Научная статья
Original article
УДК 004.85, 004.056
DOI: 10.24411/2658-4964-2021-103682

МЕТОДЫ И ПОДХОДЫ К ПРЕДОБРАБОТКИ ДАННЫХ ПЛАТЕЖЕЙ ПРИ УСЛОВИИ СИЛЬНОЙ НЕСБАЛАНСИРОВАННОСТИ КЛАССОВ

METHODS AND APPROACHES TO PREPARATION OF PAYMENT DATA UNDER CONDITIONS OF STRONG CLASS IMBALANCE

Менщиков Александр Алексеевич, кандидат технических наук, ординарный доцент, доцент факультета Безопасности информационных технологий, Национальный исследовательский университет ИТМО, Россия, г. Санкт-Петербург

Федосенко Максим Юрьевич, студент, 2 курс, факультет Безопасности информационных технологий, Национальный исследовательский университет ИТМО, Россия, г. Санкт-Петербург

Menshikov Alexander Alekseevich, menshikov@itmo.ru

Fedosenko Maxim Yurievich

Аннотация

Алгоритмы машинного обучения широко применяются во многих сферах деятельности. Задача классификации обычно направлена на минимизацию количества ложных срабатываний. Например, в банковской сфере задача классификации может быть применена к обработке транзакций для выявления мошеннических действий или мошеннических схем.

Однако, при разработке самообучающейся модели возникает проблема несбалансированности данных, так как соотношение мошеннических транзакций к общему количеству транзакций слишком мало. При использовании стандартных методов классификации в такой ситуации часто возникает проблема, что при уменьшении общей ошибки классификатор полностью относит интересуемый класс к шуму. В таком случае возникает парадокс точности, когда показатели точности отражают только распределение базового класса, игнорируя при этом интересуемый класс нелегитимных транзакций. Таким образом, появляется потребность в решении проблемы несбалансированности.

Основной целью работы является обзор и анализ существующих методов борьбы с сильной несбалансированностью классов. В работе рассматриваются основные методы обработки и анализа больших несбалансированных наборов данных платежей с целью их дальнейшего применения при обучении банковских антифрод систем. Проведен обзор основных направлений для борьбы с сильной несбалансированностью классов, среди которых упомянуты технологии over-sampling и under-sampling. Приведена авторская сравнительная характеристика технологий, где отображены их основные достоинства и недостатки практического применения. Разобран принцип работы алгоритма SMOTE, возможность его применения к данным платежей в совокупности с используемыми алгоритмами машинного обучения.

Было выявлено, что при реализации системы распознавания нелегитимных транзакций в банковской сфере главной проблемой является сильная несбалансированность классов в массиве данных. Требуется, чтобы модель машинного обучения в режиме реального времени с обучением без учителя в условиях сильной несбалансированности классов умела точно распознавать нелегитимные транзакции.

Annotation

Machine learning algorithms are widely used in many fields of activity. The classification task is usually aimed at minimizing the number of false positives. For

example, in the banking sector, the classification task can be applied to the processing of transactions to identify fraudulent activities or fraudulent schemes.

However, when developing a self-learning model, the problem of data imbalance arises, since the ratio of fraudulent transactions to the total number of transactions is too small. When using standard classification methods in such a situation, the problem often arises that when the total error is reduced, the classifier completely attributes the class of interest to noise. In this case, the accuracy paradox arises, when the accuracy indicators reflect only the distribution of the base class, while ignoring the class of illegitimate transactions of interest. Thus, there is a need to solve the problem of imbalance.

The main purpose of the work is to review and analyze existing methods of dealing with a strong imbalance of classes. The paper considers the main methods of processing and analyzing large unbalanced sets of payment data for the purpose of their further application in the training of bank anti-fraud systems. The review of the main directions for dealing with a strong imbalance of classes, among which the over-sampling and under-sampling technologies are mentioned, is carried out. The author's comparative characteristics of technologies are given, where their main advantages and disadvantages of practical application are displayed. The principle of operation of the SMOTE algorithm , the possibility of its application to payment data in conjunction with the machine learning algorithms used are analyzed.

It was revealed that when implementing a system for recognizing illegitimate transactions in the banking sector, the main problem is a strong imbalance of classes in the data array. It is required that a real-time machine learning model with unsupervised learning in conditions of severe class imbalance should be able to accurately recognize illegitimate transactions.

Ключевые слова: машинное обучение, несбалансированность классов, бинарная классификация, парадокс точности, транзакции, антифрод, фрод-мониторинг, финансовое мошенничество, 115-ФЗ, oversampling, undersampling, алгоритм SMOTE, алгоритм ASMO.

Keywords: machine learning, class imbalance, binary classification, accuracy paradox, transactions, anti-fraud, fraud monitoring, financial fraud, 115-FZ, oversampling, undersampling, SMOTE algorithm, ASMO algorithm.

В современной тенденции развития информационных технологий, огромное значения уделяется исследованиям в области разработки искусственного интеллекта и производным от него направлениям: data science (технологии обработки больших данных) и machine learning (машинное обучение). В настоящее время, алгоритмы машинного обучения широко применяются во многих сферах деятельности: наука, медицина, спорт, интернет вещей, аналитика. Задачи классификации, решаемые data science и machine learning обычно направлены на минимизацию количества ложных срабатываний, т.е. неверной классификации информационных объектов [1]. Например, в банковской сфере задача классификации может быть применена к обработке транзакций для выявления мошеннических действий или мошеннических схем. Однако, при разработке самообучающейся модели возникает проблема несбалансированности данных, так как соотношение мошеннических транзакций к общему количеству транзакций слишком мало (по данным банка России доля мошеннических транзакций составила 0,0016 за 2017 год) [2]. При использовании стандартных методов классификации в данной ситуации часто возникает проблема, что при уменьшении общей ошибки классификатор полностью относит интересуемый класс к шуму. В таком случае возникает парадокс точности, когда показатели точности отражают только распределение базового класса. Таким образом, появляется потребность в решении проблемы несбалансированности. [3]

Схематично, проблема несбалансированности для мошеннических транзакций представлена на рисунке 1:

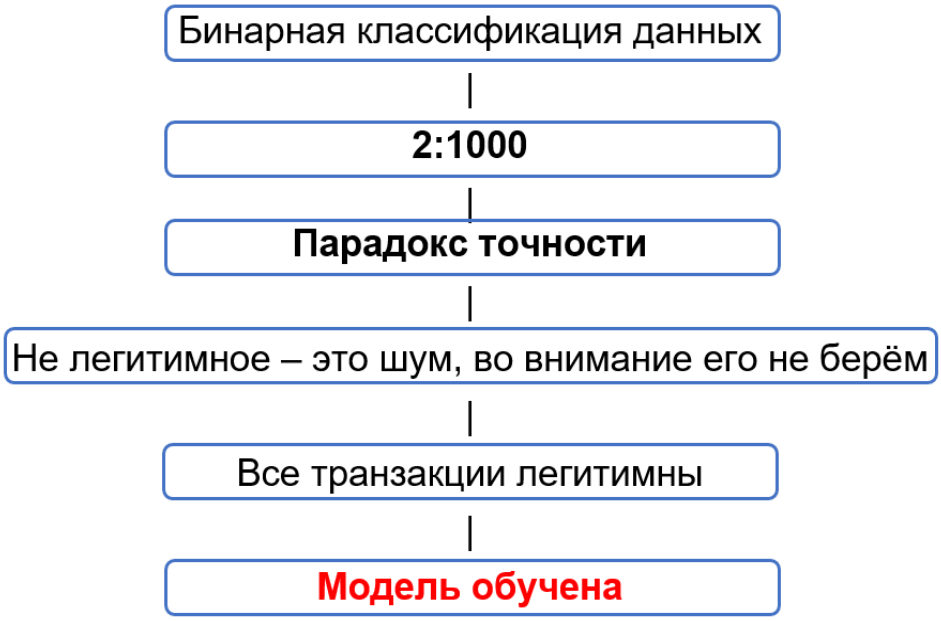

Рисунок 1 – Структурная схема парадокса несбалансированного примера [1]

Несбалансированность классов характеризуется непропорциональным соотношением наблюдений для каждого из классов. В рассматриваемом примере - для бинарной классификации, состоящей из класса легитимных и класса мошеннических транзакций. Это является одной из существенных проблем машинного обучения. Дисбаланс классов существует во многих областях и не только при обнаружении мошенничества. Большинство алгоритмов машинного обучения работают оптимальным образом, когда количество выборок в каждом классе является примерно одинаковым или имеет соотносимую пропорциональность (50:50, 70:30, 80:20). Это связано с тем, что большинство алгоритмов разработаны таким образом, чтобы максимизировать точность и уменьшить погрешность, что также является основной задачей анализа данных и последующего машинного обучения. Однако, в условиях сильного дисбаланса классов возникает угроза переобучения алгоритма (т.е. неверного обучения), что приводит к некорректным оценкам классификации объектов и последующим ошибкам в применяемой производственной области. Таким образом, чтобы избежать подобного явления и достичь хороших результатов, необходимо исследовать методы работы с несбалансированными данными, а также

разрабатывать эффективные алгоритмы их классификации. Особенно в условиях реальных данных, при которых не избежать данного явления, как в случае соотношения мошеннических и легитимных транзакций.

В банковской сфере, применение методов машинного обучения, используется для решения следующих задач производственного процесса:

- Автоматизация парсинга документооборота (персональные данные клиентов и сотрудников, различные договора и т.д.);
- Автоматизация выполнения стандартных процедур, имеющих однозначный исход (купля-продажа валюты на биржах);
- Выявление должников и потенциальных неплательщиков по услугам;
- Кредитный и потребительский скоринг;
- Аналитика платежей в рамках 115-ФЗ «О противодействии легализации (отмыванию) доходов, полученных преступным путем, и финансированию терроризма» [4];
- Системы фрод-мониторинга и предотвращение электронного мошенничества.

Фрод-мониторинг представляет собой систему, предназначенная для оценки финансовых транзакций на предмет подозрительности с точки зрения мошенничества [5]. Разновидностью мошенничества в области информационных технологий является фрод (от англ. fraud), а меры, направленные на предотвращение данных действий - антифродом. Часто с понятием фрода ассоциируют операции, связанные с хищением денежных средств с банковских карт и счетов. Существуют и другие типы фрода, например, предназначенные для получения финансовой выгоды за регистрацию новых пользователей, искусственного увеличения популярности веб-ресурса, легализацией и отмыванием полученных незаконным путем финансовых средств, получением кредитов. [6] Проблема несбалансированности классов наиболее актуальна для данных систем. В результате неверного обучения модели определения легитимности платежей, система сталкивается с **ошибками первого рода** (когда система определяет мошенническую транзакцию как

легитимную) и **ошибками второго рода** (когда система определяет легитимную транзакцию как мошенническую). [3]

Отсюда, при решении проблемы несбалансированности классов, определяют следующие виды информационных объектов [7]:

- **Мажоритарные объекты** – подавляющее большинство в бинарной классификации (в данном случае легитимные транзакции)
- **Миноритарные объекты** - подавляющее меньшинство в бинарной классификации (в данном случае нелегитимные транзакции)

На основе данных видов, выделяют следующие подходы при анализе несбалансированных данных:

1. Методы, направленные на обработку входных данных: изменяют соотношения обучающей выборки для балансировки распределения и допустимого удаления сложных для анализа образцов.
2. Методы, направленные на оптимизацию используемых алгоритмов: изменяют и модифицирует существующие алгоритмы обучения модели, в результате чего происходит смещение по отношению к мажоритарным объектам. Затем происходит адаптация алгоритмов для работы с искажённым распределением.
3. Гибридные методы: представляют собой сочетание преимуществ двух предыдущих методов, минимизируя при этом их недостатки. На практике метод является наиболее актуальным и продуктивным.

Результатом применения гибридных методов является появление следующих технологий обработки данных: **over-sampling** и **under-sampling**.

При использовании under-sampling происходит уменьшение размера мажоритарного класса, а для обучения модели берётся вся выборка миноритарного класса. Получить искусственно необходимое для модели соотношение возможно путём удаления из выборки следующих транзакций:

- Транзакции, совершённые в регионе держателя карты - исключает применения электронного мошенничества из других стран).

- Транзакции, математическое ожидание суммы которых не выходит из диапазона среднеквадратичного отклонения – вероятнее всего, не могут являться мошенническими.
- Транзакции по назначению платежа (зарплата, выплата пособий) – представляют собой периодические платежи с явно обозначенной целью и способом реализации.
- Дублирующиеся транзакции – увеличивают мажоритарный класс однотипными данными, что в свою очередь ухудшает соотношение и даёт дополнительную нагрузку на вычислительные мощности при обучении модели.

Из оставшейся выборки данные берутся случайном образом или согласно предложенному аналитиками более сложному математическому алгоритму (например, при помощи применения теория графов, случайный леса и т.д.).

При использовании over-sampling происходит увеличение размера миноритарного класса при неизменном количественном составе мажоритарного класса. Получить искусственно необходимое для рассматриваемой модели соотношение возможно при использовании следующих подходов:

- Подход, основанный на добавлении транзакций, сумма которых отличается от математического ожидания – достаточно грубый метод, однако увеличивает количество критериев для отнесения транзакции к мошеннической.
- Подход, основанный на учёте в выборке всех транзакций между регионами – позволяет обучить систему реагировать на данные, в обычном случае нехарактерные транзакции с целью их дальнейшего анализа аналитиком платежей.
- Подход, основанный на создании общих групп из элементов легитимных и нелегитимных классов согласно какой-нибудь классификации – представляет собой механизм оптимизации обработки данных с выделением общих закономерностей
- Подход, основанный на добавлении транзакций, совершённых лицами из «стоп-листа» банка и перечня «Росфинмониторинга» - определяет транзакцию в миноритарный класс согласно имеющимся законным актам.

В данном методе, самым оптимальным набором данных являются дублирующиеся составляющие нелегитимного класса.

Сравнительная характеристика применения рассмотренных технологий, представлена в таблице 1:

Таблица 1 – Сравнительная характеристика Over-Sampling и Under-Sampling [3]

	Достоинства	**Недостатки**
Over-Sampling	1. Класс нелегитимных транзакций останется неизменным; 2. Класс легитимных транзакций окажется отфильтрован, похожие транзакции будут объединены или удалены; 3. Количество необходимых вычислительных ресурсов не будет увеличено.	1. Потери важных данных вследствие уменьшения мажоритарного класса; 2. Вследствие предыдущего пункта, вероятность неверного обучения модели возрастает;
Under-Sampling	1. Все имеющиеся естественные транзакции будут сохранены; 2. Часть данных класса легитимных транзакций будет верно переопределён.	1. Класс нелегитимных транзакций будет искусственно увеличен; 2. Большие затраты вычислительных ресурсов.

Иллюстрация распределения выборки до и после применения рассмотренных алгоритмов приведена ниже на рисунке 2:

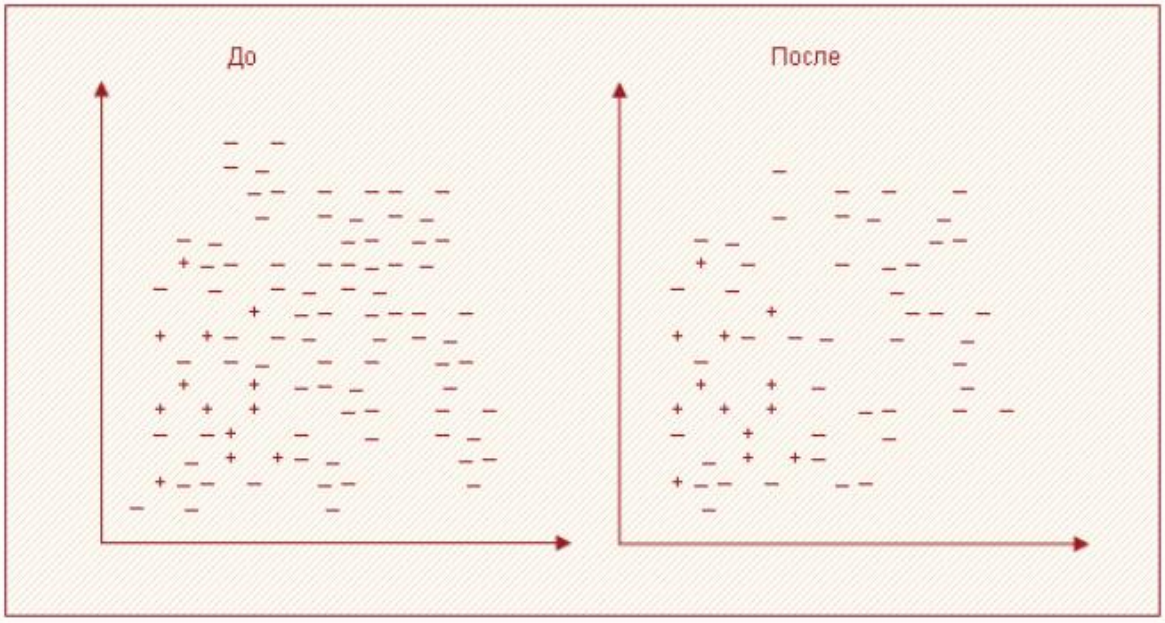

Рисунок 2 –Результат применения Sampling технологий на выборку бинарных данных [5].

Для балансировки классов наибольшей популярностью пользуется алгоритм **SMOTE** (Synthetic Minority Oversampling Technique), представляющий собой метод передискретизации синтетического меньшинства в используемой выборке. Он направлен на то, чтобы сбалансировать распределение классов при помощи случайного увеличения объектов класса меньшинства путем их тиражирования. SMOTE синтезирует новые экземпляры меньшинства между существующими экземплярами миноритарного класса, генерируя виртуальные записи обучения путем линейной интерполяции. Эти синтетические обучающие записи генерируются путем случайного выбора одного или нескольких k-ближайших соседей для каждого примера в миноритарном классе [8]. После процесса передискретизации данные восстанавливаются, и к обработанным данным можно применить несколько моделей классификации. Алгоритм имеет следующие преимущества:

- Позволят «избавиться» от дублей данных;
- Задаёт метод группировки на основе k-ближайших соседей;
- Минимизирует проблему потери «критически важных» данных из мажоритарного класса;
- Не учитывает соседние пример мажоритарного класса, что способно привести к «перекрытию» классов и добавить дополнительный шум в имеющийся набор данных [8].

Графическую интерпретацию и результаты применения данного алгоритма можно пронаблюдать на рисунке 3:

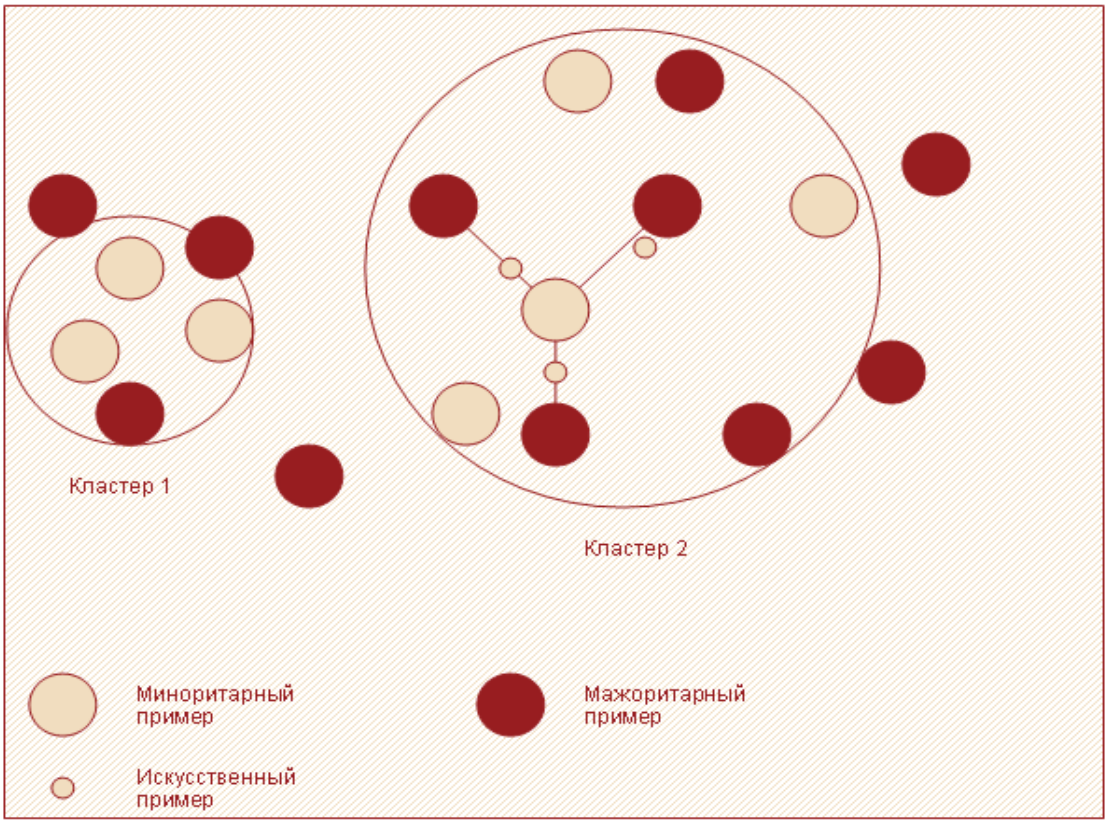

Рисунок 3 – Результат применения алгоритма SMOTE [5].

Недостаток данного подхода заключается в том, что «вслепую» увеличивает плотность объектами в области слабо представленного класса. В случае, если миноритарные объекты равномерно распределены среди мажоритарных и имеют низкую плотность, алгоритм SMOTE способен только сильнее перемешать классы.

Также, при работе над балансировкой обучающей выборки возникает ряд противоречий, который требует дополнительного анализа и работой над методами оптимизации, учитывающие следующие аспекты:

1. Увеличение соотношения класса нелегитимных транзакций к классу легитимных, например, с 2:1000 до 100:1000 требует больших вычислительных ресурсов, а также «засоряет» нелегитимный класс не всегда уместными «дублями». **Это приведет к переобучению, а также к увеличению числа ошибок первого рода.**
2. Уменьшение соотношения класса легитимных транзакций 1000:2 до 100:2 способно потерять слишком много информации об легитимных операциях. **Это**

приведет к переобучению, а также к увеличению числа ошибок второго рода.

Для борьбы с приведёнными противоречиями стоит рассмотреть гибридные методы использования алгоритмов, например, совокупность SMOTE с алгоритмом адаптивного искусственного увеличения числа объектов миноритарного класса ASMO (Adaptive Synthetic Minority Oversampling) [9].

В данной работе был произведен обзор и сравнение уже существующих методов борьбы с сильной несбалансированностью классов, на основе чего получены следующие выводы:

1. Поскольку количество мошеннических транзакций очень мало по отношению ко всем транзакциям, то проблема несбалансированности классов при анализе данных имеет место быть;
2. В наборе данных количество записей одного класса может быть очень мало, а другого класса, наоборот, велико. В результате, при использовании алгоритмов машинного обучения, чувствительные к балансу классы алгоритма переобучаются;
3. При решении задач выявления мошенничества, большое значение имеют предварительный тщательный анализ данных и выбор правильной методологии построения и валидации эффективности моделей, так как в противном случае велика вероятность их переобучения;
4. Не существует одного стандартного решения, которое бы одинаково хорошо подходило для любых задач выявления мошенничества, — в каждом конкретном случае необходим индивидуальный подход, учитывающий все особенности проблемы и требования к работе системы борьбы с мошенничеством.

Что касается банковским антифрод систем, то для обнаружения и анализа финансового мошенничества используются различные алгоритмы обучения, как с учителем (supervised learning), так и без учителя (unsupervised learning) [10]. В первом случае речь идет об алгоритмах классификации, когда обучающая выборка содержит в себе заранее известные ответы, во втором — когда таких ответов нет. Также, для оптимизации процесса обработки данных,

последовательности транзакций могут быть рассмотрены в качестве обыкновенного текста, и тогда появляется возможность применения методов анализа текстовых данных и обработки естественного языка (Natural Language Processing, NLP). Однако, при использовании алгоритмов классификации в процессе обучения модели на исторических данных, можно прогнозировать вероятность, с которой транзакция окажется мошеннической. Это, в свою очередь, будет являться полезным предиктором при дальнейшем обучении модели следующими алгоритмами машинного обучения: Случайный лес (Random Forest), Градиентный бустинг (Gradient boosted models), Логистическая регрессия (Logistic Regression), Наивный Байесовский классификатор, Машины опорных векторов (Support Vector Machines) [11]. В настоящее время, данные алгоритмы наиболее часто применяются в процессе обучения моделей в силу своих оптимизационных достоинств и высокой точности при дальнейших предсказаниях [12].

Литература

1. Махсотова Ц. В. Исследование методов классификации при несбалансированности классов // Научный журнал. 2018. № 5 (18). С 35-36
2. Обзор несанкционированных переводов денежных средств за 2017 год // Центральный банк Российской Федерации. 2018. С. 4-8.
3. Н.Б. Паклин, С.В. Уланов, С.В. Царьков. Построение классификаторов на несбалансированных выборках на примере кредитного скоринга // Искусственный интеллект. 2010. № 3. С. 528-534
4. Марахтанов А.Г., Паренченков Е.О., Смирнов Н.В. Определение электронного мошенничества методами машинного обучения в случае несбалансированного набора данных // Вестник Пермского национального исследовательского политехнического университета. электротехника, информационные технологии, системы управления. 2020. № 36. С. 80-95
5. Береговой В.Ю. Киберпреступность в финансовой сфере в России. // Банковский бизнес. 2016. №4. С. 35-39

6. Нечаев А.В. Антифрод системы и принципы их работы. // Труды Северокавказского филиала Московского технического университета связи и информатики. 2019. №2. С. 88-94.
7. Различные стратегии сэмплинга в условиях несбалансированности классов. [Электронный ресурс]. URL: https://basegroup.ru/community/articles/imbalance-datasets (дата обращения: 05.08.2021).
8. ML | обработка несбалансированных данных с помощью алгоритма SMOTE и NEAR MISS в Python. [Электронный ресурс]. URL: http://espressocode.top/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/ (дата обращения: 05.08.2021).
9. Он, Хайбо, Ян Бай, Эдвардо А. Гарсия и Шутао Ли. «ADASYN: Адаптивный подход синтетической выборки для несбалансированного обучения», в Международной объединенной конференции IEEE по нейронным сетям (IEEE World Congress по вычислительному интеллекту), 2008. С. 1322–1328
10. Осипова Т. А., Зайцев К. С., Биферт В. О. Применение алгоритмов машинного обучения к задаче выявления мошенничества при использовании пластиковых карт. // International journal of open information technologies. 2021. № 8. С. 23-29
11. Замятина Е.В., Луценко А.В. Анализ значимости алгоритмов machine learning в антифрод - системах коммерческого банка. // Материалы и методы инновационных исследований и разработок. 2018. №1. С. 129-131
12. Рыбник А.А. Применение искусственного интеллекта и машинного обучения в банковском бизнесе // Актуальные проблемы развития банковского и небанковского секторов экономики. 2018. №1. С. 250-257

Literature

1. Makhsotova Ts. V. Investigation of classification methods with unbalanced classes // Scientific journal. 2018. No. 5 (18). P. 35-36

2. Review of unauthorized money transfers for 2017 // Central Bank of the Russian Federation. 2018. P. 4-8.
3. N.B. Paklin, S.V. Ulanov, S.V. Tsar'kov. Construction of classifiers on unbalanced samples using the example of credit scoring // Artificial Intelligence. 2010. No. 3. P. 528-534
4. Marakhtanov A.G., Parenchenkov E.O., Smirnov N.V. Determination of electronic fraud by machine learning methods in the case of an unbalanced data set // Bulletin of the Perm National Research Polytechnic University. Electrical engineering, information technology, control systems. 2020. No. 36. P. 80-95
5. Beregovoy V.Yu. Cybercrime in the financial sector in Russia. // Banking business. 2016. No. 4. P. 35-39
6. Nechaev A.V. Antifraud systems and principles of their work. // Proceedings of the North Caucasus Branch of the Moscow Technical University of Communications and Informatics. 2019. No. 2. P. 88-94.
7. Different strategies of sampling in conditions of imbalance of classes. [Electronic resource]. URL: https://basegroup.ru/community/articles/imbalance-datasets (date of access: 05.08.2021).
8. ML | handling unbalanced data using the SMOTE and NEAR MISS algorithm in Python. [Electronic resource]. URL: http://espressocode.top/ml-handling-imbalanced-data-with-smote-and-near-miss-algorithm-in-python/ (date accessed 05.08.2021).
9. He, Haibo, Yang Bai, Edward A. Garcia and Shutao Li. "ADASYN: An Adaptive Synthetic Sampling Approach for Unbalanced Learning", at the IEEE World Congress on Computational Intelligence, 2008. pp. 1322–1328
10. Osipova TA, Zaitsev KS, Bifert VO Application of machine learning algorithms to the problem of detecting fraud when using plastic cards. // International journal of open information technologies. 2021. No. 8. P. 23-29
11. Zamyatina E.V., Lutsenko A.V. Analysis of the significance of machine learning algorithms in antifraud systems of a commercial bank. // Materials and methods of innovative research and development. 2018. No. 1. P. 129-131

12. Rybnik A.A. Application of artificial intelligence and machine learning in the banking business // Actual problems of development of the banking and non-banking sectors of the economy. 2018. No. 1. P. 250-257

© Менщиков А.А., Федосенко М.Ю., 2021. Научно-образовательный журнал для студентов и преподавателей «StudNet» №9/2021.