ԲԱՆԲԵՐ ԵՐԵՎԱՆԻ ՀԱՄԱԼՍԱՐԱՆԻ. ՏՆՏԵՍԱԳԻՏՈՒԹՅՈՒՆ
2021, № 3, 73-81 *Տնտեսամաթեմատիկական մոդելավորում*
https://doi.org/10.46991/BYSU:G/2021.12.3.073

ԱՐՀԵՍՏԱԿԱՆ ԲԱՆԱԿԱՆՈՒԹՅԱՆ ՄՈԴԵԼՆԵՐԻ ԿԻՐԱՌՈՒԹՅՈՒՆԸ ՖԻՆԱՆՍՆԵՐՈՒՄ (ՀՀ ՈՒՎԿ ՕՐԻՆԱԿՈՎ)

ԳԵՎՈՐԳ ՂԱԼԱՉՅԱՆ

Նորագույն տեխնոլոգիաները և մեքենայական ուսուցման ժամանակակից ալգորիթմները միկրո ֆինանսական կազմակերպությունների համար տվյալների կիրառության նոր հնարավորություններ են տալիս: Այս հոդվածում ներկայացնում ենք մեթոդներ, որոնք կարող են օգտագործվել նման կազմակերպություններում առավել արդյունավետ ֆինանսական պլանավորման համար, և դրանք կիրառել ՀՀ-ում գործող ՈՒՎԿ-ի օրինակով:

Նման մեթոդները թույլ են տալիս նվազագույն ծախսով ունենալ կանխատեսման բարձր ճշգրտության մոդելներ: Ցույց ենք տվել, որ առաջարկվող մոտեցումները լուծում են մոդելների մեկնաբանության խնդիրը և տալիս են փոփոխականների բացատրություններ բինար կլասիֆիկացիայի խնդրում: Նաև ներկայացրել ենք ալգորիթմ, որը ստեղծում է տվյալների լատենտ տարածություն վարկային դիմումների սեգմենտավորման համար:

Բանալի բառեր - *արհեստական բանականություն, մեքենայական ուսուցում, փոփոխականի դիսկրետիզացիա, չափողականության կրճատում, նմուշների ճանաչում, լատենտ տարածություն*

Ներածություն

Արհեստական բանականությունը առաջին անգամ սահմանել է Ջոն Մաքքարթին 1959 թ.՝ որպես «բանական մեքենաների ստեղծման և նախագծման գիտություն[1]»: Սակայն տվյալ ժամանակաշրջանում տեսական հետազոտությունները գործնականում կիրառելի չէին, և նորանոր տեխնոլոգիական հաջողությունները փոխում էին պատկերացումները արհեստական բանականության մասին: Բանականությունը սահմանվում է որպես պատճառային կապերի բացահայտման, սովորելու և խնդիրներ լուծելու ընդհանուր մտավոր կարողություն, իսկ արհեստական բանականությունը՝ մեքենաների կողմից նման վարքագծի դրսևորում[2]: Հետևաբար սահմանումները տարբերվում են ըստ ժամանակաշրջանի տեխնոլոգիական հնարավորությունների, առկա խնդիրների, առաջարկվող լուծումների և այլն:

[1] Տե՛ս **Newell, A., Shaw, J.C. and Simon, H. A**. Report on a general problem-solving program 1959. Proceedings of the International Conference on Information Processing, էջ 256-264:

[2] Տե՛ս **David L. Poole,** Computational intelligence 1998, Oxford University Press, էջ 174:

Վերջին տարիներին մեծացել է մեքենայական ուսուցման և արհեստական բանականության դերը ֆինանսական ծառայություններում, որպեսզի դրանց կիրառությունը բարձրացնի ռիսկերի կառավարման արդյունավետությունը: Մեծ է նաև սպառողների վերաբերյալ տվյալների ծավալը, որ հավաքագրում են ֆինանսական կազմակերպությունները՝ հետևյալ երկու գործոնների հաշվառմամբ: Նախ և առաջ, պայմանավորված օրենսդրությամբ և կանոնադրություններով, կազմակերպությունները պարտավորվում են ավելի շատ տվյալներ հավաքել սպառողների վերաբերյալ՝ միևնույն ժամանակ ապահովելով դրանց գաղտնիությունը և անվտանգությունը: Երկրորդ՝ բանկերում, ՈՒՎԿ-ներում և ապահովագրական ընկերություններում ավելացել են տեխնիկական հնարավորությունները, որոնք թույլ են տալիս իրականացնել սպառողների վարքագծի մշտադիտարկում, ընդհուպ մինչև դրանց անընդհատ թարմացման հնարավորությունը: Ըստ այդմ՝ ֆինանսական կազմակերպությունները կարիք ունեն առավել հզոր վերլուծական գործիքների՝ միաժամանակ փորձելով պահել դրանց ճշգրտությունը: Մեքենայական ուսուցումը առաջարկում է մի շարք մոդելների խմբեր, որոնք փորձում են լուծել այս խնդիրը: Այն կիրառական վիճակագրության ճյուղ է, որի զարգացումը պայմանավորված է տեխնոլոգիական բուռն աճով, ու թեև մոդելները ներկայացվել են 20-րդ դարի սկզբին, դրանք կիրառվել են մեծ տվյալների համար ոչ գծային մոդելների գնահատման հնարավորությունը և դրանց ալգորիթմների զարգացումը նկատի ունենալով:

Գործնականում առկա է շփոթություն. մեքենայական ուսուցում և արհեստական բանականություն հասկացությունները կիրառվում են նույն համատեքստում՝ նկատի ունենալով վերահսկվող ուսուցումը, մինչդեռ իրականությունն այլ է:

Արհեստական բանականությունը ներառում է հետազոտությունների շատ լայն շրջանակ, որն ընդգրկում է մեթոդներ, որոնք սովորեցնում են համակարգերին լուծել խնդիրներ, օրինակ՝ համակարգչային տեսողություն (computer vision), վերահսկվող և չվերահսկվող ուսուցումներ (supevised and unsupervised learning), ամրապնդմամբ ուսուցում և գենետիկ ալգորիթմներ (reinforcement learning and genetic algorithms) և այլն:

Մեքենայական ուսուցումը զբաղվում է տվյալների ինքնուրույն յուրացմամբ: Այն ներառում է վերլուծական մոդելների լայն շրջանակ, որոնք կարելի է դասակարգել վերահսկվող կամ չվերահսկվող մոդելների: Վերահսկվող մեքենայական ուսուցման դեպքում գնահատվում կամ կանխատեսվում է ելքային փոփոխականը (output variable) մեկ կամ մի քանի մուտքային փոփոխականների (input variables) հիման վրա՝ ունենալով և՛ ելքային, և՛ մուտքային տվյալներ: Չվերահսկվող ուսուցման դեպքում տվյալները ուսումնասիրվում են առանց ելքային փոփոխականի՝ փորձելով գտնել նմուշներ (pattern):

Մեքենայական ուսուցումը՝ որպես արհեստական բանականության ճյուղ, հզոր գործիք է հատկապես կանխատեսման նպատակների առումով. տվյալներում գտնելով կապեր կամ նմուշներ՝ այն ծառայում է ընտրանքից դուրս արժեքների համար կանխատեսումներ կատարելուն (out-of-sample prediction): Տվյալների բազմությունից առանձնացվում է ընտրանք (training sample), և մոդելի ճշգրտությունը ստուգվում է այլ ընտրանքի վրա (validation sample): Այս քայլը k անգամ կրկնելուց հետո ընտրվում կամ ընդհանրացվում է կանխատեսման լավագույն մոդելը և կիրառվում ստուգման ընտրանքի վրա (testing sample): Այսպիսի մոդելները հիմնվում են մեծ ծավալ ունեցող և հաշվողական բարձր հզորություն պահանջող տվյալների վրա, ուստի հաճախ ասոցացվում են մեծ տվյալների վերլուծության հետ: Մեքենայական ուսուցման մոդելները, ի տարբերություն ավանդական վիճակագրական մոդելների, տվյալների բազմության վերաբերյալ հաճախ ենթադրություն չեն պահանջում և կարող են գնահատել ոչ գծային և ոչ պարամետրիկ մի շարք մոդելներ:

Սույն հետազոտությամբ փորձում ենք ցույց տալ արհեստական բանականության մոդելների կիրառման մի քանի տարբերակ ՀՀ-ում գործող ՈՒՎԿ-ի օրինակով՝ նպատակ ունենալով ազդել վարկերի մարման տոկոսի վրա:

Տվյալներ, նկարագրություն և վերլուծություն

Հետազոտության համար օգտագործվել են ՀՀ-ում գրանցված և գործող ՈՒՎԿ-ի[3] կողմից 2015-2019 թթ. տրված գյուղատնտեսական վարկերի տվյալներ, որոնց խմբերը և փոփոխականները հետևյալն են.

- Դեմոգրաֆիական տվյալներ (վերաբերում են վարկառուին).
  - Սեռ
  - Տարիք
  - Բնակության մարզ
  - Վարկի տրամադրման մասնաճյուղ:
- Վարկառուին պատկանող գյուղացիական տնտեսությանը վերաբերող համախառն տվյալներ.
  - Ընդհանուր ակտիվներ
  - Ընթացիկ ակտիվներ
  - Ոչ ընթացիկ ակտիվներ
  - Ընդհանուր պարտավորություններ
  - Համախառն եկամուտներ հաշվետու ժամանակաշրջանում
  - Զուտ եկամուտներ հաշվետու ժամանակաշրջանում
  - Համախառն ծախսեր հաշվետու ժամանակաշրջանում:
- Գյուղացիական տնտեսության ֆինանսական ագրեգատներ.
  - Պարտք-սեփական կապիտալ հարաբերակցություն[4]

[3] Կազմակերպության հետ ստորագրված է անվանման և տվյալների չբացահայտման համաձայնագիր:

[4] Տե՛ս **Eugene F. Brigham** Financial Management: Theory & Practice 15th Edition, էջ 104:

- Տոկոսադրույքի ծածկման հարաբերակցություն[5]։
- Վարկառուի վարկային պատմության տվյալներ.
- Նախկինում ուշացրած վարկերի քանակ
- Նախկինում վարկերի անընդհատ ուշացման առավելագույն օրերի քանակ
- Նախկինում վարկերի ուշացման ընդհանուր օրերի քանակ։
- Սույն վարկի վերաբերյալ տվյալներ.
- Սույն վարկի գումարը
- Սույն վարկի մարման ժամկետը
- Սույն վարկի տրման ամսաթիվը
- Սույն վարկի փոխարժեքը։

Տվյալների նախնական մշակում

Նախքան հիմնական վերլուծությունը կատարվել է տվյալների նախնական մշակում։ Քայլերը հետևյալն են.

- *Կեղծ փոփոխականների ստեղծում* – դիսկրետ փոփոխականները փոխարինվել են համապատասխան կեղծ փոփոխականներով (one-hot encoding)։
- *Փոփոխականների նորմալավորում* – բացատրվող շարունակական տվյալների համար իրականացվել է նորմալավորում՝ կիրառելով պոլինոմիալ կամ լոգարիթմական ֆունկցիաներ. բաշխման նորմալությունը ստուգվել է Կոլմոգորով-Սմիրնով վիճակագրական թեստի միջոցով։
- *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։
- *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։

Ռիսկայնության դասակարգում տվյալների դիսկրետիզացիայով

Ուսումնասիրված առաջին խնդիրը վարկառուի դասակարգումն է։ Ունենալով պատմական տվյալներ տարբեր վարկառուների մասին, օգտագործելով նրանց վարքագծի նմուշները՝ փորձել ենք կանխատեսել նոր վարկառուների վարկունակությունը՝ օգտագործելով արհեստական բանականության որոշ գծային և ոչ գծային մոդելներ։

Կիրառված մոդելներն են.

- Logistic Regression[6] (**LOGIT**)

[5] Տե՛ս նույն տեղը, էջ 105։

[6] Տե՛ս **Mark Schmidt, Nicolas Le Roux, Francis Bach** Minimizing Finite Sums with the Stochastic AverageGradient, Mathematical Programming B, Springer, 2017, 162 (1-2), էջ 83-112. 10.1007/s10107-016-1030-6։

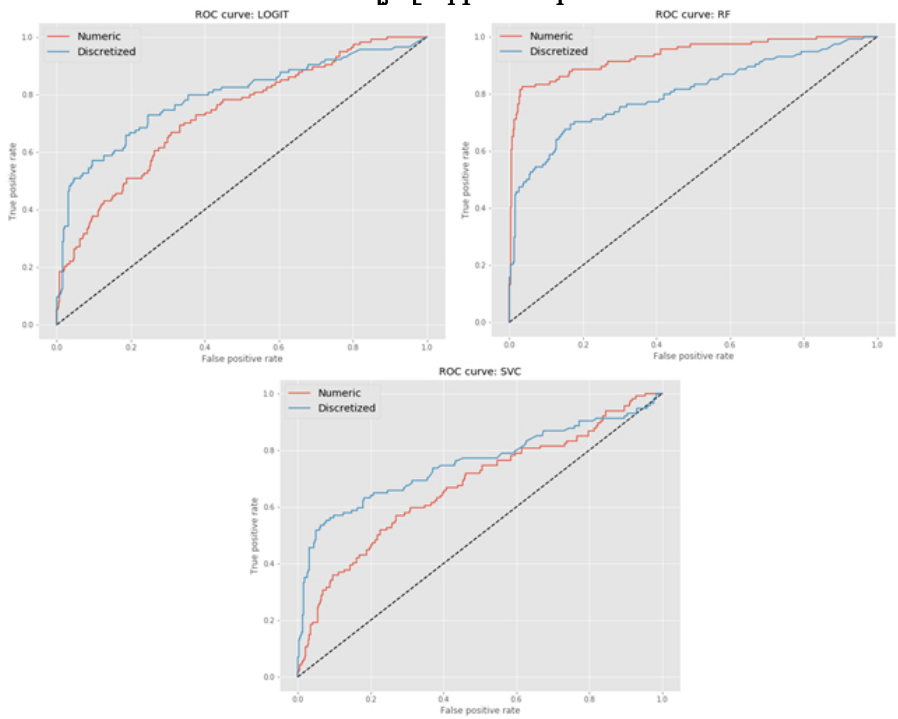
- Random Forest[7] (**RF**)
- Support Vector Classifier[8] (**SVC**):

Բացի այդ, կատարել ենք փոփոխականների դիսկրետիզացիա: Այն ենթադրում է շարունակական փոփոխականի արտապատկերումը դիսկրետ միջակայքերի՝ փորձելով առավելագույնը բացատրել ելքային փոփոխականը: Կիրառվել է C5.0[9] դասակարգման ալգորիթմ: Նպատակն է ցույց տալ, որ ոչ գծային մեթոդները կարող են լինել արդյունավետ, երբ կիրառված են դիսկրետիզացիայի հետ: Այսպիսի մոդելները, բարձր ճշգրտությունից բացի, նաև մեկնաբանելի են:

Ստացված արդյունքները ցույց են տալիս, որ նախնական տվյալների վրա կառուցված ոչ գծային մոդելը պակաս ճշգրիտ է, քան դիսկրետ տվյալների վրա կառուցված գծային մոդելը (տե՛ս աղյուսակ 1): Ինչպես նաև՝ ROC ցուցանիշը նվազում է, երբ ոչ գծային մոդելի տվյալների համար կիրառվում է դիսկրետիզացիայի ալգորիթմ (տե՛ս գծապատկեր 1):

Գծապատկեր 1

Կիրառված LOGIT, RF, SVC մոդելների ROC կորերը նախնական և դիսկրետ տվյալների համար



[7] Տե՛ս **L. Breiman,** Random Forests, Statistics Department University of California Berkeley, January 2001, available at: https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf :

[8] Տե՛ս **Wu, Lin and Weng,** Probability estimates for multi-class classification by pairwise coupling, JMLR 5:975-1005, 2004:

[9] Տե՛ս **Pang, Su-lin & Gong, Ji-zhang**. C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks. Systems Engineering - Theory & Practice. (2009). 29, էջ 94-104. 10.1016/S1874-8651(10)60092-0:

Դասակարգման մոդելների ճշգրտության ցուցանիշները

Ցուցանիշ	Նախնական տվյալներ			Դիսկրետ տվյալներ		
	LOGIT	RF	SVC	LOGIT	RF	SVC
Accuracy	0.8438	***0.8775***	0.8438	***0.8944***	0.8663	0.8438
Precision	0.8506	***0.8815***	0.8504	***0.8915***	0.8730	0.8504
Recall	0.8501	***0.8798***	0.8504	***0.8798***	0.8697	0.8504
F1	0.8502	***0.8801***	0.8502	***0.8801***	0.8700	0.8502

Փոփոխականների կարևորություն

Ունենալով կանխատեսման բարձր ճշգրտություն՝ մեքենայական ուսուցման մոդելների մեծ մասը մեկնաբանելի չէ (black-box models) նշված մոդելների բարդ կառուցվածքների պատճառով: Սակայն որոշ դեպքերում անհրաժեշտություն է առաջանում ստանալու մեկնաբանելի մոդելներ կամ մոդելների արդյունքների մեկնաբանություններ: Պատճառներից նշենք մի քանիսը:

- Բարդ կառուցվածքով մոդելները, որոնք հիմնականում արհեստական նեյրոնային ցանցեր են, խոցելի են «adversarial attack»[10] նմուշներին:
- Որոշ օրենսդրական կարգավորումներ պահանջում են, որ կազմակերպությունները ավտոմատացված որոշումների համակարգեր կիրառելիս ներկայացնեն որոշումների մեկնաբանություններ:
- Տվյալների հետագա հավաքման ռազմավարություն մշակելու տեսանկյունից դիտարկել ոչ կարևոր փոփոխականները:

Այս հետազոտությունում փոփոխականների կարևորությունը հաշվել ենք «պատահական անտառ»[11] (random forest) ալգորիթմից ստացված արդյունքներից[12], ապա նորմավորել 0-ից 100% սանդղակում: Ինչպես տեսնում ենք, ներկա վարկի ուշացումը առավելագույնը որոշվում է նախկինում ուշացրած վարկերով, նվազագույնը՝ վարկը տրամադրող մասնաճյուղով և մասնաճյուղի մարզով: Հարկ է նշել, որ որոշ ազդեցություն ունի նաև վարկառուի սեռը, որը համարվում է զգայուն փոփոխական: Հետագա հետազոտություններում նախատեսվում է մեղմել այս գործոնը:

[10] Տե՛ս **Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy** Explaining and Harnessing Adversarial Examples, ICLR 2015:

[11] Տե՛ս **Ho, Tin Kam** Random Decision Forests, Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995, էջ 278–282:

[12] Տե՛ս **L. Breiman,** Random Forests, Machine Learning, 45(1), 5-32, 2001. https://doi.org /10.1023/A:1010933404324:

Չափողականության կրճատում

Չափողականության կրճատումը (dimensionality reduction) բացատրող փոփոխականների ներկայացումն է այլ՝ լատենտ և ավելի փոքր չափողականությունում: Նպատակները տարբեր են. դրանից մեկը կանխատեսման մոդելում ուսուցման և կանխատեսման ծախսերի կրճատումն ու արագության աճն է: Այս օրինակում մեր նպատակն է կրճատել չափողականությունը մինչև $\mathbb{R}^2$ երկչափ տեսողական պատկերման համար:

Կիրառվել է t-SNE[13] ալգորիթմը. այն չափողականության կրճատման մեթոդների հարևան գրաֆների ենթախմբի ալգորիթմ է: Երկչափանի արտապատկերման համար պսևդոալգորիթմի քայլերն հետևյալն են.

1. Ընտրանքի բոլոր x_i կետերի k շրջակայքի x_j կետերի համար ալգորիթմը հաշվում է հարևան լինելու ճշմարտանմանությունը որպես պայմանական հավանականություն.

$$p_{j|i} = \frac{exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k\neq i} exp(-\|x_i - x_k\|^2/2\sigma_i^2)};$$

2. Ապա պայմանական հավանականություններից ստանում է համատեղ հավանականություններ.

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n};$$

3. Գեներացնում է y_i պատահական կետեր նախնական ընտրանքի չափով և հաշվել q_{ij} միասին հանդես գալու հավանականությունը:

4. Նվազագույնի է հասցնում $p_{ij} - q_{ij}$ տարբերությունը՝ համատեղ հավանականությունների բաշխումների տարբերություն ընդունելով Քալբեք-Լիբլեր էնթրոպիան[14].

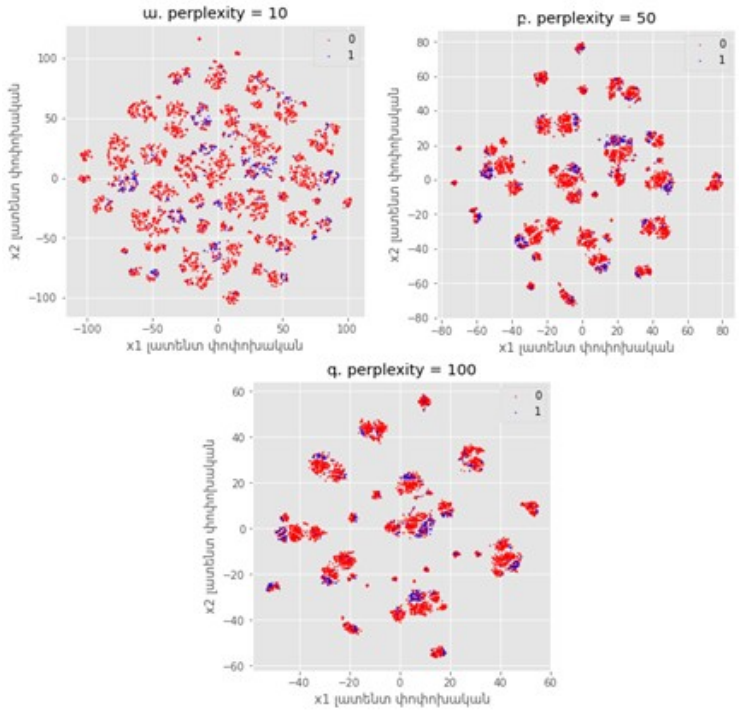
$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}};$$

Գծապատկեր 2-ում պատկերված են ստացված 2 լատենտ փոփոխականները: Երրորդ մոդելի արդյունքների վերաբերյալ կարող ենք ասել, որ ոչ վարկունակ հաճախորդները ցրված վարքագիծ չունեն և տեղակայված են հատուկ խմբերում՝ հետագա վերլուծություններում հնարավորություն տալով ուսումնասիրելու առանձին խմբեր:

[13] Տե՛ս **Laurens van der Maaten, Geoffrey Hinton,** Visualizing Data using t-SNE, Journal of Machine Learning Research 9 (2008) էջ 2579-2605

[14] Տե՛ս **Kullback, S. Leibler, R.A**. On information and sufficiency. Annals of Mathematical Statistics. (1951). 22 (1), էջ 79–86:

t-SNE մոդելի արդյունքները



Եզրակացություն

Հետազոտությամբ ստացել ենք հետևյալ արդյունքները.

- Գծային մոդելները դիսկրետիզացված տվյալների հետ ապահովում են ճշգրտության բարձր ցուցանիշներ, ինչպես նաև հնարավորություն են տալիս կանխատեսման մեկնաբանելի մոդել ստանալու:
- Այնուհանդերձ, ոչ գծային որոշ մոդելներ (օրինակ՝ «պատահական անտառ» մոդելը) թույլ են տալիս ստանալ մոդելի մեկնաբանություններ, որոնք կարող են լինել անշեղ՝ ելնելով նրանց կրկնողական բնույթից (bootstrap aggregation):
- Կիրառելով t-SNE մոդելը՝ կարող ենք տվյալները արտապատկերել ավելի փոքր չափողականությունում և ներկայացնել գծապատկերի միջոցով: Այն հետագայում օգտագործվելու է վարկառուների սեգմենտավորման համար:

ГЕВОРГ КАЛАЧЯН – *Применение моделей искусственного интеллекта в финансах (на примере УКО в РА).* – Развивающиеся технологии и современные алгоритмы машинного обучения открыли новые возможности для микрофинансовых организаций. В статье представлены методологии, которые могут быть использованы для лучшего финансового планирования в указанных организациях, дано приложение для операции в РА.

Такие методологии позволяют получать модели прогнозирования с оптимизацией затрат и высокой точностью. Более того, в статье показано, что предлагае-

мые методы решают проблему интерпретируемости модели и дают объяснение переменных для проблемы бинарной классификации; также продемонстрирован алгоритм, который создает скрытое пространство функций для визуализации данных и сегментации приложений.

Ключевые слова: *искусственный интеллект, машинное обучение, дискретизация переменных, уменьшение размерности, распознавание образов, скрытое пространство*

GEVORG GHALACHYAN – *Application of Artificial Intelligence Models in Finance (on the example of the UCO in RA).* – Evolving technologies and state-of-the-art machine learning algorithms have brought new opportunities for microfinance organizations. In this paper, we present the methodologies that can be used for better financial planning for such organizations and show the application for a UCO operation in RA.

Such methodologies allow obtaining cost-optimized and high-accuracy prediction models. Moreover, we showed that suggested techniques solve the problem of model interpretability and provide feature explanations for binary classification problems. Also, we demonstrated an algorithm that creates a latent space of features for data visualization and application segmentation.

Key words: *artificial intelligence, machine learning, feature discretization, dimensionality reduction, pattern recognition, latent space*