Liiklusõnnetuste prognoosmudel

Proof-of-concept projekti dokumentatsioon

14. Juuli 2020

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

MAANTEEAMET

Selles dokumendis on analüüs liiklusõnnetuste prognoosmudeli *proof-of-concept* projekti väljundite osas ning soovitused jätkuprojekti tulemuslikumaks läbiviimiseks juhul kui seda otsustatakse läbi viia. Käesolev dokument on valminud eelanalüüsi käigus ning täiendatud prototüübi valmimise jooksul.

Projekti eesmärk

Antud andmeanalüüsi projekti väljundiks on liiklusõnnetuste toimumise riski matemaatilise prognoosimudeli loomise eelanalüüs, mis ühtlasi sisaldab mudeli prognoosivõime hindamist ja prognoosimudeli rakenduse prototüübi loomist. Juhul kui eelanalüüsi käigus selgub, et liiklusõnnetuste prognoosimudel on empiiriliselt piisavalt täpne, tekib Tellijal võimalus hinnata kvantitatiivselt, kui palju teatud tegur (mudeli sõltumatu muutuja) mõjutab tulevikus liiklusõnnetuste toimumisriski ja nende raskusastet (mudeli sõltuvat muutujat) valitud kohas ja ajal. Mudel võimaldab Tellijal hinnata, milliseid faktoreid saab kõige efektiivsemalt „kontrolli alla võtta" selleks, et õnnetuste esinemist antud kohas vähendada. Mudeli eduka valideerimise korral saab Tellija kasutada mudelit teehoiu- ja liiklusohutustegevuste planeerimisel ning Politsei- ja Piirivalveamet (edaspidi PPA) saab täiendavaid jõudusid suunata kõrgema liiklusõnnetuse tõenäosusega piirkondadesse liiklust rahustama ning liiklusõnnetusi ennetama.

Andmed

Siin alapeatükis on lühikirjeldus andmetest ning nendega tutvumisel tehtud tähelepanekutest. Täpsem analüüs ning andmete uurimine on kirjeldatud dokumendis LISA 2. Alusandmete uurimine

Projekti läbiviimiseks andis Tellija ligipääsu järgmistele andmetele:

- Liiklusloendusandmed (liiklussagedus, liiklusvoo koosseis, sõidukiirused);
- Ilmaolud (teeilmajaamade andmed);
- Teeregistri andmed (WFS teenus);
- Politsei markeeringuga sõidukite väljapaneku kohad ja aeg;
- Liiklusjärelevalve mahud;
- Liiklusrikkumised;
- Liiklusõnnetuste andmekogu andmed (liiklusõnnetused, neis osalenud sõidukid, hukkunuid ja vigastatuid kirjeldavad andmed nt asukoht, sõiduki liik, osaleja liik jmt);
- Talihooldusmasinate töötsüklid ja maršruudid
- Ajutiste kiiruspiirangutega kohad
- Teavitus- ja koolitustegevused;
- Homogeensete teelõikude andmestik

Lisaks projekti alguses võimaldatud andmetele, hangiti juurde andmed järgnevatest allikatest:

- Ilmateenistuse ilmavaatlused 10-minutiliste ja 60-minutiliste intervallidega
- Liikluskindlustusfondi õnnetuste andmestik

Andmete ettevalmistamine masinõppe mudelite treenimiseks

Siin alapeatükis on lühikirjeldus tegevustest, mis viidi läbi andmete ettevalmistamiseks masinõppe mudelite treenimiseks ja edasiseks analüüsiks. Täpsem ja tehnilisem tegevuste kirjeldus ja ülevaade on antud dokumentides LISA 2. Alusandmete uurimine ning LISA 3. Alusandmete puhastamine ja korrastamine..

Andmete uurimine

Andmete uurimise etapis pöörati eelkõige tähelepanu andmete vormilisele kujule ning andmete sisust arusaamisele. Andmete sisulise poole ühtlustamine ja korrastamine on läbi viidud järgmises osas Andmete puhastamine ja korrastamine.

Andmete uurimise täpsemad vaatlused ning soovitused on välja toodud dokumendis LISA 2. Alusandmete uurimine. Selles dokumendis on antud soovitusi üldiselt teemal andmete hoiustamise põhimõtted ning veidi täpsemalt proleemide koha pealt mis esinesid mitme andmeallika puhul nagu näiteks tagasiühilduvus ja tulpade nimetamine. Samuti on igast andmeallikast antud lühikirjeldus ning konkreetse andmeallikaga seonduvad probleemid.

Alusandmete uurimise põhjal pandi paika plaan andmete puhastamiseks ning korrastamiseks. Samuti lisati uurimise põhjal lisaandmeid, nagu näiteks teeilmajaamade metainfo ning sarnaste teelõikude info.

Andmete puhastamine ja korrastamine

Kuna antud juhul on tegemist küllaltki suure andmehulgaga (masinõppe protsessi kaasati ca 10 miljonit andmepunkti päeva kohta, ehk ühtekokku ca 3 miljardit andmepunkti), siis läheneti projekti käigus andmete töötlemisele küllaltki minimalistlikult, et säilitada algupärased seosed. Suurandmetega masinõppes on üldiselt trend algandmeid muuta võimalikult vähe ning lasta pigem seoste leidmise ja õppimise töö ära teha masinõppemudelil, et vältida eelnevalt kallutatud või vale signaali sisseviimist andmetöötluse käigus. Seetõttu on andmetes küllaltki vähe lisatunnuseid ning samuti on andmete imputeerimisse ja puuduvate väärtuste täitmisesse lähenetud minimalistlikult. Samuti süstemaatiliste vigade parandamine on küllaltki lihtsate meetoditega lahendatud, sest testide käigus ilmnes, et need ei oma selget mõju mudeli ennustuskvaliteedile ning seetõttu pole ka tõestusprojekti raames oluline nende ülipõhjalik lahendamine..

Üks tähtsamaid aspekte andmete korrastamisest oli **puuduvate tunnuste täitmine**. Kuna erinevad mudelitüübid eeldavad erinevat puuduvate tunnuste täitmist, siis siinkohal toome ära ainult ülevaatlikult puuduvate tunnuste täitmise Random Forest algoritmi kohta, millega me lõplikud katsed tegime. Tunnuste *'katlai', 'slai', 'plaiv', 'plaip', 'inspire_functionalclass'* puudvad väärtused täideti vastavalt *'inspire_functionalclass'* tunnuse kaupa grupeeritult teede keskmiste väärtustega. Tunnuste *'inspire_functionalclass', 'valh_hoold_xv', 'tas_moots_xv', 'kpp', 'kpv', ''kptp_kptyyp_xv', 'tolm_ttmeetod_xv', 'yl_ylliik_xv', 'defosa', 'poikpr', 'kpikipr', 'lpikipr', 'kvuuk', 'lvuuk', 'vork', 'auk', 'muren', 'serv', 'lapp', 'hoolklt_hoolklt_xv', 'hoolkls_hoolkls_xv', 'defaasta'* puuduvad väärtused täideti vastavalt *'inspire_functionalclass'* tunnuse kaupa grupeeritult teede keskmiste väärtustega. Tunnuste *'precint', 'roadstat', 'snow_thick', 'precipcl', 'precipaccu', 'waterthic1', 'preciptype', 'grip', 'snow',*

'ice_thick', 'blackicethic2' puuduvad väärtused täideti väärtusega -1. Tunnuse *'icetemp1'* puuduvad väärtused täideti väärtusega 20. Selline puuduvate tunnuste asendamine diametraalselt erineva väärtusega, mis ei esine andmestikus, on otsustuspuude põhisele lähenemisele sobiv. Eriti antud juhul, kus teatud väärtuste, nagu jää temperatuur, kus vöärtuse puudumine kannab muuhulgas signaali. Kuna otsustuspuud töötavad andmestiku jaotamisel väärtuste põhjal, on oluline et sellised täidetud väärtused oleksid erinevad mitte-puuduvatest väärtustest. Tunnuste 'suund_soidusuund_xv', 'kate_kate_xv', 'rdt_raudtee_xv' täitmisel kasutati vastavalt väärtusi 9, 99, -1. Ülejäänud tunnuste puuduvate väärtuste täitmisel kasutati väärtust -1.

Lisaks puuduvate tunnuste täitmisele jäeti osad andmehulgad analüüsist välja:

- Liiklusloenduste andmehulk jäeti analüüsist välja, kuna neil lõikudel, kus loendusandmed vahetult puudusid, tekitas lähima/sarnase lõigu imputeerimine olulisel määral valepositiivseid. Nende andmete kasutamiseks oleks vaja välja mõelda hea süsteem puuduvate väärtuste imputeerimiseks.
- Kiiruste mõõdistuste andmehulk jäeti kasutusest välja, sest selle parsimine oleks kulutanud ebamõistlikult palju aega, arvestades, et tegemist on tõestusprojektiga. Täpsem selgitus on leitav LISA 2. dokumendis.
- Ennetustegevuste andmestikust kasutati ainult Meediakampaaniate infot. Koolituste info polnud piisavalt hästi ühendatav teede andmestikuga.

Treeningandmestiku koostamine

Treeninandmestiku koostamise esimeseks osaks on geomeetriliste andmete põhja koostamine. Täpsemalt tähendab see, et arvutatakse välja teelõigud, mis on arvutuslikult piisavalt erinevad, et neid eraldi teelõikudena käsitleda ning seejärel rakendatakse neile kõik ülejäänud unikaalsed geomeetrilised parameetrid, et koostada ühene geomeetriline põhi. Geomeetriline põhi koostatakse esimesena, sest geomeetriliste ühenduste loomine on arvutuslikult väga kulukas ning seetõttu pole seda otstarbekas iga kuupäeva tarvis korrata. Geomeetrilise põhja koostamine ning sellega seotud probleemid on täpsemalt kirjeldatud dokumendis LISA 7. Geomeetrilise andmestiku koostamine.

Geomeetriliste andmete ning ajas muutuvate andmete korrastamine ja ühendamine ühtseks treeningandmestikuks, on puhta koodi kujul kirjeldatud koodibaasi (https://git.mnt.ee/poc/poc) failides *train_gen_fulldata.py* ja *make_yearly_mrg_X.py*.

Treeningandmete tunnuste kirjeldused

Mudeli sisendiks sobivale kujule töötletud treeningandmete tunnuste kirjeldused on dokumendis LISA 4. Treeningandmete tunnuste kirjeldused. Neile tunnustele lisanduvad veel teelõigu asukoha määramiseks vajalikud tunnused nagu tee number, sõidutee number, algus kilomeeter, lõpp kilomeeter mida küll ei kasutata treenimisel, kuid mida on kasutatud treeningandmete loomisel ja unikaalsuse määramisel.

Mudelite treenimine

Selles alapeatükis antakse teoreetiline ülevaade mudeli treenimise käigus läbi viidavatest tegevustest. Tehniline ülevaade mudeli treenimisest (ja ennustuste tegemisest) on dokumendis LISA 5. Mudelite treenimine ja ennustuste tegemine

Treenimisprotsessi kirjeldus

Mudelite treeningprotsessi saame laias laastus jagada viieks etapiks: Andmete sisselugemine, andmete jaotamine treening-ja testandmestikuks, andmete kodeerimine ja standardiseerimine, mudeli treenimine, mudeli ennustusvõime hindamine.

Andmete sisselugemine on mudeli treenimise protsessi esimene etapp. Andmed loetakse sisse sellisel kujul nagu nad on eeltöötluse käigus salvestatud. Meie praeguses lahenduses loetakse andmed sisse serveri asukohast, kus need asuvad parquet (https://en.wikipedia.org/wiki/Apache_Parquet) failide kujul. Parquet failikuju on valitud eesmärgiga vähendada failide mahtu kettal ning samuti et kiirendada failide sisselugemist analüüsi jaoks. Kuna andmestik kogu liilusõnnetuste ja teeolude ajaloo kohta on väga mahukas, siis andmete sisselugemisel jäetakse alles ainult valim andmetest. Selleks, et suunata mudelit tugevamalt ennustama riskantseid teelõike ja olusid ning ühtlasi tekitada ka neis ennustustes rohkem eraldatust võimaldamaks riskantsete lõikude ja olude järjestamist, on kaasatud kõik õnnetustega seotud kirjed ning vähendatud valim neis andmetest, mis ei ole seotud õnnetustega ning õnnetustega seotud andmed on kõik kaasatud. Seeläbi on õnnetuste signaal andmestikus võimendatud ning ka ennustustes riskid selgemalt vaadeldavad.

Andmed jaotatakse treening-ja testandmestikuks enne andmete kodeerimist ja standardiseerimist, et vältida igasugust andmeleket, mis võib tekkida testandmete eeltöötluse protsessi lisamise tõttu. Treening- ja testandmestik jaotatakse kuupäeva põhiselt. See tähendab, et andmed ühest ja samast kuupäevast ei saa esineda nii treening- kui testandmestikus. Ka selle tegevuse eesmärgiks on vältida andmeleket. Andmed ühest kuupäevast võivad näiteks jagada omavahel ilmastikutingimuste, hooldemasinate ning politseipatrullide kohalolu infot, mistõttu võiks ühe ja sama kuupäeva kasutamine nii treening- kui testandmestikus tekitada andmeleket. Selline jaotamine on oluline ühtlasi ka simulatsioonivõimekuse hindamiseks. Kui me teame kui täpselt suudab mudel ennustada seni nägemata kuupäeva liiklusõnnetuste riskihinnanguid ja tegureid, siis annab see kindluse, et simulatsioonifunktsionaalsus toimib ligikaudu sama kvaliteediga. Kuna simulatsioonifunktsionaalsus oleks oluline vahend tegevuste planeerimisel, siis selline treening- ja testandmestiku jaotamine on meie hinnangul kõigi püstitatud eesmärkide täitmiseks kõige otstarbekam.

Andmete kodeerimine ja standardiseerimine viiakse läbi kahes jaos. Esiteks tuvastatakse andmestikust kategoorilised tunnused. Seejärel kodeeritakse need tunnused numbrilisele kujule, et need oleks mudeli poolt töödeldavad. Kodeerimise järel rakendatakse standardiseerimist. Kuigi prototüübi näidiskoodis kasutatud Random Forest algoritm ei vaja skaleerimist ja standardiseerimist, ei ole see tegevus ka mudeli ennustusvõimele kuidagi kahjulik. Seetõttu on standardiseerimise protsess eeltöötluse protsessi sisse jäetud, et muuta mudeliversiooni välja vahetamine ning teiste mudelite versioonide katsetamine

lihtsamaks. Standardiseerimine ja skaleerimine on läbi viidud kasutades scikit-learni StandardScaler meetodit.

Mudeli treenimine viiakse läbi eeltöödeldud treeningandmetel. Treeningandmed antakse mudelile töötlemiseks juhuslikul järjestuses, et vältida viimaste andmete liigse mõju avaldumist. Näiteks kui andmed anda töötlemiseks kuupäevaliselt järejstatult, siis võib vastavalt mudeli eripärale jääda mudelile "meelde" valdavalt kas siis esimese kuu või viimase kuu andmed, mis tähendab, et ennustusvõime üldandmestikul kannatab. Juhusliku järjestuse puhul on see probleem kõrvaldatud. Näidiskoodis on mudeli treenimine paralleliseeritud kiirema treenimise eesmärgil ning sisendklasside kaalud on võrdsustatud, et suunata mudelit rohkem tähelepanu pöörama õnnetustega seotud andmetele, kuna loomulikus andmestikus on neid oluliselt vähem kui neid andmepunkte, mis pole õnnetustega seotud. Kuna võib eeldada, et koguandmestik annab mudelile piisavalt esindusliku ülevaate andmete struktuurist täitmaks meie tõestusprojekti eesmärke, näidatakse treeningandmeid mudelile vaid ühe korra. Täisprojekti lahenduse käigus tuleb tõenäoliselt neid iteratsioone korrata, et saavutada parim võimalik tulemus.

Mudeli ennustusvõime hindamine viiakse läbi kahe seotud protsessina: arvutuslik hinnang tulemustele ning eksperdi hinnang tulemustele.

Arvutuslik hinnang kujutab endast mudeli töötäpsuse hinnangut ehk teisisõnu seda kui täpselt suudab mudel ennustada testandmete riskihinnanguid. Käesoleva projekti eripäraks on see, et hea tulemus tähendab seda, et mudel peab tegema parajal hulgal nii valepositiivseid kui valenegatiivseid ennustusi. Valepositiivsed ennustused on olulised toomaks välja neid teelõike ja tingimusi, mis kujutavad endast riski liiklusohutusele, kuid kus õnnetusi pole (veel) juhtunud. Selliste lõikude ja tingimuste leidmine on oluline tõhustamaks ennetustegevust. Teisalt on oluline ka paras hulk valenegatiivseid ennustusi, sest arvestatav osa liiklusõnnetustest toimub lähtuvalt teguritest, mille kohta pole meile kättesaadavates andmestikes infot, mis tähendab, et need konkreetsed liiklusõnnetused peaks olema mudeli poolt viletsasti üles leitavad ehk need peaksid tekitama ennustustes valenegatiivseid. Kuna paraja hulga valepositiivsete ja valenegatiivsete määramisel on oluline osa tunnetusel, oli mõistlik valida ennustusvõime hindamiseks küllaltki lihtsad meetrikud. Antud juhul kasutasime ennustusvõime hindamiseks segadusmaatriksit ning klassipõhist F1 skoori.

Eksperdi hinnang tulemustele viidi läbi testandmete ennustuste uurimusliku valideerimise kujul. Nende valideerimiste käigus uuriti nii ennustusi ja nende muutumist konkreetsetel teadaoleva riskihinnanguga lõikudel kui ka vaadati riskihinnangute ja juurpõhjuste üldstatistika muutumist vastavalt teelõikudele ja vaadeldud ajaperioodile. Eksperdi hinnangud tulemustele anti kokkuvõtvate vaatluste kujul, kus anti ühtlasi soovitusi millistes aspektides peaks mudeli tööd parandama ning kuidas peaks väljundite kuju muutma, et nende kasutamisest saadavat kasu ja infot maksimeerida.

Lähtuvalt arvutuslikule ja ekspertihinnangutele tehti töö käigus iteratiivselt täiendusi nii andmetesse kui ka mudelite treenimisse misjärel siis saadeti andmed uuesti valideerimiseks. Lõplik mudeli versioon kujunes kolme sellise iteratsiooni tulemusena millele lisandusid veel jooksvad konsulteerimised ekspertidega.

Mudelite versioonid

Mudeleid testiti kahes järgus. Esimeses järgus võeti laiem valik mudeleid ning testiti neid vähendatud andmete. Testi tulemused on täpsemalt kirjeldatud selle dokumendi alapeatükis "Tulemused". Samuti valiti esimeses testimise järgus mudelitüüpide baasarhitektuurid, et saada tõestusprojekti jaoks sobiv ülevaade, millised mudelid võiksid olla kõige mõistlikumad kontrollimaks olemasolevate andmete pealt riskide ennustatavust.

Esimeses järgus valiti testimiseks järgmised mudelitüübid:

- Random Forest Classifier
- XGBoost classifier
- DNN classifier
- K-Nearest Neighbor Classifier
- Decision Tree Classifier
- Ada Boost Classifier
- GaussianNB Classifier
- QuadraticDiscriminantAnalysis Classifer

Esimese järgu testimistest andsid kõige paljulubavamaid tulemusi (kirjeldatud "Tulemused" alapeatükis) Random Forest Classifier ja XGBoost Classifier (mis mõlemad põhinevad tegelikult metsadel) millega viidi läbi testid täisandmetel. Täisandmetel läbi viidud testide alusel hindasime täisanalüüsi läbiviimiseks kõige sobivamaks Random Forest Classifierit. Iseenesest ei ole selline tulemus üllatav, sest Random Forest Classifier on väga erinevate andmestike puhul hea *baseline* mudel, et kontrollida hüpoteesi pidavust. Random Forest Classifierit hindasime antud juhul XGBoostist sobivamaks eelkõige seetõttu, et testitud parameetrite põhjal olid tehtud ennustused mitte ilmtingimata täpsemad vaid pakkusid parema jaotuse õnnetustega mitteseotud teelõikude riskide hindamiseks. Samuti kallutas otsust Random Foresti kasuks ennustamise kiirus.

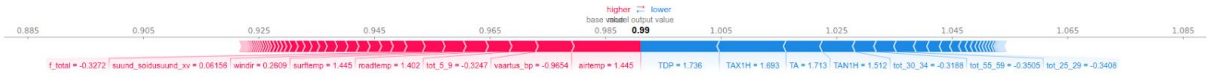
Väljundi disain

Nagu ka eelnevalt mainitud, siis käesoleva projekti eripäraks on see, et hea tulemus tähendab seda, et mudel peab tegema parajal hulgal nii valepositiivseid kui valenegatiivseid ennustusi. Valepositiivsed ennustused on olulised toomaks välja ohtlike teelõike ja tingimusi, kus õnnetusi pole (veel) juhtunud ning valenegatiivsed ennustused on olulised indikaatorid selles, et mudel ei sobita end ennustama neid õnnetusi mille juurpõhjused on väljaspool meile saadaolevaid andmeid (kriminaalne joove, oluline kiiruseületus, jne).

Eelnevat arvesse võttes hindasime kõige mõistlikumaks väljundiks jaotada õnnetused kolme gruppi: 0-õnnetuse puudumine, 1-kerge vigastatutega liiklusõnnetus (=viis või vähem vigastatut ja 0 hukkunut), 2-raske liiklusõnnetus. Mudeleid treenitigi viisil, mil andmetega viidi vastavusse üks neist klassidest. Täpsem kirjeldus treenimisest koos koodiga on ära toodud dokumendis LISA 5. Mudelite treenimine ja ennustuste tegemine. Kuna teelõikude ja tingimuste riskiastmed on erinevad, oli algusest peale selge, et lihtsalt ennustatud klassiga ei ole otstarbekas väljundit esitada. Seetõttu valisime ekspertide valideerimiseks põhiväljundiks õnnetusklasside ennustustõenäosused, mida saab

tõlgendada ka kui riskihinnangute skoori. See tähendab, et kui konkreetse klassi ennustamisel sai andmepunkt endale hinnanguks kas 0, 1, või 2, siis riskiskooriga sai see sama andmepunkt endale hinnanguks näiteks 0.37, 0.46, 0.17, kus iga väärtus tähistab siis vastavalt klassi 0, 1, 2 tõenäosust ehk riskihinnangut. Sellise väljundi alusel on võimalik iga teelõigu ja tingimuse kohta arvutada nii keskmiseid riskihinnanguid kui ka teelõike omavahel järjestada riskantsest vähemriskantsemani. Näiteks teelõik keskmise riskiskooriga 0.14, 0.54, 0.11 on mudeli väljundite alusel ilmselgelt riskantsem kui teelõik mille keskmine riskiskoor on 0.72, 0.11, 0.05. Seeläbi on võimalik valida ennetustegevuses prioriteetsemad (riskantsemad) teelõigud ning uurida läbi juurpõhjuste hinnangu ning eksperthinnangu milliste nende teelõikude puhul on riskid kõrvaldatavad ning seeläbi juba ka paremini planeerida kulutusi teede liiklusohutuse suurendamiseks.

Selleks, et riskihinnangute põhjused oleksid paremini vaadeldavad, lisasime riskihinnangutele ka juurpõhjuste analüüsi. Juurpõhjuste hinnangud anti igale ennustatud andmepunktile (Andmepunkt on üks teelõik teatud ajahetkel. Ajahetked on võetud 10 minutiliste intervallidega.). Juurpõhjuste hinnangud anti lähtuvalt kõige tõenäolisemast klassist. Näiteks kui konkreetse andmepunkti riskiklasside tõenäosused on 0.37, 0.46, 0.17, siis juurpõhjuse hinnang antakse lähtuvalt klassist 1. Juurpõhjuse hinnang näitab, et millised parameetrid mõjutasid kui palju konkreetse ennustuse väärtuse kujunemist. Positiivse väärtusega parameetrid näitavad seda, kui palju need konkreetsed parameetrid riskihinnangut tõstsid ning negatiivse väärtusega parameetrid näitavad kui palju need konkreetsed parameetrid antud riskihinnangu puhul riski leevendasid. See riski tõstmine või leevendamine on arvutatud keskmise riskiväärtuse suhtest lähtuvalt ennustuste aluseks oleva mudeli õpitud parameetritest. Näidisvisualiseering ühe konkreetse andmepunkti riskihinnangu kujunemisest on toodud alloleval joonisel.



Põhiversiooni tööparameetrid

Kuna andmestik on küllaltki mahukas, siis seetõttu on ka mudelite töö küllaltki aeganõudev. Siinkohal toome välja peamiste etappide tööparameetrid, et lõppkasutaja oskaks paremini arvestada vajaminevate ressurssidega:

1. Treeningandmete koostamisel kulub ühe päeva andmete koostamiseks keskmiselt 40 minutit ning tipphetkedel ca 120GB mälu.
2. Random Forest mudeli treenimiseks kaasa pandud versiooni juures kulub keskmiselt 4 tundi ning tipphetkedel ca 100GB mälu.
3. Random Forest mudeli abil täisandmete ennustamise puhul kulub ühe päeva ennustamisele koos juurpõhjustega keskmiselt 1 tund ning tipphetkedel ca 60GB mälu.
4. Ühe päeva andmestik koos riskihinnangute ja juurpõhjuse hinnangutega võtab ligikaudu 1.5GB salvestusruumi.

Ennustuste tegemine

Selles alapeatükis antakse teoreetiline ülevaade riskihinnangute ja juurpõhjuste ennustamise käigus läbi viidavatest tegevustest. Tehniline ülevaade neist tegevustest on dokumendis LISA 5. Mudelite treenimine ja ennustuste tegemine

Riskihinnangute ennustused

Riskihinnangute ennustused testandmetel tehakse eelnevalt treeningandmetele sobitatud mudeliga. See tähendab, et ennustused tehakse andmetel, mida treenimisel nähtud ei ole. Selleks, et lõppversioonis kogu andmestik riskihinnangute ennustused saaks, viiakse lõplik ennustamine läbi ristvalideerimise põhimõttel, kus võetakse korduvalt erinev andmehulk treenimiseks nii, et ristvalideerimise lõpuks kogu andmestik endale ennustusväärtused saaks. Samuti nagu ka treenimise puhul, viiakse siin läbi ennustus paralleliseeritud moel, et kiirendada ennustuse protsessi. Uute näidete ennustamisel (näiteks simulatsiooni funktsionaalsuses) viiakse ennustamisele eelnevalt läbi samasugune andmete eeltöötluse protsess nagu on eelnevalt kirjeldatud treeningandmete puhul. Selleks kasutatakse muuhulgas ka treeningandmete peal sobitatud kategoriseerimismudelit ning standardiseerimisparameetreid.

Nagu on kirjeldatud eelnevalt väljundite disaini juures, siis riskihinnangute ennustuse peamiseks väljundiks planeeritud kasutuse jaoks on konkreetsete näidete riskiklasside tõenäosused, kus iga individuaalse andmepunkti iga riskiklass (0, 1, 2) saab endale vastavuses riskiklassi tõenäosuse (näiteks 0.37, 0.46, 0.17). Need riskiklasside tõenäosused võimaldavad siis omakorda jälgida nii konkreetse lõigu riskihinnangute muutumist ajas, mingi hulga lõikude riskihinngangute muutumist ja statistikat, kui ka järjestada erinevaid lõike ja tingimusi vastavalt siis riskihinnangute tõenäosuste alusel.

Juurpõhjuste ennustused

Lisaks riskihinnangutele ennustatakse ka konkreetsete andmepunktide juurpõhjuste hinnangud. Juurpõhjuste hinnangud anti lähtuvalt kõige tõenäolisemast klassist. Näiteks kui konkreetse andmepunkti riskiklasside tõenäosused on 0.37, 0.46, 0.17, siis juurpõhjuse hinnang antakse lähtuvalt klassist 1. Juurpõhjuse hinnang näitab, et millised parameetrid mõjutasid kui palju konkreetse ennustuse väärtuse kujunemist. Positiivse väärtusega parameetrid näitavad seda, kui palju need konkreetsed parameetrid riskihinnangut tõstsid ning negatiivse väärtusega parameetrid näitavad kui palju need konkreetsed parameetrid antud riskihinnangu puhul riski leevendasid. Mida kõrgem on konkreetse parameetri juurpõhjuse väärtus, seda tugevamalt konkreetne parameeter tõstab riski antud andmepunkti puhul ning mida madalam on konkreetse parameetri juurpõhjuse väärtus, seda rohkem konkreetne parameeter leevendab antud andmepunkt liiklusõnnetuse riski,

Juurpõhjuste ennustuste tegemine põhineb meie näidetes riskihinnangute aluseks oleval mudelil ning seetõttu juurpõhjuste hinnangute mudel eraldi kalibereerimist ei vaja. Sõltuvalt mudelitüübist võib siiski see vajadus tekkida. Juurpõhjused arvutatakse statistilisel moel lähtuvalt sellest millised teed mudeli parameetrites ennustuse tegemisel aktiveeritakse. Juurpõhjuse määramine on arvutuslikult oluliselt kulukam kui riskihinnangute määramine,

mistõttu võtab see ka oluliselt rohkem aega (ligikaudu 100 korda kauem näidiskoodis kasutatud riskihinnangu mudeliga võrreldes).

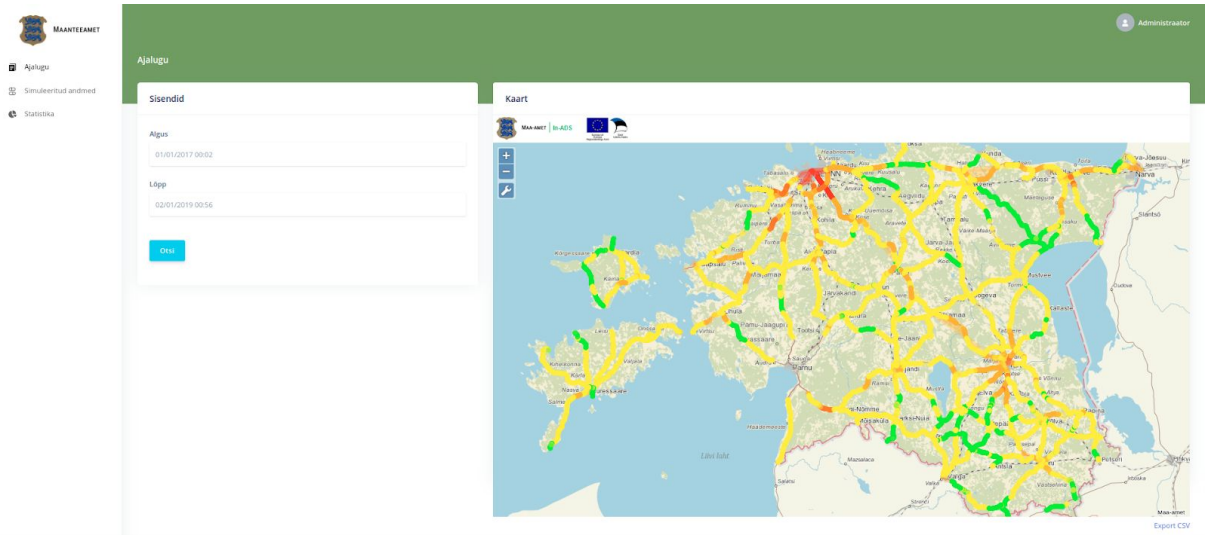
Kuna juurpõhjuste hinnangu puhul tegemist on siiski heuristilise hinnanguga juurpõhjustele, siis neid hinnanguid tasub üksikute andmepunktide puhul võtta pigem indikaatorina kui absoluutse tõena. Kõige rohkem kasu on juurpõhjuste hinnangu analüüsimisel pikema perioodi pealt. Näiteks kui vaadata kuu lõikes riskantsemate teelõikude juurpõhjusi, siis joonistub sealt oluliselt selgemalt välja, millised tegurid on konkreetse teelõigu puhul stabiilselt riski tõstvad ning millega seetõttu oleks ka põhjust tegeleda. Lisaks sellele on oluline tähele panna, et mõned tegurid (näiteks rahvastikutihedus) on universaalselt kõrgel kohal riski tõstjana juurpõhjuse alusel. Seetõttu võib olla kasulik otsida riskantsete teelõikude tugevate mõjurite hulgast tegureid, mis tavapäraselt ei ole teelõikudel kõrge juurpõhjuse väärtusega. Seeläbi on võimalik paremini tuvastada millised teelõigu iseärasused muudavad antud teelõigu võrreldes teistega riskantsemaks.

Rakenduse prototüüp

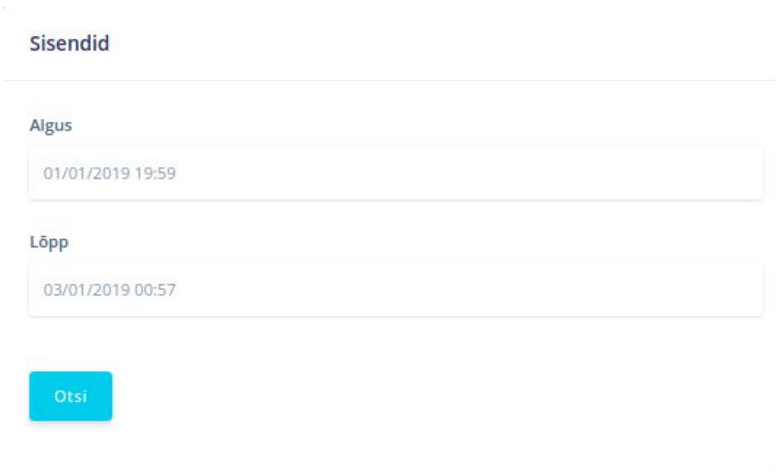
Projekti tulemite kasutamiseks ning andmete simuleerimiseks mudeli testimise eesmärgil loodi projekti käigus prototüüp. Prototüübil on kolm põhivaadet: Ajaloo vaade, Simulatsiooni vaade, Üldstatistika vaade. Raportis antakse kõrgel tasemel ülevaade prototüübi sisust. Täpsema kasutusjuhendi leiate dokumendist “Realiseeritud funktsionaalsuste kasutusjuhend”, mille link on leitav projekti confluence dokumentatsiooni esilehelt. Abiks tulemuste tõlgendamisel prototüübis on selle sama raporti tulemuste peatüki alaosa “Lõplikud tulemused” alguses olev tähelepanek nr 4.

Ajaloo vaade

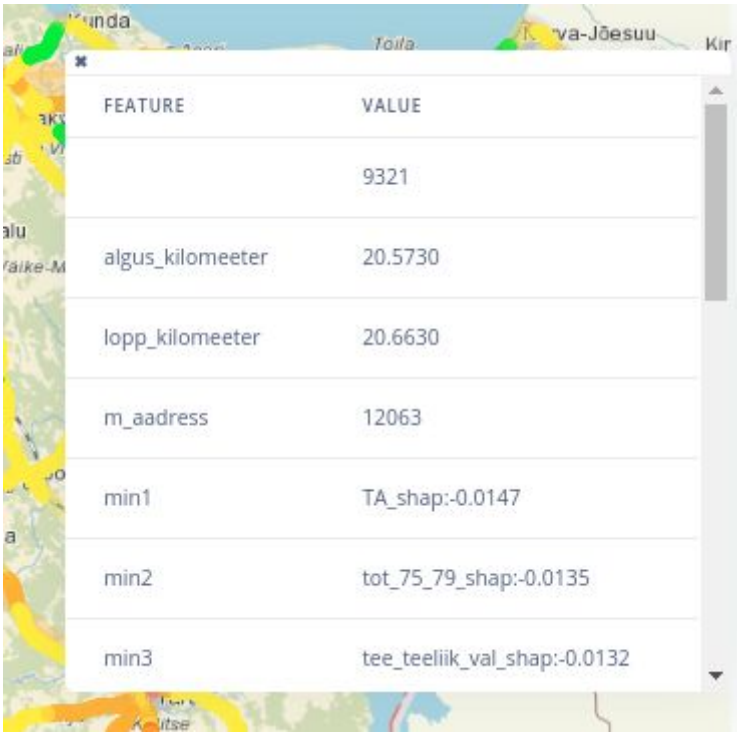
Ajaloo vaate eesmärgiks on kaardil kuvatuna vaadata mudeli poolt ennustatud väärtusi konkreetsete teelõikude kohta. Ennustatud riskiklasside väärtused on eelnevalt arvutatud reaalsete sisendandmete pealt. Kuvatakse agregaatväärtused riskidele ja juurpõhjustele valitud perioodi kohta ning tunnused teelõigu asukoha määramiseks Maanteeameti enda andmestikes.



Vaadeldavat perioodi saab kasutaja muuta vastavate parameetritega.



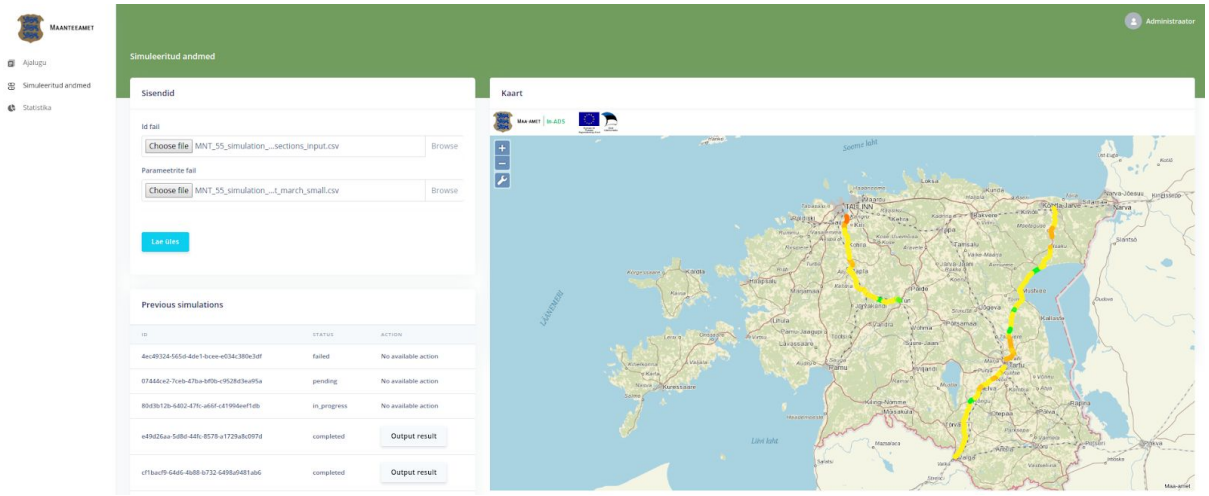
Teelõigu peale vajutades kuvatakse Teelõigu info, riskiklasside tõenäosushinnangud (klassidele 0, 1, 2) ja ennustust kõige rohkem mõjutanud (nii riski tõstvad kui langetavad) tegurid. Tegurite nimed on lahti seletatud dokumendis LISA 4. Tunnuste kirjeldused.xlsx.



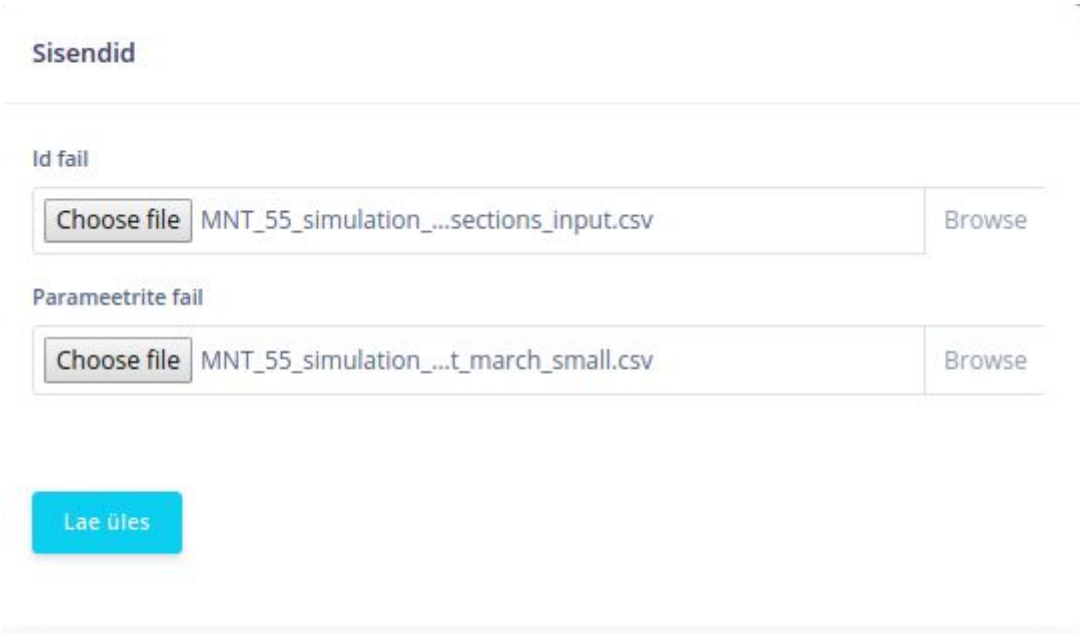
Simulatsiooni vaade

Simulatsiooni vaade võimaldab kaardil kuvatuna vaadata mudeli poolt ennustatud väärtusi konkreetsete teelõikude kohta simuleeritud tingimustel. Ennustatud riskiklasside väärtused arvutatakse simuleeritud sisendandmete pealt. Simulatsioonivaate peamine eesmärk on vastata küsimusele “Kui palju mingi ennetustegevus mõjutab liiklusohutust?”. See tähendab, et kasutaja saab nö läbi mängida erinevad “mis oleks kui” stsenaariumid, et vaadata millised muudatused valitud teelõikude riskihinnangule kõige rohkem leevendavat (või raskendavat mõju omavad). Näiteks saab kasutaja lisada ennetuskampaaniaid või

muuta teekatte parameetreid (laius, katte materjal, defektide hulk/iseloom), et kontrollida, kas mudeli hinnangul võiks nende parameetrite muutmine ja puuduste kõrvaldamine leevendada valitud teelõikude riske. Samuti saab simuleerida ilmastikutingimuse ja teehoolde seoseid ning mitmeid teisi parameetreid. Loetelu parameetritest on täpsemalt välja toodud dokumendis LISA 4. Tunnuste kirjeldused.xlsx.

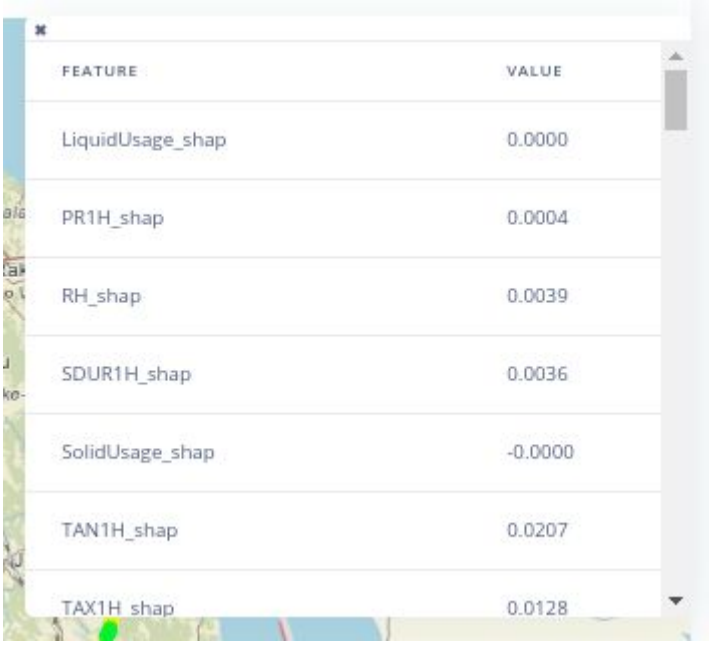


Simulatsiooni tingimused ja teelõigud defineeritakse kasutaja poolt üles laetud failidega.



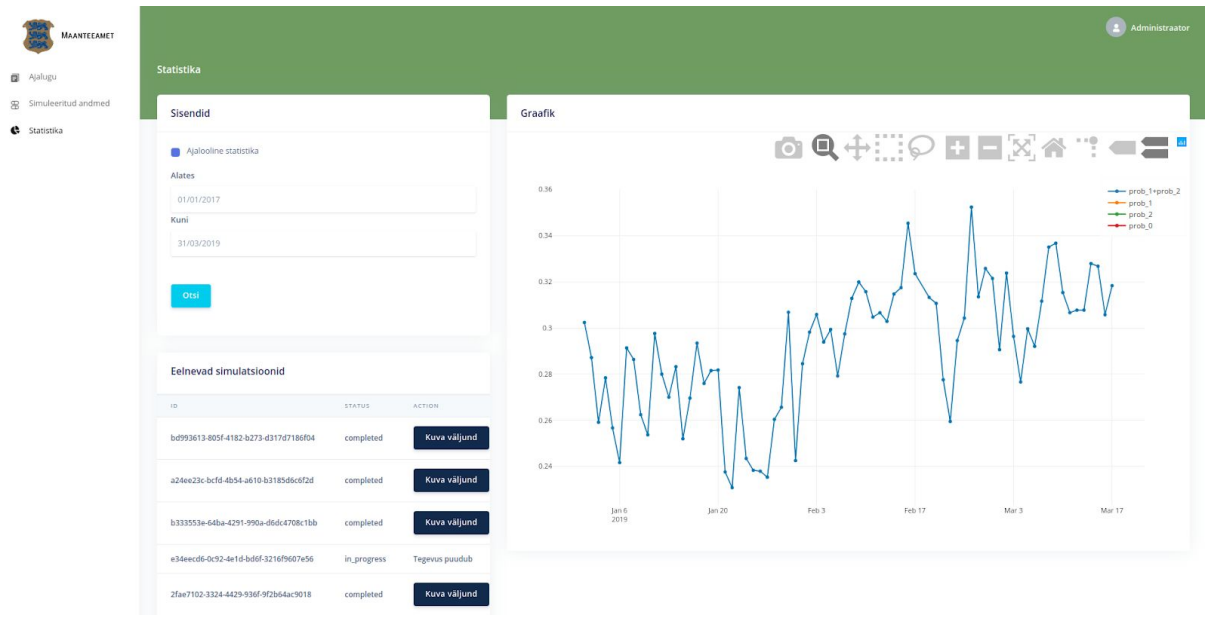
Simulatsiooni eesmärgiks on uurida, kuidas tingimuste muutudes muutuvad mudeli riskihinnangute ja juurpõhjuste ennustused. SImulatsiooni protsessis saab kasutaja ise määrata parameetrite sisendväärtused nii nagu on kirjeldatud prototüübi kasutusjuhendites. Seejärel kombineeritakse sisend vajadusel ajalooliste vaatlustega ning edasine ennustusprotsess toimub nii nagu ka ajalooliste andmete puhul riskide ennustamine. Simulatsiooni protsess on koodina kirjeldatud failis MNT_55_Full_simulation_with_root_cause_data_base_cluster.py, mida saab vaadata koodibaasis.

Simulatsiooni vaates kuvatakse agregaatväärtused riskidele ja juurpõhjustele valitud perioodi kohta valitud teelõikudele ning tunnused teelõigu asukoha määramiseks Maanteeameti enda andmestikes. Vaadeldavat perioodi ja simuleeritavaid tingimusi saab kasutaja muuta vastavate parameetritega. Teelõigu peale vajutades kuvatakse Teelõigu info, riskiklasside tõenäosushinnangud (klassidele 0, 1, 2) ja ennustust kõige rohkem mõjutanud (nii riski tõstvad kui langetavad) tegurid.



Üldstatistika vaade

Üldstatistika vaade võimaldab graafikutena kuvatuna vaadata mudeli poolt ennustatud väärtusi summaarselt valitud teelõikude kohta kas ajalooliste andmete või simuleeritud tingimuste alusel. Graafikutena kuvamise eesmärk on eelkõige vaadelda kuidas ennustatud risk muutub ajas või tingimusi simuleerides. Kuvatud on prob_0, prob_1, prob_2 ning prob_1+prob_2. Näiteks kas talvel on teepinna temperatuuril olulisem roll riski hindamisel kui suvel, või kas näiteks suvel sademete hulka suurendades riskihinnangud tõusevad.



Tulemused

Riskihinnangute ennustuste tulemused

Esimene treeningiteratsioon

Mudelite esmase valiku jaoks treeniti 8 erinevat mudelitüüpi vähendatud andmehulga peal. Andmehulka oli valitud juhuslikult 1 miljon andmerida, kus riskantsete näidete hulk oli võimendatud ca 3% peale juhtudest. Mudeleid võrreldi üldjoontes kahe meetriku põhjal: Segadusmaatriks (*confusion matrix*) ja F1 skoor. Esialgne valik tehti ainult automaatsete meetrikute pealt, et filtreerida välja kõige paljulubavamad mudeliversioonid. Esialgse valiku tegemisse maanteeameti poolseid eksperte ei kaasatud.

F1 skoorid

Antud analüüsi tarvis on F1-skoorid välja toodud klassipõhisena, kuna klasside kaal andmestikus on väga erinev ning samuti nende ennustused käituvad mudelites väga erinevalt. Tavapärane lahendus oleks esitada üks F1 skoor, kuid antud eesmärgiks ei oleks see meie hinnangul olnud piisavalt informatiivne.

Meetod	**0-klassi F1**	**1-klassi F1**	**2-klassi F1**
KNEigborClassifier	0.98794618	0.20883978	0.01965924
DTClassifier	0.99098271,	0.19096585	0.06208054

RFClassifier	0.98613269	0.28216259	0.10225198
DNN	0.99141457	0.12372014	0.0
AdaBoostClassifier	0.99048522	0.10112768	0.02008368
GaussianNB	0.95298312	0.06443739	0.04365108
QuadraticDiscriminantAnalysis	0.10441919	0.05164182	0.01232807
XGBoost	0.92683286	0.11212538	0.07223614

Segadusmaatriksid

Segadusmaatriksid on kasulikud vaatamaks detailsemalt mudelite ennustuste jaotust, et paremini hinnata mudeli sobivust.. Hea selgitus segadusmaatriksist on näiteks siin: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd

KNEigborClassifier	Pred 0	Pred 1	Pred 2
True 0	173430	1112	416
True 1	1688	378	8
True 2	1016	56	15

DecisionTreeClassifier	Pred 0	Pred 1	Pred 2
True 0	174573	317	68
True 1	1814	260	0
True 2	978	72	37

RFClassifier	Pred 0	Pred 1	Pred 2
True 0	172340	2159	459
True 1	1346	715	13
True 2	883	120	84

DNN	Pred 0	Pred 1	Pred 2
True 0	174889	69	0
True 1	1929	145	0
True 2	1031	56	0

AdaBoost Classifier	Pred 0	Pred 1	Pred 2

True 0	174471	464	23
True 1	1862	139	73
True 2	1003	72	12

GaussianNB	**Pred 0**	**Pred 1**	**Pred 2**
True 0	161199	1808	11951
True 1	1445	132	497
True 2	702	83	302

QuadraticDiscriminant	**Pred 0**	**Pred 1**	**Pred 2**

Analysis			
True 0	9644	7877	157437
True 1	69	265	1740
True 2	46	47	994

XGBoost	**Pred 0**	**Pred 1**	**Pred 2**
True 0	151649	17953	5356
True 1	540	1233	301
True 2	93	742	252

Selle iteratsiooni põhjal said välja valitud edasiseks treenimiseks Random Forest Classifier ja XGBoost. Random Forest Classifier sai valitud, sest meie hinnangul oli sellel ülesande jaoks sobivaim positiivsete ja valepositiivsete suhe. XGBoost sai valitud, sest sellel oli riskiklasside osas oluliselt kõrgem *recall* ning eesmärk oli teise iteratsiooni käigus sobitada seda nii, et riskiklasside recall jääks ligikaudu samaks, kuid valepositiivsete hulk väheneks.

Teine teratsioon

Teise iteratsiooni käigus vaadeldi mõnevõrra täpsemalt kahe välja valitud mudeliversiooni Random Forest Classifier ja XGBoost ennustusvõimet ning testiti täpsemalt erinevaid mudeli parameetreid ning disainispetsiifilisi aspekte. Kuna disainispetsiifiliste arenduste käik ei ole oluline selle projekti väljundite jaoks, siis sellel kirjeldused me pikemalt ei peatu.

Samadel testandmetel saavutati järgmised tulemused:

Meetod	**0-klassi F1**	**1-klassi F1**	**2-klassi F1**
RFClassifier	0.97817749	0.25757522	0.08591375
XGBoost	0.95261797	0.19381509	0.08309609

RFClassifier	**Pred 0**	**Pred 1**	**Pred 2**

True 0	168429	4412	2117
True 1	1071	794	209
True 2	792	159	154

XGBoost	**Pred 0**	**Pred 1**	**Pred 2**
True 0	159433	10269	5256
True 1	279	1404	393
True 2	56	739	292

Eeltoodud tulemuste põhjal valiti välja sobivaim mudel tõestusprojekti lõpptulemuste tootmiseks.

Lisaks objektiivsetele F1 meetrikutele oli üks parameeter, mida vaadati ka valenegatiivsete optimaalne arv. Kuna eeldatavasti ei ole arvestatav osa liiklusõnnetustest ennustatav meile saadaolevate parameetrite pealt, võib liiga madalat valenegatiivsete arvu tõlgendada ka kui mudeli ülesobitamist mingitele kindlatele parameetritele, mis tähendab, et näiteks simulatsioonifunktsionaalsuses ei oleks see mudel väga hästi kasutatav. Seda kaalutlust arvesse võttes on meie hinnangul Random Foresti tulemused paremad kui XGBoosti tulemused.

Kuna meie hinnangul oli lisaks valenegatiivsete hulgale ka endiselt Random Forest Classifieri valepositiivsete hulk optimaalsem kui XGBoostil, valiti parima mudelina välja Random Forest Classifer. Mudeli scikit-learni parameetriteks (nende parameetrite abil defineeritakse mudeli täpsem käitumine, täpsemad selgitused on leitavad https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html): *RandomForestClassifier(max_depth=10, random_state=0, verbose=1, n_jobs=10, n_estimators=100, class_weight="balanced")*

Lõplikud tulemused

Esiteks on oluline teha neli tähelepanekut:

1. Lõplike tulemuste analüüs genereeriti teise treening- ja testhulgaga kui eelnevad analüüsid mistõttu võib ka tulemuste jaotus olla esialgsest analüüsist mõnevõrra erinev. Samas esialgsete tulemuste analüüs viitas sellele, et antud lähenemine on ka teistsuguse andmehulga puhul sobiv. Treening- ja testandmestik on koostatud samal põhimõttel, mis esialgsete testide puhul, kuid seekord on kaasatud kõik andmed. Treening andmestik on genereeritud 80% andmete põhjal ning test andmestik ülejäänud 20% andmete põhjal.
2. Lõplike tulemuste uurimisel on segadusmaatriksil ja F1 skooril pigem informatiivne roll. Nende meetrikute toetuseks on siin välja toodud ka mõningate parameetrite täpsemad vaatlused. Peamine ennustusvõimekuse analüüs nõuab siiski

ekspertteadmisi ning selleks saab kõige sobivamalt kasutada loodud prototüüpi nii genereeritud ennustuste kui ka simulatsiooni läbiviimiseks.

3. Lõplike tulemuste genereerimise käigus läbisid mudelid mitu iteratsiooni valideerimisi ekspertidega.
4. Otseselt ei saa öelda, et mudel klassifitseerib midagi õnnetuseks. Pigem tuleks sellest mõelda kui, et mudel õpib lõikude riskantsust väljendama seni toimunud õnnetuste pealt. See tähendab, et mudeli meelest riskantsematel lõikudel võiks toimuda rohkem õnnetusi, kuid mudel "õnnetuseks" kui selliseks ei klassifitseeri, pigem oleks õige mõelda selle kohta "riskantseks". Lõigu riskantsus on antud juhul põhimõtteliselt kunstlik meetrik, kus riskantsematel lõikudel on suurem risk õnnetusteks, kuid riskantsuse väärtus ei ole 1-1 ülekantav treenigandmete õnnetusega. Seega kui mudel ennustab prob_1=0.88, siis see ei tähenda õnnetust, vaid seda, et risk õnnetuseks on kõrge. 1-1 õnnetusteks võiks need riskid üle kanda siis, kui treeningandmetes poleks õnnetuste juhud üleesindatud

   Nüüd kui tahta need riskid tõlgendada õnnetuste toimumiseks võttes aluseks mudeli 10-minuti kaupa ennustused, siis meie arvutuste kohaselt peaks kokkuvõttes toimuma ca 700 000 korda vähem õnnetusi, kui mudel seda ennustab. See erinevus tuleb sellest, et treeningandmestikus on õnnentustega seotud andmete osakaal 100 000 korda suurem kui üldandmestikus. See on omakorda võimendatud veel hinnanguliselt 7 korda märgenduste kaalumisega treenimiseprotsessis. See erinevus, mis tuleb tegelikult võrreldes selle 700 000-ga, tuleneb juba sellest, mida mudel treenimise käigus õppis. Siinkohal on ennustuseks loetud siis argmax(prob_0, prob_1, prob_2).

   Kuna prototüübis vaatleme lõigu kaupa valitud keskmiseid, siis keskmiste pealt argmax võttes tuleb hinnanguline risk ning seeläbi ka eeldatav õnnetuste arv mõnevõrra erinev (üldiselt on klassid 1 ja 2 vähemesindatud kui 10-minuti kaupa eraldi vaadates). Sama kehtib ka väljaspool prototüüpi erinevate agregeeringute kohta. Keskmiste riskide tõlgendamiseks õnnetustesse on vajalik veidi teine lähenemine. Tõenäoliselt kõige parem kompromisslahendus lihtsuse ja adekvaatsuse vahel on järgnev. Kuna keskmine kujuneb nii kõrgematest kui madalamatest väärtustest, tuleks keskmisi vaadeldes võtta hinnanguliseks õnnetuse toimumise väärtuseks mitte argmax vaid 1.0 nii prob_1 kui prob_2 puhul. See tähendab, et kui näiteks ühe päeva (24tundi, 24*6=144 vaatlust) keskmiseks riskihinnanguks on 0.74, siis selle õnnetusteks tõlgendamisel oleks see (0.74/1.0)*144/700000 klass 1 õnnetust hinnanguliselt vaadeldud perioodil, kus siis 0.74 on keskmine riskihinnang prob_1, 1.0 on siis see tinglik õnnetuse piir, 144 on vaatluste arv perioodis ning 700000 on see kordaja, mille jagu vähem juhtub eeldatavalt tegelikult õnnetusi kujutatud riskihinnangutest. Tulevikus oleks kindlasti kasulik eeldatavasse õnnetuste arvu leidmisesse kaasata ka lõigu liiklustihedus. Riskihinnangud ise on endiselt keskmise puhul järejestatavad.

Segadusmaatriks ja F1 skoor

Segadusmaatriksist ja F1 skoorist näeme, et võrreldes testhulgaga on küll valenegatiivsete osakaal kõrgem, kuid mitte oluliselt. Endiselt on oluline osa 0-klassi näidetest ennustatud

riskantseteks, kuid see on oodatud olukord, sest kõigil riskantsetel lõikudel lihtsalt ei ole (veel) õnnetust juhtunud.

Meetod	**0-klassi F1**	**1-klassi F1**	**2-klassi F1**
RFClassifier	0.98145771	0.29120849	0.07181484

RFClassifier	**Pred 0**	**Pred 1**	**Pred 2**
True 0	858590	14621	3952
True 1	10136	5270	495
True 2	3733	402	332

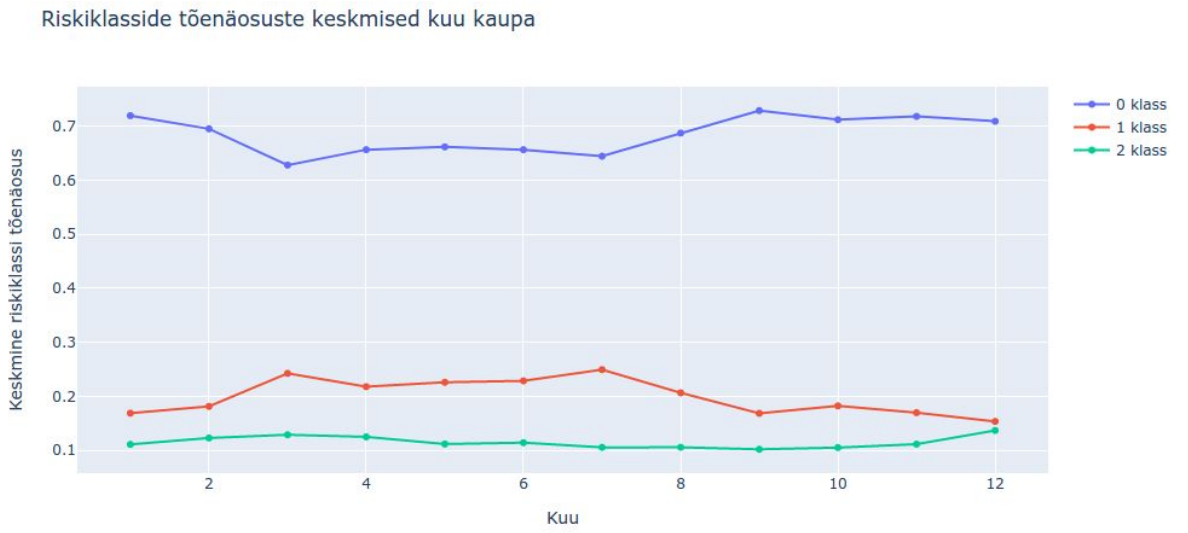
Ennustuste detailanalüüs

Järgnevalt toome välja mõned aspektid riskihinnangute ennustustest. Arusaadavalt ei ole võimalik kõiki ennustuste aspekte välja tuua, seega siinkohal toome välja mõned olulisemad, et oleks võimalik paremini aru saada ennustuste iseloomust ning sellest kuidas oleks mõistlik neid edaspidi valideerida.

Esimesel joonisel on riskiklasside tõenäosuse keskmised kuu kaupa. Siin on mõistlik jälgida kahte aspekti..

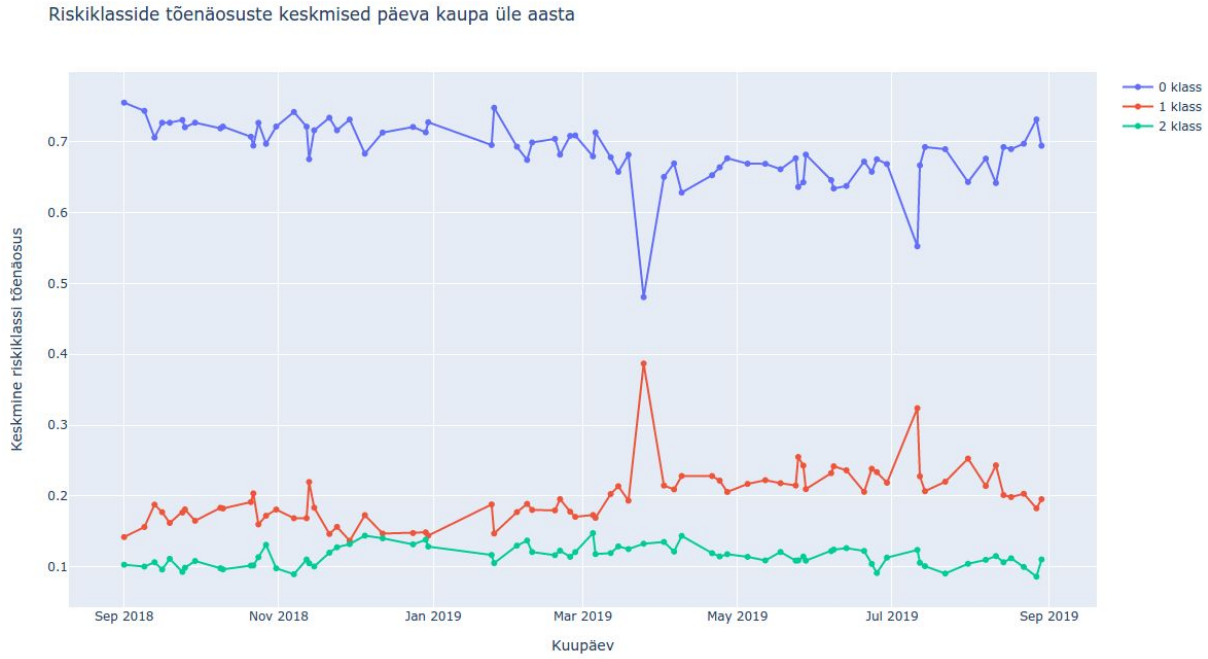
Esiteks on märgata, et alates septembrist kuni veebruarini riskihinnangud õnnetustega seotud klassidele 1 ja 2 langevad oluliselt. See kattub samuti liiklusõnnetuste arvulise jaotusega läbi aasta, seega mudeli ennustused on selles osas loogilised.

Teiseks on märgata, et selgelt kõige tõenäolisem klass on 0-klass, ehk need lõigud, kus õnnetuse risk on väga madal. Samuti on kerge liiklusõnnetuse toimumine (klass 1) oluliselt tõenäolisem kui raske liiklusõnnetuse (klass 2) toimumine. Seega on õnnetuste riskide suhe loogilises järjestuses.

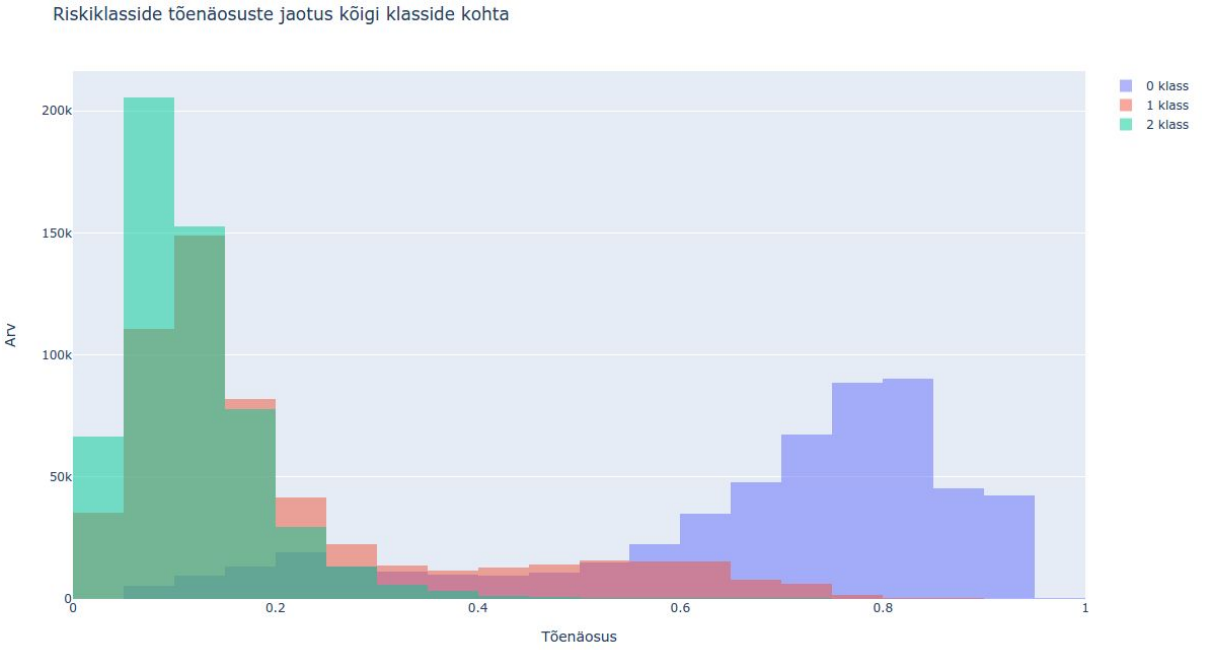


Järgmisel joonisel on jälgitud riskihinnangute liikumist päeva tundide lõikes. Siin näeme, et liiklusõnnetuste risk on päeval oluliselt kõrgem kui öösel. Samuti on näha, et hommikutundidel liiklusõnnetuste risk tõuseb järsult samaks kui õhtul see langeb küllaltki laugelt. Nagu ka ekspertidega vestlustest välja tuli, on selline käitumine ootuspärane, sest esiteks, päeval sõidabki rohkem autosid ja risk on kõrgem. Teiseks, hommikul on liiklusaktiivsuse suurenemine järsem kui õhtune liiklusaktiivsuse vähenemine. Seetõttu on ka selline riskide muutumine ootuspärane.

Alloleval joonisel on jälgitud riskide keskmise muutumist päevade kaupa aasta lõikes. Päevade järjestusse on tekkinud lüngad, sest andmestik jaotati treening-ja testhulkades kuupeävade alusel, et vältida õnnetuste sildi kohalt andmelekke teket. Siin näeme põhimõtteliselt samu liikumisi kui kuu lõikes keskmistelt.

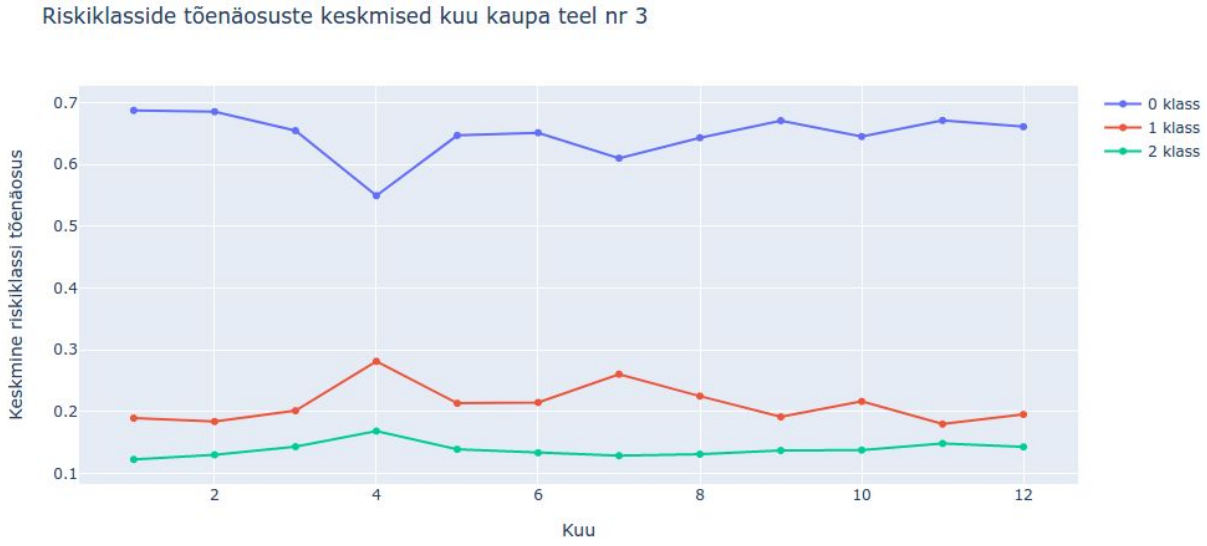


Selleks, et saada paremat aimdust riskihinnangute jaotusest, vaatasime ka iga klassi riskiklasside tõenäosuste jaotusdiagrammi. Sellelt jaotusdiagrammilt on näha, et üldiselt on teed küllaltki selgelt madala õnnetusriskiga. Samas on hea omadus ennustustel, et see tõenäosuste jaotus ei ole täielikult äärmustesse jaotatud vaid on tekkinud küllaltki sujuv lohk arvestades andmestikku. See tähendab, et neid ennustusi saab tõenäoliselt kasutada hästi teede riskantsuse põhjal järjestamiseks ning võimalik, et ka seeläbi tegevuste planeerimiseks.

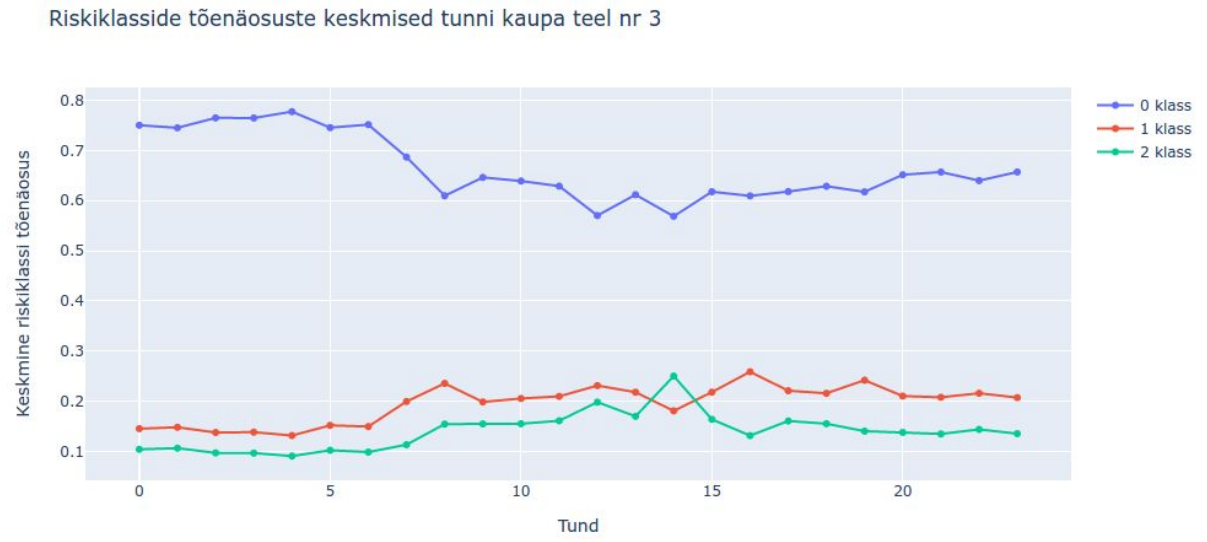


Sama statistikat sai uuritud ka ühel konkreetsel teel. Valitud sai tee number 3, sest Maanteeametil on ühes teises projektis see tee lähema uurimise all ning seega tekib selle tee puhul hea võrdlusmoment.

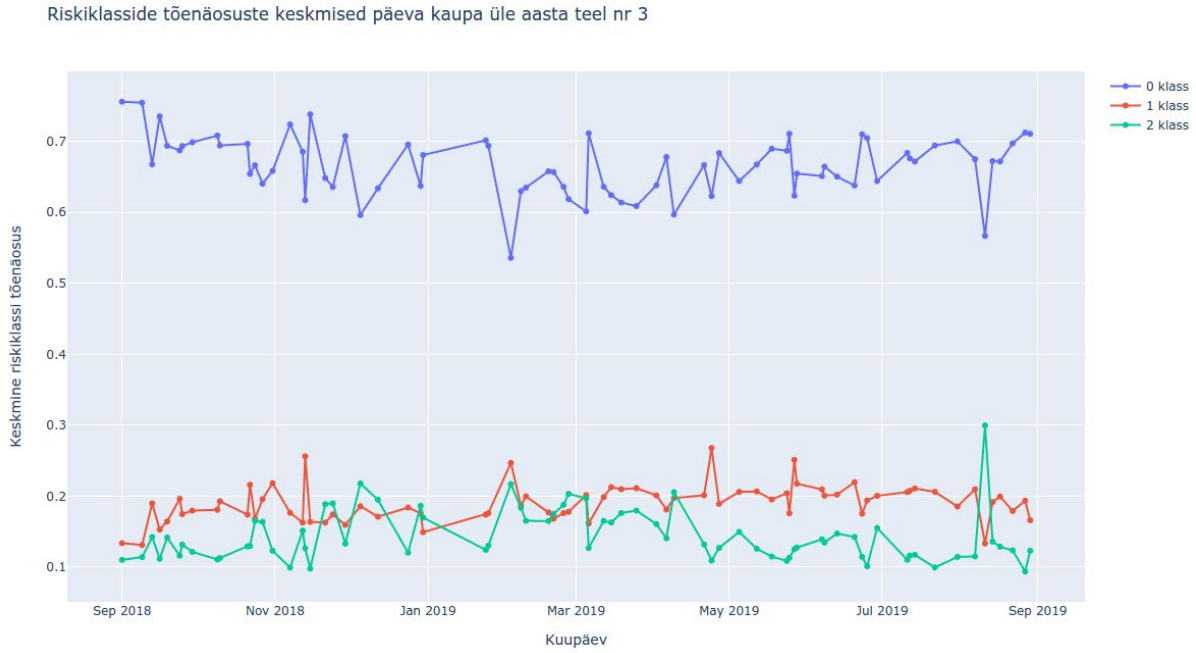
Üldjoones on tee nr 3 riskihinnangud sarnase profiiliga nagu üldine riskihinnang. Välja võiks tuua piigid aprilli ja juuli koha peal, kuid üldiselt suuri erinevusi ei ole kuude lõikes.



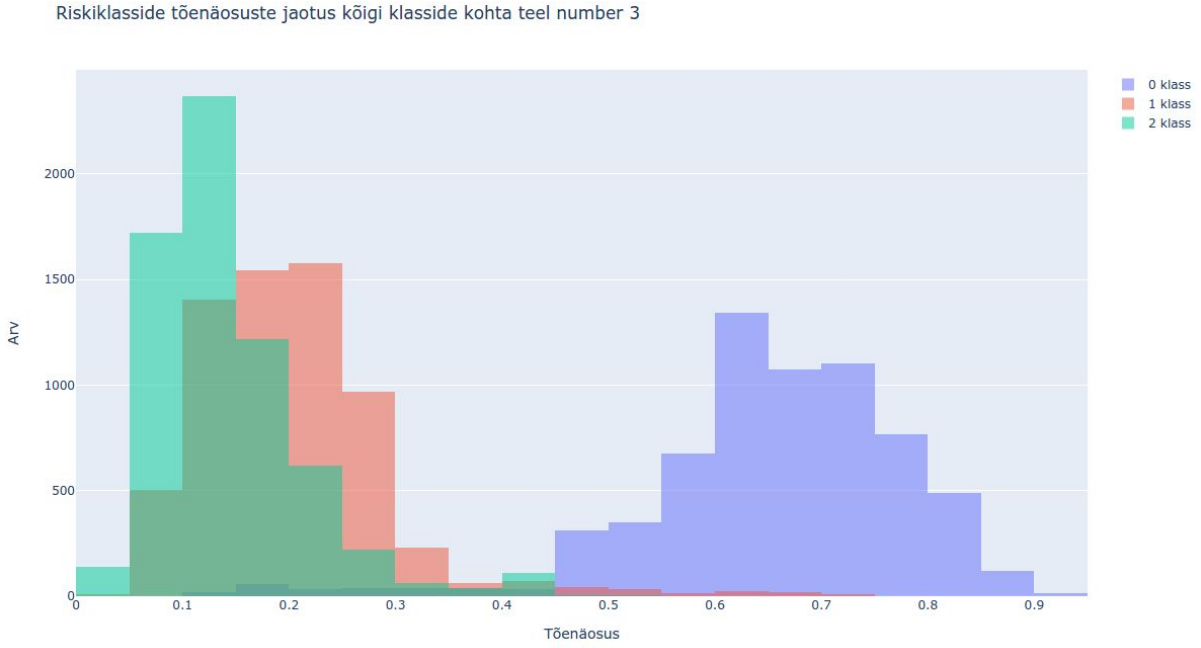
Samuti nagu kuupõhisel graafikul, ei ole ka tundidepõhisel graafikul suuri erinevusi üldisest profiilist. Küll aga võiks välja tuua, et 2. Klassi (raske õnnetus) tõenäosus on tee nr 3 lõikes 1. Klassi (kerge õnnetus) tõenäosusele lähemal kui üldises profiilis, mis võiks viidata, et raskete õnnetuste risk on teel nr 3 veidi kõrgem kui üldisel profiilil. Samas on klassi 1 tõenäosused ka veidi madalamad kui üldisel profiilil, mis võiks viidata, et tervikuna on tee nr 3 risk ligikaudu sama või isegi veidi madalam kui üldine profiil.



Päevapõhisel graafikul on üldiselt kõik oodatult paigas, kuid tekkinud on mõned huvitavad piigid Aprilli ning Augusti perioodi. Neid piike tasuks täpsemalt uurida ekspertidel, võimalik et sealt tuleb välja mõni mudeli artefakt või hoopis kasulik seos.

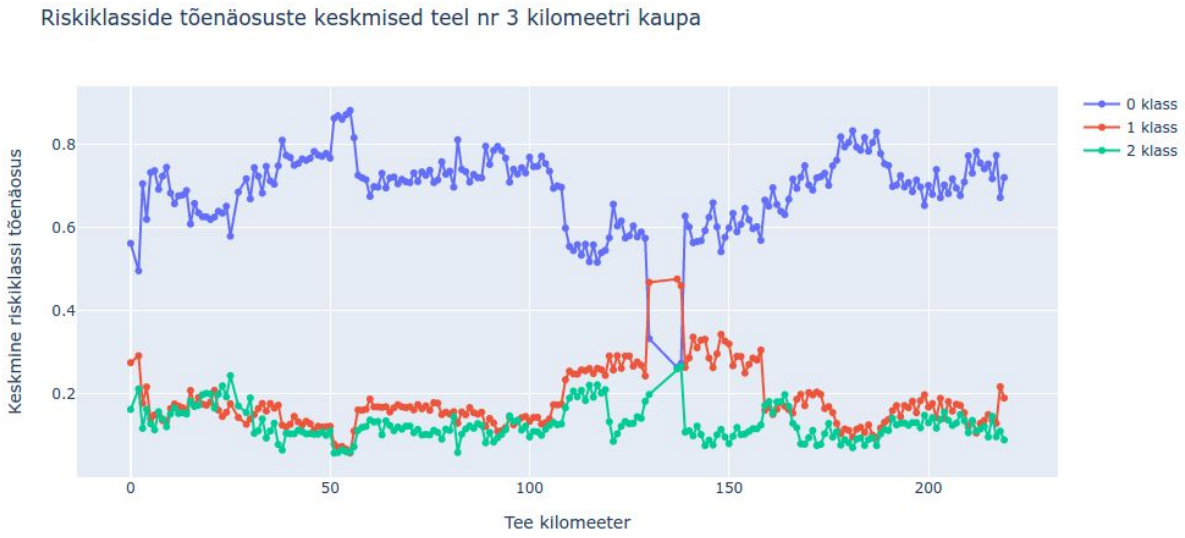


Samuti vaatasime ka riskiklasside tõenäosuste jaotust teel number 3. Nagu näha, siis on siin puhul tõenäosuste jaotus mõnevõrra rohkem eraldatud, mis viitab, et tervikuna on tee number 3 keskmisest ohutum. Samas riskiklasside jaotus on nihkunud mõnevõrra paremale, mis viitab sellele, et teel number 3 on mõõduka riskiga lõike keskmisest rohkem..



Selleks, et täpsemalt näiteks tee number 3 ohtlike lõike leida tuleks juba detailsemalt minna teelõikude tasemele. All asuval ülevaatlikul joonisel tee number 3 kilomeetrite keskmise riskihinnangu osas on näha, et tee nr 3 kilomeetrid (üks tee kilomeeter sisaldab mitut lõiku)

on küllaltki erineva riskiastmega. Tee osade eraldatud riskihinnangu osas on hea, sest see võimaldab ennustuse paikapidavuse korral paremini teelõikude ennetustegevusi prioritiseerida.



Praegu välja toodud tasemel analüüs võib isegi piisav olla ohtlikemate teelõikude tuvastamisel ja järjestamisel, kuid kindlasti on olulisi statistikuid, mida eksperdid oskavad paremini välja tuua kui meie. Samuti eeldab mudeli valideerimine ekspertidelt detailsemat tööd loodud prototüübiga.

Valideerimistagasiside ekspertidelt

Järgnevalt toome välja mõned ekspertvalideerimise käigus tehtud tähelepanekud. Ilmselgelt kogu valideerimisprotsessi dokumenteeritult esitada ei ole võimalik, seega välja on toodud ainult mõned tähelepanekud. Maanteeameti ja PPA ekspertide poolt tehtud mõned tähtsamad tähelepanekud riskihinnangute osas olid järgmised (meiepoolne kommentaar kaldkirjas):

- Tundlikkus - Mudel leidis pooled klass 1 õnnetused ja 12% klass 2 õnnetustest. Tundub, et mudel ei ole hea klass 2 tuvastamises, aga tasub tähele panna, et paljud tegelikult klass 2 tulemused on mudeli poolt liigitatud klass 1 õnnetuseks, mitte õnnetuse mitte-toimumiseks. - *Siin tasub tähele panna, et kuigi klass 2 ennustatakse kõige tõenäolisemana küllaltki harva, siis klass 2 riski tõenäosused on endiselt järjestatavad, mis tähendab, et suurema klass 2 tõenäosuse väärtusega näited on ka kõrgema raske liiklusõnnetuse riskiga.*

Detailsemalt vaadati riskihinnanguid teedel number 3 ja number 15. Nende ennustuste kohta anti selline hinnang (esitatud toimetatud versioon hinnangust):

- Positiivne:
    - Linnulennult vaadates joonistuvad riskiskoori väärtuses ilusti välja tiheasustatud alad: nt teel nr 3 Jõhvi, Lemmatsi, Külitse ja Elva.

- Prob 1 ja Prob 2 väärtuste suund on ühesugune, kui lõigul ajaga kasvab raske LÕ toimumise tõenäosus, siis ka kergema LÕ toimumisrisk tõuseb.
- Tee nr 3 riskiskoor on kevad – suvi perioodil kõrgem. Peipsi piirkond on eriti väga hooajalise liiklusega. Nt lõigud 8074-8084 (piirkiirus on 70km/h), suvel on ennustatud risk suurem, mis on loogiline, kuna suvel ilusate ilmadega liikumiste arv kasvab.
- Üks suurema Prob 2 väärtusega ala teel nr 15 (lõigud 23122;23124;23127;23128, km 26,096 - 26,38). Tegemist on ristmiku ala-ga, kus on kehtestatud madalam lubatud piirkiirus 70 km/h.
- Üldiselt on teel 15 Prob 2 väärtused väiksemad kui teel nr 3.

- Vajab kaalumist:
  - Peab kaaluma ühte klastrisse sattunud väga lühikeste lõikude ühendamist. Nt lõigud 8047 - 8061, peab vaatama, miks nad on erinevateks lõikudeks jaotatud. Klaster on neil sama. - *Siinkohal tuleb läbi viia ka edasistes tegeuvstes välja toodud sarnaste teelõikude ühendamine*
  - Kuna ristmikel on potentsiaalsete konfliktkohtade arv reeglina suurem, kui tavalisel teelõigul, siis äkki oleks mõistlik käsitleda neid teelõikudest eraldi. Tänases mudelis on lõigud ja ristmikud koos. See on vaid kaalumise koht, võib-olla mõtlen liiga kitsalt ja üritan liiga palju paralleele otside tänapäevase praktikaga. - *Ristmike eraldi käsitlemine on välja toodud edasiste tegevuste all.*
  - Jõhvi linnas olevad lõigud on kahes klassis 79 ja 117, jaotus kaheks võib olla loogiline, kuid nt lõigud 8064 ja 8066 oma olemuselt võiksid pigem ühes klastris olla. Või 8066 jääbki 79 klastris (riigitee lõik, teised hooldusnõuded jms), kuid siis jääb arusaamatuks, mille poolest erinevad lõigud 8064 ja 8072-8073. - *Siinkohal tuleb läbi viia ka edasistes tegeuvstes välja toodud sarnaste teelõikude ühendamine*

- Vajab kontrollimist:
  - Mõnedel ennustustel löövad selgelt välja ekstreemväärtused (ui ennustuse väärtustest on vaid üks eriti kõrge) nt lõik 8049 10/10/2019 või 23127, 10/05/2019, kl 22:30. Peaks vaatama koos juurpõhjustega.
  - Tundub, et ennustuse väärtused korreleeruvad lõigu pikkusega, pikematel lõikudel on suuremad väärtused.
  - 'on_riigitee' tundub, et tunnus ei tööta korralikult, andmetes on 0 seal, kus on ja ei ole riigiteed. - *probleem üleantavas versioonis lahendatud*
  - 'asulas' on ka probleemne, kuid see võib olla teeregistri andmekvaliteedi teema. - *tuleneb tõepoolest teeregistrist, seega ka üleantavas versioonis võib see viga olla*

- Teel nr 3 MAX Prob 1 on lõigul 8064 01/03/2019 ennustusel - tegemist on linnasisese lõiguga, mis ei ole riigitee. Kas oluliste sisendandmete puudumine ei vii mudelit segadusse. - *Puudulike andmetega lõikude riskide ennustamine on tõepoolest ebatäpsem. Seetõttu võeti mitte riigiteed täielikult analüüsist välja.*
- Teel nr 3 lõigul 8697 (kurv enne Kauksi risti, km 46,838 - 46,854, päris lai ala) 05/04/2019 Prob 2 riskiskoor on märgatavalt suurem kui teistel aegadel või naaberlõikudel. Küll Prob 2 natuke kasvab kevad - suve perioodil ka naaberlõigul 8695, kuid see 2m lõik füüsiliselt eriti ei erinete naaberlõikudest, aga Prob 2 on kordades suurem (0,001 ja 0,09). Sama teema ka teel nr 15 lõigul 23313 (Hagudi, tiheasustatud piirkond). - *Tuleb läbi viia detailsem analüüs ekstreemumlõikude puhul. Siinkohal võiks kindlasti abiks olla ka simulatsiooni funktsionaalsus.*

Juurpõhjuste ennustuste tulemused

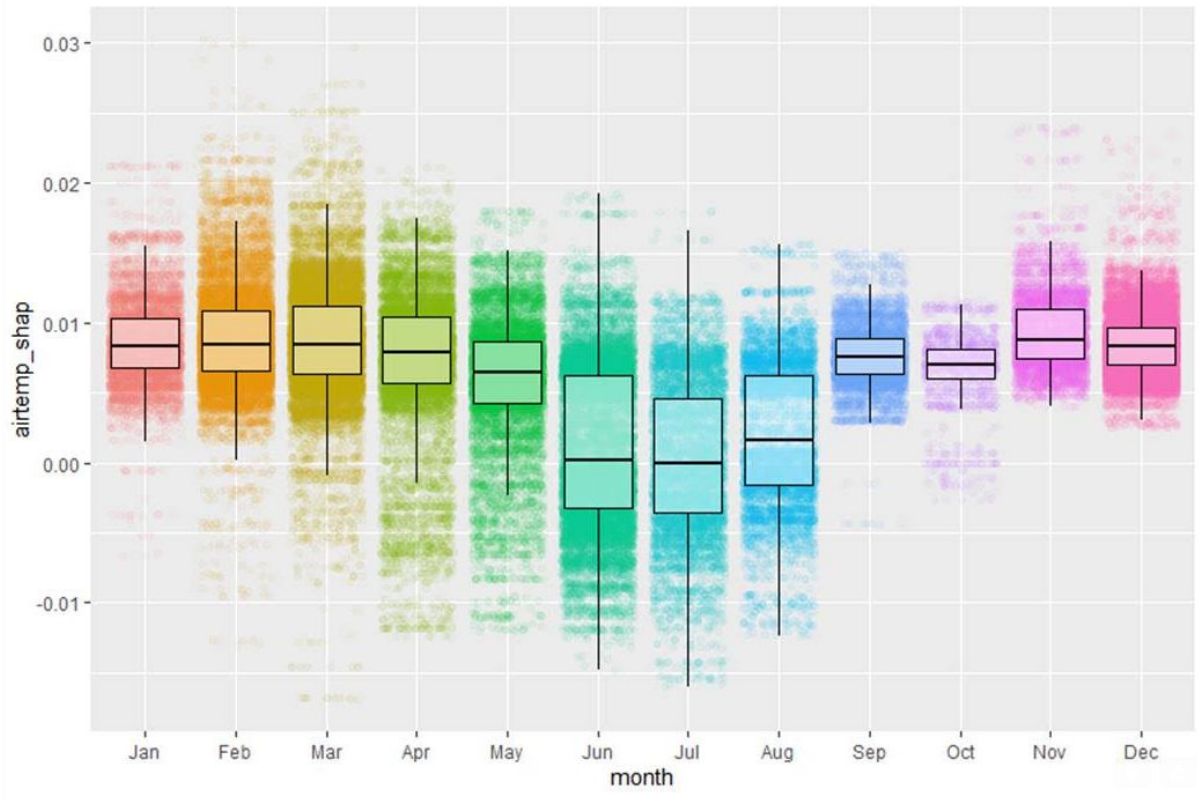
Ennustatud riskihinnangute põhjal parimale masinõppemudeli tulemustele rakendati ka juurpõhjuse analüüsi. Juurpõhjuste analüüsi eripäraks on see, et meil ei ole andmestikus enamikel juhtudel otseselt silti küljes, mis on erinevate õnnetuste juurpõhjused. On küll ülestähendused erinevatest õnnetuse ajal olnud tingimustest, kuid kuna see ei lähtu otseselt meile saadaolevatest andmetest, siis juurpõhjuse mudeli sobitamiseks sellest kasu kahjuks ei olnud (nt on keeruline siduda õnnetuse kirjeldust sellega, et tee laius ja rahvastikutihedus piirkonnas on olulised tegurid õnnetuse riski tõstmisel).

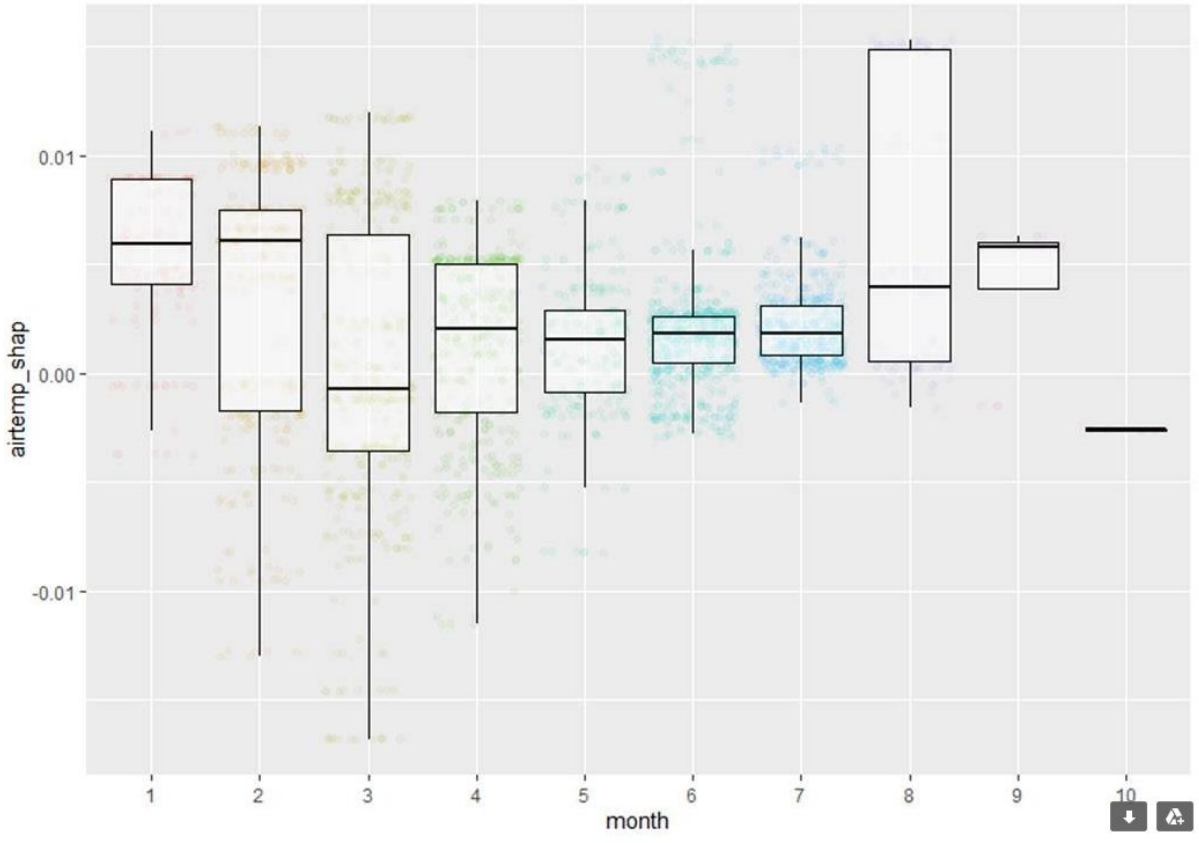
Seetõttu rakendati juurpõhjuse tulemuste valideerimisel kahte protsessi: uuriti statistilisi parameetreid ning samuti uurisid eraldiseisvalt juurpõhjuse parameetreid Maanteeamet ja PPA eksperdid.

Maanteeameti ja PPA juurpõhjuste analüüs

Maanteeameti ja PPA ekspertide poolt tehtud tähtsamad tähelepanekud olid järgmised (meiepoolne kommentaar kaldkirjas):

- Õhu temperatuuri mõju on kuude kaupa erinev. Suvel nii põhjustab kui maandab riski võrdsel määral juhtudest (mediaan on 0 lähedane) ja muidu pigem riski põhjustav tegur. Toodi välja karpdiagramm vastavalt nii kõikide juhtude kohta kui ka ainult nende kohta, millel oli reaalselt ka õnnetus toimunud. (Jooniste autoriks on Ülo Leppik)- *Arutelude käigus jõuti arvamusele, et selline temperatuuride juurpõhjuste jaotuse muutumine ajas on kindlasti loogiline kõigi näidete puhul. Õnnetustega seotud näidete puhul signaal mõnevõrra ebalegem. Eraldi tähelepanekut vääris tõsiasi, et märtsikuu juurpõhjuste väärtuste hajuvus on kõige suurem. See on indikatsioon sellele, et juurpõhjuse hinnang liigub küllaltki adekvaatselt, sest ilmateenistuse andmetel oli 2019 aasta märts temperatuuridelt väga kõikuv. Samuti tuleb tähele panna, et õnnetustega seotud juurpõhjuste näited puuduvad alates oktoobrist 2019. See tuleneb, sellest, et õnnetuste andmestik lõppes septembriga 2019.*





- snow_thick, icetemp1, blackicethic2,snow, ice_thick ja võibolla mõned tunnused veel, on katki veidral moel. Nad sisaldavad kõik väärtuseid, mis on ebaloogilised ja/või vastuolus tunnuste kirjeldusega. Nt icetemp1 on väidetavalt jää temperatuur

mõõdetuna kahe komakoha täpsusega kraadides, tegelikult on see täisarvudes tunnus, mille tavaline väärtus on +20 (päris soe jää kohta!). - *See tähelepanek on välja toodud, et taaskord mainida, et puuduvad tunnused on tähistatud ebaloogilise väärtusega, mida andmestikus muidu ei esineks. Enamikel väärtustel on selleks -1, kui näiteks jää temperatuuri puhul -1 oleks täiesti loogiline väärtus ning seega on konkreetsel juhul puuduv väärtus tähistatud 20-ga.*

- Iga riskihinnang tekib 101 tunnuse koostoimel, sellest on raske luua intuitiivselt haaratavat ülevaadet. Prooviti sellist lähenemist: Filtreeriti välja read, mille prob_1 oli suurem kui 0,75, ehk nende juhtude puhul on riskihinnang kõrge – mudel liigitab need hästi kindlalt klass 1 õnnetuseks. Siis vaadati millised tunnused säärastel juhtudel kõige rohkem panustavad, keskmise juurpõhjuse väärtuse järgi järjestatult on topp 10. Siin peamiselt tot_* tunnused, see on iseenesest hästi loogiline, need tunnused on elanike arv piirkonnas, mida rohkem elanikke, seda rohkem liiklust ja järelikult ka õnnetusi.

Tunnuse_nimi	**Keskmine juurpõhjuse väärtus**
tot_5_9_shap	0.0504
vaartus_dp_shap	0.0429
tot_30_34_shap	0.0377
tot_40_44_shap	0.0319
tot_lt5_shap	0.0304
month_shap	0.0263
tot_45_49_shap	0.0251
vaartus_bp_shap	0.0235
tot_15_19_shap	0.0221
f_total_shap	0.0220

Korrati sama protsessi uuesti ilma nende tunnusteta:

Tunnuse_nimi	Keskmine juurpõhjuse väärtus
vaartus_dp_shap	0.0429
month_shap	0.0263
vaartus_bp_shap	0.0235
iri_shap	0.0168
plaiv_shap	0.0118
plaip_shap	0.0109
tas_moots_xv_shap	0.00667
slai_shap	0.00577
ckpi_shap	0.00567
geom_length_shap	0.00543

See on palju huvitavam tulemus. Nr 1 on päevane rahvaarv, kuu on teine, millele järgneb brutopalk. Järgnevad andmed tee kohta- tasasuse väärtus, teeperve laius nii vasakult kui paremalt, sõiduosa laius, kurvilisus ja viimaks tee lõigu pikkus. (Tasasuse mõõtmissuund, tas_moots_xv, on siin veider, miks peaks mõõtmissuund mõjutama õnnetuste arvu teel?)

Kokkuvõtvalt saame öelda, et mudel andis kõrged klass 1 õnnetuse skoorid selle järgi, kui palju inimesi on piirkonnas ja tee kuju järgi.

- Vastupidi saab ka vaadata, mis olid riski maandavad tegurid nendel samadel kõrge riskihinnanguga ridadel. Tulemuseks on eklektilisem valik. Õhurõhk, teeliik, kampaaniad ja nädala päev. Huvitav on see, et kui riski tõstvad tegurid olid kõik ajast sõltumatud, siis peamised maandavad tegurid on enamuses ajast sõltuvad.

Tunnuse_nimi	Keskmine juurpõhjuse väärtus

airpres_shap	-0.00282
inspire_functionalclass_shap	-0.00255
tee_teeliik_val_shap	-0.00223
active_campaigns_shap	-0.00168
dayofweek_shap	-0.00143
dewp_shap	-0.00119
WD10M_shap	-0.00118
VIS10M_shap	-0.00109
hoolklt_hoolklt_xv_shap	-0.000690
kptp_kptyyp_xv_shap	-0.000673

- *Eelnevad kaks punkti on heaks indikaatoriks sellele, et kuidas võiks konkreetse lõigu juurpõhjuseid hinnata. Eelkõige võrsub siit idee, et juurpõhjused, mis on kõigil teelõikudel sarnaselt mõjutavad, et pruugi olla informatiivsed selles võtmes, et mis just selle konkreetse teelõigu riskantseks teeb. Kui üldiselt olulised põhjused eemaldada, siis võib saada teelõigu spetsiifiliselt oluliselt paremad hinnangud juurpõhjuse osas.*

Statistiline juurpõhjuste analüüs

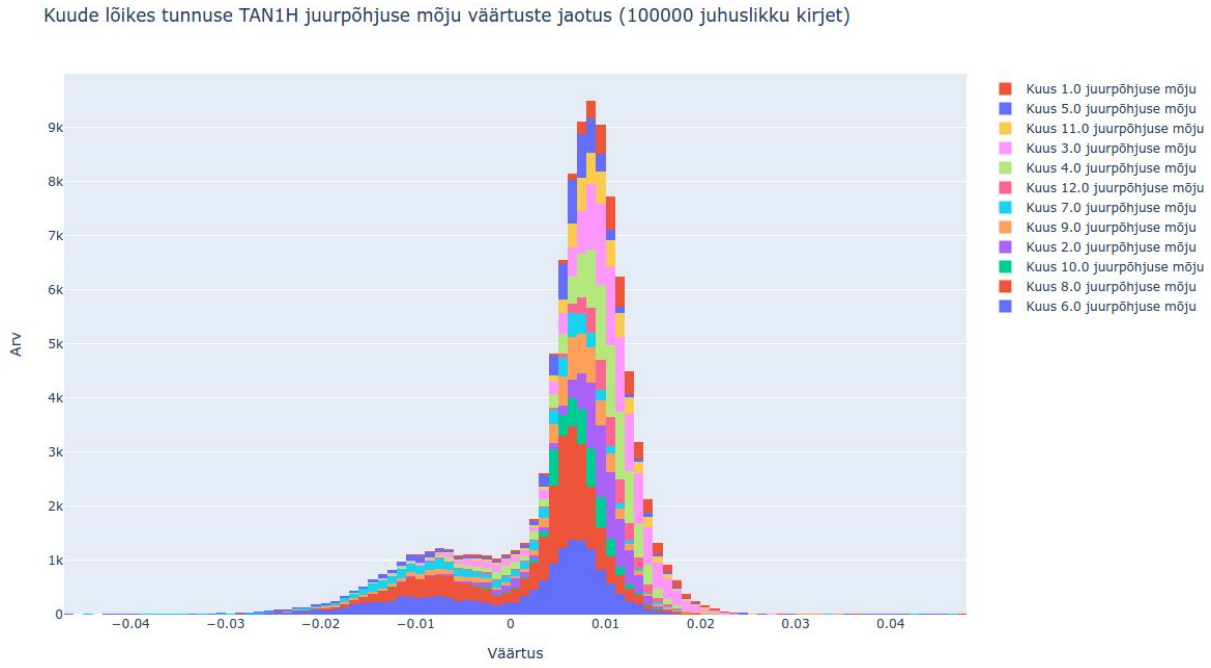
10 kõige olulisemat tunnust mudeli ennustustel olid 'hour', 'vaartus_dp', 'tot_35_39', 'airpres', 'vaartus_bp', 'tot_5_9', 'TAN1H', 'TAX1H', 'WD10M', 'RH'. Üldjoontes on nende tunnuste top 10-s asumine oodatav, kuid näiteks tunnuse "hour" kõige kõrgemal kohal asetsemine viitab sellele, et mudeli parema üldistusvõime ja tegeliku juurpõhjuse parema tuvastamise huvides võiks selle tunnuste nimekirjast välja arvata. Seeläbi võiks ka mudeli simulatsioonivõimekus kindlasti paraneda.

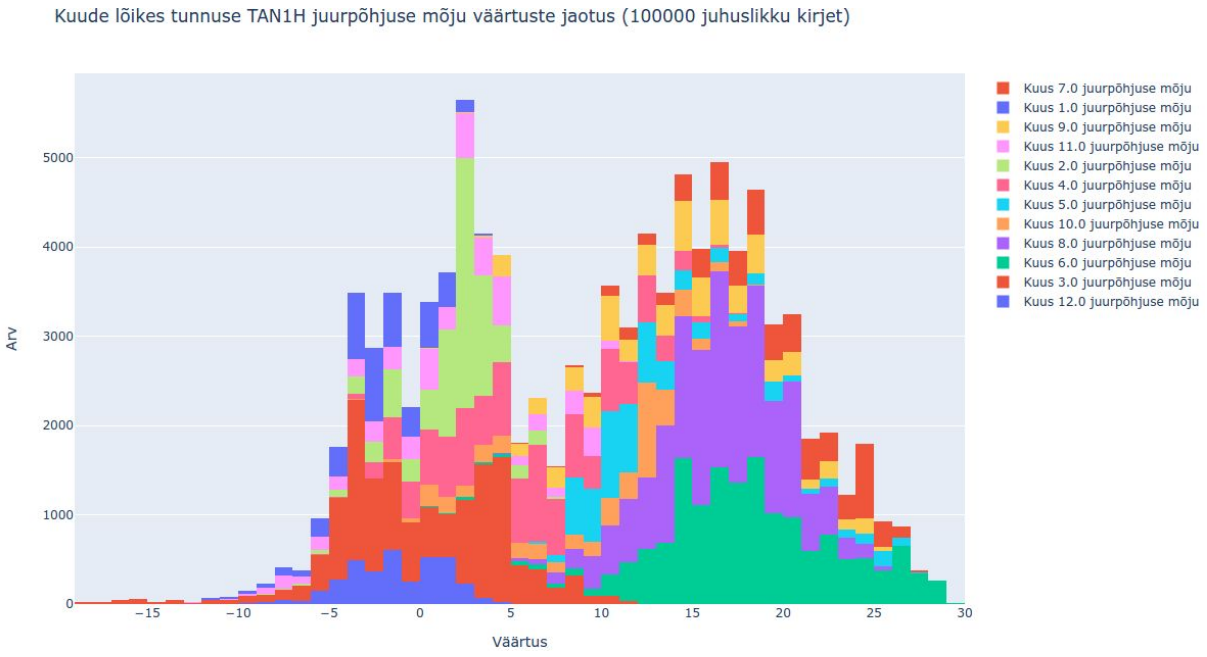
10 kõige vähem olulist tunnust mudeli ennustustel olid 'lvuuk', 'yl_ylliik_xv', 'rparv', 'rdt_raudtee_xv', 'sixth', 'tolm_ttmeetod_xv', 'SolidUsage', 'LiquidUsage', 'precipcl', 'on_katkestus'. Slinkohal on tähelepanuväärne, et kuigi tegemist on tunnustega, mis võiksid liiklusõnnetuste riski oluliselt mõjutada - tee hooldus (SolidUsage, LiquidUsage), raudtee

ülesõit ('rdt_raudtee_xv'), ülekäigud ('yl_ylliik_xv'), on nende olulisus väga madal. Selle põhjuseks on tõsiasi, et kõigi nende tunnuste puhul on andmestikus väga vähe mitte-puuduolevaid väärtusi. Sellest lähtub, et jätkuprojekti üks oluliseid väljakutseid on nende harvade väärtuste parem kaasamine riskihinnangute ennustamisse.

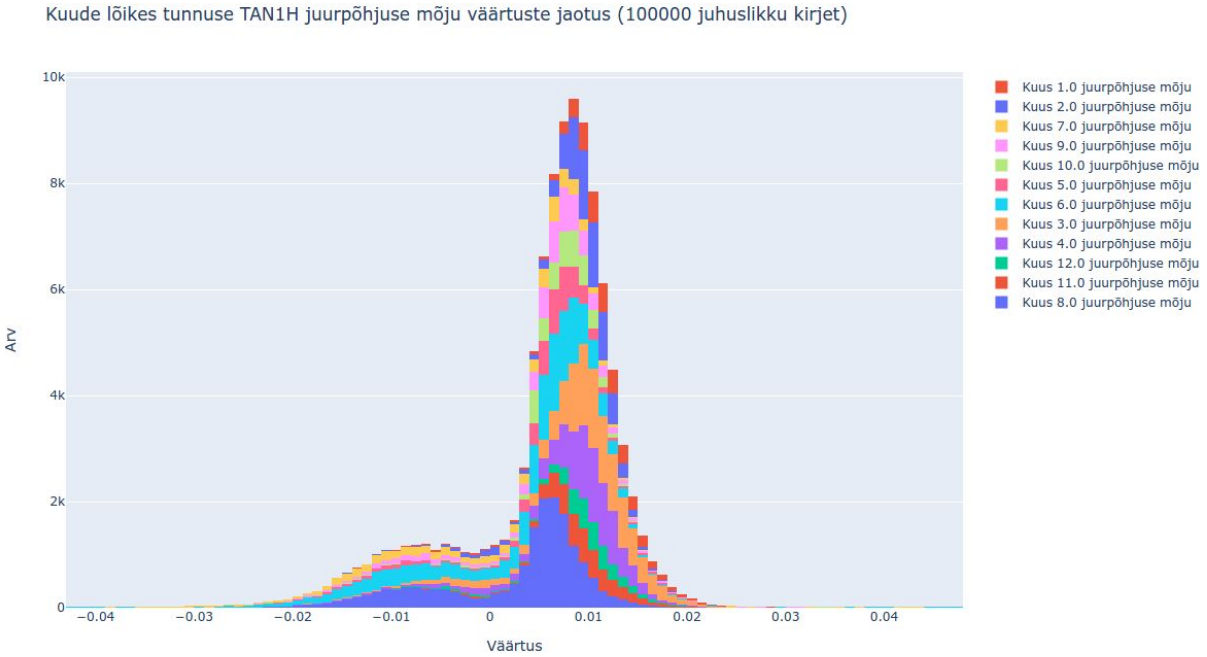
Selleks, et anda natuke tunnetust sellesse, kuidas võiks juurpõhjuse väärtusi uurida lisaks eeltoodud meetoditele, oleme järgnevalt koostanud väikese ülevaate tulemustest.

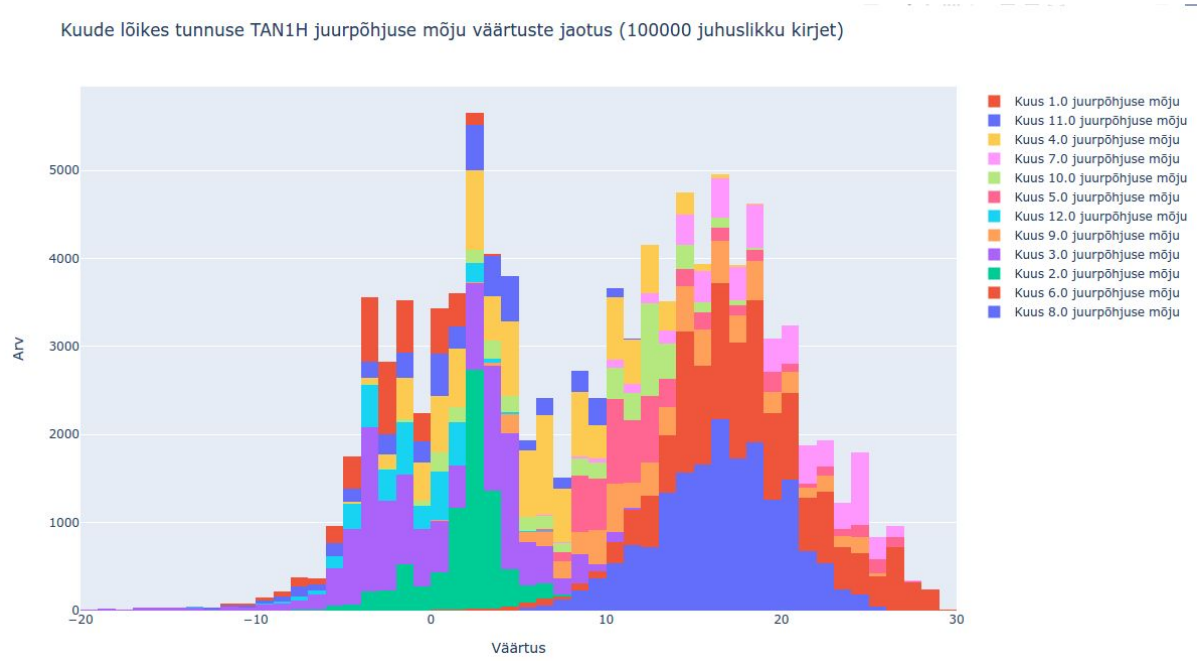
Esiteks vaatasime tunnuse TAN1H (temperatuuri tunni keskmine) juurpõhjuse väärtuste jaotust ja tunnuse enda väärtuste jaotust kuude lõikes summeritult. Kui põhjalikult kahte joonist uurida, siis on näha, et valdavalt õhutemperatuuri väärtused tõstavad õnnetuse riski. Samas on näha ka, et teatud juhtudel need väärtused langetavad riski. Väärtusliku tähelepanekuna tooksime välja, et näiteks kui suvekuude juurpõhjuse väärtusi on nii riski tõstval poolel kui leevendaval poolel, siis talvekuudel on õhutemperatuur oluliselt harvemini riski leevendava mõjuga (vt kuu 1, 2, 4 temperatuure vs juurpõhjuse väärtusi).



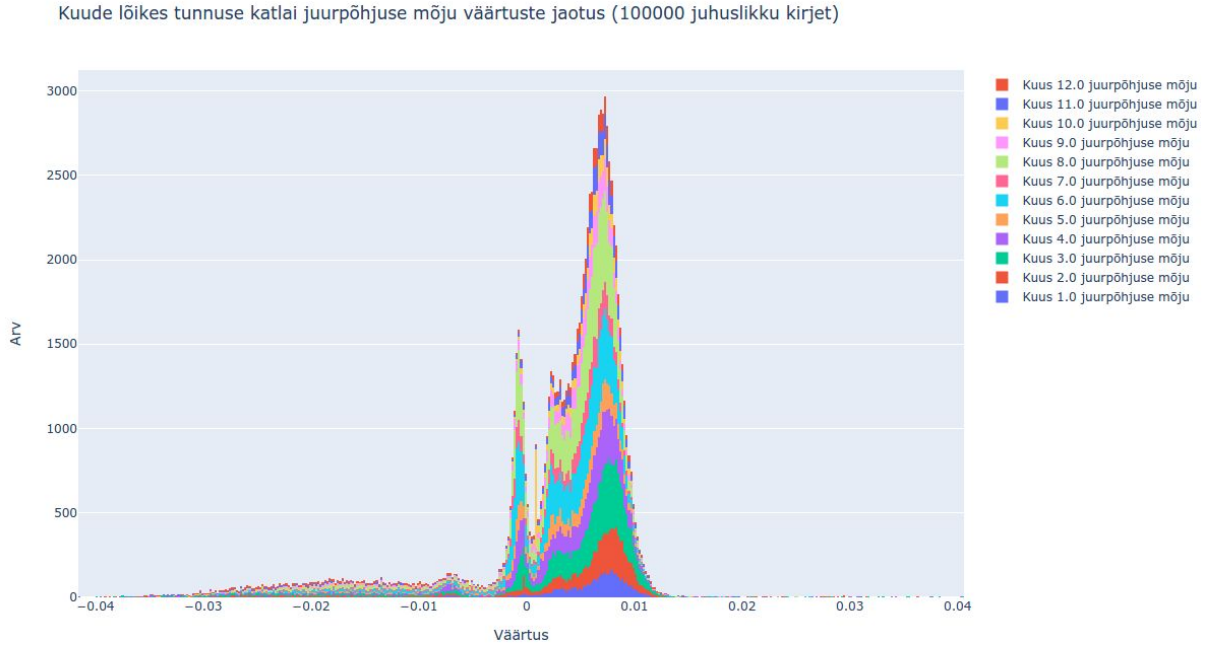


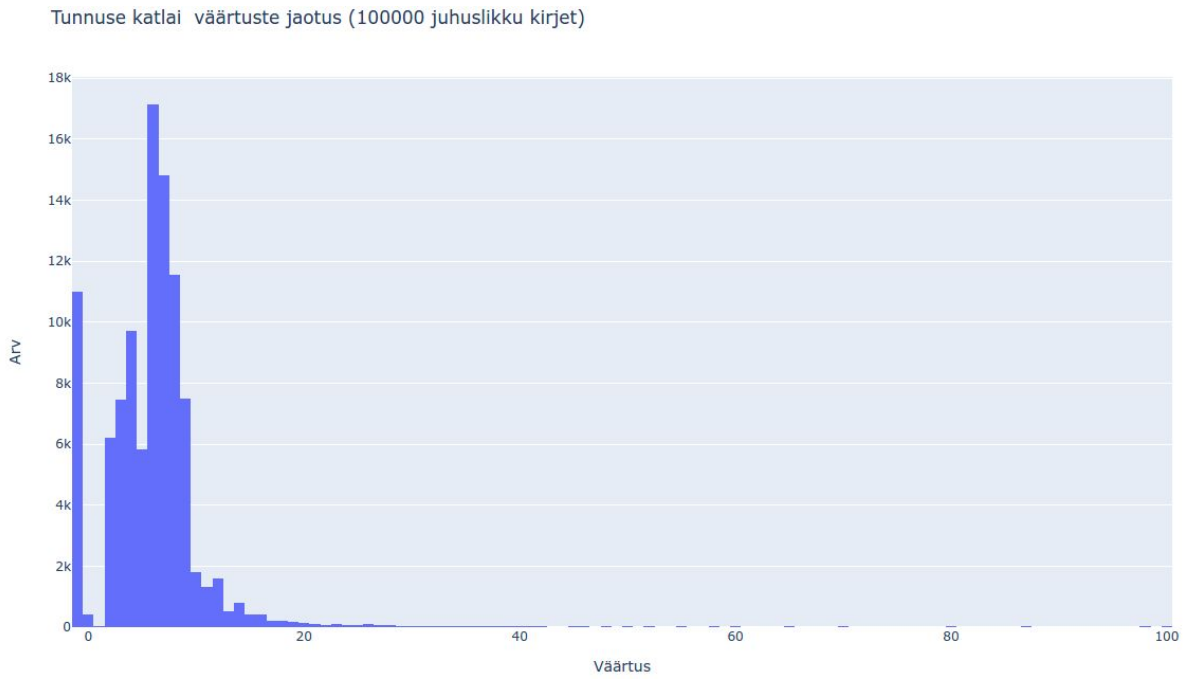
Vaatlesime täpselt samu seoseid ka konkreetselt tee number 3 puhul. Siin on üldine trend jäänud püsima, et talvekuudel on temperatuur leevendava tegurina oluliselt harvem kui suvekuudel. Antud juhul on märgata aga näiteks aprilli kuu osas anomaaliat antud teel, kus aprilli väärtused on leevendava tegurina esindatumad kui üldises valimis. Selliseid teede/teelõikude iseärasusi tulekski meie hinnangul eelkõige otsida, et leida tegurid, mis konkreetse tee ohtlikuks/ohutuks muudavad. Tõenäoliselt tuleb see välja ka mitme tunnuse juurpõhjuse kontekstis, mitte ainult ühe juurpõhjuse vaates.





Lisaks temperatuurile vaatlesime ka ühte tunnust, mis ei ole seotud ajalise muutusega. Selleks võtsime vaatluse alla tunnuse "katlai" (katte laius). Siin näeme, et kuigi katte laius varieerub olulisel määral (alumine joonis), on juurpõhjuse mõju väärtus katte laiust arvesse võttes oluliselt väiksema varieeruvusega kuude lõikes kui temperatuur. See on heaks indikaatoriks, et juurpõhjuse mudel suudab esile tuua tõenäolisema juurpõhjuse (nt kui tee on kitsas ja libedaga juhtub õnnetus, siis on õigem juurpõhjus tõenäoliselt libedus ku katte laius, kuna sama katte laiusega ja hea ilmaga ja muude tingimustega juhtub õnnetusi harva).





Üldine tulemuste analüüs

Üldiselt võib hinnata tulemuste analüüsi põhjal ennustusmudeli ja juurpõhjuse mudeli tööd lubavaks. Võimalik, et riskihinnangute mudeli tulemused on juba praegu tegevuste planeerimisel kasutatavad, kuid juurpõhjuse osas nii positiivse hinnangu andmiseks peaksid tõenäoliselt eksperdid veidi pikemalt saama aega tulemusi uurida.

Tulemuste uurimine on näidanud kätte ka mõned kitsaskohad mida tuleks parandada:

- Mudeli üldistusvõime haruldastel tunnustel vajab parandamist - hetkel näiteks jääkihi paksus on harvaesinev tunnus ning seetõttu selle mõju läheb mõningal määral kaotsi mudelis. Tõenäoliselt tuleks selle tarvis muuta mudeli parameetrite hulka suuremaks, et need juhud paremini ära õppida. Alternatiivne lahendus võiks olla ka selliste andmeridade tegelikust proportsioonist suurem esindatud treeninghulgas.
- Tuleks proovida eemaldada mõned olulisemad tunnused nagu näiteks “hour”, “month” kui tahta suunata mudelit rohkem üldistama ja vähem täppissobitust tegema. Samuti võiks selliste tunnuste kõrvaldamine aidata kaasa tegeliku juurpõhjuse leidmisel
- Juurpõhjuse mudel vajab paremini disainitud lähenemist, et üksiknäited oleks paremini analüüsitavad
- Juurpõhjuse tulemused vajavad ekspertide poolt põhjalikumat analüüsi, et nende sobivuses selgust saada

Positiivsetest külgedest tasub kindlasti välja tuua:

- Riskihinnangute parasjagu eraldatud jaotus, mis lubab teelõike järjestada riskantsuse alusel

- Riskihinnangute ja juurpõhjuste muutumised on loogiliselt põhjendatavad, mis annab kindlust, et tegemist võiks olla kasutatavate lähenemistega

Kokkuvõttes võib öelda, et tulemused kinnitavad, et liiklusõnnetuste prognoosmudeli loomine on võimalik juba olemasolevate andmete põhjal. Täiendavate andmete tekkimisel ning olemasolevate andmete parandamisel võib mudeli töö täpsus kindlasti paraneda.

Edasised tegevused

Selles alapeatükis toome välja projekti käigus kogunenud ideed, mida võiks jätkuprojektis ellu viia tulemuste parandamiseks juhul kui jätkuprojekti kasuks otsustatakse.

Täiendavad andmed

Valdavalt on andmete parandamiseks vajalikud tegevused välja toodud andmete töötlemise peatükkides ja lisades. Siin on välja toodud kas olulisemad neist tähelepanekutest või siis ideed, mis mujal analüüsis välja pole käidud.

1. Ristmikud - praegu ei ole teeinfos eraldi välja toodud ristmikke ning teisi liiklussõlme tüüpe. Tõenäoliselt väga suure ennustusväärtusega oleks ristmike eraldi ühe geograafilise objektina määratlemine. Hetkel on mitmed kohalike teedega ristumised täiesti puudu ning riigiteede liiklussõlmed on jaotatud mitmeteks teelõikudeks, mis muudab nende analüüsimise ühe tervikuna (mis oleks mõistlik) küllaltki keerukaks.
2. Kannatanuteta liiklusõnnetused - hetkel pole liikluskindlustusfondi liiklusõnnetuste info ühendatav meile saadaolevate andmetega, sest seal pole alati võimalik täpset asukohta ja ajahetke määrata. Samuti pole sealt võimalik alati lihtsalt tuvastada kannatanutega liiklusõnnetusi ning liiklusõnnetuse raskusastet. Kui ühendatavuse ja asukoha probleemid osutuvad selles andmestikus või andmete kogumises lahendatavaks, siis oleks ka see andmestik tõenäoliselt mõistlik kasutusele võtta.
3. Raskete vigastatute andmed - hetkel selgus projekti käigus, et liiklusõnnetuste üles märkimisel on raskete vigastatute märkimine ebaühtlane ning tihtipeale on need võrdsustatud lihtsalt vigastatutega. Iseenesest oleks mudeli tööloogika koha pealt mõistlik raskete vigastatute üles märkimine korda teha, sest eeldatavasti annaks see kasulikku infot nii ennustuste tegemisel kui ka ennetustegevuste planeerimisel.
4. Liiklusloenduste imputeerimine - Kuna paljudel teelõikudel puuduvad piisavalt täpsed liiklusloenduste andmed, on hetkel seda andmestikku keeruline kasutada tingituna sellest, et lähima/sarnaseima teelõigu imputeerimine soodustab valepositiivseid/valenegatiivseid tulemusi ning andmete lihtsalt puuduvaks jätmine põhjustab üldistusvõime vähenemist. See imputeerimise versioon oleks tõenäoliselt kõige otstarbekam välja pakkuda maanteeameti ekspertidel.

Sisendandmehulga vähendamine

Üheks takistuseks eelkõige juurpõhjuste arvutamisel on hetkel sisendandmehulga suurus. Hetkel on iga päeva riigiteede täisandmestik keskeltläbi 43 miljonit rida 115 tunnusega. Selline andmete hulk takistab mõnevõrra operatiivset andmete analüüsi. Samas teisest küljest on selline andmete hulk hea andmaks mudelile terviklikuma ülevaate erinevatest

teedega seotud profiilidest. Käesolevas tõestusprojektis ei ole neid andmevähendamise tegevusi rakendatud selleks et vältida potentsiaalset tahtmatut infokadu ning samuti piiratud ajaraami tõttu.

Mõned mõtted andmehulga vähendamiseks oleks:

1. Baseerida mudeli treenimine ainult õnnetustega seotud teelõikudel - see vähendab andmehulkasid olulisel määral ning samas aitab kujundada ohtlike teelõikude "profiilid". Seeläbi uutel ja simuleeritud andmetel ennustusi tehes võiks ohtlikumad teelõigud välja sõeluda ennustuskindluse (nt *prediction probability*) alusel.
2. Vähendada teelõikude arvu analüüsis - iga teelõik genereerib andmetesse päevas 24x6=144 rida. Antud lahenduses on arvutuslikult genereeritud 300 000 teelõiku. Meie hinnangul oleks võimalik ühendada olemasolevad teelõigud nii, et teelõike tekiks ligiakudu kolm korda vähem. See tähendaks ka ühtlasi kolm korda väiksemat andmestikku, mis masinõppeliselt oleks juba oluliselt kergemini hallatav andmemaht ning lubaks seetõttu ka operatiivsemat analüüsi ning detailsemat juurpõhjuse hinnangut.
3. Loobuda iga ajahetke eraldi andmepunktina väljendamisest ning ühendada sarnased andmeread üheks. Põhimõtteliselt oleks tegemist ligikaudsete duplikaatide vähendamisega. See tegevus võiks anda maksimaalselt 10x väiksemad andmestikud.

Kokkuvõttes võiks tegevused 2. ja 3. anda juba ligikaudu 30 kordse andmehulga vähendamise, mis ei tohiks ühtegi takistust tekitada andmete analüüsimisel ning oleks seega meie hinnangul ka piisavad ning sobivaimad tegevused andmete vähendamise osas.

Masinõppe mudeli täiendused

Selle tõestusprojekti mudelite arhitektuurid ei ole taotluslikult väga keeruliseks aetud. Üheks põhjuseks on piiratud ajaraam, kui teine põhjus on vältida mudeliarhitektuurist tulenevaid hälbeid riskiennustustes.

Kuigi analüüsi käigus arvati välja nii XGBoost kui ka (lihtne) DNN, siis tegelikkuses oleme me arvamusel, et gradient boostingul või ka tehisnärvivõrkudel põhineval mudelil oleks lõpplahenduses suur potentsiaal saavutada häid tulemusi, seega meie soovitus on alustada lõppprojekti arendusi neist lähenemistest.

Samuti tuleb meeles pidada, et siin tõestusprojektis on testitud ainult "puhtaid" mudeleid. Lõpp-projektis annab tõenäoliselt parima tulemuse kombinatsioon mitmest mudelist.

Lisaks mudeliarhitektuuridele tuleks lisatähelepanu pöörata ka andmete eeltöötlusele. Näiteks puuduvate tunnuste tähistamine nõuab hetkel sõltuvalt andmetüübist eritähelepanu ning seda tuleb ka edasises analüüsis pidevalt meeles pidada. Seetõttu võiks üritada leida parema lahenduse, mis eeldab vähem teadmiste kaasas kandmist eeltöötluse etapist. Samuti tuleks mõelda kas on mõistliku rakendada standardiseerimist enne või pärast puuduvate tunnuste täitmist.

Juurpõhjuse mudeli ideed

Praeguses lahenduses on kasutatud üldist juurpõhjuse määramise teeki SHAP, mis küll toimib üldjuhul hästi, kuid lõppprojekti raames väärib juurpõhjuse mudeli arendamine kindlasti olulist osa andmeteaduse arendustes. Võibolla isegi võrdset osa mudeli enda arendustega.

Keerukamaid mudeleid kasutades tasub juurpõhjust määrates vaadata täpsemalt mudelite treenimis- ja ennustusprotsessi, et sealtkaudu statistiliselt igale ennustusele adekvaatsem juurpõhjus määrata.

Samuti tasuks tõsiselt mõelda selle üle, kuidas üksikennustuse juurpõhjuse hinnang oleks võimalikult üheselt tõlgendatav. Hetkel tuli analüüsidest välja, et loogika võimaldaks suure hulga juurpõhjuse ennustusi nii kinnitada kui ka ümber lükata. Seega tuleks mõelda kuidas juurpõhjuse määramise protsessi selgemaks ja usaldusväärsemaks muuta.

Kokkuvõte

Kokkuvõtteks võib öelda, et vähemalt esmase analüüsi tulemusena paistab riskihinnangute mudel ennustavat teelõikude riske piisavalt hästi, et vähemalt andmeteaduse mõttes on projektiga edasi liikuda mõistlik.

Ühtlasi annab paremate tulemuste saavutamine veelgi parema sisendi juurpõhjuse mudelile ning seeläbi kindlasti hõlbustab ka juurpõhjuste valideerimist. Kuigi hetkel on juurpõhjuste hindamine veel veidi hägusam kui ehk on optimaalne, siis üldisest analüüsist nähtub, et tunnuste mõju liigub ajaliselt ja väärtuste pooles loogiliselt (nt temperatuuri juurpõhjus), mis viitab, et ka juurpõhjus riskide kujunemisel on ennustatav ning seetõttu võiks viimistletud mudel olla heaks sisendiks ennetustegevuste ellu viimisel.

Jätkuprojekti puhul tuleks koostöös mõelda, milliseid küsimusi mudelilt küsida tahetakse. Näiteks “eelarve 1M eur, kuidas seda paigutada et maksimeerida päästetud elusid, arvestades kriteeriume x, y, z?”. Teisisõnu, riskihinnangute ja ennustusmudelite peale saab ehitada eraldi analüütika mootori, mis nende mudelite hinnanguid maksimaalselt hästi tõlgendada oskab. Loomulikult jääb siin ruumi ka inimesest analüütikule tegeleda tulemuste tõlgendamisega nii riski ja prognoosi mudelitest, kui ka lõpp-mootorist.

Jätkuprojekti puhul tuleb meeles pidada, et projekt ei koosne vaid masinõppelisest poolest vaid tähendab ka seda, et erinevate sisendandmestike formaadid ning salvestamised peab samuti korda tegema. Suurimad andmeteaduslikud väljakutsed on tõenäoliselt juurpõhjuse mudeli viimistlemine ning masinõppemudeli simulatsioonifunktsionaalsuseks eraldi kohendamine.

Projekt on kaasrahastatud Euroopa Liidu Euroopa Regionaalarengufondist. Projekti edusse panustasid hästi ja operatiivselt Maanteeameti ja PPA spetsialistid oma domeeni spetsiifiliste teadmistega. Lisaks toome esile Statistikaameti ja Keskkonnaagentuuri, kellelt pärinesid osad projektis kasutatud andmed.

Euroopa Liit
Euroopa
Regionaalarengu Fond

Eesti
tuleviku heaks

MAANTEEAMET