TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Erki Toom 210729IAIB

Töökuulutuste reklaamide tulemuse ennustamine

Bakalaureusetöö

Juhendaja: Ants Torim
PhD

Tallinn 2024

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: Erki Toom

27.05.2024

Annotatsioon

Antud bakalaureusetöö eesmärgiks on luua ennustusmudelid, mis võimaldavad lähteandmete põhjal ennustada reklaami tulemust kahes reklaamkanalis eraldi. Kummagi reklaamkanali mudelisse laetakse sisse samad reklaamide andmed ning saadavalolevate atribuutide põhjal koostatakse mudel, mis ennustab tulemuse.

Töös kasutatakse CV.ee tööportaali vahendusel loodud Meta ja Google reklaamide andmeid. Reklaamide kuju on standardne ehk kõikidel reklaamidel on sarnane ülesehitus. Mitmest allikast kokku kogutud andmed viidi ühtsele kujule selleks loodud Python rakenduses ning salvestati andmebaasi. Andmebaasist võetud andmeid töödeldi edasi Jupyter Notebookis. Andmete paremini mõistmiseks need visualiseeriti ning seejärel töödeldi sobivale kujule erinevate ennustusmudelite katsetamiseks, et leida sobivaim.

Kesksel kohal töös on ametinimede vektorkujule teisendamine, et mudelis oleks võimalik ka ametinime arvesse võtta. Vektoriks teisendamisel kasutati saadavalolevaid keeletöötluse tööriistu.

Lõputöö väljund võimaldab tulevikus otsustada kumb reklaamkanal annab parema tulemuse kindla reklaami korral ning seeläbi võimaldab reklaami eelarvet optimeerida, suunates see ennustuse põhjal paremini toimivasse reklaamkanalisse

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 24 leheküljel, 7 peatükki, 12 joonist, 3 tabelit.

Abstract

Predicting the results of advertisements of job ads

The aim of this thesis is to create predictive models for two advertisement channels, which will predict the results of the advertisements based on the source data.. The data from the same ads is used for both channels and, based on the available attributes, prediction models are created for each of the channels.

The research paper uses data from Meta and Google ads created through the CV.ee job portal. The advertisements are created in a standardized way, which means all the advertisements have a similar structure. A Python application was created to unify the data collected form multiple sources. This data was stored in a database. Data was then exported from the database into a Jupyter Notebook for further processing. To better understand the data, it was visualized and then processed into a suitable form to try out different prediction models in order to find the most suitable one.

A key part of the thesis is the conversion of job titles into vector form, so that the job title can be used as an attribute in the prediction model. Publicly available language processing tools were used for the conversion to vector format.

The output of the thesis makes it possible to later decide which advertising channel will give the better result for a specific advertisement, thus allowing to optimize the advertising budget by directing it to a better performing advertising channel based on the prediction

The thesis is in estonian and contains 24 pages of text, 7 chapters, 12 figures, 3 tables.

Lühendite ja mõistete sõnastik

API	*Application Programming Interface*
CBOW	*Continuous Bag of Words*
CRISP-DM	*Cross-Industry Standard Process for Data Mining*
CSV	*Comma-separated values*
Id	*Identifier*
K-means	*Method of vector quantization*
MSE	*Mean Squared Error*
NLP	*Natural language processing*
R^2	*Coefficient of determination*
SQL	*Structured Query Language*
Skip-gram	loomuliku keele töötlemisel (*NLP*) kasutatav algoritm
vektor	1-mõõtmeline arvude massiiv

Sisukord

Jooniste loetelu

Tabelite loetelu

1 Sissejuhatus

Enamus tööealisi inimesi on kasutanud tööportaali abi [1] uue ametikoha leidmisel. Tööportaali kasutuspõhimõte on lihtne – tööandjad sisestavad kuulutuse, et täita saadavalolev ametipositsioon ning tööotsijad sirvivad avaldatud kuulutusi. Portaal võimaldab tööotsijal sobiva kuulutuse leidmisel sellele kandideerida ning tööandjal seejärel kandideerijate avaldusi sirvida.

Kuidas aga jõuda selleni et sobiv kandidaat just õige kuulutuse üles leiaks? Tööportaalides on saadaval erinevaid teenuseid, mille abil saab kuulutust teistest esile tuua ning seeläbi rohkemate inimesteni jõuda. Antud teenused aga toimivad ainult juhul kui veebikasutaja on juba tööportaali lehel kuulutusi sirvimas. Kuulutusi käiakse sirvimas suure tõenäosusega ainult kõige aktiivsemal tööotsingu perioodil, näiteks peale viimatise töölepingu lõppemist. See aga ei pruugi tähendada, et uued ja huvitavad pakkumised ei võiks sellegipoolest huvi äratada ka ajal kus aktiivselt uut tööd ei otsita.

Tänapäeval üks parimaid meetodeid töökuulutus kandidaadini suunata on kasutada selleks sotsiaalmeediat ja muid reklaamkanaleid, mis ei sõltu sellest, et kasutaja otseselt tööportaali lehte külastab. Populaarseimad kanalid, mida selleks kasutada on Meta ja Google [18]. Mõlemad mainitud reklaamkanalid on võimelised reklaami efektiivsust jooksvalt optimeerima ning seda etteantud andmete abil sobivate inimesteni suunama.

Antud töö eesmärgiks on luua mudel, mis võimaldab ennustada, milline reklaamkanal toob töökuulutusele parema tulemuse. Selline ennustus võimaldab reklaami aktiveerides suunata suurem osa eelarvest reklaamkanalisse, mis ennustuse põhjal toob paremat tulemust. Seeläbi saab maksimeerida kandidaatide arvu kuulutusel ning suurendada võimalust, et just õige inimene leitakse ametipositsioonile.

Reklaamide süstemaatiline analüüs on võimalik, kuna nende loomine on automatiseeritud ning kõik reklaamid vastavad standardsele etteantud mallile. Kõikidel reklaamidel on asukoht, kategooria ja pealkiri.

2 Eesmärgid

2.1 Ärieesmärgid

Tööportaali eesmärk on viia kokku tööandjad ja sobivad kandidaadid. Selleks et tööandja leiaks sobiva kandidaadi on tarvis, et sisestatud töökuulutus jõuaks võimalikult paljude tööotsijateni, kes on pakkumisest huvitatud. Lisaks tööportaali sisestele teenustele on ettevõtte huvi ka väliste kanalite abil tööpakkumiste haaret suurendada. Seejuures on tähtis, et teenus oleks võimalikult kuluefektiivne, et väikseima võimaliku kuluga tuua paremat tulemust ning seeläbi suurendada kasumit ja kliendi rahulolu. Google ja Meta on ühed kõige laialdasemalt kasutusel oleval reklaamplatvormid [2] ning lisaks sisse ehitatud automaatsele eelarve optimeerimise algoritmile on võimalik reklaami suunata ka sihtrühma- ja asukohapõhiselt. Mõlemas platvormis on võimalik reklaami eelarvet vabalt valida, mis langeb kokku ettevõtte ärihuvidega – saab pakkuda nii madala kui kõrge eelarvega reklaame ning seeläbi pakkuda teenust enamikule klientidest.

Mitme reklaamkanali puhul on tähtis aru saada, mis olukorras missugune reklaamkanal paremini toimib. Selle info põhjal on võimalik eelarvet efektiivsemalt kasutada vähendades seda ühes kanalis ning suurendades teises, selleks et reklaam saaks võimalikult hea tulemuse. Läbiv eesmärk on tuua klientide kuulutustele rahaühiku kohta võimalikult suur kogus näitamisi ja kandidaate. Seeläbi suureneb lehe külastatavus suurema koguse kandidaatide näol ning suureneb äriklientide rahulolu teenuse kasutamisel. Ka väike eelarve optimeerimine võib tuua märgatavat kasu. Lühidalt võib eesmärgid kokku võtta järgnevalt:

- Optimeerida reklaamide eelarve kasutust
- Suurendada kandidaatide arvu
- Suurendada kasumimarginaali

2.2 Andmekaeve eesmärgid

Enne andmete analüüsima hakkamist on tarvis mõista mida on vaja saavutada ning mis andmed on olulised ja asjakohased selle eesmärgini jõudmiseks [3]. Alustuseks on tarvis vaadelda saadavalolevaid andmeid – mis lähteandmed on saadaval, kuidas need seonduvad mõõdikutega ning mis tulemust me soovime antud andmete töötlusega

saavutada. Edasiste sammudega tuleb hinnata kas andmed on piisavad, kaardistada ohukohad ja piirangud, valida sobivad atribuudid edasiseks töötluseks ning seejärel alustada andmete sobivale kujule töötlemisega.

Andmete töötlemise käigus loodi uued ühikupõhised atribuudid, visualiseeriti andmed lihtsama ülevaate saamiseks, veenduti et puuduvad väärtused on asjakohaselt eemaldatud või asendatud, konverteeriti andmed sobivale kujule – vajadusel tehti uusi veerge ning leiti korrelatsioonid atribuutide ja ennustatavate väärtuste vahel. Kõik eelnevad tegevused toetasid lõppeesmärgi saavutamist - koostada korrastatud ja töödeldud andmete põhjal ennustav mudel, mida treeniti olemasolevate andmete põhjal ennustama reklaami tulemust kummagi reklaamkanali kohta eraldi. Parima tulemuse saavutamiseks valiti andmestikule sobivad masinõppe meetodid ning võrreldi neid omavahel, et leida kõige paremini toimiv. Ennustuse edukust kontrolliti jagades andmed test- ja treeningandmeteks ning hinnati kasutades R^2 skoori ja ruutkeskmist viga (MSE). Mudelit võib pidada:

- Toimivaks kui R^2 skoor seletab 25% või rohkem variatsioonist, ehk $R^2>=0.25$. Antud number on suurem kui tavapärane veavahemik [4], seega võimaldab tuua suuremat kasu kui suvalise väärtuse valimine.
- Hästi toimivaks, kui R^2 skoor seletab 50% või rohkem variatsioonist. Kuna kõiki parameetreid ja reklaamkanalite aspekte ei ole võimalik arvesse võtta, siis 50% variatsiooni seletamine on hea tulemus ning võimaldab teha arvestatava kasuteguriga ennustusi.

2.3 Metoodika ja töö struktuur

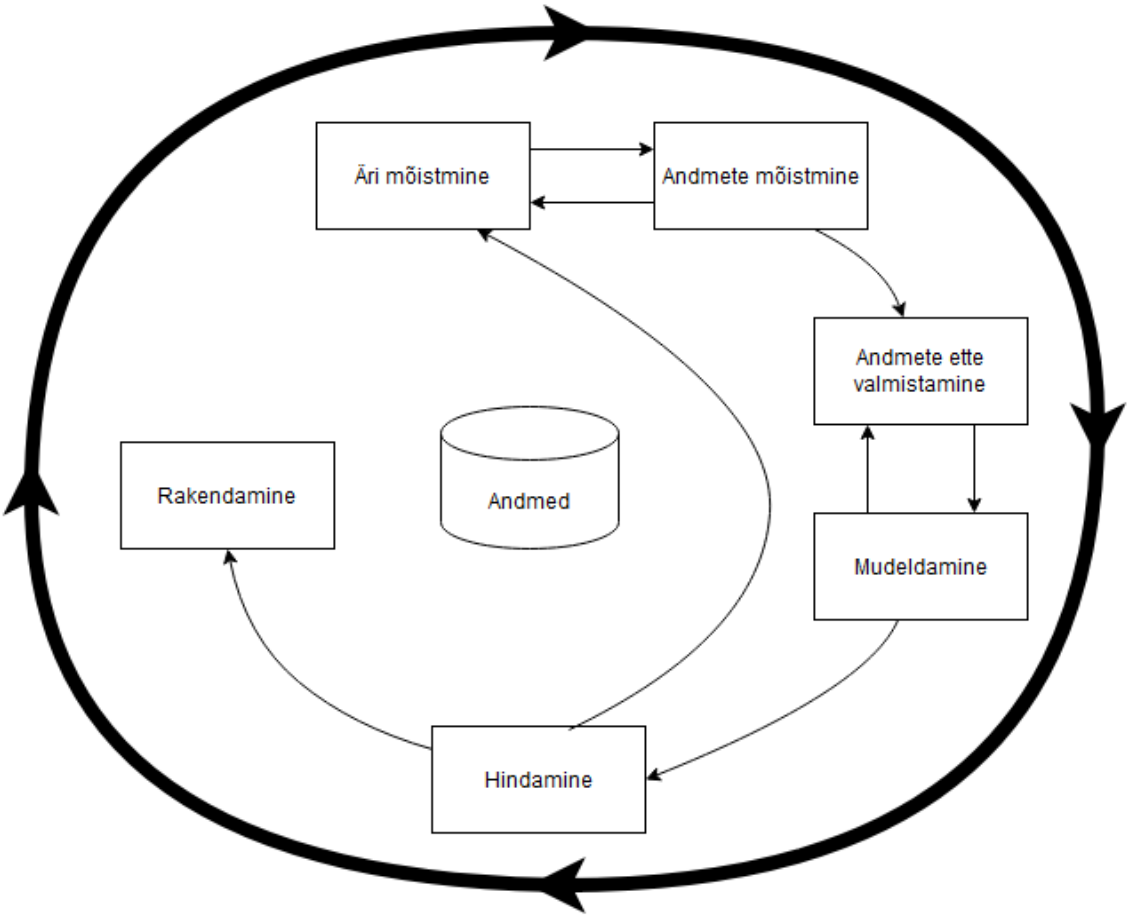
Antud probleemi lahendamise aluseks on võetud CRISP-DM andmetöötluse metoodika [5], mis on nelja tasandisse jagunev hierarhiline mudel. Kõige üldistatuma tasandi ehk faasi sammud on kujutatud Joonisel 1. Iga joonisel näha olev faas jaguneb üldisteks ülesanneteks ning seejärel konkreetsemateks tegevusteks. Mida sügavamale tasandile minna, seda täpsemad on seal kirjeldatud tegevused. Faase saab lühidalt kokku võtta järgnevalt:

1. Äri mõistmine - esimene faas keskendub sellele, et mõista projekti nõudmisi ja oodatavaid tulemusi äritasandil. Seejärel koostatakse selle teadmise põhjal

andmekaeve probleemi definitsioon ja esialgne plaan, mille alusel eesmärke saavutada.

2. Andmete mõistmine – siin algab esialgne andmete kogumine ning nendega tutvumine. Tuvastatakse andmetes olevad puudused ja seosed. Võimalik, et siin faasis õnnestub tuvastada uut kasutatavat informatsiooni.
3. Andmete ette valmistamine – kogutud andmete töötlemine ja ette valmistamine. Sisaldab kõiki samme selleks, et viia andmed mudeldamiseks sobivale kujule, sealhulgas: andmete ühisele kujule konverteerimine, puudulike andmete eemaldamine või asendamine, töötlemine, standardiseerimine jm.
4. Mudeldamine – erinevate mudelite koostamine ja katsetamine, samuti nende parameetrite optimeerimine parima tulemuse saamiseks. Eri mudeleid katsetades võib olla vaja tagasi liikuda eelmisesse sammu.
5. Hindamine – mudeli hindamine andmeanalüüsi vaatepunktist. Hinnata tehtud tööd ning selle vastavust ärieesmärkidele - tähtis on tuvastada kas midagi ärieesmärkidest jäi tähelepanuta.
6. Rakendamine – mudeli kasutuselevõtt. Kasutuselevõtt võib olla erineva keerukusega ja tihti ei vastuta selle eest andmeteadlane. Tähtis on, et mudeli rakendamise eest vastutaja mõistaks, mis on vaja teha selleks, et tehtud töö tulemus kasutusele võtta.

Kirjeldatud mudel või selle variatsioonid on tavapraktikas laialdaselt kasutatud. Antud töös on lähtutud Joonisel 1 kuvatud mudelist.



Joonis 1. CRISP-DM andmetöötluse metoodika faasid.

3 Andmete kogumine ja eeltöötlus

Uurimistöö andmed põhinevad Alma Media Estonia OÜ vahendusel tehtud reklaamidel. Andmed võeti Alma Media Estonia OÜ, Meta ja Google andmestikest. Baasandmed võeti CV.ee andmetest, kuhu on koondatud Metast ja Googlest tulevad reklaamide andmed. Antud andmete seas on tööandja id ja nimi, reklaami algus- ja lõpukuupäev, töökuulutuse id, töökuulutuse kategooria ning eraldi nii Google kui Meta jaoks id, eelarve, klikid, vaatamised, eelarve ja kulutatud raha. Sellele lisaks on CV.ee avalikust APIst juurde päritud iga kuulutuse jaoks kuulutuse positsiooninimi ja asukoht. Asukoht koosneb kolmest väljast – linn, maakond ja riik. Google Ads keskkonnast tehti eraldi CSV faili kujul väljavõte, et saada Google reklaamides kasutatav pealkiri. See on vajalik, sest Google seab rangemad piirangud reklaamide pealkirjadele ning seetõttu on mõningad Google reklaamide ametinimed lühendatud ning teatud erisümbolid on eemaldatud.

Kõik kogutud andmed töödeldi selleks loodud Python rakenduses. Esmalt koondati andmed ühte andmestruktuuri, seejärel konverteeriti ametinimed vektoriteks Word2vec tööriistal põhinevat keelemudelit kasutades. Kuna antud keelemudel põhineb Eesti keelel, siis juhul kui keelemudel ei suutnud sõna tuvastada, tõlgiti sõna inglise keelest eesti keelde ning prooviti uuesti vektoriks teisendada. Esmase töötluse läbinud andmete salvestamiseks sai loodud PostgreSQL andmebaas. Edasiseks töötluseks saadavad andmed võeti SQL päringuga, millega sorteeriti välja read kus ei õnnestunud ametinimest vektorit genereerida.

3.1 Keelemudel ja vektor

Võtmekohal ennustusmudeli loomiseks on ametinimedest genereeritavad vektorid, mis on genereeritud Google poolt väljatöötatud tööriista word2vec [6] abil. Tegemist on NLP-l (*natural language processing*) põhineva tööriistaga. NLP üks tähtsamaid aspekte on sõnade masinloetavale kujule teisendamine. Enamus masinõppe algoritme ei suuda sõnu või vabateksti tõlgendada, seega masinõppe mudeli sisend peab olema numbrilisel kujul. Laialdaselt kasutatakse selleks vektoreid, kuna vektor võimaldab sõnade vahelisi

sarnasusi väljendada. Lihtsustatult tähendab sõnade vektorkujule teisendamine seda, et on etteantud sõnastik ja igale seal esinevale sõnale vastab ettemääratud pikkusega numbrite jada, mis väljendab selle sõna omadusi sõnastikus. Kõige lihtsamas vormis on numbrite jada pikkus võrdne sõnastikus esinevate sõnade arvuga, kuid sellise mudeli mälu vajadus kasvab eksponentsiaalselt ning ei ole teatud piirist enam jätkusuutlik. Eeltreenitud mudelid nagu Word2Vec, Glove ja fastText kasutavad enda algoritmi, et sõnastikus olevate sõnade vahelisi seoseid kompaktselt ja efektiivselt väljendada [7].

Word2vec võtab sisendina treeningandmed ning nende põhjal õpib ja tagastab sõnade vektorväärtused. Selle jaoks on saadaval kaks õppimisalgoritmi – CBOW ja Skip-gram. Skip-gram on aeglasem ja sobib hästi andmestiku jaoks, mis sisaldab palju unikaalseid sõnu. CBOW (*Continuous bag of words*) on kiirem ning kuna ametinimedes sõnad tihti korduvad, siis käesolevas töös kasutati sel viisil treenitud mudelit. Nii Skip-gram kui CBOW on ühed parimad saadavalolevad masinõppe mudelid sõnadevaheliste sarnasuste tuvastamiseks [8]. Antud töös on kasutatud eeltreenitud 100-dimensioonilist CBOW algoritmi [9]. See mudel võimaldab arvutada vektori iga ametinimes esineva sõna kohta eraldi. 100-dimensiooniline mudel tähendab et üks vektor konverteeritakse hiljem 100ks eraldiseisvaks atribuudiks selleks et seda saaks ennustusmudelis kasutada. Selleks et kogu ametinimi vektorkujule saada prooviti kahte lähenemist:

1. Arvutati iga sõna kohta vektori ning seejärel liideti need kokku, et saada keskmistatud vektor. Selleks kasutati numpy mean funktsiooni [10], et arvutada keskmine iga ametinimes esineva sõna vektori dimensiooni kohta.
2. Arvutati iga sõna kohta vektor ning salvestati see andmebaasis eraldi väljas. Selline lähenemine tähendas, et lähteandmete (veergude) hulk suurenes olulisel määral.

Töö käigus selgus, et keskmistatud vektor toob kehvema tulemuse kui iga ametinimes sisalduva sõna jaoks eraldi vektori leidmine, seega ennustusmudeli tulemustes on välja toodud ainult iga sõna jaoks eraldi leitud vektoritega andmetest saadud tulemused.

Andmeid töödeldes tuli välja, et tihti kasutatakse ametinimes ingliskeelseid väljendeid. Kasutusel olev keelemudel suudab suure osa neist tuvastada vaatamata sellele et tegemist on eesti keelel põhineva keelemudeliga, kuid mõned sõnad jäävad siiski tundmatuks ning mudel ei suuda nende jaoks vektorit leida. Seetõttu sai kasutusele võetud Easygoogletranslate [11] nimeline tõlkeprogramm, mis baseerub Google Translate

tõlgetel ja võimaldab vähese vaevaga sõna tõlkida eesti keelde. Peale tõlkimist proovitakse uuesti sõna vektoriks teisendada.

3.2 Asukohad ja kategooriad

Lähteandmetest on lisaks ametinimele veel ennustamiseks sobilikud asukoht ja kategooriad. Kategooriateks on fikseeritud nimekiri 33 unikaalse väärtusega. Kategooria sümboliseerib valdkonda, milles ettevõte tegutseb või millega ametipositsioon kõige paremini seondub. Antud valdkonna on valinud tööandja reklaami sisestades.

Asukohad jagunevad kolmeks – linn, maakond ja riik ning sarnaselt kategooriale määrab asukoha klient ise. Kohustuslik neist on ainult riik. Kitsam asukoha tüüp eeldab et ka laiem tüüp on valitud, ehk kui on valitud linn, siis on alati valitud ka maakond. Asukohad on fikseeritud nimekirjast võetud, seega neid on piiratud arv – valitavad on riigid ning suuremad Eesti asulad ja maakonnad. Kõik nimekirjas esinevad asukohad ei saa antud andmestikus esindatud, seega lõplik asukohtade arv selgub hilisema töötluse käigus.

3.3 Andmebaas ja edasine töötlus

Selleks, et andmestik oleks lihtsasti hallatav ja mugavalt kättesaadav said kõik andmed peale esialgset töötlust salvestatud PostgreSQL andmebaasi. Muuhulgas said salvestatud ametinimede keskmistatud vektorid ning iga sõna kohta eraldi vektorid.

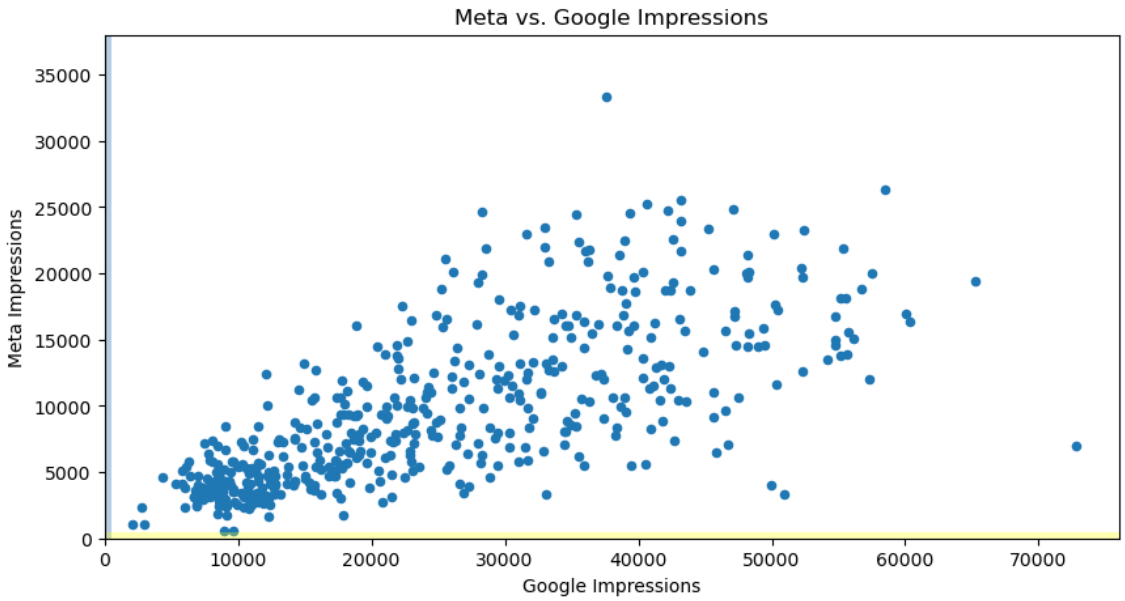
Edasiseks töötluseks kasutati Jupyter Notebooki. Tegemist on veebipõhise interaktiivse andmetöötlusplatvormiga, mis ühendab endas reaalajas koodi, võrrandite, teksti ja andmete visualiseerimise [12]. Andmebaasist võteti .csv formaadis reklaamide andmed ning imporditi need Jupyter Notebooki. Selle käigus said andmetest eemaldatud read, kus ei õnnestunud vektori genereerimine. Seega igal analüüsitaval reklaamil on asukoha, kategooria ja ametinime põhjal genereeritud vektor või vektorid.

4 Andmete visualiseerimine

Andmete visualiseerimiseks kasutati Matplotlib teeki, mis on üks levinumaid andmete visualiseerimise teeke Pythonis [13]. Selleks et saaks võrrelda erinevate eelarvetega reklaame omavahel on kõik tulemused arvestatud ühiku euro (€) kohta.

4.1 Graafikud – scatterplot ja barchart

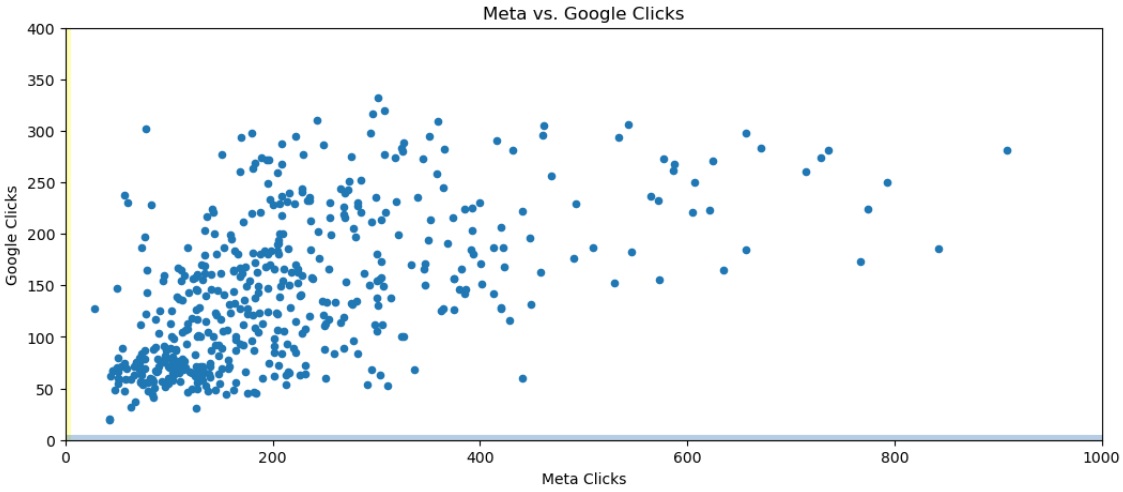
Esmalt vaatleme hajuvusdiagrammi, kus ühel teljel on Google- ja teisel Meta kuulutuste näitamised (Joonis 2).



Joonis 2. Meta ja Google reklaamide näitamiste jaotus.

Näitamiste (*impressions*) diagrammilt on näha, et Google reklaamidel on oluliselt rohkem näitamisi. Antud tulemus seab kahtluse alla kas eri keskkondade näitamised on omavahel võrreldavad suurused.

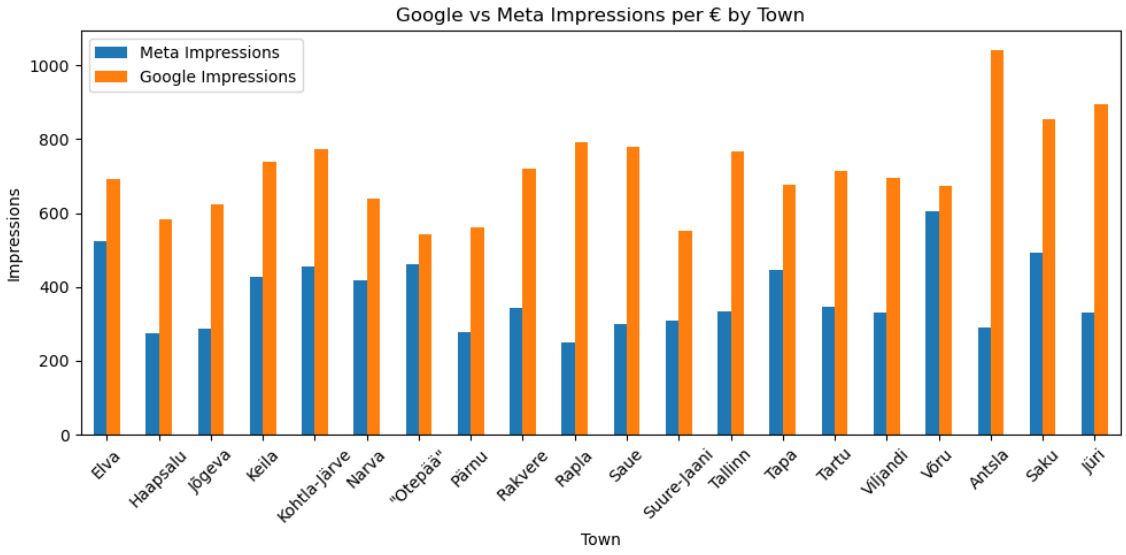
Järgmisena vaatleme klikkide hajuvusdiagrammi, kus võrreldakse Google ja Meta kuulutuste klikkide arvu (Joonis 3).



Joonis 3. Meta ja Google reklaamide klikkide jaotus.

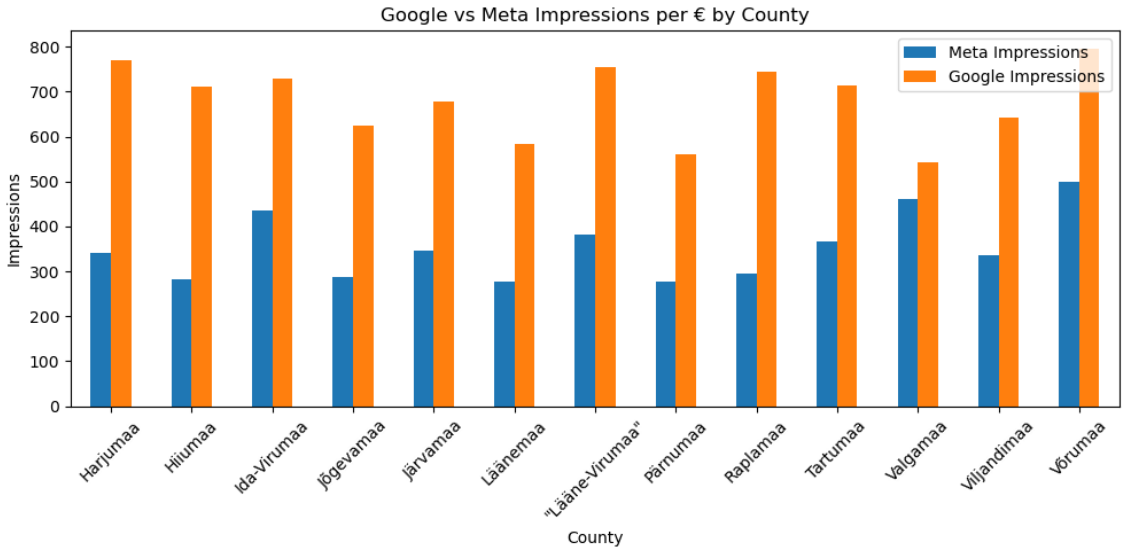
Klikkide diagrammilt on näha, et esineb arvestataval määral hajumist, mis viitab sellele et reklaamide tulemused erinevad reklaamkanaliti. See toetab hüpoteesi, et eelarve ümberjagamisel vastavalt ennustusele võime saada parema tulemuse kui eelarve võrdselt jagamisel. Klikkide arv kummaski reklaamkanalis on sarnases suurusjärgus, kuid Metas esineb üksikuid reklaame, mis on toonud erakordselt häid tulemusi.

Järgmisena vaatleme tulpdiagramme, mis illustreerivad kui suur on tulemuste varieeruvus igat parameetrit (asukohad, valdkonnad) eraldi vaadeldes. Linnade lõikes näitamised on näha järgnevalt (Joonis 4).



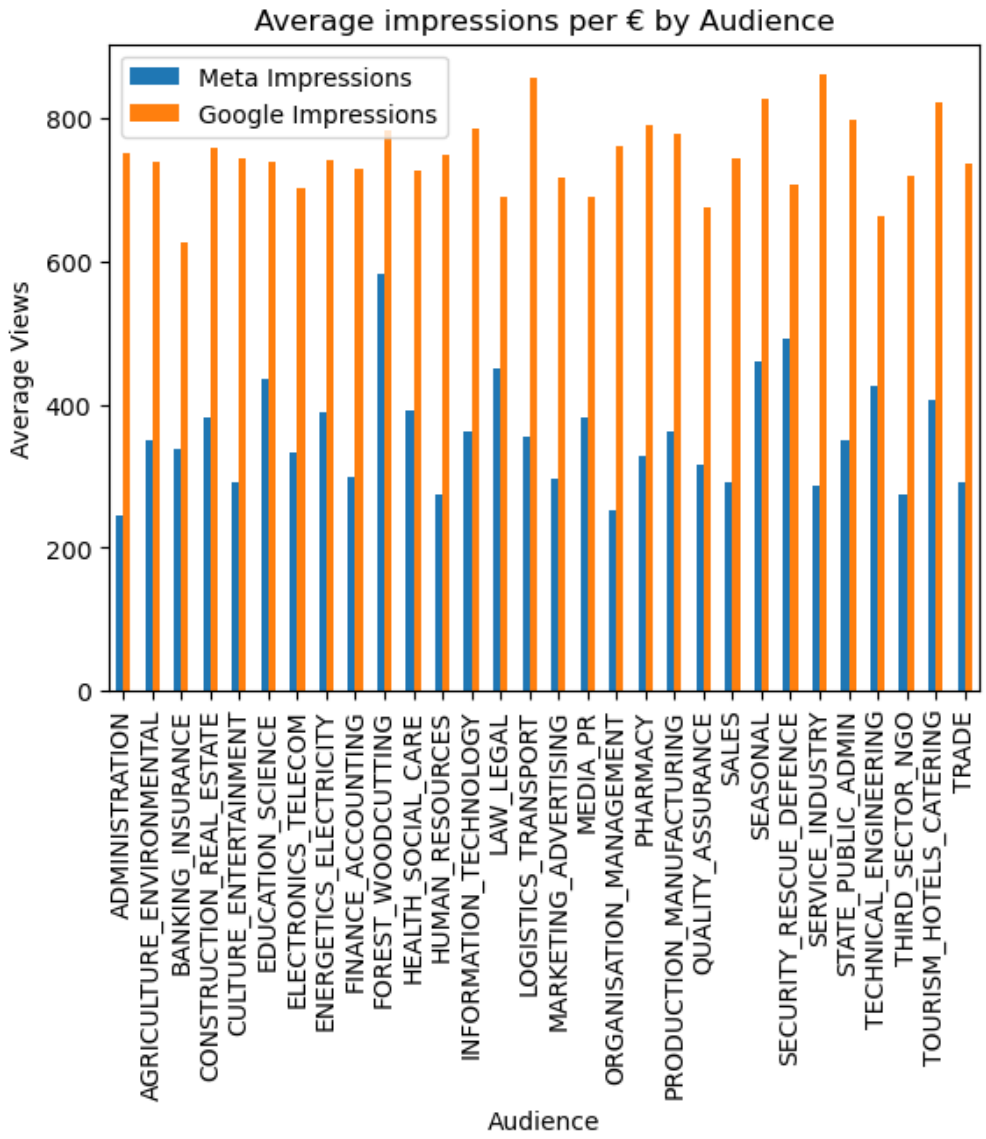
Joonis 4. Google ja Meta tulemused linna järgi.

Jooniselt selgub et vaatamised on kõigis linnades suuremad Googles. Järgmisena vaatleme näitamisi maakonniti (Joonis 5).



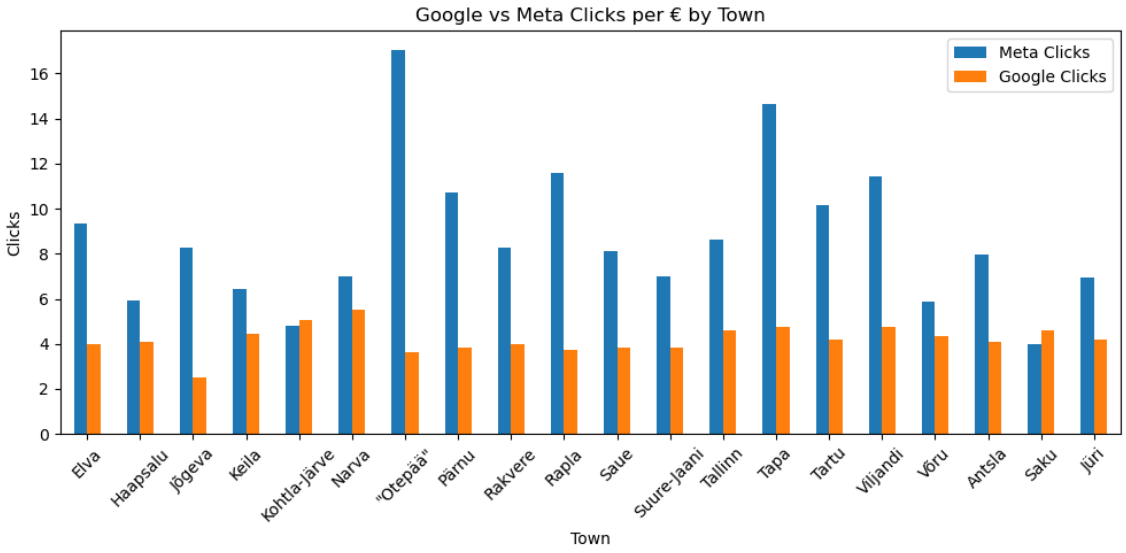
Joonis 5. Google ja Meta tulemused maakonna järgi.

Ka maakonniti on selge, et Google reklaamid saavad oluliselt rohkem näitamisi. Vaatleme ka töökuulutuste valdkondadesse kuuluvuse järgi võetud statistikat (Joonis 6).



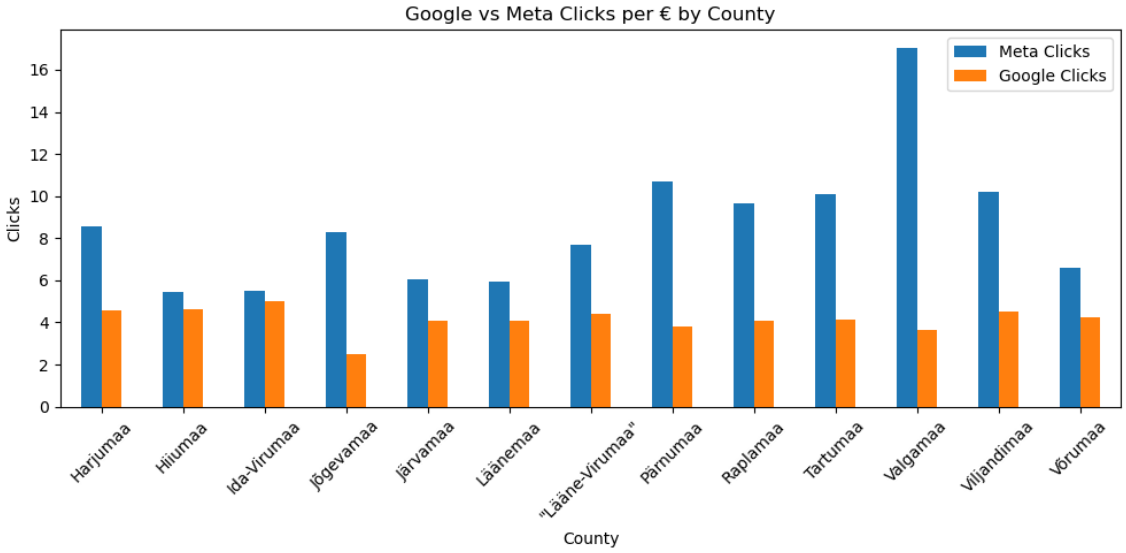
Joonis 6. Keskmine vaatamiste arv kategooriate järgi.

Näitamiste arvu osas on Google iga asukoha ja kategooria puhul oluliselt ees. Edasi vaatleme klikkide arvu, alustades linnadest (Joonis 7).



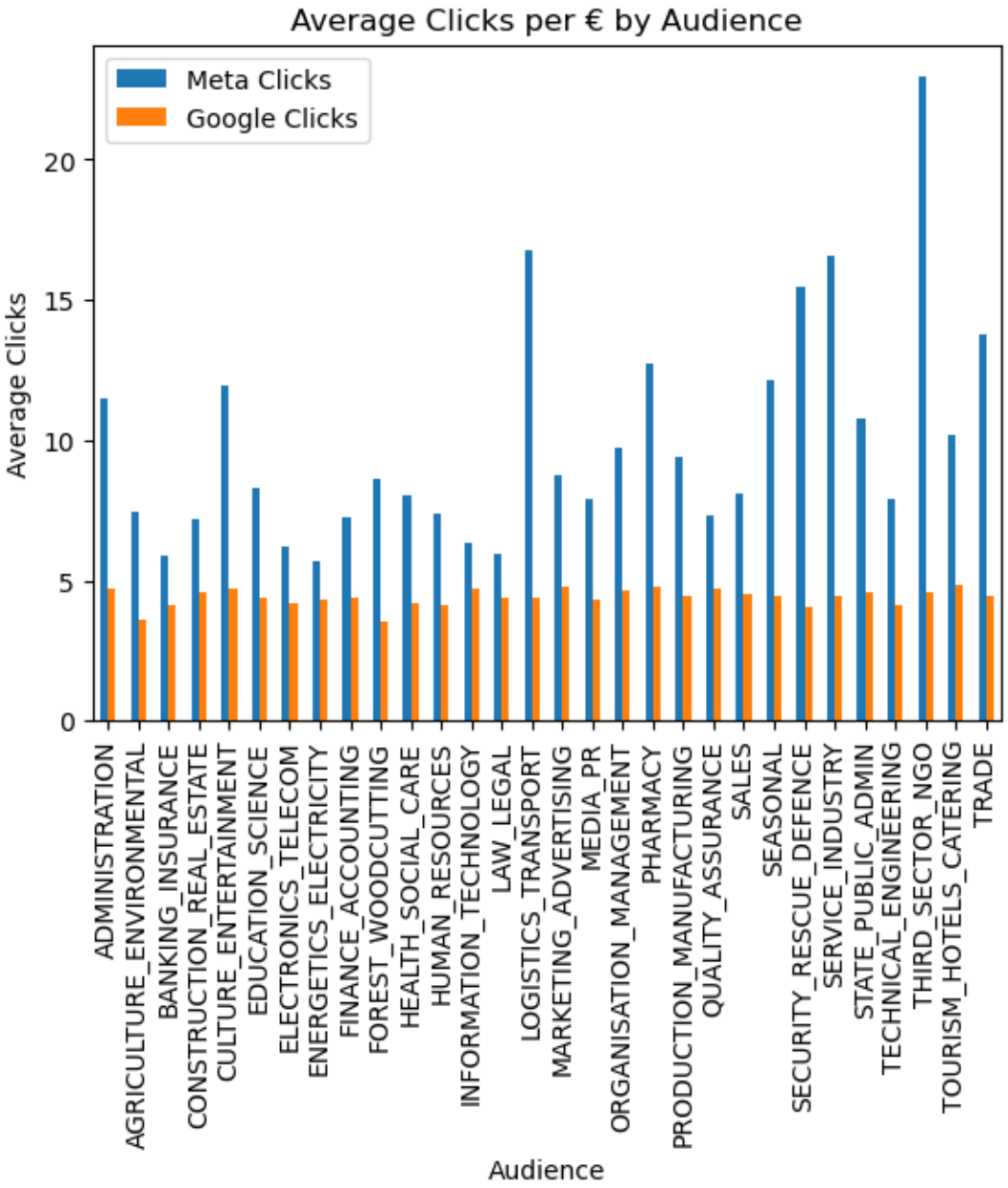
Joonis 7. Google ja Meta klikkide arv linna järgi.

Selgub et klikkide osas on Meta pea kõikides linnades tulemustega ees. Erandiks on vaid Saku ja Kohtla-Järve, kus tulemus jääb napilt konkurendile alla. Maakondade statistika on näha järgnevalt (Joonis 8).



Joonis 8. Google ja Meta klikkide arv maakonna järgi.

Jällegi on näha et Meta on klikkidega ees, edestades teist reklaamikanalit kõikides maakondades. Viimasena vaatleme klikkide arvu valdkondade järgi (Joonis 9).



Joonis 9. Keskmine klikkide arv kategooriate järgi.

Antud diagrammidelt joonistub välja, et Google näitamiste arv on alati oluliselt suurem kui Meta näitamiste arv. Selle põhjal võib järeldada, et antud mõõdiku järgi ei ole mõistlik edasist analüüsi teha. Klikkide arv on enamjaolt Meta reklaamidel suurem, kuid esineb siiski ka olukordi, kus Google toob parema tulemuse või kus tulemus on sarnane. Antud andmetest võib järeldada, et edasine töö peaks toimuma klikkide arvu põhjal, kuna näitamiste arv ei pruugi olla võrreldav.

5 Ennustusmudelite koostamine ja hindamine

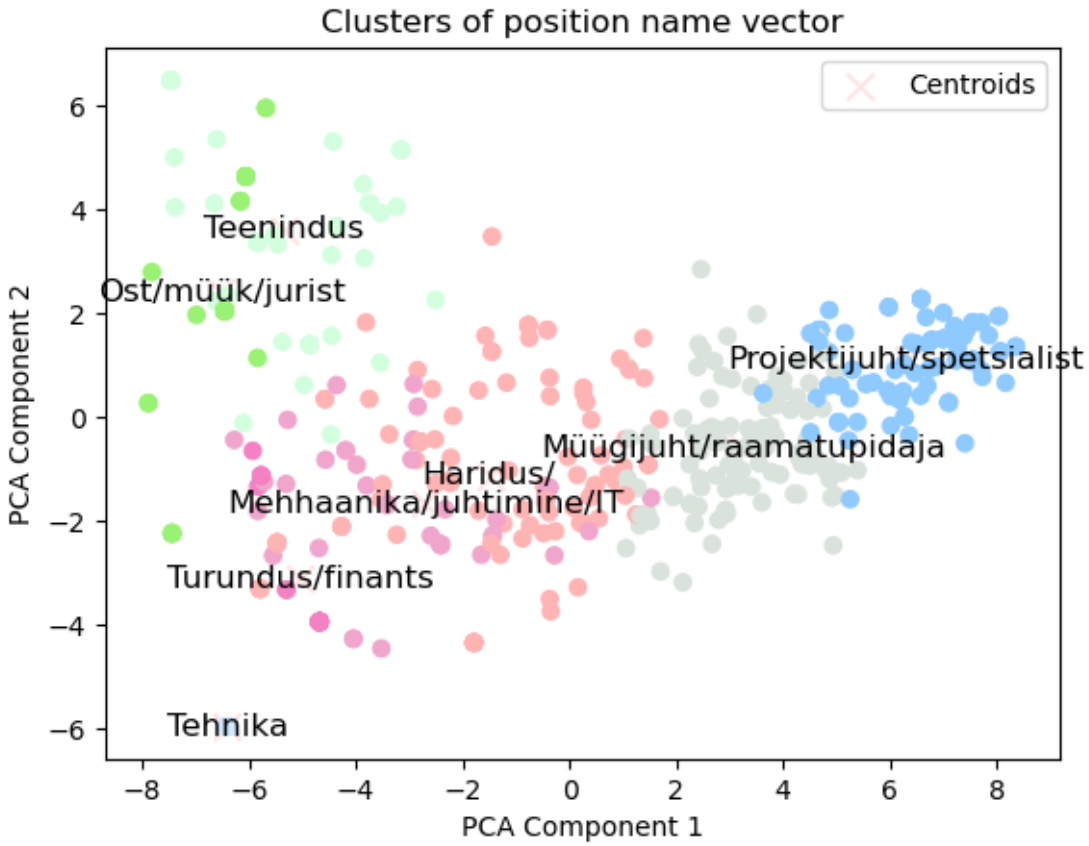
Ennustusmudel koostati kummagi reklaamkanali kohta eraldi. See võimaldab reklaamkanaleid hiljem võrrelda. Samuti võimaldab see ennustusmudeli konfigureerida eraldiseisvalt, et saada parim võimalik ennustus kummastki mudelist. Iga mudeli andmestikku jäeti alles vaid selle reklaamkanali hindamiseks vajalikud andmed ning kõik ülejäänud veerud said eemaldatud. Samuti eemaldati kõik read, kus olevad andmed olid puudulikud. Ridade eemaldamisel eemaldati read mõlema reklaamkanali andmestikust, et kumbagi andmestikku jääksid täpselt samad reklaamid. Selle tulemusena jääb alles 516 rida.

5.1 Andmestik ja korrelatsioonid

Andmestikku jäid alles klassifikaatorina linn, maakond, riik ja kategooria. 100-dimensioonilised vektorid konverteeriti eraldi veergudeks – iga dimensioon saab olema eraldi veerg. Andmed standardiseeriti ehk konverteeriti kujule kus iga atribuudi mediaanväärtus on 0 ja standardhälve on 1. See välistab ohu, et suure variatsiooniga atribuut hakkab domineerima ning mõjutab mudelit ebaproportsionaalselt palju [14].

5.2 K-means klasterdamine

Ametinimede visualiseerimiseks kasutati eelnevalt ametinimede põhjal leitud standardiseeritud keskmistatud vektoreid. Selleks et diagrammi koostamine oleks võimalik oli esmalt vaja vähendada andmete dimensioone. Vektori 100 dimensiooni konverteeriti kaheks dimensiooniks kasutades peakomponentanalüüsi. Kahe dimensioonilisi andmeid on võimalik kuvada Scatterplot diagrammis ning järgneval joonisel (Joonis 10) on näha ametinimede jaotumine loodud peakomponentide alusel. KMeans klasterdamine võimaldab jagada andmed klastriteks ning määrata igale klastrile üks keskne punkt koos vastava nimetusega. Klastri nimetus sai valitud andmete visuaalsel hindamisel.



Joonis 10. Ametinimede jaotumine loodud peakomponentide alusel.

Kuigi klastrites on näha teatud selgelt eristatavaid jooni, siis arvestataval määral on ka ametinimesid, mis ei klapi valitud klastri nimetusega, kuna klastris esinevad punktid on liialt varieeruvad.

5.3 Heatmap

Ennustatava atribuudi ja teiste atribuutide vaheliste korrelatsioonide põhjal loodi soojuskaart (*heatmap*). Soojuskaart on hea visuaalne vahend tuvastamaks mudelis suurimat korrelatsiooni omavad atribuudid. Atribuudid, mis olid suurimas korrelatsioonis klikkide arvuga on järgnevad:

Meta atribuut	**Korrelatsioon**	**Google atribuut**	**Korrelatsioon**
Audience_LOGISTICS_TRANSPORT	0.27	Town_tallinn	0.18
Audience_SERVICE_INDUSTRY	0.24	County_harjumaa	0.19
			Järgneb

Audience_THIRD_SECTOR_NGO	0.18	County_tartumaa	-0.17
Audience_TRADE	0.18	Googleword1_vec_40	-0.15
Word2_vec_37	-0.18	Town_tartu	-0.13
Word1_vec_50	0.19	Googleword1_vec_37	0.13
Word2_vec54	-0.18	Googleword2_vec_4	0.13
Word2_vec_69	-0.18	Googleword2_vec_61	0.12

Tabel 1. Suurima korrelatsiooni absoluutväärtusega atribuudid.

5.4 Katsetatud mudelid

Tähtis on valida andmetele sobiv masinõppe mudel, kuna vale mudeli valimisel on oht üle- või alatreenimiseks. Kuna lähteandmeid on väike kogus ning andmed on suhteliselt lihtsad, siis on antud kontekstis mõistlik kasutada lihtsamaid masinõppe mudeleid nagu *lineaar regression*, *random forest regressioon* jm. Seejuures on tähtis valida masinõppe meetodid, millega saab ennustada pidevat väärtust (klikid).

Katsetatud mudellid	**Meta R^2 score**	**Meta MSE**	**Google R^2 score**	**Google MSE**
Random Forest Tree	0.3627	0.5213	-0.0353	0.5578
Linear regression	-3.8665	3.1624	-8.1825	4.4087
Ridge regression	-1.4125	1.9732	-1.2221	1.9908
Gradient Boosting Regression	0.2094	0.6467	-0.0555	0.5687
Support Vector Regression	0.0764	0.7554	-0.0592	0.5707
Extra trees regression	0.2731	0.5945	-0.1032	0.5944

Tabel 2. Katsetatud mudelid ja nende valideerimise skoorid.

Tabel 2 näitab, et Random Forest Tree mudeli tulemus on R^2 skoori ja MSE poolest parim nii Meta kui ka Google reklaamide puhul.

5.5 Hüperparameetrid

Selgus, et katsetatud mudelitest parima tulemuse andis Random Forest Tree regressioon mudel. Selleks, et tulemust veelgi paremaks saada prooviti erinevaid parameetreid kolmel hüperparameetril – n_estimators, max_depth ja max_features (Tabel 3). Katsetatud sai järgnevaid väärtusi:

n_estimators	5	10	20	30	40	50	60	70	80	90	100	
max_depth	None	2	5	8	10	15	20	25	30			
max_features	1	3	5	8	10	15	20	25	30	35	40	45

Tabel 3. Hüperparameetrid.

N_estimators parameeter tähistab puude arvu mudelis. Suurem arv puid toob tihti parema tulemuse, kuid teatud hetkest saadav kasu stagneerub. Arvestades et puude lisamine suurendab oluliselt arvutusaega siis on kasulik antud piir tuvastada et vältida liigset arvutuskeerukust.. Max_depth tähistab suurimat kaugust juurelemendi ja alamelemendi vahel. Max_features on suurim atribuutide arv mida ühe puu kontekstis arvesse võetakse [15]. Parimad hüperparameetrid Random Forest Tree mudelis on:

- Meta: n_estimators=70, max_depth=15 ja max_features = 40
- Google: n_estimators=60, max_depth=5 ja max_features = 8

Antud parameetritega õnnestus parandada nii MSE kui R^2 väärtusi järgnevalt:

- Meta:
  R^2: 0.4011
  MSE: 0.4899
- Google
  R^2: 0.0913
  MSE: 0.4896

5.6 Mean squared error ja R^2 skoor (hindamine)

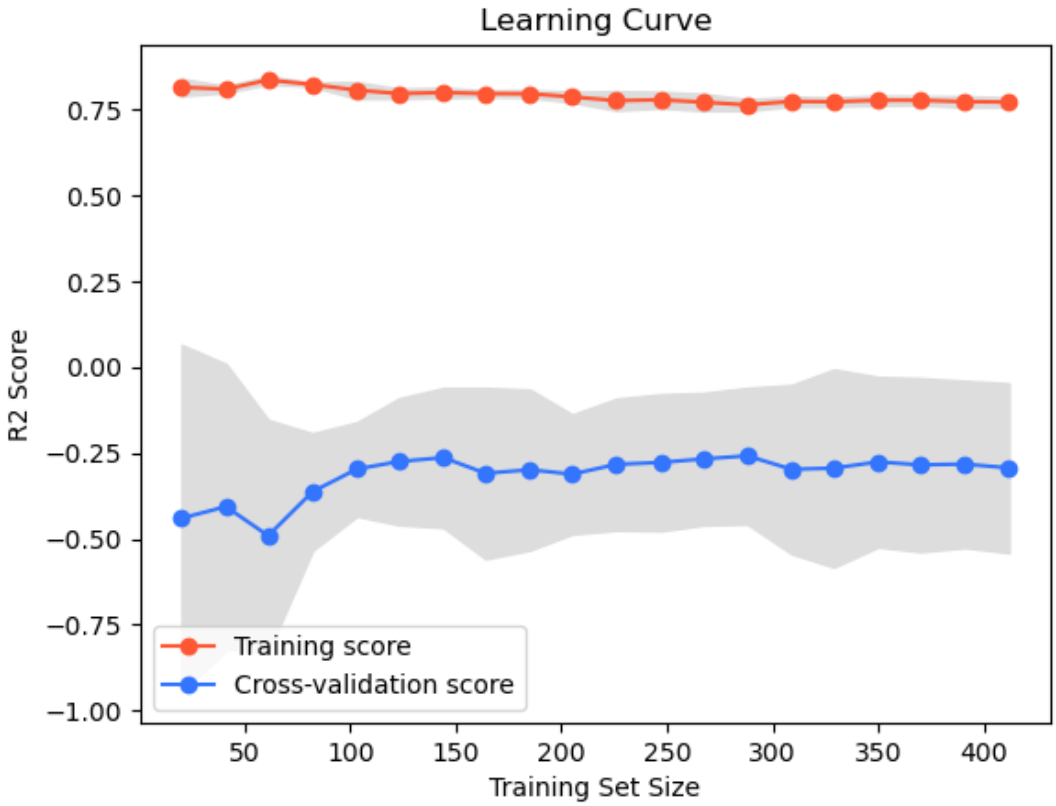
Antud töös parimaks osutunud masinõppe meetodi (Random Forest Tree regressioon) eelisteks on lihtsam valideerimine, mis muuhulgas vähendab ületreenimise ohtu, kuid otsustuspuudel põhinevad meetodid võivad olla tulemustes mõnevõrra varieeruvamad kui alternatiivid [16]. R^2 skoor on laialdaselt kasutusel olev mõõdik ennustusmudeli

toimivuse hindamiseks. Selle (kasulik) väärtus jääb vahemikku 0 ja 1, mis sümboliseerib protsentuaalselt seda kui suurt osa andmetest on seletatav regressioonimudeliga. See tähendab, et kõrgem väärtus on parem, kuid tähtis on vältida ületreenimist. Ületreenitud mudel oskab teha ennustusi treeningandmete põhjal, kuid uute andmete põhjal ei pruugi anda oodatud tulemusi. Samuti ei tähenda madal R^2 skoor, et andmetes ei esine ennustamiseks kasulikke seoseid ning tähtis on arvestada kui suur on andmete varieeruvus kuna see mõjutab oluliselt R^2 skoori [17]. Antud uuringus on suur andmete varieeruvus, mis võib ka olulisel määral mõjutada R^2 skoori.

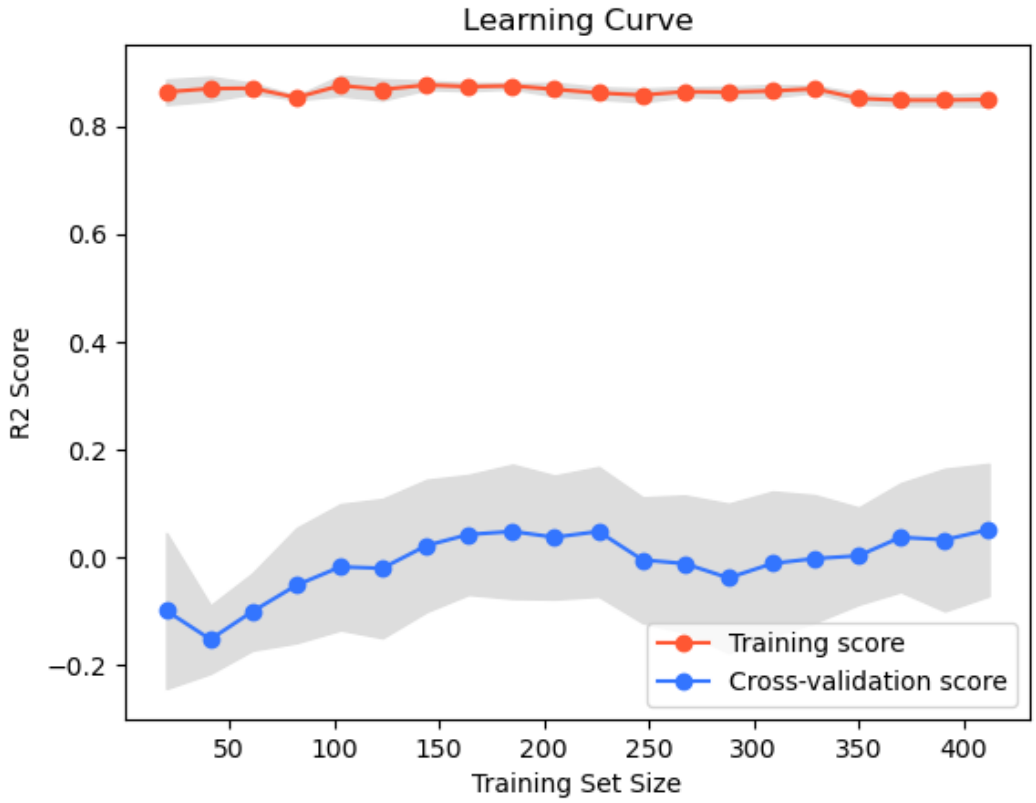
Mean Squared Error (edaspidi MSE) väljendab keskmist ruutu ennustatud väärtuse ja päris väärtuse erinevusest. Kuna MSE sõltub andmete suurusjärgust, siis peetakse seda vähemtähtsaks mudeli hindamisel kui R^2 skoori. Sellegipoolest võimaldab MSE näha ennustuse vea suurust ning seeläbi anda rohkem infot R^2 skoori mõistmiseks. Suurem MSE skoor võib vihjata et suure andmete varieeruvuse tõttu on R^2 skoor madalam [18].

5.7 Õppimiskõver

Õppimiskõver näitab treeningskoori ja valideerimise skoori suhet erinevate lähteandmete koguste kohta. Samuti võimaldab see teha järeldusi mudeli toimivuse ja edasiste tegevuste kohta. Google ja Meta õppimiskõverad on kujutatud järgnevatel joonistel (Joonis 11 ja Joonis 12).



Joonis 11. Google õppimiskõver.



Joonis 12. Meta õppimiskõver.

Mõlemalt jooniselt on näha, et valideerimise skoor on tõusvas trendis – Meta (Joonis 12) puhul veidi rohkem, Google (Joonis 11) puhul veidi vähem. Juhul kui treeningskoor on kõrge ning valideerimise skoor on madal, aga tõusvas trendis, siis on suur tõenäosus, et suurem lähteandmete hulk parandab tulemust kuni hetkeni mil treening- ja valideerimise skoor hakkavad graafikul üksteisele lähenema [19].

6 Tulemused ja edasised tegevused

Tulemuste hindamiseks sai kasutatud eelpool mainitud R^2 ja MSE mõõdikuid. Parim R^2 skoor Meta reklaamide jaoks oli .4011, mis tähendab et mudel suudab seletada 40.11% varieeruvusest. Parim MSE skoor on 0.4899, mis näitab et ennustustes esineb arvestatavas suuruses vigu. Google ennustus tuli oluliselt kehvem ning parim R^2 skoor oli kõigest 0.0912, ehk mudel suudab seletada vaid 9.12% varieeruvusest. MSE jäi samasse suurusjärku kui teisel reklaamkanalil - 0.4896.

R^2 skoori puhul ei ole kindlat arvu, alates millest mudel on hea. Alati tuleb arvesse võtta analüüsitavat andmestikku. R^2 skoor annab aimu kui hästi õnnestus testandmetes seoseid leida, kuid see ei garanteeri et päris andmetega töötades saame sama tulemuse [20]. Antud töö ärieesmärk ei põhine kindlal lävendil ning ennustuseks kasu saamiseks piisab ka sellest kui suudame ennustada suvalisest valimist paremini. Meta puhul see õnnestus, kuid paraku Google ennustus võib jääda alla ennustusmudeli vea suurusele. Ennustusmudelist võib siiski piiratud määral kasu olla, kuna ka ainult ühe reklaamkanali ennustus võib aidata otsustada kuidas on parim eelarvet jaotada.

Mudeli paremaks toimimiseks oleks tähtis koguda rohkem andmeid. Kuna lähteandmete hulk oli suhteliselt väike, siis on väga tõenäoline, et suurem andmete hulk võimaldab saada parema tulemuse. Teine oluline täiendus mudelile oleks kasutada sõnade vektorite asemel fraaside vektoreid. See tähendab vastava keelemudeli treenimist ning fraaside vektorite kasutamist atribuutidena ennustusmudelis.

7 Kokkuvõte

Tööportaal võimaldab kuulutustele lisatähelepanu saamiseks klientidel kasutada väliseid reklaamkanaleid Meta ja Google. Selleks on loodud teenus, mis võimaldab reklaamkanalitesse saata reklaami fikseeritud kujul. Kuna reklaamid on analoogsed, siis on nad võrreldavad ja on võimalik võrrelda eri reklaamkanalitest saadud tulemusi. Selleks et reklaame võrrelda tuvastati andmete seast asjakohased atribuudid. Nendeks atribuutideks olid asukoht (linn, maakond, riik), kategooria ja ametinimi. Andmete visualiseerimise käigus tuli välja, et näitamiste arv ei ole reklaamkanalite vahel võrreldav suurus ning edasine töö peaks põhinema klikkide arvul. Lisaks andis andmete visualisserimine kinnituse, et reklaamide tulemused varieeruvad reklaamkanalite vahel piisavalt et võimaldada tulemuse ennustusest ärilist kasu saada. Ametinimede vektorite klasterdamine andis kinnituse, et ametinimede vektoriline kujutuse põhjal saab teha andmete kohta järeldusi. Kuigi klastrites esines punkte mille puhul ei olnud ilmselge miks nad just sinna klastrisse kuuluvad, siis oli klastri punktide väärtuste põhjal siiski selgelt aru saada, et teatud sarnased jooned on õnnestunud tuvastada.

Ennustusmudeli koostamisel eemaldati andmetest puudulike andmetega read ning üleliigsed veerud, standardiseeriti andmed ning *one-hot encodingu* abil konverteeriti kategooriad ja asukohad veergudeks. Ametinime tähistavad 100 dimensioonilised vektorid konverteeriti nii, et iga vektori dimensioon on eraldi veerus. Lõplik ennustusmudelis kasutusel olev andmestik sisaldab 516 rida ja 574 veergu. Leiti veergude korrelatsioonid ennustatava atribuudiga ning visualiseeriti *heatmapina*. Sealt joonistus välja, et väga tugeva korrelatsiooniga atribuute ei ole kuigi palju. Kõige tihedamini oli klikkidega tugev korrelatsioon kategooria tüüpi kuuluvatel atribuutidel. Masinõppe meetoditena sai kasutatud Random Forest regression, Linear regression, Ridge regression, Gradient Boosting Regression, Support Vector regression ja Extra trees regression meetodeid. Kõige parem tulemus õnnestus saavutada Random Forest regression masinõppe meetodiga. Parima tulemuse leidmiseks sai proovitud erinevaid hüperparameetreid ning valitud sobivaimad. Parim ennustuse tulemus mis saavutada õnnestus Meta reklaamide puhul on R^2: 0.4899 ja MSE: 0.4011 ning see täidab eesmärkides seatud alumise lävendi. Google reklaamide ennustusmudel jäi lävendist allapoole ning eesmärki ei õnnestunud täita, R^2: 0.0913 MSE: 0.4896. Täiendavate

lähteandmete kogumisel ning ametinimede vektoriteks teisendamise täiendamisel on lootust tulemust parandada.

Kasutatud kirjandus

[1] „Kantar Emori uuring: CVKeskus.ee on töötajate peamine tööotsingukanal.“ cvkeskus.ee. Accessed: May 27, 2024. [Online.] Available: https://www.cvkeskus.ee/karjaarikeskus/personaliotsing/varbamine/kantar-emori-uuring-cvkeskusee-on-tootajate-peamine-toootsingukanal

[2] Linder, J. „Global Advertising Industry Statistics,“ gitnux.org. Accessed: May 27, 2024. [Online.] Available: https://gitnux.org/global-advertising-industry/

[3] J. Jackson. „Data Mining: A Conceptual Overview,“ *Communications of the Association for Information Systems,* vol. 8, pp. 267-296, March 2002. DOI: 10.17705/1CAIS.00819.

[4] R. Nau, „Linear regression models.“ people.duke.edu. Accessed: May 27, 2024. [Online.] Available: https://people.duke.edu/~rnau/rsquared.htm#punchline

[5] P. Chapman et. al. *CRISP-DM 1.0. Step-by-step data mining guide*. (2000). [Online]. Available: https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

[6] *Word2vec*. (2013). [Online]. Available: https://code.google.com/archive/p/word2vec/

[7] A. H. Wadud. M. F. Mridha. M. M. Rahman. „Word embedding methods for word representation in deep learning for natural language processing,“ *Iraqi Journal of Science,* vol. 63, no. 3, pp. 1349-1361, March 2022. DOI: 10.24996/ijs.2022.63.3.37.

[8] K. S. Brown *et. al*., „Investigating the extent to which distributional semantic models capture a broad range of semantic relations.“ *Cognitive Science,* vol. 47, no. 5, pp. 1-42, May 2023.

[9] E. Aedmaa. *Pretrained word and multi-sense embeddings for Estonian*. (2023). [Online]. Available: https://github.com/eleriaedmaa/embeddings?tab=readme-ov-file

[10] *Numpy.mean*. [Online]. Available: https://numpy.org/doc/stable/reference/generated/numpy.mean.html

[11] A. E. Odaci. Easygoogletranslate.py. (2021). [Online]. Available: https://github.com/ahmeterenodaci/easygoogletranslate/blame/main/easygoogletranslate.py

[12] „Project Jupyter Documentation.“ Docs.jupyter.org. Accessed: May 27, 2024. [Online.] Available: https://docs.jupyter.org/en/latest/

[13] A. H. Sial. S. Y. S. Rashdi & A. H. Khan. „Comparative analysis of data visualization libraries Matplotlib and Seaborn in Python,“ *International Journal of Advanced Trends in Computer Science and Engineering,* vol. 10, no 1, Jan/Feb. 2021. https://doi.org/10.30534/ijatcse/2021/391012021.

[14] *StandardScaler*. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[15] S. Saxena. *A beginner’s guide to random forest hyperparameter tuning*. (2023). [Online]. Available: https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/

[16] „What is ML?“ Ibm.com. Accessed: 27.05.2024. [Online.] Available: https://www.ibm.com/topics/machine-learning

[17] J. A. Colton & K. M. Bower. „Some misconceptions about R^2“, *Extra Ordinary Sense,* vol. 3, no. 2, pp. *20-22,* Jan. 2002. https://wserver.crc.losrios.edu/~larsenl/ExtraMaterials/MisconceptionsR2.pdf
[18] „Understanding the 3 most common loss functions for Machine Learning Regression,“ towardsdatascience.com. Accessed: May 27, 2024. [Online.] Available: https://towardsdatascience.com/understanding-the-3-most-common-loss-functions-for-machine-learning-regression-23e0ef3e14d3
[19] *Learning Curve.* [Online]. Available: https://www.scikit-yb.org/en/latest/api/model_selection/learning_curve.html
[20] F. Moksony. „Small is beautiful. The use and interpretation of R2 in social research.“ *Review of Sociology. Special issue.* pp. 130-138, Jan. 1999. https://www.researchgate.net/publication/307632302_Small_is_beautiful_The_use_and_interpretation_of_R2_in_social_research_Review_of_Sociology

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

Mina, Erki Toom,

1. Annan Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose “Töökuulutuste reklaamide tulemuse ennustamine", mille juhendaja on Ants Torim,
   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.

27.05.2024

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.