Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Hele-Liis Peedosk

Logistilise regressiooni ja otsustuspuumeetodite kasutamine otsemüügi efektiivsuse suurendamiseks

Matemaatilise statistika eriala
Bakalaureusetöö (9 EAP)

Juhendaja: prof. Kalev Pärna

Tartu 2017

Logistilise regressiooni ja otsustuspuumeetodite kasutamine otsemüügi efektiivsuse suurendamiseks

Otsemüük on viis, kuidas suurendatakse läbi klientidega vahetu kontakti loomise ettevõtte toodete tarbimist. Jättes välja kliendid, kes tõenäoliselt ei ole tootest huvitunud, saab vähendada otsemüügile kuluvaid ressursse ning suurendada selle kasutegurit. Potentsiaalsete klientide eristamiseks mittepotentsiaalsetest kasutatakse prognoosimudeleid, mille loomine oli ka antud töö eesmärgiks. Mudelid koostati kasutades logistilist regressiooni, otsustuspuid ja otsustusmetsi. Kasutatud andmestikku kuulus 8412 telefoni teel tehtud pakkumist, mis suunas kliente krediitkaardi lepingut vormistama. Andmestik sisaldas 80 tunnust, mis kirjeldasid klientide poolt tehtud arveldus- ja kaardimakseid ning pakutavate teenuste kasutamist. Töö tulemusena valmis mitu kasulikku prognoosimudelit, mille kasutamisel väheneb mitteresultatiivsete kõnede arv üle 30%, tagades samaaegselt, et vähemalt 95% potentsiaalsetele klientidele tehakse pakkumine.

Märksõnad:
Otsemüük, prognoosimudel, regressioonanalüüs, puud
CERCS teaduseriala: Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika (P160)

Using logistic regression and decision tree methods to improve efficiency of direct selling campaigns

Direct selling is a way for a company to increase consumption of its products and services through direct contact with clients. By leaving out clients who are not likely to be interested in the product, costs of the sales process can be reduced together with the gain in efficiency. The aim of this Bachelor's Thesis is to build predictive models in order to distinguish potential consumers. Data set contained 8412 phone call offerings to open a credit card account, and the variables mostly contained information about payment and card transactions, used services, and signed contracts. Methods used were logistic regression, decision trees, and random forests, and as a result multiple models were fitted. These models are able to identify 30% of clients who are not interested in the product while ensuring that at least 95% of potential consumers will get the offer.

Keywords:
Direct selling, predictive model, regression analysis, trees
CERCS research specialisation: Statistics, operation research, programming, actuarial mathematics (P160)

Sisukord

Sissejuhatus

Tänapäevases konkurentsirohkes ettevõtluskeskkonnas on peamine äritegevuse eesmärk ettevõtte väärtuse kasvatamine. Selleks tuleb tagada pikaajaline konkurentsivõimeline tootlus, mida peegeldab lühiajaline mõõdik - kasum. Pankadel, nagu ka teistel ettevõtlusasutustel, on vaja kasumi suurendamiseks kasvatada pakutavate teenuste ja toodete müüki. Kolm põhilist võimalust selle saavutamiseks on leida uusi kliente, suurendada olemasolevate klientide tarbimist või vältida klientidega koostöö lõpetamist.

Esiteks, oluline on pöörata tähelepanu kliendisuhete hoidmisele. Sanz Saiz ja Pilogre (2010) märgivad, et pangale on olemasolevate klientide hoidmine kuus korda soodsam, kui uute klientide leidmine. Lisaks toovad autorid üle Euroopa läbiviidud uuringu analüüsi tulemusena välja, et kliendisuhete lõpetamise põhjusteks olid eelkõige madal klienditeeninduse tase ja pakutavate teenuste maksumus. Selleks, et säilitada kliendi rahulolu pangaga, on vaja muuta kümnete või isegi sadade tuhandete inimeste kohta teada olevad andmed kasulikuks infoks, et läheneda igale üksikule kliendile personaalsemalt.

Pankadel tekib andmeid klientide kohta igapäevaselt, mis enamasti salvestatakse ja säilitatakse andmeaitades. Klientide kohta käivad andmed, mida pangad omavad, on nii isikut kirjeldavad demograafilised tunnused kui ka tehtud tehingute ja kasutatud toodete logid. Tehingute andmete uurimisel saab infot tarbimisharjumustest, vajadustest ja nende dünaamikast aja jooksul.

Teiseks, tähtis on ka suurendada klientide tarbimist. Pankadel on keeruline leida uusi kliente, kes poleks pakutavate teenustega varem kokku puutunud. Seetõttu peavad krediidiasutused panema rõhku just olemasolevate klientide lojaalsusele. Selleks, et klient ei otsustaks kasutada konkurentide samaväärseid teenuseid, vaid kasutaks antud ettevõtte teenuseid rohkem, peab pank pakkuma tooteid ja teenuseid, mis on klientidele eelkõige vajalikud, kuid samas ka meeldivad ja huvipakkuvad.

Tarbimist suurendatakse aktiivselt toodete pakkumise ja otsemüügi abil. Otsemüük on müügistiil, kus tooteid ja teenuseid pakutakse otse isiku poole pöördudes ning seda tehakse kasutades ebatraditsioonilist jaemüügi kanalit (Peterson & Wotruba, 1996). Levinuimad näited on meilikampaaniad, telefoni- ja ukselt uksele müük. Otsepakkumised on kasulikud, sest need võimaldavad edastada potentsiaalsele ostjale rohkem informatsiooni ning seda saab läbi viia paindliku ajagraafiku järgi, leides just kliendile sobiva aja (Peterson & Wotruba, 1996).

Klientide kohta leiduva informatsiooni kasutamine võimaldab teha otsepakkumisi efektiivsemalt. Prognoosides, millised kliendid on tõenäolisemalt teatud tootest või teenu-

sest huvitunud, võib võtta ühendust potentsiaalsemate ostjatega. Leides üles huvitunud kliendid, saab kulutada müügile vähem ressursse, suurendades samal ajal resultatiivsete pakkumiste arvu.

Käesoleva töö eesmärk on konstrueerida prognoosimudel telefonimüügi efektiivsuse suurendamiseks. Selleks kasutatakse klientide tehingute ja teenuste tarbimise andmeid, et saada infot kliendi tarbimisharjumiste ja pakutava toote praktilise vajaduse kohta. Konstrueeritava mudeli eesmärk on prognoosida võimalikult täpselt kliendi ostusoovi.

Antud töö on jagatud kaheks osaks. Esimeses pooles antakse ülevaade prognoosimudeli konstrueerimiseks kasutatud klassikalise statistika logistilise regressiooni meetodist ning masinõppe otsustuspuude meetodist. Töö teises pooles tehakse mõlema meetodiga prognoosimudel, kasutades ühe Eesti krediidiasutuse telefonimüügi andmeid, mis on kogutud aastatel 2014-2017.

1 Ülesande matemaatiline püstitus

1.1 Statistiline mudel

Käesolev peatükk põhineb autorite James, Witten, Hastie ja Tibshirani (2015: 15-24) raamatul.

Statistilise mudeli treenimiseks on vaja statistilist andmestikku, mis koosneb uuritavast tunnusest Y, mida võib nimetada ka funktsioon- või sõltuvaks tunnuseks, ning p erinevast seletavast tunnusest $X_1, X_2, \ldots, X_p$, mida nimetatakse sageli ka sõltumatuteks, kirjeldavateks või argumenttunnusteks. Eeldame, et Y ja $X = (X_1, X_2, \ldots, X_p)$ vahel esineb mingi seos

$$Y = f(X) + \varepsilon,$$

kus suurus ε on juhuslik viga.

Mudeli eesmärk on saada teadmata kujul olevale funktsioonile f selline hinnang $\hat{f}$, et $Y \approx \hat{f}(X)$. Tänu saadud hinnangule $\hat{f}$ on võimalik teadaolevate kirjeldavate tunnuste X põhjal prognoosida funktsioontunnuse Y väärtust, kasutades seost

$$\hat{Y} = \hat{f}(X).$$

Lisaks on võimalik hinnatud funktsiooni $\hat{f}$ kasutades leida seoseid uuritava ja seletavate tunnuste vahel. Saab vastata küsimustele, millised sõltumatud tunnused on seotud sõltuva tunnusega ning kui tugev on leitud seos.

Funktsiooni f kuju saab hinnata parameetriliste ja mitteparameetriliste meetoditega.

Statistilise mudeli treenimisel parameetriliste meetoditega tehakse eelnevalt oletus funktsiooni f üldise kuju kohta. Seejärel kasutatakse valitud funktsiooni määravate parameetrite hindamiseks sobivat protseduuri, mis kasutab sisendina treeningvalimit. Sellise lähenemise tulemusena peab hindama ainult loetud arvu parameetreid ning see teeb mudeli tegemise arvutuslikult kiiremaks ja lihtsamaks. Väga oluline on teha korrektne oletus, et saadud mudel sobiks andmetega.

Mitteparameetriliste meetodite kasutamisel ei tehta funktsiooni f kuju kohta eeldusi, vaid üritatakse leida hinnang $\hat{f}$, mis oleks andmepunktidele võimalikult lähedal. Selline lähenemine on kasulik just keerulisemate funktsioonide hindamisel. Mitteparameetriliste meetodite puhul ei taandata funktsiooni f hindamist loetud hulga parameetrite hindamiseks ning seetõttu on vaja suuremat treeningvalimit kui parameetrilise meetodi puhul.

1.2 Klassifitseerimismudel

Tunnust, mille väärtused on mittearvulised, nimetatakse kvalitatiivseks, mittearvuliseks või kategoriaalseks tunnuseks ning selle tunnuse väärtusi nimetatakse kategooriateks või klassideks. Kvalitatiivse uuritava tunnuse Y prognoosimudelit nimetatakse klassifitseerimismudeliks.

Klassifitseerimine on objektide määratlemine ühte eelnevalt defineeritud kategooriasse. Paljude meetodite käigus hinnatakse eelnevalt igasse klassi kuulumise tõenäosus ning tehakse selle tõenäosuse põhjal otsus, milline kategooria objektile määrata (James et al., 2015: 127-129).

Kvalitatiivset tunnust, mille erinevaid võimalikke väärtusi on kaks tükki, nimetatakse binaarseks tunnuseks. Levinuimad kaheväärtuselised tunnused on jah/ei küsimused: kas objektil esineb sündroom, kas objekt on abielus, kas tehing oli seaduslik, kas klient võttis pakkumise vastu jne. Lisaks on võimalik muuta ka teised binaarsed tunnused sündmusel põhinevaks. Näiteks tunnusel 'sugu' on tavaliselt kaks taset: 'mees' ja 'naine'. Nimetades tunnuse ümber 'kas on naine', millel on tase 1, kui tegemist on naisega, ja 0, kui tegemist ei ole naisega, vaid hoopis mehega, saame samuti binaarse tunnuse. Edaspidi nimetatakse käesolevas töös binaarse tunnuse klassi sündmuse toimumise korral positiivseteks ja mittetoimumise puhul negatiivseteks.

Kõige levinum parameetriline klassifitseerimismudel on logistiline regressioon (James et al., 2015: 127), mida kirjeldatakse peatükis 2. Mitteparameetrilisi meetodeid rakendavad mitmed masinõppe meetodid, sh otsustuspuud ja otsustusmetsad (Berry & Linoff, 2004: 8-9).

1.3 Mudelite võrdlemine

Erinevate mudelite prognoosivõime hindamiseks soovitatakse kogutud andmestik jagada kaheks: treening- ja testandmestikuks. Selline lähenemine võimaldab kontrollida mudeli töökindlust uute andmeobjektide klassifitseerimisel (Berry & Linoff, 2004: 78-80).

Binaarse uuritava tunnuse korral võib klassifitseerimismudeliga teha kahte liiki vigu. Negatiivse objekti klassifitseerimisel positiivsesse klassi tehakse I liiki viga ning saadakse väärpositiivne (FP) otsus. Kui objekt on pärit positiivsest klassist, kuid klassifitseeritakse mudeli põhjal negatiivseks, nimetatakse seda väärnegatiivseks (FN) otsuseks ja seejuures tehakse II liiki viga. Mõlemat tüüpi vead moodustavad koos õigesti klassifitseeritud positiivsete (TP) ja negatiivsete (TN) klassisiltidega eksimismaatriksi, mis on kujutatud tabelis 1.

Mudelite prognoosivõime hindamiseks toovad Sokolova ja Lapalme (2009) välja eksimismaatriksi põhjal järgmised arvutatavad statistikud, mille arvutusvalemid on toodud tabelis 1.

- Täpsus (*accuracy*, lüh. *Acc*) näitab korrektselt klassifitseeritud objektide osakaalu.
- Prognoosiviga (*error*, lüh. *Err*) näitab valesti klassifitseeritud objektide osakaalu.
- Kordustäpsus (*precision*, lüh. *Pr*) näitab mudeli poolt positiivseks klassifitseeritud objektide seas tegelike positiivsete osakaalu.
- Tegelike negatiivsete osakaalu negatiivselt klassifitseeritud andmete seas näitab negatiivsete prognooside korrektsus (*negative predictive value*, lüh. *NPV*).
- Tundlikkus (*sensitivity*, lüh. *Sens*) näitab mudeli efektiivsust klassifitseerida tegelik positiivne objekt positiivseks.
- Spetsiifilisus (*specificity*, lüh. *Spec*) näitab mudeli võimet klassifitseerida tegelik negatiivne objekt negatiivseks.
- Tasakaalustatud täpsus (*balanced accuracy*, lüh. *BA*) näitab mudeli võimet vältida valesti klassifitseerimist ja arvutatakse järgnevalt:

$$BA = \frac{1}{2}\Big(Sens + Spec\Big) = \frac{1}{2}\Big(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\Big).$$

Tabel 1. Eksimismaatriks ja sellel põhinevad statistikud

	Tegelik positiivne	Tegelik negatiivne	
Prognoos positiivne	Tõeselt positiivne (TP)	Väärpositiivne (FP)	Kordustäpsus $Pr = \frac{TP}{TP+FP}$
Prognoos negatiivne	Väärnegatiivne (FN)	Tõeselt negatiivne (TN)	$NPV = \frac{TN}{TN+FN}$
	Tundlikkus $Sens = \frac{TP}{TP+FN}$	Spetsiifilisus $Spec = \frac{TN}{TN+FP}$	Täpsus $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

1.4 Probleemid müügimudeli andmetega

1.4.1 Tasakaalustamata andmestik

Tasakaalustamata andmestikuks nimetatakse andmestikku, milles pole kvalitatiivse uuritava tunnuse klassid võrdselt esindatud. Kui ühe klassi osakaal on märkimisväärselt väiksem teistest, nimetatakse sellesse klassi kuulumist harva esinevaks sündmuseks (Ku-

bat & Matwin, 1997). Käesolevas töös eeldatakse tasakaalustamata andmestike puhul, et positiivses klassis on vähem objekte kui negatiivses.

Tihti on uuritavaks sündmuseks just see klass, mis on väiksema osakaaluga. Seejuures on vähemesindatud klassi korrektselt prognoosimine sageli isegi olulisem kui suure osakaaluga klassi puhul. Ebavõrdsete klassidega andmestikud on näiteks krediitkaardi pettuste, sõjaliste konfliktide, harva esinevate haiguste kohta (King & Zeng, 2001). Eelnevalt nimetatud andmestike korral võib olla positiivse klassi valesti prognoosimine väga kulukas ning seda tuleks vältida.

King ja Zeng (2001) toovad välja, et tasakaalustamata andmestiku puhul peab andmeid märkimisväärselt rohkem koguma, et ka väiksema klassi objekte oleks piisavalt palju statistiliselt oluliste mudelite treenimiseks. Siiski märgivad autorid, et piisab kõikide positiivsete ja väiksema hulga juhuslikult valitud negatiivsete objektide kaasamisest, et saada ligikaudselt sama efektiivne mudel kui tervet andmestikku kasutades.

Tasakaalustamata andmestiku puhul ei ole prognoosimisviga parim näitaja, mille järgi hinnata mudeli headust. See võib olla suurema klassi korral kõrge, kuid harva esineva sündmuse klassi korral väga-väga madal. Tuleb valida sobivamad statistikud vastavalt uurimisküsimusele ja kasutatud meetodile.

1.4.2 Kirjeldavate tunnuste jaotus

Kasutades mudeli treenimiseks tehinguid kirjeldavaid tunnuseid, tekib tihti probleem, et need on asümmeetriliselt hajunud ning ebaühtlaselt oma väärtuste piirkonnas jaotunud (Cadez, Smyth, Ip & Mannila, 2003). Kui klient ei oma antud toodet või pole teenust kasutanud, on väärtuseks 0. Seega võib paljude tunnuste korral olla sagedaseim väärtus 0. Samas leidub ka kliente, kellel on samade tunnuste väärtused keskmisest ja mediaanist kordades suuremad. Väga suurte erinevustega väärtused võivad mõjutada mudelile erindina ning see võib mudeli usaldusväärsust vähendada.

2 Logistiline regressioon

2.1 Logistilise regressioonimudeli kuju

Käeosoleva peatüki kirjutamisel on kasutatud autorite Hosmer ja Lemeshow (2000:6-7, 31-33) õpikut.

Binaarse tunnuse väärtuste tõenäosusi prognoositakse väga sageli logistilise regressiooni mudeliga. Olgu funktsioontunnusel Y kaks võimalikku taset, mille tähistame 0, kui sündmust ei toimunud ja 1, kui sündmus toimus. Tähistame suurusega Y_i tunnuse Y väärtuse i-ndal objektil, kus $i = 1, \ldots, n$ ja n on objektide arv andmestikus. Sündmuse esinemise ja mitteesinemise tõenäosusi tähistatakse vastavalt $\pi_i = P(Y_i = 1)$ ja $1-\pi_i = P(Y_i = 0)$.

Kuna prognoositav tõenäosus peab jääma 0 ja 1 vahele, ei saa kasutada tavalist lineaarset regressiooni. Väärtused lõigust [0, 1] teisendatakse üksüheselt reaalarvulisele skaalale kasutades logit-seosefunktsiooni:

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i}.$$

Logit-seosefunktsioon esitatakse juhusliku suuruse X_i realisatsioonide lineaarkombinatsiooniga

$$\log \frac{\pi_i}{1 - \pi_i} = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p,$$

kus $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ on funktsiooni määravad parameetrid, x_{i1},$x_{i2}, \ldots, x_{ip}$ on i-nda vaadeldud objekti kirjeldavate tunnusete väärtused ja p on argumenttunnuste arv.

Kasutades logistilist regressioonimudelit saab prognoosida sündmuse esinemise tõenäosust π_i objektil i, mis on võrdne

$$\pi_i = \frac{e^{\beta_0+x_{i1}\beta_1+x_{i2}\beta_2+...+x_{ip}\beta_p}}{1 + e^{\beta_0+x_{i1}\beta_1+x_{i2}\beta_2+...+x_{ip}\beta_p}}.$$

2.2 Parameetrite hindamine

Hosmer ja Lemeshow (2000:7-10, 33-36) on oma õpikus kirjeldanud ka parameetrite hindamist. Parameetriliste meetodite kasutamisel taandub uuritava ja kirjeldavate tunnuste seost iseloomustava funktsiooni f hindamine loetud arvu parameetrite hindamisele. Logistilise regressioonimudeli korral on hinnatavaid parameetreid $p + 1$ tükki, kus p on kirjeldavate tunnuste arv. Tähistame hinnatavate parameetrite vektori $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$.

Logistilise regressioonimudeli parameetreid $\beta_0, \beta_1, \ldots, \beta_p$ hinnatakse suurima tõepära (STP) meetodil. STP-meetodi põhimõte on leida parameetritele väärtused, mis maksimeerivad antud valimi saamise tõepära.

Maksimeerides tõepärafunktsiooni

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

kus n on valimi objektide arv, saadakse parameetrite $\beta_0, \beta_1, \ldots, \beta_p$ STP-hinnangud. Samuti võib maksimeerida log-tõepära funktsiooni $l(\boldsymbol{\beta}) = \ln[L(\boldsymbol{\beta})]$, mida on enamasti lihtsam arvutada. STP-meetodi korral leitakse parameetrite hinnangud enamasti kasutades iteratiivseid lahendamismeetodeid (Van der Paal, 2014).

2.3 Kirjeldavate tunnuste valik

Käesolev peatükk on refereeritud autorite James et al. (2015: 203-214) raamatust, kui ei ole teisiti märgitud.

Statistiliste mudelite konstrueerimisel üritatakse enamasti leida selline mudel, mis on võimalikult väheste tunnustega, kuid kirjeldab andmeid piisavalt hästi. Kui andmestikus on palju tunnuseid, mille vahel valida, on kõikide tunnuste kombinatsioonide läbi katsetamine väga ressursimahukas viis parima mudeli leidmiseks. Levinud meetod mudelisse kaasatavate kirjeldavate tunnuste automaatseks valimiseks on sammregressioon.

Sammregressiooni ideeks on konstrueerida hea mudel, valides iteratiivselt tunnuseid, mida juurde võtta või välja jätta. Sammregressiooni tehakse peamiselt kolmel erineval viisil.

Ettepoole valiku puhul alustatakse mudelist, kus on ainult vabaliige. Seejärel valitakse lisamiseks tunnus, mille lisamisel mudel paraneb kõige rohkem. Tunnuseid lisatakse niikaua, kuni mudelisse on kaasatud kõik tunnused. Kasutusele võetakse mudel, mis on erinevate tunnuste arvuga mudelite seast parim.

Tahapoole valiku korral alustatakse mudelist, kuhu on kaasatud kõik kirjeldavad tunnused. Seejärel eemaldatakse tunnus, mis on mudelis kõige ebavajalikum. Seda korratakse, kuni jõutakse ainult vabaliikmega mudelini. Kasutatav mudel valitakse ka tahapoole valiku korral nende mudelite seast, mis olid parimad mudelid erinevate parameetrite arvude korral.

Segavaliku puhul lisatakse mudelisse tunnuseid analoogselt ettepoole valiku meetodile, kuid igal sammul kontrollitakse, kas mõne tunnuse lisamine või ärajätmine parandaks mudelit.

Parimat mudelit võib logistilise sammregressiooni korral valida Akaike informatsioonikirteeriumi põhjal (AIC). AIC arvutatakse:

$$\text{AIC} = -\frac{2}{n} \cdot l(\boldsymbol{\beta}) + 2 \cdot \frac{d}{n},$$

kus $l(\boldsymbol{\beta})$ on maksimeeritud log-tõepära funktsioon, n treeningvalimi objektide arv ja d on kaasatud parameetrite arv (Hastie, Tibshirani & Friedman, 2017: 230-232). Sammregressiooni tulemuseks on mudel, mille korral on AIC väikseim.

Sammregressiooni kasutatakse küll laialdaselt, kuid kritiseeritakse aina rohkem. Ratner (2010) toob välja põhjuseid, miks ei tasu sammregressiooni kasutada. Nende hulgas on ka järgnevad väited:

- sammregressiooni tulemusena ei saada kõige paremat mudelit;
- suure multikollineaarsuse esinemisel tekib palju probleeme ning saadud mudel ei ole kasutatav;
- mudelisse võib sattuda palju müratunnuseid;
- lõplikku mudelisse kaasatakse tihti liiga palju kirjeldavaid tunnused ning see võib kaasa tuua ülesobitamise;
- parameetrite hinnangud on liiga suured;
- sammregressiooni tulemusena saadavad p-väärtused ei ole sama sisuga, kui tavalise hüpoteeside testimise puhul.

Hosmer ja Lemeshow (2000: 116-135) on samuti mõne väljatoodud puudusega nõustunud. Siiski nendivad autorid, et kui andmed ja valdkond on analüütikule uued ning ei ole võimalik hetketeadmiste põhjal välja pakkuda oodatavaid seoseid, on sammregressioon üks kasulik meetod esmaseks analüüsiks.

2.4 Logistilise regressiooni kasutamisel tekkivad probleemid

King ja Zeng (2001) toovad välja, et logistilise regressiooni kasutamisel harva esineva sündmuse prognoosimiseks on suurima tõepära meetodiga saadud parameetrite hinnangute vektor $\hat{\boldsymbol{\beta}}$ nihkega hinnang parameetrite vektorile $\boldsymbol{\beta}$. Lisaks on alahinnatud sündmuse toimumise tõenäosus π ning seda ka juhul, kui parameetritele on leitud nihketa hinnang.

Van der Paal (2014) lisab, et parameetrite nihkega hinnangute asemel võib tasakaalustamata andmestiku puhul tekkida olukord, kus suurima tõepära iteratsiooniprotsess ei koondu. Sellisel juhul parameetritele hinnanguid ei leidu ning neid nimetatakse lõpmatuteks parameetriteks. Eelnevalt kirjeldatud nähtust võib põhjustada eralduvus

ehk olukord, kui üks või mitu kirjeldavat tunnust prognoosivad üheselt uuritava tunnuse väärtust. Osalise eralduvuse korral on parameetrite hinnangud küll leitavad, kuid need on liiga suured. Eralduvus on binaarsete uuritavate tunnuste korral tihti esinev probleem, kuid tasakaalustamata andmete korral tuleb seda ette veelgi tihemini.

Kirjeldavate tunnuste valimisel võib tekkida analoogselt lineaarse regressioonimudeliga ka logistilise mudeli konstrueerimisel probleem multikollineaarsusega (Hosmer & Lemeshow, 2000: 1-7). Selleks nimetatakse olukorda, kus kaks või rohkem argumenttunnust on omavahel tugevasti seotud. Kui andmestikus on palju kirjeldavaid tunnuseid, on võimalik, et paljude tunnuste vahel on tugev seos.

James et al. (2015: 99-102) annavad oma õpikus multikollineaarsuse probleemist ja selle lahendamisest hea ülevaate. Multikollineaarsuse korral ei ole küll parameetrite hinnangud nihkega, kuid saadud hinnangud võivad olla ebastabiilsed ning seda väljendavad parameetrite hinnangute $\hat{\beta}_j$ kõrged standardvead, kus $j = 1, \ldots, p$ ja p on kirjeldavate tunnuste arv. Multikollineaarsust kontrollitakse varieeruvusindeksiga (VIF). VIF arvutatakse iga parameetri kohta ning see näitab hinnangu $\hat{\beta}_j$ varieeruvuse suhet teiste argumentidega koos hinnatud mudeli ning ainult parameetriga β_j hinnatud mudeli vahel.VIF arvutatakse igale parameetrile kasutades järgnevat valemit:

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R_{X_j|X_{-j}}}.$$

Eelnevas valemis tähistab suurus $R_{X_j|X_{-j}}$ determinatsioonikordajat, kui mudelis on j-s tunnus avaldatud teiste kirjeldavate tunnuste kaudu. Multikollineaarsus tekitab probleeme, kui VIF > 5.

3 Mitteparameetrilised meetodid

3.1 Otsustuspuud

Statistiliseks modelleerimiseks kasutatakse üha enam otsustuspuude meetodeid, mida saab kasutada nii regressioonanalüüsi kui ka klassifitseerimisülesannete lahendamise osana. Otsustuspuu on reeglite kogum, mis jagab eelnevalt teada olevad andmed gruppidesse, mis on uuritava tunnuse mõttes homogeensemad kui algandmed (Berry & Linoff, 2004: 165-166). Kõiki tehtud tükeldusi on võimalik kujutada hierarhilisel kujul puuna, mistõttu nimetataksegi selliseid meetodeid puumeetoditeks.

Otsustuspuu koosneb järgnevatest osadest (Tan, Steinbach & Kumar, 2006: 150-151):

- kaared – lülid, mis ühendavad kahte tippu;
- tipud – elemendid, kus asuvad andmed. Need jagunevad omakorda:
  - juurtipp – tipp, mis ei ole hargnenud ühestki tipust, kuid millest hargneb välja kaks või rohkem kaart alamtippudesse;
  - vahetipud – tipud, mis on kaartega seotud ühe vanemtipuga ja kahe või rohkem alamtipuga;
  - lehed – tipud, mis on ühendatud ainult ühe vanemtipuga ja ei hargne rohkemateks alamtippudeks.

Tan et al. (2006: 150-151) on selgitanud puude põhjal prognoosimist järgnevalt. Juurtipule ja igale vahetipule on määratud otsustusreegel, mille põhjal valitakse alamtipp. Liikumine vastavalt otsustusreeglitele toimub alates juurtipust läbi vahetippude, kuni jõutakse leheni. Igale lehele on määratud väärtus, mis on sõltuva tunnuse prognoos. Uuritava objekti prognoos on sellele lehele omistatud väärtus, kuhu see objekt vastavalt kirjeldavate tunnuste väärtustele jõuab.

Vahetippudest võib hargneda kaks või rohkem alamtippu. Kui see hargneb kaheks, on tegemist binaarse otsustuspuuga. Otsustusreegel on kahendmuutuja: vastates kas jah või ei, liigutakse kas vasakusse või paremasse kaarde (Berry & Linoff, 2004: 170-171). Joonise 1 vasakpoolsel osal on kujutatud binaarse otsustuspuu struktuur ja parempoolsel osal selle puu tükeldused kahemõõtmelisel tasandil.

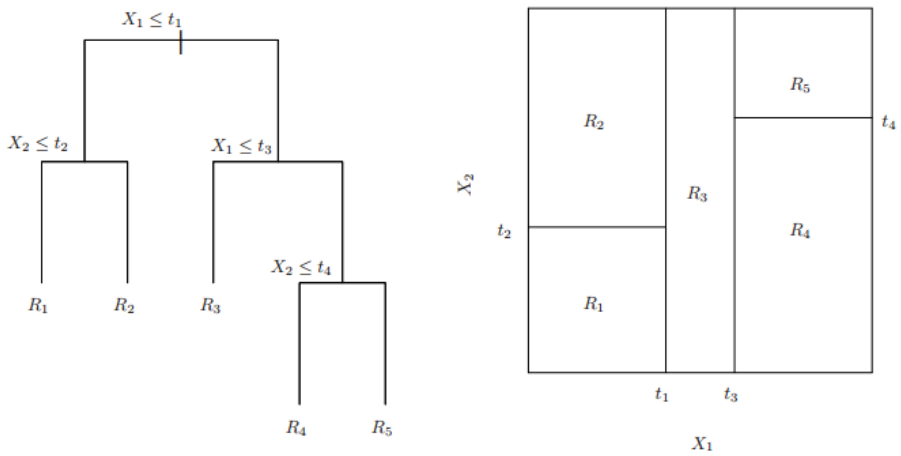
Otsustuspuude kasutamine on levinud eelkõige tänu meetodi lihtsusele. Otsustuspuude eelised on järgmised.

- Osustuspuud on kerge kasutada ja interpreteerida ka mittestatistikutel (James et al., 2015: 303).
- Uuritav ja kirjeldav tunnus võivad olla kompleksse seosega. Sellisel juhul võib

otsustuspuu anda klassikalistest meetoditest (lineaarsed ja üldistatud lineaarsed mudelid) isegi täpsemaid tulemusi (James et al., 2015: 314-315).

- Otsustuspuid saab kasutada nii kvantitatiivsete kui ka kvalitatiivsete tunnuste korral, sealjuures neile eeldusi seadmata. Seega ei teki probleeme erinditega andmestikus ega asümmeetriliselt jaotunud tunnustega (Berry & Linoff, 2004:209).
- Otsustuspuud kirjeldavad andmestikku väga täpselt ning seetõttu saab kasutada otsustuspuid ka info kogumiseks enne mõne järgmise meetodi kasutamist (Berry & Linoff, 2004: 209).

Siiski on otsustuspuudel ka puudusi. Berry ja Linoff (2004: 170) ei soovita kasutada otsustuspuid pideva uuritava tunnuse korral, kuna puu suudab prognoosida vaid diskreetseid väärtusi, mida on sama palju kui lehti. Lisaks ei kasutata arvuliste kirjeldavate tunnuste puhul kogu olemasolevat andmehulka, vaid ainult väärtusi, mille põhjal tükeldati kirjeldavaid tunnuseid. See võib olla ka põhjus, miks autorid James et al. (2015: 315-316) toovad oma raamatus välja, et otsustuspuu prognoositäpsus võib olla madalam kui muudel prognoosimeetoditel. Lisaks märgivad autorid, et ka väikesed muutused andmestikus võivad kaasa tuua märgatava erinevuse hinnatud puu kujus. Sellist ebastabiilsust saab parandada agregeerides mitmete otsustuspuude prognoose (Breiman, 1996). Neid meetodeid on kirjeldatud peatükis 3.2.



Joonis 1. Otsustuspuu ja sellele vastavad lahutused kahemõõtmelisel tasandil (James et al., 2015).

3.1.1 Klassifitseerimispuu

Kvalitatiivse uuritava tunnuse prognoosimise korral nimetatakse konstrueeritud otsustuspuud klassifitseerimispuuks.

Uuritava tunnuse prognoos on klassifitseerimispuu puhul lehele valitud klassi määratlev

silt. Kasutatakse treeningandmestikku, kus iga objekti kohta on teada nii uuritav tunnus kui ka kirjeldavad tunnused. Klassisilt on treenitava andmestiku objektide ühte lehte grupeeritud uuritava tunnuse sagedaseim väärtus (James et al., 2015: 311-314). Lisaks klassi määratlusele saab klassifitseerimispuu korral leida ka klasside proportsioonid lehes. See võimaldab järjestada objekte osakaalult suurimast väikseimani (Berry & Linoff, 2004: 169-170).

3.1.2 Klassifitseerimispuu konstrueerimine

Klassifitseerimispuu konstrueerimiseks kasutatakse mitmeid erinevaid algoritme. Idee on nendel siiski sarnane: leida juurtippu ja igasse vahetippu parimad otsustusreeglid, mis teeksid andmete hulga uuritava tunnuse suhtes aina homogeensemaks (Berry & Linoff, 2004: 172-175).

James et al. (2015: 311-314) toovad välja mitu kriteeriumi, mida võib kasutada klassifitseerimispuude otsustusreeglite headuse mõõtmiseks. Olgu $\hat{p}_{mk}$ k-ndasse klassi kuuluvate objektide osakaal lehes R_m, kus $k = 1, .., K$ ja K on uuritava tunnuse klasside arv ning $m = 1, .., M$, kus M on lehtede arv. Saab minimeerida järgmiseid suuruseid:

- klassifitseerimisviga $E = 1 - \max_k \hat{p}_{mk}$,
- Gini indeksit $G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$,
- summaarset entroopiat $D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$.

Puu konstrueerimisel kasutatakse enamasti Gini indeksit või summaarset entroopiat. Klassifitseerimisviga pole puu ehitamisel piisavalt tundlik kriteerium (James et al., 2015: 311-314).

Tan et al. (2006: 151-155, 164-166) kirjeldasid otsustuspuu konstrueerimise algoritmi järgnevalt. Esmalt tuleb leida juurtipus parim lahutus. Järgnevalt tuleb kontrollida alamtippudes lõpetamiskriteeriumi kehtivust. Kui see veel ei kehti, siis korrata alamtippudes parima lahutuse ja lõpetamiskriteeriumi kehtivuse kontrollimise samme rekursiivselt kuni lõpetamiskriteerium muutub kehtivaks. Kehtiva lõpetamiskriteeriumiga alamtipp ongi puu leht. Lõpetamiskriteeriumiks on tavaliselt üheselt määratud klassisilt ja/või võrdsed kirjeldavate tunnuste väärtused. Lisaks kontrollitakse, et objektide arv tipus poleks alla valitud miinimumi.

Eelnevalt kirjeldatud algoritmi järgides saadakse selline puu, mis kirjeldab andmeid väga detailsel tasemel. Berry ja Linoff (2004: 175-176, 184) toovad välja, et kasutades saadud puud prognoosimiseks, võib tekkida probleeme puu üldistamisvõimetuse tõttu ning tagajärjeks on uute andmete korral kõrge prognoosiviga. Seda saab vältida nii mudeli treenimisel kui ka peale treenimise protsessi. Viimasel juhul tuleb treenitud puu

pügada ehk panna kokku mitu väiksemat lehte. Selliste lehtede valik tehakse, kasutades bootstrap- ja ristvalideerimise meetodeid, kus tehakse vastavalt algoritmile juhuslik valik algsest objektide hulgast. Neid meetodeid on kirjeldanud ka Tan et al. (2006: 187-188).

3.1.3 Otsustuspuude kasutamine tasakaalustamata andmestikel

Tasakaalustamata andmestike pealt treenitud otsustuspuude prognoos on uute andmete korral negatiivse klassi suunas kallutatud (Kubat & Matwin, 1997). Otsustuspuude meetodi puhul üritatakse lehtedes suurendada klasside homogeensust. Kui positiivsed objektid on hajusad ja negatiivseid objekte on märkmisväärselt rohkem, on enamuses lehtedes siiski ülekaalus negatiivsed objektid ning seega on ka terve leht negatiivse klassisildiga (Kubat & Matwin, 1997). Positiivsete objektide suure hajususe tõttu võib tekkida olukord, kus mudelisse jääb ainult juurtipp ja prognoositakse alati negatiivset klassi.

Kotsiantis, Kanellopoulos ja Pintelas (2006) toovad välja meetodeid ebavõrdsusest tingitud probleemide lahendamiseks. Läheneda saab kas andmete või algoritmi tasemel ning lisaks on välja töötatud hübriide, mis kombineerivad mõlemat tüüpi meetodeid. Algoritmi tasemel kasutatakse laialdaselt kulumaatriksitel põhinevat meetodit, mis määrab positiivsete ja negatiivsete klasside valesti klassifitseerimise puhul erineva maksumuse. Kubat, Holte ja Matwin (1997) töötasid välja meetodi SHRINK, mis muudab otsustuspuude algoritmi klassisildi määramisel. Nimetatud meetodi puhul ei määrata klassikuuluvust osakaalu järgi, vaid positiivse klassisildi saavad kõik lehed, kus leidub vähemalt üks positiivne objekt. Meetoditeks andmete tasemel nimetatakse erinevaid taasvaliku meetodeid, mis tasakaalustavad andmestikku. Neid saab jagada ala- ja ülevalikuks[1].

Kotsiantis et al. (2006) kirjeldavad mõlemat taasvaliku meetodit. Alavaliku puhul vähendatakse suurema osakaaluga klassi objekte, jättes sellest juhusliku valiku põhjal nii palju objekte välja, et oleks saavutatud soovitud osakaal. Alavaliku puhul jäetakse kasutamata suur hulk potentsiaalselt kasulikku infot sisaldavaid andmeid. Ülevalikul sellist puudust ei ole, kogu info jääb alles. Selle meetodi puhul suurendatakse vähemesinenud klassi osakaalu, tehes selle klassi objektidest juhusliku valiku põhjal koopiaid. Seda tehakse niikaua, kuni saavutatakse soovitud osakaal. Valimi suurendamine tõstab märkimisväärselt arvutuslikku mahtu. Veelgi enam, täpsete koopiate tegemine tõstab positiivsete objektide kaalu ning seega ka ülesobitamise võimalust. Kumb valikumeetod paremaid tulemusi annab, sõltub nii uurimisküsimusest, kasutatavast andmestikust kui ka klassifitseerimis-

[1] Ametlike tõlgete puudumise tõttu tõlgib autor inglisekeelsed terminid „*undersampling*“ ja „*oversampling*“ vastavalt alavalik ja ülevalik.

algoritmist (Liu, 2004).

Ala- ja ülevalikust on väljatöötatud ka modifikatsioone, lahendamaks nende meetodite puudusi. Chawla, Bowyer, Hall ja Kegelmeyer (2002) töötasid välja meetodi SMOTE. See teostab samuti ülevalikut, kuid juhuslikult valitud vähemesindatud objekti kopeerimise asemel interpoleeritakse uus objekt, mis ei kattu eelnevate objektidega. Seda meetodit saab kombineerida ka alavalikuga. Tänu sellisele modifikatsioonile välditakse mudeli ülesobitamist, mis võib tekkida tavalise ülevaliku puhul.

3.2 Mitme puu agregeerimine

3.2.1 *Bagging*

Breiman (1996) tuli ideele teha sama andmestiku pealt mitmeid otsustuspuid ning nende prognoose agregeerides saada üksikust otsustuspuust parema prognoosivõimega mudel. Iseenesest mõistetavalt ei saa kaasata täpselt samu objekte treeningprotsessi, vaid kasutada tuleb mingil määral erinevat objektide hulka. Meetodit, mille Breiman (1996) välja töötas, nimetatakse bootstrap-agregeerimiseks (*bootstrap aggregating*) ning lühidalt *bagging*-meetodiks.

James et al. (2015: 316-317) kirjeldavad selle meetodi algoritmi järgnevalt.

1. Moodustada algandmest lihtsa juhusliku tagasipanekuga valiku abil B treeningandmestikku, mis on sama suured kui algne andmestik. Selliseid valimeid nimetatakse bootstrap-valimiteks.
2. Treenida B erinevat kärpimata otsustuspuud, kasutades erinevaid bootstrap-valimeid.
3. Fikseerida iga puu korral saadud prognoos.
4. *Bagging*-meetodi prognoos uuritavale objektile on kõikide bootstrap-prognooside keskmine kvantitatiivse uuritava tunnuse korral ning sagedaseim prognoos kvalitatiivse uuritava tunnuse korral.

Bagging-meetod vähendab otsustuspuude ebastabiilsust ning parandab prognoosivõimet. Kuna iga treenitud puu on kärpimata kujul, on need kõrge dispersiooni, kuid väikese hinnangu nihkega. Kõrge dispersiooni probleemi lahendab puude agregeerimine (James et al., 2015: 316-317). Üksikute puude ja *bagging*'u prognoosivõimet on võrreldud erinevate autorite poolt, sh Breiman (1996), Dietterich (2000), ning järeldati, et *bagging*-meetodiga saadakse väiksema prognoosiveaga mudel. Samas möönab Breiman (1996), et võites mudeli täpsuses, kaotatakse lihtsal moel visualiseeritav ja interpreteeritav struktuur.

3.2.2 Otsustusmets

Breiman (2001) arendas *bagging*-meetodit edasi, lisades sellele juhusliku kirjeldavate tunnuste valiku. Sellist meetodit nimetatakse otsustusmetsaks. Otsustusmetsa kasvatamise algoritm sarnaneb *bagging*'ule, erinevus on ainult puu treenimise meetodis. Uuema meetodi korral kaasatakse juur- ja vahetippudes parima otsustusreegli välja selgitamisel ainult piiratud hulk juhuslikult valitud argumente. Tihti valitakse selline argumentide arv m, et $m \approx \sqrt{p}$, kus p on kõikide argumentide arv andmestikus (James et al., 2015: 319-321). Ka otsustusmetsas ei kärbita saadud puid ning lõpp-prognoos saadakse kõikide prognooside agregeerimisel. *Bagging* on otsustusmetsa erandjuht, kui $m = p$.

Otsustusmetsa meetod muudab puude ehitust erinäolisemaks, kuna puude ehitus pole enam nii tugevalt seotud uuritavat tunnust kõige paremini kirjeldavate tunnustega. Väheneb treenitud puude omavaheline korrelatsioon ning seetõttu väheneb dispersioon ka puude agregeerimisel (James et al., 2015: 319-321).

3.2.3 Tasakaalustatud otsustusmets

Chen, Liaw ja Breiman (2004) pakkusid välja otsustusmetsi ning alavalikut ühendava algoritmi, mida nimetatakse tasakaalustatud otsustusmetsaks (edaspidi BRF, mis on lühend inglisekeelsest nimetusest *Balanced Random Forest*). Kuna tasakaalustamata andmete kasutamisel treenitud otsustusmetsad üritavad minimeerida üldist klassifitseerimisviga, kaasneb sellega tihti positiivse klassi prognoosivea suurenemine, kuna keskendutakse suurema osakaaluga klassile. Tasakaalustatud otsustusmets vähendab sellist võimalust.

BRF-i konstrueerimise algoritm põhineb tavaliste otsustusmetsade algoritmil, ainuke erinevus on otsustuspuude treenimiseks kasutatud valimis. Iga tasakaalustatud otsustusmetsa puu treenimiseks võetakse algsest objektide hulgast vähemesindatud klassist bootstrap-valim ning suurema osakaaluga klassist juhuslik tagasipanekuga valim, mis on sama suur kui positiivne klass. Saadud valimi pealt treenitakse klassifitseerimispuu analoogselt otsustusmetsa algoritmile ja lõplik prognoos saadakse agregeerides kõikide konstrueeritud puude prognoose.

Meetodi autorid väidavad peale empiirilisi katseid, et BRF annab paremaid tulemusi, võrreldes SMOTE ja SHRINK klassifitseerimispuu meetoditega. Tulemusi võrreldi mitmete headusnäitajate, sh täpsuse ja kordustäpsuse alusel, mida on kirjeldatud peatükis 1.3.

4 Ülesande praktiline lahendus

Töö praktilises osas antakse ülevaade andmestikust, kirjeldatakse valitud tunnuseid, rakendatakse teoreetilises osas kirjeldatud meetodeid püstitatud eesmärgi saavutamiseks ning võrreldakse saadud mudeleid headusnäitajate poolest. Mudelite tegemiseks ning võrdlemiseks on kasutatud tarkvara R ja selle lisapakette.

4.1 Andmestiku kirjeldus

Koostatava müügimudeli eesmärk on prognoosida, kui tõenäoliselt võtab klient vastu otsemüügipakkumise. Kasutatud andmestikku kuulub 8412 telefoni teel tehtud pakkumist, mille eesmärk oli propageerida krediitkaardi kasutuselevõttu. Andmed koguti klientide kohta, kellele tehti aastate 2014-2017 jooksul müügikõne.

4.1.1 Andmestiku ülevaade

Uuritavaks tunnuseks on dihhotoomne tunnus 'leping', millel on väärtus 0, kui valimisse sattunud klient ei teinud lepingut, ning 1, kui vormistati leping. Kuna lepingu sõlmimine võib võtta mõningal juhul natukene rohkem aega, tuli seada ajaline piir. Otsustati, et kui kõne saanud isik vormistab lepingu, mis hakkab kehtima hiljemalt 90 päeva jooksul, on tunnusel 'leping' tase 1.

Kogutud andmestikus on leping vormistatud 9,7% juhtudest 90 päeva jooksul peale kõnet. Kuna uuritava tunnuse tasemed ei ole võrdse osakaaluga, on tegu tasakaalustamata andmestikuga. Seda tuleb arvesse võtta prognoosimudeli konstrueerimisel ning mudelite võrdlemisel.

Mudeli treenimiseks ja selle prognoosivõime hindamiseks jagati andmestik eelnevalt juhusliku valiku põhjal treening- ja testandmestikuks, kuhu kuulus vastavalt $\approx 70\%$ ja $\approx 30\%$ algsest andmestikust. Kõikide valimite mahud ja edukate kõnede osakaalud on välja toodud tabelis 2.

Tabel 2. Andmestiku ülevaade

	Valimi suurus	Edukaid kõnesid
Terve andmestik	8412	812 (9,7%)
sh treeningandmestik	*5888*	*584 (9,9%)*
sh testandmestik	*2524*	*228 (9,0%)*

4.1.2 Kirjeldavad tunnused

Lisaks uuritavale tunnusele on andmestikus 80 kirjeldavat tunnust. Ärisaladuse kaitsmise tõttu ei ole võimalik nimetada kõiki andmestikus olevaid tunnuseid, kuid antakse siiski nendest põgus ülevaade blokkidena.

Klientide isikuandmetest on andmestikku kaasatud ainult sugu ja vanus. Äritegevuse ning klientide aktiivsuse kasvamise tõttu on kaasatud ka aasta, millal pakkumine tehti, ning selle kasutamisel eeldatakse sama tendentsi ka järgnevatel aastatel. Neid tunnuseid tähistatakse logistilise regressioonimudeli valemites suurtähega I.

Väga oluline tunnus on 'kõne', mis on tasemega 1, kui klient võttis kõne vastu ja talle oli võimalik pakkumine edastada, ning tasemega 0, kui klient ei võtnud kõnet vastu või kui tal polnud helistamise hetkel võimalik telefoniga rääkida. See tunnus on kaasatud kirjeldava tunnusena just seetõttu, et identifitseerida tunnuseid, mis mõjutavad klientide krediitkaardi kasutuselevõttu ka ilma otsepakkumist saamata. Küll aga pole teada enne, kui kõne on tehtud, kas klient võtab kõne vastu ja on nõus pakkumist ära kuulama. Tunnuse 'kõne' prognoosimudelisse kaasamisel arvutatakse tulevaste pakkumiskampaaniate käigus uute klientide tõenäosuste prognoosid selle tunnuse mõlema taseme kohta. Selline lähenemine motiveerib ka müügikonsultante, kuna on võimalik näha, kas ja kui palju võib nende tehtud töö mõjutada kliendi edasist käitumist.

Ülejäänud tunnused kirjeldavad kliendi toodete ja teenuste kasutamist. Need jagunevad järgmisteks gruppideks.

1. Lepingud – 10 tunnust, mis hõlmavad kliendilepingu kehtivuse pikkust ja erinevate lepingute, näiteks järelmaksu- ja investeerimisteenuste lepingute sõlmimise indikaatortunnuseid. Mudelites tähistatakse kirjeldatud tunnused tähega L.
2. Teenused – 3 tunnust, mis näitavad pangateenuste kasutamise aktiivsust. Siia hulka kuulub näiteks aktiivsete pangakaartide arv. Mudelites tähistatakse neid tunnuseid tähega T.
3. Kaarditehingud – 14 tunnust, mis kirjeldavad kõiki kaardimakseid. Nende hulka kuuluvad ka tunnused, mis kirjeldavad hotellides tehtud tehinguid. Kirjeldatud tunnuseid tähistatakse tähega K.
4. Kaarditehingud välismaal – 10 tunnust, mis kirjeldavad kaardimakseid välismaal. Mudelites on vastavad tunnused tähistatud tähekombinatsiooniga KV.
5. Arveldusmaksed – 19 tunnust, mis kirjeldavad sissetulevaid ja väljuvaid makseid arvelduskontodel. Siin ei ole eristatud siseriiklikke ja välismakseid. Need tunnused tähistatakse mudelites tähega M.

6. Välismaksed arvelduskontodel – 20 tunnust, mis kirjeldavad väljuvaid makseid välismaiste pankade kontodele ja sissetulevaid makseid välismaistelt kontodelt. Mudelites tähistatakse need tunnused tähekombinatsooniga MV.

Eelpool mainitud tehinguid kirjeldavad tunnused sisaldavad järgmist infot. Kõigepealt võeti välja kliendi poolt tehtud tehingute arvud ja summad kõnele eelneva 180 päeva jooksul ning summeeriti 30 päeva kaupa. Seejärel arvutati nende summade pealt nii 180 päeva kui ka 90 päeva tehingute kogusummasid, keskmisi tehingute summasid, osakaale ning aktiivseid tehingute tegemise ajavahemikke.

4.2 Mudelite võrdlemine

Käesoleva töö kontekstis on parim müügimudel selline, mille tulemusena suudetakse vähendada sellistele klientidele tehtavate kõnede arvu, kes ei ole antud tootest kindlasti huvitunud. Et helistamise nimekirjadest ei jäetaks välja kliente, kes võiksid pakkumise vastu võtta, hoitakse klassifitseerimisel väärnegatiivsete prognooside arvu minimaalsena. Selle saavutamiseks peab mudel korrektselt prognoosima enamuse tegelikest positiivsetest objektidest ning seda näitab mudeli tundlikkus. Tundlikkuse taset saab muuta klassifitseerimislävendi muutmisel. Klassifitseerimislävend on piir, millest väiksema tõenäosusega kliendid klassifitseeritakse negatiivseks ja millest suurema tõenäosusega kliendid klassifitseeritakse positiivseks. Kui valida lävend, mis maksimeerib tundlikkuse, võib jõuda olukorrani, kus mudel prognoosib alati positiivsesse klassi ehk tegelike negatiivsete objektide prognoosid on väärpositiivsed ning seetõttu ka spetsiifilisus minimaalne. Mudelite võrdlemiseks tuleb tundlikkuse ja spetsiifilisuse vahel leida sobiv tasakaal.

Enamasti on klassifitseerimislävend 0,5, mis tähendab, et objekt klassifitseeritakse positiivsesse klassi, kui sündmuse esinemise tõenäosus on üle 0,5. Tasakaalustamata andmete kasutamisel on nii logistilise regressiooni kui ka otsustuspuude meetodite prognoosid sündmuse esinemise tõenäosusele alahinnatud (ptk 3.1.3, 2.4). Seetõttu ei tasu saadud tõenäosuse prognoosi kasutada absoluutse, vaid suhtelise mõõdikuna, mis on aluseks objektide järjestamisel. Eelnimetatud põhjusel valitakse klassifitseerimislävend vastavalt mudelile.

Mudeleid otsustati hinnata järgmiselt. Esmalt valitakse testandmestikule prognoositud tõenäosuste põhjal klassifitseerimislävend, mille korral $Sens \gtrapprox 0{,}95$. Parimaks mudeliks valitakse klassifitseerimismudel, mille spetsiifilisus on suurim vastavalt valitud lävendile. Erinevate mudelite headusnäitajaid võrreldakse kasutades testandmestikule arvutatud prognoose.

4.3 Mudelite konstrueerimine logistilise regressiooni abil

Esmalt koostatakse klassifitseerimismudeleid logistilise regressiooni meetodil. Logistilise regressioonimudeli konstrueerimisel on oluline märgata kitsaskohti ja leida meetmed kas nende kõrvaldamiseks või kontrolli all hoidmiseks. Antud töös kasutatakse vastavate meetmetena argumenttunnuste väärtuste grupeerimist ja teisendamist ning seejärel tunnuste valimiseks sammregressiooni. Viimast kasutatakse punktis 2.3 välja toodud puuduste tõttu ettevaatlikusega.

4.3.1 Logistiline regressioonimudel

Andmestikus on 80 tunnust, mille seast valitakse mudelisse olulisi tunnuseid. Kirjeldavate andmete hulgas on tunnnuste blokid, mis kirjeldavad sarnaseid andmeid erineva nurga alt. Mudeli stabiilse prognoosivõime tagamiseks peavad iga parameetri multikollineaarsust kirjeldavad suurused olema väiksemad kui 5 ehk $\mathrm{VIF}(\hat{\beta}_j) < 5$, kus $j = 1, \ldots, p$ ja p on parameetrite arv andmestikus. Kuna andmestikus on tunnuseid palju, tehti multikollineaarsuse vähendamiseks tunnuste seast automaatne valik. Selle käigus eemaldati logistilisest mudelist kõige suurema varieeruvusindeksiga parameetri hinnangule vastav tunnus kuni kõik allesjäänud parameetrite hinnangute VIF-id olid väiksemad valitud piirist. Parema prognoosimudeli saavutamiseks otsustati käesolevas töös, et VIF peab olema väiksem kui 3.

Seejärel valiti sammregressiooniga multikollineaarsuse testi läbinud tunnuste seast välja sellised, mille korral oli mudeli headusenäitaja AIC kõige suurem. Sammregressiooniga koostati kaks mudelit, neist esimese puhul ei kaasatud tunnuste koosmõjusid tunnusega 'kõne', teise puhul aga kaasati.

Sammregressiooni asemel soovitab Ratner (2010) mudelisse kaasatavaid kirjeldavaid tunnuseid valida lähtudest eksperthinnangust. Seetõttu konstrueeriti lisaks sammregressioonile ka mudel, kuhu tunnuseid valiti kõikide kirjeldavate tunnuste seast.

Võrreldavates mudelites on kõik tunnused statistiliselt olulised olulisusnivool α =0,1. Saadud mudelite tulemused on koondatud tabelisse 3. Kõikide leitud prognoosimudelite abil on võimalik vähendada mitteresultatiivsete kõnede arvu üle 30%. Selleks tuleb jätta välja kliendid, kelle prognoosid on vastavalt mudelile fikseeritud klassifitseerimislävendist madalamad. Sellisel viisil suurendakse müügikonsultantide töö efektiivsust.

Tabel 3. Logistiliste regressioonimudelite võrdlus.

	Oluliste parameetrite arv	Klassifitseerimis-lävend	Tundlikkus	Spetsiifilisus
Sammregressioon koosmõjudeta	13	0,056	0,952	**0,325**
Sammregressioon koosmõjudega	17	0,060	0,952	**0,324**
Ekspertvalik	12	0,077	0,952	**0,338**

Erinevused mudeli headusenäitajates on väikesed ning selline järjestus võib olla tekkinud kasutatud testandmestiku eripärast. Lisaks olid mitmed tunnused olulised kõikides konstrueeritavates mudelites ning seetõttu on ka kõikidesse nendesse mudelisse kaasatud. Seega võivad mudelid prognoosida väga sarnaselt ka mitmete samade tunnuste tõttu.

Testandmestiku prognooside põhjal saavutas suurima spetsiifilisuse ekspertvaliku põhjal tehtud mudel, mille tunnused olid valitud automatiseeritud valikumeetodeid kasutamata. Kui multikollineaarsuse tõttu jäetakse teatud tunnused edasisest modelleerimisest välja, on võimalik, et eemaldatakse just see tunnus suure multikollineaarsusega tunnuste seast, millel oleks mudeli konstrueerimise hilisemas faasis suurem tähtsus kui allesjäetavatel tunnustel. Siiski on vaja teha selline valik tunnuste seast, et sammregressiooni ei kaasataks multikollineaarseid tunnuseid. Hilisemas faasis oluliste tunnuste välja jätmine võib olla põhjuseks, miks automatiseeritud valikuga meetodid ei suutnud teha paremaid mudeleid.

Suurima spetsiifilisusega mudelisse kaasati 12 tunnust, mis kirjeldavad kaardimaksete suurust ja nende tegemise tihedust nii Eestis kui ka välismaal, erinevate lepingute kehtivust ning aega, kui kaua kliendileping on kehtinud. Lisaks osutus oluliseks tunnus 'kõne'. Kasutades peatükis 4.1.2 toodud tähistusi ja alaindekseid $1, \ldots, k$, kus k on oluliste parameetrite arv, saame esitada välja valitud mudeli kujul

$$\begin{aligned}\text{logit}(\pi_i) = &- 3{,}91 + 2{,}05 \cdot k\tilde{o}ne + 0{,}0006 \cdot K_1 - 0{,}54 \cdot KV_2 + 1{,}48 \cdot L_3 + 0{,}26 \cdot K_4 + \\ &+ 0{,}02 \cdot M_5 + 0{,}20 \cdot L_6 - 0{,}83 \cdot KV_7 + 0{,}07 \cdot KV_8 - 0{,}06 \cdot L_9 + 0{,}21 \cdot T_{10} + \\ &+ 0{,}57 \cdot L_{11}.\end{aligned}$$

Väga olulise tunnuse 'kõne' ($p < 2e{-}16$) kordaja hinnang 2,05 näitab, et kui suudetakse kliendile pakkumine edastada, suurenevad lepingu vormistamise šansid $e^{2{,}05} = 7{,}8$ korda. See näitab, et klientide tähelepanu pööramine valitud tootele suurendab müüki.

4.3.2 Grupeeritud andmete pealt konstrueeritud mudel

Tehinguid kirjeldavad tunnused on jaotunud väga ebasümmeetriliselt, lisaks on märgataval osal klientidest erinevate tunnuste väärtuseks 0. Vastavad tunnused grupeeriti, et ekstreemsete väärtustega objektid ei mõjutaks mudeli prognoosivõimet. Tehinguid kirjeldavad tunnused, mille väärtused olid nii negatiivsed kui ka positiivsed, grupeeriti viide rühma:

- -2, kui tunnuse väärtus on negatiivsete väärtuste mediaanist väiksem;
- -1, kui tunnuse väärtus on negatiivsete väärtuste mediaanist suurem;
- 0, kui tunnuse väärtus on võrdne 0-ga;
- 1, kui tunnuse väärtus on positiivsete väärtuste mediaanist väiksem;
- 2, kui tunnuse väärtus on positiivsete väärtuste mediaanist suurem.

Tunnused, mille väärtused olid mittenegatiivsed, grupeeriti nelja rühma:

- 0, kui tunnuse väärtus on võrdne 0-ga;
- 1, kui tunnuse väärtus on positiivsete väärtuste tertsiilist $q_{\frac{1}{3}}$ väiksem;
- 2, kui tunnuse väärtus on positiivsete väärtuste tertsiilist $q_{\frac{1}{3}}$ suurem ja väiksem tertsiilist $q_{\frac{2}{3}}$;
- 3, kui tunnuse väärtus on suurem positiivsete väärtuste tertsiilist $q_{\frac{2}{3}}$.

Kui tunnuse mingis moodustatud grupis oli objekte vähem kui 100, grupeeriti uuesti kõik väärtused klassidesse positiivsed, negatiivsed ja 0-ga võrdsed.

Analoogselt grupeerimata tunnustele konstrueeriti ka grupeeritud andmete puhul kolm erinevat mudelit. Esimesse mudelisse valiti sammregressiooni abil olulised tunnused just nende eelnevalt valitud tunnuseid seast, mille puhul $\mathrm{VIF}(\hat{\beta}_j) < 3$. Teine mudel koostati sarnaselt esimesele, kuid sammregressiooni kaasati ka koosmõjud tunnusega 'kõne'. Kolmas mudel koostati mitteautomaatselt intuitiivse valiku põhjal. Saadud mudelite tulemused on esitatud tabelis 4.

Tabel 4. Grupeeritud andmete pealt treenitud logistiliste regressioonimudelite võrdlus.

	Oluliste parameetrite arv	Klassifitseerimislävend	Tundlikkus	Spetsiifilisus
Sammregressioon koosmõjudeta	9	0,052	0,952	**0,323**
Sammregressioon koosmõjudega	20	0,062	0,952	**0,324**
Ekspertvalik	9	0,080	0,952	**0,345**

Taaskord on parima spetsiifilisusega see mudel, mis konstrueeriti automatiseeritud meetodeid kasutamata. Sel korral on erinevus natuke suurem, kuid samuti ei saa olla kindel, et see väike erinevus on statistiliselt oluline.

Ka grupeeritud andmete pealt koostatud suurima spetsiifilisusega mudelis on lepingu vormistamiseks oluline, et kõne võetakse vastu ning tehakse pakkumine. Lisaks mõjutavad lepingu vormistamist ka kaarditehingud nii Eestis kui ka välismaal, kuid ka mitmed erinevad lepingud, kliendilepingu pikkus ja kliendi vanus. Prognoosimudel avaldub logit-seosefunktsiooni kaudu järgmiselt:

$$\begin{aligned}\text{logit}(\pi_i) = &-4{,}59 + 2{,}03 \cdot \textit{kõne} + 0{,}18 \cdot K_1 - 0{,}62 \cdot KV_2 + 0{,}22 \cdot T_3 + 0{,}21 \cdot T_4 + \\ &+ 0{,}60 \cdot L_5 + 0{,}01 \cdot I_6 + 1{,}51 \cdot L_7 - 0{,}06 \cdot L_8.\end{aligned}$$

Tunnus 'kõne' on ka selles mudelis statistiliselt väga oluline ($p < 2e{-}16$). Kordaja hinnang $\hat{\beta}_1 = 2{,}03$ näitab, et pakkumise edastamisel suurenevad lepingu vormistamise šansid $e^{2{,}03} = 7{,}6$ korda.

4.3.3 Teisendatud andmete pealt konstrueeritud mudel

Tunnuste teisendamist võib kasutada asümmeetriliselt jaotunud tunnuste normaliseerimiseks. Kuupjuurteisendus on levinud teisendusevorm, kuna kuupjuurt saab võtta kõikidest reaalarvudest, sealhulgas nullist ja negatiivsetest arvudest. Lisaks vähendab kuupjuurteisendus tunnuse ülisuurte väärtuste mõju mudeli hindamisel.

Konstrueeriti mudelid, valides tunnuseid algandmestikust, milles oli tehinguid kirjeldavate tunnuste väärtustest võetud kuupjuur. Ka teisendatud andmeid kasutades konstrueeriti kolm mudelit sarnaselt grupeeritud ja algandmete pealt tehtud mudelitele. Mudelite võrdlus on esitatud tabelis 5.

Tabel 5. Teisendatud andmete pealt treenitud logistiliste regressioonimudelite võrdlus.

	Oluliste parameetrite arv	Klassifitseerimis-lävend	Tundlikkus	Spetsiifilisus
Sammregressioon koosmõjudeta	10	0,056	0,952	**0,326**
Sammregressioon koosmõjudega	18	0,060	0,952	**0,328**
Ekspertvalik	10	0,062	0,956	**0,339**

Ka teisendatud andmeid kasutades konstrueeritud logistilistest regressioonimudelitest on suurima spetsiifilisusega mudel, kuhu valiti tunnused manuaalse protsessi käigus.

Kaasatavad tunnused kirjeldasid peamiselt kaartide tehinguid välismaal ja erinevaid lepinguid. Sarnaselt eelnevalt koostatud mudelitele osutus ka selle mudeli juures väga oluliseks tunnuseks 'kõne' ($p < 2e{-}16$). Saadi prognoosimudel kujul

$$\begin{aligned}\text{logit}(\pi_i) = &- 654{,}94 + 2{,}07 \cdot k\tilde{o}ne + 0{,}10 \cdot KV_1 - 0{,}87 \cdot KV_2 - 0{,}67 \cdot KV_3 + 0{,}32 \cdot I_4 + \\ &+ 0{,}07 \cdot K_5 + 0{,}53 \cdot L_6 + 1{,}42 \cdot L_7 + 0{,}17 \cdot L_8 + 0{,}26 \cdot M_9.\end{aligned}$$

4.4 Mudelite konstrueerimine otsustuspuude abil

Klassifitseerimisülesannete lahendamisel kasutatakse aina rohkem otsustuspuude meetodeid. Populaarsust koguvad need meetodid just lihtsuse tõttu. Üksikut klassifitseerimispuud kasutatakse andmestikus leiduvate seoste uurimiseks, kuid prognoosimiseks pole need alati parimad valikud. Mitme puu kombineerimisel saadakse täpsemad ja stabiilsemad prognoosid.

Käesolevas töös on konstrueeritud nii üksikuid klassifitseerimispuid kui ka otsustusmetsi. Nende ehitamisel kasutatakse treeningandmestikku ning sellest taasvaliku abil moodustatud valimeid.

4.4.1 Üksik klassifitseerimispuu

Klassifitseerimispuu konstrueeritakse kõiki treeningandmeid kasutades. Saadav otsustuspuu peaks olema lihtsalt interpreteeritav ning andma andmestikust ülevaate.

Tasakaalustamata andmestiku puhul on võimalik, et kasulikku otsustuspuud ei leidu, kuna harvemini esineva klassiga objektid on lehtedes alati vähemuses. Sellisel juhul prognoosivad kõik lehed negatiivset sündmust ning sisuliselt jääb lõplikku mudelisse ainult juurtipp, millekohaselt prognoositakse kõik objektid ühte klassi. Seda probleemi ei tohi märkamata jätta, eriti kuna prognoositäpsus on tasakaalustamata andmestiku puhul kõrge ja võib tekitada vale arusaama, et mudel toimib hästi. Tihti on olulisemad just positiivsed sündmused, mis jäävad sellisel juhul tähelepanuta.

Otsemüügi jaoks konstrueeritava mudeli treeningandmestik on tasakaalustamata. Nende andmete kasutamisel leidis klassifitseerimispuu algoritm sellise eelnevalt kirjeldatud puu, kuhu kuulus ainult juurtipp. Saadud puu on kasutu, kuna see ei võimalda eristada, millise kliendiga ühendust võtta ning millisega mitte.

4.4.2 Taasvalikuga andmestiku põhjal konstrueeritud klassifitseerimispuud

Alavaliku abil üritatakse lahendada uuritava tunnuse ebavõrdsetest klassi osakaaludest tulenevat probleemi – positiivse klassi alahindamist. Moodustatakse üks SMOTE valim

ja neli alavaliku meetodil saadud valimit, milles on positiivse klassi osakaal tõstetud kas 20, 30, 40 või 50%-ni. Kõikide alamvalimite pealt üritati konstrueerida klassifitseerimispuu. Osakaalude 0,2 ja 0,3 puhul ei olnud võimalik otsustuspuud leida, taas tekkis mudel ainult juurtipuga. Sellised mudelid ei ole sobivad ning need jäetakse võrdlusest välja. Ülejäänud puude võrdlemiseks on tulemused koondatud tabelisse 6.

Tabel 6. Taasvaliku meetodite abil saadud valimitel treenitud klassifitseerimispuude võrdlus.

	Treening-objektide arv	Klassifitseerimis-lävend	Tundlikkus	Spetsiifilisus
SMOTE	4088	0,210	0,969	**0,271**
Alavalik osakaaluga 0,5	1168	0,105	0,956	**0,319**
Alavalik osakaaluga 0,4	1460	0,128	0,965	**0,317**

Alavalikuga saadud valimi põhjal treenitud mudelid on testandmestiku peal katsetades sarnaste tulemustega – spetsiifilisus on ≈0,32. Siiski ei saa selliste tulemuste põhjal otsustada, kumb mudel paremini prognoosib. Väiksema osakaaluga valimisse kaasati rohkem objekte, mistõttu võivad selle prognoosimudeli tulemused olla stabiilsemad prognooside leidmisel uutele andmetele. Lisaks on vaja selle mudeli puhul koguda klientide kohta andmed 10 tunnuse kohta, mille põhjal teeb saadud klassifitseerimispuu otsused. Tasakaalustatud klassifitseerimispuu puhul on vaja infot 12 tunnuse kohta. Seega võiks edaspidi kasutada klassifitseerimispuud, mis on treenitud valimi põhjal, kus on positiivse klassi osakaal 0,4.

4.4.3 Otsustusmetsa prognoosimudel

Mitme otsuspuu agregeerimisel võib saada täpsema ja stabiilsema prognoosimudeli. Käesolevas töös kasutatavate andmete pealt konstrueeriti otsustusmets algoritmi järgi, mis on kirjeldatud peatükis 3.2.2. Otsustusmetsa treenimisel tuleb valida mitmele parameetrile õige väärtus, et mudel sobiks kasutatava andmestikuga. Nendeks on tunnuste arv, mille vahel iga puu igas tipus parimat tükeldamisväärtust otsida, ning minimaalne lehtede suurus (Hastie et al., 2017: 592-593).

Kõige tähtsam parameeter, mida tuleb optimeerida, on kõikide otsustuspuude parima tükeldamisväärtuse otsimisel kaasatavate kirjeldavate tunnuste arv. Kui tunnused on omavahel tugevalt korreleeritud, tasuks valida väiksem kirjeldavate tunnuste arv m

igasse tippu (James et al., 2015: 319-321). Samas toovad Hastie et al. (2017: 596-597) välja, et kui suure hulga kirjeldavate tunnuste seas on väga vähe olulisi tunnuseid, ei ole väike m hea valik, kuna sel juhul tuleks otsutusreegli valimisel valida kehvade tükeldamisväärtuste seast parim. Kasutatud andmestikus on palju tunnuseid, mis on omavahel tugevalt sõltuvuses. Samas on paljud tunnused sellised, mis ei ole uuritava tunnusega väga seotud. Kirjeldavate tunnuste arvuks katsetati nii $m = 70$, $m = 40$ kui ka $m = \sqrt{p} \approx 9$.

Minimaalne lehtede suurus on klassifitseerimismudeli korral tavaliselt 1 (Hastie et al. 2017: 592). Kui puude arv on otsustusmetsas suur, võib minimaalne lehtede suurus olla üsna väike, kuna ülesobitatud üksikud puud agregeeritakse ning seega väheneb mudeli dispersioon. Vastasel juhul võib minimaalse lehesuuruse suurendamine parandada mudeli prognoosivõimet. Seega konstrueeriti mudeleid, kus lehtede arv oli kas 1 või 5 ning puude arv oli 500 või 1000.

Antud töös otsiti erinevate parameetrite suuruste head kombinatsiooni, mille põhjal konstrueeritud otsustusmets annaks häid tulemusi. Koostatud otsustusmetsade tulemused koondati tabelisse 7.

Tabel 7. Otsustusmetsa meetodil koostatud prognoosimudelite võrdlus.

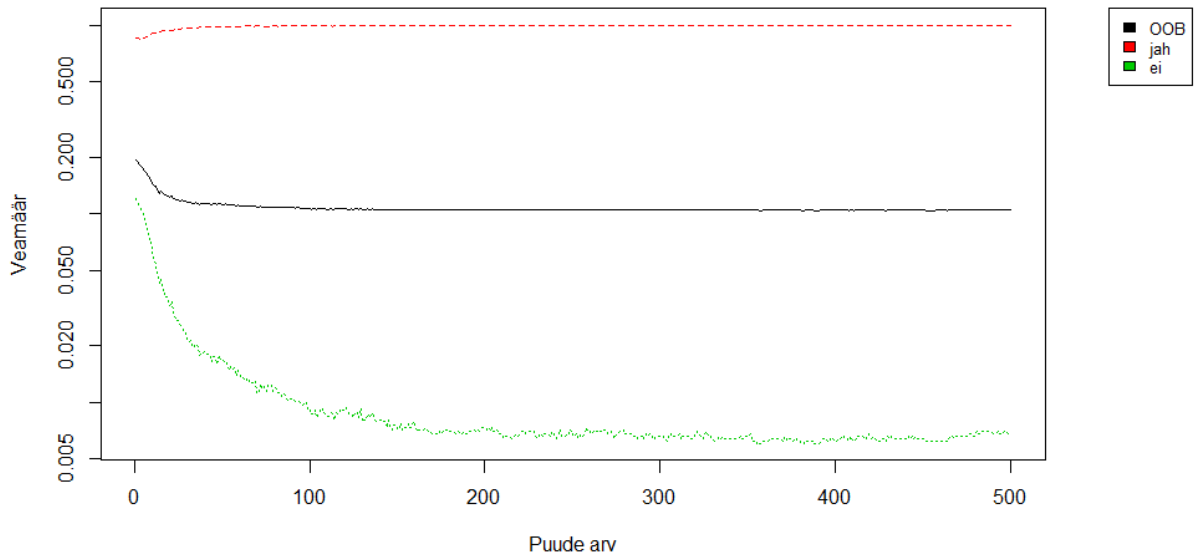
Puude arv	Kirjeldavate tunnuste arv	Minimaalne lehtede suurus	Tundlikkus	Spetsiifilisus
500	9	1	0,952	**0,136**
500	9	5	0,956	**0,147**
1000	9	1	0,952	**0,132**
1000	9	5	0,956	**0,134**
500	40	1	0,956	**0,242**
500	40	5	0,952	**0,233**
1000	40	1	0,952	**0,261**
1000	40	5	0,961	**0,230**
500	70	1	0,952	**0,272**
500	70	5	0,956	**0,248**
1000	70	1	0,956	**0,257**
1000	70	5	0,952	**0,266**

Otsustusmetsade, mille tippudes kaasati ainult 9 tunnust, tulemused on märgatavalt madalamad. Suurima spetsiifilisusega otsustusmets koosneb 500 puust, kus minimaalne objektide arv lehes on 1 ning tippudes valiti 70 tunnuse seast parim tükeldamisväärtus. Siiski pole tulemustest parameetrite muutmisel näha ühest trendi, millal on spetsiifilisus

parem. Väikesed erinevused võivad olla tekkinud näiteks testandmestiku eripärast või juhuslikkusest kirjeldavate tunnuste valimisel.

Otsustusmetsa meetodil tehtud mudeli kasutamistel tuleb olla kindel, kas valitud puude arvuga mudeli veamäär on stabiliseerunud (James et al., 2015: 319-321). Hindamiseks kasutatakse OOB-valimite (*out-of-bag*) põhjal leitud prognoosivigu. Selle käigus leiti igale treeningandmestiku objektile prognoos, kuid agregeeriti vaid selliseid puid, mille treenimisel jäi vastav objekt bootstrap-valimist välja.

Kõikide eespool välja toodud otsustusmetsade OOB-vead on enam-vähem stabiliseerunud. Joonisel 2 kujutatakse parima spetsiifilisusega mudeli stabiliseerumist. Lisaks on jooniselt näha peatükis 3.2.3 nimetatud otsustusmetsade puudust - üldise klassifitseerimisvea minimeerimisel suureneb väiksema klassi ('jah') prognoosiviga, kui klassifitseerimislävendiks võetakse 0,5. Eelnevalt nimetatud puudus on peamine põhjus, miks tasakaalustamata andmestiku uurimisel tuleks tavalistele otsustusmetsadele otsida alternatiive.



Joonis 2. Otsustusmetsa prognoosivigade stabiliseerumine puude arvu suurendamisel.

4.4.4 Taasvalikuga otsustusmetsa prognoosimudelid

Tasakaalustamata andmestike põhjal otsustusmetsa konstrueerimisel võib kasutada otsustusmetsa algoritmisiseselt erinevate puude tegemiseks mitmeid taasvaliku meetodeid. Tasakaalustatud otsustusmetsa on kirjeldatud peatükis 3.2.3. Lisaks sellele otsustati proovida analoogselt valimi tasakaalustamisele ka teisi taasvalikumeetodeid.

Otsustusmets konstrueeriti kasutades tunnuste arvu m, mille korral oli OOB-valimitelt leitud kapa kordaja suurim. Kapa kordaja arvutatakse valemiga

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)},$$

kus $P(A) = Acc = \frac{TP+TN}{TP+TN+FP+FN}$ ja

$$P(E) = \frac{(TP + FP)(TP + FN)}{TP + TN + FP + FN} + \frac{(FN + TN)(FP + TN)}{TP + TN + FP + FN}$$

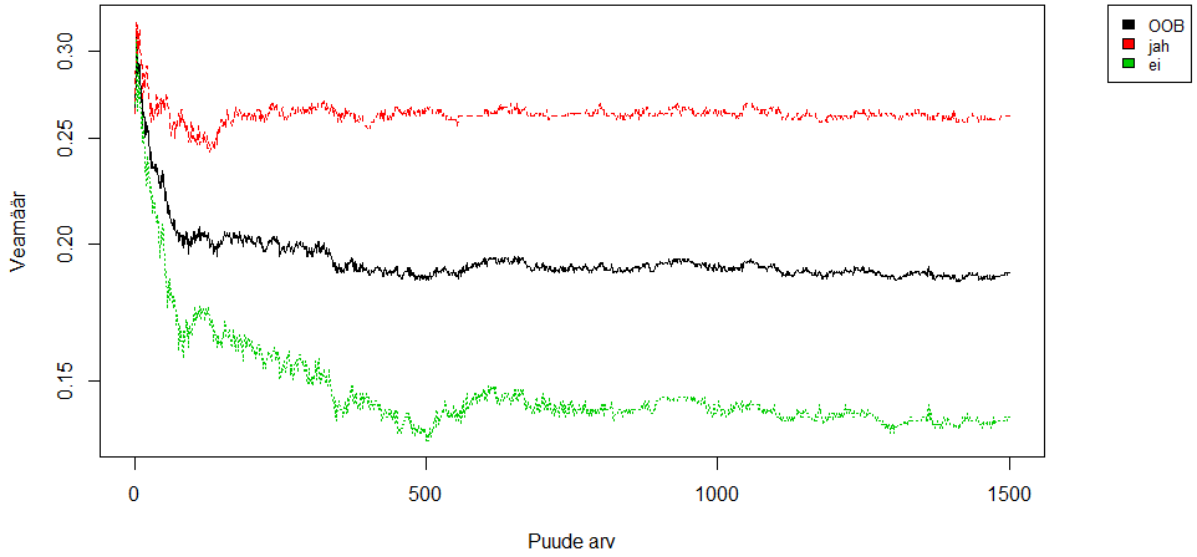
ning TP, TN, FP, FN on punktis 1.3 kirjeldatud otsuste tüübid (Warrens, 2013).

Taasvalikuga otsustusmetsa kaasati 1500 puud, kuna väiksema valimiga otsustusmetsad ei stabiliseerunud nii kiiresti nagu tavaliste otsustusmetsade korral. Iga puu korral tehti maksimaalne puu, st minimaalne lehtede suurus oli 1. Tulemusi saab võrrelda tabeli 8 põhjal.

Tabel 8. Taasvalikuga otsustusmetsade võrdlus.

Taasvaliku meetod	Puude arv	Kirjeldavate tunnuste arv	Tundlikkus	Spetsiifilisus
SMOTE	1500	53	0,952	**0,249**
Alavalik osakaaluga 0,5	1500	34	0,952	**0,253**
Alavalik osakaaluga 0,4	1500	23	0,952	**0,275**
Alavalik osakaaluga 0,3	1500	70	0,952	**0,248**
Alavalik osakaaluga 0,2	1500	59	0,952	**0,250**

Kõige rohkem paistab nende seas silma alavaliku meetod, mille korral oli valimi positiivse klassi osakaal 40%. Mudel saavutas nõutud tundlikkuse tasemel $\approx 27\%$ spetsiifilisuse. Jooniselt 3 on näha, et prognoosivead on saavutanud stabiilse taseme, kuid jäävad siiski selle taseme ümber võnkuma. Lisaks annab joonis kinnitust, et üldise vea minimeerimisel väheneb ka väiksema klassi ('jah') prognoosiviga. Seetõttu on konstrueeritud otsustusmets sobivam antud probleemi lahendamiseks kui punktis 4.4.3 leitud suurima spetsiifilisusega otsustusmets.



Joonis 3. 40% alavalikuga otsustusmetsa prognoosivigade stabiliseerumine puude arvu suurendamisel.

4.4.5 Agregeerimata andmete pealt puumeetodil koostatud prognoosimudelid

Otsustuspuid ja -metsi peetakse robustseteks meetoditeks, mida võib kasutada ka keeruliste andmete korral. Nende kasutamise eeliseks on ka see, et meetod suudab ise leida üles erinevad olulised kombinatsioonid tunnustest. Eelnevalt käsitsi agregeeritud tehingu andmeid kirjeldavate tunnustega ei saanud puumeetodid kõige paremini hakkama ning seetõttu konstrueeriti mudelid ka kasutades andmeid, kus ei ole tehingu andmed agregeeritud. Selles andmestikus on vaid tehingute arvud ja kogusummad 30 päeva lõikes, mis kombineeritakse meetodi enda poolt.

Agregeeritud tunnuste korral andsid parimaid tulemusi tasakaalustatud otsustuspuu ning 40% alavalikuga klassifitseerimispuu ja otsustusmets. Seetõttu katsetati agregeerimata tunnuseid kasutades just neid meetodeid. Tulemused on koondatud tabelisse 9.

Tabel 9. Agregeerimata andmete põhjal treenitud mudelite võrdlus.

	Puude arv	Kirjeldavate tunnuste arv	Tundlikkus	Spetsiifilisus
Tasakaalustatud klassifitseerimispuu	1	92	0,965	**0,324**
Osakaaluga 0,4 klassifitseerimispuu	1	92	0,945	**0,328**
Osakaaluga 0,4 otsustusmets	1500	8	0,952	**0,270**

Positiivse klassi 40% osakaaluga valimi põhjal treenitud klassifitseerimispuu, ei suutnud valitud tundlikkuse piires saavutada head spetsiifilisust. Kui pidada sobivaks ka tundlikkuse taset 0,945, siis saavutas see mudel spetsiilisuse ≈0,33. Käesolevas töös saavutas puumeetodil koostatud prognoosimudelite seas suurima spetsiifilisuse, mis vastab ka nõutud 95% tundlikkuse tasemele, tabelis 9 esitatud tasakaalustatud klassifitseerimispuu. Mõlema mudeli kõige tähtsamaks tunnuseks on 'kõne'. Kui pakkumist ei ole suudetud edastada, prognoosivad mõlemad mudelid lepingu vormistamise tõenäosuseks 0. Esimese mudeli kasutamiseks on vaja koguda infot 12 tunnuse kohta, mis hõlmavad infot üldiste kaardi- ja arveldusmaksete kohta. Teise kasutamiseks tuleb koguda infot 10 tunnuse kohta, mis lisaks üldistele kaardi- ja arveldusmaksetele kirjeldavad ka lepinguid ja teenuste kasutamist.

Konstrueeritud otsustusmetsa spetsiifilisus on ligikaudu samal tasemel kui agregeerimata tunnuste puhul leitud otustusmets, mis treeniti positiivse klassi 40% osakaaluga andmestiku põhjal. Spetsiifilisus 0,27 on märgatavalt väiksem kui klassifitseerimispuude puhul leitud stabiilsus ning see on ka põhjus, miks antud töös eelistatakse siiski leitud klassifitseerimispuid.

4.5 Mudelite tulemuste võrdlemine

Logistilise regressioonimudelite ning otsustuspuude ja -metsade treenimisel kasutati sama andmestikku. Otsustuspuude korral katsetati uuringu käigus tulnud ideed kasutada ka andmestikku, mille tunnused olid võetud otse andmebaasidest suurema eeltöötluseta.

Kõikide koostatud mudelite seast paistsid kõige enam silma logistilise regressiooni meetodil koostatud mudelid. Fikseeritud 95% tundlikkuse juures suutsid need mudelid saavutada stabiilselt üle 32% spetsiifilisust ning see on märkimisväärselt parem enamikust puumeetodil saadud mudelitest. Puudel põhinevate prognoosimudelite parimad tule-

mused saadi kasutades agregeerimata andmestikku. Taasvalikuga klassifitseerimispuud küündisid ainsate puumeetodil tehtud mudelitena logistilise regressiooni tulemuste tasemele, saavutades üle 32% spetsiifilisust.

Spetsiifilisusega, mis on üle 0,3, saaks vähendada üle 30% võrra kõnede arvu sellistele klientidele, kes ei ole pakkumisest huvitunud. Käesoleva töö eesmärgiks oli leida mudel, mille kasutamisel suureneb tehtud resultatiivsete kõnede suhtarv, vähendades müügikonsultantide tühja tööd. Töö tulemusena saadi mitu prognoosimudelit, mis aitavad vähendada mitteresultatiivsete kõnede arvu ligikaudu kolmandiku võrra. Seejuures võib olla kindel, et peaaegu mitte ühtegi huvitunud klienti ei jäeta valimist välja.

Kõige suurema spetsiifilisuse (34%) saavutas testandmestikul logistiliste regressioonimudelite seas grupeeritud kirjeldavate tunnustega mudel, kuhu valiti käsitsi, lähtudes intuitsoonist, kõik oluliseks osutunud 9 tunnust.

Puumeetodil leitud mudelitest saavutasid testandmestikul suurima spetsiifilisuse (üle 32%) üksikud klassifitseerimispuud. Nende treenimisel kasutati 40% ja 50% alavaliku meetodil saadud valimit andmestikust, kus tehingute info oli peaaegu töötlemata, oli vaid summeritud 30 päeva kaupa 180 päevase ajavahemiku jooksul. On üllatav, et just üksikud klassifitseerimispuud saavutasid parimad tulemused kõikide puumeetodil saadud mudelite seas. Mitme puu agregeerimine peaks parandama mudeli prognoosivõimet ja vähendama ebastabiilsust. Käesoleva töö parimaid puumudeleid kasutades tuleb arvestada sellega, et üksikute klassifitseerimispuude prognoosid ei pruugi olla stabiilselt sama heade tulemustega.

Mitme mudeli ligikaudu samaväärsete tulemuste korral tuleb parim välja valida ka teisi aspekte arvestades. Konstrueeritud logistilist regressioonimudelit kasutades on prognoosimiseks vaja koguda klientide kohta infot lisaks tunnusele 'kõne' vaid 8 tunnuse kohta, klassifitseerimispuude korral aga 10 või 12 tunnuse kohta. Samas on klassifitseerimispuude kasutamiseks vaja välja võtta iga kliendi tehingute info ainult 30 päeva kaupa summeerituna. Logistilise regressioonimudeli kasutamiseks on nendest andmetest vaja teha lisaagregeerimisi.

Andmestiku suurenedes võtab puumeetodil uue prognoosimudeli koostamine tänu automatiseeritud algoritmidele vähem aega. Logistilise regressioonimudeli koostamine samas võib võtta kordades rohkem aega, kui tunnuseid valitakse käsitsi. Prognoosimudeleid uuesti hinnates võib treeningandmestiku suurenemise tõttu saada täpsema mudeli. Mudeleid saab uuesti hinnata 90 päeva möödumisel peale iga pakkumise tegemist ning otsustuspuude puhul saab see olla ka regulaarne tegevus töökoormat märkimisväärselt suurendamata.

Kuna mõlemal mudelil on plusse ja miinuseid, tasuks mudelite töökindlust võrrelda ka

järgmise pakkumiskampaania tulemusi uurides ning seejärel otsustada, milline mudel läheb rakendusse.

4.6 Alternatiivsed võimalused tulevasteks uuringuteks

Käesolev töö ei ole antud probleemi kõikehõlmav käsitlus. Tulevaste analüüside käigus võib teha muudatusi lahenduskäigus nii andmestiku kui ka algoritmide tasemel.

Andmestikku tuleks kaasata rohkem infot isikuandmete kohta. Samuti võiks lisada rohkem infot kaardimaksete kohta erinevates valdkondades. Selliste andmete analüüs avab rohkem tehingute mahtusid kirjeldavate tunnuste sisu ning seob need infoga, millisel põhjusel on vastavad kulutused tehtud.

Mitteparameetrilistest meetoditest toovad Kotsiantis et al. (2006) välja erinevaid masinõppe meetodeid, mida tasuks kasutada klassifitseerimismudelite koostamisel tasakaalustamata andmestike põhjal.

1. Kasutades lähima naabri meetodit, tuleks klassifitseerimisprotsessi käigus kasutada kaalutud kauguse funktsiooni. See funktsioon annab positiivse ja negatiivsete klasside andmepunktide kaugustele erinevad kaalud ning see aitab vähendada tasakaalustamatusest tulenevaid probleeme andmestiku jaotust muutmata.
2. Tugivektorsüsteemide korral tuleks kasutada väiksema osakaaluga klassi suunas nihutatud hüpertasandeid.
3. Meetodi Ripper põhjal otsib algoritm andmestikust selliseid eristavaid reegleid, mis vastavad ainult huvipakkuva klassi objektidele.
4. Maksumusepõhisel õppimisel määratakse huvipakkuva klassi valesti klassifitseerimisel suurem maksumus ning seetõttu teeb uuritava klassi prognoosivead madalamaks.

Parameetrilistest meetoditest tasuks tasakaalustamata andmeid kasutades klassifitseerimisülesannete lahendamisel uurida ka muid seosefunktsioone üldistatud lineaarsete mudelite perest. Van der Paal (2014) on välja toonud mitmeid asümmeetrilisi seosefunktsioone, sealhulgas binaarsed kvantiil-, asümmeetrilised t- ja üldistatud ekstremaaljaotuse seosefunktsioonid.

5 Kokkuvõte

Ettevõtetel on kasumi teenimiseks vähemalt kaks näiliselt lihtsat võimalust: suurendada olemasolevate klientide tarbimist ning vältida nendega kliendisuhete lõppemist. Otsemüügi pakkumised, mis tehakse vastavalt kliendiprofiilile, peaksid suurendama tarbimist, luues samal ajal kliendiga personaalse kontakti. Prognoosides kliendi huvitatust ja vajadust toote vastu, saab pakkuda kliendile toodet just sellisel ajal, millal tal seda vaja võiks minna.

Antud töö eesmärk oli luua mudel, mis leiaks kirjeldavate tunnuste põhjal prognoosid, kui tõenäoliselt pakutav toode kliendi poolt aktsepteeritakse. Kasutatud andmestiku kuulus 8412 telefoni teel tehtud pakkumist, mis suunas kliente krediitkaardi lepingut vormistama. Andmestik sisaldas 80 tunnust, mis kirjeldasid klientide poolt tehtud arveldus- ja kaardimakseid ning pakutavate teenuste kasutamist. Kasutatav andmestik oli uuritava tunnuse suhtes tasakaalustamata ning seega tuli mudelite konstrureerimisel tähelepanu pöörata ka sellest tulenevatele kitsaskohtadele.

Mudeleid tehti erinevatel meetoditel. Parameetrilistest meetoditest kasutati logistilist regressiooni. Nimetatud meetodi põhjal hinnati mudel, kasutades tunnuseid muutmata kujul, kuid lisaks ka grupeeritud ja teisendatud kujul. Mitteparameetrilistest meetoditest kasutati otsustuspuudel põhinevaid meetodeid. Konstrueeriti üksikuid klassifitseerimispuid ning otsustusmetsi, mille käigus leiti prognoos paljude puude prognooside agregeerimisel. Nii üksikute puude kui ka metsade korral kasutati erinevaid taasvaliku meetodeid, et parandada tasakaalustamata andmestikust tekkivaid probleeme.

Erinevaid mudeleid võrreldes osutusid headeks kõik logistilise regressiooni mudelid ning kaks klassifitseerimispuude mudelit. Kõikide nende mudelite kasutamise korral oli testandmestiku põhjal võimalik vähendada üle 30% kõnede arvust sellistele klientidele, kes ei soovi tootelepingut vormistada. Samuti kindlustasid need mudelid, et pakkumine oleks tehtud vähemalt 95% klientidest, kes kindlasti toodet soovisid.

Selline tulemus on kasulik ettevõttele mitmel moel. Eelkõige vähendavad koostatud prognoosimudelid müügikonsultantide tühja tööd, vähendades mittehuvitunud klientidele tehtavaid kõnesid. Lisaks, resultatiivsete kõnede osakaalu suurenedes kasvab selle tulemusena ka tarbimine ning ettevõte teenib suuremat tulu.

6 Kasutatud kirjandus

Berry, M. J. A., Linoff, G. S. (2004). *Data Mining Techniques: for Marketing, Sales, and Customer Relationship Management* (2nd Edition). Indianapolis: Wiley Publishing, Inc.

Breiman, L. (1996). Bagging Predictors. *Machine Learning* 24(2): 123-140. http://dx.doi.org/10.1023/A:1018054314350

Breiman, L. (2001). Random Forests. *Machine Learning* 45(1): 5-32. http://dx.doi.org/10.1023/A:1010933404324

Cadez, I. V., Smyth, P., Ip, E., Mannila, H. (2003). *Predictive Profiles for Transaction Data using Finite Mixture Models.*. Kasutatud 24.04.2017. http://www.datalab.uci.edu/papers/profiles.pdf

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16: 321-357. http://dx.doi.org/10.1613/jair.953

Chen, C., Liaw, A., Breiman, L. (2004). *Using Random Forest to Learn Imbalanced Data*. Kasutatud 22.04.2017. http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

Dietterich, T. G. (2000). An experimental Comparison of Three Methods for Construction Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning* 40(2): 139-157. http://dx.doi.org/10.1023/A:1007607513941

Hastie, T., Tibshirani, R., Friedman, R. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd Edition, 12th printing). New York: Springer.

Hosmer, D. W., Lemeshow, S. (2000). *Applied Logistic Regression* (2nd Edition). New York: Wiley. http://dx.doi.org/10.1002/0471722146

James, G., Witten, D., Hastie, T., Tibshirani, R. (2015). *An Introduction to Statistical Learning: with Applications in R* (6th printing). New York: Springer.

King, G., Zeng, L. (2001). Logistic Regression in Rare Events Data. *Political Analysis* 9: 137–163.

Kotsiantis, S., Kanellopoulos, D., Pintelas, P. (2006). Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering* 30(1): 25-36.

Kubat, M., Holte, R., Matwin, S. (1997). Learning when negative examples abound. Someran, M. van,Widmer, G. (toim.), *Machine Learning: ECML-97*, (lk 146–153). http://dx.doi.org/10.1007/3-540-62858-4_79

Kubat, M., Matwin, S. (1997). Addressing the Curse of Imbalances Training Sets: One Sided Selection. Fisher, D. H. (toim.), *Proceedings of the Fourteenth International Conference on Machine Learning*, (lk 179-186). Kasutatud 24.04.2017. http://sci2s.ugr.es/keel/pdf/algorithm/congreso/kubat97addressing.pdf

Liu, A. Y. (2004). *The Effect of Oversampling and Undersampling on Classifying Imbalanced Text Datasets* (magistritöö). Kasutatud 20.04.2017. https://pdfs.semanticscholar.org/cade/435c88610820f073a0fb61b73dff8f006760.pdf

Peterson, R. A., Wotruba, T. R. (1996). What Is Direct Selling? – Definition, Perspectives, and Research Agenda. *The Journal of Personal Selling & Sales Management* 16(4): 1-16. Kasutatud 15.04.2017. https://pdfs.semanticscholar.org/6531/f74757e7461a4518e632b5f671c13d1b75d0.pdf

Ratner, B. (2010). Variable selection methods in regression: Ignorable problem, outing notable solution. *Journal of Targeting, Measurement and Analysis for Marketing* 18(1): 65-75. http://dx.doi.org/10.1057/jt.2009.26

Sanz Saiz, B., Pilorge, P. (2010). *Understanding customer behavior in retail banking.* Kasutatud 15.04.2017.
http://www.ey.com/Publication/vwLUAssets/Understanding_customer_behavior_in_retail_banking_-_February_2010/$FILE/EY_Understanding_customer_behavior_in_retail_banking_-_February_2010.pdf

Sokolova, M., Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management* 45: 427–437. http://dx.doi.org/10.1016/j.ipm.2009.03.002

Tan, P., Steinbach, M., Kumar, V. (2006). *Introduction to Data Mining.* Boston: Pearson/Addison Wesley.

Van der Paal, B. (2014). *A comparison of different methods for modelling rare events data* (magistritöö). Kasutatud 24.04.2017.
http://lib.ugent.be/fulltxt/RUG01/002/163/708/RUG01-002163708_2014_0001_AC.pdf

Warrens, M. J. (2013). A comparison of Cohen's Kappa and agreement coefficients by Corrado Gini. *IJRRAS* 16(3): 345-351. Kasutatud 03.05.2017.
http://www.arpapress.com/Volumes/Vol16Issue3/IJRRAS_16_3_03.pdf

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Hele-Liis Peedosk**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose ,,Logistilise regressiooni ja otsustuspuumeetodite kasutamine otsemüügi efektiivsuse suurendamiseks", mille juhendaja on prof. Kalev Pärna,

   1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

   1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 09.05.2017