Išgyvenamumo modelių taikymas personalo kaitai prognozuoti

Vilius Kavaliauskas

Vilniaus universitetas, Taikomosios matematikos institutas
Naugarduko g. 24, Vilnius
vilius.kavaliauskas@mif.stud.vu.lt

Santrauka. Personalo stabilumas yra itin svarbus įmonės sėkmės komponentas. Suprasti, kas labiausiai įtakoja darbuotojų kaitą, yra dažnai (ir prasmingai) darbdavio keliamas tikslas. Nors įprastai tam pasitelkiami klasikinės statistikos sprendimai, jie nebūtinai yra geriausias pasirinkimas. Šiame darbe standartiniam duomenų rinkiniui pritaikyti ir palyginti trys išgyvenamumo analizės metodai. Nustatyta, jog atsitiktiniai išgyvenimo miškai šiuo atveju veikia geriausiai.

Raktiniai žodžiai: personalo kaita, išgyvenamumo analizė, mašininis mokymasis.

1 Įvadas

Darbovietės personalo stabilumas yra svarbus kompanijos sėkmės komponentas. Darbdavio užduotis yra ne tik rasti aukštą potencialą turinčius individus, tačiau ir žinoti esminius faktorius, didinančius nepasitenkinimo ir, galiausiai, išėjimo riziką.

Šio uždavinio sprendimo metodų yra pakankamai. Logistinė regresija, Naivusis Bajeso, atsitiktinių miškų ar atraminių vektorių klasifikatoriai [1] dažnai minimi kaip galimi klasifikavimo metodai. Tačiau naudojant šiuos metodus neatsižvelgiama į laiko įtaką duomenims, kadangi kiekviena duomenų eilutė laikoma atskiru nepriklausomu stebėjimu [3].

Tokio kompromiso nereikia pritaikius išgyvenamumo analizės metodus, kai modeliuojamas laikas ir cenzūruoti stebėjimai naudojami kaip papildoma informacija. Tai nėra naujovė nei statistikos, nei darbuotojų kaitos uždavinio kontekste [3, 7]. Visgi, išgyvenamumo analizė nėra dažnai minima kaip galima personalo kaitos analizės alternatyva.

Šio darbo tikslas yra palyginti klasikinių ir inovatyvių išgyvenamumo analizės metodų taikymą darbuotojo išėjimo iš darbo prognozavimui.

DOI: https://doi.org/10.15388/LMITT.2024.6

2 Duomenys

Naudojamas laisvai prieinamas *Edward Babushkin* pateiktas realus duomenų rinkinys[1] apie 1129 darbuotojus iš įvairių pramonės sričių. Jame pateikti lytis, amžius, užmokesčio tipas, keliavimo į darbą būdas, įsidarbinimo šaltinis ir Didžiojo Penketo (angl. *Big Five)* asmenybės bruožų[2] įvertinimai. Duomenys išskaidyti į apmokymo ir testavimo aibes santykiu 7:3.

3 Metodai

Pirmasis metodas – parametrinė AFT (angl. *Accelerated Failure Time*) regresija [6]. Grafiniam tinkamumui nustatyti naudojamas sukurtas programinis įrankis, leidžiantis nustatyti tinkamus parametrinius skirstinius. Patikrinus eksponentinį, Veibulo, loglogistinį bei lognormalųjį skirstinius, akivaizdžiai netinka tik pastarasis. Tada skirstinių tinkamumas tikrinamas tikėtinumų santykio kriterijumi. Nustatyta, jog tinka Veibulo skirstinys.

Kitas naudojamas metodas – Kokso semiparametrinė proporcingųjų rizikų (angl. *Proportional Hazards*, PH) regresija [5]. Tai dažniausiai naudojamas ir geriausiai žinomas išgyvenamumo analizės modelis. Po stratifikavimo, proporcingųjų rizikų prielaidą modelis tenkina.

Trečias taikytas metodas – atsitiktiniai išgyvenimo miškai (angl. *Random Survival Forests*, RSF) [3]. Tai atsitiktinių miškų modifikacija, kur kiekviename medyje siekiama atskirti kuo labiau išgyvenimo charakteristika besiskiriančius individus. Esminiai hiperparametrai – medžių skaičius bei skaidymo taisyklė. Nustatyta, jog optimalią paklaidą fiksavo 500 medžių turintis miškas su lograngine skaidymo taisykle.

4 Rezultatai

Modelių palyginimui skaičiuojamas konkordancijos koeficientas [2]. Kuo jis arčiau 1, tuo modelis veikia tiksliau. Tiek apmokymo, tiek testavimo aibėje geriausiai pasirodo RSF modelis (1 lentelė). Tiesa, konkordancija testavimo aibėje pastebimai sumažėja.

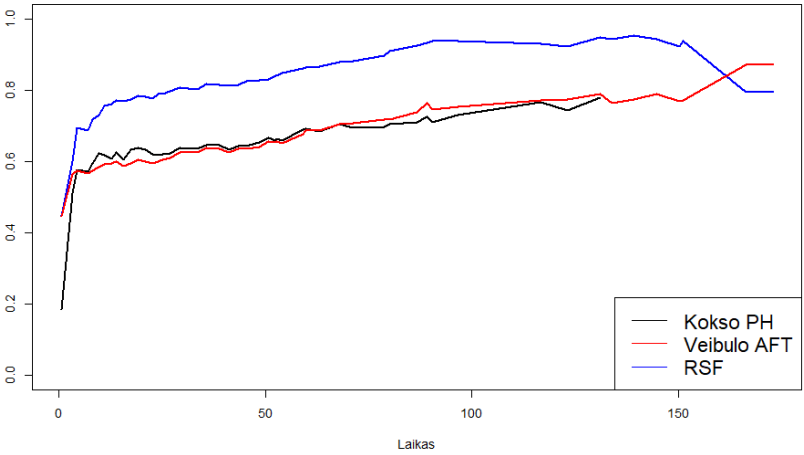
---

1 Viešai prieinamas Kaggle bei autoriaus tinklalapyje.
2 Emocinio stabilumo, ekstraversijos, atvirumo patirtims, sukalbamumo, sąmoningumo

1 lentelė. Konkordancijos indeksas apmokymo ir testavimo aibėse.

	Apmokymo aibė	Testavimo aibė
Veibulo AFT	0,659	0,565
Kokso PH	0,636	0,548
RSF	0,829	0,622

Apmokymo aibei nubraižomas dinaminis AUC (angl. *Area Under Curve*, kur *Curve* yra ROC kreivė) [4]. Kuo kreivė arčiau 1 visuose laiko taškuose, tuo modelis geresnis (1 pav.). Išvados dėl tinkamiausio modelio panašios kaip ir naudojant konkordancijos koeficientą. Veibulo AFT bei Kokso PH modeliai pasirodo labai panašiai, o RSF beveik visoje laiko skalėje pranašesnis.



1 pav. Dinaminis AUC apmokymo aibėje.

5 Išvados

Pritaikius tris išgyvenamumo analizės metodus galima teigti, jog toks požiūris personalo kaitai analizuoti yra tinkamas. Atsitiktiniai išgyvenimo miškai lenkia Kokso PH ir Veibulo AFT modelius pagal konkordancijos indeksą. Tik RSF modelis apmokymo aibėje fiksuoja rezultatą, aukštesnį nei 0,8 ir vienintelis testavimo aibėje viršija 0,6. Iš dinaminio AUC išvados tokios pat – RSF beveik visame laiko intervale pranašesnis. Papildomas šio modelio privalumas tas, kad jį ir pritaikyti yra lengviausia.

Literatūra

[1] Alamsyah, A., Salma, N. *A Comparative Study of Employee Churn Prediction Model*. 4th International Conference on Science and Technology (ICST), 2018, pp. 1-4.

[2] Collett, D. *Modelling Survival Data in Medical Research*. British Actuarial Journal, 1995, **1**(2).

[3] Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., Qiang, B. *RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis*. Web Information Systems Engineering – WISE 2020. Lecture Notes in Computer Science, 2020, **12343**.

[4] Kamarudin, A.N., Cox, T., Kolamunnage-Dona, R. *Time-dependent ROC curve analysis in medical research: current methods and applications.* BMC Med Res Methodol, 2017, **17**, 53.

[5] Kleinbaum, D., Klein, M. *The Cox Proportional Hazards Model and Its Characteristics.* Survival Analysis: A Self-Learning Text, 2005, 2, pp. 97-159.

[6] Kleinbaum, D., Klein, M. *Parametric Survival Models.* Survival Analysis: A Self-Learning Text, 2005, 2, pp. 289-361.

[7] Morita, J. G., Lee, T. W., Mowday, R. T. *The regression-analog to survival analysis: A selected application to turnover research*. Academy of Management Journal, 1993, **36**(6), pp. 1430-1464.