

TARTU ÜLIKOOL
Arvutiteaduse instituut
Andmeteaduse õppekava

Desiree Himuškin

**Andmenihke tuvastamine ja leevendamine
kõnekeskuse andmete näitel**

Magistritöö (15 EAP)

Juhendaja(d): Janika Aan,
Anna Aljanaki

Tartu 2025

Andmenihke tuvastamine ja leevendamine kõnekeskuse andmete näitel

Lühikokkuvõte:

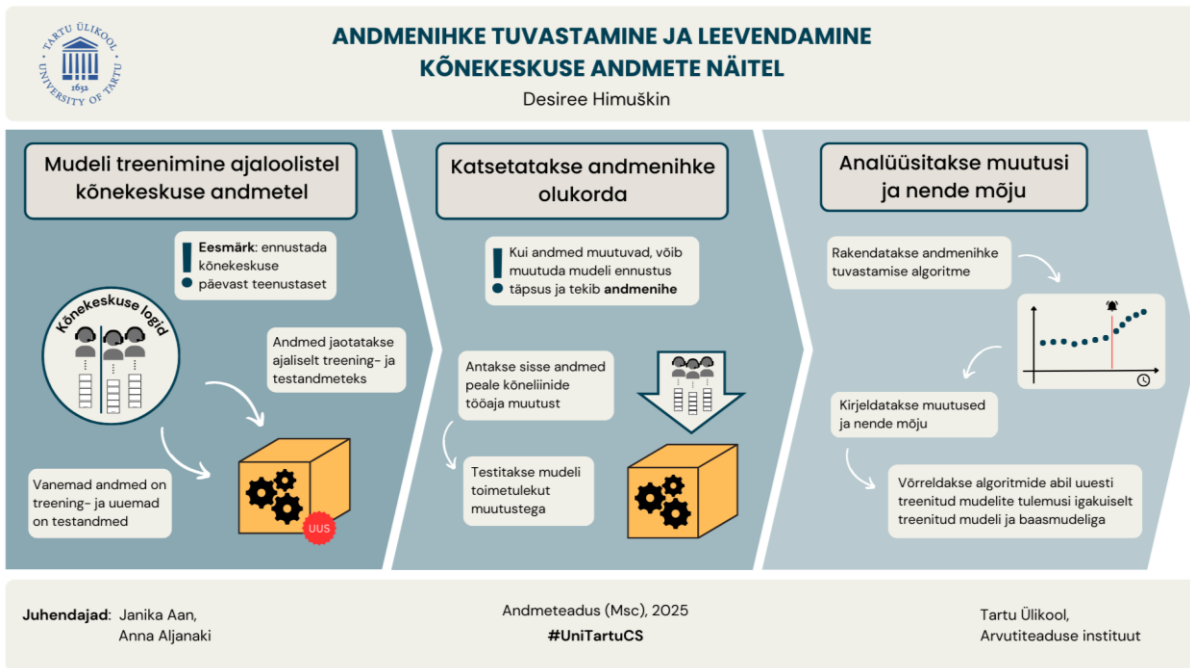
Käesoleva magistritöö eesmärk on analüüsida enda loodud mudelitel, kuidas mõjutab kõnekeskuse tööaja muutus mudeli teenustaseme ennustustäpsust, kas see toob endaga kaasa andmenihke ning kuidas vähendada selle mõju. Esmalt antakse ülevaade kirjandusest, senistest lähenemistest kõnekeskuse teenustaseme ennustamiseks ja andmenihke algoritmidest. Seejärel seletatakse lahti ja analüüsitakse kõnekeskuse andmestiku üldiselt. Kolmandana, antakse ülevaade kasutatud mudelitest ja nende tunnustest, millele järgneb praktiline osa, kus kirjeldatakse muutuse mõju tunnustele, katsetatakse andmetel kahte andmenihke algoritmi, proovitakse leevendada muutuste mõju mudeli ennustustäpsusele ja võrreldakse tulemusi. Viimaks, tehakse järeldused muutuse mõjust kõnekeskuse tööle algoritmide tulemuste põhjal ja arutletakse mudeli kasutatavuse üle teistes kõnekeskustes.

Võtmesõnad:

Kõnekeskus, SLA, teenustase, masinõpe, andmenihe, andmenihke algoritm, ADWIN, PCA-CD

CERCS: P176 - Tehisintellekt, P160 - Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Visuaalne abstrakt:



Detecting and Mitigating Data Drift in Call Center Data

Abstract:

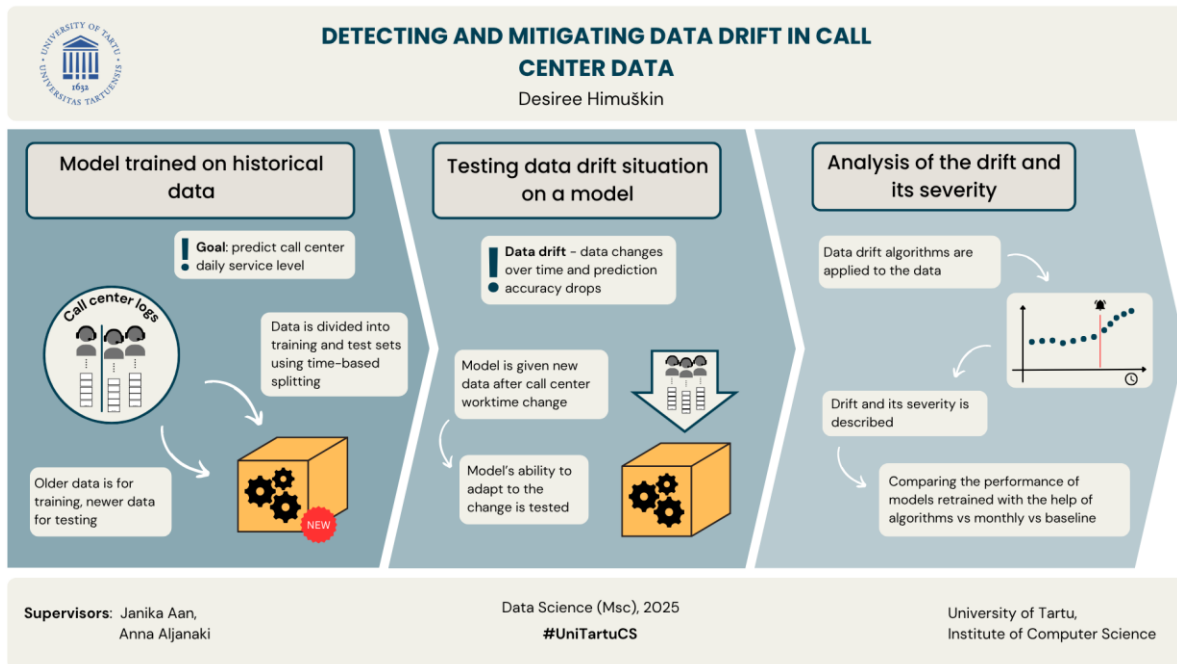
The aim of this master thesis is to analyse how call center worktime change affects the performance of created machine learning models, see if this change creates a data drift and how to mitigate its effect. First, literature, existing solutions on how to predict service level and data drift algorithms are analysed. Secondly, the call center dataset is described and analysed. Thirdly is given an overview of used models and their features, which is followed by practical work which includes describing the effects of the change on the data, testing two data drift algorithms on the dataset, trying to alleviate the effects of the change on prediction accuracy and finally comparing the results. In the final part, conclusions on the effect of the change are made based on the results of the algorithms and the usability of deploying these models in other call centers is discussed.

Keywords:

Call center, SLA, service level, machine learning, ML, data drift, data drift algorithm, ADWIN, PCA-CD

CERCS: P176 - Artificial intelligence, P160 - Statistics, operation research, programming, actuarial mathematics

Visual abstract:



Sisukord

Sissejuhatus.....	7
1 Kirjanduse ülevaade ja taust	8
1.1 Terminid.....	8
1.2 Kõnekeskuse töö	8
1.2.1 Kõnekeskuse töö hindamine	9
1.2.2 Erlang C ja Erlang A valemid	9
1.2.3 Masinõppe mudelid.....	10
1.3 Andme- ja kontseptsiooninihe	11
1.3.1 Nihke tuvastamise raamistik	11
1.3.2 Nihke tuvastamise algoritmid	11
2 Andmestik.....	14
2.1 Andmeanalüüs.....	15
3 Mudelid.....	19
3.1 Info kõnede mudel	20
3.2 Tehniliste kõnede mudel	20
4 Andmenihke olukord	22
4.1 Mõju info kõnede tunnustele	22
4.2 Info kõnede mudelil nihke tuvastamine ja leevendamine ADWIN algoritmiga.....	23
4.3 Info kõnede mudelil nihke tuvastamine ja leevendamine PCA-CD algoritmiga.....	26
4.4 Info kõnede mudelite võrdlus Erlang C-ga.....	27
4.5 Mõju tehniliste kõnede tunnustele	28
4.6 Tehniliste kõnede mudelil nihke tuvastamine ja leevendamine ADWIN algoritmiga	29
4.7 Tehniliste kõnede mudelil nihke tuvastamine ja leevendamine PCA-CD algoritmiga	32
4.8 Tehniliste kõnede mudelite võrdlus Erlang C-ga.....	32
5 Järeldused.....	34
5.1 Andmenihe info kõnede andmetes.....	34
5.2 Andmenihe tehniliste kõnede andmetes.....	34
5.3 Mudelite rakendamine teistes kõnekeskustes	34
Kokkuvõte.....	36
Viidatud kirjandus.....	37
Lisad.....	38
I. Litsents.....	38

Sissejuhatus

Kõnekeskused on lahutamatu osa igast suuremast ettevõttest, mis pakub teenust või müüb tooteid. Ja tihti võib olla nii, et kahe sarnast lahendust pakkuva ettevõtte vahe ongi ainult kliendiga suhtlus. Kokkupuude kõnekeskuse töötajaga avaldab otseselt mõju kliendikogemusele – negatiivne kogemus võib viia isegi selleni, et klient loobub teenusest ja siirdub konkurendi juurde. Hea kliendikogemuse korral on kliendid valmis ka teenuste eest rohkem maksma, kui tajuvad, et ettevõtte on kliendikeskne. Seega on ettevõtetele oluline planeerida efektiivselt kõnekeskuse tööd ehk säilitada tasakaalu teenuse kvaliteedi ja minimaalsete kulude vahel.

Teenustase on üks meetrika, mille alusel kõnekeskus hindab enda tööd ning graafikute planeerimisel jälgivad, et see oleks aktsepteeritaval tasemel. Kõnekeskustele on sätestatud SLA-d ehk teenustasemelepped, mis määravad ära aktsepteeritava teenustaseme ja kui seda ei täideta, siis lisaks rahulolematutele klientidele võivad sellega kaasned ka rahalised trahvid.

Magistritöö üks eesmärk on luua masinõppe mudelid teenustaseme ennustamiseks ja katsetada olukorda, kus kõnekeskus muudab oma tööaega, võrrelda erinevate algoritmide abil treenitud mudelite tulemusi teenuse kvaliteedi ennustamisel ja analüüsida, kuidas erinevad mudelid toimivad sellistes erakordsetes olukordades. Selle jaoks kasutatakse ühe reaalse kõnekeskuse andmeid. Selles kõnekeskuses on kaks erinevat kõneliini, mille jaoks on loodud kaks mudelit, ka töötaja muutus on kahel liinil erinev.

Teine eesmärk on analüüsida muutuse mõju andmetele ja teha kindlaks, kas antud muudatus tõi kaasa ka andmenihke ja kui suure ning mida oleks vaja, et mudelid suudaksid inimsekkumiseta andmete muutustega kohaneda. On oluline jälgida andmete muutumist ajas, sest klientide käitumisharjumused võivad muutuda või kõnekeskuses endas muutub näiteks kõneagentide hulk ja kui sellised muutused jäävad tähelepanuta, siis langeb mudeli ennustustäpsus ilma meie teadmata ning see mõjutab otseselt kõnekeskuse töö kvaliteeti.

Töö koosneb viiest sisulisest osast. Esimeses osas antakse ülevaade olemasolevast kirjandusest ja probleemi taustast. Teises osas antakse ülevaade, milline on kõnekeskuse andmestik üldiselt, analüüsitakse kasutatud andmestiku ja kolmandas osas kirjeldatakse mudelid. Neljandas keskendutakse praktilisele osale, kus katsetatakse erinevaid andmenihke algoritme, kirjeldatakse nende tulemusi ja analüüsitakse muutuse mõju. Viiendas osas tehakse järeldusi, arutledes algoritmide ja mudelite tulemuste kui ka piirangute üle.

1 Kirjanduse ülevaade ja taust

1.1 Terminid

SLA ehk teenustaseme leping¹ (ingl *Service Level Agreement*) määratleb teenuseosutajalt oodatava teenuse taseme. See koosneb erinevatest komponentidest nagu näitajad, mille alusel teenust mõõdetakse ja parandusmeetmed või karistused kui kokkulepitud teenustaset ei täideta. SLA-sid on eri tüüpi², näiteks *response time* SLA, paneb paika kui kaua kõne võiks maksimaalselt järjekorras ootel olla. *Resolution time* SLA, keskendub jällegi sellele, millise aja sees kliendi probleem peaks lahendatud saama. *Response time* SLA, mis on fookuseks antud töös, võib olla näiteks, et 80% kõnedest tuleb vastu võtta 30 sekundi jooksul.

SL ehk teenustase (ingl *Service Level*) on kõnede protsent, mis täidab SLA-s määratud tingimusi. Selle alusel hinnatakse SLA täitmist ja see näitaja peab olema vähemalt sama kõrge kui SLA-s sätestatud tase.

1.2 Kõnekeskuse töö

Kõnekeskuste efektiivse töö aluseks on hea planeerimine, mis on tavaliselt jaotatud nelja etappi. Kõige esimene etapp on eelarve planeerimine, seda tehakse kvartaalselt, millega pannakse paika eelarve, mis jõustub aasta pärast. [1]

Teises etapis, mida tehakse igakuiselt, hinnatakse vajalikku kõnekeskuse agentide hulka järgmiseks kvartaliks ning vajadusel värvatakse ja koolitatakse välja uusi töötajaid. Igakuiselt ennustatakse ka järgmise kuu töömahtu, mille põhjal planeeritakse töögraafikuid. Kõnede mahtu hindab töögraafikute koostaja, võttes arvesse nii ajaloolisi trende, kuid arvestab ka planeeritud sündmustega, millel on suurem kliendi mõju. Sündmused, millel on suurem kliendi mõju võivad olla näiteks kampaaniad, hinnatõusud või muud suuremad muudatused, mis mõjutavad kliendibaasi, kas osaliselt või täielikult. Selle hindamiseks peab olema teada teavituste valimi suurus, mille alusel oskab kogunud töötaja hinnata, kui palju võib sellest tekkida pöördumisi. [1]

Neljas etapp on igapäevane graafikute ülevaatamine, mille käigus jälgitakse, kas tööol olevate agentide arv suudab täita SLA või mitte, seda hinnatakse üldjuhul päeva täpsusega. Kui on tekkinud olukord, et sissetulevaid kõnesid on oodatust rohkem, siis kutsutakse appi lisajõude. Lisaressurss ei ole üldiselt sama efektiivne kui kõnekeskuse enda töötaja, seega tuleb arvestada, et nemad suudavad ühes tunnis vastata vähem kõnesid.

Kõnekeskustel on üldjuhul erinevad liinid erinevate teemade jaoks ning nende tööaeg võib ka olla erinev, näiteks eraklientidele on Telias kaks eri kõneliini³, üldinfo ja arved, mis on avatud tööpäevadel üheksast hommikul kuni kella kaheksani õhtul ning tehniline tugi, mis töötab lisaks tööpäevadele ka nädalavahetusel, aga lühematel tundidel kui nädala sees.

Töögraafikute planeerimise teeb keeruliseks ja ajakulukaks just see, et iga liini jaoks tuleb ennustada kõne maht, luua eraldi töögraafikud ning jaotada töötajad vahetustesse. Töötajate vahetustesse jaotamine võtab kõige rohkem aega, sest tuleb arvestada nii töötajate koormustega kui ka nende isiklike eelistustega.

Lisaks kõnede vastamisega on agentidel veel tegevusi, mille ajal nad ei saa uusi sissetulevaid kõnesid vastu võtta, nagu näiteks kõnejärgsed tegevused, paberitöö, pausid ja

¹ <https://www.liveagent.ee/klienditoe-sonastik/sla/>

² <https://central.com/resources/sla-call-center/>

³ <https://digitark.telia.ee/era/1-juulist-muutuvad-telia-konekeskuse-teenindusajad-nadalavahetustel>

väljahelistamised. Need kõik on süsteemis registreeritud, neid on võimalik jooksvalt jälgida ja ka hiljem analüüsida, et näha, mis tegevustega kõnekeskuse agendid hõivatud on.

1.2.1 Kõnekeskuse töö hindamine

Kõnekeskuse teenustaseme hindamisel jälgitakse erinevaid näitajaid. Teenustaset hinnatakse erinevatel liinidel eraldi. Töö hindamise aluseks on kõnemaht ehk kui palju on pakutud kõnesid, mis jaotud vastatud ja vastamata kõnedeks. Vastatud kõnede puhul eristatakse omakorda vastatud otsekõnesid ja väljahelistamisi. Otsekõne on sissetulev kõne, mille puhul helistab klient kõnekeskusesse. Väljahelistamise või tagasihelistamise korral helistab kõnekeskuse töötaja kliendile ise. Mõlemat tüüpi kõnesid pakub agendile kõnekeskuse süsteem. *Response time* teenustaseme hindamisel jälgitakse kui suure osa pakutud kõnedest moodustasid otsekõned, mis vastati SLA-s sätestatud ajavahemikus.

Juhul kui töö on ka lisaressurs, siis töögraafikute paremaks planeerimiseks tulevikus analüüsitakse SLA täitmist ka ilma lisaressursita, et näha kui täpne oli planeering. Ehk mis oli kõnekeskuse agentide poolt õigeaegselt vastuvõetud kõnede protsent pakutud kõnedest.

1.2.2 Erlang C ja Erlang A valemid

Klassikaline tööriist, mida kõnekeskused kasutavad personali planeerimiseks, on Erlang C ja Erlang A arvutus. Erlang C⁴ valem sai nime taani matemaatiku A.K. Erlangi järgi, kes arendas selle mudeli 1917. aastal. Sellel ajal ei eksisteerinud veel kõnekeskused tänapäevases mõttes, kuid nende tekkides avastati, et see valem on kasutatav ka kõnekeskuse töö planeerimisel. Valem loodi hoopiski telefonikeskuste personali planeerimise jaoks⁵. Sarnaselt kõnekeskustele töötasid seal operaatorid, kuid kõnedele vastamise asemel nad hoopiski ühendasid kõnesid vastavalt helistaja soovile.

Erlang C arvutab tõenäosuse P_w , et kõne ootab, kui kõne liiklusintensiivsus tunnis on A (ingl *traffic intensity*) ja olemasolevaid agente on N . Kõneliikluse intensiivsus on kõikide kõnede kestuste summa jagatud tööajaga.

$$P_w = \frac{\frac{A^N}{N!} \frac{N}{N-A}}{\left(\sum_{i=0}^{N-1} \frac{A^i}{i!}\right) + \frac{A^N}{N!} \frac{N}{N-A}}$$

Erlang C valem eeldab, et kõik kõned saavad Poisson'i protsessi kohaselt keskmise kiirusega, kõik kõned teenindatakse ja ühtegi kõne ei katkestata [2]. Kõne ootamise tõenäosust arvutatakse intervallide kaupa, intervall võib olla näiteks 30 minutit. Liiklusintensiivsuse A leidmise jaoks on olulised parameetrid kõnede saabumise määr ehk keskmine kõnede saabumise kiirus, keskmine teenindusaeg ehk keskmine aeg, mis kulub ühel agendil kõne teenindamiseks.

Uuringud on näidanud, et see mudel sobib kasutamiseks suurtes kõnekeskustes, kus on madal kuni mõõdukas hõivatuse määr, sest mudeli viga on märkimisväärselt suur kui on kõrge hõivatuse määr, vähe agente ja kannatamatud helistajad [2]. Teades P_w väärtust saab arvutada teenustaseme (SL), kus e on Euleri arv, *Target Time* on näiteks *Response Time* SLA-s sätestatud aeg sekundites ja *AHT* (ingl *Average Handling Time*), mis on ühele kõnele

⁴ <https://www.callcentrehelper.com/erlang-c-formula-example-121281.htm>

⁵ <https://www.callcentrehelper.com/celebrating-100-years-of-erlang-121278.htm>

keskmiselt kuluv aeg. AHT^6 on summa keskmisest kõneajast, ooteajast (*hold time*) ja kõnele järgnenud järeltööst.

$$SL = 1 - [P_w \times e^{-[(N - A) \times (Target\ Time / AHT)]}]$$

Kogu päeva teenustaseme leidmiseks tuleb leida keskmine üle kõigi intervallide teenustaseme.

Erlang A valemi lõi matemaatik Conny Palm 1946. aastal Erlang C edasiarendusena ja seetõttu vahest kutsutakse seda valemit ka Palm/Erlang A valemiks⁷, kus A viitab kõne katkestamisele ehk *abandonment*.

Erlang A mudelil on samad sisendparameetrid kui Erlang C mudelil, kuid sinna lisandub veel helistaja kannatlikkuse aeg, mis on eksponentsiaalselt jaotunud ja kui see ületatakse, siis helistaja lõpetab kõne. Kannatlikkuse aja eeldus on ka selle mudeli piiranguks, sest alati ei vasta eeldatav jaotus tegelikkusele. See mudel sobib kasutamiseks, kui kõnede katkestamine on oluline tegur ja soovitakse vältida vajalike agentide arvu ülehindamist.

Erlang A kannatlikkuse määramise keerukuse tõttu on Erlang C mudel populaarsem. Varasem analüüs kahe mudeli vahel on leidnud, et Erlang A on üldiselt täpsem, kuid optimistlikum ja Erlang C on pigem pessimistlikum. Ehk ühe mudeli puhul võidakse vajaliku töötajate arvu ülehinnata ja teise puhul alahinnata. Mõlemad mudelid eeldavad, et agendid on statistiliselt identsed ehk kõik agendid on sama töökiiruse ja oskustega.

1.2.3 Masinõppe mudelid

Kirjandust analüüsid, ilmneb, et kõnekeskuse töö planeerimise lihtsustamiseks kasutatakse erinevaid masinõppe (ML) lähenemisi, et ennustada erinevaid parameetreid nagu näiteks kõnede maht, vajalik agentide arv, agentide töö efektiivsus, jne. Näiteks puupõhiste algoritmidega ennustatakse kumulatiivset kõnepikkust, Bayes'i klassifikaator aitab klassifitseerida agente *onboarding*'u protsessi ajal, et eristada agente, kes täidavad ettevõtte seatud minimaalseid töö efektiivsuse nõuded ja kes mitte. [3]

Erlang A mudeli täpsust saab näiteks parandada sellega, kui asendada Poisson'i protsess närvivõrkude abil ennustatud kõnede liiklusega (ingl *call traffic*) ja stiimulõppe abil leida optimaalsed vahetuste algusajad kõnekeskuse agentidele. Selline lähenemine ületab staažika töötaja ennustustäpsust ja aitab säilitada teenuse kvaliteeti. [4]

Gradientvõimendamise masinad (ingl *gradient boosting machine, GBM*), mis on treenitud simuleeritud andmete peal on andnud häid tulemusi teenustaseme ennustamisel väikeste ja keskmise suurusega kõnekeskuste stsenaariumite puhul ning mõõdukalt hästi suure ja keeruka kõnekeskuse stsenaariumi puhul. Simuleeritud andmete peal treenitud mudelite puhul tuleb jälgida, et treeningandmed kataksid kõiki vahetuse tüüpe ja teenustaseme väljundid oleksid hästi jaotunud, et tagada usaldusväärsed tulemused ja vältida kallutatust. [5]

Testid on näidanud, et sügavõppe mudelid annavad parema ennustustäpsuse, kuna tabavad paremini keerukaid mustreid, kuid võrreldes juhendatud õppe mudelitega on need arvutuslikult keerukamad ja toimivad kui "musta kastina" ehk on keeruline mõista, miks mudel just sellise tulemuseni jõudis. Selliste lahenduste juurutamisega praktikas ei kiirustata, sest raske on anda niivõrd kriitilise teenuse juhtimisõigust mudelile. Pigem eelistatakse, et mudel toetab planeerimist, mitte ei võta kogu seda protsessi enda peale. Mudelid, mis ennustavad ainult ajalooliste andmete pealt ei oska ette näha muudatusi, mida kogenud graafikute koostaja oskab arvesse võtta.

⁶ <https://www.callcentrehelper.com/how-to-measure-average-handling-time-52403.htm>

⁷ <https://www.callcentrehelper.com/a-beginners-guide-to-the-erlang-a-formula-140998.htm>

1.3 Andme- ja kontseptsiooninihe

Kõnekeskuse töö mudeldamisel kasutatakse tavaliselt ajaloolisi andmeid, mis tähendab, et ennustusi tehakse eelnevate teadmiste põhjal ja ei võeta arvesse muutusi. Andmed ja trendid muutuvad aja jooksul ning kui seda muutust ei tuvastata ja sellega ei tegeleta, siis kasvab mudeli ennustusviga ja ühel hetkel pole see mudel enam kasutatav. Kontseptsiooninihe tähendab, et andmed muutuvad ja suhe, mis oli varasemalt sisendandmete ja ennustatava vahel on muutunud. Andmenihke puhul muutub sisendandmete jaotus.

Nihke tüüpe on nelja erisugust: järsk, järkjärguline, inkrementaalne ja korduv. Järsk nihe tähendab seda, et muutus toimub järsku lühikese aja sees. Kui uus kontseptsioon asendab vana hoopiski ajapikku, siis on tegemist järkjärgulise nihkega. Inkrementaalse nihke korral asendab ka uus vana kontseptsiooni ajapikku, kuid inkrementaalselt ehk vahepealne kontseptsioon on algus- ja lõppkontseptsiooni segu. Korduva nihke puhul võib vana kontseptsioon teatud aja möödudes uuesti ilmuda. [6]

Nihe on kolmemõõtmeline, sellel puhul saab hinnata, millal nihe toimus ja kui kaua see kestis, kui tõsine on nihe ja kus see aset leidis ehk millal, kuidas ja kus.

1.3.1 Nihke tuvastamise raamistik

Nihke tuvastamise raamistik koosneb neljast osast: andmete hankimine, andmete modelleerimine, teststatistiku arvutamine ja hüpoteesi test [6]. Andmevoost võetakse andmetükid kindlas ajalisel järjestuses, et säilitada neis peituv muster, mis jagatakse ajaliselt kaheks andmestikuks, ajaloolised ja uued andmed.

Andmete modelleerimise etapp on valikuline ja oluline siis, kui on vaja vähendada andmete dimensionaalsust või valimi suurust, selle eesmärk on eraldada need muutujad, mis nihkudes mõjutavad mudelit kõige rohkem. Andmenihke kõige keerulisem etapp on teststatistiku valimine, mis hindaks täpselt kahe andmestiku kaugust või erinevust. See kvantifitseerib nihke tõsiduse ja moodustab hüpoteesi testi jaoks teststatistiku. [6]

Viimases etapis hinnatakse, kas täheldatud muutus on statistiliselt oluline. Teststatistikul ilma hüpoteesi testita ei ole mingit tähendust, sest hüpoteesi test määrab usaldusvahemiku, mis aitab hinnata, kas muutus on tingitud andmenihkest, müra või juhuslikult sattunud valim oli kallutatud. Kõige sagedamini kasutatavad hüpoteesitesti on teststatistiku jaotuse hindamine, *bootstrapping*, permutatsioonitestid ja Hoeffding'i *inequality-based bound identification*. [6]

1.3.2 Nihke tuvastamise algoritmid

Nihke tuvastamiseks on olemas hulk erinevaid algoritme ning need võib jaotada kolme kategooriasse, vastavalt sellele, mille alusel nad nihet hindavad. Veamäära alusel (ingl *Error Rate-based*) nihke tuvastamisel jälgivad algoritmid, veamäära tõusu või languse statistilist olulisust. Kui muutus on statistiliselt oluline, siis käivitatakse mudeli uuendusprotsess. Enamik neist sobivad ainult klassifitseerimisega ja ei toimi kui ennustatakse reaalarvulist märgendit. [6]

Andmejaotusel põhinevad (ingl *Data Distribution-based*) nihke tuvastamise algoritmid jälgivad nihke algpunkti ehk andmeid. Need algoritmid hindavad kaugust ajalooliste ja uute andmete jaotuse vahel ning kui nende erinevus on statistiliselt oluline, siis käivitatakse mudeli uuendusprotsess. Seda tüüpi algoritmide puhul pole oluline, kas tegemist on klassifitseerimisprobleemiga või regressiooniga, sest jälgitakse sisendandmete muutusi. [6]

Mitmikhüpoteesi testil (ingl *Multiple Hypothesis Test*) põhinevad nihke tuvastamise meetodid kasutavad sarnaseid tehnikaid nagu eelmised kaks kategooriat, aga nende erinevus on see, et nad kasutavad mitut teststatistikut ja seega testivad mitut hüpoteesi enne otsuse langetamist. Need algoritmid jaotuvad kahte rühma: paralleelsed mitmikhüpoteesi testid ja hierarhilised

mitmikhüpoteesi testid. Paralleelsete mitmikhüpoteesi testid jälgivad mitut teststatistikut samaaegselt. Hierarhiliste mitmikhüpoteesi testide puhul on olemas nihke tuvastamise kiht ja valideerimise kiht. Tuvastamise kihis olev teststatistik peaks olema tundlik muutustele, kuid ei tohiks olla väga arvutusmahukas ja kui seal kihis tuvastatakse muutus, siis kindluse saamiseks testitakse hüpoteesi ka valideerimise kihis. [6]

Töö [6] autorid on oma uurimistöö tulemusena välja toonud kõige asjakohasemad algoritmid. Nad leidsid, et kõik nihke algoritmid suudavad hinnata millal nihe aset leidis, kuid vähesed suudavad hinnata kuidas ja kus. Mitmed mitmikhüpoteesi testid ja enamik veamääral põhinevad algoritmid suudavad tuvastada ainult nihke aega, kõige paremini kirjeldavad kõiki nihke aspekte andmejaotust hindavad nihke tuvastamise algoritmid. Enamik algoritme selles töös sobivad ainult klassifitseerimise probleemidele, sest tuvastatakse nihet, kas ennustustäpsuse või mõne muu meetrika abil, mis regressiooni puhul pole relevantne. Kuid osasid algoritme saab kohandada nii, et need oleksid kasutatavad ka regressiooni puhul.

Osade nihke algoritmide puhul on neid raske liigitada ühemõtteliselt kontseptsiooni- või andmenihke algoritmide alla ja tihti kasutatakse neid termineid läbisegi. Näiteks PCA-CD, mida töös [6] defineeritakse kui kontseptsiooninihke algoritm, on algallikas [7] kutsutud üldisemalt kui muutuse tuvastamise (ingl *change detection*) algoritm ja menelaus pakettis, kus on selle töö põhjal loodud funktsioon, on hoopiski liigitanud selle andmenihke algoritmide alla. Seega selles töös kutsutakse neid üheselt andmenihke algoritmideks.

2.3.2.1 ADWIN

ADWIN ehk *ADaptive WINDOWing* on muutuse tuvastamise algoritm, mis kasutab nihke tuvastamiseks nihkuva akna meetodit (ingl *sliding window*), et hinnata ühe parameetri dispersiooni ja keskmist väärtust. See on ka üks populaarsemaid nihkuva akna põhiselt toimiv andmenihke algoritmi. See algoritm katsetab erinevad lõikekohti, et see aken jaotada omakorda kaheks alam-aknaks [8]. Nihke hindamiseks võrreldakse kahe alam-akna paaride andmejaotust ning kui nende erinevus ületab olulisusnivood, siis vanu andmeid eemaldatakse aknast kuniks kõik alam-akende paarid jäävad alla olulisusnivoo⁸. Uues aknas olevaid andmeid saab kasutada mudeli uuesti treenimiseks. [6]

Selle algoritmi üheks eeliseks on see, et kasutaja ei pea ise defineerima akna suurust, vaid seda otsustab algoritm ise vastavalt sellele kui muutlikud on andmed [8]. Kui muutusi ei toimu on vaadeldav aken suurem, aga kui toimub järsk muudatus, siis on aken väiksem, et mitte lasta mudelit mõjutada vanadest andmetest. Kasutaja peab defineerima vaid olulisusnivoo, akna ja alam-akna minimaalse suuruse ning protsessi sammu ehk kui tihti nihke tuvastamise protsess peaks käivituma. Mida väiksemad on need väärtused, seda kiiremini saab tuvastada muutusi, kuid see võib kaasa tuua ka palju valepositiivseid nihke tuvastamisi. Samuti on ADWINi eeliseks see, et seda saab rakendada ükskõik millise numbrilise parameetri nihke tuvastamiseks, nii ennustusvea kui ka mõne tunnuse muutuse tuvastamiseks.

2.3.2.2 PCA-CD

Principal Component Analysis Change Detection (PCA-CD) on andmenihke algoritm, mis tuvastab nihet andmejaotuses, nagu nimigi viitab, PCA ehk peakomponentide analüüsi alusel ning sarnaselt ADWINle kasutab kahte andmeakent, kus ühes on vanad andmed ja teises uued. Mõlema akna suurused tuleb eelnevalt defineerida ning ajalooliste andmetega aken on ajas fikseeritud, kuid uute andmetega aken nihkub edasi. Nihke tuvastamiseks arvutatakse kõigepealt ajalooliste andmete PCA, millele seejärel projekteeritakse uute andmete PCA

⁸https://menelaus.readthedocs.io/en/latest/menelaus.change_detection.html#module-menelaus.change_detection.adwin

andmepunktid ja kui nende erinevus ületab lävendi, siis on toimunud nihe ning vana aken asendatakse uuega⁹. Lävendi määrab dünaamiliselt Page-Hinkley test. [6]

Erinevalt ADWINist, jälgib see algoritm korraga muutusi mitmes tunnuses. Kuid selle algoritmi üheks suurimaks miinuseks on see, et kasutaja peab ise akna suuruse defineerima, mis võib väga tugevalt mõjutada algoritmi efektiivsust. Uues aknas olevaid andmeid võib kasutada uuesti treenimiseks, kuid olenevalt domeenist ja defineeritud akna suurusest, võib sellest jääda väheks. Näiteks kõnekeskuse andmete nihke hindamiseks sobib umbes kolmekümne päeva pikkune aken, kuid uuesti treenimiseks sellisest andmemahust ei piisa.

Kahe algoritmi kohta on tehtud kokkuvõtlik tabel Tabel 1.

Tabel 1. Ülevaatlik tabel andmenihke algoritmidest

Algoritm	Jälgitavate parameetrite arv	Jälgitavate parameetrite tüüp	Nihke tuvastamiseks jälgitakse	Akna suuruse määramine
ADWIN	1	Täpsus, veamäär, featuur	Keskmist ja dispersiooni	Dünaamiline, määratakse miinimumsuurus
PCA-CD	n	Featuurid	PCA-d	Fikseeritud

⁹ https://menelaus.readthedocs.io/en/latest/menelaus.data_drift.html#module-menelaus.data_drift.pca_cd

2 Andmestik

Kõnekeskuse töö hindamiseks on olulised kaks andmestiku: kõnekeskuse sündmused ja agentide tegevuste logi. Kõnekeskuse sündmuste andmestik sisaldab kõiki pöördumisi kõnekeskusesse. Agentide tegevuste logist on võimalik näha iga agendi tegevusi ja millal neid tehti. Siin töös kasutati ühe Eesti ettevõtte kõnekeskuse andmeid, kuid kõnekeskuste andmete ülesehitus ja iseloom on üldjoontes sama [9].

Puhastamata kõnekeskuste sündmuste andmestik jaotub kolmeks põhiliseks osaks: kõneagendi info, kõneinfo, kanali info (Tabel 2). Kõneagendi info on üldjuhul tema kasutajatunnus, nimi ja üksus või tiim, kus ta töötab. Kõneinfo sisaldab kõne tehnilisi andmeid nagu näiteks, A ja B number või *source ja destination*, millal kõne süsteemi jõudis, millal kõne vastu võeti ja lõpetati. Kõnede kohta võib eraldi veel salvestada, järjekorras ootamise aja ja kõnekestuse ehk rääkimise aja ning milline oli selle kõne tulemus, kas see vastati õigeaegselt (ingl *answered on time*) ehk SLA-s sätestatud aja jooksul, lihtsalt vastati (ingl *handled*) või katkestatud (ingl *abandoned*).

Salvestatakse ka kanali infot, mis kanalisse kõne tuli ehk kanali nimi või tunnus. Kanalitel võib olla hierarhiline struktuur, näiteks on era- ja äriklientidel eraldi kanalid, mis omakorda jaotuvad erinevateks järjekordadeks. Järjekorrad on jaotatud temaatiliselt ja vastavalt sellele, mis sidekanali kaudu klient pöördub. Näiteks erinevatesse järjekordadesse lähevad kaks eraklienti, kellest üks helistab üldinfo numbrile ja teine võtab ühendust vestluse kaudu.

Tabel 2. Kõnekeskuse sündmuste näidisandmestik lihtsustatud kujul

Agendi tunnus	A number	B number	Saandumise kellaeg	Vastuvõtmise kellaeg	Kõne lõpetamise kellaeg	Ootamis-aeg (s)	Rääkimise aeg (s)	Kõne staatus	Järjekorra nimi	Kanali tüüp	Kanali tunnus
A1	5512345	123	24.02.2025 15:12:05	24.02.2025 15:12:20	24.02.2025 15:19:44	15	444	Answered on time	Lepingud ja arved	Kõne	Era
A2	5567890	321	24.02.2025 15:15:00	24.02.2025 15:16:02	24.02.2025 15:30:23	62	861	Handled	Tehniline abi	Kõne	Äri
A3	5654321	456	24.02.2025 15:16:17		24.02.2025 15:20:07	230	0	Abandoned	Üldinfo	Kõne	Era

Agentide sündmuste logid on salvestatud andmebaasi intervallidena, näiteks võib üks intervall olla 15 minutit (Tabel 3). Erinevate tegevuste kirjeldamiseks on andmebaasis kahte tüüpi veerge. Ühed tähistavad erinevate sündmuste toimumise arvu ning teised, mis kirjeldavad toimunud tegevuste ajalist kestust. Esimest tüüpi veerud on näiteks kui mitu kõne, meili või sõnumit agendile suunati, kui mitu korda tehti pause ja kui mitu kõne, meili või sõnumit teenindati. Teist tüüpi veergudest leiab teenindusaja, paberitööle kulunud aja, pausi pikkuse või vaba aja.

Nendest andmetest võib välja lugeda näiteks kui pikk oli reaalne tööaeg, millal ja kui pikalt oldi pausil või lõunal, kui palju võttis aega paberitöö või kui palju oli tal vaba aega kõnede vahel.

Tabel 3. Agentide tegevuste logi näidisandmestik lihtsustatud kujul

Agendi tunnus	Aeg	Staatus	Suunatud kõnede arv	Suunatud meilide arv	Vastatud kõnede arv	Vastatud meilide arv	Pauside arv	Teenindu saeg (s)	Kõneaeg	Paberitöö aeg (s)	Pausi aeg (s)	Vaba aeg (s)
A1	24.02.2025 13:15:00	Vaba	1	0	1	0	0	900	200	300	0	400
A1	24.02.2025 13:00:00	Hõivatud	0	0	0	0	1	900	0	0	600	300
A1	24.02.2025 12:45:00	Vaba	2	0	2	0	0	900	500	200	0	200

2.1 Andmeanalüüs

Antud lõputöö raames keskendutakse erakliendi kõnedele, seega andmestikust jäetakse välja ärikõned ja sündmused, mis tulid sisse muid kanaleid pidi. Eraklientidele on kaks kõneliini, üldinfo ja arved ning tehniline info. Nagu liinide nimedki viitavad, siis üks liin pakub üldist infot ja teine aitab klientidel lahendada tehnilisi probleeme. Kõnesündmuste andmeid hoitakse kaks aastat ehk töö tegemise hetkel on andmestikus kõne sündmused alates 01.02.2023. Andmestikust jäeti välja kõnesündmused, millel staatus ei ole “vastatud õigeaegselt”, “vastatud” või “hüljatud” või puudub agendi info. Selle tulemusena saadi andmestik, kus ei olnud puuduolevaid andmeid. Andmestiku kvaliteet on hea ilmselt tänu sellele, et need logib automaatselt maha kõnekeskuse süsteem.

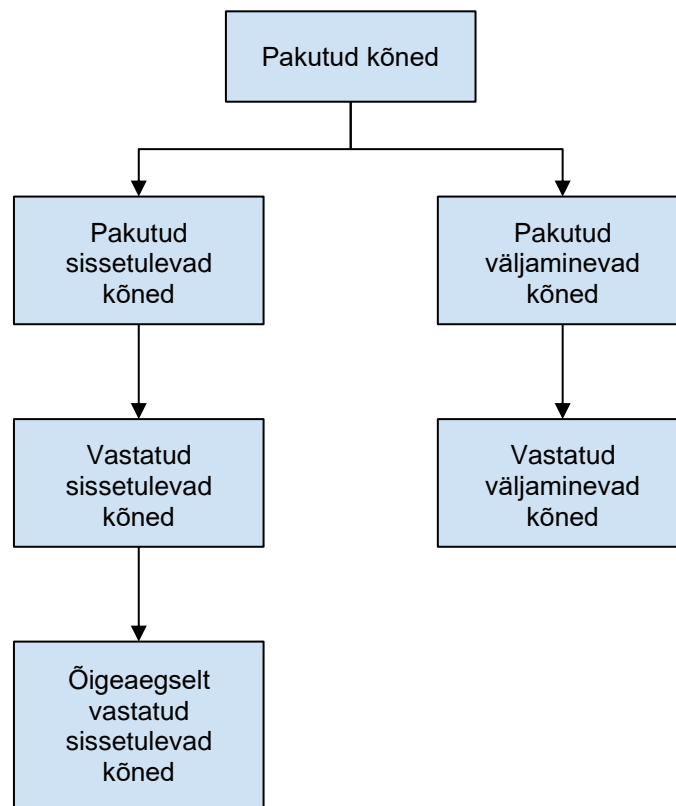
Kui vaadata pakutud kõnede trende läbi aastate ja kuude, siis tundub, et nende maht püsib suhteliselt stabiilne. 2024. aasta võrdlus 2023. aasta pakutud kõnede arvuga näitab, et kui kevadel 2024. aasta oli pakutud kõnede maht veel tunduvalt väiksem kui 2023. aastal, siis juuli ja augusti kuus teeb kõnemaht suure hüppe ja jõuab kogu aasta tipptasemele. SLA eesmärgi täitmine on infokõnede liinil läinud kolme aasta võrdluses veidi halvemaks, kuid tehnilistel kõnede liinil on trend olnud vastupidine.

Huvitav on ka asjaolu, et infokõnede liinil kasvab iga-aastaselt keskmine töötajate arv päevas, kuid tehniliste kõnede liinil just väheneb. See võib olla tingitud asjaolust, et infokõnede liinil on päevane mediaan taustatöö aeg suhteliselt muutumatu, kuid tehniliste kõnede liinil on seda mahtu suudetud stabiilselt vähendada. Tehniliste kõnede teenustase langus võib olla tingitud sellest, et vähenenud on korruga nii taustatööle kuluv aeg kui ka agentide arv ja see on tõstnud töö mahtu ühe agendi kohta.

Teenustase skaleeriti 0 ja 1 vahelt 0 ja 100 vahele, et tulemuste tõlgendamine oleks intuitiivsem. Andmeid analüüsitakse päeva lõikes, sest vaja on luua mudel, mis hindaks päevast teenustaset, seega on iga päeva kohta kaks rida, eraldi info ja tehniliste kõnede kohta. Analüüsist jäeti välja kõik lisaressursiga seonduv ehk agentide puhul võeti arvesse ainult kõnekeskuse omatöötajad ja nendega seotud kõnesündmused, kuna soovitakse hinnata just nende tööd. Visualiseeriti erinevate andmeveergude jaotust ja suhet teistega, et saada aimu andmepunktide jaotusest, et leida olulisi seoseid. Näiteks leiti, et tehnilisi kõnesid on vähem kui info kõnesid, kuid need kestavad kauem. Analüüsides tehniliste kõnede taustatöö mahtu ja agentide arvu mõju teenustasemele, selgus, et taustatöö ei suurene lineaarselt koos kõnede mahuga, kuid kui on rohkem taustatööd, siis teenustase on madalam. Erinevalt taustatööst, pikem kõnekestus teenustaset ei mõjuta.

Info kõnede puhul oli taustatöö mõju sarnane, kuid nende puhul mõjutab teenustaset negatiivselt ka pikem kõnekestus.

Kõnede puhul eristatakse sissetulevaid kõnesid, sissetulevad kõned, mis vastati SLA-s sätestatud aja jooksul ja väljaminevad kõned (Joonis 1). Väljaminevate kõnede all mõeldakse selliseid kõnesid, kui kõnekeskus helistab kliendile tagasi. Teenustaseme hindamisel jälgitaksegi õigeaegselt vastatud sissetulevate kõnede osakaalu kogu pakutud kõnedest.



Joonis 1. Kõnesündmuste jaotus

Kõnede mahtu mõjutavad erinevad faktorid nagu näiteks nädalapäev, riigipüha või kuu algus. Tsükliliste muutujate peal nagu nädalapäev, päev ja kuu kasutati *sin-cos* tsüklilist kodeeringut¹⁰ (ingl *cyclical encoding*), et nende tsükliline iseloom väljenduks. Näiteks juulikuu puhul on $a = 7$ ja $\max(a) = 12$.

$$x = \sin\left(\frac{a \times 2\pi}{\max(a)}\right) \quad y = \cos\left(\frac{a \times 2\pi}{\max(a)}\right)$$

Teenustaset päevade lõikes uurides, selgus, et kuu alguses umbes 3. kuupäevast kukub tase mitmeks päevaks madalamale tasemele kui muidu. See on seotud arvete välja saatmisega, sest arved saadetaksegi välja 2.-3.ndal kuupäeval. Seega on oluline arvestada kui kaua on möödunud arvete väljasaatmisest. Uuriti veel ebatavaliselt hea või halva tulemusega päevi ja nii avastati, et riigipühadel on tavaliselt teenustase hea, kuid muid mustreid ei leitud. Seega tuleb eeldada, et halva tulemusega päevad on mõjutatud, kas välja saadetud teavitustest või millestki, mida ei saa ette näha.

Teenustaset nädalapäevade lõikes uurides selgus, et info kõnede puhul on teenustase suurema tõenäosusega väga hea reedeti ning väga halb just kolmapäeviti. Esmaspäev võib olla kehvena teenustasemega kui arvete saatmise kuupäev satub nädalavahetusele ja kliendid hakkavad sellega tegelema alles uue nädala alguses. Tehniliste kõnede puhul on teenustase suurema tõenäosusega väga hea pühapäeviti ning nädala sees kolmapäeval ja neljapäeval, kuid seevastu

¹⁰ <https://medium.com/@axelazara6/why-we-need-encoding-cyclical-features-79ecc3531232>

on halvem tase laupäeviti. Kehv tase laupäeval võib olla tingitud sellest, et nädalavahetusel on inimestel rohkem aega televiisorit vaadata ja seega on suurem tõenäosus tehniliste probleemide avastamiseks või tekkimiseks. Lisaks võib mõjutada teenustaset ka mõni suurem hinnatõus või kampaania, kuid tavaliselt graafiku koostajad oskavad sellega arvestada. Neil on umbes teada, kui suurt kliendibaasi mõni muudatus võib mõjutada ja oskavad seega hinnata sissetulevate kõnede mahtu.

Uuriti ka töötajate staaži mõju teenustasemele, kuid sealt ei tulnud välja mingit mustrit, mis viitaks, et väiksema töökogemusega agendid oleksid madalama tulemuslikkusega. Selle uurimiseks vaadati, kas andmetes tekib mingeid selgeid lõikekohti (ingl *cut-off*) kohti, et proovida kategoriseerida töötajaid tööstaaži põhjal. Ehk kas uutel ja kogenumatel töötajatel on selgeid erinevusi kõnekestuses, vastatud kõnede arvus või isiklikus teenustasemes. Isikliku teenustaseme all mõeldakse, kui mitu protsenti ühele agendile pakutud kõnedest suudeti vastata SLA-s sätestatud aja sees.

Erakordsete päevade tuvastamiseks treeningandmetes on loodud tunnus *call_diff*, mis hindab mitu kõne keskmiselt oleks üks agent pidanud ühes tunnis vastama ja mis oli tegelikkuse vahe. See on ligikaudne hinnang, mis ei võta arvesse kõnede saabumise jaotust, vaid eeldab, et kõned saavad ühtlases jaotuses. Üks tunnus, mis on mõjutatud paljude erinevate muutujate poolt, kas otseselt või kaudselt: pakutud otsekõnede arv, agentide arv, ooteaeg, kõnekestus, taustatöö. Selle saamiseks leitakse vahe kui mitu kõne peaks üks agent õigeaegselt ära vastama ja kui palju vastati reaalsuses. Kui kahe arvu vahe on suur, siis see tähendab, et ka teenustase on kehv ja töötajate planeerimine ei olnud hea või kõnede maht oli oodatust suurem. Vastupidiselt väike või olematu vahe on päevadel, kui teenustase on 100%-lähedane. Selline erindite leidmise meetod on intuiitiivne, võrreldes näiteks *k*-keskmiste klasterdamisalgoritmiga (ingl *k-means*), mille efektiivsus sõltub valitud klastrite arvust *k* ning mis ei ole sobilik kasutamiseks andmestikel, millel on palju tunnuseid, sest selle tulemus saab kannatada “Mõõtmelise needusest”.

Tunnuste olulisust ja mõju uuriti korrelatsiooni ja XGB regressiooni mudeliga globaalsel tasemel, tunnuste olulisust hinnati *total gain* alusel. Nende abil saadi aimu, mis võiksid olla mudeli potentsiaalsed tunnused. Lokaalset ehk ühte päeva ei saa võtta tunnuste olulisuse hindamise aluseks, sest sellise lähenemise puhul võib ekslikult sisse võtta väheolulised tunnused või hoopiski välja jätta olulised. Võrreldes näiteks SHAPi XGB-ga, siis SHAP on mudeli spetsiifiline ja tunnuste olulisus on seotud valitud mudeliga, aga XGB regressorit kasutatakse reaalsest ennustavast mudelist sõltumatult, et tuvastada olulisi tunnuseid globaalsel tasemel. Lisaks, ei olnud selles etapis lõplik kasutatav mudel teadagi. Uuriti ka korrelatsioonimaatriksit ja mõnes mõttes oli sealt loetav tulemus ootuspärane, sest teenustase oli negatiivselt korreleeritud ooteaja ja väljahelistamise arvuga. Samas oli veidi üllatav, et agentide arv ei olnud tugevalt korreleerunud teenustasemega ehk rohkem agente tööl ei tähenda automaatselt paremat teenustaset.

Mudeli jaoks tunnuseid valides tuli arvestada, et kõiki väärtusi me ette ei tea ja seega tuleb need välja jätta või leida viis, kuidas anda neile hinnang ajaloo põhjal. Näiteks kõne pikkus, taustatöö ja ooteaeg on tunnused, mida me ei tea ette, aga neile saab väärtuse anda ajaloo põhjal. Nende tunnuste puhul toimib hästi viimase kolme kuu päeva keskmistest mediaani võtmine. Mediaan väärtus ei ole mõjutatud erandlikest päevadest. Taustatööle kuluv aeg on suhteliselt stabiilne päevade lõikes ja kui ka otsustatakse seda lühemaks saada, siis see muutus on mõnsekundiline ja ei juhtu üleöö. Andmete alguspunktis, kust andmed algavad, antakse nendele tunnustele väärtuseks 0.

Ka kõnekeskuse agentide arv ja otsekõnede protsent, mis mõjutavad teenustaset, ei ole ette teada, aga neile ei saa adekvaatset väärtust anda ainult ajaloo keskmistamisega. Neid tunnuseid ennustatakse kahe eri mudeliga.

3 Mudelid

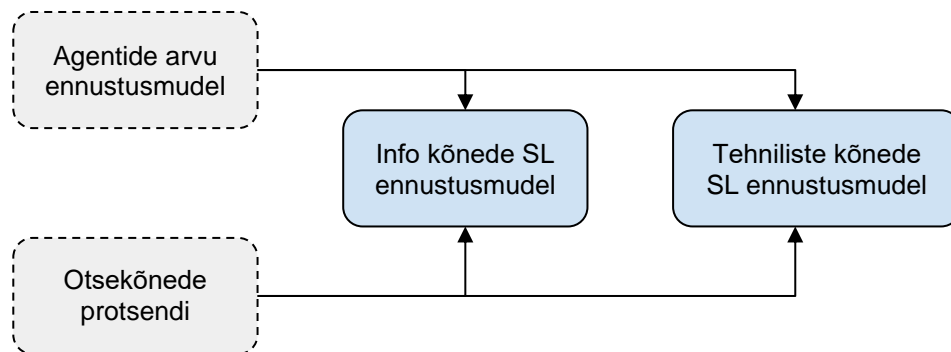
Info ja tehniliste kõnede jaoks olid eri tunnused erineva olulisusega ja kuna eraldi mudelid kahe kanali jaoks andsid paremad tulemused kui üks mudel, siis otsustati teha kaks erinevat mudelit. Andmed jagati treening-, valideerimis- ja testandmestikuks loogilises ajalisel järjestuses ehk kõige värskem olemasolev kuu on testimiseks, sest on oluline leida mudel, mis suudab kõige paremini ennustada ajalooliste andmete põhjal. Selline lähenemine aitab säilitada andmetes oleva mustri ja hoida ära olukorda, kus mõni hilisem muutus lekib treeningandmetesse ja testandmestiku satub hoopiski mõni vana kontseptsiooniga päev. Mudelite ja tunnuste sobivust hinnati R^2 -skoori, MSE ja MAE-ga. R^2 -skoori abil hinnati mudeli headust ehk kui suurt osa varieeruvustest mudel suudab ära seletada ning MSE ja MAE abil vigade suurust. Kaaluti ka MAPE valimist, kuid selle tõlgendamine ei ole kohe selgelt arusaadav äripoolle, sest see on suhteline veaprotsent reaalsest teenustasemest. MAE sobib selle jaoks paremini, sest arv on ennustatud teenustaseme ja reaalse teenustaseme vahe ehk on samal skaalal. Ülesobitamise vältimiseks katsetati eri mudeleid ja tunnuseid ka viiekordse ristvalideerimisega (ingl *cross-validation*).

Mudelite eesmärk on hinnata kõnekeskuse prognoositavat teenustaset ja ressursivajadust päeva ja kuu lõikes, mis põhineb varasemate perioodide andmetel ja arvestab näiteks pakutud, vastatud ja välja helistatud kõne mahtudega, kõne pikkuse, taustatööks kuluva aja ja ressursiga.

Mudel aitab töögraafikute koostamisel ja seega peaks vastama järgmistele küsimustele:

- Kui kõnede maht jääb sarnaseks või langeb/kasvab teatud protsendi võrra, milline on teenustase?
- Kui inimressurss jääb sarnaseks või langeb/kasvab, milline on teenustase?
- Kui kõne pikkus ja/või taustatöö aeg jääb samaks eelnevate perioodidega või muutub, milline on vajalik ressursihulk, et täita SLA?
- Kui palju pakutud kõnesid saab maksimaalselt olla, et oleks reaalne SLA eesmärk täita?

Kahe põhilise mudeli ehk SL-i ennustavate mudelite jaoks on loodud ka kaks lisamudelit, üks ennustab sissetulevate otsekõnede protsenti ja teine kõnekeskuse agentide arvu (Joonis 2).



Joonis 2. Mudelite ülevaade

Kaks lisamudelit loodi seetõttu, et need tunnused on teenustaseme ennustamisel olulised, ilma nendeta olid ennustused halvemad ja neid väärtuseid ei saanud lihtsalt võtta ajaloost, vaid on samamoodi mõjutatud erinevatest parameetritest. Teenustaseme ennustamiseks katsetati viite eri tüüpi mudelit: KNeighboursRegressor (*instance-based*), PoissonRegressor (*Generalized Linear Model*), Lasso (lineaarne mudel L1 regulatsiooniga), RandomForestRegressor (*ensemble of decision trees (bagging)*), XGBRegressor (*boosted ensemble model*). Katsetati eri tüüpi mudeleid, et aru saada, mis tüüpi algoritmid antud andmetega kõige paremini toimivad.

Kuna antud töö põhirõhk on andmenihke uurimine, siis rohkem süvitsi mudelite testimisesse ja valikusse ei minda.

3.1 Info kõnede mudel

Info kõnede jaoks oli parim mudel ElasticNet. Tabelis 4 on välja toodud info kõnede mudeli teenustaseme ennustamise tunnused.

Tabel 4. Info kõnede teenustaseme ennustamise mudeli tunnused

Tunnuse nimi	Kirjeldus	Kust saadakse väärtus
pakutud_kõned	Pakutud kõnede arv	Kasutaja sisend
total_worktime_min	Kõikide kõnekeskuse agentide tööaja summa	Kasutaja sisend
total_call_length	Eeldatav kõne pikkus (summa) sekundites	Täidetakse taustal
median_taustat66	Ennustatud taustatööle kuluv aeg sekundites, paberitöö + järeltöö	Täidetakse taustal
est_pakutud_tagasihelistamised	Ennustatud pakutud väljaminevate kõnede arv	Täidetakse taustal
is_holiday	Riigipüha tunnus	Täidetakse taustal
total_wait_time	Eeldatav ooteaeg (summa) sekundites	Täidetakse taustal
vajalik_kõne_per_person	Kui palju kõnesid keskmiselt peaks üks agent vastama, et kõik kõned saaksid vastatud	Täidetakse taustal
first_workday_after_bills_sent	Esimese tööpäeva pärast arvete väljasaatmist tunnus	Täidetakse taustal
days_since_bills_sent	Möödunud tööpäevade arv arvete välja saatmisest	Täidetakse taustal

3.2 Tehniliste kõnede mudel

Tehniliste kõnede jaoks oli samuti parim mudel ElasticNet. Tabelis 5 on välja toodud tehniliste kõnede mudeli teenustaseme ennustamise tunnused.

Tabel 5. Tehniliste kõnede teenustaseme ennustamise mudeli tunnused

Tunnuse nimi	Kirjeldus	Kust saadakse väärtus
pakutud_k6ned	Pakutud kõnede arv	Kasutaja sisend
month	Kuu väärtus	Kasutaja sisend
total_worktime_min	Kõikide kõnekeskuse agentide tööaja summa	Kasutaja sisend
est_agentide_arv	Hinnanguline kõnekeskuse agentide arv	Täidetakse taustal või kasutaja sisend
Kuu_sin_encoded	Tsükliliselt kodeeritud kuu sin-väärtus	Täidetakse taustal
Kuu_cos_encoded	Tsükliliselt kodeeritud kuu cos-väärtus	Täidetakse taustal
day_sin_encoded	Tsükliliselt kodeeritud päeva sin-väärtus	Täidetakse taustal
day_cos_encoded	Tsükliliselt kodeeritud päeva cos-väärtus	Täidetakse taustal
median_taustat66	Ennustatud taustatööle kuluv aeg sekundites, paberitöö + järeltöö	Täidetakse taustal
est_pakutud_otsek6ned	Ennustatud pakutud otsekõnede arv	Täidetakse taustal
est_pakutud_tagasihelista mised	Ennustatud pakutud väljaminevate kõnede arv	Täidetakse taustal
year	Aasta väärtus	Täidetakse taustal
est_otsek6nede%	Ennustatud protsent, otsekõnede osakaal kõikidest pakutud kõnedest	Täidetakse taustal
vajalik_k6ne_per_person	Kui palju kõnesid keskmiselt peaks üks agent vastama, et kõik kõned saaksid vastatud	Täidetakse taustal
first_workday_after_bills_ sent	Esimene tööpäev pärast arvete väljasaatmist	Täidetakse taustal
bills_sent	Arvete saatmise päeva märg	Täidetakse taustal
days_since_bills_sent	Möödunud päevade arv arvete välja saatmisest	Täidetakse taustal
median_wait_time	Ennustatud keskmine ooteaeg	Täidetakse taustal

4 Andmenihke olukord

Aja möödudes andmete iseloom muutub ning hea mudel võiks suhteliselt väikese vaevaga kohanduda uue olukorraga ehk võiks piisada mudeli uuesti treenimisest uute andmetega. Töökindluse hindamiseks, andmenihke mõju hindamiseks ning analüüsimiseks katsetatakse olukorda, kus kõnekeskus on muutnud oma tööaega.

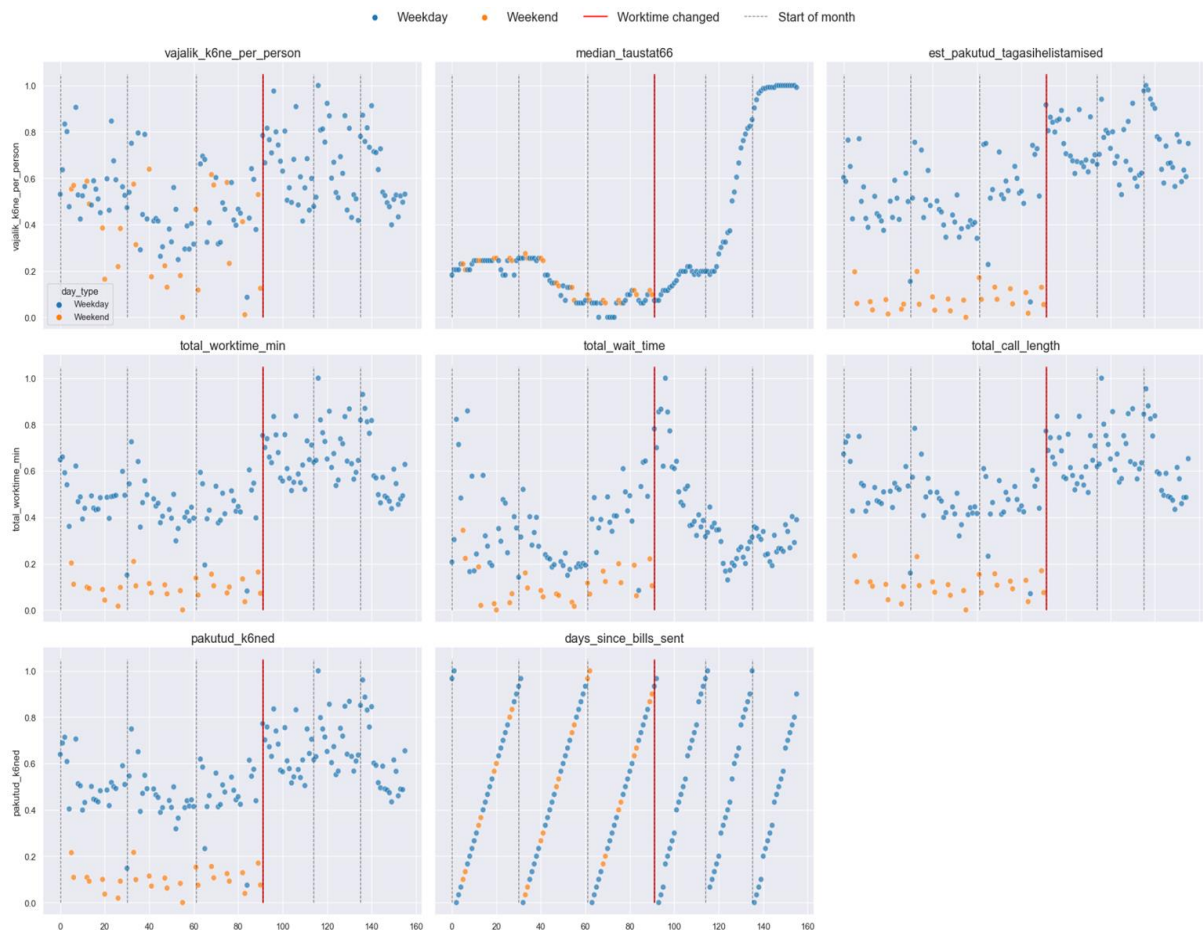
Töö tegemisel on kasutatud reaalseid andmeid ja see olukord on toimunud ühes eesti ettevõtte kõnekeskuses 2024. aasta juunikuus, kui tehniliste kõnede kanal hakkas nädalavahetusel lühemalt töötama ja info kõnede kanal oli edaspidi avatud ainult tööpäeviti. Kuna kõnekeskuse tööaja muutmine ja selle muudatuse jõustumise aeg oli ette teada, siis on andmestikust muudatuse leidmine lihtne. Selle olukorra puhul analüüsitakse ja kirjeldatakse andmete muutust, katsetatakse erinevaid andmenihke algoritme, et leida parim lähenemisviis mudeli ennustustäpsuse säilitamiseks.

Mudelite ennustatud tulemusi võrdleme baasmudeliga, Erlang C. Antud andmestikuga sobib just Erlang C paremini, sest kõnede hülgamist kasutatud andmestikus nii tihti ei esine, et see oleks oluliselt mõjutav parameeter.

4.1 Mõju info kõnede tunnustele

Tööaja muutus oli järsk nihe ja seda on näha ka tunnuste andmejaotusest. Alloleval joonisel (Joonis 3) on infomudeli tunnuste jaotused välja toodud vahemikus aprill kuni septembri lõpp 2024 ehk kolm kuud enne ja pärast tööaja muutust, kahel perioodil on erinev taustavärv. Ühe kuu asemel kolme kuu hindamine enne muutust on oluline, saada üldisem ülevaade trendidest, et mitte tahtmatult teha ennatlike järeldusi ebaharilikel päevadel. Pärast muudatust sama pika vahemiku analüüsimise eesmärk on aru saada, mis on selle muutuse lühi- ja pikaajalised mõjud.

Kuna selles perioodis on suurimad erinevused tööpäevade ja nädalavahetuse vahel, siis ei eristata joonisel kõiki päevi, vaid ainult neid kahte. Tunnuste väärtused on parema visualiseeringu ja andmekaitse eesmärgil normaliseeritud nulli ja ühe vahele *MinMaxScaler*'iga.



Joonis 3. Info kõnede mudeli tunnuste jaotused aprill kuni september 2024

Peaaegu, et iga tunnuse puhul on näha, et laupäeviti ja pühapäeviti on tunnuse väärtused väiksemad kui tööpäevadel (Joonis 3). Nädalavahetusel on sissetulevate kõnede maht väiksem, mis omakorda tähendab seda, et ka vajaminev töömaht on väiksem ja ka klientide ooteaeg on lühem. Kõige selgemalt on näha tööpäevade ja nädalavahetuste erinevused tunnustel: hinnanguline pakutud väljahelistamiste arv (`est_pakutud_tagasihelistamised`), kogu tööaeg (`total_worktime_min`), kogu kõne pikkus (`total_call_length`) ja pakutud kõnede arv (`pakutud_kõned`).

Kui enne muutust on näha, et kuu alguses on suurem sissetulevate kõnede maht ja inimesi on rohkem tööl kui kuu lõpus, siis pärast muudatust see enam nii ei ole. Pärast muutust on pakutud kõnede maht üldiselt suurem ja langeb muudatuse eelsele tasemele alles septembris. Taustatöö (`median_taustat66`) väärtus on seotud ka suurenenud kõnede ja töömahuga, kuid selle mõju avaldub alles hiljem, sest see väärtus saadakse ajaloolistest ehk viimase kolme kuu andmetel pealt.

4.2 Info kõnede mudelil nihke tuvastamine ja leevendamine ADWIN algoritmiga

Kõnekeskuse andmete puhul on parem vältida liiga lühikesi perioode, sest nagu analüüs näitas, siis tihti on kuu alguses rohkem kõnesid ja seda ei tohiks nihke alla liigitada. Kasutatakse ADWIN algoritmi, mis saadakse menelaus paketist¹¹. Valepositiivsete nihete vältimiseks

¹¹https://menelaus.readthedocs.io/en/latest/menelaus.change_detection.html#module-menelaus.change_detection.adwin

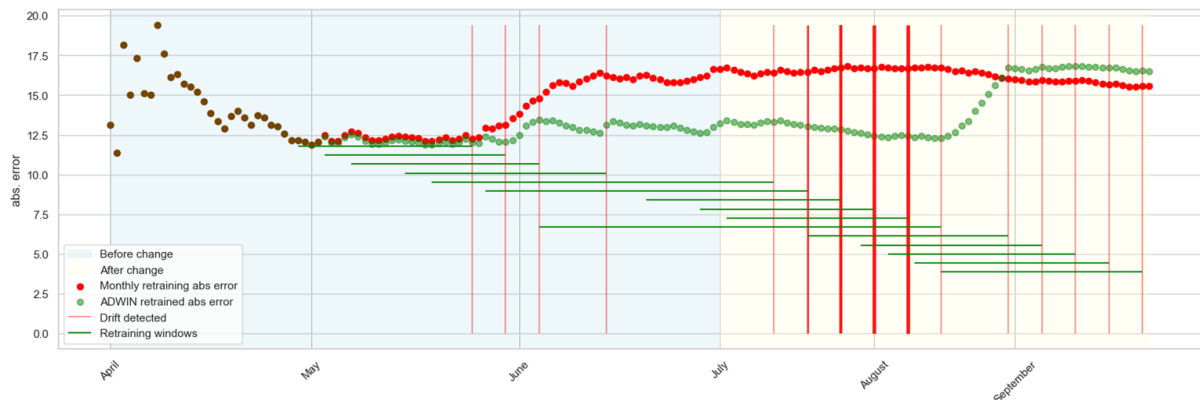
määratakse olulisusnivoo 0,001, minimaalseks akna suuruseks 10 päeva ja minimaalseks alamakna suuruseks 5 päeva. Nihke tuvastamisprotsess käivitatakse iga 5 uue andmepunkti ehk iga 5 päeva tagant.

ADWINi algoritmi katsetatakse kahel viisil, esimesel juhul jälgitakse muutusi mudeli tunnuste jaotuses ja teisel juhul jälgitakse muutusi ennustustäpsuses ehk absoluutses ennustusveas. Mõlemal juhul kasutatakse samu ADWINi parameetrite väärtusi.

Selleks, et jälgida muutusi erinevates tunnustes, peab looma igaühe jaoks eraldi instantsi, sest üks ADWIN algoritmi instants jälgib ainult ühte tunnust. Seega on loodud lahendus sarnane mitmikhüpoteesi testi põhjal nihke tuvastamise algoritmide ülesehitusega, sest jälgitakse muutusi paralleelselt mitmes tunnuses.

Info kõnede mudeli tunnuste muutustest tahetakse teada kõikide tunnuste puhul välja arvatud riigipüha tunnus (is_holiday). Mudeli uuesti treenimise käivitamiseks piisab vaid ühes tunnuses nihke tuvastamisest ja treenimiseks kasutatakse uues aknas olevaid andmeid. Kui nihe tuvastatakse mitmes tunnuses korraga, siis treenimise akna algpunktiks võetakse kõige lähem andmepunkt, sest esiteks võivad eri tunnustel olla eri suuruses aknad ning teiseks, tabatakse sellise lähenemisega kõige paremini muutusi, mis on kiirelt toimunud ja tunnused, mis on stabiilsed olnud, neile on see mõju väike. Kõige kaugema andmepunkti valimisel satuksid treeningandmete hulka ka n.ö. vanad andmed, mis oli enne nihet. Aken lõpeb alati kõige uuema andmepunktiga.

Tulemuste hindamiseks võrreldakse ADWINi abil treenitud info kõnede mudeli ennustatud teenustaseme absoluutset erinevust reaalsest teenustasemest ja igakuiselt treenitud mudeli absoluutset ennustusviga. Igakuiselt treenitud mudeli treeningandmeteks kasutatakse kogu eelnevat olemasolevat ajalugu, sest antud andmestikus hoitakse ajalugu kuni kaks aastat ja töö tegemise hetkel on sellele eelnevast ajast alles umbes aastajagu andmeid.



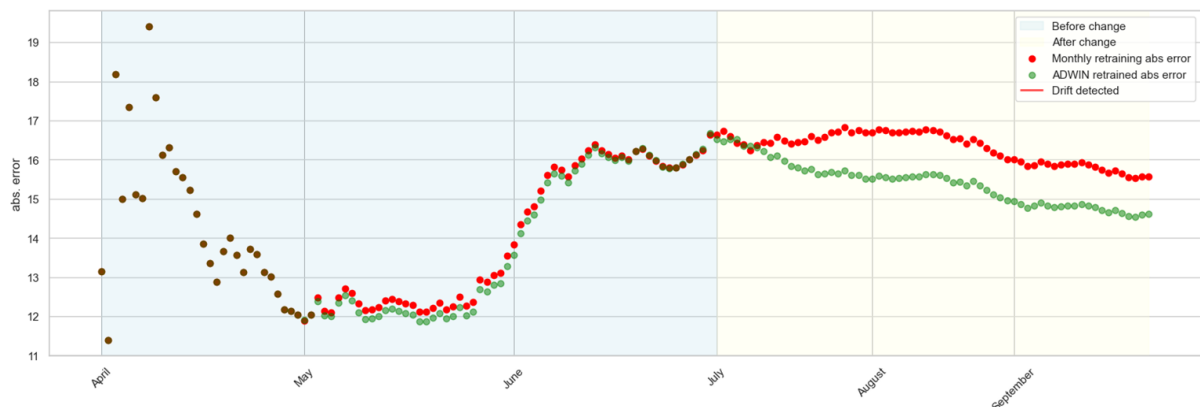
Joonis 4. Ennustuste absoluutne viga ADWIN vs igakuine treenimine vahemikus aprill kuni september, jälgides muutusi mudeli tunnustes

Joonisel 4 on punaste täppidega välja toodud igakuiselt treenitud mudeli ja roheliste täppidega ADWINi abil treenitud mudeli ennustuste absoluutne viga. Punased jooned tähistavad nihke tuvastamist, joone paksust ja värvi tugevust mõjutab nihkunud tunnuste arv. Kui samaaegselt tuvastatakse nihe mitmes tunnuses, siis on joon tugevam. Rohelised vertikaalsed jooned tähistavad andmete vahemiku, mida kasutati teise mudeli uuesti treenimisel. Taustavärvi vahetus tähistab muutuse jõustumist ehk 1. juulit, päeva millal muutus kõnekeskuse tööaeg. Vaadeldud perioodis on igakuiselt treenitud mudeli mediaan MAE 16,4 protsenti ja algoritmi abil uuesti treenitud mudeli mediaan MAE 14,8.

Kõige tihemini tuvastati nihet taustatöö (median_taustat66) tunnuses, enne tööaja muutust oli see ainuke tunnus, mille jaotus muutus tihti. Enne mai lõppu ei tuvastatud üheski tunnuses nihet. Kõige esimene nihe pärast tööaja muutust toimus ooteajast (total_wait_time). Kuigi ADWIN algoritm ei ütle, kuhu suunas ja kui palju nihe toimus, siis tuginedes tunnuste jaotuse graafikutele (Joonis 3) muutus ooteaeg pikemaks. Pärast seda tuvastatakse stabiilselt aina rohkem tunnustes nihkeid iga kord kui nihke tuvastamise protsess käivitub ehk iga viie andmepunkti tagant. Umbes poole juulikuu sees, tuvastatakse nihe hinnangulises väljahelistatud kõnede mahus (est_pakutud_tagasihelistamised) ja ooteaja (total_wait_time) tunnustes. Nende kahe tunnuse tõus on seotud, sest kui klient peab järjekorras liiga kaua ootama, siis ta lõpetab kõne ja agent peab talle ise tagasi helistama. Kuu teises poole täheldatakse nihe lisaks väljahelistamiste mahus ka taustatöös, tööajas (total_worktime_min) ja pakutud kõnede arvus. Kõige rohkem tuvastati nihkeid augustikuu alguses, kus lisaks eelnevale tuvastatakse nihe ka kogu kõneaja (total_call_length) tunnuses. Augustikuu teises pooles olukord stabiliseerub ja tööaja muutuse mõju kaob ning edaspidi muutub ainult taustatöö maht.

ADWINi abil treenitud mudel annab stabiilselt paremaid tulemusi, kuni augustikuu teise pooleni. Augustikuu teises pooles tuvastatakse nihe tunnuses, mis hindab kui mitu kõnet üks kõnekeskuse agent peaks ära vastama (vajalik_kõne_per_person). Siinkohal on tähelepanuväärne, et uuesti treenimiseks kasutatav aken on väga suur võrreldes teiste akendega ja kaasab isegi neid andmepunkte, mis jäävad enne esimest juulit, ilmselt selle tõttu kasvab viga hüppeliselt. Pikem aken on tingitud ilmselt sellest, et antud tunnuses toimus muutus aeglaselt ja seega on valitud aken pikem. Septembrikuus treenitakse mudelit veel mitu korda ja lühemate akendega, sest taustatöö mahu tunnus on väga muutlik, kuid ennustusviga ei vähene.

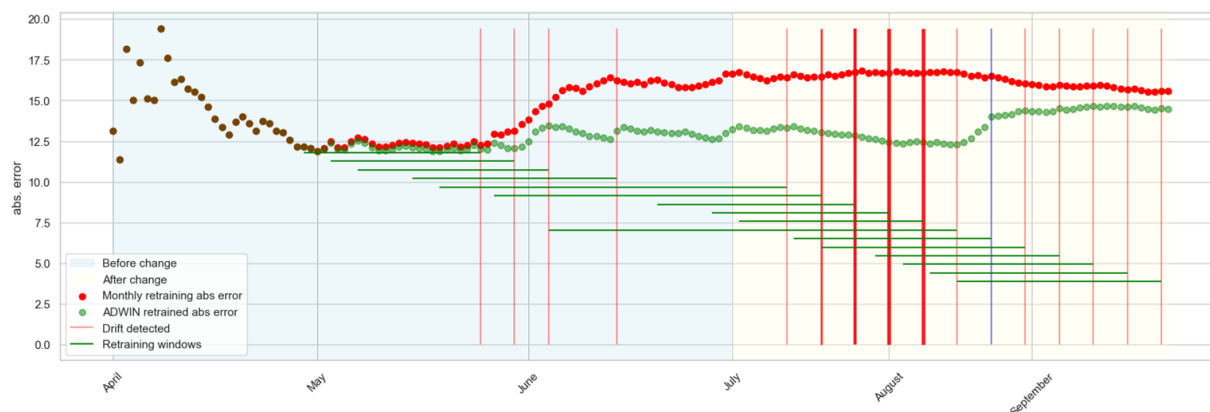
See on ka üks nõrkusi andmejaotusel põhinevatel nihke tuvastamise algoritmidel, et äkilised muutused ennustusveas jäävad tähelepanuta. Seega katsetatakse ADWINit teisel viisil ehk rakendatakse ADWINit ennustusvea jälgimiseks.



Joonis 5. Ennustuste absoluutne viga ADWIN vs igakuine treenimine vahemikus aprill kuni september, jälgides muutusi ennustusveas

ADWIN ei tuvastanud kuskil ennustusveas nihet ja seega ei treenitud mudelit kordagi selle perioodi jooksul uuesti (Joonis 5). Veidi üllatavalt on, et kui enne juulikuu muudatust olid mudelite ennustusvead põhimõtteliselt samad, siis pärast muudatust tõuseb igakuiselt treenitud mudeli ennustusviga. Absoluutse ennustusvea muutusi jälgides on ADWINi abil treenitud mudeli viga stabiilsem ja tõestab ka seda, et eelneva lähenemise puhul nähtud hüpe oli tekitatud pidevast uuesti treenimisest. Kuigi viga on antud juhul stabiilsem, on see siiski keskmiselt kõrgem kui eelneval juhul. Vaadeldud perioodis on igakuiselt treenitud mudeli mediaan MAE

16,4 protsenti ja algoritmi abil uuesti treenitud mudeli mediaan MAE 15,3. Võib olla annab nende kahe lähenemise segu parimad tulemused.



Joonis 6. Ennustuste absoluutne viga ADWIN vs igakuine treenimine vahemikus aprill kuni september, jälgides muutusi nii ennustusveas kui tunnustes. Sinine joon tähistab nihet ennustusveas

Hübriid lähenemise puhul jälgitakse nii tunnuste kui ka ennustusvea muutust. Samade ADWINi parameetrite puhul ennustusviga jälgiv algoritmi instants nihet ei tuvasta (Joonis 6), aga kui see teha veel tundlikumaks ehk akna suuruse ländiks panna 5 ja alamakna suuruse ländiks 2, siis peatatakse ennustusvea hüppeline tõus viie päeva jooksul ja ennustusviga stabiliseerub ning jääb väiksemaks kui igakuiselt treenitud mudeli ennustusviga. See on ka ainukene kord, kus ennustusveas tuvastatakse nihe ja see on joonisel kujutatud sinise joonega. Kõik muud nihke tuvastamised tunnuste puhul jäävad samaks.

See lähenemine annab kõige parema tulemuse, kuid nõuab rohkelt ressursi, sest mudeli nihke tuvastamise protsess käivitatakse iga viie päeva tagant ja absoluutse vea erinevus ei ole nii märkimisväärne, et õigustaks nii sagedast uuesti treenimist. Igakuiselt treenitud mudeli mediaan MAE on 16,4 protsenti ja algoritmi abil uuesti treenitud mudeli mediaan MAE on 13,4.

4.3 Info kõnede mudelil nihke tuvastamine ja leevendamine PCA-CD algoritmiga

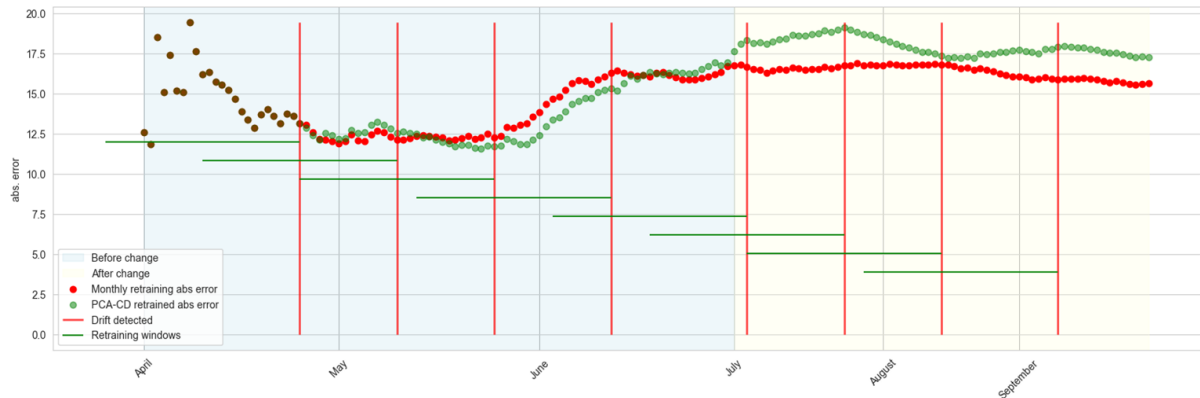
Kuigi ADWIN algoritm andis päris head tulemused, siiski katsetatakse veel teist tüüpi algoritmi, PCA-CD, et näha, kas nihke tuvastamine PCA-ga annab veel paremad tulemused. Kasutatud algoritm on samuti menelaus paketist¹².

Kuna selle algoritmi puhul on akna suurused fikseeritud ja infokõnede andmejaotus võib olla muutlik, siis määratakse akna suuruseks 10 päeva. Lisaks defineeritakse algoritmi jaoks divergentsi meetrikana (ingl *divergence metric*) Kullback-Leibleri divergents, Page Hinkley testi jaoks delta 0,001, PCA jaoks defineeritakse kui palju variatiivsusest peab säilitama 95% ja sammuks ehk kui tihti protsess käivitub 33%, mis on suhteline akna suurusega. Reaalne samm on seega umbes 3,3 päeva. Väiksem akna suurus, samm ja variatiivsuse säilitamise parameetrid annavad paremad tulemused kui suuremad väärtused.

See algoritm jälgib muudatusi mitmes tunnuses korraga. Jälgitavateks tunnusteks on võetud samad, mis ADWINi puhul, sest pole vajadust jälgida tunnuseid, mida ei saa mõjutada väljaspoolt.

¹² https://menelaus.readthedocs.io/en/latest/menelaus.data_drift.html#module-menelaus.data_drift.pca_cd

Nihke tuvastamisel kasutatakse mudeli uuesti treenimiseks eelneva 30 päeva andmeid, sest algoritm ise sobivat vahemiku ette ei anna ja see tundub piisavalt pikk vahemik, et saada tasakaalustatud andmestik, kus on ajalooliste andmete osakaal ei oleks väga suur.



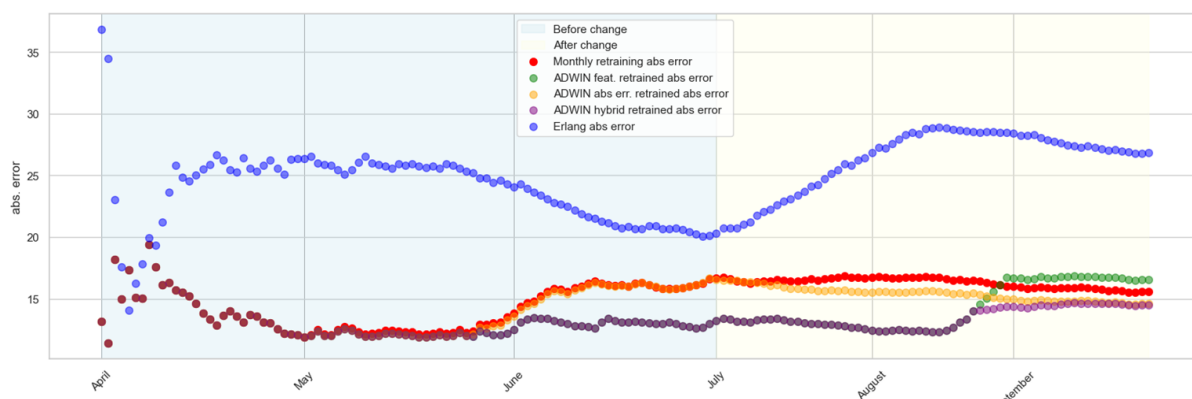
Joonis 7. Ennustuste absoluutne viga PCA-CD vs igakuine treenimine vahemikus aprill kuni september

Joonisel 7 on näha, et PCA-CD abil uuesti treenitud mudeli ennustusviga on suurem, erinevalt eelnevatest katsetest ADWINiga. Kui enne juunikuud oli mudeli ennustustäpsus peaaegu sama hea või parem kui igakuiselt uuesti treenitud mudeli tulemus, siis pärast nihke tuvastamist juulikuus ennustusviga muutus suuremaks ja jäigi nii, olenemata sellest, et mudelit treeniti korduvalt veel uuesti. Mudeli ennustusviga langes stabiilselt peale uuesti treenimist juulikuus teises pooles, kuid hakkas siis tõusma uuesti pärast uuesti treenimist augusti keskpaigas. Vaadeldud perioodis on igakuiselt treenitud mudeli mediaan MAE 16,4 protsenti ja algoritmi abil uuesti treenitud mudeli mediaan MAE 16,9.

Erinevalt ADWINist ei saa selle algoritmi puhul teada, mis tunnus täpselt nihke põhjustas.

4.4 Info kõnede mudelite võrdlus Erlang C-ga

Erlang C valemi ehk kõne ootamise tõenäosuse P_w arvutamiseks jaoks on vaja teada kõneliikluse intensiivsust A , agentide arvu N . P_w arvutatakse iga 15-minutilise intervalli kohta. A arvutamiseks on vaja teada pakutud kõnede arvu intervallis ja AHT -d ehk keskmine kõnele kuluv aeg, mis on summa kõneajast, ootel oleku ajast ja järeltööle kuluvast ajast. Selle arvutamiseks kasutatakse eelneva kolme kuu keskmist. Ka agentide arvu hindamiseks võetakse viimase eelnenud kolme kuu keskmine, kus pakutud kõnede maht jääb sarnasesse suurusjärku, täpsemalt ± 5 kõnet. Päevase teenustaseme saamiseks leitakse iga päeva kõikide intervallide keskmine.



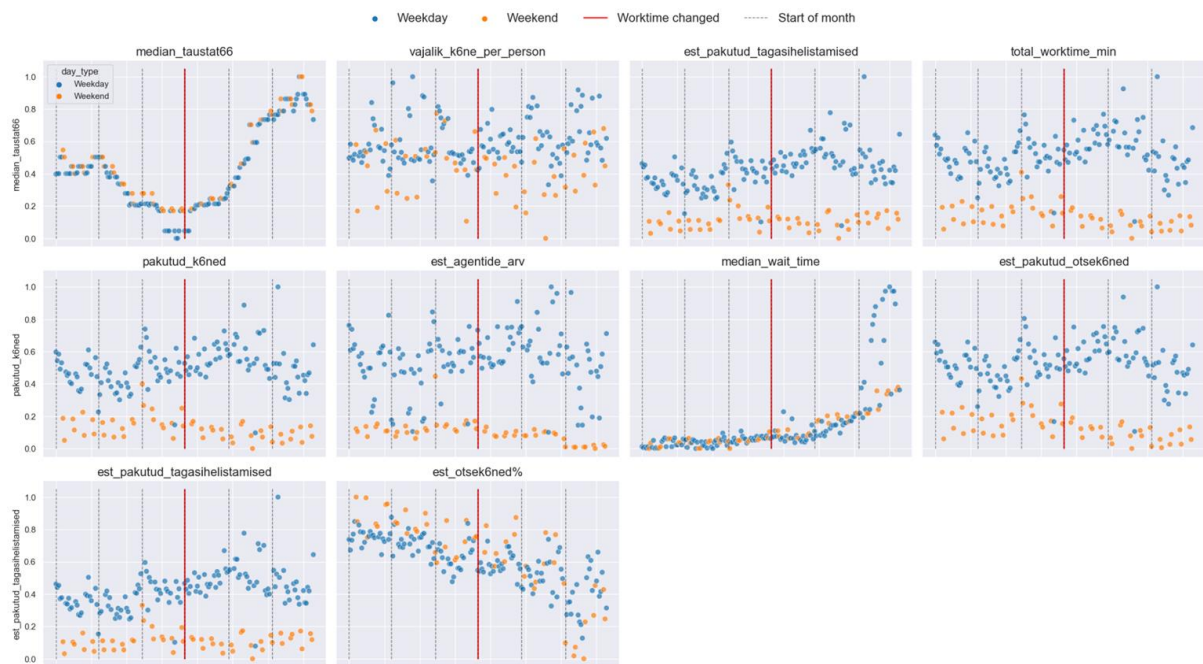
Joonis 8. Ennustuste absoluutne viga ADWIN-id vs igakuine treenimine vs Erlang C vahemikus aprill kuni september

Visuaalse selguse mõttes on andmenihke algoritmidest visualiseeritud ainult parimaid tulemusi andnud algoritmid ja PCA-CD on välja jäetud. Võrreldes masinõppe mudelite tulemusi klassikalise Erlang C valemi tulemustega (Joonis 8), siis on kohe selge, et mudelite ennustustäpsus on parem. Nihke tuvastamise algoritmidest annab kõige paremad tulemused hübriid ADWIN, kuid see on arvutuslikult ka kõige mahukam. Erlang C on arvutuslikult lihtsam, kuid selle täpsus sõltub tugevalt sellest, kui hästi oskab graafikute koostaja hinnata pakutud kõnede arvu igas intervallis, sest siin valemis on kasutatud tegelikke pakutud kõnede arvu. Samas aitab see paremini reaalset töögraafikuid koostada, sest nii saab intervallide kaupa planeerida ja arvutada teenustaset.

4.5 Mõju tehniliste kõnede tunnustele

Tehniliste kõnede liinil on muudatuse mõju väiksem, sest ainuke muudatus oli, et tööaeg on nädalavahetusesti lühem, laupäeviti kümnest viieni ja pühapäeviti kümnest neljani. Sarnaselt info kõnede jaoks on alloleval joonisel (Joonis 9) on tehniliste kõnede mudeli tunnuste jaotused välja toodud vahemikus aprill kuni septembri lõpp 2024 ehk kolm kuud enne ja pärast tööaja muutust, mis on eristatavad taustavärviga. Visualisatsioonist on välja jäetud tunnused, mida muudatused ei mõjuta nagu näiteks kuupäevaga seotud veerud, arvete väljasaatmisest möödunud päevade arv (`days_since_bills_sent`) ja *boolean* väärtustega veerud.

Kuna suurimad erinevused on tööpäevade ja nädalavahetuse vahel, siis ei eristata joonisel kõiki päevi, vaid ainult neid kahte. Tunnuste väärtused on parema visualiseeringu ja andmekaitse eesmärgil normaliseeritud nulli ja ühe vahele *MinMaxScaler*'iga.



Joonis 9. Tehniliste kõnede mudeli tunnuste jaotused aprill kuni september 2024

Sarnaselt info kõnedele on enamike tunnuste puhul on näha, et laupäeviti ja pühapäeviti on tunnuste väärtused väiksemad kui tööpäevadel (Joonis 9). Nädalavahetusel on sissetulevate kõnede maht väiksem, mis omakorda tähendab seda, et ka vajaminev töömaht on väiksem ja ka klientide ooteaeg on lühem. Kõige selgemalt on näha tööpäevade ja nädalavahetuste erinevused tunnustel: hinnanguline pakutud väljahelistatud kõnede arv (`est_pakutud_tagasihelistamised`), kogu tööaeg (`total_worktime_min`), kogu kõne pikkus (`total_call_length`) ja pakutud kõnede arv (`pakutud_kõned`).

Tunnus, mis hindab mitu kõne üks agent peaks vastama (`vajalik_kõne_per_person`) on hinnanguline näitaja, mis sõltuvalt sissetulevate kõnede ja olemasolevate agentide arvu põhjal hindab ühe agendi töökoormust. Enne muudatust on see rohkem kontsentreeritud, kuid pärast muudatust on selle väärtused rohkem hajusad, mis võib tähendada, et graafikute planeerimine ei ole enam nii täpne kui enne. Väiksem koormus on tihti nädalavahetustel ja riigipühadel. Näiteks pakutud kõnede graafikul on näha nädalavahetuse täppide hulgas kolme sinist täppi, millest esimene tähistab 1. maid, teine 24. juunit ja kolmas 20. augustit.

Enne tööaja muutust on tööpäevade ja nädalavahetuse pakutud kõnede hulga vahe väiksem ja tööpäevade kõnede maht on juulis tõusvas joones ja hakkab langema kiirelt augusti alguses.

Kiire taustatöö (`median_taustat66`) suurenemine on seotud ilmselt suurenenud kõnede ja töömahuga juunis, kuid selle mõju avaldub alles hiljem, sest see väärtus saadakse retrospektiivselt nagu ka mediaan ooteaeg (`median_wait_time`).

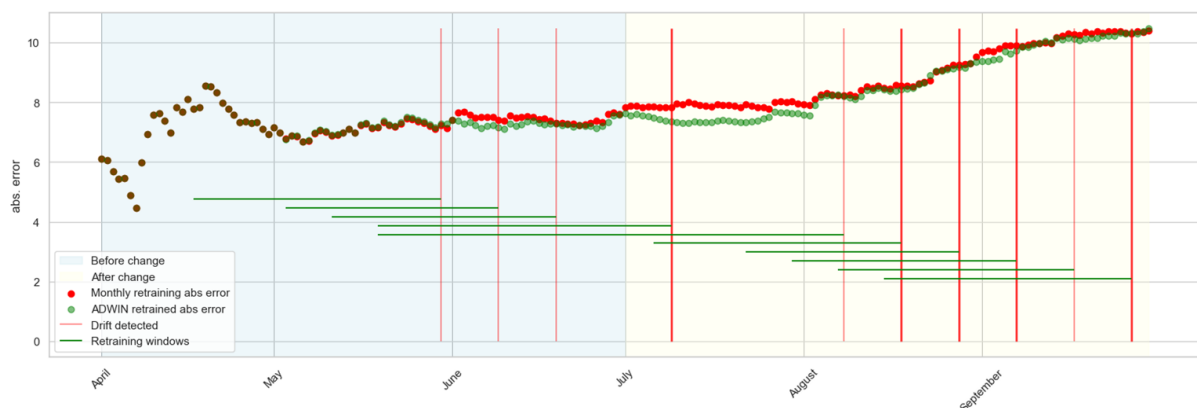
4.6 Tehniliste kõnede mudelil nihke tuvastamine ja leevendamine ADWIN algoritmiga

Sarnaselt info kõnede mudeliga, valitakse ka siin pikemad ajavahemikud nihke hindamiseks, sest muutus tehniliste kõnede liinil on väiksem kui info kõnede omal. Kasutatakse samuti ADWIN algoritmi menelaus pakettist. Valepositiivsete nihete vältimiseks määratakse olulisusnivoo 0,001, minimaalseks akna suuruseks 21 päeva ja minimaalseks alam-akna suuruseks 10 päeva. Nihke tuvastamisprotsess käivitatakse iga 10 andmepunkti tagant.

Sarnaselt eelnevale katsetatakse ka siin ADWINi algoritmi kolmel viisil, esimesel juhul jälgitakse muutusi mudeli tunnuste jaotuses, teisel juhul jälgitakse muutusi ennustustäpsuses ehk absoluutses ennustusveas ja kolmandal juhul jälgitakse muutusi mõlemas. Kõikidel juhtudel kasutatakse samu ADWINi parameetrite väärtusi.

Tehniliste kõnede mudeli tunnuste muutustest tahetakse teada kõikide tunnuste puhul välja arvatud need, mida muudatused ei mõjuta nagu näiteks kuupäevaga seotud veerud, arvete väljasaatmisest möödunud päevade arv (*days_since_bills_sent*) ja *boolean* väärtustega veerud. Mudeli uuesti treenimise käivitamiseks piisab vaid ühes tunnuses nihke tuvastamisest ja treenimiseks kasutatakse uues aknas olevaid andmeid. Kui nihe tuvastatakse mitmes tunnuses korraga, siis treenimise akna algpunktiks võetakse kõige lähem andmepunkt, sest esiteks võivad eri tunnustel olla eri suuruses aknad ning teiseks, tabatakse sellise lähenemisega kõige paremini muutusi, mis on kiirelt toimunud ja tunnused, mis on stabiilsed olnud, neile on see mõju väike. Kõige kaugema andmepunkti valimisel satuksid treeningandmete hulka ka n.ö. vanad andmed, mis oli enne nihet. Aken lõppeb alati kõige uuema andmepunktiga.

Tulemuste hindamiseks võrreldakse ADWINi abil treenitud tehniliste kõnede mudeli ennustatud teenustaseme absoluutset erinevust reaalsest teenustasemest ja igakuiselt treenitud mudeli absoluutset ennustusviga. Igakuiselt treenitud mudeli treeningandmeteks kasutatakse kogu eelnevat olemasolevat ajalugu, antud andmestikus hoitakse ajalugu kuni kaks aastat. Töö tegemisel hetkel eelneb sellele ligi aastajagu andmeid.



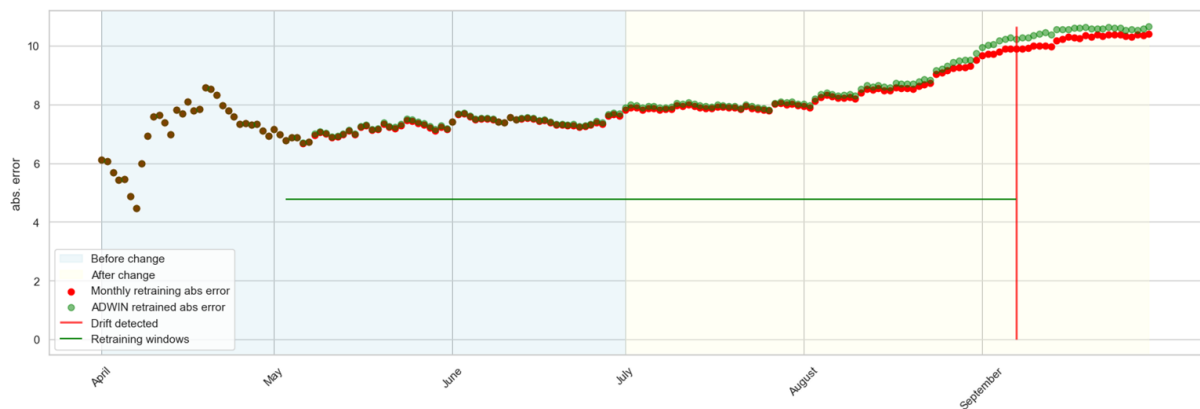
Joonis 10. Ennustuste absoluutne viga ADWIN vs igakuine treenimine vahemikus aprill kuni september, jälgides muutusi mudeli tunnustes

Joonisel 10 on punaste täppidega välja toodud igakuiselt treenitud mudeli ja roheliste täppidega ADWINi abil treenitud mudeli ennustuste absoluutne viga. Hallid vertikaalsed jooned tähistavad kuu vahetust. Punased jooned tähistavad nihke tuvastamist, joone paksust ja värvi tugevust mõjutab nihkunud tunnuste arv. Kui samaaegselt tuvastatakse nihe mitmes tunnuses, siis on joon tugevam. Rohelised vertikaalsed jooned tähistavad andmete vahemiku, mida kasutati teise mudeli uuesti treenimisel. Taustavärvide vahetus tähistab 1. juulit ehk päeva, millal muutus kõnekeskuse tööaeg.

Nihkeid tuvastataksegi peamiselt ainult taustatöös ja ooteajas, mis on ootuspärane, arvestades nende tunnuste jaotust. Ainukene kord, kus tuvastatakse nihet mõnes teises tunnuses on juuni keskel pakutud väljahelistamiste kõnede arvus (*est_pakutud_tagasihelistamised*). Enne muudatust toimusid nihked ainult taustatöös, kuid pärast muudatust ka juba ooteajas. See simulatsioon kinnitab eeldust, et tehniliste kõnede liini töömahtu tööaja muutus väga ei mõjutanud, kuigi ennustusviga hakkab kasvama augustikuus.

Kuigi juulikuus on ADWINi abil treenitud mudeli ennustustäpsus veidi parem, siis kahe mudeli ennustustäpsuse vahe pole ka piisav, et õigustada nii sagedast mudeli uuesti treenimist.

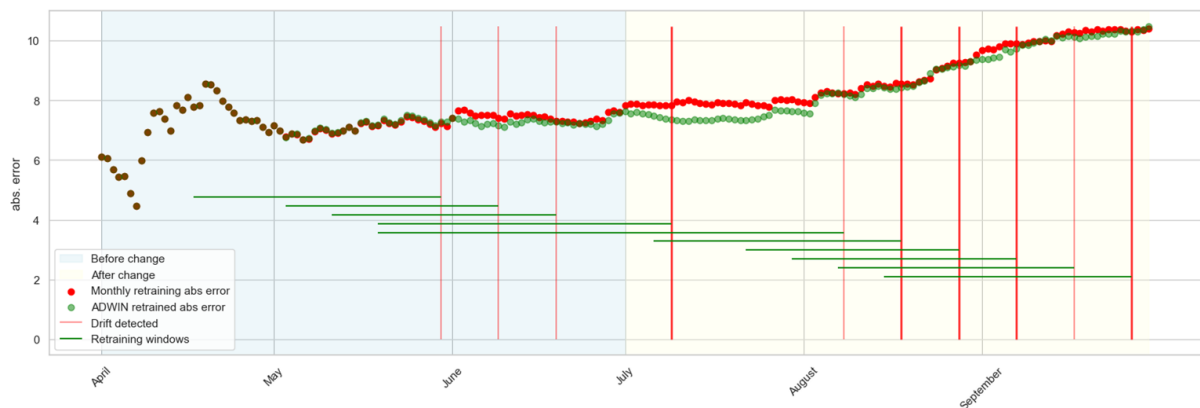
Ennustusviga on mõlemal mudelil päris madal, joonise põhjal eksib mudel maksimaalselt +/- 10 protsendiga. See võib tähendada, et tehniliste kõnede liini toimimine on stabiilsem ja seega paremini mudeldatav. Vaadeldud perioodis on igakuiselt treenitud mudeli mediaan MAE 7,8 protsenti ja algoritmi abil uuesti treenitud mudeli mediaan MAE 7,7.



Joonis 11. Ennustuste absoluutne viga ADWIN vs igakuine treenimine vahemikus aprill kuni september, jälgides muutusi ennustusveas

Jälgides ainult mudeli ennustusviga (Joonis 11), siis tuvastatakse nihe ainult septembri alguses ja uuesti treenimine seda paremaks ka ei tee. Katsetades ka väiksemate akende vahemikega teistsuguseid tulemusi ei saa. Antud juhul annab parema tulemuse igakuine mudeli treenimine, kuid joonise põhjal saab väita, et mudelit võib treenida isegi harvemini, sest mudeli vigade vahe on väga väike. Vaadeldud perioodis on igakuiselt treenitud mudeli mediaan MAE 7,8 protsenti ja algoritmi abil uuesti treenitud mudeli mediaan MAE 7,8.

Hübriid lahenduse kasutamine paremaid tulemusi ei anna (Joonis 12) ja nii joonis kui ka tuvastatud nihked jäävad samaks nagu esimese lähenemise puhul. Ennustusveas nihet ei tuvastata, isegi kui teha see veel tundlikumaks: akna suuruse lävend 7, alam-akna suuruse lävend 3 ja nihke tuvastamisprotsessi käitatakse iga 5 andmepunkti tagant. Vaadeldud perioodis on igakuiselt treenitud mudeli mediaan MAE 7,8 protsenti ja algoritmi abil uuesti treenitud mudeli mediaan MAE 7,7.



Joonis 12. Ennustuste absoluutne viga ADWIN vs igakuine treenimine vahemikus aprill kuni september, jälgides muutusi nii ennustusveas kui tunnustes

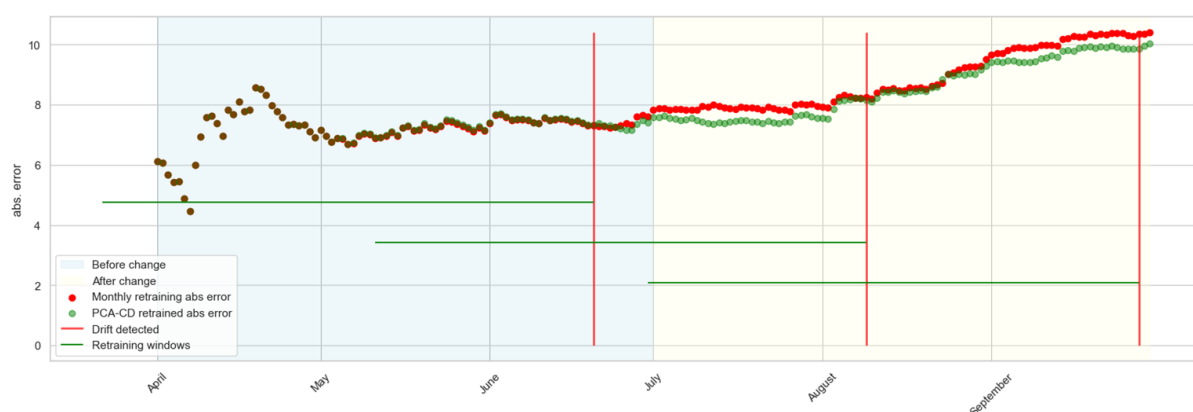
4.7 Tehniliste kõnede mudelil nihke tuvastamine ja leevendamine PCA-CD algoritmiga

Kuna ADWIN algoritm ei andnud erilisel häid tulemusi, siis katsetatakse veel PCA-CD algoritmi, mis hindab andmeid PCA alusel, et näha, kas teistsugune lähenemine annab paremaid tulemusi.

Kuna selle algoritmi puhul on akna suurused fikseeritud ja tehniliste kõnede andmejaotus on stabiilsem, siis määratakse akna suuruseks 30 ehk umbes üks kuu ja ka protsessi käivitamise samm on pikem. Lisaks defineeritakse algoritmi jaoks divergentsi meetrikana (ingl *divergence metric*) Kullback-Leibleri divergents, Page Hinkley testi jaoks delta 0,001, PCA jaoks defineeritakse kui palju variatiivsusest peab säilitama 99% ja sammuks ehk kui tihti protsess käivitub 33%, mis on suhteline akna suurusega. Reaalne samm on seega umbes 10 päeva.

See algoritm jälgib muutusi mitmes tunnuses korraga. Jälgitavateks tunnusteks on võetud samad, mis ADWINi puhul, sest pole vajadust jälgida tunnuseid, mida ei saa mõjutada väljaspoolt.

Nihke tuvastamisel kasutatakse mudeli uuesti treenimiseks eelneva kolme kuu andmeid, sest algoritm ise sobivat vahemiku ette ei anna ja see tundub piisavalt pikk vahemik, et saada tasakaalustatud andmestik, kus ajalooliste andmete osakaal ei oleks väga suur.

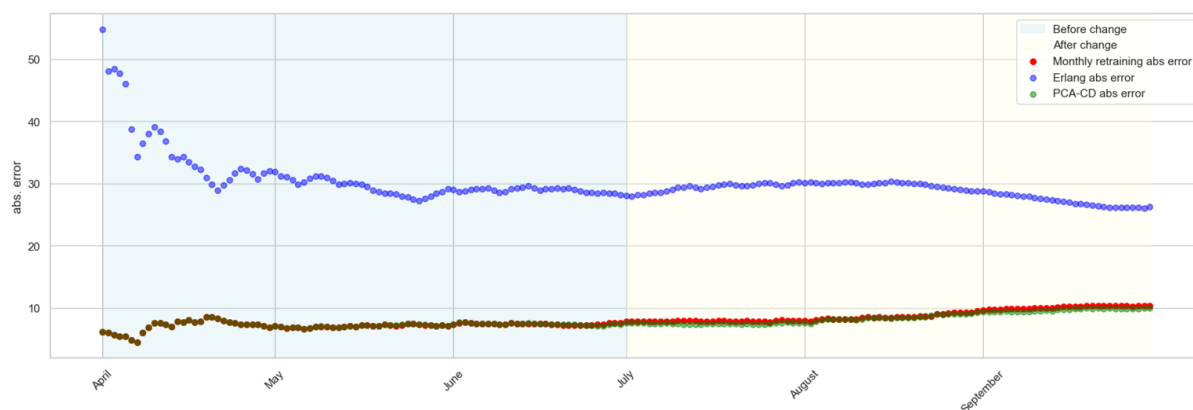


Joonis 13. Ennustuste absoluutne viga PCA-CD vs igakuine treenimine vahemikus aprill kuni september

PCA-CD annab parema tulemuse kui ADWIN (Joonis 13), sest on stabiilselt veidi parem kui igakuiselt treenitud mudel, eriti just juulis ja septembrikuus. Nihet tuvastatakse kolmel korral ning peale tööaja muutust alles augustikuu alguses. Vaadeldud perioodis on igakuiselt treenitud mudeli mediaan MAE 7,8 protsenti ja algoritmi abil uuesti treenitud mudeli mediaan MAE 7,65.

4.8 Tehniliste kõnede mudelite võrdlus Erlang C-ga

Erlang C valemi ehk kõne ootamise tõenäosuse P_w arvutamiseks jaoks on vaja teada kõneliikluse intensiivsust A , agentide arvu N . P_w arvutatakse iga 15-minutilise intervalli kohta. A arvutamiseks on vaja teada pakutud kõnede arvu intervallis ja AHT -d ehk keskmine kõnele kuluv aeg, mis on summa kõneajast, ootel oleku ajast ja järeltööle kuluvast ajast. Selle arvutamiseks kasutatakse eelneva kolme kuu keskmist. Ka agentide arvu hindamiseks võetakse viimase eelnenud kolme kuu keskmine, kus pakutud kõnede maht jääb sarnasesse suurusjärku, täpsemalt +/-5 kõnet. Päevase teenustaseme saamiseks leitakse iga päeva kõikide intervallide keskmine.



Joonis 14. Ennustuste absoluutne viga PCA-CD vs igakuine treenimine vs Erlang C vahemikus aprill kuni september

Kuna algoritmide tulemused on väga sarnased, siis visuaalse selguse mõttes on andmenihke algoritmidest visualiseeritud ainult parima tulemuse andnud algoritm, PCA-CD (Joonis 14). Erlang C ennustuste viga on peaaegu kolmekordne, kuid erinevalt mudelist selle ennustustäpsus paraneb septembris. Andmenihke algoritmid küll tuvastasid nihkeid, kuid võrreldes igakuiselt treenitud mudeliga, need paremaid tulemusi ei andnud. See võib tähendada, et kas nihe oli minimaalse mõjuga või mudel tuleb ise hästi toime muutustega, ilma, et peaks liigselt sekkuma.

5 Järeldused

Andmenihke algoritmide puhul pole ühte universaalset parimat lahendust, kahe erineva andmestiku puhul oli selgelt näha, et kui info kõnede andmestikule sobis hästi ADWIN algoritm, siis tehniliste kõnede andmestikul see nii häid tulemusi ei andnud. Sama oluline on ka jälgitava parameetri valimine ja tasub läbi mõelda, kuidas soovitakse nihet tuvastada, kas jälgida mingisugust meetrikat nagu näiteks veamäär või sisendandmete jaotust. ADWINit rakendades saab soovi korral jälgida ka mõlemat.

Sisendandmete jälgimisel jääb tähelepanuta veamäära hüppeline tõus ja samas veamäära alusel nihke tuvastamisel on keeruline öelda, kus muutus toimus ja seega ei tea kui suurt akent valida uuesti treenimiseks. Sellistes olukordades on parem kasutada algoritme, mis valivad akna suuruseid dünaamiliselt nagu ADWIN, sest see defineerib akna suuruse vastavalt olukorrale. Fikseeritud akna suuruste kasutamine annab parema kontrolli nihke tuvastamise protsessi üle, aga samas eeldab väga häid teadmisi andmete domeenist ja nende muutlikkusest.

5.1 Andmenihe info kõnede andmetes

Info kõnede andmetele sobis antud nihke olukorras kõige paremini ADWINi algoritm, kus jälgiti korraga nii tunnuste jaotuse kui ka veamäära muutust. Ainult tunnuste jaotust jälgides tekkis olukord, kus algoritmi käivitatud sage uuesti treenimine andis hoopiski vastupidise efekti ja ennustusviga kasvas hüppeliselt. Jällegi veamäära jälgimisel ei tuvastatud üldse nihet ja seega mudelit ei treenitud uuesti mitu kuud, mis andis parema küll tulemuse kui igakuine treenimine, kuid oli siiski kehvem kui hübriid lahendus. Samas nõudis see lahendus vähem ressursi ja kuna sisendandmete väärtused on muutlikud, siis see hoidis ära olukorra, kus sagedasel uuesti treenimisel mudeli tulemused muutuvad hoopiski halvemaks nagu juhtus ainult tunnuste jälgimisel.

Katsetatud algoritmide tulemuste põhjal saab öelda, et info kõnede andmetes toimus andmenihe. Eelkõige pakutud kõnede mahu, ooteaja ja kogu kõne pikkuse alusel on näha, et nädalavahetusel tegemata jäänud kõned kandusid üle tööpäeva sisse. Võrreldes sellele eelneva kolme kuuga tõusid nende tunnuste väärtused hüppeliselt ehk tegemist oli järsu andmenihkega.

5.2 Andmenihe tehniliste kõnede andmetes

Tehniliste kõnede andmetel andis paremaid tulemusi PCA-CD kuigi MAE parandus oli minimaalne, mis võib olla märk sellest, et tööaja muutus tehniliste kõnede liinil mingisugust järsku andmenihet kaasa ei toonud. Ainukene tunnus, milles tihti nihet tuvastati oli kogu tööaeg ehk tegemist võis olla ka järkjärgulise või inkrementaalse nihkega. Algoritmide abil treenitud mudelite ennustustäpsus oli peaaegu sama hea igakuiselt treenitud mudeliga. ADWINiga ennustusviga jälgides, tuvastatakse ja treenitakse mudelit uuesti ainult ühe korra kuue kuu jooksul, septembris ja joonise põhjal on näha, et selle mudeli tulemused on sama head kui igakuiselt ümber treenitud mudeli omad.

Samas ei tohi jätta tähelepanuta asjaolu, et osasid tunnuseid mõjutab ka hooajalisus. Näiteks suvel oli ühel päeval korraga rohkem agente tööl kui näiteks kevadel ja ka töötunde oli rohkem. See võib olla tingitud sellest, et suveperioodil ollakse vähem haiged.

5.3 Mudelite rakendamine teistes kõnekeskustes

Loodud mudeleid saab kasutusele võtta sarnase profiiliga kõnekeskustes ehk väiksemates ja keskmise suurusega kõnekeskustes, kus agentide arv päevas jääb 5-50 vahele ja kuna iga liini jaoks on tehtud eraldi mudel, siis peaks liinide arv jääma ka mõistlikkuse piiridesse. Suuremates kõnekeskustes võib lisaks tekkida mingisuguseid olulisi tunnuseid, mis antud

kõnekeskuses ei ole olulised ja seega pole nendega ka arvestatud. Kõnekeskustes tekkivad andmed on üldjuhul sarnased olenemata domeenist, kuid erinevaid andmeid on siiski palju ja tuleb leida üles need olulised ehk siin töös loodud mudelid ja nende tunnused võiksid olla abiks mudeli alustamisega. Kindlasti ei saa neid mudeleid kasutusele võtta täpselt selliste tunnustega nagu on kasutatud siin. Isegi sama kõnekeskuse eri liinidel on eri parameetrid erineva olulisusega. See töö võiks aidata mõista kõnekeskuse andmete seoseid ja nende olulisust. Näiteks mõnes teises sarnases kõnekeskuses võib olla väga oluline tunnus hüljatud kõned, mida nendes mudelites ei kasutata.

Eelkõige tuleb mudeli tegemisel mõelda äripoolse soovidele ja kõnekeskuse vajadustele, sest nende mudelite puhul eeldatakse, et seda kasutab kogunud kõnekeskuse töögraafikute koostaja, kes oskab hinnata sissetulevate kõnede mahtu ja töömahtu. Lihtsaks kasutamiseks võiks mudel olla kasutajale kättesaadav läbi kliendiliidese, kuhu saab lihtsasti sisestada muutujad ja tagastatakse ennustatav teenustase.

Veel mugavamaks kasutamiseks ja et näha suuremat pilti võiks mudel sisestatud andmete põhjal tagastada ühe päeva asemel näiteks terve kuu ennustatavad teenustasemed. See ei võta siiski ära võimalust päevade parameetreid ükshaaval muutmast, vaid pigem annab alguspunkti, et ei peaks iga päeva eraldi sisse toksima. Kasutaja saab iga päeva jaoks muuta nii kohustuslike sisendandmeid kui ka neid parameetreid, mis lisatakse taustal. Nii on kasutajal eelkõige läbipaistvus, kuidas mudel otsuseid teeb, kuid soovi korral saab neid ka muuta, kui tundub, et mõni tunnuse väärtus pole päris õige.

Kokkuvõte

Magistritöö üks eesmärk oli luua mudel teenustaseme ennustamiseks ja katsetada olukorda, kus kõnekeskus muudab oma tööaega, võrrelda erinevate algoritmide abil treenitud mudelite tulemusi teenuse kvaliteedi ennustamisel ja analüüsida, kuidas erinevad mudelid toimivad sellistes erakordsetes olukordades.

Kahele erinevale liinile sobis ElasticNet mudel, kuid tehti kaks eri mudelit, sest eri liinidel olid olulised erinevad tunnused ja ka ennustustäpsus oli eraldi parem. Mudelite sobivust hinnati MSE, R^2 -skoori ja MAE abil.

Katsetused erinevate andmenihke algoritmidega näitasid, et info kõnede liini andmetel andis kõige paremaid tulemusi ADWIN andmenihke algoritm, mille abil treenitud mudel oli parem nii klassikalisest Erlang C arvutusest ja igakuiselt uuesti treenitud mudelist. Kõige paremaid tulemusi andis algoritm siis, kui jälgiti nii tunnuste jaotuse muutust kui ka veamäära. Perioodis kolm kuud enne ja kolm kuud pärast muutust oli kuue kuu mediaan MAE 16,4 protsenti ja algoritmi abil uuesti treenitud mudeli mediaan MAE 13,4.

Tehniliste kõnede liini andmetele sobis kõige paremini PCA-CD algoritm, mis jälgis ainult tunnuste jaotuse muutust. Kuigi see andis paremaid tulemusi kui Erlang C valem ja igakuiselt uuesti treenitud mudel, siis ei olnud täpsuse vahe piisav, et õigustaks selle algoritmi kasutuselevõttu. Sellise muutuse jaoks piisab ka igakuisest treenimisest. Perioodis kolm kuud enne ja kolm kuud pärast muutust oli kuue kuu mediaan MAE igakuiselt treenitud mudelil 7,8 protsenti ja algoritmi abil uuesti treenitud mudeli mediaan MAE oli 7,65.

Jälgitavate parameetrite valik on sama oluline kui algoritmi valik, sest vale algoritmi ja parameetrite kombinatsioon võib mudeli treenida hoopiski halvemaks. Samamoodi tuleb arvestada, et lühemad algoritmi vaatlusaknad ja ajalugu, millega mudelit uuesti treenitakse, teeb algoritmi muutustele tundlikumaks. Aktsepteeritava lahenduse leidmiseks tuleb mõista kõnekeskuse tööd ja eelnevalt analüüsida andmeid, et mitte klassifitseerida nihkeks tavapäraseid muutusi andmetes.

Teine eesmärk oli analüüsida muutuse mõju andmetele ja teha kindlaks, kas antud muudatus tõi kaasa ka andmenihke ja kui suure ning mida oleks vaja, et mudelid suudaksid inimsekkumiseta andmete muutustega kohaneda. Info kõnede kanali andmetel kasutatud algoritmid näitasid, et seal toimus järsk andmenihe, kus muutuse mõju avaldub järkjärguliselt kuu aja jooksul. Eelkõige sai mõjutatud pakutud kõnede maht, ooteaeg ja kogu kõne pikkus, mille väärtused tõusid nähtavalt võrreldes muutusele eelneva kolme kuuga. Nädalavahetusel tegemata jäänud kõned kandusid tööpäeval sisse.

Tehniliste kõnede kanali andmetes nii tugevat andmenihet ei tuvastatud. Kõnekeskuse tööaja mõju sellele kõneliinile oli väiksem ja ainuke tunnus, kus muutus tuvastati oli kogu tööaeg. Selle tunnuse nihe võib olla seotud tööaja muutusega, kuid peab arvestama ka sellega, et seda tunnust mõjutab ka hooajalisus. Suvisel perioodil ollakse vähem haiged ja tänu sellele võib ka agentide arv ühel päeval suurem olla.

Loodud mudelid ei sobi ilmselt teistes kõnekeskustes üks-ühele samade tunnustega kasutusele võtta, vaid tuleks siiski enne analüüsida enda kõnekeskuse iseloomu, kuid need võiksid olla heaks alguspunktiks. Kõnekeskuste andmete struktuur on üldjuhul sarnane, kuid erinevates kõnekeskustes võivad olla mõned parameetrid erineva tähtsusega.

Viidatud kirjandus

- [1] G. M. Koole ja S. Li, “A Practice-Oriented Overview of call center workforce Planning,” *Stochastic Systems*, vol. 13, no. 4, pp. 479–495, juuli 2023, doi: 10.1287/stsy.2021.0008.
- [2] T. R. Robbins, “Evaluating the performance of the Erlang models for call centers,” *International Journal of Applied Science and Technology*, vol. 9, no. 1, p. 1, jaanuar 2019, doi: 10.30845/ijast.v9n1p1.
- [3] C. Yüzüak, F. Nurdağ, ja M. Tartuk, “Optimization of call center agent resources using various machine learning methods: A Systematic review,” *AIS Electronic Library (AISeL)*. <https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1020&context=confirm2024>
- [4] W. Kumwilaisak, S. Phikulngoen, J. Piriataravet, N. Thatphithakkul ja C. Hansakunbuntheung, “Adaptive call center workforce management with deep neural network and reinforcement learning,” *IEEE Access*, vol. 10, pp. 35712–35724, jaanuar 2022, doi: 10.1109/access.2022.3160452.
- [5] S. Li, Q. Wang ja G. Koole, “Predicting Call Center Performance with Machine Learning,” in *Springer proceedings in business and economics*, 2018, pp. 193–199. doi: 10.1007/978-3-030-04726-9_19.
- [6] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama ja G. Zhang, “Learning under Concept Drift: A Review,” *IEEE Transactions on Knowledge and Data Engineering*, p. 1, jaanuar 2018, doi: 10.1109/tkde.2018.2876857.
- [7] A. A. Qahtan, B. Alharbi, S. Wang ja X. Zhang, “A PCA-Based Change Detection Framework for Multidimensional Data Streams,” *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, august 2015, doi: 10.1145/2783258.2783359.
- [8] A. Bifet ja R. Gavaldà, “Learning from Time-Changing Data with Adaptive windowing,” *Proceedings of the 2007 SIAM International Conference on Data Mining*, aprill 2007, doi: 10.1137/1.9781611972771.
- [9] G. Koole, S. Li ja S. Ding, “Call center data analysis and model validation,” *Queueing Systems*, vol. 109, no. 1, veebruar 2025, doi: 10.1007/s11134-025-09935-4.

Lisad

I. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Desiree Himuškin,

(autori nimi)

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

Andmenihke tuvastamine ja leevendamine kõnekeskuse andmete näitel,

(lõputöö pealkiri)

mille juhendajad on

Janika Aan ja Anna Aljanaki,

(juhendajate nimed)

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commons'i litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Desiree Himuškin

15.05.2025