Business understanding

Background

Käesolev projekt tegeleb Tartu rattaringluse (Tartu Smart Bike) ning selle andmetega. Tartu rattaringlus on 2019. aasta 8. juunil alustanud kaasaegne ühistranspordi lahendus, mis koosneb 69-st üle Tartu linna paiknevast rattaparklast ning 750-st jalgrattast, millest 510 on elektri- ning 240 tavajalgrattad. Kasutajatel on võimalus linnas liiklemiseks rentida ühest parklast jalgratas ning hiljem tagastada see teise parklasse.

Kuna rattaringlus on küllaltki uus nähtus Tartu linna elus, siis pole see puudusteta. Seetõttu on Tartu linn jaganud Tartu Ülikooli õppeaine Sissejuhatus andmeteadusesse (LTAT.02.002) raames rattaringluse kohta andmeid, ootusega, et selle aine tudengid oskaksid antud andmestike analüüsimise ja töötlemise läbi pakkuda ideid rattaringluse täiustamiseks.

Kuna ka meie meeskonnaliikmed on rattaringluse kasutajad ja oleme tähele pannud mõningaid puudusi ja probleeme, otsustasime kasutada seda võimalust, et rattaringlust kõigi kasutajate jaoks paremaks muuta.

Business goals

Rattaringluse puhul on täheldada probleemi, et mõned rattajaamad on sageli tühjad, samal ajal kui mõned teised jaamad on jällegi täis. Hetkel küll toimub ka rataste ümberpaigutamine, kuid seda tehakse põhiliselt vaadates hetkeolukorda, mis tähendab, et reageerimine pole alati kõige kiirem.

Meie projekti põhieesmärgiks ongi rataste ümberpaigutamise süsteemi täiustada, muutes seda hetkeolukorra põhjal otsustavast tulevikku vaatavaks. Sellisel viisil toimiv süsteem suudaks ennustada mis kellaajal millisesse jaama on vaja rattaid transportida ning millistest jaamadest on võimalik selleks rattaid ära võtta. Kirjeldatud süsteem vähendaks suuresti eespool kirjeldatud probleemi, kuna see kõrvaldaks muidu reageerimisele kuluva aja.

Seejuures tuleb samuti kindlaks teha, kas sellise süsteemi toimimiseks oleks vaja rattaparklate mahutavusi (dokkide arvu parklas) muuta. Kui projekti käigus peaks selline vajadus tekkima, siis tuleb lisaeesmärgina leida parklate sellised suurused, mis toetaks loodavat süsteemi.

Business success criteria

Meie projekti võib õnnestunuks lugeda kui meie algoritm ennustab 85% protsendilise täpsusega rataste arvu dokis sõltumata kuust ja päevast. Selle sama algoritmiga saame ennustada rataste optimaalset paigutust erinevatel kellaaegadel.

Inventory of resources

Oleme saanud Tartu linnalt rattaringluse kohta andmed ajaperioodilt 2019 juuni - september.
Üldiselt linna ega rattaringluse poolt täiendavat abi ei saa, kuid kui peaks tekkima sisulisi küsimusi, saame pöörduda rattaringluse poole meili teel (info@rattaringlus.ee).
Meie poolt kasutatavaks tarkvaraks on põhiliselt Jupyter Notebook ja Python 3 ning selle erinevad paketid.

Requirements, assumptions, and constraints

Põhiline piirang on see, et me ei tohi linna poolt jagatud andmestikku jagada inimestega väljaspool meie meeskonda, ainukeseks erandiks on aine Sissejuhatus andmeteadusesse instruktorid. Selle kinnitamiseks oleme allkirjastanud lepingu.

Risks and contingencies

Projekti suurimaks riskiks on see, et meie lähenemine kirjeldatud probleemi lahendamiseks ei pruugi olla efektiivsem kui praegu kasutusel olev lahendus. Põhjus peitub selles, et rattaringlust mõjutavad väga mitmed asjaolud, nagu näiteks aastaaeg, suuremate ürituste toimumine linnas, ilmastikuolud jne. Neid faktoreid antud projekti puhul me arvestada ei saa, kuna meil puuduvad vastavad andmed.
Samuti on raskendavaks asjaoluks see, et kuna rattaringlus on küllaltki uus, siis on meil hetkel kasutada vaid nelja kuu andmed ning need andmed on suvekuude ning septembri kohta, seetõttu võivad ka lõpptulemused olla liialt suveperioodile keskendunud või halvemal juhul olla üldse rakendatavad ainult suveperioodi jaoks.

Terminology

Rattaringlus - Kaasaegne ühistranspordi lahendus, kus linna peal on rattaparklad, kust inimesed saavad rentida endale jalgratta linna peal liiklemiseks ning hiljem peavad ratta tagastama samuti ühte rattaparklasse.

Rattaparkla (rattajaam) - Dokkidest koosnev jalgrattaparkla, kust laenutatakse ja kuhu tagastatakse rattaringluses kasutusel olevad jalgrattad.

Dokk - Koht parklas kuhu on kinnitatud jalgratas.

Costs and benefits

Kui me jõuame oma seatud eesmärkideni, siis meie projektist võiks kõige rohkem kasu olla Tartu linnale, kelle valduses rattad on. Kuigi ilmselt pole see nende jaoks äriline kasu, sest tegemist on projektide toel rajatud ettevõtmisega, mille peamine eesmärk peaks olema linnaruumi hubasemaks muutmine ning loodussäästlike tulevikutehnoloogiate katsetamine.

Data-mining goals

Kasutame rattaringluse nelja kuu andmeid, et luua mudel, mille abil on võimalik ennustada kellaajaliselt igas parklas saadaval olevate jalgrataste arvu. Pärast selle mudeli loomist on meil võimalik hakata analüüsima, kas ja millised on need kellaajad, kui parklast laenutatakse kõige rohkem ja kui parkla seisab tühjana või peaaegu tühjana ning samuti neid kellaaegasid, kui parkla on suhteliselt täis. Tehes sellised mustrid kindlaks, saame pakkuda sobivaid aegasid parklasse rataste juurde toomiseks ning võimaldab leida need parklad, kust on sel hetkel võimalik rattaid ära võtta ilma, et antud parklates tekiks potentsiaalne rataste puudujääk.

Data-mining success criteria

Kõige olulisem meie loodava mudeli põhjal analüüsimisel on see, et kui leiame potentsiaalsed parklad, kust saame mingil kellaajal rattaid ära viia, siis tuleb teha kindlaks, et nendest parklatest rataste äravõtmine ei põhjustaks hiljem nendes samades parklates rataste puudujääki. Vastasel juhul oleks selline ümberpaigutamine tulutu. Seda arvestades on võimalik ka hiljem organiseerida nendesse parklatesse rataste tagasitoomine, kuid sellisel juhul peaks jällegi kaaluma, kas selline variant on piisavalt kasulik või on pigem mõttekas üldse mitte antud parklast rattaid ära viia.

Data understanding

Gathering data

Meil on vaja põhiliselt andmeid rataste paiknemise ning liikumise kohta ehk rattalaenutuse algusaeg, lõpuaeg, alguskuupäev, lõppkuupäev, alguspunkt, lõpp-punkt. Lisaks vajame ka dokkide arvu igas parklas.
Rattalaenutuste andmed juuni-septembrikuu kohta on meil juba Tartu linnalt kätte saadud. Täiendavalt on vaja andmeid parklate praeguste suuruste kohta, kuid selle saame loodetavasti rattaringluse käest. Kui sellega peaks probleeme olema, on meil võimalik ka rattaringluse veebilehelt ise vaadata parklate suurused ning selle info põhjal ise andmestik koostada.

Describing data

Meie andmestikus on järgnevad tulbad:
route_code - ID
cyclenumber - ratta number
unlockedat - ratta dokist avamise kuupäev
unlockedattime - ratta dokist avamise kellaaeg
lockedat - ratta dokki lukustamise kuupäev
lockedattime - ratta dokki lukustamise kellaaeg
startstationname - sõidu alustamise parkla
endstationname - sõidu lõpetamise parkla
rfidnumber - vahend millega ratas dokist avati (Smart Bike App või kaart)
length - sõidu pikkus kilomeetrites
DurationMinutes - sõidu kestus minutites
CycleType - ratta tüüp (tava või elektri)
costs - sõidu maksumus
Membership - kasutaja liikmelisus

Andmed on jaotatud nelja faili kuude kaupa, juunist kuni septembrini. Neis on vastavalt 61974, 60291, 58145, 56235 rida. Andmeid tundub olevat piisavalt, et nende põhjal ennustusmudel treenida.

Exploring data

Septembrikuises failis on 1067 kirjet, kus on algpunkt või lõpp-punkt undetermined. Äärmisel juhul on mõni rida täiesti kasutu: 56212 12:19:00+00 Undetermined -- Määramata -- RFID Card 0.78. Puudulike ridade puhul ei jää ilmselt muud üle, kui tuleb need eemaldada, sest alg- ja lõppjaam on mõlemad hädavajalikud. Olemasolevatest tulpadest pole meie projekti jaoks ilmselt vajalikud *rfidnumber, length, DurationMinutes, CycleType, costs* ja *Membership.*

Verifying data quality

Eelmises punktis on täheldatud mõningad puudused andmetes. Üldiselt on andmed siiski piisavalt kvaliteetsed meie projekti läbiviimiseks ning need pärinevad usaldusväärselt allikalt ehk Tartu linnalt endalt.

Projekti plaan

1) Tutvume andmetega põhjalikumalt ning otsime andmetes vigased ning puuduoleva infoga read ning eemaldame need. Samuti eemaldame ebavajalikud tulbad. Tulemuseks saame puhastatud andmestikud.
   Ajakulu: 1 tund
2) Valime mudeli treenimise ja testimise jaoks andmed (training and test dataset).
   Ajakulu: 1 tund
3) Otsustame missuguste algoritmide põhjal hakkame mudeleid treenima ning seejärel treenime mudelid ja valime neist parima täpsusega mudeli.
   Ajakulu: 5 tundi
4) Ennustame mudeli põhjal ~10-minutilise intervalli (võibolla tuleb intervalli suurust muuta) kaupa jalgrataste arvu igas parklas. Saame andmed, kus on kellaaeg, parkla nimi ning rataste arv antud parklas.
   Ajakulu: 3 tundi
5) Analüüsime punktis 4 saadud tulemusi eesmärgiga leida need kellaajad, kui parklad on tühjad (parklas on rattaid alla 10% parkla mahutavusest) ning samuti kellaajad kui parklad on täis (rattaid üle 65% parkla mahutavusest). Tühi ja täis parkla on siin kontekstis defineeritud nendele vastavate lävendite põhjal, kuid ilmselt on vaja leida optimaalsemad lävendid, praegused pakutud lävendid on oletatavad. Antud tulemuste põhjal koostada 2 andmestikku, millest ühes on kellaajad ja parklate nimed, mis on antud kellaaegadel tühjad, teises on kellaajad ja parklate nimed, mis on antud kellaaegadel täis.
   Ajakulu: 3 tundi
6) Punktis 5 loodud andmestike põhjal luua üldine süsteem rataste ümberpaigutamiseks.
   Ajakulu: 8 tundi
7) Ideaalis oleks siinkohal ka loodud süsteemi testimine Tartu rattaringluses.
8) Projekti lõpptulemusena loodud uue rataste ümberpaigutamise süsteemi tutvustamiseks plakati loomine.
   Ajakulu: 6 tundi