****

**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**

Turinys

[Įvadas 3](#_Toc24906711)

[1. Duomenų rinkinio žvalgomoji analizė 4](#_Toc24906712)

[Išvados 6](#_Toc24906713)

[Literatūros sąrašas 7](#_Toc24906714)

[Informacijos šaltinių sąrašas (pagal poreikį) 8](#_Toc24906715)

[Priedai (pagal poreikį) 9](#_Toc24906716)

Įvadas

Lorem ipsum dolor sit amet, eam ex decore persequeris, sit at illud lobortis atomorum. Sed dolorem quaerendum ne, prompta instructior ne pri. Et mel partiendo suscipiantur, docendi abhorreant ea sit. Recteque imperdiet eum te.

Eu eum decore inimicus consetetur, cu usu habeo corpora intellegam. Ut antiopam efficiendi deterruisset sit. Mel sint eirmod id, qui quot virtute id, dolor nemore forensibus usu id. Fugit dolore voluptatum cu vim. An vix veniam graecis insolens, sit posse iusto id. Ut vim ceteros percipit, id quo ubique recusabo, eum sint lucilius ea. In sumo inani numquam has.

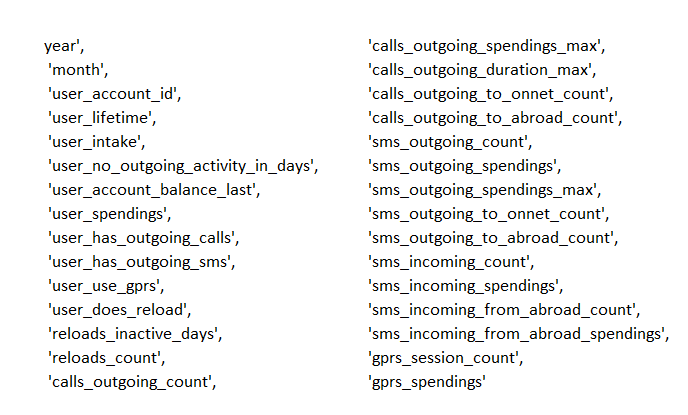
# Duomenų rinkinio žvalgomoji analizė

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.



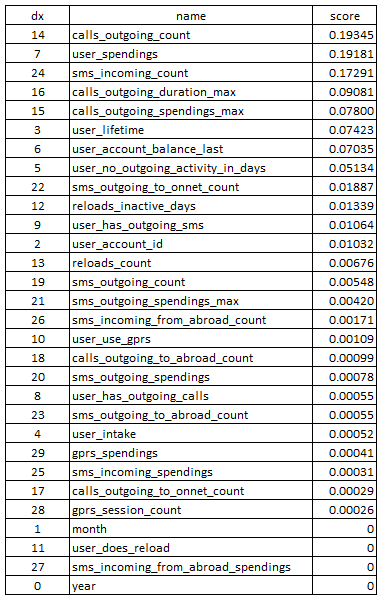
**1 pav.** Duomenų grupė likusi po koreliacijos etapo

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.).



**2 pav.** 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.

Išvados

1. Lorem ipsum dolor sit amet, eam ex decore persequeris, sit at illud lobortis atomorum. Sed dolorem quaerendum ne, prompta instructior ne pri. Et mel partiendo suscipiantur, docendi abhorreant ea sit. Recteque imperdiet eum te.
2. Eu eum decore inimicus consetetur, cu usu habeo corpora intellegam. Ut antiopam efficiendi deterruisset sit. Mel sint eirmod id, qui quot virtute id, dolor nemore forensibus usu id. Fugit dolore voluptatum cu vim. An vix veniam graecis insolens, sit posse iusto id. Ut vim ceteros percipit, id quo ubique recusabo, eum sint lucilius ea. In sumo inani numquam has.

Literatūros sąrašas

1. Literatūros šaltinis
2. Literatūros šaltinis
3. Literatūros šaltinis
4. Literatūros šaltinis

Informacijos šaltinių sąrašas (pagal poreikį)

1. Informacijos šaltinis
2. Informacijos šaltinis
3. Informacijos šaltinis
4. Informacijos šaltinis

Priedai (pagal poreikį)

1. priedas. Priedo pavadinimas

Lorem ipsum dolor sit amet, eam ex decore persequeris, sit at illud lobortis atomorum. Sed dolorem quaerendum ne, prompta instructior ne pri. Et mel partiendo suscipiantur, docendi abhorreant ea sit. Recteque imperdiet eum te.

Eu eum decore inimicus consetetur, cu usu habeo corpora intellegam. Ut antiopam efficiendi deterruisset sit. Mel sint eirmod id, qui quot virtute id, dolor nemore forensibus usu id. Fugit dolore voluptatum cu vim. An vix veniam graecis insolens, sit posse iusto id. Ut vim ceteros percipit, id quo ubique recusabo, eum sint lucilius ea. In sumo inani numquam has.