TARTU ÜLIKOOL
Arvutiteaduse instituut
Infotehnoloogia mitteinformaatikutele õppekava

Marelle Ellen

Soovitusprogrammi optimeerimine masinõppe abil idufirma Promoty näitel

Magistritöö (15 EAP)

Juhendaja: Anna Leontjeva, PhD

Tartu 2018

Referral Program Optimization Using Machine Learning Based on Promoty's Case

Abstract:

Referral programs are one of the best options for startups to achieve fast growth. Despite that, many Estonian startups have not managed to run successful referral programs. For that reason, a methodology was proposed for Estonian startups for applying supervised and unsupervised machine learning to analyze and optimize their referral program. The methodology was implemented on a startup called Promoty. For every Instagram user not connected to Promoty, the model estimates the probability of becoming so-called „useful user" who invites at least one new member to Promoty's platform. With 60% precision, the model is able to identify 23,6% of all the useful users. Therefore, additional features that should be added to improve the model's prediction power were proposed. Also, cluster analysis was performed to find out the descriptive features of Promoty's useful users and based on that, recommendations for improving the referral program were made. The results are significant for Promoty to achieve fast growth when entering new markets as well as for other startups that are implementing or planning to implement referral programs.

Keywords:

marketing, marketing management, marketing strategy, referral program, startups, lean thinking, data science, data mining, machine learning

CERCS: P160 Statistics, operation research, programming, actuarial mathematics

Soovitusprogrammi optimeerimine masinõppe abil idufirma Promoty näitel

Lühikokkuvõte:

Idufirmade jaoks üks parimaid võimalusi kiire kasvu saavutamiseks on soovitusprogrammid. Vaatamata sellele ei ole paljud Eesti idufirmad suutnud neid edukalt käivitada. Seetõttu pakuti antud magistritöös Eesti idufirmadele välja metoodika juhendatud ja juhendamata masinõppe rakendamiseks soovitusprogrammi analüüsiks ja optimeerimiseks. Metoodika näitlikustamiseks loodi idufirma Promoty jaoks mudel, mis hindab iga Promotyga mitteseotud Instagrami kasutaja tõenäosust osutuda Promoty jaoks nn. kasulikuks kasutajaks, kes kutsub platvormiga liituma vähemalt ühe uue kasutaja. Töö käigus loodud mudel suudab 60% esitustäpsuse juures tuvastada 23,6% kõikidest kasulikest kasutajatest. Mudeli ennustusvõime parendamiseks pakuti välja tunnused, mida Promoty peaks täiendavalt koguma. Samuti viidi läbi klasteranalüüs, et selgitada välja Promoty jaoks kasulikku kasutajat iseloomustavad tunnused, ning teha sellest tulenevalt ettepanekuid soovitusprog-

rammi efektiivsemaks rakendamiseks. Antud töö tulemused on olulised idufirmale Promoty kiire kasvu saavutamiseks uutele turgudele sisenemisel, samuti teistele soovitusprogramme rakendavatele või rakendada plaanivatele idufirmadele.

Võtmesõnad:
turundus, turunduse juhtimine, turundusstrateegiad, soovitusprogramm, idufirmad, kulusäästlik mõtlemine, andmeteadus, andmekaeve, masinõpe

CERCS: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Sisukord

Sissejuhatus

Eesti idufirmad on näidanud järjepidevat kasvu. Kaasatud investeeringute summa on viimase kuue aasta jooksul iga aasta kasvanud, jõudes 2017. aastal 270 miljoni euroni [14]. Lisaks mängivad idufirmad üha olulisemat rolli riigi majanduses: 2017. aastal andsid idufirmad Eestis tööd 2981 inimesele ja tasusid riigile makse 37 miljoni euro väärtuses [32]. Üks peamisi tegureid plahvatusliku kasvu taga võib olla meie riigi väiksus: esiteks muudab see lihtsamaks uute toodete lansseerimise ning valideerimise kohalikul turul, teiseks on kohalikud idufirmad juba algusest peale sunnitud rahvusvahelisel tasandil mõtlema [22]. Seetõttu on loogiline, et meie idufirmad on tugevalt kasvule orienteeritud: turunduse peamine eesmärk on uute klientide leidmine võimalikult kiiresti ja võimalikult madala kuluga [23].

Kiire ja jätkusuutliku kasvu tagamiseks kasutatakse tihti tootesse ehitatud nn kasvumootoreid nagu soovitusprogrammid või mehaanismid, mis ajendavad toodet järjepidevalt kasutama. Kasvumootori eesmärgiks on tekitada jätkusuutlik kasv, mille puhul uued kliendid tulevad eelmiste klientide tegevuse tulemusena. Paraku on mitmed uurimused [23, 31] näidanud, et paljud Eesti idufirmad ei ole suutnud soovitusprogramme ja teisi kasvumootoreid edukalt rakendada – võib-olla just seetõttu, et nendele on traditsioonilises turunduses vähe tähelepanu pööratud. Kuivõrd masinõppe abil on võimalik leida andmetest seoseid, mustreid ning ennustada sündmuste toimumise tõenäosust tulevikus, on masinõpe üks võimalusi idufirma turunduse optimeerimiseks ning seeläbi kiire kasvu tagamiseks.

Seetõttu pakutakse antud magistritöös Eesti kasvufaasi idufirmadele ettevõtetele välja metoodika, kuidas kasutada masinõppe meetodeid soovitusprogrammi kui ühe võimaliku kasvumootori analüüsiks ja optimeerimiseks. Idufirma Promoty näitel luuakse masinõppe mudel, et tuvastada soovitusprogrammi nn kasulikke kasutajaid, kes on valmis Promotyga liituma kutsuma vähemalt ühe uue kasutaja, ning aitavad seeläbi kaasa jätkusuutliku kasvu saavutamisele. Metoodika on suunatud eelkõige Eesti kasvufaasi idufirmadele, kellel on kõige suurem vajadus kiire kasvu järele, kuid sobib ka teistele sarnast soovitussüsteemi rakendavale ettevõtetele.

Magistritöö raames seati neli eesmärki:

1. Pakkuda Eesti idufirmadele välja metoodika kirjeldava analüüsi, juhendatud ja juhendamata masinõppe rakendamiseks soovitusprogrammi analüüsiks ja optimeerimiseks.

2. Luua masinõppe mudel, mis hindab iga Instagrami kasutaja puhul võimalikult täpselt tõenäosust, kas tegemist on Promoty jaoks nn kasuliku kasutajaga, kes on valmis läbi personaalse soovituslingi kutsuma Promotyga liituma vähemalt ühe uue kasutaja.

3. Selgitada välja, millised tunnused iseloomustavad Promoty jaoks kasulik-

ku kasutajat, ning sellest tulenevalt teha ettepanekuid soovitusprogrammi efektiivsemaks rakendamiseks.

4. Pakkuda välja tunnused, mida oleks mudeli täpsuse parendamiseks vajalik täiendavalt koguda.

Töö on oluline kolmel põhjusel. Esiteks, ehkki teoreetilise kirjanduse põhjal on soovitusprogrammid üks parimaid võimalusi idufirma kliendibaasi kiire kasvu saavutamiseks, on mitmed uurimused [23, 31] on näidanud, et paljud Eesti idufirmad ei ole osanud neid edukalt rakendada. Teiseks, loodud mudel on oluline idufirma Promoty jaoks turundustegevuste ja -eelarve optimeerimisel uutele turgudele sisenemisel. Välja töötatud mudeli rakendamine võimaldab turunduskommunikatsioonis esimesena sihtida kasutajaid, kellel on kõrgem tõenäosus osutuda kasulikuks kasutajaks, ning seeläbi tagada uuel turul võimalikult kiire kasv ning turunduseelarve optimaalne kasutamine. Kolmandaks, vaatamata sellele, et masinõppe rakendamisel turundusvaldkonnas on palju võimalusi, ei ole see töö autori hinnangul Eestis veel tavapärane praktika ning sellealase teadmuse tekitamine ja edasikanne on teretulnud.

Magistritöö koosneb kolmest peatükist. Peatükis 1 antakse ülevaade teoreetilistest kontseptsioonidest, mis on olulised antud töö konteksti mõistmiseks. Erialase kirjanduse ja varasemate teadustööde põhjal kirjeldatakse idufirmade turundusraamistikke, tuuakse välja erinevused traditsioonilise turundusega ning analüüsitakse idufirma võimalikke kasvumootoreid. Samuti antakse ülevaade varasematest uuringutest masinõppe rakendamisest turundusvaldkonnas nii akadeemiliste kui praktiliste allikate põhjal.

Peatükis 2 kirjeldatakse magistritöö praktilise osa tööprotsessi ning ning metoodikat masinõppe rakendamiseks soovitussüsteemi optimeerimisel. Samuti kirjeldatakse ja põhjendatakse kasutatud meetodeid ning antakse ülevaade Promoty juhtumi puhul kasutatud andmestikest.

Peatükis 3 tuuakse välja eelkirjeldatud meetodite abil läbi viidud analüüsi tulemusi. Esmalt kirjeldatakse soovitusprogrammi efektiivsust kasvumootorina ning võrreldakse seda postituste tegemisega, mis töö autori hinnangul võib samuti kasvumootoriks osutuda. Seejärel luuakse masinõppe meetodite abil mudel, mis hindab Instagrami kasutajate tõenäosust osutuda Promoty jaoks kasulikuks kasutajaks. Viimases alapeatükis rakendatakse klasteranalüüsi, et saada parem ülevaade erinevatest kasutajagruppidest ning selle põhjal turundussõnumeid paremini sihtida.

Magistritöö autor soovib tänada oma juhendajat dr. Anna Leontjevat edasiviivate mõttevahetuste ning arvukate paranduste ja täienduste eest. Samuti Promoty asutajaid Aleks Koha ja Leonardo Romanellot, kes olid abiks andmete kogumisel.

1 Teoreetilised lähtekohad ja varasemad uurimused

Magistritöö esimeses peatükis antakse ülevaade teoreetilistest kontseptsioonidest, mis on olulised antud töö konteksti mõistmiseks. Erialase kirjanduse ja varasemate teadustööde põhjal kirjeldatakse idufirmade turundusraamistikke ning tuuakse välja erinevused traditsioonilise turundusega, samuti kirjeldatakse idufirma kasvumootoreid ning masinõppe rakendamise võimalusi turundusvaldkonnas.

1.1 Ülevaade idufirmade turundusest

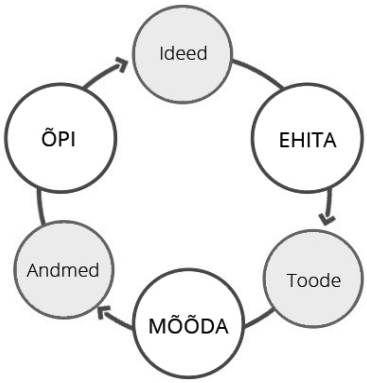
1.1.1 Nutika idufirma printsiipide rakendamine

Üks enim tunnustatud lähenemisi idufirmade loomisele on nutika (ingl *lean*) idufirma meetod, mis on ka Eesti idufirmade seas laialt kasutatud [28]. Nutika idufirma meetod [11] tugineb ehita-mõõda-õpi tsüklile (joonis 1a), mille käigus ehitatakse ideede valideerimiseks minimaalne elujõuline toode ning testitakse seda potentsiaalse kliendi peal. Saadud tagasiside pealt tehakse järeldused, mis on aluseks uute ideede genereerimisele ning toodete ehitamisele. Püstitades hüpoteese ning neid eksperimentide abil kontrollides on idufirma eesmärgiks õppida, kuidas luua jätkusuutlik ettevõte.

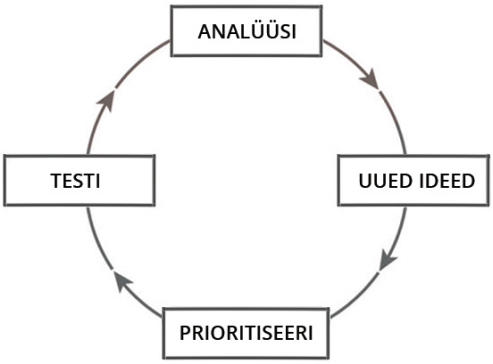
Teise olulisima idufirmade turundusmetoodikana on viimastel aastatel esile kerkinud häkkerturundus (ingl *growth hacking*). Holiday [30] hinnangul on häkkerturunduse populaarsuse taga uue generatsiooni suurettevõtted nagu Facebook, Dropbox ja Airbnb, mille ehitamisel on kombineeritud häkkerturundust traditsiooniliste turunduskanalitega. Kuigi häkkerturunduse meetodid ei ole veel laialt kasutatavad, on seda peetud ka suunaks, kuhu turundus tervikuna tulevikus areneb [5].

Häkkerturundus on äristrateegia, mis rakendab uute klientide leidmiseks tehnikaid, mida on võimalik testida, jälgida ning skaleerida [30]. Häkkerturunduse protsess on pidev tsükkel, mis sarnaneb nutika idufirma tsüklile: andmete analüüsile järgneb ideede genereerimine ja prioritiseerimine, seejärel koostatakse ning viiakse läbi eksperimendid ideede valideerimiseks. Edasi liigub tsükkel analüüsi osasse, et tõlgendada eksperimendi tulemusi ning määrata järgmine samm (joonis 1b). Süstemaatilise lähenemise eesmärk on kiirelt mõista, millised ideed on väärtuslikud ning millistest loobuda, et seeläbi ettevõtte turundust märgatavalt efektiivsemaks muuta. [12]

Traditsioonilisest turundusest erineb häkkerturundus olulisel määral. Esiteks vaatab häkkerturundus ettevõtte äristrateegiat kui tervikut: kui traditsioonilise turunduse eesmärk on tekitada teadlikkust ning juhtida tarbija ostuotsuseni, siis häkkerturundus vaatab tervet tarbija teekonda teadlikkuse tekkimisest kuni kordu-



(a) Nutika idufirma tsükkel [11]



(b) Häkkerturunduse protsess [12]

Joonis 1. Nutika idufirma tsükli ja häkkerturunduse protsesside võrdlus

vostuni, tulu tekkimise ja toote või teenuse soovitamiseni uutele kasutajatele [8]. Teiseks on häkkerturundus tugevalt orienteeritud analüütikale ja tulemustele: kui traditsiooniline turundus on pigem brändikeskne, siis häkkerturundus on mõõdikute ja investeeringute tasuvuse keskne [30].

Sidudes omavahel erinevad äri osad, nõuab häkkerturundus terviklikku arusaama äri toimimisest ning sellest, kuidas muutus kasutaja teekonna ühes etapis mõjutab teekonna teisi etappe [8]. Näiteks on eksperimentide või andmeanalüüsi läbi võimalik leida mõõdikud, mis aitavad määrata kliendi nn kasulikkuse juba tarbija teekonna varajases faasis. Mõistes, kuidas on nimetatud mõõdik seotud ettevõtte pikaajaliste eesmärkidega, saab ennustada ettevõtte eesmärkide edenemist ning astuda teadlikke samme kasutajakogemuse parendamiseks, et edendada ettevõtte pikaajaliste eesmärkide täitmist.

Eesti kasvufaasi idufirmade seas läbi viidud uurimus [23] näitas, et idufirmade turundustegevus on kuluefektiivne, orienteeritud kasutajate arvu kasvule ning turundusinvesteeringute tasuvusele. Seda kinnitavad eelkirjeldatud turundusraamistikud: idufirmadel soovitatakse turundusele läheneda süstemaatiliselt, püstitades hüpoteese ning neid eksperimentide või andmeanalüüsi abil valideerida. Veel enam – leides mõõdikud, mis aitavad määrata kliendi "kasulikkuse" juba tarbija teekonna varajases faasis, on võimalik ennustada ettevõtte pikaajaliste eesmärkide täitumist.

1.1.2 Jätkusuutliku kasvu tagamine kasvumootorite abil

Idufirmade jaoks on üks olulisimaid turunduskanaleid nende enda toode [23]. Esiteks on toode aluseks suusõnalisele turunduskommunikatsiooni levikule, teiseks on veebipõhistesse toodetesse võimalik ehitada erinevaid kasvumootoreid ning tagada sellega kasutajate arvu jätkusuutlik kasv [23]. Ka mitmed idufirma turundust käsitlevad teosed rõhutavad, et kiire ja jätkusuutliku kasvu saavutamiseks on

vajalik tootesse ehitada erinevaid kasvumootoreid [30, 11, 6]. Eric Riesi [11] loodud nutika idufirma metodoloogia defineerib jätkusuutliku kasvu kui olukorra, kus "uued kasutajad tulevad eelmiste kasutajate tegevuse tulemusena" [11, p. 218]. Järgnevalt kirjeldatakse Riesi [11] põhjal kolme kasvumootorit, mis aitavad jätkusuutlikku kasvu tagada.

Esimene võimalus on luua kinnimakstud kasvumootor, mille puhul eelmiste klientide tulu inveseeritakse uute klientide leidmisesse. Kinnimakstud kasvumootor toimib juhul, kui kliendi saamiseks tehtav kulu on keskmiselt väiksem kui kliendi eluea jooksul toodud tulu (ingl *customer lifetime value* - LTV). Kasvu kiirendamiseks on ettevõttel kaks võimalust: tõsta ühe kliendi pealt tema eluea jooksul saadavat tulu või vähendada kliendi saamiseks tehtud kulu.

Teine võimalus on külge jääv kasvumootor, mis on disainitud selliselt, et kliendid kasutaks toodet võimalikult pika ajaperioodi jooksul. Kasvu kiirendamiseks tuleb tõsta kasutajate määra, kes naasevad pärast eelmist kasutuskorda (ingl *retention rate*) ning langetada kasutajate määra, kes on toote kasutamisest loobunud (ingl *churn rate*). Kasvu kiirus sõltub uute kasutajate kaasamise määra ning eelmainitud loobunud kasutajate määra vahest: kui see on positiivne, on positiivne ka ettevõtte kasutajabaasi kogukasv.

Kolmas võimalus on viiruslik kasvumootor. Viirusliku leviku ideaalne juhtum on see, kui toode levib tootekasutuse loomuliku kõrvalmõjuna – ilma, et kasutaja tooteinfot teadlikult levitaks. Hea näitena toob ta Hotmail meilikeskkonna turundusvõtte, mille puhul lisati iga toote kaudu saadetud kirja lõppu viide tootele ning pani kasutajad seeläbi toote kohta infot jagama. Laiema definitsiooni annab samas seerias ilmunud teos "Nutikas analüütika" [8], kus viiruslikku kasvumootorit kirjeldatakse kui olukorda, kus senised kasutajad lisavad uusi kasutajaid. Sealjuures võib viiruslikkuse puhul eristada kolme tüüpi: loomulik viiruslikkus, kunstlik viiruslikkus ning suusõnaline viiruslikkus [8].

- Loomulik viiruslikkus (ingl *inherent virality*) tekib siis, kui kasutaja on motiveeritud teisi kasutajaid liituma kutsuma, sest see on kasulik tervele grupile. Loomulik viiruslikus tekib tihti mängude või sotsiaalvõrgustike puhul. Sealjuures märgitakse, et loomulik viiruslikkus on kõige siiram ning seetõttu ka eelistatuim viiruslikkuse variant [8].

- Suusõnaline viiruslikkus (ingl *word-of-mouth virality*) tekib siis, kui rahulolevad kasutajad jagavad oma kogemust ja algatavad vestluseid [8]. Seda tüüpi levikut on küll keeruline mõõta ja jälgida, kuid tegemist on ühe efektiivseima turundusmeetodiga [8]: ka traditsioonilise turunduse teooria kinnitab, et tarbijad usaldavad sõprade soovitusi märgatavalt rohkem kui brändide sõnumeid [2]. Eesti kasvufaasi idufirmad peavad samuti suusõnalist turundust ning soovitusi üheks efektiivseimaks turunduskanaliks. [23]

- Suusõnalist turunduskommunikatsiooni saab tehislikult edendada mängude, võistluste ja erinevate tasude abil [6] – seda nimetatakse kunstlikuks viiruslikuks ning siia alla kuuluvad ka soovitusprogrammid, mida antud töös käsitletakse. Kunstliku viiruslikkuse kriitikana tuuakse välja, et sellisel viisil on küll lihtne saada palju uusi kasutajaid, kuid need kasutajad võivad olla liitunud tasu saamise eesmärgil ning ei pruugi kujuneda aktiivseteks kasutajateks [8]. Seetõttu tasub tehisliku viiruslikkuse mõju hindamisel arvestada ja optimeerida ka nende uute kasutajate määra, kes toodet aktiivselt kasutama hakkavad.

Viirusliku kasvumootori olulisim mõõdik on viiruslik koefitsient, mis näitab, mitu uut kasutajat toob keskmiselt juurde iga kasutaja [11]. Koefitsendi arvutamiseks jagatakse seniste kasutajate tegevuse kaudu liitunud uute kasutajate arv (K) seniste kasutajate arvuga (N) [11]:

$$V = \frac{K}{N}$$

Mida kõrgem on koefitsient, seda kiiremini kasutajate arv kasvab – sealjuures avaldab isegi väike muutus kasutajate arvu kasvu kiirusele märgatavat mõju [11]. Jätkusuutliku kasvu tagamiseks on oluline, et koefitsient oleks üle 1,0 – see tähendab, et keskmiselt kutsub iga kasutaja liituma vähemalt ühe uue kasutaja [8]. Kiiret kasvu vajavate idufirmade puhul on soovituslik viiruslik koefitsient 1,2 [8]. Samas tõdetakse, et ehkki teoorias on see paljulubav, on tegelikkuses viiruslikku koefitsienti üle 1,0 väga keeruline saavutada [8]. Koefitsendi tõstmiseks soovitakse kaardistada tarbija teekond ning parendada ükshaaval seda iseloomustavaid mõõdikuid: lühendada tsükli aega, mis kulub kasutajal registreerumisest kuni teise kasutaja kutsumiseni; tõsta välja saadetud soovituste arvu või soovituse saanud kasutajate registreerumise määra (ingl *conversion rate*) [8].

Erinevaid viiruslikkuse tüüpe on omavahel võimalik ka edukalt kombineerida, kuid analüüsi ja optimeerimise faasis tuleb neid vaadelda eraldiseisvate kasvumeetoditena, kuna kasutajate käitumine ning erinevad mõõdikud võivad oluliselt erineda [8]. Samuti on oluline märkida, et viiruslik levik sõltub sihtgrupi ja toote iseloomust ning ei pruugi sobida igale ettevõttele [8]. Sel juhul tuleks viiruslikkust käsitleda kui kasvu võimendajat (ingl *force multiplier*), mis võimendab traditsioonilise turunduse tulemusi ning aitab suurendada turundusinvesteeringute tasuvust [8].

1.1.3 Soovitusprogrammid

Üheks idufirmade turundusmeetodiks on soovitusprogrammid (ingl *referral programs*). Soovitusprogrammide puhul tekitatakse kunstlik viiruslikkus, tasudes kasutajale soovituse eest allahindluse, tasuta kasutusvõimaluse, krediidi või muu hüvega [30]. Soovitussüsteemide efektiivsus tuleneb inimeselt-inimesele soovituste

usaldusväärsusest: eeldades, et sõbrad soovitavad neile tooteid, mille vastu neil huvi on, eelistavad enamik tarbijaid sõprade tootesoovitusi ettevõte soovitustele [6].

Soovitusprogrammidel on mitmeid eeliseid. Esiteks võib homofiilia ehk endasarnaste eelistamise põhimõtte tõttu eeldada, et nn kasulikud kasutajad kutsuvad liituma endasarnaseid inimesi, kes osutuvad samuti kasulikeks kasutajateks [34]. Seetõttu sobivad soovitusprogrammid hästi ettevõtetele, millel on väga piiratud turunduseelarve või spetsiifiline sihtgrupp, kelleni on traditsiooniliste turundusmeetodite abil keeruline pääseda [3]. Teiseks on leitud, et soovituse kaudu tulnud kasutajad on nii lühi- kui pikas perspektiivis kasumlikumad [34]. Kolmandaks, uute kasutajate kutsumine soovitusprogrammi raames tõstab kasutaja lojaalsust ettevõttele [15]. Sealjuures tugevdab kõrgem tasu soovituse eest nii hoiakulist kui ka käitumuslikku lojaalsust, samas kui väiksem tasu tugevdab ainult käitumuslikku lojaalsust [15]. See tähendab, et väiksema tasu puhul kasutajad küll jätkavad toote kasutamisega ja selle soovitamisega, kuid see ei mõjuta nende tundeid antud ettevõtte või brändi suhtes [4].

Enamikes soovitusprogramme puudutavates teadustöödes joonistub välja konkreetne kliendigrupp, kes toob kõige kõrgema väärtusega uusi kliente [34]. Mitmetes temaatilistes töödes on viidatud Feicki ja Price'i 1987.aasta uurimusele nn turu asjatundjatest (ingl *market mavens*) [13]: need on indiviidid, kellel on loomupärane huvi ning sellest tulenev teadlikkus erinevate toodete ja ostlemispaikade osas ning kes seetõttu teistele tarbijatele tihti infot jagavad. Turu arvamusliidrite identifitseerimiseks pakutakse välja Likerti-tüüpi küsimustik, mis koosneb väitest "Mulle meeldib tutvustada oma sõpradele uusi brände ja tooteid" ja viiest sarnasest väitest [13]. 2012. aasta uuring kinnitab, et turu asjatundjad teevad teiste kasutajatega võrreldes rohkem soovitusi, toovad rohkem uusi kasutajaid ja seetõttu ka rohkem tulu – seega tasub just neid sihtida [36].

Paraku ei ole ühtseid reegleid, mille põhjal kasulikud kliendid ära tunda: soovitusprogrammi edu määrab see, milliseid kasutajaid soovitusprogrammis osalemise üleskutsega sihitakse ning millist tasu neile soovituse eest pakutakse [34]. Soovitussüsteemi efektiivsuse tõstmiseks soovitatakse A/B testimise abil optimeerida nii uutele kasutajatele suunatud reklaammaterjale ja maandumislehti kui ka senistele klientidele suunatud kommunikatsiooni [8]. Samuti võib rakendada erinevaid mõjustamisvõtteid: näiteks on leitud, et kunstlik nappuse tekitamine ning isikustatud sõnumi kasutamine võib viirusliku turunduse kampaania algusfaasis soovituste määra tõsta [20]. Mitmes Eesti idufirmas on loodud spetsiaalne tehniline lahendus või rakendus, mille abil on kasutajal lihtne toodet sõpradele soovitada [23]. Häkkerturunduse printsiipidele toetudes tuleks eelkirjeldatud mõõdikute seast valida korraga üks mõõdik, mida kasutajakogemuse parendamise või A/B testimise abil parendada.

Hoolimata soovitusprogrammide efektiivsusest kasutatakse neid üllatavalt vähe

nii Eesti idufirmade seas – 2017. aasta uuringus oli kaheksast kasvufaasi idufirmast vaid üks suutnud luua toimiva soovitusprogrammi [31] – kui ka mujal maailmas [3]. Võimaliku põhjusena tuuakse välja, et 1) soovitusprogramme ei peeta efektiivseks turundusmeetodiks, 2) soovitusprogrammide tehnilist teostust peetakse liiga keerukaks või 3) kardetakse, et osad kasutajad hakkavad soovitusprogrammi finantsilise kasu saamise eesmärgil ära kasutama [3].

Nagu eelnevalt mainitud, ei ole soovitussüsteemid kõikide idufirmade jaoks efektiivne turundusstrateegia: see sõltub nii toote kui ka sihtgrupi iseloomust [23]. Samuti võib soovitussüsteem olla efektiivne vaid selle rakendamise algusperioodil: on vähetõenäoline, et kasutajad jätkavad uute kasutajate kutsumisega terve toote kasutamise perioodi ajal [6]. Olles kutsunud liituma need sõbrad, kes võiks tootest kõige suurema tõenäosusega huvitatud olla, lõpetatakse aktiivne soovitamine ning soovitussüsteemi viiruslikkuse koefitsient langeb [6]. Nimetatud tendentsile viitab ka Bassi uuenduse leviku teooria: innovatiivne toode hakkab levima aeglaselt, seejärel levik kiireneb suusõnalise turunduse tõttu ning kui enamik inimesi on tootest kuulnud, levik aeglustub taas [21, 8].

Töö autori hinnangul võib soovitusprogrammi efektiivsus varieeruda võrdlemisi palju: see sõltub konkreetsest tootest, sihtgrupist ning soovitamise eest saadavast tasust, mistõttu üheseid järeldusi on keeruline teha. See võib olla ka põhjuseks, miks erinevate kasutajagruppide efektiivsust on akadeemilisel tasandil seni vähe uuritud [3].

1.2 Masinõppe rakendamise võimalused turundusvaldkonnas

Masinõppe meetodite abil on arvutil võimalik andmeid kirjeldada või mingi tunnuse väärtust ennustada, kasutades selleks andmeid või varasemat kogemust [10]. Järgnevalt antakse akadeemilise kirjanduse põhjal ülevaade enimkasutatatud masinõppe rakendusvõimalustest turunduses.

Esiteks, masinõppe abil on võimalik tuvastada kõige suurema tõenäosusega toote ostmisest huvitatud tarbijad või tarbijate grupp [29]. Ettevõtte jaoks võib see tähendada märgatavat kokkuhoidu: väiksema, kuid täpselt sihitud sihtgrupini jõudmine on üldiselt kuluefektiivsem kui massturundus [29]. Ehkki paljud internetikasutajad võõristavad enda kohta andmete kogumist, on see suhtumine tõenäoliselt muutumas ja 2016.aasta uuringu põhjal [7] klikkavad noored hea meelega nende huvidele vastavatele reklaamidele.

Teiseks, juhendatud õppe abil on võimalik tuvastada ka need kliendid, kelle lahkumise tõenäosus on kõige suurem. Leitud tõenäosuse põhjal segmenteerimine on hea sisend lojaalsusprogrammide loomiseks ja käivitamiseks ning võib aidata vähendada klientide lahkumist [33]. Kuna vanade klientide hoidmine on ettevõt-

te jaoks soodsam kui uute klientide leidmine, tasub seniste klientide hoidmiseks ja kaasamiseks pingutada [30]. Samuti saab masinõppe abil optimeerida turundussõnumite ja -kanalite efektiivsust sihtgruppide lõikes: näiteks pole lojaalsele kasutajale mõtet sooduskupongi pakkuda, samas kui ebalojaalse kliendi lahkumist võib soodushind edasi lükata [16].

Kolmandaks, masinõppe meetodid võivad aidata tuvastada trende. See annab strateegilise eelise, mis aitab teha otsuseid õigeaegselt ning seeläbi nii kulusid kokku hoida kui ka tulusid suurendada [29]. Näiteks jaekaubanduses võimaldab turu ostukorvi analüüs (ingl *market basket analysis*) selgitada, milliseid tooteid tarbijad tihti koos ostavad [29], ning vastavalt sellele optimeerida turundussõnumeid või parendada tootepaigutust. Trendikaupade turul on edukaimad need kiirmoe brändid, mis suudavad reaalajas toodete liikumist ja trende analüüsida: see võimaldab kõige trendikamaid tooteid pakkuda ning annab seeläbi tugeva konkurentsieelise [1]. Kasutades traditsiooniliste turu-uuringute andmeid, saab teostada konkurentide analüüsi ning ennustada turuolukorra muutumist [9].

Lisaks andmete kogumisele enda veebikeskkonnast tasub kaasata ka erinevad sotsiaalmeedia kanalid: näiteks on võimalik tuvastada sotsiaalmeedias kasutaja ostukavatsus ning pakkuda talle sobivaid tooteid, otsides sarnaseid kasutajaid varem toodete arvustusi jaganud kasutajate seast [37]. Lisaks on sotsiaalmeediast eraldatava info põhjal võimalik suunata reklaame reaalajas muutuvate tunnuste põhjal nagu tarbija asukoht, uudisteportaalides aktuaalsed teemad ning meeleolu kasutaja sotsiaalmeedia postitustes. [16]. Samuti tasub jälgida, millised kasutajad suhtlevad ettevõttega rohkem kui ühes kanalis, sest need kasutajad on tihti lojaalsemad [16].

Peale eelkirjeldatud võimaluste on masinõppel veel hulgaliselt konkreetseid rakendusvõimalusi nagu näo- ja kõnetuvastus [10]. Töö autori hinnangul võivad need olla heaks võimaluseks efektsete reklaamikampaaniate läbi viimiseks, kuid turundusstrateegia seisukohast on olulisimad just eelkirjeldatud võimalused.

2 Metoodika

Antud peatükis antakse ülevaade magistritöö praktilise osa tööprotsessist ning kirjeldatakse metoodikat masinõppe rakendamiseks soovitussüsteemi optimeerimisel. Samuti kirjeldatakse ja põhjendatakse kasutatud meetodeid ning antakse ülevaade Promoty juhtumi puhul kasutatud andmestikest.

2.1 Sobiva idufirma valimine

Kuigi mitme idufirma esindajad olid masinõppe rakendamise võimalusest huvitatud, osutus takistuseks vähene andmete hulk või raskused konkreetse probleemi defineerimisel. Promoty kasuks otsustamisel sai määravaks kaks tegurit. Esiteks on töö autor ise Promoty meeskonnas, mistõttu on tal hea arusaamine idufirma ärimudelist ning tööprotsessidest. Samuti annab see võimaluse magistritöö tulemust hiljem rakendada ning vajadusel täiendavaid andmeid koguda, et mudeli täpsust parendada. Teine oluline kriteerium oli andmete olemasolu: Promotyga oli lühikese aja jooksul liitunud üle 8000 kasutaja, kelle kohta oli kogutud piisaval hulgal andmeid. Kolmas kriteerium oli mudeli loomise vajalikkus: kuna Promotyl on plaanis mõne kuu jooksul välisturgudele laieneda, on turunduskulutuste optimeerimiseks ja kiire kasvu tagamiseks oluline välja selgitada, millistel kasutajatel on kõige kõrgem tõenäosus osutuda kasulikuks kasutajaks.

Promoty arendab platvormi [27], mis ühendab ettevõtete esindajaid ja sotsiaalmeedia keskkonna Instagram [18] kasutajaid. Platvormi kaudu on ettevõtete esindajatel võimalik kontakteeruda korraga mitmete Instagrami kasutajatega ning saata neile pakkumine teha ettevõtet reklaamiv postitus. Pakkumise vastuvõtmise ning teostamise järel kantakse kasutajale kindlaksmääratud tasu, mis on määratud Promoty algoritmi poolt vastavalt kasutaja jälgijate arvule ning kaasatuse määrale. Idufirmale läheb vahendustasuna kindel protsent ettevõtete tasutud summast.

Platvormi lansseerimisel kasutati soovitusprogrammi, mis oli meeskonnaliikmete hinnangul esimeste kasutajate kaasamiseks väga efektiivne. Selleks loodi Instagrami konto, mille piltidele pandi üleskutse Promotyga liituda ning uudse platvormi abil raha teenida, ning hakati jälgima Instagrami kasutajaid. Esimesena hakati jälgima suurema jälgijaskonnaga kasutajaid ning tuntud inimesi, seejärel Instagrami algoritmi poolt soovitatud kasutajaid. Liitunud kasutajatele kuvati üleskutse kutsuda platvormiga liituma teisi kasutajaid. Preemiana said nad rahalist tasu: kui kasutaja poolt kutsutud uus kasutaja teeb Promotys mõne postituse, saab tema kutsuja kindla protsendi Promoty vahendustasust. Seega on kasutajaid kutsudes võimalus teenida passiivset tulu ning see oli töö autori hinnangul peamine põhjus, miks Promoty soovitussüsteem edukalt käivitus.

2.2 Tööprotsessi kirjeldus

Antud peatükis on meetodite kirjeldamisel ja põhjendamisel tuginetud Massachusetts Institute of Technology kirjastuse välja antud teosele "Introduction to Machine Learning" [10], kui ei ole märgitud teisiti.

Enamasti jagatakse masinõppe ülesanded juhendatud ja juhendamata õppeks. Juhendatud õppe eesmärk on luua mudel, mis ennustab ühe andmepunkti mingit tunnust vastavalt selle teistele tunnustele. Juhendatud õpe on omakorda jagatud klassifitseerimiseks, mille puhul ennustatakse andmepunkti mingisse klassi kuulumist, ja regressiooniks, mille puhul ennustatakse mingi kindla tunnuse väärtust. Juhendamata õpe ehk klasterdamine on sarnase trendide või mustritega andmepunktide grupeerimine, et andmeid paremini mõista ning selle põhjal otsuseid teha. Seega tuleb sobiva meetodi valimisel lähtuda konkreetsest probleemist ja ülesandest. Promoty näitel kasutati klassifitseerimist, kus jagati kasutajad kahte gruppi – kasulikud kasutajad ning mittekasulikud kasutajad – ning seejärel ennustati iga kasutaja kasulike kasutajate gruppi kuulumise tõenäosust. Samuti kasutati klasterdamist, et paremini mõista erinevaid kasutajagruppe ning neid iseloomustavaid tunnuseid. Järgnevalt tuuakse välja viis sammu masinõppe rakendamiseks soovitussüsteemi optimeerimisel.

1. Kasuliku kasutaja defineerimine

Kui masinõppe meetodite rakendamise eesmärk on kasulike kasutajate tuvastamine, on esimene samm antud ettevõtte jaoks kasuliku kasutaja defineerimine. Kuna jätkusuutliku kasvu tagamiseks peaks soovitusprogrammi viiruslikkuse koefitsient olema vähemalt 1,0 ehk iga kasutaja peaks kutsuma liituma vähemalt ühe uue kasutaja [8], defineeriti kasulik kasutaja Promoty puhul kui kasutaja, kes on Promotyga liituma kutsunud vähemalt ühe uue kasutaja. Mudeli ennustusvõime parendamiseks võib ka katsetada teiste kasuliku kasutaja definitsioonidega – Promoty puhul loodi ka sellised mudelid, kus kasulikuna defineeriti vähemalt 2, 5 või 10 uut kasutajat kutsunud kasutaja. Oluline on leida definitsioon, mis on nii ärilisest seisukohast põhjendatud kui ka mille puhul on mudeli ennustusvõime võimalikult kõrge.

2. Tunnuste valimine

Seejärel tuleb valida tunnused, mida mudelisse lisada. Tunnuste valikul on oluline arvestada, et tunnused peavad potentsiaalselt aitama määrata andmepunkti klassi ning olema kättesaadavad mudeli rakendamise ajal. Promoty puhul olid magistritöö valmimise faasis kättesaadavad järgmised andmed: 1) demograafilised tunnused, mis on kasutaja poolt sisestatud Promoty keskkonnas, 2) kasutaja tegevust Instagrami keskkonnas iseloomustavad tunnused, mis jõuavad Promoty keskkonda Instagrami rakendusliidese abil, ning 3) kasutaja tegevust Promoty keskkonnas iseloomustavad tunnused. Andmestiku tunnuste loetelu on välja toodud lisas Lisa II.

Kui eesmärk on sihtida uusi kasutajaid, kes ei ole veel idufirmaga seotud, tuleb mudeli koostamisel välja jätta idufirma tootega seotud tunnused nagu registreeru-

mise aeg või toote kasutamise sagedus. Samuti ei ole õige kaasata mudeli koostamisel ajas olulisel määral muutuvad tunnused, mille puhul pole võimalik määrata andmepunkti tunnuste väärtust. Näiteks Promoty puhul oli kättesaadav Instagrami jälgijate andmestik, kuid kuna ei olnud võimalik kindlaks teha, kas kasutaja oli Promoty jälgija Instagramis juba enne soovituste tegemist, ei saanud antud tunnust mudelisse kaasata. Sel põhjusel tasub võimalikule andmete muutumisele juba varem mõelda ning luua eraldi andmestik, kus kõik tunnused on määratud kasutaja liitumiskuupäeva väärtustega.

3. Treening- ja testandmete eraldamine ning tasakaalustamine

Mudeli õpetamiseks tuleb andmestik jagada treening- ja testandmeteks. Esmalt treenitakse mudelit osa andmete peal. Seejärel rakendatakse mudelit ülejäänud andmete peal, et võrrelda mudeli ennustatavat tulemust tegeliku tulemusega ning hinnata selle abil mudeli ennustusvõimet. Oluline on märkida, et andmestike puhul, kus ühe klassi osakaal on märgatavalt kõrgem kui teise klassi osakaal, võib mudel suuremat klassi rohkem arvesse võtta ning see mõjutab ennustuste täpsust. Promoty puhul oli vähemalt ühe uue kasutaja kutsunud kasutajate arv andmestikus märgatavalt madalam kui uusi kasutajaid mitte kutsunud kasutajate arv. Seetõttu prooviti mõlema meetodi puhul mudeli koostamisel mudeli ennustustäpsust alaesindamise (ingl *undersampling*) abil parandada. See tähendab, et treeningandmestiku loomisel kasutati kõiki vähemusklassi kuuluvaid andmepunkte ning lisati sama arv enamusklassi kuuluvaid andmepunkte. Testandmete osakaal tuleb samaks jätta.

4. Masinõppe meetodi valik ja rakendamine

Järgmise sammuna tuleb valida masinõppe meetod ning algoritm. Antud töös kasutati kasulike kasutajate tuvastamise mudeli loomiseks juhumetsa ning logistilist regressiooni: esiteks on tegemist suhteliselt lihtsate algoritmidega, teiseks sobivad need hästi binaarse muutuja tõenäosuse ennustamiseks suurte andmestike puhul [25]. Juhumets (ingl *random forest*) on algoritm, mis koosneb mitmetest otsustuspuudest [25]. Ehkki otsustuspuid kasutatakse ka eraldiseisva analüüsimeetodina, on juhumetsas paljude otsustuspuude ennustusvõime kombineeritud ning otsustusmetsa üldine ennustusvõime seega kõrgem [25]. Iga puu ehitamisel võetakse arvesse vaid juhuslik osa tunnustest, mistõttu sobib juhumets ka suure hulga tunnustega andmestike puhul [25]. Logistiline regressioon võimaldab lisaks mõõta iga tunnuse olulisust mudelis.

Juhendamata masinõppe ehk klasterdamise puhul võib kasutada *k-means* algoritmi, mis on kõige populaarsem eraldava klasterdamise algoritm [25]. Nimetatud algoritm genereerib andmestikule K-eralduse (ingl *K-partition*) [25]. Iga klastrit iseloomustab vastava klastri kese, mis on ühtlasi ka kõige tüüpilisem antud klastri esindaja [25]. Optimaalse klastrite arvu leidmiseks on kõige lihtsam võimalus *Elbow* meetod, mille rakendamisel tekkivalt jooniselt saab lihtsasti välja lugeda sobivaima klastrite arvu. Promoty puhul rakendati sobivaima klastrite arvu leidmiseks *Elbow*

meetodit ning kui parimat klastrite arvu selgelt välja ei joonistunud, *NbClust* paketti. Viimase eelis on see, et see kombineerib parima klastrite arvu välja selgitamiseks 30 erinevat algoritmi [26].

5. Loodud mudelite hindamine

Kõrgeima ennustusvõimega mudeli leidmiseks tuleb leida mõõdikud, mille alusel erinevaid mudeleid omavahel võrrelda. Antud töös tugineti peamiselt neljale mõõdikule: 1) esitustäpsus (ingl *precision*) ehk tõeste positiivsete tulemuste osakaal tõeste positiivsete ja valepositiivsete tulemuste summast, 2) saagis (ingl *recall*) ehk tõeste positiivsete tulemuste osakaal tõeste positiivsete ja valenegatiivsete summast, 3) maksimaalne saagis (ingl *max recall*) tingimusel, et esitustäpsus on üle 60 %, ning 4) ROC kõvera alune ala (AUC - ingl *area under the curve*, mis näitab ennustuse täpsust: mida rohkem erineb AUC 0,5-st, seda parem on mudel. Samuti joonistati olulisimate mudelite ROC kõver (ingl *receiver operating characteristic curve*), et visualiseerida tundlikkuse ehk tõeste positiivsete määra ja spetsiifilisuse ehk tõeste negatiivsete määra võimalikud kombinatsioonid. Lisaks toodi olulisimate mudelite puhul välja eksimisgraafik (ingl *confusion matrix*), mis näitab tõeste positiivsete, valepositiivsete, tõeste negatiivsete ja valenegatiivsete tulemuste hulka.

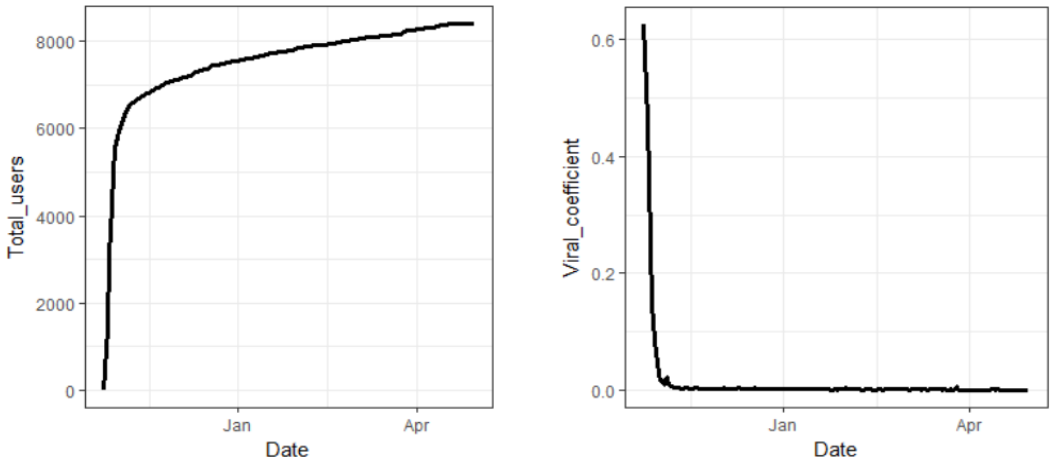
3 Tulemused

Antud peatükis tuuakse välja eelkirjeldatud meetodite abil läbi viidud analüüsi tulemused. Esimese alapeatükis analüüsitakse soovitusprogrammi kui kasvumootorit ning võrreldakse seda teise potentsiaalse kasvumootoriga. Seejärel hinnatakse kasutajate kasulikkuse hindamiseks loodud masinõppe mudeleid ning kirjeldatakse klasteranalüüsi tulemusi. Kasutatud kood R keeles on avalikult kättesaadav keskkonnas GitHub [24].

3.1 Soovitusprogramm kui kasvumootor

Visualiseerides Promoty kasutajate arvu muutmise ajas (joonis 2a), on selgelt näha, et soovitussüsteemi abil saavutati toote lansseerimisel küll viiruslik levik, kuid ligikaudu nädala möödudes aeglustus kasutajate arvu kasv järsult. Seda põhjendab nii teoreetilises kirjanduses välja toodud tendents, et soovitusprogrammi efektiivsus väheneb aja möödudes [6], kui ka Bassi uuenduslike toodete leviku teooria [21]. Samas peab märkima, et kui Bassi teooria kohaselt on uuenduslikel toodetel peale lansseerimist nn aeglase leviku periood, millele järgneb kiire kasv, siis Promoty puhul algas kiire kasv kohe pärast lansseerimist.

Soovitussüsteemi viiruslikkuse koefitsiendi muutumine peegeldab kasutajate arvu kasvu kõverat (joonis 2b): see on kõrgeim toote lansseerimisel ning langeb seejärel kiirelt, jäädes stabiilselt nulli lähedale. Seega võib öelda, et soovitusprogramm toimis eduka kasvumootorina vaid esimestel nädalatel pärast selle käivitamist.



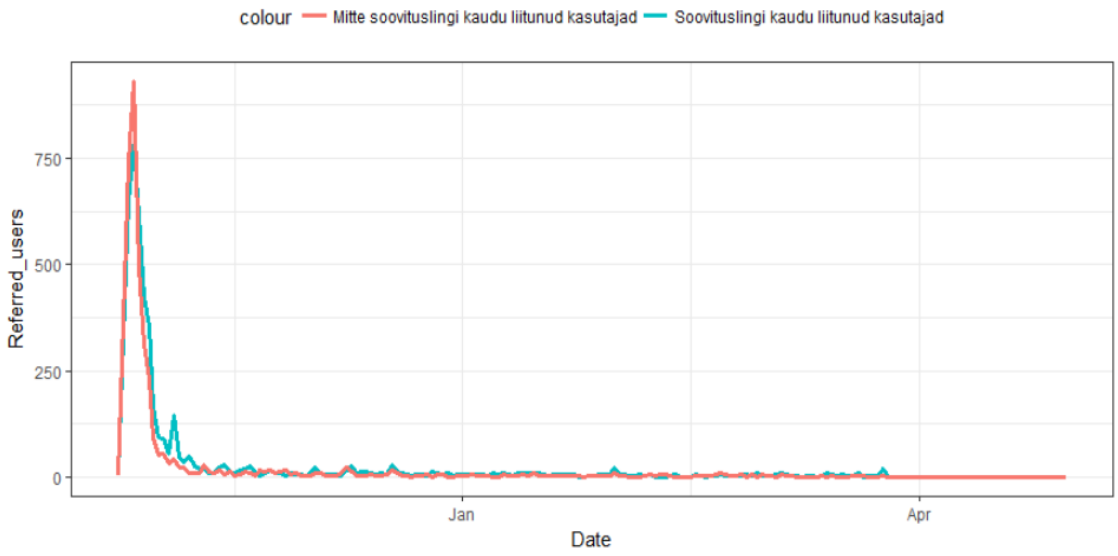
(a) Promoty kasutajate arv

(b) Promoty viiruslik koefitsent

Joonis 2. Promoty kasutajate arvu ja viiruslikkuse koefitsendi muutmine

Soovituslingi kaudu on Promotyga liitunud 4551 kasutajat, mis on 55,3% kõikidest kasutajatest. Võrreldes uute registreerunud kasutajate lisandumist ajas (joonis

3), on näha, et soovituslingi kaudu liitunud kasutajate arv on terve Promoty tegutsemisperioodi jooksul suhteliselt võrdne orgaaniliselt liitunud kasutajate arvuga. Kuna igal kasutajal oli unikaalne soovituslink kujul *promoty.eu/unikaalne-kood* ja soovituslinki jagati ka keskkondades, kus otse lingile klikata polnud võimalik (näiteks Instagrami pildi kirjelduses), on võimalik, et osad soovituslinki näinud kasutajad liikusid lingile klikkamise asemel otse aadressile *promoty.eu* ning liitusid sealtkaudu. Seetõttu võib soovitusprogrammi mõju uute kasutajate arvule olla tegelikult isegi suurem.



Joonis 3. Soovituslingi kaudu ja mitte soovituslingi kaudu liitunud kasutajate arv

Tuginedes peatükis 1 välja toodud idufirma kasvumootorite kirjeldusele, võib Promoty puhul toimida ka teine, tootesse mitteteadlikult ehitatud kasvumootor: kuna toode on uuenduslik ning reklaampostituste tegemine sotsiaalmeedia kasutajate poolt ei ole tavapärane, võib eeldada, et iga Promoty kaudu tehtud reklaampostitus tekitab suusõnalist turundust ning toob seeläbi uusi kasutajaid. Viimane variant vastab Riesi [11] kirjeldusele ideaalsest kasvumootorist, kus uued kliendid tulevad tootekasutuse kõrvalmõjuna. Kuna aga suusõnalise turunduse mõju pole võimalik täpselt mõõta, leiti võimaliku kasvumootori hindamiseks korrelatsiooniseosed postitusi iseloomustavate mõõdikute ning uute kasutajate arvu vahel.

Kuna töö autori hinnangul ei pruugi seos tunnuste vahel olla lineaarne ning tegemist ei pruugi olla normaaljaotusega, kasutati seoste analüüsimiseks Spearmani kordajat [35], mis suudab mõõta ka mittelineaarset seost. Korrelatsioonitabel (tabel 1) näitas, et postituste tegemise ja uute kasutajate liitumise vahel on positiivne, kuid nõrk seos. Tugevaim on seos Promoty kaudu tehtud reklaampostituste arvu ning kodulehe külastuste arvu vahel (Spearmani korrelatsioonikoefitsient 0.3044). Üllataval

Tabel 1. Korrelatsiooniseosed postituste ja uute kasutajate arvu vahel

	Kodulehe külastuste arv	Liitunud kasutajate arv kokku	Soovituslingi kaudu liitunud kasutajate arv	Mitte soovituslingi kaudu liitunud kasutajate arv
Postituste arv	0.3044	0.2474	0.2986	0.1103
Postituste levik	0.2988	0.2305	0.2911	0.0875
Meeldimiste arv	0.2987	0.2325	0.2927	0.0886
Kommentaaride arv	0.2874	0.2179	0.2827	0.0553

kombel on seos postituste tegemise ja soovituslingi kaudu liitunud kasutajate arvu vahel märgatavalt tugevam kui postituste tegemise ja mitte soovituslingi kaudu liitunud kasutajate vahel. See tähendab, et postituste tegemine on Promoty puhul seotud soovitussüsteemiga ning seda ei saa analüüsida eraldiseisva kasvumootorina.

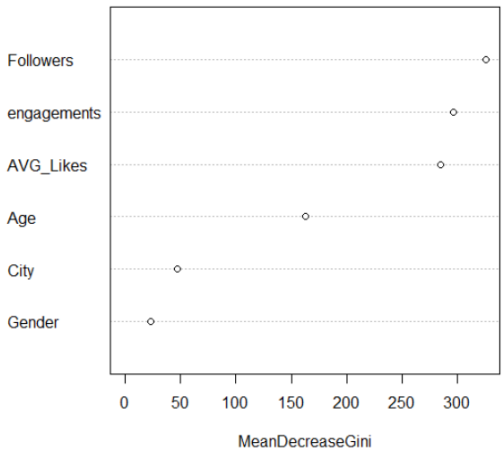
3.2 Uute kasutajate kasulikkuse ennustamine

Antud magistritöö üks eesmärke on hinnata masinõppe mudelite abil võimalikult täpselt iga Promoty kasutaja puhul tõenäosust, kas tegemist on nn kasuliku kasutajaga, kes on valmis läbi personaalse soovituslingi kutsuma Promotyga liituma vähemalt ühe uue kasutaja. Mudelite loomiseks rakendati juhumetsa ning logistilist regressiooni. Mõlemal juhul kasutati kuut tunnust: kasutaja vanust, sugu ning elukohta, samuti jälgijate arvu, keskmist meeldimiste arvu ning kaasatuse määra Instagrami keskkonnast. Tunnuste kirjeldus on leitav lisast Lisa I.

3.2.1 Juhumetsa abil

Juhumetsa muutujate olulisuse graafikult selgub, et kasuliku kasutaja ennustamise mudelis on demograafilistest tunnustest olulisemad tunnused, mis iseloomustavad kasutaja aktiivsust Instagrami keskkonnas: kasutaja jälgijate arv, keskmine kaasatuse määr ning keskmine meeldimiste arv (tabel 4). Mudeli rakendamise seisukohast on see positiivne, kuna võimaldab kasulikuks osutumise tõenäosust hinnata ka nende Instagrami kasutajate puhul, kelle demograafilised andmed ei ole teada.

Promoty jaoks on oluline jõuda võimalikult paljude potentsiaalsete kasulike kasutajateni. Seetõttu otsiti masinõppe mudelite hindamisel maksimaalset saagist, mille juures esitustäpsus on üle 60% ehk mudel suudab kasulikele kasutajatele anda õige hinnangu enamatel kordadel kui vale hinnangu. Paraku on juhumetsa mudeli ennustusvõime suhteliselt madal: tasakaalustamata treeningandmete põhjal loodud mudeli maksimaalne saagis on 60% ennustustäpsuse juures 19,3% (tabel 2) ning tasakaalustatud treeningandmete põhjal loodud mudeli puhul 12,4%. ROC kõvera



Joonis 4. Tunnuste olulisus tasakaalustatud treeningandmete põhjal loodud mudelis

Tabel 2. Juhumetsa mudelite võrdlus

Tasakaalustamata treeningandmete põhjal loodud mudel			Tasakaalustatud treeningandmete põhjal loodud mudel		
	0	1		0	1
0	3293	770	0	3340	333
1	57	32	1	2241	458
Precision: 35,96%			Precision: 16,97%		
Recall: 3,99%			Recall: 57,90%		
Max. recall: 19,32%			Max. recall: 12,43%		
AUC: 0.5546			AUC: 0.6212		

alune ala on esimesel juhul 0,59 ning teisel juhul 0,62, olles mõlemal juhul lähedal 0,5-le ning kinnitades mudeli suhteliselt madalat ennustusvõimet.

3.2.2 Logistilise regressiooni abil

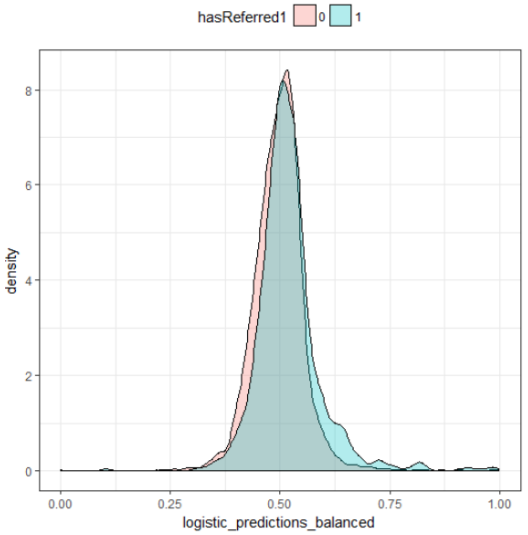
Teise masinõppe meetodina kasutati logistilist regressiooni, kus sisendparameetriteks võeti samad tunnused. Koefitsientide analüüs eksponentfunktsiooni rakendamise järel (tabel 3) näitab, et statistiliselt olulisim tunnus on kasutaja sugu (koefitsient 0,74), kus meessoost kasutajate tõenäosus osutuda kasulikuks kasutajaks väiksem kui naissoost kasutajate puhul. Samuti selgub nimetatud tabelist, et tõenäosus on kõrgem kasutajate puhul, kes on pärit keskmistest või väikestest linnadest (koefitsendid vastavalt 1.14 ja 1.18). Oluline on märkida, et erinevalt juhumetsa

Tabel 3. Tunnuste koefitsendid logististilise regressiooni mudelis

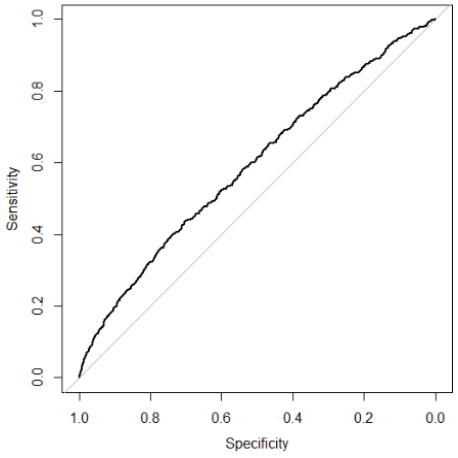
Vanus	Sugu1	Linn1	Linn2	Jälgijaid	Meeldimisi	Kaasatus
0.9797	0.7422	1.1421	1.1788	0.9999	1.0012	0.9904

mudelist pole Instagramiga seotud mõõdikutest statistiliselt oluline mitte ükski.

Mudeli ennustustäpsuse hindamiseks loodi tihedusgraafik, et hinnata testandmete peal tehtud ennustuste vastavust tegelikule tulemusele. Joonisel 5a) on näha, et punane ja roheline ala kattuvad olulisel määral. See tähendab, et tegelikult kasulikuks osutunud kasutajatele ennustati sarnast tõenäosust kui tegelikult mittekasulikuks osutunud kasutajatele. Madalat ennustusvõimet näitab ka mudeli ROC kõver (joonis 5b), mis on suhteliselt sirge. Mudeli parendamiseks kasutati treeningandmete tasakaalustamist, kuid nagu ka juhumetsa mudeli puhul, vähendas tasakaalustamine maksimaalset saagist (tabel 4).



(a) Tihedusgraafik



(b) ROC kõver

Joonis 5. Logistilise regressiooni mudeli tihedusgraafik ning ROC kõver

Parema ennustusvõimega mudeli leidmiseks loodi täiendavaid logistilise regressiooni mudeleid erinevate kasuliku kasutaja definitsioonidega (tabel 5). Selleks lisati igale andmepunktile binaarsed tunnused *hasReferred2*, *hasReferred5*ja *hasReferred10* vastavalt sellele, kas kasutaja on kutsunud Promoty platvormile vähemalt kaks, viis või kümme uut kasutajat. Sealjuures selgus, et mida kõrgem on kasuliku kasutaja defineerimise kriteerium, seda suurem on ROC kõvera alune ala ning seda väiksem on maksimaalne saagis 60 % esitustäpsuse juures.

Tabel 4. Logistilise regressiooni mudelite võrdlus

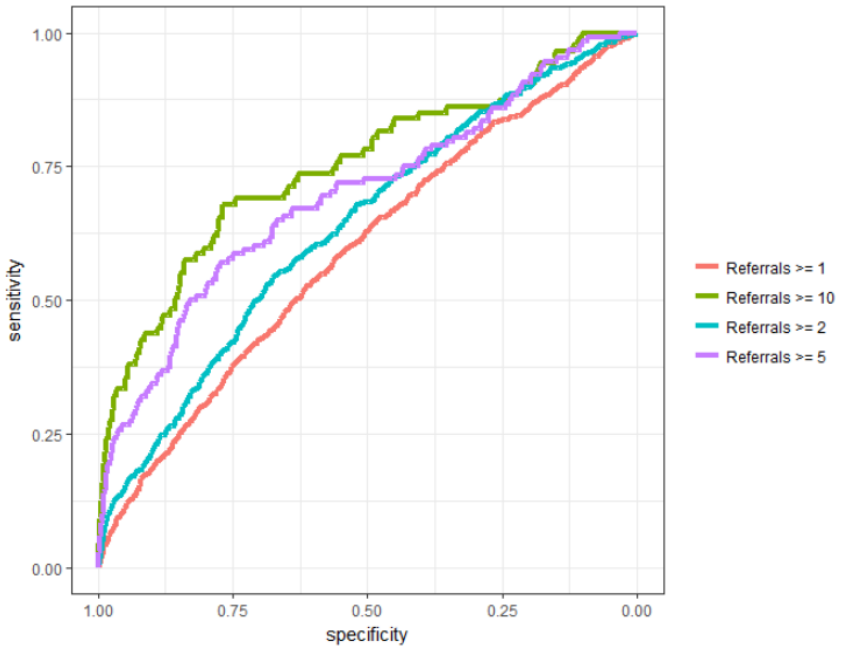
Tasakaalustamata treeningandmete põhjal loodud mudel			Tasakaalustatud treeningandmete põhjal loodud mudel		
	0	1		0	1
0	2077	1273	0	2702	2879
1	385	417	1	280	511
Precision: 24,67%			Precision: 15,07%		
Recall: 51,99%			Recall: 64,60%		
Max. recall: 23,57%			Max. recall: 15,07%		
AUC: 0.5919			AUC: 0.5919		

Tabel 5. Mudeli ennustusvõime sõltuvus kasuliku kasutaja definitsioonist

Kasuliku kasutaja definitsioon	Soovituste arv >= 1	Soovituste arv >= 2	Soovituste arv >= 5	Soovituste arv >= 10
Kasulike kasutajate arv	1582	791	259	179
Kasulike kasutajate osakaal	18,85%	9,43%	3,09%	2,13%
Max. recall	23,57 %	13,36 %	5,51 %	4,10 %
AUC	0.5919	0.6345	0.6898	0.7130

Nimetatud tendentsi võib põhjendada kasulike kasutajate osakaalu vähenemisega: kuna andmestik on tugevalt tasakaalust väljas, ennustab mudel peamiselt enamusklassi ehk uusi kasutajaid platvormile mitte kutsunud kasutajate põhjal. Ka tunnuste *hasReferred5*ja *hasReferred10* põhjal loodud mudelite ROC kõverad 6 näitavad, et mudel suudab suhteliselt täpselt ennustada mittekasulikke kasutajaid, kuid ei suuda ennustada kasulikke kasutajaid. Seega võib Promoty puhul teha järelduse, et mida kõrgem on kasuliku kasutaja kriteerium, seda väiksema täpsusega suudab mudel kasulikke kasutajaid ennustada.

Võrreldes kõiki töö käigus loodud mudeleid, osutus kõige kõrgema ennustusvõimega mudeliks logistilise regressiooni mudel, kus kasulik kasutaja on defineeritud kui kasutaja, kes on platvormile kutsunud vähemalt ühe uue kasutaja. Antud mudeli maksimaalne saagis on 23,6%, mis tähendab, et mudel suudab Promotyga mitteseotud Instagrami kasutajate puhul tuvastada 23,6% kõikidest kasulikest kasutajatest. Sama kasuliku kasutaja kriteeriumiga juhumetsa meetodil loodud mudelil oli mõnevõrra madalam ennustusvõime (maksimaalne saagis 19,32%; ROC kõvera alune ala 0.55). Sealjuures olid mudelites oluliseks peetud tunnused täiesti erinevad: kui logistilise regressiooni puhul mõjutasid tulemust kõige rohkem kasutaja sugu



Joonis 6. Logistilise regressiooni ROC kõver erineva kasuliku kasutaja definitsiooni korral

ja elukoht, siis juhumetsa mudelis olid olulisimad tunnused kasutaja jälgijate arv, kaasatuse määr ning keskmine meeldimiste arv.

3.3 Kasutajate segmenteerimine

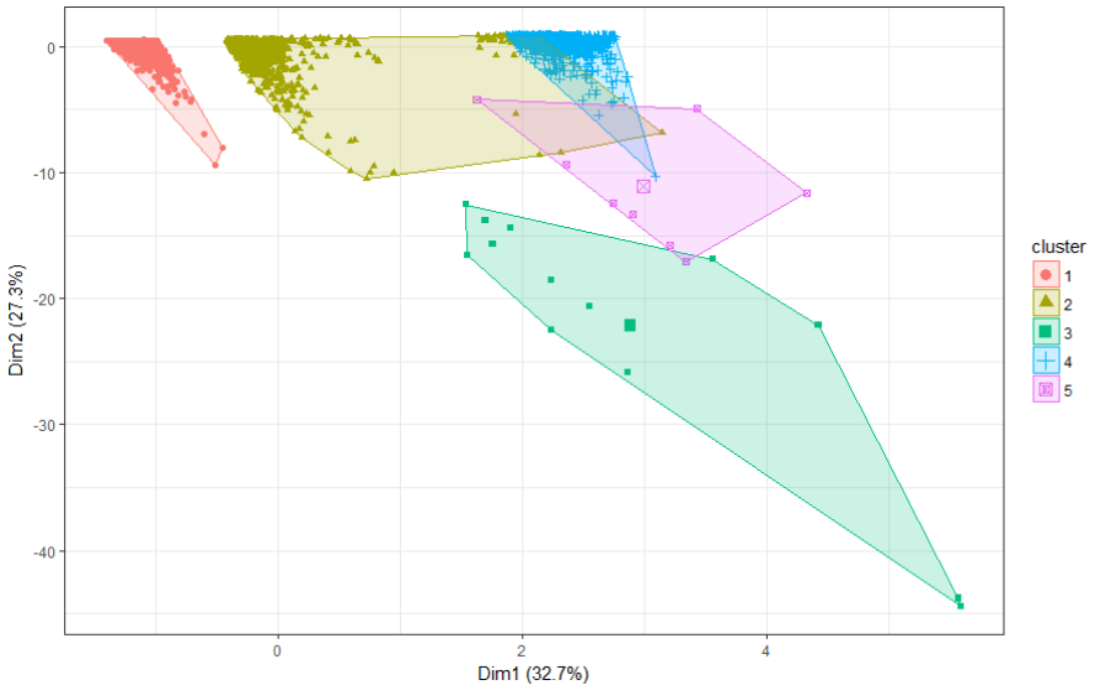
Kuna juhendatud õppe ennustusvõime osutus pigem madalaks, rakendati andmete mõistmiseks ning kasulikke kasutajaid iseloomustavate tunnuste välja selgitamiseks klasteranalüüsi.

Erinevate meetodite võrdlemisel kasutati mediaani. Erandiks oli kutsutud uute kasutajate arv kasutaja kohta, mille puhul mediaankeskmine oli mitmel grupil 0 ning kus võrdluse tekitamiseks kasutati aritmeetilist keskmist.

Esmalt võeti klasteranalüüsis arvesse kõik tunnused, mida kasutati uute kasutajate mudeli loomisel: vanus, sugu ja linn ning jälgijate arv, keskmine meeldimiste arv ning kaasatus Instagrami keskkonnas. Sel viisil tekkis kolm klastrit, mis on üksteisest selgelt eristuvad: esimeses klassis on teistest oluliselt vähem kasutajaid ning nende aktiivsus Instagrami platvormil on märgatavalt kõrgem (jälgijate arvu mediaan 19 118, teistel klastritel vastavalt 404 ja 533). Esimene ja kolmas klaster on sarnased, kuid esimeses klastris on ülekaalus naissoost kasutajad ning kolmandas klastris meessoost kasutajad. Võrreldes soovitussüsteemiga seotud mõõdikuid nagu keskmine soovituste arv kasutaja kohta, on erinevused minimaalsed. Seega on kõiki

tunnuseid kaasav klasterdamine küll kasulik andmete paremaks mõistmiseks, kuid ei aita kaasa töö eesmärgi saavutamisele.

Eesmärgiga selgitada, millised tunnused iseloomustavad kasulikke kasutajaid, vähendati tunnuste arvu ning võeti klasterdamisel arvesse vaid need tunnused, mis on otseselt seotud kasutaja tegevusega Instagrami keskkonnas: jälgijate arv, keskmine meeldimiste arv ning kaasatuse määr. Kuivõrd Instagrami platvormil tarbijatega suheldes oluline ka inimlik faktor ning kuna kasutaja vanus ja elukoht ei pruugi kohe selguda, on lihtsam manuaalselt hinnata kasutaja kasulikkust eelnimetatud tegurite põhjal. Lisaks kaasati tunnusena mudelisse sugu, sest seda tunnust on profiilipildi või nime alusel lihtne ka manuaalselt tuvastada. *Elbow* meetodit rakendades joonistus selgelt välja, et optimaalne klastrite arv on viis. *Kmeans*-i abil leitud viis klastrit on visualiseeritud joonisel 7) ning klastreid iseloomustavate tunnuste mediaanid on välja toodud lisas Lisa II.



Joonis 7. Promoty kasutajate klastrid

Eelkirjeldatud viisil klasterdades tekib selge erisus idufirma jaoks oluliste mõõdikute osas: kui esimesel kolmel klastril jääb keskmine kutsutud kasutajate arv ühe kasutaja kohta alla 1,0, siis neljanda klastri puhul on see 7,8 ning viienda klastri puhul 130 (tabel 6). Seega on klastermeetodi abil identifitseeritud kaks gruppi aktiivseid soovitajaid:

- Klastrisse 4 kuuluvatel kasutajatel on Instagramis ligikaudu 5000 jälgijat

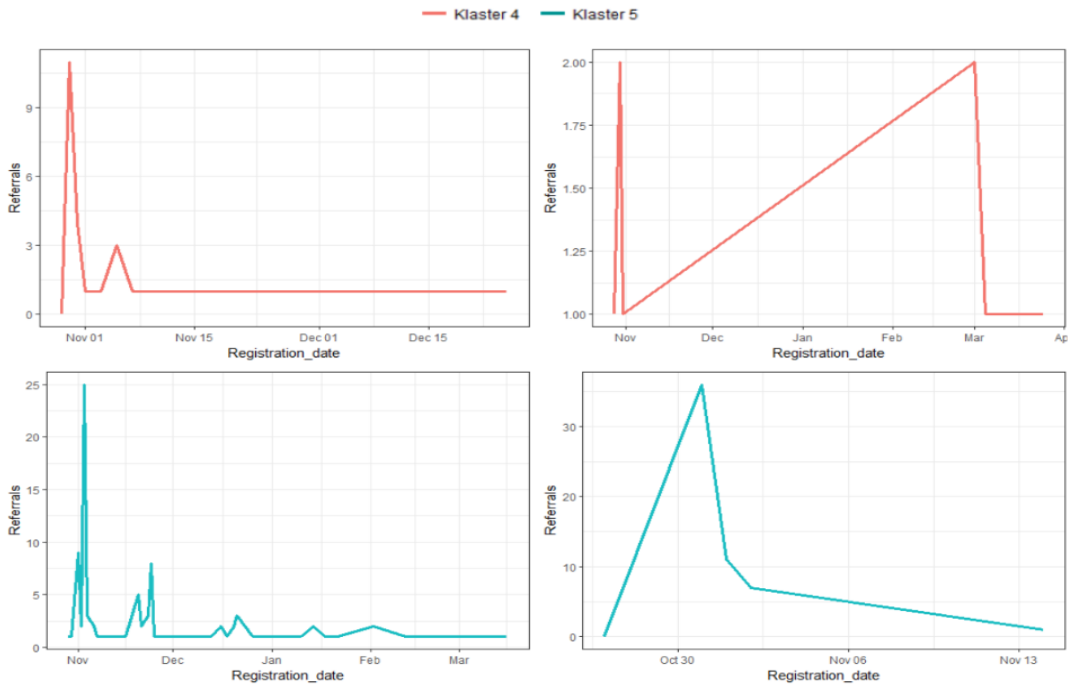
Tabel 6. Aktiivseid soovitajaid ja brändisaadikuid iseloomustavad tunnused

	Aktiivsed soovitajad	Brändisaadikud
Kasutajate arv	167	7
Keskmine jälgijate arv	4933	10687
Keskmine meeldimiste arv	818	1254
Keskmiselt kutsutud uusi kasutajaid ühe kasutaja kohta	7,8	130
Kasutajate arv ja osakaal, kes on kutsunud vähemalt 1 uue kasutaja	84 (50 %)	7 (100 %)
Kasutajate arv ja osakaal, kes on kutsunud vähemalt 10 uut kasutajat	56 (34 %)	7 (100 %)

ning piltidel keskmiselt 800 meeldimist. Tegemist on aktiivsete soovitajatega, keskmiselt on iga kasutaja kutsunud Promotyga liituma 8 uut kasutajat. Sealjuures on oluline märkida, et pooled antud klastrisse kuuluvatest kasutajatest ei ole oma soovituslingi kaudu Promotysse veel ühtegi uut kasutajat kutsunud. Samas on nad sarnased teistele samasse klastrisse kuuluvatele kasutajatele, mistõttu on neil töö autori hinnangul potentsiaali osutuda kasulikuks kasutajaks ning neid tuleks turundussõnumitega sihtida.

- Klastrisse 5 kuuluvad kasutajad on Instagramis väga aktiivsed, neil on ligikaudu 10 000 jälgijat ning piltidel keskmiselt 1300 meeldimist. Kõik klastrisse kuuluvad kasutajad on kutsunud Promotyga liituma vähemalt 10 uut kasutajat, sealjuures on keskmine kutsutud uute kasutajate arv 130. Seega võiks neid nimetada Promoty brändisõnumi edasikandjateks ehk brändisaadikuteks [17]. Töö autori hinnangul on see kasutajagrupp ettevõttele kõige väärtuslikum, mistõttu tuleks hoida häid suhteid seniste brändisaadikutega ning sihtida turundussõnumeid sarnastele kasutajatele, kes võiks samuti brändisaadikuteks osutuda.

Eelkirjeldatud kasutajagruppide paremaks tundma õppimiseks valiti mõlemast klastrist kaks tüüpilist esindajat, kelle tunnused vastasid klastri keskmistele tunnustele, ning vaadeldi, millise ajaperioodi jooksul ning millise sagedusega on liitunud nende soovituslingi kaudu kutsutud kasutajad. Graafikuid võrreldes selgub, et graafikud on langeva trendiga: enamik kasutajaid kutsus kõige rohkem uusi kasutajaid liituma mõne päeva jooksul pärast registreerumist. Samas on uute kasutajate kutsumine erinev isegi samasse klastrisse kuuluvate kasutajate puhul ning nii väikese valimi puhul tüüpilist käitumismustrit välja ei joonistu (joonis 8).



Joonis 8. Huvipakkuvate klastrite tüüpiliste kasutajate soovituste arv ajas

4 Arutelu

Järgnevas peatükis tuuakse välja olulisimad tulemused ning kirjeldatakse nende tähendust idufirma turunduse kontekstis. Samuti tehakse ettepanekud mudeli ennustusvõime tõstmiseks ja edasiseks analüüsiks.

4.1 Tulemuste tõlgendamine

Kirjeldav analüüs näitas, et Promoty näitel oli soovitusprogramm efektiivne küll esimese nädala jooksul, kuid seejärel aeglustus kasutajate arvu kasv järsult. See tähendab, et uutele turgudele sisenedes tasub esimeste kasutajate kaasamiseks kindlasti soovitusprogramm käivitada, kuid seejärel on vajalik täiendavate turundustegevuste läbiviimine. Eesti kontekstis on Promoty puhul praeguses faasis oluline mõelda, kuidas ergutada kasutajaid taas aktiivselt soovituslinki jagama. Kuna postituste tegemise ning uute kasutajate lisandumise vahel on nõrk positiivne seos (Spearmani korrelatsioonikoefitsient 0,25), võib eeldada seost platvormi kasutavate brändide arvuga: mida rohkem tehakse kampaaniapostitusi, seda rohkem tekib suusõnalist viiruslikkust ning see mõjub positiivselt uute kasutajate arvule.

Ennustava mudeli ehitamisel otsiti mudelit, mis suudaks Promotyga mitteseotud kasutajate puhul hinnata võimalikult täpselt tõenäosust, kas tegemist on Promoty

jaoks nn kasuliku kasutajaga. Kõrgeima saagisega oli tasakaalustamata treeningandmete põhjal loodud logistilise regressiooni mudel, mis suudab Promotyga mitteseotud Instagrami kasutajate puhul tuvastada 60% esitustäpsuse juures 23,6% kõikidest kasulikest kasutajatest. Töö autori hinnangul aitab selline mudel küll uutele turgudele sisenedes ressursse mõnevõrra efektiivsemalt suunata, kuid soovituslik on mudeli ennustusvõime parendamiseks koguda täiendavaid tunnuseid. Võimalikud potentsiaalselt andmepunkti klassi määravad tunnused pakutakse välja peatükis 4.2.

Oluline on märkida, et mudeli ennustusvõime sõltub kasuliku kasutaja definitsioonist. Promoty puhul oli nii, et mida kõrgem on kasuliku kasutaja defineerimise kriteerium, seda väiksem on kasulike kasutajate osakaal andmestikus ning seda rohkem ennustab mudel enamusklassi ehk uusi kasutajaid platvormile mitte kutsunud kasutajate põhjal. See tähendab, et mudel suudab suhteliselt täpselt ennustada mittekasulikke kasutajaid, kuid ei suuda ennustada kasulikke kasutajaid. Samas võib juhtuda, et rohkem soovitusi teinud kasutajad on mingite tunnuste põhjal selgelt eristatavad, mistõttu on mudelil neid kergem ennustada – seetõttu soovitab töö autor teistel idufirmadel samuti erinevate kasulike kasutajate definitsioonidega katsetada ning leida definitsioon, mis on nii ärilisest seisukohast põhjendatud kui ka mille puhul on mudeli ennustusvõime kõrge.

Klasteranalüüsi tulemusel joonistus välja kaks kasutajate gruppi, kelle keskmine soovituste arv inimese kohta on märgatavalt kõrgem kui teistel klastritel. Töö autori hinnangul on need kasutajagrupid Promoty jaoks kõige väärtuslikumad, mistõttu tuleb nendega aktiivselt suhelda ning turunduskommunikatsioonis sarnaseid kasutajaid sihtida. Oluline on märkida, et ehkki kõik kirjeldatud klastritesse kuuluvatest kasutajatest ei ole Promotysse veel kasutajaid kutsunud, on nad sarnased teistele samasse klastrisse kuuluvatele kasutajatele, mistõttu on neil töö autori hinnangul potentsiaali osutuda kasulikuks kasutajaks ning neid tuleks turundussõnumitega sihtida nii Eesti turul kui ka välisturgudele sisenedes. Teistel idufirmadel soovitab töö autor samuti katsetada klasteranalüüsi teostamisel erinevate tunnuste kombinatsioonidega, et leida nende ettevõtte jaoks kõige tähenduslikumad kasutajagrupid.

4.2 Edasiarenduse võimalused

Antud töö puhul osutus peamiseks piiranguks tunnuste vähesus. Vaatamata sellele, et kättesaadavaid tunnuseid oli rohkem, otsustati peatükis 2.2 kirjeldatud põhimõtetele kaasata lõplikku mudelisse kuus tunnust. Kuna saadud mudeli ennustusvõime jääb suhteliselt madalaks, soovitab töö autor Promotyl koguda kasutajate kohta täiendavaid tunnuseid ning kasutada neid mudeli täiendamiseks. Sel viisil on suurem tõenäosus leida tunnus või tunnuste kombinatsioon, mis kasutaja kasulikkust kõige rohkem mõjutab.

Üks parimaid kohti andmete kogumiseks potentsiaalsete kasutajate kohta on

sotsiaalmeedia: lisaks demograafilistele andmetele saab sealt infot kasutaja huvide, tarbimisharjumuste ja ostukavatsuste kohta [37]. Laia valiku täiendavaid võimalusi annab pildi- ja tekstituvastustehnoloogiate rakendamine, näiteks piltide värvigamma hindamine ning piltidel olevate objektide tuvastamine ja analüüs. Kui pildid on sarnases värvigammas, võib eeldada, et kasutaja on oma jagatud sisu hoolikalt läbi mõelnud ning pühendab piltide valikule ja töötlemisele aega, mistõttu võivad Promoty-taolised sotsiaalmeediaga seotud teenused talle keskmisest enam huvi pakkuda. Samuti võib analüüsida piltide kirjelduste pikkust, selles sisalduvaid märksõnu ja tundmuseid, teemaviidete kasutust ja profiilil olevas tutvustavas tekstis sisalduvaid märksõnu.

Lisaks tasub uurida võrgustiku mõju: kui paljud kasutaja poolt jälgitud kasutajad soovitavad antud ettevõtet, võib eeldada, et kasutaja on rohkem aldis ettevõtte teenuseid tarbima ning seda ka teistele soovitama. Promoty puhul saab tekstituvastustehnoloogiate abil vaadata, kui paljudel kasutaja sõpradel on tutvustavas sektsioonis Promoty soovituslink – kuna töö autori hinnangul on see üks populaarsemaid kanaleid personaalse soovituslingi levitamiseks, võib see võrgustiku puhul olulist mõju omada.

Esimesena soovitab töö autor lisada Promoty kasutajate andmestikule postituste arvu ning tema poolt jälgitud kasutajate arvu: need tunnused on kõikidel avalikel profiilidel selgelt välja toodud, mistõttu on neid lihtsam koguda. Lisaks võib eraldi tunnusena leida kasutaja jälgijate ning jälgitud kasutajate arvu suhte, mis võib samuti klassi määramisele kaasa aidata. Samuti soovitab töö autor Promotyl uurida lähemalt klasteranalüüsis välja tulnud aktiivsete soovitajate ning brändisaadikute kasutajagruppe – mõistes, miks antud kasutajad Promotyt aktiivselt soovitanud on, saab edasises turunduskommunikatsioonis just neid aspekte rõhutada. Kasutajate tundma õppimiseks võib klastrite esindajatega läbi viia poolstruktureeritud intervjuusid või saata laiali küsimustik. Sealjuures võib küsimuste koostamisel toetuda Feicki ja Price'i välja töötatud küsimustikule turu asjatundjate tuvastamiseks [13].

Instagrami puhul võib piiranguks osutuda ka Instagrami rakendusliideste pidev muutumine: näiteks aprillis suleti kasutajate andmete kaitsmiseks mitmed Instagrami rakendusliidesed, mis põhjustas mitmete Instagramiga seotud platvormide sulgemise või ajutise töö katkemise [19]. Seetõttu on võimalik, et kõiki eelkirjeldatud andmeid Instagrami rakendusliidese kaudu kätte ei saa ning on vajalik kasutada täiendavaid tööriistu andmete kraapimiseks.

Antud töö valmimise hetkel Promotys teisi turundustegevusi olulisel määral läbi ei viidud, mistõttu polnud nende mõju hindamine magistritöö raames relevantne. Kuna aga soovitussüsteemi mõju on tugevalt vähenenud, on idufirmal vajalik alustada täiendavate turundustegevustega. Seega soovitab töö autor Promotyl koguda andmeid ettevõtte turundustegevuse kohta, et analüüsida erinevate kanalite ning sõnumite mõju nii soovituslingi kui ka mitte soovituslingi kaudu liitunud uute

kasutajate arvule. Kuivõrd Promoty puhul on peamised turunduskanalid Instagram ning blogi, tasub koostada eraldi andmestik tehtud Instagrami postituste kuupäevade, leviku, meeldimiste ja kommentaaride arvuga ning blogiartiklite avaldamise kuupäevade ja lugemiste arvuga. Oluline on silmas pidada, et turundustegevusena arvestataks ka potentsiaalsete kasutajatega suhtlemist Instagramis – selle puhul saab mõõta panustatud aega või uute kasutajate arvu, kellega suheldi.

Kuna Promoty puhul on maksvaks kliendiks ettevõtted, kes Promoty abil oma tooteid või teenuseid saavad reklaamida, tasub Promotyl tulevikus kindlasti luua ennustavad mudelid ärikasutajate kasulikkuse hindamiseks.

Kokkuvõte

Idufirmade turundust käsitleva kirjanduse põhjal on soovitusprogrammid üks parimaid võimalusi kiire kasvu saavutamiseks. Vaatamata sellele ei ole paljud Eesti idufirma suutnud neid edukalt rakendada. Seetõttu pakutakse antud magistritöös Eesti idufirmadele välja metoodika juhendatud ja juhendamata masinõppe rakendamiseks soovitusprogrammi analüüsiks ja optimeerimiseks. Masinõppe abil on võimalik leida ja sihtida neid idufirmaga mitteseotud kasutajaid, kellel on suurim tõenäosus saada kasulikuks kasutajaks. Täpsem sihtimine võimaldab ressursse efektiivsemalt suunata ning seeläbi kiirema kasvu saavutada.

Magistritöö raames seati neli eesmärki:

1. Pakkuda Eesti idufirmadele välja metoodika kirjeldava analüüsi, juhendatud ja juhendamata masinõppe rakendamiseks soovitusprogrammi analüüsiks ja optimeerimiseks.

2. Luua masinõppe mudel, mis hindab iga Instagrami kasutaja puhul võimalikult täpselt tõenäosust, kas tegemist on Promoty jaoks nn kasuliku kasutajaga, kes on valmis läbi personaalse soovituslingi kutsuma Promotyga liituma vähemalt ühe uue kasutaja.

3. Selgitada välja, millised tunnused iseloomustavad Promoty jaoks kasulikku kasutajat, ning sellest tulenevalt teha ettepanekuid soovitusprogrammi efektiivsemaks rakendamiseks.

4. Pakkuda välja tunnused, mida oleks mudeli täpsuse parendamiseks vajalik täiendavalt koguda.

Kasulike kasutajate tuvastamiseks juhendatud õppe abil pakuti töös välja viis sammu: 1) leida üks või mitu ärilisest seisukohast relevantset kasuliku kasutaja definitsiooni, 2) valida tunnused, mis potentsiaalselt aitavad määrata andmepunkti klassi ning olema kättesaadavad mudeli rakendamise ajal, 3) jagada andmestik treening- ja testandmeteks ning vajadusel tasakaalustada treeningandmestik, 4) valida sobiv masinõppe meetod ning seda rakendada, kasutades korraga üht kasuliku kasutaja definitsiooni, 5) võrrelda loodud mudeleid ning leida kõrgeima ennustusvõimega mudel.

Välja pakutud metoodika näitlikustamiseks rakendati kirjeldatud masinõppe meetodeid idufirma Promoty kasulike kasutajate leidmiseks. Kasutades juhumetsa ja logistilise regressiooni algoritme, loodi mudel, mis hindab Promotyga mitteseotud Instagrami kasutajate tõenäosust osutuda Promoty jaoks kasulikuks kasutajaks. Sealjuures oli kasulik kasutaja defineeritud kui kasutaja, kes kutsub läbi personaalse soovituslingi Promotyga liituma vähemalt ühe uue kasutaja. Kõige kõrgema

ennustusvõimega mudeliks osutus tasakaalustamata treeningandmete põhjal loodud logistilise regressiooni mudel, mis suudab Promotyga mitteseotud Instagrami kasutajate puhul tuvastada 60% esitustäpsuse juures 23,6% kõikidest kasulikest kasutajatest.

Kuna juhumetsa ja logistilise regressiooni abil loodud mudelid pidasid oluliseks erinevaid tunnuseid, leiti kasulikke kasutajaid iseloomustavad tunnused k-means algoritmi abil. Klasteranalüüsi tulemusel leiti kaks kasutajate gruppi, kelle keskmine soovituste arv inimese kohta on märgatavalt kõrgem kui teistel klastritel, ning nimetati need vastavalt: aktiivsed soovitajad ja brändisaadikud. Sealjuures on aktiivsete soovitajate klastrisse kuuluvatel kasutajatel Instagrami keskkonnas ligikaudu 5000 jälgijat, piltidel 800 keskmiselt meeldimist ning nad on kutsunud liituma keskmiselt kaheksa uut kasutajat iga senise kasutaja kohta. Brändisaadikutel on Instagramis 10 000 jälgijat, piltidel keskmiselt 1300 meeldimist ning keskmine kutsutud uute kasutajate arv on 130. Töö autori hinnangul on need kasutajagrupid Promoty jaoks kõige väärtuslikumad, mistõttu tuleb nendega aktiivselt suhelda ning turunduskommunikatsioonis sarnaseid kasutajaid sihtida.

Ennustava mudeli ennustusvõime tõstmiseks on soovituslik koguda täiendavaid tunnuseid. Esimese sammuna soovitab töö autor lisada Promoty kasutajate andmestikule postituste arvu ning tema poolt jälgitud kasutajate arvu: nimetatud tunnused on Instagrami keskkonnas kõikidel avalikel profiilidel selgelt välja toodud, mistõttu on neid lihtsam koguda. Lisaks võib eraldi tunnusena leida kasutaja jälgijate ning jälgitud kasutajate arvu suhte, mis võib samuti klassi määramisele kaasa aidata. Samuti soovitab töö autor Promotyl uurida lähemalt klasteranalüüsis välja tulnud aktiivsete soovitajate ning brändisaadikute kasutajagruppe – mõistes, miks ja kuidas need kasutajad Promotyt soovitanud on, saab edasises turunduskommunikatsioonis just neid aspekte rõhutada.

Vaatamata sellele, et masinõppe rakendamisel turundusvaldkonnas on palju võimalusi, ei ole see töö autori hinnangul Eestis veel tavapärane praktika. Lisaks sellele, et loodud mudel aitab Promotyl välisturgudele laienedes sihtida potentsiaalseid kasulikke kasutajaid ning seeläbi efektiivsemalt kiire kasv saavutada, loodab töö autor, et koostatud magistritöö on abiks Eesti idufirmadele ning teistele soovitusprogramme rakendavatele ettevõtetele. Samuti loodab töö autor, et sellealase teadmuse tekitamine ja edasikanne avaldab positiivset mõju Eesti idufirmade kasvule ning seeläbi ka Eesti majandusele.

Viidatud Kirjandus

[1] Sorescu A. "Data-Driven Business Model Innovation". *Journal of Product Innovation Management* (2017). 34, p 691-696.

[2] Kotler P. Armstrong S. *Principles of Marketing*. Boston: Pearson, 2014. ISBN: 978-0273786993.

[3] Berman B. "Referral Marketing: Harnessing the Power of your Customers". *Business Horizons* (2016). 59, p 19-28.

[4] Hayes E. B. *TCE: Total Customer Experience. Building your Business around your Customers*. 2013.

[5] Skiera B. "Data, Data and Even More Data: Harvesting Insights From the Data Jungle". *Challenges Data Science* (2016). 8, p 11-17.

[6] Dorf B. Blank S. *The Startup Owner's Manual: The Step-by-Step Guide for Building a Great Company*. K & S Ranch Inc, 2012. ISBN: 978-0984999309.

[7] *College Students Want Targeted Social Ads*. https://www.emarketer.com/Article/College-Students-Want-Targeted-Social-Ads/1013756 (12.05.2018).

[8] Yoskovitz B. Croll A. *Lean Analytics: Use Data to Build a Better Startup Faster*. Sebastopol: O'Reilly Media, 2013. ISBN: 978-1-449-33567-0.

[9] Mezzina P. Ducange P. Pecori R. "A Glimpse on Big Data Analytics in the Framework of Marketing strategies". *Soft Computing* (2018). 22, p 325-342.

[10] Alpaydın E. *Introduction to Machine Learning*. London: The MIT Press, 2010. ISBN: 978-0-262-01243-0.

[11] Ries E. *Nutikas idufirma*. Tallinn: AS Äripäev, 2011. ISBN: 978-9949-523-18-4.

[12] Brown P. M. Ellis S. *Hacking Growth : How Today's Fastest-Growing Companies Drive Breakout Success*. New York: Crown Business, 2017. ISBN: 9780451497222.

[13] Price L. L. Feick L. F. "The Market Maven: A Diffuser of Marketplace Information". *Journal of Marketing* (1987). 51, p 83-97.

[14] *Funding, Failures & Exits of Estonian Tech Startups 2006-2017 #EstonianMafia*. http://bit.ly/estonianstartups (17.05.2018).

[15] Helm S. V. Tax S. S. Garnefeld I. Eggert A. "Growing Existing Customers' Revenue Streams Through Customer Referral Programs". *Journal of Marketing* (2013). 77, p 17-32.

[16] Fletcher H. *7 Best Uses for Predictive Analysis and Modelling in Multichannel Marketing*. http://www.targetmarketingmag.com/article/7-best-uses-predictive-analytics-modeling-multichannel-marketing/all/ (08.05.2018).

[17] Riivits-Arkonsuo I. "Consumer's Journey as Ambassador of Brand Experiences. Tarbija teekond brändielamuste saadikuna". *Tallinna Tehnikaülikooli Majandusteaduskonna doktoritöö* (2015). https://digi.lib.ttu.ee/i/?3849.

[18] *Instagram.com.* http://www.instagram.com (12.05.2018).

[19] Constine J. *Facebook Restricts APIs, Axes Old Instagram Platform Amidst Scandals.* https://techcrunch.com/2018/04/04/facebook-instagram-api-shutdown/ (18.05.2018).

[20] Benlian A. Koch O. F. "Promotional Tactics for Online Viral Marketing Campaigns: How Scarcity and Personalization Affect Seed Stage Referrals". *Journal of Interactive Marketing* (2015). 32, p 37-52.

[21] Bass F. M. "A New Product Growth Model for Consumer Durables". *Management Science* (1969). 15, p 215-227.

[22] Dumas M. "The Rise of the Estonian Startupsphere". *IT Professional Magazine* (2014). July/August, p 8-11.

[23] Ellen M. "Digitaalmeedia kasutus kasvufaasis olevate veebi- ja tarkvarapõhiste Eesti idufirmade turundustegevuses". *Tallinna Tehnikaülikooli Majandusteaduskonna bakalaureusetöö* (2016). https://digi.lib.ttu.ee/i/?6088&lang=en.

[24] Ellen M. *Soovitusprogrammi optimeerimine masinõppe abil (töös kasutatud kood).* https://github.com/marelleellen/soovitusprogrammi-optimeerimine (20.05.2018).

[25] Devi S. Murty M. N. *Introduction to Pattern Recognition and Machine Learning.* Singapur: World Scientific Co. Ple. Ltd., 2015. ISBN: 978-9814335454.

[26] *NbClust function.* https://www.rdocumentation.org/packages/NbClust/versions/3.0/topics/NbClust (20.05.2018).

[27] *Promoty.eu.* http://www.promoty.eu (16.05.2018).

[28] Aavisto R. "Nutika idufirma meetodi kasutamine Eestis". *Tallinna Tehnikaülikooli Majandusteaduskonna bakalaureusetöö* (2014). https://digi.lib.ttu.ee/i/file.php?DLID=6088&t=1.

[29] Groth R. *Data Mining: a Hands-on Approach for Business Professionals.* Harlow: Prentice Hall, Inc., 1998. ISBN: 0-13-756412-0.

[30] Holiday R. *Growth Hacker Marketing.* London: Profile Books Ltd, 2014. ISBN: 978 1 78125 436 3.

[31] Vunk R. "Growth Hacking in Estonian Startups". *Estonian Business School magistritöö* (2017). http://mi.ee/sites/default/files/blogid/raili_vunk.pdf.

[32] Sillavee S. Riistop R. *How Estonian startups rocked Estonia and the whole world in 2017*. http://startupestonia.ee/blog/how-estonian-startups-rocked-estonia-and-the-whole-world-in-2017 (17.05.2018).

[33] Gallardo L. Rodriguez-Cañamero S. García-Unanue J. San Emeterio I. C. Iglesias-Soler E. "A Prediction Model of Retention in a Spanish Fitness Centre". *Managing Sport and Leisure* (2016). 21, p 300-318.

[34] Van den Bulte C. Schmitt P. Skiera B. "Referral Programs and Customer Value". *Journal of Marketing* (2011). 75, p 56-59.

[35] *Spearman's Rank-Order Correlation*. https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php (17.05.2018).

[36] Elsner R. Walsh G. "Improving Referral Management by Quantifying Market Mavens Word of Mouth Value". *European Management Journal* (2012). 30, p 74-81.

[37] He Y. Jiang H. Wu Y. Li X. Zhao W. X. Guo Y. "We Know What You Want to Buy: A Demographic-based System for Product Recommendation On Microblogs". *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (2014). 2014, p 1935-1944.

Lisa I Kasutatud andmestikud

Tabel 7. Kõik Promoty kasutajate kohta teada olevad tunnused (8392 andmepunkti)

Parameeter	Andmetüüp	Täiendav selgitus
ID	numbriline	kasutaja unikaalne ID
Gender	kategooriline	0 - naine; 1 - mees
Age	numbriline	kasutaja vanus
City	kategooriline	0 - suurlinn; 1 - keskmise suurusega linn; 2 - väikelinn või asula
Followers	numbriline	kasutaja jälgijate arv Instagrami keskkonnas
AVG_Likes	numbriline	keskmine meeldimiste arv Instagrami keskkonnas
Engagement	numbriline	keskmine kaasatuse määr Instagrami keskkonnas
Registration_date	kuupäev	registreerumise kuupäev Promoty keskkonnas
Registration_time	kategooriline	registreerumise aeg Promoty keskonnas
isReferred	kategooriline	0 - kasutaja ei ole registreerunud soovituslingi kaudu; 1 - kasutaja on registreerunud soovituslingi kaudu
Referrals	numbriline	kasutaja kutsutud uute kasutajate arv Promotys
hasReferred1	kategooriline	0 - kasutaja soovituslingi kaudu ei ole ühtegi kasutajat Promoty keskkonnaga liitunud; 1 - kasutaja soovituslingi kaudu on Promotyga liitunud vähemalt üks uus kasutaja

Tabel 8. Juhumetsa ja logistilise regressiooni mudelites kasutatud tunnused

Parameeter	Andmetüüp	Täiendav selgitus
ID	numbriline	kasutaja unikaalne ID
Gender	kategooriline	0 - naine; 1 - mees
Age	numbriline	kasutaja vanus
City	kategooriline	0 - suurlinn; 1 - keskmise suurusega linn; 2 - väikelinn või asula
Followers	numbriline	kasutaja jälgijate arv Instagrami keskkonnas
AVG_Likes	numbriline	keskmine meeldimiste arv Instagrami keskkonnas
Engagement	numbriline	keskmine kaasatuse määr Instagrami keskkonnas
hasReferred1	kategooriline	0 - kasutaja soovituslingi kaudu ei ole ühtegi kasutajat Promoty keskkonnaga liitunud; 1 - kasutaja soovituslingi kaudu on Promotyga liitunud vähemalt üks uus kasutaja

Tabel 9. Promoty kaudu tehtud reklaampostituste andmestik (119 andmepunkti)

Parameeter	Andmetüüp	Täiendav selgitus
ID	numbriline	postituse unikaalne ID
Campaign	faktoriaalne	kampaania, mille raames postitus tehti
Likes	numbriline	postituse meeldimiste arv
Comments	numbriline	postituse kommentaaride arv
Reach	numbriline	postituse levik
Date	kuupäev	postituse kuupäev

Lisa II Klasteranalüüsi tulemused

Tabel 10. Klasteranalüüsi tulemused

	Klaster 1	Klaster 2	Klaster 3	Klaster 4	Klaster 5
Kasutajate arv	10	4861	3104	167	7
Meessoost1	4	1233	840	38	2
Naissoost0	6	3628	2264	129	5
Keskmine jälgijate arv	45 173	504	476	4933	10 687
Keskmine meeldimiste arv	3999	72	149	818	1254
Keskmine kaasatuse määr	10,65 %	16 %	32,2 %	20,2 %	13,5 %
Keskmine kutsutud kasutajate arv (aritmeetiline keskmine)	0	0,45	0,43	7,78	130
Kasutajate arv, kes on kutsunud liituma vähemalt ühe uue kasutaja	0	911	572	84	7
Kasutajate arv, kes on kutsunud liituma vähemalt kümme uut kasutajat	0	75	41	56	7

Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Marelle Ellen**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

   Soovitusprogrammi optimeerimine masinõppe abil idufirma Promoty näitel

   mille juhendaja on Dr. Anna Leontjeva

   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, 21.05.2018