ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

ԱՐՍԻՆԵ ԱՐՍԵՆԻ ՍԱՌԻԿՅԱՆ

ՀԱՐԿԱՅԻՆ ԵԿԱՄՈՒՏՆԵՐԻ ԿԱՆԽԱՏԵՍՈՒՄՆ ԱՅԼԸՆՏՐԱՆՔԱՅԻՆ ՏՎՅԱԼՆԵՐԻ ՄԻՋՈՑՈՎ (ՀՀ ՕՐԻՆԱԿՈՎ)

Ը.00.08 - «Տնտեսության մաթեմատիկական մոդելավորում» մասնագիտությամբ տնտեսագիտության թեկնածուի գիտական աստիճանի հայցման ատենախոսության

ՍԵՂՄԱԳԻՐ

Երևան 2024

Ատենախոսության թեման հաստատվել է Հայաստանի պետական տնտեսագիտական համալսարանում:

Գիտական ղեկավար՝	Տնտեսագիտության դոկտոր, պրոֆեսոր՝ Թավադյան Աշոտ Աղասու
Պաշտոնական ընդդիմախոսներ՝	Տեխնիկական գիտությունների դոկտոր, պրոֆեսոր՝ Առաքելյան Արամ Հմայակի
	Տնտեսագիտության թեկնածու Հակոբյան Ենոք Բենիամինի
Առաջատար կազմակերպություն՝	Հայաստանի ազգային պոլիտեխնիկական համալսարան(ՀԱՊՀ)

Ատենախոսության պաշտպանությունը կայանալու է 2024թ. սեպտեմբերի 13-ին՝ ժամը 13:30-ին, Երևանի պետական համալսարանում գործող, ՀՀ ԲԿԳԿ-ի՝ Տնտեսագիտության 015 մասնագիտական խորհրդի նիստում:

Ատենախոսությանը կարելի է ծանոթանալ Երևանի պետական համալսարանի գրադարանում:

Սեղմագիրն առաքված է 2024 թ. հուլիսի 9-ին:

Մասնագիտական խորհրդի
գիտական քարտուղար
տնտեսագիտության թեկնածու, դոցենտ Ա. Հ. Հակոբջանյան

Ատենախոսության թեմայի արդիականությունը: Հարկային եկամուտների՝ այդ թվում թաքցված եկամուտների, կանխատեսումն առաջնային կարևորություն ունի կառավարությունների համար՝ հարկաբյուջետային քաղաքականության, բյուջետային գործընթացի շրջանակներում հարկային եկամուտների և ծախսերի ծրագրավորման, ինչպես նաև ընդհանուր տնտեսական կայունության վրա ունեցած զգալի հետևանքների պատճառով: Հարկային պարտավորությունների կատարումից խուսափումը խաթարում է հարկային համակարգի ամբողջականությունը, ինչը նվազեցնում է հանրային վստահությունը, և հանգեցնում է եկամուտների կորստի:

Հարկային եկամուտների հավաքագրման և հարկային իրավախախտումների միջև կապը բարդ է և փոխկապակցված, իսկ եկամուտների հստակ կանխատեսումը կախված է եկամուտների արտահոսքի հնարավոր աղբյուրների, այդ թվում՝ հարկերից խուսափելու և խարդախության դեպքերի հայտնաբերումից և վերացումից: Ավելին, չբացահայտված հարկային իրավախախտումները կարող են հարկային բեռի անհավասար բաշխման պատճառ դառնալ, ընդ որում կորցրած եկամտի փոխհատուցման բեռն ազնիվ հարկ վճարողների համար ավելի մեծ կլինի, ինչը կխաթարի հարկային համակարգի համաչափության ու հավասարության սկզբունքները:

Ատենախոսության թեմայի արդիականությունը պայմանավորված է վերոշարադրված հանգամանքներով, ինչից էլ բխում է այդ գործընթացների գնահատման և կանխատեսման անհրաժեշտությունը:

Ատենախոսության նպատակը և խնդիրները: Հետազոտության նպատակն է մշակել Հայաստանում հարկային եկամուտների կանխատեսման նոր մեթոդաբանություն՝ կիրառելով ավանդական և այլընտրանքային տվյալների աղբյուրների վրա հիմնված Մեքենայական ուսուցման (այսուհետ՝ ՄՈՒ) մոդելներ: Հարկային հաշվետվությունների տվյալներին ավանդական ապավինումն ընդլայնվում է այլընտրանքային տվյալների հաշվին: Դրանք այն տվյալներն են, որոնք ներկայումս չեն օգտագործվում ՀՀ պետական եկամուտների կոմիտեի (այսուհետ՝ ՊԵԿ) կողմից՝ կարևորություն տալով, մասնավորապես, տեղաբաշխման հետևանքներին և հսկիչ դրամարկղային մեքենաների (այսուհետ՝ ՀԴՄ) բարձր հաճախականության տվյալներին: Հետազոտությունն ուղղված է հարկերից խուսափման դեպքերի հայտնաբերման և եկամուտների գնահատման ճշգրտության բարձրացմանը ՄՈՒ ալգորիթմների կիրառմամբ և նշված այլընտրանքային տվյալների ներառմամբ:

Վերոնշյալ նպատակին հասնելու համար առաջադրվել և լուծվել են հետևյալ խնդիրները՝

- ճշգրտությունն ու հուսալիությունն ապահովելու նպատակով տարբեր աղբյուրներից տվյալների, այդ թվում՝ ՀԴՄ-ների կտրոնների բարձր հաճախականության և աշխարհագրական դիրքի տվյալների սահմանում, հավաքագրում և մաքրում,

- կանխատեսումներ իրականացնելու նպատակով ՄՈՒ մի շարք մոդելների օգտագործում, ինչպիսիք են լոգիստիկ ռեգրեսիան, որոշումների ծառերը, պատահական անտառները, գրադիենտ խթանումը և դրա տարատեսակ LightGBM-ը,
- կանխատեսումների ընդհանրականացումն ու հուսալիությունն ապահովելու համար ROC-AUC չափանիշների հիման վրա մոդելների գնահատում և ընտրություն,
- տվյալների բազայում առկա բոլոր հարկ վճարողների համար հարկերից խուսափելու հավանականությունը գնահատելու նպատակով լավագույն մոդելի ստացում,
- առավել նույնական ընկերություններին հայտնաբերելու նպատակով KNN-ի ներդրում՝ հիմնված ֆինանսական տվյալների վրա,
- համեմատելով փաստացի հավաքագրված եկամուտները կանխատեսվող եկամուտների հետ՝ թաքցված հարկային եկամուտների հնարավոր վճարումների գնահատում՝ եկամուտների կորուստների բացահայտման և նվազեցման նպատակով:

Ատենախոսության օբյեկտը և առարկան: Ատենախոսության հետազոտության օբյեկտը հարկ վճարողների գործունեությունն է, իսկ հետազոտության առարկան՝ թաքցված հարկային եկամուտների բացահայտումը և պոտենցիալ հարկային եկամուտների գնահատումը:

Ատենախոսության տեսական, տեղեկատվական և մեթոդական հիմքերը: Առաջադրված խնդիրների լուծման համար տեսական հիմք են ծառայել հայրենական ու արտասահմանյան մի շարք հեղինակների աշխատություններ: Ատենախոսության համար տեղեկատվական հիմք են ծառայել ՊԵԿ-ի տվյալները՝ որոնք հասանելի են դարձել հեղինակի կողմից իրականացված հարցումների միջոցով: Կիրառվել են մի շարք տվյալների վերափոխման և պատրաստման մոտեցումներ և փորձարկվել են տարբերը ՄՈՒ մոդելներ: Առաջադրված խնդիրների լուծման, հաշվարկների և գծապատկերների կառուցման համար օգտագործվել են Python ծրագրավորման լեզուն և Microsoft Excel ծրագիրը:

Ատենախոսության հիմնական գիտական արդյունքներն ու նորույթը: Ատենախոսության շրջանակներում կատարված ուսումնասիրությունների ու վերլուծությունների հիմնական գիտական արդյունքները և գիտական նորույթը հետևյալն են.

– Հստակ կանխատեսումներ իրականացնելու համար փորձարկվել և բացահայտվել են բարձր հաճախականության տվյալների փոփոխականների ստեղծման և օգտագործման մեթոդներ:
– Հարկ վճարողների՝ ստուգումներով պայմանավորված արձագանքները բնորոշելու համար, գնահատվել է տեղաբաշխման հետևանքների ազդեցությունը:
– Վերլուծվել են հարկերից խուսափելու դեպքերի հայտնաբերման ՄՈՒ մի շարք մոդելներ և մանրակրկիտ գնահատման միջոցով բացահայտվել է հարկերից խուսափելու դեպքերի ամենաարդյունավետ մոդելը:

– Գնահատվել են թաքցված հնարավոր հարկային եկամուտները՝ օգտագործելով հարկերից խուսափելու հավանականությունները և KNN մոդելը, որի արդյունքում բացահայտվել է, որ առաջարկվող մոտեցումը կարող է վերականգնել առավելագույնը 1.3 անգամ ավելի շատ հարկային վճարումներ՝ օգտագործելով ստուգման ենթարկված ընկերությունների 72%-ը:

Ատենախոսության արդյունքների տեսական և գործնական նշանակությունը: Առաջարկվող մեթոդաբանությունները և մոդելները ՊԵԿ-ի տվյալների հավաքագրման և նախնական մշակման ռազմավարությունների առումով կարող ենք համարել առանցքային առաջընթաց են՝ ուղղված Բիզնեսի վերլուծության (BI) կարողությունների ընդլայնմանը: Հարկերից խուսափման դեպքերի հայտնաբերման համար ՄՈՒ մոդելների ներդրման արդյունքում ՊԵԿ-ն ակնկալում է հարկերից խուսափելու դեմ պայքարի ավելի պարզ և արդյունավետ մոտեցում: Ավելին, օրենսդրական կարգավորումն արդեն ընդունված է, և նպատակային համալիր հարկային ստուգումների որոշակի տոկոսը[1] հնարավոր է իրականացնել առաջարկվող ՄՈՒ մոդելի հիման վրա ընտրվող հարկ վճարողների մոտ: Բացի այդ, առկա համակարգերում անխափան ինտեգրման նպատակով հարկային փորձագետների կողմից ներկայումս այլ բաղադրիչներ ևս ենթարկվում են ներքին գնահատման՝ տվյալների համապարփակ վերլուծության և որոշումների կայացման գործընթացների համար: Ըստ էության, այս նախաձեռնությունները ենթադրում են ՊԵԿ-ի գործառնական ենթակառուցվածքի արդիականացմանն ու ամրապնդմանն ուղղված համատեղ ջանքեր՝ հնարավորություն տալով հաղթահարել հարկերից խուսափելու դեպքերի հայտնաբերման և եկամուտների պաշտպանության աճող մարտահրավերները:

Ատենախոսության արդյունքների փորձարկումը և հրապարակումները: Ատենախոսության հրապարակումները արտացոլված են «Կիրառական արհեստական բանականություն» (Applied Artificial Intelligence) և «Զարգացման ուսումնասիրություններ ամսագիր» (The Journal of Development Studies) Սկոպուսի ցանկի միջազգային ամսագրերում, ինչպես նաև «Այլընտրանք», «Տարածաշրջան և աշխարհ» գիտական հանդեսներում:

Ատենախոսության կառուցվածքը և ծավալը: Ատենախոսությունը կազմված է ներածությունից, երեք գլուխներից, եզրակացությունից, օգտագործված գրականության ցանկից (73 անուն) և հավելվածներից: Ատենախոսությունը կազմված է 121 էջից:

[1] https://www.arlis.am/DocumentView.aspx?docid=191233

Ներածությունում ներկայացվել է թեմայի հետազոտության նպատակները, տեսական բազան, մեթոդաբանական հիմքերը, գիտական նորարարությունները, հիմնական եզրակացությունները, ինչպես նաև կիրառական գիտական և գործնական մեթոդները:

Ատենախոսության առաջին **«Հարկային եկամուտների կանխատեսման և տվյալների այլընտրանքային աղբյուրներին վերաբերող տեսական բազա»** գլխում մանրամասն անդրադարձ է կատարվում հարկային եկամուտների կանխատեսման և հարկային խարդախության հետ փոխկապակցվածության կարևորությանը, տվյալների այլընտրանքային աղբյուրների սահմանմանը և տարբեր տեսակներին, ինչպես նաև այն հարցին, թե ինչու են դրանք կարևոր, ինչպես նաև իրականացվել է ՀՀ-ի ինստիտուցիոնալ կառուցվածքի ուսումնասիրություն:

Հարկային եկամուտների կանխատեսումը վճռական նշանակություն ունի կառավարության ֆինանսական պլանավորման և քաղաքականության մշակման համար, որի նպատակն է ճշգրիտ կանխատեսել ապագա հարկերի հավաքագրումները՝ տնտեսական անարդյունավետությունից խուսափելու համար: Ճշգրիտ կանխատեսումներն օգնում են բացահայտել հարկային եկամուտների փոփոխությունների ոլորտները, օգնում են քաղաքականության վերլուծության և հարկաբյուջետային պլանավորման հարցում: Սակայն, հարկային եկամուտների կանխատեսումը բավականին բարդ է, արտաքին գործոնների, հարկաբյուջետային մշակույթի և հարկերի պահանջների չկատարման պատճառով[2]: Ավելին՝ հարկային ճեղքի առկայության պատճառով պոտենցիալ հարկային եկամուտների զգալի մասը մնում է չիրացված[3]: Այս ճեղքը հարկային օրենքների և կանոնակարգերի համաձայն հավաքագրման ենթակա հարկի գումարի և հարկային մարմինների կողմից փաստացի հավաքագրված գումարի միջև տարբերություն է: Ավելի պարզ, դա հարկ վճարողների կողմից վճարման ենթակա գումարի և կառավարության կողմից փաստացի ստացված գումարի միջև՝ հիմնականում հարկային եկամուտները թաքցնելու պատճառով, առաջացած տարբերությունն է:

Առկա կանխատեսման մեթոդները ոչ ամբողջական են և չեն արտացոլում տնտեսական[4], քաղաքական և իրավական միջավայրի բարդությունները[5]: Այդ մարտահրավերների հաղթահարման հիմնական մոտեցումներից մեկը տվյալների ոչ ավանդական աղբյուրների և դրանց մշակման մեթոդների օգտագործումն է[6]:

2 Chatagny, F., & Siliverstovs, B. (2013). Rationality of Direct Tax Revenue Forecasts under Asymmetric Losses: Evidence from Swiss Cantons. European Economics: Political Economy & Public Economics eJournal.

3 Murphy, R. (2021). Reappraising the Tax Gap. Combating Fiscal Fraud and Empowering Regulators, 61–74. P. 51

4 Heinemann, A., Kotina, H., & Stepura, M. (2017). An Interdisciplinary View on Tax Revenue Estimates and Forecasts and its Impacts on a Multilevel Public Budget System, 127-128.

5 Xiong, F., Chapple, L., & Yin, H. (2018). The use of social media to detect corporate fraud: A case study approach. Business Horizons, p. 4-8

6 Warner, G., Wijesinghe, S., Marques, U., Badar, O., Rosen, J., Hemberg, E., & O'Reilly, U. (2014). Modeling tax evasion with genetic algorithms. Economics of Governance, 16, 165-178.

Ներկայումս հնարավոր է գտնել բազմաթիվ տվյալներ, որոնք կարող են օգտագործվել թաքցված եկամուտների կանխատեսման նպատակով: Վերջիններս հայտնի են որպես «այլընտրանքային տվյալներ»՝ տվյալներ, որոնք այդ պահին չեն օգտագործվում հարկային մարմնի կողմից՝ հարկերից խուսափելու հավանականությունը կանխատեսելու համար:

Սույն ատենախոսության շրջանակներում հիմնականում կարևորվում են երկու տեսակի տվյալներ, այն է՝ ՀԴՄ-ների կտրոններից ստացվող բարձր հաճախականության և վարքագծային տվյալների, որոնք ցույց են տալիս հարկ վճարողների կողմից խարդախություն կատարելու դրդապատճառները: ՀԴՄ-ների կտրոններից ստացվող բարձր հաճախականության տվյալները թույլ են տալիս հարկային մարմիններին իրական ժամանակում պատկերացում կազմել տնտեսական գործունեության և հարկ վճարողների վարքագծի մասին[7]: Այս տվյալները թույլ են տալիս համապատասխան մարմիններին հայտնաբերել նախազգուշացնող նշաններ, ինչպիսիք են անկանոն գործարքները կամ հակասական հաշվետվությունները, որոնք կարող են վկայել հարկերից խուսափման հնարավոր դեպքերի մասին: Միևնույն ժամանակ, հարևան ընկերություններում իրականացվող ստուգումը նույնպես, կարող է ազդել ստուգման չենթարկված ընկերությունների վարքագծի վրա՝ հանգեցնելով վարքագծային փոփոխությունների[8]: Հետազոտության շրջանակներում առաջարկվում է երկու կանխավարկած, որոնք բացատրում են հարկ վճարողների արձագանքները՝ ստուգումների ընթացքում պահանջների կատարման բարելավման ակնկալիքը կամ ստուգումից հետո պահանջների կատարման նվազումը: Բացի այդ, ուսումնասիրությունները ցույց են տալիս, որ որևէ տարածքում ստուգումների քանակի ավելացումը կարող է ազդել հարևան տարածքներում հարկային պահանջների կատարման մակարդակի վրա: Սա վկայում է այն մասին, որ հարկ վճարողների վարքագիծը հասկանալու համար հարևան տարածքների տվյալները հաշվի առնելը նշանակալի կարևորություն ունի: Թեև նման տվյալների օգտագործումն ինքնին մարտահրավեր է, այնուամենայնիվ, այն արժեքավոր պատկերացում է տալիս հարկային խարդախության դեպքերի կանխատեսման մասով՝ լրացնելով ավանդական տվյալների աղբյուրները իրական ժամանակի տնտեսական և վարքագծային ազդանշաններով:

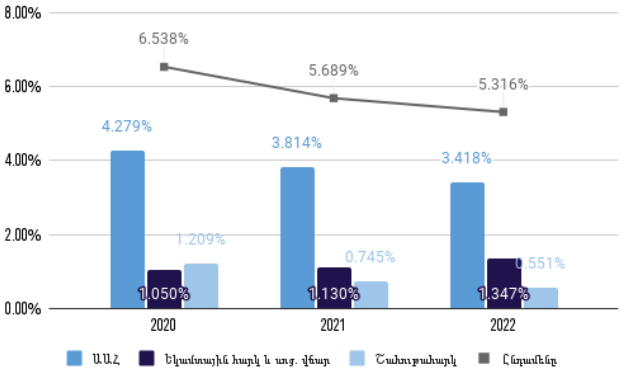
ՀՀ տնտեսությունում հարկային եկամուտների կանխատեսումները հիմնված են տնտեսական աճի ցուցանիշների, հարկային վարչարարության գործոնների և օրենսդրական փոփոխությունների վրա: Սկզբում կանխատեսումը սկսվում է տնտեսական աճի կանխատեսմամբ, որի հիման վրա այնուհետև գնահատվում է ՀՆԱ-ն, որից հետո գնահատվում է հարկեր-ՀՆԱ հարաբերակցությունը: Նպատակային վարչարարություն իրականացնելու և ռեսուրսների արդյունավետ

7 Gemmell, N., & Hasseldine, J. (2014). Taxpayers' Behavioural Responses and Measures of Tax Compliance "Gaps": A Critique and a New Measure. Fiscal Studies, 35(3), 275–296.

8 Lediga, C., Riedel, N., & Strohmaier, K. (2020). Tax enforcement spillovers – evidence from South Africa. SSRN Electronic Journal. p.12-16

օգտագործումն ապահովելու նպատակով ՊԵԿ-ը գնահատում է հարկային ճեղքը: Գնահատված հարկային ճեղքը բաղկացած է երկու բաղադրիչից՝ կարգապահական ճեղք (օրենսդրության պահանջների կատարման մասով ճեղքը) և քաղաքականության մասով ճեղքը (հարկային պոտենցիալի և հավաքագրված հարկերի տարբերության այն հատվածը, որն առաջանում է վարած հարկային քաղաքականության արդյունքում): Ելնելով ճեղքերի գնահատման արդյունքներից՝ ՊԵԿ-ն ընտրում է առավել ռիսկային (հարկային կարգապահության առավել ցածր մակարդակ ունեցող) խմբերը կամ ոլորտները, մշակում և իրականացնում է հարկ վճարողների կարգապահության բարելավմանն ուղղված ծրագրեր՝ կարգապահության 4 հիմնական բաղադրիչներով (գրանցում, ժամանակին հաշվետվությունների ներկայացում, պարտավորությունների ճշգրիտ հայտարարագրում և հարկերի վճարում) կարգապահությունը բարելավելու և կարգապահական ճեղքը նվազեցնելու նպատակով[9]:

Գծապատկեր 1-ում ներկայացված է յուրաքանչյուր հարկատեսակի մասով հարկային ճեղք/ՀՆԱ հարաբերակցության ամփոփ վիճակագրությունը 2020-2022 թվականների համար:



Գծապատկեր 1. 2020թ.-2022թ. հարկային ճեղքի և երկրի ՀՆԱ-ի հարաբերակցություն:[10]

Երկրորդ **«ՄՈՒ-ի վրա հիմնված հարկային խարդախության դեպքերի կանխատեսման մեջ այլընտրանքային տվյալների ինտեգրման մեթոդաբանական հիմունքներ»** գլուխը վերաբերում է հարկային խարդախության դեպքերի կանխատեսման մեջ օգտագործվող ՄՈՒ մոդելների տեսական

9 Հարկային պահանջների կատարման ճեղքի ՊԵԿ-ի գնահատականը

10 https://www.arlis.am/Annexes/7/2024_N266hav.pdf

հիմունքներին, ինչպես նաև տվյալների այլընտրանքային աղբյուրների ինտեգրման եղանակներին:

Հարկային խարդախությունների դրսևորման բազմազան եղանակներով ու դրանց մշտապես փոփոխվող բնույթով պայմանավորված՝ հաճախ այդ խարդախությունների հայտնաբերման համար կիրառվող ավանդական մեթոդները բավականաչափ արդյունավետ չեն: ՄՈՒՆ առաջարկում է լուծում, որը շատ արդյունավետ է տվյալների լայնածավալ վերլուծության, նշանների ճանաչման և անոմալիաների հայտնաբերման գործում[11]: ՄՈՒ ալգորիթմները կարող են բացահայտել խարդախության անգամ փոքր նշանները, կանխատեսել նոր միտումները և հետաքննության համար առաջնահերթություն տալ բարձր ռիսկային դեպքերին:

Գոյություն ունի հարկային խարդախության դեպքերի կանխատեսման մեջ կիրառվող երկու հիմնական ՄՈՒ կատեգորիա՝ վերահսկվող և չվերահսկվող ուսուցում: Վերահսկվող ուսուցումը տարածված մեթոդ է հարկային խարդախության դեպքերի հայտնաբերման մոդելներում և ներառում է մոդելին հայտնի տվյալների և պիտակների փոխանցում, նաև թույլ է տալիս ուսումնասիրել նշանները և դրանց հիման վրա կանխատեսումներ կատարել ուսուցման փուլում: Մյուս կողմից, չվերահսկվող ուսուցումը ներառում է տվյալների մեջ նշանների որոնում՝ առանց նախապես սահմանված պիտակների: Այս մեթոդի նպատակն է բացահայտել մուտքային տվյալների աննկատ կառուցվածքները կամ թաքնված օրինաչափությունները՝ առանց հստակ արձագանքների[12]:

Այս գլխում քննարկվում են նաև վերահսկվող ուսուցման տարածված կիրառելիություն ունեցող կանխատեսող մոդելավորման հիմնական տեսակները: Քննարկվող մոդելները ներառում են լոգիստիկ ռեգրեսիա, որոշումների ծառ, պատահական անտառ, գրադիենտ խթանում և դրա հատուկ LightGBM տեսակը, դրանց ներքին մեխանիզմներն ու տարբերությունները: Այս մոդելներից յուրաքանչյուրը փորձարկման փուլում ենթարկվում է մանրակրկիտ փորձաքննության, որի վերջնական նպատակը թաքցված եկամուտների դեպքերի հայտնաբերման ամենաարդյունավետ մոդելի որոշումն է:

Գրադիենտի խթանումը ՄՈՒ գործիք է, որը համախմբում է թույլ ուսուցման մոդելները, սովորաբար՝ որոշման ծառերը, որպեսզի ստեղծվի առավել ուժեղ կանխատեսման սովորելով նախկին սխալներից և ճշգրտելով մոդելի բարդությունը՝ մոդելի հատկություններին մոդել՝ ավելորդ հարմարեցումը կանխելուն ուղղված ուսուցման արդյունավետության պարամետրեր կիրառելու միջոցով: Տվյալների առանձին առանձին օրինաչափությունների վրա կենտրոնանալով՝ հնարավոր է նվազեցնել մոդելի բարդությունը՝ միաժամանակ հարմարեցնելով այն բարդ

11 Abrantes, P. C., & Ferraz, F. (2016). Big data applied to tax evasion detection: A systematic review. 2016 International Conference on Computational Science and Computational Intelligence (CSCI). p. 436-437

12 James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). An introduction to statistical learning: With applications in R. p.127-151,373-285

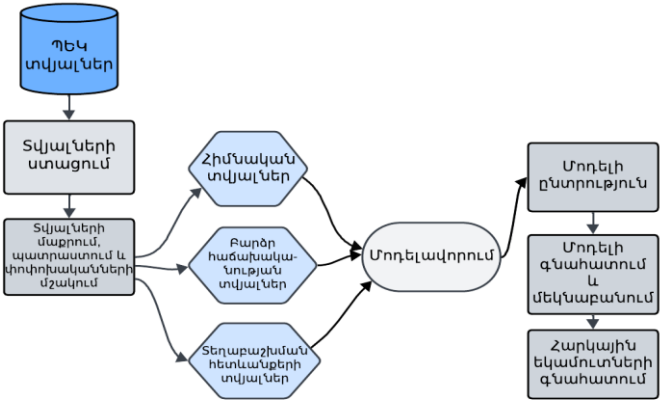
հարաբերություններին: Կորուստների բարդ ֆունկցիան ուղղակի նվազեցնելու փոխարեն այն պարզեցնում է օպտիմալացման գործընթացը[13]:

$$(\beta_m, a_m) = argmin \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + \beta h(x_i; a))$$

$$F_m(x) = F_{m-1}(x) + \beta_m h(x; a_m)$$

որտեղ, $F_{m-1}(x_i)$-ը մոդելի կանխատեսումն է m-1 կրկնություններից հետո, $L(y_i, F(x))$-ը կորստի ֆունկցիան է, $h(x; a_m)$-թույլ մոդելն է m-րդ կրկնության ժամանակ, a_m-ը թույլ մոդելի պարամետրերն են, β_m-ը թույլ մոդելի կշիռներներն են:

Երրորդ **«Այլընտրանքային տվյալների ինտեգրումը հարկային խարդախության դեպքերի կանխատեսման մեջ. կատարողականի գնահատում և գործնական եզրակացություններ»** գլուխը ներառում է ՊԵԿ-ից ստացված իրական տվյալների վերլուծություն և ուսումնասիրություն, ներդրված փոփոխականների մշակում, ՄՈՒ տարբեր մոդելների համեմատություն, ինչպես նաև ակնկալվող հարկային տուգանքների վերջնական գնահատում: Գործընթացի սխեման ներկայացված է Գծապատկեր 2-ում:



Գծապատկեր 2: ՄՈՒ մոդելի գործընթացների շղթան[14]

Տվյալները պահվում են ՊԵԿ-ի տվյալների բազայում՝ տարբեր աղյուսակներում և հաշվետվություններում, և դրանք ստանալու համար հարկավոր է ՊԵԿ-ի հատուկ թույլտվությունը: Տարբեր ձևաչափերի միջև տվյալների կառուցվածքի և որակի

13 Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5). p. 1192

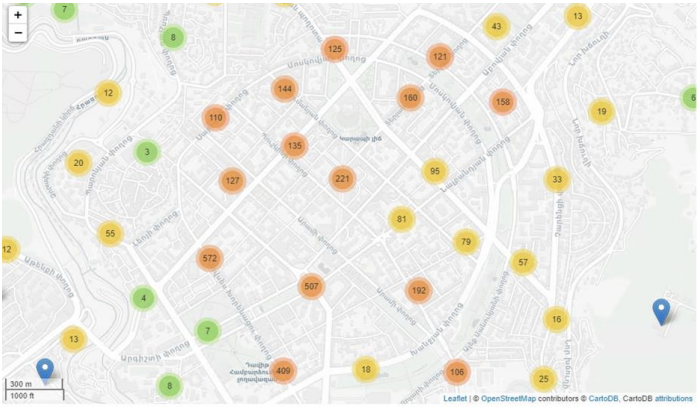
14 Գծապատկերը կազմվել է հեղինակի կողմից

անհամապատասխանությունների համար պահանջվում են ինտենսիվ մաքրման և նորմալացման ընթացակարգեր: Ընկերության մեկ նույնացուցիչի ներքո ի վերջո համախմբվելու ընդհանուր նպատակը որոշում է բազմաթիվ հաշվետվությունների առանձին-առանձին նախնական մշակման գործընթացը՝ արդյունավետ կերպով միավորելով տարբեր տվյալների աղյուսակները՝ հաշվի առնելով ժամանակային անհամապատասխանությունները և աղյուսակների կառուցվածքը:

Հաջորդ քայլը փոփոխականների մշակումն է, որը բաղկացած է երեք ենթախմբերից: Առաջինը ավանդական տվյալներն են, որոնց մեջ են մտնում հարկային հաշվետվություններում առկա հիմնական ֆինանսական տվյալները, ընկերությանը վերաբերող հատուկ տեղեկությունները, օրինակ՝ ոլորտը և հիմնադրման տարին, աշխատողներին վերաբերող տվյալները, օրինակ՝ ժամանակի ընթացքում աշխատողների ընդհանուր թվի փոփոխությունները կամ միջին աշխատավարձը, կուտակված հաշիվ-ապրանքագրերը և ստուգումների տվյալները:

Հաջորդը աշխարհագրական կամ գտնվելու վայրին վերաբերող տվյալներն են, որոնց նպատակն է բացահայտել հարևան ընկերություններում անցկացվող ստուգումների ազդեցությունը հարկ վճարողների ակնկալիքների և վարքագծային արձագանքի վրա: Բարդությունն այն էր, որ գտնվելու վայրին վերաբերող տվյալներն անհատների կողմից ձեռքով են մուտքագրվում: Այս խնդիրը լուծելու համար տվյալները ենթարկվել են տեքստային մաքրման, ինչպես նաև API-ով[15] գեոկոդավորվել են՝ հասցեները վերածելով եզակի գտնվելու վայրի կոորդինատների:

Քարտեզում ցուցադրվում է ՀԴՄ-ների բաշխման օրինակ՝ լայնության և երկայնության մասին տեղեկատվությունն ավելացնելուց հետո:



Գծապատկեր 3: ՀԴՄ-ների տարածական բաշխումը[16]

15 https://yandex.com/maps-api/products/geocoder-api

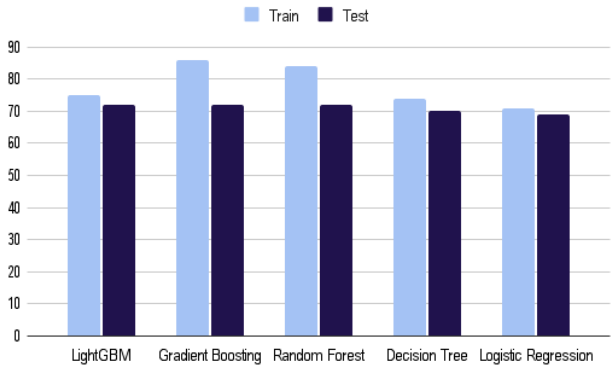
16 Գծապատկերը կազմվել է հեղինակի կողմից

Այնուհետև KNN (k-մոտակա հարևաններ) ալգորիթմի օգնությամբ որոշվել են 30 ամենամոտ ընկերությունները, վերջիններս հետագայում ճշգրտվել են երկրորդական ստուգման փուլում, որի ընթացքում չափվել է հեռավորությունը մղոններով: Այս մոտեցումն ապահովել է մոտակա ձեռնարկությունների ճշգրիտ բացահայտումը 200 մետր շառավղով: Արդյունքում, այս բազմաքայլ մոտեցման միջոցով հնարավոր է դարձել հստակ և արդյունավետ բացահայտել մոտակա ընկերությունները: Այս տեղեկատվությունը ՊԵԿ-ի համար կարելի է համարել բավականին արժեքավոր, քանի որ նախկինում ՊԵԿ-ի կողմից տարածքային վերլուծություն իրականացնելու հնարավորություն չի եղել, իսկ ներկայացված վերլուծությունը կարող է օգտակար լինել նախկինում անհայտ օրինաչափությունները և միտումները բացահայտելու գործընթացում:

Մոտակայքում գտնվող ընկերությունները բացահայտելուց հետո նրանց ստուգման և խարդախության մասին տեղեկությունները համակցվել են, այնուհետև մշակվել՝ ընկերության մակարդակի տվյալների հետ համատեղելու համար, օրինակ՝ աուդիտի ենթարկված հարևան ընկերությունների ընդհանուր թիվը կամ խարդախության համամասնությունը:

Երրորդ քայլում, մշակվել են բարձր հաճախականության տվյալները ՀԴՄ-ներից և իրականացվել է նոր փոփոխականների ստեղծման մի շարք մոտեցումներ: Դրանք ներառում են այնպիսի գործարքների համախմբում, ինչպիսիք են ընդհանուր վաճառքը կամ թողարկված ՀԴՄ կտրոնների միջին քանակը, ինչպես նաև անոմալիաների հայտնաբերումը՝ միջքառորդական միջակայքի վրա հիմնված խիստ շեղումների քանակի բացահայտումը և մասնակի ռեգրեսիայի հիման վրա միտումների հայտնաբերումը: Մասնակի ռեգրեսիան արդյունավետ մեթոդ է տվյալների վերլուծության մեջ միտումները գտնելու համար, քանի որ այն կարող է հայտնաբերել ժամանակի ընթացքում միտումների փոփոխությունները և ֆիքսել ոչ գծային հարաբերակցությունները: Այն օգտագործվել է յուրաքանչյուր ՀԴՄ-ի համար առանձին և արտահանել է այնպիսի փոփոխականներ, ինչպիսիք են բեկման կետերի քանակը, հատվածի միջին տևողությունը և թեքության ինտենսիվությունը:

Վերջնական տվյալները ստանալուց հետո ստեղծվել են տարբեր մոդելներ լավագույնն ընտրելու նպատակով՝ կիրառելով ցանցային որոնման հիպերպարամետրային օպտիմիզացում բոլոր մոդելների համար՝ կանխատեսվել են խարդախության այն դեպքերը, որոնք տեղի կունենան 2023 թվականին՝ օգտագործելով 2022 թվականի և ավելի վաղ ժամանակահատվածի պատմական տվյալները: Մոդելի կայունությունը և ընդհանրացման կարողությունները գնահատելու համար ուսումնասիրությունն օգտագործել է գնահատման ճշգրտությունը, F1-ցուցիչը և ROC-AUC չափանիշները: Մոդելների համեմատությունը ըստ ROC-AUC-ի ցուցիչի ներկայացված են գծապատկեր 4-ում.



Գծապատկեր 4: ՄՈՒ մոդելների գնահատականները[17]

Արդյունքներից կարող ենք տեսնել, որ ամենաարդյունավետ մոդելը LightGBM-ն է, քանի որ չնայած նրան որ թեստային արդյունքներում ունի հավասար ROC-AUC ցուցիչ ինչ պատահական անտառները կամ գրադիենտi խթանման մոդելը, այն ունի ամենացածր գերուսուցումը, ինչը բավականին կարևոր է մոդելի ընդհանրական աշխատանքի և պրակտիկ կիրառման համար:

Այնուհետև մոդելի ընտրությունից հետո այն օպտիմալացնելու համար կիրառվում է Ռեկուրսիվ փոփոխականների վերացում մեթոդը (RFE)՝ պահպանելով միայն ամենաարդյունավետ փոփոխականները: RFE-ի ներդրման արդյունքում օգտագործվել են 30 կարևոր փոփոխականներ և ստացվել հետևյալ ՄՈՒ մոդելը:

$$model = LGBMClassifier(random\_state = 42, \max\_depth = 4,$$
$$n\_estimators = 25, min\_samples\_leaf = 15,$$
$$boosting\_type = 'dart', class\_weight = 'balanced')$$

որտեղ՝ $random\_state$-ը պատահականության գործակից է արդյունքների վերարտադրելիության համար, max_depth-ը առանձին ծառերի առավելագույն խորությունը, $n\_estimators$-ը ծառերի քանակը, $min\_samples\_leaf$-ը տերևի մեջ պահանջվող նվազագույն նմուշների քանակը, $boosting\_type$-ը բուստինգի տիպը, $class\_weight$-ը կշիռների բալանսավորումը ըստ դասի հաճախականությունների:

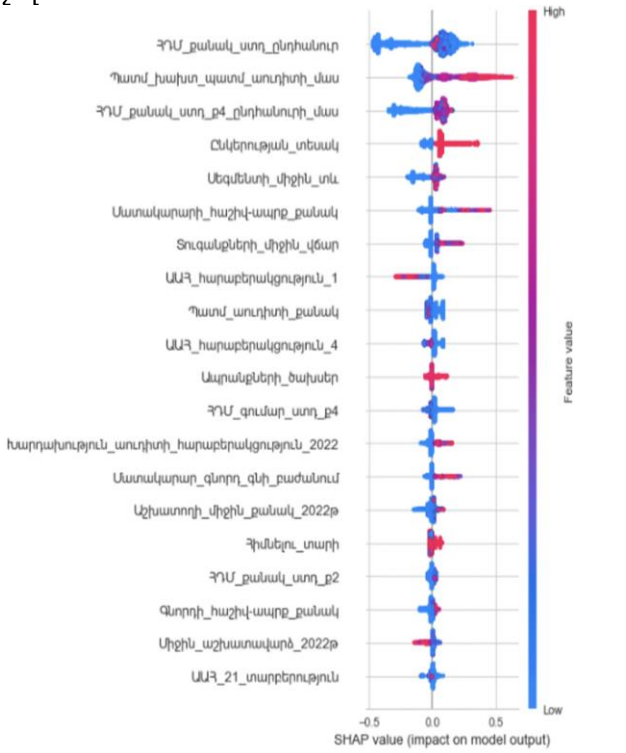
Մոդելի աշխատանքն ավելի լավ հասկանալու նպատակով մոդելի ելքային տվյալները բացատրելու համար օգտագործվում են SHapley Additive exPlanations

17 Աղյուսակը կազմվել է հեղինակի կողմից

(SHAP)[18] արժեքները՝ կանխատեսումը վերագրելով դրա յուրաքանչյուր առանձին փոփոխականին:

Ելնելով մոդելի SHAP փոփոխականի կարևորությունից՝ նկատելի է, որ ստացված կանխատեսման մոդելները ցուցադրում են կանխատեսման ճշտության մեծ կախվածություն բարձր հաճախականության և վարքագծային գործակիցներից:

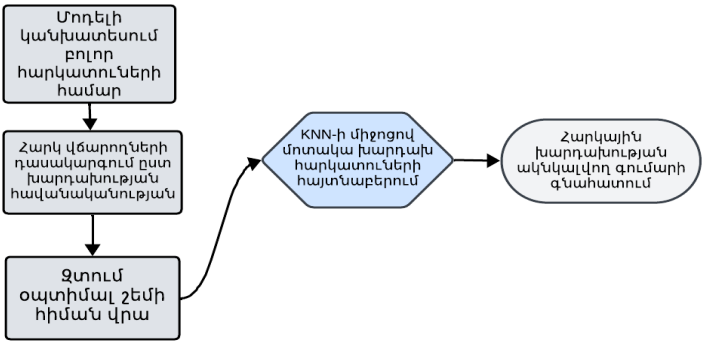
Ամենաբարձր արդյունավետությամբ մոդելը, օգտագործելով ընդամենը 30 փոփոխական, հասել է գրեթե նույն ROC-AUC-ի, ինչ բոլոր փոփոխականներն օգտագործող մոդելը: Հատկանշական է, որ այլընտրանքային փոփոխականները, օրինակ՝ վաճառքի ստանդարտ շեղումը, կանխատեսման նշանակալի գործոն են: Բացի այդ, 10-պատիկ խաչաձև ստուգման միջոցով ստացված միջին արդյունքները հաստատել են մոդելի կայուն աշխատանքը՝ 10 կրկնությունների դեպքում 0,72 ROC-AUC միջին ցուցիչով:



18 Lundberg Scott M. and Lee Su-In. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems. 4768–4777, p.4772-4776

Գծապատկեր 5: Փոփոխականի կարևորությունը ըստ SHAP արժեքների[19]

Մոդելի արդյունավետությունը համեմատվել է Բաղդասարյանի և այլոց կողմից իրականացված նախորդ հետազոտության հետ (2022թ., էջ 985),[20] որում հիմնական ուշադրությունը դարձվում էր 2018 և 2019 թվականների խարդախության կանխատեսումներին: Թեև 2019թ.-ին վերջիններիս մոդելը որոշակի բարելավում է ցույց տվել ROC-AUC-ի ցուցիչներում, այն ունի բավականին բարձր գերուսուցում և ունի ավելի քիչ տվյալներ: Ի հակադրություն, ներկայիս մոդելը, թեև ունի ավելի մեծ տվյալների հավաքածու, ցույց տվեց, որ ունի ավելի ցածր գերուսուցում և ապահովում է այնպիսի արդյունքներ, ինչպիսիք ստացվում են ավելի փոքր տվյալների հավաքածուների վրա: Սա ընդգծում է փոփոխականների մշակման կարևորությունը կանխատեսման արդյունավետության բարձրացման և ընդհանրացման համար:
Ստորև ներկայացված կառուցվածքի հիման վրա կատարվել է մոդելի արդյունավետության համեմատություն 2023թ.-ի փաստացի արդյունքների հետ:



Գծապատկեր 6: Ակնկալվող հարկերի վճարման գնահատման սխեմա[21]

Խարդախությամբ զբաղվող ընկերություններ ճանաչելու վերաբերյալ վերջնական որոշում է կայացվել ըստ խարդախության հավանականության՝ օգտագործելով F1-ցուցիչի 0,7 օպտիմալ շեմը: Հատկանշական է, որ դրոշակավորված ընկերությունների միայն 52%-ն է համապատասխանում հարկային մարմինների կողմից ստուգման ենթարկված ընկերություններին, ինչը ցույց է տալիս ՄՈՒ-ի և ավանդական ստուգման վրա հիմնված մոտեցումների միջև առկա բացը: Հաշվի առնելով հարկային ստուգումներին բնորոշ կողմնակալությունները, այսինքն

19 Գծապատկերը կազմվել է հեղինակի կողմից

20 Baghdasaryan, V., Davtyan, H., Sarikyan, A., & Navasardyan, Z. (2022). Improving tax audit efficiency using machine learning: The role of taxpayer's network data in Fraud Detection. Applied Artificial Intelligence, 36(1). P.985

21 Գծապատկերը կազմվել է հեղինակի կողմից

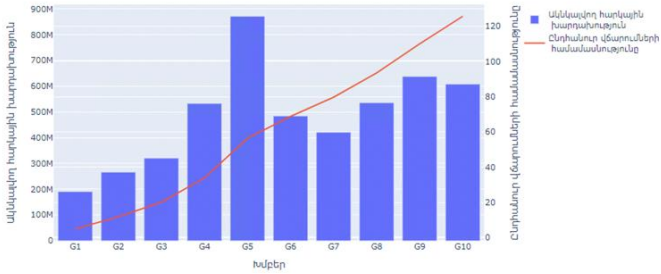
այն, որ ստուգումները գուցե չընդգրկեն պոտենցիալ խարդախության դեպքերի զգալի մասը, ինչը կհանգեցնի թաքցված եկամուտների իրական ծավալի թերագնահատմանը՝ մշակվել է նոր մեթոդ, որը կիրառում է հարկային խարդախությունների հավանականությունը և KNN ալգորիթմը: Որպես լրացուցիչ ստուգում, ընտրվել են միայն այն ընկերությունները, որոնց համար գտնվել է մերձավոր հարևան ըստ Էվկլիդեսյան հեռավորության: Վերջնական ակնկալվող հարկային տուգանքը ստանալու համար հարևանի տուգանքի գումարն այնուհետև բազմապատկվում է խարդախության հավանականության հետ.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

որտեղ՝ $d -$ և Էվկլիդյան հեռավորությունն է, իսկ (x,y)-ը կետի կոորդինատները:

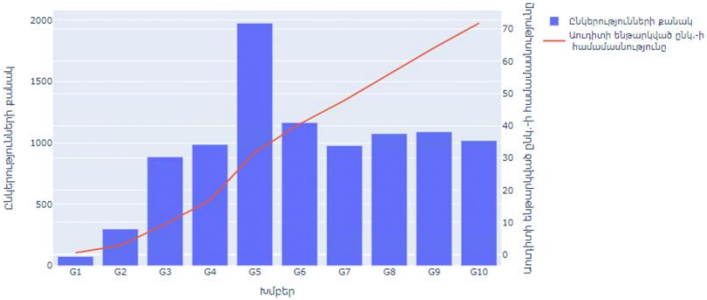
Նմանատիպ ընկերությունների օգտագործումը համադրելի ֆինանսական պրոֆիլներով, ինչպիսիք են եկամուտները, ծախսերը և աշխատողների աշխատավարձերը, հնարավորություն է տալիս ստեղծել ավելի համապարփակ մեթոդ՝ ընդհանուր հարկային խարդախության վճարումները գնահատելու համար: Սա ոչ միայն ներկայացնում է անօրինական գործունեության հետևանքով առաջացած ֆինանսական կորուստը, այլ նաև ցույց է տալիս պոտենցիալ հարկային եկամուտները, որոնք կարող էին հավաքագրվել, եթե խարդախության սխեմաները հայտնաբերվեին:

Արդյունքները գնահատելու համար տվյալները բաժանվում են 10 տարբեր մասերի՝ ելնելով խարդախության հավանականությունից, որտեղ 1-ին խումբը ներառում է խարդախության ամենաբարձր հավանականությունը: Ամենաբարձր հավանականության խմբում կանխատեսվող վճարումները, եթե առաջարկվող ռազմավարությունը կյանքի կոչվեր, կկազմեին մոտ 5,9 մլրդ դրամ՝ էականորեն ավելին, քան փաստացի հավաքագրված 3,9 մլրդ դրամը: Հետևաբար, կանխատեսվող ընդհանուր տուգանքները շուրջ 1,52 անգամ գերազանցում են հավաքագրված գումարը: Խմբի կատարողականի արդյունքները ցուցադրված են ստորև բերված գծապատկերում: Յուրաքանչյուր խմբի համար ակնկալվող ընդհանուր հարկային խարդախությունը (դրամով) ցուցադրվում է սյունյակավոր դիագրամով, իսկ կուտակային տուգանքները՝ բաժանված հավաքագրված վճարումների ընդհանուր գումարի վրա՝ կարմիր գծով: Ակնհայտ է, որ 8-րդ խումբն անվնասաբերության կետն է, քանի որ ընդհանուր ակնկալվող տուգանքները նույնիսկ մի փոքր ավելի են, քան փաստացի հավաքագրված գումարը:



Գծապատկեր 7: Ակնկալվող հարկային վճարումները և դրանց համամասնությունը՝ հիմնված խարդախության հավանականության խմբերի վրա[22]

Այս արդյունքներին հակադրվող տեսակետների համաձայն կարելի է պնդել, որ ավելի մեծ թվով հարկ վճարողներ կարող էին նպաստել տուգանքների ավելացմանը, ինչը կարող է բացատրել անհամապատասխանությունը: Այս մտահոգությունը լուծելու և լրացուցիչ հստակություն մտցնելու նպատակով իրականացվել է համեմատական վերլուծություն՝ առանձնակի ուշադրություն դարձնելով մասնակից առանձին ընկերությունների քանակի վրա: Այս վերլուծության ընթացքում յուրաքանչյուր խմբի եզակի ընկերությունների կուտակային թիվը համեմատվել է կարմիր գծով ներկայացված ստուգման ենթարկված ընկերությունների ընդհանուր թվի հետ և յուրաքանչյուր խմբի ընկերությունների եզակի թիվը ցուցադրվել է սյունյակավոր դիագրամում: Վերլուծությունը ցույց է տվել, որ ընկերությունների ընդհանուր թիվը ոչ միայն սուգման ենթարկված ընկերությունների թվից ցածր է, ինչը կազմում է փաստացի ընդհանուր թվի ընդամենը 85%-ը, այլև, անվնասաբերության սցենարի համար, անհրաժեշտ էր ընդամենը 62%-ը (8,305): Այս վերլուծությունն օգնում է վերացնել տուգանքների նկատվող աճի վրա ընդհանուր հարկ վճարողների պոտենցիալ ազդեցության վերաբերյալ անորոշությունը:



22 Գծապատկերը կազմվել է հեղինակի կողմից

Գծապատկեր 8: Հարկ վճարողների թիվը և նրանց համամասնությունը՝ հիմնված խարդախության հավանականության խմբերի վրա[23]

Իրականացված վերլուծությունների արդյունքում հեղինակը հանգել է մի **շարք եզրակացությունների**, որոնք ամփոփ ներկայացված են ստորև:

1. մշակվել է տարածական վերլուծություն իրականացնելու համար բաց հասցեի ձևաչափից եզակի կոորդինատների նույնականացման գործիք,
2. գնահատվել է տեղակայման բաշխվածության ազդեցությունը, որը բացահայտում է թե ինչպես են նույն տարածաշրջանի հարկ վճարողների մոտ իրականացված ստուգումներն ազդել հարկ վճարողների վարքագծի վրա՝ վերհանելով 30 ամենամոտ ընկերությունները և միավորելով նույն տարածաշրջանի հարկ վճարողների կողմից խախտումների դեպքերի վերաբերյալ տվյալներն ու ստուգումների արդյունքները,
3. ձևավորվել է հսկիչ-դրամարկղային մեքենաների բարձր հաճախականության տվյալների հիման վրա բարդ փոփոխականների մշակում, ինչպիսիք են գործարքների համախմբումը, անոմալիաների հայտնաբերումը և մասնակի ռեգրեսիայի միտումների հայտնաբերման ալգորիթմները,
4. մշակվել է այլընտրանքային տվյալների ինտեգրման մեթոդաբանություն, ներառյալ ՀԴՄ-ներից բարձր հաճախականության տվյալները և տեղակայման բաշխվածության ազդեցությունը:
5. ստեղծվել է ՄՈՒԻ-ի վրա հիմնված խարդախության դեպքերի հայտնաբերման մոդել՝ ընդամենը 30 փոփոխականներով և 0.72 ROC-AUC-ի կայուն և ընդունելի կատարողականությամբ: Արդյունքների հետագա ստուգումը, որը հիմնված է 10-ակի խաչաձև վավերացման վրա, միջինում ապահովում է ROC-AUC 0,72 միավոր:
6. մշակվել է հնարավոր հարկային խարդախության գնահատման նոր մեթոդը՝ հիմնված խարդախության հավանականությունների և KNN ալգորիթմի վրա (օգտագործվում է առավել միանման ընկերությունները բացահայտելու համար): Ելնելով մեթոդից՝ առաջարկվող գործիքով հարկային եկամուտները կկազմեն մոտավորապես 5,9 մլրդ ՀՀ դրամ, ինչն էականորեն մեծ է, քան 3,9 մլրդ դրամը, որը փաստացի հավաքագրվել է ստուգման ենթարկված ընկերությունների միայն 85%-ից:

23 Գծապատկերը կազմվել է հեղինակի կողմից

Ատենախոսության հիմնական արդյուքները հրապարակվել են գիտական 4 հրապարակումներում.

1. Baghdasaryan, V., & Sarikyan, A. (2023). Location-based tax incentives for non-farm rural enterprises in Armenia. *The Journal of Development Studies*, *60*(4), 553–573. https://doi.org/10.1080/00220388.2023.2286894
2. Baghdasaryan, V., Davtyan, H., Sarikyan, A., & Navasardyan, Z. (2022). Improving tax audit efficiency using machine learning: The role of taxpayer's network data in Fraud Detection. *Applied Artificial Intelligence*, *36*(1), 963-985. https://doi.org/10.1080/08839514.2021.2012002
3. Enterprise Zone Tax Exemption Policy Impact Evaluation in Rural Regions of Armenia *ALTERNATIVE* Quarterly Academic Journal on Economy and Management 2024, pages 100-108.
4. Sarikyan, A. (2024). A multifaceted approach to Amazon's financial performance: Time Series, difference in difference, and regression discontinuity analysis of R&D, marketing, and mobile adoption. *Регион и Мир / Region and the World*, 116–121. https://doi.org/10.58587/18292437-2024.2-116

САРИКЯН АРСИНЕ АРСЕНОВНА

“ПРОГНОЗИРОВАНИЕ НАЛОГОВЫХ ДОХОДОВ, ИСПОЛЬЗУЯ АЛТЕРНАТИВНЫХ ИСТОЧНИКОВ ДАННЫХ, ПРИМЕНЕНИЕ В АРМЕНИИ”.

Автореферат диссертации на соискание ученой степени кандидата экономических наук по специальности 08.00.08-
“Математическое моделирование экономики“.
Защита диссертации состоится 13-го сентября 2024 года, в 13:30 часов, на заседании специализированного совета по экономике 015 КВОН РА, действующего в Ереванском государственном университете по адресу г. Ереван, 0025, ул. Абовяна 52”.

РЕЗЮМЕ

Прогнозирование налоговых поступлений является основой налогово-бюджетной политики, а также государственного финансового планирования: оно проливает свет на ожидаемые источники дохода, что позволяет принимающим решения лицам разумно распределять средства на общественные нужды. Одной из проблем при прогнозировании налоговых поступлений является скрытие дохода или налоговое мошенничество, которое отвлекает финансовые ресурсы, предназначенные для правительства и, в конечном итоге, для общественных нужд.

Цель данного исследования - разработать методологии прогнозирования налоговых поступлений в контексте Армении с использованием передовых методов машинного обучения (МО). Это позволит создать надежные модели, учитывающие налоговое мошенничество, которое равносильно потере доходов. Наряду с традиционным использованием данных в налоговых отчетах, подробно излагаются альтернативные и нетрадиционные источники, в частности обращая внимание на побочные эффекты местоположения и высокочастотные данные из кассовых аппаратов.

Для достижения поставленных целей были выявлены следующие задачи:

– выявление лучших практик путем проведения тщательного обзора литературы по подходам, используемым в настоящее время для прогнозирования налоговых поступлений,
– определение, сбор, очистка и предварительная обработка данных из различных источников, включая высокочастотные и геолокационные данные,
– экспериментирование и выбор моделей МО, таких как логистическая регрессия, дерево решений, случайный лес, градиентный бустинг и LightGBM, с целью определения наиболее эффективной модели. Определение наиболее эффективной модели, ее тщательная оценка и проверка,
– оценка производимых в результате налогового мошенничества платежей путем выявления наиболее сопоставимых компаний, используя сопоставимые финансовые профили и вероятности мошенничества.

Основные результаты и научная новизна диссертации включают следующее:

1. разработка детальных и сложных методов подготовки данных для устранения несоответствий в данных и обеспечение выработки подходящего для моделирования,

2. разработка уникальных подходов для конструирования признаков высокочастотных данных по налоговым поступлениям. Учитывая обилие информации, содержащейся в этих данных, исследование направлено на создание более точных и надежных моделей обнаружения мошенничества путем выявления скрытых закономерностей и аномалий,
3. изучение и оценка побочных эффектов местоположения для понимания того, каким образом проверки соседей влияют на компании. Результаты могут дать новое представление о взаимозависимости и воздействиях социальных сетей между различными компаниями, а также о том, как аудиторские вмешательства влияют на соблюдение налогоплательщиками налоговых обязательств,
4. проверка ряда моделей МО с целью обнаружения мошенничества и определения наиболее эффективной модели, которая обеспечивает приблизительно 0,72 ROC-AUC, подтвержденную 10-кратной перекрестной проверкой,
5. разработка нового метода оценки потенциальных налоговых мошенничеств с использованием вероятностей мошенничества и модели К-ближайших соседей (KNN) для выявления наиболее похожих налогоплательщиков. Предложенный подход позволит вернуть примерно в 1,3 раза больше налоговых платежей, а количество компании составят 72% от фактическо проверяемых.

ARSINE ARSEN SARIKYAN

"TAX REVENUE FORECASTING, USING ALTERNATIVE DATA SOURCES, APPLICATION IN ARMENIA"

The abstract of the dissertation submitted for the pursuing degree of PhD in Economics in the field of 08.00.08 – "Mathematical Modeling of the economy".
The defense of the dissertation will take place at 13:30 on September 13, 2024, at the meeting of the Specialized Council 015 in Economics of the RA HESC, acting at the Yerevan State University.
Address: 0025, 52 Abovyan Street, Yerevan.

ABSTRACT

Tax revenue forecasting constitutes a cornerstone of fiscal policy and government financial planning; it sheds light on expected income sources, enabling decision-makers to allocate funds wisely for public needs. One of the challenges of revenue forecasts is tax fraud, as financial resources that ought to go to the government and, ultimately, to public needs are diverted through dishonest means.

This research aims to advance tax revenue forecasting methodologies in the context of Armenia by leveraging cutting-edge Machine Learning (ML) techniques to develop robust forecasting models that account for tax fraud, which is the same as the lost revenue. Traditional reliance on tax reports data is expanded upon by incorporating alternative and non-traditional sources, mainly focusing on location spillover effects and high frequency data from cash register machines.

The study seeks to enhance the accuracy of revenue estimates by utilizing ML algorithms, drawing insights from the intricate patterns within tax reports, and considering alternative data sources such as location-based influences and high frequency data from tax cash register machines. Currently, a vast array of data sources, termed "alternative data sources," are available for tax fraud prediction purposes.

Last but not least, the dissertation aims to implement an advanced tax payment estimation method that will help the government of the Republic of Armenia to optimize its operations. By effectively identifying fraudulent activities, the model reduces the non-compliance tax gap and is a solid tool against tax fraud.

In order to achieve these goals, the following problems have been identified.

– identification of the best practices by undertaking a thorough literature review of the approaches currently used to forecast tax revenue,
– data definition, collection, cleaning, and preprocessing from a variety of sources, including high-frequency cash receipt data and geolocation data,
– experimentation and selection of ML models, such as logistic regression, decision tree, random forest, gradient boosting, and LightGBM, to identify the best-performing model. Identification of the best performing model, its thorough evaluation and validation,
– estimation of expected tax fraud payments through the discovery of the most comparable companies using comparable financial profiles and fraud probabilities.

The main results and scientific novelty of the dissertation are as follows;

1. development of detailed and intricate data preparation techniques for addressing data inconsistencies and guaranteeing the production of a comprehensive dataset appropriate for modeling,
2. creation of unique approaches for high-frequency tax receipt data feature engineering. With the abundance of information found in this data, the research aimed to create more accurate and dependable fraud detection models by identifying latent patterns and anomalies that might be indicators of possible fraudulent activity,
3. examination and assessment of the location spillover effects to understand how neighbor audits affected companies, with the goal of spotting fraudulent trends in taxpayer behavior. The results can provide new insights into the interdependence and social network effects among different companies, as well as the ways audit interventions impact taxpayer compliance,
4. exploration and examination of a range of machine learning models for tax fraud detection and identification of the most effective model through rigorous evaluation for accurate fraud detection, which provides around 0.72 ROC-AUC, validated by 10-fold cross-validation,
5. design of the novel method for estimating potential tax fraud payments using fraud probabilities and the KNN model for identifying the most similar taxpayers. The suggested approach could recover approximately 1.3 times more tax payments with only 72% of audited companies.