****

**KAUNO TECHNOLOGIJOS UNIVERSITETAS**

**MATEMATIKOS IR GAMTOS MOKSLŲ FAKULTETAS**

**Savarankiško projekto ataskaita**

Studijų modulio

P160M132 DIDŽIŲJŲ DUOMENŲ RINKINIŲ TYRYBOS METODAI

|  |  |
| --- | --- |
|  |  |
| **Darbą parengė:**  Anatolijus Nachumovas  Martynas Ragaišis  Ugnė Butkutė | (parašas) (data) |
|  |  |
| **Darbą priėmė:**  lekt. D. Aliulis | (parašas) (data) |
|  |  |

**Vilnius, 2019**

Contents

[1. Duomenų rinkinio žvalgomoji analizė 4](#_Toc24919481)

[1.1. Duomenų eliminavimas 4](#_Toc24919482)

[2. Atrinktų duomenų klasterizavimas 6](#_Toc24919483)

[2.1. Duomenų apžvalga prieš klasterizavimą 6](#_Toc24919484)

[2.2. Duomenų klasterizavimas 6](#_Toc24919485)

[Išvados 8](#_Toc24919486)

[Literatūros sąrašas 9](#_Toc24919487)

DARBO TIKSLAS

Panaudojus išmoktus duomenų tyrybos metodus ir Apache Spark bei Python programines priemones, sudaryti mobilaus ryšio operatoriaus klientų segmentavimo ir klientų praradimo prognozavimo modelius. Atlikti žvalgomosios duomenų analizės, modeliavimo, modelių įvertinimo programinę realizaciją. Parengti duomenų tyrybos rezultatų aiškinamąjį raštą su išvadomis.

DUOMENYS

Duomenų rinkiniai pateikti Moodle sistemoje nurodytu adresu. Duomenų rinkinį kiekvienam studentui sudaro 2 failai:

1. customer\_usage\_**<varianto\_nr>**.csv – agreguoti vartotojų birželio-rugpjūčio mėnesių paslaugų vartojimo duomenys. Kintamasis user\_account\_id nurodo kliento identifikacinį ID.

2. customer\_churners\_**<varianto\_nr>**.csv – klientų praradimo istorija. Kintamojo churn reikšmė 1 nurodo, kad klientas rugsėjo mėnesį paliko kompaniją, 0 reiškmė nurodo, kad klientas toliau naudojosi paslaugomis.

UŽDUOTIS

1. Atlikti duomenų rinkinio žvalgomąją analizę.

2. Panaudojus duomenų kokybės gerinimo ir K–vidurkių arba GMM metodus sudaryti telekomunikacijų kompanijos klientų geriausią segmentavimo modelį ir parinkti optimalų segmentų skaičių pagal pasirinktą klasterizavimo metriką.

3. Interpretuoti gautus segmentus.

4. Panaudojus klasifikavimo metodus sudaryti klientų praradimo prognozavimo modelius visiems segmentams. Įvertinti sudarytų modelių prognozavimo kokybę pasirinktomis metrikomis.

PAPILDOMA UŽDUOTIS

5. Parinkti optimalų segmentų skaičių pagal klientų praradimo prognozavimo kokybę (kiekvienam segmentui sudarius po klasifikavimo modelį ir jį įvertinus pagal pasirinktą klasifikavimo metriką).

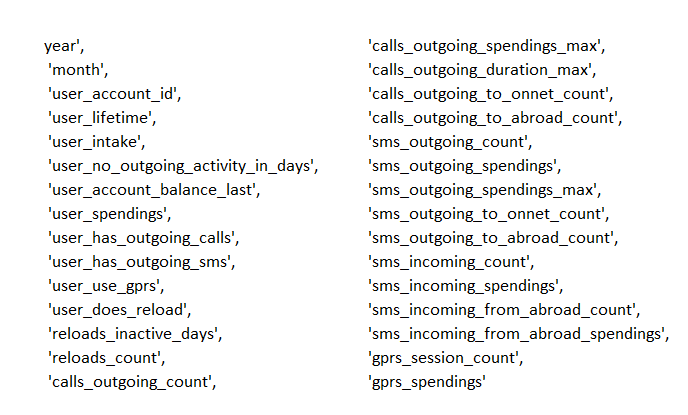
# Duomenų rinkinio žvalgomoji analizė

Mūsų grupės gauto duomenų rinkinio numeris – 4. Pradinėje darbo stadijoje turime du skirtingus .csv tipo failus: customer\_churn\_00004.csv bei customer\_usage\_00004.csv. Duomenų rinkinį sudaro – 49885 skirtingų vartotojų id.

## Duomenų eliminavimas

Atlikus duomenų apžvalgą, buvo nuspręsta pasirinkti 8 skirtingas duomenų grupes, kurių pagalba įgyvendinsime savo projektą. Tam, kad įvertintume, kurie duomenys yra mums naudingi, o kurių eliminavimas nesukeltų didelių kokybės pokyčių projektui, atlikome duomenų koreliacijos tyrimą.

Koreliacijos koeficientus skaičiavome Spearmano bei Pearsono metodais, tačiau buvo priimtas sprendimas atsižvelgti į pastarojo metodo gautus sprendinius. Didesnis duomenų kiekis turėjo aukštesnį nei 0.75 koreliacijos koeficiantą. Remiantis gautais rezultatais duomenų rinkinys buvo padalintas į dvi dalis: duomenys, kurių tolimesnioje analizėje nenaudosime ir duomenys, kurie gali būti naudojami tolimesniuose žingsniuose.



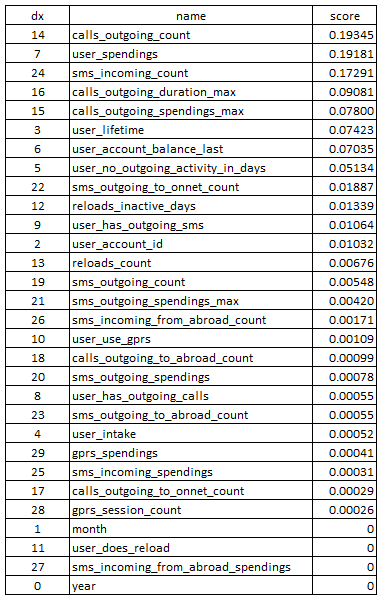
**1 pav.** Duomenų grupė likusi po koreliacijos etapo

Suskirsčius turimą duomenų rinkinį į dvi grupes buvo nuspręsta galutiniam projekto variantui išsirinkti 8 svarbiausius ir daugiausia informacijos galinčius suteikti duomenų stulpelius. Šiame žingsnyje buvo nuspręsta įvertinti mūsų grupės įžvalgas bei sprendimo medžio pateiktus rezultatus. Tokie duomenys, kaip „user\_spending“, „user\_no\_outgoing\_activity\_in\_days“, „sms\_incoming count“ bei „calls outgoing\_count“ buvo išskirti tiek mūsų grupės įžvalgose tiek sprendimo medžio rezultatuose.

***RandomForestClassifier(labelCol="churn", featuresCol="initial\_features", seed = 8464,***

***numTrees=10, cacheNodeIds = True, subsamplingRate = 0.7)***

Aukščiau pateikiame sukurtą sprendimo medžio apmokymo kodą. Pasinaudojus juo gauname sprendinius, kuriuos palyginame su savo grupės padarytomis įžvalgomis ir galutinai nusprendžiame, kuriuos duomenis įtraukti tarp 8 analizei naudotinų (žr. 2 pav.).



**1 lentelė** Sprendimų medžio rezultatai

Atsižvelgiant į gautus sprendinius, matome, kad „calls\_outgoing\_count“, „user\_spendings“ bei „sms\_incoming\_count“ surinko didžiausius balus. Šie trys duomenų rinkiniai taip pat buvo įtraukti per mūsų grupės diskusiją.

Tiesa, nors koreliacijos vertinimo metu buvo atmesti ,,calls\_outgoing\_spending“ duomenys, tačiau vėliau nutarta šiuos duomenis taip pat įtraukti tarp 8 svarbiausių ir tolimesniuose žingsniuose naudotinų duomenų. Pastebėta, kad šie duomenys koreliuoja su duomenų rinkiniais, kurie nebus naudojami sekančiuose analizės žingsniuose.

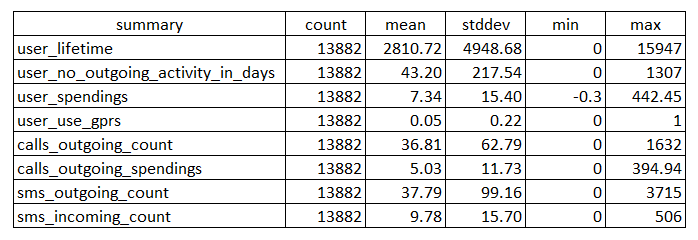
Taigi, sekančiuose skyriuose analizuosime šiuos duomenis: ,,calls\_outgoing\_count“, „user\_spending“, „sms\_incoming\_count“, „user\_use\_gprs“, „sms\_outgoing\_count“, „user\_lifetime“, „user\_no\_outgoing\_activity\_in\_days“ bei „calls\_outgoing\_spending“. Panaudojus sprendimo medžio gautus rezultatus ir mūsų išvadas manome, jog šie duomenys yra tinkamiausi gauti išsamias įžvalgas šios telekomunikacijų paslaugas teikiančios įmonės vartotojų įpročių tendencijas.

# Atrinktų duomenų klasterizavimas

Šioje darbo dalyje pateikiame išsirinktų duomenų klasterizavimo metodus ir gautus rezultatus. Klasterizavimas bus atliekamas su 8 skirtingais duomenų rinkiniais, kurių pasirinkimai buvo atlikti po pirminės duomenų apžvalgos.

## Duomenų apžvalga prieš klasterizavimą

Žemiau pateiktoje lentelėje matome išrinktų duomenų imtį (count), vidurkį (mean), standartinį nuokrypį (stddev) bei minimalias (min) ir maksimalias (max) reikšmes (žr. 3 pav.). Matome, kad didesnį nuokrypį turi „user\_no\_outgoing\_activity\_in\_days“ bei „user\_use\_gprs“. Tiesa, išskirčių galima tikėtis atsižvelgiant į maksimalias reikšmes kiekviename iš nagrinėjamų duomenų rinkinių.

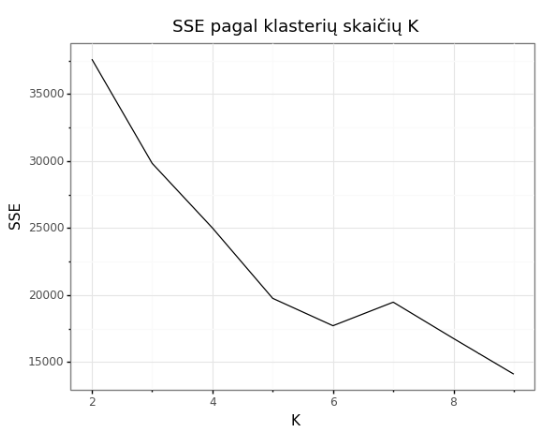
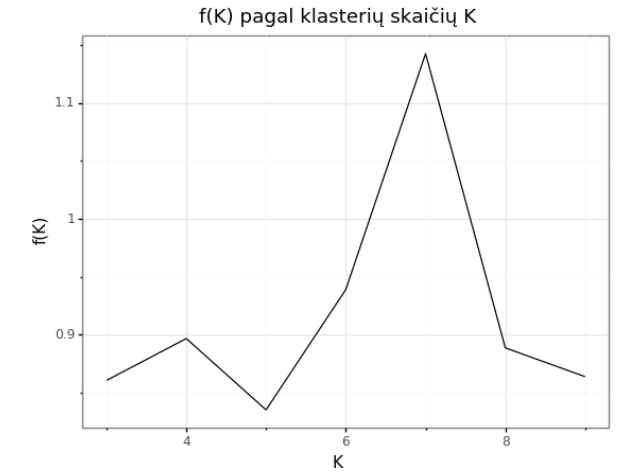


**2 lentelė** Atrinktų duomenų statistinė analizė

Didžiausią „calls\_outgoing\_count“ reikšmę sudaro 44.3 vidurkio reikšmes. Tuo tarpu „sms\_incoming\_count“ maksimali reikšmė vidurkį viršyja net 51.74 kartus. Šie duomenys atlikus klasterizavimą padės įvertinti vartotojų įpročius siunčiant ir priimėjant trumpasias žinutes (sms), taip pat skambučių įpročius ir su šiomis paslaugomis susijusias išlaidas.

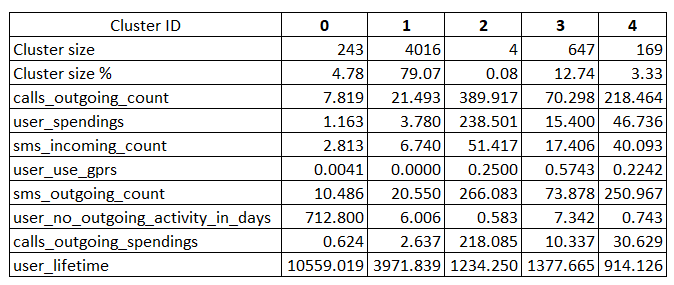
## Duomenų klasterizavimas

Norėdami tiksliai sužinoti, kiek klasterių būtų efektyviausia išskirti su mūsų turimais duomenimis, buvo nuspręsta atlikti SSE bei f(K) metrikų analizę. Pasirinkome abiem atvejam modelius nuo 2 iki 9 klasterių, kad grafikų pagalba būtų lengviau priimti mums labiausiai tinkantį sprendimą. Mūsų atveju abi analizės parodė skirtingus rezultatus (žr. 4 pav.). Pirmuoju atveju, SSE metrikos identifikuoja, kad efektyvu būtų pasirinkti 6 skirtingus klasterius. Visgi, antruoju metodu aiškiai matome, kad 5 klasteriai – efektyviausias sprendimas.



**2 pav.** SSE ir f(K) analizių rezultatai

Turėdami aukščiau pateiktus grafikus pasitarus mūsų grupėje buvo nuspręsta, kad 5 klasteriai visgi būtų efektyviausias ir priimtiniausias sprendimas. Todėl, tolimesniame projekto žingsyje atliekame vartotojų segmentaciją su 5 klasteriais.



**3 lentelė** Klasteriai ir duomenų centrai

Suskirtyti duomenys yra į 5 klasterius su ID kodais nuo 0 iki 4. Svarbu atkreipti dėmesį į klasterį, kurio ID yra 2. Šis klasteris savyje talpina išskirtis tokias kaip „calls\_outgoing\_spendings“. Matome, kad vidutiniškai kituose klasteriuose išlaidos skambučiams sudaro nuo 0.62 iki 30.63 piniginių vienetų. Visgi, 2-ame klasteryje šių duomenų rinkinio centras yra 218.09. Ta pati situacija pastebima ir su kitu duomenų rinkiniu - „user\_spendings“ (bendros kliento išlaidos). Tuo tarpu klasterio 0 „user\_lifetime“ centras beveik 10 kartų lenkia 2, 3 bei 4 klasterio to pačio duomenų rinkinio centrus.

Šį k-means klasterizavimo modelį išsaugojome Apache Parquet formatu, kad prireikus ateityje, būtų lengvai pasiekiamas visiems grupės nariams. Visi duomenų failai bei aprašai išsaugoti GitHub repozitorijoje.

Išvados

1. Atlikus pirminę duomenų analizę bei pasitelkus sprendimų medžio sudarytų duomenų reitingavimu buvo išrinkti 8 mūsų nuomone svarbiausi duomenų rinkiniai, kurie toliau panaudoti klasterių sudarymui.
2. Išanalizavus SSE bei f(K) pateiktus siūlymus buvo priimtas sprendimas duomenų rinkinius klasterizuoti į 5 atskirus klasterius.
3. Atlikus k-means klasterizavimą pastebėta, kad vienas iš klasterių išsiskiria iš kitų dėl ypač mažo duomenų kiekio ir didelių išskirčių. Visig, daugiau nei 79 proc. duomenų pateko į klasterį pažymėtu 1 ID.

Literatūros sąrašas

1. Han J., Kamber M. Data Mining: Concepts and Techniques (3 ed.). Burlington, MA, USA: Elsevier Science & Technology, 2011.-496 pp. Prieiga per http://site.ebrary.com/lib/ktu/Doc?id=1039930

2. Didžiųjų duomenų rinkinių tyrybos metodai. Studijų medžiaga. Kaunas: KTU, 2019. Prieiga per http://vma.ktu.lt/

3. PySpark programinės įrangos dokumentacija [interaktyvus]. 2019 [žiūrėta 2019-10-11]. Prieiga per: http://spark.apache.org/docs/latest/api/python/index.html

4. Python 3.7 programavimo kalbos dokumentacija [interaktyvus]. 2019 [žiūrėta 2019-10-11]. Prieiga per: https://docs.python.org/3/

5. Exacaster White paper: Essential Guide for Predicting Customer Churn [interaktyvus]. 2019 [žiūrėta 2019-10-11]. Prieiga per: http://www.exacaster.com/whitepaper-essential-guide-predicting-customer-churn/

6. Segmentation for Prepaid: An Exacaster White Paper. 2019 [žiūrėta 2019-10-11]. Prieiga per: http://www.exacaster.com/whitepaper-segmentation-for-prepaid/

7. *Berry M. J., Linoff G. S. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management (3rd Edition). Hoboken, NJ, USA, Wiley, 2011*. Prieiga per http://site.ebrary.com/lib/ktu/docDetail.action?docID=10513818