Samdomųjų darbuotojų darbo užmokesčio prognozavimas taikant mašininio mokymosi metodus

Dominykas Vilčinskas, Rūta Levulienė

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

Santrauka. Darbo užmokesčio modeliavimas leidžia įvertinti įvairių rodiklių daromą įtaką pajamoms, gali padėti įmonėms nustatyti optimalų atlyginimą. Be to, didelis skirtumas tarp darbdavio deklaruojamo ir modelio prognozuojamo darbo užmokesčio reikšmių galėtų identifikuoti galimus sukčiavimo atvejus, kai siekiama išvengti mokestinių prievolių sąmoningai nurodant mažesnį atlyginimą. Šiame darbe nagrinėjami gyventojų užimtumo statistinio tyrimo duomenys ir taikomi skirtingi mašininio mokymosi modeliai atlyginimo prognozavimui. Rezultatai rodo, kad darbo užmokestį tiksliausiai prognozuoja medžių ansambliu paremtas gradientinio auginimo metodas – XGBoost.

Raktiniai žodžiai: Darbo užmokestis, prognozavimas, mašininis mokymasis, XGBoost.

1 Įvadas

Darbo užmokesčio prognozavimas leidžia ne tik įvertinti skirtingų rodiklių įtaką atlyginimams, bet gali suteikti ir papildomą praktinę naudą. Įmonės, naudodamos modelius, galėtų tiksliau nustatyti konkurencingą ar optimalų atlyginimą, atsižvelgiant į rinkos sąlygas bei darbuotojų kvalifikaciją. Be to, valstybės institucijoms darbo užmokesčio modeliai gali padėti identifikuoti galimus sukčiavimo atvejus. Pavyzdžiui, kai oficialiai deklaruojamas mažesnis atlyginimas nei yra iš tikrųjų, siekiant išvengti mokestinių prievolių. Atlyginimo, kaip ir kitų ekonominių rodiklių prognozei naudojami statistiniai metodai pavyzdžiui, tiesinė regresija [1], ar jos modifikacijos: LASSO, RIDGE [2]. Tačiau vis dažniau taikomi mašininio mokymosi modeliai, pavyzdžiui atsitiktiniai miškai [3]. Pastaruoju metu, įvairiose srityse, siekiant pasiekti kuo didesnį prognozavimo tikslumą, plačiai taikomi medžiais paremti gradien-

DOI: https://doi.org/10.15388/LMITT.2025.28

tinio auginimo metodai. Atlikti tyrimai, kuriuose parodyta, kad standartiniams - struktūrizuotiems duomenims, gradientinio medžių auginimo metodai dažnai lenkia bet kuriuos kitus, įskaitant ir giliojo mokymosi modelius [4,5]. Pagrindinis šio tyrimo tikslas – pritaikyti gradientinio auginimo metodą darbo užmokesčio prognozavimui ir palyginti jo tikslumą su kitais dažnai naudojamais regresijos modeliais.

2 Duomenys

Darbe naudojami gyventojų užimtumo statistinio tyrimo duomenys. Tačiau apklausoje maža dalis respondentų nurodo savo pajamas. Todėl apklausos duomenys susiejami su kitu šaltiniu – naudojame požymius iš apklausos kaip prediktorius, o darbo užmokesčio reikšmės ir kiti darbovietės kintamieji gaunami iš administracinio šaltinio.

Nagrinėjami 2024 metų samdomųjų darbuotojų duomenys. Atlikus pradinę analizę, tyrimui buvo atrinkti tokie asmens požymiai kaip išsilavinimas, lytis, profesija, baigtų studijų sritis, amžius, stažas darbovietėje ir dirbamų valandų skaičius, o taip pat papildomi darbovietės kintamieji: darbuotojų skaičius, apskritis, ekonominės veiklos grupė ir vidutinis atlyginimas.

Padalinus duomenis į mokymo ir testavimo aibes santykiu 80:20, mokymo aibę sudaro 4785 įrašai, o testavimo – 1196. Be to, kategoriniai kintamieji perkoduojami pseudokintamaisiais.

3 Metodologija

Tyrime naudojami mašininio mokymosi regresijos modeliai: atraminių vektorių regresorius (AVR) (angl. support vector machine) [6], medžių ansamblių metodai: atsitiktiniai miškai (AM) (angl. random forest) [7] ir XGBoost [8]. Tiesinė regresija taikoma kaip bazinis modelis palyginimui. Siekiant sumažinti išsiskiriančių stebėjimų įtaką rezultatams, modeliuojamos logaritmuotos darbo užmokesčio reikšmės. Ši transformacija yra monotoninė, todėl tinkama interpretavimo atveju, o taip pat imant prognozuotų reikšmių eksponentę, gaunama pradinių darbo užmokesčio reikšmių prognozė. Mašininio mokymosi metodams optimalūs parametrai parenkami taikant gardelės metodą su kryžmine patikra. AVR atveju nagrinėjami skirtingi branduoliai, reguliarizacijos, branduolio koeficiento ir kiti parametrai. AM modelyje optimizuojami tokie hiperparametrai: medžių skaičius, maksi-

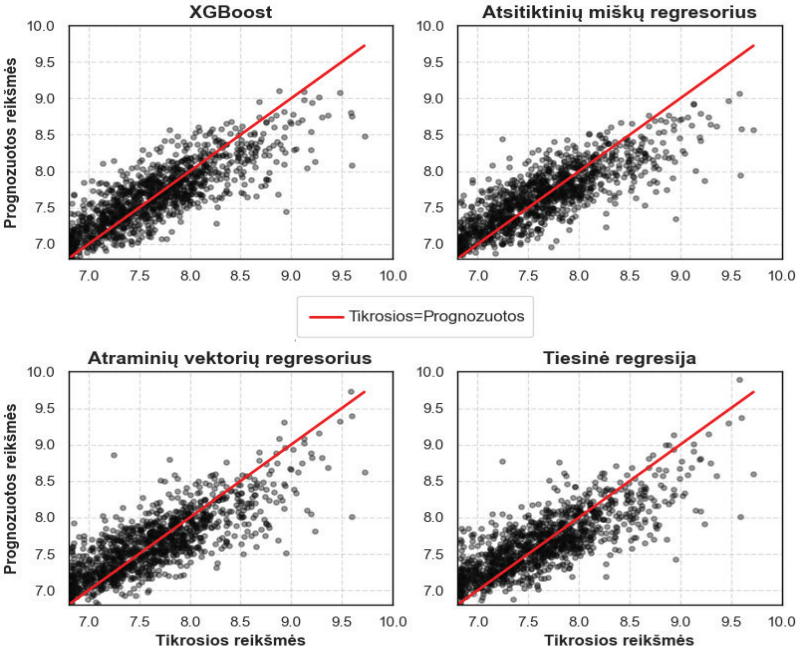
malus medžių gylis, mažiausias stebinių skaičius lape ir dalijime. XGBoost atveju nagrinėjami tokie parametrai kaip medžių skaičius, maksimalus medžių gylis, mokymosi greitis, reguliarizacijos parametras ir kiti. Modelių hiperparametrų parinkimas ir apmokymas atliekamas mokymo aibėje, o jų tinkamumas vertinimas testavimo aibėje pagal keturias metrikas: vidutinė kvadratinė paklaida (MSE), vidutinė absoliutinė paklaida (MAE), Pirsono (angl. Pearson) koreliacijos koeficientas (ρ) ir determinacijos koeficientas (R^2).

4 Rezultatai

Toliau pateikiamas modelių tinkamumo vertinimas testavimo aibėje pagal anksčiau nurodytas metrikas (žr. 1 lentelę ir 1 pav.). Raudonos linijos 1 pav. vizualiai parodo modelių tinkamumą, t.y. linija vaizduoja situaciją, kai prognozuojamos reikšmės idealiai sutampa su tikrosiomis, kuo arčiau taškai šios linijos, tuo modelis tinkamesnis. Pastebime, jog mašininio mokymosi metodų atveju gauname geresnius rezultatus nei naudojant tiesinę regresiją. Taip yra todėl, nes šie metodai gali įvertinti sudėtingesnius, netiesinius sąryšius tarp kintamųjų. Tiksliausiai prognozuojama naudojant XGBoost modelį, kurio visų metrikų reikšmės testavimo aibėje yra geriausios. Vis dėlto nė vienas modelis neprognozuoja darbo užmokesčio labai tiksliai, kas gali būti susiję su nepakankamai informatyviais požymiais. Pavyzdžiui, darbo stažas apibūdino tik patirtį dabartinėje darbovietėje, bet ne visą profesinę patirtį. Be to, nebuvo turima tiksli darbovietės vietovė – naudojant tik apskrities informaciją, gali būti neišskirtos didžiųjų miestų darbovietės, kuriose atlyginimai dažniausiai yra didesni.

1 lentelė. Modelių tinkamumo vertinimas testavimo aibėje

Modelis	MSE	MAE	ρ	R^2
XGBoost	**0,0933**	**0,2320**	**0,8279**	**0,6840**
Atsitiktiniai miškai	0,0993	0,2387	0,8165	0,6637
Atraminių vektorių regresorius	0,1073	0,2497	0,7979	0,6366
Tiesinė regresija	0,1103	0,2540	0,7922	0,6263



1 pav. Modelių tikrosios ir prognozuotos reikšmės testavimo aibėje

5 Išvados

Remiantis gautais rezultatais, galima teigti, kad mašininio mokymosi metodai yra tinkamesni darbo užmokesčio prognozavimui nei tiesinė regresija. Geriausią tikslumą pasiekė medžiais paremtas gradientinio auginimo metodas XGBoost, taigi šį modelį siūlome naudoti darbo užmokesčio prognozavimui. Norint pasiekti didesnį tikslumą reikėtų įtraukti daugiau kintamųjų, tačiau tam reikėtų papildomų duomenų, pavyzdžiui, apklausų metu rinkti detalesnę informaciją, įtraukiant klausimus dėl visos profesinės patirties, geografinės vietovės.

Literatūra

[1] Anuj More, et al. (2021). „PREDICT-NATION Skills Based Salary Prediction for Freshers," Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021).

[2] Guangqi Wang. (2022). „Employee Salaries Analysis and Prediction with Machine Learning," International Conference on Machine Learning and Intelligent Systems Engineering (MLISE), 2022.
[3] Babasaheb S. Satpute, et al. (2023). ,,Machine Learnig Approach for Prediction of Employee Salary using Demographic Information with Experience" 4th IEEE Global Conference for Advancement in Technology, 1-5.
[4] Tim Januschowski, et al. (2022). „Forecasting with trees," International Journal of Forecasting, 38, 1473-1481.
[5] Ravid Shwartz-Ziv & Amitai Armon. (2022). ,,Tabular data: Deep learning is not all you need." Information Fusion, 81, 84-90.
[6] Harris Drucker, et al. (1997). Support vector regression machines. Advances in Neural Information Processing Systems. 28. 779-784.
[7] Leo Breiman. (2001) Random Forests. Machine Learning 45, 5–32, 2001.
[8] Tianqi Chen & Carlos Guestrin (2016). XGBoost: „A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.