jamk

Asiakaspoistuman ennustaminen koneoppimisen avulla

Jesse Hyvärinen

Opinnäytetyö, AMK
Toukokuu 2024
Tieto- ja viestintätekniikan koulutusohjelma

jamk | Jyväskylän ammattikorkeakoulu
University of Applied Sciences

jamk

Kuvailulehti

Hyvärinen, Jesse

Asiakaspoistuman ennustaminen koneoppimisen avulla

Jyväskylä: Jyväskylän ammattikorkeakoulu. Toukokuu, 2024, 50 sivua.

Tieto- ja viestintätekniikan tutkinto-ohjelma. Opinnäytetyö AMK.

Julkaisun kieli: suomi

Julkaisulupa avoimessa verkossa: kyllä

Tiivistelmä

Opinnäytetyö tarkastelee koneoppimista ja koneoppimismallien ennustamiskykyä asiakaspoistuman suhteen. Työn toimeksiantajana toimi tamperelainen kameratarvikkeiden kansainvälinen verkkokauppa. Työssä ennustettiin toimeksiantajan asiakasaineiston perusteella mahdollisia poistuvia asiakkaita koneoppimismalleja käyttäen. Työn tavoitteena oli vertailla käytettyjen mallien välisiä eroavaisuuksia tarkkuuksissa, löytää poistumaan vaikuttavia muuttujia sekä tutkia voidaanko poistumaa ylipäänsä ennustaa koneoppimisen menetelmin. Työ toteutettiin tutkimuksellisena kehittämistyönä.

Empiirinen tutkimus toteutettiin kvantitatiivisena tutkimuksena yritykselle kertyneen asiakasdatan pohjalta. Teoriapohja kerättiin koneoppimista ja asiakaspoistuman ennustamista käsittelevistä tutkimuksista, joiden pohjalta valittiin käytettävät koneoppimismallit.

Tutkimustuloksista käy ilmi, että käytetyt mallit pystyivät ennustamaan poistuvia verkkokauppa-asiakkaita suurehkolla tarkkuudella. Mallit olivat kuitenkin merkittävästi huonompia ennustamaan pysyviä asiakkaita. Käytettyjen mallien välille ei juuri kuitenkaan saatu merkittäviä eroja. Satunnaismetsä näytti kuitenkin tunnistavan poistuvat asiakkaat hieman paremmin kuin logistinen regressio. Poistumaan vaikuttavia muuttujia onnistuttiin löytämään ja tehdyt löydökset olivat linjassa aikaisempien tutkimuksien löydöksien kanssa.

Työn tavoitteet täyttyivät ja toimeksiantajalle saatiin tuotua lisää substanssiosaamista aiheen suhteen. Ennustemalleja käyttämällä voidaan tulevaisuudessa tehdä myös tarkempia toimia asiakassäilyvyyden parantamiseksi ja tehty tutkimus voi toimia osana yrityksen asiakasymmärtämisen parantamista.

Avainsanat (asiasanat)

asiakaspoistuma, koneoppiminen, asiakaspoistuman ennustaminen

Muut tiedot (salassa pidettävät liitteet)

-

jamk

Description

Hyvärinen, Jesse

Predicting customer churn using machine learning

Jyväskylä: JAMK University of Applied Sciences, May 2020, 50 pages

Degree Programme in Information and Communications Technology. Bachelor´s thesis.

Permission for open access publication: Yes

Language of publication: Finnish

Abstract

Thesis examines machine learning and the predictive capabilities of machine learning models regarding customer churn. The commissioning party for the study was an international online retailer of used cameras and camera accessories based in Tampere. The study predicted potential churn among the client's customers based on their data, using machine learning models. The objectives of the study were to compare the differences in accuracy between the models used, identify variables influencing churn, and investigate whether churn can be predicted using machine learning methods. The study was conducted as a research and development project.

The empirical research was carried out as a quantitative study based on customer data accumulated by the company. The theoretical framework was derived from studies on machine learning and customer churn prediction, guiding the selection of machine learning models.

The research findings indicate that the models used were able to predict departing online retail customers with a relatively high degree of accuracy. However, the models did not perform that well when predicting not churning customers. Minimal differences were observed between the models used, though random forest succeeded a bit better predicting churning customer. Variables impacting churn were successfully identified, and the findings were consistent with previous research.

The objectives of the study were achieved, providing the commissioning party with additional expertise in the subject matter. By using predictive models, more precise actions can be taken in the future to improve customer retention, and the conducted research can contribute to enhancing the company's understanding of its customer base.

Keywords/tags (subjects)

Customer churn, machine learning, predicting customer churn

Miscellaneous (Confidential Information)

n/a

Sisältö

Kuviot

Taulukot

1 Johdanto

Kun asiakas vaihtaa Elisalta Telialle, tai Netflixistä Amazon Primeen, ja päättää tilauksensa ”lopeta tilaus” -napista, on asiakassuhde loppunut. Asiakkaan ja palveluntarjoajan välinen sopimus on päättynyt ja asiakas on ”churnannut”, eli tuttavallisemmin poistunut. Kun asiakas tilaa tammikuussa kameratarvikkeiden verkkokaupasta keskiformaatin filmikameran, eikä edes kirjaudu asiakastililleen ennen seuraavaa syksyä – onko kyseessä poistunut asiakas? Tilauspalveluihin verrattaessa, verkkokaupassa tilanne ei ole yhtä mustavalkoinen. Asiakas saattaa tilata uuden filmin elokuussa, uuden objektiivin syyskuussa, tai asiakkaan ainoaksi ostokseksi jää tammikuussa tilattu keskiformaatin kamera. Asiakas voi viettää pitkiäkin aikoja niin sanottuna väliaikaisena poistujana, testata kilpailijan palveluita ja palata tämän jälleen aktiiviseksi ostajaksi. Määritelmän lopullisesta poistumisesta tekee loppupeleissä yritys itse. Mutta olisiko yrityksien mahdollista ennustaa potentiaaliset poistujat jo kertyneen tilaushistorian perusteella?

Asiakasdata on tänä päivänä yksi yrityksien tärkeimmistä omistuksista ja sitä on tarjolla enemmän kuin koskaan. Asiakasdata voi vastata kysymyksiin siitä keitä asiakkaat – tai potentiaaliset asiakkaat – ovat, mistä he pitävät, mitkä heidän tapansa ovat tai mikä heitä motivoi. Mitä enemmän dataa, sitä tarkempi on kokonaiskuva asiakkaasta (Stephenson 2018, 85). Asiakaspoistuman ennustamisessa voidaan käyttää hyväksi yrityksille kertynyttä asiakasdataa; uutiskirjetilaus sekä ostohistoria voivat toimia syötteinä, joiden pohjalta saadaan aikaiseksi ennuste asiakaspoistumasta.

Vuonna 2023 jo 15 %:alla Suomalaisista yrityksistä oli käytössään tekoälyteknologioita. Vastaava lukema yli 100 hengen yrityksissä oli jo 42 prosenttia, joka tarkoittaa, että lähes puolet yli 100 hengen yrityksistä käytti tekoälyä toiminnassaan. Vastaavasti, koneoppimista datan analysoinnissa käytti seitsemän prosenttia yrityksistä. Yli 100 hengen yrityksessä jo lähes neljännes, 24 prosenttia, oli omaksunut koneoppimisen osaksi datan analysointia. (Tietotekniikan käyttö yrityksissä 2023.) Uudet teknologiat ovat vähitellen valtaamassa alaa osana päätöksentekoa ja tietojohtamista, ja ehkäpä ne voisivat mahdollistaa myös ennakoivia toimia asiakassuhteen säilyttämiseksi?

Asiakaspoistuman ennustaminen koneoppimisen avulla on kuitenkin haastavaa, ja ennustamisen kohteena oleva ala vaikuttaa merkittävästi ennusteisiin ja niiden tarkkuuksiin. Viimeisten vuosien aikana on tuotettu kuitenkin runsaasti tutkimuksia asiakaspoistuman ennustamisesta koneoppimisen keinoin ja useita eri malleja käyttäen. Suurin osa tutkimuksista käsittelee poistumaa kuitenkin joko pankki-, telekommunikaatio- tai vakuutusalan saralla, joissa asiakkuus perustuu yhteiseen sopimukseen. Myös verkkokauppojen asiakaspoistumaa on tutkittu, ei kuitenkaan yhtä paljon kuin jo mainittujen pankki-, telekommunikaatio- tai vakuutusalan suhteen. Tämä opinnäytetyö pyrkii löytämään sopivia koneoppimismalleja verkkokaupan asiakaspoistuman ennustamiseen toimeksiantajayrityksen asiakas- ja tilausdataa hyödyntäen. Työssä tarkastellaan ja vertaillaan valittujen mallien tarkkuutta ja pyritään löytämään poistumaan vaikuttavia muuttujia. Tämä opinnäytetyö pyrkii löytämään vastauksia seuraaviin tutkimuskysymyksiin:

1. Voidaanko verkkokaupan asiakkaiden poistumista ennustaa koneoppimisen menetelmillä?

2. Mikä käytetyistä koneoppimismalleista tuottaa parhaan lopputuloksen?

3. Mitkä käytetyistä muuttujista näyttävät vaikuttavan poistuman syntymiseen?

Jos asiakkaista osataan ennustaa tarkasti mahdolliset poistujat, voidaan resursseja ohjata ja kohdentaa paremmin mahdollisiin poistujiin. Tätä tietoa voidaan käyttää hyödyksi asiakassuhteiden säilyttämisessä ja markkinointitoimenpiteissä. Vallitsevan tiedon mukaan, jo viiden prosentin parannus poistumassa voi parantaa yrityksen tulosta jopa kaksikymmentäviisi prosenttia - liiketoiminnan kannalta hyöty on kiistämätön (Reicheld 2001, 1).

Toimeksiantajana tälle opinnäytetyölle toimii vuonna 2010 perustettu tamperelainen Kameratori Oy. Kameratori myy, ostaa ja huoltaa käytettyjä kameroita sekä kameratarvikkeita, ja on yksi Euroopan suurimmista käytettyjen kameroiden kauppapaikoista. Tuotteita myydään yli 90 maahan ympäri maapalloa. Kameratori Oy on viimeisten vuosien aikana herännyt toden teolla tiedolla johtamisen aikakauteen, tästä parhaimpana esimerkkinä vuonna 2023 Kasperi Heikkilän julkaisema diplomityö: ”Liiketoimintatiedon hallinta kiertotalousliiketoiminnassa”, joka tarkastelee tiedolla johtamista niin yleisellä tasolla kuin toimeksiantajayrityksessäkin. Asiakasdatan suhteen tiedon prosessointia ja analysointia voidaan helposti parantaa, juuri vaikkapa asiakaspoistuman paremman ymmärtämisen kannalta. Opinnäytetyön tavoitteena on myös tuoda lisää ymmärrystä koneoppimisesta, ja sen käyttömahdollisuuksista, toimeksiantajayritykselle.

2 Toimeksiantaja

Opinnäytetyön toimeksiantajana on tamperelainen, vuonna 2010 perustettu Kameratori Oy, joka on käytettyjen kameroiden, objektiivien sekä niihin liittyvien lisätarvikkeiden verkkokauppa, jonka palveluita on vuosien saatossa käyttänyt jo yli 40 000 asiakasta (Tervetuloa Kameratorille n.d.). Yritys on panostanut vuosia vahvasti kiertotalouteen, ja sen työllistämät kamerateknikot korjaavat ja restauroivat sadoittain kameroita uusiokäyttöön, antaen uuden elämän jopa viisikymmentä vuotta vanhoille kameroille (Storås 2023). Kameratori Oy on kasvanut merkittävästi viime vuosien aikana, niin liikevaihdollisesti kuin henkilöstömäärältään, ja yritys tähtää myös vahvaan kasvuun tulevaisuudessa. Jopa 85–90 prosenttia Kameratorin liikevaihdosta tulee nykyisin ulkomailta ja yritys onkin tunnettu globaali toimija alallansa. (Runsas 2024.)

Kameratori on laajentanut katalogiaan kattamaan VALOI-filmiskannaustarvikkeiden valmistuksen ja myynnin sekä valmistanut SantaColor-värifilmiä kattamaan kuluttajien kysyntää (Runsas 2024). Jotta Kameratorilla olisi myydä kameroita myös tulevaisuudessa, kouluttaa yritys itse kamerateknikkoja mestari - kisälli periaatteella korjaamaan ja huoltamaan hankittuja tuotteita (mt.).

Kameratori on onnistunut kasvamaan vuosien saatossa kansainvälistä verkkokauppaa operoivaksi kiertotalouden toimijaksi, jolla on vakaa aikomus ja halu jatkaa kasvua myös tulevaisuudessa.

3 Tutkimusmenetelmä

3.1 Tutkimusasetelma

Opinnäytetyön yksi keskeisimpiä tavoitteita on tuottaa uutta tietoa valitusta ilmiöstä ja tämän tiedon tuottaminen täytyy perustua aiheellisesti valitun tutkimusmenetelmän soveltamiseen sekä systemaattiseen tiedonkeruuseen (Bister 2019, 26). Valittuun ilmiöön liittyy aina jonkin ongelma, joka pyritään ratkaisemaan ja juuri tämän tutkimusongelman määrittäminen toimii lähtökohtana kaikelle opinnäytetyössä tapahtuvalle työskentelylle (Bister 2019, 27; Kananen 2010, 18). Bister toteaa (2019, 27) luotettavan teoriapohjan rakentamisen oleva avainasemassa tutkimusongelman ratkaisemiselle ja näkee tutkimuskysymysten lähinnä ohjaavan koko työprosessia, kun taas Kananen (2010, 18) painottaa tutkimuskysymysten oikeanlaisen asettelun tärkeyttä itse tutkimusongelman ratkaisemisessa.

3.2 Tutkimusmenetelmän valinta

Opinnäytetyön tutkimusmenetelmäksi valikoitui kehittämistutkimus. Kehittämistutkimus on luonteeltaan työelämälähtöistä kehittämistyötä ja sen tuloksena syntyy tuote, tai sen osa, joissain tapauksissa tuotoksena voi olla kehittämissuunnitelma. Tuotos ei välttämättä vaikuta heti liiketoimintaan, vaan vaikutukset voivat näkyä vasta pidemmällä aikavälillä. (Bister 2019, 46.) Kehittämistutkimus kytkeytyy usein liiketoiminnan tavoitteisiin ja sen tarkoituksena on ratkaista jokin työelämälähtöinen ongelma.

Tutkimuksen empiirinen osuus toteutettiin määrällisenä, eli kvantitatiivisena, tutkimuksena toimeksiantajalta saamalla aineistolla. Määrällisessä tutkimuksessa edetään deduktiolla, eli tutkimuksessa edetään teoriasta kohti käytäntöä, joka tuottaa lukuja vastauksena kysymyksiin. (Kananen 2012, 31–32.) Kanasen mukaan (2021, 33) tiukka jako määrälliseen ja laadulliseen ei kuitenkaan ole ehdoton, vaikka työn alussa päätös käytettävästä tutkimusotteesta olisi tehty.

Kvantitatiivisen tutkimuksen pohjana ovat muuttujat, jotka ovat ominaisuuksia, joista tutkimuksessa ollaan kiinnostuneita. Muuttujien jako voidaan tehdä määrällisiin ja laadullisiin muuttujiin. (Kananen 2010, 78–79.) Kananen (2010, 78) toteaa muuttujan olevan määrällisen tutkimuksen pe-

ruskäsite, jonka ymmärtäminen vaikuttaa merkittävästi tutkimuksen tuottamaan tietoon. Tutkimuksen analyysi- ja tulkintavaiheessa voidaan kohdata ylitsepääsemättömiä ongelmia, jos käsitettä ei ymmärretä kunnolla (Kananen 2010, 78).

3.3 Tutkimuskysymykset

Bisterin mukaan (2019, 27) tutkimuskysymysten pohdinta auttaa jäsentämään opinnäytetyön ydinsisältöä. Tähän avuksi soveltuu myös miellekartta, joka auttaa keskeisten käsitteiden välisten suhteiden ymmärtämisessä. Kanasen mukaan (2010, 18) tutkimuskysymykset tulisi johtaa suoraan tutkimusongelmasta, jotta tutkimusongelma saadaan ratkaistua - kysymykset ns. kuorivat ilmiön auki. Tutkimuskysymyksen kysymysmuoto myös määrittää vastauksen, joten kysymysasettelun on oltava tarkkaa (Kananen 2010, 18). Miksi-kysymyksellä ei saada vastausta Paljonko-kysymykseen.

Bister (2019, 28) arvelee kolmen tutkimuskysymyksen oleva sopiva määrä moneen tarkoitukseen, yhden ollessa liian vähän. Useimmiten ensimmäinen kysymys on pääkysymys/peruskysymys, jota seuraavat alakysymykset (Bister 2019, 28; Kananen 2010, 17). Kysymyksiin tulee myös muistaa vastata työn lopussa, ja jos työn fokus muuttuu, tulee kysymykset sovittaa vastaamaan uudistunutta työn sisältöä (Bister 2019, 28). Hirsjärven, Remeksen ja Sajavaaran (1997, 126) mukaan perinteisen kaavan mukaan etenevässä tutkimuksessa pyritään esittämään ongelma mahdollisimman selkeästi ja tarkasti. Hirsjärvi ja muut (1997, 126) esittelevät myös osaongelmat, jotka voidaan johtaa pääongelmaa analysoimalla, ja jotka vastannevat tutkimuskysymyksiä. Tämän työn tutkimuskysymykset on esitelty johdannossa.

3.4 Luotettavuuden arviointi

Jotta tutkimus ja opinnäytetyön tulokset voisivat olla merkittäviä, tulee kiinnittää erityistä huomiota sen luotettavuuteen (Bister 2019, 61). Bister (2019, 61) mainitsee luotettavuuden suurina tekijöinä lähteiden laadun; millaisia lähteitä on käytetty ja miten paljon. Luotettavuutta heikentävinä tekijöinä voidaan pitää esimerkiksi aineiston vähyyttä, vääriä analyysimenetelmiä, virheellisiä johtopäätöksiä tai opinnäytetyön tekijästä johtuvia vaikutuksia. Tutkimuksessa käytetyn mittaus- ja tutkimusmenetelmän tulisi mitata juuri sitä ilmiön ominaisuutta, jota on ollutkin tarkoitus mitata, jotta sitä voidaan pitää validina. (Mts. 61–62.)

Kanasen (2012, 166) mukaan kehittämistutkimuksen laadun arviointi tulisi tehdä jokaisen käytetyn menetelmän luotettavuuskriteereillä, jolloin määrälliset tutkimusosat tulisi arvioida määrällisen luotettavuuskriteeristön avulla. Kananen (2012, 166) nostaa laadun arvioinnin kärkiteemoiksi mahdollisimman tarkan dokumentoinnin tutkimuksessa tehdyistä asioista; mitä tehtiin, miksi tehtiin ja miten tehtiin.

Määrällisessä tutkimuksessa luotettavuuden määrittelevät reliabiliteetti ja validiteetti, joilla tarkoitetaan tutkimuksen pysyvyyttä ja oikeiden asioiden tutkimista (Kananen 2012, 167).

Tutkimuksen ollessa reliaabeli, toistettaessa sama tutkimus, saataisiin siis samanlaiset tulokset kuin aikaisemmassa tutkimuksessakin. Ilmiöt saattavat kuitenkin muuttua ajan kuluessa, jolloin uusintamittauskaan ei takaa reliabiliteettia. Uusintatutkimusten teko on usein myös myös kallista ja vaikeaa. (Kananen 2012, 167–170.) Kananen (2012, 167) nostaa esiin myös, ettei reliabiliteetti takaa validiteettia, sillä väärä mittari tuottaa saman tuloksen myös uusintatutkimuksissa.

Validiteetin tärkein alalaji on ulkoinen validiteetti, sillä määrällinen tutkimus pyrkii yleistykseen ja ulkoinen validiteetti mittaa tutkimustulosten yleistettävyyttä. Yleistettävyydellä tarkoitetaan, että tutkimustulokset pysyvät samanlaisina samanlaisissa tilanteissa. Sisältövaliditeetilla tarkoitetaan oikeiden mittareiden käyttämistä tutkittavan asian suhteen. Käytettävät mittarit on syytä dokumentoida ja perustella hyvin ja on suositeltavaa käyttää mittareita, joiden toimivuus on todettu aikaisemmissa tutkimuksissa. (Kananen 2012, 167–170.) Kananen (2012, 170) toteaa, että on varsin vaikeaa näyttää sisältövaliditeetin toteutumista omassa työssään. Sisäisen ja ulkoisen validiteetin summana saadaan kokonaisvaliditeetti (mts. 171).

3.5 Raportointi

Opinnäytetyön raportointi noudattaa tutkimusraportin peruskaavaa (IMRD). IMRD-lyhenne koostuu sanoista Introduction (johdanto), Methods (menetelmät), Results (tulokset), Discussion (pohdinta) ja opinnäytetöiden arviointikriteerit perustuvat juuri tähän rakenteeseen (Bister 2019, 54) ja nämä rakenteet toistuvat myös tässä opinnäytetyössä. Yleiset kansainväliset tavoitteet ohjaavat opinnäytetyölle asetettuja korkeakoulutasoisen raportoinnin vaatimuksia ja yksityiskohtaisemmat, ja yksilöllisemmät, raportointiohjeet ovat saatavilla korkeakoulujen verkkosivuilta (Kananen 2012,

193). Kananen (2012, 193) nostaa esille myös dokumentaation ja raportoinnin tarkkuuden merkityksen, sillä ne vaikuttavat suoraan ulkopuolisen lukijan arvioon esitettyjen tulosten luotettavuudesta. Riittävä ja laadukas dokumentaatio varmistaa sen, että lukija ymmärtää tehtyjen ratkaisujen perustelut ja sen, että tehdyt ratkaisut ovat olleet oikeita (Kananen 2012, 194).

4 Asiakaspoistuma

4.1 Asiakaspoistuma yrityksien näkökulmasta

Asiakaspoistumasta on tullut viimeisten vuosien aikana niin sanottu kuuma peruna tutkimuskirjallisuudessa. Vuoteen 2008 asti, asiakaspoistumaa on käsitelty vain hyvin rajallisessa määrässä tieteellisiä julkaisuja - kun taas vuosien 2008 ja 2020 välillä sitä käsiteltiin 1305 eri julkaisussa, julkaisutahdin kiihtyessä kuljettaessa kohti 2020-lukua (Bhale & Bedi 2023, 3). Yritysten keräämän datamäärän jatkaessa kasvuaan, kiitos kehittyneen teknologian ja enenevissä määrin verkkoon siirtyneen kaupankäynnin, lienee melko perusteltua sanoa, että asiakaspoistuman tutkimukselle on tarvetta. Bhalen ja Bedin (2023) tutkimus korostaa aiheen merkitystä ja ajankohtaisuutta globaalissa mittakaavassa ja tarjoaa arvokkaita lähteitä pureuduttaessa asiakaspoistuman ytimeen.

Yritysten kannalta kasvava kiinnostus asiakaspoistumaa kohtaan on ymmärrettävää. Uuden asiakkaan hankkiminen voi maksaa viisi tai jopa kaksikymmentäviisi kertaa enemmän kuin olemassa olevan asiakkaan säilyttäminen (Gallo 2014), puhumattakaan Reicheldin (2001, 1) mainitsemasta kahdenkymmenenviiden prosentin tulosparannuksesta vain viiden prosentin pienennyksellä asiakaspoistumaan. Myös Bhalen ja Bedin (2023, 7) tutkimus tunnistaa asiakaspoistuman näyttelevän keskeistä roolia menestyvän liiketoiminnan osana. Tyytymättömät asiakkaat ovat alttiimpia vaihtamaan palvelun tai tuotteen tarjoajaa, joka taas merkitsee yritykselle vähemmän kassavirtaa (Bhale & Bedi 2023, 7).

Asiakaspoistuma on nousemassa varteenotettavaksi, ja tärkeäksi, yrityksen toimintaa kuvaavaksi mittariksi. Asiakaspoistuman ymmärtäminen ja sen valjastaminen yrityksen hyötykäyttöön lähtee itse termin ymmärtämisestä, jota käsitellään luvussa 4.2. Luvussa 4.3 käsitellään tarkemmin asiakaspoistumaan vaikuttavia syitä. Itse poistuman ennustamista käsitellään luvussa 4.4. Luku 4.5 käsittelee asiakaspoistumaan oleellisesti linkittyvää asiakkuudenhallintaa.

4.2 Asiakaspoistuma terminä

Asiakaspoistuma on mittari, joka kertoo niiden asiakkaiden prosenttiosuuden, jotka päättävät lopettaa asiakassuhteensa tiettynä ajanjaksona. Yksinkertaistettuna; “poistujat” ovat tilaajia, jotka peruuttavat tilauksensa tai asiakkaita, jotka eivät enää palaa kaupoille. (Danao 2023.) Poistumisnopeutta voidaan mitata vuositasolla, kuukausittain, kvartaaleittain tai tarvittaessa vaikkapa viikoittain, riippuen toimialasta ja myytävästä tuotteesta. Tilauspohjaisilla yrityksillä keskimääräinen asiakaspoistuma on noin 5 prosenttia (What is good churn rate n.d.), kun taas keskimääräinen kertaostoksiin painottuva yritys, kuten esimerkiksi kauneus- tai muotibrändi, voi kärsiä jopa 75 prosentin poistumasta (Ecommerce Churn Rate: How To Calculate and Reduce Churn 2022). Kuvio 1 selventää asiakaspoistuman laskentakaavaa.

Asiakaspoistuma = (Poistuneet asiakkaat valittuna ajanjaksona / Asiakkaita valitun ajanjakson alkaessa) x 100

Kuvio 1. Kaava asiakaspoistuman laskemiselle (Ecommerce Churn Rate: How To Calculate and Reduce Churn 2022, muokattu)

Tilauspalveluissa, esim. Netflix, Viaplay, asiakaspoistuman laskeminen on yksinkertaista - kun asiakas on lopettanut tilauksensa, on hän poistunut. Liikuttaessa verkkokaupan puolella, ei samanlaista määritystä poistumisesta voi tehdä. Onko asiakas poistunut, jos uutta tilausta ei tule kuukauden kuluessa? Yu, S. Guo, J. Guo & Huang (2011, 1426) sijoittavat asiakaspoistuman asiakkaan elinkaaren viimeiseen vaiheeseen ja määrittelevät termin tarkoittamaan tilannetta, jossa asiakas on vaihtanut palvelun/tuotteen tarjoajaa, kun taas Shobana, Gangadhar, Renjith, Bamini sekä Chincholkar (2023, 3) nostavat esiin myös termin “väliaikaiset poistujat”, jotka määritellään asiakkaiksi, joiden ostofrekvenssi on vähentynyt normaalista. Nämä, väliaikaiset poistujat, saattavat kuitenkin palata aktiivisiksi asiakkaiksi tietyn ajan kuluessa, eivätkä täten kuulu varsinaisesti asiakaspoistuman joukkoon. Shobana ja muut (2023, 3) toki nostavat esiin myös ”kokonaan menetetyt” asiakkaat, joilla tarkoitetaan lopullisesti poistuneita asiakkaita. Kokonaan menetetyt asiakkaat saattavat kuitenkin kirjautua jopa asiakastileilleen, mutta eivät enää osta yrityksen tarjoamia tuotteita tai palveluita. Transaktioiden täydellinen katoaminen toimii siis indikaattorina poistumiselle.

Tarkkaa määrittelyä “kokonaan menetetyille” asiakkaille on vaikea löytää, johtuen erilaisista tuotteista ja keskimääräisten tilausaikojen vaihtelusta. Tarkahkon määritelmän tarjoaa Shopifyn ylläpitämä blogi (Ecommerce Churn Rate: How To Calculate and Reduce Churn 2022), joka määrittelee asian seuraavasti: “*For any customer cohort, their churn rate is the percentage of customers who didn’t reorder over a timeline of two times the average repeat purchase timeline.”*, joka vapaasti suomennettuna, ja yksinkertaistettuna, tarkoittaa asiakaspoistuman olevan se prosentti asiakkaista, jotka eivät ole tilanneet tietyn määritellyn jakson aikana. Edellä mainittua kaava voi toimia lähtökohtana asiakaspoistuman laskemiselle, koska keskimääräinen tilausaika ottaa huomioon yritys- ja alakohtaiset eroavaisuudet. Täydellinen vaihtoehto tämä ei kuitenkaan ole, sillä asiakas saattaa palata vielä myöhemmin asiakkaaksi.

Jahromin, Stakhovychin ja Ewingin tutkimuksessa (2014, 1260) asiakaspoistuman määritteleminen “ei-sopimuksellisessa” ympäristössä todetaan myös haastavaksi. Jahromi ja muut (mts. 1260–1261) nostavatkin tutkimuksessaan tärkeämmäksi tietyn ajanjakson epäaktiivisten asiakkaiden ennustamisen, kuin lopullisesti poistuneiden asiakkaiden ennustamisen. Ajanjakso voidaan määritellä mielivaltaisesti, kunhan se on liiketoimintapäätösten kannalta hyödyllinen (mts. 1260). Esimerkiksi, jos asiakas ei tilaa seuraavan puolen vuoden aikana, ne asiakkaat, jotka tilasivat vain ensimmäisen puolivuotisjakson aikana, lasketaan poistuneiksi.

Tiivistettynä, asiakaspoistumalla tarkoitetaan niitä asiakkaita, jotka eivät enää aktiivisesti käytä yrityksen palveluita. Tarkempi määrittely on hyvä tehdä tapaus- ja yrityskohtaisesti liiketoiminnalliset näkökulmat huomioon ottaen. Tutkimuksista ei nouse esiin yhtä oikeaa tapaa, tai käytettyä rajaarvoa poistuman määrittämiselle, joka toimisi jokaisella alalla.

4.3 Syitä asiakaspoistumaan

Asiakaspoistumaan ei ole aina yhtä selkeää syytä - kyseessä voi olla monen asian summa. Näiden syiden ymmärtäminen on kuitenkin yritysten kannalta elintärkeää, ja yrityksen, jotka ymmärtävät asiakaspoistumaan vaikuttavia syitä, ovat paremmassa kilpailutilanteessa. Poistumisen syiden kirjo on laaja, syy voi olla heikoksi koetussa asiakaspalvelussa, korkeassa hintatasossa, tuotteen huonossa varastotilanteessa, ongelmat tuotepalautuksissa, heikkolaatuisissa tuotteissa tai kanta-asiakasohjelman puuttuminen (How to Reduce Customer Churn for Retail Succes 2023). Usein ensikontaktin työntekijät, jotka ovat suorassa vuorovaikutuksessa asiakkaan kanssa, ovat avainroolissa

asiakaspoistuman kannalta, joka kertoo asiakaspalvelun laatuun panostamisen tärkeydestä (Subramanian 2023).

4.4 Poistuman ennustaminen

Asiakaspoistuman ennustamisessa pyritään tunnistamaan potentiaaliset poistujat perustuen aikaisempaan asiakasdataan sekä käytökseen (Zhu, Baesens, Backiel & vanden Broucke 2017, 3). Zhun & muiden (2017, 1) mukaan tarkat asiakaspoistuman ennustemallit auttavat yrityksiä kehittämään tehokkaita asiakassäilyvyysohjelmia. Mitä tarkempi ennustemalli, sitä paremmin yritykset voivat sekä suunnata saatavilla olevia resursseja asiakassuhteen säilyttämiseen, että ymmärtää mahdollisia poistuman syitä.

Matuszelański ja Kopczewska (2022, 193) nostavat tutkimusartikkelissaan esiin tilattujen tuotteiden määrän vaikutuksen asiakaspoistumaan. Asiakkaat, jotka tilasivat ensimmäisellä kerralla kolme, tai sitä useamman tuotteen, jättivät suuremmalla todennäköisyydellä tilaamatta tulevaisuudessa. Matuszelańskin ja Kopczewskan (2022, 193) mukaan myös asiakkaiden ensimmäiseen tilaukseen käytetty rahasumma indikoi seuraavaa tilausta, tosin vain tiettyyn pisteeseen asti, jonka jälkeen uuden tilauksen todennäköisyys pieneni.

Miguéis, Dirk Van den Poel, Camanho sekä João Falcão e Cunha (2012, 11252) esittelevät tutkimuksessaan ennustemallin, jota varten asiakkaat luokitellaan osittain poistuneiksi, jos he sijoittuvat laskevalle kulutuskäyrälle kvartaalikulutuksensa mukaan. Miguéisin ja muiden (2012, 11252–11254) malli keskittyy ennustamaan mahdollista poistumaa asiakkaan ensiostoksen kategorian mukaan. Muita käytettyjä muuttujia olivat ostostiheys, käytetty raha sekä aika viimeisimmästä ostoksesta. Koneoppimismallina käytettiin logistista regressiota (Miguéis & muut 2012, 11252). Tutkimus toteutettiin eurooppalaisen vähittäiskauppaketjun asiakasdatan pohjalta.

Sweidan, Johansson, Gidenstam ja Alenljung (2022, 2) käyttävät tutkimuksessaan pohjana muuttujien jakamista RFM-mallin mukaisesti. RFM tulee sanoista recency, frequency ja monetary ja se voidaan suomentaa tarkoittamaan viimeaikaisuutta (recency), asiointitiheyttä (frequency) sekä rahallista arvoa (monetary). Sweidanin ja muiden tutkimus (2022) on tämän opinnäytetyön kannalta erityisen kiinnostava, koska tutkimuksessa on käytetty verkkokauppadataa ja se on verrattain

tuore. Tutkimuksen tulokset ovat myös hyvin lupaavia, poistuvia asiakkaita onnistuttiin ennustamaan kohtuullisella tarkkuudella ja huomattiin, että 75 prosenttia niistä asiakkaista, joilla on vain yksi tilaus, tulevat poistumaan (mts. 6).

Fridrichin ja Dostalin (2022, 2) mukaan ennustemallin muuttujina tulisi käyttää niin sanottuja perusominaisuuksia, jotka muodostuvat asiakkaan ja tuotteiden välisen vuorovaikutuksen ympärille, ja ovat yleisesti saatavilla kaikille verkkokaupassa toimiville vähittäismyyjille. RFM-muuttujien rinnalle tutkimuksessa nostettiin myös asiakkaan selailemat tuotekategoriat, keskimääräinen sessiopituus sekä selailuajankohta (mts. 3–4). Fridrich & Dostalin (2022, 10) tutkimus tunnistaa viimeaikaisuuden sekä asiointitiheyden tärkeinä tekijöinä poistuman kannalta.

Kuten tutkimuksista käy ilmi, useat erilaiset tekijät vaikuttavat asiakaspoistuman ennustamiseen. Useassa tutkimuksessa (Sweidan & muut 2022; Fridrich & Dostal 2022) esiin noussut RFM-malliin nojaava muuttujien valinta vaikutti olevan suosittu ja suositeltava tapa lähestyä asiakaspoistuman ennustamista. Saadut tulokset olivat myös riippuvaisia valituista ja käytetyistä koneoppimismalleista. Vaikka itse poistuman ennustaminen on binäärinen ongelma, voi lähestymisen sen suhteen tehdä usean eri koneoppimismallin ja muuttujan kautta.

4.5 Asiakkuudenhallinta (CRM)

Customer Relationship Management, tuttavallisemmin CRM ja suomeksi asiakkuudenhallinta, ei määrittelynsä puolesta ole helpoimmasta päästä. Hayleyn (2016, 26) mukaan CRM:ää ei voi tarkasti edes määritellä, koska määritelmä on täysin sidonnainen valittuun näkökulmaan. Yhteisiä piirteitä määritelmistä voi kuitenkin löytää, ne kaikki käsittelevät pitkäaikaisten asiakkuussuhteiden vaalimista menestyvän liiketoiminnan varmistamiseksi (mts. 27). Lähimmäksi kaiken kattavaa määritelmää päästään määritelmässä, jossa CRM voidaan pitää ensisijaisesti strategiana ja yritysfilosofiana, jossa asiakas asetetaan kaiken liiketoiminnan keskipisteeseen, jotta yritys voisi kasvattaa voittoja parantamalla asiakkaiden hankintaa ja säilyttämistä (mts. 27).

CRM saatetaan useasti mieltää itse teknologiaksi ja usein törmää puhuttavan CRM-alustoista. Tässä ei varsinaisesti mitään väärää ole, mutta on tärkeää selventää, ettei CRM ole teknologia sinänsä, mutta se hyödyntää teknologiaa haluttujen tavoitteiden saavuttamiseksi. (Hayley 2016, 27). Kerätty asiakasdata on teknologian avulla mahdollista integroida haluttuun alustaan ja muuttaa

hyödylliseksi tiedoksi (mts. 27). Hayleyn mukaan (2016, 27) teknologia mahdollistaa myös yrityksen vuorovaikutuksen asiakkaiden kanssa tavalla, joka tuottaa arvoa asiakkaalle helpottaen liiketoimintaa heidän kanssaan.

5 Koneoppiminen

5.1 Koneoppimisen lyhyt historia

Koneoppiminen on saattanut käsitteenä nousta suuremman yleisön tietoisuuteen vasta 2000-luvulla, mutta sen historia juontaa juurensa aina 1950-luvulle. Käsitteen keksijänä voidaan pitää matemaatikko A. L. Samuelia, joka vuonna 1959 julkaistussa teoksessaan “Some Studies in Machine Learning Using The Game of Checkers”, tutki onko konetta mahdollista opettaa pelaamaan shakkia. (Samuel 1959). Samuelin mukaan (1959, 221) kone voi oppia pelaamaan shakkia paremmin kuin ohjelman kirjoittanut henkilö.

Hypätään ajassa eteenpäin aina vuoteen 1997; maailma pidättää hengitystä suurmestari Garri Kasparovin sekä IBM:n Deep Blue supertietokoneen otellessa shakkimaailman herruudesta. Deep Blue voittaa ja maailma kohahtaa. Vaikka Deep Blue hyödynsi enemmän raakaa koodiriviä ja silkkaa laskentatehoa, edustaa se silti merkittävää virstanpylvästä tekoälylle ja koneoppimiselle. (Deep Blue n.d.)

Seuraava, todella järisyttävä virstanpylväs, koettiin vuonna 2016 tekoälypohjaisen AlphaGo:n voitettua GO-pelin suurmestarin Lee Sedolin GO-pelissä. Sakkilauden koon ollessa 8 x 8, on GO-lauta kooltaa 19 x 19, joka mahdollistaa suuremman määrän siirtoja kuin universumissa on atomeja (AlphaGo n.d.). Kun Deep Blue ohjelmoitiin pelaamaan parhaalla mahdollisella tavalla, AlphaGo koulutti itse itsensä pelaamalla pelejä, ja päivittämällä itseään niistä saaduilla kokemuksilla (Alpaydin 2021, 98). Kone oli oppinut pelaamaan peliä paremmin kuin hallitseva suurmestari ja uusi aikakausi oli alkanut.

Koneoppimista voidaan kutsua tekoälyn alalajiksi, ja yksinkertaisimmillaan sillä tarkoitetaan järjestelmän autonomista oppimista datan avulla. Valitulle koneoppimismallille syötetään tietoa, jonka avulla pyritään ennustamaan/tunnistamaan haluttu lopputulema. Mitä enemmän dataa valitulle

koneoppimismallille syötetään, sen paremmin se osaa hienosäätää algoritmejaan, parantaen mallin suorituskykyä. Karkeasti jaotellen koneoppisessa käytetyt mallit voidaan kolmen eri kategorian alle: ohjattuun oppimiseen, ohjaamattomaan oppimiseen, sekä vahvistusoppimiseen. (What is machine learning? n.d; What is Machine Learning (ML) n.d.)

5.2 Ohjattu Oppiminen

Ohjattu oppiminen, joka tunnetaan myös termillä supervised learning, on koneoppiminen kategoria, jossa hyödynnetään merkittyjä datakokoelmia, joiden avulla algoritmi koulutetaan luokittelemaan dataa tai ennustamaan tuloksia. Mallia voidaan käyttää vaikkapa a-kirjaimen tunnistamiseen kuvasta. Haluttu ”tulos” on tiedossa, ja tämän saamiseksi mallille syötetään opetusdatana kuvia halutusta tuloksesta. Jos haluaa tunnistaa A-kirjaimen, mallille syötetään kuvia A-kirjaimesta. Mitä suurempi lähdemateriaali on käytössä, sitä parempia ovat myös koulutettu algoritmit ja tulokset. Kertaalleen koulutetulle algoritmille voidaan syöttää uutta dataa, jonka lopputuleman se ennustaa pohjautuen vanhoihin kokemuksiinsa. Yleisiä ohjatun oppimisen malleja ovat Lineaarinen Regressio, Naiivi bayesilainen luokittelu, Päätöspuut sekä K-lähintä naapuria. (What is Machine Learning (ML)? n.d.)

Kelleher ja Tierney (2021, 100–102) määrittelevät ohjatun oppimisen tavoitteeksi funktion oppimisen, jolla tapausta kuvaavien piirteiden arvot vastaavat tapauksen kohdepiirteen arvoja mahdollisimman hyvin. Esimerkkinä Kelleher ja Tierney (2021, 100) käyttävät roskapostisuodattimen mallia, jossa algoritmi pyrkii oppimaan funktion, joka kuvaa sähköpostia kuvaavat piirteet joko roskapostiksi tai ei roskapostiksi.

Ohjattu oppiminen edellyttää, että tietoaineiston jokainen tapaus on nimettävä kohdepiirteen arvolla – usein syy kohdepiirteestä kiinnostumiselle on kuitenkin se, ettei kyseistä arvoa voida mitata suoraan. Tämä aiheuttaa sen, että datan esikäsittely vie paljon aikaa ennen kuin ohjattua oppimista voidaan käyttää malliin. (Kelleher & Tierney 2021, 102.)

Kanasen ja Puolitaipaleen mukaan (2016, 50) ohjattu oppiminen on kaikkein suosituin koneoppimisen muoto liiketoiminnassa ja sitä voidaan käyttää useaan eri tarkoitukseen; aina kuvien tunnistamisesta, laadunvalvontaan ja suosittelukoneisiin. Kananen ja Puolitaival (2016, 50) mainitsevat

hyvänä esimerkkinä kaupan todennäköisyyksien ennustamisen myyntivihjeiden pohjalta. Tavoitearvoina täytyy vain tietää, syntyikö kauppa vai ei, jonka pohjalta algoritmia voidaan kouluttaa.

Ohjatussa oppimisessa koulutusdatalla on äärimmäisen suuri merkitys mallin toimivuuden kannalta. Koska dataa joudutaan usein luokittelemaan manuaalisesti mallia varten, on riskinä datan virheellinen luokittelu. Mikäli malli koulutetaan virheellisesti luokitellulla datalla, tekee kone tästä omat virheelliset tulkintansa. Tarkkuus voi näyttää päällisin puolin hyvältä, mutta luokittelut saattavat olla todellisuudessa vääriä. (Kananen & Puolitaival 2016, 49.)

5.3 Ohjaamaton Oppiminen

Ohjaamaton oppiminen (unsupervised learning) on koneoppimismalli, jossa käytetään ei-strukturoitua dataa mallien tunnistamiseen - eli valmiiksi merkittyjä suuria tietojoukkoja ei tarvita. Suurena erona ohjattuun oppimiseen on myös se, ettei ns. oikeaa toistettavaa vastausta ole. Tavoitteena on, että kone organisoi datan itsenäisesti, tunnistaa poikkeukset, ja löytää säännönmukaisuudet (Kananen & Puolitaival 2016, 51). Kämäräinen (2023, 96) toteaa Internetin olevan pullollaan merkitsemätöntä ja ”villiä dataa”, esimerkkinä kuvat, joista ohjaamaton oppiminen voi oppia ilman ihmistyötä, jolloin datan manuaalinen merkitseminen, eli datan kuratointi, voidaan jättää välistä.

Algoritmi oppii datasta ilman ihmisen ohjausta ja kategorisoi, sekä ryhmittelee, sitä ominaisuuksien mukaan. Algoritmien käyttämä oppimisprosessi perustuu toistuvasti esiintyvien kaavojen tunnistamiseen datan seasta. Jos algoritmille syötetään kasa kuvia olutpulloista ja pähkinöistä; rupeaa se lopulta erottamaan kuvia löytämiensä ominaisuuksien perusteella. Kanasen ja Puolitaipaleen (2019, 52) mukaan ohjaamattoman oppimisen avulla datasta voidaan löytää sellaisia yhteneväisiä ominaisuuksia, jotka ihmiseltä jäisivät helposti havaitsematta. Mallia, ja sen hyperparametreja, voidaan säätää myös siten, että ne algoritmi kiinnittää huomiota tarkemmin eri ominaisuuksiin tai jättää toisia täysin huomioimatta (Kananen & Puolitaival 2019, 53).

Yleisimpiä käytettyjä ohjaamattoman oppimisen algoritmeja ovat K-Means klusterointi, Hierarkinen klusterointi sekä pääkomponenttianalyysi (PCA). Ihanteellisia käyttökohteita ohjaamattomalle

oppimiselle ovat esimerkiksi kuvantunnistus sekä asiakassegmentointi. (What is machine learning? n.d; What is Machine Learning (ML) n.d.; What is machine learning? 2017.)

Eräs globaalisti merkittävimmistä ohjaamattoman oppimisen algoritmeista on PageRank-algoritmi, joka pyöri Stanfordin yliopiston palvelukoneella nimellä google.stanford.edu. Vuonna 1998 Computer Networks and ISND Systems-lehdessä julkaistu PageRank-artikkeli on hyvästä syystä edelleen lehden ladatuin ja viitatuin algoritmi. (Kämäräinen 2023, 97). PageRank-algoritmi loi pohjan Google-nimisen yrityksen menestykselle arvottamalla verkkosivuja.

5.4 Vahvistettu Oppiminen

Vahvistettu oppiminen (reinforcement learning) on koneoppimismalli, jossa malli oppii erehdyksen ja virheen kautta. Malli oppii suorittamaan määritellyn tehtävän niin sanotun “palautesilmukan” avulla, kunnes sen suorituskyky on toivottavalla tasolla. Malli saa positiivista palautetta, kun se suoriutuu tehtävästä hyvin ja negatiivista palautetta, kun suoriutuminen on huonoa. Esimerkki vahvistusoppimisesta on koneen kuljettaminen paikasta A paikkaan B. Päämäärä on selkeästi asetettu, mutta reittiä ei ole määritelty. Kone saattaa kohdata matkalla muuttujia, sortuneita tienpätkiä, hevosia tai muita koneita, jotka tekevät tehtävästä kaikkea muuta kuin yksinkertaisen. Malli suorittaa tehtävän, jää jumiin kymmenen kertaa ja saa tehtävästään negatiivista palautetta. Seuraavalla kerralla malli jää jumiin vain kerran ja saa positiivista palautetta. Lopulta malli oppii valitsemaan saatujen miinus- ja pluspisteiden pohjalta parhaimman mahdollisen suoritustavan ja maksimoi pisteet. Algoritmi on vihdoin oppinut mitä kannattaa tehdä ja mitä kannattaa olla tekemättä. (What is machine learning (ML)? n.d.)

Vahvistusoppiminen on koneoppimisen muodoista kaikista hienostunein, mutta samalla sen käytännön hyödyntäminen on ollut haastavampaa kuin ohjaamattoman- ja ohjatun oppimisen (Kananen & Puolitaival 2019, 158). Vaikka metodi ei vaadi paljoa dataa opettamiseen, on Kanasen ja Puolitaipaleen (2019, 159) mukaan vahvistusoppimien käyttö liike-elämässä vaikeasti hyödynnettävää, sillä useimmat algoritmit oppivat menestyksekkäästi vain silloin kun tavoitteet, olosuhteet ja säännöt ovat muuttumattomia. Vahvistusoppimista on tästä huolimatta hyödynnetty onnistuneesti kuitenkin jo itseohjautuvissa autoissa sekä dynaamisessa hinnoittelussa. (Kananen & Puolitaival 2019, 159).

5.5 Käytetyt koneoppimismallit ja niiden valinta

Vaikka asiakaspoistuman ennustamista koneoppimisen keinoin on tutkittu useasta eri näkökulmasta eri aloilla, ei konsensusta ole löydetty siitä, mikä algoritmi olisi yksiselitteisesti paras asiakaspoistuman ennustamiseen Jain, Yadavin & Rajapandyn (2021, 155) mukaan Logistinen regressio on paras algoritmi IT-sektorilla 90% tarkkuudella, satunnaismetsä finanssialalla 86% tarkkuudella ja XGBoost televiestintäalalla 83% tarkkuudella, kun taas Geilerin, Affeldtin & Nadifin (2022, 14-20) mukaan tarkimmin toiminut malli lähes jokaisessa aineistossa oli satunnaismetsän, logistisen regression sekä XGBoostin yhdistelmä. Wagh, Andhale, Wagh, Pansare, Ambadekar & Gawande (2024, 15) havaitsivat päätöspuun saavan huonoja tuloksia, mutta vastaavasti satunnaismetsä suoriutui poistujien luokittelusta televiestinnän alalla lähes 99 prosentin tarkkuudella. Vähittäiskaupan alalla suoritetussa tutkimuksessa tarkimmaksi ennustemalliksi todettiin Gradient Boost, kun huonoiten suoriutui päätöspuu (Sweidan, Johansson, Gidenstam & Alenljung 2022). Sweidan & muut (2022, 4) lisäävät päätöspuun kuitenkin sopivan hyvin poistuman ennustamiseen, koska sen rakennetta voidaan tulkita ja ymmärtää logiikka ennustusten takana. Tutkimuksissa löytyy myös yhteneväisyyksiä, sillä Fridrichin ja Dostalin (2022, 10) tutkimus nostaa esille jo edellä mainitun Gradient Boost-menetelmän toimivana mallina asiakaspoistuman ennustamiseen. Kyseinen tutkimus mainitsee tämän lisäksi hyvin suoriutuviksi malleiksi myös Logistisen Regression sekä tukivektorikone-mallin.

Tutkimuksien, ja aineistojen vertailu suoraan keskenään on kuitenkin melko hankalaa johtuen aineistojen ja käytettyjen mallien eroavaisuuksista. Tutkittavat aineistot ovat useilta eri aloilta, ja jopa samaa alaa tutkivat aineistot saattavat olla aineistojen ja käytettyjen muuttujien osalta huomattavan erilaisia. Ahn, Hwang, Kim, Choi & Kang (2022) näkevät, että saatavilla olevan loki- ja asiakasdatan määrä määrittää käytettävää ennustamismallia, ja koska jokaisen yrityksen käyttämät lokitietotyypit ja määrät vaihtelevat - ovat tietyt mallit soveltuvampia tietyille aloille. Vaikka suositellusta mallista ei päästä täyteen yhteisymmärrykseen, tutkimuksissa toistuu kuitenkin tärkeä pääsanoma; asiakaspoistuman ennustamisella on ensisijaisen tärkeä rooli asiakassuhteiden hallinnassa toimialasta riippumatta (Fridrich & Dostal 2022; Lazarov & Capota 2007). Keinoja tämän toteuttamiseen on useita ja uusia algoritmeja sekä yhdistelmämalleja kehitetään jatkuvasti lisää.

Taulukossa 1 on eritelty tutkimuksissa asiakaspoistuman ennustamiseen käytettyjä koneoppimismalleja, taulukosta käy ilmi myös tutkimuksen toimiala.

Taulukko 1. Tutkimuksissa käytettyjä ennustemalleja ja tutkimusten toimialat

Artikkeli & julkaisuvuosi	Toimiala	Käytetyt ennustemallit
Sweidan, Johansson, Gidenstam & Alenljung, 2022	Vaatekauppa	Gradient Boosting, Logistinen Regressio, Päätöspuu
Jain, Yadav, Rajapandy, 2021	IT, Televiestintä, Finanssiala	Logistinen Regressio, Satunnaismetsä, Tukivektorikone, XGBoost
Miguéis, Dirk Van den Poel, Camanho & João Falcão e Cunha, 2012	Vähittäiskauppa	Logistinen Regressio
Matuszelański, K & Kopczewska, 2022	Vähittäiskauppa	XGBoost, Logistinen Regressio
Wagh, Andhale, Wagh, Pansare, Ambadekar & Gawande, 2024	Televiestintä	Päätöspuu, Satunnaismetsä
Jahromi, Stakhovych & Ewing, 2014	B2B	Päätöspuu, Logistinen Regressio, Boosting
Geiler, Affeldt & Nadif, 2022	Finanssiala	Bayssian Classifier, Logistinen Regressio, Tukivektorikone, Päätöspuut, XGBoost

Käytettäviksi malleiksi valikoituivat logistinen regressio sekä päätöspuupohjainen satunnaismetsä. Molempia malleja oli käytetty useassa tutkimuksessa, ja ne ovat yksiä yleisimpiä käytössä olevia koneoppimismalleja.

5.5.1 Logistinen Regressio

Logistinen regressio on yksi yleisimpiä käytössä olevista koneoppimismalleista. Sitä käytetään löytämään muuttujien välisiä yhteyksiä ja ennustamaan niiden perusteella binaarisia vastauksia haluttuihin kysymyksiin. Vakuutusyhtiö voi käyttää logistista regressiota ennustamaan/arvioimaan onko kyseessä luottopetos, tai vaihtoehtoisesti terveydenhuollossa voidaan arvioida, onko asiakas jonkin taudin osalta riskiryhmässä. (What is logistic regression n.d.b.)

Logistiset regressiomallit voidaan jakaa kahteen eri ryhmään, on olemassa binäärinen logistinen regressio, sekä multinominen logistinen regressio. Binäärisissä malleissa muuttuja on aina kaksiluokkainen, kun taas multinomisissa malleissa muuttujien määrää ei ole rajoitettu. Multinomista logistista regressiomallia voi käyttää periaatteessa aina binäärisen mallin sijasta. (Nummenmaa 2004, 319.)

Koska logistinen regressiomalli on matemaattisesti vähemmän monimutkainen, kuin useampi muu koneoppimismalli, voidaan se ottaa käyttöön melko matalalla kynnyksellä. Malli on yksinkertaisuutensa vuoksi myös nopeahko käsittelemään suuria määriä dataa ja sen vahvuuksiksi luetaankin sen nopeus. Logistista regressiota voidaan käyttää myös hyväksi datan esikäsittelyssä, ennen kuin käyttöön otetaan joku monimutkaisempi koneoppimismalli. (What is logistic regression n.d.a.)

Logistisen regression saavuttamaa suosiota on helppo ymmärtää. Mallia voidaan pitää ns. aloitustason mallina, mutta sitä voidaan silti käyttää tehokkaasti laajalla rintamalla eri sovellutuksia

5.5.2 Satunnaismetsä

Satunnaismetsä on yksi suosituimmista nykyisin käytettävistä päätöspuumalleista ja se soveltuukin hyvin logistisen regression rinnalla käytettäväksi perusmalliksi, ennen siirtymistä monimutkaisempiin algoritmeihin (Alpaydin 2021, 99; Random forests n.d.).

Kämäräisen (2023, 147) mukaan päätöspuu on oppimismenetelmänä erityisen tehokas ja sietää virheitä opetusnäytteissä. Toisinaan päätöspuu saattaa kärsiä kuitenkin ylioppimisesta ja opetusai-

neiston luokittelutarkkuus paranee kohti täydellistä, tämä voidaan kuitenkin havaita testausaineistolla, jonka luokittelutarkkuus alkaa heikentyä ylioppimispisteen jälkeen (Kämäräinen 2023, 147–148).

Päätöspuut itsessään koostuvat päätössolmuista, oksista ja lehdistä, joiden toimintaa voidaan kuvata seuraavasti: solmukohta on ominaisuus, haara tätä seuraavaa päätössääntö ja jokainen lehti päättelyn tulos. (Kananen & Puolitaival 2019, 125). Puiden tulkinta on yksi päätöspuiden käytön eduista, puu on mahdollista muuntaa jos-niin säännöiksi, joita lukijan on helppo ymmärtää (Alpaydin 2021, 97). Päätöspuiden opetus on myös parametritöntä, ne kasvavat tarpeen vaatiessa eikä ennakolta oletettua rakennemallia ole. Yksinkertaisen tehtävän puu on pieni, kun taas monimutkaisemmassa tehtävässä puun koko voi kasvaa merkittävästi. (Alpaydin 2021, 98.) Kelleher ja Tierney (2021, 137–138) toteavat päätöspuiden toimivan hyvin nominaalisella ja ordinaalisella aineistolla. Nominaalisella tarkoitetaan kategorisia piirteitä, josta esimerkkinä voisi olla viinin eri kategoriat kuivasta makeaan. Ordinaalinen piirre voi muistuttaa hyvinkin nominaalista, erona on kuitenkin se, että ordinaaliset piirteet voidaan luokitella arvon mukaan. Esimerkkinä ordinaalisesta käy hyvin kyselyn vastausta kuvaavat piirteet: samaa mieltä, eri mieltä, todella eri mieltä. (Kelleher & Tierney 2021, 50–51.) Haasteita voi ilmetä, kun aineisto muuttuu numeraaliseksi. Ongelmien kertyminen voidaan kuitenkin estää muuntamalla numeeriset piirteet ordinaalisiksi. (Kelleher & Tierney 2021, 137–138.)

Satunnaismetsässä mallin toiminta perustuu useisiin käytössä oleviin päätöspuihin, joita opetetaan satunnaisesti valituilla opetusaineiston osajoukoilla. Käytössä olevat päätöspuut kasvavat tilanteen niin vaatiessa: monimutkaisen tehtävän puu on suurempi kuin yksinkertaisen. (Alpaydin 2021, 98.) Lopullinen mallin ennustus tapahtuu enemmistön äänillä; päätöspuiden ennustukset yhdistetään äänestämällä (Alpaydin 2021, 98).

5.6 Käytettävien mallien arviointi

Koneoppimismallien käyttöön liittyy vahvasti myös mallien arviointi. Jos halutaan luokitella tai ennustaa asioita, tulisi ainakin tietää kuinka hyvin luokittelut tai ennusteet osuvat kohdilleen. Kanasen & Puolitaival (172–173) jakavat mallien antamat tulokset neljään eri kategoriaan:

- Pieni harha ja pieni varianssi, mitatut arvot ovat lähellä todellisia arvoja ja pienessä kasassa

- Pieni harha ja suuri varianssi, mitatut arvot ovat lähellä todellisia arvoja, mutta hajallaan
- Suuri harha ja pieni varianssi, mitatut arvot ovat kaukana todellisista arvoista, mutta pienessä kasassa
- Suuri harha ja suuri varianssi, mitatut arvot ovat kaukana todellisista arvoista sekä hajallaan

Jos mallin arvot ovat pienessä kasassa ja päällekkäin todellisten arvojen kanssa, on tilanne ihanteellinen ja käytetty malli toimii hyvin. Jos mallin antamat tulokset ovat huonoja, tulee niitä analysoida ja selvittää millä tavalla tulokset hajaantuvat. Tämän tiedon avulla voidaan määritellä, kuinka mallin toimintaa voidaan korjata. (Kananen & Puolitaival 2019, 173.)

5.6.1 Luokittelu

Yksinkertainen keino koneoppimismallin suorituskyvyn kuvaamiseen on käyttää luokittelijaa, joka jakaa asioita kahteen ryhmään. Luokittelija jakaa asioita oikein tunnistettuihin (true positive), väärin tunnistettuihin (false positive), oikein hylättyihin (true negative) sekä väärin hylättyihin (false negative). (Kananen & Puolitaival 2019, 174.) Asiakaspoistuman kohdalla luokittelijan käyttämien arvojen logiikkaa selvennetään taulukossa 2.

Taulukko 2. Luokittelumatriisi

True Positive Todellisuus: Asiakas poistunut Malli ehdottaa: Asiakas poistunut	False Positive Todellisuus: Asiakas ei poistunut Malli ehdottaa: Asiakas poistunut
False Negative: Todellisuus: Asiakas poistunut Malli ehdottaa: Asiakas ei poistunut	True Negative Todellisuus: Asiakas ei poistunut Malli ehdottaa: Asiakas ei poistunut

Näiden arvojen pohjalta voidaan laskea mallin luokittelutäsmäävyys kuviossa 2 esitetyn laskukaavan mukaan (Kananen & Puolitaival 2019, 175). Luokittelutäsmäävyys on siis tosien positiivisten ja

tosien negatiivisten osuus kaikista luokitteluista (Alpaydin 2021, 74). Vastaavasti tästä voidaan johtaa luokitteluvirheen kaava, joka on väärät negatiiviset ja väärät positiiviset jaettuna kaikkien luokiteltujen määrällä (mts. 74).

$$\text{Luokittelutäsmäävyys} = \frac{\text{Todet Positiiviset} + \text{Todet Negatiiviset}}{\text{Todet Positiiviset} + \text{Väärät Positiiviset} + \text{Todet Negatiiviset} + \text{Väärät Negatiiviset}}$$

Kuvio 2. Luokittelutäsmäävyys

Pelkkä luokittelutäsmäävyys on kuitenkin yksin huono mittari kertomaan mallin hyvästä toimivuudesta, varsinkaan jos luokkien koko on merkittävästi epätasapainossa. Väärien päätösten arvo voi vaihdella myös rajusti toimialasta riippuen. Väärä negatiivinen päätös esimerkiksi lääketieteellisessä diagnoosissa on huomattavasti kalliimpi kuin väärä positiivinen, koska tällöin hoitoon pääsy myöhästyy ja tauti voi jäädä kokonaan diagnosoimatta. (Alpaydin 2021, 72.) Se, mikä on oikea suhde väärien positiivisten ja väärien negatiivisten välillä, on tapauskohtaista. Siihen vaikuttaa vahvasti toimiala, jolla operoidaan, mutta yleisesti ottaen väärät negatiiviset arvot ovat ei halutuimpia ennusteita. (Kananen & Puolitaival 2019, 178.)

5.6.2 Täsmäävyys

Täsmäävyys (precision) tarkastelee tosien positiivisten ja tosien positiivisten ja väärien positiivisten välistä suhdetta. Esimerkkinä voidaan käyttää roskapostia; jos malli ennustaa kuusi viestiä roskapostiksi, joista oikein ennustettuja olisi kolme, olisi mallin täsmäävyys 50 prosenttia. Kuviossa 3 on esitelty täsmäävyyden laskentakaava. Täsmäävyys voi saada minkä tahansa arvon nollan ja yhden väliltä. Jos täsmäävyys on yksi, ovat kaikki positiiviksi luokitellut tulokset tosia. (Alpaydin 2021,

75.) Yksinkertaistettuna täsmäävyys vastaa kysymykseen siitä, kuinka suuri tunnistetuista positiivista oli oikein (Kananen & Puolitaival 2019, 176).

$$\text{Täsmäävyys} = \frac{\text{Todet Positiiviset}}{\text{Todet Positiiviset} + \text{Väärät Positiiviset}}$$

Kuvio 3. Täsmäävyyden laskentakaava

5.6.3 Osumien määrä

Osumien määrä on tosien positiivisten ja kaikkien positiivisten välinen suhde (Alpaydin 2021, 75). Vaikka luokittelija tunnistaisi asiakaspoistujat 80 prosentin tarkkuudella, ei sitä yksinään voi pitää vielä merkkinä mallin toimivuudesta. Tämän lisäksi täytyy ymmärtää täsmäävyyden ja osumien määrän välinen suhde, joka voidaan laskea F1 scorella. Osumien määrän laskentakaava on esiteltynä kuviossa 4. Jos osumien määrä on yksi, on kaikki todet positiiviset tunnistettu, mutta mukana saattaa olla myös vääriä positiivia arvoja. (Alpaydin 2021, 76.) F1 scoren kaava esitelty kuviossa 5.

$$\text{Osumien määrä} = \frac{\text{Todet Positiiviset}}{\text{Todet Positiiviset} + \text{Väärät Negatiiviset}}$$

Kuvio 4. Osumien määrän kaava

$$\text{F1 Score} = \frac{2 \times \text{Todet Positiiviset}}{2 \times \text{Todet Positiiviset} + \text{Väärät Positiiviset} + \text{Väärät Negatiiviset}}$$

Kuvio 5. F1 Score-kaava

5.6.4 ROC-kuvaaja & AUC

ROC-kuvaaja, receiver operator characteristics, auttaa tulkitsemaan todennäköisyyksiin perustuvan testin luokittelukykyä. Kuvaaja muodostetaan oikein tunnistettujen määrän ja väärien tunnistettujen määrän funktiona. ROC-kuvaajan alle jäävästä alueesta käytetään nimitystä AUC (Area

Under The ROC-Curve) ja AUC-tunnusluvun avulla voidaan nopeasti hahmottaa, kuinka hyvästä luokittelijasta on kysymys. (Kananen & Puolitaival 2019, 179.) Mitä lähempänä kuvaaja on vasenta ylänurkkaa, sen tarkempi se on. Arvoja AUC voi saada nollasta yhteen, jos arvo on 0.1 on tarkkuus kymmenen prosentin luokkaa. Jos arvo on yksi, ollaan 100 prosentin tarkkuudessa. 0.5 arvolla saadaan kolikonheittoa vastaava tarkkuus. (Classification: ROC Curve and AUC n.d.)

6 Opinnäytetyön toteutus

6.1 Asiakasdata

Tässä projektissa käytetty asiakasdata on koostettu kahdesta erillisestä tietokantataulusta. Ensimmäinen tauluista on ”customers”, joka koostuu seitsemästä eri muuttujasta ja on saatu toimeksiantajayrityksen edustajalta csv-muodossa. Aineisto on alun perin kerätty toimeksiantajayrityksen verkkokaupan asiakasdatasta. Aineisto on valmiiksi sanitoitu toimeksiantajan puolesta, eli siitä on siivottu pois henkilökohtaiset yksityiskohdat, jotka mahdollistaisivat yksilöiden tunnistamisen. Aineistosta on valmiiksi karsittu pois myös joitain toimeksiantajan ja opinnäytetyön tekijän mielestä tarpeettomia muuttujia. Käytetty aineisto on vuosilta 2022–2024.

Kuviossa 6 on esitelty asiakasdata-taulun rakennetta.

```
Customer ID                    object
Accepts Email Marketing         int64
Default Address Country Code    int64
Total Spent                   float64
Total Orders                    int64
Tags                           object
dtype: object
Index(['Customer ID', 'Accepts Email Marketing',
       'Default Address Country Code', 'Total Spent', 'Total Orders', 'Tags'],
      dtype='object')
```

Kuvio 6. Asiakasdata-taulun rakenne

Kategorisia muuttujia taulussa on viisi, “Customer ID”, “Accepts Email Marketing”, “Default Address City”, “Default Address Country Code” sekä “Tags”. Numeraalisia muuttujia ovat “Total Spent” sekä “Total Orders”. Taulukossa 3 on esitelty muuttujat.

Taulukko 3. Asiakasdata-taulun muuttujat ja selitykset

Customer ID	Asiakkaan uniikki id, kategorinen muuttuja
Accepts Email Marketing	Hyväksyykö asiakas sähköpostimarkkinoinnin, Totuusarvo, joko faulty (ei hyväksy) tai true (hyväksyy). Kategorinen muuttuja.
Default Address City	Kaupunki, jonka asiakas on merkinnyt asuinpaikakseen. Kategorinen muuttuja
Default Address Country Code	Kategorinen muuttuja, asiakkaan maakoodin lyhenne
Tags	Asiakkaan liitetyt tunnisteet. Kategorinen muuttuja.
Total Spent	Asiakkaan käyttämä rahasumma. Numeerinen muuttuja.
Total Orders	Asiakkaan tekemien tilauksien kokonaismäärä. Numeerinen muuttuja.

Uniikkeja arvoja on eniten “Default Address City” muuttujassa. Tyhjiä arvoja on eniten “Tags” muuttujassa, mutta niitä on merkittävät määrät niin “Default Address City” kuin “Default Addres Country Code” muuttujissakin.

6.2 Tilausdata

Tilausdatassa muuttujia on 12. Tilausdata oli ennakolta suodatettu käsittämään vain verkkokaupan asiakkaat, jotka eivät ole B2B asiakkaita. Asiakkaalla saattaa olla useampi kuin yksi tilaus, jolloin sama asiakas voi toistua usealla rivillä Kuviossa 7 on esitelty tilausdatan rakenne ja muuttujatyypit. Kategorisia muuttujia on viisi ja numeerisia seitsemän kappaletta. Tilausdata on otettu ajallisesti samalta aikajaksolta asiakasdatan kanssa. Taulukossa 4 on eriteltynä tarkemmin muuttujien nimiä, sekä niiden merkityksiä.

```
Order status              object
Customer Id                int64
Order Id                   int64
Referrer Source           object
Customer Type             object
Day                       object
Year                       int64
Quarter                   object
Gross Sales              float64
Discounts                float64
Returns                  float64
Units Per Transaction      int64
dtype: object
Index(['Order status', 'Customer Id', 'Order Id', 'Referrer Source',
       'Customer Type', 'Day', 'Year', 'Quarter', 'Gross Sales', 'Discounts',
       'Returns', 'Units Per Transaction'],
      dtype='object')
```

Kuvio 7. Tilausdatan rakenne ja muuttujat

Taulukko 4. Tilausdatan muuttujien selitykset

Order status	Kertoo tilauksen statuksen, onko palautettu, maksettu, maksettu osittain
Customer ID	Asiakas id
Order ID	Tilaus id
Referrer Source	Viittaajan lähde, kertoo kuinka asiakas on löytänyt verkkokauppaan
Customer Type	Palaava vai ensiasiakas
Day	Tilauksen päivämäärä
Year	Tilauksen vuosi
Quarter	Tilauksen kvartaali

Gross Sales	Tilauksen arvo ilman toimitusmaksuja ja alennuksia
Discounts	Alennuksien määrä tilauksessa
Returns	Palautettujen tuotteiden rahallinen arvo tilauksessa
Units Per Transaction	Tilattuja tuotteita per tilaus

6.3 Datan esikäsittely

Data harvemmin on sellaisenaan syötettävissä koneoppimismalleille, vaan vaatii yleensä esikäsittelyä. Esikäsittelyn on tarkoitus varmistaa, että koneoppimismallille syötettävä data on asianmukaisessa muodossa. Syötetyn datan laadulla on selkeä yhteys käytetyn algoritmin suoriutumiseen. Esikäsittelyyn kuuluu tarvittavien kirjastojen tuominen, aineiston lataaminen, tyhjien arvojen tarkistaminen, poikkeavuuksien käsittely sekä datan normalisointi. (Data Preprocessing in Python n.d.)

6.3.1 Puuttuvat arvot

Puuttuvia arvoja oli asiakasdatassa kolmessa sarakkeessa. Sarake “Default Addres City” pudotettiin joukosta, koska maantieteellinen sijainti saadaan selville myös Default Addres Country Code sarakkeesta. Puuttuvien arvojen kohdalla täytettäväksi tuli siis kaksi saraketta. Molempien sarakkaiden osalta päädyttiin puuttuvat arvot korvaamaan joko “no tag” tai “unknown” arvoilla.

Tilausdatassa tyhjiä käsiteltäviä arvoja oli vain Customer Type sarakkeessa, ja tässäkin sarakkeessa vain kahden arvon verran. Tarkemmassa tarkastelussa selvisi, että kyseessä oli saman asiakkaan kaksi tilausta, joista ensimmäinen sai arvon 0 (ensitilaus) ja toinen arvon 1 (palaava tilaaja). Tyhjien arvojen lukumäärä on nähtävillä kuviossa 8.

```
None
Order status             0
Customer Id              0
Order Id                 0
Referrer Source          0
Customer Type            2
Day                      0
Year                     0
Quarter                  0
Gross Sales              0
Discounts                0
Returns                  0
Units Per Transaction    0
```

Kuvio 8. Tilausdatan tyhjät arvot

6.3.2 Ääriarvojen käsittely

Koska datassa saattaa olla asiakkaita, joiden tilausmäärä on nolla, poistettiin nämä rivit datasta. Jos tilauksia on nolla, ei asiakassuhdetta ole syntynyt eikä asiakaspoistuman ennustaminen ole hedelmällistä. Esikäsittelyn tässä vaiheessa riveistä oli poistettu noin 31 prosenttia, joka tarkoittaa, että noin 69 prosenttia asiakastilin luoneista on tehnyt vähintään yhden tilauksen. Tilausdatassa oli myös useita rivejä customer id nollalla, jotka poistettiin aineistosta.

Myös kaikki ensimmäiset tilaukset vuodelta 2024 poistettiin, koska niiden osalta ei vielä voitu laskea poistumaa. Units Per Transcation sarakkeesta suodatettiin pois joukko ääriarvoja, jotka poikkesivat merkittävästi normaalijakaumasta. Todennäköisesti arvoja oli sekoittamassa vielä joukko B2B-tilauksia, koska B2B-ominaisuus tuli käytettyyn verkkokauppa-alustaan vasta vuoden 2023 puolella – eli kaikki ennen vuotta 2023 tehdyt B2B-tilaukset näkyivät normaaleina tilauksina. Aineistosta poistettiin myös tilaus statukseltaan pending, partially paid sekä voided-tiloissa olleet tilaukset. Quarter sarake pudotettiin pois myös lopullisesta taulusta.

6.3.3 Muuttujien käsittely

Asiakasdatassa kolmen käytettävän sarakkeen arvot olivat vielä kategorisina muuttujina, jotka muutettiin numeeriseen muotoon. Accepts Email Marketing sarakkeen arvot “no” ja “yes” muutettiin numeeriseen muotoon. Default Address Country Code muutettiin numeeriseen muotoon. Tags sarakkeen kanssa käytettiin one-hot-encoding-koodausta, jossa jokainen sarakkeen kategorinen muuttuja sai oman sarakkeensa, josta kävi ilmi, oliko kyseinen tagi merkitty asiakkaalla vai ei.

Arvoilla yksi ja nolla esitetään näkyykö kyseinen arvo rivillä. Tätä ennen “Tags” saraketta kuitenkin käsiteltiin, jotta löydettiin yleisimmät käytetyt tagit. Viisi käytetyintä tagia nähtävissä kuviossa 9.

```
Top 5 'Tags':
Tags
no tags
shopify-forms-21498
Inbox online store chat
forms-email-signup
newsletter
```

Kuvio 9. Käytetyimmät tagit

6.3.4 Feature Engineering

Uusiksi sarakkeiksi käsiteltiin olemassa olevasta aineistosta ensitilaukseen liittyviä tietoja. Ensitilauksen arvo, tilattujen tavaroiden määrä, ensitilauksen palautuksien määrä, ensitilauksen viittavan liikenteen määrä sekä ensitilauksessa käytettyjen alenuksien määrä. Asiakkaille laskettiin myös keskiarvot kerralla tilattujen tuotteiden määrästä, keskimääräinen tilauksen arvo, sekä kaikkien asiakkaan tilauksien kokonaismäärä sekä kokonaisarvo. Asiakkaille laskettiin myös kulunut aika viimeisestä tilauksesta, jota käytettiin poistuman määrittelemiseksi. Poistuneiksi määriteltiin kaikki asiakkaat, joiden viimeisimmästä tilauksesta oli kulunut yli 200 päivää. Asiakkaille laskettiin myös keskimääräinen tilausfrekvenssi, jos tilauksia oli enemmän kuin yksi. Jos tilauksia oli vain yksi, täytettiin keskimääräisen tilausfrekvenssin sarake kaikkien rivien keskiarvolla. Lopulta asiakas ja tilausdata taulut yhdistettiin yhdeksi tauluksi, jonka muuttujia käytettiin koneoppimismallien kouluttamiseen. Kuviossa 10 esiteltynä lopullinen taulu ja taulukossa 5 esiteltynä käytetyt muuttujat sekä niiden selitteet.

```
Average Frequency              float64
Churn                            int64
First Order Value              float64
First Order QTY                float64
First Order Discounts          float64
First Order Returns            float64
AVG Items Per Transaction      float64
Total Order Qty                  int64
Total Spent                    float64
AVG Order Value                float64
Total Returns                  float64
Total Discounts                float64
referrer_Direct                   bool
referrer_Email                    bool
referrer_Search                   bool
referrer_Social                   bool
referrer_Unknown                  bool
Accepts Email Marketing          int64
Default Address Country Code     int64
No Tags                          int64
Shopify Forms                    int64
Online Store Chat                int64
Forms Email Signup               int64
Newsletter                       int64
```

Kuvio 10. Lopullinen taulu

Taulukko 5. Käytetyt muuttujat ja niiden selitteet

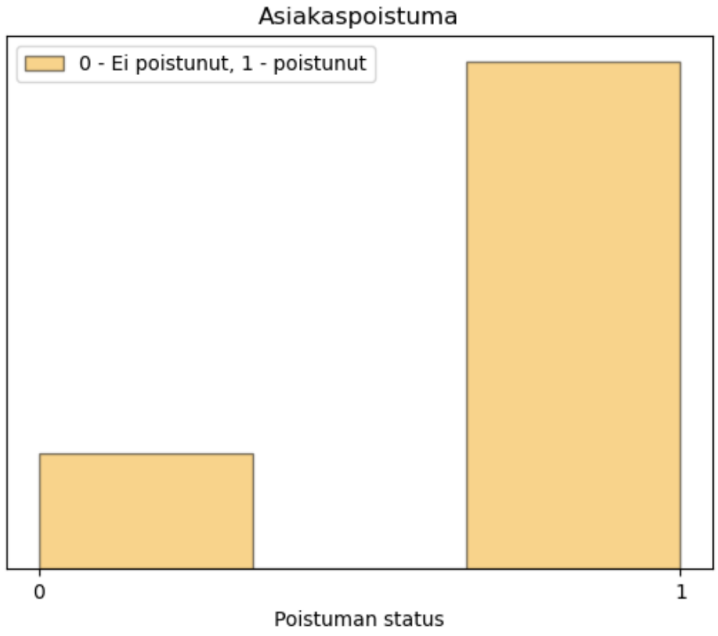
Muuttuja	Selite
Average Frequency	Keskimääräinen tilauksien välinen aika
Churn	Onko asiakas poistunut vai ei
First Order Value	Asiakkaan ensimmäisen tilauksen arvo

First Order QTY	Asiakkaan ensimmäisen tilauksen tuotteiden lukumäärä
First Order Discounts	Asiakkaan ensimmäisen tilauksen kokonaisalennukset
First Order Returns	Asiakkaan ensimmäisen tilauksen palautukset
AVG Items Per Transaction	Asiakkaan keskimääräinen tilattujen tuotteiden lukumäärä per tilaus
Total Order QTY	Asiakkaan tilauksien kokonaislukumäärä
Total Spent	Asiakkaan tilauksiin käyttämä kokonaisrahasumma
AVG Order Value	Asiakkaan keskimääräinen tilaussuma
Total Returns	Asiakkaan tekemien palautuksien rahallinen kokonaismäärä
Total Discounts	Asiakkaan saamien alennuksien kokonaismäärä
referrer_Direct	Viittaava lähde on “Suora”
referrer_Email	Viittaava lähde sähköposti
referrer_Invalid	Ei viittavaa lähdettä

referrer_Search	Viittaava lähde “haku”
referrer_Social	Viittaava lähde sosiaalinen media
referrer_Unknown	Viittaava lähde tuntematon
Default Address Country Code	Asiakkaan maakoodi
Accepts Email Marketing	Asiakas hyväksyy sähköpostimarkkinoinnin
No Tags	Asiakkaalle ei ole merkitty tunnisteita, kyllä/ei arvo
Shopify Forms	Asiakas täyttänyt Shopifyn lomakkeen
Online Store Chat	Onko asiakas keskustellut asiakaspalvelun kanssa verkkokaupan chatissa
Forms Email Signup	Asiakas täyttänyt lomakkeen sähköpostista
Newsletter	Tilaako asiakas uutiskirjettä vai ei

6.3.5 Jako testi- ja koulutusaineistoon

Aineisto jaettiin esikäsittelyn jälkeen kahteen osaan, joista toista käytetään koneoppimismallien kouluttamiseen ja toista mallien suorituskyvyn testaamiseen. Jakona käytettiin yleistä 20/80 jakoa, jossa testausaineistona toimii 20 prosenttia alkuperäisestä aineistosta, loppujen 80 prosentin toimiessa koulutusdatana. Käsitellystä aineistosta 81 prosenttia oli poistuneita asiakkaita, poistumattomia 19 prosenttia. Luokkien epätasapaino havainnollistettuna pylväskaaviossa kuviossa 11.



Kuvio 11. Asiakaspoistuman luokkajakauma

Luokat olivat huomattavassa epätasapainossa, joten niitä tasapainotettiin käyttämällä ”Class Weights”-parametria. Tekniikka antaa erilaiset painoarvot molemmille koulutusaineiston luokille ja rankaisee vähemmistöluokan virheellisestä luokittelusta enemmän kuin enemmistöluokan kohdalla tehdystä virheestä (Yenigun 2023). Parametrille annettiin arvo ”balanced”, jossa scikit-learn kirjasto laskee painotukset epätasapainoisuuden mukaan.

6.3.6 Käytetyt työkalut ja kirjastot

Tämän tutkimuksen toiminnallinen osuus on tehty Python-ohjelmointikielellä ja sille kehitetyillä kirjastoilla. Perinteisessä ohjelmistokehityksessä suosiota nauttiva Python on eräs käytetyimpiä

kieliä tekoälyn ohjelmoinnissa (Kolari & Kallio 2023, 154). Python-ohjelmointikieltä käytettiin Jupyter Lab ympäristössä, joka on datatieteitä varten suunniteltu avoimen lähdekoodin projekti (About Us n.d.). Tutkimuksessa käytettyjä kirjastoja olivat numpy, datan käsittelyyn tarkoitettu pandas, visualisoinnissa käytetty seaborn, matplotlib sekä koneoppimismalleja tarjoava scikit-learn. Mainitut kirjastot ovat avoimia ja ilmaisia ja tarjoavat näin matalan kynnyksen alkuun pääsemiselle.

7 Eettisyys ja luotettavuus

7.1 Eettisyys

Tutkimus on toteutettu noudattaen hyvää tieteellistä käytäntöä ja JAMKIN eettisiä periaatteita. Tutkimuksen teoreettisessa viitekehyksessä on kerrottu täsmällisesti, keiden julkaisemiin materiaaleihin on viitattu, kuten Bisterin (2019, 63) mukaan kuuluukin. Tutkimuksen tulokset on julkaistu vääristelemättä, eikä toimeksiantaja ole vaikuttanut niiden analysointiin tai raportointiin.

Toimeksiantajalla on ollut mahdollisuus tutustua tutkimukseen ennen sen julkaisua ja antaa kommentteja. Toimeksiantajan kanssa oli sovittu ennakolta työskentelytavoista ja osapuolten vastuista.

Tutkimusprosessi, tiedonhankinta sekä aineistonkäsittely on kuvattu yksityiskohtaisesti ja selkeästi. Käytettyä tutkimusaineistoa on tuotu julki asiayhteyteen sopivalla tavalla.

7.2 Luotettavuus

Koska asiakaspoistumalle ei ole olemassa ei-sopimuksellisessa tilanteessa yhtä universaalia määritelmää, ei tutkimus täytä ulkoisen validiteetin kriteeristöä. Tulokset eivät ole yleistettäviä koskemaan verkkokauppaa yleisesti. Sisältövaliditeettia puoltaa työn tarkka dokumentaatio, sekä se, että käytetyt mittarit ja muuttujat on valikoitu aikaisemmissa tutkimuksissa käytettyjen mittarien pohjalta (Kananen 2012, 170). Kokonaisvaliditeetin osalta työ jää puolitiehen, sisältövaliditeetti toteutuu, mutta ulkoisen validiteetin kriteeristö ei.

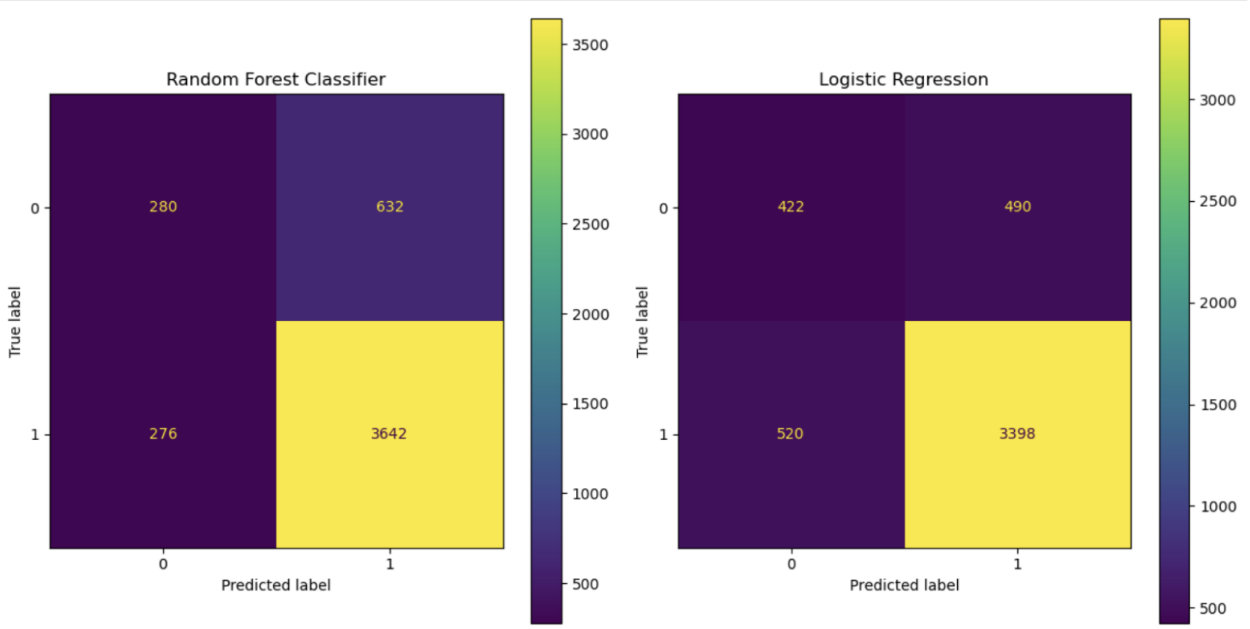
Tutkimuksen reliabiliteettia on vaikea arvioida, koska tutkittu ilmiö liittyy asiakkaiden ostokäyttäytymiseen, joka saattaa muuttua vuosien mittaan. Reliabiliteetin stabiilius ei siis ole kovin hyvä. Yrityksen ns. “niche”-markkina itsessään aiheuttaa haasteita; kameroiden kulutustrendit vaihtelevat vuosittain ja nyt trendikäs filmikamera saattaa olla viiden vuoden päästä asiakkaiden silmissä roskaa. On perusteltua kuitenkin mainita, että jos tutkimus kuitenkin tehtäisiin uudestaan samalla otannalla, saataisiin samoin keinoin todennäköisesti samat tulokset.

Tutkimuksen luotettavuuden puolesta puhuu se, että sen aikana esiin tulleita asioita, datan analysoinnista ja visualisoinnista kertyneitä huomioita, on huomioitu toimeksiantajayrityksen puolesta. Saatuihin tuloksiin ja löytöihin luotetaan toimeksiantajan osalta.

8 Tulokset ja Pohdinta

8.1 Tulokset

Malleille syötettiin testiaineisto, josta malli luokitteli asiakkaita joko poistuviksi tai ei-poistuviksi. Kuviossa ?? on molempien käytettyjen mallien sekaantumismatriisi. Ylhäällä vasemmalla näkyvät oikein luokitellut ei poistujat (todet negatiiviset), alhaalla vasemmalla väärin luokitellut poistujat (väärät positiiviset), ylhäällä oikealle väärin luokitellut ei poistujat (väärät negatiiviset) ja alhaalla vasemmalla näkyvät oikein luokitellut poistujat (todet positiiviset). Kuviossa 12 esitellään käytettyjen mallien sekaannusmatriisit visualisoituina.

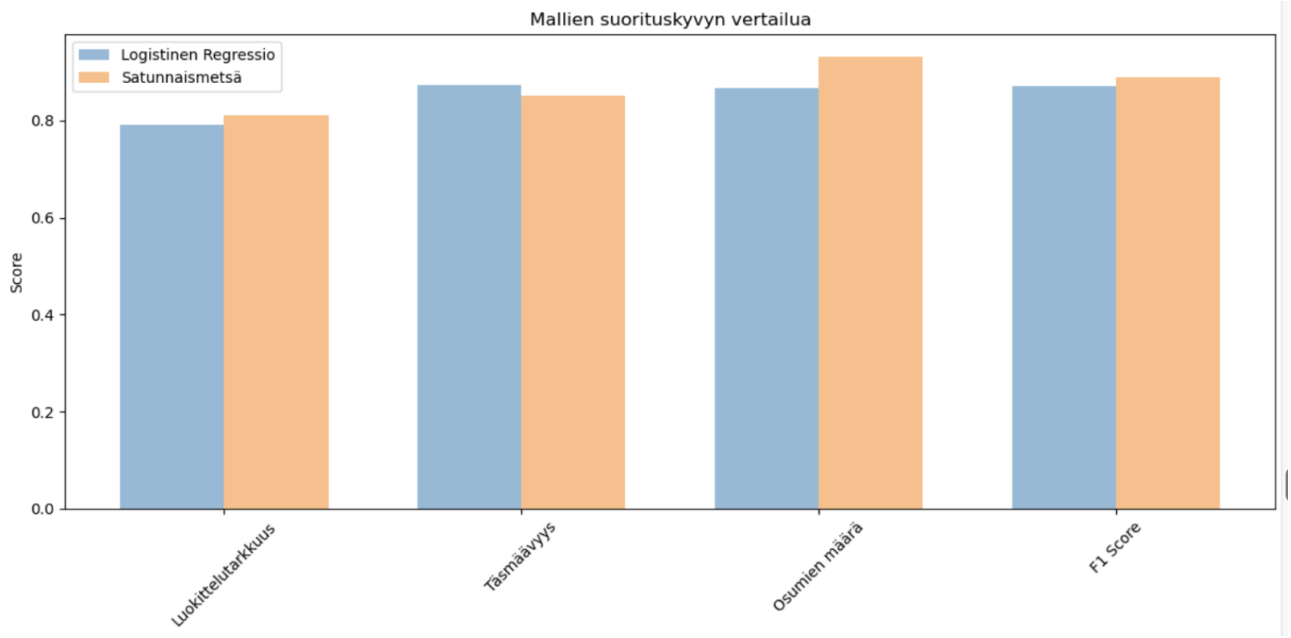


Kuvio 12. Sekaannusmatriisi

Logistisen regression luokittelutarkkuus oli 79 prosenttia, joten 520 kappaletta poistujista jäi tunnistamatta. Malli luokitteli poistujiksi 490 ei poistujaa ja onnistui tunnistamaan 422 kappaletta tosia ei-poistuneita asiakkaita. F1 Score oli poistuneiden asiakkaiden osalta 87 prosenttia. Ei poistuneiden osalta F1 Score oli 46 prosenttia. Täsmäävyys poistujien suhteen oli 87 prosenttia, kuten myös osumien määrä. Ei-poistujien suhteen täsmäävyys oli 45 prosenttia ja osumien määrä 46 prosenttia. ROC AUC score oli 66 prosenttia. Tosia positiivisia oli 3398 kappaletta, tosia negatiivisia 422 kappaletta, vääriä positiivia 490 kappaletta ja vääriä negatiivisa 520 kappaletta.

Satunnaismetsä tunnisti 3642 poistujaa oikein ja sai näiden osalta luokittelutarkkuudeksi 81 prosenttia. 276 poistujaa jäi tunnistamatta. Malli luokitteli poistujiksi 632 ei poistujaa ja onnistui tunnistamaan 280 kappaletta todellisia ei poistuneita asiakkaita. F1 score poistuneiden osalta oli 89 prosenttia, kun taas ei poistuneiden osalta F1 score oli 38 prosenttia. ROC AUC score oli 62 prosenttia. Tosia positiivisia oli 3642 kappaletta, tosia negatiivisia 276 kappaletta, vääriä positiivisia 632 kappaletta ja vääriä negatiivisia 276 kappaletta. Täsmäävyys poistujien osalta oli 85 prosenttia ja osumien määrä oli 93 prosenttia. Ei poistuneiden osalta osumien määrä oli 31 prosenttia ja täsmäävyys 50 prosenttia.

Kuviossa 13 on visualisoitu testattujen mallien tarkkuuksia luokittelutarkkuuden (accuracy), täsmäävyyden (precision), osumien määrän (recall) sekä F1 Scoren mukaan. Taulukossa 6 on eriteltynä mallien arvioinnissa käytettyjen mittarien arvoja.



Kuvio 13. Testattujen koneoppimismallien suorituskyvyn vertailua

Taulukko 6. Mallien luokittelutarkkuus, täsmäävyys, osumien määrä sekä F1-pisteet poistujille

Malli	Luokittelutarkkuus	Täsmäävyys	Osumien Määrä	F1 Score
Logistinen Regressio	0.79	0.87	0.87	0.87
Satunnaismetsä	0.81	0.85	0.93	0.89

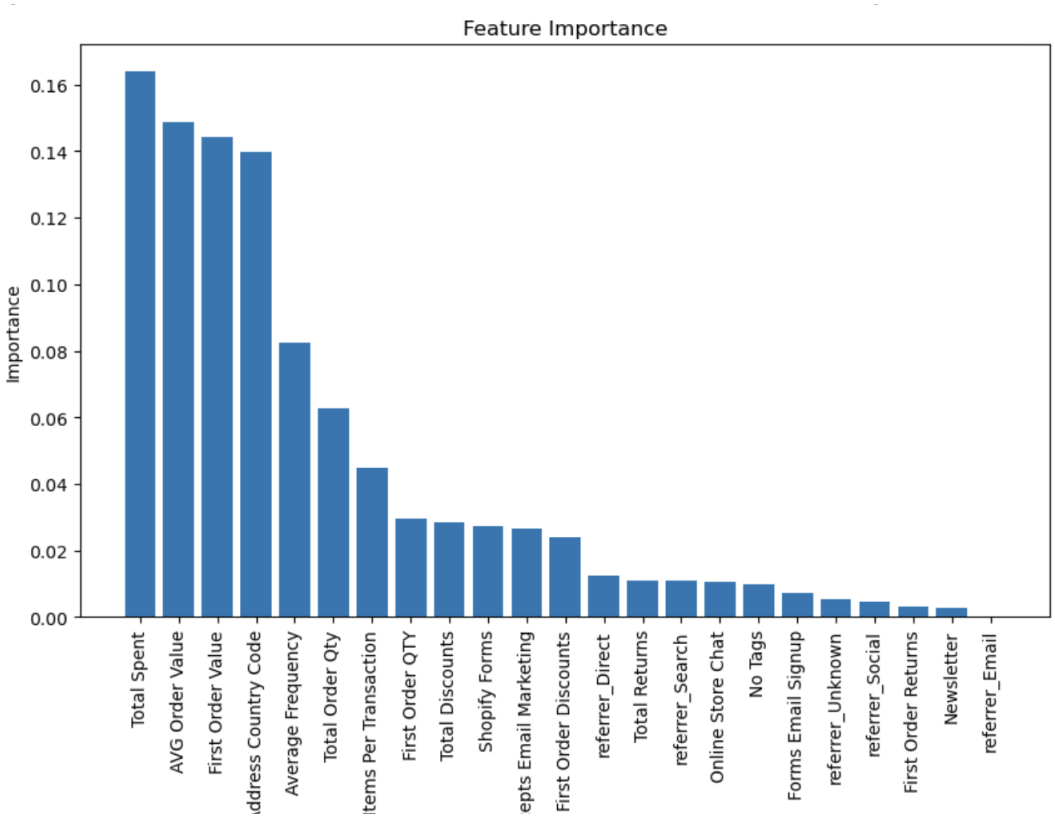
Molemmat mallit suoriutuivat verrattain hyvin todellisten poistujien ennustamisesta. Hieman paremmat tulokset tämän luokan osalta sai kuitenkin satunnaismetsä tarkkuuden noustessa 81 prosenttia. Malli tunnisti oikein siis 81 prosenttia luokan kaikista poistuneista asiakkaista. Logistinen regressio pääsi hyvin lähellä tätä tarkkuuttaa tarkkuuden ollessa 79 prosenttia. Satunnaismetsä

näyttää tunnistavan poistuvat asiakkaat paremmin, mutta tekee vääriä positiivia arvioita useammin kuin logistinen regressio. Logistinen regressio tunnistaa paremmin ei poistujat, mutta merkitsee lähes tuplasti enemmän poistujia ei poistujiksi. Molempien mallien suorituskyky ROC AUC arvoa mitaten on vajavainen, mallit suoriutuvat tämän osalta noin 60 prosentin tarkkuudella

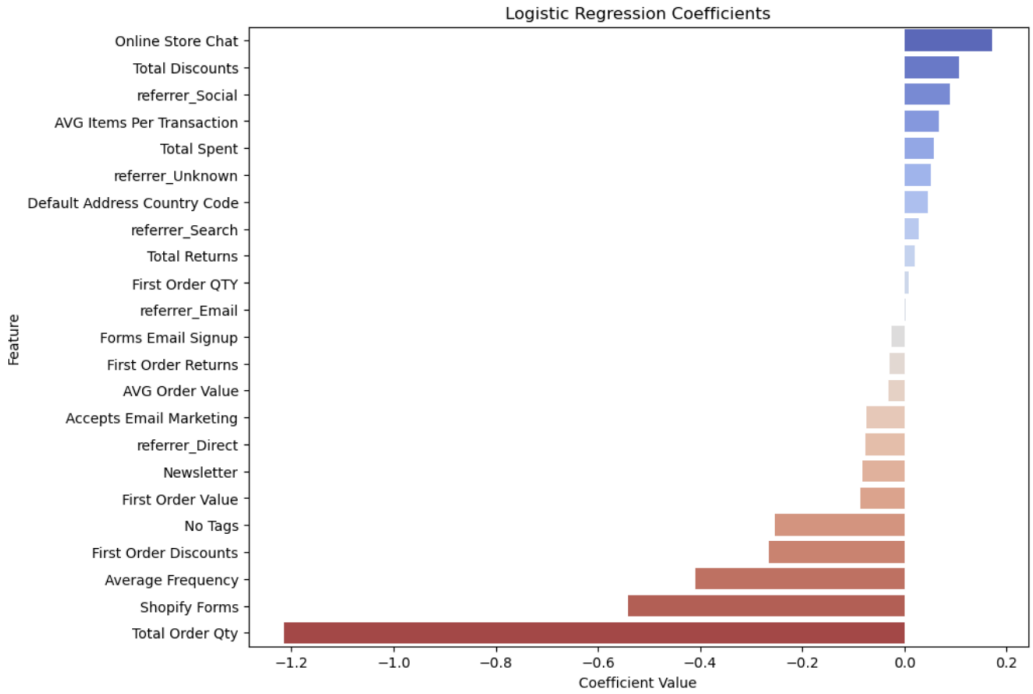
Väärien positiivisten ennusteiden suuri määrä voi olla ongelmallinen resurssien allokoinnin kanssa. On turhaa kohdistaa suuria markkinointitoimenpiteitä joukkoon, jonka osana on suuri joukko ennakoidusti pysyviä asiakkaita. Asiakaskokemus voi myös heikentyä, jos kampanjointi on suoritettu väärällä tavalla.

Jain ja muiden (2021, 154) saama logistisen regression luokittelutarkkuus oli 85 prosentin luokkaa ja satunnaismetsän noin 86 prosentin. Sweidan ja muut (2022, 6) saivat satunnaismetsän luokittelutarkkuudeksi 73 prosenttia ja logistisen regression 85 prosenttia. Saadut tarkkuudet ovat lähelle toteutetussa tutkimuksessa saatuja arvoja.

Muuttujien välisiä suhteita tulkittaessa käytettiin scikit learning feature importance työkalua. Satunnaismetsän osalta havaittiin, että suurimmat vaikuttavat tekijät olivat asiakkaan käyttämä rahamäärä, keskimääräinen tilausarvo, ensimmäisen tilauksen arvo, keskimääräinen tilausväli sekä tuotteiden tilattu kokonaismäärä. Logistista regressiota tutkittiin scikit learning coefficient mallilla ja havaittiin, että vahvimmin poistumaa indikoi Online Store Chat-muuttuja sekä referrer_Social-muuttuja. Vielä vahvempana näkyy kuitenkin kokonaistilausmäärän vaikutus poistumaa pienentävänä tekijänä. Myös Shopify Forms-muuttuja sekä Average Frequency näyttävät pienentävän poistumaa. Kuvioissa 14 ja 15 on esiteltynä muuttujien vaikutuksia asiakaspoistumaan.



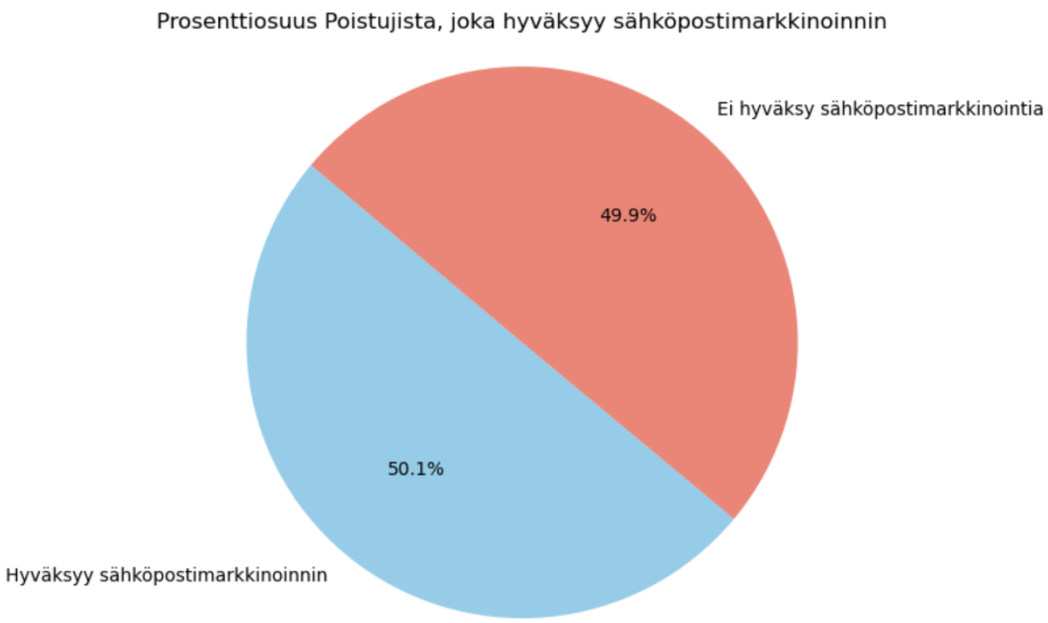
Kuvio 14. Muuttujien vaikutusarvot poistumaan satunnaismetsä-mallissa



Kuvio 15. Muuttujien vaikutuksia poistumaan Logistisessa Regressiossa

Sweidanin ja muiden (2021, 6) tutkimuksessa havaittiin, että tiheämpi ostotiheys korreloi negatiivisesti poistuman kanssa ja poistujat käyttivät vähemmän rahaa kuin ei poistujat. Nämä löydökset ovat hyvin linjassa tutkimuksessa tehtyjen löytöjen kanssa. Myös Fridrich & Dostal (2022, 10) nostavat poistumaan vaikuttavien tekijöiden joukkoon keskimääräisen asiointitiheyden. Asiointitiheyden tarkempi analysointi paljastaa kuitenkin, että poistujien keskimääräinen tilausväli oli tutkitussa joukossa tiheämpi kuin ei-poistujien. Tämä johtuu suurelta osin aineiston käsittelystä, sillä vain yhden tilauksen tehneiden asiakkaiden puuttuvat tilaustiheysarvot korvattiin keskiarvolla.

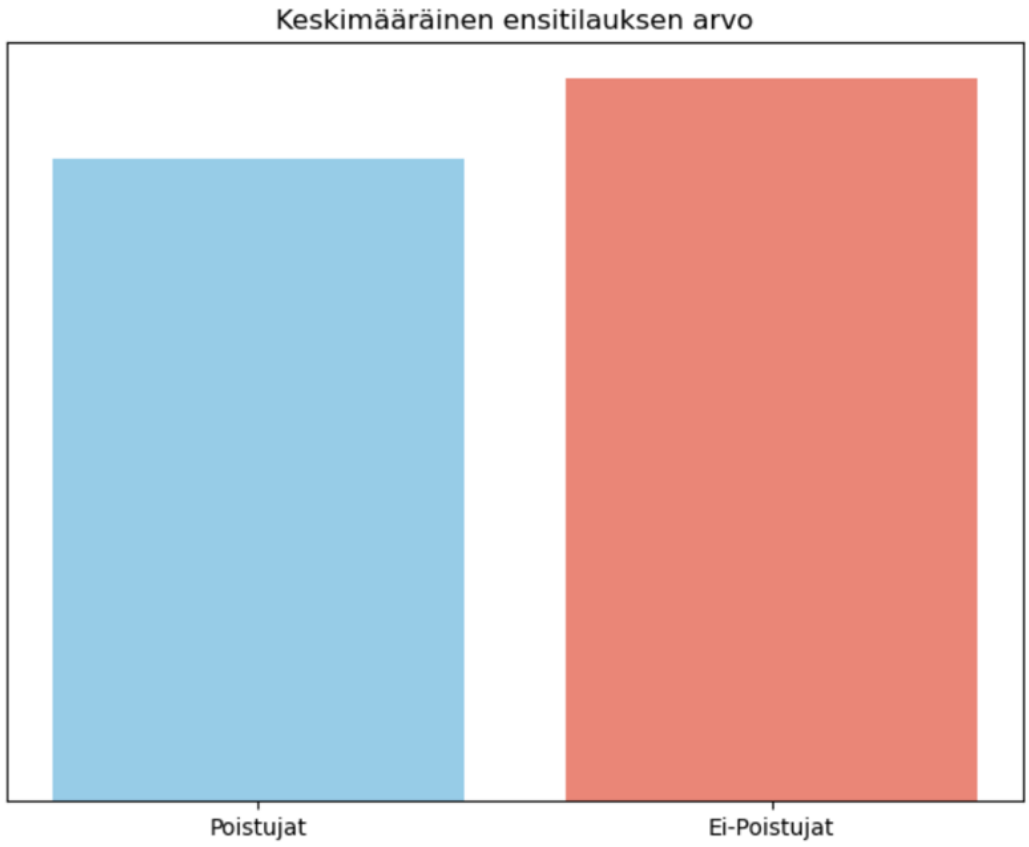
Aineiston analyysista selvisi, että jopa puolet poistujista hyväksyy sähköpostimarkkinoinnin, joka avaa ovia asiakkaiden säilyttämiseksi. Tätä on havainnollistettu kuviossa 16.



Kuvio 16. Poistujat ja sähköpostimarkkinointi

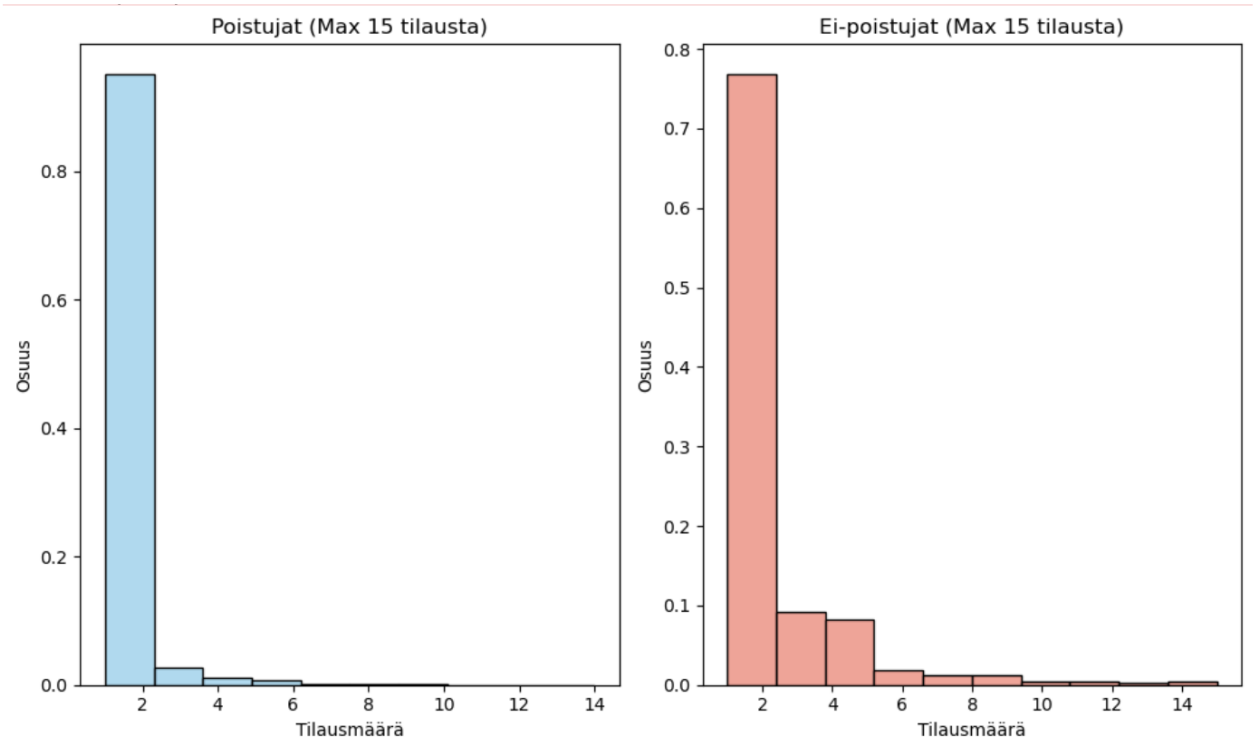
Toinen mielenkiintoinen havainto muuttujista on, että valtaosa sosiaaliseen mediaan viittaavasta liikenteestä linkittyy poistujiin. Poistujissa on 581 merkintää sosiaalisesta mediasta liikenteen lähteenä ja ei poistujilla 62. Vastaavanlainen havainto on tehtävissä myös Online Store Chat-muuttujasta, joka on 1190 poistujalla ja vain 174 ei-poistujalla.

Ei poistujista huomattiin myös, että heidän keskimääräinen tilausarvonsa, sekä ensitilauksensa arvo on suurempi, kuin poistujien. Ensitilauksien arvojen suhteita on visualisoitu pylväsdiagrammein kuviossa 17.



Kuvio 17. Ensitilauksen arvojen suhde poistujien ja ei-poistujien välillä

Valtaosa poistujista näyttää myös tilaavan vain kerran, jonka jälkeen asiakkuussuhde päättyy. Tehty löytö poistujien tilausmäärästä, on hyvin linjassa Sweidanin ja muiden (2022, 6) mainitsemaan löytöön poistujien tilauksista – noin 75 prosenttia yhden tilauksen asiakkaista poistuu. Poistujien ja ei-poistujien tilausmäärien jakaumaa on havainnollistettu kuviossa 18. Shopify Forms muuttujan jakauma poistujien ja ei-poistujien suhteen oli lähes tasan, voidaan siis sanoa, että lomakkeen täyttäminen näyttää pienentävän asiakaspoistuman riskiä.



Kuvio 18. Tilausjakaumat poistujien ja ei-poistujien kesken

8.2 Tekoälyn käyttö tässä opinnäytetyössä

Opinnäytetyön kappaleiden ideoinnissa on käytetty apuna tekoälysovellus ChatGPT:tä. Myös tekstiä on kirjoitettu uudelleen tekoälyn avustuksella, jotta kieli olisi selkeämpää, tiiviimpää ja paremmin ymmärrettävää. ChatGPT:tä käytettäessä on otettu huomioon vastuullisuus tietosuojaan liittyen.

8.3 Pohdinta

Tutkimuskysymyksiin saatiin vastaukset. Verkkokaupan asiakaspoistumaa voidaan pyrkiä ennustamaan koneoppimismallien avulla. Jo tehdyt tutkimukset osoittavat, että parhaimmillaan mallit voivat yltää jopa 90 prosentin tarkkuuksiin ennustuksissa. Tässä tutkimuksessa ennusteissa saatiin tästä hieman alhaisempia arvoja. Kysymykseen siitä, kumpi tarkastelluista algoritmeista on parempi poistuman ennustamisessa, saatiin vastaus. Satunnaismetsä vaikuttaa olevan käytettävissä olevan aineiston osalta tarkempi ennustamaan poistujia. Kolmantena kysymyksenä olleeseen ”Mitkä käytetyistä muuttujista näyttävät vaikuttavan eniten poistuman syntymiseen?”, saatiin myös vastauksia. Alhainen tilausmäärä, alhaisempi kokonaiskulutustaso sekä pienempi keskiostos

näyttävät vaikuttavan poistuman syntyyn. Myös asiakkaat, jotka olivat keskustelleet tilausta tehdessään verkkokaupan chatissa asiakaspalvelijan kanssa olivat todennäköisempiä poistujia.

Vieläkin tarkempi datan analysointi ja dokumentointi olisi voinut olla hyvinkin hedelmällistä. Jo pelkästään poistujien dataa tutkimalla kävi ilmi, että poistuneiden asiakkaiden joukossa suuri osa on hyväksynyt sähköpostimarkkinoinnin, joten korjaavat toimenpiteet asiakkuuden säilyttämiseksi olisi helppo aloittaa sähköpostikampanjoinnilla. Tässä voisi olla ainakin yksi konkreettinen suunta jatkokehitykselle.

Tarkempi analyysi poistuneista useamman tilauksen asiakkaista voi avata myös ovia pidempiin asiakkuussuhteisiin. Onko mahdollista löytää yhteneväisiä muuttujia jo useamman tilauksen tehneelle poistujalle? Ensiasiakkaille voisi myös suunnitella erilaisia asiakaspolkuja ensitilauksen arvon perusteella. Jos asiakas määritellään ensitilauksen tilausarvon perusteella tulevaksi poistujaksi ja tämä täyttää muut määritellyt ehdot, segmentoidaan asiakas eri ryhmään kuin raja-arvot ylittävä todennäköisesti ei-poistuva asiakas. Tämä mahdollistaisi kohdennetun sähköpostimarkkinoinnin potentiaalisille poistujille.

Usealla poistujalla toistui muuttuja, joka kertoi asiakkaan keskustelleen asiakaspalvelun kanssa. Kyseistä muuttujaa ei enää käytetä, joten on vaikea sanoa voiko poistuminen liittyä vaikeaksi koettuun asiakaspalvelutilanteeseen tai tilauksesta koituneeseen pettymykseen. Merkintä asiakas/tilaustietoihin asiakaspalvelun kanssa keskustelusta ei varmasti tekisi haittaa asiakasdatalle ja sen tulkitsemiselle tulevaisuudessa.

Jos aiheesta tehtäisiin jatkotutkimuksia, voitaisiin arvioinnissa käytettävien mallien määrää lisätä kahdesta useampaan ja vertailla niiden toimintaa. Ennusteita voitaisiin tehdä myös erilaisilla aikaväleillä, eikä pyrkiä ennustamaan koko saatavilla olevan datan perusteella. Muuttujien vaikutusta ennusteisiin voisi myös testata, niitä lisäämällä tai poistamalla. Muuttujien lisääminen voisi tapahtua helposti rikastamalla dataa saatavilla olevilla keinoilla. Asiakkaiden sessiodata voisi olla yksi mahdollinen lisättävä muuttuja. Yksityiskohtainen tuotedata voisi tuoda myös uuden ulottuvuuden ennusteisiin. Olisi mielenkiintoista tietää olisiko esimerkiksi tiettyä tuoteryhmää ostava segmentti toisia alttiimpi poistumalle, tai vaihtoehtoisesti olisiko jotain muuta tuoteryhmää tilaava asiakasryhmä immuunimpi poistumiselle. Asiakkaiden ennakoidut elinkaaren arvot saattaisivat

myös rikastaa dataa mielenkiintoisilla tavoilla. Pelkän binäärisen luokittelun rinnalle, voisi asiakkaita pyrkiä segmentoimaan poistumariskin perusteella. Käytettyjen muuttujien määrää olisi ollut myös hyvä testata ja muuttujien kategorisointia olisi voinut tehdä rohkeammin. Mahdollisesti parempia tuloksia olisi ollut saatavissa pienemmällä määrällä paremmin käsiteltyjä muuttujia. Mielenkiintoisena jatkotutkimusaiheena koen myös toimeksiantajayrityksen/yleisesti yrityksien suhtautumisen koneoppimiseen ja sen hyödyntämiseen osana liiketoimintaa. Koetaanko ja nähdäänkö tässä oikeasti mahdollista lisäarvoa ja kuinka hankalaksi sen käytäntöön viemistä pidetään?

Tutkimuksessa jätettiin kokonaan huomiotta neuroverkot, syväoppiminen ja niiden avulla ennustaminen. Ennusteiden tekeminen laajemmalla skaalalla malleja olisi voinut tuoda siihen myös lisäarvoa. Tarkemmalla perehtymiselle ja vertailulla jo tehtyihin tutkimuksiin, olisi käytettyjen mallien ja muuttujien valinnat saanut perusteltua paremmin ja todennäköisesti tuloksena olisi ollut myös tarkemmin toimivat ennustemalli.

Käytettyjen mallien parametrien optimointi jäi nyt myös kokonaan pois eikä käytettyjen mallien tehokkuutta mitattu, jonka osalta työ jäi puolitiehen.

Opinnäytetyössä käytetyt ennustemallit eivät sellaisinaan myöskään ole vietävissä tuotantoon, vaan vaatisivat ympärilleen toimivan graafisen käyttöliittymän. Nyt dataa ja ennusteita, on käsitelty pelkästään Jupyter Notebook-ympäristössä. Datan noutoa ei ole myöskään automatisoitu, vaan se vaatii käyttäjältä toimia. Ennustemallien ajaminen itsessään on käsin tehtävää työtä – olisiko tämän automatisoinnissa ja ennustetunnelin luomisessa ideaa?

Suurimmat kysymykset projektin aikana nousivat poistuman määrittelyn hankaluudesta – milloin ei sopimuksellisessa ympäristössä voidaan sanoa jonkun poistuneen? Ratkaisu tähän löytyy varmasti yrityksen ja erehdyksen kautta, eikä oikeaa vastausta ole välttämättä olemassa. Tämän olisi voinut määritellä myös yhdessä toimeksiantajan kanssa, joka jää nyt tulevaisuudessa tehtäväksi. Tutkimuksessa käytettyjä muuttujia olisi voinut kategorisoida ja ryhmitellä paremmin ja vertailla vaikuttaisiko tämä saatuihin tuloksiin. Ennustemalli ei kerro myöskään suoraan, ketä kannattaisi lähestyä toimenpiteillä asiakkuuden säilyttämiseksi. Osa poistujista on varmasti arvokkaampia asiakkaan arvioidun elinkaaren arvolla mitattu, jolloin tallaiselle ominaisuudelle olisi hyötynäkökulma liikevaihdollisesta näkökulmasta.

Aikataulutuksen osalta opinnäytetyö eteni suunnitellusti, eikä varsinaisia suvantovaiheita ehtinyt tulla. Alkuperäinen ajatus poukkoili, kokeili rajojaan, ja lopulta rajautui käsittelemään poistuman ennustamista muutaman eri koneoppimismallin pohjalta. Aihe vei äkkiä mukaansa ja tutkimuksesta toiseen hyppiminen oli aiheuttaa liiallista aiheen laajentumista. Kirjoitettavaa ja tutkittavaa olisi varmasti loppuelämäksi.

Lähteiden käytön osalta jopa vähempi määrä olisi varmasti riittänyt, nyt harvassa lähteessä päästin alkuperäisen lähteen ja tiedon juurille. Lähteissä on käytetty niin ulkomaisia, kuin kotimaisiakin lähteitä ja lähteet ovat monipuolisesti eri formaateista kerättyjä.

Jälkiviisaana on helppoa sanoa, että toimeksiantajayrityksen kanssa olisi pitänyt pitää useampia palavereita, ja ehkä mahdollisesti pallotella ideoita ja ajatuksia. Toisaalta nyt opinnäytetyö on varmasti tekijänsä näköinen ja vastaa tutkimuskysymyksiin vähintäänkin kohtuullisella tasolla. Heikkoudeksi työn osalta voidaan todeta, ettei sillä heti ole suoraa liikevaihdollista tai tuloksellista hyötyä toimeksiantajan kannalta. Sen avulla voidaan kuitenkin ymmärtää asiakaspoistumaa ja siihen vaikuttavia syitä hieman paremmin, eikä sovi unohtaa kerrytettyä osaamista koneoppimisen puolelta. Toimeksiantajayritykselle tarjotaan kuitenkin konkreettisia ideoita seuraavia vaiheita varten, joka voitaneen laskea ansioksi.

Datan visualisointi olisi voinut olla parempaa ja informatiivisempaa ajatellen valittua aihetta, visualisoinnin yhdenmukaisuus jäi myös puolitiehen ja käytetyt värimallit eivät olleet konsistenttejä. Jokaisesta vaiheesta ei ole myöskään avattuja kuvia tai kuvioita, jota voidaan pitää heikkoutena. Binäärinen kategorisointi poistuneisiin ja ei poistuneisiin asiakkaisiin saattoi olla liian raju, ja ehkä mukaan olisi mahtunut myös osittain poistuneiden joukko. Myös pureutuminen koneoppimismalleihin olisi voinut olla lähtökohdiltaan “matemaattisempi” ja sisältää mukanaan myös kaavoja auki kirjoitettuna. Nyt malleja lähestyttiin melko yleisellä tasolla, eikä yksityiskohtiin asti päästy.

Käsitelty aihe herätti aidon mielenkiinnon ja syvensi tietämystä sen ympäriltä. Aina kun luuli ymmärtävänsä jotain, paljastui uusi kerros, joka laajensi ennestään jo käsiteltyä aihetta. Näin jälkiviisaana on kuitenkin helppoa sanoa, että aihe olisi tullut rajata vielä tarkemmin, jotta siitä olisi saatu vielä enemmän irti. Toisaalta jo tällaisenaan työ tuo jotain uutta alalle, sillä käsitellystä aiheesta ei ole vielä merkittävää määrää tehtyjä opinnäytetöitä.

Lähteet

About Us. N.d. About Us-infosivu Jupyter-verkkosivustolla. Viitattu 21.4. https://jupyter.org/about.

Ahn, K., Hwang, D., Kim, Choi, H. & Kang, S. A. 2022. Survey on a Churn Analysis in Various Business Domains. IIEE Access, Volume 8, January, 220816-220839. Viitattu 19.3.2024. https://ieeexplore.ieee.org/document/9281029.

Alpaydin, E. 2021. Koneoppiminen. Helsinki: Terra Cognita.

AlphaGo. N.d. Google deepmind-sivu. Viitattu 21.4. https://deepmind.google/technologies/alphago/.

Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. 1959. IBM Journal of Research and Development, 3, 3, 210-229. Viitattu 28.2.2024. https://ieeexplore.ieee.org/document/5392560.

Bister, T. Tietojenkäsittelyn Opinnäytetyö, Viittoja ja karttoja tutkimisen ja kehittämisen tielle. 2019. Jyväskylä: Jyväskylän Ammattikorkeakoulu.

Bhale, A. U. & Bedi, H. S. 2023. Customer Churn Construct: Literature Review and Bibliometric Study. Viitattu 25.3. https://www.researchgate.net/publication/372564864_Customer_Churn_Construct_Literature_Review_and_Bibliometric_Study.

Brownlee, J. 3.10.2023. Supervised and Unsupervised Machine Learning Algorithms. Artikkeli Machine Learning Mastery-verkkosivustolla. Viitattu 29.2.2024. https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

Classification: ROC Curve and AUC. N.d. Opintomateriaali Googlen Machine Learning-kurssilta. Viitattu 23.4.2024. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

Danao, M. 25.11.2023. What Is Churn Rate & How Do You Calculate It. Artikkeli Forbes-lehden verkkosivulta. Viitattu 19.3.2024. https://www.forbes.com/advisor/business/churn-rate/.

Deep Blue. N.d. Artikkeli IBM-verkkosivustolla. Viitattu 28.2.2024. https://www.ibm.com/history/deep-blue.

Ecommerce Churn Rate: How To Calculate and Reduce Churn. 21.10.2022. Artikkeli Shopify-verkkosivustolla. Viitattu 29.2.2024. https://www.shopify.com/blog/churn-rate-in-ecommerce#.

Fridrich, M & Dostal, P. 2022. User Churn Model in E-Commerce Retail. Scientific Papers of the University of Pardubice Series D Faculty of Economics and Administration 30,1. Viitattu 28.2.2024. https://www.researchgate.net/publication/359739936_User_churn_model_in_e-commerce_retail.

Gallo, A. 29.10.2014. The Value of Keeping the Right Customers. Artikkeli Harvard Business Review -sivustolta. Viitattu 4.3.2024. https://hbr.org/2014/10/the-value-of-keeping-the-right-customers.

Geiler, L. Affeldt, S & Nadif, M. 2022. A survey on machine learning methods for churn prediction. International Journal of Data Science and Analytics, 14, 3, 221-242. https://hal.science/hal-03824873/document.

Hayley, M. 2016. A literature review on CRM – definitions, benefits, components, andimplementation. Australian Journal of Management and Financial Research, 1, 26–34. Viitattu 28.2.2024. https://www.researchgate.net/publication/304785150_Australian_Journal_of_Management_and_Financial_Research.

Hirsjärvi, S., Remes, P. & Sajavaara, P. 1997. Tutki ja kirjoita. Kustannusosakeyhtiö Tammi: Helsinki.

How to Reduce Customer Churn for Retail Success. 13.10.2023. Blogi-julkaisu Shopify-verkkosivustolla. Viitattu 18.3.2024. https://www.shopify.com/blog/customer-churn#.

Jain, H., Yudav, G. & Rajapandy, M. 2021. Churn Prediction and Retention in Banking, Telecom, and IT Sectors Using Machine Learning Techniques. Viitattu 14.3.2024. https://www.researchgate.net/publication/343223435_Churn_Prediction_and_Retention_in_Banking_Telecom_and_IT_Sectors_Using_Machine_Learning_Techniques.

Kelleher, J.D & Tierney, B. 2021. Datatiede. Helsinki: Terra Cognita.

Kananen, H. & Puolitaival, H. 2019. Tekoäly: Bisneksen uudet työkalut. Helsinki: Alma Talent.

Kananen, J. 2010. Opinnäytetyön kirjoittamisen käytännön opas. Jyväskylä: Jyväskylän Ammattikorkeakoulu.

Kananen, J. 2012. Kehittämistutkimus opinnäytetyönä: Kehittämistutkimuksen kirjoittamisen käytännön opas. Jyväskylä: Jyväskylän Ammattikorkeakoulu.

Kämäräinen, J. Koneoppimisen perusteet. 2023. Otatieto/Gaudeamus.

Lazarov, V & Capota, M. 2007. Churn Prediction. Viitattu 18.3.2024. http://www.vladislav.lazarov.pro/files/research/papers/churn-prediction.pdf.

Matuszelański, K & Kopczewska, K. Customer Churn in Retail E-Commerce Business: Spatial and Machine Learning Approach. Journal of Theoretical and Applied Electronic Commerce Research. 2022, 17, 1, 165-198. Viitattu 24.4.2024. https://doi.org/10.3390/jtaer17010009.

Miguéis, V.L., Dirk Van den Poel, Camanho, A.S. & João Falcão e Cunha. 2012. Modeling partial customer churn: On the value of first product-category purchase sequences. Expert Systems with Applications, 39, 12, 11250-11256. https://doi.org/10.1016/j.eswa.2012.03.073.

Data Preprocessing in Python. 10.6.2023. Artikkeli Geeks For Geeks-sivustolla. Viitattu 14.5.2024. https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/.

Nummenmaa, L. 2004. Käyttäytymistieteiden Tilastolliset Menetelmät. Kustannuosakeyhtiö Tammi: Helsinki

Reicheld, F. 2001. Prescription for cutting cost. Viitattu 18.3.2024. https://media.bain.com/Images/BB_Prescription_cutting_costs.pdf.

Random forests. N.d. Oppimateriaali Googlen sivustolta. Viitattu 19.3.2024.https://developers.google.com/machine-learning/decision-forests/random-forests.

Runsas, J. 29.2.2024. Kameratori on analogisten kameroiden Nooan arkki. Viitattu 29.2.2024. https://tampereenkauppakamarilehti.fi/fi-fi/article/kauppakamarilehti/kameratori-on-analogisten-kameroiden-nooan-arkki/1405/.

Shobana, J., Gangadhar, Ch., Arora, R.K., Renjith, P.N., Bamini, J. & Chincholkar, Y. 2023. E-commerce customer churn prevention using machine learning-based business intelligence strategy. Measurement: Sensors, Volume 27, 2023, 2665-9174. Viitattu 23.4.2024. https://www.sciencedirect.com/science/article/pii/S2665917423000648.

Stephenson, D. 2018. Big Data Demystified, how to use big data, data science and AI to make better business decisions and gain competitive advantage. Harlow: Pearsons.

Storås, N. 27.6.2023. Filmikameroiden renessanssi. Artikkeli HS Visiossa. Viitattu 29.2.2024. https://www.hs.fi/visio/art-2000009644607.html.

Subramanian, K. 9.8.2023. Churn Matters: How To Better Manage Yours. Artikkeli Forbes-lehden sivustolla. Viitattu 21.4. https://www.forbes.com/sites/forbesfinancecouncil/2023/08/09/churn-matters-how-to-better-manage-yours/.

Sweidan, D, Johansson, U, Gidenstam, A & Alenljung, B. 2022. Predicting Customer Churn In Retailing. 21st IEE International Conference on Machine Learning and Applications (ICMLA). Viitattu 18.3.2024. https://ieeexplore.ieee.org/document/10068870.

Jahromi, A.T., Stakhovych, S. & Ewing, M. Managing B2B customer churn, retention and profitability. Industrial Marketing Management, 43, 7, 1258-1268. Viitattu 18.3.2024. https://doi.org/10.1016/j.indmarman.2014.06.016.

Tervetuloa Kameratorille. N.d. Esittelysivu Kamerastoren verkkosivustolla. Viitattu 29.2.2024. https://kamerastore.com/fi/pages/tervetuloa-kameratorille.

Tietotekniikan käyttö yrityksissä. 2023. Helsinki: Tilastokeskus. Viitattu 29.4.2024. https://stat.fi/tilasto/icte.

Wagh, S., Andhale, A., Wagh, K., Pansare, J., Ambadekar, S. & Gawande, S. 2024. Customer churn prediction in telecom sector using machine learning techniques. Results in Control and Optimization, 14,2024,100342. Viitattu 24.4.2024. https://doi.org/10.1016/j.rico.2023.100342.

What is good churn rate. N.d. Artikkeli Recurly-verkkosivustolla. Viitattu 18.3.2024. https://recurly.com/research/churn-rate-benchmarks/.

What is logistic regression? N.d.a. Artikkeli AWS-verkkosivustolla. Viitattu 29.4.2024. https://aws.amazon.com/what-is/logistic-regression/.

What is logistic regression? N.d.b. Artikkeli IBM-verkkosivustolla. Viitattu 29.4.2024. https://www.ibm.com/topics/logistic-regression.

What is machine learning?. 3.5.2017. Artikkeli The Conversation-verkkosivustolla. Viitattu 29.2.2024. https://theconversation.com/what-is-machine-learning-76759.

What is machine learning?. N.d. Opetusmateriaali IBM-verkkosivustolla. Viitattu 28.2.2024. https://www.ibm.com/topics/machine-learning.

What is Machine Learning (ML). N.d. Opetusmateriaalia Google-verkkosivustolta. Viitattu 28.2.2024. https://cloud.google.com/learn/what-is-machine-learning.

Zhu, B., Baesens, B., Backiel, A. & vanden Broucke S.K.L.M. 2017. Benchmarking sampling technique for imbalance learning in churn prediction. Journal of the Operational Research Society, 69(17), March 2017, 1-17. Viitattu 28.3.2024. https://www.researchgate.net/publication/314283433_Benchmarking_sampling_techniques_for_imbalance_learning_in_churn_prediction#pf11.

Yenigün, O. Handling Class Imbalance in Machine Learning. 28.2023. Artikkeli Medium-verkkosivustolla. Viitattu 14.5.2024. https://python.plainenglish.io/handling-class-imbalance-in-machine-learning-cb1473e825ce.

Yu, X., Guo, S., Guo, J. & Huang, X. An extended support vector machine forecasting framework for customer churn in e-commerce. Expert Systems with Applications, 38, 3, 1425-1430. Viitattu 28.3.2024. https://doi.org/10.1016/j.eswa.2010.07.049.(https://www.sciencedirect.com/science/article/pii/S0957417410006779.