https://github.com/erkirb/PAT2023_projekt

Kaggle- Spaceship Titanic

Spaceship Titanic | Kaggle

Projekti nimi: SS-Titanic

Projekti teostaja: Erki Jõekalda

Task 2. Business understanding

Antud juhul on tegemist täiel määral fantaseeritud Kaggle võistluse jaoks fabritseeritud situatsiooniga mille kirjeldusega on võimalik tutvuda ülal viidatud kodulehel.

Eesmärgid

Ülesande püstituse ja eesmärgi seadmise vaatenurgast on oluline faktoloogia, et väljamõeldud kosmoselaeval Titanic toimus kauges tulevikus kokkupõrge ajaanomaalia nähtusega mille tulemusena teleporteerus umbes pool reisijaskonnast teise dimensiooni. Ülesande alusandmestik kannab endas detailseid atribuute reisijate kohta, mida on võimalik kasutada tuvastamaks dimensiooni vahetanud reisijate tunnuste hulgas mustrite või ühisjoonte olemasolu. Paraku on aga andmdestik poolik ning peamiseks eesmärgiks ongi püüda leida mudel mis maksimaalse täpsusega suudaks ennustada test-hulka kuuluvate reisijate puhul, millised reisijaist tõenäoliselt liikusid üle teise dimensiooni**.**

- Analüüs ja klassifitseerimine – kasutades ette antud andmestikku tuleb püüda tuvastada selgeid mustreid mis viitaksid dimensiooni vahetanud reisijatele
- Teha kindlaks riskifaktorid mis enim mõjutavad reisijate dimensiooni vahel liikumise tõenäosust

Antud ülesande puhul on edukuse kriteerium võistluse käigus test-andmestiku pealt teostatud reisijate staatuse ennustuste täpsus.

Olukorra hindamine

Kuna tegemist on Kaggle võistlusega, siis vajalik andmestik on esitatud juba võistluse ülesande püstituses ning lisaandmeid ei ole võimalik koguda. Niimoodi tagatakse ka tulemuste hindamise õiglane ja ühene protsess.

Projekti peamiseks piiranguks on andmete puhastamiseks, eelanalüüsiks ning alternatiivsete mudelite treenimiseks vajaminev ajaline ressurss. Andmete hulk peaks olema piisav asjakohase mudelini jõudmiseks.

Välja mõeldud andmeteaduse ülesande tõttu jäävad käesoleval juhul kõrvale tavapärased äriprojekti eesmärgid ja motivaatorid, nagu näiteks projekti tulemusena potentsiaalselt saavutatav kulude kokkuhoid või tulude kasvatamine. Samuti jäävad kõrvale küsimused potentsiaalsete lisaandmete kogumise võimalustest, probleemid vajalikele andmetele ligipääsuga või väga levinud äriettevõtete andmeladude kvaliteedi- ning koherentsuse probleemid.

Andmekaeve eesmärkide määratlemine

Käesoleva ülesande lahendamise käigus võib eesmärkidena välja tuua järgnevat:

- Andmestikus leiduvate atribuutide klassifitseerimine tuvastamaks tunnuseid mis võiksid aidata reisijate dimensioonided vahelise liikumise tõenäosust ennustada.
- Korrelatsiooni analüüs reisijate tunnuste ja nende staatuste vahel.
- Reisijate demograafiline analüüs (koduplaneet, vanus) leidmaks seoseid transpordi staatusega.

Antud ülesande lahendamise edukuse kriteeriumina võib tinglikult määratleda lõpliku mudeli ennustuse tulemuse paiknemise Kaggle võistluse edetabelis. Võime võtta realistliku eesmärgina paigutumise vähemalt edetabeli keskpaika.

Task 3. Data understanding

Antud Kaggle võistluse ülesande andmestik koosneb infost kosmoselaeva reisijate päritolu ja vanuse kohta, nende reisi sihtmärgist, pileti klassist ning andmetest reisi jooksul tarbitud teenuste kohta.

Andmete detailsem kirjeldus on järgnev:

PassengerId: iga reisija unikaalne tunnus fromaadis gggg_pp mis moodustub reisija grupi koodist (gggg) ja nende järjekorra numbrist grupis

HomePlanet: reisija päritolu planeet

CryoSleep: määratleb kas reisija valis oma reisi ajaks vegetatiivse *cyrosleep* oleku, mis viitab, et reisija veedab kogu reisi aja oma kajutis

Cabin: kajuti number mis viitab ka kajuti asukohale kosmoselaevas

Destination: vastava isiku reisi sihtkoht

Age: reisija vanus

VIP: tõeväärtuse formaadis tunnus reisija VIP staatuse kohta

RoomService, FoodCourt, ShoppingMall, Spa, VRDeck: atribuudid mis näitavad reisija poolt tehtud kulutusi kosmoselaeva erinevates teenuspunktides

Name: reisija nimi

Transported: tõeväärtusena esitatud tunnus kas reisija liikus üle teise dimensiooni või mitte (antud ülesandes 'target label')

Andmete vaatlusuuringud

- Kui tavapäraselt on reisija tunnus ebaoluline väärtus, siis antud juhul esineb selles tunnuses grupeerimise faktor. Seega vajab lähemat uurimist asjaolu, kas grupi tunnus omab mingisugust olulisust ülesande lahendamise vaatenurgast.

- Üle poole reisijate koduplaneediks on Maa ja teised planeedid on oluliselt vähem esindatud. Maalt pärit isikute käitumine võib erineda teistest ning vajab seetõttu lähemat uurimist.
- Reisi sihtkohtadest on selgelt kõige levinum TRAPPIST-1e, seega on mõsitlik uurida kas antud asjaolu omab mingisugused korrelatsioone teiste tunnustega.
- Kuna kajuti number kannab endas infot asukoha kohta laevas, siis vajab kindlasti analüüsi kas kajuti asukoht omab mingisugust olulisust reisija dimensiooni vahetuse kontekstis.
- Teenuste kasutamise mustrid annavad võimaluse uurida peidetud seoste olemasolu reisija staatuse suhtes

Andmete kvaliteet

Esmaste vaatluste tulemusena võib järeldada, et andmete üldine kvaliteet on üsna hea. Koheselt ei hakka silma selgelt vigaseid väärtuseid. Küll leidub andmestikus üsna palju tühjasid väljasid millede puhul edasise analüüsi käigus tuleb jõuda järelduseni, kas ja millise reegli alusel oleks võimalik neid auke täita.

Andmete järejepidevus tundub olevat asjakohane ning ei esine esmapilgul olulisi anomaaliaid.

Kõik tabelis leiduvad andmed (va. reisija nimi) paistavad olema asjakohased ülesande lahendamise vaatenurgast. Põhjalikum analüüs võib küll viidata, et mõned veerud ei kanna endas lisaväärtust parema tulemsue savutamise juures.

Task 4. Planning your project

Käesoleva projekti plaanin läbi viia üksi. Projekti käigus plaanitavad tegevused võib jaotada järgnevatesse gruppidesse:

- Projekti teema valimine, alternatiivsete projekti suundade uurimine, analüüs ja kaalumine (20h)
- Valitud projekti andmestiku ettevalmistus (9h)
    - Puuduvate väärtuste haldus (3h)
    - Kategooriliste muutujate transformeerimine (2h)
    - Andmete normaliseerimine ja võimalik skaleerimine (2h)
    - Kõrvalekallete tuvastamine ja haldus (2h)
- Uuriv andmete analüüs (13h)
    - Andmestiku jaotusanalüüs (näit. atribuudid *HomePlanet, Age, Destination etc*.) (3h)
    - Erinevate atribuutide korrelatsioonide analüüsid (5h)
    - Andmestiku mustrite uurimine ja presenteerimine erinevate graafikute kujul (5h)
- Modelleerimine (13h)
    - Tunnuste valmine ja olulisuse määratlemine (3h)
    - Ennustavate algoritmide loomine ja treenimine (8h)
    - Mudelite tulemuste täpsuse hindamine (2h)
- Optimeerimine ja validatsioon (7h)
    - Paremat potentsiaali näitavate mudelite hüperparameetrite tuunimine (4h)
    - Ristvalideerimine mudelite töökindluse tagamiseks (3h)
- Raporti ja dokumentatsiooni koostamine (8h)

- Protsessi ja analüüsi tulemuste kirjeldamine (5h)
- Lõpliku raporti ja dokumentatsiooni vormistamine (3h)

Meetodid ja vahendid

Projekti progammilise ja analüüsi põhiosa plaanin lahendada Jupyter Notebook baasil kasutades asjakohaseid Pythoni mooduleid.

Kuna antud projekti läbiviimise üheks isiklikuks eesmärgiks olen seadnud kursuse käigus läbitud materjali ja meetodite maksimaalse praktiseerimise, siis plaanin teostada andmestiku üsna põhjaliku eelanalüüsi ning püüda proovida treenida ja optimeerida võimalikult palju erinevaid masinõppe mudeleid. See aitab loodetavasti koguda paremat tunnetust, kuidas erinevad meetodid toimivad teatud tüübilise andmestiku puhul.

Kindlasti on plaan rakendada mudeleid *Logistic Regression, k-Nearest Neighbors, Decision Trees, Support Vector Machines* ja *Random Forest*, kuid ilmselt püüan proovida ka veel mõnda muud lähenemist.

Hetkel koostatud plaani alusel võiks kogu projekti läbi viimisele kuluda umbes 70h aega.