, kun konkurssia mallinnettiin kaksi vuotta ennen sen tapahtumista. Taulukossa 2 on vihreällä pohjalla merkittynä lisäksi edellä mainitut viisi tunnuslukua lopullisesta mallista.

Taulukko 2: Kaikki esivalitut tunnusluvut ja vihreällä pohjalla lopullisessa mallissa käytetyt tunnusluvut.

| Nro | Tunnusluku | Kategoria |
|---|---|---|
| 1 | Sijoitetun pääoman tuottoprosentti (ROI) | Kannattavuus |
| 2 | Oman pääoman tuottoprosentti (ROE) | Kannattavuus |
| 3 | Kokonaispääoman tuottoprosentti (ROA) | Kannattavuus |
| 4 | Liiketulos (EBIT) | Kannattavuus |
| 5 | Käyttökate (EBITDA) | Kannattavuus |
| 6 | Net gearing | Kannattavuus |
| 7 | Omavaraisuusasteprosentti (Oma pääoma / taseen loppusumma) | Vakavaraisuus |
| 8 | Vieraan pääoman takaisinmaksukyky (Rahoitustulos / [vastattavaa yhteensä – oma pääoma]) | Vakavaraisuus |
| 9 | Lyhytaikainen vieras pääoma / taseen loppusumma | Vakavaraisuus |
| 10 | Oma pääoma / vieras pääoma | Vakavaraisuus |
| 11 | Quick ratio (Rahoitusomaisuus / lyhytaikainen vieras pääoma) | Maksuvalmius |
| 12 | Current ratio | Maksuvalmius |
| 13 | Rahoitustulosprosentti | Maksuvalmius |
| 14 | Käyttöpääoma / taseen loppusumma | Maksuvalmius |
| 15 | Myyntisaamisten kiertoaika | Maksuvalmius |
| 16 | Liikevaihdon kasvuprosentti | Kasvu |
| 17 | Jalostusarvon kasvuprosentti | Kasvu |

Tässä työssä käytetyn tutkimusaineiston pohjalta taulukossa 2 lueteltuja sijoitetun pääoman tuottoprosenttia (ROI) ja myyntisaamisten kiertoaikaa ei voitu laskea, joten ne joudutaan jättämään pois valinnasta. Kuvailevat tunnusluvut tutkimusaineistosta kaksi vuotta ennen konkurssia on taulukossa 3.

Taulukko 3: Kuvailevat tunnusluvut kaksi vuotta ennen konkurssia.

| **Tunnusluku** | **Aineiston ominaisuus** | **Konkurssit** | **Aktiiviset** |
|---|---|---|---|
| Oman pääoman tuottoprosentti (ROE) | Keskiarvo | -4,03 | 11,03 |
| | Mediaani | 3,38 | 8,88 |
| | Keskihajonta | 66,68 | 24,08 |
| Kokonaispääoman tuottoprosentti (ROA) | Keskiarvo | -0,22 | 4,12 |
| | Mediaani | 1,18 | 2,75 |
| | Keskihajonta | 6,17 | 6,60 |
| Liiketulos (EBIT) | Keskiarvo | 140,75 | 413,01 |
| | Mediaani | 47,81 | 168,06 |
| | Keskihajonta | 632,20 | 1033,85 |
| Käyttökate (EBITDA) | Keskiarvo | 381,77 | 692,84 |
| | Mediaani | 125,21 | 337,25 |
| | Keskihajonta | 969,51 | 1206,87 |
| Net gearing | Keskiarvo | 170,56 | 142,65 |
| | Mediaani | 120,18 | 87,32 |
| | Keskihajonta | 151,20 | 157,12 |
| Omavaraisuusasteprosentti | Keskiarvo | 0,31 | 0,37 |
| | Mediaani | 0,26 | 0,35 |
| | Keskihajonta | 0,18 | 0,20 |
| Vieraan pääoman takaisinmaksukyky | Keskiarvo | -0,02 | -0,01 |
| | Mediaani | -0,01 | -0,01 |
| | Keskihajonta | 0,03 | 0,05 |
| Lyhytaikainen vieras pääoma / taseen loppusumma | Keskiarvo | 0,15 | 0,13 |
| | Mediaani | 0,12 | 0,09 |
| | Keskihajonta | 0,13 | 0,11 |
| Oma pääoma / vieras pääoma | Keskiarvo | 0,66 | 0,91 |
| | Mediaani | 0,35 | 0,54 |
| | Keskihajonta | 0,99 | 1,32 |
| Quick ratio | Keskiarvo | -0,03 | -0,02 |
| | Mediaani | -0,02 | -0,01 |
| | Keskihajonta | 0,04 | 0,08 |
| Current ratio | Keskiarvo | 1,42 | 1,77 |
| | Mediaani | 1,16 | 1,43 |
| | Keskihajonta | 0,92 | 1,42 |
| Rahoitustulosprosentti | Keskiarvo | -1,50 | -0,78 |
| | Mediaani | -0,99 | -0,53 |
| | Keskihajonta | 1,92 | 8,97 |
| Käyttöpääoma / taseen loppusumma | Keskiarvo | 0,32 | 0,28 |
| | Mediaani | 0,32 | 0,27 |
| | Keskihajonta | 0,23 | 0,20 |
| Liikevaihdon kasvuprosentti | Keskiarvo | 19,57 | 11,72 |
| | Mediaani | 4,79 | 5,40 |
| | Keskihajonta | 64,57 | 388,48 |
| Jalostusarvon kasvuprosentti | Keskiarvo | 3,77 | 21,41 |
| | Mediaani | 1,28 | 5,85 |
| | Keskihajonta | 27,14 | 2368,46 |

## 4.2 Malli

Aineistoon ja muuttujiin perehtymisen jälkeen aloitetaan ennustemallin rakentaminen. Konkurssiriskin arvioimiseksi aineiston pohjalta rakennetaan kaksi erilaista mallia, joista ensimmäinen sisältää kaikki esivalinnan kautta määritetyt muuttujat ja toinen malli ne viisi muuttujaa, joita Lappalainen (2021) päätyi käyttämään omassa tutkimuksessaan. Konkurssiyritysten suhteellinen osuus kaikista yrityksistä on hyvin pieni. Tutkimukseen sopivia konkurssiyrityksiä oli 62 kappaletta, kun aktiivisten yritysten määrä oli 41813. Aktiivisista yrityksistä valitaan satunnaisotannalla 310 eri yritystä mallien rakentamista varten, jolloin konkurssiyritysten osuus aineistosta on 20 %. Myös Lappalainen (2021) toteutti tutkimuksessaan aktiivisten yritysten otannan samoin. Tämä valittu aineisto jaetaan sitten satunnaisesti koulutusdataan ja erilliseen testausdataan, jota käytetään mallien validointiin. Tässä työssä käytettävän scikit-learn-kirjaston oletusasetuksena on käyttää 25 % aineistosta mallin testaamiseen (Hackeling, 2017). Muuttujat skaalataan käyttämällä scikit-learnin standardscaler-funktiota, jotta mallista saadut tulokset regressiokertoimien osalta ovat vertailukelpoisia keskenään (Hackeling, 2017). Skaalauksen tuloksena muuttujien keskiarvo on 0 ja keskihajonta 1, mikä helpottaa myös yhtälön (13) mukaisen uskottavuusfunktion määrittämisessä.

Koulutusdataan sovitetaan logistinen regressiomalli, jonka toteutuksessa hyödynnetään scikit-learn-kirjastoa. Selittävinä muuttujina ensimmäisessä mallissa käytetään taulukon 2 mukaisia esivalittuja muuttujia ja toisen mallin sovituksessa käytetään taulukossa 2 korostettua viittä muuttujaa. Molempien mallien toteutuksessa konkurssiriskiä arvioidaan kaksi vuotta ennen konkurssia. Mallien regressiokertoimet ovat taulukoissa 4 ja 5.

Taulukko 4: Regressiokertoimet kaikki esivalitut muuttujat sisältävässä mallissa.

| **Tunnusluku** | **Regressiokerroin** |
|---|---|
| ROE | 0,2599 |
| ROA | -0,7882 |
| EBIT | 0,3403 |
| EBITDA | -0,1925 |
| Net gearing | 0,1317 |
| Omavaraisuusasteprosentti | -0,0163 |
| Current ratio | -0,6610 |
| Vieraan pääoman takaisinmaksukyky | -0,3598 |
| Lyhytaikainen vieras pääoma / taseen loppusumma | -0,2844 |
| Oma pääoma / vieras pääoma | 0,4636 |
| Quick ratio | -0,1107 |
| Rahoitustulosprosentti | -0,1325 |
| Käyttöpääoma / taseen loppusumma | 0,4825 |
| Liikevaihdon kasvuprosentti | 0,4079 |
| Jalostusarvon kasvuprosentti | 0,0111 |

Taulukko 5: Regressiokertoimet viisimuuttujaisessa mallissa.

| **Tunnusluku** | **Regressiokerroin** |
|---|---|
| EBITDA | -0,6691 |
| Omavaraisuusasteprosentti | -0,1545 |
| Vieraan pääoman takaisinmaksukyky | -0,5959 |
| Lyhytaikainen vieras pääoma / taseen loppusumma | 0,1235 |
| Rahoitustulosprosentti | -0,7992 |

Regressiokertoimista voidaan arvioida muuttujan merkitystä konkurssiriskin kasvuun. Itseisarvoltaan suuret muuttujat vaikuttavat eniten konkurssiriskiin ja vastaavasti itseisarvoltaan pienet muuttujat vaikuttavat vähän konkurssiriskiin. Positiivinen kertoimen arvo tarkoittaa, että muuttujan kasvu lisää konkurssiriskiä, kun taas negatiivinen arvo tarkoittaa, että muuttujan kasvu vähentää konkurssiriskiä.

Taulukon 4 valossa konkurssiriskiä alentavat eniten kokonaispääoman tuottoprosentin (ROA:n) kasvu ja current ration kasvu, joiden regressiokertoimet ovat negatiivisimmat. Mallin valossa konkurssiriskiä kasvattavat eniten oman pääoman kasvu vieraaseen pääomaan nähden sekä käyttöpääoman kasvu taseen loppusummaan nähden. Mallin valossa muita konkurssiriskiä kasvattavia tekijöitä ovat liikevaihdon

kasvuprosentin positiiviset muutokset sekä liiketuloksen (EBIT) kasvu.

Taulukon 5 viisimuuttujaisen mallin valossa konkurssiriski kasvaa lyhytaikaisen vieraan pääoman kasvaessa taseen loppusummaan nähden. Mallin valossa konkurssiriskiä pienentävät eniten rahoitustulosprosentin sekä käyttökatteen (EBITDA) kasvu. Muita konkurssiriskiä alentavia tekijöitä ovat vieraan pääoman takaisinmaksukyvyn parantuminen sekä omavaraisuusasteen kasvu.

Kun tarkastellaan molempien mallien regressiokertoimia, voidaan huomata, että kertoimien etumerkit pysyvät samoina lukuun ottamatta muuttujaa lyhytaikainen vieras pääoma / taseen loppusumma. Tämän muuttujan kertoimen etumerkki riippuu muista mallin sisältävistä muuttujista. Lisäksi kertoimien itseisarvot vaihtelevat eri mallien välillä, mikä vaikuttaa myös kertoimien suuruusjärjestykseen.

Mallien hyvyyttä voidaan arvioida mallista johdetun luokittelutaulukon sekä erottelukykykäyrän avulla. Kaikki muuttujat sisältävän mallin luokittelutaulukko on taulukossa 6.

Taulukko 6: Luokittelutaulukko kaikki muuttujat sisältävälle mallille.

| | Ennustettu 0 | Ennustettu 1 |
|---|---|---|
| Todellinen 0 | 78 | 2 |
| Todellinen 1 | 11 | 2 |

Taulukosta 6 saadaan mallin tarkkuus, sensitiivisyys sekä positiivinen ennustearvo:

$$\text{tarkkuus} = 0{,}86$$

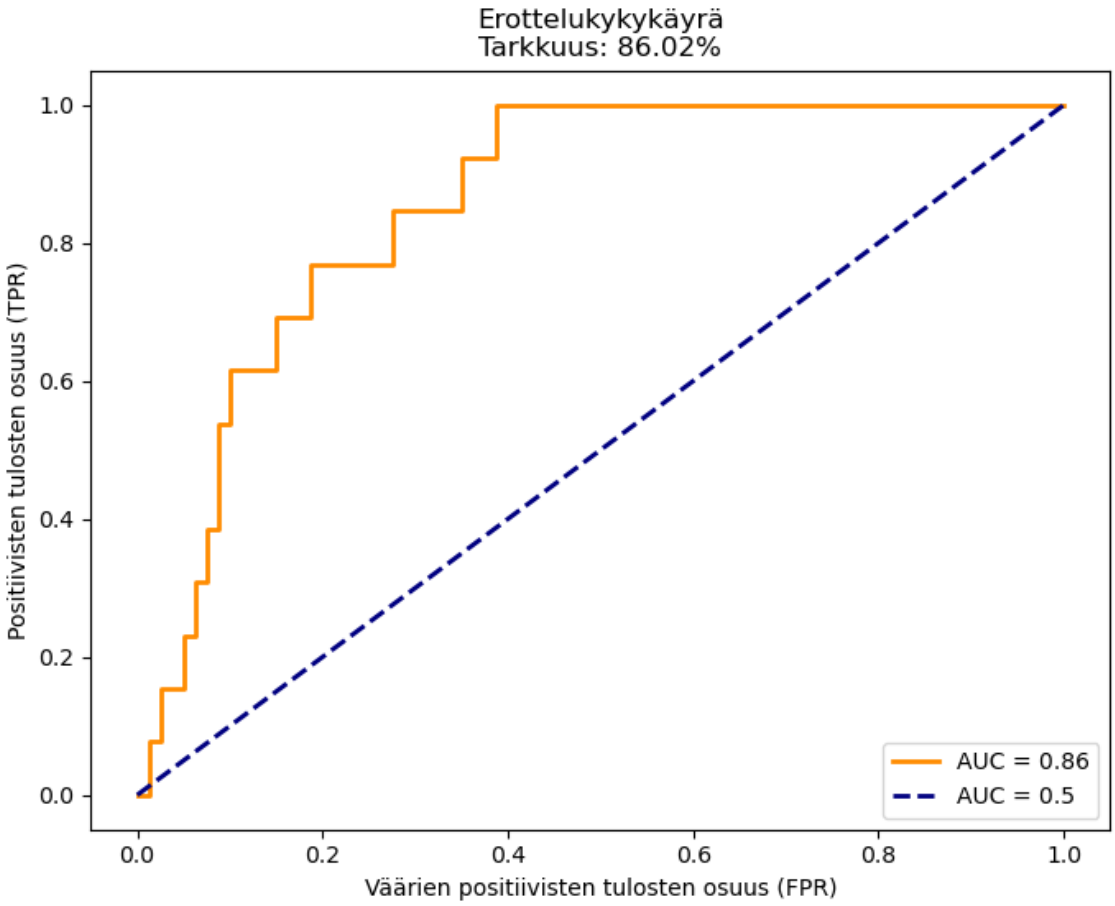
$$\text{sensitiivisyys} = 0{,}15$$

$$\text{positiivinen ennustearvo} = 0{,}50$$

Kaikki muuttujat sisältävä malli luokittelee kaikki tapaukset oikein noin 86 % tarkkuudella. Todellisista konkurssiyrityksistä se kykenee tunnistamaan oikein 15 %. Puolet yrityksistä, jotka malli on ennustanut konkursseiksi, ovat todella olleet konkurssiyrityksiä. Mallin sensitiivisyys, eli olennaisin kyky tunnistaa konkurssiyritykset aktiivisista on verraten heikko.

Kaikki muuttujat sisältävän mallin erottelukykykäyrä on kuvassa 2. Käyrän alle

jäävä pinta-ala (AUC) on 0,86, joka on hyvällä tasolla (Hosmer Jr et al., 2013). Sininen katkoviiva edustaa AUC:n arvoa 0,5.




Kuva 2: Kaikki muuttujat sisältävän mallin erottelukykykäyrä.

Viisimuuttujaisen mallin luokittelutaulukko on taulukossa 7.

Taulukko 7: Viisimuuttujaisen mallin luokittelutaulukko.

| | Ennustettu 0 | Ennustettu 1 |
|---|---|---|
| Todellinen 0 | 75 | 0 |
| Todellinen 1 | 17 | 1 |

Taulukosta 7 saadaan mallin tarkkuus, sensitiivisyys sekä positiivinen ennustearvo:

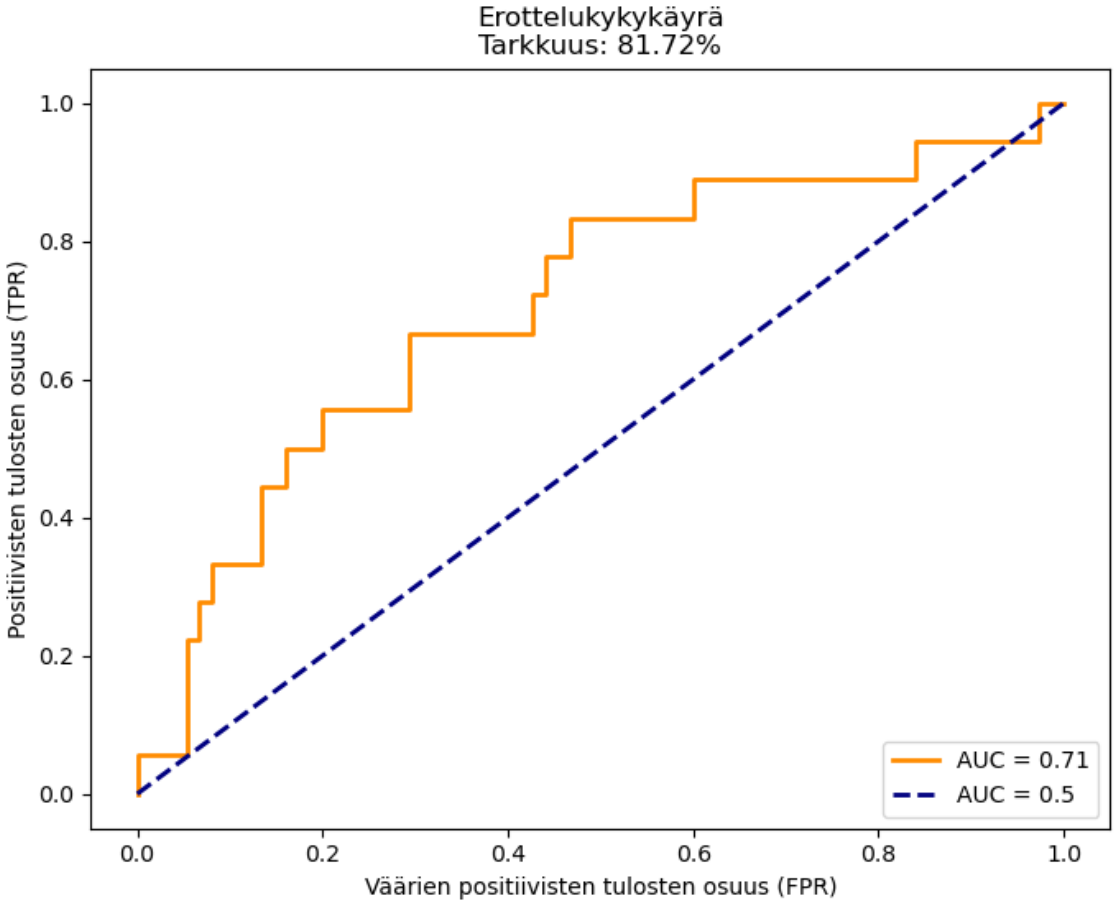
$$\text{tarkkuus} = 0{,}82$$

$$\text{sensitiivisyys} = 0{,}06$$

$$\text{positiivinen ennustearvo} = 1{,}0$$

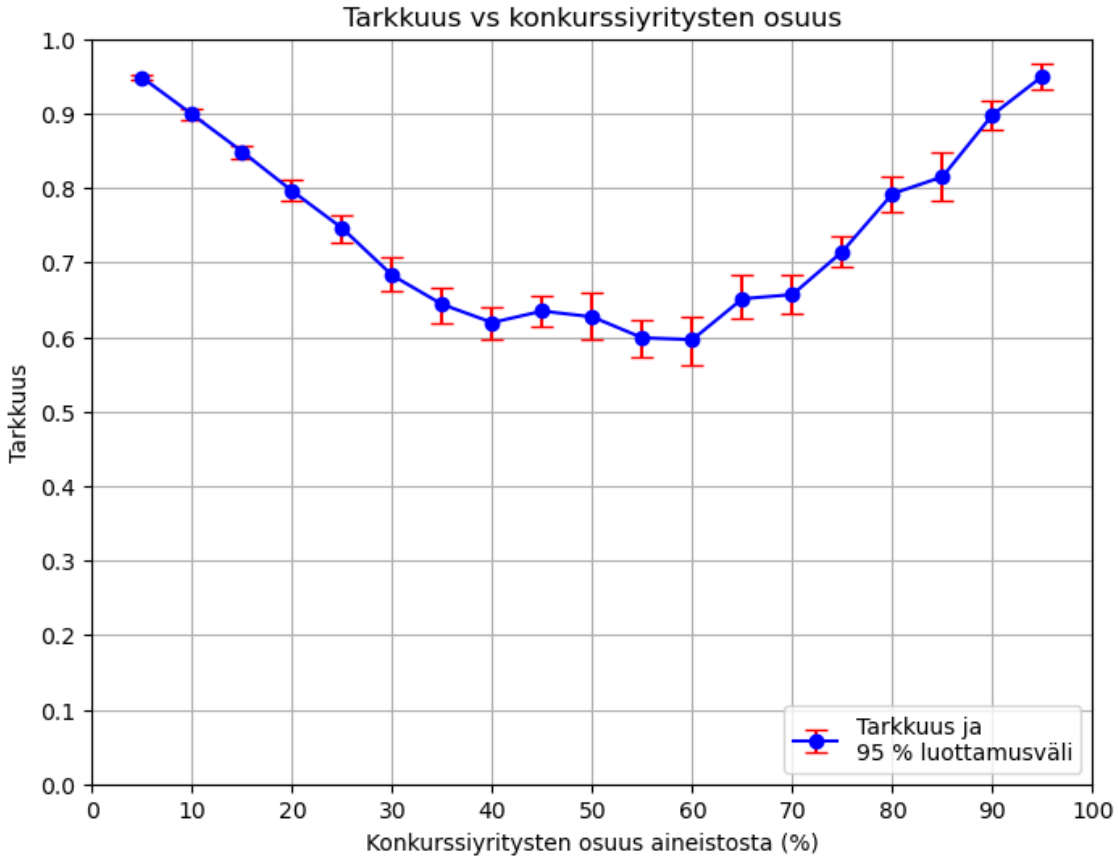
Viisimuuttujainen malli luokittelee kaikki tapaukset oikein noin 82 % tarkkuudella. Todellisista konkurssiyrityksistä se kykenee tunnistamaan oikein 6 %. Kaikki ennustetuista konkurssiyrityksistä on luokiteltu oikein konkurssiyrityksiin. Verrattuna kaikki muuttujat sisältävään malliin, on sensitiivisyys edelleen heikolla tasolla.

Viisimuuttujaisen mallin erottelukykykäyrä on kuvassa 3. Käyrän alle jäävä pinta-ala (AUC) on 0,71, joka on hyväksyttävällä tasolla (Hosmer Jr et al., 2013).

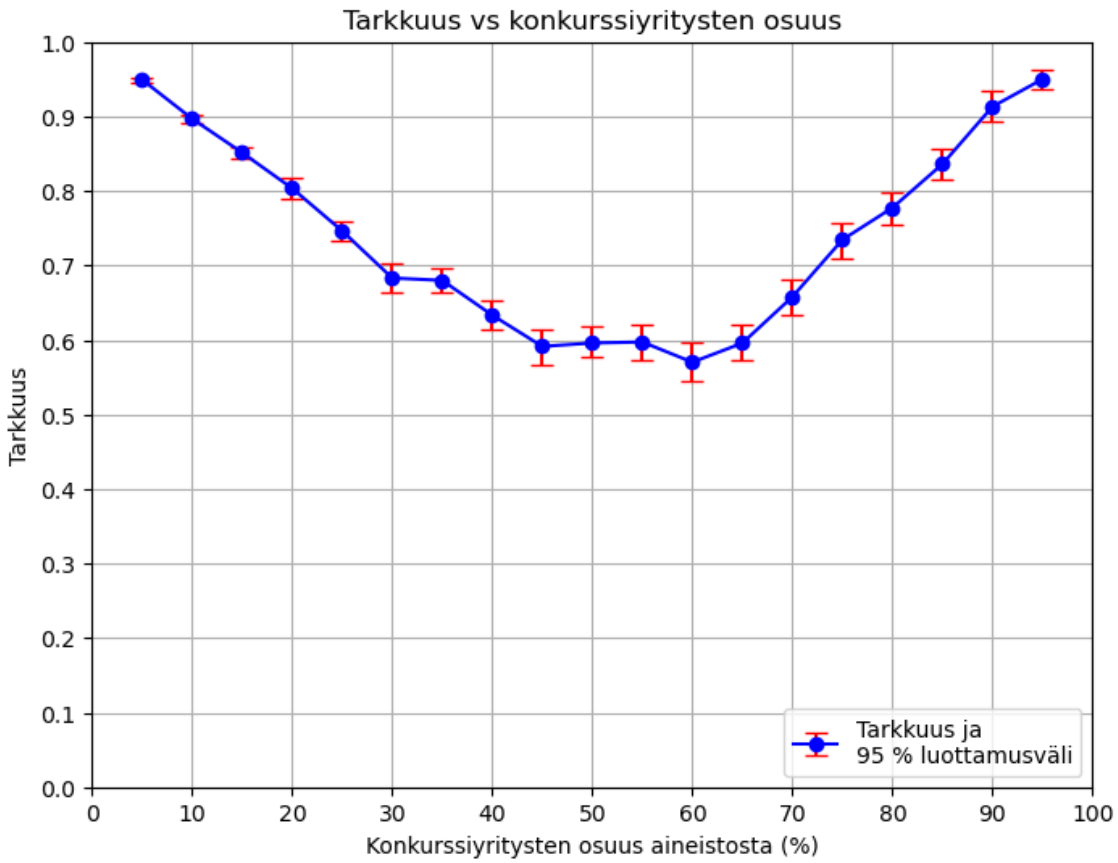



Kuva 3: Viisimuuttujaisen mallin erottelukykykäyrä.

Aktiiviset yritykset valittiin mallien rakentamista varten satunnaisotannalla. Kuvissa 4 ja 5 on lisäksi esitetty mallien tarkkuudet, kun konkurssiyritysten osuutta aineistosta on muutettu. Satunnaisotannalla valittiin tarvittava määrä yrityksiä ja niiden avulla muodostetun mallin tarkkuus arvioitiin. Keskiarvo neljänkymmenen mallin tarkkuudesta on kuvattu graafisesti suhteessa konkurssiyritysten osuuteen. Kuvaajiin on lisätty myös 95 % luottamusvälit.




Kuva 4: Kaikki esivalitut muuttujat sisältävän mallin tarkkuus konkurssiyritysten osuuden funktiona.




Kuva 5: Viisimuuttujaisen mallin tarkkuus konkurssiyritysten osuuden funktiona.

Kuvista 4 ja 5 nähdään, että mallien tarkkuudet laskevat noin 60 % tasolle, kun konkurssiyritysten osuus aineistosta kasvaa noin 50 % saakka. Tämän jälkeen tarkkuudet alkavat nousta nopeasti takaisin lähtötasoilleen, kun konkurssiyritysten osuutta kasvatetaan molemmissa malleissa.

# 5 Yhteenveto

Tässä työssä tutkittiin teollisuusyritysten konkurssien ennustamista logistisella regressiolla. Ennusteet tuotettiin rakentamalla tilinpäätöstietoihin tukeutuva logistinen regressiomalli, jolla konkurssia ennustettiin kaksi vuotta ennen sen tapahtumista. Työn alussa perehdyttiin aikaisempaan konkurssien ennustamista käsittelevään tutkimukseen ja tapoihin mallintaa konkurssiriskiä. Aikaisempien tutkimustulosten lisäksi tarkasteltiin konkurssitutkimukseen liittyviä haasteita mallinnuksessa. Tämän jälkeen esiteltiin menetelmäksi valittu logistinen regressio ja mallin arvioinnissa käytettäviä menetelmiä. Tutkimusaineistoa muokattiin mallin rakentamisen kannalta sopivaksi, jonka jälkeen rakennettiin kaksi logistista regressiomallia konkurssiriskin arvioimiseksi. Mallien rakentamisen jälkeen niiden antamia tuloksia ja ennustekykyä tulkittiin ja arvioitiin.

Aineistona työssä käytettiin EU:n alueella toimivien teollisten pk-yritysten tilinpäätöstietoja. Aineisto koostui vuosina 2020-2021 konkurssin tehneistä sekä toimintaansa edelleen jatkavista aktiivisista yrityksistä. Yhteensä aineisto koostui noin 42 000 yrityksestä, joista 62 oli konkurssiyrityksiä. Tilinpäätöstiedoista valittiin aikaisemman tutkimuksen perusteella 15 selittävää muuttujaa. Vastemuuttujana mallissa oli konkurssitapahtuma.

Työssä rakennettujen mallien perusteella konkurssiriskiä alentavat eniten kokonaispääoman tuottoprosentin (ROA), maksuvalmiutta kuvaavan current ration sekä käyttökatteen eli EBITDA:n kasvut riippuen mallissa käytetyistä muuttujista. Konkurssiriskiä eniten kasvattavia tekijöitä ovat oman pääoman kasvu vieraaseen pääomaan nähden sekä käyttöpääoman kasvu taseen loppusummaan nähden. Mallien kyvyt erotella konkurssiyritykset oikein oli kuitenkin heikohko.

Työn yhtenä tavoitteena oli arvioida aikaisemmin kehitetyn rakennusalan yrityskonkursseja ennnustavan mallin soveltuvuutta teollisuusyritysten konkurssien ennustamiseen. Mallien kriittinen ominaisuus konkurssiyritysten erottelulle oikein on heikko. Malli, joka on kehitetty yhdelle aineistolle soveltuu heikosti konkurssiennustukseen toiselle aineistolle.

Heikkoa erottelukykyä voi osaltaan selittää mallin rakentamisessa käytetty epätasapainoinen aineisto. Konkurssiyritysten osuus kaikista tutkittavista yrityksistä on tässä aineistossa pieni, mikä voi aiheuttaa ongelmia sekä mallin opettamisessa että testauksessa. Näitä ongelmia voitaisiin välttää hyödyntämällä erilaisia otantamenetelmiä. Lisäksi mallin erottelukykyyn voitaisiin vaikuttaa etsimällä optimaalinen kynnysarvo, jota suuremmat arvot luokitellaan konkurssiyrityksiin. Toisaalta mallin tarkkuus voi heikentyä pienempää kynnysarvoa käytettäessä.

Tässä työssä ei otettu kantaa mallissa käytettävien muuttujien valinnalle. Toimialakohtaista optimaalista muuttujien valintaa voitaisiin tutkia enemmän myös eri toimialojen yrityksillä. Konkurssiriskiä voitaisiin arvioida myös muilla koneoppimisen menetelmillä ja vertailla eri menetelmien sopivuutta konkurssien ennustamiseen.

# Viitteet

William H. Beaver. Financial Ratios as Predictors of Failure. *Journal of Accounting Research*, pages 71–111, 1966.

William H. Beaver, Maria Correia, Maureen F. McNichols, et al. Financial Statement Analysis and the Prediction of Financial Distress. *Foundations and Trends in Accounting*, 5(2):99–173, 2011.

Mohamed Bekkar, Hassiba K. Djemaa, ja Taklit A. Alitouche. Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, 3(10):27–38, 2013.

Jodi L. Bellovary, Don E. Giacomino, ja Michael D. Akers. A Review of Bankruptcy Prediction Studies: 1930 to Present. *Journal of Financial Education*, 33(12):1–42, 2007.

Edward B. Deakin. A Discriminant Analysis of Predictors of Business Failure. *Journal of Accounting Research*, 10(1):167–179, 1972.

Euroopan komissio. *Käyttöopas pk-yrityksen määritelmä*. Luxemburg: Euroopan unionin julkaisutoimisto, 2015.

Executive Office of the President Office of Management and Budget. 2017 NAICS Manual, 2017. URL https://www2.census.gov/library/reference/naics/publications/2017-NAICS-Manual.pdf.

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

John S. Grice ja Michael T. Dugan. The Limitations of Bankruptcy Prediction Models: Some Cautions for the Researcher. *Review of Quantitative Finance and Accounting*, 17:151–166, 2001.

Gavin Hackeling. *Mastering Machine Learning with scikit-learn*. Packt Publishing Ltd, 2017.

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

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

Erkki K. Laitinen. *Konkurssin ennustaminen*. Vaasan Yritysinformaatio Oy, 1990.

Otto Lappalainen. *EU:n rakennusalan yritysten konkurssien ennustaminen*. Pro gradu -tutkielma, Itä-Suomen yliopisto, 2021.

Leo Lehtinen. *Konkurssin ennustaminen tilinpäätöstiedoilla*. Pro gradu -tutkielma, Aalto-yliopisto, 2016.

Jarmo Leppiniemi. Pien- ja mikroyrityksen tilinpäätökset: yhtäläisyydet ja erot, 2017. URL https://tilisanomat.fi/yritysjuridiikka/pien-ja-mikroyrityksen-tilinpaatokset-yhtalaisyydet-ja-erot. Luettu 11.03.2024.

Peter McCullagh. *Generalized Linear Models*. Routledge, 2019.

Scott Menard. *Applied Logistic Regression Analysis*. Sage, 2002.

Douglas C. Montgomery, Elizabeth A. Peck, ja Geoffrey G. Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 2021.

Moody's. Orbis, 2024. URL https://www.moodys.com/web/en/us/capabilities/company-reference-data/orbis.html. Luettu 10.03.2024.

Moody's Analytics. Modeling default risk. 2003. URL https://www.moodysanalytics.com/-/media/whitepaper/before-2011/12-18-03-modeling-default-risk.pdf. Luettu 10.02.2024.

James A. Ohlson. Financial Ratios and the Probabilistic Prediction of Bankruptcy. *Journal of Accounting Research*, 18(1):109–131, 1980.

Harlan D. Platt ja Marjorie B. Platt. Development of a Class of Stable Predictive Variables: The Case of Bankruptcy Prediction. *Journal of Business Finance & Accounting*, 17(1):31–51, 1990.

Harlan D. Platt ja Marjorie B. Platt. Predicting Corporate Financial Distress: Reflections on Choice-based Sample Bias. *Journal of Economics and Finance*, 26 (2):184–199, 2002.