Vartotojų praradimo identifikavimas taikant mašininio mokymosi ir statistinius metodus

Martynas Taparauskas, Rūta Levulienė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Taikomosios matematikos institutas, Naugarduko g. 24, Vilnius, Lietuva
martynas.taparauskas@mif.stud.vu.lt

Santrauka. Vartotojų praradimo identifikavimas leidžia programėles valdančioms įmonėms įvertinti įvairų rodiklių įtaką klientų išlaikymui ir planuoti artėjančią tinklų apkrovą. Šiame tyrime analizuojami vartotojų aktyvumo bei programėlės techninės kokybės duomenys ir taikomi statistiniai bei mašininio mokymosi modeliai siekiant įvertinti, kurie vartotojai atsisakys paslaugos pasibaigus abonementui. Rezultatai rodo, kad tinkamiausias modelis nagrinėjamo duomenų rinkinio klasifikavimui – atsitiktiniai miškai.

Raktiniai žodžiai: vartotojų praradimas, klasifikavimas, mašininis mokymasis, klasių balansavimas.

1 Įvadas

Vartotojų praradimo identifikavimas padeda įmonėms ne tik suprasti, kokie veiksniai daro didžiausią įtaką vartotojų išlaikymui, bet ir įvertinti, kiek vartotojų prasitęs abonementus artėjančiu laikotarpiu, kas gali būti naudinga planuojant tinklų apkrovą. Šių prognozių patikimumui bei tikslumui reikia tinkamai parinktų ir informatyvių modelių. Ankstesniuose tyrimuose prarastų vartotojų identifikavimui naudojami statistiniai metodai su tiesiogine kintamųjų įtakos interpretacija, pavyzdžiui logistinė regresija bei jos LASSO modifikacija [1]. Taip pat plačiai pritaikomi ir mašininio mokymosi modeliai, gebantys pateikti kintamųjų svarbos (angl. feature importance) įverčius, pvz., atsitiktiniai miškai ir sprendimų medžiai [2] ar medžiais paremti gradientinio auginimo metodai [3]. Pagrindinis šio tyrimo tikslas – apmokyti interpretuojamus prognozės modelius klasifikuoti vartotojus, kurie atsisakys paslaugos pasibaigus abonementui.

DOI: https://doi.org/10.15388/LMITT.2026.30

2 Metodologija

Tyrime buvo taikyti šie mašininio mokymosi ir statistiniai modeliai: LASSO logistinė regresija [4], sprendimų medžiai [5], atsitiktiniai miškai [6] ir XGBoost [7].

Optimalūs parametrai parenkami taikant gardelės metodą su kryžmine patikra. LASSO regresijai optimizuojamas reguliarizacijos stiprumas ir maksimalus iteracijų skaičius, sprendimų medžiams optimizuojamas maksimalus medžio gylis, minimalus imties dydis reikalingas mazgo padalinimui ir lapams, atsitiktinių miškų modeliui ir XGBoost optimizuojamas medžių kiekis bei maksimalus medžių gylis. Taip pat XGBoost optimizacija papildoma ir mokymosi greičiu.

Siekiant padidinti klasifikavimo tikslumą tyrime taip pat išbandyti klasių balansavimo metodai SMOTE, ADASYN, parodę teigiamus rezultatus tobulinant atsisakančių vartotojų klasifikavimo modelius, kai yra nebalansuotas duomenų rinkinys (žr. [8]).

Klasifikavimo tikslumas vertintas pagal jautrumą, tikslumą, bendrą tikslumą, F1 matą, bei lyginant modelių ROC kreives.

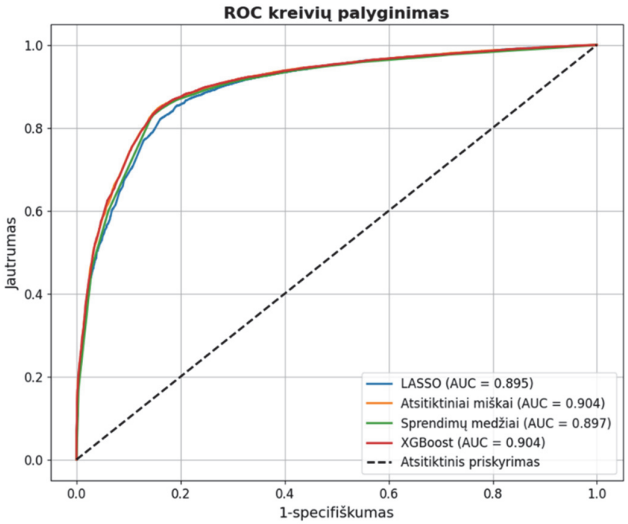
3 Rezultatai

Tyrime naudoti viešai neprieinami, anonimizuoti mobilios programėlės vartotojų aktyvumo ir techninės kokybės duomenys. Siekiant užtikrinti anonimiškumą, duomenų rinkinyje intervaliniai kintamieji buvo standartizuoti, o kategoriniai kintamieji perkoduoti pseudokintamaisiais. Duomenys padalinti į mokymo ir testavimo aibes 80:20 santykiu, mokymo aibėje yra 96359 įrašai, testavimo – 24090. Tyrime sprendžiama dviejų klasių klasifikavimo problema, atsisakiusių paslaugos vartotojų duomenų rinkinyje yra ~59 %.
Toliau pateikiamas modelių klasifikavimo tikslumo palyginimas naudojant testavimo duomenų aibę ir anksčiau minėtas metrikas (žr. 1 lentelę ir 1 pav.). Juoda punktyrinė linija 1 pav. iliustruoja, kaip atrodytų ROC kreivė atsitiktinai priskiriant kiekvieną vartotoją į vieną ar kitą grupę su lygiomis tikimybėmis. Lyginant pasirinktų modelių ROC kreives galima pastebėti, kad visi modeliai yra geresni, negu atsitiktinis priskyrimas, o XGBoost ir atsitiktiniai miškai pasiekė aukščiausią AUC rodiklį, kuris sutapo abiem atvejais. Remiantis išplėstinėmis modelių klasifikavimo kokybės metrikomis 1 lentelėje galima pastebėti, kad atsitiktiniai miškai pasiekė aukščiausius rodiklius visais atvejais, išskyrus tikslumo, kur rezultatas sutapo su XGBoost modeliu.

Sintetiniai klasių balansavimo metodai SMOTE ir ADASYN klasifikavimo rodiklių reikšmingai nepakeitė. XGBoost modelis su SMOTE balansavimu pasiekė aukščiausią bendra tikslumą ir ROC AUC įverti tarp jų - 84,49 % ir 0,903 atitinkamai, tačiau šie rodikliai yra mažesni negu atsitiktinių miškų modelio apmokyto naudojant nekoreguotą mokymo aibę.

1 **lentelė.** Modelių klasifikavimo kokybės rezultatai.

Modelis	Jautrumas	Tikslumas	F1	Bendras tikslumas
LASSO logistinė regresija	81,85 %	82,85 %	82,23 %	83,01 %
Sprendimų medžiai	83,85 %	83,77 %	83,81 %	84,28 %
Atsitiktiniai miškai	**84,31 %**	**84,04 %**	**84,16 %**	**84,58 %**
XGBoost	84,17 %	**84,04 %**	84,10 %	84,55 %



1 pav. Pritaikytų modelių ROC kreivės ir AUC įverčiai.

4 Išvados

Remiantis tyrimo rezultatais galime teigti, kad pasirinkti modeliai yra tinkami klasifikuoti vartotojus į atsisakysiančius paslaugos ir toliau vartosiančius,

o sintetinis klasių balansavimas klasifikavimo kokybei įtakos neturėjo. Atsitiktinių miškų modelis parodė geriausius klasifikavimo rezultatus tirtam duomenų rinkiniui, todėl šį modelį ir rekomenduojame prarastų vartotojų klasifikavimui esant panašaus pobūdžio duomenims.

Literatūra

[1] Nababan, A. M., Purnamasari, F., Elveny, M., Fransiska, C., Zendrato, N., Nasution, U. R. P., & Rahmat, R. F. (2024). Logistic Regression for Merchant Customer Churn Prediction: A Data-Driven Approach. Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI), 72–77. https://doi.org/10.1145/3696271.3696283

[2] Ou, L. (2023). Customer Churn Prediction Based on Interpretable Machine Learning Algorithms in Telecom Industry. 644–647. https://doi.org/10.1109/csmis60634.2023.00120

[3] Xu, Y., & Zhang, S. (2025). Application of XGBoost Algorithm in E-commerce Platform User Loss Prediction and Precise Marketing Intervention in Marketing. 202–207. https://doi.org/10.1145/3730436.3730469

[4] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. http://www.jstor.org/stable/2346178

[5] Rokach, L., & Maimon, O. (2005). Decision Trees. In *Data Mining and Knowledge Discovery Handbook* (pp. 165–192). https://doi.org/10.1007/0-387-5465-x_9

[6] Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

[7] Chen, T., & Guestrin, C. (2016). XGBoost: a Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, *1*(1), 785–794. https://doi.org/10.1145/2939672.2939785

[8] Imani, M., Zahra Ghaderpour, Majid Joudaki, & Beikmohammadi, A. (2024). The Impact of SMOTE and ADASYN on Random Forest and Advanced Gradient Boosting Techniques in Telecom Customer Churn Prediction. https://doi.org/10.1109/icwr61162.2024.10533320