http://dx.doi.org/10.35596/1729-7648-2025-23-5-66-74

УДК 004.93

ЭКСПЕРИМЕНТАЛЬНЫЕ ИССЛЕДОВАНИЯ ПО ПРИМЕНЕНИЮ МЕТОДОВ БАЛАНСИРОВКИ ДАННЫХ В ЗАДАЧАХ КЛАССИФИКАЦИИ

М. М. ЛУКАШЕВИЧ, Е. КЛИЦУНОВА

Белорусский государственный университет (Минск, Республика Беларусь)

Аннотация. Рассмотрены методы работы с несбалансированными данными при построении моделей машинного обучения для решения задачи классификации. Проведено исследование методов балансировки с определением их влияния на эффективность классических и ансамблевых моделей. Выбраны пять наборов данных различного объема и степени дисбаланса, выполнена их предобработка. Изучено влияние реализованных в библиотеке imbalanced-learn методов увеличения меньшего класса, уменьшения большего класса как при изолированном применении, так и при их комбинации. Определен диапазон оптимального соотношения классов после балансировки (от 1:1 до 2:1, где первое число соотносится с количеством объектов изначально меньшего класса) и оценено влияние подбора гиперпараметров при помощи Optuna. Установлено, что оптимизация гиперпараметров не компенсирует отсутствие балансировки данных, а наилучшие показатели качества моделей достигаются применением комплексного подхода с комбинацией двух методов балансировок различных типов, использованием ансамбля и подбором гиперпараметров. Наибольший вклад в качество моделей дало применение одного метода балансировки вместе с использованием ансамбля, поэтому такую комбинацию можно рекомендовать в условиях ограниченных временных и вычислительных ресурсов. Добавление метода уменьшения большего класса и подбор гиперпараметров целесообразно проводить при достаточном количестве ресурсов и высоких требованиях к качеству модели.

Ключевые слова: классификация, машинное обучение, несбалансированные данные, балансировка данных, сравнительный анализ, классические модели, ансамбли, оптимизация гиперпараметров.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Лукашевич, М. М. Экспериментальные исследования по применению методов балансировки данных в задачах классификации / М. М. Лукашевич, Е. Клицунова // Доклады БГУИР. 2025. Т. 23, № 5. С. 66–74. http://dx.doi.org/10.35596/1729-7648-2025-23-5-66-74.

EXPERIMENTAL STUDIES ON THE APPLICATION OF DATA BALANCING METHODS IN CLASSIFICATION PROBLEMS

MARINA M. LUKASHEVICH, KATERYNA KLITSUNOVA

Belarusian State University (Minsk, Republic of Belarus)

Abstract. This article examines methods for working with imbalanced data when building machine learning models for classification problems. Balancing methods are studied to determine their impact on the performance of classical and ensemble models. Five datasets of varying sizes and degrees of imbalance are selected and preprocessed. The impact of the imbalanced-learn library's methods of increasing the smaller class and decreasing the larger class is studied, both when used separately and in combination. The optimal class ratio after balancing is determined (from 1:1 to 2:1, where the first number corresponds to the number of objects in the initially smaller class), and the impact of hyperparameter selection using Optuna is assessed. It is established that hyperparameter optimization does not compensate for the lack of data balancing, and the best model performance is achieved by using an integrated approach combining two different types of balancing methods, using an ensemble, and hyperparameter selection. The greatest impact on model quality was achieved by using a single balancing method in conjunction with ensemble modeling, so this combination is recommended for limited time and computational resources. Adding a larger class reduction method and hyperparameter tuning is advisable when resources are sufficient and model quality requirements are high.

Keywords: classification, machine learning, imbalanced data, data balancing, comparative analysis, classical models, ensembles, hyperparameter tuning.

Conflict of interests. The authors declare no conflict of interests.

For citation. Lukashevich M. M., Klitsunova K. (2025) Experimental Studies on the Application of Data Balancing Methods in Classification Problems. *Doklady BGUIR*. 23 (5), 66–74. http://dx.doi.org/10.35596/1729-7648-2025-23-5-66-74 (in Russian).

Введение

Несбалансированные данные – это наборы данных, в которых количество объектов одного класса существенно превышает количество объектов других классов. Наиболее характерна проблема дисбаланса классов для задач классификации: она может привести к неудовлетворительному качеству моделей машинного обучения, поскольку классические и ансамблевые алгоритмы классификации рассчитаны на равное соотношение классов. При этом стандартные метрики оценки качества моделей, такие как точность (accuracy), могут давать излишне оптимистичную оценку, даже если все объекты меньшего класса классифицируются неверно. Особенно актуальна проблема несбалансированных данных для задач, где подобная неправильная классификация может иметь серьезные последствия: например, пропуск опасного заболевания при диагностике или мошеннических действиях.

Существует большое число исследований [1–4], направленных на разработку новых методов работы с несбалансированными данными. Эти методы можно условно разделить на следующие категории: модификация моделей (как классических, так и ансамблевых) и предварительная балансировка данных.

Среди балансировок наибольшее распространение получили методы увеличения меньшего класса (oversampling) (Random Oversampling, SMOTE, B-SMOTE, B-SMOTE SVM, ADASYN) и методы уменьшения большего класса (undersampling) (Random Undersampling, NearMiss, TomekLinks, CNN, ENN, OSS, NCR). Модификации ансамблей заключаются во внедрении некоторых из указанных балансировок, а модификации алгоритмов зачастую подразумевают обучение с учетом издержек классификации, взвешивание классов или изменения алгоритма, тесно связанные с особенностями реализации конкретной модели машинного обучения.

Цель исследования – экспериментальное сравнение эффективности различных методов балансировки. Изучение изолированного применения методов уже проводилось в [5, 6], и, хотя есть исследования их комбинаций [7], им все еще уделено мало внимания. Поэтому, помимо изолированного применения балансировок с классическими и ансамблевыми моделями классификации, обращено внимание на их сочетания. Оценен вклад настройки гиперпараметров в итоговое качество моделей. Отдельное внимание уделено подбору оптимального соотношения классов после балансировки.

Методы проведения исследования

Для сравнительного анализа были выбраны пять наборов данных с платформы Kaggle с различными объемами, степенью дисбаланса, размерностью признакового пространства и тематикой. Описание выбранных наборов данных с указанием размера, типа классификации и соотношения классов приведено в табл. 1.

Таблица 1. Описание выбранных наборов данных
Table 1. Description of selected datasets

Номер набора	Набор данных	Размер набора, тыс.	Тип классификации	Соотношение классов
1	Классификация здоровья плода	~2	Мультиклассовая	8:100 и 13:100
2	Прогнозирование церебрального инсульта	~43	Бинарная	2:100
3	Кредитный риск	~32		22:100
4	Классификация кредитного рейтинга	100	Мультиклассовая	17:100 и 29:100
5	Показатели здоровья при диабете	~254		2:100 и 15:100

На предварительном этапе выполнены разведочный анализ данных и предобработка. Каждый набор проверялся на наличие пропущенных значений, повторений и выбросов; пропуски заполнялись медианным значением, либо соответствующие записи удалялись, а выбросы корректировались значениями соседних точек или исключались. Категориальные признаки преобразовывались в числовой формат с использованием методов LabelEncoder. Для отбора наиболее значимых признаков применялся метод mutual_info_classif, нормализация числовых признаков выполнялась с помощью метода MinMaxScaler. Итоговая выборка делилась на тренировочную и тестовую в соотношении 70:30.

Основным языком программирования выбран Python благодаря наличию библиотеки imbalanced-learn, в которой реализованы методы балансировки данных. Вспомогательными библиотеками служили Pandas, NumPy и Scikit-learn для работы с наборами данных, Matplotlib и Seaborn для визуализации, XGBoost и LightGBM для использования ансамблевых моделей, а также фреймворк Optuna для подбора гиперпараметров.

В качестве основной метрики оценки применялась сбалансированная точность, вспомогательной – матрица ошибок. Сбалансированная точность определялась по формуле

$$BalancedAccuracy = \frac{1}{N}\sum_{i=1}^{N} Recall_i = \frac{1}{N}\sum_{i=1}^{N} \frac{TP_i}{TP_i + FN_i}, \quad (1)$$

где N – количество классов в наборе данных.

Для сравнения использовались как классические алгоритмы (дерево решений, MLP, *k*-ближайших соседей, наивный байесовский классификатор), так и ансамблевые методы (случайный лес, градиентный бустинг, XGBoost, AdaBoost и LightGBM).

Изолированное применение балансировок

Классические модели классификации после применения методов балансировки в среднем демонстрировали результаты не хуже моделей, полученных на исходных данных.

Для методов увеличения меньшего класса улучшение оказалось наиболее выраженным и составляло от 1 до 56 % со средним улучшением по всем наборам данных 11 %. Особенно существенное повышение качества наблюдалось для небольших и средних наборов с ярко выраженным дисбалансом классов: для них среднее улучшение составило 14 % против 8 % в наборах данных с менее выраженным дисбалансом.

Для методов уменьшения большего класса могла наблюдаться отрицательная динамика: результат варьировался от ухудшения на 41 % до улучшения на 54 %, при этом среднее значение полученных изменений по всем наборам данных близко к нулю.

Ансамблевые модели оказались менее чувствительны к легкому и умеренному дисбалансу в наборах данных: без применения балансировок на таких наборах данных результаты оказались на 3–24 % лучше, чем у классических моделей. При этом после балансировок сбалансированная точность улучшилась в среднем на 5 %. Однако на наборах данных с сильно выраженным дисбалансом ансамблевые модели показывали значительно худшие результаты, чем классические модели: ухудшение достигало 13–23 %, его среднее значение составило 5 %. После добавления балансировки ситуация значительно менялась: результаты становились лучше, чем на исходном наборе данных с классической моделью, на 13–18 %. Однако они все еще могли не превышать результатов, полученных при использовании балансировки и классической модели. Важно отметить и то, что для больших наборов данных с заметным дисбалансом итоговая сбалансированная точность оставалась достаточно низкой (около 0,5), что ограничивает возможность практического применения таких моделей. Также для наборов данных характерно то, что чем меньше степень дисбаланса, тем больше возможных сочетаний балансировки и модели можно подобрать для получения хороших результатов.

В процессе экспериментов для соотношения классов получены показатели сбалансированной точности по выбранным наборам данных, приведенные в табл. 2. Для всех случаев улучшение (значения выделены зеленым цветом) или ухудшение (значения выделены красным цветом) в процентах определялось по сравнению со сбалансированной точностью, полученной при применении классической модели, что отражено в табл. 2. Более подробно результаты описаны в [8].

Таблица 2. Результаты экспериментов для выбранных наборов данных
Table 2. Experimental results for selected datasets

Соотношение классов	Исходный набор		Сбалансированный набор		Лучшая модель и техника балансировки
	Классическая модель	Ансамбль	Классическая модель	Ансамбль	
8:100 и 13:100	0,873	0,920 +5,38 %	0,880 +0,80 %	0,934 +6,99 %	RF – Random Oversampling RF – SMOTE RF – B-SMOTE GB – Random Undersampling GB – OSS XGBoost – Random Oversampling LightGBM – TomekLinks LightGBM – OSS
2:100	0,650	0,502 –22,77 %	0,753 +15,85 %	0,779 +19,85 %	AdaBoost – Random Oversampling
22:100	0,839	0,864 +2,98 %	0,839 +0 %	0,878 +4,65 %	XGBoost – Random Oversampling
17:100 и 29:100	0,732	0,804 +9,84 %	0,770 +5,19 %	0,825 +12,70 %	RF – Random Oversampling RF – SMOTE RF – B-SMOTE RF – B-SMOTE SVM RF – ADASYN RF – Random Undersampling
2:100 и 15:100	0,453	0,391 –13,69 %	0,490 +8,17 %	0,468 +3,31 %	GaussianNB – SMOTE GaussianNB – B-SMOTE SVM GaussianNB – ADASYN

Комбинированное применение балансировок

Для данного этапа исследования формировались пары из всех алгоритмов увеличения меньшего класса и тех алгоритмов уменьшения большего класса, что показали наилучший результат при изолированном применении. Также проводилось исследование сочетаний методов балансировки, которые описываются в литературе[1] как наиболее универсальные вне зависимости от набора данных, а именно: SMOTE с Random Undersampling, SMOTE с TomekLinks и SMOTE с ENN. Результаты представлены на рис. 1, при этом для сравнения приведены и те значения, когда использовался изолированно метод увеличения большего класса – таким значениям соответствует прочерк вместо названия метода уменьшения большего класса.

По результатам экспериментов можно говорить о том, что использование упомянутых ранее универсальных комбинаций хоть и дает неплохие результаты, однако не гарантирует получение наилучшего качества модели. В целом прирост сбалансированной точности при комбинации алгоритмов из разных категорий оказался не очень высоким и составлял до 4 %, в среднем 1–2 %. Иногда он мог быть и отрицательным, когда выбирался алгоритм уменьшения большего класса, который не приводил к значительному улучшению качества модели при изолированном применении. Так, при выборе алгоритма уменьшения большего класса, который ухудшал качество модели на первом этапе, наблюдалась значительная отрицательная динамика (до –21 %, в среднем – (–2) %) при использовании этого алгоритма даже в комбинации с более удачными вариантами. Несмотря на то что в среднем улучшение оказалось не очень значительным, пиковые значения сбалансированной точности достигались именно при комбинации балансировок.

Определение оптимального соотношения классов

Для задачи мультиклассовой классификации в качестве начального соотношения классов после балансировки бралось наибольшее из доступных значений: к примеру, для соотношений 8:100 и 13:100 начальным выбиралось 13:100. В бинарной классификации этот параметр определяет-

[1] Brownlee, J. Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning / J. Brownlee // Machine Learning Mastery. 2020.

ся единственным возможным образом. До значения 100:100 соотношение увеличивалось с шагом 10, после чего шаг принимался равным 100, пока не будет получено соотношение 1000:100. График зависимости сбалансированной точности модели от соотношения классов и выбранного метода балансировки для наборов данных с бинарной классификацией представлен на рис. 2.

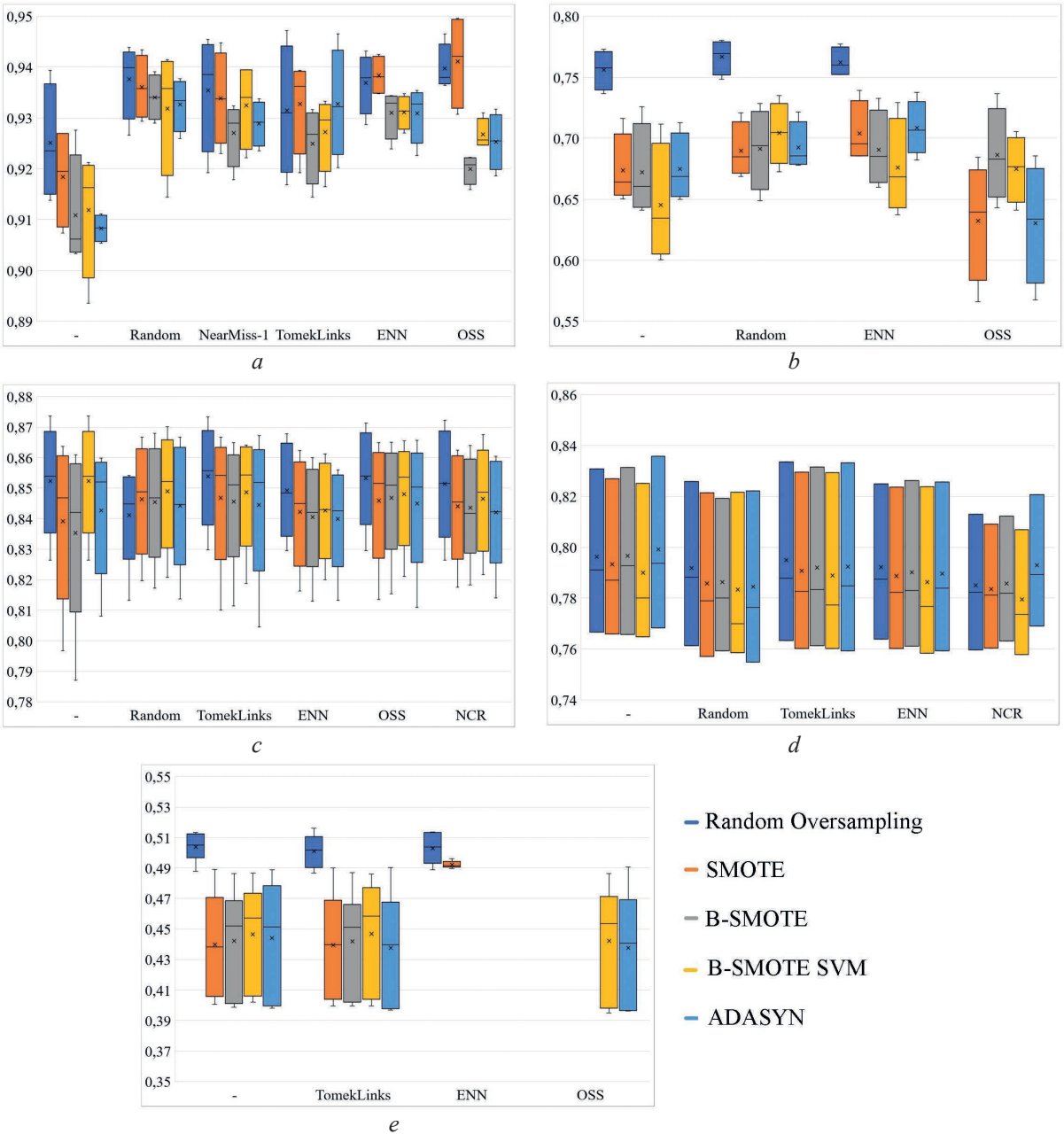
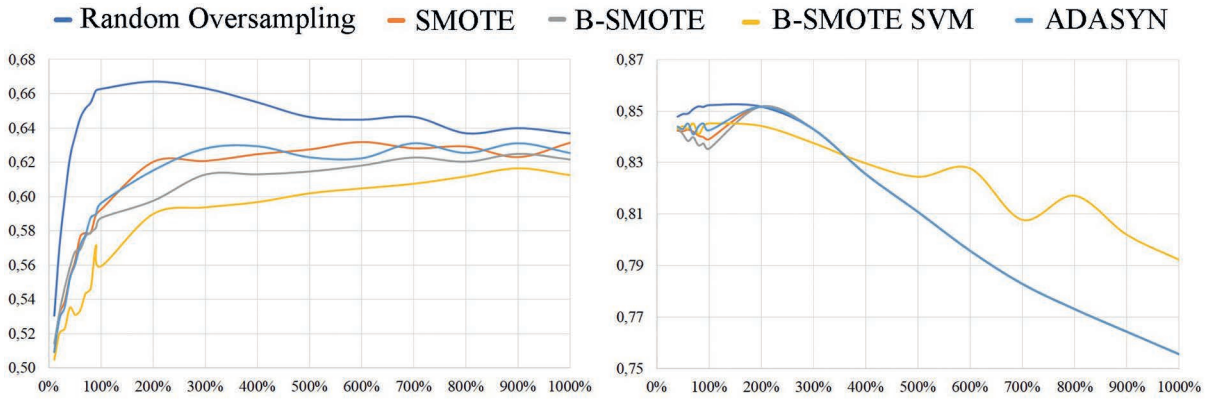


Рис. 1. Результаты комбинирования методов балансировки для набора данных:
a – «Классификация здоровья плода»; *b* – «Прогнозирование церебрального инсульта»;
c – «Кредитный риск»; *d* – «Классификация кредитного рейтинга»;
e – «Показатели здоровья при диабете»

Fig. 1. Results of combining balancing methods for a data set: *a* – “Fetal health”; *b* – “Cerebral stroke”;
c – “Credit risk”; *d* – “Credit rating”; *e* – “Diabetes health indicators”

Оптимальное соотношение классов как для бинарной, так и для мультиклассовой классификации располагается между 1:1 и 2:1 для тех методов балансировки, которые показали в целом наилучший результат. Соотношение для методов, показавших не такой хороший результат, может быть и больше (от 6:1 и более), однако такие результаты нельзя назвать показательными, если цель заключается в том, чтобы добиться глобально наилучшего качества модели на каждом шаге

ее построения. Соотношения менее 1:1 оказались неоптимальными для всех исследуемых наборов данных.



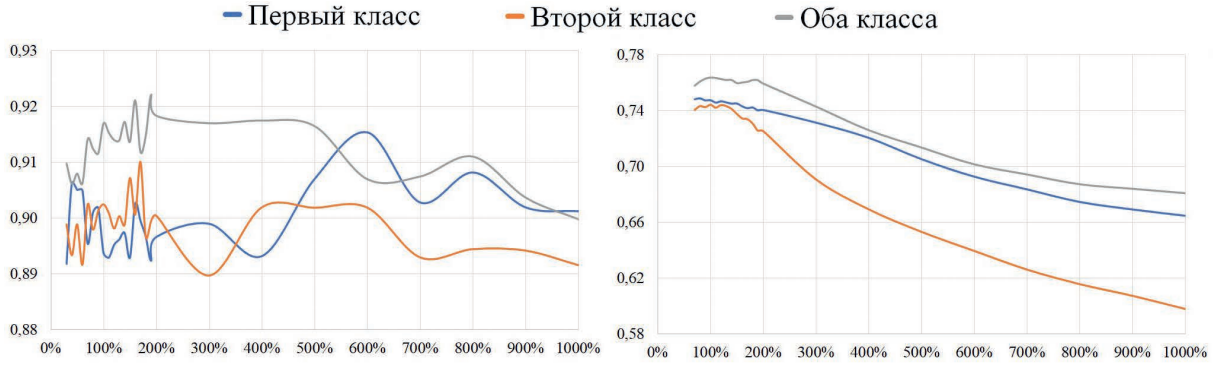
a *b*

Рис. 2. Зависимости сбалансированной точности модели от соотношения классов для наборов данных: *a* – «Прогнозирование церебрального инсульта»; *b* – «Кредитный риск»

Fig. 2. Dependences of the balanced accuracy of the model on the class ratio for data sets: *a* – "Cerebral stroke"; *b* – "Credit risk"

Следует отметить, что для наборов данных не удалось получить единственное универсальное соотношение классов – оно оказалось вариативным в зависимости от набора данных. При этом хоть и наблюдается прирост качества модели при подборе соотношения для балансировки, разница между сбалансированной точностью на пиковом значении и на соотношении 1:1 сравнительно мала и составляет в среднем 1–3 %. В связи с чем целесообразно подбирать соотношение только при очень высоких требованиях к качеству моделей. При этом нет и явной очевидной зависимости между выраженностью исходного дисбаланса классов в наборе и смещением оптимального соотношения правее, чем 1:1.

Дополнительно была изучена эффективность двух стратегий балансировки в задаче мультиклассовой классификации: изолированная балансировка одного из классов и одновременная балансировка обоих классов. Зависимости качества модели от соотношения классов для балансировки представлены на рис. 3. Для оценки использовались наборы данных с умеренным дисбалансом и с техникой балансировки Random Oversampling.



a *b*

Рис. 3. Зависимость качества модели от соотношения классов для балансировки для наборов данных: *a* – «Классификация здоровья плода»; *b* – «Классификация кредитного рейтинга»

Fig. 3. Model quality dependence on class ratio for balancing datasets: *a* – "Fetal health"; *b* – "Credit rating"

При балансировке одного класса наилучшее качество моделей наблюдалось в случае увеличения численности того из меньших классов, который изначально содержал больше объектов. Однако одновременная балансировка обоих классов обеспечивает значительно лучшие результаты по сравнению с балансировкой только одного из них. Поэтому можно рекомендовать стратегию балансировки сразу обоих классов, поскольку в данном случае качество моделей оказывается гарантированно не хуже, чем при балансировке только одного из них.

Также набор данных маленького размера реагировал на малое изменение соотношения балансировки большими скачками качества модели, а вот для набора данных большего размера графики оказались более сглаженными. Поэтому для небольших наборов данных уместным можно назвать меньший шаг при подборе соотношения классов.

Анализ влияния подбора гиперпараметров

Для анализа влияния подбора гиперпараметров использовалась байесовская оптимизация с 300 итерациями и применением алгоритма Tree-structured Parzen Estimator (TPE) фреймворка Optuna. Для всех наборов данных характерно то, что подбор гиперпараметров не смог скомпенсировать отсутствие балансировки данных – балансировка давала результат на 11–18 % лучше для наборов с выраженным дисбалансом и на 1–2 % лучше для наборов с легким дисбалансом по сравнению с подбором гиперпараметров. Однако качество моделей при комбинировании методов балансировки данных, ансамблевых алгоритмов и подбора гиперпараметров оказалось наилучшим.

Также можно сделать вывод о том, что чувствительность моделей к подбору гиперпараметров зависит от исходного уровня дисбаланса данных: чем сильнее выражен дисбаланс, тем более значимым оказывается эффект от настройки гиперпараметров, особенно для алгоритмов градиентного бустинга.

Алгоритмы балансировки данных, за исключением OSS, в целом оказались не очень чувствительны к подбору гиперпараметров, кроме как на наименьшем наборе данных. В среднем улучшение по всем наборам данных составило 1–4 %. Поэтому в условиях ограниченного времени можно рекомендовать подбор гиперпараметров для алгоритмов балансировки только в том случае, когда используется OSS или когда набор данных очень мал – особенно, если балансировка выполняется только одним методом.

Следует отметить также, что если модель изначально показывала худшие результаты по сравнению с другими, то даже после подбора гиперпараметров ее результаты не становились лучше, чем у тех моделей, которые изначально были сильнее и тоже проходили подбор гиперпараметров.

Вклад балансировок и настройки модели в итоговый результат

Для наглядности и визуальной оценки вклада каждого из этапов работы с набором данных была построена гистограмма, приведенная на рис. 4 и отражающая достигнутую сбалансированную точность для каждого набора. Номера наборов данных на рис. 4 приведены согласно порядку их представления в табл. 1. Видно, что применение каждой последующей техники (метод увеличения меньшего класса, ансамбль вместо классической модели, метод уменьшения большего класс и подбор гиперпараметров) позволяет увеличить сбалансированную точность в сравнении с классической моделью.

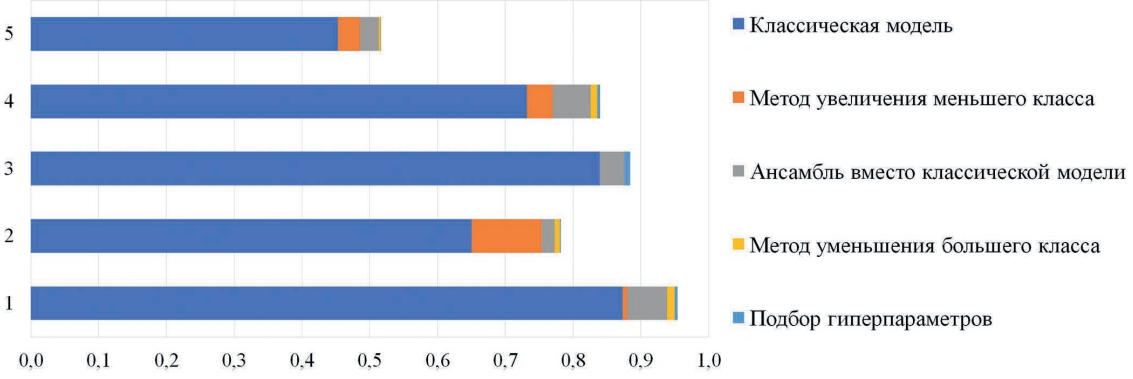


Рис. 4. Вклад этапов работы с набором данных в итоговый результат
Fig. 4. Contribution of the stages of working with a dataset to the final result

Наибольший вклад в итоговую модель давало применение одного метода балансировки. А учитывая, что пиковые значения достигались на методах увеличения меньшего класса, для решения практических задач может быть достаточно объединения одной балансировки с ансам-

блем. Вклад добавления еще одного метода балансировки (а именно – уменьшения большего класса) и подбора гиперпараметров оказался сравнительно низким, а на наборах данных большего размера еще и крайне затратным касательно временных и вычислительных ресурсов. Причем закономерности, что именно добавление второго метода или именно подбор гиперпараметров дает строго больший вклад, на рассмотренных данных нет. Поэтому нельзя рекомендовать ограничиться только одним из этих способов.

Заключение

1. Применение методов увеличения меньшего класса при использовании классических способов классификации улучшило качество моделей в среднем на 11 %, при этом наблюдалась неотрицательная динамика для наборов. Методы уменьшения большего класса дали нестабильные результаты: наблюдались как ухудшение до 41 %, так и улучшение до 54 %, среднее значение по всем наборам близко к нулю. Без балансировки ансамблевые методы показывали результаты в среднем на 5 % хуже классических, с балансировкой улучшение достигало 13–18 % по сравнению с классическими моделями на исходном наборе.

2. Более выраженный эффект дали методы увеличения меньшего класса, это характерно для всех рассмотренных наборов данных. Комбинирование методов увеличения меньшего класса и уменьшения большего класса целесообразно в тех случаях, когда есть временные и вычислительные ресурсы, и требуется как можно лучшее качество модели. Причем выбирать для комбинации следует те алгоритмы, которые изолированно дают наилучший прирост качества модели. В ином случае даже использование широко распространенных и известных сочетаний может обернуться ухудшением качества модели, если изолированно эти алгоритмы из сочетания не дают значимого положительного результата.

3. Оптимальное соотношение классов после балансировки для большинства рассмотренных наборов данных находится в диапазоне между 1:1 и 2:1, где первое число соотносится с количеством объектов изначально меньшего класса. Нецелесообразными оказались балансировка с соотношением менее 1:1 и балансировка только одного меньшего класса вместо всех в мультиклассовой классификации. При этом прирост качества моделей при подборе соотношения классов оказался не очень выраженным и составил 1–3 %. Поэтому в условиях ограниченных временных ресурсов можно рекомендовать балансировку до уровня 1:1 или 2:1 без более тщательного подбора точного соотношения.

4. Для всех рассмотренных наборов данных отсутствие балансировки не компенсировалось подбором гиперпараметров. Качество моделей при комбинировании метода увеличения меньшего класса, метода уменьшения большего класса, ансамблевого алгоритма и подбора гиперпараметров оказалось наилучшим. Наибольший вклад при этом внесли увеличение меньшего класса и использование ансамблевой модели вместо классической.

Список литературы

1. Classification of Imbalanced Data: Review of Methods and Applications / P. Kumar [et al.] // IOP Conference Series: Materials Science and Engineering. IOP Publishing. 2021. Vol. 1099, No 1.
2. Krawczyk, B. Learning from Imbalanced Data: Open Challenges and Future Directions / B. Krawczyk // Progress in Artificial Intelligence. 2016. Vol. 5, No 4. P. 221–232.
3. Branco, P. A Survey of Predictive Modeling on Imbalanced Domains / P. Branco, L. Torgo, R. Ribeiro // ACM Computing Surveys (CSUR). 2016. Vol. 49, No 2. P. 1–50.
4. Sun, Y. Classification of Imbalanced Data: A Review / Y. Sun, A. K. C. Wong, M. S. Kamel // International Journal of Pattern Recognition and Artificial Intelligence. 2009. Vol. 23, No 4. P. 687–719.
5. Kim, M. An Empirical Evaluation of Sampling Methods for the Classification of Imbalanced Data / M. Kim, K. B. Hwang // PLoS One. 2022. Vol. 17, No 7.
6. Dube, L. Enhancing Classification Performance in Imbalanced Datasets: A Comparative Analysis of Machine Learning Models / L. Dube, T. Verster // Data Science in Finance and Economics. 2023. Vol. 3, No 4. P. 354–379.
7. Khan, A. A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced Problems: Combination, Implementation and Evaluation / A. Khan, O. Chaudhari, R. Chandra // Expert Systems with Applications. 2024. Vol. 244.

8. Клицунова, Е. Сравнительный анализ методов балансировки данных для задач машинного обучения / Е. Клицунова, М. М. Лукашевич // BIG DATA и анализ высокого уровня: сб. науч. ст. XI Междунар. науч.-практ. конф. Минск: Белор. гос. ун-т информ. и радиоэлек., 2025. С. 74–83.

Поступила 18.06.2025 Принята в печать 18.07.2025

References

1. Kumar P., Bhatnagar R., Gaur K., Bhatnagar A. (2021) Classification of Imbalanced Data: Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering. IOP Publishing.* 1099 (1).
2. Krawczyk B. (2016) Learning from Imbalanced Data: Open Challenges and Future Directions. *Progress in Artificial Intelligence.* 5 (4), 221–232.
3. Branco P., Torgo L., Ribeiro R. (2016) A Survey of Predictive Modeling on Imbalanced Domains. *ACM Computing Surveys (CSUR).* 49 (2), 1–50.
4. Sun Y., Wong A. K. C., Kamel M. S. (2009) Classification of Imbalanced Data: A Review. *International Journal of Pattern Recognition and Artificial Intelligence*. 23 (4), 687–719.
5. Kim M., Hwang K. B. (2022) An Empirical Evaluation of Sampling Methods for the Classification of Imbalanced Data. *PLoS One.* 17 (7).
6. Dube L., Verster T. (2023) Enhancing Classification Performance in Imbalanced Datasets: A Comparative Analysis of Machine Learning Models. *Data Science in Finance and Economics.* 3 (4), 354–379.
7. Khan A., Chaudhari O., Chandra R. (2024) A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced Problems: Combination, Implementation and Evaluation. *Expert Systems with Applications.* 244.
8. Klitsunova K., Lukashevich M. M. (2025) Comparative Analysis of Data Balancing Methods. *BIG DATA and Advanced Analytics, Collection of Scientific Articles of XI International Scientific and Practical Conference.* Minsk, Belarusian State University of Informatics and Radioelectronics. 74–83 (in Russian).

Received: 18 June 2025 Accepted: 18 July 2025

Вклад авторов

Лукашевич М. М. разработала концепцию исследования, провела постановку задачи, подбор ключевых источников и инструментальных методов исследования, выбрала метрики оценки качества моделей, подготовила текст рукописи.

Клицунова Е. осуществила подбор наборов данных, провела экспериментальное исследование, выполнила анализ результатов, подготовила текст рукописи.

Authors' contribution

Lukashevich M. M. developed the concept of the study, formulated the problem, selected key sources and instrumental research methods, chose metrics for assessing the quality of models, prepared the text of the paper.

Klitsunova K. selected datasets, conducted the experimental study, analyzed the results, prepared the text of the paper.

Сведения об авторах

Лукашевич М. М., канд. техн. наук, доц., доц. каф. информационных систем управления, Белорусский государственный университет

Клицунова Е., бакалавр информатики, Белорусский государственный университет

Адрес для корреспонденции

220030, Республика Беларусь,
Минск, просп. Независимости, 4
Белорусский государственный университет
Тел.: +375 29 709-06-08
E-mail: lukashevichmm@bsu.by
Лукашевич Марина Михайловна

Information about the authors

Lukashevich M. M., Cand. Sci. (Tech.), Associate Professor, Associate Professor at the Department of Information Management Systems, Belarusian State University

Klitsunova K., Bachelor of Computer Science, Belarusian State University

Address for correspondence

220030, Republic of Belarus,
Minsk, Nezavisimosti Ave., 4
Belarusian State University
Tel.: +375 29 709-06-08
E-mail: lukashevichmm@bsu.by
Lukashevich Marina Mikhailovna