U 2922

ՀՀ ԿՐԹՈՒԹՅԱՆ, ԳԻՏՈՒԹՅԱՆ, ՄՇԱԿՈՒՅԹԻ ԵՎ ՍՊՈՐՏԻ
ՆԱԽԱՐԱՐՈՒԹՅՈՒՆ
ԵՐԵՎԱՆԻ ՊԵՏԱԿԱՆ ՀԱՄԱԼՍԱՐԱՆ

ՂԱԼԱՉՅԱՆ ԳԵՎՈՐԳ ՀՈՎՍԵՓԻ

ԱՐՀԵՍՏԱԿԱՆ ԲԱՆԱԿԱՆՈՒԹՅԱՆ ՄՈԴԵԼՆԵՐԻ ԳՆԱՀԱՏՄԱՆ ԱՐԴՅՈՒՆԱՎԵՏՈՒԹՅՈՒՆԸ ՍՈՑԻԱԼ-ՏՆՏԵՍԱԿԱՆ ԽՆԴԻՐՆԵՐՈՒՄ

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

ՍԵՂՄԱԳԻՐ

ԵՐԵՎԱՆ – 2022

© National Library of Armenia OCR by Portmind

Ատենախոսության թեման հաստատվել է Երևանի պետական համալսարանում:

Գիտական ղեկավար՝ տնտեսագիտության դոկտոր, պրոֆեսոր
Ռուբեն Ալբերտի Գևորգյան

Պաշտոնական ընդդիմախոսներ՝ տնտեսագիտության դոկտոր, պրոֆեսոր
Աշոտ Աղասու Թավադյան

տնտեսագիտության թեկնածու, դոցենտ
Արամ Տալվորիկի Կարապետյան

Առաջատար կազմակերպություն՝ **Հայաստանի ազգային պոլիտեխնիկական համալսարան**

Ատենախոսության պաշտպանությունը կայանալու է 2022թ. հունիսի 24-ին՝ ժամը 15:00-ին, Երևանի պետական համալսարանում գործող ՀՀ ԲՈԿ-ի տնտեսագիտության թիվ 015 մասնագիտական խորհրդի նիստում:

Հասցե՝ 0009, ք. Երևան, Աբովյան փ. 52:

Ատենախոսությանը կարելի է ծանոթանալ Երևանի պետական համալսարանի գրադարանում:

Սեղմագիրն առաքված է 2022թ. մայիսի 13-ին:

015 մասնագիտական խորհրդի
գիտական քարտուղար,
տնտեսագիտության
թեկնածու, դոցենտ **Ա. Հ. Հակոբջանյան**

11-2922

ԱՏԵՆԱԽՈՍՈՒԹՅԱՆ ԸՆԴՀԱՆՈՒՐ ԲՆՈՒԹԱԳԻՐԸ

Ատենախոսության թեմայի արդիականությունը: Արհեստական բանականությունը միջդիսցիպլինար հետազոտության ճյուղ է, որը ուսումնասիրում է հաշվողական մեքենաները և համակարգերը, որոնք օժտված են այնպիսի մարդկային հատկանիշներով, ինչպիսիք են ընկալումը, ուսուցումը, պատճառահետևանքային կապերի ստեղծումը, պլանավորումը և խնդիրների լուծումը: Չնայած 20-րդ դարում այն հայտնի է եղել տարբեր անուններով՝ կիբեռնետիկա, ավտոմատա, արհեստական բանականության հետազոտության սկիզբ է համարվում 1956 թվականը՝ պայմանավորված Դարթմութի արհեստական բանականության ուսումնասիրության նախագծով: Երկրորդ համաշխարհային պատերազմից ի վեր այս ճյուղում կատարվել են մեծաթիվ տեսական հետազոտություններ, սակայն դրանց գործնական կիրառությունը հնարավոր է դարձել ընդամենը 1970-ականների վերջին, իսկ 2000-ականներից արդեն արհեստական բանականության համակարգերը դարձել են լայն կիրառության գործիքներ՝ շնորհիվ հաշվողական տեխնոլոգիաների արտադրության մասայականացման: Նման համակարգերը այժմ լայն կիրառելի են ոլորտներում, ինչպիսիք են բժշկությունը, մեդիա տեխնոլոգիաները, ֆինանսները, տիեզերական ուսումնասիրությունները: Մասնավորապես ֆինանսական տարբեր կազմակերպություններում՝ բորսաներ, բանկեր, ներդրումային ընկերություններ, ունիվերսալ վարկային կազմակերպություններ, առկա են մեծ հաճախականությամբ և ծավալի տվյալների հոսք, որոնց օպտիմալ մշակման և գիտելիքի արտածման մեխանիզմները մեծապես ազդում են որոշումների կայացման արդյունավետությանը: Արհեստական բանականության ֆինանսական կազմակերպություններում օգտագործվող մոդելները ունեն բարձր ճշգրտություն, կանխատեսման կարճ ժամանակահատված և տվյալների ընդհանրացման բարձր մակարդակ:

Այնուհանդերձ, նմանատիպ համակարգերում շատ հաճախ բացակայում է վստահության տարրը, ասել է թե լինել իրավական՝ օրենքներին և այլ իրավական նորմերին համապատասխան, էթիկական՝ հասարակական էթիկական նորմերին համապատասխան, սոցիալական տարբեր խմբերի համար նմանապես գործող: Քանի որ արհեստական բանականության մոդելները առավելապես մաթեմատիկական են՝ դիսկրետ կամ շարունակական ելքային փոփոխականներով, իրավական տարրի պահպանումը տեխնիկական խնդիր է, որը լուծվում է մաթեմատիկական սահմանափակումների ավելացմամբ: Պատկերը այլ է էթիկական տարրը պահպանելու խնդրում. այստեղ հաշվի են առնվում հասարակական տարբեր երևույթներ, որոնք հստակ կարգավորումներ չունեն կամ կարգավորված են մասնակի, օրինակ՝ վարկային կազմակերպության արհեստական բանականությամբ աշխատող համակարգի անվտանգությունը կեղծ նմուշներից, նրա կողմից վարկի հաստատման/մերժման բացատրություն, տվյալների արդար մշակում: Այս և այլ նմանատիպ տարրերի ուսումնասիրությամբ, շեղումների մաթեմատիկական սահմանմամբ, բացահայտմամբ և մեղմացմամբ է զբաղվում էթիկական արհեստական բանականությունը: Նաև հարկ է նշել, որ վերջին տարիներին նաև ուսումնասիրություններ են կատարվում էթիկական արհեստական բանականության օրենսդրական ամրագրումների և պահանջների ուղղությամբ, օրինակ՝ Եվրոպական

© National Library of Armenia

հանձնաժողովի 2019 թվականին հրապարակած ուղեցույցը վստահելի արհեստական բանականության վերաբերյալ:

Այս հետազոտության շրջանակներում ուսումնասիրվել է էթիկական արհեստական բանականության երեք տարր՝ արդարություն, բացատրելիություն, անվտանգություն: Արդար արհեստական բանականությունը կարողանում է նման սուբյեկտների և/կամ խմբերի համար կայացնել նման որոշումներ՝ առանց որևէ սոցիալական խտրականության, օրինակ՝ սեռ, ռասա, տարիքային խումբ, ազգություն: Բացատրելի է համարվում արհեստական բանականության այն մոդելը, որը ելքային փոփոխականի կանխատեսված արժեքի համար կարող է տալ մեկնաբանություն, թե ինչ չափանիշներով է կատարվել կանխատեսումը: Անվտանգության համատեքստում ուսումնասիրվել է արհեստական բանականության մոդելի վրա հարձակումները՝ կեղծ նմուշներով և մոլորեցմամբ: Նման տարրերի ցուցանիշների սահմանումը, կշռումը և ագրեգացումը ճշգրտության ցուցանիշների հետ թույլ կտա կազմակերպություններին կատարել մոդելի բազմաչափ գնահատում և ընտրություն:

Հետազոտության հիմնական նպատակը և խնդիրները: Հետազոտության հիմնական նպատակներն են ուսումնասիրել արհեստական բանականության մոդելների էթիկական ցուցանիշների էմպիրիկ վերլուծությունը, քանակական գնահատման մեթոդների սահմանումը և շեղման մեղմացման կիրառությունների ուսումնասիրությունը: Այս նպատակին հասնելու համար առաջադրվել են հետևյալ խնդիրները.

- ուսումնասիրել արհեստական բանականության՝ մասնավորապես երկուական դասակարգման մոդելների տարբեր դասերի, դրանց ճշգրտության ցուցանիշների վերաբերյալ հետազոտությունները, ինչպես նաև կիրառությունը ֆինանսական կազմակերպություններում,
- մոդելների տարբեր դասերից ընտրել մեքենայական ուսուցման մեկական մոդել, ունիվերսալ վարկային կազմակերպության օրինակով կիրառել նշված մոդելները, սահմանել և գնահատել դրանց ճշգրտության ցուցանիշները, ընտրել լավագույն մոդելը ըստ այդ ցուցանիշի, գտնել փոփոխականներ, որոնք առավելագույնն են բացատրում վարկառուի ռիսկայնությունը,
- ուսումնասիրել վստահելի արհեստական բանականության վերաբերյալ նախկին ուսումնասիրությունները, օրենսդրական կարգավորումները, մասնավորապես՝ արդարության, բացատրելիության ցուցանիշների քանակական վերլուծության մեթոդաբանությունը, առկա տեխնիկական հնարավորությունները և մաթեմատիկական գրադարանները դրանց հաշվարկման համար,
- գնահատված մոդելների համար ընտրել էթիկական արհեստական բանականության համապատասխան ցուցանիշները, կառուցել ճշգրտության և էթիկական պայմանների ագրեգատ ցուցանիշ, ապա կատարել մոդելների համեմատական վերլուծություն,
- ուսումնասիրել արհեստական բանականության էթիկական ցուցանիշների մեղմացման ալգորիթմների ուսումնասիրություն, գնահատված մոդելներին համապատասխան ալգորիթմի ընտրություն, ապա դրա կիրառումից հետո, վերահաշվարկել ճշգրտության և էթիկական, ինչպես նաև ագրեգատ ցուցանիշը և կատարել մոդելների ցուցանիշների վերլուծություն,

- կատարել եզրահանգումներ ցուցանիշների, մոդելների ընտրության և ագրեգատի կառուցման վերաբերյալ:

Հետազոտության օբյեկտը և առարկան: Հետազոտության օբյեկտը սոցիալ-տնտեսական խնդիրներում արհեստական բանականության մոդելների արդյունավետության գնահատման գործընթացն է, իսկ առարկան արհեստական բանականության մոդելների էթիկական ցուցանիշների քանակական գնահատականներն են, նրանց կառուցման և հետազոտման մեթոդները:

Հետազոտական տեսական, մեթոդաբանական և տեղեկատվական հիմքերը: Հետազոտական տեսական հիմքը արհեստական բանականության, նրա էթիկական տարրերի, մեքենայական ուսուցման էթիկական ցուցանիշների ու դրանց մեղմացման ալգորիթմների վերաբերյալ արտասահմանյան հեղինակների աշխատություններն ու այլ աղբյուրներն են: Հետազոտության շրջանակներում կիրառվել են մաթեմատիկական մոդելավորման, կիրառական ինֆորմատիկայի, տնտեսաչափական և վիճակագրական վերլուծության մեթոդներ, այդ թվում՝ տարածական տվյալների դասակարգման վերլուծություն, փոփոխականների և ցուցանիշների նորմավորում, երկու օբյեկտների միջև հեռավորության և նմանության չափման վիճակագրական վարկածների ստուգման, կշռման մեթոդներ, ինչպես նաև տվյալների վիզուալիզացիա, ագրեգատ ցուցանիշի կառուցում, համեմատական վերլուծություն, արտաքին այլ պայմանների անտեսում և դրանից վերացարկում:

Հետազոտության տեղեկատվական հիմք են ծառայել Հայաստանի Հանրապետության տարածքում գրանցված և գործող ունիվերսալ վարկային կազմակերպություններից մեկի կողմից տրամադրված տվյալները գյուղատնտեսական վարկերի և դրանց վճարման, ինչպես նաև վարկառուի դեմոգրաֆիայի, սոցիալ-տնտեսական կարգավիճակի, վճարունակության վերաբերյալ: Կազմակերպության հետ ստորագրվել է չբացահայտման համաձայնագիր՝ ըստ որի տեղեկատվության և տվյալների որոշ ձևեր և տեսակներ չեն կարող բացահայտվել ընթերցողի համար:

Ատենախոսության գիտական նորույթը: Ատենախոսության հիմնական արդյունքները և գիտական նորույթները հետևյալն են.

- ստացվել է արհեստական բանականության մոդելների գնահատման սոցիալ-տնտեսական ցուցանիշներ արդարության, բացատրելիության և անվտանգության համար, որոնց և ճշգրտության ցուցանիշների հիման վրա կառուցել ենք ագրեգատ ցուցանիշ: Այնուհետև յուրաքանչյուր ցուցանիշի բարելավման համար առաջարկվել է մոդելի դասից անկախ մեղմացման ալգորիթմ, ներկայացվել է ագրեգատ ցուցանիշի բարելավումը,
- առաջարկվել է տվյալների հավաքագրման և մշակման մեթոդաբանություն, որը առավել համապատասխանում է արդարության, բացատրելիության և անվտանգության գնահատման խնդիրներին: Այն փորձարկվել է համագործակցող ունիվերսալ վարկային կազմակերպության օրինակով, առաջարկվել են փոփոխություններ տվյալների հավաքագրման մեխանիզմում,
- առաջարկվել է ՀՀ-ում վարկառուների վճարունակությունը գնահատող ցուցանիշներ և փոփոխականներ: ՀՀ բնակչության նկատմամբ

© National Library of Armenia

ներկայացուցչական ընտրանքի վրա պարզվել է, որ գյուղատնտեսական վարկերի ռիսկի վրա առավելապես ազդում են նախորդ վարկերի ուշացման օրերը, տնտեսության ընդհանուր ամսական եկամուտը, տոկոսադրույքի ծածկման գործակիցը:

Հետազոտության արդյունքների տեսական և գործնական նշանակությունը: Հետազոտության տեսական և գործնական նշանակությունը այն է, որ ստացված արդյունքները և առաջարկված մեթոդները կնպաստեն արհեստական բանականության մոդելների առավել արդյունավետ գնահատմանը, ցուցանիշներում շեղումների բացահայտմանը և դրանց մեղմացմանը: Կառուցված մոդելների օրինակները և ցուցանիշները կարող են օգտագործվել ունիվերսալ վարկային կազմակերպությունների, բանկերի կամ այլ ֆինանսական կազմակերպությունների կողմից տվյալների մշակման, որոշումների կայացման համակարգերի ստեղծման, նաև հետազոտությունում նշված օրենքներին համապատասխանեցման պարագաներում:

Ատենախոսության արդյունքների փորձաքննությունը և հրապարակումները: Ատենախոսության հիմնական դրույթները քննարկվել են

- ԵՊՀ տնտեսագիտության մեջ մաթեմատիկական մոդելավորման ամբիոնում,
- ԵՊՀ կողմից գիտաարտադրական գործուղման ժամանակ՝ «IBM Research Zurich GmbH», Ցյուրիխ, Շվեյցարիա (2019 օգոստոս – 2020 հունվար),
- KDD2021 գիտաժողովի RAI@KDD2021 բաժնում, Սինգապուր (առցանց, 2021 օգոստոս 12-15)

Ատենախոսության կառուցվածքը և ծավալը: Ատենախոսությունը բաղկացած է ներածությունից, երեք գլուխներից, եզրակացություններից և առաջարկություններից, օգտագործված գրականության ցանկից և հավելվածներից: Ատենախոսության ծավալը 118 էջ է (առանց հավելվածների 110 էջ):

ԱՏԵՆԱԽՈՍՈՒԹՅԱՆ ՀԻՄՆԱԿԱՆ ԲՈՎԱՆԴԱԿՈՒԹՅՈՒՆԸ

Ատենախոսության «**Ներածություն**»-ում հիմնավորվել է թեմայի արդիականությունը, առաջադրվել են հետազոտության նպատակներն ու խնդիրները, սահմանվել են ատենախոսության ուսումնասիրության օբյեկտը և առարկան, ներկայացվել են ստացված արդյունքների գիտագործնական նշանակությունը, նրանց փորձաքննությունը և հրապարակումները, ինչպես նաև աշխատանքի կառուցվածքը և ծավալը:

Ատենախոսության առաջին՝ «**Արհեստական բանականությունը, նրա էթիկական պայմանները և կիրառությունը ֆինանսներում**» գլխում ներկայացված են արհեստական բանականության սահմանումը, գնահատումը և շեղումների տեսակները, էթիկական պայմանները, ինչպես նաև քննարկվում են մասնավոր դեպքեր՝ արհեստական բանականության կիրառությունը ֆինանսներում, վարկային ռիսկի գնահատման խնդիրը, անդրադարձ է կատարվել այլ հետազոտողների կողմից կատարված աշխատանքներին, ընդգծված են սույն հետազոտության համար հիմնարար աշխատությունները:

Արհեստական բանականությունը միջդիսցիպլինար հետազոտության ճյուղ է, որը ուսումնասիրում է մեքենաներ, որոնք օժտված են մարդկային հատկանիշներով, ինչպիսիք են ընկալումը, ուսուցումը, պատճառահետևանքային կապերի ստեղծումը, պլանավորումը և խնդիրների լուծումը: **Մեքենայական ուսուցումը** ուսումնասիրում է ալգորիթմներ, որոնք դիտարկումներից վերցնում են տվյալներ և ինֆորմացիա՝ որպես մուտքային փոփոխականներ, կատարում որոշակի տվյալների ընդհանրացում՝ արտահայտելու մարդկային մտավոր, զգայական, ինչպես նաև ֆիզիկական հատկանիշներ: **Ընդհանրացումը** պրոցես է, որտեղ հատուկ նմուշների խմբերը վերացարկվում են ընդգրկուն հայեցակարգերի և/կամ որոշման կանոնների:

Վստահելի մեքենայական ուսուցումը (trustworthy machine learning) ուսումնասիրում է մեքենայական ուսուցման համակարգերի համապատասխանությունը սոցիալական արդարություն, հավասարության կանոններին, օրենսդրական պահանջներին, ինչպես նաև հասարակական և բարոյական նորմերին: Վստահությունը հարաբերություն է վստահորդի և հոգաբարձուի միջև՝ վստահորդը վստահում է հոգաբարձուին, և նրա կառավարչական սահմանումը համապատասխանում է մեքենայական ուսուցման համակարգի վստահություն սահմանմանը: Վստահությունը կառավարչական հարաբերություններում կողմերից մեկի պատրաստակամությունն է լինել խոցելի մյուս կողմի գործողություններին՝ ակնկալելով, որ մյուս կողմը, կանի վստահորդի համար կարևոր գործողություն՝ անկախ դրա դիտարկման և վերահսկման հնարավորությունից:Այս սահմանմամբ կառաջնորդվենք մեքենայական ուսուցման համակարգի վստահության տարրերը քննարկելիս, և աշխատանքում վստահության ասելով, ի նկատի ունենք վերոնշյալ սահմանումը: Վստահելի մեքենայական ուսուցման համակարգը պետք է ունենա բավականաչափ արտադրողականություն, հուսալիություն, մարդկային փոխազդեցություն, նպատակ: Վստահելի մոդելի կառուցումը ունի երեք ազդեցության երեք հնարավոր փուլեր

- ուսուցման տվյալների նախնական մշակում (pre-processing),
- մեքենայական ուսուցման ալգորիթմի կիրառում (training կամ inprocessing),

© National Library of Armenia

- մոդելի արդյունքների հետմշակում (post-processing):

Գծապատկեր 1: Մեքենայական ուսուցման մոդելի կառուցումը ըստ տվյալների մշակման փուլի:

Հետազոտության շրջանակներում ուսումնասիրել ենք վարկային ռիսկի գնահատման տնտեսամաթեմատիկական մոդելավորումը: Խնդրի ձևակերպումը հետևյալն է. վարկային կազմակերպության կողմից տրամադրած վարկերի վերաբերյալ պատմական տվյալների հիման վրա կառուցել մեքենայական ուսուցման մոդել, որը վարկի նոր դիմումի համար կկանխատեսի տրվող վարկի ուշացման հավանականությունը: Հետազոտության համար օգտագործվել են ՀՀ-ում գրանցված և գործող ՈՒՎԿ-ի կողմից 2015-2019 թթ. տրված գյուղատնտեսական վարկերի տվյալներ, որոնց խմբերը և փոփոխականները հետևյալն են՝

- դեմոգրաֆիական տվյալներ,
- վարկառուին պատկանող գյուղացիական տնտեսությանը վերաբերող համախառն տվյալներ,
- գյուղացիական տնտեսության ֆինանսական ագրեգատներ,
- վարկառուի վարկային պատմության տվյալներ,
- սույն վարկի վերաբերյալ տվյալներ:

Ատենախոսության երկրորդ՝ «**Արհեստական բանականության սոցիալ-տնտեսական ցուցանիշների տնտեսագիտական-մաթեմատիկական սահմանումը եվ մեկնաբանությունը**» գլխում ներկայացված են վարկային ռիսկի գնահատման մոդելը, նրա գնահատման ճշգրտության և էթիկական ցուցանիշերը:

$X = \{x_{ij}\}_{mxn}$-ը կնշանակենք մուտքային փոփոխականների մատրիցը՝

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mn} \end{bmatrix},$$

որտեղ՝

- $i = 1, 2, \ldots, m$
- $j = 1, 2, \ldots, n$
- x_{ij}-ն i-րդ փոփոխականի j-րդ նմուշի արժեք է:

X մատրիցում սյուները և տողերը համապատասխանում են վարկի տրամադրման նմուշներին և վարկառուին բնութագրող փոփոխականներին: Հետագա քայլերում առանձին նմուշները կդիտարկվեն որպես սյուն վեկտորներ:

$Y = \{y_j\}_n$-ը կնշանակենք ելքային փոփոխականների սյուն վեկտորը՝

$$Y = [y_1 \ldots y_n]:$$

$y_j \in \{0,1\}$, այն է՝ ելքային փոփոխականը ունի երկու հնարավոր արժեք. 0, երբ j-րդ վարկի դիմումը մերժվել է, և 1, երբ j-րդ վարկի դիմումը հաստատվել է:

X մատրիցի յուրաքանչյուր տող վեկտորի համապատասխանում է Y-ի մեկ սկալյար արժեք. այս զույգը կոչվում է ուսուցման տվյալների բազմություն (learning set)՝

$$D_{train} = \{(x_{ij}, y_j) \in \mathbb{R}^{m+1}\}:$$

Երկուական դասակարգման մոդելը սահմանվում է որպես $\mathbb{R}^m$ բազմության մուտքային տվյալների արտապատկերում $\mathbb{R}^1$ բազմության ելքային տվյալներ[1]:

Վարկային ռիսկի մոդելավորման համար օգտագործել ենք հետևյալ ալգորիթմները՝

- Logistic Regression[2]
- Random Forest[3]
- Gradient Tree Boosting[4]
- Multi-layer Perceptron[5]

Հետազոտվել են 10485 վարկերի վերաբերյալ նախնական տեսքի նմուշներ. տվյալների նախնական մշակման արդյունքում մնացել են 8279 նմուշներ: Քայլերը հետևյալն են՝

1. Հեռացված արժեքների որոշում և անտեսում (outlier detection)

Բաշխումից հեռացված տվյալները որոշվել են ներքառորդային միջակայքի մեթոդով (IQR). այս արժեքները չեն ներգրավվել հետագա հետազոտությունում ոչ ներկայացուցչական լինելու պատճառով:

2. Կեղծ փոփոխականների ստեղծում

Դիսկրետ փոփոխականները փոխարինվել են համապատասխան կեղծ փոփոխականներով (one-hot encoding):

3. Փոփոխականների նորմավորում

Բացատրվող շարունակական տվյալների համար իրականացվել է նորմավորում՝ կիրառելով պոլինոմիալ կամ լոգարիթմական ֆունկցիաներ. բաշխման նորմալությունը ստուգվել է Կոլմոգորով-Սմիրնով վիճակագրական թեստի միջոցով:

4. Ուսուցման-ստուգման տվյալների բաժանում (train-validation split)

Տվյալները պատահական սկզբունքով բաժանվել են ուսուցման-ստուգման խմբերի 70/30 հարաբերակցությամբ: Նպատակն է կատարել ուսուցում տվյալների 70 տոկոսի հիման վրա և ստուգել մոդելի արդյունավետությունը՝ օգտագործելով 30 տոկոսը:

Արդարության գաղափարը տնտեսագիտական-քաղաքագիտական ուսումնասիրություններում խմբավորվում է հետևյալ կերպ.

- Բաշխիչ արդարությունը մարդկանց կողմից ստացված բարիքների կամ արդյունքների հավասարությունն է:
- Ընթացակարգային արդարությունը մարդկանց կողմից բարիքների ստացման միանմանությունն է:
- Վերականգնող արդարությունը վերականգնում է կատարված վնասները:
- Հատուցող արդարությունը մեղավորներին պատժելն է:

---

[1] Ghalachyan G.. Multidimensional Evaluation of Binary Classification Models with Socio-Economic Metrics. Messenger of ASUE #6, 2021

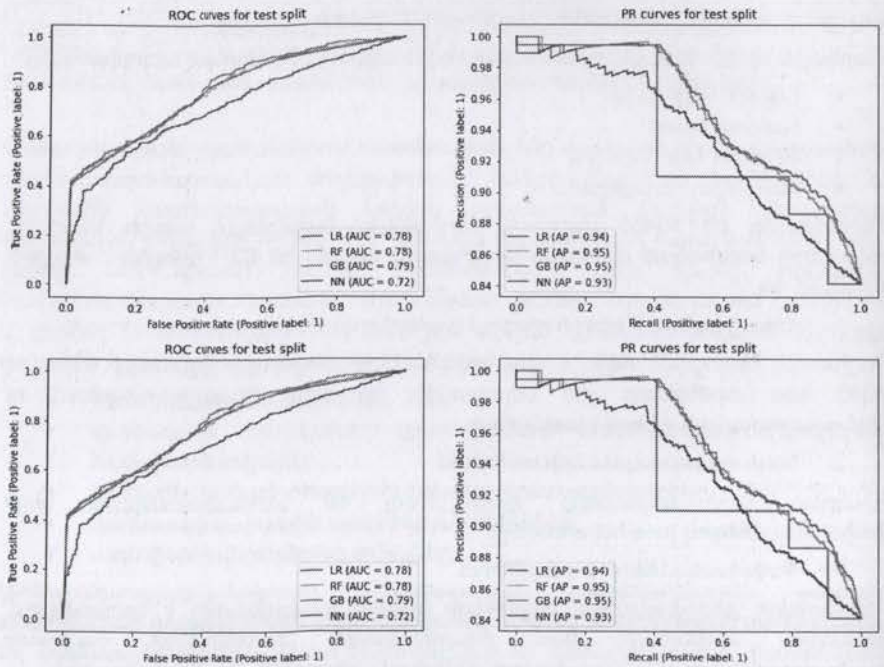
[2] Christopher M. Bishop: Pattern Recognition and Machine Learning, Chapter 4.3.4

[3] G. Louppe, "Understanding Random Forests: From Theory to Practice", PhD Thesis, U. of Liege, 2014.

[4] Friedman, J. H. "Stochastic Gradient Boosting" March 1999

[5] Popescu, Marius-Constantin & Balas, Valentina & Perescu-Popescu, Liliana & Mastorakis, Nikos. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems. 8.

© National Library of Armenia



Գծապատկեր 2: Մոդելների ճշգրտության գնահատումը ըստ ROC և PR ցուցանիշների ուսուցման և ստուգման տվյալներում

Մեքենայական ուսուցման մեջ կարևորվում է բաշխիչ արդարությունը, քանի որ կանխատեսման մոդելները ստեղծված են վերջնական արդյունքների ճշգրիտ տեղաբաշխման համար: Տեղաբաշխման համար կարևոր դասեր են համարվում սեռը, ռասան, հասարակական դասը, տարիքային խումբը, ֆիզիկական ունակությունը և այլն: Ուսումնասիրվում է բաշխման արդարության երկու տեսակ՝ խմբային և անհատական: Խմբային արդարությունը դասակարգչի պաշտպանված հատկանիշի երկու տարբեր խմբերում նույն վարքի դրսևորումն է: Անհատական արդարությունը երկու տարբեր, բայց միևնույն հատկանիշներով անձանց դասակարգումն է ըստ նույն դասերի:

Որպես խմբային արդարության քանակական ցուցանիշ օգտագործվել է տարբերակող ազդեցությունը (**disparate impact**):

$$DI = \frac{P(\hat{Y} = 1 | S \neq 1)}{P(\hat{Y} = 1 | S = 1)} > 1 - \varepsilon,$$

որտեղ

- S – պաշտպանված հատկանիշը (protected attribute),
- $S = 1$ – արտոնյալ խումբը (privileged group),

- $S \neq 1$ – ոչ արտոնյալ խումբը (unprivileged group),
- $\hat{Y} = 1$ – դրական արդյունքը (positive outcome),
- $\varepsilon \in [0,1]$ – թույլատրելի շեղումը:

Ցուցանիշ	Սեռ	Տարիք
Տարբերակող ազդեցություն	0.82214	0.919722

Աղյուսակ 1: DI ցուցանիշը պատմական տվյալների համար՝ սեռ և տարիք պաշտպանված փոփոխականի համար

Աղյուսակ 1-ում տեսնում ենք, որ նշված փոփոխականներում իրապես առկա է առկա է արդարության շեղում, մասնավորապես վարկառուի սեռի պարագայում:

Ցուցանիշ \ Մոդել	LR	RF	GB	NN
accuracy	0.759058	0.785483	0.845411	0.798913
true_positive_rate	0.791966	0.846069	0.98924	0.891679
true_negative_rate	0.583969	0.483511	0.480153	0.235344
disparate_impact	1.006646	0.755051	0.870591	0.854277
statistical_parity_difference	0.004851	-0.22781	-0.02922	-0.15622

Աղյուսակ 2: Գնահատված չորս մոդելների ճշգրտության ցուցանիշները և արդարության ցուցանիշները gender երկուական փոփոխականի նկատմամբ male արտոնյալ խմբով

Մեքենայական ուսուցման մոդելների վրա հարձակումները ուսումնասիրելիս կարևոր է դիտարկել 3 հարց՝

1. Համակարգի ո՞ր հատվածն է հարձակման ենթակա՝ ուսուցում, թե՞ տեղակայում:
2. Ի՞նչ հնարավորություններ ունի հարձակվողը, այն է՝ ի՞նչ տեղեկություններ են տվյալների և մոդելի մասին հայտնի, ի՞նչ տվյալներ և մոդելներ կարող են փոխվել և ի՞նչ եղանակով:
3. Ո՞րն է հակառակորդի նպատակը, օրինակ նմուշների դասակարգման շփոթում կամ կոնկրետ դասի ուղղորդում:

Առաջին հարցին պատասխանելով ստանում ենք, թե որն է հարձակման **թիրախը**: Հակառակորդները կարող են թիրախավորել կամ մոդելավորման փուլը կամ տեղակայման փուլը: Հարձակվելով մոդելավորման փուլի վրա՝ նրանք կարող են փչացնել ուսուցման տվյալները կամ մոդելը, որպեսզի այն լինի անհամապատասխան տեղակայման ժամանակ տեսած տվյալների նկատմամբ: Սրանք հայտնի են որպես **կեղծ նմուշների կամ թունավորման հարձակումներ (poision attack)**[6] և ունեն նմանություններ բաշխման տեղաշարժի հետ, քանի որ դրանք փոխում են ուսուցման տվյալների վիճակագրությունը:

[6] Singh M., Ghalachyan G., Kush R. , Reginald E. . An Empirical Study of Accuracy, Fairness, Explainability, Distributional Robustness, and Adversarial Robustness. RAI@KDD21, 2021. https://arxiv.org/abs/2109.14653

© National Library of Armenia

Մոլորեցման կամ խուսափող (evasion) հարձակումները, որոնք ուղղված են տեղակայման փուլին, այլ դասի ալգորիթմներ են, որոնք չունեն անմիջական ազդեցություն բաշխման տեղաշարժին, բայց ունեն անուղղակի նմանություն նախորդ ենթագլխում սահմանված անհատական արդարության հետ: Այս հարձակումները ուղղված են առանձին նմուշների փոփոխմանը, որոնք մեքենայական ուսուցման համակարգը պետք է գնահատի:

Հաջորդ հարցի պատասխանը ցույց է տալիս հակառակորդի **կարողությունը** հարձակում իրականացնելու խնդրում. որոշ հակառակորդներ ավելի մեծ հնարավորություններ ունեն, քան մյուսները: Կեղծ նմուշների հարձակումների դեպքում հակառակորդները փոխում են ուսուցման տվյալները կամ մոդելը ինչ-որ կերպ, ենթադրելով, որ նրանք ունեն հասանելիություն տեխնոլոգիական ենթակառուցվածքին: Սա մեծապես կախված է մեքենայական ուսուցման ենթակառուցվածքի անվտանգությունից, սակայն ամենահեշտ ու կիրառվող եղանակն է հանդես գալ որպես օգտատեր ու ստեղծել նոր նմուշներ. սա հայտնի է որպես **տվյալների ներարկում (data injection)**: Ավելի դժվար է **տվյալների փոփոխումը (data modification)**, որում հակառակորդը փոխում է պիտակները կամ առանձնահատկությունները առկա ուսուցման տվյալների բազայում: Ամենադժվարը սակայն **տրամաբանության կեղծումն (logic corruption)** է, որի դեպքում հակառակորդը փոխում է մեքենայական ուսուցման ալգորիթմի կոդը, վարքագիծը կամ մոդելը: Տվյալների ներարկման և տվյալների փոփոխման հարձակումները նախամշակման ընթացքում կատարվող հարձակումներ են, իսկ տրամաբանական կոռուպցիան մոդելի ուսուցման և դրանից հետո կատարվող հարձակումներ:

Երրորդ հարցի պատասխանը մեզ տալիս է տեղեկություն, թե որն է հարձակվողի **նպատակը**, որը վերաբերում է և՛ կեղծ նմուշների, և՛ մոլորեցման հարձակումներին: Տարբեր հակառակորդներ փորձում են տարբեր նպատակների հասնել: Ամենահեշտ նպատակը **վստահության նվազեցումն (confidence reduction)** է՝ դասակարգիչը գնահատում ելքային փոփոխականը այնպես, որ տարբեր դասերի հանդես գալու հավանականությունը լինեն հավասար, այն է՝ էնթրոպիան լինի առավելագույնը: Հաջորդ նպատակը **սխալ դասակարգումն (misclassification)** է՝ փորձելով ստիպել դասակարգչին սխալ կանխատեսումներ անել, որն էլ իր հերթին կարող է լինել **թիրախավորված (targeted)**՝ դասակարգչին ուղղորդի կոնկրետ դասի կանխատեսման, և **չթիրախավորված (not targeted)**:

Ներկայացնենք կեղծ նմուշների ստեղծման գրադիենտային մեթոդի մաթեմատիկական տեսքը $h_\theta: X \to \mathbb{R}^k$ k-ական դասակարգչի համար, որը ունի $\ell: \mathbb{R}^k \times \mathbb{Z}_+ \to \mathbb{R}_+$ օպտիմիզացիոն ֆունցիան: Այն ունի հետևյալ տեսքը՝

$$\ell(h_\theta(x), y) = log(\sum_{j=1}^{k} exp(h_\theta(x)_j)) - h_\theta(x)_y:$$

Լավագույն θ պարամետրեր գտնելու համար նվազեցնում ենք $\ell(h_\theta(x_i), y_i)$ կորստի ֆունկցիան ըստ θ-ի՝ $\underset{\theta}{minimize} \frac{1}{m} \sum_{i=1}^{m} \ell(h_\theta(x_i), y_i)$, որը սովորաբար լուծում ենք ըստ նվազող ստոխաստիկ գրադիենտների մեթոդի կամ դրա այլ մոդիֆիկացիայի՝ յուրաքանչյուր քայլում թարմացնելով θ-ն կորստի գրադիենտի ինչ-որ չափով՝

$$\theta := \theta - \frac{\alpha}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \nabla_\theta \ell(h_\theta(x_i), y_i),$$

որտեղ $\mathcal{B}$ – ն տվյալ քայլում օգտագործվող ընտրանքն է: Սակայն խնդիրը վերաձևակերպվում է հակառակորդային ուսուցման ժամանակ հետևյալ տրամաբանությամբ. մաքսիմալացնել կորստի ֆունցիան ըստ կեղծ նմուշի այնպես, որ կանխատեսված և իրական ելքային արժեքների տարբերությունը լինի առավելագույնը՝

$$\underset{x}{maximize}\ \ell(h_\theta(\hat{x}), y),$$

Թիրախավորված հարձակման ժամանակ, երբ ցանկանում ենք, որ կեղծ նմուշը վերադարձնի հստակ արժեք, օպտիմիզացիոն ֆունկցիան կսահմանենք հետևյալ կերպ:

$$\underset{\delta \in \Delta}{maximize}(\ell(h_\theta(x+\delta), y) - \ell(h_\theta(x+\delta), y_{target}))$$

Գնահատված մոդելների համար HopSkipJump[7] հարձակման միջոցով ստացել ենք 300-ական կեղծ նմուշներ, ապա հաշվարկել Empirical Robustness ցուցանիշը:

$$Empirical\ Robustness = \frac{1}{n} \frac{\|X - X'\|_2}{\|X\|_2},$$

որտեղ

- X-ն իրական նմուշների մատրիցն է,
- X'-ը կեղծված նմուշների մատրիցն է:

մոդել	LR	RF	GB	NN
Empirical Robustness	1	.1524	.2411	.0905

Աղյուսակ 3: Empirical Robustness ցուցանիշի արժեքները գնահատված մոդելների համար (արժեքները նորմավորված են LR մոդելի նկատմամբ)

Լոջիթ մոդելը, պայմանավորված գծային բնույթով, պաշտպանված է նման հարձակումներից: Ի հակադրումն՝ նեյրոնային ցանցերում որոշումներ կայացման սահմանը ունի բարդ պոլինոմիալ մոտարկման տեսք, և կեղծ նմուշները իրական նմուշներին գտնվում են շատ մոտ: Պատահական անտառ և խթանիչ գրադիենտներ մոդելները փոքր ինչ ավելի պաշտպանված են նմանատիպ դեպքերից, սակայն դա հետևանք է նմուշների ընտրման պատահական բնույթի:

Ատենախոսության երրորդ՝ **«Արհեստական բանականության սոցիալ-տնտեսական ցուցանիշների շեղումների մեղմացման ալգորիթմներ»** գլխում ներկայացված են մոդելների շեղման մեղմացման ալգորիթմները, դրանց կիրառությունը վարկային ռիսկի մոդելավորման խնդրում, տնտեսամաթեմատիկական մեկնաբանությունը:

[7] Classifiers C. J. Simon-Gabriel, N. Sheikh, A. Krauseln. PopSkipJump: Decision-Based Attack for Probabilistic Proc. International Conference on Machine Learning (ICML), 2021

© National Library of Armenia

Վերակշռման (reweighing) մեթոդը[8] դասեր-պաշտպանված հատկանիշ խմբերից յուրաքանչյուրի նմուշներին տալիս է կշիռներ հետևյալ մեթոդով՝

$$W_{positive\ privileged} = \frac{N_{privileged} * N_{positive}}{N_{all} * N_{positive\ privileged}},$$

որտեղ

- $N_{privileged}$-ը ուսուցման տվյալներում արտոնյալ խմբին պատկանող նմուշների քանակն է,
- $N_{positive}$ը ուսուցման տվյալներում ելքային փոփոխականի դրական դասին պատկանող նմուշների քանակն է,
- N_{all}-ը ուսուցման տվյալներում առկա նմուշների քանակն է,
- $N_{positive\ privileged}$-ը ուսուցման տվյալներում արտոնյալ խմբին պատկանող նմուշների քանակն է, որոնք պատկանում են ելքային փոփոխականի դրական դասին
- $W_{positive\ privileged}$-ը դրական արտոնյալ դասին տրվող կշռի մեծությունն է:

Մոդելների պաշտպանվածությունը անմիջականորեն կապված է դրա կանխատեսման ճշգրտության ռիսկից՝ $R(h_\theta) = \boldsymbol{E}_{(x,y)\sim\mathcal{D}}[\ell(h_\theta(x)), y)]$, որտեղ $\mathcal{D}$-ն ցույց է տալիս նմուշների իրական բաշխումը: Սակայն քանի որ մեզ հասանելի է տվյալների նմուշներ՝ $D = \{(x_i, y_i) \sim \mathcal{D}\}, i = 1, \dots, m$, ուստի կօգտագործենք ռիսկի էմպիրիկ գնահատականը, իսկ մեքենայական ուսուցման խնդիրը կլինի ուսուցման տվյալների վրա մինիմալացնել այդ ռիսկը՝

$$\hat{R}(h_\theta, D) = \frac{1}{|D|} \sum_{(x,y)\in D} \ell(h_\theta(x)), y),$$

$$\underset{\theta}{minimize}\ \hat{R}(h_\theta, D_{train}).$$

Ճշգրտության ռիսկի այլընտրանք է հակառակորդային ռիսկը (adversarial risk) և դրա գնահատականը՝

$$R_{adv}(h_\theta) = \boldsymbol{E}_{(x,y)\sim\mathcal{D}}[\max_{\delta\in\Delta(x)} \ell(h_\theta(x+\delta)), y)],$$

$$\hat{R}_{adv}(h_\theta, D) = \frac{1}{|D|} \sum_{(x,y)\in D} \max_{\delta\in\Delta(x)} \ell(h_\theta(x+\delta)), y):$$

Հակառակորդային ռիսկը ցույց է տալիս, թե նմուշի նվազագույն շեղումը թույլատրելի δ շրջակայքում ինչքանով է ազդում իրական և կանխատեսված արժեքների տարբերությանը: Հակառակորդը փորձում է մաքսիմալացնել հակառակորդային ռիսկը հնարավորինս փոքր դելտա շրջակայքում և ստանալ հակառակորդ նմուշներ: Պաշտպանության տեսանկյունից մենք փորձում ենք միաժամանակ լուծել 2 խնդիր՝ ստանալ բարձր ճշգրտություն և բարձր անվտանգություն, ինչը համապատասխանում է ճշգրտության և հակառակորդային ռիսկի միաժամանակյա մինիմալացման: Մաթեմատիկական բանաձևումը հետևյալն է՝

[8] F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification without Discrimination," Knowledge and Information Systems, 2012.

$$\underset{\theta}{minimize}\ \hat{R}_{adv}(h_\theta, D_{train}) \equiv \underset{\theta}{minimize} \frac{1}{|D_{train}|} \sum_{(x,y)\in D_{train}} \max_{\delta\in\Delta(x)} \ell(h_\theta(x+\delta)), y):$$

Նշված $min - max$ օպտիմիզացիայի խնդիրը կոչվում է **հակառակորդային ուսուցում** և կարելի է հեշտությամբ լուծել կիրառելով Դանկինսի թեորեմը[9,10]: Չնայած տեսական նմանությանը՝ հակառակորդ հարձակումների տեսակների համար պաշտպանության մեթոդները տարբերվում են գործնականում կիրառելիս:

Որոշումների կայացման համակարգերում կարևոր նշանակություն ունի կայացրած **որոշման բացատրությունը**: Այն կարևորագույն հատկանիշ է ինչպես մարդկանց կողմից կայացրած որոշումներում, այնպես էլ ավտոմատացված համակարգերում, և բացառություն չեն մեքենայական ուսուցման մոդելները: Տնտեսամաթեմատիկական մոդելավորման պարզագույն խնդիրների լուծման եղանակները բացատրելի են, օրինակ՝ գծային ծրագրում կամ բազմագործոն ռեգրեսիա: Նման մոդելների գնահատման արդյունքում ստանում ենք պարամետրեր, որոնք օգտագործվում են և՛ նոր իրավիճակների գնահատման, և՛ դրանց բացատրության համար: Տնտեսամաթեմատիկական կանխատեսման մոդելը կոչվում է բացատրելի, եթե դրա կանխատեսումները կարելի է ներկայացնել մուտքային փոփոխականի նախնական տեսքի միաքայլ ձևափոխության միջոցով: Առավել բարդ մոդելները, որոնցից երեքը քննարկել ենք այս աշխատանքում, բացատրելի չեն: Այսպիսով առաջանում է խնդիր ստեղծելու օժանդակ մոդելներ, որոնք կբացատրեն կանխատեսման մոդելի արդյունքները:

Բացատրությունների առանձնացումը մեքենայական ուսուցման հիմնական մոդելից (model-agnostic explaination models) ունի որոշ առավելություններ , օրինակ՝ նրանց ճկունությունը: Մեքենայական ուսուցման համակարգեր մշակողները ազատ են օգտագործել մեքենայական ուսուցման ցանկացած դասի մոդել՝ գծային, անսամբլային, նեյրոնային ցանցեր և այլն, և տեխնիկապես նպատակահարմար է ունենալ մեկնաբանման մեթոդներ, որ կարող են կիրառվել ցանկացած մոդելի վրա: Պատճառը հետևյալն է. որպես կանոն, մեքենայական ուսուցման մոդելների ոչ թե մեկ, այլ շատ դասեր են գնահատվում առաջադրանքը լուծելու համար, և մոդելները բացատրելիության տեսանկյունից համեմատելիս ավելի հեշտ է աշխատել մոդել-ագնոստիկ բացատրությունների հետ, քանի որ նույն մեթոդը կարող է օգտագործվել ցանկացած դասի համար: **Տեղային փոխնակ մոդելները** (local surrogate models) ահմանվում են որպես մեկնաբանելի մոդելներ, որոնք օգտագործվում են մեքենայական ուսուցման մոդելների անհատական կանխատեսումները բացատրելու համար: **Տեղային մեկնաբանելի մոդել-ագնոստիկ բացատրություններ (LIME)** մոդելի հեղինակներն առաջարկում են տեղային փոխնակ մոդելների հետևյալ տարբերակը. փոխնակ մոդելի ուսուցման հիմքում ընկած է չմեկնաբանվող մոդելի կանխատեսումները մոտարկելու գաղափարը: Այստեղ հաշվի չեն առնվում ուսուցման տվյալները և օգտագործվում է միայն մեքենայական ուսուցման այդ տվյալների վրա սովորած մոդելը, որտեղ կարող ենք տեղադրել մուտքային տվյալների արժեքները և

[9] Danskin, John M. The theory of Max-Min and its application to weapons allocation problems New York: Springer. 1967

[10] Bertsekas, Dimitri P. Nonlinear programming (Second ed.). Belmont, Massachusetts. 1999

© National Library of Armenia

ստանալ մոդելի կանխատեսումները: Այս գործողությունը կարող ենք կատարել վերջավոր քանակությամբ, և նպատակն է հասկանալ, թե ինչու է մեքենայական ուսուցման մոդելը որոշակի կանխատեսում արել: LIME-ը փորձարկում է, թե ինչ է տեղի ունենում կանխատեսումների հետ, երբ կատարում ենք մուտքային տվյալների որոշակի տատանումներ մեքենայական ուսուցման մոդելում: Ապա այն ստեղծում է նոր տվյալների բազա, որը բաղկացած է կեղծված նմուշներից և չբացատրվող մոդելի համապատասխան կանխատեսումներից: Այս նոր տվյալների վրա LIME-ն այնուհետև պատրաստում է մեկնաբանելի մոդել, որը կշռվում է ընտրված նմուշների նմանությանբ. մեկնաբանելի մոդելը կարող է լինել կամայական մեկնաբանելի մոդել: LIME մոդելի բացատրությունը ունի հետևյալ մաթեմատիկական ձևակերպումը՝

$$explanation(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g),$$

որտեղ

- x-ը բացատրվող նմուշն է,
- f-ը՝ մեքենայական ուսուցման չմեկնաբանվող մոդելը,
- g-ն՝ մեկնաբանելի մոդելը,
- π_x-ը՝ բացատրվող նմուշի շրջակայքի չափը,
- G-ն՝ հնարավոր մեկնաբանելի մոդելների բազմությունը,
- $\Omega(g)$-ն՝ բացատրվող մոդելի բարդությունը,
- L-ը՝ օպտիմիզացիոն ֆունկցիան:

Մոդելների համար բացատրություններ տալուց զատ անհրաժեշտ է դրա քանակական ներկայացումը: Faithfulness[11] (հավատարմություն) ցուցանիշը գնահատում է գծային կորելյացիա մուտքային փոփոխականի՝ բացատրելիության ալգորիթմի կողմից տրված կարևորության և կանխատեսող մոդելում տվյալ հատկանիշի ազդեցության միջև: Առավել բարձր կորելյացիայի բացարձակ արժեքը ցույց է տալիս, որ տվյալ մոդելը բացատրելի է, այն է՝ նշված կանխատեսման և բացատրության մոդելների տրամաբանությունները մոտ են: LIME բացատրության մոդելի և Faithfulness ցուցանիշի համադրմամբ կառուցված կանխատեսամն մոդելների համար ստացել ենք բացատրելիության ցուցանիշի գնահատական:

Հետազոտության էմպիրիկ վերլուծությունը կատարվել է հետևյալ քայլերով[12]՝

1. խնդրի տնտեսամաթեմատիկական սահմանում,
2. նախնական տվյալների համար արհեստական բանականության մոդելների կառուցում և ուսուցում,
3. արհեստական բանականության մոդելների ճշգրտության և էթիկական ցուցանիշների գնահատում,
4. արհեստական բանականության մոդելների էթիկական ցուցանիշների շեղման մեղմացում,

[11] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Advances in Neural Information Processing Systems 31, pages 7775-7784. 2018.

[12] Ղալաչյան Գ.: Արհեստական բանականության մոդելների կիրառությունը ֆինանսներում (ՀՀ ՈՒՎԿ օրինակով): Բանբեր Երևանի համալսարանի. Տնտեսագիտություն #3(36), 2021, էջ 73-81

5. մշակված տվյալների համար արհեստական բանականության մոդելների կառուցում և ուսուցում:

Ստացված չորս մոդելների ճշգրտության և էթիկական չափանիշների միաժամանակյա գնահատման համար կառուցել ենք ընդհանրացված (**aggregate** – AGG) կշռված ցուցանիշ, ըստ հետևյալ առանձին ցուցանիշների՝

- ճշգրտություն – **accuracy** – ACC,
- տարբերակող ազդեցություն – **disparate impact** – DI,
- էմպիրիկ անշեղություն – empirical **robustness** – ER,
- հավատարմություն – **faithfulness** – FF:

Ընդհանրացված ցուցանիշը կառուցվել է հետևյալ եղանակով՝

$$AGG = 0.7 * ACC + 0.1 * DI + 0.1 * ER + 0.1 * FF:$$

Մոդել / Ցուցանիշ	LR	RF	GB	NN	կշիռ
ACC	0.76	0.79	0.85	0.8	0.7
DI	1	0.76	0.87	0.86	0.1
ER	1	0.15	0.24	0.09	0.1
FF	0.84	0.75	0.68	0.22	0.1
AGG	*0.816*	*0.719*	*0.774*	*0.677*	

Աղյուսակ 4: Ընդհանրացված ցուցանիշի գնահատականները նախնական տվյալների համար կառուցված մոդելների համար (արդարության ցուցանիշի համար վերցված է «սեռ» պաշտպանված հատկանիշը)

Մոդել / Ցուցանիշ	LR	RF	GB	NN	կշիռ
ACC	0.7	0.75	0.84	0.8	0.7
DI	1	0.84	0.94	0.92	0.1
ER	1	0.15	0.86	0.54	0.1
FF	0.86	0.72	0.7	0.44	0.1
AGG	*0.776*	*0.696*	*0.824*	*0.75*	

Աղյուսակ 5: Ընդհանրացված ցուցանիշի գնահատականները փոփոխված տվյալների համար հակառակորդային ուսուցմամբ կառուցված մոդելների համար (արդարության ցուցանիշի համար վերցված է «սեռ» պաշտպանված հատկանիշը)

© National Library of Armenia

Ատենախոսության «**Եզրակացություններ**» բաժնում ամփոփվել են հետազոտության արդյունքում ստացված եզրահանգումները, որոնք ներկայացված են ստորև:

- ՈՒՎԿ կողմից տրամադրված տվյալների միջոցով սահմանել ենք գյուղատնտեսական վարկի ռիսկայնության մակարդակը. տվյալ վարկի վճարման ուշացման օրերի հիման վրա. վարկի չորս և ավելի օրերի ուշացումը սահմանել ենք ռիսկային, իսկ երեք և քիչ օրերը՝ ոչ ռիսկային:
- Կիրառված արհեստական բանականության մոդելները մեծ ճշգրտությամբ կարող են կանխատեսել գյուղատնտեսական վարկերի ռիսկայնությունը՝ օգտագործելով երկուական դասակարգման ալգորիթմներ: Գնահատված չորս մոդելներից ճշգրտության լավագույն ցուցանիշ ունի Gradient Boosting մոդելը (~85%), որը բացատրվում է մոդելի սխալ կանխատեսումներից սովորելու տրամաբանությամբ. նմանատիպ մոդելները կարող են ապահովել բարձր արդյունավետություն՝ օգտագործելով քիչ տվյալներ:
- Ի հակադրումն նախորդի՝ նեյրոնային ցանցերի հիման վրա կառուցված մոդելների ընդհանրացումը սակավ կիրառելի է ստուգման տվյալների համար՝ ~80% ճշգրտություն: Մյուս կողմից Logistic Regression և Random Forest մոդելների վատ ցուցանիշները բացատրվում են համապատասխանաբար մոդելի գծային բնույթով և ենթակա մոդելներում փոփոխականների ընտրության պատահական բնույթով:
- Բացի ճշգրտության ցուցանիշներից կարևոր է մոդելների արդյունավետության գնահատումը ըստ էթիկական չափանիշների՝ արդարություն, բացատրելիություն, անվտանգություն: Քանի որ մոդելի ավանդական ուսուցման ընթացքում նպատակ է դրվում զուտ բարձր ճշգրտություն ստանալու, էթիկական ցուցանիշների շեղումները հնարավոր է կանխել մոդելի ուսուցման ավարտից հետո մեքենայական ուսուման համակարգի կառուցման ցիկլի երեք տարբեր հատվածներում՝ տվյալների նախնական մշակում, ալգորիթմի ուսուցուցում, հետմշակում:
- Մոդելի արդարության համար սահմանվում են խմբային և անհատական ցուցանիշներ: Խմբային ցուցանիշները դիտարկվում են որևէ սոցիալապես զգայուն փոփոխականի նկատմամբ՝ սեռ, ռասա, տարիք և այլն. նպատակն է ունենալ մոդել որը նման կանխատեսումներ կանի տվյալ փոփոխականի տարբեր խմբերի համար: Հետազոտությունում դիտարկվել էր վարկառուի սեռը և տարիքային խումբը որպես զգայուն փոփոխական: Պատմական տվյալներում ավելի մեծ շեղում նկատվել է սեռի պարայագայում արական սեռ արտոնյալ խմբով: Կիրառված նախամշակման ալգորիթմով ելքային փոփոխական-զգայուն փոփոխական խմբերի համար սահմանվել են կշիռներ, ապա այն օգտագործելով ուսուցման օպտիմիզացիոն ֆունկցիայում՝ GB մոդելի DI ցուցանիշը բարձրացրել ենք 87%-ից 93%՝ տալով ճշգրտության ընդամենը 1%-ի կորուստ:
- Բացատրելիության ցուցանիշները առաջնային են այնպիսի համակարգերի համար, որտեղ կանխատեսման արժեքները առաջնային ազդեցություն ունեն կյանքի որակի վրա: Գյուղատնտեսական վարկերի պարագայում, ի հավելումն բարձր ճշգրտության մոդելների, առաջարկում ենք կիրառել տեղային մոդել-ագնոստիկ բացատրության մոդել. նրա արդյունքները ցույց են տալիս նմուշի

յուրաքանչյուր փոփոխականի ազդեցությունը որոշման կայացման համար: Մեքենայական մոդելի բացատրելությունը չափվել է faithfulness ցուցանիշով:
- Անվտանգության տեսանկյունից մեքենայական ուսուցման մոդելների խոցելիությունը մեծապես կախված է դրա ենթակառուցվածքից և մոդելի տեղակայումից. մոդելի մոլորեցման համար ստեղծվում են կեղծ նմուշներ, ապա դրանք կիրառվում ցանկալի արդյունք ստանալու համար: Ուսումնասիրության արդյունքում պարզել ենք, որ գծային մոդելները առավել անվտանգ են, քան նեյրոնային ցանցերը և անսամբլները, այնուհանդերձ՝ կիրառելով հակառակորդային ուսուցում և տվյալների հարթեցում այս մոդելների դասի համար ևս հնարավոր է ապահովել անվտանգություն պայմանները: Մոդելների անվտանգությունը չափվել empirical robustness ցուցանիշով:
- Ճշգրտության և էթիկական ցուցանիշների հիման վրա կառուցված ագրեգատ ցուցանիշը ցույց է տալիս նրա արդյունավետության ընդհանուր մակարդակը:

Ագրեգատը նշված ցուցանիշների նորմալացված և կշռված գումարն է. կշիռները որոշվում են ըստ առանձին ցուցանիշի կարևորության:

Ամփոփելով կարող ենք ասել, որ ստացել ենք արհեստական բանականության արդյունավետություն ագրեգատ ցուցանիշ: Ինչպես նաև պարզել ենք, որ գյուղատնտեսական վարկերի ռիսկի վրա առավելապես ազդում են նախորդ վարկերի ուշացման օրերը, տնտեսության ընդհանուր ամսական եկամուտը, տոկոսադրույքի ծածկման գործակիցը:

Ատենախոսության հիմնական դրույթները հրապարակվել են հեղինակի յոթ գիտական հոդվածներում՝

1. Ղալաչյան Գ.: Մեքենայական ուսուցման կիրառությունը ՀՀ ֆինանսական ռիսկի կառավարման ոլորտում: Ֆինանսներ և էկոնոմիկա #8 (216), 2018, էջ 7-9
2. Ղալաչյան Գ.: Արհեստական բանականության տնտեսագիտական ընկալումը և կիրառումը պետական կառույցներում: Պատմություն և քաղաքականություն #2(13), 2021, էջ 153-161
3. Ghalachyan G.. Defining and Detecting Fairness Bias for Binary Classification Problem in Financial Analysis. Scientific Artsakh #2 (9), 2021, pp. 183-191
4. Ղալաչյան Գ.: Արհեստական բանականության բացատրելիությունը և կիրառման իրավական պահանջները: Պատմություն և քաղաքականություն #6(17), 2021, էջ 140-147
5. Ղալաչյան Գ.: Արհեստական բանականության մոդելների կիրառությունը ֆինանսներում (ՀՀ ՈՒՎԿ օրինակով): Բանբեր Երևանի համալսարանի. Տնտեսագիտություն #3(36), 2021, էջ 73-81
6. Ghalachyan G.. Multidimensional Evaluation of Binary Classification Models with Socio-Economic Metrics. Messenger of ASUE #6, 2021
7. Ghalachyan G.. Data Interchange Biases and Its Impact on Algorithmic Fairness. History and Politics #1(18), 2022, pp 89-98

© National Library of Armenia

ГЕВОРГ ОВСЕПОВИЧ КАЛАЧЯН
ОЦЕНКА ЭФФЕКТИВНОСТИ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В СОЦИАЛЬНО-ЭКОНОМИЧЕСКИХ ЗАДАЧАХ

Автореферат диссертации на соискание ученой степени кандидата экономических наук по специальности: Ը.00.08 - «Математическое моделирование экономики»

Зашита диссертации состоится 24-го июня 2022 г. в 15:00 часов на заседании Специализированного совета по экономике 015 ВАК РА, действующего в Ереванском государственном университете.
Адрес: 0009, г. Ереван, ул. Абовяна 52.

РЕЗЮМЕ

Применение систем искусственного интеллекта широко распространено в различных социально-экономических задачах, поскольку растущие вычислительные мощности позволяют обучать все более точные модели. Тем не менее, есть опасения по поводу его эффективности в отношении этических аспектов — справедливости, безопасности, объяснимости. Сложные нелинейные модели имеют тенденцию быть предвзятыми и не следовать социальным нормам, таким как несправедливые решения по отношению к социальным группам, уязвимость к враждебным атакам, неспособность объяснить решения. Основная цель этого исследования - оценить модель искусственного интеллекта с точки зрения надежности и смягчить продемонстрированные предубеждения.

Для достижения цели были поставлены и решены следующие задачи:

- изучить классы моделей бинарной классификации, методологию ее оценки с точки зрения точности и применения финансовыми учреждениями;
- оценить различные классы моделей для модели прогнозирования кредитных дефолтов, используя набор данных одобренных сельскохозяйственных кредитов от универсальной кредитной организации, зарегистрированной и действующей в РА;
- изучить прежние исследования в области надежного искусственного интеллекта, законодательных норм о справедливых, безопасных и интерпретируемых системах принятия решений, а также, что наиболее важно, методологию оценки объективности, объяснимости и устойчивости модели бинарной классификации;
- оценить точность и достоверность оценочных моделей и предложить конкретную методологию для решения проблемы прогнозирования дефолта по кредиту;
- интерпретировать алгоритмы смягчения социально-экономических отклонений и применять методы для оценочных моделей;

- создавать совокупную метрику моделей искусственного интеллекта на основе исходных и скорректированных результатов оценки моделей и сравнивать производительность различных алгоритмов.

Основные результаты исследования и научная новизна заключаются в следующем

- Были предложены особенности кредитоспособности: поскольку выборка кредитов была репрезентативной по РА, мы констатируем, что сельскохозяйственный кредитный риск в РА в основном зависит от предыдущих дефолтов, общего дохода домохозяйства, коэффициента покрытия процентов.
- Были обнаружены проблемы с методологией сбора и обработки данных для сотрудничающего УКО, и были сделаны соответствующие предложения на основе статистического моделирования и показателей значимости.
- На основе четырех отдельных показателей производительности модели – точности, разрозненное воздействия, эмпирической надежности и достоверности, а также был введен новый взвешенный совокупный показатель: модели повышения градиента показали наименьшую компенсацию точности за достоверность.

© National Library of Armenia

GEVORG HOVSEP GHALACHYAN

EVALUATION EFFICIENCY OF ARTIFICIAL INTELLIGENCE MODELS IN SOCIO-ECONOMIC PROBLEMS

The abstract of the Dissertation for pursuing the degree of PhD in Economics in the field Ը.00.08 – "Mathematical Modeling of Economy"

The defense of the Dissertation will take place on June 24, 2022, at 15:00 at the meeting of the Specialized Council 015 in Economics of the Supreme Certifying Committee of the Republic of Armenia acting at the Yerevan State University.
Address: 52 Abovyan St., Yerevan, 0009

ABSTRACT

Application of artificial intelligence systems is widespread in different socio-economic problems as the growing computing power allows to train more and more accurate models. Yet there is a concern about its performance regarding the ethical aspects – fairness, security, interpretability. Complex non-linear models tend to be biased and not to follow social norms, such as unfair decisions towards social groups, vulnerability to adversarial attacks, inability to explain decisions. The main objective of this research is to evaluated artificial intelligence model in terms of trustworthiness and mitigate the demonstrated biases.

To achieve the objective the following tasks were defined and solved:

- study the classes of binary classification models, its evaluation methodology in terms of accuracy and application by financial institutions;
- estimate different classes of models for credit default prediction model using the dataset of approved agricultural loans from universal credit organization registered and operating in RA;
- study the former research on trustworthy artificial intelligence, legislative norms of fair, secure and interpretable decision systems, most importantly methodology for evaluation of binary classification model fairness, explainability and adversarial robustness;
- evaluate the accuracy and trustworthiness of the estimated models and suggest specific methodology for the credit default prediction problem;
- interpret the algorithms for socio-economic bias mitigation, and apply the methods for the estimated models;
- create an aggregate metric of artificial intelligence models based on initial and adjusted results of model evaluation, and compare the performance of different algorithms.

The following findings constitute the primary results and the main scientific novelty of dissertation:

- Features of credit solvency were suggested: as the sample of loans were representative across RA we state that agricultural credit risk in RA mainly depends on previous defaults, total income of household, interest coverage ratio.
- Data collection and processing methodology issues were discovered for the collaborating UCO, and respective suggestions were made based on the statistical modeling and significance measures
- Based on four separate metrics of model performance – accuracy, disparate impact, empirical robustness and faithfulness, and new weighted aggregate metric was introduced, and following the evaluated model results gradient boosting models showed least compensation of accuracy for trustworthiness.

© National Library of Armenia

ՀՀ Ազգային գրադարան
NL1823696

© National Library of Armenia

© National Library of Armenia