УДК 044.8
DOI: 10.18523/2617-3808.2023.6.30-35

Ліп'яніна-Гончаренко Х. В.

МЕТОД ФОРМУВАННЯ НАВЧАЛЬНОЇ ВИБІРКИ ДЛЯ МАСИВІВ ДАНИХ НА ОСНОВІ МАШИННОГО НАВЧАННЯ

У цій роботі запропоновано новий метод формування навчальної вибірки на базі машинного навчання, що об'єднує дані з RFM-аналізу та кластерного аналізу. Метод застосовано до даних, отриманих з аукціонів українського сайту ProZorro Продажі. Запропонована вибірка охоплює 92 638 аукціонів, 29 164 унікальні аукціони та 39 747 унікальних організаторів. У процесі RFM-аналізу дані розбито на групи: «Найкращі організатори тендерів», «Вірні організатори тендерів» та ін. Далі, методом K-means, дані були поділено на кластери, що дало змогу відокремити різні категорії організаторів. Результати тестування, проведеного з використанням Logistic Regression і Naive Bayes, засвідчили високу точність для обох методів. Продемонстровано, що вибірка та групування за допомогою запропонованого методу допомагають відрізняти організаторів тендерів за їхніми характеристиками та результатами. Подальші дослідження мають бути у напрямі розроблення автоматизованої системи для вибору організаторів тендерів на основі машинного навчання, що сприятиме оптимізації участі у тендерних процедурах.

Ключові слова: навчальна вибірка, машинне навчання, RFM-аналіз, кластерний аналіз, тендери.

Вступ

Нині постає необхідність розроблення ефективного методу формування навчальної вибірки для задач машинного навчання, зокрема в контексті аналізу даних RFM-аналізу та кластерного аналізу. Наявні методи формування вибірок часто не забезпечують оптимальний відбір репрезентативних та інформативних зразків для навчання моделей.

Ключові аспекти проблеми такі:

- Недостатня репрезентативність. Традиційні методи формування навчальних вибірок подеколи не враховують різноманітність даних і їх динамічні зміни. Внаслідок цього моделі можуть бути погано навчені на нових, раніше не відомих даних.
- Незначущі дані. Велика кількість шумових або незначущих даних може негативно впливати на якість навчання моделей, призводячи до перенавчання або недонавчання.
- Обмежені ресурси. Для ефективного навчання моделей важливо вибирати найінформативніші дані, особливо якщо обчислювальні або часові ресурси є обмеженими.
- Змінність даних. Динаміка зміни даних може потребувати адаптивності методу формування вибірки, щоби було відображено актуальну структуру даних.

Отже, постановка проблеми передбачає розроблення нового методу формування навчальної вибірки, який би поєднував підходи RFM-аналізу та кластерного аналізу для досягнення оптимальної репрезентативності, враховував значущість даних і був адаптивним до змін. Такий метод має покращити ефективність навчання моделей машинного навчання та забезпечити точніші результати прогнозування на нових даних.

Огляд літератури

У статті [6] розглянуто методи формування класифікованої навчальної вибірки, що генерується лише за допомогою активних перешкод, для адаптації вагових коефіцієнтів просторових фільтрів за умов поєднаної наявності перешкод. У статті [7] запропоновано адаптивний метод формування класифікованої навчальної вибірки на основі використання порогової оцінки коефіцієнта кореляції міжканалів комбінованих перешкод. У статті [4] представлено новий непараметричний лінійний метод вилучення ознак для класифікації гіперспектральних зображень, який використовує ідеї вікон Парзена для визначення локального середнього значення сусідніх зразків і нові вагові функції для формування матриць розкиду міжкласової та внутрішньокласової варіації. У статті [3] запропоновано алгоритм роз-

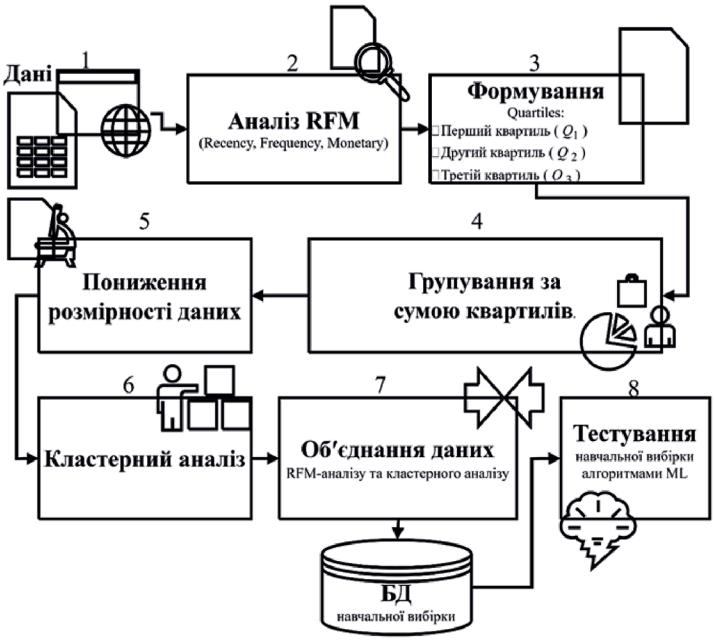

Рис. 1. Структура формування навчальної вибірки для сегментації організаторів тендерів на основі машинного навчання

роблення навчального набору, який найкраще описує об'єкти розпізнавання.

Метою цієї статті є розроблення методу формування навчальної вибірки на основі машинного навчання.

На відміну від аналогів, розроблений метод формування навчальної вибірки на основі машинного навчання дасть змогу автоматизовано та ефективно вибирати найбільш репрезентативні та інформативні зразки для побудови моделей машинного навчання. Основна ідея полягає в тому, щоб використовувати здобуті знання зі сфери RFM-аналізу та кластерного аналізу, групуючи схожі об'єкти разом. Це дає можливість підібрати найбільш репрезентативних представників різних груп даних і їхні характеристики для формування навчальної вибірки.

Застосування такого методу в контексті аналізу та оброблення даних може значно покращити результати навчання моделей, зменшити вплив шуму та незначущих даних, а також зробити процес навчання швидшим і ефективнішим. Характерною особливістю цього підходу є його адаптивність, тобто здатність адекватно реагувати на зміни в навчальних даних і змінювати склад навчальної вибірки відповідно до нової інформації.

Метод

Авторка розробила метод формування навчальної вибірки на основі машинного навчання. Цей метод (рис. 1) передбачає такі кроки:

Крок 1. Ввід вхідних даних (Блок 1).

Крок 2. RFM-аналіз (Блок 2). Аналіз RFM [1] (Recency, Frequency, Monetary) — це техніка сегментації даних.

Крок 3. Формування Quartiles (Блок 3). Quartiles ділять число точок даних на чотири частини, або чверті, більш-менш однакового розміру. Дані повинні бути впорядковані від найменшого до найбільшого для обчислення квартилів. Три основні квартилі такі:

- Перший квартиль (Q_1) — це середнє число між найменшим числом (мінімальним) і медіаною набору даних. Він також відомий як *нижчий* або *25-й емпіричний* квартиль, оскільки 25 % даних лежать нижче цієї точки.
- Другий квартиль (Q_2) є медіаною набору даних, 50 % даних лежать нижче цієї точки.
- Третій квартиль (Q_3) — це середнє значення між медіаною та найвищим значенням (максимумом) набору даних. Він відомий як *верхній* або *75-й емпіричний* квартиль, оскільки 75 % даних лежать нижче цієї точки.

Крок 4. Групування за сумою квартилів (Блок 4).

Крок 5. Пониження розмірності даних (Блок 5). Зниження розмірності є процесом скорочення кількості випадкових змінних шляхом отримання множини головних змінних. Виділення ознак і зниження розмірності можна об'єднати в один етап за допомогою методу головних компонент (МГК), лінійного розділювального аналізу (ЛРА), канонічного кореляційного аналізу (ККА) або розкладення невід'ємних матриць (РНМ).

Крок 6. Кластеризація (Блок 6). Кластерний аналіз [5] — задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, які називають кластерами, так, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися. Цей аналіз поділяють на такі етапи:

- проведення дослідження;
- підготовка даних до кластерного аналізу;
- вибір методу кластерного аналізу;
- вибір міри відстані між об'єктами та її обчислення;
- вибір стратегії кластеризації;
- застосування вибраної стратегії для утворення кластерів;
- перевірка результатів кластерного аналізу на осмисленість і їх інтерпретацію.

Крок 7. Об'єднання (Блок 7) даних RFM-аналізу та кластерного аналізу та внесення до БД. На основі цих даних можна провести навчання класифікації.

Крок 8. Тестування (Блок 8) навчальної вибірки на основі алгоритмів класифікації машинного навчання.

Для підтвердження розробленого методу формування навчальної вибірки на основі машинного навчання в наступному розділі проведено його реалізацію для сегментації організаторів тендерів.

Реалізація. Для формування навчальної вибірки для сегментації організаторів тендерів на основі машинного навчання вибрано мову Python. При цьому використано такі бібліотеки: pandas, numpy, train_test_split, KMeans, PCA.

Як вхідні дані використано завершені угоди учасників тендерів в Україні з сайту ProZorro Продажі [8]. Вибірка після очищення становить 93 336 значень відносно 10 параметрів. У процесі оцінювання кількісних показників виявлено 92 638 аукціонів, 29 164 унікальних аукціонів та 39 747унікальних організаторів.

Далі проведено RFM-аналіз, який допоміг розділити організаторів на різні категорії або кластери, щоб визначити організаторів, які частіше проводять аукціони з найбільшими сумами. Це три атрибути клієнта для кожного організатора (рис. 2).

ID_Orzanizer	Recency	Frequency	Monetary
1	952	2	1134546
2	412	1	23414
3	154	1	7
4	421	14	8625230
5	1214	1	24225
6	122	9	27834555
7	1	1	16368
8	1086	17	5746360
9	436	4	307177
10	647	4	209575

Рис. 2. RFM-аналіз (head(10))

Щоб розрахувати Recency, потрібно вибрати дату, з якої буде проведено оцінювання.

Frequency угод дасть можливість дізнатися, скільки разів організатор провів угоди. Для цього проведено перевірку, скільки рахунків зареєстровано тим самим організатором.

Monetary визначає, скільки грошей зароблено на угодах організатором.

Найпростіший спосіб просегментувати організаторів — це використати Quartiles, а саме надати оцінки від 1 до 4 Recency, Frequency and Monetary (рис. 3, чотири — найвище значення, одиниця — найнижче значення).

Із рисунку 3 видно, що організатор з ID 408 і 1632 отримали найвищу оцінку, тобто: R_Quartile = 4, нещодавно проведена угода; F_Quartile = 4, проведено найбільшу кількість угод; M_Quartile = 4, зароблено найбільше коштів. Відповідно у цих організаторів тендерів RFMScore = 444.

Оцінку загальної вибірки (рис. 4) проведено за такими критеріями: The best organizers of tenders (RFMScore = 444), Loyal organizers of tenders (F_Quartile = 4), Large consumers (M_Quartile = 4), Tenders are seldom held, but for a

ID_Orzanizer	Recency	Frequency	Monetary	Rank	R_Quartile	F_Quartile	M_Quartile	RFMScore
1745	374	21	15920922384	1.0	3	4	4	344
453	353	52	9056844144	2.0	3	4	4	344
408	275	33	5848400325	3.0	4	4	4	444
1290	426	48	4465808177	4.0	2	4	4	244
1632	31	24	3856853376	5.0	4	4	4	444

Рис. 3. RFM-Quartiles

large sum (RFMScore = 114) і Weak tender organizers (RFMScore = 111).

```
The best organizers of tenders:  119
Loyal organizers of tenders:  385
Large consumers:  455
Tenders are seldom held, but for a large sum:  15
Weak tender organizers:  48
```

Рис. 4. Оцінка RFM організаторів тендерів

Тепер, коли є сегментація наших організаторів тендерів, можна оцінити кожну групу окремо та проаналізувати, як витрачаються кошти і які організатори найчастіше проводять тендери.

Щоби більше зрозуміти поведінку організатора тендерів, треба глибше вивчити взаємозв'язок між змінними RFM. Тому потрібно поєднати отримані результати з певними прогнозуючими моделями, як-от кластеризація K-means clustering, логістична регресія або рекомендаційна система, для отримання кращих інформативних результатів щодо поведінки організаторів тендерів.

Для групування обрано K-means clustering, оскільки цей простий метод широко використовують для сегментації ринку.

Перед кластеризацією зменшено розмірність даних методом PCA з двома вимірними векторами (компонентами).

За методом ліктя визначено кількість значень на другому кластері. Оцінка Silhoutte також є найвищою для другого кластеру. Також спостерігається значне зменшення помилки кластеру з 2 до 5, а після 6 зменшення не є великим. Відповідно обрано n_clusters = 5, щоб правильно сегментувати організаторів тендерів.

На рис. 5 подано кластеризацію K-means організаторів тендерів, де кількість кластерів дорівнює 5. Графік представлено відносно 2-компонентного PCA методу. На boxplot діаграмі зображено викиди до кожного кластеру, теж у розріз двокомпонентного PCA методу. До кожного кластеру віднесено таку кількість значень: кластер за номером 0 — 494; кластер за номером 3 — 475; кластер за номером 2 — 352; кластер за номером 1 — 345; кластер за номером 4 — 155.

Аналіз boxplot показав, що перший компонент (кластер 0) має найменше викидів, що вказує на детальний розподіл. Проведено докладний аналіз цього кластеру: мінімум — 0,6, Q1 — 1,2, медіана — 1,5, Q3 — 2, максимум — 3,2; є декілька викидів. Другий кластер (1): мінімум — -2,5, Q1 — -1,5, медіана — -1,4, Q3 — -0,9, максимум — -0,1; викидів немає. Третій кластер (2): мінімум — 0, Q1 — 1,3, медіана — 1,8, Q3 — 2,4, максимум — 3,8; викидів немає. Четвертий кластер (3): мінімум — -2,5, Q1 — -1,3, медіана — -1, Q3 — -0,5, максимум — 0,6; один викид. П'ятий кластер (4): мінімум — -0,6, Q1 — -0,2, медіана — 0,3, Q3 — 0,5, максимум — 1,2; викидів немає.

При порівнянні оцінок RTF та груп K-means з організаторами трендів, виявлено збіг між групою, що активно проводить тендери і залучає значні кошти, та групою, яка мало активна у цьому напрямі і не досягає великих обсягів фінансування. Інші групи організаторів тендерів демонструють лише частковий збіг.

На основі цих даних можна проводити передбачення груп за допомогою методів машинного навчання. Для цього використано метод Logistic Regression [9] та Naive Bayes [2], адже ці методи мають найпростішу логіку кваліфікації та хороші результати оцінки моделювання.

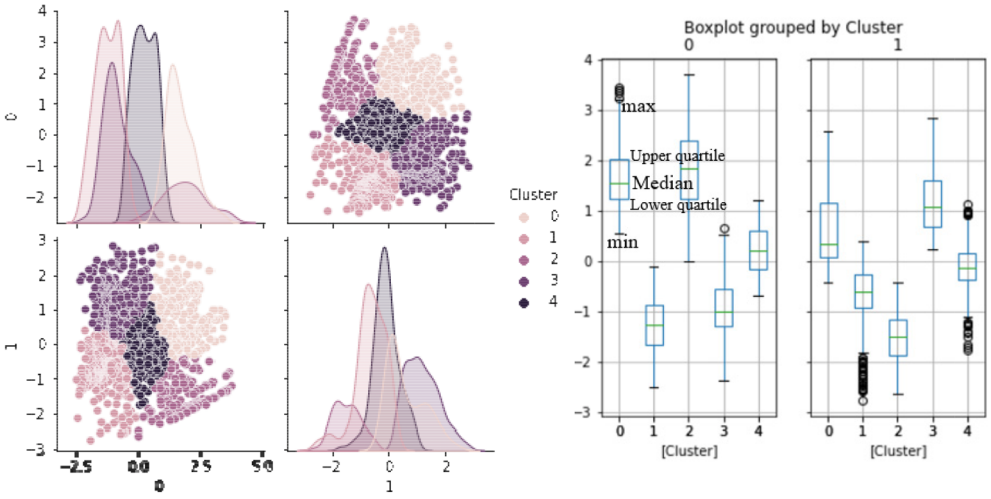

Рис. 5. Кластеризація K-means організаторів тендерів

ID_Orzanizer	Recency	Frequency	Monetary	Cluster	RFMScore
1551	30.0	1.0	3060000.0	3	413
695	309.0	1.0	2598260.0	3	413
954	1.0	1.0	1401975.0	4	413
598	309.0	1.0	1140000.0	3	413
579	30.0	1.0	1070856.0	3	413
...	...	...	...	...	...
1349	317.0	1.0	43.0	0	411
1754	30.0	1.0	36.0	0	411
1214	317.0	1.0	32.0	0	411
300	290.0	1.0	24.0	0	411
3	154.0	1.0	7.0	0	411

Рис. 6. Результат кластеризації RTF оцінки та K-means

Для навчання взято 70 % вибірки. Проведено навчання алгоритмами Logistic Regression і Naive Bayes. Після проведення тестування результатами оцінювання є, для обох методів:

```
Train Set Accuracy for Power Transformed Data: 100.0 %
Test Set Accuracy for Power Transformed Data: 100.0 %

Bias Error: 0.0
Variance Error: 0.0
```

Результати моделювання свідчать, що RTF оцінки та K-means дають стовідсоткову точність групування, відповідно за цими даними у подальшому можна проводити класифікацію організаторів тендерних проектів, що дає змогу визначати більш привабливих організаторів тендерів.

Отже, вибірка містить 92 638 аукціонів, 29 164 унікальні аукціони та 39 747 унікальних організаторів. На основі RFM-аналізу сформовано такі групи: The best organizers of tenders — 119; Loyal organizers of tenders — 385; Large consumers — 455; Tenders are seldom held, but for a large sum — 15; Weak tender organizers — 48. На основі кластеризації за методом K-means віднесено таку кількість значень: кластер за номером 0 — 494; кластер за номером 3 — 475; кластер за номером 2 — 352; кластер за номером 1 — 345; кластер за номером 4 — 155. Після проведення тестування алгоритмами Logistic Regression та Naive Bayes, результатами оцінювання є для обох методів: Train Set Accuracy for Power Transformed Data — 100 %; Test Set Accuracy for Power Transformed Data — 100 %.

Висновок. Авторка розробила метод формування навчальної вибірки на основі машинного навчання, що дає змогу сформувати вибірку на основі об'єднання даних RFM-аналізу та кластерного аналізу. Метод реалізований на основі вхідних даних за завершеними угодами учасників тендерів в Україні з сайту ProZorro Продажі. Вибірка містить 92 638 аукціонів, 29 164 унікальні аукціони та 39 747 унікальних організаторів. На основі RFM-аналізу сформовано такі групи: The best organizers of tenders — 119; Loyal organizers of tenders — 385; Large consumers — 455; Tenders are seldom held, but for a large sum — 15; Weak tender organizers — 48. На основі кластеризації методом K-means віднесено таку кількість значень: кластер за номером 0 — 494; кластер за номером 3 — 475; кластер за номером 2 — 352; кластер за номером 1 — 345; кластер за номером 4 — 155. Порівнюючи оцінки RTF та групи K-means з організаторами трендів, виявлено збіг між групою, що активно організовує тендери та залучає значні кошти, і групою, яка мало залучена до проведення тендерів та не отримує значних сум. Тестування даних проведено алгоритмами Logistic Regression і Naive Bayes. Після проведення тестування результатами оцінювання є, для обох методів: Train Set Accuracy for Power Transformed Data — 100 %; Test Set Accuracy for Power Transformed Data — 100 %.

До напрямів подальших наукових досліджень слід віднести розроблення автоматизованої системи для вибору організаторів тендерів на основі машинного навчання, що також дасть можливість автоматизувати процес участі у тендері.

Список літератури

1. Anitha P. RFM model for customer purchase behavior using K-Means algorithm [Electronic resource] / P. Anitha, Malini M. Patil // Journal of King Saud University — Computer and Information Sciences. — 2019. — https://doi.org/10.1016/j.jksuci.2019.12.011.
2. Classification Method of Fictitious Enterprises Based on Gaussian Naive Bayes [Electronic resource] / Andriy Krysovatyy [et al.] // 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT), LVIV, Ukraine, 22–25 September 2021. — [S. l.], 2021. — https://doi.org/10.1109/csit52700.2021.9648584.
3. Kamilov M. Algorithm for the Development of a Training Set that Best Describes the Objects of Recognition [Electronic resource] / M. Kamilov, M. Hudayberdiev, A. Khamroev // Procedia Computer Science. — 2019. — Vol. 150. — Pp. 116–122. — https://doi.org/10.1016/j.procs.2019.02.024 (date of access: 18.08.2023).
4. Kianisarkaleh A. Nonparametric feature extraction for classification of hyperspectral images with limited training samples [Electronic resource] / Azadeh Kianisarkaleh, Hassan Ghassemian // ISPRS Journal of Photogrammetry and Remote Sensing. — 2016. — Vol. 119. — Pp. 64–78. — https://doi.org/10.1016/j.isprsjprs.2016.05.009.
5. Li G. Application of Improved K-Means Clustering Algorithm in Customer Segmentation [Electronic resource] / Gang Li // Applied Mechanics and Materials. — 2013. — Vol. 411–414. — Pp. 1081–1084. — https://doi.org/10.4028/www.scientific.net/amm.411-414.1081.
6. Method of Forming Classified Training Sample in Case of Spacial Signal Processing under Influence of Combined Interference

[Electronic resource] / D. M. Piza [et al.] // Radioelectronics and Communications Systems. — 2018. — Vol. 61, no. 7. — Pp. 325–331. — https://doi.org/10.3103/s0735272718070051.

7. Piza D. M. Methods of Forming Classified Training Sample for Adaptation of Weight Coefficient of Automatic Interference Compensator [Electronic resource] / D. M. Piza, G. V. Moroz // Radioelectronics and Communications Systems. — 2018. — Vol. 61, no. 1. — Pp. 32–37. — https://doi.org/10.3103/s0735272718010041.
8. ProZorro Продажі. — https://bi.prozorro.sale/#/participantsCard.
9. Recognizing the Fictitious Business Entity on Logistic Regression Base [Electronic resource] / Andriy Krysovatyy [et al.] // Proceedings of the 3rd International Workshop on Intelligent Information Technologies & Systems of Information Security Khmelnytskyi. — 2022. — Vol. 3156. — Pp. 218–227. — https://ceur-ws.org/Vol-3156/paper15.pdf.

References

Anitha, P., & Patil, M. M. (2019). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University — Computer and Information Sciences.* https://doi.org/10.1016/j.jksuci.2019.12.011

Kamilov, M., Hudayberdiev, M., & Khamroev, A. (2019b). Algorithm for the Development of a Training Set that Best Describes the Objects of Recognition. *Procedia Computer Science*, *150*, 116–122. https://doi.org/10.1016/j.procs.2019.02.024.

Kianisarkaleh, A., & Ghassemian, H. (2016). Nonparametric feature extraction for classification of hyperspectral images with limited training samples. *ISPRS Journal of Photogrammetry and Remote Sensing*, *119*, 64–78. https://doi.org/10.1016/j.isprsjprs.2016.05.009.

Krysovatyy, A., Lipianina-Honcharenko, H., Sachenko, S., Desyatnyuk, O., Banasik, A., & Lukasevych-Krutnyk, I. (2022). Recognizing the Fictitious Business Entity on Logistic Regression Base. *Proceedings of the 3rd International Workshop on Intelligent Information Technologies & Systems of Information Security Khmelnytskyi*, *3156*, 218–227. https://ceur-ws.org/Vol-3156/paper15.pdf.

Krysovatyy, A., Lipyanina-Goncharenko, H., Desyatnyuk, O., Sachenko, S., Lukasevych-Krutnyk, I., & Butrin-Boka, N. (2021). Classification Method of Fictitious Enterprises Based on Gaussian Naive Bayes. In *2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT). IEEE*. https://doi.org/10.1109/csit52700.2021.9648584.

Li, G. (2013). Application of Improved K-Means Clustering Algorithm in Customer Segmentation. *Applied Mechanics and Materials*, *411–414*, 1081–1084. https://doi.org/10.4028/www.scientific.net/amm.411-414.1081.

Piza, D. M., & Moroz, G. V. (2018). Methods of Forming Classified Training Sample for Adaptation of Weight Coefficient of Automatic Interference Compensator. *Radioelectronics and Communications Systems*, *61* (1), 32–37. https://doi.org/10.3103/s0735272718010041.

Piza, D. M., Bugrova, T. I., Lavrentiev, V. N., & Semenov, D. S. (2018). Method of Forming Classified Training Sample in Case of Spacial Signal Processing under Influence of Combined Interference. *Radioelectronics and Communications Systems*, *61* (7), 325–331. https://doi.org/10.3103/s0735272718070051.

ProZorro. https://bi.prozorro.sale/#/participantsCard.

Kh. Lipianina-Honcharenko

METHOD FOR FORMING TRAINING SAMPLES FOR DATA ARRAYS BASED ON MACHINE LEARNING

The study introduces an innovative methodology for crafting training samples through the integration of machine learning techniques. This method encompasses a fusion of RFM (Recency, Frequency, Monetary) analysis and cluster analysis, offering a comprehensive approach to sample formation. The application of this approach is demonstrated on a dataset derived from concluded tender agreements by participants in Ukraine, sourced from the ProZorro Sales platform. The compiled dataset encompasses an impressive volume, encompassing a total of 92,638 auctions, which further breaks down into 29,164 distinct auctions and an assemblage of 39,747 unique organizers.

The utilization of RFM analysis within this framework yields the categorization of the dataset into distinct groups, each characterized by its own distinct attributes. These groupings include designations such as "The Best Organizers of Tenders," "Loyal Organizers of Tenders," "Large Consumers," "Tenders Held Infrequently but with Substantial Sums," and "Weak Tender Organizers." Following the RFM analysis, the K-means clustering methodology is implemented, resulting in the division of the data into five clusters, each contributing to a nuanced differentiation of diverse organizer profiles.

Intriguingly, a comparative analysis involving RTF (Relative Total Frequency) scores and the K-means groupings reveals congruence between clusters representing organizers who actively orchestrate numerous tenders with significant monetary value, as well as clusters characterized by minimal tender activity with less substantial monetary implications. To validate the efficacy of the proposed method, rigorous testing is conducted employing Logistic Regression and Naive Bayes algorithms. Encouragingly, the results consistently showcase impressive accuracy for both methods, highlighting their robustness.

An outlook towards future research endeavors suggests a promising avenue of developing an automated system for the selection of tender organizers, underpinned by machine learning principles. Such a system would undoubtedly revolutionize the optimization of participation strategies within the domain of tender processes, fostering efficiency and accuracy in decision-making.

Keywords: training sample, machine learning, RFM analysis, cluster analysis, tenders.

Матеріал надійшов 20.08.2023