TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Elis Rahulaan 182948IABM

Kliendi lahkumise ennustamine kliendikontakti põhjal

Magistritöö

Juhendaja: Ahti Lohk
PhD
Janika Aan
MSc

Tallinn 2021

Autorideklaratsioon

Kinnitan, et olen koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autor: Elis Rahulaan

10.05.2021

Annotatsioon

Klientide lahkumise ennustamine kujutab endas tuvastamist, kas klient kavatseb kliendisuhte ettevõttega lõpetada või mitte. Kliendi lahkumise ennetamine on iga ettevõtte jaoks potentsiaalne täiendav tuluallikas, sest uue kliendi leidmine on mitmeid kordi kulukam kui olemasoleva kliendi hoidmine. Sellest tulenevalt on käesoleva töö eesmärgiks leida meetod, mis aitaks vähendada klientide lahkumist, kombineerides selleks struktureeritud ja struktureerimata andmeid. Eesmärgi saavutamiseks jaguneb töö kaheks: kliendikontakti meelestatuse hindamine ja kliendi lahkumise ennustamine.

Töö esimeses osas kasutatakse masinõppel põhinevat lähenemist ning luuakse meelestatuse hindamise mudel, kus kliendikontaktid klassifitseeritakse negatiivseks või mittenegatiivseks. Potentsiaalsete teenuste lõpetajate tuvastamiseks loodi kliendi lahkumise ennustusmudel täpsusega 68%, mis tuvastab 63% lahkujatest. Töö tulemusena on võimalik tuvastada potentsiaalseid lahkujaid.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 43 leheküljel, 5 peatükki, 8 joonist, 10 tabelit.

Abstract

Customer churn prediction based on customer contact

Customer churn prediction means detecting whether a customer is going to leave the company during a given period or not. Preventing the churn represents a potential additional source of revenue for any business, as it costs much less than acquiring the new customer. Therefore, the aim of this thesis is to find a method that would help to reduce customer churn using customer contact information. Therefore, the work is divided into two parts: predicting the sentiment of a call and predicting customer churn.

In order to achieve this goal, a machine learning approach was used for sentiment analysis. 2000 calls were labeled as negative or non-negative. Text data was represented in numbers using two models: Bag of Words (BoW) and TF-IDF. In the modelling phase, models were trained using three algorithms: Naïve Bayes classifier, logistic regression, and support vector machine. It was managed to train a model with accuracy of 86% and recall of 77%, which means the model identifies 77% of the negative contacts. As the result, it is possible to identify the calls where customers express their negative thoughts about the company and its services.

In the next phase, customer churn prediction models were trained using four algorithms: Naïve Bayes classifier, logistic regression, support vector machine and CatBoost classifier. A model with accuracy of 68% and recall of 63% was trained, which means the model identifies 63% of the churners. As a result, CatBoost technique was suggested as the method which would help to reduce customer churn.

The thesis is in Estonian and contains 43 pages of text, 5 chapters, 8 figures, 10 tables.

Lühendite ja mõistete sõnastik

AUC	*Area under the ROC curve*
BoW	*Bag of Words*
CB	*CatBoost*
Kliendikontakt	Kliendi ja ettevõtte omavaheline suhtlus
LR	*Logistic regressioon*
NB	*Naïve Bayes*
NLP	*Natural language processing*
ROC	*Receiver operating characteristic*
SVM	*Support vector machines*
TF-IDF	*Term frequency-inverse document frequency*

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Klientide lahkumine on oluline teema iga ettevõtte jaoks, mis tähistab kliendisuhte lõpetamist ettevõtte ja kliendi vahel. Kliendi lahkumisega seotud kulud hõlmavad ettevõtte jaoks nii saamata jäänud tulu kui ka kulusid turundusele, mis tehakse uue kliendi hankimiseks. Uute klientide hankimine on aga keskmiselt kolm korda kulukam kui olemasoleva kliendi hoidmine [1]. Kliendi lahkumise ennetamine on oluliseks algpunktiks klientide lojaalsuse kasvatamisel. Pikaaegsed kliendid on lojaalsemad ettevõtte suhtes ning nende pealt on võimalik teenida suuremat tulu. Nad on vähem hinnatundlikumad konkurentide hindade suhtes. Sellised kliendid pole ettevõtte jaoks ka kulukad [1]. Uute klientide hankimine nõuab aga kulutusi turundusse, reklaami ja lisatööjõudu. Klientide lahkumine on väga aktuaalne probleem iga ettevõtete jaoks, kuna see mõjutab otseselt ettevõtte rahalist seisu.

Iga ettevõtte pikaajaliseks eesmärgiks on kasumi teenimine. Mida rohkem kliente lahkub, seda enam on tarvis kasumi teenimiseks pingutada. Siinjuures on oluliseks mõõdikuks on antud juhul klientide lahkumise määr ehk osakaal klientidest, kes mingil konkreetsel aja perioodil lahkusid.

Harvard Business Schooli poolt koostatud raportis on leitud, et keskmisel 5% väiksem klientide lahkumise määr, võrreldes eelneva perioodiga, põhjustab 25-95%-lise kasvu kasumis [2]. Seega sunnib klientide lahkumine sunnib ettevõtteid juurutama kliendisuhte hoidmisele suunatud sihipäraseid ja ennetavaid tegevusi ning pakkumisi. Tavapärane lähenemisviis on sihtida kliente tulenevalt nende tõenäosusest lahkuda või nende reageerimise võimest turunduskampaaniatele [3]. Kliendi lahkumise ennustusmudelites kasutatakse enamasti struktureeritud andmeid - näiteks demograafilisi (vanus, sugu, haridustase, perekonnaseis jne) ning kliendi teenuste ja tehingutega seotud informatsiooni (tellimused, laekumised jne). Struktureeritud andmed on ettevõtte jaoks üsna hõlpsasti kogutavad ning seetõttu on ka nende andmete töötlemine ettevõtte jaoks lihtne lähenemine.

Struktureeritud andmed moodustavad üldiselt ligikaudu 20% ärilisest teabest, seejuures struktureerimata andmed moodustavad 80% andmetest. Viimase alla kuuluvad igapäevasest kliendisuhtlusest saadavad andmed: nt e-kirjad, kõned ning suhtlus vestlusrobotiga jne. Need andmed on sageli korrastamata ja muudavad andmetöötluse keerulisemaks. Kuigi struktureerimata andmete töötlemine nõuab rohkem eeltööd, võib see omada siiski olulist lisainfot võrreldes struktureeritud andmetega, mis on osa igapäevasest andmetöötlusest [1].

Klientide kõned sisaldavad rohkelt struktureerimata teavet, nt vabatekstilist teavet, mida saaks kasutada ära nii ärilistel eesmärkidel kui ka kliendikogemuse parandamiseks. Sellise informatsiooni kogumine käsitsi, kasutades inimressurssi, oleks liiga aeganõudev. Oskuslikul kõneandmete töötlemisel on võimalik automaatselt kindlaks teha erinevaid murekohti ning saada tagasisidet ettevõtte teenuste kohta, mis annaks sisendi teenuste parendamiseks.

1.1 Taust ja probleem

Klientide lahkumine on suureks probleemiks paljudes tööstusharudes. Nende seas kerkib eriliselt esile telekommunikatsioonisektor, kus kliendil on kerge teenusepakkujat vahetada, kuna teenusepakkuja vahetamisega seotud kulud on madalad. Uuringu [4] kohaselt ei ole telekommunikatsiooni ettevõtted, võrreldes teiste tööstusharudega, tarbijate seas kõige populaarsemad. See ilmneb ka Kantar Emori uuringus, kus parimat teeninduskogemust pakkuvate ettevõtete edetabelis on telekommunikatsioonisektor pigem madalal positsioonil [5]. Tihti väljendatakse pahameelt telekommunikatsiooni teenusepakkuja mõne aspekti suhtes – olgu selleks arvega seotud teemad, soovimatud turunduslikud pakkumised jne [4].

Kui ettevõttel on soov kliente hoida, siis on tarvis probleemiga tegeleda ning mõista, miks kliendid lahkuvad. Kõik kliendid ei lahku samadel põhjustel. Kliendi lahkumise taga võib olla mitmeid erinevaid põhjuseid. Peamised neist on seotud toote maksumuse, toote kvaliteeti ja teeninduse kvaliteediga [6]. Seejuures võib kliendi lahkumised liigitada järgnevalt:

- tahtmatu lahkumine (ingl *involuntary churn*) – ettevõtte poolt algatatud kliendi lahkumine ehk ettevõte lõpetab kliendiga lepingu. Põhjuseks võib olla

strateegiline teenuse muutmine või kliendiga suhte lõpetamine lepingutingimuste rikkumise tõttu. Selliste klientide tuvastamine on kõige lihtsam [7];

- vabatahtlik (ingl *voluntary churn*) – neid on kõige keerulisem tuvastada. Vabatahtlik lahkumine jaguneb kaheks:

    - juhuslik (ingl *incidental churn*) – kliendi soov loobuda teenusest ja liikuda konkurendi juurde põhjustel, mis ei kuulu ettevõtte kontrolli alla. Teenuse lõpetamine võib olla seotud suuremate sotsiaalsete ja majanduslike teguritega, nt elukohavahetus, muutus finantsseisundis [7]. Selle informatsiooni alusel, mis ettevõttel on, pole võimalik juhuslikku lahkujat ette ennustada

    - tahtlik (ingl *deliberate churn*) – olukord, kus klient on rahulolematu ehk pole rahul ettevõtte teenuse või teenindusega ning otsustab kas lõpetada täielikult teenuse kasutamine või liigub konkurendi juurde. Tahtlikult lahkujad on ettevõtte jaoks suurim probleem [7].

Tahtlikult liikuvate klientide tuvastamine on keeruline protsess. Ennekõike on suureks probleemiks klientide tuvastamine, kes ei ole teenusega rahul ning kelle kavatsust ei ole alati struktureeritud andmetest näha. Seejuures annaks nende klientide tuvastamine kõige rohkem kasu. See julgustab analüüsima struktureerimata andmeid.

Kliendisuhtluse analüüsimine annab võimaluse saada tagasisidet klientidelt varem kui see on võimalik tagasiside küsitluste kaudu ning kliendi lahkumise kaudu, mis on otsene viide sellele, et klient ei olnud rahul, aga siis on juba liiga hilja reageerida. Seni ei ole struktureerimata teavet kasutatud ära piisavalt ning selle asemel on kliendianalüüsis valitud lihtsam variant, mis ei nõua suuremat andmete eeltöötlust. Kliendikontakti käigus tekkinud struktureerimata andmete töötlemine on ajamahukas ning selle laiapõhjalisem käsitlus on siiani veel üsna puudulik.

Sellise info töötlemiseks ei piisa tavalistest ärianalüütika vahenditest. Klassikaliselt kasutatakse selleks ennustusmudeleid. Kliendi lahkumise ennustusmudelite eesmärgiks on tuvastada kõige suurema tõenäosusega teenuse lõpetajad. Selline tegevus annab ettevõttele võimaluse sihtida konkreetseid kliente ning rakendada nende peal strateegiat,

mis peaks vähendama klientide lahkumist, mis tähendab ennetada finantsilist kaotust, mida klientide lahkumine tekitab.

Eelnimetatud probleemiga on varasemalt juba tegeletud. *Wonderflow* on üks näide, kus struktureerimata andmeid on ära kasutatud ärilistel eesmärkidel. *Wonderflow* on tehisintellektipõhine analüüsiplatvorm, mis võimaldab ettevõttel teha andmetest juhitud otsuseid, analüüsides erinevatest kanalitest saadavat tekstilist infot. *Wonderflow* kajastab oma juhtumiuuringus (ingl *case study*), et tänu nende analüüsiplatvormile ning oluliste tegurite välja toomisel, on kuue kuu jooksul ettevõtte klientide lahkumise määr langenud keskmiselt 12% [8].

Häid tulemusi on saadud ka struktureerimata ja struktureeritud andmete kombineerimisel. Artiklis [9] loodi kliendi lahkumist ennustav mudel, kombineerides nii struktureerimata kui ka struktureeritud andmeid. Struktureeritud andmetena kasutati kliendi demograafilisi ja tehingutega seotud andmeid. Struktureerimata andmete kaasamine parandas mudeli täpsust 5% [9]. Antud töös tuginetaksegi peamiselt artiklis [9] esitatud teadustööle.

1.2 Eesmärk

Töö peamiseks eesmärgiks leida meetod, mis aitaks vähendada klientide lahkumist, kombineerides struktureeritud ja struktureerimata andmeid. Eesmärgiks on kasutada kliendikontaktipõhist informatsiooni, milleks on tekstiks transkribeeritud kõne ning struktureeritud andmetena kõne kestvus, kõne ooteaeg ja kõne sagedus ehk mitu korda on klient pöördunud konkreetse perioodi jooksul kontaktikeskusesse. Sellest tulenevalt on eesmärgi saavutamiseks seatud kaks uurimisülesannet:

- luua **meelestatuse hindamise mudel**, mis hindaks kliendikontakti meelestatust kas negatiivseks või mittenegatiivseks, kasutades selleks transkribeeritud kõneandmeid;
- luua **kliendi lahkumise ennustusmudel** kasutades selleks kliendikontaktipõhist infot. Lisaks meelestatusele kaasatakse mudelisse konkreetsed atribuudid, mis on seotud kliendikontaktiga – kõne kestvus, kõne ooteaeg ning kõne sagedus. Viimased atribuudid on ka olulised kontaktikeskuse efektiivsuse mõõtmisel.

1.3 Töö struktuur

Eesmärgi saavutamiseks on käesolev töö jaotatud viieks peatükiks. Töö esimeses peatükis kirjeldatakse tausta ja probleemi ning töö eesmärki. Teises peatükis antakse ülevaade metoodikast, selle raames antakse ülevaade tekstikaeve protsessi erinevatest etappidest ning klassifitseerimismudelitest, mida hiljem töös kasutatakse. Kolmandas peatükis kirjeldatakse tööprotsessi – milliseid andmeid kasutati ning milliseid operatsioone läbi viidi. Neljandas peatükis esitatakse meelestatuse hindamise ning kliendi lahkumise ennustusmudeli tulemused. Viies peatükk keskendutakse töö olulistele tulemustele ja järeldustele.

2 Metoodika

Antud peatükis annab autor ülevaate tekstitöötluse protsessist ning metoodikast, mida töös jälgitakse. Peatükis tuuakse välja teksti eeltöötluse ning vektoriseerimise erinevad võimalused. Lisaks selgitatakse meelestatuse hindamise olemust ning kirjeldatakse klassifitseerimismudeleid, mida kasutatakse edasises töös nii meelestatuse hindamisel kui ka kliendi lahkumise ennustamisel.

2.1 Tööriistad

Jupyter Notebook [10] on avatud lähtekoodiga veebirakendus, mis võimaldab luua, jagada ja redigeerida dokumente, mis sisaldavad reaalajas koodi, võrrandeid, jooniseid ning teksti. Tegemist on interaktiivse märkmikuga, mis võimaldab programmeerida erinevates keeltes ning dokumenteerida kõike. Võimalik on: andmete puhastamine, andmete transformatsioon, statistiline modelleerimine, andmete visualiseerimine, masinõpe jne [10].

Oluliseks eeliseks on kasutajamugavus – kogu protsessi on võimalik väga hästi dokumenteerida nii piltide, jooniste, teksti ja koodiga. Teiseks on võimalik rakendust kasutada ilma midagi arvutisse installeerimata [11].

Python on üks enim kasutatavaid programmeerimiskeeli maailmas. Tegemist on väga mitmekülgse keelega, mis toetab mooduleid ja pakette, mis soodustavad programmi modulaarsust ja koodide taaskasutust. Samuti on tasuta kasutamiseks erinevatel platvormidel laialdane moodulite kogu. Uusim versioon 11.04.2021 seisuga Windowsile kasutamiseks on *Python* 3.9.4 [12].

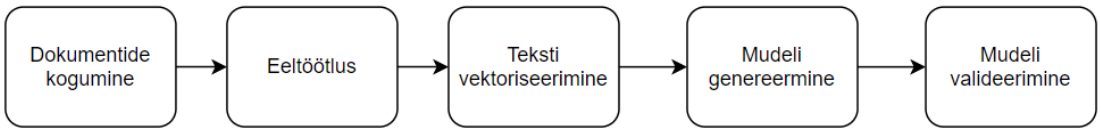
Käesoleva töö analüüs on teostatud programmeerimiskeeles *Python* kasutades *Jupyter Notebook* rakendust, mis töötab *Anaconda Navigator* abil. Töös kasutati *Python* 3.7.3 versiooni ning *Jupyter Notebook* 5.7.8 versiooni. Mudeli koostamisel kasutati *sklearn*, *pandas* ning *CatBoost* mooduleid. *Sklearn* on *Python*i programmeerimiskeele jaoks mõeldud tasuta masinõppe moodul. See sisaldab erinevaid juhendamata ja juhendatud masinõppe algoritme [13]. *Pandas* on samuti *Python*i programmeerimiskeele jaoks mõeldud moodul, mis sisaldab vahendeid andmete lugemiseks, kirjutamiseks, ümberkujundamiseks jne [14].

2.2 Tekstikaeve

Viimasel aastakümnel on andmete maht kasvanud kiires tempos infotehnoloogia arengu tõttu. Samal ajal on ka toimunud areng andmekaevandamises, on arendatud erinevaid meetodeid ja tehnikaid andmete töötlemiseks ning teabe kogumiseks. Analüüsitavaid andmeid on võimalik jagada struktureeritud ja struktureerimata andmeteks, seda nende kogumise ja hoiustamise viisi järgi. Esimese puhul on andmete kogumiseks välja töötatud süsteem ning struktureerimata andmete puhul kogutakse neid tekkimisele vastaval kujul [15].

Andmekaeve üheks valdkonnaks on tekstikaeve, mis on tänapäeva maailmas üks olulisemaid viise struktureerimata tekstiliste andmete analüüsimiseks ja töötlemiseks. Tekstikaeve abil on võimalik tuvastada erinevaid fakte, seoseid ning väiteid, mis võivad jääda tekstiliste suurandmete hulgas märkamata, samas omades kasulikku teavet. Loomuliku keele töötlus (NLP) on üks peamiseid meetodeid tekstitöötluses, mille abil teisendatakse andmed struktureeritud kujule. Struktureeritud kujul informatsiooni on juba võimalik erinevatel viisidel analüüsida.

Tekstikaeve protsessi (vt Joonis 1) esimeseks etapiks on informatsiooni hankimine ehk dokumentide kogumine. Sellele järgneb teksti eeltöötlus, kus tekstis vabanetakse üleliigsetest osadest. Seejärel viiakse tekst numbrilisele kujule ning järgnevalt on andmed analüüsiks ning mudeli genereerimiseks valmis. Mudeli loomisele järgneb selle valideerimine.



Joonis 1. Tekstikaeve protsess.

2.3 Teksti eeltöötlus

Tekstiandmete eeltöötluse käigus eemaldatakse tekstist üleliigne informatsioon. Ühed levinumad meetodid on järgmised [16]:

- tähtede muutmine väiketähtedeks [17];

- numbrite ja kirjavahemärkide eemaldamine [17];
- lemmatiseerimine ehk algvormistamine (ingl *lemmatization*) – protsess, mille käigus viiakse sõnavormid algvormi. Näiteks viiakse sõnad „lauad“ ja „laualt“ vormi „laud“ [17];
- sõnalõppude eemaldamine (ingl *stemming*) – protsess, mille käigus kaotatakse Näiteks sõnad “töötaja” ja “töötama” viiakse vormi “töö” [17];
- stopp-sõnade eemaldamine (ingl *stop word removal)* – protsess, mille käigus eemaldatakse sõnad, mida esineb tekstides sageli ning, mis ei anna teksti sisu kohta mingit infot. Näiteks asesõnad („sina“, „mina“, „me“), sidesõnad („siis“, „kui“, „ja“) ning lühendid („jne“, „jrk“, „ptk“) [18];

Sõnalõppude eemaldamise või algvormistamise puhul võib mõnel juhul kaduda tekstist oluline informatsioon, kui sõnade tähendus kontekstis muutub. Seejuures nende eemaldamisel väheneb olulisel sõnade varieeruvus [16].

2.4 Teksti vektoriseerimine

Tekstikorpuseks nimetatakse dokumentide kogumit, mida analüüsitakse. Seejuures dokument tekstikaeve mõistes viitab tekstiosale, lause või failile. Näiteks, kui analüüsitakse konkreetset kirjandusteost, siis korpuseks nimetatakse seda teost ning iga peatükk eraldi tekstifailina on üks dokument [18].

Tekstikorpus võib sisaldada tuhandeid unikaalseid sõnu, dokumendi tasandil on neid oluliselt vähem. Masinõppe algoritmid ei suuda lugeda teksti. Selleks, et teksti oleks võimalik aga analüüsida, on vaja see eelnevalt viia numbrilisele kujule ehk masinale arusaadavasse vormingusse. Sellest tulenevalt on järgnevalt välja toodud kaks viisi, kuidas teksti vektoriseerida.

2.4.1 Sõnahulgad

Sõnahulgad (ingl *bag of words*, BoW) on NLP teksti modelleerimise meetod, mis fikseerib tekstikorpuses sõna esinemissagedusi. Siinkohal ei ole oluline, mis järjekorras sõnad tekstis on, vaid see, millised sõnad tekstis üldse esinevad. Sõnad ja nende

sagedused esitatakse dokumendi termini maatriksina [17]. Näitena on välja toodud kaks dokumenti, mis kuuluvad samasse tekstikorpusesse:

- Dokument 1 (d_1) vektor: "aknalaual on kollased lilled" $= [1, 1, 1, 1, 0, 0]$;
- Dokument 2 (d_2) vektor: "köögis on ilusad lilled" $= [0, 1, 0, 1, 1, 1]$

Unikaalsed sõnad antud tekstikorpuses: aknalaual, on, kollased, lilled, köögis, ilusad.

Dokumendi termini maatriks on kujutatud tabelis 1. Maatriks on suurusega $d \times n$, kus n on dokumentide arv ning d unikaalsete terminite arv. Iga rida tähistab tekstikorpusest pärit dokumenti, veergudes on kõik dokumentide sõnad ning ridades konkreetsete sõnade sagedused igas dokumendis [17].

Tabel 1. Dokumendi termini maatriks

	aknalaual	on	kollased	lilled	köögis	ilusad
d_1	1	1	1	1	0	0
d_2	0	1	0	1	1	1

BoW algoritmi on võimalik rakendada, kasutades *Python* programmeerimiskeelt ning *sklearn* moodulist funktsiooni *CountVectorizer* [17], [13]. Funktsiooni abil teisendatakse dokument termini maatriksiks (vt Joonis 1).

2.4.2 TF-IDF

TF-IDF (ingl *Term Frequency – Inverse Document Frequency*) abil kaalutakse sõnade/terminite olulisust dokumendis. TF-IDF valem koosneb kahest komponendist, kus TF esindab termini lokaalset kaalu ning IDF globaalset kaalu. TF (ingl *Term Frequency*) mõõdab termini esinemissagedust dokumendis ja IDF (ingl *Inverse Document Frequency*) mõõdab termini olulisust dokumentide kollektsioonis. Mida väiksem on IDF väärtus, seda vähem olulisem on sõna ehk teisisõnu, mida rohkemates dokumentides sõna esineb, seda vähem informatiivsem on sõna [17], [19].

$$TF - IDF_{t,d} = TF_{t,d} \times \log \frac{N}{DF_t} \quad (1),$$

kus:

- t – termin/sõna;
- d – dokument;
- N – dokumentide arv korpuses;
- TF_t – termin t esinemissagedus dokumendis;
- DF_t – dokumentide arv, kus esineb termin t;
- $TF\text{-}IDF_{t,d}$ – sõna t TF-IDF dokumendis.

K. Uiboaed on märkinud, et kui TF-IDF väärtus on null või nullilähedane, siis võib lisada need sõnad stopp-sõnade hulka ning tekstist eemaldada [18].

TF-IDF vektoreid on võimalik genereerida *Python*i *sklearn* funktsiooniga *TfidfVectorizer* [13].

2.5 Meelestatuse hindamine

Meelestatuse analüüs on automatiseeritud protsess, mille käigus hinnatakse hoiakute, arvamuste ja emotsioonide polaarsust tekstis või kõnes loomuliku keeletöötluse kaudu. Protsessi käigus klassifitseeritakse tekst/kõne kas positiivseks, negatiivseks või neutraalseks [20].

Meelestatuse analüüsil on kaks peamist lähenemist: leksikonil ning masinõppel põhinev lähenemine. Esimese puhul kujuneb polaarsus kasutades leksikoni, mis sisaldab endas meelestatuse skooridega sõnu ja väljendeid. Leksikonis sisalduvad polaarsused väljendavad inimeste subjektiivseid tundeid ja arvamusi. Leksikaalne lähenemine on juhendamata meetod, mis ei nõua klassifitseerimiseks eelnevat andmete treenimist. Teksti klassifitseerimine positiivseks või negatiivseks tuleb võrdlusest sõnastikus olevate sõnade/väljenditega. Teksti analüüsimisel positiivseks või negatiivseks jälgitakse kumma meelestatusega sõnu on tekstis rohkem [20].

Masinõppepõhise meetodi puhul on tegemist juhendatud lähenemisega, kus andmed jaotatakse treening- ja testandmeteks. Algoritmi treenimiseks kasutatakse treeningandmeid, selle tulemusena saadakse mudel, mis ennustab väljundit vastavalt sisendile ehk infole, mis oli treeningandmestikus. Enim kasutatud masinõppealgoritmid meelestatuse ennustamiseks, mida [20], [21], [22] kui ka [23] autorid on artiklites välja, on SVM ning NB.

Artikli [20] autorid leiavad, et juhendatud masinõppe meetodid on näidanud täpsemaid tulemusi võrreldes leksikaalse lähenemisega. Samas suureks miinuseks on, et masinõppepõhised meetodid nõuavad suurel hulgal märgendatud treeningandmeid.

Varasemates artiklites kasutatud meelestatuse analüüsi kliendi lahkumise ennustamiseks. Näiteks artiklis [21] prognoositi klientide lahkumise määra sõltuvalt meelestatuse, mis sotsiaalvõrgustikus on ettevõtte suhtes. Analüüsis kasutati *Twitter*i säutsusid, et hinnata inimeste hoiakuid skaalal -2 kuni +2.

2.6 Klassifitseerimismudelid

Klassifitseerimine on teatud objekti liigitamine valikus olevasse klassi. Klassifitseerimismudelite valikul tugines autor varasematele artiklitele, kus kasutati nii meelestatuse analüüsi kui ka kliendi lahkumise ennustusmudelites järgnevaid klassifitseerimismudeleid, mis on näidanud häid tulemusi

2.6.1 Logistiline regressioon

Logistiline regressioon (ingl *logistic regression,* LR) on klassifitseerimis algoritm, millega ennustatakse uuritava sündmuse toimumist ehk Y=0 või Y=1. Logistilist regressiooni kasutatakse klassifitseerimisülesande lahendamiseks. LR modelleerib lineaarsena logaritmilist tõepärasussuhet. LR on näidanud häid tulemusi kasutamiseks teksti klassifitseerimisel [24].

Kategooriate arvu põhjal on võimalik jagada LR kolmeks:

- *Bionominal logistic regression* – uuritaval tunnusel on kaks väärtust, näiteks 1 ja 0 [25]
- *Multinominal logistic regression* – uuritaval tunnusel võib olla kolm või rohkem sõltumatut väärtust [25]
- *Ordinal logistic regression* – uuritav tunnus on kategooriline ning, näiteks väga halb, halb, hea ja väga hea. Sel juhul antakse igale kategooriale skoor, näiteks 0, 1, 2 ja 3 [25]

Logistiline regressioon on modelleeritav *Python Sklearn* funktsiooniga *LogisticRegression* [13].

2.6.2 *Naïve Bayes*

Naïve Bayes (NB) klassifikaatori puhul on tegemist tõenäosusliku klassifitseerimisalgoritmiga, mis põhineb *Bayes*i teoreemil. Klassikaline NB teoreem on valemi kujul kujutatud järgmiselt [26]:

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)} \quad (3),$$

kus:

- A ja B on kaks sündmust;
- P (A) ja P (B) on sündmuse A ja B tõenäosused;
- P (A | B) on sündmuse A tingimuslik tõenäosus, et sündmus B toimub;
- P (B | A) on sündmuse B tingimuslik tõenäosus, et sündmus A toimub.

Klassifikaatorit kasutatakse objektide jagamiseks klassidesse. NB kasutamisel tuuakse välja, et tegemist on väga lihtsa ning kiire algoritmiga. Samuti töötab algoritm hästi ka probleemidega, mis ei ole lineaarsed. Oluliseks mudeli eelduseks on, et parameetrid oleksid üksteisest sõltumatud ehk ei oleks korrelatsioonis, seda nimetatakse ka mudeli naiivseks eelduseks [27], [26].

Lisaks tavapärase klassifitseerimisülesande lahendamisel on NB näidanud praktikas väga häid tulemusi ka teksti klassifitseerimisel [27]. Näiteks on algoritm kasutusel meelestatuse analüüsis ning rämpsposti filtreerimisel. Järgneva valemiga on leitav, kas tunnus kuulub konkreetsesse klassi või mitte [26]:

$$P\left(y_j \mid x_i\right) = \frac{P(x_i \mid y_j) \times P(y_j)}{P(x_i)} \quad (4),$$

kus:

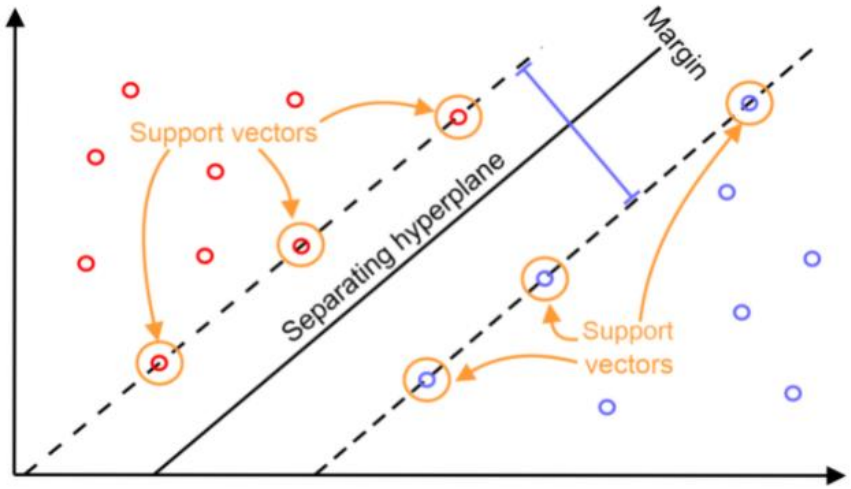
- $P\left(y_j \mid x_i\right)$ - tõenäosus, et parameetrid kuuluvad konkreetsesse klassi;
- y_j – tunnus, mis tähistab klassi kuulumist (kas 1 või 0);
- x_i – parameetrid ($x_i = x_1 + x_2 + x_3, \ldots, + x_n$);
- $P(x_i)$ – parameetri x_i esinemistõenäosus;
- $P(y_j)$ – klassi y_j esinemistõenäosus.

NB tüüpi mudeleid saab jagada kolmeks :

- *Multinomial Naïve Bayes* – sobib klassifitseerimiseks diskreetsete tunnustega. Algoritm on enim kasutusel just teksti klassifitseerimisel. *Python sklearn* moodulis on algoritm kättesaadav funktsiooniga *MultinominalNB* [13]
- *Bernoulli Naïve Bayes* – sobib klassifitseerimiseks binaarsete tunnusega: 1 või 0 (*sklearn* funktsioon *BernoulliNB* [13])
- *Gaussian Naïve Bayes* – sobib klassifitseerimiseks pidevate tunnustega ning eeldatakse, et tunnused on normaaljaotusega (*sklearn GaussianNB* funktsioon [13])

2.6.3 Tugivektormasin

Tugivektorimasin (ingl *support vector machine, SVM)* on populaarne klassifitseerimisalgoritm, mille eesmärgiks on leida optimaalne hüpertasand (ingl *hyper plane*), mis maksimeerib treeningandmete vahelist marginaali, jagades vektorruumi kaheks. SVM vajab klassifitseerimiseks märgendatud treeningandmeid, et selle pealt teha ennustus, kummale poole hüpertasandit vektor jääb [28]. Joonisel 2 on kujutatud on kujutatud kahedimensioonilist ruumi, kus sinised ruudud kujutavad üht ennustatavat klassi ning kollased ringid teist klassi.



Joonis 2. Tugivektormasin.[1]

[1]https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789345070/3/ch03lvl1sec30/svm-for-churn-prediction

2.6.4 CatBoost

CatBoost classifier (CB) on algoritm otsustuspuude gradiendi suurendamiseks, mis kasutab sümmeetrilisi puusid, mis vähendavad ennustusaega. Gradiendi suurendamine (ingl *gradient boosting*) on masinõppe meetod, mille eesmärgiks on kaofunktsiooni minimeerimine.

Mitmes artiklis [29], [30] toodi välja, et parema täpsuse saavutamiseks katsetati lisaks traditsioonilistele klassifitseerimismeetoditel *CatBoost*i, millega saadi tulemus, mis oli tugevam kui juhusliku valiku teel saadud otsus, mida loetakse heaks tulemuseks andmeteaduses kõige üldisemas kontekstis.

2.7 Klassifitseerimismudelite valideerimine

Klassifitseerimismudelite võrdlemiseks on vajalik neid hinnata ühtsetel põhimõtetel. Valideerimise käigus selgitatakse välja, milline mudel on täpseim. Järgnevalt on välja toodud erinevad võimalused mudeli täpsuse hindamiseks.

Juhul, kui prognoositakse mingi sündmuse toimumist ehk binaarset tunnust, siis on võimalik tulemused esitada 2x2 maatriksina, mida nimetatakse veamaatriksiks (ingl *confusion matrix*). Maatriks on kuvatud tabelis 2.

Tabel 2. Veamaatriks.

	Tegelik klass		
Ennustatud klass	Positiivne	Negatiivne	Kokku
Positiivne	TP	FP	TP +FP
Negatiivne	FN	TN	FN + TN
Kokku	TP + FN	FP + TN	TP + FP + FN + TN

- TP (ingl *true positive*) – tõeselt positiivne on sündmus, kus prognoositi, et sündmus toimub ning tegelikkuses ka toimus ehk nii tegelik klass kui ka ennustatud klass on positiivne [28]

- TN (ingl *false positive)* – tõeselt negatiivne on sündmus, kus prognoositi, et sündmus ei toimu ning tegelikkuses ka ei toimunud ehk nii tegelik klass kui ka ennustatud klass on negatiivne [28]

- FN (ingl *false negative)* – valenegatiivne on sündmus, kus prognoositi selle mittetoimumist, aga tegelikkuses siiski toimus ehk tegelik ning ennustatud klass ei ühtinud [28]

- FP (ingl *false positive)* - valepositiivne on sündmus, kus ennustati selle toimumist, aga sündmus siiski ei toimunud ehk tegelik ja ennustatud klass ei ühtinud [28]

Vastavalt tulemustele veamaatriksis on võimalik leida ka järgnevad mõõdikud, mille abil saab hinnata mudeli täpsust [28].

Esitustäpsus (ingl *precision*) ehk positiivne ennustusväärtus näitab, kui palju on positiivseks klassifitseeritud juhtudes tegelikult positiivsed [28]. Mõõdik on leitav valemiga (5).

$$Esitustäpsus = \frac{TP}{TP+FP} \tag{5}$$

Saagis (ingl *recall*) või teise nimega sensitiivsus (ingl *sensitivity*) ehk tõeselt positiivsete määr näitab, kui palju juhtudest ennustab mudel õigesti [28]. Määr leitakse jagades tõeselt positiivsete ennustuste arv tõeselt positiivsete ja valenegatiivsete ennustuste summaga (6):

$$Saagis = \frac{TP}{TP+FN} \tag{6}$$

Spetsiifilisus (ingl *specificity*) ehk tõeselt negatiivsete määr näitab, kui palju ennustab antud mudel uuritava sündmuse mittetoimumisest ennustab [28]. Saagis ja spetsiifilisus on seotud pöörvõrdeliselt ehk ühe väärtuse suurenedes väheneb teine ning vastupidi.

$$Spetsiifilisus = \frac{TN}{FP+TN} \tag{7}$$

Täpsus (ingl *accuracy*) näitab, mitu palju juhtudest suudab mudel õigesti klassifitseerida, arvutatakse valemiga (8) [28].

$$Täpsus = \frac{TN+TP}{TP+FP+TN+FN} \tag{8}$$

F1 skoor (ingl *F1 score*) arvestab nii saagise kui ka esitustäpsuse väärtuste erinevusi. F1 skoor on esitustäpsuse ja saagise harmooniline keskmine. F1 skoori väärtus on vahemikus 0 kuni 1. Leitav valemiga (9) [13].

$$F1\ skoor\ = 2 \times \frac{Esitustäpsus \times Saagis}{Esitustäpsus + Saagis} \quad (9)$$

ROC (ingl *receiver operating characteristic*) kõver visuaalne mõõdik kajastamaks klassifitseerimise tulemuslikkust. Graafiku horisontaalteljel kujutatakse tõeselt positiivsete määra ning vertikaalteljel valepositiivsete määra. Täiuslikku klassifikaatorit tähistab ROC-graafikul punkt (0, 1), mis klassifitseerib kõik positiivsed ja negatiivsed juhtumid õigesti [28].

Kõvera alune pindala ehk AUC (ingl *area under the ROC curve*) on samuti üks mudeli valideerimise mõõdik. AUC väärtused langevad 0 ja 1 vahele. Kui AUC väärtus on 0.5, siis ta ühtib ROC-kõvera diagonaaliga. Kui AUC väärtus on 0, siis viitab see ebatäpsusele ning väärtus 1 väga täpsele tulemusele. Vastuvõetav tulemus on vahemikus 0.7-0.8. Suuremat väärtust kui 0.8 võib nimetada juba väga heaks tulemuseks [31].

3 Tööprotsess

Käesolevas peatükis antakse ülevaade uuringu protsessi erinevatest etappidest ning mida tehes ja milliste vahenditega jõuti eesmärgini. Töö eesmärgiks on luua kliendi lahkumist ennustav mudel kombineerides struktureerimata andmed struktureeritud andmetega. Sellest tulenevalt teostati teksti eeltöötlus, vektoriseeriti tekst ehk viidi numbrilisele kujule. See järel treeniti meelestatuse hindamise mudel ning kliendi lahkumist ennustav mudel.

3.1 Uurimisobjektide analüüs

Analüüsi aluseks võeti klientide telefonitsi tehtud pöördumised kontaktikeskusesse, mis sisaldasid nii infopäringuid kui ka tehnilisi küsimusi. Antud töö raames ei töödeldud isikuandmeid, kogu alusandmestik oli pseudonümiseeritud ning ei ole otseselt kokku viidav kliendiga. Seega oli töö sisendiks mitmetest andmeallikatest kombineeritud pseudonümiseeritud andmestik. Andmestik sisaldas kahe kuu jooksul tehtud kõnesid kontaktikeskusesse, millele teenindaja vastas.

Samuti sisaldas andmestik kõnede transkriptsioone (andmeväli *text*). Kõne transkriptsioonid loodi kolmanda osapoole poolt ainult eestikeelsetele kõnedele. See piiras kogumahtu, st välja jäid kõik kõned, mis teenindati inglise ja vene keeles. Iga kõne koosneb infost, mis pärineb kahest kanalist: helistaja ja teenindaja kanalist. Mõlema kanali kasutamise eesmärgiks on asjaolu, et teenindaja võib vajadusel probleemi kliendile kõnes täpsustada.

Andmestikust jäeti kõrvale kontaktid, kus puudus seos kontakti ja kliendi vahel ehk polnud võimalik tuvastada, kas klient lõpetas teenuse või mitte kontakti järgselt kuni 1 kuu jooksul. Arvesse võeti vaid põhiteenuse lõpetamised. Samuti jäeti kõrvale need kontaktid, kus klient oli helistanud juba peale teenuse lõpetamist. Kontaktid eemaldati, kuna antud juhul nende klientide lõpetamist ei saaks enam ära hoida. Sellest tulenevalt said kõik kontaktid külge binaarse tunnuse *churn*, kus *churn* väärtus viitab sellele, et kliendi kontaktile järgnes teenuse lõpetamine ning *no_churn*, kui kontaktile ei järgnenud lõpetamine.

Andmestikust eemaldati ka kõned, mille pikkus oli lühem kui 30 sekundit või mille transkriptsiooni tähemärkide arv oli väiksem kui 250 tähemärki. See aitas välistada erinevad valeühendused, kus kõne jookseb, aga midagi ei räägita (nt teenindaja suhtleb, kuid klient ei vasta). Samuti võimaldas see lahendada osaliselt andmekvaliteediga seotud probleeme, kus kõnetranskriptsioonid sisaldavad vaid üksikuid tähemärke, ent kõne pikkused on oluliselt pikemad.

Töö sisendiks oli kaks Exceli faili, kus üks sisaldas kontakte, millele järgnes lõpetamine (*churn*) ning teine kontakte, millele ei järgnenud teenuse lõpetamist (*no churn*). Koguandmestiku mahuks jäi 33481 kontakti, millest 2554 on *churn* ja 34148 on *no churn* kontakti. Kahest andmefailist võeti juhuvalim - 2000 kontakti.

Lisaks kõne transkriptsioonile ja kliendi lahkumise tunnusele kaasati analüüsi ka järgnevad näitajad, mis on otseselt seotud kontaktiga:

- *talking_time* – arvuline tunnus, mis viitab kõne kestvusele teenindaja ja kliendi vahel

- *queue_time* – arvuline tunnus, mis näitab, kui kaua klient ootas liinil teenindajat ehk aeg kõne alustamisest kuni teenindaja vastamiseni

- *sum_calls* – arvuline tunnus, mis viitab kliendi kontaktide arvule konkreetsel perioodil. Mõõdik sisaldab ka kontakte, mis tehti vene või inglise keele ehk kui kliendi kõik kontaktid on vene või inglise keeles, siis lõplik andmestik neid ei kajasta, aga kui kliendi kontaktid on nii eesti, vene või inglise keeles, siis mõõdik arvestab ka neid kontakte. Samuti eemaldati andmestikust need kontaktid, kus kõne transkriptsioon puudus isegi eestikeelse kontakti puhul

Eelnevad näitajad on olulised mõõdikud, mille põhjal on võimalik hinnata kontaktikeskuse efektiivsust ning mis võivad oluliselt mõjutada kliendi edasisi tegevusi. Näiteks võib tekitada kliendi rahulolematust pikk ooteaeg. Näitena tabelis 4 on kujutatud juhuvalimi kirjeldav statistika meelestatuse lõikes ning on näha, et negatiivse kontakti puhul on nii kõne ooteaeg kui ka kõne pikkus keskmiselt pikemad kui mittenegatiivsel kontaktil. Juhuvalimi üldine kirjeldav statistika on kujutatud tabelis 3.

Tabel 3. Juhuvalimi kirjeldav statistika.

	talking_time	*queue_time*	*sum_calls*
count	2000.0	2000.0	2000.0
mean	358.1	190.3	6.3
std	324.5	601.9	20.6
min	30.0	1.0	1.0
25%	142.0	5.0	1.0
50%	251.0	7.0	2.0
75%	468.0	104.0	4.0
max	2449.0	8741.0	243.0

Tabel 4. Juhuvalimi kirjeldav statistika meelestatuse lõikes.

	talking_time		*queue_time*		*sum_calls*	
	mittenegatiivne	negatiivne	mittenegatiivne	negatiivne	mittenegatiivne	negatiivne
count	1251.00	749.00	1251.00	749.00	1251.00	749.00
mean	291.04	470.14	178.43	210.10	5.85	7.03
std	282.16	358.04	547.09	683.74	21.51	18.94
min	20.00	32.00	1.00	1.00	1.00	1.00
25%	115.00	225.00	5.00	5.00	1.00	2.00
50%	201.00	376.00	7.00	7.00	2.00	3.00
75%	263.50	615.00	93.50	113.00	4.00	5.00
max	1883.00	2449.00	8741.00	8434.00	243.00	243.00

Tabelis 5 ilmneb, et lahkujate ja mitte-lahkujate väga suuri erinevusi pole, vaid lahkujate keskmine ooteaeg on veidi kõrgem võrreldes mitte-lahkujatega.

Tabel 5. Juhuvalimi kirjeldav statistika (*churn/no_churn*).

	talking_time		*queue_time*		*sum_calls*	
	churn	*no_churn*	*churn*	*no_churn*	*churn*	*no_churn*

	talking_time		*queue_time*		*sum_calls*	
count	1000.00	1000.00	1000.00	1000.00	1000.00	1000.00
mean	358.50	357.74	204.13	176.46	6.23	6.35
std	334.28	314.48	592.97	610.76	13.15	25.99
min	31.00	30.00	1.00	1.00	1.00	1.00
25%	130.00	146.00	5.00	5.00	1.00	2.00
50%	235.00	266.00	7.00	7.00	2.00	2.00
75%	268.50	465.75	104.00	102.50	4.00	4.00
max	2044.00	2449.00	5635.00	8741.00	72.00	243.00

Kui võrrelda omavahel (Vt Tabel 6) mittenegatiivse ning negatiivse meelestatuse kontakte mitte-lahkujate ja lahkujat lõikes, siis ilmneb, et kõige kõrgem keskmine kõne pikkus on negatiivse kontakti puhul ning olukorras, kus klient ei lahku. Samuti on ka keskmine ooteaeg sel juhul kõige kõrgem. Seejuures on kõne sagedus kõige kõrgem lahkujate puhul, kellel on olnud negatiivne kontakt.

Tabel 6. Keskväärtused churn/no churn ning meelestatuse lõikes

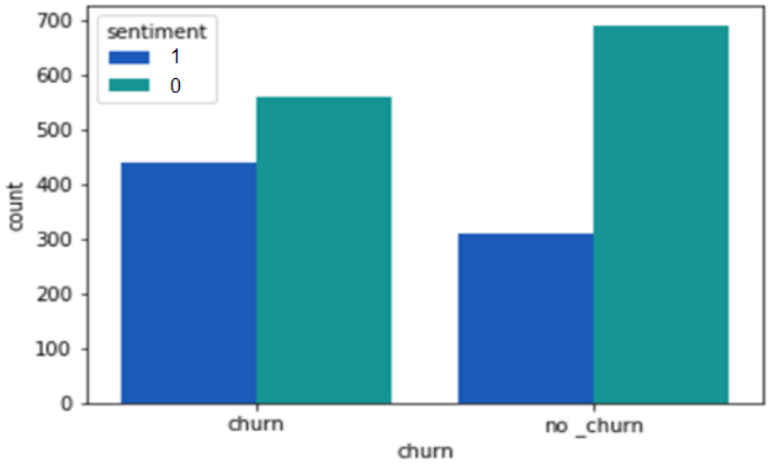
		talking_time	*queue_time*	*sum_calls*
churn	negatiivne	447.68	183.01	7.75
no churn	negatiivne	502.13	248.67	6.00
churn	mittenegatiivne	228.43	220.72	5.04
no churn	mittenegatiivne	293.17	144.17	6.50

Tabelis 7 on kujutatud mittenegatiivsete ja negatiivsete ning *churn* ja *no churn* kontaktide jaotus valimis. 2000-st kontaktist 62.6% moodustasid mittenegatiivsed kontaktid ning 37.5% negatiivsed. Mõlemal juhul, nii *churn* kui ka *no churn* kontaktide puhul, olid ülekaalus mittenegatiivsed kontaktid. Samas negatiivsete kontaktide osakaal *churn* (58.7%) puhul oli kõrgem kui *no churn* (41.3%) puhul. Joonisel 3 on kujutatud kliendi lahkumise ning meelestatuse suhe graafiliselt.

Tabel 7. Kliendi lahkumise tunnuse ja meelestatuse suhe.

	mittenegatiivne		negatiivne		**kokku**
churn	560	44.8%	440	58.7%	**1000**
no churn	691	55.2%	309	41.3%	**1000**
kokku	**1251**	**100.0%**	**749**	**100.0%**	**2000**

Muutujate mõju ning nende vaheliste seoste hindamise eesmärgil leiti šansside suhe (ingl *odds ratio*, OR) [32], millest tulenes, et negatiivse meelestatusega kliendikontakti puhul on šanss kliendil lahkuda 1.76 korda kõrgem kui mittene kliendikontakti puhul ehk tõenäosus on 64% suurem. Samuti leiti, et seos on statistiliselt oluline, kuna OR väärtus jäi usaldusvahemikku.



Joonis 3. Kliendi lahkumise tunnuse ja meelestatuse suhe.

3.2 Teksti eeltöötlus

Teksti eeltöötluse eesmärgiks oli tekstist eemaldada kõik ebaoluline. Tekstitöötluse raames muudeti kõik tähed kõnede transkriptsioonis väiketähtedeks. Samuti eemaldati kõik numbrid nii numbrilisel (1, 2, 3 jne) kui ka sõnalisel kujul (üks, kaks, kolm jne).

Sõnahulga vähendamiseks eemaldati tekstikorpusest mitte-sisulised sõnad ehk stopp-sõnad. Stopp-sõnade eemaldamisel tugineti K. Uiboaed avalikule sõnaloendile, mis on kättesaadav keskkonnas *Github* [33]. Kuna käesoleva töös on käsitletud lemmatiseerimata teksti, siis eemaldati sõnad, mis olid järgnevates failides: "adps.csv",

"advs-etc.csv", "konj.csv", interj.csv ning "estonian-stopwords.txt" [33]. Kokku sisaldasid eelnevad failid 6610 sõna.

Lisaks eemaldati autori poolt tekstist veel hulk sõnu, mis olid antud kontekstis sisutühjad. Näiteks jäeti välja ettevõte nimi erinevates vormides ning teenindajate nimed, kuna need kordusid liiga tihedalt. Kokku loodi loend 196 sõnast.

3.3 Teksti märgendamine

Kliendikontakti meelestatuse hindamisel kasutatakse masinõppel põhinevat meetodit, seda just ettevõtte spetsiifikast ning kasutatavast sõnavarast tulenevalt. Sellest tulenevalt märgendati kontaktid. Kõnede märgendamisega tegeles töö autor. Kokku märgendati 2000 kontakti: 1000 *churn* ja 1000 *no churn* kontakti. Märgendamise eesmärgiks oli hinnata kliendikontakti meelelestatust kas negatiivseks või mittenegatiivseks ning tuvastada kliendikontaktid, millele võiks järgneda tahtlik teenuse lõpetamine.

Negatiivne meelestatus peaks viitama negatiivsele tagasisidele ettevõtte brändi või teenuste suhtes. Näiteks pöördutakse murega seoses teenuse toimimisega. Antud hinnang ei ole seotud kontakti lõpliku meelestatusega ehk kui klient andis negatiivset tagasisidet ning tal teenus ei tööta, ent sai sellele lahenduse või vastuse kontakti jooksul, ei tohiks muuta kontakti meelestatust.

Mittenegatiivse meelestatuse alla kuuluvad ka kontaktid, kus klient pöördub sooviga lahkuda näiteks seoses kolimisega. Seda võib arvestada juhuslikuks kliendi lahkumiseks ning antud juhul pole eesmärgiks selliseid kontakte ennustada. Oluline on just tagasiside probleemide ja tõrgete kohta. Kõik kontaktid, mida ei saa klassifitseerida negatiivseks on mittenegatiivsed.

3.4 Modelleerimine

Meelestatuse hindamisel eemaldati kõik meelestatuse analüüsi jaoks ebavajalik info. Jäeti alles väljad *text* ning *sentiment*. Meelestatuse hindamisel kasutati masinõppepõhist lähenemist. Mudeli loomisel kasutati märgendatud andmestikku. Kogu märgendatud andmestik jagati kaheks osaks: treeninghulk (ingl *training set*) ja valideerimishulk (ingl *validation set*). Mudel loodi kasutades treeningandmeid. Seejärel valideeriti see ülejäänud märgendatud andmete ehk valideerimishulga peal. Teksti vektoriseerimisel kasutati kaht

meetodit (BoW ja TF-IDF), mis toodi ka välja peatükis 2. Meelestatuse analüüsis kasutati kolme algoritmi:

- *Naïve Bayes* (NB)
- Tugivektormasin (SVM)
- Logistiline regressioon (LR)

Kliendi lahkumise ennustamine on kahe klassiga klassifitseerimisülesanne. Kliendi lahkumise ennustusmudeli loomisel kasutati kolme peamist algoritmi, mis olid enim mainitud varasemates teadusartiklitest. Lisaks kasutati algoritmi *CatBoost*, et paremat tulemust saada:

- *Naïve Bayes* (NB)
- Tugivektormasin (SVM)
- Logistiline regressioon (LR)
- *CatBoost* (CB)

Tabelis 8 on kujutatud väljad, mida kasutati mudeli loomisel (*row_id* välistati). *Sentiment* 0 väärtus viitab mittenegatiivsele meelestatusele ning 1 negatiivsele meelestatusele. Samuti *churned* näitaja puhul kujutab 0 *no_churn* ehk kliendi mitte lahkumist ning 1 *churn* ehk kliendi lahkumist.

Tabel 8. Kliendi lahkumise ennustusmudelis kasutatud väljad.

row_id	*talking_time*	*queue_time*	*sum_calls*	*sentiment*	*churned*
1	655	121	1	1	1
2	221	4	17	0	0
3	319	8	4	1	0

4 Peamised tulemused

Ennustusmudelite koostamisel kasutati peatükis 2.6 välja toodud algoritme, mis on hästi toiminud ka varasemates töödes nii meelestatuse hindamisel kui ka kliendi lahkumise ennustamisel. Ennustusmudeleid valideeriti alapeatükis 2.7 esitatud valideerimismeetoditega.

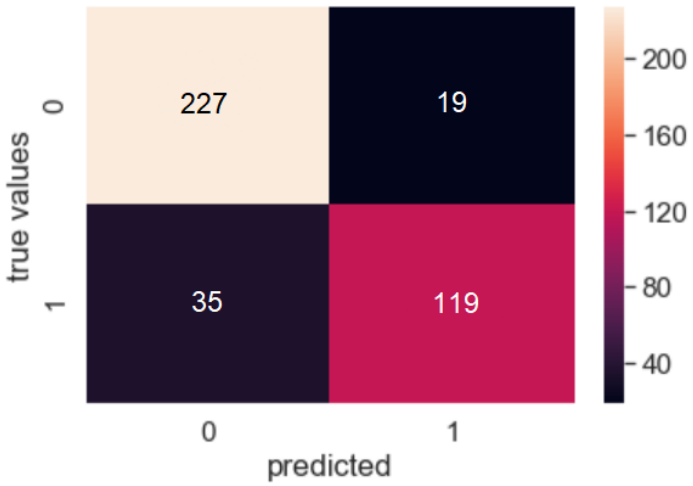
4.1 Meelestatuse hindamise ennustusmudel

Tabelis 9 esitab meelestatuse hindamise mudeli tulemused. Kõige parem tulemus saadi kasutades BoW mudelit ning NB klassifitseerijat, kus mudeli täpsuseks saadi 87%, esitustäpsuseks 86%, saagiseks 77%, F1 skooriks 78% ning AUC 0.85. BoW + SVM kombinatsiooni puhul oli täpsus kõige madalam (78%). Ülejäänud algoritmide täpsused jäid vahemikku 81-83%.

Tabel 9. Meelestatuse hindamise mudeli valideerimine.

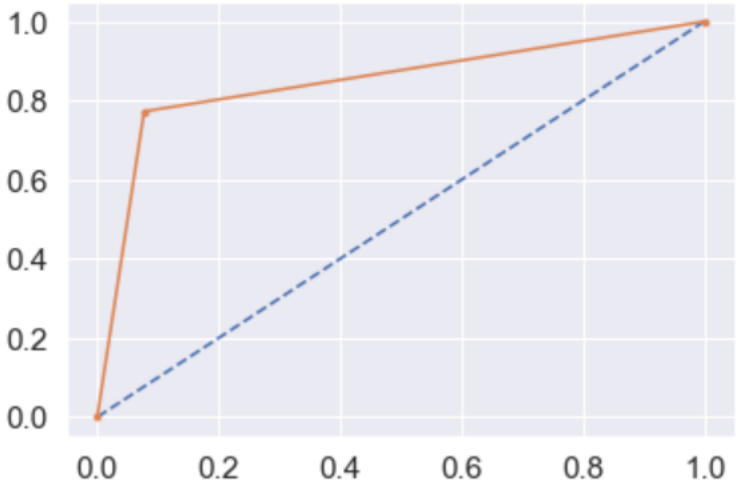
	Accuracy	*Precision*	*Recall*	*F1-score*	*AUC*
BoW + NB	**0.87**	0.86	**0.77**	**0.78**	**0.85**
BoW + LR	0.83	0.84	0.69	0.76	0.80
BoW + SVM	0.78	0.72	0.67	0.69	0.76
TF-IDF + NB	0.81	0.81	0.60	0.69	0.76
TF-IDF + LR	0.83	**0.90**	0.62	0.73	0.79
TF-IDF + SVM	0.81	0.78	0.67	0.72	0.78

BoW + NB kombinatsiooni puhul klassifitseeritud õigesti 227 + 119 ehk 346 juhtumit. Mudel klassifitseeris valesti 54 juhul (vt Joonis 4). Konkreetse algoritmi puhul ennustab mudel kõikidest negatiivsetest kontaktidest negatiivseks 77% ning 86% nendest kontaktidest, mis klassifitseeriti negatiivseks olid ka tegelikkuses negatiivsed ehk vaid 14% klassifitseeriti mittenegatiivseks, kuigi tegemist oli negatiivse kontaktiga.



Joonis 4. Veamaatriks (BoW + NB).

Joonisel 5 on kujutatud ROC-kõverat (BoW + NB), kus AUC ehk kõvera alune pindala on 0.84, mida võib vastavalt [31] pidada väga heaks tulemuseks.



Joonis 5. ROC-kõver (BoW + NB).

4.2 Kliendi lahkumise ennustusmudel

Järgnevalt on välja toodud kliendi lahkumise ennustusmudeli tulemused. Täpsuse põhjal sai parima tulemuse *CatBoost* algoritmiga genereeritud mudel (68%). Samuti oli *CatBoost* algoritmiga kõige kõrgem esitustäpsus, F1-skoor ja AUC näitajad (vt Tabel 10).

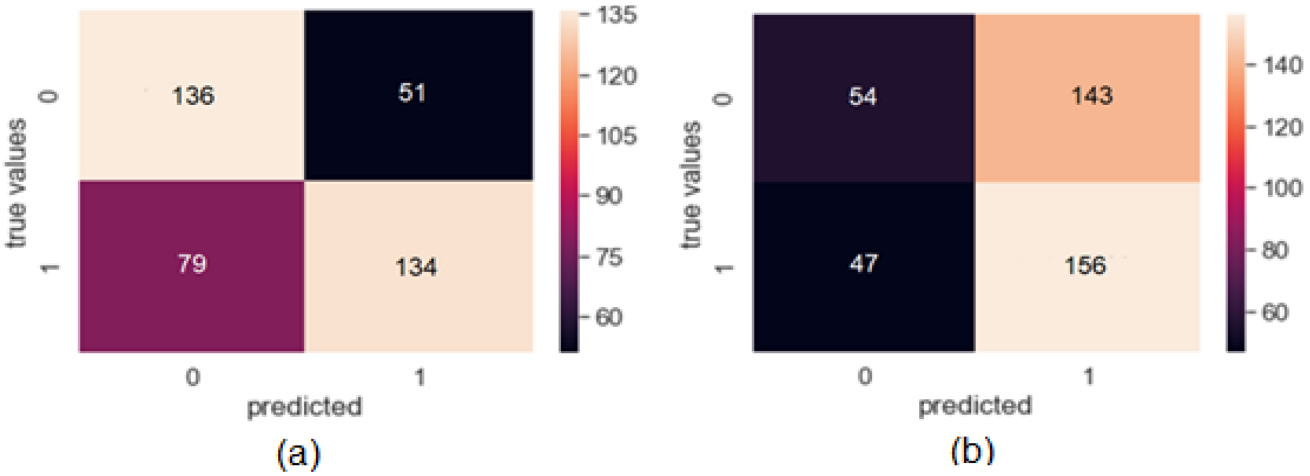
Tabel 10. Kliendi lahkumise ennustusmudeli valideerimine.

	Accuracy	*Precision*	*Recall*	*F1-score*	*AUC*
NB	0.55	0.62	0.45	0.52	0.56
LR	0.58	0.54	0.46	0.50	0.57
SVM	0.53	0.52	**0.77**	0.62	0.52
Catboost	**0.68**	**0.72**	0.63	**0.67**	**0.68**

Saagis näitab antud kontekstis, mitu protsenti kõikidest lahkujatest (*churn*) ennustab mudel lahkujaks. Kasutades *CatBoost* algoritmi ennustab mudel lahkujaks 63%. Sel juhul jääb 37% lahkujatest mudelil tuvastamata ehk need klassifitseeritakse mitte-lahkujaks (*no churn*). Esitustäpsus näitab, mitu protsenti neist, kes klassifitseeriti lahkujateks, tegelikult lahkusid. *CatBoost* algoritmi kasutades, oli selleks väärtuseks 72% ehk 28% klientidest olid tegelikult mitte-lahkujad.

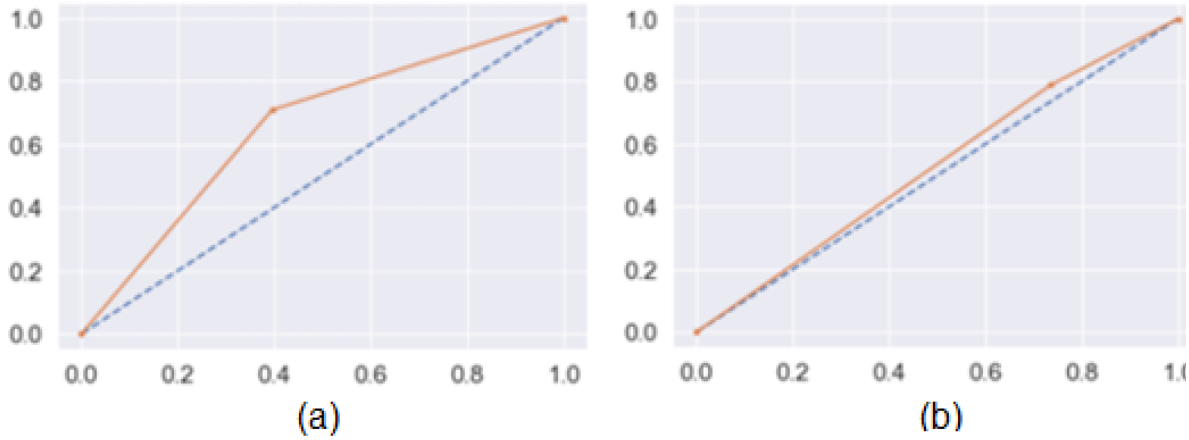
Joonisel 6 (a) on kujutatud veamaatriks kasutades *CatBoost* algoritmi, kus õigesti klassifitseeritud 136+134 ehk 270 juhtumit ning valesti 130 juhtumit. 185 klienti klassifitseeriti lahkujaks ja 215 klienti mitte-lahkujaks.Vastavalt saagisele toimisid kõige kehvemini NB ja LR algoritmid, mudelid jätsid vastavalt tuvastamata 55% ja 56% lahkujatest. NB esitustäpsus oli 62%, mis järeldub, et 38% klientidest, kes tegelikult olid mitte-lahkujad klassifitseeriti lahkujateks.

Seejuures saagis oli kõige kõrgem kasutades SVM algoritmi (77%) ehk mudel jättis ainult 23% lahkujatest tuvastamata. Seejuures esitustäpsus SVM algoritmiga oli 52%, mis tähendab, et 48% klienditest olid tegelikult mitte-lahkujad. SVM veamaatriks on kujutatud joonisel 6 (b).



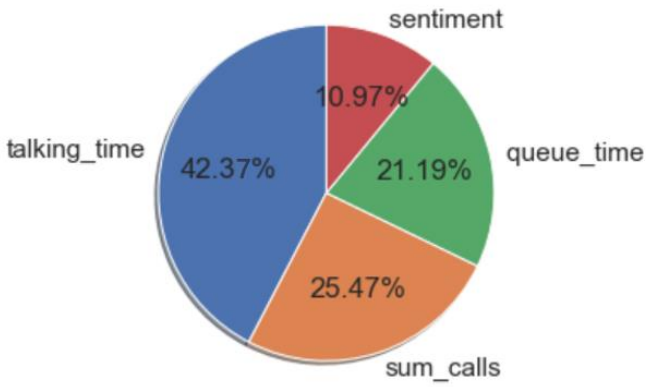
Joonis 6. Veamaatriksid – *CatBoost* (a) ja SVM (b).

Joonisel 7 on kujutatud kaks ROC-kõverat. Vasakul on *CatBoost* mudeli kõver, mille alumise pindala ehk AUC väärtuseks on 0.68, mis vastavalt [31] on aktsepteeritav tulemus. Paremal on SVM mudeli kõver, mille AUC väärtus on 0.52.



Joonis 7. ROC-kõverad - *CatBoost* (a) ja SVM (b).

Joonisel 8 on kujutatud tunnuste osatähtsused loodud ennustusmudelis, kasutades *CatBoost* algoritmi. Kõige suurema mõjuga on *talking_time* (42.36%), millele järgnevad *sum_calls* (25.47%), *queue_time* (21.19%) ning *sentiment* (10.97%). Väärtused on ümardatud kahe komakohani.



Joonis 8. Tunnuste osatähtsus mudelis %

5 Analüüs ja järeldused

Töö aluseks oli 2000 kontaktiga juhuvalim kontaktikeskuse kõneandmetest. Kõik kliendikontaktid märgendati autori poolt skaalal negatiivne ning mittenegatiivne. Töös kasutati meelestatuse hindamisel masinõppel põhinevat lähenemist, treeniti NB-põhine klassifitseerimismudel, mille täpsuseks saadi 86%. Antud mudeliga on võimalik klassifitseerida kontaktikeskuse kõnesid skaalal negatiivne ning mittenegatiivne. Mudel suudab tuvastada 77% kõikidest negatiivsetest kontaktidest.

Samuti loodi kliendi lahkumise ennustusmudel *CatBoost* algoritmiga, mille ennustustäpsus oli 68%, seejuures saagise väärtus oli 63% ehk mudeliga tuvastatakse 63% lahkujatest ning tuvastamata jääb 37% lahkujatest. Seejuures SVM algoritmiga tuvastatakse 77% lahkujatest ning tuvastamata jääb 23%. Seejuures *CatBoost* puhul klassifitseeriti 72% lahkujateks, millest järeldub, et 27% oli tegelikult mitte-lahkujad. Samuti ei ole mitte-lahkujate klassifitseerimine antud kontekstis kõige halvem variant, kuna kliendi proaktiivne kõnetamine isegi mitte-lahkuja puhul võib kasvatada kliendi lojaalsust. SVM puhul klassifitseeriti lahkujateks 52%, kes ka tegelikult lahkusid, mis tähendab, et 48% neist klientidest olid mitte-lahkujad.

Kuigi tugivektormasin algoritmiga loodud mudel suudab tuvastada 77% kõikidest lahkujatest, siis ülejäänud mudeli valideerimisnäitajad olid paremad *CatBoost* algoritmiga ning seega võib öelda, et lahkuja tuvastamisel toimib kõige paremini *CatBoost* algoritm. Seda peamiselt seetõttu, kuna eesmärgiks on luua stabiilne mudel, mis töötaks ka suuremate andmemahtude peal.

CatBoost algoritmiga loodud mudelis mängis kõige olulisemat rolli kõne pikkus, mille osatähtsus mudelis oli ligikaudu 42%. Meelestatuse osatähtsus oli mudelis ligikaudu 11%. Sellest järeldub, et kliendi lahkumisel on oluliseks mõjutajaks ka kõne pikkus. Mis tõttu saame öelda, et kliendi lahkumine ei sõltu vaid tema kõnes väljendatud meelestatusest.

Oluliseks piiranguks on antud juhul asjaolu, et tuvastada on võimalik ainult neid lahkujaid, kelle kõned olid eestikeelsed. Tuvastamata jäävad kõik kontaktid, kus suhtlus käis inglise, vene keeles või puudus kõne transkriptsiooni sisu. Samuti pole võimalik tuvastada kliente, kus andmekvaliteedi tõttu pole võimalik siduda klienti ja kõnet. Lisaks

katab mudel ainult kindlat osa ettevõtte kogu kliendibaasist ehk arvestatakse neid kliente, kes pöörduvad kontaktikeskusesse.

Töö tulemustest lähtuvalt pakub autor välja järgnevad soovitused:

- töövahend klienditeeninduses - meelestatuse hindamise mudeli abil on võimalik tuvastada negatiivsed kliendikontaktid. Üheks variandiks oleks kasutada saadud infot klienditeeninduses, kus oleks võimalik näha, kas sissetuleva kliendi eelmine kontakt oli negatiivne või mittenegatiivne. Kuigi eelnevate kõnede kohta võib olla registreeritud sündmus (mitte alati), siis lihtne mittenegatiivne/negatiivne meelestatus annaks teenindajale kiire ülevaate;

- proaktiivne sekkumine – klientide lahkumise ennustusmudel annab võimaluse leida potentsiaalsed lahkujad kliendikontakti kaudu. Sellest tulenevalt on ettevõttel võimalus sekkuda proaktiivselt ning luua tegevusplaan lahkumise ennetamiseks, näiteks proaktiivsed kõned ning pakkumised võimalikele lahkujatele.

Kokkuvõte

Käesolevas töös on käsitletud klientide lahkumist kui ettevõtte jaoks olulist probleemi, mis on otseselt seotud ettevõtte tulemuslikkusega. Kliendikontakt on üks osa klienditeekonnast, mis omab olulist infot, mida ei kasutata ära piisavalt. Kliendikontakti meelestatus võib viidata kliendi edasistele tegevustele. Töös kasutatud valimi põhjal negatiivse meelestatusega kontakti alusel on kliendil 64% suurem tõenäosus lahkuda kui mittenegatiivse kontakti puhul. Sellest tulenevalt oli töö eesmärgiks leida meetod, mis aitaks vähendada klientide lahkumist. Seega jagunes töö kaheks: kliendikontakti meelestatuse hindamine ja kliendi lahkumise ennustamine.

Töö käigus märgendati autori poolt 2000 kontakti, kasutades kontaktikeskuse kõneandmeid, kas mittenegatiivseks või negatiivseks, mille eesmärgiks oli tuvastada kliente, kus teenuse kasutamisega esineb tõrkeid või kliendil on negatiivne hinnang ettevõttele. Kõne transkriptsioonid viidi numbrilisele kujule kasutades kaht mudelit: BoW ja TF-IDF. Mudelid treeniti kasutades kolme algoritmi: *Naïve Bayes*, logistiline regressioon ja tugivektormasin. Parima täpsusega meelestatuse hindamise mudel saadi kombineerides BoW mudelit ning *Naïve Bayes* algoritmi. Saadi mudel täpsusega 86%, mis suudab tuvastada 77% negatiivsetest kliendikontaktidest.

Teiseks loodi kliendi lahkumise ennustusmudel. Mudeli treenimiseks kasutati nelja algoritmi: *Naïve Bayes*, logistiline regressioon, tugivektormasin ja *CatBoost*. Kasutades *CatBoost* algoritmiti saadi mudel täpsusega 68%. Mudelite valideerimisel osutus kõige olulisemaks punktiks valideerimisnäitajate stabiilsus.

Oluliseks töö panuseks oli meelestatuse hindamise mudel, mis loodi kombineerides struktureeritud andmeid struktureerimata andmetega, mille abil saab tuvastada negatiivseid kliendikontakte. Samuti on võimalik kasutatud meelestatuse infot kliendi lahkumise ennustusmudelis. Töö tulemusena loodi meetod, mis aitab vähendada klientide lahkumist tuvastades kontaktid, mille järgselt võiks klient lahkuda. Antud teadmine annab ettevõttele võimaluse reageerida varem ning planeerida proaktiivseid tegevusi klientide hoidmiseks.

Kasutatud kirjandus

[1] A. M. Almana, M. S. Aksoy ja A. Rasheed, „A survey on data mining techniques in customer churn analysis for telecom industry,“ *Journal of Engineering Research and Applications,* kd. 4, nr 5, pp. 165-171, 2014.

[2] F. F. Reichheld ja P. Schefter, „The Economics of E-loyalty,“ *Harvard Business School Working Knowledge, 10,* 2000.

[3] A. Lemmens ja S. Gupta, „Managing Churn to Maximize Profits,“ *Marketing Science,* kd. 39, nr 5, pp. 956-973, 2020.

[4] The European Business Review, „How Costly Is Customer Churn in the Telecom Industry?,“ 18 August 2020. [Võrgumaterjal]. Available: https://www.europeanbusinessreview.com/how-costly-is-customer-churn-in-the-telecom-industry/. [Kasutatud 4 May 2021].

[5] Kantar Emor, „Parimat teeninduskogemust pakub tänavu Smartpost Itella,“ 24 March 2021. [Võrgumaterjal]. Available: https://www.kantaremor.ee/pressiteated/parimat-teeninduskogemust-pakub-tanavu-smartpost-itella/. [Kasutatud 10 May 2021].

[6] M. B. Joolfoo, R. A. Jugurnauth ja K. M. Joolfoo, „Customer Churn Prediction in Telecom Using Machine Learning in Big Data Platform,“ *Journal of Critical Reviews,* kd. 7, nr 11, pp. 1991-2001, 2020.

[7] E. Shaaban, Y. Helmy, A. Khedr ja M. Nasr, „A proposed churn prediction model,“ *International Journal of Engineering Research and Applications,* kd. 2, nr 4, pp. 693-697, 2012.

[8] Wonderflow, „How AI and VOC analytics helped a global B2C service provider to reduce churn,“ 2021.

[9] N. N. Vo, S. Liu, J. Brownlow, C. Chu, B. Culbert ja G. Xu, „Client Churn Prediction with Call Log Analysis,“ *In International Conference on Database Systems for Advanced Applications,* pp. 752-763, 2018.

[10] „Project Jupyter,“ 2021. [Võrgumaterjal]. Available: https://jupyter.org/. [Kasutatud 11 April 2021].

[11] V. Kiisk, „PYTHON JA JUPYTER NOTEBOOK,“ 2020. [Võrgumaterjal]. Available: http://kodu.ut.ee/~kiisk/python.html. [Kasutatud 11 April 2021].

[12] „The Python Software Foundation,“ 2021. [Võrgumaterjal]. Available: https://www.python.org/doc/essays/blurb/. [Kasutatud 11 April 2021].

[13] Pedregosa et al, „Scikit-learn: Machine Learning in Python,“ *Journal of Machine Learning Research 12,* pp. 2825-2830, 2011.

[14] W. &. o. McKinney, „Data structures for statistical computing in python,“ *In Proceedings of the 9th Python in Science Conference,* kd. 445, pp. 51-56, 2010.

[15] J. Puusalu, Suurandmed: olemus ja kasutamise kitsaskohad, 2020.

[16] U. Kamath, J. Liu ja J. Whitaker, Deep learning for NLP and speech recognition, Cham: Springer, 2019.

[17] J. Whitaker, K. Uday ja J. Liu, Deep learning for NLP and speech recognition, Springer, 2019.

[18] K. Uiboaed, „Tekstikaevega seotud termineid,“ [Võrgumaterjal]. Available: https://kristel.gitbooks.io/sissejuhatus-tekstikaevesse/content/tekstikaeve-terminid.html.

[19] C. C. Aggarwal ja Z. ChengXiang, Mining text data, Springer, 2012.

[20] S. Vohra ja J. Teraita, „A comparative study of sentiment analysis techniques,“ *Journal Jikrce,* kd. 2, nr 2, pp. 313-317, 2013.

[21] S. Ranjan ja S. Sood, „Sentiment Analysis Based Telecom Churn Prediction,“ *Journal of Web Engineering & Technology,* kd. 7, nr 1, pp. 6-12, 2020.

[22] O. Shepelenko, „Opinion Mining and Sentiment Analysis using Bayesian and Neural Networks,“ Master thesis, University of Tartu, Institute of Computer Science, 2017.

[23] H. Pajupuu, R. Altrov ja J. Pajupuu, „Identifying Polarity in Different Text Types,“ *Folklore: Electronic Journal of Folklore,* kd. 64, pp. 125-142, 2016.

[24] I. Konstantinos, K. I. Diamantaras, G. Sarigiannidis ja K. C. Chatzisavvas, „A Comparison of Machine Learning Techniques for Customer Churn Prediction,“ *Simulation Modelling Practice and Theory,* kd. 55, pp. 1-9, 2015.

[25] D. Barron, „Logistic Regression: Binary, ordinal and multinominal,“ University of Oxford, 2018. [Võrgumaterjal]. Available: http://users.ox.ac.uk/~jesu0073/Lecture%203/LogisticRegression.pdf. [Kasutatud 24 April 2021].

[26] D. Berrar, „Bayes’ theorem and naive Bayes classifier,“ %1 *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics,* Amsterdam, Elsevier Science Publisher, 2018, pp. 403-412.

[27] S. L. Ting, W. H. Ip ja A. H. Tsang, „Is Naïve Bayes a Good Classifier for Document Classification?,“ *International Journal of Software Engineering and Its Applications,* kd. 5, nr 3, pp. 37-46, 2011.

[28] K. B. Subramanya ja S. Arun, „Enhanced feature mining and classifier models to predict customer churn for an E-retailer,“ %1 *7th International Conference on Cloud Computing, Data Science & Engineering-Confluence,* 2017.

[29] P. Lalwani, M. K. Mishra, P. Sethi ja J. S. Chadha, „Customer churn prediction system: a machine learning approach.,“ *Computing,* pp. 1-24, 2021.

[30] Y. Deng, D. Li, L. Yang ja J. Zhao, „Analysis and prediction of bank user churn based on ensemble learning algorithm,“ *IEEE International Conference on Power Electronics, Computer Applications (ICPECA),* pp. 288-291, 2021.

[31] J. N. Mandrekar, „Receiver operating characteristic curve in diagnostic test assessment,“ *Journal of Thoracic Oncology,* kd. 5, nr 9, pp. 1315-1316, 2010.

[32] M. Sõmer ja I. Seppo, „Regressioon,“ Eesti Rakendusuuringute Keskus CentAR, 2018.

[33] K. Uiboaed, „estonian-stopwords,“ GitHub repository, 2018. [Võrgumaterjal]. Available: https://github.com/kristel-/estonian-stopwords. [Kasutatud 3 May 2021].

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

Mina, Elis Rahulaan

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Kliendi lahkumise ennustamine kliendikontakti põhjal“, mille juhendajad on Ahti Lohk ja Janika Aan
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Olen teadlik, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

10.05.2021

1 Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.