ՀՏԴ 004.8

Վարկային ռիսկի մոդելավորման համար գենետիկական ալգորիթմների վրա հիմնված սինթետիկ տվյալների գեներացման մեթոդ

Գառնիկ Առաքելյան

DOI: https://di.org/10.58726/27382923-ne2024.1-8

Հանգուցային բառեր. *լոգիստիկ ռեգրեսիա, kNN, գենետիկական ալգորիթմ, մուտացիա, տվյալների խմբավորում, կորելյացիա*

Նախաբան

Յուրաքանչյուր կազմակերպություն, այդ թվում նաև բանկերը և վարկային կազմակերպությունները, գործում են անկայուն միջավայրում և չունենալով այդ միջավայրի մասին ամբողջական տեղեկատվություն՝ կարող են կրել էական վնասներ: Այդպիսի կորուստների հիմնական աղբյուր է համարվում վարկային ռիսկը, որի կառավարման համար մշակվում են տարբեր մաթեմատիկական մոդելներ: Սակայն մոդելավորման ժամանակ հաճախ առաջանում են խնդիրներ, որոնք կապված են բավականաչափ դիտարկումների բացակայության հետ, և դրանք կարող են լուծվել ժամանակակից տեխնոլոգիաների միջոցով, մասնավորպես մեքենայական ուսուցման մեթոդներով:

Ներկայումս մեքենայական ուսուցման տեխնոլոգիաները ակտիվ կիրառվում են տարբեր խնդիրներ լուծելու համար: Օրինակ՝ 2019 թ. Google-ը ներկայացրել է Teachable Machine 2.0 տեխնոլոգիան, որը իրենից ներկայացնում է ինքնուրույն սովորող նեյրոնային ցանց, որը կարող է ճանաչել խոսք [11]: IBM-ն ունի Watson for Oncology [9], տեխնոլոգիա, որը մշակում է մեծ ծավալի բժշկական տվյալներ, ներառյալ պատկերները, քաղցկեղի ճշգրիտ ախտորոշման համար: Watson for Oncology ներկայումս կիրառվում է Նյու Յորքի, Բանգկոկի և Հնդկաստանի հիվանդանոցներում:

Կարելի է բերել նմանատիպ բազմաթիվ օրինակներ: Բանկերում և վարկային կազմակերպություններում մեքենայական ուսուցման տեխնոլոգիաները կարող են կիրառվել վարկային ռիսկի կառավարման համար: Նմանատիպ որակյալ մոդելներ ստեղծելու համար անհրաժեշտ են մեծ քանակությամբ տվյալներ: Հաճախ լինում են դեպքեր, երբ մոդելավորման համար բավականաչափ տվյալներ առկա չեն: Տվյալների հավաքագրումը հաճախ պահանջում է զգալի ֆինանսական ռեսուրսներ, ինչպես նաև այն

կարող է անիրագործելի լինել տարբեր պատճառներով, օրինակ՝ տեղեկատվական անվտանգության նկատառումներով:

Միջազգային պրակտիկայում մեքենայական ուսուցման մոդելի ստեղծման համար բավականաչափ տվյալների բացակայության դեպքում հաճախ կիրառում են սինթետիկ գեներացված տվյալները: Օրինակ՝ American Express-ը սինթետիկ տվյալների միջոցով մշակել է խարդախության հայտնաբերման մոդել [3]:

Սույն հետազոտության շրջանակում դիտարկվել են այլ հետազոտողների աշխատանքները: Աշխատանքում փորձ է կատարվել թվով փոքր քանակի իրական վարկային դիտարկումների հիման վրա գեներացնել սինթետիկ տվյալներ, որոնք կարող են օգտագործվել մեքենայական ուսուցման այնպիսի մոդելների ստեղծման համար, որոնք պահանջում են մեծ ծավալի տվյալների բազա: Սինթետիկ տվյալների գեներացման համար կիրառվել են գենետիկական ալգորիթմի տրամաբանությունը, Դարվինի էվոլյուցիայի տեսության հայեցակարգը և մեքենայական ուսուցման մեթոդներ, որոնք մեծ քանակի տվյալներ չեն պահանջում: Ստացված տվյալների որակը գնահատվել է վիճակագրական մեթոդներով:

Ստացված արդյունքները գործնականում կիրառելի են և ցույց են տալիս, որ ցանկացած բանկ կամ վարկային կազմակերպություն կարող է մշակել վարկային ռիսկի կառավարման որակյալ լուծումներ նույնիսկ առկա փոքր քանակությամբ տվյալների առկայության դեպքում:

Գրականության ակնարկ: Հետազոտության շրջանակում ուսումնասիրվել են թեմային առնչվող տարբեր հոդվածներ, թեզեր և գրքեր: Դիտարկենք դրանցից մի քանիսը:

Ժոան Ֆոնսեկայի և Ֆերնանդո Բակաոյի կողմից գրված «Tabular and latent space synthetic data generation: a literature review» հոդվածում իրականացվել է սինթետիկ աղյուսակային տվյալների ստեղծման ալգորիթմների գրական վերլուծություն: Աշխատանքում ուսումնասիրվել են աղյուսակային տվյալների գեներացման 70 ալգորիթմներ, որոնք սովորաբար անտեսվում են [6]:

Մանհար Ուալիայի, Բրենդան Թիրնիի և Սյուզան ՄաքՔևերի հեղինակած «Synthesising Tabular Data using Wasserstein Conditional GANs with Gradient Penalty (WCGAN-GP)» հոդվածում դիտարկվում է Generative Adversarial Networks (GANs) մոդելների կիրառումը աղյուսակային տվյալների ստեղծման համար: Իրական տվյալներից չտարբերվող և առանց տվյալների կորուստի աղյուսակային սինթետիկ տվյալների գեներացման խնդիրը լուծելու համար կիրառվել է Wasserstein Conditional Generative

Adversarial Network (WCGAN-GP) մոդելը: Ուսումնասիրության արդյունքները ցույց են տալիս, որ սինթետիկ տվյալները պահպանում են իրական տվյալների մեջ առկա բաշխումները և հարաբերությունները [12]:

Բորիս վան Բրեյգելի, Ժաոժի Քյանի և Միհայելա վան դեր Շարի կողմից գրված «Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic Data» երկում ուսումնասիրվում է գեներատիվ գործընթացի ազդեցությունը մեքենայական ուսուցման հետագա խնդիրների վրա: Ուսումնասիրությունը ցույց է տալիս, որ իրականի նման սինթետիկ տվյալների կիրառումը հանգեցնում է հետագա մոդելների և վերլուծությունների ստեղծմանը, որոնք վատ են ընդհանրացնում իրական տվյալները: Մոդելների որակը բարձրացնելու համար հետազոտությունում կիրառվում է Deep Generative Ensemble (DGE) մեխանիզմը: DGE-ն բարելավում է հետագա մոդելներում ուսուցումը, քանակական գնահատման որակը [2]:

Խալեդ Էլ Էմամի, Լյուսի Մոսկերայի և Ռիչարդ Հոպտրոֆի հեղինակած «Practical Synthetic Data Generation» գրքում ներկայացված են սինթետիկ տվյալների ստեղծման տարբեր մեթոդներ՝ իրական տվյալների հիման վրա ստեղծված մտացածին տվյալներ: Գիրքը նկարագրում է սինթետիկ տվյալների ստեղծման քայլերը՝ օգտագործելով բազմաչափ նորմալ բաշխում, բաշխումներ ընտրելու մեթոդներ, տվյալների կառուցվածքի մոդելավորման մոտեցումներ, ինչպես նաև մեթոդներ և ցուցիչներ, որոնք կարող են օգտագործվել տվյալների օգտակարությունը գնահատելու համար [5]:

Ընդհանուր առմամբ, կարելի է եզրակացնել, որ սինթետիկ աղյուսակային տվյալների գեներացման հարցը հիմնականում մնում է չուսումնասիրված, մինչդեռ անհրաժեշտ է նշել, որ ներկայումս կատարվում են աշխատանքներ, ինչի արդյունքում մշակվում են սինթետիկ տվյալների գեներացման տարբեր տեսակի մոդելներ: Հարկ է նշել, որ այդ մոդելները հաճախ չեն բավարարում որոշ տեսակի զգայուն տվյալներ գեներացման համար, օրինակ՝ վարկերի վերաբերյալ տվյալների գեներացում: Այդ իսկ պատճառով հետազոտությունում փորձ է կատարվում ուսումնասիրել սինթետիկ աղյուսակային տվյալների բնույթը գենետիկական ալգորիթմերի կիրառմամբ:

Հետազոտության մեթոդաբանություն: Սինթետիկ տվյալները իրենցից ներկայացնում են արհեստական ստեղծված տվյալներ, որոնք ընդօրինակում են իրական աշխարհի դիտարկումները և կիրառվում են մեքենայական ուսուցման մոդելների ուսուցման ժամանակ, երբ իրական տվյալների ստանալը բարդ է կամ թանկ: Սինթետիկ տվյալների գեներա-

ցումը թույլ է տալիս ստեղծել բոլորովին նոր տվյալներ, որոնք կունենան առկա իրական տվյալներին նման բնութագրեր: Այս պրոցեսի արդյունքում կարելի է գեներացնել ցանկացած քանակի տվյալներ, որոնք համարժեք են իրական տվյալներին և երևույթներին:

Ըստ էության, սինթետիկ տվյալներ ստեղծելով, մենք վերստեղծում ենք մի բան, որը գոյություն ունի իրական աշխարհում՝ ֆիքսելով դրա բնութագրերը՝ առանց դրանք ուղղակիորեն պատկերելու:

Կախված իր կառուցվածքից՝ սինթետիկ տվյալները կարող են լինել մասնակի կամ ամբողջական: Մասնակի տեսակը պարունակում է ինչպես գեներացված սինթետիկ, այնպես էլ իրական արժեքներ, մինչդեռ ամբողջական տեսակը պարունակում է միայն գեներացված սինթետիկ արժեքներ [1]:

Սույն հետազոտության շրջանակում փորձ է կատարվել ստեղծել սինթետիկ տվյալների գեներացման մեթոդաբանություն՝ կիրառելով Դարվինի էվոլյուցիայի տեսության հայեցակարգը:

Կարևոր է նշել, որ գեներացված սինթետիկ տվյալները պետք է ունենան նույն վիճակագրական հատկությունները, ինչ իրական տվյալները:

Ներկայումս գոյություն ունեն սինթետիկ տվյալների գեներացման համար տարբեր գործիքներ: Սակայն այդ որոշ լուծումները ենթադրում են տվյալների ներբեռնումը web սերվեր, ինչը կարող է առաջացնել խնդիրներ անվտանգության հետ, իսկ մյուսները պահանջում են արդեն իսկ մեծ ծավալի տվյալների առկայություն:

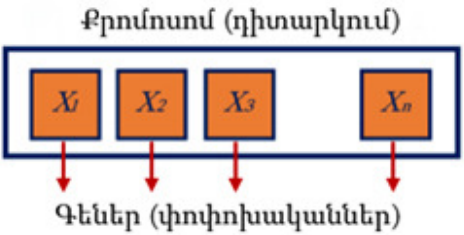
Սույն հետազոտությունում առաջարկվում է սինթետիկ տվյալների գեներացման մեթոդ, որը բաղկացած է մի քանի հաջորդաբար կատարվող քայլերից: Նշենք, որ այս պարագայում դիտարկվում է բինար դասակարգման դեպքը, որը ենթադրում է վարկային ռիսկերի գնահատման և կառավարման մոդել ստեղծելու անհրաժեշտություն, որը կկարողանա կանխատեսել վարկառուի դեֆոլտի առկայությունը կամ բացակայությունը: Դիտարկենք այս քայլերը.

1. Առկա իրական տվյալների հիման վրա անհրաժեշտ է ստեղծել դասակարգման հեշտ մոդել, օրինակ՝ լոգիստիկ ռեգրեսիա կամ kNN: Այս ուսումնասիրության համատեքստում այս մոդելները պետք է կանխատեսեն վարկառուի դեֆոլտը՝ հիմնվելով այլ հատկանիշների վրա: Ենթադրվում է, որ առկա իրական տվյալների ծավալը բավարար է վերոնշյալ մոդելը ստեղծելու համար, սակայն դրանք բավարար չեն առավել բարդ մոդելներ կառուցելու համար, օրինակ՝ նեյրոնային ցանցեր:

2. Սինթետիկ տվյալների գեներացումը անհրաժեշտ է իրականացնել կախյալ փոփոխականի յուրաքանչյուր խմբի համար: Առկա իրական

տվյալներում կախյալ փոփոխականը հանդիսանում է վարկառուի դեֆոլտը, որը կարող է ընդունել երկու արժեք՝ 0 (իրադարձության բացակայություն) և 1 (իրադարձության առկայություն)։ Հետևաբար սինթետիկ տվյալների գեներացման համար առկա իրական տվյալները անհրաժեշտ է բաժանել երկու խմբի՝ մեկը պարունակում է դեֆոլտի իրադարձության բացակայող դիտարկումները, իսկ մյուսը՝ դեֆոլտի իրադարձության առկա դիտարկումները։

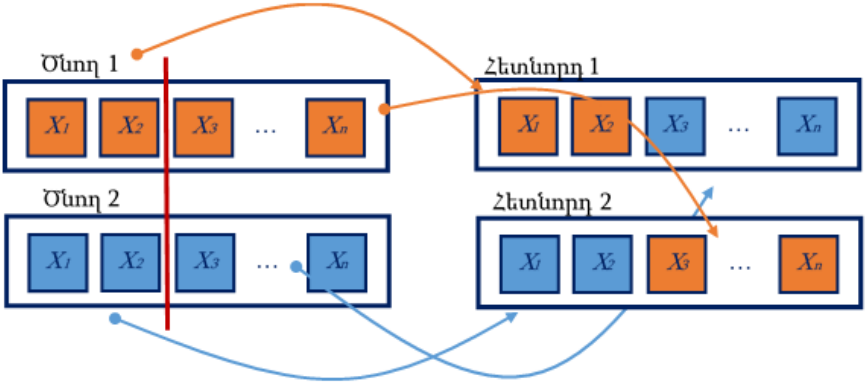
3. Դարվինի էվոլյուցիայի տեսության սկզբունքների համաձայն՝ երկրորդ կետում նշված իրական տվյալների յուրաքանչյուր խումբը դիտարկվում է որպես պոպուլյացիա։ Յուրաքանչյուր խմբի համար պատահականության սկզբունքով անհրաժեշտ է ընտրել երկու դիտարկում, որոնք կհանդիսանան ծնողներ իրենց հետնորդների համար, որոնք էլ հանդիսանում են սինթետիկ գեներացված տվյալներ։ Էվոլյուցիայի տեսության համաձայն՝ յուրաքանչյուր դիտարկում կհամարվի քրոմոսոմ, իսկ յուրաքանչյուր փոփոխական՝ գեն (նկար 1)։



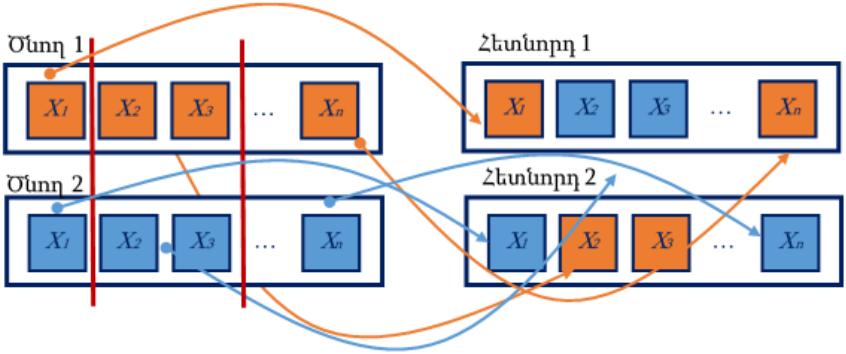
Նկար 1. Քրոմոսոմի կառուցվածք

Յուրաքանչյուր վերոնշյալ ընտրված զույգ դիտարկումների համար իրականացվում է ծնող քրոմոսոմների խաչաձև բեղմնավորումը հետնորդներ ստանալու համար։ Նշենք, որ այստեղ կախյալ Y փոփոխականը չի դիտարկվում։ Սույն մեթոդոլոգիայի շրջանակում հնարավոր է գեների ժառանգման երկու տարբերակ.

- Գեների ռեկոմբինացիա մեկ կետով, երբ հետազոտողը նախապես սահմանում է գեների այն տոկոսը, որը հետնորդը կժառանգի առաջին ծնողից, իսկ մյուս գեները՝ երկրորդ ծողից (նկար 2)։
- Գեների ռեկոմբինացիա երկու կետով, երբ հետազոտողը նախապես սահմանում է յուրաքանչյուր ռեկոմբինացիայի կետի գեների այն տոկոսը, որը հետնորդը կժառանգի առաջին ծնողից։ Մյուս գեները ժառանգվելու են երկրորդ ծնողից (նկար 3)։



Նկար 2. Գեների ռեկոմբինացիա մեկ կետով



Նկար 3. Գեների ռեկոմբինացիա երկու կետով

4. Ստացված հետնորդների շրջանակում իրականացվում է մուտացիա: Համաձայն սույն մեթոդի՝ մուտացիան իրենից ներկայացնում է ցածր հավանականությամբ ժառանգի քրոմոսոմում որևիցե գենի փոփոխություն: Մուտացիայի առաջացման առավելագույն հավանականությունը (x) որոշվում է էքսպերտային գնահատմամբ: Մուտացիայի ժամանակ գենը ընդունում է տվյալ փոփոխականի բոլոր արժեքների միջինը, որը կլորացվում է մինչև մոտակա ամբողջ թիվը: Տեղին է.

$$0 \le P(M) \le x$$
$$P(M) = \frac{x}{100} - \frac{x}{100 * n}$$

$$X_i = \frac{1}{n}\sum_{i=1}^{n} g_i$$
$$g_i \in G$$

որտեղ x՝ X_i գենի մուտացիայի առաջացման առավելագույն հավանականությունն է, P(M)՝ մուտացիայի իրադարձության հավանականությունն է, G՝ X_i գենի արժեքների բազմությունն է $\{g_1, g_2, \dots, g_n\}$, n՝ հետնորդների քանակ:

5. Վերոնշյալ երրորդ և չորրորդ կետերը պետք է կրկնվեն այնքան անգամ, որքան անհրաժեշտ է ցանկալի քանակությամբ սինթետիկ տվյալներ ստանալու համար: Այսպիսով, եթե անհրաժեշտ է գեներացնել F ծավալի սինթետիկ տվյալներ, ապա գործընթացն ավարտելու պայմանը կլինի

$$\sum_{i=0}^{n} S_i - S_0 \geq F$$

որտեղ S_i՝ գեներացված տվյալների քանակն է պրոցեսի ընթացքում, S_0՝ իրական տվյալների քանակն է, n՝ պրոցեսի կրկնության քանակ, F՝ տվյալների քանակ, որը անհրաժեշտ է գեներացնել:

6. Յուրաքանչյուր գեներացված տվյալների համար անհրաժեշտ է առաջին կետի ստեղծված մոդելի հիման վրա հաշվարկել կախյալ Y փոփոխականը:

7. Գեներացված սինթետիկ տվյալների որակը ստուգելու համար դրանք վիճակագրորեն տարբեր գործակիցների և թեստերի միջոցով համեմատվում են իրական տվյալների հետ:

Վերլուծություն: Հետազոտությունը իրականացվել է իրական վարկային տվյալների հիման վրա (3000 դիտարկում), ՀՀ տարածքում գործող Յունիբանկ ԲԲԸ-ի օրինակով: Առկա բազան բաղկացած է թվով 13 փոփոխականներից՝ վարկառուի տարիք վարկը վերցնելու պահին (20-ից մինչև 65 տարեկան), վարկառուի սեռ (իգական, արական), վարկառուի ամուսնական կարգավիճակ (միայնակ, ամուսնացած, բաժանված, այրի), վարկառուի կրթության աստիճան (գիտական աստիճան, բարձրագույն կրթություն, միջին մասնագիտական կրթություն, դպրոցական կրթություն), սեփականության առկայություն (անշարժ գույքի առկայություն, շարժական գույքի առկայություն, անշարժ և շարժական գույքերի առկայություն, սեփականության բացակայություն), պայմանագրային գումար (30,000-ից մինչև 1,336,000 ՀՀ դրամ), վերջին 12 ամիսների ընթացքում

ժամկետանց օրերի քանակ (0-ից մինչև 160), ուշացումների քանակ (0-ից մինչև 67), ռիսկի դասերի փոփոխությունների քանակ (0-ից մինչև 28), վարկային բեռ (0-ից մինչև 2,984,303 ՀՀ դրամ), վարկային պատմության երկարություն (0-ից մինչև 488 օր), առավելագույն մարած վարկ (0-ից մինչև 18,000,000 ՀՀ դրամ), դեֆոլտ (0 – ոչ, 1 – այո)։

Ինչպես տվյալների վերլուծությունը, այնպես էլ սինթետիկ տվյալների գեներացումը իրականացվել է՝ համաձայն վերոնշյալ մեթոդաբանության՝ կիրառելով Excel ծրագիրը, Python ծրագրավորման լեզուն և դրա համապատասխան գրադարանները՝ Pandas, Matplotlib, Seaborn և Scikit-learn։

Իրական տվյալների որոշ թվային փոփոխականների նկարագրական վիճակագրությունը ներկայացված է ստորև (աղյուսակ 1)։ Աղյուսակից երևում է, որ բաց թողնված տվյալներ առկա չեն՝ բոլոր փոփոխականներում տվյալների քանակը 3000 հատ է։

Աղյուսակ 1

Նկարագրական վիճակագրություն

Indicator	Վարկ. բեռ	Վարկ. պատմ. երկ.	Առավ. մարած վարկ	Պայմ. գումար
count	3000	3000	3000	3000
mean	436443	58,8	338093	258077
std	507191	62,7	875317	145527
min	-	-	-	30000
25 %	-	12	-	183800
50 %	257514	38,5	150000	200000
75 %	758630	84	360000	384850
max	2984303	488	18000000	1336000

Աղյուսակ 1-ից երևում է, որ վարկային բեռ, վարկային պատմության երկարություն, առավելագույն մարած վարկ և պայմանագրային գումար փոփոխականների համար ստանդարտ շեղումը բավականին բարձր է։ Դա պայմանավորված է բանկի վարկային քաղաքականության հետ և այն փաստի հետ, որ տվյալներում առկա են տարբեր վարկային պատմություն ունեցող հաճախորդներ, օրինակ՝ մի հաճախորդը կարող է չունենալ վարկային պատմություն, իսկ մյուսը՝ ունի երկար վարկային պատմություն և մեծ ծավալի բեռ։ Հաշվի առնելով առկա տվյալների հատկու-

թյունները՝ այլ փոփոխականների ստանդարտ շեղումը գտնվում է ընդունելի միջակայքում:

Վերլուծության ընթացքում կառուցվել է նաև կորելյացիոն մատրիցը (աղյուսակ 2), որտեղից երևում է, որ առկա է կապ «ռիսկի դասերի փոփոխությունների քանակ» և «ուշացումների քանակ», «վարկային պատմության երկարություն» և «ուշացումների քանակ», «ռիսկի դասերի փոփոխությունների քանակ» և «վերջին 12 ամիսների ընթացքում ժամկետանց օրերի քանակ», «ուշացումների քանակ» և «վերջին 12 ամիսների ընթացքում ժամկետանց օրերի քանակ» փոփոխականների միջև:

Աղյուսակ 2

Կորելյացիոն վերլուծություն
(առավել ուժեղ կապ ունեցող փոփոխականների մասով)

	Վերջին 12 ամ. ընթ. ժամ. օր. քանակ	**Ուշա-ցումների քանակ**
Ուշացումների քանակ	0,35	1,00
Ռիսկի դասերի փոփոխ. քանակ	0,39	0,61
Վարկ. պատմ. երկ.	0,11	0,45

Նկարագրական վիճակագրության և կորելյացիոն վերլուծության դիտարկումից հետո իրականացվել է սինթետիկ տվյալների գեներացումը: Դրա համար, համաձայն վերոնշյալ մեթոդաբանության առաջին կետի՝ իրական տվյալների հիման վրա կառուցվել են լոգիստիկ ռեգրեսիայի և kNN մոդելները Scikit-learn գրադարանի միջոցով, և այդ մոդելներից ընտրվել է լավագույնը:

Մոդելները կառուցելու համար տվյալները խմբավորվել են, քանի որ տվյալների նախնական մշակումը և խմբավորումը կարող են բարելավել մեքենայական ուսուցման մոդելի որակը [7]: Խմբավորված տվյալները ստանդարտացվել են՝ օգտագործելով Weight of Evidence (WOE) մեթոդը [4]: Յուրաքանչյուր խմբի համար WOE-ն հաշվարկվում է հետևյալ բանաձևով

$$WOE = \ln \frac{Percentage\ of\ good\ in\ the\ class}{Percentage\ of\ bad\ in\ the\ class}$$

որտեղ ln-ը բնական լոգարիթմն է [8]:

Աղյուսակ 3-ում և 4-ում ներկայացված են իրական տվյալների խմբավորումները և ստանդարտացման արդյունքները:

Աղյուսակ 3

3,000 իրական տվյալների խմբավորում և ստանդարտացում

Խումբ	Գումար (հազ. դրամ)	WOE
Սեռ		
Իգական	401,760	0,222
Արական	372,471	-0,236
Կրթություն		
Գիտական աստիճան	88,819	-0,172
Բարձրագույն	185,635	0,058
Միջին մասնագիտական	498,783	0,007
Միջնակարգ	994	1,309
Ամուսնական կարգավիճակ		
Բաժանված	11,594	1,419
Ամուսնացած	603,498	0,055
Միայնակ	15,157	-0,130
Այրի	143,982	-0,310
Ուշացումների քանակ		
1. 0	327,328	0,335
2. 1	446,903	-0,241
Վարկային բեռ		
1. 0	229,977	2,281
2. 1-300,000	138,009	-0,143
3. 300,001+	406,245	-0,865
Պայմանագրային գումար		
1. Մինչև 200,000	333,626	0,402
2. 200,001-400,000	269,056	-0,091
3. 400,001+	171 549	-0,630
Վերջին 12 ամիսների ընթացքում ժամկետանց օրերի քանակ		
1. 0-30	761,429	0,036
2. 31+	12,802	0,000

Խումբ	Գումար (հազ. դրամ)	WOE
Սեփականությանառկայություն		
Անշարժ գույքի առկայություն	42,343	-0,040
Շարժական գույքի առկայություն	180,294	0,057
Անշարժ և շարժական գույքի առկայություն	37,095	-0,136
Բացակայություն	514,499	-0,007
Տարիք		
1. 20-25	123,702	-0,220
2. 26-35	242,899	-0,143
3. 36-50	251,022	0,117
4. 51+	156,608	0,211
Ռիսկի դասերի փոփոխությունների քանակ		
1. 0	652,277	-0,097
2. 2+	121,954	0,540
Վարկային պատմության երկարություն		
1. 0-270	763,328	0,002
2. 271-365	7,778	-0,258
3. 366+	3,124	0,132
Առավելագույն մարած վարկ		
1. 0-350,000	544,095	0,115
2. 350,001+	230,136	-0,269

Սույն հետազոտությունում մոդելավորման համար որպես անկախ փոփոխականներ ընդունվում են աղյուսակ 2-ում հաշվարկված WOE արժեքները յուրաքանչյուր խմբի համար, իսկ որպես կախյալ փոփոխական, որը անհրաժեշտ է կանխատեսել՝ «Դեֆոլտ» փոփոխականը, որը կարող է ընդունել 0 (իրադարձության բացակայություն) և 1 (իրադարձության առկայություն) արժեքները։

Աշխատանքի ընթացքում կիրառվել է cross-validation գործիքը, որը թույլ է տալիս շրջանցել մոդելների գերուսուցման խնդիրը, իսկ մոդելների լավագույն պարամետրեր ընտրելու համար կիրառվել է GridSearchCV գործիքը։ Լոգիստիկ ռեգրեսիայի համար որպես հիպերպա-

րամետրեր ընդունվել են «penalty» (L1 և L2 կանոնավորացում) և «C» (կանոնավորացման հակադարձ ուժը), իսկ kNN մոդելի համար՝ «n_neighbours» (հարևանների քանակ 1-ից մինչև 10) and «P» (հզորության պարամետր Մինկովսկու չափման համար: Երբ p=1, մոդելը հեռավորությունը հաշվարկելու համար օգտագործում է Մանհեթենի հեռավորությունը, իսկ երբ p=2՝ էվկլիդեսյան հեռավորությունը):

Մոդելավորման համար տվյալները բաժանվել են 2 խմբի՝ ուսուցման համար նախատեսված խումբ (x_train, y_train) և թեստավորման համար նախատեսված խումբ (x_test, y_test) համապատասխանաբար 65 % և 35 % հարաբերակցությամբ: Մոդելները ուսուցվել են և լոգիստիկ ռեգրեսիայի համար որպես լավագույն պարամետրեր ընդունվել են «penalty» – L2 և «C» – 1000, իսկ kNN մոդելի համար՝ «n_neighbours» – 7 և «P» – 1:

Վերոնշյալ պարամետրերով ստեղծված մոդելները պահպանվել են, և այդ մոդելների համար հաշվարկվել են Scikit-learnscor մեթոդով որակական ցուցանիշը: Այն իրենից ներկայացնում է մոդելի որակը նկարագրող հասարակ մետրիկա, որը ցույց է տալիս ճիշտ գուշակածների մասնաբաժինը: Այսպիսով լոգիստիկ ռեգրեսիայի համար այդ ցուցանիշը, որը հաշվարկվել է՝ հիմք ընդունելով ուսուցման համար նախատեսված բազան, կազմել է 78,5 %, իսկ թեստի համար նախատեսված բազային հիման վրա՝ 80,8 %: KNN մոդելի համար այդ ցուցանիշը, որը հաշվարկվել է՝ հիմք ընդունելով ուսուցման համար նախատեսված բազան, կազմել է 84,1 %, իսկ թեստի համար նախատեսված բազային հիման վրա՝ 79,2 %:

Թեստային տվյալների հիման վրա հաշվարկվել են նաև Precession և Recall գործակիցները [10], որոնք ցույց են տալիս մոդելի որակը: Այնուհետև յուրաքանչյուր մոդելի համար հաշվարկվել է F1 մետրիկան, որը գաղափարապես միավորում է վերը նշված ցուցանիշների տեղեկատվությունը (աղյուսակ 4): Ստորև բերված են նշված մետրիկաների հաշվարկման բանաձևերը.

$$Precesion = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall},$$

որտեղ TP (True positive)՝ ճիշտ կանխատեսված դիտարկումների թիվն է, FP (False positive)՝ սխալ դրական կանխատեսված դիտարկումների թիվն

է, FN (False negative)՝ սխալ բացասական կանխատեսված դիտարկումների թիվն է:

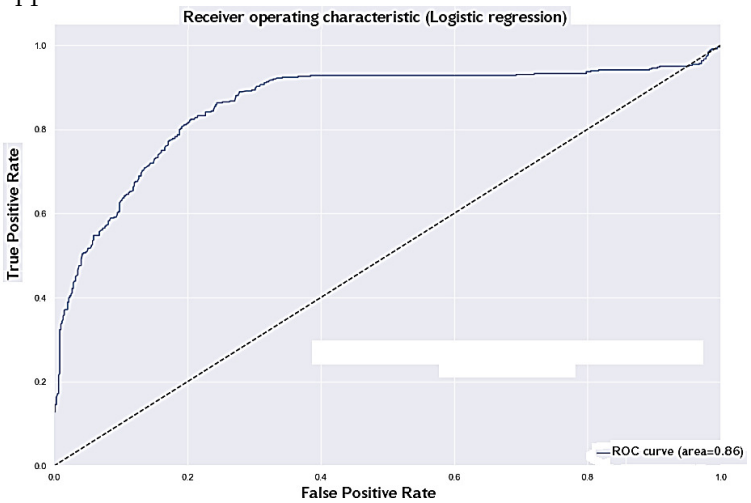
Աղյուսակ 4

Թեստային տվյալների հիման վրա հաշվարկված որակական մետրիկաները

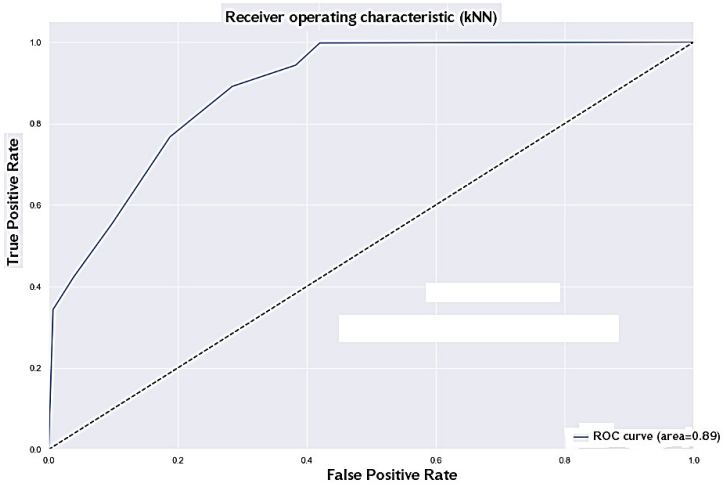
Մոդել	Թեստային տվյալներ	
	F1 Score	AUC
Logistic regression	0.79	0.86
kNN	0.76	0.89

Յուրաքանչյուր մոդելի համար կառուցվել է ROC կորերը: Այն իրենից ներկայացնում է գրաֆիկ, որը թույլ է տալիս գնահատել դասակարգչի (մոդելի) որակը՝ ցույց տալով ճիշտ դասակարգված և սխալ դասակարգված օբյեկտների հարաբերակցությունը (նկար 4, 5): ROC-ի քանակական մեկնաբանությունը ցույց է տալիս AUC (Area under Curve) գործակիցը: Որքան AUC-ը մեծ է, այնքան մոդելի որակը բարձր է (սակայն բարձր արժեքը կարող է գերուսուցման նշան լինել): Եթե AUC-ը 0,5 է, ապա մոդելը որակապես վատն է, իսկ 0,5-ից ցածր արժեքը ցույց է տալիս, որ դասակարգիչը ճիշտ հակառակն է կանխատեսում:

Կատարված աշխատանքի արդյունքում, համաձայն աղյուսակ 5-ի, մշակված մոդելներից լավագույնն է ճանաչվում kNN մոդելը, քանի որ այն կարողանում է առավել ճշգրիտ բացահայտել դեֆոլտ գնացած հաճախորդներին:



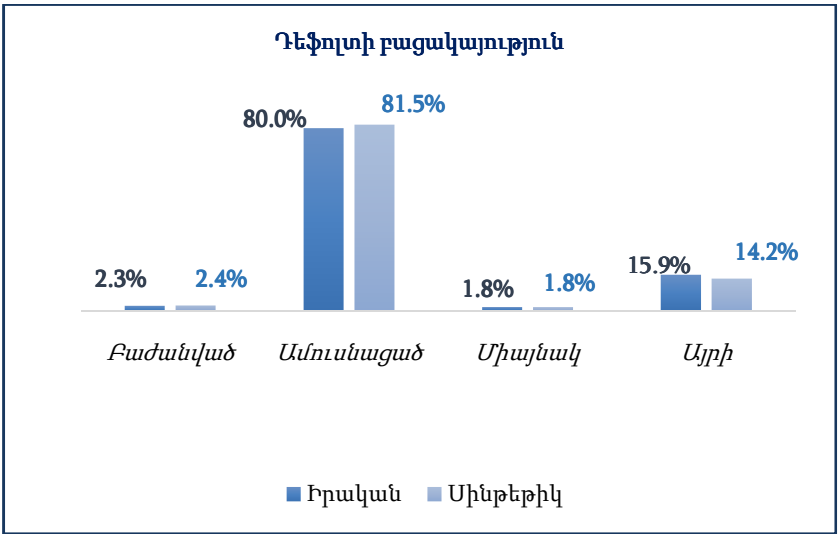
Նկար 4. Լոգիստիկ ռեգրեսիայի մոդելի ROC կոր

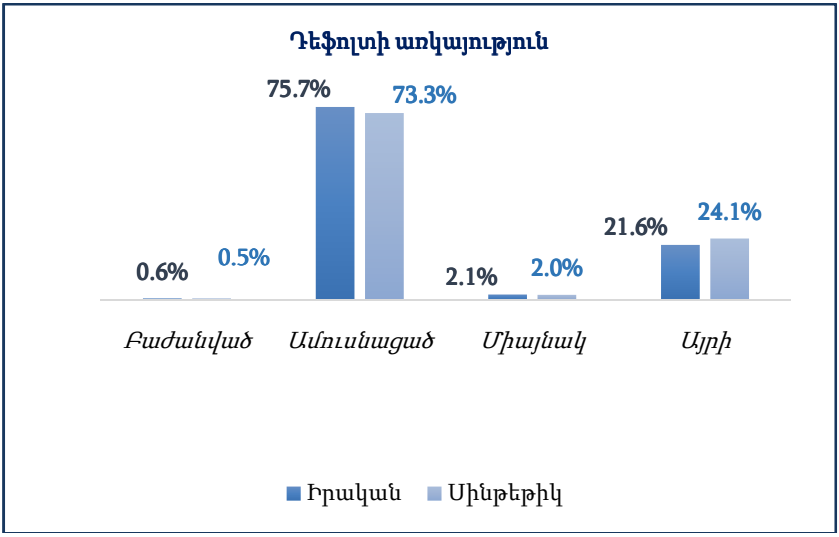


Նկար 5. kNN ռեգրեսիայի մոդելի ROC կոր

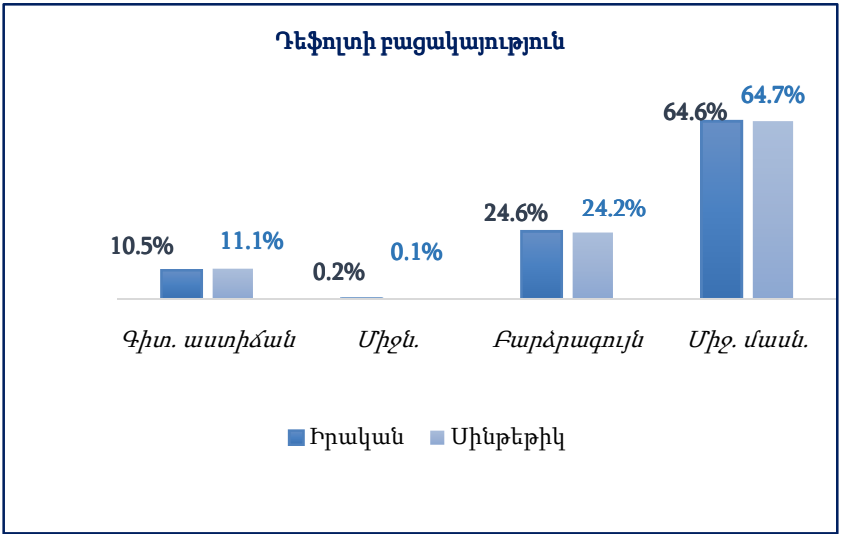
Վերոնշյալ մեթոդաբանության համաձայն՝ 1-ին քայլից հետո 2-5 քայլերը իրականացվել են՝ համաձայն մշակված Python սկրիպտի: 3,000 իրական տվյալների հիման վրա գեներացվել են 100,000 սինթետիկ տվյալներ: Յուրաքանչյուր սինթետիկ գեներացված դիտարկման համար «Դեֆոլտ» կախյալ փոփոխականը կանխատեսվել է 1-ին կետում մշակված և ընտրված որպես լավագույն՝ kNN մոդելի միջոցով:

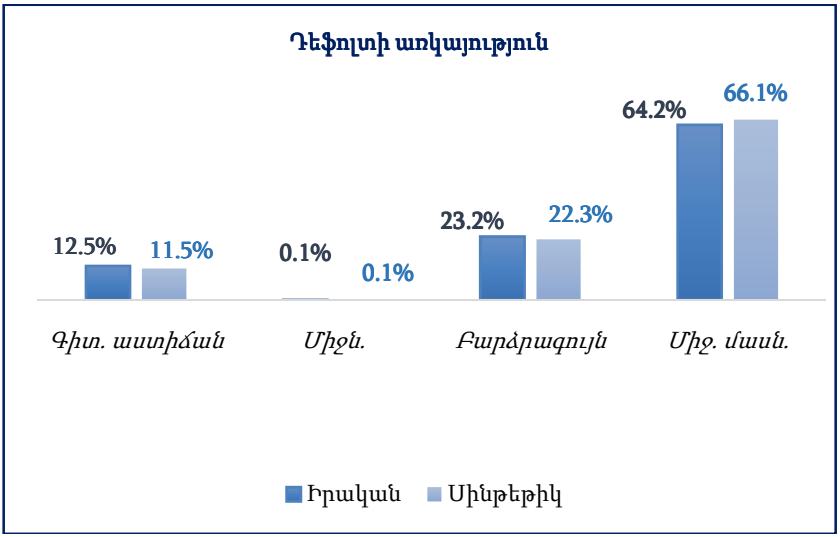
Գեներացված կատեգորիալ փոփոխականները գնահատվել են իրական և գեներացված տվյալների համապատասխան խմբերում կշիռների համեմատման միջոցով: Համեմատությունը իրականացվել է առանձին՝ համապատասխանաբար դեֆոլտի բացակայության և առկայության հիման վրա: Համեմատելով ստացված տվյալները՝ կարելի է պնդել, որ գեներացված կատեգորիալ փոփոխականները համապատասխանում են իրական տվյալներին (նկար 6, 7, 8, 9-ում):



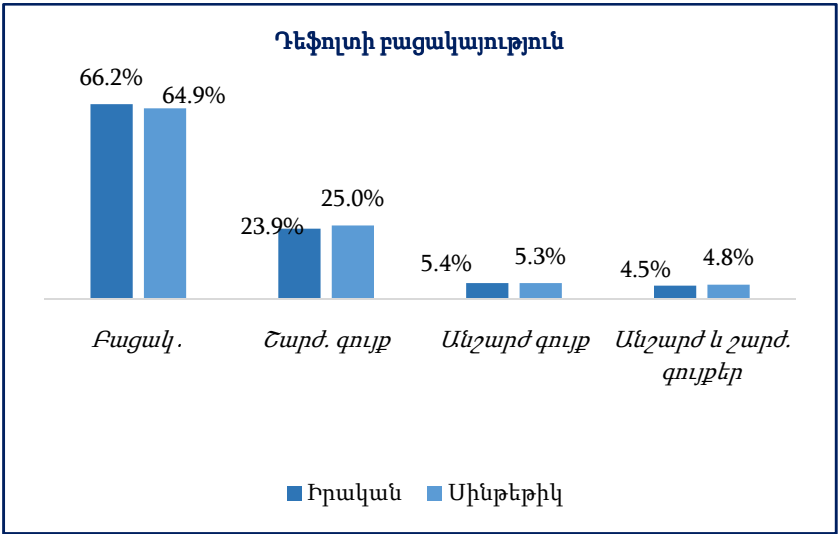


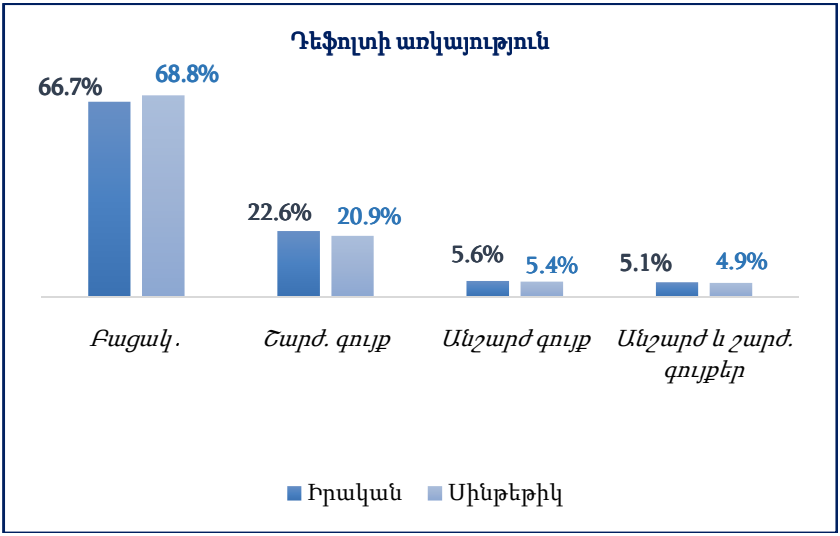
Նկար 6. Իրական և գեներացված տվյալների համեմատություն «Ամուսնական կարգավիճակ» փոփոխականի համար



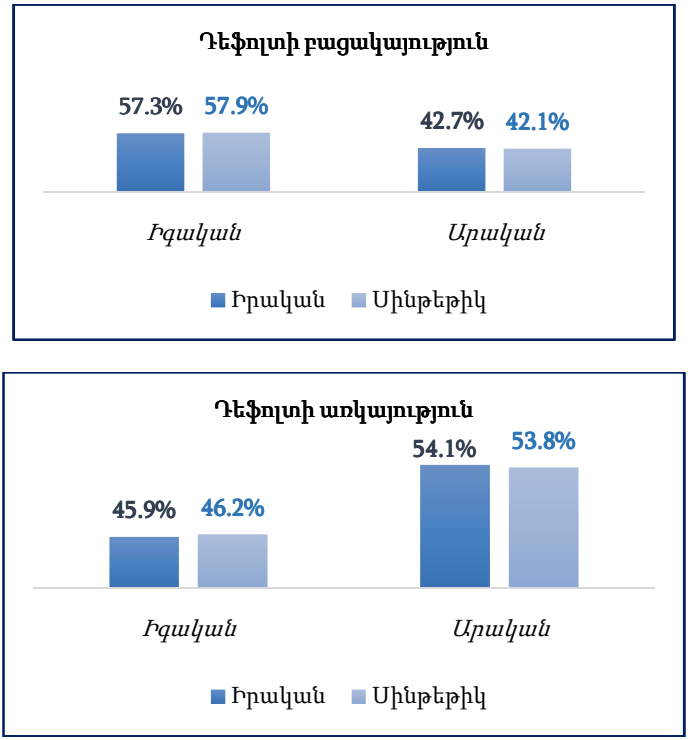


Նկար 7. «Կրթություն» փոփոխականի իրական և գեներացված տվյալների համեմատություն





Նկար 8. «Մեփականության առկայություն» փոփոխականի իրական և գեներացված տվյալների համեմատություն



Նկար 9. Իրական և գեներացված տվյալների համեմատություն «Սեռ» փոփոխականի համար

Գեներացված թվային փոփոխականները գնահատվել են t-վիճակագրության միջոցով: Ստացված արդյունքները ներկայացվել են աղյուսակ 5-ում, ինչի արդյունքում կարելի է նշել, որ բոլոր գեներացված փոփոխականները համապատասխանում են իրական տվյալներին, բացառությամբ «Վարկային բեռ» փոփոխականի: Այդ փոփոխականը կարող է տարբերվել մուտացիայի գործոնի առկայության պատճառով, քանի որ այդ փոփոխականի արժեքները գտնվում են 0-ից մինչև 2,984,303 ՀՀ դրամ միջակայքում:

Աղյուսակ 5

T-վիճակագրության արդյունքներ

Փոփոխական	Statistic	Pvalue
Տարիք	0,856516785	0,391713985
Վերջին 12 ամիսների ընթացքում ժամկետանց օրերի քանակ	(0,681300611)	0,495682828
Ուշացումների քանակ	(0,527846142)	0,597607265
Ռիսկի դասերի փոփոխությունների քանակ	0,732714716	0,463734114
Վարկային բեռ	(3,497259066)	0,000470263
Վարկային պատմության երկարություն	(0,467802758)	0,639926638
Առավելագույն մարած վարկ	(0,466592493)	0,640792446
Պայմանագրային գումար	(1,167098726)	0,243173204

Եզրակացություններ: Բանկերը և վարկային կազմակերպությունները կարող են կրել զգալի կորուստներ, քանի որ դրանք գործում են անկայուն միջավայրում, որի մասին ամբողջական տեղեկատվություն առկա չէ: Վարկային ռիսկը հանդիսանում է այն հիմնական ռիսկերից, որոնց հետ առնչվում են ֆինանսական ինստիտուտները: Բանկերի և վարկային կազմակերպությունների համար կարևոր է հստակ գիտակցել այս ռիսկի ճանաչման, գնահատման, վերահսկման և կառավարման կարևորությունը:

Վարկային ռիսկի կառավարումը կարևոր տեղ է զբաղեցնում ֆինանսաբանկային համակարգում: Այդ ռիսկը նվազեցնելու համար անհրաժեշտ է ներդնել որոշումների կայացման հուսալի համակարգ: Այդ նպատակով ստեղծվում են տարբեր մաթեմատիկական մոդելներ: Մոդելավորման ժամանակ հետազոտողները հաճախ բախվում են խնդիրների՝ բավարար դիտարկումների բացակայության պատճառով:

Սույն աշխատանքի շրջանակում վերոնշյալ խնդիրը լուծելու համար մշակվել է փոքր քանակի իրական տվյալների հիման վրա մեթոդաբանություն, որը թույլ է տալիս գեներացնել սինթետիկ տվյալներ:

Ստացված արդյունքները գործնականում կիրառելի են և ցույց են տալիս, որ ցանկացած բանկ կամ վարկային կազմակերպություն կարող է մշակել վարկային ռիսկի կառավարման հուսալի համակարգ նույնիսկ փոքր քանակի տվյալների հիման վրա:

DOI: https://di.org/10.58726/27382923-ne2024.1-8

Գրականություն

1. AltexSoft, Synthetic Data for Machine Learning: Its Nature, Types, and Means of Generation, AltexSoft software r&d engineering, 22.03.2022, https://www.altexsoft.com/blog/synthetic-data-generation/(Date of last access 21.02.2024)
2. Breugel B., Qian Z., Schaar M. Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic Data. PMLR., 2023, https://proceedings.mlr. press/v202/van-breugel23a/van-breugel23a.pdf (Date of last access 21.02.2024)
3. Castellanos S. Fake it to Make it: Companies Beef up AI Models with Synthetic Data. WSJ PRO., 23.07.2021, https://www.wsj.com/articles/fake-it-to-make-it-companies-beef-up-ai-models-with-synthetic-data-11627032601(Date of last access 27.02.2024)
4. Chakraborty A. Information Value (IV) — how to use it in EDA and Model Building? Medium, 05.09.2021 https://medium.com/mlearning-ai/weight-of-evidence-woe-and-information-value-iv-how-to-use-it-in-eda-and-model-building-3b3b98efe0e8 (Date of last access 18.02.2024
5. Emam K., Mosquera L., Hoptroff R. Practical Synthetic Data Generation. O'Reilly Media, Inc., 2020, p. 175.
6. Fonseca J., Bacao F. Tabular and latent space synthetic data generation: a literature review. Springer Open., 10.07.2023, https://journalofbigdata. springeropen.com/articles/10.1186/s40537-023-00792-7(Date of last access 27.02.2024)
7. Geeks for Geeks, Binning in Data Mining, Geeks for Geeks, 11.01.2023, https://www.geeksforgeeks.org/binning-in-data-mining/(Date of last access 18.02.2024)
8. Ghazaryan A., Grigoryan L., Arakelyan G. Implementation of Machine Learning in the Credit Risk Management System of Individuals. Messenger of ASUE, 5(71), 2022, pp. 123-138.
9. IBM, 5725-W51 IBM Watson for Oncology, IBM, 01.08.2023, https://www.ibm.com/docs/en/announcements/watson-oncology?region= CAN (Date of last access 27.02.2024)
10. Mattman Ch., Machine Learning with Tensor Flow, Manning Publications Co., 2020, p. 456
11. Phillips K. Teachable Machine 2.0 makes AI easier for Everyone, 07.11.2019, https://blog.google/technology/ai/teachable-machine/(Date of last access 01.03.2024)

12. Walia M., Tierney B., McKeever S. Synthesising Tabular Data using Wasserstein Conditional GANs with Gradient Penalty (WCGAN-GP). CEUR Workshop Proceedings, 18.12.2020, https://ceur-ws.org/Vol-2771/AICS2020_paper_57.pdf (Date of last access 21.02.2024)

Метод генерации синтетических данных на основе генетических алгоритмов для моделирования кредитного риска

Гарник Аракелян

Резюме

Ключевые слова: *логистическая регрессия, kNN, генетический алгоритм, мутация, группировка данных, корреляция*

Любая компания, в том числе банки и кредитные организации, осуществляют свою деятельность в нестабильной среде и не имея о ней полной информации, могут понести значительные убытки. Одним из основных источников подобных убытков является кредитный риск, для управления которого создаются различные математические модели. Однако при моделировании часто возникают проблемы, связанные с отсутствием необходимого количества наблюдений.

В рамках темы исследования были изучены работы других исследователей. В рамках данной работы предпринята попытка создания синтетических данных на основе имеющихся в малом количестве реальных наблюдений о кредитах, которые могут быть использованы для создания моделей машинного обучения, которые требуют наличие большого набора данных.

Для генерации синтетических данных была использована логика генетических алгоритмов, концепции теории эволюции Дарвина, а также методы машинного обучения, не требующие наличие большого количества данных. Качество сгенерированных данных было оценено статистическими методами.

Полученные результаты практически применимы и показывают, что любой банк или кредитная организация может разработать качественное решение для управления кредитными рисками даже при наличии имеющихся в малом количестве данных.

A Method for Generating Synthetic Data based on Genetic Algorithms for Modeling Credit Risk

Garnik Arakelyan

Summary

Key words: *logistic regression, kNN, genetic algorithm, mutation, data grouping, correlation*

Any company, including banks and credit organizations, operates in an unstable environment and may incur significant losses without having complete information about it. One of the main sources of such losses is credit risk, for the management of which various mathematical models are created. However, modeling often faces challenges related to the lack of a sufficient number of observations.

Within the research topic, studies by other researchers have been examined. In this work, an attempt was made to create synthetic data based on a small number of real credit observations, which can be used to create machine learning models that require a large dataset.

To generate synthetic data, the logic of genetic algorithms, the concepts of Darwin's theory of evolution, as well as machine learning methods that do not require a large amount of data were used. The quality of the generated data was assessed using statistical methods.

The results obtained are practically applicable and demonstrate that any bank or credit organization can develop a high-quality solution for managing credit risks even with a small amount of available data.

Ներկայացվել է 28.02.2024 թ.
Գրախոսվել է 02.04.2024 թ.
Ընդունվել է տպագրության 30.05.2024 թ.