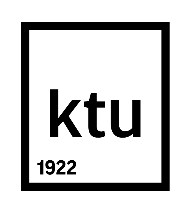
****

**KAUNO TECHNOLOGIJOS UNIVERSITETAS**

**MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS**

**P160M132 DIDŽIŲJŲ DUOMENŲ RINKINIŲ TYRYBOS METODAI**

Kursinis darbas

Atliko: MGDVDAM-9 gr. stud.

Martynas Jonaitis

Edgaras Dvariškis

Priėmė: lekt. Darius Aliulis

**KAUNAS, 2019**

1. **Duomenų paruošimas**

**Duomenų apžvalga**

Projekte buvo naudojami duomenys apie mobilaus ryšio operatoriaus vartotojus. Duomenų faile yra 65 kintamieji, kurie kiekvienas turi 13844 įrašus. Kiekvieną unikalų vartotoją apibūdina trys kintamieji: *<user\_account\_id>*, *<year>* ir *<month>.* Šie kintamieji nurodo kurio periodo ir kurio vartotojo duomenys pateikti. Likę 62 kintamieji apibūdina periodo rodmenis. Juos galima suskirstyti taip:

|  |  |  |
| --- | --- | --- |
| **Kintamieji** | **Apibūdinimas** | **Skaitinė reikšmė** |
| calls\_outgoing\_count | Išeinančių skambučių kiekis | interger |
| calls\_outgoing\_duration | Išeinančių skambučių trukmė | numeric |
| calls\_outgoing\_duration\_max | Ilgiausias skambutis per periodą | numeric |
| calls\_outgoing\_inactive\_days | Dienų, be išeinančių skambučių, kiekis | interger |
| calls\_outgoing\_spendings | Išlaidos išeinantiems skambučiams | numeric |
| calls\_outgoing\_spendings\_max | Brangiausias skambutis per periodą | numeric |
| calls\_outgoing\_to\_abroad\_count | Skambučių į kitas šalis kiekis | interger |
| calls\_outgoing\_to\_abroad\_duration | Trukmė skambučių į uždienį | numeric |
| calls\_outgoing\_to\_abroad\_inactive\_days | Dienų, be išeinančių skambučių į kitas šalis, kiekis | interger |
| calls\_outgoing\_to\_abroad\_spendings | Išlaidos skambučiams į užsienį | numeric |
| calls\_outgoing\_to\_offnet\_count | Skambučių įkitus tinklus, kiekis | interger |
| calls\_outgoing\_to\_offnet\_duration | Išeinančių skambučių į kitus tinklus trukmė | interger |
| calls\_outgoing\_to\_offnet\_inactive\_days | Dienų, be išeinančių skambučių į kitus tinklus, kiekis | interger |
| calls\_outgoing\_to\_offnet\_spendings | Kaina, išleista skambučiams į kitus tinklus | interger |
| calls\_outgoing\_to\_onnet\_count | Skambučių, tinklo viduje, kiekis | interger |
| calls\_outgoing\_to\_onnet\_duration | Trukmė skambučių tinklo viduje | numeric |
| calls\_outgoing\_to\_onnet\_inactive\_days | Dienų, be išeinančių skambučių tinklo viduje, kiekis | interger |
| calls\_outgoing\_to\_onnet\_spendings | Išlaidos skambučiams tinklo viduje | numeric |
| gprs\_inactive\_days | Dienų be duomenų naudojimo kiekis | interger |
| gprs\_session\_count | Prisijungimų skaičius | interger |
| gprs\_spendings | Išlaidos už duomenų naudojimą | numeric |
| gprs\_usage | Išnaudotų kb kiekis | numeric |
| last\_100\_calls\_outgoing\_duration | Paskutinių 100 dienų išeinančių skambučių trukmė | numeric |
| last\_100\_calls\_outgoing\_to\_abroad\_duration | Paskutinių 100 dienų išeinančių skambučių š kitas šalis trukmė | numeric |
| last\_100\_calls\_outgoing\_to\_offnet\_duration | Paskutinių 100 dienų išeinančių skambučių į kitus tinklus trukmė | numeric |
| last\_100\_calls\_outgoing\_to\_onnet\_duration | Paskutinių 100 dienų išeinančių skambučių tiklo viduje trukmė | numeric |
| last\_100\_gprs\_usage | Paskutinių 100 dienų naudotas kb kiekis | numeric |
| last\_100\_reloads\_count | Paskutinių 100 dienų papildymų skaičius | interger |
| last\_100\_reloads\_sum | Paskutinių 100 dienų papildymų suma | numeric |
| last\_100\_sms\_outgoing\_count | Paskutinių 100 dienų SMS žinučių kiekis | interger |
| last\_100\_sms\_outgoing\_to\_abroad\_count | Paskutinių 100 dienų SMS žinučių kiekis į kitas šalis | interger |
| last\_100\_sms\_outgoing\_to\_offnet\_count | Paskutinių 100 dienų SMS žinučių kiekis į kitus tinklus | interger |
| last\_100\_sms\_outgoing\_to\_onnet\_count | Paskutinių 100 dienų SMS žinučių kiekis tinklo viduje | interger |
| month | Mėnesis, kada pradėjo klientas naudotis telefono kompanijos paslaugomis | interger |
| reloads\_count | Papildymų kiekis | interger |
| reloads\_inactive\_days | Dienų, be papildymų, skaičius | interger |
| reloads\_sum | Papildymų bendra suma | numeric |
| sms\_incoming\_count | Įeinančių SMS žinučių skaičius | interger |
| sms\_incoming\_from\_abroad\_count | Įeinančių SMS žinučių iš kitų tinklų skaičius | interger |
| sms\_incoming\_from\_abroad\_spendings | Išlaidos įeinančioms SMS žintumės iš kitų šalių | numeric |
| sms\_incoming\_spendings | Išlaidos įeinančioms SMS žinutėms | numeric |
| sms\_outgoing\_count | Išeinančių SMS žinučių kiekis | interger |
| sms\_outgoing\_inactive\_days | Dienų, be išeinančių SMS žinučių, kiekis | interger |
| sms\_outgoing\_spendings | Išlaidos išeinančios SMS žinutėms | numeric |
| sms\_outgoing\_spendings\_max | Brangiausia SMS žinutė | numeric |
| sms\_outgoing\_to\_abroad\_count | Išeinančių SMS žinučių į kitas šalis skaičius | interger |
| sms\_outgoing\_to\_abroad\_inactive\_days | Dienų, be išeinančių SMS žinučių į kitas šalis, skaičius | interger |
| sms\_outgoing\_to\_abroad\_spendings | Išlaidos išeinančioms SMS žinutėms į kitas šalis | numeric |
| sms\_outgoing\_to\_offnet\_count | Išeinančių SMS žinučių į kitus tinklus kiekis | interger |
| sms\_outgoing\_to\_offnet\_inactive\_days | Dienų, be išeinančių SMS žinučių į kitus tinklus, skaičius | interger |
| sms\_outgoing\_to\_offnet\_spendings | Išlaidos išeinančioms SMS žinutėms į kitus tinklus | numeric |
| sms\_outgoing\_to\_onnet\_count | Išeinančių SMS žinučių tinklo viduje kiekis | interger |
| sms\_outgoing\_to\_onnet\_inactive\_days | Dienų, be išeinančių SMS žinučių tinklo viduje, kiekis | interger |
| sms\_outgoing\_to\_onnet\_spendings | Išlaidos išeinančioms SMS žinutėms tinklo viduje | numeric |
| user\_account\_balance\_last | Kliento sąskaitos likutis periodo gale | numeric |
| user\_account\_id | Unikalus kliento indentifikavimo numeris | interger |
| user\_does\_reload | Klientas padarė bent vieną papildymą | interger |
| user\_has\_outgoing\_calls | Klientas padarė bent vieną skambutį | interger |
| user\_has\_outgoing\_sms | Klientas išsiuntė bent vieną SMS žinutę | interger |
| user\_intake | Naujas klientas | interger |
| user\_lifetime | Kliento buvimo laikas dienomis | interger |
| user\_no\_outgoing\_activity\_in\_days | Dienų, kai vartotojas neatliko jokių operacijų, kiekis | interger |
| user\_spendings | Išlaidos per periodą | numeric |
| user\_use\_gprs | Klientas panaudojo duomenis bent vieną kartą | interger |
| year | Metai, kada pradėjo klientas naudotis telefono kompanijos paslaugomis | interger |

28 *numeric* (realieji skaičiai) tipo kintamieji:

34 *interger* (sveikieji skaičiai) tipo kintamieji:

**DUOMENŲ KOKYBĖS GERINIMAS**Prieš pradėdami duomenų paruošimo modeliavimui žingnius (istornių paslaugų vartojimo duomenų agregravimą pagal vartotojo ID) ir modeliavimą, turime atlikti žvalgomają analizę ir duomenų kokybės gerinimą. Duomenų kokybės gerinimas buvo vykdomas tiukrinant kiekvieną kintamąjį ir ieškant nekorektiškų reikšmių. Randant tokias rekšmes jas šaliname iš imties ir tolimesnį darbą darome su likusia imtimi. Tai darysime naudodmai python funkciją „Describe“, kuri padės matyti: minimumo ir maksimumo taškus, kvantilius, standartinį nuokrypį ir vidurkį ; naudosime histogramas, kuriuose matysis, kaip pasiskirstę duomenys. Toliau aprašysime tuos kintamuosius, kuriuos reikėjo sutvarkyti:

***user\_lifetime*** - šis kintamasis apibūdina kiek dienų klientas jau yra vartotojas. Kadangi turi kelių periodų (tiksliai trijų mėnesių) duomenis, matom, jog po kiekvieno periodo (mėnesio) šis kintamasis padidėja 31 vnt. Todėl teigiame, jog kintamasis yra matuojamas dienomis ir kiekvienas periodas (mėnesis) laikomas 31 dienos ilgio. Nusibrėžus šio kintamojo histogramą (1 pav.) matome, jog dalis reikšmių yra didesnės negu 15000. Laikantis ankstesnių prielaidų gauname, jog tokie vartotojai yra klientai jau daugiau negu 40 metų t.y. nuo ankščiau negu 1973 metai (mūsų nagrinėjame duomenys yra 2013 metų). Kadangi pirmasis komercinis mobilusis ryšys pradėtas teikti tik 1979 metais, galite teigti, jog visi šie įrašai yra klaidingi. Panaikinta 16550 įrašai.

***user\_intake*** - šis kintamasis yra binarinis ir nurodo ar klientas yra naujas. Pagal histogramą (2 pav.) matome, jog tokių yra mažuma – tiksliai 1344. Tačiau didžioji dalis, tiksliai 1322, yra nekorektiški, nes jiems kintamojo *<user\_lifetime>* reikšmė yra nelygi nuliui t.y. nors klientai ir nauji, jie jau yra vartotojai iš ankščiau. Kadangi atmetus šiuos įrašus lieka 20 naujų klientų įrašų, panaikiname ir juos, laikydami jog tolimesniam klaisifikavimui jie neturės įtakos. Bendrai pašaliname 1344 įrašus.